THE UNIVERSITY *of* EDINBURGH

THE UNIVERSITY *of* EDINBURGH

Deep unsupervised machine learning in the presence of missing data

Vaidotas Šimkus

A thesis submitted in fulfilment of the requirements for the degree of
Philosophiae Doctor in Data Science and Artificial Intelligence

University of Edinburgh

2024

School of Informatics,
University of Edinburgh,
10 Crichton Street,
Edinburgh,
EH8 9AB

Declaration

I declare that this thesis was composed by me, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Vaidotas Šimkus
November 2024

Acknowledgements

I feel very fortunate to have been advised by my supervisor Michael Gutmann, whose exemplary mentorship has made my PhD journey truly rewarding. Under your guidance, I have learnt the invaluable lessons and skills that define a good researcher, including the importance of occasionally stepping back to move forward with greater clarity. Your patient mentorship was key in developing my analytical abilities and research acumen.

I extend my gratitude to Chris Williams and Arno Onken, who served on my PhD committee, for their thoughtful feedback and invaluable guidance during our annual review meetings. I would also like express my thanks to the entire community within Informatics, especially Simon, Titas, Samuel, Ben, Steven, Michał, and Sandy. You have not only broadened my academic perspectives, but also inspired me to be more ambitious.

The PhD journey during a global pandemic would not have been nearly as enjoyable without my wonderful friends. Mantas, Ugnė, Simon, Sofija, Ivo, Titas—your stimulating and impassioned discussions recharged my mind and spirit, providing the much-needed balance to the PhD life. Edita, Indrė, Tomas, Giedrė, Emilis—thank you for the annual hikes and for finding time for our re-unions. Your friendship filled my journey with cherished moments that I will forever treasure.

I will also be forever grateful to my family and my inspiring grandparents for the steadfast support throughout my endeavours. My deepest gratitude goes to my beloved mother, Skirmutė, who sadly did not get to see this journey unfold. Yet her lasting legacy of kindness, curiosity, and belief in me was the wind at my back, propelling me forward.

Finally, I would like to thank my partner, Ana, for her gentle support throughout this endeavour. Your unconditional love and kindness provided me a sanctuary that relieved the stresses and frustrations inherent to such an endeavour. Thank you, Ana, for being my pillar of strength and for gently reminding me to embrace life's joys even when the PhD journey became overwhelming. This achievement would have been much more daunting without you by my side.

Abstract

Advances in deep statistical models have re-shaped modern data-driven applications, demonstrating remarkable empirical success across diverse domains. However, while some domains benefit from an abundance of clean and fully-observed data, enabling the practitioners to reap the full-benefits of deep models, other domains often grapple with incomplete data, hindering the effective application of these powerful models. In this thesis, we aim to investigate and address important challenges caused by missing data that hinder the use of deep models, focusing on two key statistical tasks: parameter estimation from incomplete training data sets and missing data imputation.

Firstly, we explore the problem of missing data imputation using pre-trained models, focusing on deep statistical models in the class of variational autoencoders (VAEs). Our exploration reveals limitations of existing methods for conditional sampling of VAEs, identifying pitfalls, related to commonly desired properties of learnt VAEs, that hinder the methods' performance in certain scenarios. To mitigate the pitfalls, we propose two novel methods based on Markov chain Monte Carlo and importance sampling. Our evaluation shows that the proposed methods improve missing data imputation using pre-trained VAEs across diverse data sets.

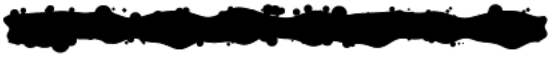
Subsequently, we shift our attention to the estimation of VAEs from incomplete training data sets. While this area has received substantial attention in the literature, we report a previously unknown phenomenon caused by missing data that hinders the effective fitting of VAEs. To overcome the adverse effects and improve VAE estimation from incomplete data, we introduce two strategies based on variational mixture distributions that trade-off computational efficiency, model accuracy, and learnt latent structure. We demonstrate that the proposed approaches improve VAE estimation from incomplete data compared to existing approaches that do not use variational mixtures.

Expanding our focus to the broader challenges of estimating general statistical models, we observe an uneven progress across different classes of deep models. To advance the adoption of all deep statistical models, we introduce variational Gibbs inference (VGI), a general-purpose method for maximum-likelihood-based estimation of general statistical models with tractable likelihood functions. We show that the method is capable of accurate model estimation from incomplete data, including VAEs and normalising flows. Importantly, VGI is one of the few probabilistically-principled methods in the current literature for normalising flow estimation from incomplete data, achieving state-of-the-

art performance. By providing a unified framework for handling missing data in model estimation, VGI paves the way for leveraging the full potential of deep statistical models across diverse domains grappling with missing data.

Lay Summary

Imagine collecting a huge stack of documents with the intention of doing research on it. You leave the documents in your office overnight, intending to start your research in the morning. However, when you return, you find that a clumsy colleague had spilled black ink all over the pages, making parts of the text completely illegible. The ink-stained areas can be considered *"missing data"*—we know there was originally meaningful information there, but we can no longer read it. You may be tempted to discard all the documents and start over from scratch. But that would waste all the hard work you put into collecting them in the first place. So what can be done?

One approach is to try filling in the missing portions of the pages with plausible completions based on the readable parts surrounding them. For example, if a sentence read "In this zoo we saw  a single reef", you could fill in the blank with something like "various aquatic sea animals cohabiting" or "fish, sharks, and dolphins happily sharing", as those would be plausible completions that make sense given the context of a zoo and a reef. However, filling it with "cats, dogs, monkeys, and fish living in" would likely be an implausible completion. Manually filling in all the missing data in this way would be extremely laborious. Hence, as a first step in this thesis, we investigate machine learning techniques to automatically generate plausible imputations or completions for the missing data. These techniques use the mathematical language of probability along with powerful machine learning models capable of telling what is plausible or sensible given the context, and what is not.

Next, we focus on learning this "plausibility model" from the stack of ink-stained documents itself. However, these models are usually created for use with clean and complete data. So how can we learn them from data with missing parts? Depending on the model, different approaches may work. We first explore a special type of models that are capable of simply "ignoring" the missing parts and still be able to learn something useful from the remaining readable portions, like that there might be an (artificial) reef in a zoo. We then investigate more complicated models that cannot "ignore" the missing data. Instead, iterative procedures are needed to learn these models that aim to first fill in the missing data using the current model, then re-train the model on this completed data, and repeat this process. The key idea is to leverage the models' understanding of plausibility to fill in the blanks, while simultaneously learning that understanding from the data itself, through an iterative process.

Finally, we note that the aforementioned tasks have a long history of dedicated traditional methods. However, applying those traditional approaches to modern extremely flexible machine learning models that use artificial neural networks presents new challenges. The core aim of this thesis is to better understand these challenges that arise when working with modern models and incomplete data, and then mitigate them by developing novel methods for these tasks.

Contents

Chapter 1

Introduction

Data-driven technologies hold a promise to transform many aspects of our world from everyday technology to science, owing their potential to rapid advances in unsupervised machine learning. A key element of such machine learning techniques often involves deep statistical models. These models are either used in a pre-training step of the machine learning pipeline or serve as an end goal in their own right. Their remarkable capability to represent complex and inherently random real-world phenomena can be viewed as a proxy for the "understanding" of the world—an invaluable asset when developing such data-driven systems that aim to act in the real world. However, the models are typically formulated in terms of fully-observed data and often require large amounts of such data during training, therefore limiting their practical applicability to domains where data is abundant and clean.

While domains where data is fully-observed are relatively scarce, scenarios involving incomplete data are abundant and sources of missingness are diverse. For example, data may be partially missing due to participants withdrawing from long-term studies, data being pooled from different sources with varying data schemas, or the use of faulty or unreliable measurement devices causing invalid measurements. Consequently, these scenarios result in data sets containing "gaps" with unknown values, commonly referred to as *missing data*. Missing data introduces challenges when applying standard machine learning techniques that assume complete observations during training and inference.

These challenges need to be addressed to realise the full potential of the information contained in the incomplete data sets, provided they contain sufficient information to start with as is typical in machine learning scenarios. For example, if the values in the data set are missing infrequently, it may be tempting to simply discard data-points that contain missing values, and work with the remaining fully-observed data. However, this simplistic strategy can lead to a significant loss of information. As illustrated in Figure 1.1, even a seemingly low missingness probability of 0.01 can result in the removal of more than half of the original data-points in a 100-dimensional case, which gets exponentially worse as the dimensionality and missingness probability grow. An alternative strategy may involve filling-in (or imputing) the missing values to treat the data as if it were fully-observed, provided that the missing data distribution is identifiable (more in section 2.2.3). But,

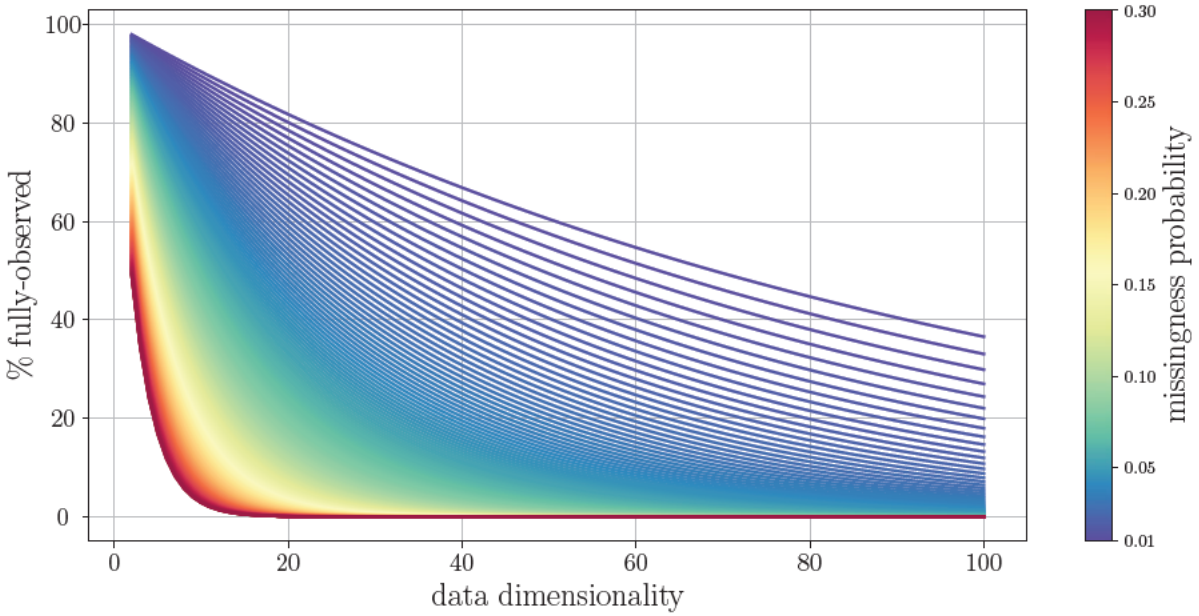


Figure 1.1: Percentage of fully-observed data-points as a function of data dimensionality, for various missingness probabilities from 0.01 to 0.3, when the missingness is uniformly random, corresponding to MCAR missingness (see section 2.2.3).

caution must be exercised in order to avoid introducing statistical biases into the machine learning pipeline.

Potential biases arising from missing data have been a core research subject over the last 50 years in the field of statistics (*e.g.* Little and Rubin, 2020, Chapters 2–5), leading to principled procedures for handling incomplete data sets (Dempster et al., 1977; Rubin, 1987). These studies primarily focused on traditional statistical models where the necessary computations could often, at least partially, be framed in a well-understood form that allows efficient calculation of the exact solution, or in other words, the solution can be computed with pen-and-paper. Modern (deep) statistical models, in contrast, do not generally allow such efficient computation. Consequently, applying traditional procedures for incomplete data to deep models is not without novel challenges, which warrants further investigation.

In this thesis, we aim to better *understand the core challenges* affecting deep models in the presence of incomplete data sets and in doing so develop methods that *mitigate these challenges* for two important statistical tasks: missing data imputation and model estimation. By addressing these goals, we hope to take a step towards increasing the adoption of deep models in diverse domains where data is often affected by missingness.

1.1 Contents of this thesis

In chapter 2, we provide the key background on deep statistical models and missing data. The chapter covers these two topics individually and mostly discusses textbook-style material. Hence, readers familiar with these topics may skim or skip chapter 2. In chapter 3, we explore the intersection of these two topics, discussing the relevant recent literature and establishing the research questions for this thesis. The main chapters of this thesis each contain the following publications:

- Chapter 4 contains the following published journal paper:

> Vaidotas Simkus and Michael U. Gutmann. Conditional Sampling of Variational Autoencoders via Iterated Approximate Ancestral Sampling. *Transactions on Machine Learning Research*, 2023

- Chapter 5 contains the following published journal paper:[1]

> Vaidotas Simkus and Michael U. Gutmann. Improving Variational Autoencoder Estimation from Incomplete Data with Mixture Variational Families. *Transactions on Machine Learning Research*, 2024

- Chapter 6 contains the following published journal paper:[2]

> Vaidotas Simkus, Benjamin Rhodes, and Michael U. Gutmann. Variational Gibbs Inference for Statistical Model Estimation from Incomplete Data. *Journal of Machine Learning Research*, 24(196):1–72, 2023

Finally, we conclude this thesis and discuss future outlook in chapter 7.

Attributions

As the main author of the three papers presented in this thesis, I was responsible for designing the research, conducting the experiments, analysing the results, and writing the papers. My advisor, Michael Gutmann, guided me through the research process by assisting in the design of the experiments, analysing the results, providing methodological insights, and giving detailed feedback on the paper manuscripts. Ben Rhodes, the second author of the paper in chapter 6, contributed to the initial design of the proposed method and provided feedback on the final version of the manuscript.

[1] An earlier version of this paper was accepted to the Data-centric Machine Learning Research (DMLR) workshop at ICLR 2024.

[2] This paper was also presented as a poster at NeurIPS 2023 via the journal-to-conference track.

Chapter 2

Background

This chapter lays down the foundational concepts for this thesis, including prominent deep statistical models (section 2.1) and key statistical procedures with missing data (section 2.2). The two areas are covered individually, focusing largely on topics found in standard textbooks. Readers who are familiar with these foundational topics may proceed to the next chapter, which details recent research in the intersection of deep statistical models and missing data (chapter 3).

2.1 Deep statistical models

The overarching objective in unsupervised machine learning is to learn about the complex and inherently random phenomena in the real- and digital-worlds. By doing so, we seek to understand how these phenomena work, create synthetic systems that emulate the natural behaviour, and automate decision making in the presence of uncertainty. The foundation of these unsupervised machine learning methods often lies in statistical models.

A parametric statistical model is a set of probability distributions $\{p_{\boldsymbol{\theta}}(\boldsymbol{x})\}_{\boldsymbol{\theta}\in\Omega_{\boldsymbol{\theta}}}$, where $\boldsymbol{x} \in \mathcal{X}^D$ is a D-dimensional a random variable, $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ is a probability distribution with parameters $\boldsymbol{\theta}$, and $\Omega_{\boldsymbol{\theta}}$ is the parameter space. Typically, users specify the statistical model based on prior assumptions and domain knowledge about the random phenomenon in question. The parameters $\boldsymbol{\theta}$ are then estimated using a data set comprised of observations from the target phenomenon. This process yields a probabilistic model $p_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x})$ that represents the unknown random phenomenon, which can then be used for various tasks, such as simulation, decision making, and missing data imputation (*e.g.* Goodfellow et al., 2016, Section 5.1.1; Murphy, 2023, Section 20.3).

A deep statistical model, also called a deep generative model, is a parametric statistical model that uses neural networks to parametrise its components. The use of neural networks endows these models with remarkable representational capabilities. Over the past decade, owing to their representational capacity, deep models have excelled at modelling various natural phenomena, with applications spanning natural language, images, video, and scientific domains, among others. In the following sections, we will review two prominent classes of deep models—variational autoencoders and normalising flows—that are

particularly relevant to this thesis.

2.1.1 Variational autoencoders

Variational autoencoders (VAEs, Kingma and Welling, 2013; Rezende et al., 2014) represent a large family of deep statistical models that aim to model the target phenomenon using a well-structured latent space. A VAE model is typically specified via a decoder distribution $p_{\boldsymbol{\theta}}(\boldsymbol{x} \mid \boldsymbol{z})$, and a prior distribution $p_{\boldsymbol{\theta}}(\boldsymbol{z})$ over the latent variables $\boldsymbol{z} \in \mathcal{Z}^L$:

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \int p_{\boldsymbol{\theta}}(\boldsymbol{x} \mid \boldsymbol{z}) p_{\boldsymbol{\theta}}(\boldsymbol{z}) \, \mathrm{d}\boldsymbol{z}. \tag{2.1}$$

Both the decoder and prior distributions are generally prescribed to be in a parametric family of distributions with a known likelihood function, for example, Gaussian or Bernoulli. Then, to parametrise $p_{\boldsymbol{\theta}}(\boldsymbol{x} \mid \boldsymbol{z})$, a neural network is used, taking $\boldsymbol{z}$ as input and predicting the parameters of the distribution over $\boldsymbol{x}$, such as the mean and the covariance matrix of a Gaussian. Moreover, to simplify model specification and improve learning efficiency, the prior distribution $p_{\boldsymbol{\theta}}(\boldsymbol{z})$ is often fixed to a standard Gaussian distribution, and the decoder distribution often factorises as $p_{\boldsymbol{\theta}}(\boldsymbol{x} \mid \boldsymbol{z}) = \prod_{j=1}^{D} p_{\boldsymbol{\theta}}(x_j \mid \boldsymbol{z})$.

Since the integral in eq. (2.1) defining the marginal likelihood is generally intractable, a VAE model is typically estimated using the variational evidence lower-bound (ELBO):

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}) \geq \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x} \mid \boldsymbol{z}) p_{\boldsymbol{\theta}}(\boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x})} \right], \tag{2.2}$$

or the importance-weighted ELBO (IWELBO, Burda et al., 2015):

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}) \geq \mathbb{E}_{\{\boldsymbol{z}_l\}_{l=1}^{I} \sim q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})} \left[\log \frac{1}{I} \sum_{l=1}^{I} \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x} \mid \boldsymbol{z}_l) p_{\boldsymbol{\theta}}(\boldsymbol{z}_l)}{q_{\boldsymbol{\phi}}(\boldsymbol{z}_l \mid \boldsymbol{x})} \right], \tag{2.3}$$

where $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x})$ is a variational distribution in a parametric family with a known likelihood function that is specified by the user, whose parameters $\boldsymbol{\phi}$ are fitted jointly with $\boldsymbol{\theta}$ by maximising the objective.

The objectives encourage the variational distribution $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x})$ to approximate the intractable model posterior $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x})$, in a process known as variational inference (Jordan et al., 1999).[3] This allows for efficient estimation of model parameters $\boldsymbol{\theta}$ without directly evaluating the intractable marginal likelihood in eq. (2.1), as the objectives are equal to the marginal log-likelihood $\log p_{\boldsymbol{\theta}}(\boldsymbol{x})$ when the variational distribution $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x})$ matches

[3]Note that when using the IWELBO, $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x})$ is not directly matched to $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x})$, but is instead encouraged to produce a good importance sampling proposal distribution for approximating $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x})$ via importance sampling (Cremer et al., 2017; Rainforth et al., 2019).

the model posterior $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x})$ exactly for all data-points $\boldsymbol{x}$. To improve the efficiency of the inference process, amortised variational inference (Gershman and Goodman, 2014) is typically used. Here, the variational distribution $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x})$ is parametrised using a neural network, called an encoder. This enables "sharing" (or amortising) the conditional distribution across all data-points $\boldsymbol{x}$ in the data set, rather than computing it separately for each $\boldsymbol{x}$.

Since the above seminal papers, VAEs have been recognised as flexible statistical models, with a particular capability to learn compact and useful representations of data across many domains (Higgins et al., 2017; Gómez-Bombarelli et al., 2018; Miao et al., 2022). Typical representations of a data-point $\boldsymbol{x}$ are samples drawn from the model posterior $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x})$ or the approximate posterior $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x})$, or point-estimates such as $\arg\max_{\boldsymbol{z}} q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x})$. On the other hand, despite their representational capabilities, the applications of VAEs can sometimes be limited due to the intractable marginal likelihood in eq. (2.1).

2.1.2 Normalising flows

Normalising flows (Rezende and Mohamed, 2015; Papamakarios et al., 2021) represent a family of deep models that model the target phenomenon with a simple base distribution and a learnable non-linear transformation of random variables, enabling efficient evaluation of the density. Formally, flows represent a distribution $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ via a simple base distribution $p_{\boldsymbol{u}}(\boldsymbol{u})$ of a random variable $\boldsymbol{u} \in \mathcal{X}^D$ and a transformation $T_{\boldsymbol{\theta}} : \mathcal{X}^D \mapsto \mathcal{X}^D$ that maps $\boldsymbol{u}$ to $\boldsymbol{x}$:

$$\boldsymbol{x} = T_{\boldsymbol{\theta}}(\boldsymbol{u}), \quad \text{with} \quad \boldsymbol{u} \sim p_{\boldsymbol{u}}(\boldsymbol{u}). \tag{2.4}$$

The transformation $T_{\boldsymbol{\theta}}$ must be differentiable and invertible, which ensures that the density of $\boldsymbol{x}$ is well-defined, and can be computed via the change of variables formula (*e.g.* Murphy, 2021, Section 2.8.3; Papamakarios et al., 2021):

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = p_{\boldsymbol{u}}(T_{\boldsymbol{\theta}}^{-1}(\boldsymbol{x})) \left| \det J_{T_{\boldsymbol{\theta}}^{-1}}(\boldsymbol{x}) \right|, \tag{2.5}$$

where $J_{T_{\boldsymbol{\theta}}^{-1}}(\boldsymbol{x})$ denotes the Jacobian of the inverse transformation $T_{\boldsymbol{\theta}}^{-1}(\boldsymbol{x})$.

In practice, the base distribution $p_{\boldsymbol{u}}(\boldsymbol{u})$ is typically specified to be a standard Gaussian and the transformation $T_{\boldsymbol{\theta}}$ is a composition of many simple transformations $T_{\boldsymbol{\theta}} = T_{\boldsymbol{\theta},1} \circ \ldots \circ T_{\boldsymbol{\theta},L}$, such as affine (Rezende and Mohamed, 2015; Dinh et al., 2015) or spline (Durkan et al., 2019) functions, that are parametrised using neural networks.

Normalising flow models offer extreme flexibility and enable both tractable sampling and

evaluation of the density $p_{\boldsymbol{\theta}}(\boldsymbol{x})$. These characteristics have led to a wide adoption of normalising flows across various scientific domains (Satorras et al., 2021; Williams et al., 2021; Dai and Seljak, 2022).

2.1.3 Other deep models

In the previous two sections we discussed two prominent families of deep statistical models: variational autoencoders and normalising flows. These two models are the central to the thesis. However, the landscape of deep models is vast and encompasses a broader range of techniques. A comprehensive review of all these models is beyond the scope of this thesis. For those keen on learning about additional deep models, we refer to Murphy (2023, Part IV) for a detailed review.

2.2 Missing data

Missing data is a fundamental challenge in statistical data analysis, with significant implications for the applicability and validity of standard statistical procedures. In this context, Rubin (1976) was the first to formally outline various assumptions about missing data and establish a statistically valid framework for its analysis, marking the beginning of an extensive field of research over the last 50 years.

In this section, we briefly review the necessary background on missing data since understanding this fundamental challenge is essential for implementing appropriate statistical procedures on data sets with missing values. We begin by defining the notation to describe incomplete data (section 2.2.1). We then discuss the main assumptions about missing data (section 2.2.3) and their implications for statistical tasks (section 2.2.4). Finally, we discuss key statistical tasks with missing data, including missing data imputation (section 2.2.5) and model estimation (section 2.2.6).

2.2.1 Notation for missing data

Let $\boldsymbol{x} \in \mathcal{X}^D$ denote the D-dimensional random variable of interest. Additionally, we define a random variable $\boldsymbol{m} \in \{0, 1\}^D$, referred to as the missingness pattern or mask, which can take on 2^D possible configurations. We denote the j-th dimension of $\boldsymbol{x}$ and $\boldsymbol{m}$ as x_j and m_j, respectively.

We can then view the generation of incomplete data as a two-step generative process. First, we generate $\boldsymbol{x}$ from a complete-data distribution $p^*(\boldsymbol{x})$ and then generate the missingness pattern $\boldsymbol{m}$ from the missingness process $p^*(\boldsymbol{m} \mid \boldsymbol{x})$, also known as the missingness mechanism. This pattern $\boldsymbol{m}$ determines which variables in $\boldsymbol{x}$ are observed and which are

missing.

We define $\boldsymbol{x}_{\text{obs}}$ to be the observed elements of $\boldsymbol{x}$, and $\boldsymbol{x}_{\text{mis}}$ to be the missing elements of $\boldsymbol{x}$, as determined by the mask $\boldsymbol{m}$. Formally, we define $\boldsymbol{x}_{\text{obs}}$ and $\boldsymbol{x}_{\text{mis}}$ as

$$\boldsymbol{x}_{\text{obs}} \stackrel{\text{def}}{=} \boldsymbol{x}[\boldsymbol{m}], \quad \text{and} \quad \boldsymbol{x}_{\text{mis}} \stackrel{\text{def}}{=} \boldsymbol{x}[1 - \boldsymbol{m}], \tag{2.6}$$

where $\boldsymbol{x}[\boldsymbol{m}]$ denotes the indexing of $\boldsymbol{x}$ with a mask $\boldsymbol{m}$, for example, if $\boldsymbol{x} = (1, 2, 3, 4)$ and $\boldsymbol{m} = (0, 1, 1, 0)$ then $\boldsymbol{x}[\boldsymbol{m}] = (2, 3)$. Note that the dimensionality of $\boldsymbol{x}_{\text{obs}}$ and $\boldsymbol{x}_{\text{mis}}$ depends on $\boldsymbol{m}$, and that the sum of their dimensionalities must equal to D. This results in 2^D possible configurations of the observed $\boldsymbol{x}_{\text{obs}}$ and missing $\boldsymbol{x}_{\text{mis}}$ variables.

2.2.2 Modelling incomplete data

In the general case, modelling incomplete data requires capturing both $\boldsymbol{x}$ and $\boldsymbol{m}$. In the literature, the two main families of models for incomplete data are selection models and pattern-mixture models.

Selection models, as introduced by Heckman (1976), specify the joint distribution over $\boldsymbol{x}$ and $\boldsymbol{m}$ as follows:

$$p_{\boldsymbol{\theta},\gamma}(\boldsymbol{x}, \boldsymbol{m}) = p_{\boldsymbol{\theta}}(\boldsymbol{x}) p_{\gamma}(\boldsymbol{m} \mid \boldsymbol{x}). \tag{2.7}$$

Pattern-mixture models, on the other hand, as proposed by Glynn et al. (1986), specify the joint as:

$$p_{\boldsymbol{\nu},\boldsymbol{\lambda}}(\boldsymbol{x}, \boldsymbol{m}) = p_{\boldsymbol{\lambda}}(\boldsymbol{x} \mid \boldsymbol{m}) p_{\boldsymbol{\nu}}(\boldsymbol{m}). \tag{2.8}$$

The two models are formally equivalent via the probability chain rule on the joint $p(\boldsymbol{x}, \boldsymbol{m})$. However, as argued by Fitzmaurice et al. (2008, Section 18.3), the selection model is a more natural approach to modelling the distribution of interest $p^*(\boldsymbol{x})$. Moreover, under certain assumptions outlined in the following section, it can substantially simplify model estimation and imputation (see section 2.2.4). We will therefore follow the selection modelling approach in this thesis.

2.2.3 Taxonomy of assumptions about missing data

Making certain assumptions about the missing data can greatly simplify the statistical procedures for handling it. But, discrepancies and ambiguity in missingness assumptions within the statistics literature have previously led to disagreement and confusion (Seaman et al., 2013). To avoid such ambiguity, we adopt the clarified definitions for describing

assumptions about missing data, as proposed by Seaman et al. (2013) in their work aimed at resolving this lack of consensus.

To define the missingness assumptions, we introduce additional notation for this subsection. We typically work, in the machine learning field, with data sets containing N i.i.d. samples. Thus, we generalise the notation from section 2.2.1 to accommodate this. Let $\boldsymbol{X}$ be a set of N random variables $\{\boldsymbol{x}^1, \dots, \boldsymbol{x}^N\}$, and let $\boldsymbol{M}$ be a set of N mask variables $\{\boldsymbol{m}^1, \dots, \boldsymbol{m}^N\}$. Moreover, we denote $\widetilde{\boldsymbol{X}}$ to be the realised values of the random variable $\boldsymbol{X}$ and $\widetilde{\boldsymbol{M}}$ to be the realised values of $\boldsymbol{M}$.

With the generalised notation and assuming realised values $\widetilde{\boldsymbol{M}}$ and $\widetilde{\boldsymbol{X}}[\widetilde{\boldsymbol{M}}]$, we now define the three common missingness assumptions:

Definition 1 *The data are missing completely at random (MCAR) if*

$$p_{\boldsymbol{\gamma}}(\boldsymbol{M} = \widetilde{\boldsymbol{M}} \mid \boldsymbol{X} = \widetilde{\boldsymbol{X}}') = p_{\boldsymbol{\gamma}}(\boldsymbol{M} = \widetilde{\boldsymbol{M}} \mid \boldsymbol{X} = \widetilde{\boldsymbol{X}}'') \quad \forall \boldsymbol{\gamma}, \widetilde{\boldsymbol{X}}', \widetilde{\boldsymbol{X}}'', \tag{2.9}$$

where $\widetilde{\boldsymbol{X}}'$ and $\widetilde{\boldsymbol{X}}''$ are any possible values of $\boldsymbol{X}$.

The MCAR assumption implies that the missingness probability is independent of the data. For brevity, we may sometimes use the notation $p_{\boldsymbol{\gamma}}(\boldsymbol{m} \mid \boldsymbol{x}) = p_{\boldsymbol{\gamma}}(\boldsymbol{m})$ to denote the MCAR assumption.

Definition 2 *The data are missing at random (MAR) if*

$$\begin{aligned} &p_{\boldsymbol{\gamma}}(\boldsymbol{M} = \widetilde{\boldsymbol{M}} \mid \boldsymbol{X} = \widetilde{\boldsymbol{X}}') = p_{\boldsymbol{\gamma}}(\boldsymbol{M} = \widetilde{\boldsymbol{M}} \mid \boldsymbol{X} = \widetilde{\boldsymbol{X}}'') \\ &\forall \boldsymbol{\gamma}, \widetilde{\boldsymbol{X}}', \widetilde{\boldsymbol{X}}'' \text{ such that } \widetilde{\boldsymbol{X}}'[\widetilde{\boldsymbol{M}}] = \widetilde{\boldsymbol{X}}''[\widetilde{\boldsymbol{M}}] = \widetilde{\boldsymbol{X}}[\widetilde{\boldsymbol{M}}], \end{aligned} \tag{2.10}$$

where $\widetilde{\boldsymbol{X}}'$ and $\widetilde{\boldsymbol{X}}''$ are possible values of $\boldsymbol{X}$.

The MAR assumption implies that the missingness probability does not vary with different values of the missing data. For brevity, similar to the previous case we may use the informal notation $p_{\boldsymbol{\gamma}}(\boldsymbol{m} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}) = p_{\boldsymbol{\gamma}}(\boldsymbol{m} \mid \boldsymbol{x}_{\text{obs}})$ to denote the MAR assumption.[4]

Definition 3 *The data are missing not at random (MNAR) if*

$$\exists \boldsymbol{\gamma}, \widetilde{\boldsymbol{X}}', \widetilde{\boldsymbol{X}}'' \text{ such that } \widetilde{\boldsymbol{X}}'[\widetilde{\boldsymbol{M}}] = \widetilde{\boldsymbol{X}}''[\widetilde{\boldsymbol{M}}] = \widetilde{\boldsymbol{X}}[\widetilde{\boldsymbol{M}}] \text{ and } \widetilde{\boldsymbol{X}}'[1 - \widetilde{\boldsymbol{M}}] \neq \widetilde{\boldsymbol{X}}''[1 - \widetilde{\boldsymbol{M}}],$$

[4]Note that the split of a variable $\boldsymbol{x}$ into the observed $\boldsymbol{x}_{\text{obs}}$ and $\boldsymbol{x}_{\text{mis}}$ missing variables, as defined in eq. (2.6), uses the missingness mask $\boldsymbol{m}$. This notation does not imply that the missing dimensions are known from $\boldsymbol{x}_{\text{obs}}$ alone—this information is provided by $\boldsymbol{m}$. As such, the distribution $p_{\boldsymbol{\gamma}}(\boldsymbol{m} \mid \boldsymbol{x}_{\text{obs}})$ in the informal notation for MAR missingness is not degenerate and follows the definition 2. This informal notation is common in the literature as it substantially simplifies definitions and derivations.

$$\textit{and } p_\gamma(\boldsymbol{M} = \widetilde{\boldsymbol{M}} \mid \boldsymbol{X} = \widetilde{\boldsymbol{X}}') \neq p_\gamma(\boldsymbol{M} = \widetilde{\boldsymbol{M}} \mid \boldsymbol{X} = \widetilde{\boldsymbol{X}}''), \tag{2.11}$$

where $\widetilde{\boldsymbol{X}}'$ *and* $\widetilde{\boldsymbol{X}}''$ *are possible values of* $\boldsymbol{X}$.

The MNAR assumption implies that the missingness probability may depend on the missing data, and in general means that no simplifying assumptions are made about the missingness process. While MNAR is the least restrictive assumption, it presents additional challenges, such as a general lack of identifiability, which is beyond the scope of this thesis and we refer the curious readers to Mohan et al. (2013), Nabi et al. (2020), and Ma and Zhang (2021) for further details.

We note that the above definitions are the weakest versions of MCAR, MAR, and MNAR. This is because the conditions in the definitions must only hold for the realised values $\widetilde{\boldsymbol{M}}$ and $\widetilde{\boldsymbol{X}}[\widetilde{\boldsymbol{M}}]$ and not for all possible values of $\boldsymbol{M}$ and $\boldsymbol{X}$. Consequently, the above definitions are sometimes called the *realised* MCAR/MAR/MNAR (Seaman et al., 2013).

Stronger conditions also exist that require definitions 1 to 3 additionally hold for all possible values of $\boldsymbol{M}$ and $\boldsymbol{X}[\boldsymbol{M}]$, which subsumes the realised case. These stronger conditions are known as the *everywhere* MCAR/MAR/MNAR (Seaman et al., 2013).

In this thesis, we will adopt the *realised* MAR assumption in definition 2. The conditions of realised MAR must only hold for the missingness patterns observed in the data set, which is bounded by the number of data-points N, whereas the everywhere case would require holding for all 2^D patterns, which can be much larger than the typical N for even moderate D. Additionally, the MAR assumption is commonly used in practice, due to ignorable missingness results outlined in the following section, and significantly simplifies the handling of missing data, as well as, mitigates the non-identifiability issues mentioned above that come with the more general MNAR assumption.

2.2.4 Ignorable missingness

In the previous section we introduced three missingness assumptions that are used to determine the validity of statistical procedures on incomplete data. Specifically, these assumptions help us assess whether it is appropriate to ignore the missingness mechanism in our procedures. In this section, we discuss when the missingness mechanism is ignorable, focusing on the two main statistical tasks addressed in this thesis: missing data imputation and model estimation.

Definition 4 *For missing data imputation, the missingness mechanism is ignorable if*

(van Buuren, 2018, Section 2.2.6)

$$p^*(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{m}) = p^*(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}). \tag{2.12}$$

This means that the missingness is ignorable for missing data imputation if the data are MAR (or MCAR). We show this using the informal MAR notation, $p^*(\boldsymbol{m} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}) = p^*(\boldsymbol{m} \mid \boldsymbol{x}_{\text{obs}})$, as follows:

$$p^*(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{m}) = \frac{p^*(\boldsymbol{m} \mid \boldsymbol{x}_{\text{mis}}, \boldsymbol{x}_{\text{obs}}) p^*(\boldsymbol{x}_{\text{mis}}, \boldsymbol{x}_{\text{obs}})}{\int p^*(\boldsymbol{m} \mid \boldsymbol{x}_{\text{mis}}, \boldsymbol{x}_{\text{obs}}) p^*(\boldsymbol{x}_{\text{mis}}, \boldsymbol{x}_{\text{obs}}) \, \mathrm{d}\boldsymbol{x}_{\text{mis}}} \tag{2.13}$$

$$= \frac{p^*(\boldsymbol{m} \mid \boldsymbol{x}_{\text{obs}}) p^*(\boldsymbol{x}_{\text{mis}}, \boldsymbol{x}_{\text{obs}})}{p^*(\boldsymbol{m} \mid \boldsymbol{x}_{\text{obs}}) \int p^*(\boldsymbol{x}_{\text{mis}}, \boldsymbol{x}_{\text{obs}}) \, \mathrm{d}\boldsymbol{x}_{\text{mis}}} \tag{2.14}$$

$$= p^*(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}). \tag{2.15}$$

This means that in order to perform imputation of the missing data we do not need to know or estimate the missingness mechanism, thus greatly simplifying the statistical procedure.

Definition 5 *For model estimation, the missingness mechanism is ignorable if (Little and Rubin, 2020, Theorem 6.1A; Seaman et al., 2013, Theorem 1)*

I. the model parameters $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ are distinct, such that the joint parameter space of $(\boldsymbol{\theta}, \boldsymbol{\gamma}) \in \Omega_{\boldsymbol{\theta},\boldsymbol{\gamma}}$ is the product space of each individual parameter $\Omega_{\boldsymbol{\theta}} \times \Omega_{\boldsymbol{\gamma}}$,

II. and, the full likelihood $p_{\boldsymbol{\theta},\boldsymbol{\gamma}}(\boldsymbol{x}_{\text{obs}}, \boldsymbol{m})$ can be factored into two factors, such that each parameter appears in only one of the factors.

Hence, for a correctly specified model, the missingness mechanism is ignorable in likelihood-based estimation if the data are MAR (or MCAR). We show that the full likelihood $p_{\boldsymbol{\theta},\boldsymbol{\gamma}}(\boldsymbol{x}_{\text{obs}}, \boldsymbol{m})$ factorises under the MAR assumption, using the informal MAR notation:

$$p_{\boldsymbol{\theta},\boldsymbol{\gamma}}(\boldsymbol{x}_{\text{obs}}, \boldsymbol{m}) = \int p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}) p_{\boldsymbol{\gamma}}(\boldsymbol{m} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}) \, \mathrm{d}\boldsymbol{x}_{\text{mis}} \tag{2.16}$$

$$= \underbrace{\int p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}) \, \mathrm{d}\boldsymbol{x}_{\text{mis}}}_{\text{Depends on } \boldsymbol{\theta}} \cdot \underbrace{p_{\boldsymbol{\gamma}}(\boldsymbol{m} \mid \boldsymbol{x}_{\text{obs}})}_{\text{Depends on } \boldsymbol{\gamma}}. \tag{2.17}$$

And hence, assuming distinctness of the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$, and that we are typically only interested in $\boldsymbol{\theta}$, we have

$$p_{\boldsymbol{\theta},\boldsymbol{\gamma}}(\boldsymbol{x}_{\text{obs}}, \boldsymbol{m}) \propto \int p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}) \, \mathrm{d}\boldsymbol{x}_{\text{mis}}, \tag{2.18}$$

where the r.h.s. is often called, perhaps counter-intuitively, the "ignorable likelihood"

(Little and Rubin, 2020, Section 6.2).

In this thesis, we will assume that the missingness is ignorable to simplify the handling of missing data, which also means that we assume that the data are MAR.

2.2.5 Missing data imputation

Traditional statistical procedures have been primarily developed for fully-observed data sets. But many real-world data sets contain missing data. This raises an important question: how can we use the statistical procedures developed for fully-observed data on incomplete data sets? Furthermore, the missing values themselves may be of key interest, so how can we learn about them?

Multiple imputation (MI, Rubin, 1987; Little and Rubin, 2020, Section 5.4) aims to address both questions. Firstly, we note that the missing data are inherently random and typically follow an unknown distribution $p^*(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{m})$, which under the MAR assumption in definition 2 becomes $p^*(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$, see section 2.2.4. Hence, the first step in MI is to produce $K > 1$ independent imputations of the missing data that approximately follow the unknown distribution of missing variables $p^*(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$. In the second step of MI, the incomplete data set is copied K times, and each copy is completed using one of the K imputations. Each of the K imputed data sets is then analysed independently using the statistical procedures for fully-observed data. Finally, in the third step, the analysis results are pooled using Rubin's rules (Rubin, 1987, Result 3.2; Little and Rubin, 2020, Section 5.4) to obtain a single result and accurately estimate its uncertainty.

The key step to the MI procedure is the choice of an imputation method that aims to draw approximate samples from the unknown conditional distribution $p^*(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$. In this thesis, we will use deep statistical models for this task.

2.2.6 Model estimation from incomplete data

In unsupervised machine learning, a key objective is often to estimate a parametric statistical model that approximates an unknown ground truth distribution from samples. For incomplete data, the ground truth distribution is

$$p^*(\boldsymbol{m}, \boldsymbol{x}_{\text{obs}}) = \int p^*(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}) p^*(\boldsymbol{m} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}) \, \mathrm{d}\boldsymbol{x}_{\text{mis}}. \tag{2.19}$$

Here, the dimensionality of $\boldsymbol{x}_{\text{obs}}$ varies between 0 and D depending on the missingness pattern $\boldsymbol{m}$ (see section 2.2.1). To represent this distribution, we specify a statistical model

with an equivalent factorisation using the selection modelling approach in eq. (2.7):

$$p_{\boldsymbol{\theta},\boldsymbol{\gamma}}(\boldsymbol{m}, \boldsymbol{x}_{\text{obs}}) = \int p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}) p_{\boldsymbol{\gamma}}(\boldsymbol{m} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}) \, \mathrm{d}\boldsymbol{x}_{\text{mis}}. \tag{2.20}$$

In the general case, our goal is to estimate the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ such that $p_{\boldsymbol{\theta},\boldsymbol{\gamma}}(\boldsymbol{m}, \boldsymbol{x}_{\text{obs}}) \approx p^*(\boldsymbol{m}, \boldsymbol{x}_{\text{obs}})$, given only samples from $p^*(\boldsymbol{m}, \boldsymbol{x}_{\text{obs}})$.

Given N i.i.d. data-points $\{\boldsymbol{m}^i, \boldsymbol{x}^i_{\text{obs}}\}_{i=1}^N$ drawn from $p^*(\boldsymbol{m}, \boldsymbol{x}_{\text{obs}})$, a principled approach to parameter estimation is maximum-likelihood estimation (MLE, *e.g.* Murphy, 2021, Section 4.2). MLE seeks to maximise the log-likelihood of the parameters

$$\ell(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \frac{1}{N} \sum_{i=1}^{N} \log p_{\boldsymbol{\theta},\boldsymbol{\gamma}}(\boldsymbol{m}^i, \boldsymbol{x}^i_{\text{obs}}). \tag{2.21}$$

By inserting the ignorable missingness assumption, we can factorise the log-likelihood into two terms, one for each parameter $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ (section 2.2.4)

$$\ell(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \underbrace{\frac{1}{N} \sum_{i=1}^{N} \log \int p_{\boldsymbol{\theta}}(\boldsymbol{x}^i_{\text{obs}}, \boldsymbol{x}^i_{\text{mis}}) \, \mathrm{d}\boldsymbol{x}^i_{\text{mis}}}_{\ell(\boldsymbol{\theta})} + \underbrace{\frac{1}{N} \sum_{i=1}^{N} \log p_{\boldsymbol{\gamma}}(\boldsymbol{m}^i \mid \boldsymbol{x}^i_{\text{obs}})}_{\ell(\boldsymbol{\gamma})}. \tag{2.22}$$

We can use the full log-likelihood in the equation above to optimise the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ of our model. Moreover, since the full log-likelihood factorises into two terms, one for $\boldsymbol{\theta}$ and one for $\boldsymbol{\gamma}$, maximising $\ell(\boldsymbol{\theta}, \boldsymbol{\gamma})$ jointly is the same as optimising $\ell(\boldsymbol{\theta})$ and $\ell(\boldsymbol{\gamma})$ separately. Hence, because our primary interest lies in the distribution $p_{\boldsymbol{\theta}}(\boldsymbol{x})$, we only need to optimise $\boldsymbol{\theta}$ with respect to the log-likelihood $\ell(\boldsymbol{\theta})$:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \Omega_{\boldsymbol{\theta}}} \ell(\boldsymbol{\theta}). \tag{2.23}$$

When this maximisation is not analytically tractable, we can typically resort to approximations, such as stochastic gradient ascent (SGA, *e.g.* Spall, 2003, Chapter 5), to find a local maximum.

Importantly, the log-likelihood $\ell(\boldsymbol{\theta})$ requires solving an integral over the missing variables $\boldsymbol{x}_{\text{mis}}$:

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}}) = \int p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}) \, \mathrm{d}\boldsymbol{x}_{\text{mis}}. \tag{2.24}$$

While the marginalisation of the missing variables $\boldsymbol{x}_{\text{mis}}$ in the above equation may be feasible for models with specific simplifying assumptions, it is generally intractable for more general models. In the next section, we review a principled approach to tackle this

challenge.

Expectation-maximisation

A classical method that circumvents the need for directly marginalising the missing variables in eq. (2.24) is the expectation-maximisation algorithm (EM, Dempster et al., 1977).[5] To introduce the EM algorithm we adopt the variational free-energy view of the EM by Neal and Hinton (1998), which better aligns with some useful variations of the algorithm discussed later. The algorithm works by first establishing an evidence lower-bound (ELBO) to the marginal log-likelihood, which, apart from a sign change, is analogous to the variational free energy in statistical physics:

$$\ell(\boldsymbol{\theta}) \geq \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{f^i(\boldsymbol{x}^i_{\text{mis}}|\boldsymbol{x}^i_{\text{obs}})}\left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}^i_{\text{obs}}, \boldsymbol{x}^i_{\text{mis}})}{f^i(\boldsymbol{x}^i_{\text{mis}} \mid \boldsymbol{x}^i_{\text{obs}})}\right], \tag{2.25}$$

where $f^i(\boldsymbol{x}^i_{\text{mis}} \mid \boldsymbol{x}^i_{\text{obs}})$ is an auxiliary distribution for the i-th data-point. Then, starting with some initial parameters $\boldsymbol{\theta}^t$ at iteration $t = 0$, the algorithm iterates between two steps:

- E-step: Maximises the bound w.r.t. $f^i(\boldsymbol{x}^i_{\text{mis}} \mid \boldsymbol{x}^i_{\text{obs}})$ for each data-point $\boldsymbol{x}^i_{\text{obs}}$ in the data set, which yields $f^i(\boldsymbol{x}^i_{\text{mis}} \mid \boldsymbol{x}^i_{\text{obs}}) = p_{\boldsymbol{\theta}^t}(\boldsymbol{x}^i_{\text{mis}} \mid \boldsymbol{x}^i_{\text{obs}})$.
- M-step: Maximises the bound w.r.t. $\boldsymbol{\theta}$ to obtain $\boldsymbol{\theta}^{t+1}$, with $f^i(\boldsymbol{x}^i_{\text{mis}} \mid \boldsymbol{x}^i_{\text{obs}}) = p_{\boldsymbol{\theta}^t}(\boldsymbol{x}^i_{\text{mis}} \mid \boldsymbol{x}^i_{\text{obs}})$ for each $\boldsymbol{x}^i_{\text{obs}}$ in the data set.

The iterations of EM are guaranteed to converge to a local optimum of $\ell(\boldsymbol{\theta})$ (Dempster et al., 1977, Theorem 2). The local convergence of EM is similar to the guarantees of SGA when used to find the MLE solution in eq. (2.23).

Informally, the reason why the EM algorithm works, is that the E-step transforms the inequality in eq. (2.25) into an equality, achieving a kind of "local" marginalisation of the missing variables. As a result, the gradients of the ELBO w.r.t. $\boldsymbol{\theta}$ are exactly the gradients of the marginal log-likelihood $\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}})$. In fact, if the M-step was replaced with a "minimal" M-step that performed a single step of SGA, hence increasing but not necessarily maximising the ELBO,[6] then the iterates would correspond to those obtained by directly maximising eq. (2.23) with SGA (see appendix A for more details). Importantly, to fit the parameters $\boldsymbol{\theta}$, the EM algorithm does not require directly evaluating the integral in eq. (2.24) to marginalise the model $p_{\boldsymbol{\theta}}(\boldsymbol{x})$.

[5] Although we discuss the EM in the context of models where marginalisation is intractable, it is also relevant for certain models with tractable marginalisation, such as finite mixture models.

[6] An EM algorithm that increases the ELBO in the M-step, but not necessarily maximises, is also known as the generalised EM (Dempster et al., 1977).

However, the distribution $p_{\boldsymbol{\theta}^t}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ in the E-step is often complex, making analytical expressions for it generally intractable, and the expectation in eq. (2.25) required in the M-step is generally also computationally demanding. To address these limitations, more flexible extensions of the EM algorithm have been proposed:

Monte Carlo EM (Wei and Tanner, 1990): This variant aims to sample the distribution $p_{\boldsymbol{\theta}^t}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ using, for example, iterative approximate inference methods such as Markov chain Monte Carlo (MCMC, *e.g.* Barber, 2017, Chapter 27.4). Then, the required expectation in the M-step is approximated using a sample average. The Monte Carlo approach can hence be seen as a decomposition of the model estimation objective into two iterative tasks: missing data imputation using the model $p_{\boldsymbol{\theta}^t}(\boldsymbol{x})$, followed by model parameter estimation using the imputed data. A limitation of the Monte Carlo EM, is that it involves sampling of the conditional distribution $p_{\boldsymbol{\theta}^t}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ for each observed data-point at every iteration of the EM algorithm, which can be computationally expensive.

Variational EM (Beal and Ghahramani, 2003) : This variant uses variational inference to approximate the intractable distribution $p_{\boldsymbol{\theta}^t}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ with a more tractable surrogate distribution $q(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ in some variational family $\mathcal{Q}$. That is, instead of setting $f^i(\boldsymbol{x}^i_{\text{mis}} \mid \boldsymbol{x}^i_{\text{obs}}) = p_{\boldsymbol{\theta}^t}(\boldsymbol{x}^i_{\text{mis}} \mid \boldsymbol{x}^i_{\text{obs}})$, variational EM sets $f^i(\boldsymbol{x}^i_{\text{mis}} \mid \boldsymbol{x}^i_{\text{obs}}) = \arg\max_{q^i(\boldsymbol{x}^i_{\text{mis}}) \in \mathcal{Q}} \mathbb{E}_{q^i(\boldsymbol{x}^i_{\text{mis}})}[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}^i_{\text{obs}}, \boldsymbol{x}^i_{\text{mis}}) / q^i(\boldsymbol{x}^i_{\text{mis}})]$ for each $\boldsymbol{x}^i_{\text{obs}}$ in the data set. Hence, the variational approach can improve the computational efficiency by fitting a variational distribution $q(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ that can be efficiently sampled (or, alternatively, a $q(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ that enables efficient computation of the required expectation). However, this approximation might come at the cost of some bias if the variational distribution $q(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ is not flexible enough to well-approximate $p_{\boldsymbol{\theta}^t}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$.

We refer the interested readers to McLachlan and Krishnan (2007) for more details on the large family of EM algorithms.

In chapter 6 of this thesis, we will address the limitations of the Monte Carlo EM and variational EM algorithms by using a hybrid approach for general deep statistical model estimation.

Chapter 3

Prior work and research gap

After establishing the fundamental background on deep statistical models and missing data in chapter 2, we now review the recent research in the intersection of the two areas—specifically focusing on the application of deep statistical models in the context of missing data.

In section 3.1, we review recent work on the application of deep models to imputation of missing data. Subsequently, in section 3.1.3 we discuss the research gap and formulate a research question within the area of missing data imputation, which this thesis aims to explore.

Turning our attention to deep model estimation from incomplete data, section 3.2 provides an overview of the relevant developments in this area. Following this, in section 3.2.3 we outline the research gap and specific research questions to be addressed in this thesis concerning deep model estimation from incomplete data.

3.1 Missing data imputation using deep models

Missing data imputation is a key step in data analysis (section 2.2.5) and is closely interlinked with model estimation from incomplete data (section 2.2.6). Moreover, it can be a goal in itself, such as in pharmaceutical design, where it can facilitate the creation of new drugs with desired properties by generating new molecules *in-silico* conditional on a functional group (*e.g.* Griffiths and Hernández-Lobato, 2020).

As covered in section 2.2.5, under the assumption of MAR data (see definition 2), the key to multiple imputation of missing data is drawing samples that closely resemble the unknown ground truth conditional distribution $p^*(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$. Motivated by the success of deep statistical models in capturing complex distributions of natural data, researchers have made efforts to harness this capability for missing data imputation. In the following sections, we review the existing research on using deep models for multiple imputation of missing data.

3.1.1 Using conditionally-specified models

One way to leverage deep models for missing data imputation is grounded in *conditional* deep models. Unlike standard deep models that aim to approximate an unconditional distribution $p_{\boldsymbol{\theta}}(\boldsymbol{x}) \approx p^*(\boldsymbol{x})$ of the data, conditional models learn a distribution of the form $p_{\boldsymbol{\theta}}(\boldsymbol{x} \mid \boldsymbol{y}) \approx p^*(\boldsymbol{x} \mid \boldsymbol{y})$, where $\boldsymbol{y}$ is the conditioning variable. The conditioning is generally achieved by simply providing $\boldsymbol{y}$ as an additional input to the neural networks of the deep model.

In the context of multiple imputation we are interested in learning conditional distributions of the form $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$. But, as discussed in section 2.2.1, the challenge arises from the potential existence of up to 2^D missingness patterns, rendering the learning of 2^D conditional deep models impractical. Instead, the existing works aim to learn a model capable of arbitrary conditioning on any subset of the random variable $\boldsymbol{x}$. The task of learning an arbitrarily-conditional model can be viewed as that of learning a set of conditional models $\{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}) \mid \boldsymbol{x}_{\text{obs}} = \boldsymbol{x}[\boldsymbol{m}], \boldsymbol{x}_{\text{mis}} = \boldsymbol{x}[1-\boldsymbol{m}], \forall \boldsymbol{m} \in \{0,1\}^D\}$ with shared parameters $\boldsymbol{\theta}$. While the set of distributions learnt in this way may not have a single well-defined joint distribution (*e.g.* Arnold et al., 1999, Section 1.6), parameter sharing between all the conditional distributions can make the learning computationally efficient.

The training procedure of an arbitrarily-conditional model ideally involves using a fully-observed data set. At each iteration of the algorithm, data-points $\boldsymbol{x}$ are randomly partitioned into the target $\boldsymbol{x}_{\text{targ}}$ and conditioning $\boldsymbol{x}_{\text{cond}}$ parts based on a heuristic distribution (see *e.g.* Tashiro et al., 2021, Section 4.3 for an example of the heuristics). The model is then trained to learn the conditional distribution $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{targ}} \mid \boldsymbol{x}_{\text{cond}})$ for all $\boldsymbol{x}_{\text{targ}}$ and $\boldsymbol{x}_{\text{cond}}$. Furthermore, when dealing with incomplete training data, the target and conditioning parts are chosen as subsets of the observed incomplete data-point $\boldsymbol{x}_{\text{obs}}$.

Upon training the arbitrarily-conditional model, multiple imputation is performed via standard sampling of the conditional model. This approach relies on the (fairly strong) assumption that the model can generalise from the conditional distributions seen during training $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{targ}} \mid \boldsymbol{x}_{\text{cond}})$ to the distribution of the missing variables $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$. For example, when the training data is incomplete, it assumes that the model can generalise from being trained on $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{targ}} \mid \boldsymbol{x}_{\text{cond}})$, where $(\boldsymbol{x}_{\text{targ}}, \boldsymbol{x}_{\text{cond}}) = \boldsymbol{x}_{\text{obs}}$, to the distribution of missing variables $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$, where $(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}) = \boldsymbol{x}$. While the assumption may not always hold, especially in high-missingness settings, the method's simplicity and scalability remain appealing in many use-cases.

The arbitrarily-conditional modelling approach as described in this section was first pro-

posed by Ivanov et al. (2019) and applied for multiple imputation using *variational autoencoders* as the conditional model, leveraging the conditional factorisation of the decoder distribution in VAEs (see below eq. (2.1)). Building on this work, Li et al. (2020) introduced several normalising flow architectures capable of arbitrary conditioning, thereby extending the approach to the family of *normalising flow* models. In a more recent development, Tashiro et al. (2021) further extended this approach to *conditional score-based diffusion models*, and demonstrated its effectiveness on time-series imputation. Additionally, a related approach has been proposed by Yoon et al. (2018) leveraging *generative adversarial networks* tailored for multiple imputation of missing data.

3.1.2 Using jointly-specified models

An alternative approach to multiple imputation of missing data is based on *conditional sampling* of a *jointly-specified* deep model.[7] Given a joint deep model $p_{\boldsymbol{\theta}}(\boldsymbol{x}) \approx p^*(\boldsymbol{x})$, multiple imputation could be performed via conditional sampling, where the imputations $\boldsymbol{x}_{\text{mis}}$ are sampled from $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$, conditional on the observed data-point $\boldsymbol{x}_{\text{obs}}$. However, exact sampling of the conditional distribution $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ is generally intractable for deep models, and hence approximations are needed.

Approximate inference-based approaches with asymptotic guarantees

Traditional alternatives for approximate sampling of such intractable conditional distributions include Markov chain Monte Carlo (MCMC, *e.g.* Barber, 2017, Chapter 27.4) and importance sampling (IS, *e.g.* Chopin and Papaspiliopoulos, 2020, Chapter 8). While both approaches offer asymptotically exact sampling of the target distribution $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$, they differ in their theoretical properties.

In MCMC, the practitioner specifies a stochastic transition kernel $\tau(\boldsymbol{x}_{\text{mis}}^{t+1} \mid \boldsymbol{x}_{\text{mis}}^{t})$ that defines the move from the current state $\boldsymbol{x}_{\text{mis}}^{t}$ of the Markov chain to the next $\boldsymbol{x}_{\text{mis}}^{t+1}$. As long as the kernel is appropriately specified, iteratively applying this kernel starting from an arbitrary initial state $\boldsymbol{x}_{\text{mis}}^{0}$ converges to samples $\boldsymbol{x}_{\text{mis}}^{\infty}$ from the target distribution $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ as the number of iterations t goes to infinity.

In IS, a practitioner first specifies a proposal distribution $q(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ that can be easily evaluated and sampled. Then, a set of samples $\{\boldsymbol{x}_{\text{mis}}^{k}\}_{k=1}^{M}$ is drawn from the proposal $q(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$. These samples are re-sampled with probabilities proportional to $p(\boldsymbol{x}_{\text{mis}}, \boldsymbol{x}_{\text{obs}})/q(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ such that they (approximately) follow the target conditional

[7] For the purposes of this discussion, we assume that the model has been trained on fully-observed data, although it could also have been trained using the EM or related algorithms (see section 2.2.6), unless otherwise noted.

distribution $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$. The importance sampling-resampling approach typically introduces bias of order of $\mathcal{O}(1/M)$ (Owen, 2013; Paananen et al., 2021), where M is the number of samples taken from the proposal distribution $q(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$. The bias diminishes as $M \to \infty$, ultimately approaching zero.

While both MCMC and IS promise asymptotically exact sampling, their practical efficiency greatly depends on the specification of the transition kernel for MCMC and the proposal distribution for IS. Moreover, the varying target distribution for each observation $\boldsymbol{x}_{\text{obs}}$ can complicate the tuning of these methods to achieve optimal sampling efficiency, if the number of observations is large.

Nevertheless, both MCMC and IS methodologies have found application in multiple imputation of missing data using deep models. In the seminal paper on *variational autoencoders* (VAEs) by Rezende et al. (2014, Appendix F), the authors proposed pseudo-Gibbs sampling to draw approximate samples from the conditional distribution $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ defined by the VAE. The method alternates between sampling from $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$ and $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{z})$, drawing samples that progressively approach $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ as the iterations proceed. While reminiscent of the traditional Gibbs sampler (Geman and Geman, 1984), the unavailability of the model posterior distribution $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$ in VAEs leads to its replacement with sampling from $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$ in the pseudo-Gibbs approach. As a result, the asymptotic accuracy of the pseudo-Gibbs sampler depends on how well the encoder distribution $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$ approximates the model posterior $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$.

Building on this, Mattei and Frellsen (2018, Section 3.2) proposed a simple modification, known as Metropolis-within-Gibbs (MWG), which incorporates a Metropolis–Hasting accept-reject step (Metropolis et al., 1953; Hastings, 1970) to correct for the mismatch between $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$ and $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$. This simple modification ensures that the algorithm asymptotically convergences to samples from $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$, defined by the VAE, even in the presence of imperfect approximation $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$ of the intractable model posterior $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$.

In the context of *normalising flows*, Cannella et al. (2021) introduced projected latent MCMC (PL-MCMC) for conditional sampling of flow models. Similar to pseudo-Gibbs and MWG, PL-MCMC leverages the latent space of the normalising flow in a Metropolis–Hastings-type algorithm, in order to efficiently sample the conditional distribution of the normalising flow model.

More recently, Wu et al. (2023) proposed a sequential Monte Carlo method, an IS-type approach, for conditional sampling of *denoising diffusion models*, promising asymptoti-

cally exact sampling of the model. They demonstrated the effectiveness of the approach for protein design.

Approximate ad-hoc approaches

In addition to the approaches grounded in traditional approximate inference methodology, the diverse landscape of deep models has influenced the development of numerous ad-hoc methods for approximate conditional sampling of jointly-specified models. While these methods typically lack asymptotic guarantees, they often excel in providing scalable approximations. In this section, we highlight a few notable approaches in this area.

Strauss and Oliva (2022) introduced a method for *variational autoencoders* that estimates the posterior distribution over the latents given partial observations, enabling conditional sampling. In addition to the encoder distribution $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x})$ they train an auxiliary model $q_{\boldsymbol{\psi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{cond}})$ by minimising $\mathbb{E}_{p^*(\boldsymbol{x}_{\text{targ}}|\boldsymbol{x}_{\text{cond}})}[D_{\text{KL}}(q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{targ}}, \boldsymbol{x}_{\text{cond}}) \parallel q_{\boldsymbol{\psi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{cond}}))]$, where $(\boldsymbol{x}_{\text{cond}}, \boldsymbol{x}_{\text{targ}}) = \boldsymbol{x}$ are obtained by randomly partitioning the fully-observed data-point $\boldsymbol{x}$. The auxiliary model can then be used to sample imputations by first sampling from $q_{\boldsymbol{\psi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$, followed by $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{z})$. This is tractable when sampling $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{z})$ is tractable, for example, when the decoder distribution factorises such that $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{z}) = p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{z})$.

Moreover, most VAE models that are trained on incomplete data also learn a partially-observed posterior distribution $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}) \approx p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ that can be readily used for (approximate) missing data imputation by first sampling $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$, followed by $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{z})$, just like the aforementioned approach.

Whang et al. (2021) explored the use of variational inference for conditional inference in the latent space of a *normalising flow*. Recognising that the deterministic mapping between the latent and the observed space in flow models makes the variational objective challenging to optimise, they instead opt for approximate conditioning. Their approach involves smoothing the observations using a user-defined smoothing kernel to facilitate a more tractable optimisation objective at the cost of approximation accuracy.

Li et al. (2019) proposed an approach for conditional sampling of *generative adversarial networks*. Their method involves learning a separate imputation network using an adversarial objective that matches the conditional distribution of an unconditional GAN. Experimentally, they show that their approach can result in better-calibrated imputations compared to the aforementioned approach by Yoon et al. (2018).

3.1.3 Research question

As outlined in the preceding sections, the research on missing data imputation using deep models is diverse, encompassing a wide array of models and techniques, each presenting various advantages and trade-offs.

The generality and flexibility of the VAE framework has influenced a substantial interest in applying it to the task of data imputation. As discussed in section 3.1.2, popular methods for conditional sampling of VAEs are based on Gibbs sampling, a type of MCMC approach. The appeal of these methods stems from their relative simplicity, as they cleverly reuse parts of the trained VAE to perform conditional sampling, without introducing additional hyperparameters to tune. However, despite offering asymptotic guarantees, the Gibbs sampler is not without its limitations, which can present challenges in practical applications.

Remarkably, despite the popular use of Gibbs-like approaches for conditional sampling of VAEs, little research has been devoted to understanding the potential pitfalls of these methods within the context of VAEs. Hence, to better understand the challenges of conditional sampling of VAEs, we formulate the following research question:

Research Question 1: What are the challenges when using pre-trained VAEs for missing data imputation, and how might these challenges be overcome?

We address the above question in chapter 4.

3.2 Deep model estimation from incomplete data

We now shift our attention to the estimation of deep statistical model from incomplete data. We assume that we have an incomplete training data set $\{\boldsymbol{x}_{\text{obs}}^i\}_{i=1}^N$, comprising N i.i.d. realisations from an unknown distribution $p^*(\boldsymbol{x})$. Additionally, we assume that the data is missing-at-random (MAR, see definition 2) and the missingness mechanism is ignorable (see section 2.2.4). Our objective is to estimate the joint model $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ such that it approximates the ground truth distribution $p^*(\boldsymbol{x})$ as closely as possible. Once the model is estimated, multiple imputation techniques from the previous section can be used to perform downstream tasks on incomplete test-time data sets.

In this section, we provide a brief overview of the notable methods for estimating deep models from incomplete data.

3.2.1 Variational autoencoders

As discussed in section 2.2.6, a probabilistically-principled approach to handling missing data during training involves marginalising the missing variables from the likelihood. In VAEs, we can marginalise the likelihood as follows:

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}}) = \int p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})\, \mathrm{d}\boldsymbol{x}_{\text{mis}} \tag{3.1}$$

$$= \iint p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}} \mid \boldsymbol{z}) p_{\boldsymbol{\theta}}(\boldsymbol{z})\, \mathrm{d}\boldsymbol{z}\, \mathrm{d}\boldsymbol{x}_{\text{mis}} \tag{3.2}$$

$$= \int p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}} \mid \boldsymbol{z}) p_{\boldsymbol{\theta}}(\boldsymbol{z})\, \mathrm{d}\boldsymbol{z}. \tag{3.3}$$

As shown above, marginalisation of the likelihood in VAEs corresponds to marginalising the decoder distribution $p_{\boldsymbol{\theta}}(\boldsymbol{x} \mid \boldsymbol{z})$. Fortunately, in VAEs, the marginalisation $\int p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}} \mid \boldsymbol{z})\, \mathrm{d}\boldsymbol{x}_{\text{mis}}$ is often computationally tractable due to a common assumption that the visible variables $\boldsymbol{x}$ are conditionally independent given the latents $\boldsymbol{z}$, that is, $p_{\boldsymbol{\theta}}(\boldsymbol{x} \mid \boldsymbol{z}) = \prod_{j=1}^{D} p_{\boldsymbol{\theta}}(x_j \mid \boldsymbol{z})$, hence $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}} \mid \boldsymbol{z}) = \prod_{j \in \text{obs}} p_{\boldsymbol{\theta}}(x_j \mid \boldsymbol{z})$. Then, following the standard derivation, the variational ELBO or importance-weighted ELBO (section 2.1.1) is used to estimate the model parameters:

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}}) \geq \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}}, \boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})} \right]. \tag{3.4}$$

The aforementioned marginalisation assumption and the variational ELBO objective are the key ingredients of the existing methods for VAE estimation from incomplete data. Here, we briefly review the distinguishing features of several notable approaches.

Methods adapting the encoder model

Several studies have investigated the necessary adaptations to the encoder model to accommodate incomplete data points.

Vedantam et al. (2017) and Wu and Goodman (2018) advocated for the use of a Gaussian product-of-experts variational distribution, with one expert per observed dimension, to handle the missing variables when specifying the variational distribution $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$. This choice was motivated by an investigation in the factor analysis case, a linear version of Gaussian VAEs, where it was observed that the posterior of incomplete data can be accurately described by a product-of-Gaussians (Williams et al., 2018). Notably, a product-of-Gaussians is another Gaussian distribution, rendering the variational distribution in these works a Gaussian.

In contrast, Ma et al. (2019) proposed an alternative approach by using a permutation

invariant neural network for the encoder. This method enables handling of arbitrary subsets of observed variables and facilitates the use of general distribution families beyond Gaussian distributions.

Other works, as reviewed below, chose a simpler approach for parametrising the encoder. They use a neural network with fixed dimensionality, similar to the fully-observed case, and set the values of the missing variables to a constant.

Methods adapting the generative model or the learning objective

Another avenue of research focuses on addressing the missing data problem by adapting either the generative model or the learning objective.

Motivated by the effectiveness of the importance-weighted ELBO in the fully-observed data case (Burda et al., 2015), Mattei and Frellsen (2019) proposed adopting the bound for training VAEs in the presence of incomplete data, which yielded significant improvements in model estimation.

In the context of heterogeneous incomplete data, Nazábal et al. (2020) suggested using a hierarchical prior to more "easily capture the (sometimes weak) statistical dependencies in the data", thereby aiding reconstruction of the missing variables. In the same line of research, Ma et al. (2020) introduced a two-stage approach consisting of (i) fitting univariate VAEs for each data dimension and subsequently (ii) fitting a dependency VAE model on top of the univariate models. This methodology, initially motivated by data heterogeneity, is also capable of handling incomplete data by omitting the missing dimensions in the first stage, as demonstrated by the authors. This method was extended by Peis et al. (2022) who introduced hierarchical latent variables and a Hamiltonian Monte Carlo-approach to increase the flexibility of the model.

Instead of the standard selection modelling approach, Collier et al. (2021) proposed a pattern-mixture VAE (see section 2.2.2). Here, the authors suggested learning a pattern-conditional VAE model $p_{\boldsymbol{\theta}}(\boldsymbol{x} \mid \boldsymbol{m})$ rather then the marginal data model $p_{\boldsymbol{\theta}}(\boldsymbol{x})$. The marginal model can then be derived by marginalising over the 2^D missingness patterns:

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{\boldsymbol{m} \in \{0,1\}^D} p_{\boldsymbol{\theta}}(\boldsymbol{m}) p_{\boldsymbol{\theta}}(\boldsymbol{x} \mid \boldsymbol{m}). \tag{3.5}$$

3.2.2 Normalising flows

When estimating normalising flows from incomplete data, additional challenges arise compared to VAEs. Specifically, unlike in VAEs, which often allow for efficient marginalisa-

tion of the missing variables, this task is generally infeasible in normalising flows due to the deterministic transformations that map between the base and data distributions (see section 2.1.2). Therefore, estimating flows from incomplete data may require iterative approaches akin to the EM algorithm (see section 2.2.6).

To this end, Cannella et al. (2021) proposed a Monte Carlo EM (Wei and Tanner, 1990) approach for normalising flow estimation from incomplete data. Given the intractability of conditional sampling in normalising flows, they use the Metropolis–Hastings algorithm with a flow-specific acceptance probability. This approach leverages the latent space of the flow model to efficiently sample conditional distributions, thus enabling model estimation from incomplete data by Monte Carlo EM. However, the sampling performance depends on the choice of a few hyper-parameters. Since the target distribution changes throughout the learning iterations, maintaining good sampling efficiency would require fine-tuning of these hyperparameters iteratively. Consequently, the practical utility of this approach for model estimation can be limited.

A few other existing methods (Richardson et al., 2020; Bernal, 2021) use a "hard" EM-type approach (Beal and Ghahramani, 2003, Section 3.1, referred to as EM for MAP estimation) where the E-step of the EM algorithm (section 2.2.6) is replaced by $\arg\max_{\boldsymbol{x}_{\text{mis}}} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$. The hard EM algorithm can be understood to be approximating the model's conditional distribution $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ with a point-mass distribution at its mode. Hence, this type of approach ignores the uncertainty about the missing variables. If the missingness fraction is small, then these simple approaches may be sufficient, but for larger missingness fractions they may not be applicable as they would often produce significantly biased estimates of the model.

3.2.3 Research questions

In the preceding sections we reviewed the existing literature on the estimation of deep statistical models from incomplete data. This allows us to formulate several research questions to be explored in this thesis.

In section 3.2.1, we highlighted that variational autoencoders are often amenable to useful assumptions that allow for marginalisation of the missing variables—the recommended approach for dealing with missing data (see section 2.2.6). As a result, a substantial body of work has leveraged this assumption for VAE estimation from incomplete data. However, the empirical evidence in these works suggests that adaptions to the encoder or the generative models may be needed to effectively handle the missing data.

Little research has been devoted to understanding *why* VAE estimation from incomplete

data is more challenging compared to the fully-observed case. To improve our understanding of the missing data problem in the context of VAEs, we raise the following research question:

Research Question 2: What are the factors that contribute to the increased complexity of VAE estimation from incomplete data, and what strategies can be used to effectively address this additional complexity due to missing data?

We address these question in detail in chapter 5.

In contrast to the extensive literature on VAE estimation from incomplete data, the literature on normalising flow estimation in this context is small (see section 3.2.2). Moreover, existing methods for VAEs heavily rely on the critical assumption that the decoder distribution can be efficiently marginalised (see the paragraph below eq. 3.3). However, this assumption breaks down in scenarios where, for example, the decoder distribution is specified to be autoregressive or involves a conditional normalising flow (*e.g.* Gulrajani et al., 2016; Praljak et al., 2023). Hence, there is a disparity between the well-explored area of VAE estimation under the key assumption that missing variables can be marginalised and the under-explored areas on more general deep models where this assumption does not hold. This necessitates a deeper investigation into general approaches for deep statistical model estimation from incomplete data.

As discussed in section 2.2.6, the Monte Carlo EM framework (Wei and Tanner, 1990) provides a general framework for statistical model estimation from incomplete data. It involves two iterative steps: conditional sampling to impute the missing values and model estimation using the completed data. However, as mentioned in section 3.1.2, exact conditional sampling is often intractable for most deep statistical models, requiring approximations. Moreover, using approximate methods such as MCMC or importance sampling in the context of Monte Carlo EM presents two key challenges: First, the target distribution $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ varies across the potentially large set of observed data-points $\boldsymbol{x}_{\text{obs}}$ in the training data. Second, this distribution changes throughout the learning iterations as the model is updated. These factors make efficient sampling difficult in the inner loop of the Monte Carlo EM algorithm when applied to deep models since the methods' sampling efficiency often heavily relies on hyper-parameter tuning which is difficult in this iterative setting. Hence, efficiently estimating general deep statistical models from incomplete data remains an open problem.

To further the understanding of the problem of general model estimation we pose the following question:

Research Question 3: How do we efficiently estimate general statistical models for which simplifying assumptions, like marginalisation of the missing variables, may not be applicable?

This question is the main focus of chapter 6, where we explore a strategy based on Monte Carlo EM, variational inference, and traditional imputation techniques to construct a general-purpose method for deep statistical model estimation from incomplete data.

Chapter 4

Missing data imputation with variational autoencoders

Multiple imputation of missing data, which involves approximately sampling the conditional distribution of missing variables $p^*(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$, is a key step for both incomplete data analysis (section 2.2.5) and model estimation from incomplete data (section 2.2.6). Variational autoencoders (VAEs) have attracted a substantial interest for this application domain, as highlighted in section 3.1. However, since conditional sampling of VAEs is generally intractable, existing approaches rely on approximations. As discussed in section 3.1.3, there is a lack of understanding of the effectiveness of these approximate methods in the context of VAEs models for missing data imputation. This motivates the following research question:

Research Question 1: What are the challenges when using pre-trained VAEs for missing data imputation, and how might these challenges be overcome?

In this chapter, we explore the aforementioned research question, with a specific focus on VAEs that have been trained on fully-observed (or completed) data. Our publication, included verbatim in section 4.1, reveals that existing methods are subject to pitfalls which are associated with commonly-desired properties of learnt VAEs. Subsequently, the publication introduces two original iterative conditional sampling methods, based on adaptive Markov chain Monte Carlo and importance sampling, that improve missing data imputation with pre-trained VAEs.

4.1 Publication

This section includes a verbatim copy of the following publication:

Vaidotas Simkus and Michael U. Gutmann. Conditional Sampling of Variational Autoencoders via Iterated Approximate Ancestral Sampling. *Transactions on Machine Learning Research*, 2023

Published in Transactions on Machine Learning Research (11/2023)

Conditional Sampling of Variational Autoencoders via Iterated Approximate Ancestral Sampling

Vaidotas Simkus
Michael U. Gutmann
School of Informatics
University of Edinburgh

Reviewed on OpenReview: *https://openreview.net/forum?id=I5sJ6PU6JN*

Abstract

Conditional sampling of variational autoencoders (VAEs) is needed in various applications, such as missing data imputation, but is computationally intractable. A principled choice for asymptotically exact conditional sampling is Metropolis-within-Gibbs (MWG). However, we observe that the tendency of VAEs to learn a structured latent space, a commonly desired property, can cause the MWG sampler to get "stuck" far from the target distribution. This paper mitigates the limitations of MWG: we systematically outline the pitfalls in the context of VAEs, propose two original methods that address these pitfalls, and demonstrate an improved performance of the proposed methods on a set of sampling tasks.

1 Introduction

Conditional sampling of modern deep probabilistic models is an important but generally intractable problem. Variational autoencoders (VAEs, Kingma & Welling, 2013; Rezende et al., 2014) are a family of deep probabilistic models that *capture the complexity of real-world data distributions via a structured latent space*. The impressive modelling capability and the usefulness of the structured latent space make VAEs a model of choice in a broad range of domains from healthcare (Han et al., 2019) and chemistry (Gómez-Bombarelli et al., 2018) to images (Child, 2021) and audio (van den Oord et al., 2017). Ancestral sampling can be used for efficient unconditional sampling of VAEs, but many downstream tasks, for example, prediction or missing data imputation (e.g. Goodfellow et al., 2016, Chapter 5.1.1), instead require *conditional sampling*. However, *for VAEs, this is intractable*, and hence approximate methods are needed.

A canonical approximate method is Markov chain Monte Carlo (MCMC, e.g. Barber, 2017, Chapter 27.4) but the general lack of knowledge about the learnt VAE may make tuning, for example, picking a good proposal distribution, and hence successfully using MCMC samplers challenging. To make sampling easier, an approach called Metropolis-within-Gibbs (MWG, Mattei & Frellsen, 2018) re-uses the encoder, an auxiliary component from the training of the VAE, to construct a suitable proposal distribution in a Metropolis–Hastings-type algorithm (Metropolis et al., 1953; Hastings, 1970). The simplicity of MWG and its asymptotic convergence guarantees make it a compelling choice for conditional sampling of VAEs.

While a structured latent space is often a desirable property of VAEs, enabling the modelling of complex distributions, we *notice that this latent structure can cause the Markov chains of MWG to get "stuck"* hence impeding conditional sampling. In this paper we

- Detail the potential pitfalls of Metropolis-within-Gibbs in the context of VAEs (section 3).
- Propose a modification of MWG, called adaptive collapsed-Metropolis-within-Gibbs (AC-MWG, section 4.1), that mitigates the outlined pitfalls and prove its convergence.
- Introduce an alternative sampling method, called latent-adaptive importance resampling (LAIR, section 4.2), which demonstrates an improved sampling performance in our experiments.

- Evaluate the samplers on a set of conditional sampling tasks: (semi-)synthetic, where sampling from the ground truth conditional distributions is computationally tractable, and real-world missing data imputation tasks, where the ground truth distribution is not available.

With the proposed methods we address the conditional sampling problem of VAEs, a key challenge to downstream application of this flexible family of models. Our methods build and improve upon the limitations of MWG enabling more accurate use of VAEs in important tasks like missing data imputation.

2 Background: Conditional sampling of VAEs

We here describe the conditional sampling problem and the existing Gibbs-like methods that have been used to draw conditional samples.

2.1 Problem and assumptions

Given a pre-trained variational autoencoder, whose generative model we denote as $p(\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{x} \mid \boldsymbol{z})p(\boldsymbol{z})$, where $\boldsymbol{x} = (\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$ are the visible and $\boldsymbol{z}$ are the latent variables, we would like to sample:

$$p(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}) = \frac{\int p(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}, \boldsymbol{z}) \, \mathrm{d}\boldsymbol{z}}{p(\boldsymbol{x}_{\text{obs}})} = \int p(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{z}) p(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}) \, \mathrm{d}\boldsymbol{z}. \tag{1}$$

The variables $\boldsymbol{x}_{\text{mis}}$ and $\boldsymbol{x}_{\text{obs}}$ are respectively the target/missing and conditioning/observed variables. This choice of notation is motivated by the correspondence between conditional sampling and probabilistic imputation of missing data (Rubin, 1987; 1996).[1] Unlike unconditional generation, *ancestral sampling of* $p(\boldsymbol{x}_{mis} \mid \boldsymbol{x}_{obs})$ *is generally intractable since the posterior distribution* $p(\boldsymbol{z} \mid \boldsymbol{x}_{obs})$ *is not accessible* and hence approximations are required.

In the rest of the paper we assume that the generative model is such that computation of $p(\boldsymbol{x}_{\text{obs}} \mid \boldsymbol{z})$ and sampling of $p(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{z})$ is tractable. This is typically the case for most VAE architectures due to conditional independence assumptions (i.e. $x_j \perp\!\!\!\perp \boldsymbol{x}_{\setminus j} \mid \boldsymbol{z}$ for all $\forall j$) or the use of a Gaussian family for the decoder distribution $p(\boldsymbol{x} \mid \boldsymbol{z})$. Moreover, we assume that the encoder distribution, or the amortised variational posterior, $q(\boldsymbol{z} \mid \boldsymbol{x})$ (Gershman & Goodman, 2014) which approximates the model posterior $p(\boldsymbol{z} \mid \boldsymbol{x})$, is available.[2]

2.2 Pseudo-Gibbs (Rezende et al., 2014)

Rezende et al. (2014, Appendix F) have proposed a procedure related to Gibbs sampling (Geman & Geman, 1984), also called pseudo-Gibbs (Heckerman et al., 2000; Mattei & Frellsen, 2018), that due to its generality and simplicity has been regularly used for missing data imputation with VAEs (e.g. Rezende et al., 2014; Li et al., 2016; 2017; Rezende et al., 2018; Boquet et al., 2019). Starting with some random imputations $\boldsymbol{x}_{\text{mis}}^0$ the procedure iteratively samples latents $\boldsymbol{z}^t \sim q(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}^{t-1})$ and imputations $\boldsymbol{x}_{\text{mis}}^t \sim p(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{z}^t)$.[3] This iterative procedure generates a Markov chain that subject to some conditions on the closeness of the variational posterior $q(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$ and the intractable model posterior $p(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$ converges asymptotically in t to a distribution that approximately follows $p(\boldsymbol{x}_{\text{mis}}, \boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ (Rezende et al., 2014, Proposition F.1). The sampler corresponds to an exact Gibbs sampler if $q(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}) = p(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$.

However, the equality $q(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}) = p(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$ generally does not hold due to at least one of the following issues: insufficient flexibility of the variational distributional family, amortisation gap, or inference generalisation gap (Cremer et al., 2018; Zhang et al., 2021). Hence, pseudo-Gibbs sampling may produce sub-optimal samples even in the asymptotic limit or completely fail to converge due to an incompatibility of $q(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$ and $p(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{z})$.

[1] Equation (1) corresponds directly to missing data imputation with missing-at-random (MAR) missingness pattern.

[2] The variational posterior is typically available after fitting the VAE on complete data using standard variational Bayes (Rezende et al., 2014; Kingma et al., 2014), or can be fitted afterwards using a real or generated complete data set.

[3] Superscript t represents the sampler iteration.

2.3 Metropolis-within-Gibbs (Mattei & Frellsen, 2018)

Mattei & Frellsen (2018, Section 3.2) have proposed a simple modification of the pseudo-Gibbs sampler that can asymptotically in t generate exact samples from $p(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$. The method incorporates a Metropolis–Hastings accept-reject step (Metropolis et al., 1953; Hastings, 1970) to correct for the mismatch between $q(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$ and $p(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$ followed by sampling from $p(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{z})$, hence yielding a sampler in the Metropolis-within-Gibbs (MWG) family (Gelman & Rubin, 1992, Section 4.4). Specifically, at each iteration t it generates the proposal sample $\tilde{\boldsymbol{z}} \sim q(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}^{t-1})$ and accepts it as $\boldsymbol{z}^t = \tilde{\boldsymbol{z}}$ with probability

$$\rho^t(\tilde{\boldsymbol{z}}, \boldsymbol{z}^{t-1}; \boldsymbol{x}_{\text{mis}}^{t-1}) = \min\left\{1, \frac{p(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}^{t-1} \mid \tilde{\boldsymbol{z}})p(\tilde{\boldsymbol{z}})}{p(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}^{t-1} \mid \boldsymbol{z}^{t-1})p(\boldsymbol{z}^{t-1})} \frac{q(\boldsymbol{z}^{t-1} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}^{t-1})}{q(\tilde{\boldsymbol{z}} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}^{t-1})}\right\}. \tag{2}$$

If the proposal $\tilde{\boldsymbol{z}}$ is rejected, the latent sample from the previous iteration is used, so that $\boldsymbol{z}^t = \boldsymbol{z}^{t-1}$. Given $\boldsymbol{z}^t$, a new imputation $\boldsymbol{x}_{\text{mis}}^t$ is then sampled as in standard Gibbs sampling: $\boldsymbol{x}_{\text{mis}}^t \sim p(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{z}^t)$. By incorporating the Metropolis–Hastings acceptance step, the pseudo-Gibbs sampler is transformed into an asymptotically exact MCMC sampler with $p(\boldsymbol{x}_{\text{mis}}, \boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ as stationary distribution even if $q(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}) \neq p(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$.

Importantly, as noted by the authors, the asymptotic exactness of MWG comes, compared to the pseudo-Gibbs sampler, at little additional computational cost in each iteration: the quantities required for computing ρ^t are also computed in the pseudo-Gibbs sampler, except for the often cheap prior evaluations $p(\boldsymbol{z})$.

In summary, MWG has several desirable properties which make it an attractive choice for conditional sampling of VAEs: (i) it provides theoretical guarantees of convergence to the correct conditional distribution, (ii) it is simple to implement, and (iii) its per-iteration computational cost is relatively small, i.e. one standard evaluation of a VAE, and is comparable to the cost of pseudo-Gibbs. However, as we will see next, MWG is not free of important pitfalls.

3 Pitfalls of Gibbs-like samplers for VAEs

Although the Gibbs-like samplers from sections 2.2 and 2.3 are often used to conditionally sample from a VAE model, the structure of the latent space can cause poor non-asymptotic sampling behaviour. We here detail, in a form of three pitfalls, how this structure can affect the aforementioned samplers. While the reported pitfalls are related to the known limitations of the classical Gibbs (Geman & Geman, 1984) and Metropolis-within-Gibbs samplers (Gelman & Rubin, 1992), we here work out their significance in the context of VAEs. In fig. 1 we exemplify these pitfalls in an archetypical scenario using a synthetic 2-dimensional VAE model (for details about the model see appendix C.1).[4] The proposed methods in the following section, AC-MWG (section 4.1) and LAIR (section 4.2), provide remedies for the reported pitfalls.

Pitfall I. Strong relationship between the latents and the visibles can cause poor mixing. We often train VAEs to learn a structured latent space that captures the complexity of the data. This is typically achieved by using a decoder with a simple, often conditionally-independent, distribution. For example, to fit a binarised MNIST data set well with a Bernoulli decoder distribution $p(\boldsymbol{x} \mid \boldsymbol{z}) = \prod_d \text{Bernoulli}(x_d \mid \boldsymbol{z})$, the digits in the image space must be well-represented in the latent space and the variance of the decoder must be nearly 0, otherwise the model would produce noisy samples due to random "flips" of the pixels. Hence, in VAEs with simple decoders the complexity of modelling the visibles $\boldsymbol{x}$ is often converted to learning a complex structure in the latent space along with a near-deterministic mapping between the latents $\boldsymbol{z}$ and the visibles $\boldsymbol{x}$ as given by the decoder $p(\boldsymbol{x} \mid \boldsymbol{z})$. But this strong, near-deterministic, relationship can substantially inhibit the convergence and mixing properties of a sampler like Metropolis-within-Gibbs. This is because the proposed samples $\tilde{\boldsymbol{z}} \sim q(\boldsymbol{z})$ will be rejected with a high probability if the conditional distribution $p(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{z}}) \propto p(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}} \mid \tilde{\boldsymbol{z}})$ places little density/mass on the *previous* value of $\boldsymbol{x}_{\text{mis}} = \boldsymbol{x}_{\text{mis}}^{t-1}$, as a small value of $p(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}^{t-1} \mid \tilde{\boldsymbol{z}})$ will make the Metropolis–Hastings acceptance probability in eq. (2) small. This small acceptance probability leads to Markov chains that get "stuck" in a mode and prevents the sampler

[4] We note that the variational distribution $q(\boldsymbol{z} \mid \boldsymbol{x})$ in this section is constructed to be slightly wider than the model conditional $p(\boldsymbol{z} \mid \boldsymbol{x})$ to differentiate the different modes of failure.

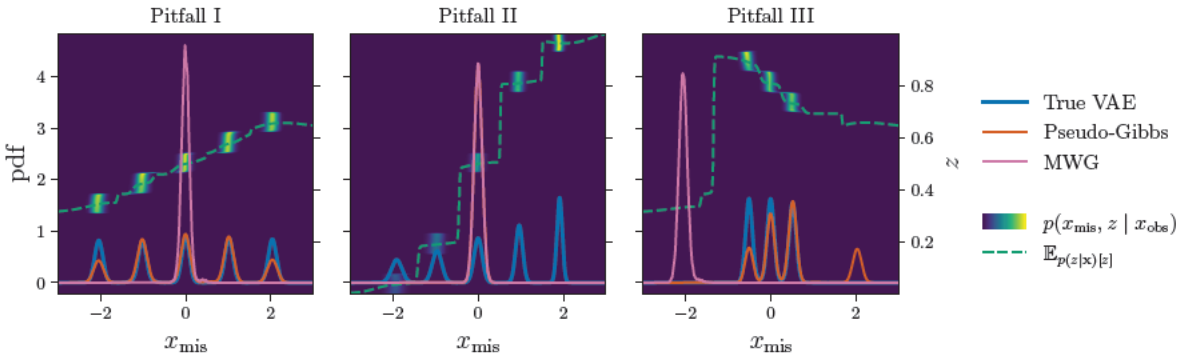


Figure 1: *Pitfalls of Gibbs-like samplers for VAE models.* (The figure is best viewed in colour.) Each panel corresponds to a distinct sampling problem, where the the observed variable $x_{\text{obs}} \in \{x_0, x_1\}$ is, from left to right, $x_1 = 0$, $x_0 = 0$, and $x_1 = 1$. The line plots show the ground truth density $p(x_{\text{mis}} \mid x_{\text{obs}})$ (blue) and the density of the samples obtained from the two Gibbs-like methods, pseudo-Gibbs (orange) and MWG (pink). The contour plot shows the conditional joint density $p(x_{\text{mis}}, z \mid x_{\text{obs}})$ of the VAE model over the missing variable x_{mis} (bottom axis) and the latent z (right axis), and the dashed green curve shows the expected value of z given x_0 and x_1. Both samplers were initialised with the same state and run for 50k iterations. *Left:* MWG fails to mix between nearby modes (in the space of z; right axis) due to high rejection probability in eq. (2). *Center:* both pseudo-Gibbs and MWG fail to find modes that are far apart (in the space of z; right axis) due to narrow proposal distribution. (We note that MWG and pseudo-Gibbs lines overlap in this plot.) *Right:* poor initialisation may leave MWG "stuck" far from the target distribution. Appendix D.1 contains an additional view of the pitfalls.

from moving to nearby modes that are close in the latent space. We illustrate this pitfall in fig. 1 (left). In this example, MWG (pink) fails to mix between the modes that are close in the space of latents. This failure occurs despite the proposal distribution generating samples from the neighbouring modes because such proposed samples are rejected by the Metropolis–Hastings step. On the other hand, pseudo-Gibbs (orange) can mix between the modes since it does not use the Metropolis–Hastings step.

Pitfall II. The encoder distribution generates proposals that are insufficiently exploratory. A further complication of the structured latent space is illustrated in fig. 1 (center). Here, the modes of the target distribution are sparsely dispersed in the latent space. In this example, we see that both MWG (pink) and pseudo-Gibbs (orange) fail to find distant modes. This is because the proposal distribution, as given by the encoder that approximates the model posterior $p(\boldsymbol{z} \mid \boldsymbol{x})$, is too "narrow" to propose values from the alternative modes. For example, given the upper-half of an MNIST image of number "8", it may not be possible to tell if the completed image should be an "8" or a "9", representing two modes of imputations. If the latent space representation of "8" and "9" are sufficiently far, then an encoder conditioned on a current imputation state, for example, $\boldsymbol{x}_{\text{obs}} \cup \boldsymbol{x}_{\text{mis}}^{t-1} \equiv$ "9", is unlikely to propose a $\tilde{\boldsymbol{z}}$ that would decode into $\tilde{\boldsymbol{x}}_{\text{mis}}$ in the alternative mode, that is, $\boldsymbol{x}_{\text{obs}} \cup \tilde{\boldsymbol{x}}_{\text{mis}} \equiv$ "8". On the other hand, even if the proposal distribution were wide enough to propose jumps to distant modes, MWG would still reject such proposals with high probability due to pitfall I and thus prevent effective exploration.

Pitfall III. Poor initialisation can cause sampling of the wrong mode. As noted by Mattei & Frellsen (2018) MWG for VAEs is extremely sensitive to initialisation, and to alleviate this they suggest initialising by first sampling using pseudo-Gibbs before switching to MWG. But, deciding when to stop the "warm-up" is not easy, and poor initialisation can make MWG get stuck. Moreover, initialisation via an (approximate) MAP using stochastic gradient ascent may also suffer from the multimodality issues described above. In fig. 1 (right) we demonstrate a case where MWG (pink) fails due to a poor initialisation.

40

The limitations of Gibbs-like samplers described in pitfalls I-III motivate our development of improved samplers. Interestingly, despite pseudo-Gibbs being theoretically inferior to MWG, we have seen in this section that pseudo-Gibbs can under some conditions perform better than MWG (fig. 1). In the following sections we propose two different methods that, like pseudo-Gibbs and MWG, utilise the encoder of the VAE to propose transitions in the latent space, whilst mitigating pitfalls I-III and having stronger theoretical guarantees than the simple pseudo-Gibbs method.

4 Remedies

The Metropolis-within-Gibbs (MWG) sampler for conditional sampling of VAEs has several desirable properties (see section 2.3). However, as discussed in the previous section, the Gibbs-like sampler can have poor non-asymptotic performance. In this section we propose two methods for conditional sampling of VAEs inspired by MWG that also mitigate its potential pitfalls (section 3). The key idea of the proposed methods is akin to ancestral sampling of eq. (1); first, the methods approximately sample the intractable posterior over the latents $p(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$, improve this approximation iteratively, and then sample from the decoder distribution $p(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{z})$ conditional on the produced latent samples. In section 4.1 we propose a few simple modifications to the MWG sampler and demonstrate on a synthetic example how this mitigates the pitfalls of MWG. In section 4.2 we propose an alternative method based on adaptive importance sampling and likewise demonstrate on a synthetic example how it mitigates the pitfalls of MWG. Detailed evaluation of the proposed methods is provided in section 5 and the code to reproduce the experiments is available at https://github.com/vsimkus/vae-conditional-sampling.

4.1 Adaptive collapsed-Metropolis-within-Gibbs

We propose several modifications to the MWG sampler from section 2.3 to mitigate the pitfalls outlined in section 3. The proposed sampler is summarised in algorithm 1.

First, to improve exploration and reduce the effects of poor initialisation (see pitfalls II and III) we introduce a prior–variational mixture proposal[5]

$$\tilde{q}_\epsilon(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}) = (1-\epsilon) q(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}) + \epsilon p(\boldsymbol{z}), \tag{3}$$

where $q(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$ is the variational encoder distribution, $p(\boldsymbol{z})$ is the prior distribution of the VAE, and $\epsilon \in (0, 1)$ is the probability to sample from the prior. Clearly this modification alone would not resolve the pitfalls of MWG, since proposals $\tilde{\boldsymbol{z}}$ sampled from the prior $p(\boldsymbol{z})$ would be rejected with high probability at the Metropolis–Hastings step due to disagreement with the current imputation $\boldsymbol{x}_{\text{mis}}^{t-1}$ in eq. (2).

Hence, we next propose changing the target distribution of the Metropolis–Hastings step from $p(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$ to $p(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$, such that a good proposal $\tilde{\boldsymbol{z}}$ would not be rejected due to a disagreement with an imputation $\boldsymbol{x}_{\text{mis}}^{t-1}$ (see pitfall I). The modified Metropolis–Hastings acceptance probability is defined as

$$\rho^t(\tilde{\boldsymbol{z}}^t, \boldsymbol{z}^{t-1}; \tilde{\boldsymbol{x}}_{\text{mis}}) = \min\left\{1, \frac{p(\boldsymbol{x}_{\text{obs}} \mid \tilde{\boldsymbol{z}}^t) p(\tilde{\boldsymbol{z}}^t)}{p(\boldsymbol{x}_{\text{obs}} \mid \boldsymbol{z}^{t-1}) p(\boldsymbol{z}^{t-1})} \frac{\tilde{q}_\epsilon(\boldsymbol{z}^{t-1} \mid \boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{x}}_{\text{mis}})}{\tilde{q}_\epsilon(\tilde{\boldsymbol{z}}^t \mid \boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{x}}_{\text{mis}})}\right\}. \tag{4}$$

Marginalising the missing variables $\boldsymbol{x}_{\text{mis}}$ out of the likelihood $p(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}} \mid \tilde{\boldsymbol{z}}^t)$ corresponds to reducing the conditioning (or collapsing) in Gibbs samplers which is a common approach to improve mixing and convergence (van Dyk & Park, 2008; van Dyk & Jiao, 2015). In our case, if the optimal proposal distribution $p(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ were known, the sampler would become a standard ancestral sampler and would be maximally efficient, i.e. it would draw an independent sample at each iteration. Moreover, rather than using the imputation $\boldsymbol{x}_{\text{mis}}^{t-1}$ from the previous iteration to condition the proposal distribution, as in MWG, we are here going to re-sample a random imputation $\tilde{\boldsymbol{x}}_{\text{mis}}$ from an available set of historical imputations $\mathcal{H}_{\text{mis}}^{t-1}$ that is updated adaptively with iterations t.

We now combine the proposed changes in eqs. (3) and (4) to introduce the algorithm called adaptive collapsed-Metropolis-within-Gibbs (AC-MWG), which can be seen as an instance of the class of adaptive independent

[5]Our mixture proposal is related to the small-world proposal of Guan et al. (2006), which has been shown to improve performance in complicated heterogeneous and multimodal distributions.

Published in Transactions on Machine Learning Research (11/2023)

Algorithm 1 Adaptive collapsed-Metropolis-within-Gibbs

Input: VAE model $p(\boldsymbol{x}, \boldsymbol{z})$, variational posterior $q(\boldsymbol{z} \mid \boldsymbol{x}_{\text{mis}}, \boldsymbol{x}_{\text{obs}})$, mixture prob. ϵ, and data-point $\boldsymbol{x}_{\text{obs}}$

$\mathcal{H}^0_{\text{mis}} = \varnothing$ ▷ Initialise imputation history

$(\boldsymbol{z}^0, \boldsymbol{x}^0_{\text{mis}}) \sim p(\boldsymbol{z})p(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{z})$ ▷ Sample the initial values

for $t = 1$ **to** T **do**

  $\tilde{\boldsymbol{x}}_{\text{mis}} \sim \text{Uniform}(\mathcal{H}^{t-1}_{\text{mis}})$ ▷ Choose random $\boldsymbol{x}_{\text{mis}}$ from the history

  $\tilde{\boldsymbol{z}} \sim \tilde{q}_\epsilon(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{x}}_{\text{mis}})$ ▷ Sample proposal value $\tilde{\boldsymbol{z}}$

  $\rho^t = \rho^t(\tilde{\boldsymbol{z}}, \boldsymbol{z}^{t-1}; \tilde{\boldsymbol{x}}_{\text{mis}})$ ▷ Calculate acceptance probability using eq. (4)

  if $u < \rho^t$, with $u \sim \text{Uniform}(0, 1)$ **then** ▷ Accept $\tilde{\boldsymbol{z}}$ with probability ρ^t

    $\boldsymbol{z}^t = \tilde{\boldsymbol{z}}$

    $\mathcal{H}^t_{\text{mis}} = \{\boldsymbol{x}^\tau_{\text{mis}}\}^{t-1}_{\tau=0}$

  else ▷ Reject $\tilde{\boldsymbol{z}}$ with probability ρ^t

    $\boldsymbol{z}^t = \boldsymbol{z}^{t-1}$

    $\mathcal{H}^t_{\text{mis}} = \mathcal{H}^{t-1}_{\text{mis}}$

  end if

  $\boldsymbol{x}^t_{\text{mis}} \sim p(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{z}^t)$ ▷ Sample $\boldsymbol{x}_{\text{mis}}$

end for

return $\{(\boldsymbol{x}^0_{\text{mis}}, \boldsymbol{z}^0), \ldots, (\boldsymbol{x}^T_{\text{mis}}, \boldsymbol{z}^T)\}$ ▷ Return all samples

Metropolis–Hastings algorithms (Holden et al., 2009). Assume we start with an initial latent state $\boldsymbol{z}^0$ and an imputation history $\mathcal{H}^0_{\text{mis}} = \{\hat{\boldsymbol{x}}^0_{\text{mis}}\}$, such that $\boldsymbol{z}^0$ and $\hat{\boldsymbol{x}}^0_{\text{mis}}$ are mutually independent (for example, $\boldsymbol{z}^0$ and $\hat{\boldsymbol{x}}^0_{\text{mis}}$ are generated via independent short runs of pseudo-Gibbs, see section 2.2, or LAIR, see section 4.2). Then a single iteration t of the sampler is as follows:

1. **Proposal sampling.** First, a historical sample $\tilde{\boldsymbol{x}}_{\text{mis}}$ is re-sampled uniformly at random from the available imputation history $\mathcal{H}^{t-1}_{\text{mis}}$.[6] We then use the proposal distribution from eq. (3) to sample a single proposal $\tilde{\boldsymbol{z}}$.

2. **Metropolis–Hastings acceptance.** The proposed sample $\tilde{\boldsymbol{z}}$ is then either accepted as $\boldsymbol{z}^t = \tilde{\boldsymbol{z}}$ with probability $\rho^t(\tilde{\boldsymbol{z}}, \boldsymbol{z}^{t-1}; \tilde{\boldsymbol{x}}_{\text{mis}})$ in eq. (4) or rejected leaving $\boldsymbol{z}^t = \boldsymbol{z}^{t-1}$.

3. **Imputation sampling.** The imputation $\boldsymbol{x}^t_{\text{mis}}$ is updated by sampling the conditional $\boldsymbol{x}^t_{\text{mis}} \sim p(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{z}^t)$.

4. **Adaptation (history update).** The available history $\mathcal{H}^t_{\text{mis}}$ is updated as follows: if a new $\tilde{\boldsymbol{z}}$ has been accepted then all imputations $\{\boldsymbol{x}^\tau_{\text{mis}}\}^{t-1}_{\tau=0}$ up to step $t-1$ are made available at the next iteration, i.e. $\mathcal{H}^t_{\text{mis}} = \{\boldsymbol{x}^\tau_{\text{mis}}\}^{t-1}_{\tau=0}$, otherwise it is left unchanged $\mathcal{H}^t_{\text{mis}} = \mathcal{H}^{t-1}_{\text{mis}}$.

Step 4 of the sampler constructs the available history $\mathcal{H}^{t-1}_{\text{mis}}$ for the next iteration such that it does not contain imputations that depend on the current state $\boldsymbol{z}^{t-1}$, which ensures that the proposed values $\tilde{\boldsymbol{z}}$ are independent of $\boldsymbol{z}^{t-1}$ and thus guarantees that the stationary distribution of the independent Metropolis–Hastings remains correct as the history $\mathcal{H}^{t-1}_{\text{mis}}$ changes (Roberts & Rosenthal, 2007; Holden et al., 2009). However, the dependence on the sample history $\mathcal{H}^{t-1}_{\text{mis}}$ makes AC-MWG non-Markovian, and hence convergence needs to be verified. Adapting proofs by Holden et al. (2009), we prove in appendix A that the Markov chain of AC-MWG correctly converges to the stationary distribution $p(\boldsymbol{z}, \boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ with probability arbitrarily close to 1 as the number of iterations T grows.

Finally, we note that the per-iteration computational cost of AC-MWG and MWG (section 2.3) are nearly the same. The differences are: re-sampling $\tilde{\boldsymbol{x}}_{\text{mis}}$ from the history $\mathcal{H}^{t-1}_{\text{mis}}$, which should be negligible compared to the cost of evaluating the model, and marginalising the missing variables from the likelihood $p(\boldsymbol{x}_{\text{obs}} \mid \boldsymbol{z}) = \int p(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}} \mid \boldsymbol{z})\, \mathrm{d}\boldsymbol{x}_{\text{mis}}$, which is often free if the standard conditional independence assumption holds.

[6] In this paper, we re-sample $\tilde{\boldsymbol{x}}_{\text{mis}}$ from all the past samples in the available history $\mathcal{H}^{t-1}_{\text{mis}}$, however other strategies might be devised to improve the computational and convergence properties of the algorithm (see e.g. Holden et al., 2009; Martino et al., 2018). For example, by using a shorter window of past samples instead of the full length of the history.

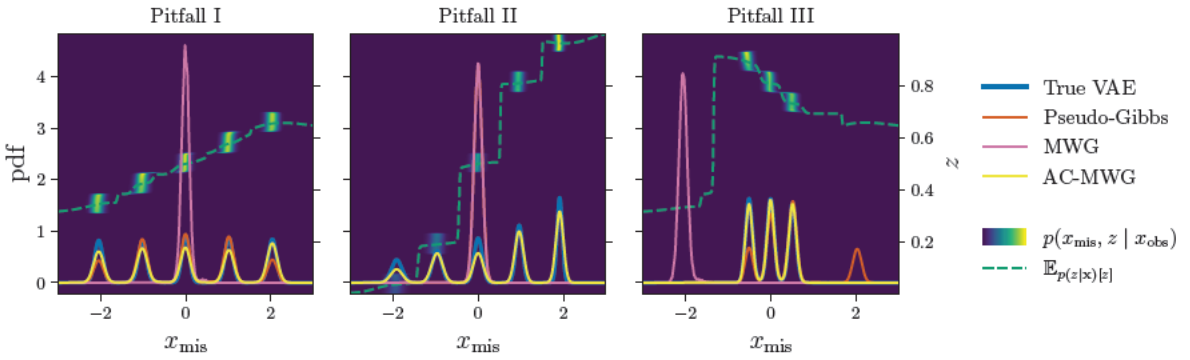


Figure 2: *The proposed AC-MWG sampler (yellow) with* $\epsilon = 0.01$ *on 2D VAE sampling problems, same as in fig. 1.* (The figure is best viewed in colour.) AC-MWG (yellow) samples the target distribution (blue) more accurately than MWG (pink) and pseudo-Gibbs (orange). All three samplers were initialised with the same state and run for 50k iterations.

4.1.1 Verification of AC-MWG on synthetic VAE

We verify the proposed AC-MWG method on the synthetic VAE example in section 3 (see additional details in appendix C.1). The results are shown in fig. 2 (see also additional figures in appendix D.1). With the proposed modifications, AC-MWG samples the target distribution more accurately by exploring modes that are close in the latent space (left) due to the modified acceptance probability in eq. (4), as well as distant modes (center) due to the modified proposal distribution in eq. (3). The modified method is also less sensitive to poor initialisation (right). Moreover, we perform ablation studies in appendices D.2 and D.4 to further validate that both modifications, the mixture proposal in eq. (3) and the collapsed-Gibbs target in eq. (4), are key to the performance of the method.

4.2 Latent-adaptive importance resampling

Instead of MCMC, we can sample from eq. (1) via importance resampling (IR, see appendix B for details on standard importance resampling and Chopin & Papaspiliopoulos, 2020, for a comprehensive introduction). However, like MCMC, the efficiency of IR significantly depends on the choice of the proposal distribution. Our goal in this section is to design an *adaptive* importance resampling method that efficiently samples $p(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ of a joint VAE model $p(\boldsymbol{x})$, and we achieve this by constructing an adaptive proposal distribution $q^t(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ using the encoder distribution $q(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$. The proposed method is summarised in algorithm 2.

As for AC-MWG, we aim to promote exploration and reduce the effects of poor initialisation (see pitfalls II and III). We thus start with the prior–variational mixture proposal $\tilde{q}_\epsilon(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$ from eq. (3) and use it to construct the following *adaptive* mixture proposal distribution $q^t(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$,

$$q^t(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}) = \mathbb{E}_{f^t(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})}\left[\tilde{q}_\epsilon(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})\right] \quad \text{with} \quad f^t(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}) = \frac{1}{K}\sum_{k=1}^{K} \delta_{\boldsymbol{x}_{\text{mis}}^{(t-1,k)}}(\boldsymbol{x}_{\text{mis}}),$$

where $f^t(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ is an imputation distribution represented as a mixture of Dirac masses at K particles $\{\boldsymbol{x}_{\text{mis}}^{(t-1,k)}\}_{k=1}^K$, which we will use to adapt the proposal distribution at each iteration t. We further rewrite the proposal by inserting the definition of $f^t(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ and $\tilde{q}_\epsilon(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$, and re-parametrise it by setting $\epsilon = \frac{R}{K+R}$, where R is a non-negative integer, to obtain

$$q^t(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}) = \frac{1}{K+R}\left(\sum_{k=1}^{K} q(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}^{(t-1,k)}) + \sum_{r=1}^{R} p(\boldsymbol{z})\right). \tag{5}$$

43

Published in Transactions on Machine Learning Research (11/2023)

Algorithm 2 Latent-adaptive importance resampling

Input: VAE model $p(\boldsymbol{x}, \boldsymbol{z})$, variational posterior $q(\boldsymbol{z} \mid \boldsymbol{x})$, data-point $\boldsymbol{x}_{\text{obs}}$, number of imputation particles K, number of iterations T

$\boldsymbol{x}_{\text{mis}}^{(0,1)}, \ldots, \boldsymbol{x}_{\text{mis}}^{(0,K)} \sim \mathbb{E}_{p(\boldsymbol{z})}\left[p(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{z})\right]$ ▷ Sample the initial imputation particle values

for $t = 1$ **to** T **do**

  $\tilde{\boldsymbol{z}}^{(t,k)} \sim q(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}^{(t-1,k)})$ for $\forall k \in \{1, \ldots, K\}$ ▷ Draw a sample for each particle.

  $\tilde{\boldsymbol{z}}^{(t,K+r)} \sim p(\boldsymbol{z})$ for $\forall r \in \{1, \ldots, R\}$ ▷ Draw R prior proposals.

  $w(\tilde{\boldsymbol{z}}^{(t,k)}) = \frac{p(\boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{z}}^{(t,k)})}{q^t(\tilde{\boldsymbol{z}}^{(t,k)} \mid \boldsymbol{x}_{\text{obs}})}$ for $\forall k \in \{1, \ldots, K+R\}$ ▷ Unnormalised importance weights.

  $\tilde{w}(\tilde{\boldsymbol{z}}^{(t,k)}) = \frac{w(\tilde{\boldsymbol{z}}^{(t,k)})}{\sum_{j=1}^{K+R} w(\tilde{\boldsymbol{z}}^{(t,j)})}$ for $\forall k \in \{1, \ldots, K+R\}$ ▷ Normalise importance weights.

  $\boldsymbol{z}^{(t,1)}, \ldots, \boldsymbol{z}^{(t,K)} \sim \text{Multinomial}\big(\{\tilde{\boldsymbol{z}}^{(t,k)}, \tilde{w}(\tilde{\boldsymbol{z}}^{(t,k)})\}_{k=1}^{K+R}\big)$ ▷ Resample $\boldsymbol{z}^{(t,k)}$ from the proposed set.

  $\boldsymbol{x}_{\text{mis}}^{(t,k)} \sim p(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{z}^{(t,k)})$ for $\forall k \in \{1, \ldots, K\}$. ▷ Update imputation particles.

end for

$\bar{w}(\tilde{\boldsymbol{z}}^{(t,k)}) = \frac{w(\tilde{\boldsymbol{z}}^{(t,k)})}{\sum_{\tau=1}^{T}\sum_{j=1}^{K+R} w(\tilde{\boldsymbol{z}}^{(\tau,j)})}$ for $\forall k \in \{1, \ldots, K+R\}$ and $\forall t \in \{1, \ldots, T\}$ ▷ Re-norm. all proposals.

$\boldsymbol{z}^i \sim \text{Multinomial}\Big(\{\tilde{\boldsymbol{z}}^{(t,k)}, \bar{w}(\tilde{\boldsymbol{z}}^{(t,k)})\}_{t=1,k=1}^{(T,K+R)}\Big)$ for $\forall i \in \{1, \ldots, T \cdot K\}$ ▷ Resample proposals from all iter.

$\boldsymbol{x}_{\text{mis}}^i \sim p(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{z}^i)$ for $\forall i \in \{1, \ldots, T \cdot K\}$. ▷ Sample imputations

return $\{\boldsymbol{x}_{\text{mis}}^i\}_{i=1}^{T \cdot K}$

The above proposal can be interpreted to have a total of $K+R$ components of which K components depend on the imputation particles $\{\boldsymbol{x}_{\text{mis}}^{(t-1,k)}\}_{k=1}^{K}$, which encourage exploitation, and R are "replenishing" prior components $p(\boldsymbol{z})$, which encourage exploration and mitigate particle collapse. Moreover, we sample the proposal distribution using stratified sampling (Robert & Casella, 2004; Owen, 2013, Section 9.12; Elvira et al., 2019, Appendix A), a well-known variance-reduction technique that draws one sample from each of the $K+R$ components.

Using the mixture proposal distribution in eq. (5) we now introduce the new algorithm that we call latent-adaptive importance resampling (LAIR), which belongs to the class of adaptive importance sampling algorithms (AIS) of Elvira & Martino (2022, Section 4). The algorithm starts with K imputation particles $\{\boldsymbol{x}_{\text{mis}}^{(0,k)}\}_{k=1}^{K}$ that may come from a simple distribution such as the empirical marginals, another multiple imputation method, or simply the unconditional marginal of the VAE $p(\boldsymbol{x}_{\text{mis}})$. An iteration t of the algorithm then performs the following three steps:

1. **Proposal sampling.** Sample the proposal distribution $q^t(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ in eq. (5) using stratified sampling. That is, for each particle $\boldsymbol{x}_{\text{mis}}^{(t-1,k)}$ draw a sample $\tilde{\boldsymbol{z}}^{(t,k)}$ from the proposal $q(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}^{(t-1,k)})$ and draw R proposals $\tilde{\boldsymbol{z}}^{(t,K+r)}$ from the prior $p(\boldsymbol{z})$, for a total of $K+R$ proposals.

2. **Weighting.** Compute the unnormalised importance weights $w(\tilde{\boldsymbol{z}}^{(t,k)})$.[7][8]

$$w(\tilde{\boldsymbol{z}}^{(t,k)}) = \frac{p(\boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{z}}^{(t,k)})}{q^t(\tilde{\boldsymbol{z}}^{(t,k)} \mid \boldsymbol{x}_{\text{obs}})}. \tag{6}$$

3. **Adaptation.**

   3.I. Resample a set $\{\boldsymbol{z}^{(t,k)}\}_{k=1}^{K}$ with replacement from the proposal set $\{\tilde{\boldsymbol{z}}^{(t,k)}\}_{k=1}^{K+R}$ proportionally to the weights $w(\tilde{\boldsymbol{z}}^{(t,k)})$.[9]

   3.II. Update the imputation particles $\{\boldsymbol{x}_{\text{mis}}^{(t,k)}\}_{k=1}^{K}$ by sampling $p(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{z})$ conditional on each $\boldsymbol{z} \in \{\boldsymbol{z}^{(t,k)}\}_{k=1}^{K}$ from step 3.I.

[7] We here use deterministic-mixture MIS (DM-MIS) weights but alternative weighting schemes can also be used that enable a more fine-grained control of the cost-variance trade-off, see Elvira et al. (2019, Section 7.2).

[8] By marginalising the variables $\boldsymbol{x}_{\text{mis}}$ in the numerator of the weights we address pitfall I, similar to eq. (4) of AC-MWG.

[9] Alternative resampling schemes may also be used, see Chopin & Papaspiliopoulos (2020, Section 9.4).

Each iteration t at step 3.II. (accordingly, line 8 of algorithm 2) produces (approximate) samples $\{\boldsymbol{x}_{\text{mis}}^{(t,k)}\}_{k=1}^{K}$ from the target distribution $p(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$, since any iteration t of the algorithm corresponds to standard importance resampling and hence inherits its properties (Cappé et al., 2004), see appendix B. In particular, the sampler monotonically approaches the target distribution as the number of proposed samples $K + R$ tends to infinity. Hence, the algorithm may be used in settings where the target distribution changes across iterations t, for instance, when fitting a model from incomplete data via a Monte Carlo EM (Wei & Tanner, 1990; Simkus et al., 2023).

However, unlike MCMC methods, a finite set of samples at any iteration t are not generally guaranteed to convergence to the target distribution as t grows large (Cappé et al., 2004; Douc et al., 2007). In particular, for finite sample sizes, $K + R \ll \infty$, the sampler bias is of the order $\mathcal{O}(\frac{1}{K+R})$ (Owen, 2013; Paananen et al., 2021) at any iteration t and depends on the disparity between the proposal and the target distributions. To improve the approximation, after the algorithm completes all T iterations, we can use samples from all iterations $t \in \{1, \dots, T\}$ to construct a more accurate estimator (Cappé et al., 2004).

4. **Draw final samples after completing all T iterations.**
   - 4.I. Re-normalise the weights of $\tilde{\boldsymbol{z}}^{(t,k)}$ over all iterations $t \in \{0, \dots, T\}$ and all $k \in \{1, \dots, K + R\}$ to obtain $\bar{w}(\tilde{\boldsymbol{z}}^{(t,k)}) = \frac{w(\tilde{\boldsymbol{z}}^{(t,k)})}{\sum_{\tau=1}^{T}\sum_{j=1}^{K+R} w(\tilde{\boldsymbol{z}}^{(\tau,j)})}$.
   - 4.II. Resample $T \cdot K$ samples $\boldsymbol{z}^i$ with replacement from the set $\{\tilde{\boldsymbol{z}}^{(t,k)}\}_{t=1,k=1}^{(T,K+R)}$ using the weights $\bar{w}(\tilde{\boldsymbol{z}}^{(t,k)})$ from the previous step.
   - 4.III. Sample imputations $\{\boldsymbol{x}_{\text{mis}}^{i}\}_{i=1}^{T\cdot K}$ via ancestral sampling by sampling $p(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{z})$ conditional on each $\boldsymbol{z} \in \{\boldsymbol{z}^i\}_{i=1}^{T\cdot K}$ from step 4.II.

The advantage of resampling from the re-weighted full sequence of samples is that the bias of the self-normalised importance sampler goes down with T (in addition to $K + R$) and hence more accurate samples can be obtained. In particular, the sampler now monotonically approaches the target distribution as the *total* number of proposed samples approaches infinity, $T(K + R) \to \infty$, and the bias is of the order $\mathcal{O}(\frac{1}{T(K+R)})$.

We note that the per-iteration computational cost of LAIR is comparable to running $K + R$ parallel chains of MWG, with the exception of: marginalising the missing variables from the likelihood, as in AC-MWG, which may often be cheap, and evaluating the denominator of the importance weights $w(\tilde{\boldsymbol{z}})$ in eq. (6), which requires that each of the $K + R$ proposed samples $\tilde{\boldsymbol{z}}$ must be evaluated with the densities of all $K + R$ components in the mixture proposal in eq. (5), hence needing $(K + R)^2$ evaluations. However, since the components of the proposal distribution in eq. (5) are typically all simple distributions, such as a diagonal Gaussians, the computational cost is often negligible for moderate number of proposals $K + R$. Moreover, the cost may be reduced by trading-off for a higher-variance of the estimator, see footnote 7. Finally, the computational cost of the final resampling in step 4 (accordingly, lines 10 to 12 in algorithm 2) is negligible since all the required quantities have already been computed in the past iterations.

4.2.1 Verification of LAIR on synthetic VAE

We now verify the proposed method, LAIR, on the synthetic VAE example in section 3 (see additional details in appendix C.1). The results are demonstrated in fig. 3 (see also additional figures in appendix D.1), where we have used $K = 19$ particles and $R = 1$ replenishing components (corresponding to $\epsilon = 0.05$). We can see that the method mitigates the three main pitfalls: poor mixing (left), poor exploration (center), and is less sensible to poor initialisation (right). Moreover, in ablation studies performed in appendices D.2 and D.4 we further investigate the sensitivity of the method to choices of $\epsilon = \frac{R}{K+R}$ and find that the method performs well as long as $0 < \epsilon < 1$.

5 Evaluation

In sections 4.1 and 4.2 we have introduced our methods, AC-MWG and LAIR, for conditional sampling of VAEs which mitigate the potential pitfalls of Gibbs-like samplers (section 3) as verified in sections 4.1.1

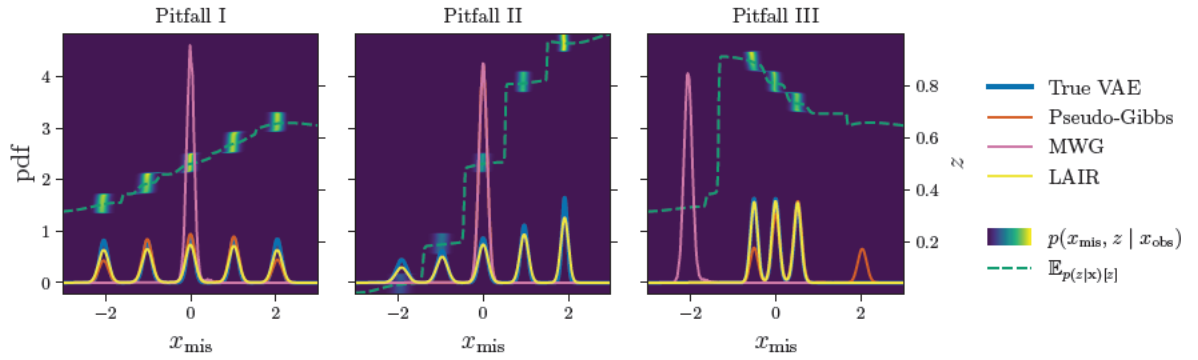


Figure 3: *The proposed LAIR sampler (yellow) with $K = 19$ particles and $R = 1$ replenishing components on 2D VAE sampling problems, same as in fig. 1.* (The figure is best viewed in colour.) LAIR (yellow) samples the target distribution (blue) more accurately than MWG (pink) and pseudo-Gibbs (orange). MWG and pseudo-Gibbs were run for 50k iterations, and LAIR was run for 2.5k iterations to match the number of generative model evaluations.

and 4.2.1. As motivated in section 2.1, conditional sampling is a fundamental tool for multiple imputation of missing data (Rubin, 1987; 1996), where the goal is to generate plausible values of the missing variables with correct uncertainty representation. We here evaluate the newly proposed methods for missing data imputation. We assume that we have a pre-trained VAE model, trained on complete data, and aim to generate imputations of the missing variables at test time.

5.1 Mixture-of-Gaussians MNIST

Evaluating the quality of imputations from data alone is a difficult task since the imputations represent guesses of unobserved values from an unknown conditional distribution (Abayomi et al., 2008; van Buuren, 2018, Section 2.5). Hence, to accurately evaluate the proposed methods, in this section we first fit a mixture-of-Gaussians (MoG) model to the MNIST data set, which we then use as the ground truth to simulate a semi-synthetic data set that is subsequently fitted by a VAE model (see appendix C.2 for more details). Using an intermediate MoG model enables us to tractably sample the reference conditional distribution (which would otherwise be unknown) when evaluating the accuracy of the conditional VAE samples obtained using the proposed and existing methods.

In fig. 4 we demonstrate the performance of the methods on 10 sampling problems (see appendix D.3 for additional figures and metrics). We measure the performance using the Fréchet inception distance (FID, Heusel et al., 2017), where for the inception features we use the final layer outputs of the encoder network. The figures show that the proposed methods, AC-MWG (pink) and LAIR (yellow), significantly outperform the performance of the Gibbs-like samplers from sections 2.2 and 2.3 (blue and green). In appendix D.4 we further perform an ablation study for the proposed methods, where: we validate that both the mixture proposal in eq. (3) and the collapsed-Gibbs target in eq. (4) are key to the good performance of AC-MWG; and we find that LAIR can perform well for a number of values of $\epsilon = \frac{R}{K+R}$, as long as $0 < \epsilon < 1$.

The results for MWG (green) use pseudo-Gibbs warm-up, as suggested by the authors Mattei & Frellsen (2018), to mitigate the effects of poor initialisation. We further investigated two different warm-up methods for MWG: an approximate MAP initialisation using stochastic gradient ascent on the log-likelihood, and LAIR. Both schemes improved over the base MWG but we found that the initialisation using LAIR generally performed better (see fig. 12 in the appendix). MWG with LAIR initialisation is denoted in fig. 4 as MWG′ (orange). We observe that with better initialisation the performance of MWG can be significantly improved, hence confirming the sensitivity of MWG to poor initialisation as discussed in section 3. However, with few exceptions MWG′ (orange) still generally performs worse than the proposed methods (pink and yellow),

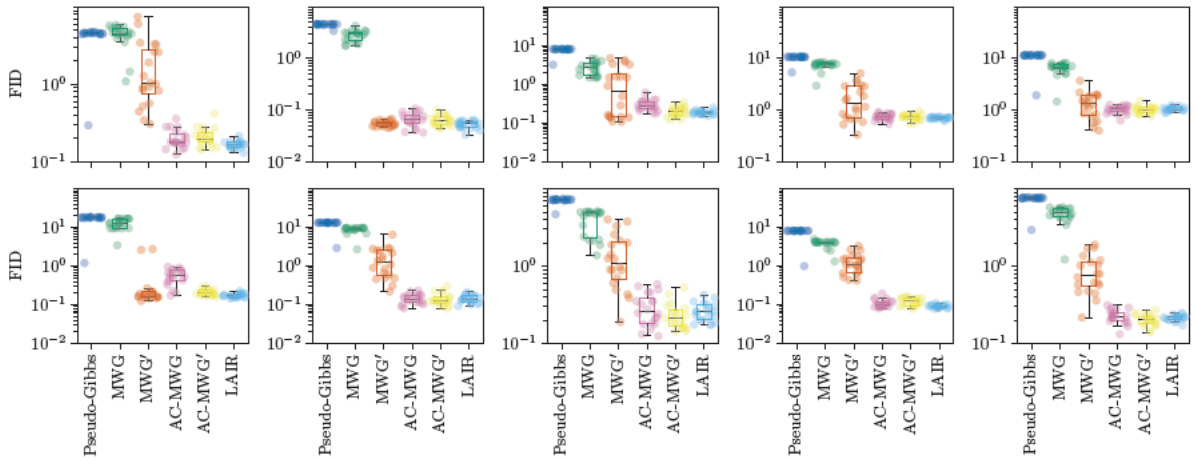


Figure 4: Fréchet inception distance (FID) between samples from the ground truth conditional $p(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$, and samples from the imputation methods. Each panel in the figure corresponds to a different conditional sampling problem $p(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$. Each evaluation is repeated 20 times, and the box-plot represents the inter-quartile range, including the median, and the whiskers show the overall range of the results.

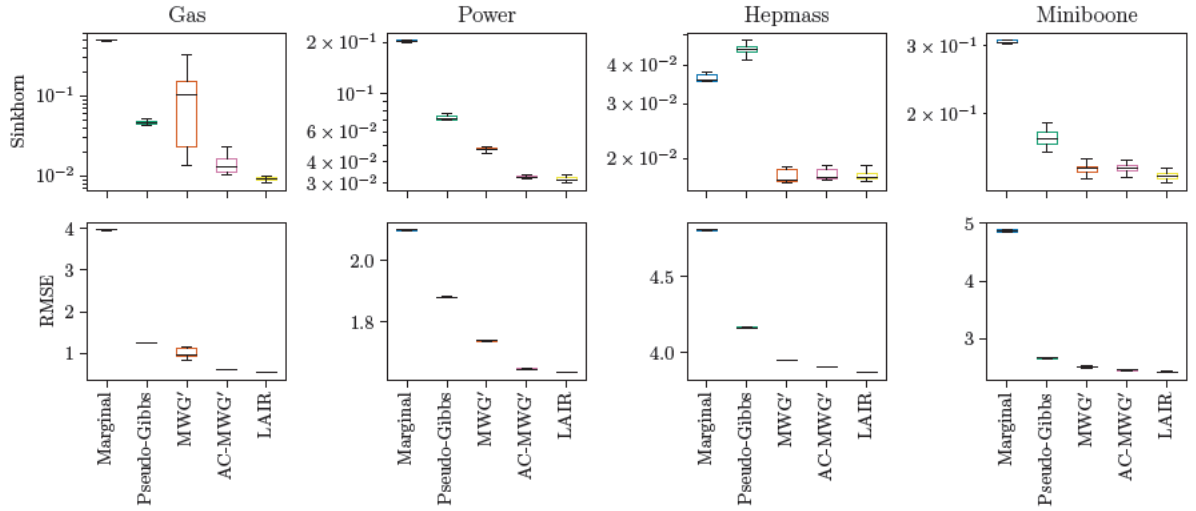


Figure 5: *Sampling performance on four real-world UCI data sets. Top:* Sinkhorn distance of the imputed data sets evaluated on a 50k data-point subset of test data (except for Miniboone where the full test data set was used). *Bottom:* Average RMSE of the imputations on the whole test data set. In both rows imputations from the final iteration of each algorithm are used and uncertainty is shown over different runs.

hence suggesting that the poor performance of MWG can be in part explained by the poor mixing of the sampler as discussed in section 3, that is addressed by the proposed methods.

5.2 Real-world UCI data sets

We now evaluate the proposed methods on real-world data sets from the UCI repository (Dua & Graff, 2017; Papamakarios et al., 2017). We train a VAE model with ResNet architecture on complete training data and evaluate the sampling accuracy of the existing and proposed methods on incomplete test data with 50%

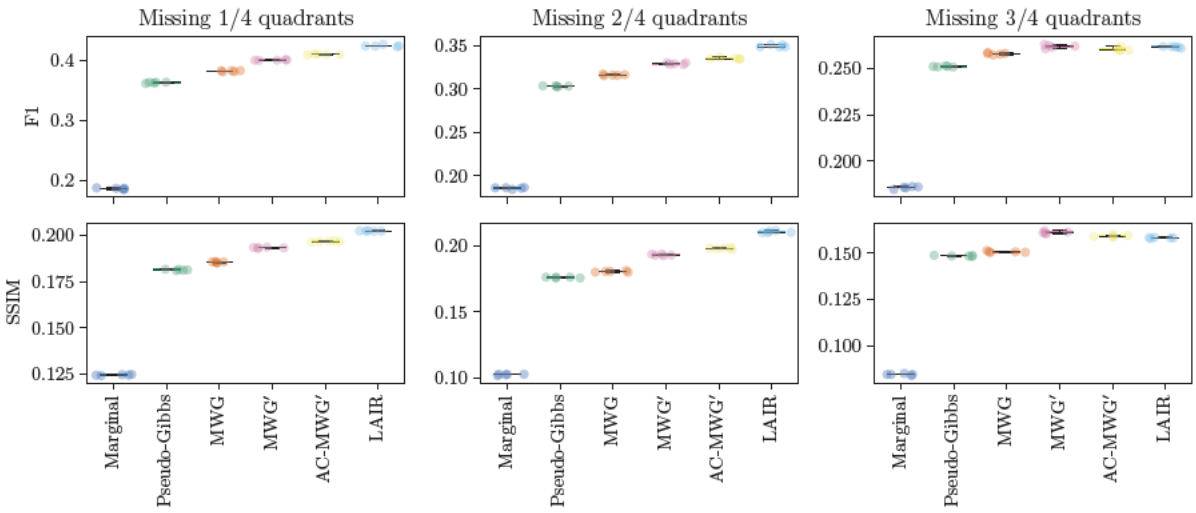


Figure 6: *Imputation accuracy on the binarised Omniglot test set with 1-3 randomly missing quadrants.* The top and bottom rows show F1 and average SSIM scores (higher is better for both metrics) respectively between the imputed and the ground truth values. In both rows imputations from the final iteration of each algorithm are used and uncertainty is shown over different runs.

missingness (see appendix C.3 for more details). We also include a simple baseline where imputations are sampled from the marginal distribution $p(\boldsymbol{x}_{\text{mis}})$ of the VAE. Moreover, in line with the observations from section 5.1 for MWG and AC-MWG we use LAIR initialisation as we have found it to considerably improve the performance of both methods. We here assess the performance using two metrics: Sinkhorn distance (Cuturi, 2013) between the imputed and ground truth data sets (computed using `geomloss` package by Feydy et al., 2019), and average RMSE of the imputations (for additional metrics, see appendix D.5).

The results are shown in fig. 5. First, the figure shows that all methods outperform marginal imputations (blue), with one exception of pseudo-Gibbs (green) on Hepmass data, where the Sinkhorn distance is slightly higher than the baseline. Second, as before, pseudo-Gibbs (green) is typically improved-upon by MWG (orange). The only exception is the Gas data with the Sinkhorn distance as metric (first row, first column) where the performance shows high variability. Other metrics (second row, first column, and appendix D.5) do not display this behaviour. Third, we see that the proposed methods, AC-MWG (pink) and LAIR (yellow), show better or comparable performance to the existing methods in terms of Sinkhorn distance (top row), and always improve on the existing methods in terms of the point-wise RMSE (bottom row). In summary, the results in this section match our findings from section 5.1, and hence further highlight the importance of mitigating the pitfalls in section 3 when dealing with real-world tasks.

5.3 Omniglot data set

In this section we evaluate the methods for conditional sampling of a VAE model trained on fully-observed binarised Omniglot data of handwritten characters (Lake et al., 2015). For the VAE model we use a convolutional ResNet encoder and decoder networks with 50 latent dimensions (see appendix C.4 for more details). We then evaluate the existing and proposed methods for conditional imputation of test set images that miss 1, 2, and 3 random quadrants. Similar to the previous section, we include a simple baseline where imputations are sampled from the marginal distribution $p(\boldsymbol{x}_{\text{mis}})$ of the VAE. The accuracy of the imputations on the binarised Omniglot is assessed using F1 score (Mattei & Frellsen, 2018) and structural similarity index measure (SSIM, Wang et al., 2004) between the ground truth and imputed values.

The results are shown in fig. 6. We first note that all conditional sampling methods perform better than marginal imputations (deep blue). Furthermore, we see that the metrics for the existing methods imply

the ranking pseudo-Gibbs (green) < MWG (orange) < MWG′ (pink), as before. Finally, we observe that the proposed methods, AC-MWG (yellow) and LAIR (light blue), further improve the accuracy of the imputations over the existing methods.

6 Discussion

Conditional sampling is a key challenge for downstream applications of VAEs and imprecise or inefficient samplers can cause unreliable results. We have examined the potential pitfalls of using Gibbs-like samplers, such as MWG, to conditionally sample from unconditional VAE models. While the outlined pitfalls are related to the well-known limitations of standard Gibbs sampler, we work out their significance in the context of VAEs. Pitfalls I and II outline two reasons for poor mixing of MWG: strong relationship between the latents $\boldsymbol{z}$ and visibles $\boldsymbol{x}$, and lack of exploration when the variational encoder distribution is used as proposal. Pitfall III highlights the importance of good initialisation for the performance of the sampler.

We introduced two samplers for conditional sampling of VAEs that address the pitfalls and show improved performance when compared to MWG and other baselines. The proposed methods, adaptive collapsed-Metropolis-within-Gibbs (AC-MWG) and latent-adaptive importance resampling (LAIR), mitigate pitfall I by marginalising the missing variables $\boldsymbol{x}_{\text{mis}}$ when (approximately) sampling the latents $\boldsymbol{z}$, and then sample the missing values $\boldsymbol{x}_{\text{mis}} \sim p(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{z})$. Therefore, in contrast to Gibbs sampling, the two methods can be seen as approximate ancestral sampling methods with asymptotic exactness guarantees. To mitigate pitfall II we have constructed proposal distributions from a mixture composed of the variational encoder distribution and the prior, which balances exploitation and exploration. Finally, we have found that poor initialisation (pitfall III) affects LAIR much less than the MCMC methods due to its ability to use information from multiple points in the latent space, and hence using LAIR to initialise MWG and AC-MWG can further improve their respective performances.

Depending on the task, computational budget, and accuracy requirements one may choose to use either AC-MWG or LAIR for conditional sampling of VAEs. For example, in tasks where the target distribution is changing between iterations, such as learning a VAE model from incomplete data (Simkus et al., 2023), LAIR could be more efficient than AC-MWG; this is because LAIR produces valid (although potentially biased) samples from the target distribution at any iteration, while AC-MWG requires a "burn-in" period until the sampler converges to the target distribution. On the other hand, on a strict computational budget AC-MWG might be preferred over LAIR: while the cost of AC-MWG is comparable to MWG (and hence also pseudo-Gibbs), each iteration of LAIR involves equivalent computations on $K + R$ particles and hence the computational cost and memory requirements is about $K + R$ times the cost of MWG. Finally, the convergence properties of the two methods are distinct: AC-MWG converges asymptotically in number of iterations, whereas the convergence in LAIR additionally scales in the number of particles $K + R$ and therefore parallelisation may be used to improve the speed of convergence at the cost of additional memory usage.

We have focused on conditional sampling of VAE models with moderate-dimensional latent spaces. To this end, we have addressed the "exploration–exploitation" dilemma by constructing the proposal distribution from the prior and variational encoder distributions. But, what works well in moderate dimensions might not work well in high dimensions, a direct consequence of the infamous "curse of dimensionality". This means that exploring the posterior by sampling the prior distribution might become impractical in higher dimensions. To scale the methods, alternative exploration strategies could be constructed by replacing the mixture proposal in eq. (3) with, for example, a mixture composed of annealed versions of the variational encoder distribution. Moreover, since the proposed methods belong to the large and general families of adaptive MCMC (Haario et al., 2001; Warnes, 2001; Roberts & Rosenthal, 2007; Holden et al., 2009; Liang et al., 2010) and adaptive importance sampling (AIS, Cappé et al., 2004; Bugallo et al., 2017), our work opens up additional opportunities to further improve the conditional sampling of VAEs.

Published in Transactions on Machine Learning Research (11/2023)

References

Kobi Abayomi, Andrew Gelman, and Marc Levy. Diagnostics for multivariate imputations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 57(3):273–291, 2008. ISSN 1467-9876. doi: 10.1111/j.1467-9876.2007.00613.x. (Cited on pg. 10)

David Barber. *Bayesian Reasoning and Machine Learning*. Cambridge University Press, 2017. ISBN 978-0-511-80477-9. doi: 10.1017/CBO9780511804779. (Cited on pg. 1, 19)

Guillem Boquet, Jose Lopez Vicario, Antoni Morell, and Javier Serrano. Missing Data in Traffic Estimation: A Variational Autoencoder Imputation Method. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2882–2886, May 2019. doi: 10.1109/ICASSP.2019.8683011. (Cited on pg. 2)

Monica F. Bugallo, Victor Elvira, Luca Martino, David Luengo, Joaquin Miguez, and Petar M. Djuric. Adaptive Importance Sampling: The past, the present, and the future. *IEEE Signal Processing Magazine*, 34(4):60–79, July 2017. ISSN 1558-0792. doi: 10.1109/MSP.2017.2699226. (Cited on pg. 13)

Oliver Cappé, Arnaud Guillin, Jean-Michel Marin, and Christian P. Robert. Population Monte Carlo. *Journal of Computational and Graphical Statistics*, 13(4):907–929, December 2004. ISSN 1061-8600. doi: 10.1198/106186004X12803. (Cited on pg. 9, 13)

Rewon Child. Very Deep VAEs Generalize Autoregressive Models and Can Outperform Them on Images. In *International Conference on Learning Representations (ICLR)*, March 2021. (Cited on pg. 1)

Nicolas Chopin and Omiros Papaspiliopoulos. *An Introduction to Sequential Monte Carlo*. Springer Series in Statistics. Springer, 2020. (Cited on pg. 7, 8, 22)

Chris Cremer, Xuechen Li, and David Duvenaud. Inference Suboptimality in Variational Autoencoders. In *International Conference on Machine Learning (ICML)*, May 2018. (Cited on pg. 2)

Marco Cuturi. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2013. (Cited on pg. 12)

Randal Douc, Arnaud Guillin, Jean-Michel Marin, and Christian P. Robert. Convergence of Adaptive Mixtures of Importance Sampling Schemes. *The Annals of Statistics*, 35(1):420–448, 2007. ISSN 0090-5364. (Cited on pg. 9)

Dheeru Dua and Casey Graff. UCI Machine Learning Repository, 2017. (Cited on pg. 11, 24)

Víctor Elvira and Luca Martino. Advances in Importance Sampling, March 2022. (Cited on pg. 8)

Víctor Elvira, Luca Martino, David Luengo, and Mónica F. Bugallo. Generalized Multiple Importance Sampling. *Statistical Science*, 34(1), February 2019. ISSN 0883-4237. doi: 10.1214/18-STS668. (Cited on pg. 8)

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between Optimal Transport and MMD using Sinkhorn Divergences. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019. doi: 10.48550/arXiv.1810.08278. (Cited on pg. 12)

Andrew Gelman and Donald B. Rubin. Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, 7(4):457–472, November 1992. ISSN 0883-4237, 2168-8745. doi: 10.1214/ss/1177011136. (Cited on pg. 3)

Stuart Geman and Donald Geman. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741, November 1984. doi: 10.1109/TPAMI.1984.4767596. (Cited on pg. 2, 3)

Samuel J. Gershman and Noah D. Goodman. Amortized Inference in Probabilistic Reasoning. In *Annual Meeting of the Cognitive Science Society*, volume 36, 2014. (Cited on pg. 2)

Rafael Gómez-Bombarelli, Jennifer N. Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D. Hirzel, Ryan P. Adams, and Alán Aspuru-Guzik. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Central Science*, 4(2):268–276, February 2018. ISSN 2374-7943. doi: 10.1021/acscentsci.7b00572. (Cited on pg. 1)

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, Cambridge, MA, USA, 2016. (Cited on pg. 1)

Yongtao Guan, Roland Fleißner, Paul Joyce, and Stephen M. Krone. Markov Chain Monte Carlo in small worlds. *Statistics and Computing*, 16(2):193–202, June 2006. ISSN 1573-1375. doi: 10.1007/s11222-006-6966-6. (Cited on pg. 5)

Heikki Haario, Eero Saksman, and Johanna Tamminen. An Adaptive Metropolis Algorithm. *Bernoulli*, 7 (2):223–242, 2001. ISSN 1350-7265. doi: 10.2307/3318737. (Cited on pg. 13)

Kuan Han, Haiguang Wen, Junxing Shi, Kun-Han Lu, Yizhen Zhang, Di Fu, and Zhongming Liu. Variational autoencoder: An unsupervised model for encoding and decoding fMRI activity in visual cortex. *NeuroImage*, 198:125–136, September 2019. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2019.05.039. (Cited on pg. 1)

Wilfred Keith Hastings. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, 57(1):97–109, 1970. ISSN 0006-3444. doi: 10.2307/2334940. (Cited on pg. 1, 3)

David Heckerman, David Maxwell Chickering, Christopher Meek, Robert Rounthwaite, and Carl Kadie. Dependency Networks for Inference, Collaborative Filtering, and Data Visualization. *Journal of Machine Learning Research*, 1(Oct):49–75, 2000. (Cited on pg. 2)

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. (Cited on pg. 10)

Lars Holden. Convergence of Markov Chains in the Relative Supremum Norm. *Journal of Applied Probability*, 37(4):1074–1083, 2000. ISSN 0021-9002. (Cited on pg. 22)

Lars Holden, Ragnar Hauge, and Marit Holden. Adaptive independent Metropolis–Hastings. *The Annals of Applied Probability*, 19(1):395–413, February 2009. ISSN 1050-5164, 2168-8737. doi: 10.1214/08-AAP545. (Cited on pg. 6, 13, 18, 20, 22)

Diederik P. Kingma and Jimmy Lei Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*, December 2014. (Cited on pg. 24, 25)

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations (ICLR)*, December 2013. (Cited on pg. 1)

Diederik P. Kingma, Danilo Jimenez Rezende, Shakir Mohamed, and Max Welling. Semi-Supervised Learning with Deep Generative Models. *Advances in Neural Information Processing Systems (NeurIPS)*, June 2014. (Cited on pg. 2)

Brenden M. Lake, Ruslan Salakhutdinov, and Joshua B. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 2015. doi: 10.1126/science.aab3050. (Cited on pg. 12, 25)

Chongxuan Li, Jun Zhu, and Bo Zhang. Learning to Generate with Memory. In *International Conference on Machine Learning (ICML)*, June 2016. (Cited on pg. 2)

Yingzhen Li, Richard E. Turner, and Qiang Liu. Approximate Inference with Amortised MCMC, May 2017. (Cited on pg. 2)

Faming Liang, Chuanhai Liu, and Raymond Carroll. *Advanced Markov Chain Monte Carlo Methods: Learning from Past Samples*. John Wiley & Sons, Incorporated, New York, 2010. ISBN 978-0-470-66973-0. (Cited on pg. 13, 20)

Luca Martino, Roberto Casarin, Fabrizio Leisen, and David Luengo. Adaptive independent sticky MCMC algorithms. *EURASIP Journal on Advances in Signal Processing*, 2018(1):5, January 2018. ISSN 1687-6180. doi: 10.1186/s13634-017-0524-6. (Cited on pg. 6)

Pierre-Alexandre Mattei and Jes Frellsen. Leveraging the Exact Likelihood of Deep Latent Variable Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, February 2018. (Cited on pg. 1, 2, 3, 4, 10, 12, 35)

Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6):1087–1092, June 1953. ISSN 0021-9606. doi: 10.1063/1.1699114. (Cited on pg. 1, 3)

Art B. Owen. *Monte Carlo Theory, Methods and Examples*. https://artowen.su.domains/mc/, 2013. (Cited on pg. 8, 9, 23)

Topi Paananen, Juho Piironen, Paul-Christian Bürkner, and Aki Vehtari. Implicitly adaptive importance sampling. *Statistics and Computing*, 31(2):16, March 2021. ISSN 0960-3174, 1573-1375. doi: 10.1007/s11222-020-09982-2. (Cited on pg. 9, 23)

George Papamakarios, Theo Pavlakou, and Iain Murray. Masked Autoregressive Flow for Density Estimation. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017. (Cited on pg. 11, 24)

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic Backpropagation and Approximate Inference. In *International Conference on Machine Learning (ICML)*, Beijing, China, 2014. (Cited on pg. 1, 2)

Danilo Jimenez Rezende, S. M. Ali Eslami, Shakir Mohamed, Peter Battaglia, Max Jaderberg, and Nicolas Heess. Unsupervised Learning of 3D Structure from Images. In *Advances in Neural Information Processing Systems (NeurIPS)*, June 2018. doi: 10.48550/arXiv.1607.00662. (Cited on pg. 2)

Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer, 2004. ISBN 0-387-21239-6. (Cited on pg. 8)

Gareth O. Roberts and Jeffrey S. Rosenthal. Coupling and Ergodicity of Adaptive Markov Chain Monte Carlo Algorithms. *Journal of Applied Probability*, 44(2):458–475, 2007. ISSN 0021-9002. (Cited on pg. 6, 13)

Geoffrey Roeder, Yuhuai Wu, and David K. Duvenaud. Sticking the Landing: Simple, Lower-Variance Gradient Estimators for Variational Inference. *Advances in Neural Information Processing Systems*, 30, 2017. (Cited on pg. 24, 25)

Donald B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, New York, 1987. ISBN 0-471-08705-X. doi: 10.2307/3172772. (Cited on pg. 2, 10)

Donald B. Rubin. Multiple Imputation After 18+ Years. *Journal of the American Statistical Association*, 91(434):473–489, 1996. ISSN 0162-1459. doi: 10.2307/2291635. (Cited on pg. 2, 10)

Vaidotas Simkus, Benjamin Rhodes, and Michael U. Gutmann. Variational Gibbs Inference for Statistical Model Estimation from Incomplete Data. *Journal of Machine Learning Research*, 24(196):1–72, 2023. ISSN 1533-7928. (Cited on pg. 9, 13)

Stef van Buuren. *Flexible Imputation of Missing Data*. CRC Press LLC, 2 edition, 2018. ISBN 978-1-138-58831-8. (Cited on pg. 10)

Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural Discrete Representation Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. (Cited on pg. 1)

52

David A. van Dyk and Xiyun Jiao. Metropolis-Hastings Within Partially Collapsed Gibbs Samplers. *Journal of Computational and Graphical Statistics*, 24(2):301–327, 2015. ISSN 1061-8600. (Cited on pg. 5)

David A. van Dyk and Taeyoung Park. Partially Collapsed Gibbs Samplers. *Journal of the American Statistical Association*, 103(482):790–796, June 2008. ISSN 0162-1459. doi: 10.1198/016214508000000409. (Cited on pg. 5)

Zhou Wang, Alan Conrad Bovik, Hamid Rahim Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, April 2004. ISSN 1941-0042. doi: 10.1109/TIP.2003.819861. (Cited on pg. 12)

Gregory R. Warnes. The Normal Kernel Coupler: An Adaptive Markov Chain Monte Carlo Method for Efficiently Sampling From Multi-Modal Distributions. Technical Report 39, University of Washington, March 2001. (Cited on pg. 13)

Greg C. G. Wei and Martin A. Tanner. A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms. *Journal of the American Statistical Association*, 85(411):699–704, September 1990. doi: 10.1080/01621459.1990.10474930. (Cited on pg. 9)

Mingtian Zhang, Peter Hayes, and David Barber. Generalization Gap in Amortized Inference. In *Workshop on Bayesian Deep Learning at Neural Information Processing Systems (NeurIPS)*, pp. 6, 2021. (Cited on pg. 2)

A AC-MWG proofs

Informally, showing convergence of MCMC samplers generally boils down to answering two questions: (i) does the Markov chain (asymptotically) reach the unique stationary distribution, and (ii) does the sampler remain in the stationary distribution after reaching it.[10]

First, we will focus on the latter question: does the AC-MWG sampler remain in the stationary distribution once it has been reached? Let p^t denote the distribution after t iterations, and $\pi(\boldsymbol{z}, \boldsymbol{x}_{\text{mis}}) = p(\boldsymbol{z}, \boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ denote the target distribution. The following theorem formalises the answer to the question.[11]

Theorem A.1. *The limiting distribution of the AC-MWG sampler conditioned on the history $\mathcal{H}_{mis}^{t-1}$ is invariant, that is $p^{t-1}(\boldsymbol{z}^{t-1}, \boldsymbol{x}_{mis}^{t-1} \mid \mathcal{H}_{mis}^{t-1}) = \pi(\boldsymbol{z}^{t-1}, \boldsymbol{x}_{mis}^{t-1})$ implies $p^t(\boldsymbol{z}^t, \boldsymbol{x}_{mis}^t \mid \mathcal{H}_{mis}^t) = \pi(\boldsymbol{z}^t, \boldsymbol{x}_{mis}^t)$.*

Proof. Let us denote $w^t = \mathcal{H}_{\text{mis}}^t \setminus \mathcal{H}_{\text{mis}}^{t-1}$ the new variables made available in the history $\mathcal{H}_{\text{mis}}^t$ after iteration t of the algorithm. Note that w^t is a random variable since it depends on the accept/reject decision in lines 9 and 12 of algorithm 1. In the proof we will show that by construction of the algorithm w^t and the new state $(\boldsymbol{z}^t, \boldsymbol{x}_{\text{mis}}^t)$ are independent, and that the statement in the theorem then follows.

Following algorithm 1 we now work out what are the new historical values of w^t at each iteration t. If a proposal $\tilde{\boldsymbol{z}}$ is *rejected* then line 12 of algorithm 1 corresponds to setting $w = \varnothing$. More generally, we allow adding to the history variables that depend on the rejected state $\tilde{\boldsymbol{z}}$ but not on the current state $\boldsymbol{z}^{t-1}$. If a proposal $\tilde{\boldsymbol{z}}$ is *accepted* then line 9 of algorithm 1 corresponds to setting w to be the set of imputations that were generated using the previous value of $\boldsymbol{z} = \boldsymbol{z}^{t-1}$. For instance, if new proposals were rejected for the last r iterations, then $\boldsymbol{z}^{t-1-r} = \boldsymbol{z}^{t-1-r+1} = \ldots = \boldsymbol{z}^{t-1}$, and hence $\boldsymbol{x}_{\text{mis}}^{t-1-r}, \boldsymbol{x}_{\text{mis}}^{t-1-r+1}, \ldots, \boldsymbol{x}_{\text{mis}}^{t-1}$ would all depend on $\boldsymbol{z}^{t-1}$, i.e. $\boldsymbol{x}_{\text{mis}}^{t-1-r}, \boldsymbol{x}_{\text{mis}}^{t-1-r+1}, \ldots, \boldsymbol{x}_{\text{mis}}^{t-1} \sim \pi(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{z}^{t-1})$. Thus, in the case of proposal acceptance, the variable w^t will contain the set of imputations $\{\boldsymbol{x}_{\text{mis}}^\tau\}_{\tau=t-1-r}^{t-1}$ that were drawn from $\pi(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{z}^{t-1})$ in the past iterations. We define the conditional distribution of w^t as $\pi(w^t \mid \hat{\boldsymbol{z}}) \overset{!}{=} \prod_{\tau=t-1-r}^{t-1} \pi(\boldsymbol{x}_{\text{mis}}^\tau \mid \hat{\boldsymbol{z}})$ where $\hat{\boldsymbol{z}}$ is $\boldsymbol{z}^{t-1}$ if a new proposal was accepted, or $\tilde{\boldsymbol{z}}$ if a proposal was rejected. This construction of the history ensures that the proposal distribution $\tilde{q}_\epsilon(\tilde{\boldsymbol{z}} \mid \boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{x}}_{\text{mis}})$ in lines 4 and 5 of algorithm 1 is *independent of the current* $\boldsymbol{z}^{t-1}$, and hence is a key ingredient to the proof.

We denote the transition kernel of AC-MWG as $k(\boldsymbol{z}^t, \boldsymbol{x}_{\text{mis}}^t, w^t \mid \boldsymbol{z}^{t-1}; \mathcal{H}_{\text{mis}}^{t-1})$.[12] The kernel, which depends on the history $\mathcal{H}_{\text{mis}}^{t-1}$, takes the current state of $\boldsymbol{z}^{t-1}$ and produces the new state $(\boldsymbol{z}^t, \boldsymbol{x}_{\text{mis}}^t)$ and the new historical variable w^t. We further use $f_{\mathcal{H}}^{t-1}(\tilde{\boldsymbol{x}}_{\text{mis}})$ to denote the probability of sampling a historical imputation $\tilde{\boldsymbol{x}}_{\text{mis}}$ from the available history $\mathcal{H}_{\text{mis}}^{t-1}$ in line 4 of algorithm 1. The kernel of AC-MWG is then defined as follows

$$
\begin{aligned}
&k(\boldsymbol{z}^t, \boldsymbol{x}_{\text{mis}}^t, w^t \mid \boldsymbol{z}^{t-1}; \mathcal{H}_{\text{mis}}^{t-1}) \\
&\quad = \pi(\boldsymbol{x}_{\text{mis}}^t \mid \boldsymbol{z}^t) \sum_{\tilde{\boldsymbol{x}}_{\text{mis}} \in \mathcal{H}_{\text{mis}}^t} f_{\mathcal{H}}^{t-1}(\tilde{\boldsymbol{x}}_{\text{mis}}) \int \Big(\tilde{q}_\epsilon(\tilde{\boldsymbol{z}} \mid \boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{x}}_{\text{mis}}) \rho_t(\tilde{\boldsymbol{z}}, \boldsymbol{z}^{t-1}; \tilde{\boldsymbol{x}}_{\text{mis}}) \delta(\boldsymbol{z}^t, \tilde{\boldsymbol{z}}) \pi(w^t \mid \boldsymbol{z}^{t-1}) \\
&\qquad\qquad + \tilde{q}_\epsilon(\tilde{\boldsymbol{z}} \mid \boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{x}}_{\text{mis}}) \left[1 - \rho_t(\tilde{\boldsymbol{z}}, \boldsymbol{z}^{t-1}; \tilde{\boldsymbol{x}}_{\text{mis}})\right] \delta(\boldsymbol{z}^t, \boldsymbol{z}^{t-1}) \pi(w^t \mid \tilde{\boldsymbol{z}}) \Big) \, \mathrm{d}\tilde{\boldsymbol{z}} \\
&\quad = \pi(\boldsymbol{x}_{\text{mis}}^t \mid \boldsymbol{z}^t) \sum_{\tilde{\boldsymbol{x}}_{\text{mis}} \in \mathcal{H}_{\text{mis}}^t} f_{\mathcal{H}}^{t-1}(\tilde{\boldsymbol{x}}_{\text{mis}}) \Big(\tilde{q}_\epsilon(\boldsymbol{z}^t \mid \boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{x}}_{\text{mis}}) \rho_t(\boldsymbol{z}^t, \boldsymbol{z}^{t-1}; \tilde{\boldsymbol{x}}_{\text{mis}}) \pi(w^t \mid \boldsymbol{z}^{t-1}) \\
&\qquad\qquad + \delta(\boldsymbol{z}^t, \boldsymbol{z}^{t-1}) \int \tilde{q}_\epsilon(\tilde{\boldsymbol{z}} \mid \boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{x}}_{\text{mis}}) \left[1 - \rho_t(\tilde{\boldsymbol{z}}, \boldsymbol{z}^{t-1}; \tilde{\boldsymbol{x}}_{\text{mis}})\right] \pi(w^t \mid \tilde{\boldsymbol{z}}) \, \mathrm{d}\tilde{\boldsymbol{z}} \Big)
\end{aligned}
$$

[10] The proofs in this section will consider a single observed data-point $\boldsymbol{x}_{\text{obs}}$, and hence nearly all quantities would depend on it. To ease the notation we will therefore suppress the conditioning on $\boldsymbol{x}_{\text{obs}}$ in all quantities, except for the proposal distribution $\tilde{q}_\epsilon$ in eq. (3) to keep it consistent with algorithm 1.

[11] The theorem is analogous to Theorem 1 by Holden et al. (2009) but we extend their proof to the component-wise setting of AC-MWG that involves an additional sampling step $\boldsymbol{x}_{\text{mis}}^t \sim p(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{z}^t)$, and where the history is maintained on $\boldsymbol{x}_{\text{mis}}$.

[12] Note that the kernel does not depend on the current $\boldsymbol{x}_{\text{mis}}^{t-1}$, since the new state only depends on the new $\boldsymbol{z}^t$, i.e. $\boldsymbol{x}_{\text{mis}}^t \sim \pi(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{z}^t)$.

The term in the parentheses corresponds to the standard Metropolis–Hastings kernel (see e.g. Barber, 2017, Section 27.4.2) with the addition of w^t to denote the new variables to be appended to the history at iteration t.

Assuming that at iteration $t-1$ the sampler is already at the stationary distribution $\pi(\boldsymbol{z}^{t-1}, \boldsymbol{x}_{\text{mis}}^{t-1})$, we now integrate the kernel with respect to the distribution of the current state $(\boldsymbol{z}^{t-1}, \boldsymbol{x}_{\text{mis}}^{t-1})$ to obtain the marginal over $\boldsymbol{z}^t$, $\boldsymbol{x}_{\text{mis}}^t$, and w^t

$$p^t(\boldsymbol{z}^t,\boldsymbol{x}_{\text{mis}}^t, w^t \mid \mathcal{H}_{\text{mis}}^{t-1}) = \int k(\boldsymbol{z}^t, \boldsymbol{x}_{\text{mis}}^t, w^t \mid \boldsymbol{z}^{t-1}; \mathcal{H}_{\text{mis}}^{t-1}) \pi(\boldsymbol{z}^{t-1}, \boldsymbol{x}_{\text{mis}}^{t-1}) \,\mathrm{d}\boldsymbol{z}^{t-1} \,\mathrm{d}\boldsymbol{x}_{\text{mis}}^{t-1}$$

Marginalising the $\boldsymbol{x}_{\text{mis}}^{t-1}$

$$= \int k(\boldsymbol{z}^t, \boldsymbol{x}_{\text{mis}}^t, w^t \mid \boldsymbol{z}^{t-1}; \mathcal{H}_{\text{mis}}^{t-1}) \pi(\boldsymbol{z}^{t-1}) \,\mathrm{d}\boldsymbol{z}^{t-1}$$

Inserting the definition of the kernel k and pushing the integral w.r.t. $\boldsymbol{z}^{t-1}$ inside the sum over $\tilde{\boldsymbol{x}}_{\text{mis}}$

$$\begin{aligned}
&= \pi(\boldsymbol{x}_{\text{mis}}^t \mid \boldsymbol{z}^t) \sum_{\tilde{\boldsymbol{x}}_{\text{mis}} \in \mathcal{H}_{\text{mis}}^t} f_{\mathcal{H}}^{t-1}(\tilde{\boldsymbol{x}}_{\text{mis}}) \Bigg(\\
&\quad \int \tilde{q}_\epsilon(\boldsymbol{z}^t \mid \boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{x}}_{\text{mis}}) \rho_t(\boldsymbol{z}^t, \boldsymbol{z}^{t-1}; \tilde{\boldsymbol{x}}_{\text{mis}}) \pi(w^t \mid \boldsymbol{z}^{t-1}) \pi(\boldsymbol{z}^{t-1}) \,\mathrm{d}\boldsymbol{z}^{t-1} \\
&\quad + \int \delta(\boldsymbol{z}^t, \boldsymbol{z}^{t-1}) \int \tilde{q}_\epsilon(\tilde{\boldsymbol{z}} \mid \boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{x}}_{\text{mis}}) \left[1 - \rho_t(\tilde{\boldsymbol{z}}, \boldsymbol{z}^{t-1}; \tilde{\boldsymbol{x}}_{\text{mis}})\right] \pi(w^t \mid \tilde{\boldsymbol{z}}) \,\mathrm{d}\tilde{\boldsymbol{z}} \pi(\boldsymbol{z}^{t-1}) \,\mathrm{d}\boldsymbol{z}^{t-1} \Bigg)
\end{aligned}$$

Marginalising the $\boldsymbol{z}^{t-1}$ in the second integral

$$\begin{aligned}
&= \pi(\boldsymbol{x}_{\text{mis}}^t \mid \boldsymbol{z}^t) \sum_{\tilde{\boldsymbol{x}}_{\text{mis}} \in \mathcal{H}_{\text{mis}}^t} f_{\mathcal{H}}^{t-1}(\tilde{\boldsymbol{x}}_{\text{mis}}) \Bigg(\\
&\quad \int \tilde{q}_\epsilon(\boldsymbol{z}^t \mid \boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{x}}_{\text{mis}}) \rho_t(\boldsymbol{z}^t, \boldsymbol{z}^{t-1}; \tilde{\boldsymbol{x}}_{\text{mis}}) \pi(w^t \mid \boldsymbol{z}^{t-1}) \pi(\boldsymbol{z}^{t-1}) \,\mathrm{d}\boldsymbol{z}^{t-1} \\
&\quad + \int \tilde{q}_\epsilon(\tilde{\boldsymbol{z}} \mid \boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{x}}_{\text{mis}}) \left[1 - \rho_t(\tilde{\boldsymbol{z}}, \boldsymbol{z}^t; \tilde{\boldsymbol{x}}_{\text{mis}})\right] \pi(w^t \mid \tilde{\boldsymbol{z}}) \pi(\boldsymbol{z}^t) \,\mathrm{d}\tilde{\boldsymbol{z}} \Bigg)
\end{aligned}$$

Expanding the second summand

$$\begin{aligned}
&= \pi(\boldsymbol{x}_{\text{mis}}^t \mid \boldsymbol{z}^t) \sum_{\tilde{\boldsymbol{x}}_{\text{mis}} \in \mathcal{H}_{\text{mis}}^t} f_{\mathcal{H}}^{t-1}(\tilde{\boldsymbol{x}}_{\text{mis}}) \Bigg(\\
&\quad \int \tilde{q}_\epsilon(\boldsymbol{z}^t \mid \boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{x}}_{\text{mis}}) \rho_t(\boldsymbol{z}^t, \boldsymbol{z}^{t-1}; \tilde{\boldsymbol{x}}_{\text{mis}}) \pi(w^t \mid \boldsymbol{z}^{t-1}) \pi(\boldsymbol{z}^{t-1}) \,\mathrm{d}\boldsymbol{z}^{t-1} \\
&\quad - \int \tilde{q}_\epsilon(\tilde{\boldsymbol{z}} \mid \boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{x}}_{\text{mis}}) \rho_t(\tilde{\boldsymbol{z}}, \boldsymbol{z}^t; \tilde{\boldsymbol{x}}_{\text{mis}}) \pi(w^t \mid \tilde{\boldsymbol{z}}) \pi(\boldsymbol{z}^t) \,\mathrm{d}\tilde{\boldsymbol{z}} \\
&\quad + \pi(\boldsymbol{z}^t) \int \tilde{q}_\epsilon(\tilde{\boldsymbol{z}} \mid \boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{x}}_{\text{mis}}) \pi(w^t \mid \tilde{\boldsymbol{z}}) \,\mathrm{d}\tilde{\boldsymbol{z}} \Bigg)
\end{aligned}$$

Using detailed balance $\tilde{q}_\epsilon(\tilde{\boldsymbol{z}} \mid \boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{x}}_{\text{mis}}) \rho_t(\tilde{\boldsymbol{z}}, \boldsymbol{z}^t; \tilde{\boldsymbol{x}}_{\text{mis}}) \pi(\boldsymbol{z}^t) = \tilde{q}_\epsilon(\boldsymbol{z}^t \mid \boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{x}}_{\text{mis}}) \rho_t(\boldsymbol{z}^t, \tilde{\boldsymbol{z}}; \tilde{\boldsymbol{x}}_{\text{mis}}) \pi(\tilde{\boldsymbol{z}})$ on the second summand above to obtain two identical integrals that cancel

$$\begin{aligned}
&= \pi(\boldsymbol{x}_{\text{mis}}^t \mid \boldsymbol{z}^t) \sum_{\tilde{\boldsymbol{x}}_{\text{mis}} \in \mathcal{H}_{\text{mis}}^t} f_{\mathcal{H}}^{t-1}(\tilde{\boldsymbol{x}}_{\text{mis}}) \Bigg(\\
&\quad \int \tilde{q}_\epsilon(\boldsymbol{z}^t \mid \boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{x}}_{\text{mis}}) \rho_t(\boldsymbol{z}^t, \boldsymbol{z}^{t-1}; \tilde{\boldsymbol{x}}_{\text{mis}}) \pi(w^t \mid \boldsymbol{z}^{t-1}) \pi(\boldsymbol{z}^{t-1}) \,\mathrm{d}\boldsymbol{z}^{t-1} \\
&\quad - \int \tilde{q}_\epsilon(\boldsymbol{z}^t \mid \boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{x}}_{\text{mis}}) \rho_t(\boldsymbol{z}^t, \tilde{\boldsymbol{z}}; \tilde{\boldsymbol{x}}_{\text{mis}}) \pi(w^t \mid \tilde{\boldsymbol{z}}) \pi(\tilde{\boldsymbol{z}}) \,\mathrm{d}\tilde{\boldsymbol{z}} \\
&\quad + \pi(\boldsymbol{z}^t) \int \tilde{q}_\epsilon(\tilde{\boldsymbol{z}} \mid \boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{x}}_{\text{mis}}) \pi(w^t \mid \tilde{\boldsymbol{z}}) \,\mathrm{d}\tilde{\boldsymbol{z}} \Bigg)
\end{aligned}$$

Cancelling the integral terms and rearranging we obtain the marginal distribution

$$p^t(\boldsymbol{z}^t, \boldsymbol{x}^t_{\text{mis}}, w^t \mid \mathcal{H}^{t-1}_{\text{mis}}) = \pi(\boldsymbol{x}^t_{\text{mis}}, \boldsymbol{z}^t) \sum_{\tilde{\boldsymbol{x}}_{\text{mis}} \in \mathcal{H}^t_{\text{mis}}} f^{t-1}_{\mathcal{H}}(\tilde{\boldsymbol{x}}_{\text{mis}}) \int \tilde{q}_\epsilon(\tilde{\boldsymbol{z}} \mid \boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{x}}_{\text{mis}}) \pi(w^t \mid \tilde{\boldsymbol{z}}) \, \mathrm{d}\tilde{\boldsymbol{z}}.$$

Importantly, the factorisation shows that $(\boldsymbol{x}^t_{\text{mis}}, \boldsymbol{z}^t)$ and w^t are independent, and hence

$$p^t(w^t \mid \mathcal{H}^{t-1}_{\text{mis}}) = \int p^t(\boldsymbol{z}^t, \boldsymbol{x}^t_{\text{mis}}, w^t \mid \mathcal{H}^{t-1}_{\text{mis}}) \, \mathrm{d}\boldsymbol{z}^t \, \mathrm{d}\boldsymbol{x}^t_{\text{mis}} = \sum_{\tilde{\boldsymbol{x}}_{\text{mis}} \in \mathcal{H}^t_{\text{mis}}} f^{t-1}_{\mathcal{H}}(\tilde{\boldsymbol{x}}_{\text{mis}}) \int \tilde{q}_\epsilon(\tilde{\boldsymbol{z}} \mid \boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{x}}_{\text{mis}}) \pi(w^t \mid \tilde{\boldsymbol{z}}) \, \mathrm{d}\tilde{\boldsymbol{z}}.$$

Therefore it immediately follows that

$$p^t(\boldsymbol{z}^t, \boldsymbol{x}^t_{\text{mis}} \mid \mathcal{H}^t_{\text{mis}} = \mathcal{H}^{t-1}_{\text{mis}} \cup \{w^t\}) = \frac{p^t(\boldsymbol{z}^t, \boldsymbol{x}^t_{\text{mis}}, w^t \mid \mathcal{H}^{t-1}_{\text{mis}})}{p^t(w^t \mid \mathcal{H}^{t-1}_{\text{mis}})} = \pi(\boldsymbol{x}^t_{\text{mis}}, \boldsymbol{z}^t),$$

which validates that the algorithm remains in the stationary distribution once it has reached it. □

Given that the sampler remains in the stationary distribution as shown in the above proof, we now show that the sampler can reach it. As discussed in section 4.1 the AC-MWG sampler corresponds to an ancestral sampler, which draws samples from $p(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ using non-Markovian adaptive Metropolis–Hastings, and then draws from $p(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{z})$ to obtain joint samples $(\boldsymbol{z}, \boldsymbol{x}_{\text{mis}}) \sim p(\boldsymbol{z}, \boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$. Therefore, to prove that the sampler reaches the stationary distribution (question (i) from the start of the section) we only need to show that the Metropolis–Hastings sampler reaches $p(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$. Let $\pi(\boldsymbol{z}) = p(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ denote the target distribution, $a^t(\tilde{\boldsymbol{x}}^t_{\text{mis}}) \in [0, 1]$ a function that depends on the historical sample $\tilde{\boldsymbol{x}}^t_{\text{mis}} \sim f^{t-1}_{\mathcal{H}}(\tilde{\boldsymbol{x}}^{t-1}_{\text{mis}})$ re-sampled from the history $\mathcal{H}^{t-1}_{\text{mis}}$ in line 4 of algorithm 1 at iteration t, and $\tilde{\boldsymbol{X}}^t_{\text{mis}} = (\tilde{\boldsymbol{x}}^1_{\text{mis}}, \ldots, \tilde{\boldsymbol{x}}^t_{\text{mis}})$ which denotes all those $\tilde{\boldsymbol{x}}_{\text{mis}}$ drawn up to iteration t, whose distribution we denote with $p^t_{\mathcal{H}}(\tilde{\boldsymbol{X}}^t_{\text{mis}})$. We formalise the answer to question (i) in the following theorem.[13]

Theorem A.2. *If the likelihood of the model is bounded and the prior–variational mixture proposal in eq. (3) uses an $\epsilon > 0$, then there is a function $a^\tau(\tilde{\boldsymbol{x}}^\tau_{mis}) \in (0, 1]$ that satisfies the strong Doeblin condition*

$$\tilde{q}_\epsilon(\tilde{\boldsymbol{z}} \mid \boldsymbol{x}_{obs}, \tilde{\boldsymbol{x}}^\tau_{mis}) \geq a^\tau(\tilde{\boldsymbol{x}}^\tau_{mis}) \pi(\tilde{\boldsymbol{z}}), \qquad \text{for } \forall \tilde{\boldsymbol{z}} \text{ and } \forall \tilde{\boldsymbol{x}}^t_{mis}, \tag{7}$$

and the total variation distance is bounded from above

$$\|p^t(\boldsymbol{z}^t) - \pi(\boldsymbol{z}^t)\|_{\text{TV}} \leq \mathbb{E}_{p^t_{\mathcal{H}}(\tilde{\boldsymbol{X}}^t_{mis})} \left[\prod_{\tau=1}^{t} (1 - a^\tau(\tilde{\boldsymbol{x}}^\tau_{mis})) \right]. \tag{8}$$

Hence the algorithm samples the target distribution within a finite number of iterations with a probability arbitrarily close to 1.

Proof. The key observation for this proof is to note that, conditionally on the history $\mathcal{H}^{t-1}_{\text{mis}}$, each iteration t of the sampler corresponds to one iteration of a (generalised) rejection sampler (see e.g. Liang et al., 2010, Section 3.1.1). Let us denote $\alpha^t \in \{0, 1\}$ a Bernoulli random variable with probability distribution $p(\alpha^t \mid \tilde{\boldsymbol{z}}, \tilde{\boldsymbol{x}}^t_{\text{mis}}) = \mathcal{B}(\alpha^t; s^t(\tilde{\boldsymbol{z}}, \tilde{\boldsymbol{x}}^t_{\text{mis}}))$ that signifies acceptance or rejection of a proposal $\tilde{\boldsymbol{z}}$ with a success probability $s^t(\tilde{\boldsymbol{z}}, \tilde{\boldsymbol{x}}^t_{\text{mis}})$ of a rejection sampler. We obtain s^t by lower-bounding the MH acceptance probability ρ^t in eq. (4). We first rewrite the MH acceptance probability

$$\begin{aligned}
\rho^t(\tilde{\boldsymbol{z}}, \boldsymbol{z}^{t-1}; \tilde{\boldsymbol{x}}^t_{\text{mis}}) &= \min \left\{ 1, \frac{\pi(\tilde{\boldsymbol{z}})}{\pi(\boldsymbol{z}^{t-1})} \frac{\tilde{q}_\epsilon(\boldsymbol{z}^{t-1} \mid \boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{x}}^t_{\text{mis}})}{\tilde{q}_\epsilon(\tilde{\boldsymbol{z}} \mid \boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{x}}^t_{\text{mis}})} \right\} \\
&= \frac{\pi(\tilde{\boldsymbol{z}})}{\tilde{q}_\epsilon(\tilde{\boldsymbol{z}} \mid \boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{x}}^t_{\text{mis}})} \min \left\{ \frac{\tilde{q}_\epsilon(\tilde{\boldsymbol{z}} \mid \boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{x}}^t_{\text{mis}})}{\pi(\tilde{\boldsymbol{z}})}, \frac{\tilde{q}_\epsilon(\boldsymbol{z}^{t-1} \mid \boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{x}}^t_{\text{mis}})}{\pi(\boldsymbol{z}^{t-1})} \right\},
\end{aligned}$$

[13] Our theorem here is analogous to Theorem 2 by Holden et al. (2009) but we extend the proof to the case where the proposal distribution is sampled stochastically using the history.

Lower-bounding the second term above to get $a^t(\tilde{\boldsymbol{x}}^t_{\text{mis}})$

$$a^t(\tilde{\boldsymbol{x}}^t_{\text{mis}}) = \min_{\tilde{\boldsymbol{z}}} \min \left\{ \frac{\tilde{q}_\epsilon(\tilde{\boldsymbol{z}} \mid \boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{x}}^t_{\text{mis}})}{\pi(\tilde{\boldsymbol{z}})}, \frac{\tilde{q}_\epsilon(\boldsymbol{z}^{t-1} \mid \boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{x}}^t_{\text{mis}})}{\pi(\boldsymbol{z}^{t-1})} \right\} = \min_{\tilde{\boldsymbol{z}}} \frac{\tilde{q}_\epsilon(\tilde{\boldsymbol{z}} \mid \boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{x}}^t_{\text{mis}})}{\pi(\tilde{\boldsymbol{z}})},$$ [14]

We finally obtain a lower-bounded acceptance probability s^t to the MH acceptance probability ρ^t

$$\rho^t(\tilde{\boldsymbol{z}}, \boldsymbol{z}^{t-1}; \tilde{\boldsymbol{x}}^t_{\text{mis}}) \geq a^t(\tilde{\boldsymbol{x}}^t_{\text{mis}}) \frac{\pi(\tilde{\boldsymbol{z}})}{\tilde{q}_\epsilon(\tilde{\boldsymbol{z}} \mid \boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{x}}^t_{\text{mis}})} = s^t(\tilde{\boldsymbol{z}}, \tilde{\boldsymbol{x}}^t_{\text{mis}}).$$ [15]

We will now show that with a probability of at least $a^t(\tilde{\boldsymbol{x}}^t_{\text{mis}})$ the sampler can jump to the stationary distribution $\pi(\boldsymbol{z})$ at any iteration t.

The conditional distribution of accepted samples of a rejection sampler is

$$p(\tilde{\boldsymbol{z}} \mid \tilde{\boldsymbol{x}}^t_{\text{mis}}, \alpha^t = 1) = \frac{\tilde{q}_\epsilon(\tilde{\boldsymbol{z}} \mid \boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{x}}^t_{\text{mis}}) p(\alpha^t = 1 \mid \tilde{\boldsymbol{z}}, \tilde{\boldsymbol{x}}^t_{\text{mis}})}{p(\alpha^t = 1 \mid \tilde{\boldsymbol{x}}^t_{\text{mis}})} \propto \tilde{q}_\epsilon(\tilde{\boldsymbol{z}} \mid \boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{x}}^t_{\text{mis}}) p(\alpha^t = 1 \mid \tilde{\boldsymbol{z}}, \tilde{\boldsymbol{x}}^t_{\text{mis}}).$$

Inserting $p(\alpha^t = 1 \mid \tilde{\boldsymbol{z}}, \tilde{\boldsymbol{x}}^t_{\text{mis}}) = s^t(\tilde{\boldsymbol{z}}, \tilde{\boldsymbol{x}}^t_{\text{mis}})$ we obtain

$$p(\tilde{\boldsymbol{z}} \mid \tilde{\boldsymbol{x}}^t_{\text{mis}}, \alpha^t = 1) \propto \tilde{q}_\epsilon(\tilde{\boldsymbol{z}} \mid \boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{x}}^t_{\text{mis}}) a^t(\tilde{\boldsymbol{x}}^t_{\text{mis}}) \frac{\pi(\tilde{\boldsymbol{z}})}{\tilde{q}_\epsilon(\tilde{\boldsymbol{z}} \mid \boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{x}}^t_{\text{mis}})} = a^t(\tilde{\boldsymbol{x}}^t_{\text{mis}}) \pi(\tilde{\boldsymbol{z}})$$

$$p(\alpha^t = 1 \mid \tilde{\boldsymbol{x}}^t_{\text{mis}}) = \int \tilde{q}_\epsilon(\tilde{\boldsymbol{z}} \mid \boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{x}}^t_{\text{mis}}) p(\alpha^t = 1 \mid \tilde{\boldsymbol{z}}, \tilde{\boldsymbol{x}}^t_{\text{mis}}) \, \mathrm{d}\tilde{\boldsymbol{z}} = \int a^t(\tilde{\boldsymbol{x}}^t_{\text{mis}}) \pi(\tilde{\boldsymbol{z}}) \, \mathrm{d}\tilde{\boldsymbol{z}} = a^t(\tilde{\boldsymbol{x}}^t_{\text{mis}})$$

Hence it follows that the accepted samples follow the target distribution

$$p(\tilde{\boldsymbol{z}} \mid \tilde{\boldsymbol{x}}^t_{\text{mis}}, \alpha^t = 1) = \frac{a^t(\tilde{\boldsymbol{x}}^t_{\text{mis}}) \pi(\tilde{\boldsymbol{z}})}{\int a^t(\tilde{\boldsymbol{x}}^t_{\text{mis}}) \pi(\hat{\boldsymbol{z}}) \, \mathrm{d}\hat{\boldsymbol{z}}} = \pi(\tilde{\boldsymbol{z}}).$$

The analogy between AC-MWG and rejection sampling allows us to conclude that the conditional probability to jump to the stationary distribution at any iteration t is (at least) $a^t(\tilde{\boldsymbol{x}}^t_{\text{mis}})$. This conditional probability depends on the historical sample $\tilde{\boldsymbol{x}}_{\text{mis}} \sim p^{t-1}_{\mathcal{H}}(\tilde{\boldsymbol{x}}_{\text{mis}})$ but is independent of the current distribution of $\boldsymbol{z}^{t-1}$.

We can now show that the probability to be in the stationary distribution within a finite number of iterations t can be made arbitrarily close to 1. Let b^t be the probability that the sampler *does not* jump to the stationary distribution in t iterations

$$b^t(\tilde{\boldsymbol{X}}^t_{\text{mis}}) = \prod_{\tau=1}^{t} \left(1 - a^\tau(\tilde{\boldsymbol{x}}^\tau_{\text{mis}})\right),$$

where $\tilde{\boldsymbol{X}}^t_{\text{mis}} = (\tilde{\boldsymbol{x}}^1_{\text{mis}}, \ldots, \tilde{\boldsymbol{x}}^t_{\text{mis}})$, and let $p^t(\boldsymbol{z}^t \mid \tilde{\boldsymbol{X}}^t_{\text{mis}})$ denote the conditional distribution of $\boldsymbol{z}^t$ after t iterations

$$p^t(\boldsymbol{z}^t \mid \tilde{\boldsymbol{X}}^t_{\text{mis}}) = \pi(\boldsymbol{z}^t)(1 - b^t(\tilde{\boldsymbol{X}}^t_{\text{mis}})) + \nu^t(\boldsymbol{z}^t \mid \tilde{\boldsymbol{X}}^t_{\text{mis}}) b^t(\tilde{\boldsymbol{X}}^t_{\text{mis}}),$$

which can be seen as a mixture of the stationary distribution $\pi(\cdot)$ with probability $(1 - b^t(\tilde{\boldsymbol{X}}^t_{\text{mis}}))$ and non-stationary distribution $\nu^t(\cdot)$ with probability $b^t(\tilde{\boldsymbol{X}}^t_{\text{mis}})$. The marginal distribution at the t-th iteration is then

$$p^t(\boldsymbol{z}^t) = \int p^t(\boldsymbol{z}^t \mid \tilde{\boldsymbol{X}}^t_{\text{mis}}) p^t_{\mathcal{H}}(\tilde{\boldsymbol{X}}^t_{\text{mis}}) \, \mathrm{d}\tilde{\boldsymbol{X}}^t_{\text{mis}}$$

We now derive a bound on the total variation distance

$$\|p^t(\boldsymbol{z}^t) - \pi(\boldsymbol{z}^t)\|_{\text{TV}}$$

[14] Note that taking the min over $\tilde{\boldsymbol{z}}$ makes $a^t(\tilde{\boldsymbol{x}}^t_{\text{mis}})$ independent of the current state $\boldsymbol{z}^{t-1}$!

[15] Note that due to $\rho^t \in [0, 1]$ we also have that the Bernoulli success probability $s^t \in [0, 1]$.

$$= \int \left| \int p^t(\boldsymbol{z}^t \mid \tilde{\boldsymbol{X}}^t_{\text{mis}}) p^t_{\mathcal{H}}(\tilde{\boldsymbol{X}}^t_{\text{mis}}) \, \mathrm{d}\tilde{\boldsymbol{X}}^t_{\text{mis}} - \pi(\boldsymbol{z}^t) \right| \mathrm{d}\boldsymbol{z}^t$$

Inserting the definition of $p^t(\boldsymbol{z}^t \mid \tilde{\boldsymbol{X}}^t_{\text{mis}})$ and using linearity of expectation to take $\pi(\boldsymbol{z}^t)$ into the expectation over $\tilde{\boldsymbol{X}}^t_{\text{mis}}$

$$= \int \left| \int \left(\pi(\boldsymbol{z}^t)(1 - b^t(\tilde{\boldsymbol{X}}^t_{\text{mis}})) + \nu^t(\boldsymbol{z}^t \mid \tilde{\boldsymbol{X}}^t_{\text{mis}}) b^t(\tilde{\boldsymbol{X}}^t_{\text{mis}}) - \pi(\boldsymbol{z}^t) \right) p^t_{\mathcal{H}}(\tilde{\boldsymbol{X}}^t_{\text{mis}}) \, \mathrm{d}\tilde{\boldsymbol{X}}^t_{\text{mis}} \right| \mathrm{d}\boldsymbol{z}^t$$

Expanding $\pi(\boldsymbol{z}^t)(1 - b^t(\tilde{\boldsymbol{X}}^t_{\text{mis}}))$ and cancelling terms

$$= \int \left| \int \left(-\pi(\boldsymbol{z}^t) + \nu^t(\boldsymbol{z}^t \mid \tilde{\boldsymbol{X}}^t_{\text{mis}}) \right) b^t(\tilde{\boldsymbol{X}}^t_{\text{mis}}) p^t_{\mathcal{H}}(\tilde{\boldsymbol{X}}^t_{\text{mis}}) \, \mathrm{d}\tilde{\boldsymbol{X}}^t_{\text{mis}} \right| \mathrm{d}\boldsymbol{z}^t$$

Applying Jensen's inequality to the (convex) norm function

$$\leq \iint \left| -\pi(\boldsymbol{z}^t) + \nu^t(\boldsymbol{z}^t \mid \tilde{\boldsymbol{X}}^t_{\text{mis}}) \right| \mathrm{d}\boldsymbol{z}^t b^t(\tilde{\boldsymbol{X}}^t_{\text{mis}}) p^t_{\mathcal{H}}(\tilde{\boldsymbol{X}}^t_{\text{mis}}) \, \mathrm{d}\tilde{\boldsymbol{X}}^t_{\text{mis}}$$

Applying triangle inequality $\int |\nu(\boldsymbol{z}^t) - \pi(\boldsymbol{z}^t)| \, \mathrm{d}\boldsymbol{z}^t \leq \int |\nu(\boldsymbol{z}^t)| \, \mathrm{d}\boldsymbol{z}^t + \int |-\pi(\boldsymbol{z}^t)| \, \mathrm{d}\boldsymbol{z}^t = 2$

$$\leq 2 \int b^t(\tilde{\boldsymbol{X}}^t_{\text{mis}}) p^t_{\mathcal{H}}(\tilde{\boldsymbol{X}}^t_{\text{mis}}) \, \mathrm{d}\tilde{\boldsymbol{X}}^t_{\text{mis}} = 2\mathbb{E}_{p^t_{\mathcal{H}}(\tilde{\boldsymbol{X}}^t_{\text{mis}})} \left[\prod_{\tau=1}^{t} (1 - a^\tau(\tilde{\boldsymbol{x}}^\tau_{\text{mis}})) \right]$$

Hence, the algorithm converges almost everywhere if the product goes to zero with $t \to \infty$. Therefore, if $a^\tau(\tilde{\boldsymbol{x}}^\tau_{\text{mis}})) > 0$ infinitely often then the sampler samples the target distribution $\pi(\boldsymbol{z})$ with probability arbitrarily close to 1.

To complete the proof we now show that the strong Doeblin condition (Holden, 2000; Holden et al., 2009) in eq. (7) holds, which requires that there exists $a^t(\tilde{\boldsymbol{x}}^t_{\text{mis}}) > 0$ for all $\tilde{\boldsymbol{z}}$ and $\tilde{\boldsymbol{x}}^t_{\text{mis}}$. Informally, the condition requires that the proposal distribution has heavier tails than the target distribution. We rewrite the condition in eq. (7) in its equivalent form as follows

$$\frac{\pi(\tilde{\boldsymbol{z}})}{\tilde{q}_\epsilon(\tilde{\boldsymbol{z}} \mid \boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{x}}^t_{\text{mis}})} \leq \frac{1}{a^t(\tilde{\boldsymbol{x}}^t_{\text{mis}})}. \tag{9}$$

Inserting the definition of $\pi(\tilde{\boldsymbol{z}}) = p(\tilde{\boldsymbol{z}} \mid \boldsymbol{x}_{\text{obs}})$ and $\tilde{q}_\epsilon(\tilde{\boldsymbol{z}} \mid \boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{x}}^t_{\text{mis}})$ from eq. (3) to the left side we obtain

$$\begin{aligned}
\frac{\pi(\tilde{\boldsymbol{z}})}{\tilde{q}_\epsilon(\tilde{\boldsymbol{z}} \mid \boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{x}}^t_{\text{mis}})} &= \frac{p(\tilde{\boldsymbol{z}} \mid \boldsymbol{x}_{\text{obs}})}{(1-\epsilon) q(\tilde{\boldsymbol{z}} \mid \boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{x}}^t_{\text{mis}}) + \epsilon p(\tilde{\boldsymbol{z}})} \\
&= \frac{p(\tilde{\boldsymbol{z}}, \boldsymbol{x}_{\text{obs}})}{p(\boldsymbol{x}_{\text{obs}}) \left((1-\epsilon) q(\tilde{\boldsymbol{z}} \mid \boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{x}}^t_{\text{mis}}) + \epsilon p(\tilde{\boldsymbol{z}}) \right)} \\
&= \frac{p(\boldsymbol{x}_{\text{obs}} \mid \tilde{\boldsymbol{z}})}{p(\boldsymbol{x}_{\text{obs}}) \left((1-\epsilon) \frac{q(\tilde{\boldsymbol{z}} \mid \boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{x}}^t_{\text{mis}})}{p(\tilde{\boldsymbol{z}})} + \epsilon \right)} \\
&= \frac{p(\boldsymbol{x}_{\text{obs}} \mid \tilde{\boldsymbol{z}})}{p(\boldsymbol{x}_{\text{obs}})} \left((1-\epsilon) \frac{q(\tilde{\boldsymbol{z}} \mid \boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{x}}^t_{\text{mis}})}{p(\tilde{\boldsymbol{z}})} + \epsilon \right)^{-1}.
\end{aligned}$$

Hence the ratio is bounded if $\epsilon > 0$ and if the likelihood is bounded, which we can safely assume since this is already a necessary condition to well learn the model. Since the left hand side of eq. (9) is bounded it follows that $a^t(\tilde{\boldsymbol{x}}^t_{\text{mis}}) > 0$, which completes the proof.

□

B Background: Importance resampling

We can generate samples following eq. (1) by using importance resampling (IR, e.g., Chopin & Papaspiliopoulos, 2020) to (approximately) sample $p(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ and then sampling $p(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{z})$ as in standard ancestral

58

sampling. We start with the standard importance sampling formulation for approximating the marginal $p(\boldsymbol{x}_{\text{obs}})$:

$$p(\boldsymbol{x}_{\text{obs}}) = \int p(\boldsymbol{x}_{\text{obs}}, \boldsymbol{z})\, \mathrm{d}\boldsymbol{z} = \int q(\boldsymbol{z}) \frac{p(\boldsymbol{x}_{\text{obs}}, \boldsymbol{z})}{q(\boldsymbol{z})}\, \mathrm{d}\boldsymbol{z} = \mathbb{E}_{q(\boldsymbol{z})}\left[w(\boldsymbol{z})\right], \tag{10}$$

where $q(\boldsymbol{z})$ is a proposal distribution that is assumed easy to sample and evaluate, and $w(\boldsymbol{z}) = p(\boldsymbol{x}_{\text{obs}}, \boldsymbol{z})/q(\boldsymbol{z})$ are the (unnormalised) importance weights, which are also computationally tractable.

The importance weight function $w(\cdot)$ can then be used to re-weigh the samples from the proposal distribution $q(\boldsymbol{z})$ to follow the model posterior $p(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$. We denote $\bar{w}(\tilde{\boldsymbol{z}}) = w(\tilde{\boldsymbol{z}})/\mathbb{E}_{q(\tilde{\boldsymbol{z}})}\left[w(\tilde{\boldsymbol{z}})\right]$ to be the (self-)normalised importance weights, and show that samples from the proposal can be re-weighted to follow the target distribution

$$\pi(\boldsymbol{z}) = \mathbb{E}_{q(\tilde{\boldsymbol{z}})}\left[\bar{w}(\tilde{\boldsymbol{z}})\delta_{\tilde{\boldsymbol{z}}}(\boldsymbol{z})\right] = \mathbb{E}_{q(\tilde{\boldsymbol{z}})}\left[\frac{w(\tilde{\boldsymbol{z}})}{\mathbb{E}_{q(\tilde{\boldsymbol{z}})}\left[w(\tilde{\boldsymbol{z}})\right]}\delta_{\tilde{\boldsymbol{z}}}(\boldsymbol{z})\right] = \mathbb{E}_{q(\tilde{\boldsymbol{z}})}\left[\frac{\frac{p(\boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{z}})}{q(\tilde{\boldsymbol{z}})}}{p(\boldsymbol{x}_{\text{obs}})}\delta_{\tilde{\boldsymbol{z}}}(\boldsymbol{z})\right] = p(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}), \tag{11}$$

where $\delta_{\tilde{\boldsymbol{z}}}(\cdot)$ is the Dirac delta distribution centred at point $\tilde{\boldsymbol{z}}$.

In practice, self-normalised importance resampling is generally implemented in four steps:

1. Draw M samples from a proposal $\tilde{\boldsymbol{z}}^1, \ldots, \tilde{\boldsymbol{z}}^M \sim q(\boldsymbol{z})$.
2. Compute the (unnormalised) importance weights $w(\tilde{\boldsymbol{z}}^m) = \frac{p(\boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{z}}^m)}{q(\tilde{\boldsymbol{z}}^m)}$ for all $\forall m \in [1, M]$.
3. Self-normalise the weights $\bar{w}(\tilde{\boldsymbol{z}}^m) = \frac{w(\tilde{\boldsymbol{z}}^m)}{\sum_{l=1}^{M} w(\tilde{\boldsymbol{z}}^l)}$ for all $\forall m \in [1, M]$.
4. Resample $\boldsymbol{z}^m$ with replacement from the set $\{\tilde{\boldsymbol{z}}^m\}_{m=1}^M$ using the normalised probabilities $\bar{w}(\tilde{\boldsymbol{z}}^m)$.

Self-normalised importance sampling is consistent in the number M of proposed samples and hence samples $p(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ exactly as $M \to \infty$ but has a bias of the order of $\mathcal{O}(1/M)$ (Owen, 2013; Paananen et al., 2021). Samples $\boldsymbol{x}_{\text{mis}} \sim p(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ can then be obtained by sampling $p(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{z})$.

In standard importance sampling applications the proposal distribution $q(\boldsymbol{z})$ is traditionally chosen heuristically using the domain knowledge of the target distribution. However, in the context of VAEs specifying a good proposal can be difficult due to poor prior knowledge about the latent space of the model. Moreover, the efficiency of the sampler depends on the quality of the proposal distribution $q(\boldsymbol{z})$ and a poor proposal distribution can cause weight degeneracy in the non-asymptotic regime ($M \ll \infty$), where only a few of the proposed samples have non-zero weights, and hence poorly approximate the target distribution.

C Experiment details

In this appendix we provide additional details on the experiments.

C.1 Synthetic 2D VAE

To investigate and illustrate the pitfalls of MWG we constructed a simple synthetic VAE model that approximates mixture-of-Gaussians data, see fig. 7. The visibles $\boldsymbol{x}$ are 2-dimensional and parametrised with a diagonal Gaussian decoder $p(\boldsymbol{x} \mid z)$, the latents z are 1-dimensional with a uniform prior $p(z) = \text{Uniform}(0, 1)$, and the variational proposal $q(z \mid \boldsymbol{x})$ is a Beta distribution amortised with a neural network. The low-dimensional example lets us compute, via numerical integration, and visualise the conditional distributions $p(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$, $p(\boldsymbol{x}_{\text{mis}}, \boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$, and $p(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$. As demonstrated in the two right-most panels of fig. 7 mixing in the joint space of the missing variable and the latent (x_0, z) may be poor due to low probability valleys between the modes (third panel), but could be easier in the marginal space of z (last panel).

59

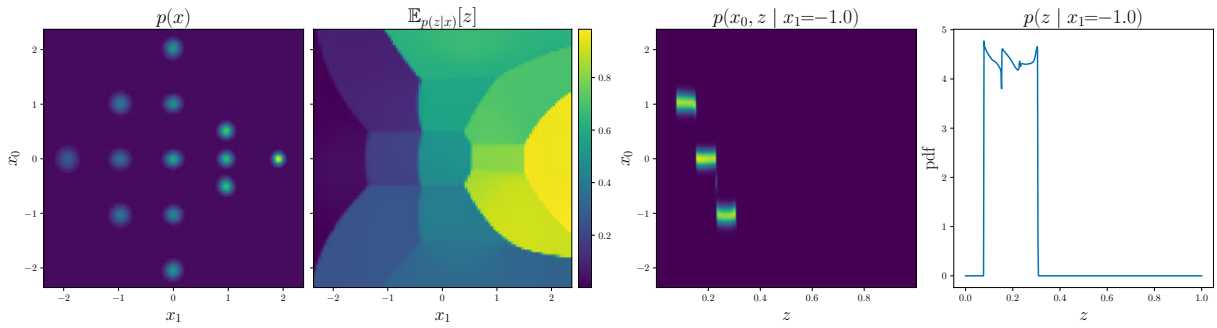


Figure 7: Left-to-right: the marginal distribution of the visibles $p(\boldsymbol{x})$ of the VAE; the posterior expected value of the latents z, i.e. $\mathbb{E}_{p(z|\boldsymbol{x})}[z]$; joint conditional distribution of x_0 and z for an observed x_1; conditional distribution of z for an observed x_1.

For pseudo-Gibbs and MWG in figs. 1 to 3 we perform a single run of each algorithm for 50k iterations, with both methods initialised at the same location using a sample drawn from the marginal distribution $p(\boldsymbol{x}_{\text{mis}})$ of the VAE. Similarly, in fig. 2 the proposed method AC-MWG performs a single run of the algorithm for 50k iterations with mixture coefficient $\epsilon = 0.01$, initialised at the same location as pseudo-Gibbs and MWG. Finally, in fig. 3 the proposed method LAIR performs a single run of the algorithm for 2.5k iterations using 19 imputation particles ($K = 19$) and 1 replenishing mixture component ($R = 1$), the algorithm is initialised with $K = 19$ samples from the marginal distribution $p(\boldsymbol{x}_{\text{mis}})$ of the VAE.

C.2 Mixture-of-Gaussians MNIST

We construct a mixture-of-Gaussians (MoG) ground truth model with 10 multivariate Gaussian components and uniform component probability $\pi(c) = \frac{1}{10}$. Each Gaussian component is fitted on all samples from the MNIST data set (downsampled to 14x14 and transformed with a logit transformation) with a particular label $c \in [1, 10]$. We then generated a semi-synthetic data set of 18k samples and fit a VAE model with a latent space dimensionality of 25. For the VAE, we have used a diagonal Gaussian decoder using ConvResNet architecture with 4 convolutional residual blocks of feature map depths of 128, 64, 32, and 32, and a dropout of 0.2. The prior distribution over the latents is a standard normal distribution. The variational distribution is parametrised with a diagonal Gaussian encoder using ConvResNet architecture with 4 convolutional residual blocks of feature map depths 32, 64, 128, and 256, and dropout of 0.2. To optimise the VAE model we have used the sticking-the-landing gradients (Roeder et al., 2017) and fit the model using batch size of 200 for 6000 epochs using Adam optimiser (Kingma & Ba, 2014) with a learning rate of 10^{-4}.

For pseudo-Gibbs we ran 5 independent chains for 10k iterations each, and to stabilise the sampler, the imputations were clipped to the minimum and maximum values of the data set for each dimension multiplied by 2. For MWG we have initialised 5 independent chains by running pseudo-Gibbs for 120 iterations with clipping and then run the MWG sampler for 9880 iterations on each chain. For MWG′ we have initialised 5 independent chains by running LAIR with K=4 and R=1 for 120 iterations for each chain, and then run the MWG sampler for 9880 iterations on each chain. For AC-MWG we have initialised 5 independent chains from the marginal distribution of the VAE and then run AC-MWG with $\epsilon = 0.05$ for 10k iterations. For AC-MWG′ we have initialised 5 independent chains by running LAIR with K=4 and R=1 for 120 iterations for each chain, and then run the AC-MWG sampler for 9880 iterations on each chain with $\epsilon = 0.05$. For LAIR we have initialised $K = 4$ particles from the marginal distribution of the VAE and then run the sampler with $K = 4$ and $R = 1$ for 10k iterations.

C.3 UCI data sets

We fit VAEs on four data sets from the UCI repository (Dua & Graff, 2017) with the preprocessing of (Papamakarios et al., 2017). For all models, the variational and the generator (decoder) distributions were

60

Published in Transactions on Machine Learning Research (11/2023)

fitted to be in the diagonal Gaussian family. For the encoder and decoder networks of the VAEs we fit MLP neural networks with residual block architecture using Adam optimiser (Kingma & Ba, 2014) with learning rate of 10^{-3} for a total of 200k stochastic gradient ascent steps (except for Miniboone where 22k steps were used) using batch size of 512 (except for Miniboone where batch size of 1024 was used), while using 8 Monte Carlo samples in each iteration to approximate the variational ELBO and sticking-the-landing gradients to reduce variance (Roeder et al., 2017). For Gas, Power, and Hepmass data the encoder and decoder networks used 2 residual blocks each with hidden dimensionality of 256, ReLU activation functions, and a latent space of 16. In addition, for Power data we add small Gaussian noise to each batch with a standard deviation of 0.001. For Miniboone data the encoder used 5 residual blocks with hidden dimensionality of 256 and decoder networks used 2 residual blocks with hidden dimensionality of 256, ReLU activation functions, a latent space of 32, and dropout of 0.5.

For pseudo-Gibbs we ran 5 independent chains for 3k iterations each, and to stabilise the sampler on Gas and Hepmass data sets imputations were clipped to the minimum and maximum values of the data set for each dimension multiplied by 2. For MWG we have initialised 5 independent chains by running LAIR with K=4 and R=1 for 100 iterations for each chain, and then run the MWG sampler for 2900 iterations on each chain. For AC-MWG we have initialised 5 independent chains by running LAIR with K=4 and R=1 for 100 iterations for each chain, and then run the AC-MWG sampler for 2900 iterations on each chain with $\epsilon = 0.3$. For LAIR we have initialised $K = 4$ particles from the marginal distribution of the VAE and then run the sampler with $K = 4$ and $R = 1$ for 3k iterations. Each method evaluations were repeated with 5 different seeds, and the uncertainty reported in the figures reflects the uncertainty over different runs.

C.4 Handwritten character Omniglot data set

We fit a VAE on a statically binarised Omniglot data set (Lake et al., 2015) downsampled to 28×28 pixels. We have used a fixed standard Gaussian prior distribution over the latents $p(\boldsymbol{z})$ with a dimensionality of 50, an encoder distribution $q(\boldsymbol{z} \mid \boldsymbol{x})$ in the diagonal Gaussian family, and a decoder distribution $p(\boldsymbol{x} \mid \boldsymbol{z})$ in a Bernoulli family. For the encoder and decoder networks we have used convolution neural networks with ReLU activations, dropout probability of 0.2, and residual block architecture with 4 residual blocks in each networks. For the encoder the residual block hidden dimensionalities were $32, 64, 128$, and 256, and for the decoder they were $128, 64, 32$, and 32. We used Adam optimiser (Kingma & Ba, 2014) with a learning rate of 10^{-4} and a cosine annealing schedule, for a total of 3k stochastic gradient ascent steps using a batch size of 200. Moreover sticking-the-landing gradients were used to reduce encoder network gradient variance (Roeder et al., 2017).

For pseudo-Gibbs we ran 5 independent chains for 5k iterations each. For MWG we have initialised 5 independent chains by running pseudo-Gibbs for 120 iterations, and then running the MWG sampler for 4880 iterations on each chain. For MWG′ we have initialised 5 independent chains by running LAIR with K=4 and R=1 for 120 iterations for each chain, and then run the MWG sampler for 4880 iterations on each chain. For AC-MWG′ we have initialised 5 independent chains by running LAIR with K=4 and R=1 for 120 iterations for each chain, and then run AC-MWG for 4880 iterations on each chain with $\epsilon = 0.05$. For LAIR we have initialised $K = 4$ particles from the marginal distribution of the VAE and then run the sampler with $K = 4$ and $R = 1$ for 5k iterations. The above evaluations were repeated with 5 different seeds, and the uncertainty reported in the figures reflects the uncertainty over different runs.

D Additional figures

In this appendix we provide additional figures for the experiments in this paper.

D.1 Synthetic 2D VAE

To aid with the understanding of the pitfalls in section 3 and our remedies in section 4, we here include additional figures on the synthetic VAE model (see details in appendix C.1). Specifically, in the top row of fig. 8 we plot the marginal distributions of the latents $p(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ that provide an additional perspective of the failure modes described in section 3: A method that is able to sample the joint distribution $p(\boldsymbol{x}_{\text{mis}}, \boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$

61

must also be able to effectively sample the marginal $p(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$, and if it is able to do so, then the joint $p(\boldsymbol{x}_{\text{mis}}, \boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ and the marginal of the missing variables $p(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ are recovered via ancestral sampling of eq. (1).

In the left-most column (pitfall I) we can see that MWG fails to explore the unimodal posterior $p(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$. As described in section 3 this is because the decoder distribution $p(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}} \mid \tilde{\boldsymbol{z}})$ places little density/mass on the *previous* value of $\boldsymbol{x}_{\text{mis}} = \boldsymbol{x}_{\text{mis}}^{t-1}$, which in turn gets such latent proposals $\tilde{\boldsymbol{z}}$ rejected. As a result, the MWG sampler remains "stuck" in a small part of the (marginal) posterior. The middle column provides an additional view of pitfall II. In particular, we see that the posterior distribution $p(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ in this case is multi-modal. However, an encoder $q(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$ conditioned on a specific completed data-point $\boldsymbol{x}_{\text{obs}} \cup \boldsymbol{x}_{\text{mis}}$ is unlikely to propose a latent value $\tilde{\boldsymbol{z}}$ that would reach one of the alternative modes. As a result, the pseudo-Gibbs and MWG samplers never reach the alternative modes of the posterior $p(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$, and remain stuck in a single mode. Finally, the right-most column reinforces the understanding of pitfall III. Specifically, we see that if MWG is initialised in a low-probability location of $p(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$, it may fail to reach the high-probability mode.

The second and third rows of fig. 8 show the posterior approximations obtained using AC-MWG (section 4.1) and LAIR (section 4.2). As we can see, similar to the results in sections 4.1.1 and 4.2.1 the proposed methods are able to avoid the pitfalls of pseudo-Gibbs and MWG. The proposed methods remedy pitfall I by targeting the marginal $p(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ instead of the joint $p(\boldsymbol{x}_{\text{mis}}, \boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$. Once approximate samples from the marginal $p(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ are obtained then the methods use this approximation to perform ancestral sampling of the joint $p(\boldsymbol{x}_{\text{mis}}, \boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$. Moreover, the methods address pitfall II by using the prior–variational mixture proposals in eqs. (3) and (5), which enable exploration of the latent space. The remedy to pitfall III is related to the remedies for pitfalls I and II: the prior–variational mixture proposal enables a search of the latent space and targeting the marginal distribution $p(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ allows the sampler to move from the poor initial location to a better one by *not* conditioning on the previous imputation value $\boldsymbol{x}_{\text{mis}} = \boldsymbol{x}_{\text{mis}}^{t-1}$.

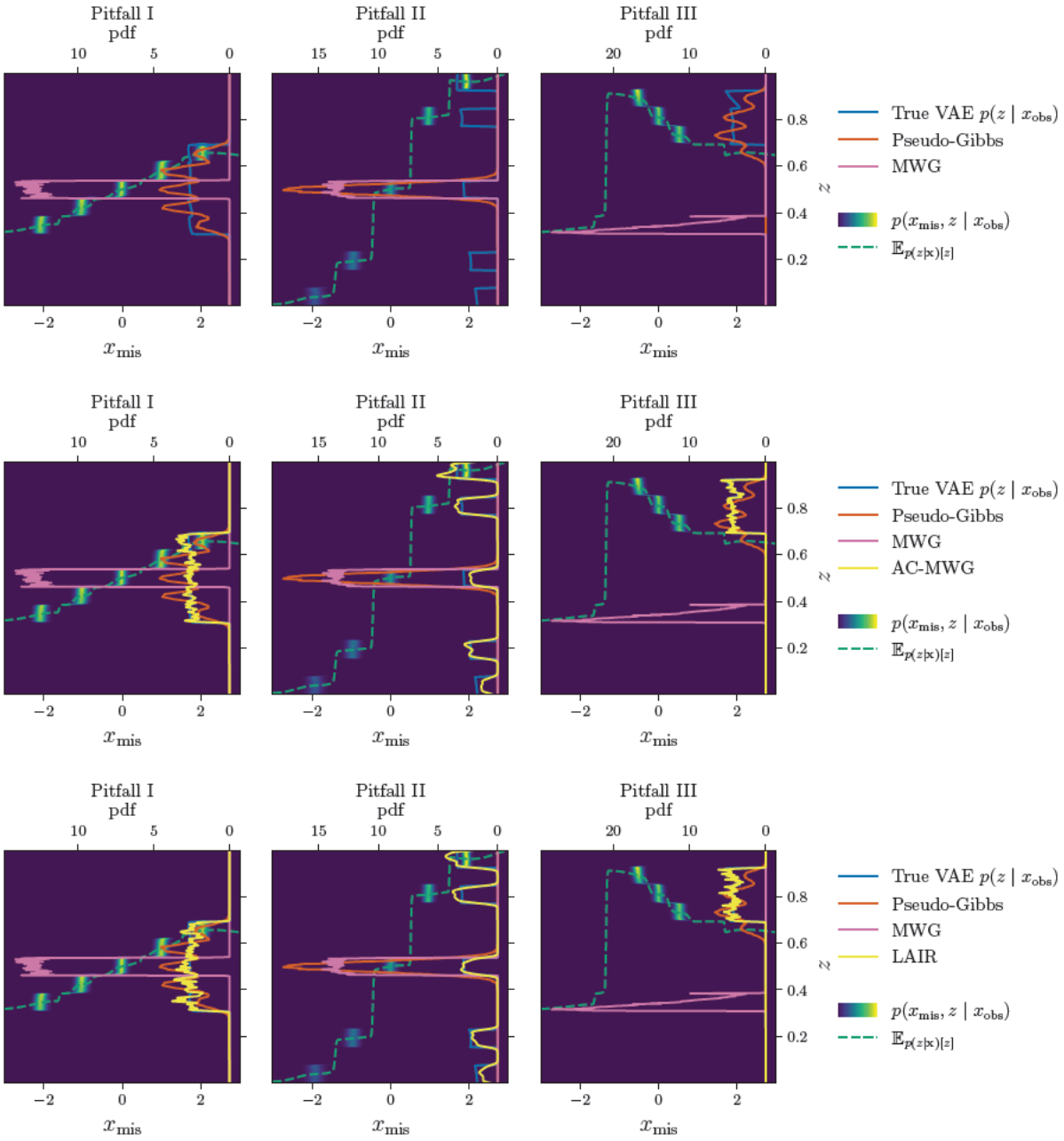


Figure 8: *Additional figures on the synthetic VAE, showing the marginals* $p(z \mid x_{obs})$. (The figure is best viewed in colour.) *Top:* showing only the true marginal $p(z \mid x_{\text{obs}})$ in blue colour and the marginals of pseudo-Gibbs (orange) and MWG (pink). *Middle:* showing the marginal of AC-MWG (yellow). *Bottom:* showing the marginal of LAIR (yellow).

D.2 Ablation study: Synthetic 2D VAE

In this section we perform an ablation study of AC-MWG and LAIR that supplements the results in sections 4.1.1 and 4.2.1.

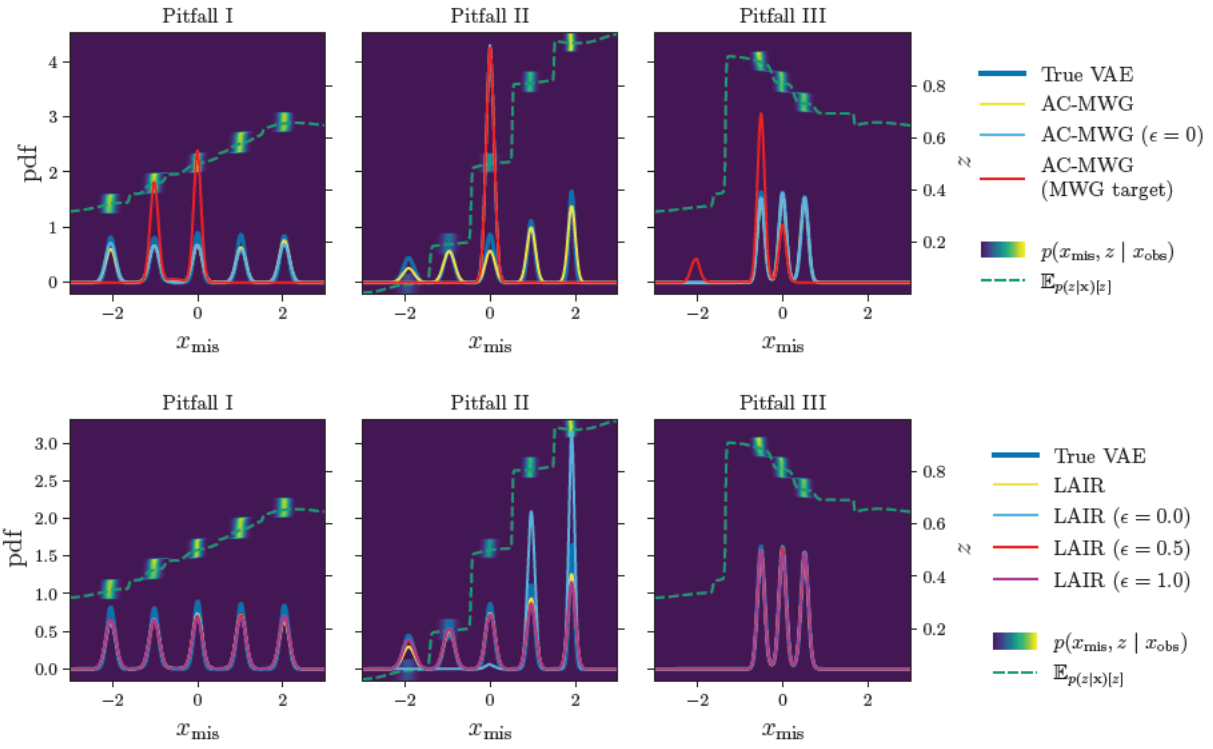


Figure 9: *Ablation studies on the 2D VAE sampling problems, same as in figs. 1 to 3.* (The figure is best viewed in colour.) *Top:* the same AC-MWG as before with $\epsilon = 0.01$ (yellow), AC-MWG with $\epsilon = 0.0$ (light blue) that does not "explore" the latent space via samples from the prior, and AC-MWG with $\epsilon = 0.01$ but the Metropolis–Hastings target of $p(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$ (red), same as MWG. *Bottom:* the same LAIR as before with $K = 19$ and $R = 1$ (i.e. $\epsilon = 0.05$, in yellow), LAIR with $K = 20$ and $R = 0$ (i.e. $\epsilon = 0.0$, in light blue) that does not "explore" the latent space using samples from the prior, LAIR with $K = 10$ and $R = 10$ (i.e. $\epsilon = 0.5$, in red), and LAIR with $K = 0$ and $R = 20$ (i.e. $\epsilon = 1.0$, in purple), which corresponds to standard (non-adaptive) importance resampling using the prior distribution as proposal.

In the top row of fig. 9 we show two ablation cases of AC-MWG. In the first case (light blue) we set $\epsilon = 0.0$ in the prior–variational mixture proposal in eq. (3). Without the prior component the sampler fails to "explore" the latent space (see the middle panel in the figure, where the light blue and red curves overlap) due to insufficiently exploratory proposal distribution (i.e. pitfall II). In the second case (red) we change the target distribution of the Metropolis–Hastings step from $p(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ in eq. (4) to $p(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$ used in standard MWG, that is, using the acceptance probability in eq. (2). We observe that with the MH target changed to $p(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$ the sampler also fails to mix between nearby modes in the latent space in the left-most panel (i.e. pitfall I), similar to MWG, which also affects the other two cases (middle and right panels). We therefore validate that the two modifications (the mixture proposal and the collapsed-Gibbs MH target $p(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$) introduced in section 4.1 are key components of the AC-MWG sampler.

In the bottom row of fig. 9 we show three ablation cases of LAIR by varying the prior probability $\epsilon = \frac{R}{K+R}$ in the mixture proposal in eq. (5). The first case (light blue) corresponds to LAIR with $\epsilon = 0.0$ (or $R = 0$) and performs sub-optimally (see the middle panel) due to lack of exploration in the latent space (see pitfall II). The second case (red) corresponds to LAIR with $\epsilon = 0.5$ (or $R = K = 10$) and performs similarly to our

64

Published in Transactions on Machine Learning Research (11/2023)

base LAIR case (yellow). The third case (purple) is LAIR with $\epsilon = 1.0$ and corresponds to a standard non-adaptive importance resampling with the prior distribution as the proposal. The standard importance sampling (purple) performs equally-well because of the simplicity and low-dimensionality of the latent space, however as the latent space gets more complex and higher dimensional the adaptive LAIR sampler will perform better (see results in appendix D.4).

D.3 Mixture-of-Gaussians MNIST

Figure 10 shows the conditional mean and standard deviation at each "missing" pixel of the image, conditional on the "observed" pixels surrounded by a red border. Top-left shows the ground truth values, and the rest show values estimated from samples produced using the VAE and the (approximate) samplers. Furthermore, fig. 11 shows the absolute error in the conditional means (black is better) and signed error on the standard deviations (blue is underestimated, red is overestimated, white is perfect). The figures show a complementary view of the results in section 5.1. Interestingly, we can see that pseudo-Gibbs and MWG can overestimate the variance at some pixels while at the same time underestimating it at other pixels. The proposed methods, AC-MWG and LAIR, are less affected by this issue.

Figure 12 corresponds to fig. 4 in the main text but we additionally show MWG with MAP initialisation using stochastic gradient ascent with 5 random restarts (red). Furthermore, fig. 13 shows the experiment results using additional metrics. The additional metrics mirror the results in the main text.

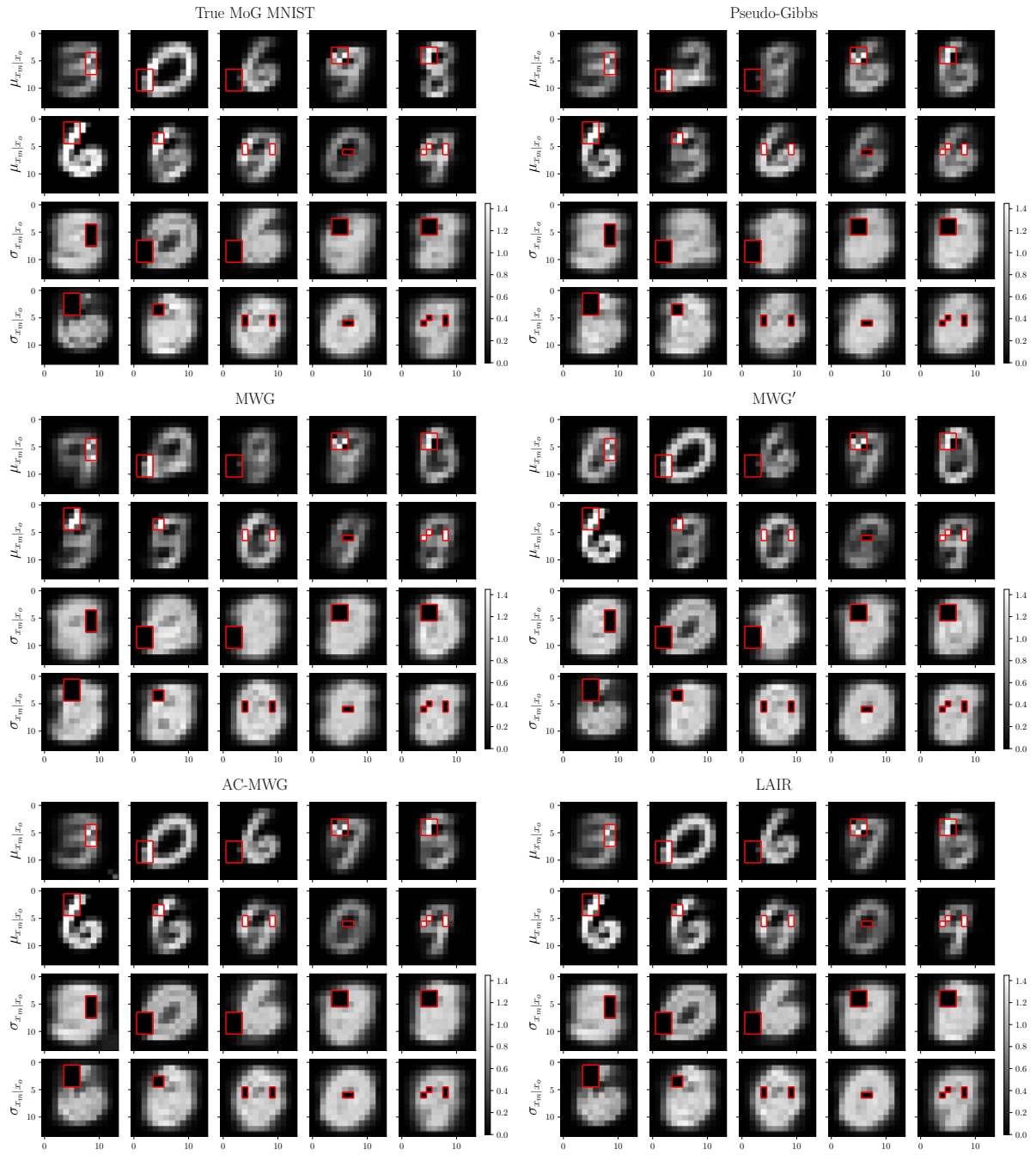


Figure 10: *Conditional mean $\mu_{\boldsymbol{x}_{mis}|\boldsymbol{x}_{obs}}$ and standard deviation $\sigma_{\boldsymbol{x}_{mis}|\boldsymbol{x}_{obs}}$ on the mixture-of-Gaussians MNIST.* The top-left panel shows the ground-truth values, and the other panels show estimates from imputations generated by the evaluated samplers. The pixels surrounded by a red border are the observed values $\boldsymbol{x}_{\text{obs}}$.

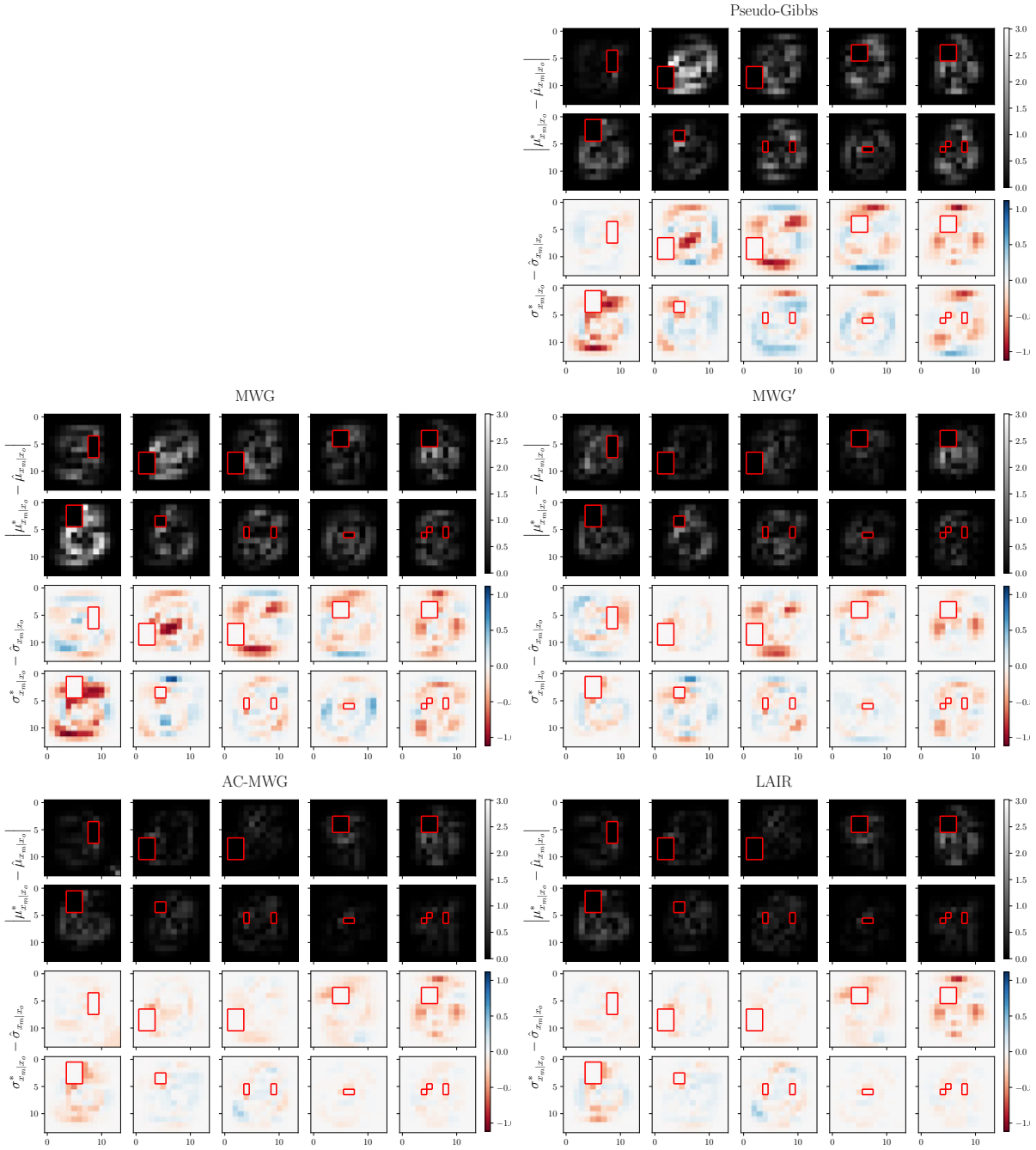


Figure 11: *The absolute error on the conditional mean* $\mu_{\boldsymbol{x}_{mis}|\boldsymbol{x}_{obs}}$ *and the signed error on the standard deviation* $\sigma_{\boldsymbol{x}_{mis}|\boldsymbol{x}_{obs}}$ *on the mixture-of-Gaussians MNIST.* We can clearly see that the proposed methods (bottom row) outperform the existing samplers.

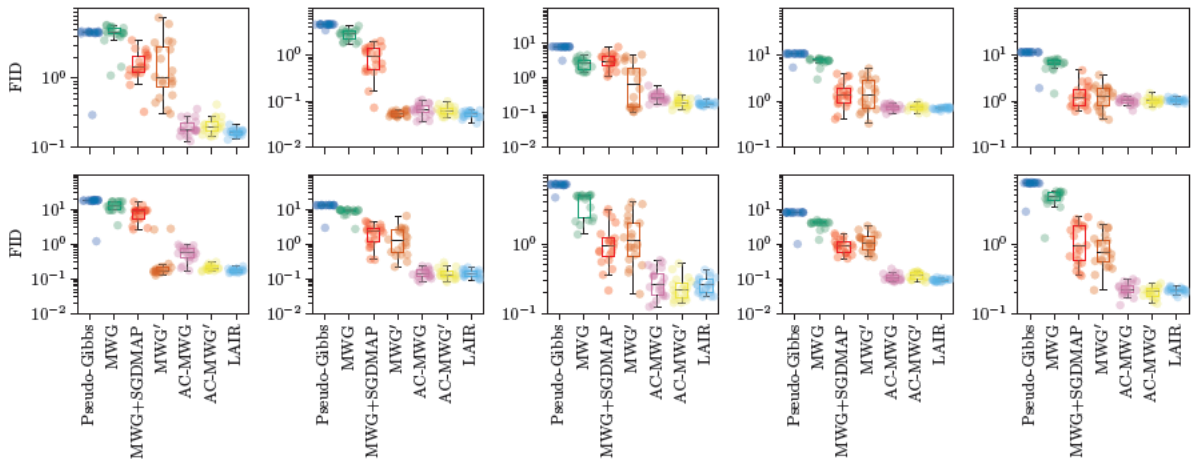
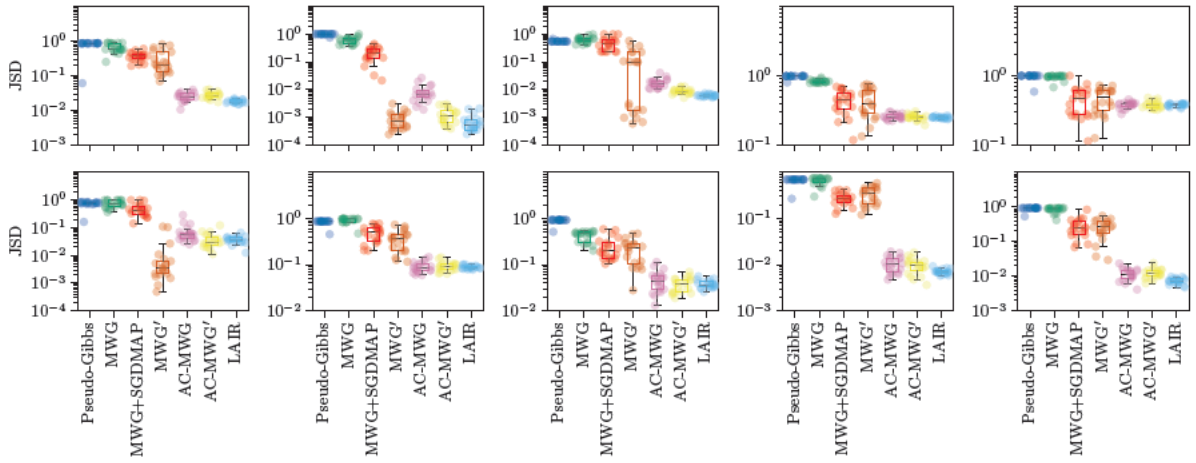
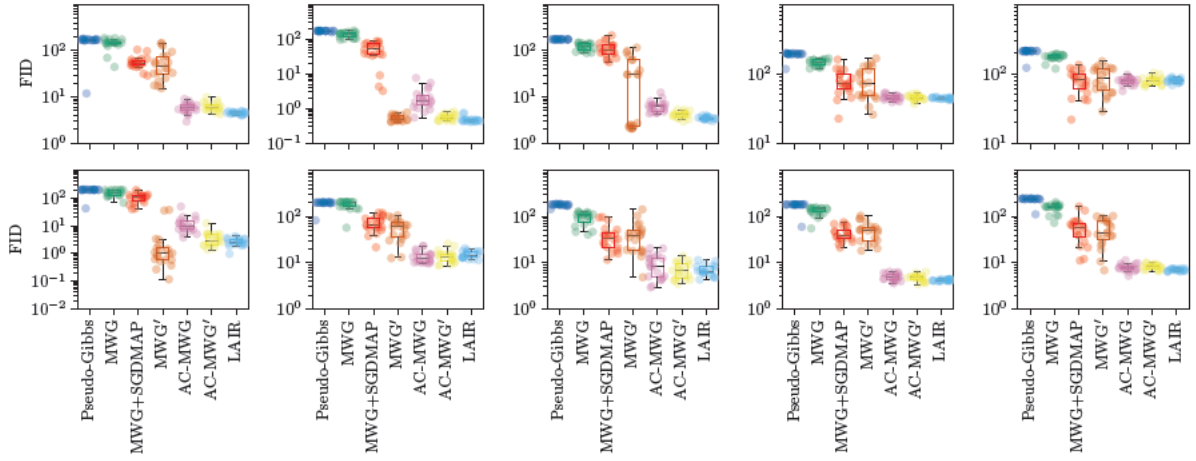


Figure 12: Same as fig. 4, but with additional method included, MWG+SGDMAP (red), which initialises MWG using stochastic gradient ascent on the log-likelihood (with 5 restarts).



(a) Jensen–Shannon divergence (JSD) between the ground truth conditional $p(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$, and estimator $\hat{p}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}) = \frac{1}{N}\sum_{i=1}^{N} p(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}^{i})$, where $\boldsymbol{x}_{\text{mis}}^{i}$ come from the imputation methods and N is the total number of imputations.



(b) Fréchet inception distance (FID) between samples from the ground truth conditional $p(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$, and samples obtained from the imputation methods. The inception model features used in FID computation are the final layer outputs of a classifier neural network.

Figure 13: *Additional metrics on the mixture-of-Gaussians MNIST.* Each panel in the subfigures corresponds to a different conditional sampling problem $p(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$. Each evaluation is repeated 20 times, and the box-plot represents the inter-quartile range, including the median, and the whiskers show the overall range of the results.

D.4 Ablation study: Mixture-of-Gaussians MNIST

This section shows an ablation study of AC-MWG and LAIR on the mixture-of-Gaussians MNIST data set that supplements the results in section 5.1.

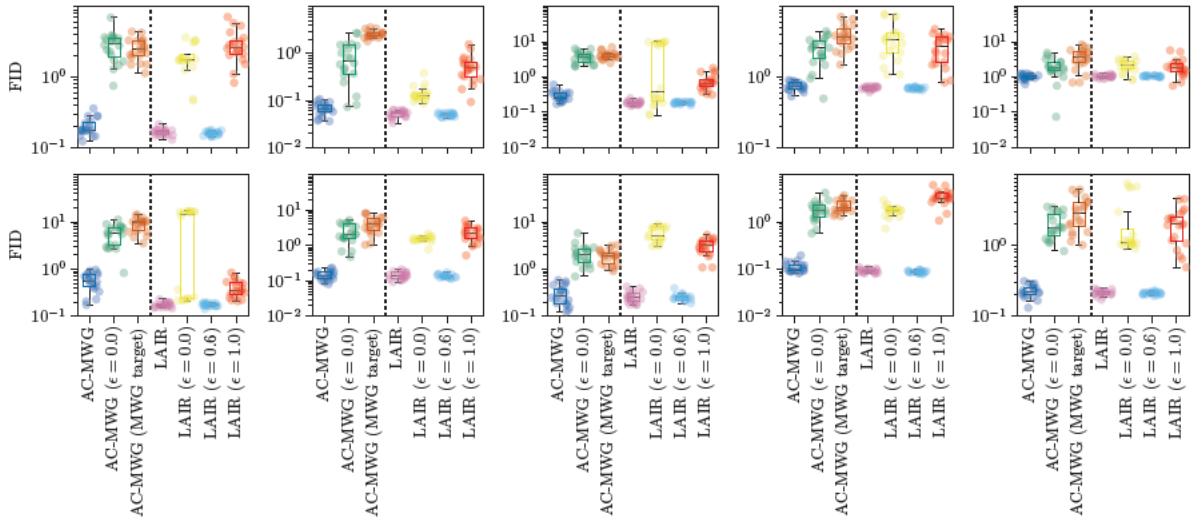


Figure 14: *Ablation studies on the MoG MNIST sampling problems, same as in fig. 4.* The FID score is computed using the final layer outputs of the encoder network as the inception features. *Left part of each panel:* AC-MWG (deep blue) is the same AC-MWG as in fig. 4 with $\epsilon = 0.05$, AC-MWG with $\epsilon = 0.0$ (green) corresponding to no prior component in the proposal distribution in eq. (3), and AC-MWG with $\epsilon = 0.05$ but the Metropolis–Hastings target of $p(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$ (orange), same as MWG. *Right part of each panel:* LAIR (pink) is the same LAIR as in fig. 4 with $K = 4$ and $R = 1$ (i.e. $\epsilon = 0.2$), LAIR with $K = 5$ and $R = 0$ (i.e. $\epsilon = 0.0$, yellow), LAIR with $K = 2$ and $R = 3$ (i.e. $\epsilon = 0.6$, light blue), and LAIR with $K = 0$ and $R = 5$ (i.e. $\epsilon = 1.0$, red) corresponding to standard (non-adaptive) importance resampling with the prior distribution as the proposal.

The left part of each panel in fig. 14 shows two ablation cases of AC-MWG. In the first case (green) we set $\epsilon = 0.0$ in the prior–variational mixture proposal in eq. (3). As explained in pitfall II the sampler fails to explore the latent space and hence exhibits degraded performance compared to AC-MWG with $\epsilon > 0$ (deep blue). In the second case (orange) we change the target distribution of the Metropolis–Hastings step from $p(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ in eq. (4) to $p(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$ used in standard MWG, that is, using the acceptance probability in eq. (2). Similarly, we see that this ablation significantly reduces the performance of the sampler (orange versus deep blue). With this evaluation, similar to the results in appendix D.2, we validate that the proposed components (mixture proposal and collapsed-Gibbs MH target) are key to the performance of the AC-MWG method.

The right-hand part of each panel in fig. 14 shows three ablation cases of LAIR with varying prior probabilities $\epsilon = \frac{R}{K+R} \in \{0.0, 0.6, 1.0\}$ (or equivalently, varying K and R) in the mixture proposal in eq. (5). The first case (yellow) is LAIR with $\epsilon = 0.0$ (i.e. $R = 0$), which corresponds to not using the prior distribution in the mixture proposal in eq. (5), and exhibits a significantly downgraded performance over LAIR with $\epsilon = 0.2$ (or $K = 4$ and $R = 1$, pink). The second case (light blue) is LAIR with $\epsilon = 0.6$ and performs similarly to LAIR with $\epsilon = 0.2$ (pink), hence showing that the method is not highly sensitive to the choice of ϵ as long as the edge cases ($\epsilon = 0$ and $\epsilon = 1$) are avoided. The third case (red) is LAIR with $\epsilon = 1.0$ and corresponds to a standard non-adaptive importance resampling with the prior distribution as the proposal. As we see here, the non-adaptive importance resampling (red) performs sub-optimally and hence validates that the adaptation in LAIR is important for good performance of the method.

Published in Transactions on Machine Learning Research (11/2023)

D.5 UCI data sets

In fig. 15 we show additional metrics of the experiments in section 5.2. We also include MWG with pseudo-Gibbs initialisation (red) as originally proposed in Mattei & Frellsen (2018). The first two rows show energy-distance MMD and Laplacian MMD between the imputed data sets and the ground truth data. We observe a similar behaviour to the results in the main text. The main exception is the Hepmass data where MWG′ (orange) seems to be preferred. However, we note that part of the good performance of MWG′ (orange) on Hepmass data is due to the use of LAIR initialisation, while using pseudo-Gibbs initialisation (red) performs similarly to LAIR (yellow). Moreover, the final row shows the average mean absolute error, and the proposed methods, AC-MWG (pink) and LAIR (yellow), are preferred over the existing methods on all data sets.

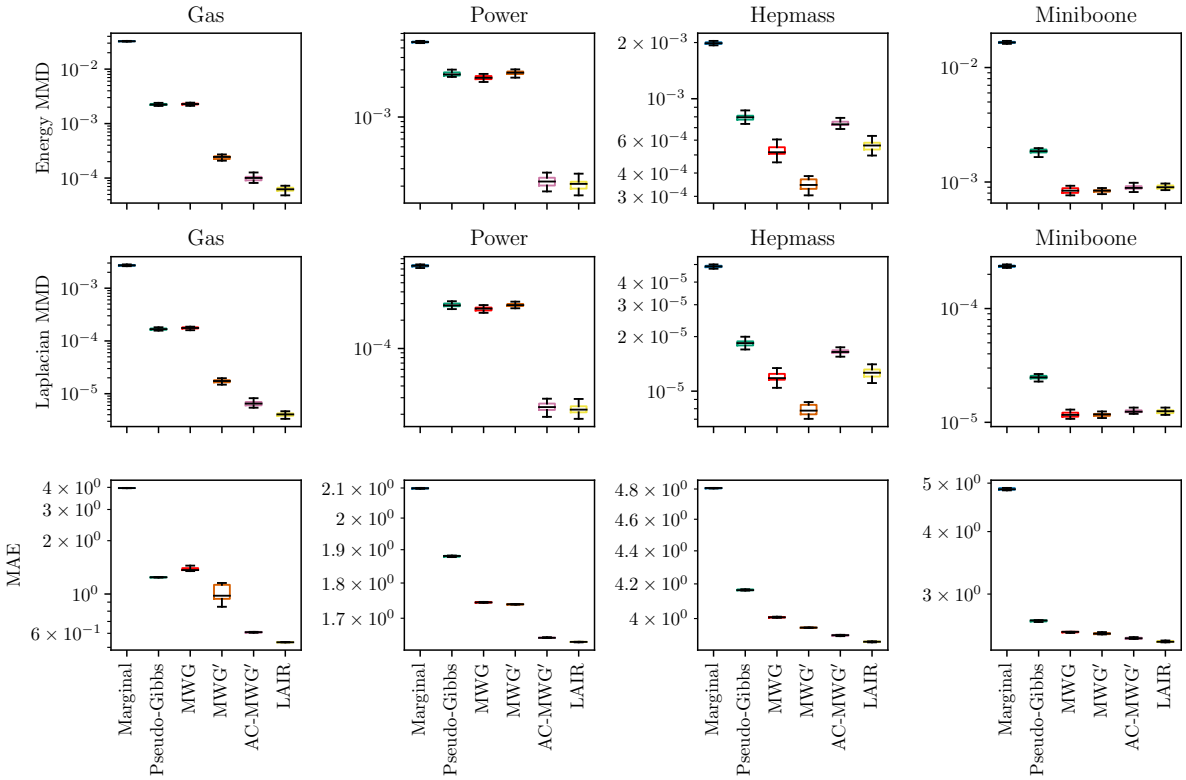


Figure 15: *Additional metrics on sampling performance on four real-world UCI data sets. Top:* energy MMD. *Middle:* Laplacian MMD. *Bottom:* average MAE of the imputations. The divergences are evaluated on a 50k data-point subset of test data (except for Miniboone where the full test data set was used), and the MAE is averaged over the full test data set. In all rows imputations from the final iteration of the corresponding algorithms are used and uncertainty is shown over different runs.

4.2 Additional discussion

Pitfalls when relationship between the latent and visible variables is weak. In section 3 of our publication, we identified that the potential pitfalls in conditional sampling of VAEs often arise due to a *strong* relationship between the latent and the visible variables, leading to poor mixing in Gibbs Markov chains (pitfall I). This naturally raises the question: What happens when this relationship is *weak*? A weak relationship between the variables implies that the values of the latents $\boldsymbol{z}$ have minimal impact on the conditional distribution $p(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}} \mid \boldsymbol{z})$. In the extreme case where the variables have no relationship, the distribution becomes independent of $\boldsymbol{z}$, resulting in $p(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}} \mid \boldsymbol{z}) = p(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$. Consequently, assuming we can sample from $p(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{z})$ as before, which becomes $p(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ in the extreme case, we can expect improved mixing of Gibbs-like samples as the relationship between latents and visibles weakens. However, as mentioned in the publication, learning a VAE model that accurately represents the observed data is a common goal in the VAE literature, and hence cases where the relationship between latents and visibles is strong are highly important to improve the applicability of VAE models for missing data imputation.

The effect of variational posterior mode collapse. In section 3 of the publication, we also identified that using the complete-data variational distribution $q(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$ that approximates the model posterior $p(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$ may be "insufficiently exploratory" for conditional sampling of the missing data distribution $p(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ (pitfall II). The insufficiency arises because the variational proposal $q(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$, when conditioned on an imputation $\boldsymbol{x}_{\text{mis}}$ from one mode of the imputation distribution, may fail to produce a proposal $\tilde{\boldsymbol{z}}$ that decodes into an alternative mode. This issue is especially pronounced when the different modes are encoded into latent encodings $\boldsymbol{z}$ that are very distant in the latent space. Moreovoer, this problem may be further amplified when the variational distribution $q(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$ fails to accurately capture the model posterior $p(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$, for example, due to posterior mode collapse (Zhao et al., 2018; Lucas et al., 2019; Wang et al., 2021). Consequently, methods relying solely on the variational distribution as the proposal may never sample imputations corresponding to the collapsed modes. To address these issues, our proposed methods use a prior–variational mixture distribution as the proposal. This approach aims to enhance the exploratory capabilities of the sampling process and mitigate the effects of posterior mode collapse.

Chapter 5

VAE estimation from incomplete data

Model estimation from incomplete data is a central task that enables many downstream tasks, including missing data imputation. As highlighted in section 3.2.1, variational autoencoders (VAEs) often use a conditional independence assumption in the decoder distribution. This assumption enables efficient marginalisation of the missing variables and, as a result, efficient model estimation (section 2.2.6). Consequently, a substantial body of literature has used this assumption for VAE estimation from incomplete data. However, as discussed in section 3.2.3, despite facilitating efficient marginalisation, research indicates that adaptations may be needed when handling incomplete data versus fully-observed data. The specific reasons behind the increased complexity of estimation from incomplete data remain unanswered. This gap motivates the following question:

Research Question 2: What are the factors that contribute to the increased complexity of VAE estimation from incomplete data, and what strategies can be used to effectively address this additional complexity due to missing data?

In this chapter, we attempt to further the understanding of VAE estimation from incomplete data by investigating this research question. Our manuscript, included verbatim in section 5.1, reveals a previously unknown phenomenon caused by missing data that impacts the estimation of the model. The paper motivates a framework based on variational mixtures, which effectively improves model estimation from incomplete data.

5.1 Manuscript

This section includes a copy of the following publication with minor additions:[8]

Vaidotas Simkus and Michael U. Gutmann. Improving Variational Autoencoder Estimation from Incomplete Data with Mixture Variational Families. *Transactions on Machine Learning Research*, 2024

[8] An earlier version of this paper was accepted to the DMLR workshop at ICLR 2024.

Published in Transactions on Machine Learning Research (06/2024)

Improving Variational Autoencoder Estimation from Incomplete Data with Mixture Variational Families

Vaidotas Simkus
Michael U. Gutmann
School of Informatics
University of Edinburgh

Reviewed on OpenReview: *https://openreview.net/forum?id=lLVmIvZfry*

Abstract

We consider the task of estimating variational autoencoders (VAEs) when the training data is incomplete. We show that missing data increases the complexity of the model's posterior distribution over the latent variables compared to the fully-observed case. The increased complexity may adversely affect the fit of the model due to a mismatch between the variational and model posterior distributions. We introduce two strategies based on (i) finite variational-mixture and (ii) imputation-based variational-mixture distributions to address the increased posterior complexity. Through a comprehensive evaluation of the proposed approaches, we show that variational mixtures are effective at improving the accuracy of VAE estimation from incomplete data.

1 Introduction

Deep latent variable models, as introduced by Kingma & Welling (2013); Rezende et al. (2014); Goodfellow et al. (2014); Sohl-Dickstein et al. (2015); Krishnan et al. (2016); Dinh et al. (2017), have emerged as a predominant approach to model real-world data. The models excel in capturing the intricate nature of data by representing it within a well-structured latent space. *However, they typically require large amounts of fully-observed data at training time, while practitioners in many domains often only have access to incomplete data sets.*

In this paper we focus on the class of variational autoencoders (VAEs, Kingma & Welling, 2013; Rezende et al., 2014) and investigate the implications of incomplete training data on model estimation. Our contributions are as follows:

- We show that data missingness can add significant complexity to the model posterior of the latent variables, hence requiring more flexible variational families compared to scenarios with fully-observed data (section 3).
- We propose finite variational-mixture approaches to deal with the increased complexity due to missingness for both standard and importance-weighted ELBOs (section 4.1).
- We further propose an imputation-based variational-mixture approach, which decouples model estimation from data missingness problems, and as a result, improves model estimation when using the standard ELBO (section 4.2).
- We evaluate the proposed methods for VAE estimation on synthetic and realistic data sets with missing data (section 6).

The proposed methods achieve better or similar estimation performance compared to existing methods that do not use variational mixtures. Moreover, the mixtures are formed by the variational families that are used

in the fully-observed case, which allows us to seamlessly re-use the inductive biases from the well-studied scenarios with fully-observed data (see e.g. Miao et al., 2022, for the importance of inductive biases in VAEs).

2 Background: Standard approach for VAEs estimation from incomplete data

We consider the situation where some part of the training data-points might be missing. We denote the observed and missing parts of the i-th data-point $\boldsymbol{x}^i$ by $\boldsymbol{x}^i_{\text{obs}}$ and $\boldsymbol{x}^i_{\text{mis}}$, respectively, where $\boldsymbol{x}^i$ is D-dimensional and the dimensions of $\boldsymbol{x}^i_{\text{obs}}$ and $\boldsymbol{x}^i_{\text{mis}}$ must add to D. This split into observed and missing components corresponds to a missingness pattern $\boldsymbol{m}^i \in \{0,1\}^D$ with $m^i_j = 1$ if the j-th dimension is observed and $m^i_j = 0$ if the dimension is missing. The missingness pattern $\boldsymbol{m}^i$ is generally different for each data-point and is a realisation of a random variable $\boldsymbol{m}$ that follows a typically unknown missingness distribution $p^*(\boldsymbol{m} \mid \boldsymbol{x}^i)$. We make the common assumption that the missingness distribution does not depend on the missing variables, which is known as the ignorable missingness or missing-at-random assumption (MAR, e.g. Little & Rubin, 2002, Section 1.3).[1] The MAR assumption allows us to ignore the missingness pattern $\boldsymbol{m}^i$ when fitting a model $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ of the true distribution $p^*(\boldsymbol{x})$ from incomplete data (see e.g. Seaman et al., 2013, Theorem 1), as well as when performing multiple imputation of the missing data (see e.g. van Buuren, 2018, Section 2.2.6).

The VAE model with parameters $\boldsymbol{\theta}$ is typically specified using a decoder distribution $p_{\boldsymbol{\theta}}(\boldsymbol{x} \mid \boldsymbol{z})$, parametrised using a neural network, and a prior $p_{\boldsymbol{\theta}}(\boldsymbol{z})$ over the latents $\boldsymbol{z}$ that can either be fixed or learnt. A principled approach to handling incomplete training data is then to marginalise the missing variables from the likelihood $p_{\boldsymbol{\theta}}(\boldsymbol{x})$, which yields the marginal likelihood

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}^i_{\text{obs}}) = \int p_{\boldsymbol{\theta}}(\boldsymbol{x}^i_{\text{obs}}, \boldsymbol{x}^i_{\text{mis}})\, \mathrm{d}\boldsymbol{x}^i_{\text{mis}} = \iint p_{\boldsymbol{\theta}}(\boldsymbol{x}^i_{\text{obs}}, \boldsymbol{x}^i_{\text{mis}} \mid \boldsymbol{z}) p_{\boldsymbol{\theta}}(\boldsymbol{z})\, \mathrm{d}\boldsymbol{z}\, \mathrm{d}\boldsymbol{x}^i_{\text{mis}} = \int p_{\boldsymbol{\theta}}(\boldsymbol{x}^i_{\text{obs}} \mid \boldsymbol{z}) p_{\boldsymbol{\theta}}(\boldsymbol{z})\, \mathrm{d}\boldsymbol{z}, \quad (1)$$

where the inner integral $\int p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}} \mid \boldsymbol{z})\, \mathrm{d}\boldsymbol{x}_{\text{mis}}$ is often computationally tractable in VAEs due to standard assumptions, such as the conditional independence of $\boldsymbol{x}$ given $\boldsymbol{z}$ or the use of the Gaussian family for the decoder $p_{\boldsymbol{\theta}}(\boldsymbol{x} \mid \boldsymbol{z})$. Similar to existing work, we also make the assumption that the marginalisation of the missing variables is tractable. However, the marginal likelihood above remains intractable to compute as a consequence of the integral over the latents $\boldsymbol{z}$.

Due to the intractable integral, VAEs are typically fitted via a variational evidence lower-bound (ELBO)

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{y}) \geq \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{y})}\left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{y} \mid \boldsymbol{z}) p_{\boldsymbol{\theta}}(\boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{y})}\right] = \log p_{\boldsymbol{\theta}}(\boldsymbol{y}) - D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{y}) \,||\, p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{y})), \quad (2)$$

where $\boldsymbol{y}$ refers to $\boldsymbol{x}^i$ in the fully-observed case, and to $\boldsymbol{x}^i_{\text{obs}}$ in the incomplete-data case, and $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{y})$ is an (amortised) variational distribution with parameters $\boldsymbol{\phi}$ that is shared for all data-points in the data set (Gershman & Goodman, 2014). The amortised distribution is parametrised using a neural network (the encoder), which takes the data-point $\boldsymbol{y}$ as the input and predicts the distributional parameters of the variational family. Moreover, when the data is incomplete, i.e. $\boldsymbol{y} = \boldsymbol{x}^i_{\text{obs}}$, sharing of the encoder for any pattern of missingness is often achieved by fixing the input dimensionality of the encoder to twice the size of $\boldsymbol{x}$ and providing $\gamma(\boldsymbol{x}^i_{\text{obs}})$ and $\boldsymbol{m}^i$ as the inputs,[2] where $\gamma(\cdot)$ is a function that takes the incomplete data-point $\boldsymbol{x}_{\text{obs}}$ and produces a vector of length D with the missing dimensions set to zero[3] (Nazábal et al., 2020; Mattei & Frellsen, 2019).

From eq. (2), we see that the training objective for incomplete and fully-observed data has the same form, and therefore it may seem that fitting VAEs from incomplete data would be similarly difficult to the fully-observed case. However, as we will see next, data missingness can make model estimation much harder than in the complete data case.

75

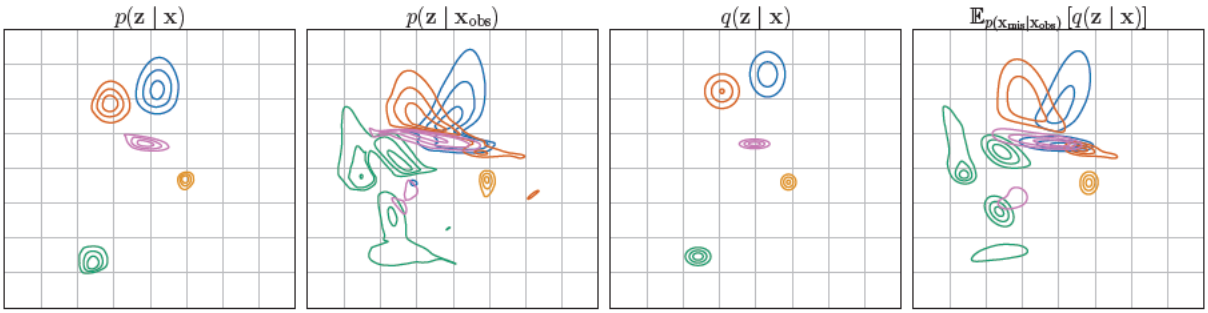


Figure 1: *Illustration of the posterior complexity due to missing data.* Each colour represents a different data-point $\boldsymbol{x}^i$. First: the model posterior $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x})$ under complete data $\boldsymbol{x}$. Second: the model posterior $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ under incomplete data $\boldsymbol{x}_{\text{obs}}$. Third: variational approximation $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x})$ of the complete-data posterior $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x})$. Fourth: an imputation-mixture variational approximation $\mathbb{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}}|\boldsymbol{x}_{\text{obs}})}[q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})]$ of the incomplete posterior $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$. In these figures, we use a VAE with Gaussian variational, prior, and decoder distributions fitted on complete data, then the incomplete data-points $\boldsymbol{x}_{\text{obs}}$ are obtained by randomly masking 50% of the values from the complete data-points $\boldsymbol{x}$. The data was generated from a 5-dimensional mixture-of-Gaussians model with 15 components, see appendix E.1 for more details.

3 Implications of incomplete data for VAE estimation

The decomposition of the ELBO in eq. (2) emphasises that accurate estimation of the VAE model requires the variational distribution $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ to accurately approximate the model posterior $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$. While it might appear that the marginalisation of the missing variables in eq. (1) comes at no cost since the ELBO maintains the same form as in the complete case, we here illustrate that his is not the case.

In the two left-most columns of fig. 1 we illustrate the model posteriors $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \cdot)$ under fully-observed data $\boldsymbol{x}$ and partially-observed data $\boldsymbol{x}_{\text{obs}}$.[4] We discover that the model posteriors $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x})$, which exhibited a certain regularity in the complete-data scenario, have become irregular multimodal distributions $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ in the case of incomplete data.[5] Hence, accurate estimation of VAEs from incomplete data may require more flexible variational families than in the fully-observed case: while a family may sufficiently well approximate the model posterior in the fully-observed case, it may no longer be sufficiently flexible in the incomplete data case. We provide a further explanation when this situation may occur in appendix A. As a result of the mismatch between the model posterior $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ and the variational distribution $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$, the incomplete-data KL divergence term $D_{\text{KL}}(q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}) \,||\, p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}))$ in eq. (2) may be large compared to the analogous KL term $D_{\text{KL}}(q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}) \,||\, p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}))$ in the complete-data case, subsequently introducing a bias to the fit of the model.

In the two right-most columns of fig. 1 we illustrate the variational approximations of the aforementioned model posterior distributions, $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x})$ and $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$. The first of the two plots shows the complete-data variational distribution $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x})$ obtained after training, which well-approximates the model posterior $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x})$. In the second of the two plots, we construct the incomplete-data posterior approximation as follows: $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}) = \mathbb{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}}|\boldsymbol{x}_{\text{obs}})}[p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})] \approx \mathbb{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}}|\boldsymbol{x}_{\text{obs}})}[q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})]$, which well-approximates the model posterior $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ too. Taken together, the two plots show that if the variational family used in the fully-observed case well-approximates the model posterior, i.e. $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}) \approx p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x})$, then the imputation-mixture $\mathbb{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}}|\boldsymbol{x}_{\text{obs}})}[q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})]$ will also be a good approximation of the incomplete-data posterior $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$. This observation suggests that we can work with the same variational family

[1]While there is some historical disparity between MAR assumptions in the statistics literature, we can here adopt the weakest MAR assumption, also known as the *realised* MAR (Seaman et al., 2013).

[2]Alternative encoder architectures, such as, permutation-invariant networks (Ma et al., 2019) are also used.

[3]Equivalent to setting the missing dimensions to the empirical mean for zero-centered data.

[4]In fig. 1 we use a VAE with Gaussian variational, prior, and decoder distributions fitted on complete data.

[5]A related phenomenon, called posterior inconsistency, has been recently reported in concurrent work by Sudak & Tschiatschek (2023), relating $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ and $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}\setminus u})$, where u is a subset of the observed dimensions (see section 5).

in both the fully-observed and incomplete data scenarios if we adopt a mixture approach. In the rest of this paper, we investigate opportunities to improve VAE estimation from incomplete data by constructing variational mixture approximations of the incomplete-data posterior.

4 Fitting VAEs from incomplete data using mixture variational families

We propose working with mixture variational families to mitigate the increase in posterior complexity and to improve the estimation accuracy of VAEs when the training data are incomplete. This allows us to use families of distributions for the mixture components that are known to work well when the data is fully-observed, and use the mixtures to handle the increased posterior complexity due to missing data.

We propose two approaches for constructing variational mixtures. In section 4.1 we specify $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ as a finite-mixture distribution that can be learnt directly using the reparametrisation trick. In section 4.2 we investigate an imputation-based variational-mixture to approximate $\mathbb{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}}|\boldsymbol{x}_{\text{obs}})}[q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})]$. Detailed evaluation of the proposed methods is provided in section 6.

4.1 Using finite mixture variational distributions to fit VAEs from incomplete data

In section 3 we saw that the imputation-mixture $\mathbb{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}}|\boldsymbol{x}_{\text{obs}})}[q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})]$ is a good approximation of the incomplete-data posterior $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ and would thus be a suitable variational distribution $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$. However, estimation of the imputation distribution $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ is generally intractable for VAEs (Rezende et al., 2014; Mattei & Frellsen, 2018a; Simkus & Gutmann, 2023). Hence, we here consider a more tractable approach and specify the variational distribution $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ in terms of a finite-mixture distribution:

$$q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}) = \sum_{k=1}^{K} q_{\boldsymbol{\phi}}(k \mid \boldsymbol{x}_{\text{obs}}) q^k_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}), \tag{3}$$

where $q_{\boldsymbol{\phi}}(k \mid \boldsymbol{x}_{\text{obs}})$ is a categorical distribution over the components $k \in \{1, \dots, K\}$ and each component distribution $q^k_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ belongs to any reparametrisable distribution family. Both $q_{\boldsymbol{\phi}}(k \mid \boldsymbol{x}_{\text{obs}})$ and $q^k_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ are amortised using an encoder network, similar to section 2.

The "reparametrisation trick" is typically used in VAEs to efficiently optimise the parameters $\boldsymbol{\phi}$ of the variational distribution $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$. This requires that the random variable $\boldsymbol{z}$ can be parametrised as a learnable differentiable transformation $t(\boldsymbol{\epsilon}; \boldsymbol{x}_{\text{obs}}, \boldsymbol{\phi})$ of another random variable $\boldsymbol{\epsilon}$ that follows a distribution with no learnable parameters. However, reparametrising mixture-families requires extra care: sampling the mixture $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ in eq. (3) is typically done via ancestral sampling by first drawing $k \sim q_{\boldsymbol{\phi}}(k \mid \boldsymbol{x}_{\text{obs}})$ and then $\boldsymbol{z} \sim q^k_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$, but the sampling of the categorical distribution $q_{\boldsymbol{\phi}}(k \mid \boldsymbol{x}_{\text{obs}})$ is non-differentiable, making the direct application of the "reparametrisation trick" generally infeasible.

To make the fitting of VAEs using mixture-variational distributions feasible we consider two objectives based on the variational ELBO (Kingma & Welling, 2013; Rezende et al., 2014):

$$\mathcal{L}_{\text{ELBO}}(\boldsymbol{x}_{\text{obs}}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}_{\text{obs}})} \left[\log w(\boldsymbol{z})\right], \quad \text{and} \tag{4}$$

$$\mathcal{L}_{\text{SELBO}}(\boldsymbol{x}_{\text{obs}}) = \sum_{k=1}^{K} q_{\boldsymbol{\phi}}(k \mid \boldsymbol{x}_{\text{obs}}) \mathbb{E}_{q^k_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}_{\text{obs}})} \left[\log w(\boldsymbol{z})\right], \tag{5}$$

$$\text{where} \quad w(\boldsymbol{z}) = \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}}, \boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})}. \tag{6}$$

The first objective $\mathcal{L}_{\text{ELBO}}$ corresponds to the standard ELBO, while $\mathcal{L}_{\text{SELBO}}$ is the stratified ELBO (Roeder et al., 2017, Section 4; Morningstar et al., 2021).[6] When working with $\mathcal{L}_{\text{ELBO}}$, due to the mixture variational family, we will need to optimise $\boldsymbol{\phi}$ with *implicit* reparametrisation (Figurnov et al., 2019). Implicit

[6]Note that the two objectives $\mathcal{L}_{\text{ELBO}}$ and $\mathcal{L}_{\text{SELBO}}$ are equal in expectation but their variance with limited Monte Carlo samples will be different, thus presenting some potential trade-offs during optimisation

reparametrisation of mixture distributions requires that the component distributions $q_{\boldsymbol{\phi}}^k(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ can be factorised using the chain rule, i.e. $q_{\boldsymbol{\phi}}^k(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}) = \prod_d q_{\boldsymbol{\phi}}^k(z_d \mid \boldsymbol{z}_{<d}, \boldsymbol{x}_{\text{obs}})$, and that we have access to the CDF (or other standardisation function) of each factor $q_{\boldsymbol{\phi}}^k(z_d \mid \boldsymbol{z}_{<d}, \boldsymbol{x}_{\text{obs}})$. However, the chain rule requirement can be difficult to satisfy for some highly flexible variational families, such as normalising flows (e.g. Papamakarios et al., 2021), and finding the (conditional) CDF of the factors can also be hard if not already known in closed form. Consequently, $\mathcal{L}_{\text{ELBO}}$ with implicit reparametrisation may not be usable with all distribution families as components of the variational mixture.

The second objective $\mathcal{L}_{\text{SELBO}}$, on the other hand, samples the mixture distribution with stratified sampling,[7] which avoids the non-differentiability of sampling $q_{\boldsymbol{\phi}}(k \mid \boldsymbol{x}_{\text{obs}})$, and as a result allows us to use any family of reparametrisable distributions as the mixture components.

The importance-weighted ELBO (IWELBO, Burda et al., 2015) is often used as an alternative to the standard ELBO as it can be made tighter. We here also consider an ordinary version, $\mathcal{L}_{\text{IWELBO}}$, and a stratified version, $\mathcal{L}_{\text{SIWELBO}}$ (Shi et al., 2019, Appendix A; Morningstar et al., 2021):

$$\mathcal{L}^I_{\text{IWELBO}}(\boldsymbol{x}_{\text{obs}}) = \mathbb{E}_{\{\boldsymbol{z}_j\}_{j=1}^I \sim q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}_{\text{obs}})} \left[\log \frac{1}{I} \sum_{j=1}^{I} w(\boldsymbol{z}_j) \right], \quad \text{and} \tag{7}$$

$$\mathcal{L}^I_{\text{SIWELBO}}(\boldsymbol{x}_{\text{obs}}) = \mathbb{E}_{\{\{\boldsymbol{z}_j^k\}_{j=1}^I \sim q_{\boldsymbol{\phi}}^k(\boldsymbol{z}|\boldsymbol{x}_{\text{obs}})\}_{k=1}^K} \left[\log \sum_{k=1}^{K} q_{\boldsymbol{\phi}}(k \mid \boldsymbol{x}_{\text{obs}}) \frac{1}{I} \sum_{j=1}^{I} w(\boldsymbol{z}_j^k) \right], \tag{8}$$

[8]

where I is the number of importance samples in $\mathcal{L}_{\text{IWELBO}}$ and the number of samples per-mixture-component in $\mathcal{L}_{\text{SIWELBO}}$.[9]

When the number of mixture-components is $K = 1$ the lower-bounds in eqs. (4-5) and eqs. (7-8) correspond to the MVAE and MIWAE bounds in Mattei & Frellsen (2019) which are among the most popular bounds for fitting VAEs from incomplete data. However, as $K > 1$ the proposed bounds can be tighter due to an increased flexibility of the variational distribution $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ (Morningstar et al., 2021, Appendix A), which potentially mitigates the problems caused by the missing data (see section 3). Finally, the importance-weighted bounds in eqs. (7) and (8) maintain the asymptotic consistency guarantees of Burda et al. (2015) and approaches the true marginal log-likelihood $\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}})$ as $K \cdot I \to \infty$, allowing for more accurate estimation of the model with increasing computational budget.

We denote the four methods based on eqs. (4), (5), (7) and (8) by **MissVAE**, **MissSVAE**, **MissIWAE**, and **MissSIWAE** respectively. In practice, the above objectives are approximated using Monte Carlo integration, and when optimising the variational distribution we have used "sticking-the-landing" (STL) gradients (Roeder et al., 2017) to reduce gradient variance.[10]

[7] Stratified sampling of mixture distributions typically draws an equal number of samples from each component and weighs the samples by the component probabilities $q_{\boldsymbol{\phi}}(k \mid \boldsymbol{x}_{\text{obs}})$ when estimating expectations. It is commonly used to reduce Monte Carlo variance (Robert & Casella, 2004).

[8] In multimodal-domain VAE literature, Shi et al. (2019) proposed a looser bound related to $\mathcal{L}_{\text{SIWELBO}}$:

$$\mathcal{L}^I_{\text{SIWELBO}}(\boldsymbol{x}_{\text{obs}}) \geq \tilde{\mathcal{L}}^I_{\text{SIWELBO}}(\boldsymbol{x}_{\text{obs}}) \overset{\text{def}}{=} \sum_{k=1}^{K} q_{\boldsymbol{\phi}}(k \mid \boldsymbol{x}_{\text{obs}}) \mathbb{E}_{\{\boldsymbol{z}_j^k\}_{j=1}^I \sim q_{\boldsymbol{\phi}}^k(\boldsymbol{z}|\boldsymbol{x}_{\text{obs}})} \left[\log \frac{1}{I} \sum_{j=1}^{I} w(\boldsymbol{z}_j^k) \right] \overset{I=1}{=} \mathcal{L}_{\text{SELBO}}(\boldsymbol{x}_{\text{obs}}),$$

and empirically showed that it may alleviate potential mixture collapse to a subset of the mixture components. Therefore, the looser bound may be useful when variational mixture collapse is observed.

[9] Note that in contrast to the ELBO objectives, the two importance-weighted objectives $\mathcal{L}_{\text{IWELBO}}$ and $\mathcal{L}_{\text{SIWELBO}}$ for any finite budget may have different expected values and variances. Our findings indicate that preference for either objective may be problem-dependent and hence both objectives should be evaluated when budget allows.

[10] We have also evaluated the doubly-reparametrised gradients (DReG, Tucker et al., 2018) for IWELBO objectives but found STL to perform similar or slightly better.

4.2 Using imputation-mixture distributions to fit VAEs from incomplete data

In section 4.1, we jointly dealt with both the inference of the latents $\boldsymbol{z}$ (section 2) and the posterior complexity increase due to missing data (section 3) by learning a finite-mixture variational distribution. Here, we propose a second "decomposed" approach to deal with the complexities of missing data.

Intuitively, if we had an oracle that were able to generate imputations of the missing data from the ground truth conditional distribution $p^*(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$, then the VAE estimation task would reduce to the case of complete-data. This suggests that an *effective strategy is to decompose the task of model estimation from incomplete data into two (iterative) tasks: (i) data imputation and (ii) model estimation*, akin to the Monte Carlo EM algorithm (Wei & Tanner, 1990; Dempster et al., 1977). However, access to the oracle $p^*(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ is unrealistic and the exact sampling of $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$, as required in EM, is generally intractable. To address this, we resort to (i) approximate but computationally cheap conditional sampling methods for VAEs to generate imputations (Rezende et al., 2014; Mattei & Frellsen, 2018a; Simkus & Gutmann, 2023) and (ii) separate learning objectives for the model $p_{\boldsymbol{\theta}}$ and the variational distribution $q_{\boldsymbol{\phi}}$ to compensate for potential sampling errors. We call the proposed approach **DeMissVAE** (**de**composed approach for handling **miss**ing data in **VAE**s).

We construct the variational distribution $q_{\boldsymbol{\phi},f^t}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ for an incomplete data-point $\boldsymbol{x}_{\text{obs}}$ using a completed-data variational distribution $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$ and an (approximate) imputation distribution $f^t(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}) \approx p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$:

$$q_{\boldsymbol{\phi},f^t}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}) = \mathbb{E}_{f^t(\boldsymbol{x}_{\text{mis}}|\boldsymbol{x}_{\text{obs}})}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})\right]. \tag{9}$$

The intuition for this construction comes from the decomposition of the model posterior $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}) = \mathbb{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}}|\boldsymbol{x}_{\text{obs}})}[p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})]$. Assuming that the completed-data variational distribution $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$ well-represents the model posterior $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$, and that the imputation distribution $f^t(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ draws plausible imputations of the missing variables, then $q_{\boldsymbol{\phi},f^t}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ will reasonably represent $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ (see the two right-most columns of fig. 1). A more technical justification is provided below, in the paragraph after eq. (13). In contrast to section 4.1 we here use a continuous-mixture variational distribution, which is more flexible than a finite-mixture distribution, albeit at an extra computational cost due to sampling the (approximate) imputations (see appendix D).

We now derive the DeMissVAE objectives for fitting the generative model $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ and the completed-data variational distribution $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$, see appendix C for a more in-depth treatment.

Objective for $p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z})$. With the variational distribution in eq. (9), we derive an ELBO on the marginal log-likelihood, similar to eq. (2), to learn the parameters $\boldsymbol{\theta}$ of the generative model:

$$\begin{aligned}\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}}) &\geq \mathbb{E}_{f^t(\boldsymbol{x}_{\text{mis}}|\boldsymbol{x}_{\text{obs}})q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}_{\text{obs}},\boldsymbol{x}_{\text{mis}})}\left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}}, \boldsymbol{z})}{\mathbb{E}_{f^t(\boldsymbol{x}_{\text{mis}}|\boldsymbol{x}_{\text{obs}})}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})\right]}\right] \\ &= \underbrace{\mathbb{E}_{f^t(\boldsymbol{x}_{\text{mis}}|\boldsymbol{x}_{\text{obs}})q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}_{\text{obs}},\boldsymbol{x}_{\text{mis}})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}}, \boldsymbol{z})\right]}_{\stackrel{\text{def}}{=}\mathcal{L}^{\boldsymbol{\theta}}_{\text{CVI}}(\boldsymbol{x}_{\text{obs}};\boldsymbol{\phi},\boldsymbol{\theta},f^t)} + \underbrace{\mathcal{H}\left[q_{\boldsymbol{\phi},f^t}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})\right]}_{\text{Const. w.r.t. } \boldsymbol{\theta}}.\end{aligned} \tag{10}$$

This lower-bound can be further decomposed into log-likelihood and KL divergence terms

$$\mathcal{L}^{\boldsymbol{\theta}}_{\text{CVI}}(\boldsymbol{x}_{\text{obs}}; \boldsymbol{\phi}, \boldsymbol{\theta}, f^t) + \mathcal{H}\left[q_{\boldsymbol{\phi},f^t}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})\right] = \log p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}}) - D_{\text{KL}}(q_{\boldsymbol{\phi},f^t}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}) \,||\, p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})), \tag{11}$$

which means that if $q_{\boldsymbol{\phi},f^t}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}) \approx p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ then maximising eq. (10) w.r.t. $\boldsymbol{\theta}$ performs approximate maximum-likelihood estimation. Importantly, the missing variables $\boldsymbol{x}_{\text{mis}}$ are marginalised-out, which adds robustness to the potential sampling errors in $f^t(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$.

Objective for $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x})$. We obtain the objective for learning the variational distribution $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x})$ by marginalising the missing variables $\boldsymbol{x}_{\text{mis}}$ from the complete-data ELBO in eq. (2) and then lower-bounding

79

the integral using $f^t(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ (see appendix B):

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}}) \geq \underbrace{\mathbb{E}_{f^t(\boldsymbol{x}_{\text{mis}}|\boldsymbol{x}_{\text{obs}})q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}_{\text{obs}},\boldsymbol{x}_{\text{mis}})}\left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}, \boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})}\right]}_{\stackrel{\text{def}}{=}\mathcal{L}^{\boldsymbol{\phi}}_{\text{LMVB}}(\boldsymbol{x}_{\text{obs}};\boldsymbol{\phi},\boldsymbol{\theta},f^t)} + \underbrace{\mathcal{H}\left[f^t(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})\right]}_{\text{Const. w.r.t. } \boldsymbol{\phi}}. \tag{12}$$

This lower-bound can also be decomposed into the log-likelihood term and two KL divergence terms

$$\begin{aligned}\mathcal{L}^{\boldsymbol{\phi}}_{\text{LMVB}}(\boldsymbol{x}_{\text{obs}};\boldsymbol{\phi},\boldsymbol{\theta},f^t) + \mathcal{H}\left[f^t(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})\right] = \log p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}}) - D_{\text{KL}}(f^t(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}) \,||\, p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})) \\ - \mathbb{E}_{f^t(\boldsymbol{x}_{\text{mis}}|\boldsymbol{x}_{\text{obs}})}\left[D_{\text{KL}}(q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}) \,||\, p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}))\right],\end{aligned} \tag{13}$$

which means that the bound is maximised w.r.t. $\boldsymbol{\phi}$ iff $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}) = p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$ for all $\boldsymbol{x}_{\text{mis}}$. Therefore, using the above objective to fit $q_{\boldsymbol{\phi}}$ corresponds directly to the complete-data case, and hence avoids having to approximate complex posteriors that arise due to missing data (see 3).

If $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}) = p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$ for all $\boldsymbol{x}_{\text{mis}}$, then maximising either of the bounds in eqs. (10) or (12) w.r.t. the imputation distribution $f^t(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ would correspond to setting $f^t(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}) = p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$. However, directly learning an imputation distribution $f^t(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}) \approx p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ is challenging (Simkus et al., 2023, Section 2.2). This motivates using sampling methods to approximate the optimal imputation distribution $f^t(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}) \approx p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ with samples. We draw samples from $f^t(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ using (cheap) approximate conditional sampling methods for VAEs to obtain K imputations $\{\boldsymbol{x}^k_{\text{mis}}\}^K_k$ and then use them to approximate the expectations w.r.t. $f^t(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ in the above objectives. We discuss the implementation of the algorithm in detail in appendix D.

Finally, we note that the $\mathcal{L}^{\boldsymbol{\theta}}_{\text{CVI}}$ and $\mathcal{L}^{\boldsymbol{\phi}}_{\text{LMVB}}$ objectives in eqs. (10) and (12) are based on the standard ELBO. Extensions to the importance-weighted ELBO might improve the method further by increasing the flexibility of the variational posterior. However, unlike the standard ELBO used in eq. (10) where the density of the imputation-based variational-mixture $q_{\boldsymbol{\phi},f^t}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ can be dropped, IWELBO requires computing the density of the proposal distribution $q_{\boldsymbol{\phi},f^t}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$, which is generally intractable. We hence leave this direction for future work.

5 Related work

Fitting VAEs from incomplete data. Since the seminal works of Kingma & Welling (2013) and Rezende et al. (2014), VAEs have been widely used for density estimation from incomplete data and various downstream tasks, primarily due to the computationally efficient marginalisation of the model in eq. (1). Vedantam et al. (2017) and Wu & Goodman (2018) explored the use of product-of-experts variational distributions, drawing inspiration from findings in the factor analysis case with incomplete data (Williams et al., 2018). Mattei & Frellsen (2019) used the importance-weighted ELBO (Burda et al., 2015) for training VAEs on incomplete training data sets. Ma et al. (2019) proposed the use of permutation invariant neural networks to parametrise the encoder network instead of relying on zero-masking. Nazábal et al. (2020) introduced hierarchical priors to handle incomplete heterogeneous training data. Simkus et al. (2023) proposed a general-purpose approach that is applicable to VAEs, not requiring the decoder distribution to be easily marginalisable. Here, we further develop the understanding of VAEs in the presence missing values in the training data set, and propose variational-mixtures as a natural approach to improve VAE estimation from incomplete data, building upon the motivation from imputation-mixtures discussed in section 3.

Variational mixture distributions. Mixture distributions have found widespread application in variational inference and VAE literature. Roeder et al. (2017) introduced the stratified ELBO corresponding to eq. (5). In the context of VAEs in multimodal domains, Shi et al. (2019, Appendix A) introduced the stratified IWELBO corresponding to eq. (8), but opted to use a looser bound instead, see footnote 8. These bounds were subsequently rediscovered by Morningstar et al. (2021) and Kviman et al. (2023), who investigated their use for VAE estimation in fully-observed data scenarios. Figurnov et al. (2019) introduced

80

implicit reparametrisation, enabling gradient estimation for ancestrally-sampled mixtures, allowing the estimation of variational mixtures using eqs. (4) and (7). Here, we build on this prior work, asserting that variational-mixtures are well-suited for handling the posterior complexity increase due to missing data (see section 3). Moreover, the imputation-mixture distribution used in DeMissVAE is a novel type of variational mixtures specifically designed for incomplete data scenarios.

Posterior complexity increase due to missing data. Concurrent to this study, Sudak & Tschiatschek (2023) have brought attention to a phenomenon related to the increase in posterior complexity due to incomplete data, discussed in section 3. They noted that, for any $\boldsymbol{x}_{\text{obs}}$ and $\boldsymbol{x}_{\text{obs}\setminus u}$, where u is a subset of the observed dimensions, the model posteriors $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ and $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}\setminus u})$ should exhibit a strong dependency. However, because of the approximations in the variational posterior (see e.g. Cremer et al., 2018; Zhang et al., 2021), the variational approximations $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ and $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}\setminus u})$ may not consistently capture this dependency. They refer to the lack of dependency between $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ and $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}\setminus u})$, compared to $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ and $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}\setminus u})$, as posterior inconsistency. Focused on improving downstream task performance, they introduce regularisation into the VAE training objective to address posterior inconsistency. In contrast to their work, we compare the fully-observed and incomplete-data posteriors, $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x})$ and $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$, respectively. Observing that the incomplete-data posterior $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ can be expressed as a mixture of fully-observed posteriors $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x})$, that is, $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}) = \mathbb{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})}\left[p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x})\right]$, we propose using variational-mixtures to improve the match between the variational and model posteriors when dealing with incomplete data in order to improve model estimation performance.

Marginalised variational bound. In the standard ELBO derivation for incomplete data in eq. (2) the missing variables are first marginalised (collapsed) from the likelihood, and then a variational ELBO is established. This approach is sometimes referred to as collapsed variational inference (CVI). In contrast, in the derivation of the DeMissVAE encoder objective in eq. (12) we swap the order of marginalisation and variational inference. Specifically, we start with the variational ELBO on completed-data, and then marginalise the missing variables (see appendix B). This approach bears similarity to the marginalised variational bound (MVB, or KL-corrected bound) in exponential-conjugate variational inference literature (King & Lawrence, 2006; Lázaro-Gredilla & Titsias, 2011; Hensman et al., 2012). In these works, MVB has been preferred over CVI due to improved convergence and guarantees that, for appropriately formulated conjugate models, MVB is analytically tractable in cases where CVI is not (Hensman et al., 2012, Section 3.3). While MVB remains intractable in the VAE setting with incomplete data, similar to how the standard ELBO is intractable in fully-observed case, we find the motivation behind MVB and DeMissVAE to be similar.

6 Evaluation

We here evaluate the proposed methods, MissVAE, MissSVAE, MissIWAE, MissSIWAE (section 4.1), and DeMissVAE (section 4.2), on synthetic and real-world data, and compare them to the popular methods MVAE and MIWAE that do not use mixture variational distributions (Mattei & Frellsen, 2019). The methods are summarised in table 1 and the code implementation is available at `https://github.com/vsimkus/demiss-vae`.

6.1 Mixture-of-Gaussians data with a 2D latent VAE

Evaluating log-likelihood on held-out data is generally intractable for VAEs due to the intractable integral in eq. (1). We hence here choose a VAE with 2D latent space, where numerical integration can be used to estimate the log-likelihood of the model accurately (see appendix E.1 for more details). We fit the model on incomplete data drawn from a mixture-of-Gaussians distribution. By introducing uniform missingness of 50% in the mixture-of-Gaussians data we introduce multi-modality in the latent space (see fig. 1), which allows us to verify the efficacy of mixture-variational distributions when the posteriors are multi-modal due to missing data.

Results are shown in fig. 2. We first note that the stratified MissSVAE approach performed better than MissVAE that uses ancestral sampling. The reason for this is likely that stratified sampling reduces Monte

Method	p_θ objective	q_ϕ objective	# of components	Mixture sampling
MVAE†	eq. (4)	eq. (4)	$K = 1$	—
MissVAE	eq. (4)	eq. (4)	$K > 1$	Ancestral
MissSVAE	eq. (5)	eq. (5)	$K > 1$	Stratified
MIWAE†	eq. (7)	eq. (7)	$K = 1$	—
MissIWAE	eq. (7)	eq. (7)	$K > 1$	Ancestral
MissSIWAE	eq. (8)	eq. (8)	$K > 1$	Stratified
DeMissVAE	eq. (10)	eq. (12)	$K > 1$	Conditional VAE

Table 1: *Summary of the proposed and baseline methods.* The non-mixture baselines (†) are based on Mattei & Frellsen (2019) and the other methods are proposed in this paper. Moreover, the methods using ancestral sampling require implicit reparametrisation (Figurnov et al., 2019), whereas the other methods work with the standard reparametrisation trick.

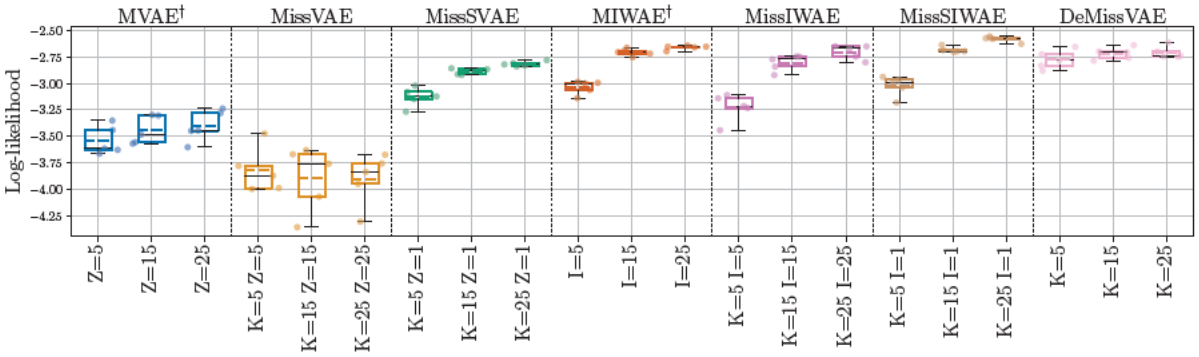


Figure 2: *Log-likelihood on held out data evaluated by numerically integrating the 2D latent variables.* VAEs were fitted on mixture-of-Gaussians data with 50% missingness. Each model is fitted with a computational budget of 5/15/25 samples from the variational distribution. The box plots show 1st and 3rd quartiles, the black lines are the medians, the dashed lines are the means, and the whiskers show the data range over 5 independent runs. MVAE and MIWAE (†) are baseline methods by Mattei & Frellsen (2019). The other five methods are proposed in this paper.

Carlo variance of the gradients w.r.t. ϕ and hence enables a better fit of the variational distribution $q_\phi(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ (see a further investigation in appendix F.1.1). In line with this intuition, the MissVAE results exhibit significantly larger variance than MissSVAE. Similarly, we observe that the stratified MissSIWAE approach performed better than MissIWAE. Importantly, we see that the use of mixture variational distributions in MissSVAE and MissSIWAE improve the model fit over the MVAE and MIWAE baselines that do not use mixtures to deal with the increased posterior complexity due to missingness. Finally, we observe that DeMissVAE is capable of achieving comparable performance to MIWAE and MissSIWAE, despite using a looser ELBO bound, which shows that the decomposed approach to handling data missingness can be used to achieve an improved fit of the model.

In appendix F.1.2, we analyse the model and variational posteriors of the learnt models. We observe that the mixture approaches better-approximate the incomplete-data posteriors, compared to the approaches that do not use variational-mixtures. Moreover, we also observe that the structure of the latent space is better-behaved when fitted using the decomposed approach in DeMissVAE.

6.2 Real-world UCI data sets

We here evaluate the proposed methods on real-world data sets from the UCI repository (Dua & Graff, 2017; Papamakarios et al., 2017). We train a VAE model with ResNet architecture on incomplete data sets with

82

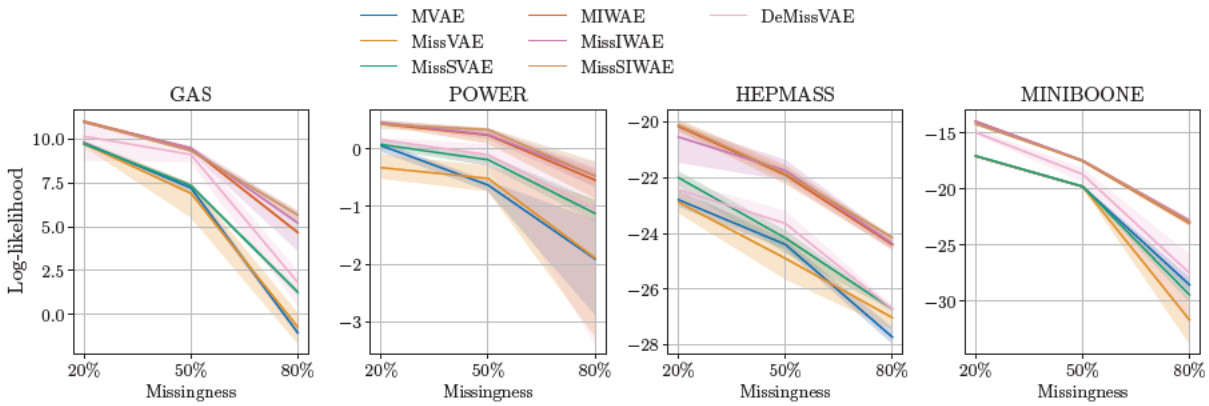


Figure 3: *Estimate of the test log-likelihood using the IWELBO with $I = 50000$, on four UCI data sets.* Each data set was rendered incomplete by applying uniform missingness of 20/50/80%. The curves show average performance over 5 independent runs of the algorithms and the intervals show the 90% centered interval.

20/50/80% uniform missingness (see appendix E.2 for more details). We then estimate the log-likelihood on complete test data set using the IWELBO bound with $I = 50$K importance samples.[11] For additional metrics see appendix F.2.

The results are shown in fig. 3. We first note that, similar to before, the stratified MissSVAE approach performed better than MissVAE which uses ancestral sampling. Importantly, we observe that using mixture variational distributions in MissSVAE improves the fit of the model over MVAE (with the exception on the Miniboone data set) that uses non-mixture variational distributions. Furthermore, the gains in model accuracy typically increase with data missingness, which verifies that MissSVAE performs better because it handles the increased posterior complexity due to missing data better (see fig. 1). Next, we observe that the performance of MIWAE, MissIWAE, and MissSIWAE is similar, although we can note a small improvement by using MissIWAE and MissSIWAE in large missingness settings. We observe only a relatively small difference between the IWAE methods because the use of importance weighted bound already corresponds to using a more flexible semi-implicitly defined variational distribution (Cremer et al., 2017), which here seems to be sufficient to deal with the complexities arising due to missingness. Finally, we note that DeMissVAE results are in-between MissSVAE and MIWAE. This verifies that the decomposed approach can be used to deal with data missingness and, as a result, can improve the fit of the model. Nonetheless, DeMissVAE is surpassed by the IWAE methods, which is likely due to using the ELBO in DeMissVAE versus IWELBO in IWAE methods that can tighten the bound more effectively.

6.3 MNIST and Omniglot data sets

In this section we evaluate the proposed methods on binarised MNIST (Garris et al., 1991) and Omniglot (Lake et al., 2015) data sets of handwritten characters. We fit a VAE model with a convolutional ResNet encoder and decoder networks (see appendix E.3 for more details). The data is made incomplete by masking 2 out of 4 quadrants of an image at random. Similar to the previous section, we estimate the log-likelihood on a complete test data set using the IWELBO bound with $I = 1000$ importance samples. In appendix F.3 we report additional results with varying dimensionality of the latent variables.

On the MNIST data set we see that MVAE $\leq$ MissVAE $<$ MissSVAE similar to the previous results but MIWAE $<$ MissSIWAE $<$ MissIWAE. This suggests that MissIWAE, which uses ancestral sampling, was able to tighten the bound more effectively compared to stratified MissSIWAE, and was able to fit the variational distribution $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ well despite the potentially larger variance w.r.t. $\boldsymbol{\phi}$. Moreover, we also see that

[11] As $I \to \infty$ IWELBO approaches $\log p_{\boldsymbol{\theta}}(x)$. Moreover, as suggested by Mattei & Frellsen (2018b), to improve the estimate on held-out data we fine-tune the encoder on complete test data before estimating the log-likelihood.

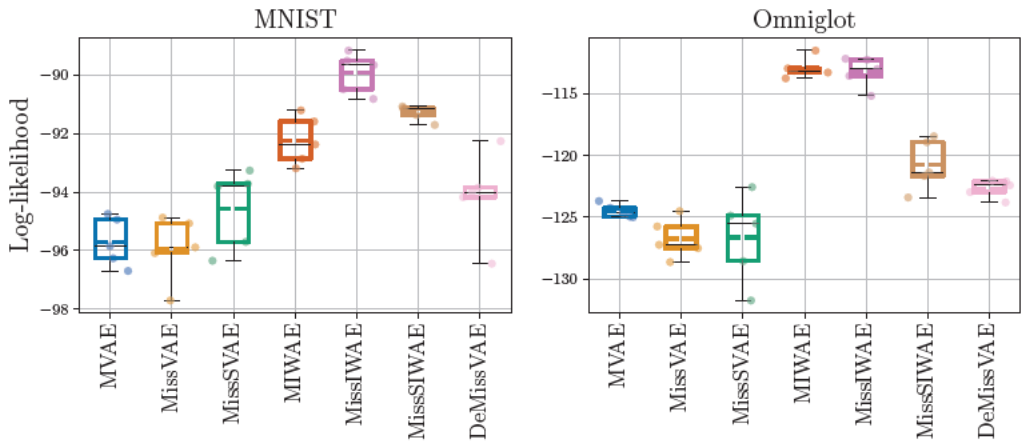


Figure 4: *Estimate of the test log-likelihood using the IWELBO with $I = 1000$, MNIST and Omniglot data sets.* Each image in the training data set was missing 2 out of 4 random quadrants. The box plots show 1st and 3rd quartiles, the black lines are the medians, the dashed lines are the means, and the whiskers show the data range over 5 independent runs.

MVAE < DeMissVAE < MIWAE, which further verifies that the decomposed approach is able to handle the data missingness well.

On the Omniglot data we observe that the mixture approaches perform similarly to MVAE and MIWAE, which do not use mixture variational distributions. This suggests that either the posterior multi-modality is less prominent in the Omniglot data set or that due to the reverse KL optimisation of the variational distribution all mixture components have degenerated to a single mode. Finally, DeMissVAE slightly outperforms MVAE, MissVAE, and MissSVAE, but is surpassed by the importance-weighted approaches.

Interestingly, in this evaluation the stratified approaches (MissSVAE and MissSIWAE) were outperformed by the approaches using standard ELBO and implicit reparametrisation (MissVAE and MissIWAE). This suggests that the performance of each approach can be data- and model-dependent and hence both should be evaluated when possible.

7 Discussion

Handling missing data is a key challenge in modern machine learning, as many real-world applications involve incomplete data. In the context of variational autoencoders, we have shown that incomplete data increases the complexity of the latent variables' posterior distribution. Therefore, accurately fitting models from incomplete data requires more flexible variational families than in the complete-data case. We stipulated that variational-mixtures are a natural approach for handling missing data. One benefit is that it allows us to work with the same variational families as in the fully-observed case, which enables the transfer of useful known inductive biases (Miao et al., 2022) from the fully-observed to the incomplete data scenario.

Subsequently, we have introduced two methodologies grounded in variational mixtures. First, we proposed using finite variational mixtures with the standard and importance-weighted ELBOs using ancestral and stratified sampling of the mixtures. Second, we have proposed a novel "decomposed" variational-mixture approach, that uses cost-effective yet often coarse conditional sampling methods for VAEs to generate imputations and ELBO-based objectives that are robust to the sampling errors.

Our evaluation shows that using variational mixtures can improve the fit of VAEs when dealing with incomplete data, surpassing the performance of models without variational mixtures. Moreover, our observations indicate that, although stratified sampling of the finite mixtures often yields better results compared to ancestral sampling, the effectiveness of these methods can be data- and model-dependent and hence both

84

Published in Transactions on Machine Learning Research (06/2024)

Method	Budget	Variational families	Latent structure*	Evaluation rank		
				MoG	*UCI*	*MNIST+Omniglot*
MissVAE	*small*	*limited*	*well-behaved*	5	5	5
MissSVAE	*medium*	*any*	*well-behaved*	4	4	4
MissIWAE	*medium*	*limited*	*potentially irregular*	3	2	1
MissSIWAE	*medium*	*any*	*potentially irregular*	1	1	2
DeMissVAE	*medium/high*	*any*	*well-behaved*$^+$	2	3	3

Table 2: *A coarse summary of advantages and disadvantages of the proposed methods. Budget*: small/medium/high depending on the number of latent samples required or whether conditional sampling of VAEs is needed. *Variational families*: which families of distributions can be used as mixture components—any reparametrisable families, or limited families, as discussed in section 4.1. *Latent structure*: methods with potential to learn irregular latent spaces may have decreased downstream performance in certain tasks. We have found (+) that DeMissVAE is able to achieve the most well-behaved latent structures on the MoG data in appendix F.1.2. Please note (*) that the learnt latent structure will depend on the chosen model architecture. *Evaluation rank*: the rank of the proposed methods in the evaluations in sections 6.1 to 6.3.

approaches should be evaluated when possible. Our results further indicate that variational mixtures provide relatively little improvement with the IWELBO-based methods compared the ELBO-based methods. We believe that this is mainly because IWELBO can be seen to be working with semi-implicitly defined variational distributions that are flexible enough to handle the posterior complexity increase due to missing data. Alternatively, this may be related to an observation by Shi et al. (2019, Appendix A) that the reverse-KL formulation of the importance-weighted bound may lead to situations where the mixture components collapse to a single mode. Hence, a future direction would be to investigate alternative formulations of importance-weighted bounds that avoid the mode-seeking nature (Bornschein & Bengio, 2015; Mnih & Rezende, 2016; Wan et al., 2020). Furthermore, we note that the decomposed approach in DeMissVAE outperforms all ELBO-based methods but falls short of surpassing IWELBO-based methods. These results point towards promising research avenues, suggesting potential improvements in VAE model estimation from incomplete data. Future directions include extending the DeMissVAE approach to incorporate IWELBO-based objectives and developing improved cost-effective conditional sampling methods for VAEs.

The choice between the proposed methods for fitting VAEs from incomplete data depends on various factors such as computational budget, variational families, model accuracy goals, and the specific requirements of downstream tasks, discussed next and summarised in table 2.

Computational and memory budget. The standard ELBO with ancestral sampling is the most suitable method for small computational and memory budgets, since the objective can be estimated using a single latent sample for each data point. On the other hand, methods using stratified sampling or the importance-weighted ELBO require multiple latent samples for each data-point and hence may only be used if the memory and compute budget allows. Moreover, for a fixed budget, stratified approaches may limit the number of components K that may be used. Lastly, akin to the standard ELBO, the DeMissVAE objectives can be estimated using a single latent sample, but the approach incurs extra cost in sampling the imputations.

Variational families. While the stratified and DeMissVAE approaches can use any reparametrisable distribution family for the mixture components, the ancestral sampling methods require the use of *implicit* reparametrisation (Figurnov et al., 2019) and as a result may not work with all distribution families (see discussion in section 4.1).

Model accuracy. Stratified sampling of mixtures can improve the model accuracy, compared to ancestral sampling, by reducing Monte Carlo gradient variance. Additionally, methods using the importance-weighted ELBO, compared to the standard ELBO, are often able to tighten the bound more effectively by using multiple importance samples, leading to improved model accuracy. DeMissVAE

85

performance lies in between the standard ELBO and importance-weighted ELBO approaches. Although the introduced DeMissVAE objectives exhibit robustness to some imputation distribution error, improved model accuracy can often be achieved by improving the accuracy of imputations by using a larger budget for the imputation step.

Latent structure. Different downstream tasks may prefer distinct latent structures, for example, conditional generation from unconditional VAEs is often easier if the latent space is well-structured (Engel et al., 2017; Gómez-Bombarelli et al., 2018). To this end, observations in appendix F.1.2 show that the latent space of DeMissVAE behaves well, and is comparable to a model fitted with complete data. This characteristic makes it preferable for downstream tasks requiring well-structured latent spaces. On the other hand, as noted by Burda et al. (2015, Appendix C) and Cremer et al. (2018, Section 5.4), the use of importance-weighted ELBO to mitigate the increased posterior complexity due to missing data may make the latent space less regular, compared to a model trained on fully-observed data set, which potentially decreases the model's performance on downstream tasks.

Finally, we step back to note that this paper is focused on the class of variational autoencoder models, a subset of the broader family of deep latent variable models (DLVMs). Much like VAEs, DLVMs usually aim to efficiently represent the intricate nature of data through a well-structured latent space, implicitly defined by a learnable generative process. Building on our findings in VAEs, where incomplete data led to an increased complexity in the posterior distribution compared to the fully-observed case, we conjecture that a similar effect may occur within the wider family of DLVMs, affecting the fit of the model. We therefore believe that there is substantial scope to explore the implications of incomplete data in other DLVM classes, particularly focusing on the effects of marginalisation on latent space representations and the associated generative processes. Investigating decomposed approaches, similar to DeMissVAE or Monte Carlo EM (Wei & Tanner, 1990), presents promising avenues for further research in this direction.

References

Jörg Bornschein and Yoshua Bengio. Reweighted Wake-Sleep. In *International Conference on Learning Representations (ICLR)*, April 2015. doi: 10.48550/arXiv.1406.2751. (Cited on 12)

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance Weighted Autoencoders. In *International Conference on Learning Representations (ICLR)*, San Juan, Puerto Rico, September 2015. (Cited on 5, 7, 13, 17, 20, 25)

Chris Cremer, Quaid Morris, and David Duvenaud. Reinterpreting Importance-Weighted Autoencoders. In *ICLR Workshop*, February 2017. (Cited on 10, 17, 25, 27)

Chris Cremer, Xuechen Li, and David Duvenaud. Inference Suboptimality in Variational Autoencoders. In *International Conference on Machine Learning (ICML)*, May 2018. (Cited on 8, 13, 25)

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977. doi: 10.1111/j.2517-6161.1977.tb01600.x. (Cited on 6)

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using Real NVP. In *International Conference on Learning Representations (ICLR)*, February 2017. (Cited on 1)

Dheeru Dua and Casey Graff. UCI Machine Learning Repository, 2017. (Cited on 9, 23)

Jesse Engel, Matthew Hoffman, and Adam Roberts. Latent Constraints: Learning to Generate Conditionally from Unconditional Generative Models. In *International Conference on Learning Representations (ICLR)*, December 2017. (Cited on 13)

Michael Figurnov, Shakir Mohamed, and Andriy Mnih. Implicit Reparameterization Gradients. In *Advances in Neural Information Processing Systems (NeurIPS)*, January 2019. (Cited on 4, 7, 9, 12)

86

Michael D. Garris, R. A. Wilkinson, and Charles L. Wilson. Methods for enhancing neural network handwritten character recognition. In *International Joint Conference on Neural Networks (IJCNN)*, volume 1, pp. 695–700, Seattle, WA, USA, 1991. IEEE. ISBN 978-0-7803-0164-1. doi: 10.1109/IJCNN.1991.155265. (Cited on 10)

Samuel J. Gershman and Noah D. Goodman. Amortized Inference in Probabilistic Reasoning. In *Annual Meeting of the Cognitive Science Society*, volume 36, 2014. (Cited on 2)

Rafael Gómez-Bombarelli, Jennifer N. Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D. Hirzel, Ryan P. Adams, and Alán Aspuru-Guzik. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Central Science*, 4(2):268–276, February 2018. ISSN 2374-7943. doi: 10.1021/acscentsci.7b00572. (Cited on 13)

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, June 2014. (Cited on 1)

James Hensman, Magnus Rattray, and Neil D. Lawrence. Fast Variational Inference in the Conjugate Exponential Family. In *Advances in Neural Information Processing Systems (NeurIPS)*, December 2012. (Cited on 8)

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. (Cited on 28)

Niels Bruun Ipsen, Pierre-Alexandre Mattei, and Jes Frellsen. Not-MIWAE: Deep Generative Modelling with Missing not at Random Data. In *International Conference on Learning Representations (ICLR)*, June 2020. (Cited on 17)

Nathaniel J. King and Neil D. Lawrence. Fast Variational Inference for Gaussian Process Models Through KL-Correction. In *European Conference on Machine Learning (ECML)*, 2006. (Cited on 8)

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations (ICLR)*, December 2013. (Cited on 1, 4, 7)

Rahul G. Krishnan, Uri Shalit, and David Sontag. Structured Inference Networks for Nonlinear State Space Models. In *AAAI Conference on Artificial Intelligence*, December 2016. doi: 10.48550/arXiv.1609.09869. (Cited on 1)

Oskar Kviman, Ricky Molén, Alexandra Hotti, Semih Kurt, Víctor Elvira, and Jens Lagergren. Cooperation in the Latent Space: The Benefits of Adding Mixture Components in Variational Autoencoders. In *International Conference on Machine Learning (ICML)*, July 2023. doi: 10.48550/arXiv.2209.15514. (Cited on 7)

Brenden M. Lake, Ruslan Salakhutdinov, and Joshua B. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 2015. doi: 10.1126/science.aab3050. (Cited on 10, 23)

Miguel Lázaro-Gredilla and Michalis K. Titsias. Variational heteroscedastic Gaussian process regression. In *International Conference on Machine Learning (ICML)*, June 2011. (Cited on 8)

Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data: Second Edition*. Wiley-Interscience, 2002. ISBN 0-471-18386-5. (Cited on 2)

Chao Ma and Cheng Zhang. Identifiable Generative Models for Missing Not at Random Data Imputation. In *Advances in Neural Information Processing Systems (NeurIPS)*, October 2021. (Cited on 17)

Chao Ma, Sebastian Tschiatschek, Konstantina Palla, José Miguel Hernández-Lobato, Sebastian Nowozin, and Cheng Zhang. EDDI: Efficient dynamic discovery of high-value information with partial VAE. In *International Conference on Machine Learning (ICML)*, pp. 7483–7504, 2019. ISBN 9781510886988. (Cited on 3, 7)

87

Pierre-Alexandre Mattei and Jes Frellsen. Leveraging the Exact Likelihood of Deep Latent Variable Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, February 2018a. (Cited on 4, 6, 20, 23)

Pierre-Alexandre Mattei and Jes Frellsen. Refit your Encoder when New Data Comes by. In *Workshop on Bayesian Deep Learning at Neural Information Processing Systems (NeurIPS)*, pp. 4, Montreal, Canada, 2018b. (Cited on 10)

Pierre-Alexandre Mattei and Jes Frellsen. MIWAE: Deep Generative Modelling and Imputation of Incomplete Data Sets. In *International Conference on Machine Learning (ICML)*, 2019. (Cited on 2, 5, 7, 8, 9, 17)

Xiao-Li Meng. On the Rate of Convergence of the ECM Algorithm. *The Annals of Statistics*, 22(1):326–339, March 1994. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1176325371. (Cited on 18)

Ning Miao, Emile Mathieu, N. Siddharth, Yee Whye Teh, and Tom Rainforth. On Incorporating Inductive Biases into VAEs. In *International Conference on Learning Representations (ICLR)*, February 2022. doi: 10.48550/arXiv.2106.13746. (Cited on 2, 11)

Andriy Mnih and Danilo J. Rezende. Variational inference for Monte Carlo objectives. *arXiv:1602.06725 [cs, stat]*, June 2016. (Cited on 12)

Warren Morningstar, Sharad Vikram, Cusuh Ham, Andrew Gallagher, and Joshua Dillon. Automatic Differentiation Variational Inference with Mixtures. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 3250–3258. PMLR, March 2021. (Cited on 4, 5, 7)

Alfredo Nazábal, Pablo M. Olmos, Zoubin Ghahramani, and Isabel Valera. Handling Incomplete Heterogeneous Data using VAEs. *Pattern Recognition*, 107, 2020. ISSN 0031-3203. doi: 10.1016/j.patcog.2020.107501. (Cited on 2, 7)

George Papamakarios, Theo Pavlakou, and Iain Murray. Masked Autoregressive Flow for Density Estimation. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017. (Cited on 9, 23)

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing Flows for Probabilistic Modeling and Inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021. (Cited on 5)

Tom Rainforth, Adam R. Kosiorek, Tuan Anh Le, Chris J. Maddison, Maximilian Igl, Frank Wood, and Yee Whye Teh. Tighter Variational Bounds are Not Necessarily Better. In *International Conference on Machine Learning (ICML)*, March 2019. (Cited on 24)

Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of Adam and beyond. In *International Conference on Learning Representations (ICLR)*, pp. 1–23, 2018. (Cited on 22, 23)

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic Backpropagation and Approximate Inference. In *International Conference on Machine Learning (ICML)*, Beijing, China, 2014. (Cited on 1, 4, 6, 7, 20, 23)

Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer, 2004. ISBN 0-387-21239-6. (Cited on 5)

Geoffrey Roeder, Yuhuai Wu, and David K. Duvenaud. Sticking the Landing: Simple, Lower-Variance Gradient Estimators for Variational Inference. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017. (Cited on 4, 5, 7, 22, 23)

Shaun Seaman, John Galati, Dan Jackson, and John Carlin. What Is Meant by "Missing at Random"? *Statistical Science*, 28(2):257–268, May 2013. ISSN 0883-4237, 2168-8745. doi: 10.1214/13-STS415. (Cited on 2, 3)

88

Yuge Shi, N. Siddharth, Brooks Paige, and Philip Torr. Variational Mixture-of-Experts Autoencoders for Multi-Modal Deep Generative Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. (Cited on 5, 7, 12)

Vaidotas Simkus and Michael U. Gutmann. Conditional Sampling of Variational Autoencoders via Iterated Approximate Ancestral Sampling. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. (Cited on 4, 6, 19, 20, 23)

Vaidotas Simkus, Benjamin Rhodes, and Michael U. Gutmann. Variational Gibbs Inference for Statistical Model Estimation from Incomplete Data. *Journal of Machine Learning Research*, 24(196):1–72, 2023. ISSN 1533-7928. (Cited on 7, 21)

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In *International Conference on Machine Learning (ICML)*, November 2015. doi: 10.48550/arXiv.1503.03585. (Cited on 1)

Timur Sudak and Sebastian Tschiatschek. Posterior Consistency for Missing Data in Variational Autoencoders. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD)*, October 2023. doi: 10.48550/arXiv.2310.16648. (Cited on 3, 8)

Tijmen Tieleman. Training restricted Boltzmann machines using approximations to the likelihood gradient. In *International Conference on Machine Learning (ICML)*, pp. 1064–1071, 2008. ISBN 9781605582054. doi: 10.1145/1390156.1390290. (Cited on 21)

George Tucker, Dieterich Lawson, Shixiang Gu, and Chris J. Maddison. Doubly Reparameterized Gradient Estimators for Monte Carlo Objectives. In *International Conference on Learning Representations (ICLR)*, November 2018. (Cited on 5, 22)

Stef van Buuren. *Flexible Imputation of Missing Data*. CRC Press LLC, 2 edition, 2018. ISBN 978-1-138-58831-8. (Cited on 2)

Ramakrishna Vedantam, Ian Fischer, Jonathan Huang, and Kevin P. Murphy. Generative Models of Visually Grounded Imagination. *International Conference on Learning Representations (ICLR)*, May 2017. (Cited on 7)

Neng Wan, Dapeng Li, and Naira Hovakimyan. F-Divergence Variational Inference. In *Advances in Neural Information Processing Systems*, volume 33, pp. 17370–17379. Curran Associates, Inc., 2020. (Cited on 12)

Greg C. G. Wei and Martin A. Tanner. A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms. *Journal of the American Statistical Association*, 85(411):699–704, September 1990. doi: 10.1080/01621459.1990.10474930. (Cited on 6, 13)

Christopher K. I. Williams, Charlie Nash, and Alfredo Nazábal. Autoencoders and Probabilistic Inference with Missing Data: An Exact Solution for The Factor Analysis Case. *arXiv preprint*, 1801.03851, January 2018. (Cited on 7)

Mike Wu and Noah D. Goodman. Multimodal Generative Models for Scalable Weakly-Supervised Learning. In *NeurIPS 2018*, February 2018. (Cited on 7)

Laurent Younes. On the convergence of markovian stochastic algorithms with rapidly decreasing ergodicity rates. *Stochastics and Stochastic Reports*, 65(3-4):177–228, February 1999. ISSN 1045-1129. doi: 10.1080/17442509908834179. (Cited on 21)

Mingtian Zhang, Peter Hayes, and David Barber. Generalization Gap in Amortized Inference. In *Workshop on Bayesian Deep Learning at Neural Information Processing Systems (NeurIPS)*, pp. 6, 2021. (Cited on 8)

A Posterior complexity due to missing information

The complexity increase of the model posterior due to missing data, shown in fig. 1, explains why flexible variational distributions (Burda et al., 2015; Cremer et al., 2017) have been preferred when fitting VAEs from incomplete data (Mattei & Frellsen, 2019; Ipsen et al., 2020; Ma & Zhang, 2021). We here define the increase of the posterior complexity via the expected Kullback–Leibler (KL) divergence as follows

$$\begin{aligned}\mathbb{E}_{p^*(\boldsymbol{x})}\left[D_{\mathrm{KL}}(p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}) \mid\mid p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\mathrm{obs}}))\right] &= \mathbb{E}_{p^*(\boldsymbol{x}_{\mathrm{obs}}, \boldsymbol{x}_{\mathrm{mis}})}\mathbb{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}_{\mathrm{obs}}, \boldsymbol{x}_{\mathrm{mis}})}\left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\mathrm{obs}}, \boldsymbol{x}_{\mathrm{mis}})}{p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\mathrm{obs}})}\right] \\ &= \mathbb{E}_{p^*(\boldsymbol{x}_{\mathrm{obs}})}\mathbb{E}_{p^*(\boldsymbol{x}_{\mathrm{mis}}|\boldsymbol{x}_{\mathrm{obs}})p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}_{\mathrm{obs}}, \boldsymbol{x}_{\mathrm{mis}})}\left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\mathrm{obs}}, \boldsymbol{x}_{\mathrm{mis}})p^*(\boldsymbol{x}_{\mathrm{mis}} \mid \boldsymbol{x}_{\mathrm{obs}})}{p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\mathrm{obs}})p^*(\boldsymbol{x}_{\mathrm{mis}} \mid \boldsymbol{x}_{\mathrm{obs}})}\right] \\ &= \mathcal{I}(\boldsymbol{z}, \boldsymbol{x}_{\mathrm{mis}} \mid \boldsymbol{x}_{\mathrm{obs}}).^{12}\end{aligned}$$

As shown above the expected KL divergence equals the (conditional) mutual information (MI) between the learnt latent encodings $\boldsymbol{z} \sim p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\mathrm{obs}}, \boldsymbol{x}_{\mathrm{mis}})$ and the missing data $\boldsymbol{x}_{\mathrm{mis}} \sim p^*(\boldsymbol{x}_{\mathrm{mis}} \mid \boldsymbol{x}_{\mathrm{obs}})$.

The mutual information interpretation allows us to reason when a more flexible variational family may be necessary to accurately estimate VAEs from incomplete data. Specifically, when the MI is small then the two posterior distributions, $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x})$ and $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\mathrm{obs}})$ are similar, in which case a simple variational distribution may work sufficiently well. This situation might appear when the observed $\boldsymbol{x}_{\mathrm{obs}}$ and unobserved $\boldsymbol{x}_{\mathrm{mis}}$ variables are highly related and $\boldsymbol{x}_{\mathrm{mis}}$ provides little additional information about $\boldsymbol{z}$ over just $\boldsymbol{x}_{\mathrm{obs}}$, for example, when random pixels of an image are masked it is "easy" to infer the complete image due to strong relationship between neighbouring pixels. On the other hand, when the MI is high then $\boldsymbol{x}_{\mathrm{mis}}$ provides significant additional information about $\boldsymbol{z}$ over just $\boldsymbol{x}_{\mathrm{obs}}$, in which case a more flexible variational family may be needed, for example, when the pixels of an image are masked in blocks such that it introduces significant uncertainty about what is missing.

B DeMissVAE: Encoder objective derivation

The standard (complete-data) ELBO in eq. (2) gives the inequality

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathrm{obs}}, \boldsymbol{x}_{\mathrm{mis}}) \geq \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}_{\mathrm{obs}}, \boldsymbol{x}_{\mathrm{mis}})}\left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathrm{obs}}, \boldsymbol{x}_{\mathrm{mis}}, \boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\mathrm{obs}}, \boldsymbol{x}_{\mathrm{mis}})}\right],$$

which, together with the identity

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathrm{obs}}) = \log \int \exp\left\{\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathrm{obs}}, \boldsymbol{x}_{\mathrm{mis}})\right\} \mathrm{d}\boldsymbol{x}_{\mathrm{mis}},$$

yields

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathrm{obs}}) \geq \log \int \exp\left\{\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}_{\mathrm{obs}}, \boldsymbol{x}_{\mathrm{mis}})}\left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathrm{obs}}, \boldsymbol{x}_{\mathrm{mis}}, \boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\mathrm{obs}}, \boldsymbol{x}_{\mathrm{mis}})}\right]\right\} \mathrm{d}\boldsymbol{x}_{\mathrm{mis}}.$$

As the integral on the r.h.s. is intractable, we lower-bound it using the imputation distribution $f^t(\boldsymbol{x}_{\mathrm{mis}} \mid \boldsymbol{x}_{\mathrm{obs}})$ and Jensen's inequality

$$\begin{aligned}\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathrm{obs}}) &\geq \log \int \frac{f^t(\boldsymbol{x}_{\mathrm{mis}} \mid \boldsymbol{x}_{\mathrm{obs}})}{f^t(\boldsymbol{x}_{\mathrm{mis}} \mid \boldsymbol{x}_{\mathrm{obs}})} \exp\left\{\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}_{\mathrm{obs}}, \boldsymbol{x}_{\mathrm{mis}})}\left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathrm{obs}}, \boldsymbol{x}_{\mathrm{mis}}, \boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\mathrm{obs}}, \boldsymbol{x}_{\mathrm{mis}})}\right]\right\} \mathrm{d}\boldsymbol{x}_{\mathrm{mis}} \\ &= \log \mathbb{E}_{f^t(\boldsymbol{x}_{\mathrm{mis}}|\boldsymbol{x}_{\mathrm{obs}})}\left[\exp\left(-\log f^t(\boldsymbol{x}_{\mathrm{mis}} \mid \boldsymbol{x}_{\mathrm{obs}})\right) \exp\left\{\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}_{\mathrm{obs}}, \boldsymbol{x}_{\mathrm{mis}})}\left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathrm{obs}}, \boldsymbol{x}_{\mathrm{mis}}, \boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\mathrm{obs}}, \boldsymbol{x}_{\mathrm{mis}})}\right]\right\}\right]\end{aligned}$$

[12] Where notation of $\boldsymbol{m}$ is suppressed due to MAR assumption. In case of missing-not-at-random (MNAR) assumption there would be an additional dependency on $\boldsymbol{m}$.

$$
\begin{aligned}
&= \log \mathbb{E}_{f^t(\boldsymbol{x}_{\text{mis}}|\boldsymbol{x}_{\text{obs}})} \left[\exp \left\{ \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}_{\text{obs}},\boldsymbol{x}_{\text{mis}})} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}, \boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}) f^t(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})} \right] \right\} \right] \\
&\geq \mathbb{E}_{f^t(\boldsymbol{x}_{\text{mis}}|\boldsymbol{x}_{\text{obs}}) q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}_{\text{obs}},\boldsymbol{x}_{\text{mis}})} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}, \boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}) f^t(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})} \right] \\
&= \underbrace{\mathbb{E}_{f^t(\boldsymbol{x}_{\text{mis}}|\boldsymbol{x}_{\text{obs}}) q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}_{\text{obs}},\boldsymbol{x}_{\text{mis}})} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}, \boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})} \right]}_{\stackrel{\text{def}}{=} \mathcal{L}^{\boldsymbol{\phi}}_{\text{LMVB}}(\boldsymbol{x}_{\text{obs}}; \boldsymbol{\phi}, \boldsymbol{\theta}, f^t)} + \underbrace{\mathcal{H}\left[f^t(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}) \right]}_{\text{Const. w.r.t. } \boldsymbol{\phi}}.
\end{aligned}
$$

C DeMissVAE: Motivating the separation of objectives

The two DeMissVAE objectives $\mathcal{L}^{\boldsymbol{\theta}}_{\text{CVI}}$ and $\mathcal{L}^{\boldsymbol{\phi}}_{\text{LMVB}}$ in eqs. (10) and (12) correspond to valid lower-bounds on $\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}})$ irrespective of $f^t(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$. Moreover, both of them are tight at $f^t(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}) = p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ and $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}) = p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$. *So, a natural question is why do we prefer* $\mathcal{L}^{\boldsymbol{\theta}}_{\text{CVI}}$ *to learn* $p_{\boldsymbol{\theta}}$ *and* $\mathcal{L}^{\boldsymbol{\phi}}_{\text{LMVB}}$ *to learn* $q_{\boldsymbol{\phi}}$*?*

Why use $\mathcal{L}^{\boldsymbol{\theta}}_{\text{CVI}}$ in eq. (10) over $\mathcal{L}^{\boldsymbol{\phi}}_{\text{LMVB}}$ in eq. (12) to learn $p_{\boldsymbol{\theta}}(\boldsymbol{x})$? Maximisation of the objective $\mathcal{L}^{\boldsymbol{\phi}}_{\text{LMVB}}$ in iteration t w.r.t. $\boldsymbol{\theta}$ would have to compromise between maximising the log-likelihood $\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}})$ and keeping the other two KL divergence terms in eq. (13) low. Specifically, the compromise between maximising $\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}})$ and keeping $D_{\text{KL}}(f^t(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}) \,||\, p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}))$ low is equivalent to the compromise in the EM algorithm, which is known to affect the convergence of the model (Meng, 1994). Moreover, if $f^t(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}) \neq p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ then minimising the $D_{\text{KL}}(f^t(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}) \,||\, p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}))$ will fit the model $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ to the biased samples from $f^t(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$. On the other hand, in $\mathcal{L}^{\boldsymbol{\theta}}_{\text{CVI}}$ the missing variables $\boldsymbol{x}_{\text{mis}}$ are marginalised from the model, therefore it avoids the compromise with $D_{\text{KL}}(f^t(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}) \,||\, p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}))$ and the potential bias of the imputation distribution $f^t(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ affects the model *only* via the latents $\boldsymbol{z} \sim q_{\boldsymbol{\phi}, f^t}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$, increasing the robustness to sub-optimal imputations.

Why use $\mathcal{L}^{\boldsymbol{\phi}}_{\text{LMVB}}$ in eq. (12) over $\mathcal{L}^{\boldsymbol{\theta}}_{\text{CVI}}$ in eq. (10) to learn $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x})$? In the case of $\mathcal{L}^{\boldsymbol{\theta}}_{\text{CVI}}$, if $f^t(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}) = p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ then the bound is tightened when $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}) = p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$ for all $\boldsymbol{x}_{\text{mis}}$, which is the same optimal $q_{\boldsymbol{\phi}}$ if we used $\mathcal{L}^{\boldsymbol{\phi}}_{\text{LMVB}}$. But, there is also at least one more possible optimal solution $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}) = p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$, which ignores the imputations and corresponds to the optimal solution of the standard approach in section 2, and thus it means that the optimum is (partially) unidentifiable and can make optimisation of $q_{\boldsymbol{\phi}}$ using $\mathcal{L}^{\boldsymbol{\theta}}_{\text{CVI}}$ difficult. Moreover, if $f^t(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}) \neq p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ then in order to minimise $D_{\text{KL}}(q_{\boldsymbol{\phi}, f^t}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}) \,||\, p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}))$ w.r.t. $\boldsymbol{\phi}$ the variational distribution $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$ would have to compensate for the inaccuracies of $f^t(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ by adjusting the probability mass over the latents $\boldsymbol{z}$, such that $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$ is correct on average, i.e. $q_{\boldsymbol{\phi}, f^t}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}) = \mathbb{E}_{f^t(\boldsymbol{x}_{\text{mis}}|\boldsymbol{x}_{\text{obs}})}[q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})] \approx p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$. These two issues make optimising $\boldsymbol{\phi}$ via $\mathcal{L}^{\boldsymbol{\theta}}_{\text{CVI}}$ such that $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}) \approx p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$ difficult. On the other hand, in $\mathcal{L}^{\boldsymbol{\phi}}_{\text{LMVB}}$ the optimal $q_{\boldsymbol{\phi}}$ is always at $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}) = p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$, irrespective of the imputation distribution $f^t(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$, hence the $\mathcal{L}^{\boldsymbol{\phi}}_{\text{LMVB}}$ objective in eq. (12) is well-defined and more robust to inaccuracies of $f^t(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ for the optimisation of $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$.

In fig. 5 we verify the efficacy of DeMissVAE via a control study on a small VAE model $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ with 2D latent space fitted on incomplete samples from a ground truth mixture-of-Gaussians (MoG) distribution $p^*(\boldsymbol{x})$. We evaluate fitting the VAE using only $\mathcal{L}^{\boldsymbol{\theta}}_{\text{CVI}}$ in eq. (10) (CVI-VAE, blue), only $\mathcal{L}^{\boldsymbol{\phi}}_{\text{LMVB}}$ in eq. (12) (MVB-VAE, yellow), and using the proposed two-objective approach (DeMissVAE, green). In the left-most figure we evaluate the three methods where we represent the imputation distribution $f^t(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}) = p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ using rejection sampling, which corresponds to the optimal imputation distribution w.r.t. $D_{\text{KL}}(f^t(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}) \,||\, p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})) = 0$. We see that the proposed approach (green) dominates over the other two control methods (blue and yellow), and importantly that marginalisation of the missing variables in DeMissVAE (green) improves the model accuracy compared to an EM-type handling of the missing

91

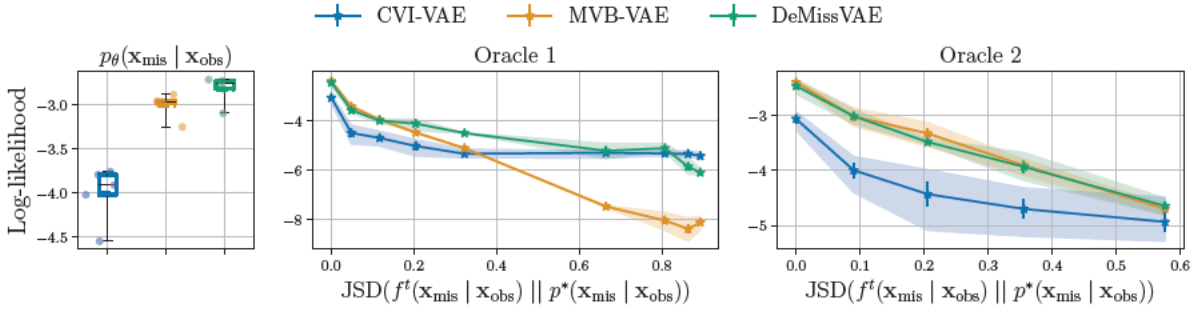


Figure 5: *A control study on a VAE model with 2D latent space (see additional details in appendix E.1), examining the sensitivity of the proposed method (DeMissVAE, green) and two control methods (blue and yellow) to the accuracy of the imputation distribution* $f^t(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$. Left: $f^t(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}) = p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ represented using rejection sampling. Center: an oracle imputation function that gets progressively "wider" from left-to-right of the figure. Right: an oracle imputation distribution that towards the right of the figure more significantly oversamples low-probability posterior modes. The log-likelihood is computed on a held-out test data set by numerically integrating the 2D latent space of the VAE. The horizontal axis on the two right-most figures shows the Jensen–Shannon divergence between the imputation distribution and the ground-truth conditional $p^\star(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$.

variables (yellow). Furthermore, in the remaining two figures we investigate the sensitivity of the methods to the accuracy of imputations in $f^t(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$. In Oracle 1 we start with the ground-truth conditional $p^\star(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ and, along the x-axis of the figure, investigate how the methods perform when the imputation distribution becomes "wider": first interpolating from $p^\star(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ to an independent unconditional distribution $\prod_{d \in \text{idx}(\boldsymbol{m})} p^\star(x_d)$ and then further towards and independent Gaussian distribution. And in Oracle 2 we investigate what happens when the sampler "oversamples" posterior modes: we interpolate the imputation distribution from $p^\star(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ to $\frac{1}{C}\sum_c^C p^\star(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, c)$, where c is the component of the mixture distribution with a total of C components. As we see in the figure, the proposed DeMissVAE approach (green) performs similar or better than the MVB-VAE (yellow) and CVI-VAE (blue) control methods, with an exception when the $f^t(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ are extremely inaccurate (last two points on the middle figure) which is expected since $q_{\boldsymbol{\phi}, f^t}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ in eq. (9) can be arbitrarily far from $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ when $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}) = p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$ but $f^t(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}) \not\approx p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$.

Finally, in fig. 6 we investigate what happens if we used only $\mathcal{L}^{\boldsymbol{\theta}}_{\text{CVI}}$ in eq. (10) or $\mathcal{L}^{\boldsymbol{\phi}}_{\text{LMVB}}$ in eq. (12) to fit the VAE model, in contrast to the two separate objectives for encoder and decoder in DeMissVAE. We use the LAIR sampling method (Simkus & Gutmann, 2023) as detailed in appendix D to obtain approximate samples from $f^t(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}) \approx p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$. And, we observe that DeMissVAE achieves a better fit of the model, in line with our motivation in this section.

D DeMissVAE: Implementing the training procedure

DeMissVAE requires optimising two objectives $\mathcal{L}^{\boldsymbol{\theta}}_{\text{CVI}}$ and $\mathcal{L}^{\boldsymbol{\phi}}_{\text{LMVB}}$ in eqs. (10) and (12) and drawing (approximate) samples to represent $f^t(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}) \approx p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$. Our aim is to implement this efficiently to minimise redundant computation.

The algorithm starts with a randomly-initialised target VAE model $p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z})$, an amortised variational distribution $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x})$, and an incomplete data set $\mathcal{D} = \{\boldsymbol{x}^i_{\text{obs}}\}_i$. And then, to represent the imputation distribution $f^0(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$, K imputations $\{\boldsymbol{x}^{ik}_{\text{mis}}\}_{k=1}^K$ are generated for each $\boldsymbol{x}^i_{\text{obs}} \in \mathcal{D}$ using some simple imputation function such as sampling the marginal empirical distributions of the missing variables. The algorithm then iterates between the following two steps:

92

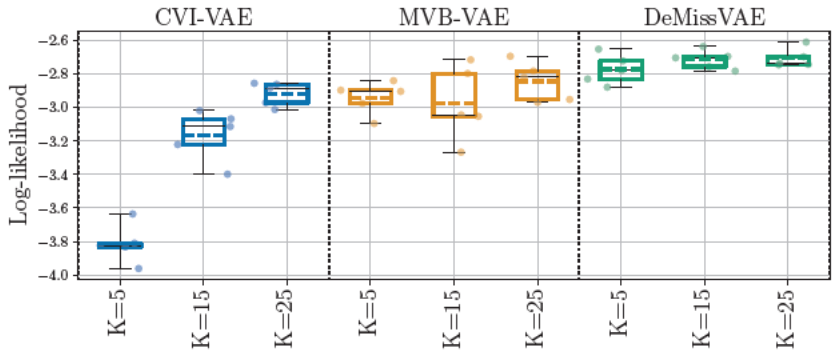


Figure 6: *A control study on a VAE model with 2D latent space (see additional details in appendix E.1), investigating the importance of the two-objective approach in DeMissVAE (green) and two control methods (blue and yellow).* In CVI-VAE (blue) we fit both the encoder and decoder using eq. (10), and in MVB-VAE (yellow) we fit both the encoder and decoder using eq. (12). The log-likelihood is computed on a held-out test data set by numerically integrating the 2D latent space of the VAE.

Algorithm 1 Shared computation of the DeMissVAE learning objectives

Input: parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, number of latent samples L, completed data-point $(\boldsymbol{x}^i_{\text{obs}}, \boldsymbol{x}^{ik}_{\text{mis}})$

$\boldsymbol{\psi}^{ik} \leftarrow \texttt{Encoder}(\boldsymbol{x}^i_{\text{obs}}, \boldsymbol{x}^{ik}_{\text{mis}}; \boldsymbol{\phi})$ ▷ Compute parameters of the variational distribution

$\boldsymbol{z}_1, \ldots, \boldsymbol{z}_L \sim q(\boldsymbol{z} \mid \boldsymbol{x}^i_{\text{obs}}, \boldsymbol{x}^{ik}_{\text{mis}}; \boldsymbol{\psi}^{ik})$ ▷ Sample latents $\boldsymbol{z}$

$\boldsymbol{\eta}_l \leftarrow \texttt{Decoder}(\boldsymbol{z}_l; \boldsymbol{\theta})$ for $\forall l \in [1, L]$ ▷ Compute parameters of the generative distribution

def $\mathcal{L}^{\boldsymbol{\theta}}_{\text{CVI}}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_L, \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_L)$: ▷ Procedure for estimating eq. (10)

  return $\frac{1}{L}\sum_{l=1}^{L} \log p(\boldsymbol{x}^i_{\text{obs}}, \boldsymbol{z}_l; \boldsymbol{\eta}_l)$

def $\mathcal{L}^{\boldsymbol{\phi}}_{\text{LMVB}}(\boldsymbol{\psi}^{ik}, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_L, \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_L)$: ▷ Procedure for estimating eq. (12)

  return $\frac{1}{L}\sum_{l=1}^{L} \log p(\boldsymbol{x}^i_{\text{obs}}, \boldsymbol{x}^{ik}_{\text{mis}}, \boldsymbol{z}_l; \boldsymbol{\eta}_l) - \log q(\boldsymbol{z}_l \mid \boldsymbol{x}^i_{\text{obs}}, \boldsymbol{x}^{ik}_{\text{mis}}; \boldsymbol{\psi}^{ik})$

return $\mathcal{L}^{\boldsymbol{\theta}}_{\text{CVI}}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_L, \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_L), \mathcal{L}^{\boldsymbol{\phi}}_{\text{LMVB}}(\boldsymbol{\psi}^{ik}, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_L, \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_L)$

1. **Imputation.** Update the K imputations $\{\boldsymbol{x}^{ik}_{\text{mis}}\}_{k=1}^{K}$ representing samples from the imputation distribution $f^t(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$, such that $f^t(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ is "closer" to $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$. For this, we use cheap approximate iterative sampling methods such as pseudo-Gibbs (Rezende et al., 2014, Appendix F), Metropolis-within-Gibbs (MWG, Mattei & Frellsen, 2018a), or latent-adaptive importance resampling (LAIR, Simkus & Gutmann, 2023). Moreover, since the model and the variational distributions are initialised randomly, we skip the imputation step during the first epoch over the data.

2. **Parameter update.** Update the parameters using stochastic gradient ascent on $\mathcal{L}^{\boldsymbol{\theta}}_{\text{CVI}}$ and $\mathcal{L}^{\boldsymbol{\phi}}_{\text{LMVB}}$ in eqs. (10) and (12) with the imputations from $f^t(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$.

Efficient parameter update. While the two objectives for $p_{\boldsymbol{\theta}}$ and $q_{\boldsymbol{\phi}}$ in eqs. (10) and (12) are different, a major part of the computation can be shared, as shown in algorithm 1. As usual, the objectives are approximated using Monte Carlo averaging and require only one evaluation of the generative model, including the encoder, decoder, and prior, for each completed data-point $(\boldsymbol{x}^i_{\text{obs}}, \boldsymbol{x}^{ik}_{\text{mis}})$. Therefore, only backpropagation needs to be performed separately and the overall per-iteration computational cost of optimising the two objectives is about 1.67 times the cost of a fully-observed VAE optimisation (instead of 2 times if implemented naïvely).[13]

Efficient imputation. To make the imputation step efficient, the imputation distribution $f^t(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ is "persitent" between iterations, that is, the imputation distribution from the previous iteration

[13] The cost of backpropagation is about 2 times the cost of a forward pass (Burda et al., 2015).

$f^{t-1}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ is used to initialise the iterative approximate VAE sampler at iteration t.[14] Moreover, an iteration of a pseudo-Gibbs, MWG, or LAIR samplers uses the same quantities as the objectives $\mathcal{L}^{\boldsymbol{\theta}}_{\text{CVI}}$ and $\mathcal{L}^{\boldsymbol{\phi}}_{\text{LMVB}}$ in eqs. (10) and (12), and hence the cost of one iteration of the sampler in the imputation step can be shared with the cost of computation of the learning objectives. However, it is important to note that the accuracy of imputations affects the accuracy of the estimated model, and hence better estimation can be achieved by increasing the computational budget for imputation or using better imputation methods.

[14] Persistent samplers have been used in the past to increase efficiency of maximum-likelihood estimation methods (Younes, 1999; Tieleman, 2008; Simkus et al., 2023).

Published in Transactions on Machine Learning Research (06/2024)

E Experiment details

In this appendix we provide additional details on the experiments.

E.1 Mixture-of-Gaussians data with a 2D latent VAE

We generated a random 5D mixture-of-Gaussians model with 15 components by sampling the mixture covariance matrices from the inverse Wishart distribution $\mathcal{W}^{-1}(\nu = D, \Psi = \mathbf{I})$, means from the Gaussian distribution $\mathcal{N}(\mu = \mathbf{0}, \sigma = \mathbf{3})$ and the component probabilities from Dirichlet distribution $\mathrm{Dir}(\alpha = \mathbf{1})$ (uniform). The model was then standardised to have a zero mean and a standard deviation of one. The pairwise marginal densities of the generated distribution is visualised in fig. 7 showing a highly-complex and multimodal distribution, and the generated parameters and data used in this paper are available in the shared code repository. We simulated a 20K sample data set used to fit the VAEs.

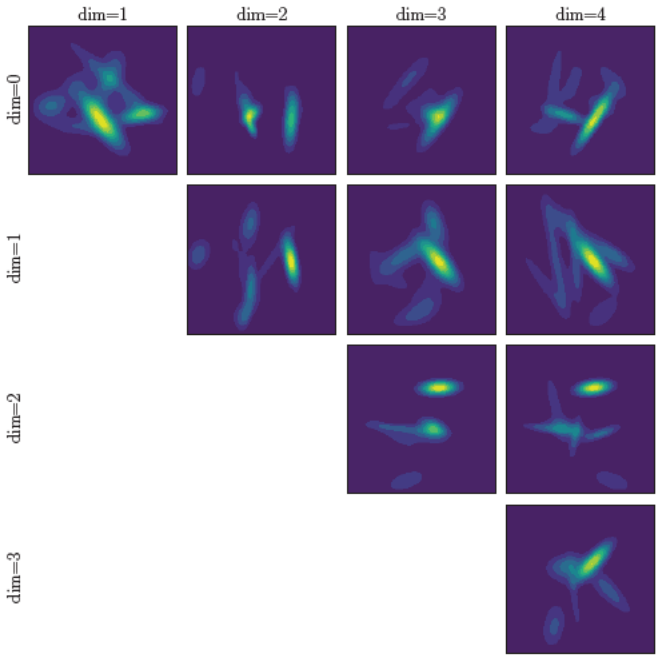


Figure 7: The pairwise marginals of the ground-truth Mixture-of-Gaussians distribution.

We then fitted a VAE model with 2-dimensional latent space using diagonal Gaussian encoder and decoder distributions, and a fixed standard Normal prior. For the decoder and encoder networks we used fully-connected residual neural networks with 3 residual blocks, 200 hidden dimensions, and ReLU activations. To optimise the model parameters we have used AMSGrad optimiser (Reddi et al., 2018) with a learning rate of 10^{-3} for a total of 500 epochs.

The hyperameters are listed in table 3, note that the total number of samples was the same for all methods (i.e. 5/15/25). Moreover, we have used "sticking-the-landing" (STL) gradients (Roeder et al., 2017) to reduce gradient variance for all methods.[15]

[15]We have also evaluated the doubly-reparametrised gradients (DReG, Tucker et al., 2018) for IWAE methods but found STL to perform similar or better.

Method	Z	K	I	**Mixture sampling**
MVAE	5/15/25	1	—	—
MissVAE	5/15/25	5/15/25	—	Ancestral
MissSVAE	1	5/15/25	—	Stratified
MIWAE	1	1	5/15/25	—
MissIWAE	1	5/15/25	5/15/25	Ancestral
MissSIWAE	1	5/15/25	1	Stratified
DeMissVAE	1	5/15/25	—	LAIR (1 iteration, $R = 0$) (Simkus & Gutmann, 2023)

Table 3: *Method hyperparameters on MoG data.*

E.2 UCI data sets

We fit the VAEs on four data sets from the UCI repository (Dua & Graff, 2017) with the preprocessing of (Papamakarios et al., 2017). The VAE model uses diagonal Gaussian encoder and decoder distributions regularised such that the standard deviation $\geq 10^{-5}$ (Mattei & Frellsen, 2018a), and a fixed standard Normal prior. The latent space is 16-dimensional, except for the MINIBOONE where 32 dimensions were used.

The encoder and decoder networks used fully-connected residual neural networks with 2 residual blocks (except for on the MINIBOONE dataset where 5 blocks were used in the encoder) with 256 hidden dimensionality, and ReLU activations. A dropout of 0.5 was used on the MINIBOONE dataset. The parameters were optimised using AMSGrad optimiser (Reddi et al., 2018) with a learning rate of 10^{-3} and cosine learning rate schedule for a total of 200K iterations (or 22K iterations on MINIBOONE). As before, STL gradients (Roeder et al., 2017) were used to reduce the variance for all methods. DeMissVAE used the LAIR sampler (Simkus & Gutmann, 2023) with $K = 5$ $R = 1$ and 10 iterations. Moreover we have used gradient norm clipping to stabilise DeMissVAE with the maximum norm set to 1 (except for POWER dataset where we set it to 0.5).

E.3 MNIST and Omniglot data sets

We fit a VAE on statically binarised MNIST and Omniglot data sets (Lake et al., 2015) downsampled to 28x28 pixels. The VAE uses diagonal Gaussian decoder distributions regularised such that the standard deviation $\geq 10^{-5}$ (Mattei & Frellsen, 2018a), a fixed standard Normal prior, and a Bernoulli decoder distribution. The latent space is 50-dimensional.

For both MNIST and Omniglot we have used convolutional ResNet neural networks for the encoder and decoder with 4 residual blocks, ReLU activations, and dropout probability of 0.3. For MNIST, the encoder the residual block hidden dimensionalities were 32, 64, 128, 256, and for the decoder they were 128,64,32,32. For Omniglot, the encoder the residual block hidden dimensionalities were 64, 128, 256, 512, and for the decoder they were 256,128,64,64. We used AMSGrad optimiser (Reddi et al., 2018) with 10^{-4} learning rate, cosine learning rate schedule, and STL gradients (Roeder et al., 2017) for 500 epochs for MNIST and 200 epochs for Omniglot.

For MVAE, we use 5 latent samples and for MIWAE we use 5 importance samples. For MissVAE we use $K = 5$ mixture components and sample 5 latent samples. For MissSVAE we use $K = 5$ mixture components and sample 1 sample from each component, for a total of 5 samples. For MissIWAE we use $K = 5$ components and sample 5 importance samples. for MissSIWAE we use $K = 5$ components and sample 1 sample from each component. For DeMissVAE we use $K = 5$ imputations and update them using a single step of pseudo-Gibbs (Rezende et al., 2014).

F Additional figures

In this appendix we provide additional figures for the experiments in this paper.

F.1 Mixture-of-Gaussians data with a 2D latent VAE

In this section we show additional analysis on the mixture-of-Gaussians data, supplementing the results in section 6.1.

F.1.1 Analysis of gradient variance with ancestral and stratified sampling

In section 6.1 we observed that the model estimation performance can depend on whether ancestral sampling (with implicit reparametrisation) or stratified sampling is used to approximate the expectations in eqs. (4), (5), (7) and (8), corresponding to MissVAE/MissIWAE and MissSVAE/MissSIWAE, respectively.

We analyse the signal-to-noise ratio (SNR) of the gradients w.r.t. $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ for the two approaches, which is defined as follows (Rainforth et al., 2019)

$$\mathrm{SNR}(\boldsymbol{\phi}) = \left|\frac{\mathbb{E}\left[\Delta(\boldsymbol{\phi})\right]}{\sigma[\Delta(\boldsymbol{\phi})]}\right|, \quad \text{and} \quad \mathrm{SNR}(\boldsymbol{\theta}) = \left|\frac{\mathbb{E}\left[\Delta(\boldsymbol{\theta})\right]}{\sigma[\Delta(\boldsymbol{\theta})]}\right|,$$

where $\Delta(\cdot)$ denotes the gradient estimate, and $\sigma[\cdot]$ is the standard deviation of a random variable. We estimate the SNR by computing the expectation and standard deviation over the entire training epoch.

The SNR for $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ is plotted in fig. 8. We observe that the stratified approaches (MissSVAE and MissSIWAE) generally have higher SNR. This is possibly the reason why MissSVAE and MissSIWAE have achieved better model accuracy than the ancestral approaches (MissVAE and MissIWAE) in section 6.1.

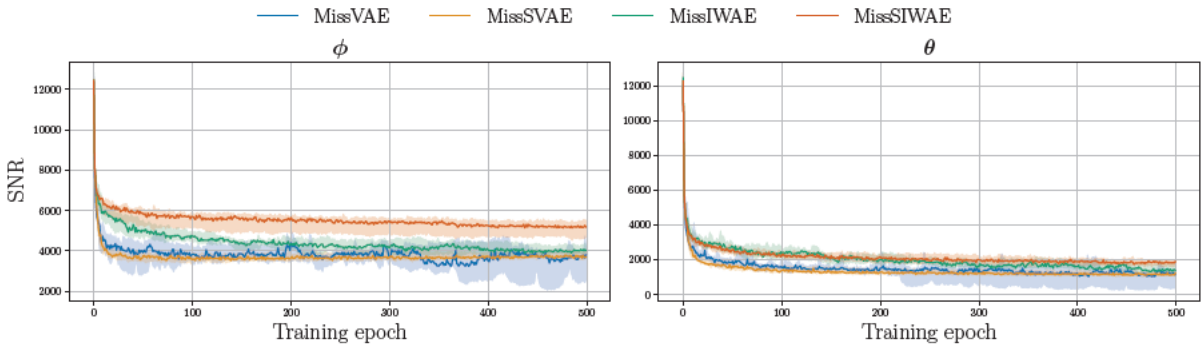


Figure 8: *Signal-to-noise ratio (SNR, higher is better) of the gradients w.r.t. encoder parameters $\boldsymbol{\phi}$ (left) and decoder parameters $\boldsymbol{\theta}$ (right).* For all methods we used a budget of 5 samples from the variational distribution (see appendix E.1 for more details). We show the median SNR over 5 independent runs and a 90% confidence interval.

F.1.2 Analysis of the model posteriors

In figs. 9 to 11 we visualise the model posteriors with complete and incomplete data, $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x})$ and $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$, respectively, and the variational distribution $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \cdot)$ that was used to fit the model via the variational ELBO. For each method we have used a budget of 25 samples from the variational distribution during training (additional details are in appendix E.1). Each figure shows the posteriors for 5 training data-points using distinct colours.

Figure 9 shows MVAE, MissVAE, and MissSVAE model posteriors $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x})$ and $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$, as well as the variational distribution $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$, which approximates the incomplete-data posterior. As motivated

97

in section 3 we observe that the Gaussian posterior in MVAE (first row) is not sufficiently flexible to approximate the complex incomplete-data posteriors $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$. On the other hand, the mixture-variational approaches, MissVAE (second row) and MissSVAE (third row), are able to well-approximate the incomplete-data posteriors.

Figure 10 shows MIWAE, MissIWAE, and MissSIWAE model posteriors $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x})$ and $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$, as well as the variational proposal $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ and the importance-weighted semi-implicit distribution $q^{\text{IW}}_{\boldsymbol{\phi},\boldsymbol{\theta},I=25}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ that arises from sampling the variational proposal $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ and re-sampling using importance weights $w(\boldsymbol{z}) = p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}}, \boldsymbol{z})/q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ (Cremer et al., 2017). Similar to the MVAE case above, the variational proposal $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ in MIWAE (first row) is quite far from the model posterior $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$, but the importance-weighted bound in eq. (7) is able to re-weigh the samples to sufficiently-well match the model posterior, as shown in the fourth column. However, as efficiency of importance sampling depends on the discrepancy between the proposal and the target distributions, we can expect that more flexible variational distributions may improve the performance of the importance-weighted ELBO methods. Importantly, we show that the variational-mixture approaches, MissIWAE (second row) and MissSIWAE (third row), are able to adapt the variational proposals to the incomplete-data posteriors well, and as a result achieve better efficiency than MIWAE.

Figure 11 shows DeMissVAE model posteriors $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x})$ and $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$, the variational distribution $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x})$, which approximates the *complete*-data posterior, and the imputation-mixture $q_{\boldsymbol{\phi},f^t}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ approximated using the 25 imputations in $f^t(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ at the end of training. We observe similar behaviour to fig. 1, where the complete data posteriors $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x})$ are close to Gaussian but the incomplete-data posteriors $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ are irregular. As we show in section 6.1, DeMissVAE is capable of fitting the model well by learning the completed-data variational distribution $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x})$ (third column) and using the imputation-mixture in eq. (9) to approximate the incomplete data posterior $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$. Moreover, we observe that the imputation-mixture $q_{\boldsymbol{\phi},f^t}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ (fourth column) captures only one of the modes of the model posterior $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$. This is a result of using a small imputation sampling budget, that is, using only a single iteration of LAIR to update the imputations (see more details in appendix D), and hence better accuracy can be achieved by trading-off some computational cost to obtain better imputations that would ensure a better representation of the imputation distribution. Nonetheless, as observed in fig. 2, DeMissVAE achieves good model accuracy despite potentially sub-optimal imputations, further signifying the importance of the two learning objectives for the encoder and decoder distributions in section 4.2 and appendix C.

Interestingly, by comparing the complete-data posteriors $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x})$ (first column) in figs. 9 to 11, we observe that they are slightly more irregular than in the complete case in fig. 1, except for DeMissVAE whose posteriors are nearly Gaussian. The irregularity is stronger for the importance-weighted ELBO-based methods in fig. 10. This is in line with the observation by Burda et al. (2015, Appendix C) and Cremer et al. (2018, Section 5.4) that VAEs trained with more flexible variational distributions tend to learn a more complex model posterior. This means that using the importance-weighted bounds, and to a lesser extent the finite variational-mixture approaches from section 4.1, to fit VAEs on incomplete data may result in worse-structured latent spaces, compared to models fitted on complete data. On the other hand, we observe that DeMissVAE learns a better-structured latent space, with the posterior $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x})$ close to a Gaussian, that is comparable to the complete case. This suggests that the decomposed approach in DeMissVAE may be important in cases where the latent space needs to be regular, at the additional cost of obtaining missing data imputations (see appendix D).

Published in Transactions on Machine Learning Research (06/2024)

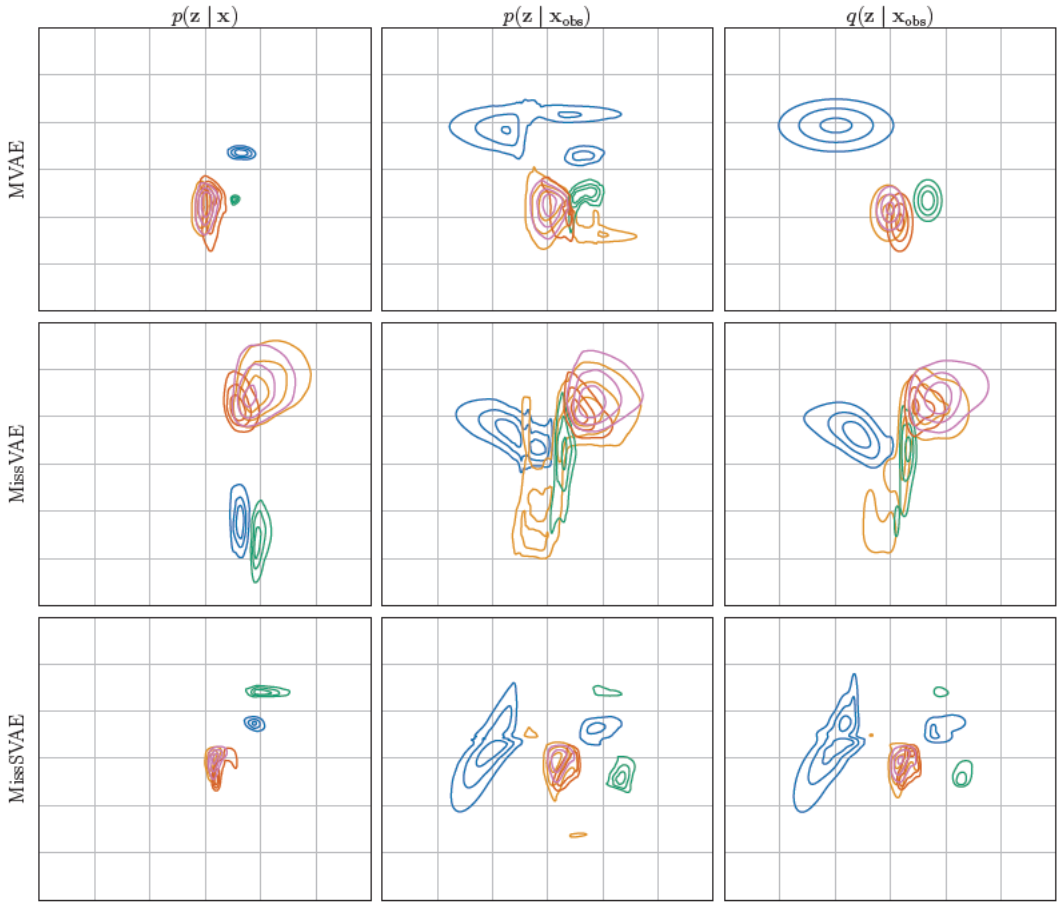


Figure 9: *Posterior distributions of MVAE, MissVAE, and MissSVAE.* First column: the model posterior $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x})$ under complete data $\boldsymbol{x}$. Second column: the model posterior $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ under incomplete data $\boldsymbol{x}_{\text{obs}}$. Third column: variational approximation $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ of the incomplete posterior $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ obtained at the end of training.

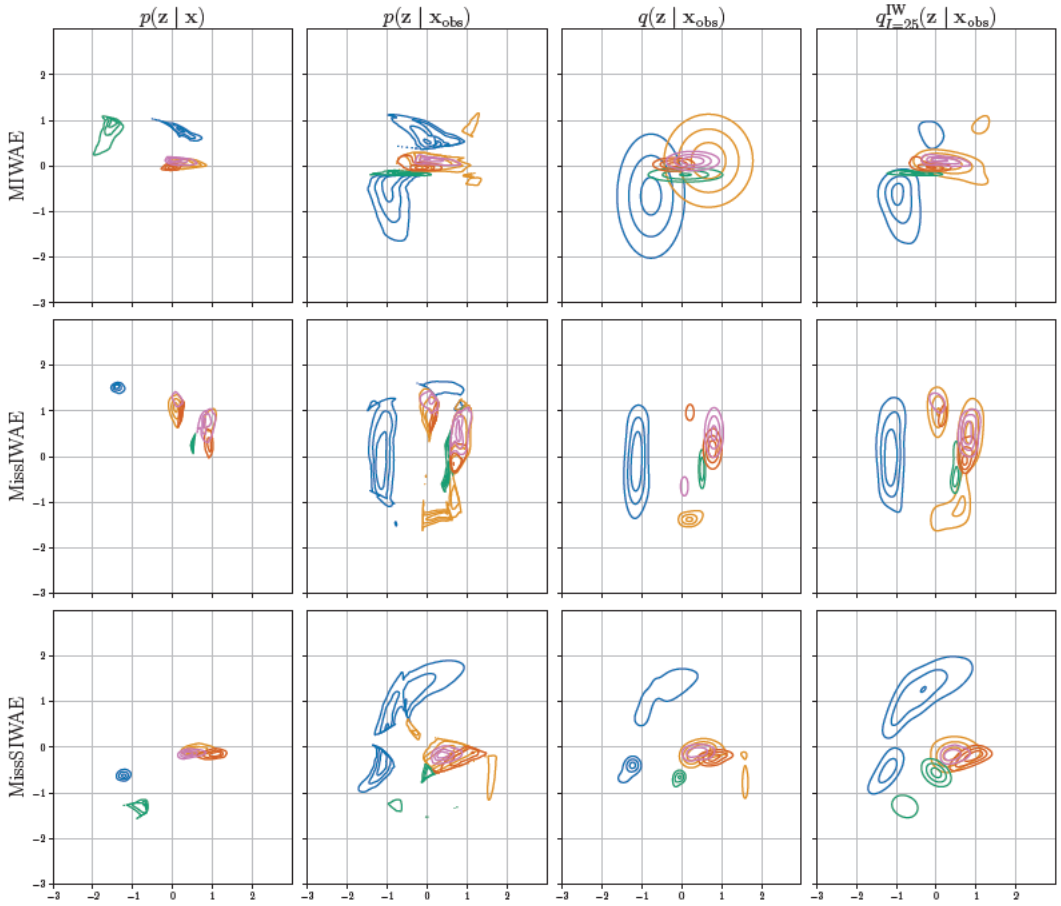


Figure 10: *Posterior distributions of MIWAE, MissIWAE, and MissSIWAE.* First column: the model posterior $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x})$ under complete data $\boldsymbol{x}$. Second column: the model posterior $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ under incomplete data $\boldsymbol{x}_{\text{obs}}$. Third column: variational proposal $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ for an incomplete data-point $\boldsymbol{x}_{\text{obs}}$ obtained at the end of training. Fourth column: importance-weighted variational distribution $q^{\text{IW}}_{\boldsymbol{\phi} I=25}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ for an incomplete data-point $\boldsymbol{x}_{\text{obs}}$ obtained after re-weighting samples from $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ (Cremer et al., 2017).

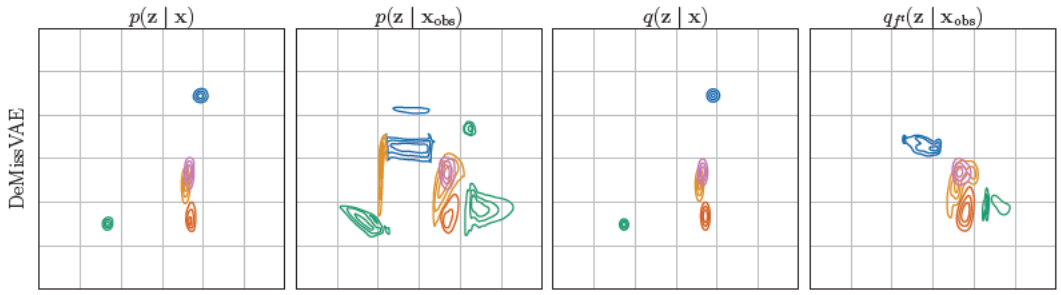


Figure 11: *Posterior distributions of DeMissVAE.* First: the model posterior $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x})$ under complete data $\boldsymbol{x}$. Second: the model posterior $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ under incomplete data $\boldsymbol{x}_{\text{obs}}$. Third: variational approximation $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x})$ of the complete-data posterior $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x})$ obtained at the end of training. Fourth: the variational imputation-mixture distribution in eq. (9) using the imputation distribution $f^t(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ obtained at the end of training, approximated using a Monte Carlo average with 25 imputations.

100

Published in Transactions on Machine Learning Research (06/2024)

F.2 UCI data sets

In fig. 12 we plot the Fréchet inception distance (FID, Heusel et al., 2017) versus training iteration on the UCI datasets. The results closely mimic the log-likelihood results in section 6.2. Importantly, we observe that using mixture variational distributions becomes more important as the missingness fraction increases, causing the posterior distributions to be more complex.

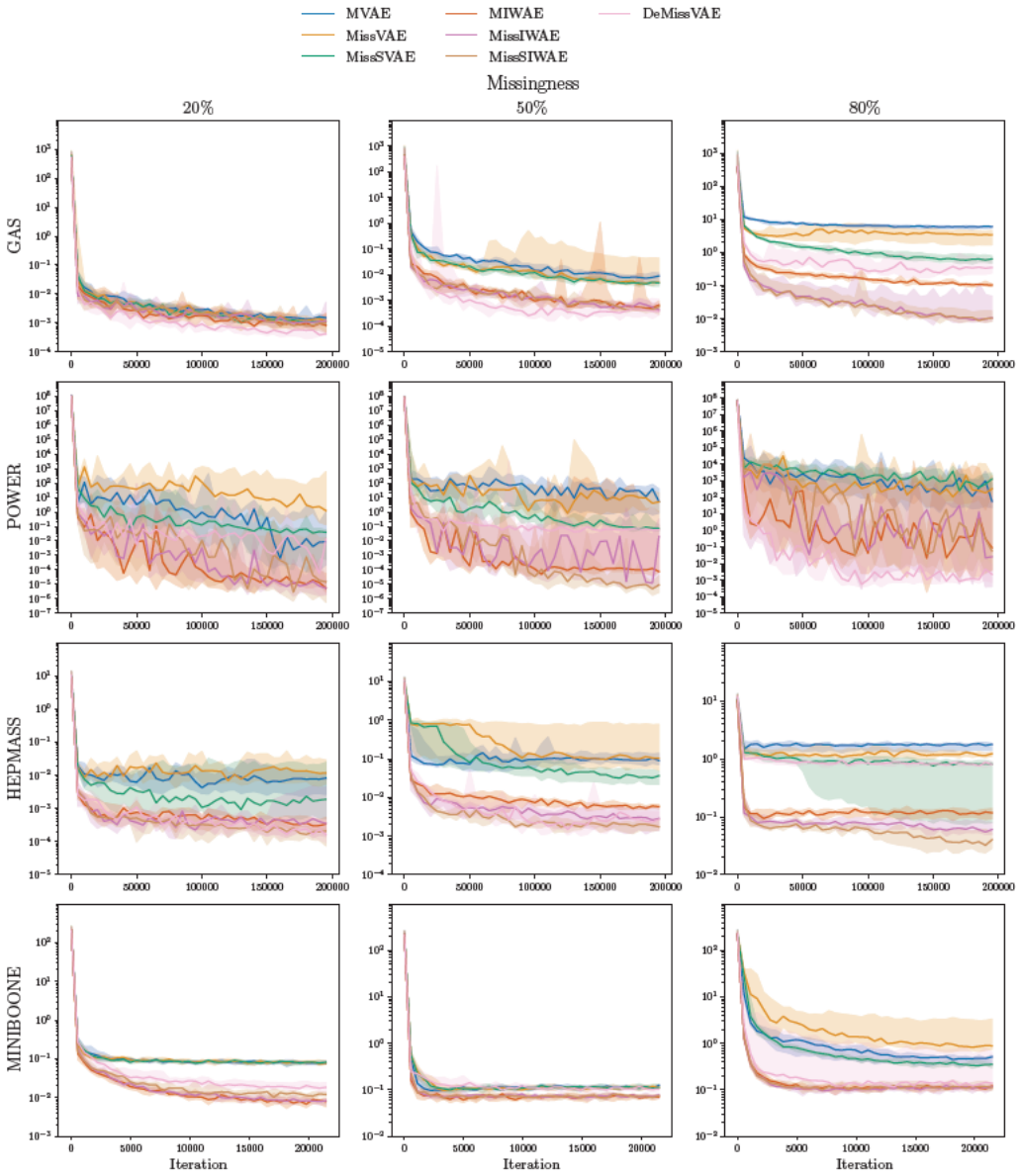


Figure 12: *FID (lower is better) between the model and the complete test data versus training iterations.* The FID is computed using features of the last encoder layer of an independent VAE model trained on complete data. Lines show the median of 5 independent runs and the intervals show 90% confidence.

101

Published in Transactions on Machine Learning Research (06/2024)

F.3 MNIST and Omniglot data sets, latent dimensionality

Rather than handling the posterior complexity due to missingness with variational mixtures, an alternative approach may be to increase the capacity of the model by making the latent space dimensionality larger. In fig. 13 we plot the estimated test log-likelihood against latent variable dimensionality.

On MNIST data, the effect from increasing the latent space is small, with the exception of DeMissVAE. The reason why DeMissVAE improves quite significantly with latent space dimensionality may be more due to easier sampling than a larger capacity of the model. This again highlights that DeMissVAE, when used with efficient sampling methods, can be an efficient approach to handle data missingness.

On Omniglot data, we observe a small improvement for MIWAE, MVAE, and the stratified MissSIWAE, while the other methods either remained at about the same accuracy or declined. However, there is no significant change from the results in section 6.3.

Hence, while increasing latent dimensionality generally increases the capacity of the model, enabling the modelling of more complex distributions, overall we observe that it provides minimal effect for dealing with data missingness.

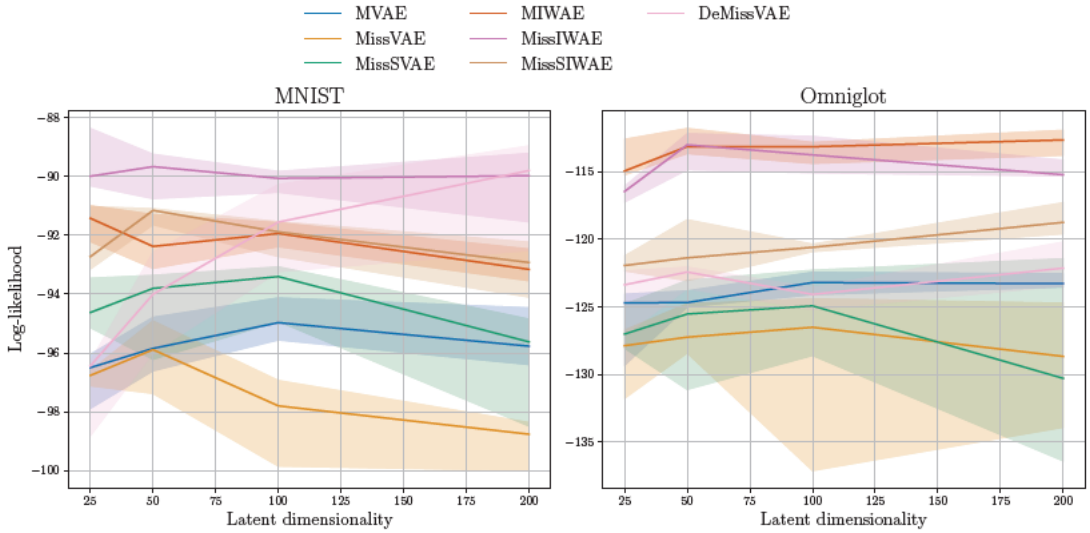


Figure 13: *Estimate of the test log-likelihood against latent dimensionality on MNIST and Omniglot data sets.* The log-likelihood was estimated on complete test data using the IWELBO with $I = 1000$. The curves show median performance over 5 independent runs of the methods and the intervals show the 90% centered interval.

Chapter 6

Towards general-purpose deep statistical model estimation from incomplete data

In the previous chapter, we introduced a method to improve VAE estimation from incomplete data. However, like much of the existing literature, this method relies on a key assumption that the missing variables can be efficiently marginalised (section 3.2.1). Yet, as we highlighted in section 3.2.3, marginalisation of missing variables is not tractable for all deep models, and even breaks down for some VAEs models with decoder distributions that do not factorise.

In section 2.2.6, we detailed Monte Carlo EM (Wei and Tanner, 1990) as an alternative for model estimation when marginalisation is not tractable. To train the model on an incomplete data set, this algorithm iterates between conditional sampling to impute the missing values and model estimation using the completed data. But, as discussed in section 3.2.3, exact conditional sampling of deep models is generally intractable, and approximate approaches such as MCMC and importance sampling can be inefficient due to challenges in hyper-parameter tuning across different data-points and throughout the learning iterations. Additionally, while variational inference (*e.g.* Jordan et al., 1999) is often used as an alternative to MCMC and importance sampling, its application for deep model estimation from incomplete data remains largely unexplored.

These observations motivate the following research question:

Research Question 3: How do we efficiently estimate general statistical models for which simplifying assumptions, like marginalisation of the missing variables, may not be applicable?

In this chapter, we aim to address the above research question and develop a new method for estimating *general* statistical models from incomplete data. Building on advances in variational inference and traditional multiple imputation techniques, our publication, included verbatim in section 6.1, introduces variational Gibbs inference (VGI), a novel

general-purpose approximate MLE method from incomplete data.

6.1 Publication

This section includes a verbatim copy of the following journal publication:[9]

> Vaidotas Simkus, Benjamin Rhodes, and Michael U. Gutmann. Variational Gibbs Inference for Statistical Model Estimation from Incomplete Data. *Journal of Machine Learning Research*, 24(196):1–72, 2023

[9] This paper was also presented as a poster at NeurIPS 2023 via the journal-to-conference track.

Journal of Machine Learning Research 24 (2023) 1-72 Submitted 11/21; Revised 3/23; Published 6/23

Variational Gibbs Inference for Statistical Model Estimation from Incomplete Data

Vaidotas Simkus
Benjamin Rhodes
Michael U. Gutmann
School of Informatics
University of Edinburgh

Editor: Mohammad Emtiyaz Khan

Abstract

Statistical models are central to machine learning with broad applicability across a range of downstream tasks. The models are controlled by free parameters that are typically estimated from data by maximum-likelihood estimation or approximations thereof. However, when faced with real-world data sets many of the models run into a critical issue: they are formulated in terms of fully-observed data, whereas in practice the data sets are plagued with missing data. The theory of statistical model estimation from incomplete data is conceptually similar to the estimation of latent-variable models, where powerful tools such as variational inference (VI) exist. However, in contrast to standard latent-variable models, parameter estimation with incomplete data often requires estimating exponentially-many conditional distributions of the missing variables, hence making standard VI methods intractable. We address this gap by introducing variational Gibbs inference (VGI), a new general-purpose method to estimate the parameters of statistical models from incomplete data. We validate VGI on a set of synthetic and real-world estimation tasks, estimating important machine learning models such as variational autoencoders and normalising flows from incomplete data. The proposed method, whilst general-purpose, achieves competitive or better performance than existing model-specific estimation methods.

Keywords: statistical model estimation, variational inference, Gibbs sampling, missing data, amortised inference

1. Introduction

This paper introduces a new general-purpose method to estimate statistical models from incomplete data that is well-suited for modern (deep) statistical models. Estimating statistical models is one of the core tasks in machine learning because the fitted model can be used in many practical downstream tasks, such as, classification, prediction, anomaly detection, data augmentation, and missing data imputation (e.g. Goodfellow et al., 2016, Chapter 5.1.1). However, most of the current methods require large amounts of fully-observed data at training time, and hence they remain largely impractical in many real-world domains that are overwhelmed with incomplete data. For example, the vast amounts of data gathered by online systems is sparse, with ratings data used in recommender systems often missing 95-99% of the total data (Marlin et al., 2011). Similarly, a review of medical trial studies has identified that 95% of studies contained missing data, with as much as 70%

of the data values missing in some studies (Bell et al., 2014). This prevalence of missing data in real-world scenarios warrants a need for principled approaches to efficiently handle missing data in machine learning.

A principled classical approach to estimating statistical models from incomplete data is expectation-maximisation (EM, Dempster et al., 1977) which aims at maximising the likelihood, however, it is mostly limited to simple models. Monte Carlo EM (MCEM, Wei and Tanner, 1990) is a less limited version of classical EM that can be understood as an iterative method that fits the statistical model on imputations derived from itself. However, exact conditional sampling for modern statistical models is typically impossible. One way to (approximately) sample imputations from a joint statistical model is via Markov chain Monte Carlo methods (MCMC, e.g. Barber, 2017, Chapter 27.4), however they tend to be computationally intensive and hence scale poorly to larger data sets (Blei et al., 2017). Variational inference (VI, Jordan et al., 1999) is often a computationally more performant alternative, while sometimes lacking the asymptotic exactness guarantees of MCMC. In case of missing data, however, as we will elaborate in the paper (Sections 2.1-2.2), existing (amortised) VI methods would require $2^d - 1$ variational distributions, one for each non-trivial pattern of missingness, and thus scale poorly with the dimension of the data d. Their applicability to estimating statistical models from incomplete data has thus been strongly limited and our paper addresses this gap in the literature.

1.1 Main Contributions

Our main contribution is a novel general-purpose method for estimating statistical models from incomplete data. The method combines the computational performance of VI and the expressiveness of Markov chains, which makes it well-suited for modern (deep) statistical models. Crucially, the method only requires d rather than $2^d - 1$ variational distributions and thereby overcomes the limitations of existing (amortised) VI methods that prevented their use for model estimation from incomplete data. We achieve this reduction from exponential to linear growth by leveraging techniques that are related to those used by popular imputation methods (see Section 2.3). As the proposed method is based on variational inference (VI) and the Gibbs sampler, we call it variational Gibbs inference (VGI).

Figure 1 illustrates VGI.[1] The method starts with K initial random imputations of the missing values of each incomplete data-point. Each iteration of our algorithm has two steps—a learning and an imputation step. In the learning step, the statistical model and the variational distributions are updated by maximising a variational lower-bound on the log-likelihood (ⓐ and ⓒ in the figure). In the imputation step, the missing values are imputed via pseudo-Gibbs sampling using the learnt variational distributions (ⓑ in the figure). These imputations are *persistent* and updated in subsequent iterations. In this way, the initial imputations iteratively adapt to the target statistical model, such that they follow the joint distribution of the missing variables conditional on the observed data. Moreover, our method facilitates parameter sharing across missingness patterns by using an amortised inference model and is amenable to parallelisation, which further increases computational efficiency.

1. An interactive demo is available at github.com/vsimkus/variational-gibbs-inference.

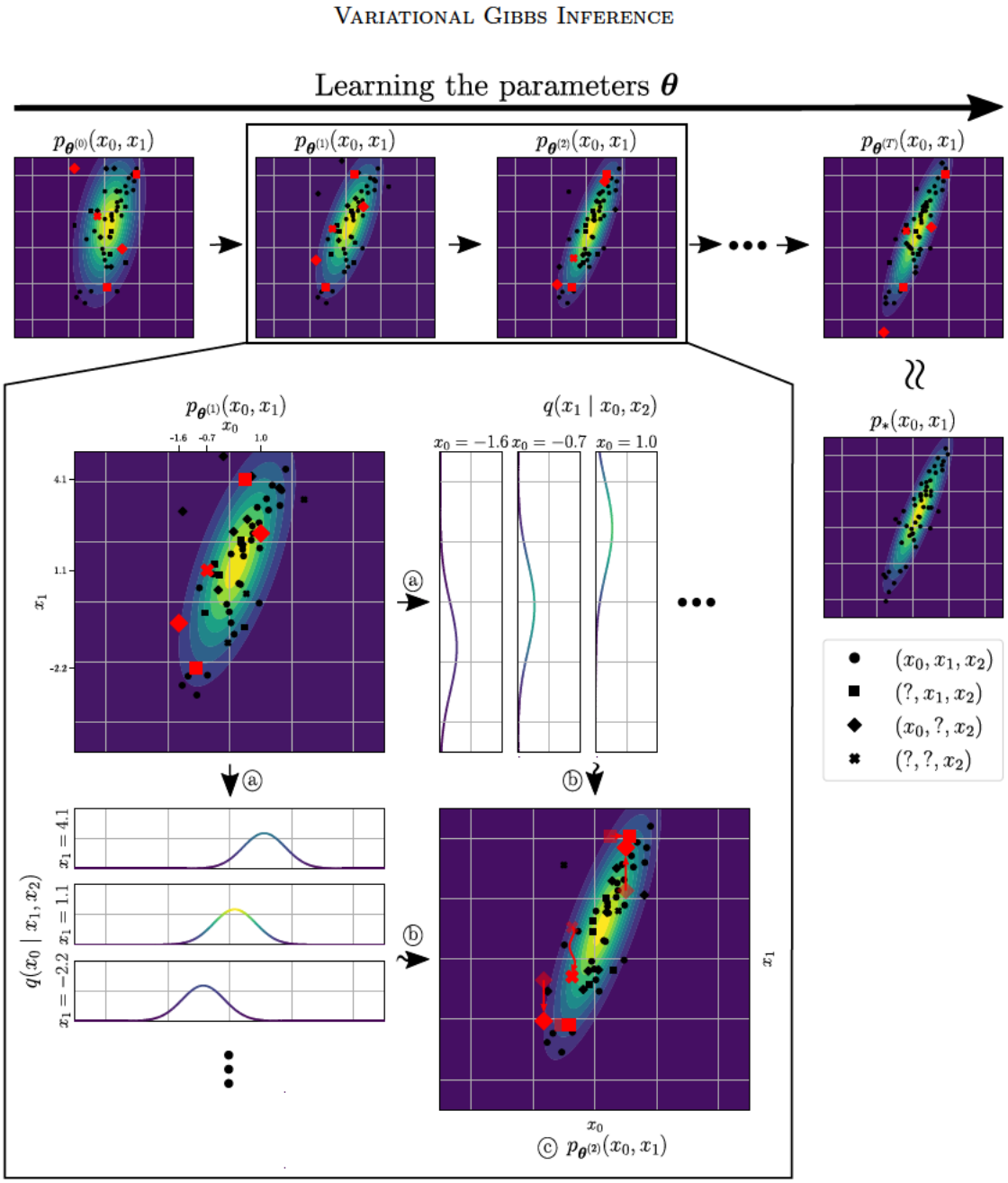


Figure 1: A schematic illustration of our method on a 3D toy model $p_{\boldsymbol{\theta}}(x_1, x_2, x_3)$, with missingness only in the first two dimensions for ease of exposition.
The top row along the arrow shows contours of the model $p_{\boldsymbol{\theta}^{(t)}}$ with parameters $\boldsymbol{\theta}^{(t)}$ learnt iteratively as it approaches the complete-data estimate p_*.
The zoomed-in pane details an iteration of the algorithm, where:
ⓐ univariate variational conditionals $q(x_j \mid \boldsymbol{x}_{\setminus j})$ are learnt from $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ and
ⓑ used to update the imputations via pseudo-Gibbs sampling;
ⓒ $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ is learnt using the imputed data via a variational objective.
Compared to standard VI, our approach reduces the number of variational distributions that need to be learnt from $2^d - 1$ to d.

1.2 Overview of the Paper

The paper is organised as follows. Section 2 provides background on statistical model estimation and standard variational inference for incomplete data, discusses the technical gap in the literature on amortised variational inference of missing data, and describes a classical approach for missing data imputation that is related to our work.

In Section 3, we derive the VGI optimisation objective (Section 3.1), present the VGI algorithm (Section 3.2), and discuss practical considerations for modelling the variational Gibbs conditionals (Section 3.3). We further introduce a block-Gibbs version of VGI (Section 3.4), which we later use in Section 5.3 to adapt our method specifically to the variational autoencoder (VAE, Kingma and Welling, 2013; Rezende et al., 2014). The details of evaluating models fitted with VGI for model selection using incomplete held-out data are presented next (Section 3.5). A discussion in Section 3.6 on the similarities and differences between VGI and related methods closes Section 3.

In Section 4, we validate and analyse our method on low- and high-dimensional toy problems where analytical solutions exist. In Section 5 we compare VGI against model-specific estimation methods on a VAE model and show that VGI produces competitive results in terms of model accuracy. In Section 6 we apply our general-purpose method to normalising flow estimation (Rezende and Mohamed, 2015) and show that it can outperform a flow-specific estimation method.

In Section 7 we summarise our findings and discuss possible future research directions.

2. Background

In this section we provide the background on statistical model estimation and variational inference with missing data, amortised variational inference, and a popular missing data imputation method based on conditional modelling. We highlight the shortcomings of those methodologies that the proposed method addresses.

2.1 Model Estimation and Variational Inference with Incomplete Data

Statistical model estimation from observed data is typically solved via maximum-likelihood estimation (MLE). Following Little and Rubin (2002) the marginal observed likelihood for incomplete data is defined via a joint model $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}}, \boldsymbol{m})$ of the observed variables of interest $\boldsymbol{x}_{\text{obs}}$ and the binary missingness mask $\boldsymbol{m}$,

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}}, \boldsymbol{m}) = \int p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}) p(\boldsymbol{m} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}) \, \mathrm{d}\boldsymbol{x}_{\text{mis}}, \tag{1}$$

where the complete variable of interest $\boldsymbol{x}$ is defined by its observed and missing components $\boldsymbol{x} = (\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$, $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$ is the statistical model of the data, and $p(\boldsymbol{m} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$ is the missingness model representing the missing data mechanism. In this paper we assume that data are missing at random (MAR) or missing completely at random (MCAR) so that the missingness model can be ignored when estimating the statistical model of the data $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ (Rubin, 1976). The maximum likelihood estimate of the parameters $\boldsymbol{\theta}$ is then given

VARIATIONAL GIBBS INFERENCE

										$q(\boldsymbol{x}^i_{\text{mis}} \mid \boldsymbol{x}^i_{\text{obs}})$
$\boldsymbol{x}^1$	x^1_1	?	x^1_3	x^1_4	$\boldsymbol{m}^1$	1	0	1	1	$q(x^1_2 \mid x^1_1, x^1_3, x^1_4)$
$\boldsymbol{x}^2$	?	x^2_2	x^2_3	?	$\boldsymbol{m}^2$	0	1	1	0	$q(x^2_1, x^2_4 \mid x^2_2, x^2_3)$
$\boldsymbol{x}^3$	?	?	?	x^3_4	$\boldsymbol{m}^3$	0	0	0	1	$q(x^3_1, x^3_2, x^3_3 \mid x^3_4)$
⋮		⋮			⋮		⋮			⋮

Figure 2: Left: Observed incomplete data. Middle: missingness mask. There are potentially as many as 2^d different missingness patterns. Right: corresponding variational posterior distributions.

by

$$\arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{N} p_{\boldsymbol{\theta}}(\boldsymbol{x}^i_{\text{obs}}, \boldsymbol{m}^i) = \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{N} \int p_{\boldsymbol{\theta}}(\boldsymbol{x}^i_{\text{obs}}, \boldsymbol{x}^i_{\text{mis}}) \, \mathrm{d}\boldsymbol{x}^i_{\text{mis}}, \tag{2}$$

where N is the number of data-points and i denotes the index of a data-point. A brief review of the different missingness mechanisms and a proof of the above is provided in Appendix A.

The integral over the missing components in (2) is typically intractable, which renders the likelihood and hence standard maximum-likelihood estimation intractable. Exactly the same problem occurs when estimating latent-variable models. Indeed, we can consider the missing components to be latent variables and hence obtain a tractable lower bound on the log-likelihood as done in variational inference (VI, Jordan et al., 1999). Following the standard derivation of the evidence lower-bound (ELBO) (e.g. Barber, 2017, Chapter 11.2), the bound on the log-likelihood for N incomplete data points is

$$\frac{1}{N} \sum_{i=1}^{N} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}^i_{\text{obs}}) \geq \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{q(\boldsymbol{x}^i_{\text{mis}} | \boldsymbol{x}^i_{\text{obs}})} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}^i_{\text{obs}}, \boldsymbol{x}^i_{\text{mis}})}{q(\boldsymbol{x}^i_{\text{mis}} \mid \boldsymbol{x}^i_{\text{obs}})} \right],$$

where $q(\boldsymbol{x}^i_{\text{mis}} \mid \boldsymbol{x}^i_{\text{obs}})$ are the variational distributions. Maximising the ELBO with respect to the model parameters $\boldsymbol{\theta}$ and variational distributions $q \in \mathcal{Q}$ in some distributional family $\mathcal{Q}$ yields an approximate MLE solution to (2). If the variational distributions are equal to the model conditionals $p_{\boldsymbol{\theta}}(\boldsymbol{x}^i_{\text{mis}} \mid \boldsymbol{x}^i_{\text{obs}})$ for all $\boldsymbol{x}^i_{\text{obs}}$ then the bound is tight and the solution can be made exact.[2]

Whilst it is straightforward to obtain a tractable lower bound on the log-likelihood, there is a crucial complication in the missing data problem that sets it apart from standard (amortised) VI problems: for d-dimensional data, there is not one but possibly $2^d - 1$ such variational conditional distributions, namely one for each non-trivial pattern of missingness,

2. Note that to be able to satisfy this condition the variational family should include the model conditionals $p_{\boldsymbol{\theta}}(\boldsymbol{x}^i_{\text{mis}} \mid \boldsymbol{x}^i_{\text{obs}}) \in \mathcal{Q}$. Still, such specification may not guarantee the exactness of the MLE solution, since due to local optima in the ELBO the variational distributions may not perfectly match the model conditionals.

as illustrated in Figure 2. This raises fundamental technical issues that have prevented the application of VI to parameter estimation from incomplete data. We discuss these complications in the next subsection.

2.2 The Difficulty of Amortising Missing Variable Inference

Classical variational inference requires the specification of one variational distribution q per observed data-point, but such an approach is computationally inefficient in modern large-scale use-cases due to a lack of parameter sharing. An amortised version of VI (Gershman and Goodman, 2014) deals with this issue by incorporating global parameter sharing. The key idea in amortised VI is to parametrise the conditional variational distribution by a deterministic function, or an inference network, of the observed inputs with globally shared parameters $\boldsymbol{\phi}$. However, in the missing data setting we want to represent all $2^d - 1$ conditional distributions caused by the different missingness patterns. The exponential growth in the number of missingness patterns means that the naïve approach of using one inference network per pattern results in a lack of parameter sharing across data-points even for moderate-dimensional data, thus cancelling the computational advantages of amortised VI.

Efficiently amortising a variational distribution $q_{\boldsymbol{\phi}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ for any $(\boldsymbol{x}_{\text{mis}}, \boldsymbol{x}_{\text{obs}}) \in \boldsymbol{x}$ entails simultaneously dealing with two problems: (i) handling all possible combinations of variables in the conditioning set, that is, for all $\boldsymbol{x}_{\text{obs}} \in \boldsymbol{x} \setminus \boldsymbol{x}_{\text{mis}}$, and (ii) constructing pdfs/pmfs for arbitrary sets of target variables $\boldsymbol{x}_{\text{mis}} \in \boldsymbol{x} \setminus \boldsymbol{x}_{\text{obs}}$. Existing work has focused on the first problem, approaching it by either simply fixing the input dimensionality of the inference network to d and then padding the missing inputs with zeroes (Nazábal et al., 2020; Mattei and Frellsen, 2019), or using permutation-invariant network architectures (Ma et al., 2019). However, there is no work in the VI literature that addresses the second problem.[3] Dealing with this problem requires care—placing restrictions on the variational family may result in a biased estimate of the target statistical model (as shown below). Hence, to match the possibly complex conditional distributions of the target model we would like to use unrestricted probabilistic models for the variational family whilst still being able to take advantage of the increased efficiency of amortised VI.

We illustrate how restricting the variational family can reduce the quality of the fitted target model. For example, taking inspiration from mean-field VI (e.g. Bishop, 2006, Section 10.1.1) one may assume independence of the missing variables given the observed and work with variational distributions of the form

$$q_{\boldsymbol{\phi}}(\boldsymbol{x}^i_{\text{mis}} \mid \boldsymbol{x}^i_{\text{obs}}) = \prod_{j \in \text{idx}(\boldsymbol{m}^i)} q_{\boldsymbol{\phi}}(x^i_j \mid \boldsymbol{x}^i_{\text{obs}}),$$

where $\text{idx}(\boldsymbol{m}^i)$ denotes the set of indices of the missing values in the data-point. However, in Figure 3 we show that such independence assumption can significantly bias the learnt probabilistic model by artificially reducing correlations in the completed data, with the bias increasing with the fraction of missingness.

In this work, we introduce a novel variational method that can fit statistical models using only d variational conditionals, thus facilitating parameter sharing without introducing strong statistical assumptions or restricting their modelling capability.

3. But see Section 3.6 for related work.

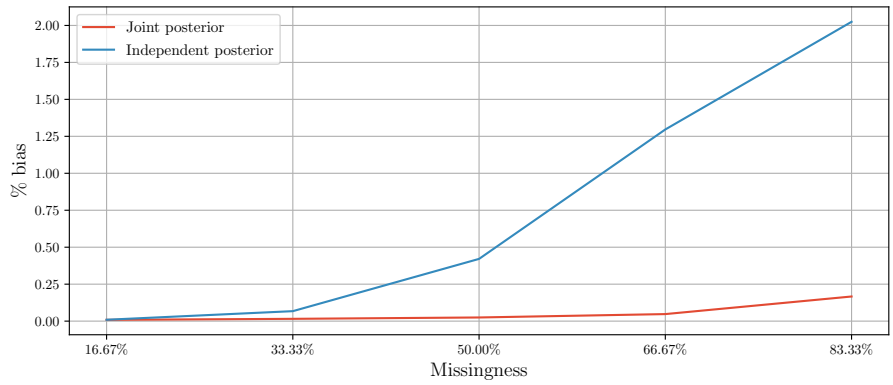


Figure 3: Percent bias measured on the test log-likelihood of a toy factor analysis model with correct (red) and incorrect (blue) modelling assumptions. Log-likelihood was computed on complete test data. We observe that with the incorrect independence assumption $x_j \perp\!\!\!\perp \boldsymbol{x}_{\text{mis}\setminus j} \mid \boldsymbol{x}_{\text{obs}}$ for $\forall j \in \text{idx}(\boldsymbol{m})$ the bias is significantly larger compared to the correctly-specified joint model. Percent bias is computed as $100 \cdot \left|(\mathbb{E}\left[\ell(p_{\boldsymbol{\theta}})\right] - \ell(p_*))/\ell(p_*)\right|$ (e.g. van Buuren, 2018, Chapter 2.5), where $\ell(p_{\boldsymbol{\theta}})$ and $\ell(p_*)$ are the average log-likelihood on test data using the fitted $p_{\boldsymbol{\theta}}$ and the ground truth p_* models, and the expectation is over different runs of the learning algorithm.

2.3 Imputing Missing Data with Fully-Conditional Models

For multivariate missing data, imputations can be generated by iteratively sampling univariate conditional distributions $f(x_j \mid \boldsymbol{x}_{\setminus j}; \boldsymbol{\phi}_j)$ for $\forall j \in \text{idx}(\boldsymbol{m})$ with local parameters $\boldsymbol{\phi}_j$. Imputation methods in this family, which learn the conditionals and impute the data iteratively, are known as fully-conditional specification (FCS) or sequential iterative regression methods (Brand, 1999; Rubin, 2003; van Buuren et al., 2006). They have been essential tools in statistical analysis of incomplete data for over two decades.

The most popular of the FCS imputation methods is multivariate imputation by chained equations (MICE, van Buuren and Oudshoorn, 2000). MICE is an iterative framework that starts with a random imputation from the observed data and then sequentially imputes each incomplete variable one-by-one. Each step consists of fitting a conditional distribution $f(x_j \mid \boldsymbol{x}_{\setminus j}; \boldsymbol{\phi}_j)$ on the data-points where x_j is observed via regression and then imputing the x_j in data-points where it is not observed by sampling the learnt conditional (see Appendix C for more details about MICE). To capture the uncertainty of the missing variables, the MICE procedure is usually independently repeated K number of times to obtain K imputations of the missing data, an approach known as multiple-imputation (MI, Rubin, 1987a). After imputation, statistical models $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ can be fit on the imputed data

sets using standard methods for complete data.[4] We provide more background on such an "impute-then-fit" approach in Appendix D.

Whilst the sampling procedure bears some similarity to Gibbs sampling (Geman and Geman, 1984), see Appendix B for a short review, an important theoretical difference between Gibbs sampling and FCS is that in standard Gibbs sampling one starts with a joint model of the target distribution and then, by decomposing it into full-conditionals, samples a Markov chain that will eventually converge to the joint target distribution. On the other hand, the univariate distributions with disjoint parameters $\boldsymbol{\phi}_j$ in FCS are not guaranteed to have a joint distribution (Arnold et al., 1999, Section 1.6), which is why this approach has been called pseudo-Gibbs sampling (Heckerman et al., 2000).[5] Nevertheless, despite the possibly incompatible univariate distributions, it has been empirically found that the procedure can generate good imputations of multivariate missing values in many practical settings (Rubin, 2003; van Buuren et al., 2006), which motivates us to use a similar factorisation to represent the variational distributions of the missing variables.

FCS has several desirable properties: (i) it easily lends itself to the specification of flexible imputation models since even for univariate Gaussian conditionals the joint distribution can be complex and multi-modal (e.g. Arnold et al., 1999, Section 3.4), (ii) heterogenous data (mixed continuous and discrete) can be handled by using different distribution families for each conditional, (iii) it alleviates the problem of having to learn $2^d - 1$ joint missing variable distributions into one that requires learning only d univariate Gibbs conditionals, and (iv) univariate conditional distributions can be easily constrained to prevent invalid imputation values.

Inspired by the empirical success of FCS methods on multiple-imputation tasks, in the next section we propose a variational approach that characterises the joint variational distributions $q_{\boldsymbol{\phi}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ via d variational full-conditional models. Thus, the joint variational distribution $q_{\boldsymbol{\phi}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ can be made very flexible since we neither require strong statistical independence assumptions nor restrictive distributional assumptions on the variational families (see Section 2.2), which enables us to well optimise the ELBO and hence obtain a good (approximate) MLE estimate of the statistical model $p_{\boldsymbol{\theta}}(\boldsymbol{x})$. In this way, we generalise variational inference of univariate missing data to the multivariate case, akin to how FCS generalises univariate regression-based imputation to the multivariate setting (e.g. van Buuren, 2018, Chapter 4.5.1).

3. Variational Gibbs Inference

We present variational Gibbs inference (VGI) to estimate the parameters of statistical models from incomplete data by maximising a variational lower-bound on the log-likelihood. The method requires only d variational conditionals and uses an iterative algorithm that alternates between two steps: (i) the learning step which fits the statistical model and the variational conditionals by optimising the variational objective, and (ii) the imputation-step which updates persistent pseudo-Gibbs chains that provide the imputations for each incom-

4. However, the standard multiple-imputation workflow for parameter estimation is generally not applicable to statistical models when their parameters are not identifiable, as is most often the case in deep generative models, and hence caution must be taken (see Appendix D).

5. Sometimes also called incompatible Gibbs sampling or compound conditional specification (Rubin, 2003).

Variational Gibbs Inference

Term	Reference	Description
$p_{\boldsymbol{\theta}}(\boldsymbol{x})$	Sec. 2.1	Statistical model of the data
$q_{\boldsymbol{\phi}_j}(x_j \mid \boldsymbol{x}_{\text{mis}\setminus j}, \boldsymbol{x}_{\text{obs}})$	Sec. 3.3	Variational conditionals
$\tau_{\boldsymbol{\phi}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{x}}_{\text{mis}})$	Eq. (4)	Gibbs transition kernel
$\tilde{\tau}_{\boldsymbol{\phi}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{x}}_{\text{mis}})$	Eq. (15)	Extended(-Gibbs) transition kernel (see Appendix E)
$f^t_{\boldsymbol{\phi}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$	Eq. (5)	Marginal distribution of a Markov chain at step t
$\mathcal{J}^t_{\text{VGI}}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}_{\text{obs}})$	Eq. (9)	The VGI objective for a single data-point $\boldsymbol{x}_{\text{obs}}$ at step t

Table 1: A glossary of key terms used in VGI.

plete data-point. The following sections introduce the variational objective (Section 3.1) and the VGI algorithm (Section 3.2), discuss practicalities to consider when modelling the variational conditionals (Section 3.3), explain how VGI can be adapted to specific statistical models, such as latent-variable models (Section 3.4), describe the details of evaluating the method on incomplete held-out data (Section 3.5), and finally discusses the related work (Section 3.6).

3.1 The Variational Objective

We derive a variational ELBO to estimate the statistical model $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ from incomplete data using the marginal distributions $f^t_{\boldsymbol{\phi}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ of Markov chains with learnable parameters $\boldsymbol{\phi}$ as the variational distribution of the missing variables. Maximising the objective allows us to learn the parameters $\boldsymbol{\theta}$ of the model $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ and the parameters $\boldsymbol{\phi}$ of the Gibbs transition *kernel* $\tau_{\boldsymbol{\phi}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{x}}_{\text{mis}})$,[6] or equivalently, the parameters of the d univariate Gibbs conditionals (*variational conditionals*) $q_{\boldsymbol{\phi}_j}(x_j \mid \tilde{\boldsymbol{x}}_{\text{mis}\setminus j}, \boldsymbol{x}_{\text{obs}})$ which characterise the kernel. By learning the kernel $\tau_{\boldsymbol{\phi}}$ we can match the marginal distributions of the imputation Markov chains $f^t_{\boldsymbol{\phi}}$ to the conditional distributions of the target model $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$. Consequently, the variational conditionals $q_{\boldsymbol{\phi}_j}(x_j \mid \boldsymbol{x}_{\text{mis}\setminus j}, \boldsymbol{x}_{\text{obs}})$ are learnt to approximate the true conditionals $p_{\boldsymbol{\theta}}(x_j \mid \boldsymbol{x}_{\text{mis}\setminus j}, \boldsymbol{x}_{\text{obs}})$.

Representing the variational distribution implicitly via the samples of a Markov chain allows the VGI method to work with a flexible variational model for adaptive imputations, which in turn enables tightening of the variational lower-bound and thus the maximisation of the likelihood. Moreover, by representing the kernel with d variational conditionals, VGI achieves efficient amortisation of the exponentially-many distributions of missing data, similar to fully-conditional imputation methods discussed in Section 2.3. A summary of the key terms in VGI is provided in Table 1.

We start our derivation with the standard variational ELBO, but with the marginal distribution $f^t_{\boldsymbol{\phi}}$ of a Markov chain as the variational distribution

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}}) \geq \mathcal{L}(\boldsymbol{\theta}, f^t_{\boldsymbol{\phi}}; \boldsymbol{x}_{\text{obs}}), \quad \mathcal{L}(\boldsymbol{\theta}, f^t_{\boldsymbol{\phi}}; \boldsymbol{x}_{\text{obs}}) = \mathbb{E}_{f^t_{\boldsymbol{\phi}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}}, \boldsymbol{x}_{\text{obs}})}{f^t_{\boldsymbol{\phi}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})} \right], \quad (3)$$

where t denotes the iteration of our algorithm. To maximise the likelihood of the model $p_{\boldsymbol{\theta}}$ we want the bound to be sufficiently tight, and hence $f^t_{\boldsymbol{\phi}}$ needs be flexible and not biased

6. We use $\tilde{\boldsymbol{x}}_{\text{mis}}$ to denote the previous imputation value before the transition.

by the choice of the initial imputation distribution f^0. The flexibility of the marginal distribution of a Markov chain depends on the kernel, and the bias induced by the choice of f^0 decreases with the number of steps t in the Markov chain (e.g. Cover and Thomas, 2006, Chapter 4.4). Hence we want $\tau_{\boldsymbol{\phi}}$ to be flexible and t to become sufficiently large.

However, we cannot evaluate the above lower-bound since the imputation distribution $f^t_{\boldsymbol{\phi}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ is implicitly-defined by the Markov chain. Moreover, maximising the bound with respect to the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ with gradient-based methods would be expensive for large t and might suffer from gradient instability. Rather than estimating the gradients of the above lower-bound with respect to $\boldsymbol{\phi}$ over the full length of the Markov chain, we propose "cutting" the chain just before the current transition and optimising the transition kernel $\tau_{\boldsymbol{\phi}}$ greedily as we sample. The imputations obtained in iteration t will then be re-used in the next iteration to further improve $\tau_{\boldsymbol{\phi}}$ and $p_{\boldsymbol{\theta}}$. This allows us to derive a tractable and efficient variational method.

We choose the kernel, which transforms the imputations from the previous iteration, $\tilde{\boldsymbol{x}}_{\text{mis}}$, and produces updated imputations $\boldsymbol{x}_{\text{mis}}$, to be Gibbs (hence the name of the method) and define it as follows

$$\tau_{\boldsymbol{\phi}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{x}}_{\text{mis}}) = \sum_{j \in \text{idx}(\boldsymbol{m})} \boldsymbol{\pi}(j) q_{\boldsymbol{\phi}_j}(x_j \mid \boldsymbol{x}_{\text{mis}\setminus j}, \boldsymbol{x}_{\text{obs}}) \delta(\boldsymbol{x}_{\text{mis}\setminus j} - \tilde{\boldsymbol{x}}_{\text{mis}\setminus j}), \tag{4}$$

where the dimension index $j \in \text{idx}(\boldsymbol{m})$ is controlled by a fixed uniform selection distribution $\boldsymbol{\pi}(j)$ of a random-scan Gibbs sampler, x_j and $\boldsymbol{x}_{\text{mis}\setminus j}$ denote the j-th missing dimension and the remaining missing dimensions of $\boldsymbol{x}_{\text{mis}}$ respectively, and the kernel is specified by the d variational conditionals $q_{\boldsymbol{\phi}_j}(x_j \mid \boldsymbol{x}_{\text{mis}\setminus j}, \boldsymbol{x}_{\text{obs}})$. Alternatively, we can also let the updated imputation x_j depend on the previous imputation value $\tilde{x}_j$, which gives variational conditionals of the form $q_{\boldsymbol{\phi}_j}(x_j^t \mid \boldsymbol{x}_{\text{obs}}, \tilde{x}_j^{t-1}, \tilde{\boldsymbol{x}}^{t-1}_{\text{mis}\setminus j})$. We call a kernel $\tilde{\tau}_{\boldsymbol{\phi}}$ (or subsequently the variational model) that uses this form of conditionals an *extended-Gibbs* kernel (or, for conciseness, *extended* kernel) and provide an analogous derivation of this section in Appendix E.

Marginalising out $\tilde{\boldsymbol{x}}_{\text{mis}}$ in (4) with respect to the *imputation distribution* f^{t-1} from the previous step in the Markov chain gives the imputation distribution after a single Gibbs update

$$\begin{aligned} f^t_{\boldsymbol{\phi}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}) &= \int \tau_{\boldsymbol{\phi}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{x}}_{\text{mis}}) f^{t-1}(\tilde{\boldsymbol{x}}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}) \, \mathrm{d}\tilde{\boldsymbol{x}}_{\text{mis}} && (5) \\ &= \sum_{j \in \text{idx}(\boldsymbol{m})} \boldsymbol{\pi}(j) f^t_{\boldsymbol{\phi}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, j) \\ &= \mathbb{E}_{\boldsymbol{\pi}(j)} \left[f^t_{\boldsymbol{\phi}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, j) \right], && (6) \end{aligned}$$

where $f^t_{\boldsymbol{\phi}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, j)$ is the imputation distribution f^{t-1} updated in dimension j,

$$f^t_{\boldsymbol{\phi}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, j) = q_{\boldsymbol{\phi}_j}(x_j \mid \boldsymbol{x}_{\text{mis}\setminus j}, \boldsymbol{x}_{\text{obs}}) f^{t-1}(\boldsymbol{x}_{\text{mis}\setminus j} \mid \boldsymbol{x}_{\text{obs}}). \tag{7}$$

Note that the absence of $\boldsymbol{\phi}$ in f^{t-1} corresponds to the aforementioned "cutting" of the chain.

Variational Gibbs Inference

Now, we continue with the standard ELBO from (3) and use (6) and (7) to derive the variational Gibbs ELBO at iteration t

$$
\begin{aligned}
\mathcal{L}^t(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}_{\text{obs}}) &\overset{(6)}{=} \mathbb{E}_{\boldsymbol{\pi}(j) f^t_{\boldsymbol{\phi}}(\boldsymbol{x}_{\text{mis}} | \boldsymbol{x}_{\text{obs}}, j)} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})}{f^t_{\boldsymbol{\phi}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})} \right] \\
&= \mathbb{E}_{\boldsymbol{\pi}(j) f^t_{\boldsymbol{\phi}}(\boldsymbol{x}_{\text{mis}} | \boldsymbol{x}_{\text{obs}}, j)} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})}{f^t_{\boldsymbol{\phi}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, j)} + \log \frac{f^t_{\boldsymbol{\phi}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, j)}{f^t_{\boldsymbol{\phi}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})} \right] \\
&= \mathbb{E}_{\boldsymbol{\pi}(j) f^t_{\boldsymbol{\phi}}(\boldsymbol{x}_{\text{mis}} | \boldsymbol{x}_{\text{obs}}, j)} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})}{f^t_{\boldsymbol{\phi}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, j)} \right] \\
&\quad + \mathbb{E}_{\boldsymbol{\pi}(j)} D_{\text{KL}}(f^t_{\boldsymbol{\phi}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, j) \mid\mid f^t_{\boldsymbol{\phi}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})) \\
&\geq \mathbb{E}_{\boldsymbol{\pi}(j) f^t_{\boldsymbol{\phi}}(\boldsymbol{x}_{\text{mis}} | \boldsymbol{x}_{\text{obs}}, j)} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})}{f^t_{\boldsymbol{\phi}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, j)} \right] \\
&\overset{(7)}{=} \mathbb{E}_{\boldsymbol{\pi}(j) f^{t-1}(\boldsymbol{x}_{\text{mis}\setminus j} | \boldsymbol{x}_{\text{obs}}) q_{\boldsymbol{\phi}_j}(x_j | \boldsymbol{x}_{\text{mis}\setminus j}, \boldsymbol{x}_{\text{obs}})} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})}{q_{\boldsymbol{\phi}_j}(x_j \mid \boldsymbol{x}_{\text{mis}\setminus j}, \boldsymbol{x}_{\text{obs}})} \right] \\
&\quad - \mathbb{E}_{\boldsymbol{\pi}(j) f^{t-1}(\boldsymbol{x}_{\text{mis}\setminus j} | \boldsymbol{x}_{\text{obs}})} \left[\log f^{t-1}(\boldsymbol{x}_{\text{mis}\setminus j} \mid \boldsymbol{x}_{\text{obs}}) \right],
\end{aligned} \tag{8}
$$

where the inequality follows from the non-negativity of KL divergence. If $\boldsymbol{x}_{\text{mis}}$ is independent of j, which holds for the stationary distribution of the Markov chain characterised by the variational conditionals (see Appendix F), then the KL divergence term is zero and maximising (8) is equivalent to maximising (3).

The ELBO in (8) can be optimised efficiently with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ since only samples from the imputation distribution $f^{t-1}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ are needed and the intractable entropy term in the last line of the equation does not depend on the parameters $\boldsymbol{\theta}$ or $\boldsymbol{\phi}$, and hence does not need to be computed. Removing the entropy term, we obtain the variational Gibbs inference (VGI) objective $\mathcal{J}^t_{\text{VGI}}$ at iteration t for one incomplete data-point $\boldsymbol{x}_{\text{obs}}$:[7]

$$
\mathcal{J}^t_{\text{VGI}}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}_{\text{obs}}) = \mathbb{E}_{\boldsymbol{\pi}(j) f^{t-1}(\boldsymbol{x}_{\text{mis}\setminus j} | \boldsymbol{x}_{\text{obs}}) q_{\boldsymbol{\phi}_j}(x_j | \boldsymbol{x}_{\text{mis}\setminus j}, \boldsymbol{x}_{\text{obs}})} \left[\log \frac{p_{\boldsymbol{\theta}}(x_j, \boldsymbol{x}_{\text{mis}\setminus j}, \boldsymbol{x}_{\text{obs}})}{q_{\boldsymbol{\phi}_j}(x_j \mid \boldsymbol{x}_{\text{mis}\setminus j}, \boldsymbol{x}_{\text{obs}})} \right] \tag{9}
$$

Importantly, while cutting the chain removes the entropy of f^{t-1} from the objective function, the entropy of the variational conditionals remains. This prevents the collapse of the variational conditionals to point masses and consequently also prevents the collapse of the joint imputation distribution f^t that is sampled using the fitted conditionals.[8]

For N incomplete data-points we will maximise the averaged objective to obtain the parameter estimates $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\phi}}$ at iteration t

$$
\hat{\boldsymbol{\theta}}^t, \hat{\boldsymbol{\phi}}^t = \arg\max_{\boldsymbol{\theta}, \boldsymbol{\phi}} \frac{1}{N} \sum_{i=1}^{N} \mathcal{J}^t_{\text{VGI}}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}^i_{\text{obs}}). \tag{10}
$$

7. The objective using the extended kernel (Appendix E) is almost the same, but the variational conditional additionally includes a dependency on $\tilde{x}_j$ and hence depends on all of the imputed variables from the previous iteration $(\tilde{x}_j, \boldsymbol{x}_{\text{mis}\setminus j}) \sim f^{t-1}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$.
8. Note that the imputation distribution $f^{t-1}(\boldsymbol{x}_{\text{mis}\setminus j} \mid \boldsymbol{x}_{\text{obs}})$ is kept fixed when we maximise the objective $\mathcal{J}^t_{\text{VGI}}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}_{\text{obs}})$ with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, due to the "cutting" of the Markov chain. After updating the parameters, we update $f^{t-1}(\boldsymbol{x}_{\text{mis}\setminus j} \mid \boldsymbol{x}_{\text{obs}})$ via the updated transition kernel according to (5).

In practice, we optimise (10) using stochastic gradient ascent and approximate the expectations in $\mathcal{J}^t_{\text{VGI}}$ with Monte Carlo integration

$$\hat{\mathcal{J}}^t_{\text{VGI}}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}_{\text{obs}}) = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{M}\sum_{m=1}^{M}\left[\log \frac{p_{\boldsymbol{\theta}}(x^k_{j^m}, \boldsymbol{x}^k_{\text{mis}\setminus j^m}, \boldsymbol{x}_{\text{obs}})}{q_{\boldsymbol{\phi}_j}(x^k_{j^m} \mid \boldsymbol{x}^k_{\text{mis}\setminus j^m}, \boldsymbol{x}_{\text{obs}})}\right], \tag{11}$$

where $x^k_{j^m} \sim q_{\boldsymbol{\phi}_j}(x^k_{j^m} \mid \boldsymbol{x}^k_{\text{mis}\setminus j^m}, \boldsymbol{x}_{\text{obs}})$, $j^m \sim \boldsymbol{\pi}(j)$, $\boldsymbol{x}^k_{\text{mis}\setminus j^m} \sim f^{t-1}(\boldsymbol{x}_{\text{mis}\setminus j^m} \mid \boldsymbol{x}_{\text{obs}})$, and K is the number of imputations for each incomplete data-point, M is the number of samples used to approximate the expectation with respect to j (and x_j), and f^{t-1} is represented via samples from a fixed number of Markov chains. We empirically found that using small M and K was sufficient in most of our experiments.[9]

To summarise, we have derived the key VGI objective $\mathcal{J}^t_{\text{VGI}}$, which maximises a lower-bound on the log-likelihood using samples from a marginal distribution of a variational Markov chain at any iteration t. Importantly, the objective uses only d variational conditionals to represent all $2^d - 1$ possible conditional distributions of missing data. We next describe the VGI algorithm which integrates the optimisation of the iteration-dependent variational objective $\mathcal{J}^t_{\text{VGI}}$ and sampling from the Markov chains $f^t(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ into an iterative procedure.

3.2 The VGI Algorithm

The variational Gibbs inference (VGI) algorithm is shown in Algorithm 1 with subroutines summarised in Algorithms 2-5.[10] The core objective of the algorithm is to fit a statistical model $p_{\boldsymbol{\theta}}$ on incomplete data set $\mathcal{D}$ using a variational model $q_{\boldsymbol{\phi}}$ of the Gibbs conditionals that is learnt jointly with $p_{\boldsymbol{\theta}}$. The method uses stochastic gradient optimisation (e.g. Spall, 2003; Ruder, 2017) to learn the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ efficiently by processing the data set in mini-batches, which are randomly chosen subsets of the full data set $\mathcal{D}$. Algorithm 1 consists of three stages: initialisation, model warm-up, and the main iterative stage, which we describe below.

Initialisation. In line 1 the algorithm starts with an incomplete data set $\mathcal{D} = \{(\boldsymbol{x}^i_{\text{obs}}, \boldsymbol{m}^i)\}_i$ and produces a K-times imputed data set $\mathcal{D}_K = \{(\boldsymbol{x}^i_{\text{obs}}, \boldsymbol{m}^i, \boldsymbol{x}^{(i,1)}_{\text{mis}}, \ldots, \boldsymbol{x}^{(i,K)}_{\text{mis}})\}_i$, where each incomplete data-point is imputed using an initial imputation distribution $\boldsymbol{x}^{(i,k)}_{\text{mis}} \sim f^0(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}^i_{\text{obs}})$. The f^0 should be chosen such that it provides a good starting distribution for the Gibbs sampler. Empirically we found that choosing f^0 to be the marginal empirical distribution worked well, which corresponds to the maximum-entropy distribution of the missing data.

Warm-up. The warm-up stage (lines 2-3) has two parts: a variational model $q_{\boldsymbol{\phi}}$ initialisation stage (line 2) and a statistical model $p_{\boldsymbol{\theta}}$ warm-up stage (line 3), which we describe below in the given order. We note that the warm-up stage can be optional, however we empirically found that it allows the models to take reasonable initial values, which can sometimes improve the model fit or stabilise the learning for a small additional initialisation cost.

9. We found $K \in [5, 10]$ and $M \in [1, 10]$ sufficient in our experiments.
10. Note that in our experiments we use optimised implementation of the algorithms by parallelising the loops where possible. Our implementation is available at `https://github.com/vsimkus/variational-gibbs-inference`.

VARIATIONAL GIBBS INFERENCE

Algorithm 1 Variational Gibbs inference (VGI) algorithm

Input: $p_{\boldsymbol{\theta}}(\boldsymbol{x})$, statistical model with parameters $\boldsymbol{\theta}$
$q_{\boldsymbol{\phi}_j}(x_j \mid \boldsymbol{x}_{\setminus j})$ for $j \in \{1 \dots d\}$, variational conditional models with parameters $\boldsymbol{\phi}$
$\mathcal{D}$, incomplete data set
K, number of imputations of each incomplete data-point
$f^0(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$, initial imputation distribution
$\alpha_{\boldsymbol{\theta}}$ and $\alpha_{\boldsymbol{\phi}}$, the parameter learning rates
max_epochs, number of epochs
Output: $\boldsymbol{\theta}, \boldsymbol{\phi}$, and K-times imputed data $\mathcal{D}_K$

Create K-times imputed data $\mathcal{D}_K$ using f^0
Warm up the parameters $\boldsymbol{\phi}$ of the variational model $q_{\boldsymbol{\phi}}$ using Algorithm 2
Warm up the parameters $\boldsymbol{\theta}$ of the statistical model $p_{\boldsymbol{\theta}}$ using Algorithm 4
for t in $[1, \textit{max\_epochs}]$ **do**
  for mini-batch $\mathcal{B}_K$ in $\mathcal{D}_K$ **do**
    Update imputations in $\mathcal{B}_K$ using pseudo-Gibbs sampler (see Algorithm 5)
    Store the updated imputations in $\mathcal{B}_K$ for use in the next epoch
    Compute $\hat{\mathcal{J}}^t_{\text{VGI}}$ in (11) using Algorithm 3
    $\boldsymbol{\theta}^t = \boldsymbol{\theta}^{t-1} + \alpha_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}} \hat{\mathcal{J}}^t_{\text{VGI}}$ ▷ Update params of $p_{\boldsymbol{\theta}}$ with a stochastic gradient step
    $\boldsymbol{\phi}^t = \boldsymbol{\phi}^{t-1} + \alpha_{\boldsymbol{\phi}} \nabla_{\boldsymbol{\phi}} \hat{\mathcal{J}}^t_{\text{VGI}}$ ▷ Update params of $q_{\boldsymbol{\phi}}$ with a stochastic gradient step
  end for
end for

The warm-up procedure for the variational model is outlined in pseudo-code in Algorithm 2. In standard variational inference for latent variable models, the variational model is often initialised randomly. However, in the missing-data case, starting-out with randomly-initialised variational conditionals may be sub-optimal since there is usually some observed data that could be used to reasonably initialise the variational model. Therefore, to make use of the available observed data, we suggest pre-training the variational inference networks on the observed values using maximum-likelihood estimation and stochastic gradient ascent (SGA) via

$$\hat{\boldsymbol{\phi}} = \arg\max_{\boldsymbol{\phi}} \mathcal{J}_{\boldsymbol{\phi}}(\mathcal{D}_K), \tag{12}$$

$$\mathcal{J}_{\boldsymbol{\phi}}(\mathcal{D}_K) = \sum_{j=1}^{d} \frac{1}{|\text{obs}(\mathcal{D}_K, j)|} \sum_{i \in \text{obs}(\mathcal{D}_K, j)} \frac{1}{K} \sum_{k=1}^{K} \log q_{\boldsymbol{\phi}_j}(x^i_j \mid \boldsymbol{x}^{(i,k)}_{\text{mis}}, \boldsymbol{x}^i_{\text{obs}\setminus j}),$$

where $\text{obs}(\mathcal{D}_K, j)$ represents the set of indices of data-points that are observed in the j-th dimension. This pre-training procedure is probabilistic regression for all dimensions $j \in [1, d]$ using the data where the j-th dimension is observed, with the missing values imputed with a baseline imputation method (for example, samples from the empirical marginals).[11] We find that a small number of pre-training iterations is often sufficient to favourably

11. Care must be taken when warming-up the extended-Gibbs kernel to avoid poor initialisation of the conditionals, see technical aside in Appendix E.

initialise the variational model, which can boost the performance of the method and stabilise training in the initial iterations of the main stage of the algorithm.

Next, the algorithm performs SGA on the parameters $\boldsymbol{\theta}$ of the statistical model $p_{\boldsymbol{\theta}}$ using the variational Gibbs objective $\hat{\mathcal{J}}_{\mathrm{VGI}}$ in (11) (see Algorithms 3 and 4). This kind of "warm-up" allows the parameters $\boldsymbol{\theta}$ to take on reasonable initial values before the main iterative stage starts. We qualitatively find that the warm-up stage generally needs to be performed for a small number of iterations until the change in $\hat{\mathcal{J}}_{\mathrm{VGI}}$ between consecutive iterations falls below a threshold. Note that we keep the variational parameters $\boldsymbol{\phi}$ fixed at this stage so that the variational model does not deteriorate while the model $p_{\boldsymbol{\theta}}$ is being initialised. Also, in this stage the imputed data in $\mathcal{D}_K$ are not yet updated, they still follow the initial imputation distribution f^0. This is because we empirically found that early Gibbs sampling can cause divergent imputation chains, training instability, or getting stuck in a local optima.

Main stage. The main stage of the algorithm iterates between updating the imputations in $\mathcal{D}_K$ and fitting the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ (see lines 6-10). For each mini-batch $\mathcal{B}_K$ we first update the imputed values $\boldsymbol{x}_{\mathrm{mis}}^{(i,1)}, \ldots, \boldsymbol{x}_{\mathrm{mis}}^{(i,K)}$ using G Gibbs updates with the variational conditionals (see Algorithm 5) and then update the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ using a single stochastic gradient of the variational Gibbs objective $\hat{\mathcal{J}}_{\mathrm{VGI}}^t$ in (11), see Algorithm 3 on computing the objective. These two steps (imputation and parameter update) are then repeated until convergence or until the computational budget is exhausted.

We find that using a small number of Gibbs updates G is generally sufficient and preferable in terms of the trade-off between compute cost and convergence rate.[12] Moreover, a large value of G may cause training instability due to the generalisation gap of the variational conditionals (see Section 3.5 for more details). Updating the imputations via Gibbs sampling using the variational conditionals and storing them for the next iteration corresponds to updating the marginal Markov chain distribution $f^t(\boldsymbol{x}_{\mathrm{mis}} \mid \boldsymbol{x}_{\mathrm{obs}})$. Reusing the imputations from the previous iteration, rather than re-sampling from scratch at each iteration, is akin to using persistent chains that have previously been used in a different context in Markov chain Monte Carlo methods (e.g. persistent contrastive divergence, Younes, 1999; Tieleman, 2008).

3.3 Choosing the Variational Model

The performance of amortised variational methods, and hence also VGI, critically depends on the choice of the functional family and the expressivity of the inference networks used to parametrise them. To maximise the compatibility of the statistical model $p_{\boldsymbol{\theta}}$ and the variational model $q_{\boldsymbol{\phi}}$, we suggest the variational family should be chosen as one that includes or is close to the family of the statistical model. If the target model $p_{\boldsymbol{\theta}}$ is a deep model, the variational inference networks should use architectural blocks in the neural network that are similar to the ones used in the target model.

We must also consider how to specify and parametrise the d variational conditionals. A straightforward way to specify them is to use d inference networks with parameters $\boldsymbol{\phi}_j$, one for each variational conditional. However, such an approach would be parameter-inefficient

12. We used $G \in [1, 5]$ in our experiments.

Variational Gibbs Inference

Algorithm 2 Variational model warm-up

Input: $q_{\boldsymbol{\phi}_j}(x_j \mid \boldsymbol{x}_{\setminus j})$ for $j \in \{1 \dots d\}$, variational conditionals; $\mathcal{D}_K$, K-times imputed data; $\alpha_{\boldsymbol{\phi}}$, the parameter learning rate; *var_warmup_epochs*, number of epochs.
Output: $\boldsymbol{\phi}$ initialised on observed data using (12)

for t_w in $[1, \textit{var\_warmup\_epochs}]$ **do**
  for mini-batch $\mathcal{B}_K$ in $\mathcal{D}_K$ **do**
    $\mathcal{J}_{\boldsymbol{\phi}} = 0$
    for j in $[1, d]$ **do**
      $\mathcal{J}_{\boldsymbol{\phi}} \mathrel{+}= \frac{1}{|\text{obs}(\mathcal{B}_K, j)|} \mathcal{J}^j_{\boldsymbol{\phi}}(\mathcal{B}_K, j)$
    end for
    $\boldsymbol{\phi} \leftarrow \boldsymbol{\phi} + \alpha_{\boldsymbol{\phi}} \nabla_{\boldsymbol{\phi}} \mathcal{J}_{\boldsymbol{\phi}}$
  end for
end for

def $\mathcal{J}^j_{\boldsymbol{\phi}}(\mathcal{B}_K, j)$:
  $\mathcal{J}^j_{\boldsymbol{\phi}} = 0$
  for $\boldsymbol{x}^i_{\text{obs}}, \boldsymbol{m}^i, \boldsymbol{x}^{(i,1)}_{\text{mis}}, \dots, \boldsymbol{x}^{(i,K)}_{\text{mis}}$ in $\mathcal{B}_K$ **do**
    if x^i_j is observed **then**
      $\mathcal{J}^j_{\boldsymbol{\phi}} \mathrel{+}= \frac{1}{K} \sum_{k=1}^{K} \log q_{\boldsymbol{\phi}_j}(x^i_j \mid \boldsymbol{x}^{(i,k)}_{\text{mis}}, \boldsymbol{x}^i_{\text{obs}\setminus j})$
    end if
  end for
  return $\mathcal{J}^j_{\boldsymbol{\phi}}$

Algorithm 3 Compute VGI objective in (11)

Input: $p_{\boldsymbol{\theta}}(\boldsymbol{x})$, statistical model; $q_{\boldsymbol{\phi}_j}(x_j \mid \boldsymbol{x}_{\setminus j})$ for $j \in \{1 \dots d\}$, variational conditionals; $\mathcal{B}_K$, K-times imputed mini-batch; M, number of missing dimensions to sample.
Output: $\hat{\mathcal{J}}_{\text{VGI}}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}_{\text{obs}})$ averaged over all $\boldsymbol{x}^i_{\text{obs}}$ in mini-batch $\mathcal{B}_K$

$\hat{\mathcal{J}}_{\text{VGI}} = 0$
for $\boldsymbol{x}^i_{\text{obs}}, \boldsymbol{m}^i, \boldsymbol{x}^{(i,1)}_{\text{mis}}, \dots, \boldsymbol{x}^{(i,K)}_{\text{mis}}$ in $\mathcal{B}_K$ **do**
  for k in $[1, K]$ **do**
    $\hat{\mathcal{J}}_{\text{VGI}} \mathrel{+}= \frac{1}{|\mathcal{B}_K| K} \hat{\mathcal{J}}^{(i,k)}_{\text{VGI}}(\boldsymbol{x}^i_{\text{obs}}, \boldsymbol{x}^{(i,k)}_{\text{mis}})$
  end for
end for

def $\hat{\mathcal{J}}^{(i,k)}_{\text{VGI}}(\boldsymbol{x}^i_{\text{obs}}, \boldsymbol{x}^{(i,k)}_{\text{mis}})$:
  $\hat{\mathcal{J}}^{(i,k)}_{\text{VGI}} = 0$
  for m in $[1, M]$ **do**
    $j^m \sim \boldsymbol{\pi}(j)$ ▷ Sample from idx($\boldsymbol{m}^i$)
    $x^k_{j^m} \sim q_{\boldsymbol{\phi}_j}(x^k_{j^m} \mid \boldsymbol{x}^{(i,k)}_{\text{mis}\setminus j^m}, \boldsymbol{x}^i_{\text{obs}})$
    $\hat{\mathcal{J}}^{(i,k)}_{\text{VGI}} \mathrel{+}= \log \frac{p_{\boldsymbol{\theta}}(x^k_{j^m}, \boldsymbol{x}^{(i,k)}_{\text{mis}\setminus j^m}, \boldsymbol{x}^i_{\text{obs}})}{q_{\boldsymbol{\phi}_j}(x^k_{j^m} \mid \boldsymbol{x}^{(i,k)}_{\text{mis}\setminus j^m}, \boldsymbol{x}^i_{\text{obs}})}$
  end for
  return $\frac{1}{M} \hat{\mathcal{J}}^{(i,k)}_{\text{VGI}}$

Algorithm 4 Statistical model warm-up

Input: $p_{\boldsymbol{\theta}}(\boldsymbol{x})$, statistical model
$q_{\boldsymbol{\phi}_j}(x_j \mid \boldsymbol{x}_{\setminus j})$ for $j \in \{1 \dots d\}$
$\mathcal{D}_K$, K-times imputed data
$\alpha_{\boldsymbol{\theta}}$, the parameter learning rate
model_warmup_epochs, # of epochs
Output: $\boldsymbol{\theta}$ initialised for the main loop

for t_w in $[1, \textit{model\_warmup\_epochs}]$ **do**
  for mini-batch $\mathcal{B}_K$ in $\mathcal{D}_K$ **do**
    Estimate $\hat{\mathcal{J}}_{\text{VGI}}$ using Algorithm 3
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}} \hat{\mathcal{J}}_{\text{VGI}}$
  end for
end for

Algorithm 5 (Pseudo-)Gibbs sampling

Input: $q_{\boldsymbol{\phi}_j}(x_j \mid \boldsymbol{x}_{\setminus j})$ for $j \in \{1 \dots d\}$
$\mathcal{B}_K$, K-times imputed mini-batch
G, number of Gibbs update steps
Output: $\mathcal{B}_K$ with updated imputations

for $\boldsymbol{x}^i_{\text{obs}}, \boldsymbol{m}^i, \boldsymbol{x}^{(i,1)}_{\text{mis}}, \dots, \boldsymbol{x}^{(i,K)}_{\text{mis}}$ in $\mathcal{B}_K$ **do**
  for k in $[1, K]$ **do**
    for g in $[1, G]$ **do**
      $j \sim \boldsymbol{\pi}(j)$ ▷ Sample from idx($\boldsymbol{m}^i$)
      $x_j \sim q_{\boldsymbol{\phi}_j}(x_j \mid \boldsymbol{x}^{(i,k)}_{\text{mis}\setminus j}, \boldsymbol{x}^i_{\text{obs}})$
      $\boldsymbol{x}^{(i,k)}_{\text{mis}} \leftarrow (\boldsymbol{x}^{(i,k)}_{\text{mis}\setminus j}, x_j)$
    end for
  end for
end for

and scale poorly to higher dimensional data. To address this, we suggest using the extended-Gibbs conditionals $q_{\boldsymbol{\phi}_j}(x_j^t \mid \boldsymbol{x}_{\text{obs}}, \tilde{x}_j^{t-1}, \tilde{\boldsymbol{x}}_{\text{mis}\setminus j}^{t-1})$ (Appendix E), which allow the imputation x_j^t to depend on the previous imputation value $\tilde{x}_j^{t-1}$. Then, we can use a single partially-shared neural network where the parameters and the computations are shared in the first part of the network for all conditional distributions. These two simple modifications allows us to scale VGI to higher-dimensional data (see Appendix G for a more detailed discussion). We investigate the effects of parameter-sharing and extended conditionals in Section 4.7.

Finally, an important caveat of our approach is that, due to approximation errors, the fitted variational conditionals may not correspond to a joint distribution, which mirrors a similar caveat of the FCS methods outlined in Section 2.3. Hence, pseudo-Gibbs sampling using the fitted conditionals may diverge (we revisit this in Section 3.5 when discussing model selection).[13] Nevertheless, we show in the experimental sections that the VGI method works well despite the lack of such convergence guarantees, similar to the existing FCS imputations methods (see Section 2.3).

3.4 Variational Block-Gibbs Inference and Latent-Variable Models

Thus far we have considered the general case where the number of possible missingness patterns is 2^d, with d being the dimensionality of the data. In some cases, missingness in a block of dimensions may be coupled such that either all of the dimensions in the block are missing or all are observed, which reduces the number of possible missingness patterns to 2^s, where s is the number of such missing variable blocks, with $s = d$ if the missingness in no dimension is coupled.

We can adapt the VGI lower-bound in (8) and equivalently the VGI objective in (9) to this scenario by letting j denote the index of a block of missing dimensions, rather than the index of a single missing dimension. Hence, the univariate x_j in (8) and (9) become potentially multivariate random variables $\boldsymbol{x}_j$. We must then specify a variational model for each block of missing dimensions, where we may choose to specify independent variational models or partially-shared models as discussed in Section 3.3. The Gibbs update step in line 6 of Algorithm 1 then corresponds to the update of a block-Gibbs sampler. Hence, we refer to this method as the variational block-Gibbs inference (VBGI).

A particularly important instance of coupled missingness is the latent-variable model $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_z) p_{\boldsymbol{\theta}}(\boldsymbol{x}_z)$, where all of the latent-variables $\boldsymbol{x}_z$ are missing together. If the missingness of no other dimension is coupled, then $j \in \text{idx}(\boldsymbol{m}) \cup \{z\}$ refers to either any one of the missing dimensions or all of the latents. Rewriting $\mathcal{J}_{\text{VGI}}^t$ for this model we get the following objective:[14]

$$\mathcal{J}_{\text{VBGI}}^t(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}_{\text{obs}}) = \mathbb{E}_{f^{t-1}(\tilde{\boldsymbol{x}}_{\text{mis}}, \tilde{\boldsymbol{x}}_z \mid \boldsymbol{x}_{\text{obs}})} \Bigg(\sum_{j \in \text{idx}(\boldsymbol{m})} \boldsymbol{\pi}(j) \mathbb{E}_{q_{\boldsymbol{\phi}_j}(x_j \mid \tilde{\boldsymbol{x}}_{\text{mis}\setminus j}, \tilde{\boldsymbol{x}}_z, \boldsymbol{x}_{\text{obs}})} \left[\log \frac{p_{\boldsymbol{\theta}}(x_j, \tilde{\boldsymbol{x}}_{\text{mis}\setminus j}, \boldsymbol{x}_{\text{obs}} \mid \tilde{\boldsymbol{x}}_z) p_{\boldsymbol{\theta}}(\tilde{\boldsymbol{x}}_z)}{q_{\boldsymbol{\phi}_j}(x_j \mid \tilde{\boldsymbol{x}}_{\text{mis}\setminus j}, \tilde{\boldsymbol{x}}_z, \boldsymbol{x}_{\text{obs}})} \right]$$

13. Note that, after learning, we advocate using the estimated p_θ for generating imputations if they were required and not the variational conditionals.

14. Similar to the standard VGI case, the VBGI objective can optionally be adapted to use the extended-Gibbs conditionals (see Appendix E).

$$+ \boldsymbol{\pi}(j = z)\mathbb{E}_{q_{\boldsymbol{\phi}_z}(\boldsymbol{x}_z|\tilde{\boldsymbol{x}}_{\text{mis}},\boldsymbol{x}_{\text{obs}})}\left[\log \frac{p_{\boldsymbol{\theta}}(\tilde{\boldsymbol{x}}_{\text{mis}}, \boldsymbol{x}_{\text{obs}} \mid \boldsymbol{x}_z)p_{\boldsymbol{\theta}}(\boldsymbol{x}_z)}{q_{\boldsymbol{\phi}_z}(\boldsymbol{x}_z \mid \tilde{\boldsymbol{x}}_{\text{mis}}, \boldsymbol{x}_{\text{obs}})}\right]\Bigg), \tag{13}$$

where $q_{\boldsymbol{\phi}_z}(\boldsymbol{x}_z \mid \tilde{\boldsymbol{x}}_{\text{mis}}, \boldsymbol{x}_{\text{obs}})$ is the variational conditional for the latents $\boldsymbol{x}_z$. The grouping of missing dimensions allows us to adapt VGI to latent-variable models. It forms the basis for the variational autoencoder-specific version of VGI that we introduce in Section 5.

3.5 Evaluating VGI on Incomplete Held-Out Data

A common step in a machine learning workflow is the validation of the estimated model on a held-out data set, which is used in early stopping or model selection (e.g. Goodfellow et al., 2016). In the incomplete data case, we can expect that the held-out data are also incomplete. One approach is to estimate the marginal log-likelihood $\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}})$ using importance sampling or sequential Monte Carlo methods (e.g. Barber, 2017), however these approaches can be too computationally expensive to obtain reliable estimates of the marginal on the held-out data, especially if we wanted to evaluate many models. Alternatively, a validation loss, analogue to the training loss, is commonly used to select a model that minimises it. The evaluation of the VGI objective in (9) on incomplete held-out data requires the iterative evaluation of the variational conditionals (Gibbs sampling) on unseen data. However, evaluating these conditionals on unseen data has known pitfalls. We will first review these pitfalls, then explain how it affects the evaluation of VGI, and finally suggest a simple procedure that enables the evaluation of the VGI objective on held-out data and produces sample imputations.

Cremer et al. (2018) have shown that amortised variational distributions may be significantly biased compared to the optimal variational distribution in the same variational family, a phenomenon they called the amortisation gap. They have also demonstrated that the amortisation gap is typically larger on held-out data if the variational model is fitted only to the training data; we will refer to this as the *inference generalisation gap* (see also the concurrent work by Zhang et al., 2021). The reason for the generalisation gap is overfitting of the variational inference network to its inputs, that is the training data, which is generally a consequence of using finite-sized training data sets. Importantly, the inference generalisation gap can appear even when the target model $p_{\boldsymbol{\theta}}$ has not been overfitted.

The same considerations apply to VGI since it is an amortised variational method. In particular, the generalisation gap in VGI can cause the imputation Markov chains to diverge due to the compounding effect of the generalisation errors. This prevents a meaningful evaluation of VGI on the held-out data. We have empirically observed that the chances of divergent behaviour increases with the number of variational conditional models. Mattei and Frellsen (2018) and Cremer et al. (2018) suggested to deal with the inference generalisation gap by fine-tuning the variational models on the held-out data. To avoid information leakage from the held-out data into the model $p_{\boldsymbol{\theta}}$, the estimated model and the corresponding parameters $\boldsymbol{\theta}$ are kept fixed during fine-tuning. Moreover, the fine-tuned variational conditionals should not be used as part of the VGI training. A simple way to achieve this is to make a copy of the variational model before fine-tuning and discarding it afterwards.

In Appendix H, Algorithm 6 we show the VGI fine-tuning procedure. The fine-tuning starts with incomplete held-out data and then fills-in the missing values with random draws from the initial imputation distribution f^0 as in Algorithm 1. Then, during the first iter-

ation it performs several rounds of Gibbs sampling over all missing values using the learnt variational conditionals to improve the imputations. To mitigate possible divergent behaviour due to the inference generalisation gap in the first iteration, the procedure rejects any values outside of the observed-data hypercube, defined by the minimum and maximum values in the observed data. After the imputation warm-up in the first iteration, the algorithm continues akin to the learning of $\boldsymbol{\phi}$ in standard in VGI, fine-tuning the variational kernel $\tau_{\boldsymbol{\phi}}$ and updating the imputations with a small number of Gibbs updates in each iteration, where all proposed imputations are accepted—the acceptance region is no longer necessary, since the kernel is being adapted to the (imputed) held-out data. At the end of the procedure, we obtain the VGI loss on the held-out data as well as imputations of the missing values.

The iterative validation procedure described in this section would be expensive if it were used to continually monitor the validation loss during training. However, the validation loss does not need to be computed at every iteration, computing it only every few iterations is often sufficient. Moreover, we have empirically found that the cost of fine-tuning was often only a small fraction of the training cost.

3.6 Related Work

We here discuss work that is closely related or shares similarities to VGI.

Monte Carlo expectation-maximisation. Monte Carlo EM (MCEM, Wei and Tanner, 1990) has been proposed as an extension of the classical expectation-maximisation (EM, Dempster et al., 1977) algorithm to the setting where the required expectation with respect to the conditional distributions of missing data does not have a closed form expression but is approximated by a Monte Carlo sample average. The method requires sampling from the conditional distribution of missing data at each iteration. Then, like VGI, MCEM iteratively maximises the ELBO using sample imputations of the missing data. In contrast to VGI, MCEM does not use a variational approximation but instead attempts to sample the true conditional distribution, for example, with methods such as rejection sampling (e.g. Barber, 2017, Chapter 27.1.2).

However, sampling from the conditional distributions in higher dimensions usually requires Markov chain Monte Carlo methods (MCMC, e.g. Barber, 2017, Chapter 27.4), which only asymptotically sample from the exact target distribution. In practice, MCMC is used to sample chains of fixed length for computational reasons and hence it may not sample the true conditional distribution. As a consequence, using samples from an unconverged Markov chain in MCEM may adversely affect the learnt model $p_{\boldsymbol{\theta}}$.

Similar to MCEM with MCMC, VGI also samples a Markov chain of imputations using the variational kernel. However, in contrast to MCEM, where the MCMC sampler needs to be restarted at each iteration and run for a sufficient amount of time, in VGI we use a single persistent Markov chain that we update throughout training. This allows our method to eventually sample better imputations of the missing data given that the algorithm is run for long enough. In Section 6 we estimate a normalising flow model (Rezende and Mohamed, 2015; Papamakarios et al., 2021) from incomplete data with VGI and MCEM using flow-specific MCMC and find that VGI outperforms MCEM both in terms of the computational performance and accuracy. In an attempt to improve the performance of MCEM, we have

empirically investigated the use of persistent chains in a Metropolis-Hastings version of MCMC on normalising flows, but found that this drastically reduced the acceptance rate of the proposed transitions and therefore also reduced the accuracy of the estimated statistical model.

Markov chain variational inference. Salimans et al. (2015) proposed a powerful framework for combining Markov chain Monte Carlo methods and variational inference for latent-variable models called Markov chain variational inference (MCVI). Similar to VGI they propose a lower-bound on the log-likelihood using a Markov chain characterised via a variational transition kernel τ. However, the two methods differ fundamentally in their goals: VGI attempts to *learn* a statistical model $p_{\boldsymbol{\theta}}$ from incomplete data while MCVI attempts to efficiently and accurately approximate a conditional distribution under a *fixed* latent-variable model. To achieve their goal, MCVI aims to optimise the variational kernel over the full length of the Markov chain. In order to avoid computing the intractable integral required to obtain the marginal distribution of a Markov chain $f^t_{\boldsymbol{\phi}}$ they propose a further lower-bound on the ELBO in (3) using a "reverse" transition kernel r, which predicts the reverse path of the Markov chain sampled using the variational kernel τ. A naïve application of MCVI to learn a statistical model would be expensive, since it requires simulating long Markov chains of imputations and then backpropagating the gradients through the sampling path. To simplify the optimisation problem, the authors of MCVI have also briefly considered a sequential (greedy) approach similar to ours. However, as shown in Section 3.1, VGI does not need to learn an auxiliary model of the "reverse" kernel r, and in fact, replacing r in the MCVI lower-bound with the true reverse transition (e.g. Murray, 2007, Section 1.4) recovers the tighter lower-bound in (3) (see Appendix I), which is approximately marginalised in VGI.

A further difference between VGI and MCVI is that MCVI has been developed for the latent-variable and not the missing data setting. Hence, unlike VGI, it does not deal with the problem of how to handle the $2^d - 1$ possible patterns of missingness and the consequent exponential growth in the number of required variational distributions.

Coordinate ascent variational inference. The VGI objective in (9) is related to coordinate ascent variational inference (CAVI, e.g. Bishop, 2006, Chapter 10.1.1; Blei et al., 2017, Section 2.4), which also considers the update of only a single dimension using a univariate variational distribution. However, the previous works only considered fully-factorised variational distributions (mean-field assumption), which can significantly bias the estimate of the statistical model as we have demonstrated in Section 2.2. In contrast, our method works without introducing the mean-field factorisation using a highly-flexible implicit variational distribution given by the marginal of a pseudo-Gibbs sampler, which is adapted to the model $p_{\boldsymbol{\theta}}$ by learning the transition kernel (univariate variational conditionals) of the sampler.

Arbitrarily-conditional models. A separate line of research focuses on constructing models such that any conditional distribution $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathrm{u}} \mid \boldsymbol{x}_{\mathrm{obs}})$, where $\boldsymbol{x}_{\mathrm{u}} \subseteq \boldsymbol{x} \setminus \boldsymbol{x}_{\mathrm{obs}}$ may be arbitrarily chosen, is directly available (Li et al., 2020). One could use such models to construct an amortised variational distribution for arbitrary missingness patterns $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathrm{mis}} \mid \boldsymbol{x}_{\mathrm{obs}})$ and use it in the variational ELBO to fit a target statistical model. However, unlike the variational conditionals in VGI that we can choose freely, such arbitrarily-conditional

models typically have restricted modelling capabilities compared to their non-conditional counterparts.

Substantive model compatible FCS. As discussed in Section 2.3 our solution to the 2^d-1 conditional missing data distributions bears similarity to existing imputation methods in the fully-conditionally specified (FCS) family. Standard FCS methods are independent of the target analysis model. However, Bartlett et al. (2015) have proposed a modified version of the method, called substantive model compatible FCS (SMC-FCS), to generate imputations that are congenial with target (nonlinear) regression model. There is thus some similarity to VGI in the sense that both methods take the target model into account. However, SMC-FCS is about imputing data compatible with regression models (and hence is for supervised learning), while our method estimates joint statistical models (and hence is focused on unsupervised learning). From our understanding SMC-FCS does not apply or readily extend to our setting.

Bayesian data augmentation. Alternatively to maximum-likelihood estimation (MLE) methods, such as the EM or VI, one can also use Bayesian inference to learn models from incomplete data. The advantage of Bayesian methods is that they provide a natural way to evaluate the epistemic uncertainty of the model. One classical method is the data augmentation (DA) algorithm (Tanner and Wong, 1987). In DA, one must first specify a prior over the parameters of the model $p(\boldsymbol{\theta})$ and assume initial imputations $X_{\text{mis}}^{(0)}$.[15] Then, the algorithm iteratively samples the posterior over the parameters $\boldsymbol{\theta}^{(t)} \sim p(\boldsymbol{\theta} \mid X_{\text{mis}}^{(t-1)}, X_{\text{obs}})$ and updates imputations (or augmentations) $X_{\text{mis}}^{(t)} \sim p(X_{\text{mis}}^{(t)} \mid X_{\text{obs}}, \boldsymbol{\theta}^{(t)})$. After an initial warm-up (or burn-in) period, the algorithm produces samples of the missing values and model parameters from the posterior distribution $p(X_{\text{mis}}, \boldsymbol{\theta} \mid X_{\text{obs}})$. Similar to VGI, the iterative procedure in DA is a Gibbs sampler. The main difference between the two is that DA treats the parameters of the model just like the missing variables and samples them accordingly. However, for complex models, just like in the MCEM case, the required distributions are usually only known up to a proportionality and hence computationally expensive MCMC methods would be required to sample them. Moreover, Bayesian inference for modern statistical models, such as VAEs and flows, suffers from scalability issues and is still an active area of research (Gal, 2016; Maddox et al., 2019; Izmailov et al., 2021; Abdar et al., 2021).

4. Experiments on Toy Models

In this section we demonstrate VGI on low- and high-dimensional factor analysis models. We analyse the accuracy of the learnt statistical models and the variational conditionals as well as the effect of the extended-Gibbs variational conditionals. The code for this and the following experimental sections is available at `https://github.com/vsimkus/variational-gibbs-inference`.

15. We use the capital X to denote all data-points $\boldsymbol{x}^i$ in the data set.

4.1 Factor Analysis Model

Factor analysis (FA, e.g. Barber, 2017, Chapter 21.1) is a linear latent-variable model that is often used to discover unobserved factors $\boldsymbol{z}$ from observed data $\boldsymbol{x}$. The prior distribution of $\boldsymbol{z}$ is assumed to be a multivariate standard Gaussian $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{0}, \boldsymbol{I})$, and the distribution of $\boldsymbol{x}$ given a value of $\boldsymbol{z}$ is

$$p_{\boldsymbol{\theta}}(\boldsymbol{x} \mid \boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{F}\boldsymbol{z} + \boldsymbol{\mu}, \boldsymbol{\Psi}),$$

and the marginal distribution is

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{F}\boldsymbol{F}^\top + \boldsymbol{\Psi}),$$

where the parameters $\boldsymbol{\theta}$ are the mean vector $\boldsymbol{\mu}$, the factor matrix $\boldsymbol{F}$, and the diagonal matrix $\boldsymbol{\Psi}$ defining the variances of the observation noise.

FA is a linear version of the variational autoencoder (VAE, Kingma and Welling, 2013; Rezende et al., 2014) that is often used as a toy model to analyse new methods (e.g. Williams et al., 2018). Hence, we start our analysis on a FA model before proceeding to evaluate VGI on more complex models in the next sections. Given its relative simplicity, it is possible to fit a FA model on incomplete data with the expectation-maximisation (EM) algorithm (see Appendix J for details). The EM algorithm provides a best-case solution that we can use to gauge the accuracy of the model estimated by VGI (see below). Moreover, reference conditional distributions $p_{\boldsymbol{\theta}}(x_j \mid \boldsymbol{x}_{\setminus j})$ can be computed analytically (Petersen and Pedersen, 2012), such that we can evaluate the accuracy of the variational conditionals, which is instrumental to producing good imputations and subsequently—to achieving an accurate fit of the model.

4.2 Data

We evaluate VGI on two data sets:

Toy data. A synthetic data set generated with a 6-dimensional FA model with a 2-dimensional latent space (see Appendix K for the ground truth parameters). The training data set has 6400 data-points and the test data set has 5000 data-points.

FA-Frey. A synthetic 560-dimensional data set based on the Frey data,[16] where a FA model with 43 latent dimensions was first fitted on the original data and then used to synthesise a new data set. The training data set has 2400 data-points and the test data set has 3000 data-points.

We consider five fractions of missingness in the training data, ranging from 16.6% to 83.3%, and simulate incomplete training data by generating a binary missingness mask uniformly at random (MCAR). The data-points that are rendered fully missing are removed from the training set, since doing so does not affect the maximum-likelihood estimate under MAR or MCAR missingness.

Where the experiments are repeated multiple times to obtain confidence intervals, the underlying training data are kept constant and only the missingness mask is re-sampled.

16. The original data set is available at `https://cs.nyu.edu/~roweis/data/frey_rawface.mat`.

Thus, the results demonstrate the accuracy of the estimation methods in the presence of missing data, rather than the variability of the model estimate on different realisations from the ground truth data generating distribution.

4.3 Experimental Settings

In our evaluation, we assume that the statistical model $p_{\boldsymbol{\theta}}$ is well-specified. That is, we fit a FA model to data following a FA model with the same latent dimensionality, so that the evaluation can focus on the effect of missing data, rather than on robustness to model specification. We parametrise $\boldsymbol{\Psi}$ as $\boldsymbol{\Psi} = \exp(\boldsymbol{\gamma})$, initialise the parameter $\boldsymbol{F}$ with samples from a standard normal distribution, and set $\boldsymbol{\mu}$ and $\boldsymbol{\gamma}$ to $\mathbf{0}$ and $\mathbf{1}$ respectively.

We specify the variational conditionals in VGI to be univariate Gaussians whose parameters $\log \sigma_j^2$ and μ_j are given by the outputs of a fully-connected neural network. For the toy data we use the standard variational conditionals with an independent network of two hidden layers for each conditional. To make the computations more efficient on the FA-Frey data, we use the extended variational conditionals with a partially-shared network: the first two hidden layers share parameters and computations for all conditional distributions and the last hidden layer has independent parameters for each distribution. The networks use leaky ReLU activation functions with negative slope of 0.01 and the weights are initialised using Kaiming initialisation (He et al., 2015).

We compute the univariate Gaussian entropy terms in (9) analytically using (e.g. Norwich, 1993)

$$-\mathbb{E}_{\mathcal{N}(x_j;\mu_j,\sigma_j^2)}\left[\log \mathcal{N}(x_j;\mu_j,\sigma_j^2)\right] = \frac{1}{2}\log(2\pi\sigma_j^2) + \frac{1}{2},$$

where μ_j and $\log \sigma_j^2$ are given by the inference networks with input $(\tilde{\boldsymbol{x}}_{\text{mis}\setminus j}, \boldsymbol{x}_{\text{obs}})$, or $(\tilde{\boldsymbol{x}}_{\text{mis}}, \boldsymbol{x}_{\text{obs}})$ in the case of the extended variational conditionals.

We fit the model $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ and the variational model using Algorithm 1. We use $K = 5$ imputation chains for each incomplete data-point, and $G = 3$ (toy-data) and $G = 5$ (FA-Frey) Gibbs updates. In the Monte Carlo averaging in $\hat{\mathcal{J}}_{\text{VGI}}$ we select $M = 1$ (toy-data) and $M = 10$ (FA-Frey) missing dimensions. To compute the gradients with respect to the variational parameters we use the reparametrisation trick (Kingma and Welling, 2013). The model parameters $\boldsymbol{\theta} = (\boldsymbol{F}, \boldsymbol{\mu}, \boldsymbol{\gamma})$ are optimised using the Adam optimiser (Kingma and Ba, 2014), whereas the variational parameters $\boldsymbol{\phi}$ are fitted using AMSGrad (Reddi et al., 2018), since we found that using Adam on the variational parameters caused training instability.

4.4 Comparison Methods

We compare VGI against the following methods:

EM (Complete). The ideal case where no data is missing, fitted using EM for FA.

Empirical sample imputation. A weak baseline where the incomplete data is $K = 5$ times imputed with random draws from the empirical distribution of the observed values, then the FA model is fitted as described in Appendix D.

EM (Dempster et al., 1977). An optimal method where the model $p_{\boldsymbol{\theta}}$ is fitted using EM for FA with missing data (see Appendix J). This method presents the best-case

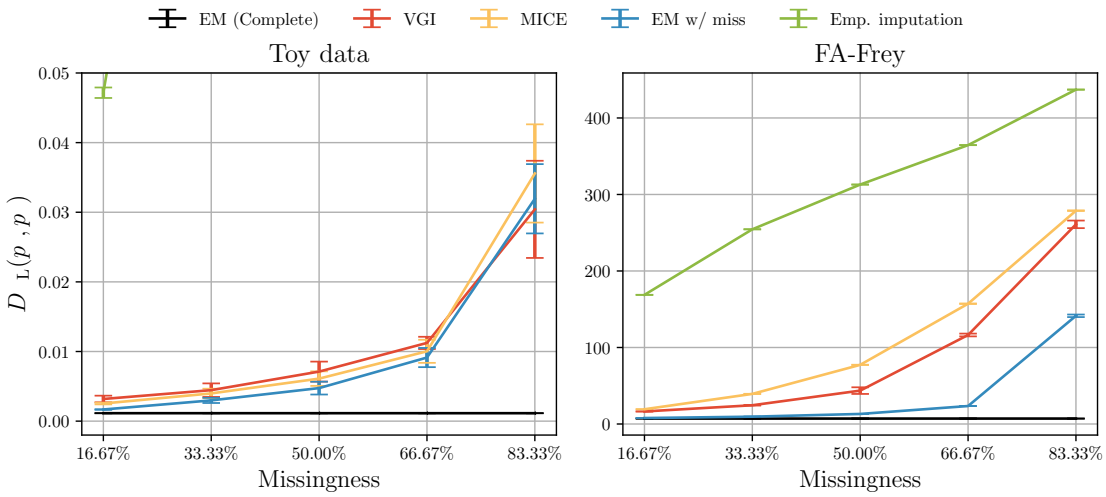


Figure 4: The accuracy of the fitted models $p_{\boldsymbol{\theta}}$ as measured by $D_{\mathrm{KL}}(p_* \,||\, p_{\boldsymbol{\theta}})$ on toy (left) and FA-Frey (right) data. Smaller values are better. The error bars show standard error on the mean of five experiments with different missingness masks and model initialisation. In the left figure, empirical sample imputation performed significantly worse than the posterior approximating methods and hence is not shown for missingness larger than 16.67%.

performance that could be achieved with a variational method, which is equivalent to setting the variational distribution in VI equal to the true conditional distribution (e.g. Barber, 2017, Chapter 11.2.2). Also, in contrast to the other methods, where SGA is used, we here compute the updated distribution parameters analytically.

MICE (van Buuren and Oudshoorn, 2000). A pseudo-Gibbs sampler as described in Section 2.3 that has been widely used in statistical analysis of incomplete data. MICE code has originally been made available in R (van Buuren and Oudshoorn, 2000). We have implemented MICE in Python using the IterativeImputer and Bayesian linear ridge regression as the conditional imputation models, as implemented in the scikit-learn package (Pedregosa et al., 2011). MICE is used to produce $K = 5$ imputations and then the FA model is fitted as described in Appendix D. MICE is another strong baseline because it should be able to produce imputations that are congenial to the target model, since both MICE conditionals and the target FA model are linear Gaussian models.

4.5 Accuracy of the Fitted FA Model

In this toy setting no over-fitting was observed, hence the model parameters $\hat{\boldsymbol{\theta}}$ from the final training iteration were used in the following evaluation for all methods.

We evaluate the accuracy of the fitted FA model using the Kullback–Leibler divergence $D_{\mathrm{KL}}(p_*(\boldsymbol{x}) \mid\mid p_{\boldsymbol{\theta}}(\boldsymbol{x}))$ between the ground truth model p_* and the fitted statistical model $p_{\boldsymbol{\theta}}$ on the two synthetic data sets, shown in Figure 4. It can be immediately seen that the posterior-approximating methods (VGI, MICE, and EM) perform significantly better than the simple empirical imputation baseline showing the clear advantage of these methods over ad-hoc approaches commonly used in practice as a quick fix for missing data.

As expected, MICE performs well on both data sets since the linear imputation model is congenial with the data distribution and the target model. VGI performs comparably to MICE on the toy data (note the overlapping error bars) and shows significant improvement on the FA-Frey data. The better performance on FA-Frey can be attributed to the fact that contrary to MICE, where missing value imputations are generated prior and independently of the model $p_{\boldsymbol{\theta}}$, in VGI the imputations are generated with respect to the model $p_{\boldsymbol{\theta}}$ and are not static throughout training, thus better representing the uncertainty of the imputed values.[17] On the toy data, both VGI and MICE achieved a performance that is comparable to the optimal EM solution, but on the FA-Frey data, the gap between them and EM increases with missingness. We show in Figure 15 of Appendix M that a similar gap appears between EM and Monte Carlo EM (MCEM, Wei and Tanner, 1990) with SGA. Hence, we attribute the performance gap to the stochasticity in the optimisation—Monte Carlo averaging and stochastic gradient ascent—used within MCEM, MICE, and VGI, and hence the gap could be reduced, given sufficient compute resources, via standard means in stochastic optimisation.

4.6 Assessing Estimation Consistency

We further evaluate the statistical consistency of VGI on the toy FA model. The learning rate was decayed according to a cosine schedule in order to satisfy the convergence conditions of stochastic gradient ascent (SGA) (e.g. Spall, 2003, Chapter 4.3.2). Note that in these experiments the variational family of the variational conditionals includes the true conditional distributions, and hence the estimator is expected to be unbiased.

In Figure 5 we plot the log-log curves of the number of incomplete training data-points versus the root mean squared error (RMSE) of the fitted model parameters (left subfigure) and the KL divergence (right subfigure).[18] The plots show a linear behavior in the logarithmic domain, which indicates consistency and conforms with the asymptotic normality of the MLE theory (e.g. Wasserman, 2005, Chapter 9.7). We note that a few points in the figure fall above the linear curve. We attribute this to normal sample variation and to false convergence of SGD due to a potentially sub-optimal learning rate decay schedule (Spall, 2003, Chapter 4.3.2) that could be addressed via hyper-parameter search.

4.7 Accuracy of the Variational Model

The imputation accuracy depends on the fit of the variational conditionals, which subsequently influences the accuracy of the fitted model. Hence, we evaluate, and show in

17. We note that the performance of MICE could be improved by generating more imputations, however with additional computations.
18. The factor loading matrix of the fitted model has been rotated using the (orthogonal) Procrustes rotation (Ten Berge, 1977) before computing the RMSE to resolve the partial non-identifiability of the parameter.

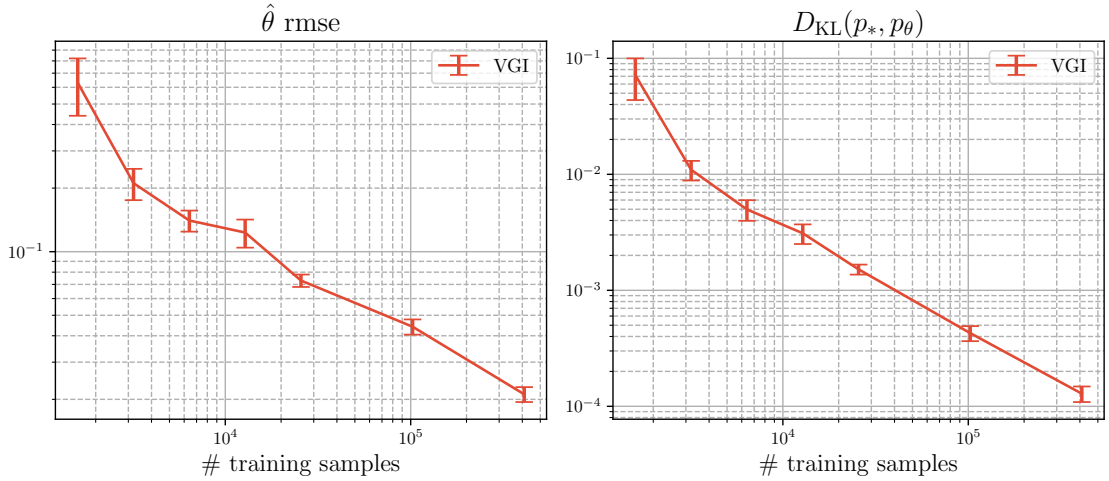


Figure 5: The accuracy of the fitted toy FA model in terms of RMSE with the ground truth parameters (left) and KL divergence to the ground truth model (right) as a function of the number of training samples (on a log-log scale). The error bars show standard error on the mean of five experiments with different missingness masks and model initialisation.

Figure 6, the quality of the fitted variational conditional distributions using the KL divergence between the learnt conditionals and the ground truth distribution $D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(x_j \mid \boldsymbol{x}_{\setminus j}) \mid\mid p_*(x_j \mid \boldsymbol{x}_{\setminus j}))$ in blue and the conditionals of the learnt statistical model $D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(x_j \mid \boldsymbol{x}_{\setminus j}) \mid\mid p_{\boldsymbol{\theta}}(x_j \mid \boldsymbol{x}_{\setminus j}))$ in red, where the conditional distributions are computed on test data.[19]

The KL divergence to the ground truth indicates how good the imputations sampled from the variational conditionals are. It can be seen that the approximation of the ground truth conditionals gets worse as the missingness increases, which is also in line with the accuracy of the target model in Figure 4. This is expected since the variational conditionals approximate the conditional distribution of the target model $p_{\boldsymbol{\theta}}$.

In contrast, the KL divergence to the learnt model $p_{\boldsymbol{\theta}}$ shows how well the variational conditionals approximate the true conditional distribution under the fitted model, this is the objective that is minimised by $\mathcal{J}_{\mathrm{VGI}}$ so we expect it to be low. We observe that the variational conditionals are good at the majority of test data-points (see red curves), however we see a long tail where the variational approximations are poor (see the tail of the red violin plot). This under-performance of the variational conditionals without fine-tuning corresponds to the approximation gap discussed in Section 3.5. Hence, when performing Gibbs sampling with the variational conditionals for model selection, some light fine-tuning of the variational model might be necessary in order to prevent divergent Gibbs imputation chains, as discussed in Section 3.5.

19. For the toy data, we also show results dissected by dimension in Appendix M, Figure 16.

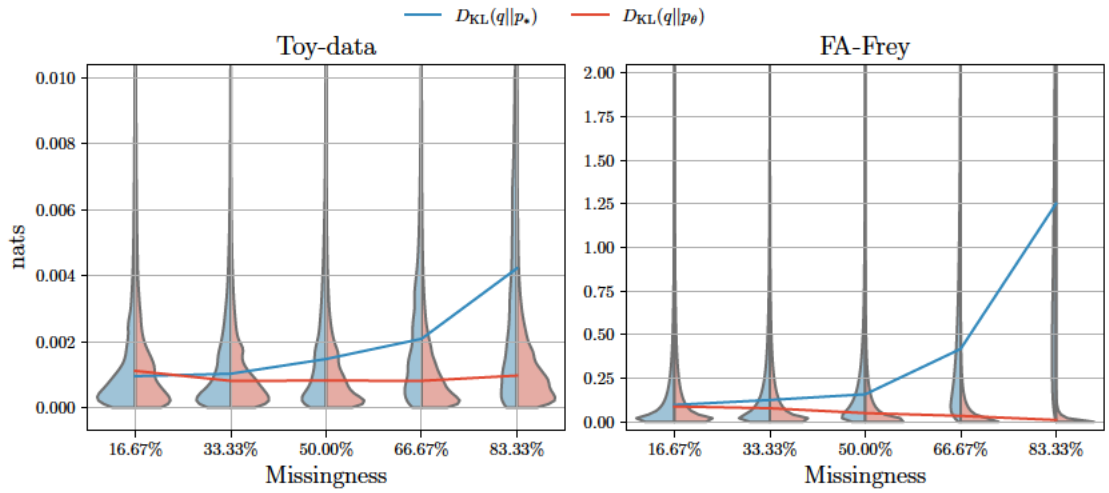


Figure 6: KL divergence, conditioned on the test set (left: toy data, right: FA-Frey data), between the univariate variational conditional distributions and the ground truth distribution $D_{\mathrm{KL}}(q_{\phi_j}(x_j \mid \boldsymbol{x}_{\setminus j}) \,||\, p_*(x_j \mid \boldsymbol{x}_{\setminus j}))$ (blue), and the posterior under the learnt model $D_{\mathrm{KL}}(q_{\phi_j}(x_j \mid \boldsymbol{x}_{\setminus j}) \,||\, p_{\boldsymbol{\theta}}(x_j \mid \boldsymbol{x}_{\setminus j}))$ (red). The lines show the median KL divergence conditioned on the test set. In Appendix M, Figure 17 we show a comparison using Wasserstein distance that displays a similar behaviour.

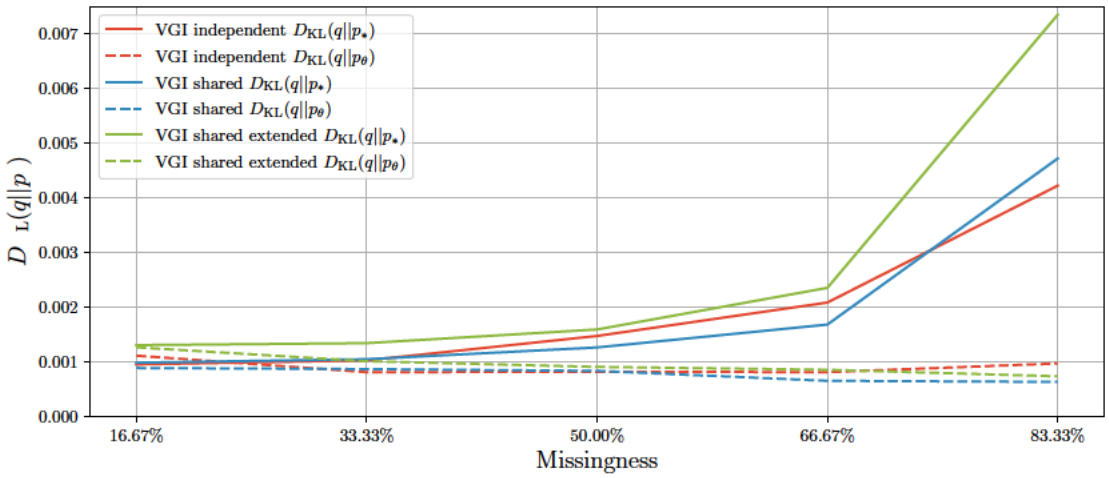


Figure 7: Median KL divergence of the univariate variational conditional distributions, conditioned on the toy test set. Comparing independent against shared-parameter variational models.

Furthermore, we see that the median KL divergence to the conditionals of the fitted model remains constant for all fractions of missingness on the toy data, but it goes down with increasing missingness on the FA-Frey data (solid red lines). We attribute this to the partial weight sharing used in the variational model on FA-Frey data. We verify this by additionally comparing an independent model (red) against a shared, but not extended, model (blue) on the toy data in Figure 7.[20] We can see that with the partially-shared model the KL divergence on toy data also improves with missingness (dashed blue line), whilst it remains almost constant for the independent model (dashed red line). The improved approximation of the conditionals with increasing missingness and partially-shared variational models can be explained by additional amortisation—it is akin to having more data to learn the same number of parameters, which results in a better fit of the variational model.

We further investigated the effect of the extended variational model on the conditional distributions using the toy data (Section 3.3): In Figure 7 we compare a shared variational model with the extended conditionals (green) to models with standard conditionals (red and blue). We observe that the extended conditionals are close approximations of the true Gibbs conditionals, and only slightly worse than the standard variational conditionals. Moreover, the estimate of the model $p_{\boldsymbol{\theta}}$ is only slightly affected for missingness larger than 66% (see Figure 18 in Appendix M). Hence, we conclude that conditioning on an additional variable $\tilde{x}_j$ in the extended-Gibbs variational model (Appendix E) results in inconsequential additional variability ("noise") of the variational distributions. Importantly, in higher dimensional problems, the noise due to conditioning on a single extra variable decreases. There is no significant difference between VGI estimation accuracy with the standard variational conditionals and the extended conditionals when the data dimensionality is larger (see Figure 19 in Appendix M). Hence, when the dimensionality is small and the computational cost is low it is best to use the standard variational conditionals, however, when the dimensionality is large, it is more favourable to use the extended variational model for computational reasons.

5. Experiments on VAE Models

In this section we estimate VAE models from incomplete data. We compare the general-purpose VGI method, VAE-specific methods based on VBGI, and existing VAE-specific methods in the literature in terms of estimation accuracy and computational efficiency.

5.1 Variational Autoencoder Model

The variational autoencoder (VAE) is a non-linear descendant of the factor analysis model. It is a latent-variable model with observables $\boldsymbol{x}$ and latent variables $\boldsymbol{x}_z$, defined via

$$p_{\boldsymbol{\theta}}(\boldsymbol{x} \mid \boldsymbol{x}_z) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_z), \Psi_{\boldsymbol{\theta}}(\boldsymbol{x}_z)), \quad \text{and} \quad p(\boldsymbol{x}_z) = \mathcal{N}(\boldsymbol{x}_z; \mathbf{0}, \boldsymbol{I}),$$

where the $\boldsymbol{\mu}_{\boldsymbol{\theta}}$ and $\Psi_{\boldsymbol{\theta}}$ are deep neural networks with parameters $\boldsymbol{\theta}$, and $\Psi_{\boldsymbol{\theta}}(\boldsymbol{x}_z)$ is diagonal. We use the Gaussian family for the generator, which is the most common case in the VAE literature, although another family of distributions could also be used. The model is

20. We also compare the accuracy of the fitted FA models with both variational models on toy data in Appendix M Figure 18, where we find both approaches comparable in the model $p_{\boldsymbol{\theta}}$ accuracy.

typically optimised using variational inference (Kingma and Welling, 2013; Rezende et al., 2014), where a variational distribution $q_{\boldsymbol{\phi}}(\boldsymbol{x}_z \mid \boldsymbol{x})$ is used to approximate the intractable posterior $p_{\boldsymbol{\theta}}(\boldsymbol{x}_z \mid \boldsymbol{x})$. The parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are then optimised with stochastic gradient ascent by maximising the following lower-bound

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}) \geq \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{x}_z|\boldsymbol{x})}\left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x} \mid \boldsymbol{x}_z)p(\boldsymbol{x}_z)}{q_{\boldsymbol{\phi}}(\boldsymbol{x}_z \mid \boldsymbol{x})}\right]. \tag{14}$$

5.2 Data

We evaluate VGI on a synthetic 560-dimensional VAE-Frey data set, where a VAE model with one hidden layer in the encoder and decoder and a 10-dimensional latent space was first fitted on the Frey data set, and then it was used to synthesise a new data set that we call VAE-Frey. We used the same VAE architecture as Kingma and Welling (2013). The training data set has 2400 data-points and the test data set has 3000 data-points. The rest of the data setup is identical to Section 4.2.

5.3 Experimental Settings

As before, we specify a target VAE model to fit data following a ground truth VAE model with the same architecture, so that the evaluation can focus on the effect of model estimation from incomplete data, rather than on the robustness of the specified model.

We consider three VGI-based methods for fitting the VAE model:

VGI. This method does not use any assumptions of the statistical model $p_{\boldsymbol{\theta}}$ and hence uses the univariate variational conditional formulation of VGI, as before. We specify the variational encoder architecture to be equivalent to the encoder of the ground truth and replace the $\log p_{\boldsymbol{\theta}}(\boldsymbol{x})$ in $\mathcal{J}_{\text{VGI}}$ in (9) with the VAE ELBO from (14).

We use the same univariate variational model with extended conditionals and partially-shared parameters as in the FA-Frey data experiments in Section 4.3 and optimise the variational parameters using AMSGrad (Reddi et al., 2018). The other VGI hyper-parameters are also equivalent to those used in the FA-Frey data experiments: $K = 5$ imputation chains, $G = 5$ Gibbs update steps, and $M = 10$ sampled missing dimensions in $\hat{\mathcal{J}}_{\text{VGI}}$.

VBGI-VAE. Here we use the structure and assumptions of a VAE to adapt the VGI method to the model $p_{\boldsymbol{\theta}}$.

First, VAE is a latent-variable model so we can treat the latent variables as missing and group them together as we discussed in Section 3.4. We then specify a joint variational distribution $q_{\boldsymbol{\phi}_z}(\boldsymbol{x}_z \mid \tilde{\boldsymbol{x}}_{\text{mis}}, \boldsymbol{x}_{\text{obs}})$, which corresponds to the encoder of the standard VAE.

Second, the missing variables $\boldsymbol{x}_{\text{mis}}$ are assumed to be conditionally independent Gaussians given the latents $\boldsymbol{x}_z$, hence we can easily specify the joint distribution of the conditionals $q_{\boldsymbol{\phi}_j}(x_j \mid \boldsymbol{x}_{\text{mis}\setminus j}, \tilde{\boldsymbol{x}}_z, \boldsymbol{x}_{\text{obs}}), j \in \text{idx}(\boldsymbol{m})$ using a shared variational model $q_{\boldsymbol{\phi}_{\text{mis}}}(\boldsymbol{x}_{\text{mis}} \mid \tilde{\boldsymbol{x}}_z, \boldsymbol{x}_{\text{obs}})$ for all patterns of missingness, similar to the target generative model. To parametrise $q_{\boldsymbol{\phi}_{\text{mis}}}$ using a neural network, we pad the $\boldsymbol{x}_{\text{obs}}$ with zeros for the missing dimensions to get a fixed-size vector for all patterns of missingness.

Whilst the VAE model assumes that $\boldsymbol{x}_{\text{mis}}$ and $\boldsymbol{x}_{\text{obs}}$ are independent given $\boldsymbol{x}_z$, in the variational conditionals $q_{\boldsymbol{\phi}_{\text{mis}}}$ we do not simplify the conditioning set but condition on $\boldsymbol{x}_{\text{obs}}$ as

in the general VGI formulation. Conditioning on $\boldsymbol{x}_{\text{obs}}$ compensates for possible information loss about the true $\boldsymbol{x}_z^* \sim p_{\boldsymbol{\theta}}(\boldsymbol{x}_z \mid \boldsymbol{x}_{\text{obs}})$ due to the use of an approximation $q_{\boldsymbol{\phi}_z}$, and hence eases the learning of $q_{\boldsymbol{\phi}_{\text{mis}}}$. We empirically validated this theoretical argument against a simplification of the variational conditional and observed that conditioning $q_{\boldsymbol{\phi}_{\text{mis}}}$ on $\boldsymbol{x}_{\text{obs}}$, as dictated by the general VGI methodology, is indeed crucial to the performance of the method.

Then, the index $j \in \{\text{mis}, z\}$ in the VGI objective refers to either all the missing variables $\boldsymbol{x}_{\text{mis}}$ or the latents $\boldsymbol{x}_z$. Incorporating the VAE assumptions into the VBGI objective for latent-variable models from (13) we get the VAE-specific objective

$$\begin{aligned}\mathcal{J}^t_{\text{VBGI-VAE}}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}_{\text{obs}}) = \mathbb{E}_{f^{t-1}(\tilde{\boldsymbol{x}}_{\text{mis}}, \tilde{\boldsymbol{x}}_z \mid \boldsymbol{x}_{\text{obs}})} \Bigg(\\ \boldsymbol{\pi}(j = \text{mis}) \mathbb{E}_{q_{\boldsymbol{\phi}_{\text{mis}}}(\boldsymbol{x}_{\text{mis}} \mid \tilde{\boldsymbol{x}}_z, \boldsymbol{x}_{\text{obs}})} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}}, \boldsymbol{x}_{\text{obs}} \mid \tilde{\boldsymbol{x}}_z) p(\tilde{\boldsymbol{x}}_z)}{q_{\boldsymbol{\phi}_{\text{mis}}}(\boldsymbol{x}_{\text{mis}} \mid \tilde{\boldsymbol{x}}_z, \boldsymbol{x}_{\text{obs}})} \right] \\ + \boldsymbol{\pi}(j = z) \mathbb{E}_{q_{\boldsymbol{\phi}_z}(\boldsymbol{x}_z \mid \tilde{\boldsymbol{x}}_{\text{mis}}, \boldsymbol{x}_{\text{obs}})} \left[\log \frac{p_{\boldsymbol{\theta}}(\tilde{\boldsymbol{x}}_{\text{mis}}, \boldsymbol{x}_{\text{obs}} \mid \boldsymbol{x}_z) p(\boldsymbol{x}_z)}{q_{\boldsymbol{\phi}_z}(\boldsymbol{x}_z \mid \tilde{\boldsymbol{x}}_{\text{mis}}, \boldsymbol{x}_{\text{obs}})} \right] \Bigg),\end{aligned}$$

where $\boldsymbol{\pi}(j) = \frac{1}{2}$ for all j, meaning that we choose a uniform Gibbs selection probability. We use $K = 5$ imputation chains and $G = 2$ Gibbs updates corresponding to one full update of all unobserved variables $\boldsymbol{x}_{\text{mis}}$ and $\boldsymbol{x}_z$, and optimise the above objective using the Adam optimiser.

VBGI-VAE-M. Alternatively, we can use the conditional independence of the observable variables given the latents for the VAE model to marginalise the missing observable dimensions $\boldsymbol{x}_{\text{mis}}$ from the likelihood, which is commonly done in the VAE literature for incomplete data. Then, $j = \text{mis}$ and the objective above simplifies to

$$\mathcal{J}^t_{\text{VBGI-VAE-M}}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}_{\text{obs}}) = \mathbb{E}_{f^{t-1}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}) q_{\boldsymbol{\phi}_z}(\boldsymbol{x}_z \mid \boldsymbol{x}_{\text{mis}}, \boldsymbol{x}_{\text{obs}})} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}} \mid \boldsymbol{x}_z) p(\boldsymbol{x}_z)}{q_{\boldsymbol{\phi}_z}(\boldsymbol{x}_z \mid \boldsymbol{x}_{\text{mis}}, \boldsymbol{x}_{\text{obs}})} \right] \Bigg).$$

To update the Markov chain distribution f^t we perform block-Gibbs sampling using $q_{\boldsymbol{\phi}_z}(\boldsymbol{x}_z \mid \boldsymbol{x}_{\text{mis}}, \boldsymbol{x}_{\text{obs}})$ for the latents and $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_z)$ for the missing observables. This imputation procedure was initially proposed by Rezende et al. (2014) for missing data imputation on test data, however to the best of our knowledge it has not been considered for learning from incomplete data. We use $K = 5$ imputation chains and $G = 2$ Gibbs updates corresponding to one full update of all unobserved variables $\boldsymbol{x}_{\text{mis}}$ and $\boldsymbol{x}_z$, and optimise the above objective using the Adam optimiser.

In all of the above methods we use the reparametrisation trick (Kingma and Welling, 2013) to compute the gradients of the variational models.

5.4 Comparison Methods

We compare the VGI-based methods against:

VAE (Complete). The ideal case where no data is missing, fitted using standard variational inference for VAEs (Kingma and Welling, 2013; Rezende et al., 2014).

MICE (van Buuren and Oudshoorn, 2000). The same as in Section 4.4.

missForest (Stekhoven and Bühlmann, 2012). An iterative FCS method that uses random forest to model the conditionals. Unlike MICE, the imputations are not sampled but instead are replaced with the deterministic values predicted by the random forest. Hence, while the method can be used on non-linear data, the imputations may be biased by lack of uncertainty representation. We implemented the method in Python using IterativeImputer and RandomForestRegressor for the conditional models from the scikit-learn package (Pedregosa et al., 2011). The default settings for this method proved to be too computationally expensive on this data set, hence to make this method computationally feasible we traded-off the expressivity of the random forest for better computational performance by limiting the number of decision trees to 20 and the maximum depth of the trees to 20.[21] Random forest methods are typically piecewise-constant regressors, which means that by limiting the depth and the number of trees, for reasons above, we may get a too coarse an approximation of the target variable, on the other hand, if the computational budget allows, the trees could be "fully grown" such that each leave corresponds to a unique training data point and hence, in the idealistic scenario of a finely sampled training set, the random forest regressor could be made arbitrarily accurate. Like MICE, missForest was used to produce $K = 5$ imputations and then the VAE model was fitted as described in Appendix D.

MICEForest. One classical way to calibrate the uncertainty of imputations for methods like missForest is predictive mean matching (Little, 1988; van Buuren, 2018, Chapter 3.4). We hence use the MICEForest Python package that provides an implementation of missForest enhanced with predictive mean matching.[22] The uncertainty is controlled via a hyperparameter that specifies the number of nearest neighbours (based on the predictive mean) considered for a random imputation draw. As with missForest above, to make this method feasible on the high-dimensional data in this section, the number of estimators was set to 20, the maximum depth of the decision trees was set to 20, and the number of nearest neighbours was set to the default 5.[23]

MVAE (Nazábal et al., 2020; Mattei and Frellsen, 2019). The missing dimensions are marginalised in the generator of the VAE and the encoder masks the missing dimensions with zeros.

HIVAE (Nazábal et al., 2020). The missing dimensions are marginalised in the generator of the VAE. The model uses a hierarchical prior and zero masking in the encoder for missing data.

PartialVAE+ (Ma et al., 2019). The missing dimensions are marginalised in the generator of the VAE. Uses a permutation-invariant encoder network with learnable location embeddings, which can handle partially-observed inputs.

21. One imputation of the data set with missForest took about 27 hours on CPU with the scikit-learn implementation.
22. MICEForest implementation can be found here `https://github.com/AnotherSamWilson/miceforest`.
23. One imputation of the data set with MICEForest took about 17 hours on CPU.

MIWAE (Mattei and Frellsen, 2019). The missing dimensions are marginalised in the generator of the VAE. Uses a tighter importance weighted lower-bound and zero masking in the encoder.

The above methods are all fitted using the Adam optimiser.

5.5 Accuracy of the Fitted VAE Models

As is common in machine learning the accuracy of the VAE models increased with training iterations and then decreased due to over-fitting, hence we used model selection on an incomplete held-out validation data set to mitigate this effect (for more details see Appendix L and Figure 20). We then use the selected model parameters $\hat{\boldsymbol{\theta}}$ to assess the accuracy of the fitted model.

In order to evaluate the accuracy of the selected VAE models, we estimated the log-likelihood on complete test data. In general, computing the marginal log-likelihood of a VAE is intractable, therefore we estimate it using the IWAE bound (Burda et al., 2015) that uses samples from a variational proposal distribution q and self-importance weighting,

$$\mathcal{L}_{\text{IWAE}}(\boldsymbol{x}) = \log\left(\frac{1}{L}\sum_{l=1}^{L}\frac{p_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x},\boldsymbol{x}_z^l)}{q(\boldsymbol{x}_z^l \mid \boldsymbol{x})}\right), \quad \text{where} \quad \boldsymbol{x}_z^1,\ldots,\boldsymbol{x}_z^L \sim q(\boldsymbol{x}_z \mid \boldsymbol{x}).$$

The IWAE bound approaches $\log p_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x})$ monotonically from below as the number of importance samples L increases (Burda et al., 2015, Theorem 1) irrespective of the proposal distribution (subject to minor conditions). However, we note that the non-asymptotic bias of the estimator is in the order of $\mathcal{O}(1/L)$ (Owen, 2013; Paananen et al., 2021) and depends on the accuracy of the proposal distribution $q(\boldsymbol{x}_z \mid \boldsymbol{x})$, and is zero if $q(\boldsymbol{x}_z \mid \boldsymbol{x}) = p_{\boldsymbol{\theta}}(\boldsymbol{x}_z \mid \boldsymbol{x})$. Hence, to attain a good proposal and mitigate the bias we fine-tune the variational encoder $q_{\boldsymbol{\phi}}$ from the training stage on the complete test data (Mattei and Frellsen, 2018).[24] We then estimate the log-likelihood using $\mathcal{L}_{\text{IWAE}}$ and the fine-tuned encoder with a large number $L = 50000$ of importance samples.

In Figure 8 we show the estimated marginal log-likelihood averaged over the test data. We see that the VGI-based methods perform consistently better than the existing VAE-specific methods, particularly in settings with missingness fraction greater than 50%. It is interesting that VGI and VBGI-VAE, which learn a variational imputation model of the missing observable variables, both outperform VBGI-VAE-M, which marginalises the missing variables from the likelihood. This suggests that forcing the VAE model to reconstruct a completed data-point is overall better for the accuracy of the generator than marginalising the missing variable dimensions in the generator as it is often done in the VAE-specific methods and VGI-VAE-M. We conjecture that marginalising the missing variable dimensions allows the latent representation in VAEs to systematically forget (collapse) some of the information presented in the inputs. In contrast, forcing the VAE model to reconstruct a completed data-point provides an adaptive regularisation that prevents such loss of information.

24. We have also considered fitting a randomly-initialised encoder on the complete test data but due to poor initialisation the fitted variational approximations were worse (also observed in previous works e.g. Altosaar et al., 2017).

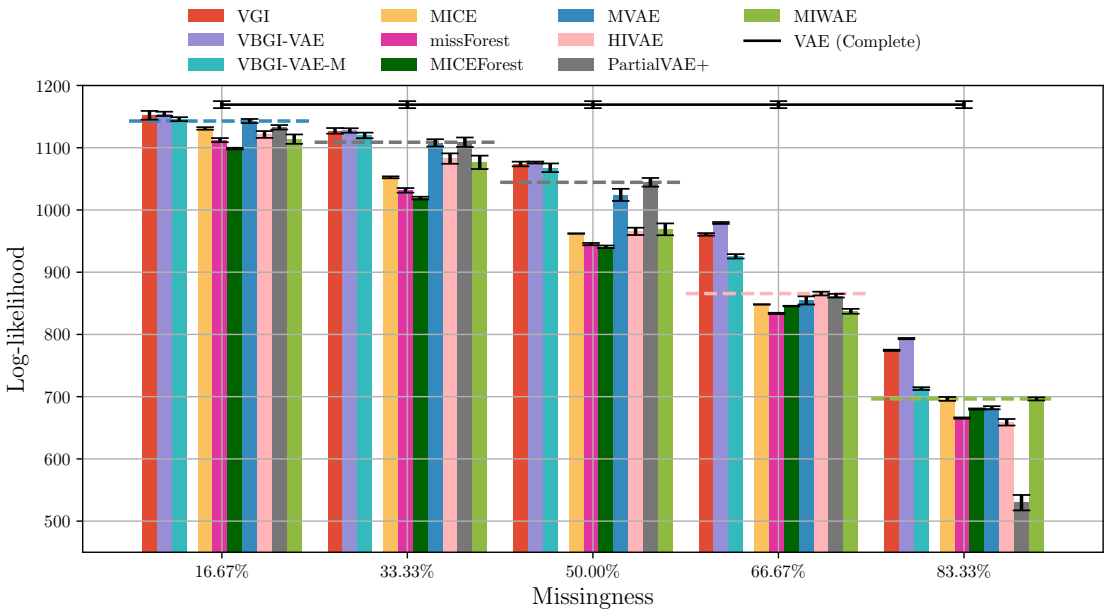


Figure 8: Importance weighted estimates of the marginal log-likelihood on complete test data. The error bars show standard error on the mean of five experiment repetitions with different missingness masks and model initialisation. The dashed horizontal lines show the best performing competitor model, which we note is always outperformed by the VGI-based methods.

The learning curves for the different methods are shown in Figure 21 in Appendix M. We note that for missingness fractions greater than 50%, the methods that marginalise the missing dimensions from the likelihood (VAE-specific and VBGI-VAE-M) start to over-fit on the training data (compare solid and dash-dotted curves) early, whereas the VGI and VBGI-VAE monotonically improve throughout the training, which allows them to surpass the other methods in terms of the model accuracy, especially at greater than 50% missingness, which we also observe in Figure 8.

Moreover, we show in Figure 22 in Appendix M that working with the extended variational model in VGI had no adverse effect on the estimation accuracy while greatly reducing the computational cost. This adds additional evidence to the results in Section 4.7.

Regarding the impute-then-fit methods, we notice that MICE performs well despite being a linear method. This can be explained as follows: the generator is only slightly non-linear since it is parametrised by a one-hidden layer neural network and the data is based on the Frey data set, which can be fitted rather well by linear methods, such as factor analysis. Figure 8 shows that the non-linear imputation methods, missForest or MICEForest, did not perform well in this experiment. We attribute the poor performance of these methods to two factors: modelling limitations due to computational cost restrictions, see Section 5.4, and, possibly, an incorrect representation of imputation uncertainty. If the computational budget

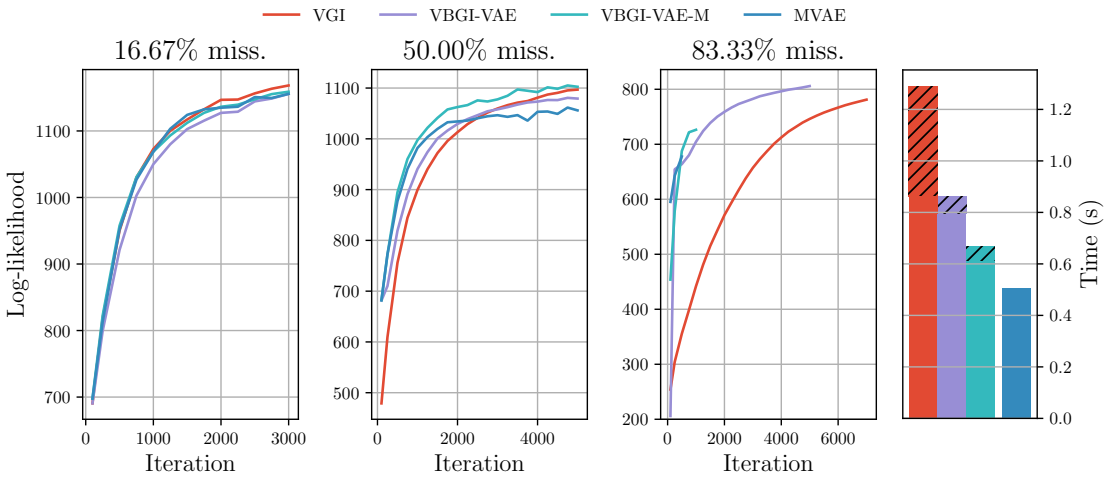


Figure 9: Left: Estimated test log-likelihood on complete data against training iteration. Right: Average duration of one training iteration in seconds. The hatched part of the bars indicates the time spent on updating the imputations via pseudo-Gibbs sampling. Plots including other fractions of missingness and the other VAE-specific methods can be found in the Appendix M Figures 23 and 24.

allows, the former issue can be addressed by using more or larger decision trees although preventing overfitting will then become important. The second issue could be addressed in MICEForest by tuning the number of nearest neighbours in predictive mean matching, although current recommendations are more based on heuristics than firm principles (e.g. van Buuren, 2018, Chapter 3.4), which can make finding an optimal setting difficult.

Finally, we observe that MVAE performs almost always better than MIWAE, even though MIWAE optimises a tighter importance-weighted lower-bound. This suggests that when the statistical model $p_{\boldsymbol{\theta}}$ matches the true generative model but, due to missingness, we do not have enough data to actually fit the model, MIWAE might be more prone to overfitting since it can tighten the bound more effectively. To confirm this intuition, we fitted an under-specified model to the incomplete data (by reducing the latent dimension to 5), and in line with our explanation above, MIWAE was then performing better than MVAE. Moreover, PartialVAE+ performed considerably well on missingness levels up to 83.33%, however the performance went down significantly at the largest missingness which points to a difficulty of approximating the variational distribution with a permutation-invariant encoder at very large missingness.

5.6 Computational Aspects

We here discuss the computational aspects of VGI-based methods for VAE estimation, and contrast it with the VAE-specific methods.

In the left-hand side of Figure 9 we plot the estimated test log-likelihood against the training iteration and compare the VGI-based methods to MVAE. We see that out of all VGI-based methods VBGI-VAE-M has the highest rate of likelihood improvement, which is comparable to the rate of MVAE, while achieving consistently better accuracy of the fit. However, the methods that marginalise the missing variables from the likelihood (MVAE and VBGI-VAE-M) quickly over-fit to the training data in the higher missingness settings and hence perform worse than VGI and VBGI-VAE. VBGI-VAE closely follows VBGI-VAE-M at a slightly slower rate of improvement, due to having to learn an additional variational model of the missing observables, but it offers a generally better accuracy of the fit, as seen in the previous section, and does not over-fit to the training data like MVAE and VBGI-VAE-M. The general-purpose VGI method, which does not use the assumptions available in a VAE, has the slowest rate of improvement, which is most clearly seen in high missingness settings. However, as VBGI-VAE, VGI generally outperforms the other methods that marginalise the missing variables from the likelihood in terms of the accuracy of the fit.

The average duration of a training iteration is shown in the right-hand side of Figure 9. We used the same VGI hyper-parameters for all levels of missingness in our experiments, which means that the average per-iteration cost is about the same. While the cost of an iteration for VAE-specific methods does not depend on the missingness, for VGI-based methods, the iteration cost could be adjusted to the level of missingness by, for example, adapting the number of Gibbs update steps (G in Algorithm 5), the number of imputation chains K, or the number of sampled missing dimensions (M in (11)). The figure thus shows results that were not optimised for computational efficiency. The average duration of an iteration of the VGI-based methods are about 2.4 (VGI), 1.6 (VBGI-VAE), and 1.25 (VBGI-VAE-M) times longer than that of the other methods. Moreover, about 0.6 (VGI), 0.2 (VBGI-VAE), and 0.45 (VBGI-VAE-M) of the excess iteration cost over the other methods is due to the Gibbs sampling of imputations, denoted by the hatched bars in the figure. We attribute the rest of the excess cost to the fitting of the variational kernel and evaluating the generative model on multiple imputed samples. We also note that for the experiments in this section we used a shared variational model with extended variational conditionals as discussed in Section 3.3 which resulted in about 2.8 times lower per-iteration cost than VGI with the standard-Gibbs conditionals (see Appendix M Figure 22).

In summary, VGI-based methods offer better estimation accuracy over the other methods at (i) a slightly higher per-iteration computational cost and (ii) a potentially slower rate of likelihood improvement in the high-missingness settings.

6. Experiments on Normalising Flows

In this section we estimate normalising flow models from incomplete UCI data using the general-purpose VGI and a flow-specific Monte Carlo EM method. We compare them in terms of model accuracy and computational efficiency.

6.1 Normalising Flow Model

A normalising flow is a probabilistic model that models complex distributions $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ via a simple base distribution $p(\boldsymbol{u})$ and an invertible deterministic transformation $T_{\boldsymbol{\theta}} : \boldsymbol{u} \mapsto \boldsymbol{x} =$

$T_{\boldsymbol{\theta}}(\boldsymbol{u})$ (Rezende and Mohamed, 2015). The density of $\boldsymbol{x}$ can be obtained from the change of variables formula (e.g. Murphy, 2021) as

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = p(\boldsymbol{u}) |\det J_{T_{\boldsymbol{\theta}}}(\boldsymbol{u})|^{-1}, \quad \text{where} \quad \boldsymbol{u} = T_{\boldsymbol{\theta}}^{-1}(\boldsymbol{x}),$$

and $J_{T_{\boldsymbol{\theta}}}$ is the Jacobian matrix of the transformation. In practice, the transformation $T_{\boldsymbol{\theta}}$ is often composed of multiple chained transformations $T_{\boldsymbol{\theta}} = T_L \circ \cdots \circ T_1$, where we suppress the dependency of the decomposed transformations T_l on $\boldsymbol{\theta}$ for notational simplicity. The parameters of each transformation are given by a neural network, thus enabling complex overall transformations $T_{\boldsymbol{\theta}}$. Moreover, both transformation $T_{\boldsymbol{\theta}}$ and $T_{\boldsymbol{\theta}}^{-1}$ must be differentiable, which allows us to fit the parameters $\boldsymbol{\theta}$ using stochastic gradient ascent.

The expressivity of the flow model greatly depends on the choice of the transformations T_l, hence to show that our method can fit expressive models, we specify the statistical model $p_{\boldsymbol{\theta}}$ as a rational-quadratic neural spline flow (RQ-NSF, Durkan et al., 2019), which has been shown to be able to model complex distributions efficiently. In RQ-NSF, each transformation T_l is composed of an invertible linear transformation and a monotonic rational-quadratic spline transformation. The spline transformation maps a fixed interval $[-B, B]$ to $[-B, B]$ using a spline function and an identity mapping outside of this range. The rational-quadratic splines are monotonic piece-wise functions with P sections (*bins*) that are characterised by the boundary coordinates (*knots*) and the (positive) derivatives at the knots, where both the knot coordinates and the derivative values are parametrised by a residual neural network.

6.2 Data

We evaluate VGI on a selection of tabular data sets from the UCI machine-learning repository (Dua and Graff, 2017), which are commonly used to evaluate normalising flow models, and follow the pre-processing in (Papamakarios et al., 2017).

POWER. Measurements of electric power consumption in one household with a one-minute sampling rate collected over a period of 47 months. Contains 6 dimensions and ~2M samples.

GAS. A collection of gas sensor measurements in several gas mixtures. Contains 8 dimensions and ~1M samples.

HEPMASS. A collection of measurements from high-energy physics experiments to detect a new particle of unknown mass. Contains 21 dimensions and ~525K samples.

MINIBOONE. Data taken from a MiniBooNE experiment used to distinguish electron neutrinos from background noise. Contains 43 dimensions and ~36K samples.

The missingness is rendered uniformly at random (MCAR) at three levels of missingness from 16.6% to 83.3%. The rest of the data setup is the same as in the previous sections.

6.3 Experimental Settings

We implement the rational-quadratic neural spline flow (RQ-NSF) with coupling following Durkan et al. (2019, Appendix B.1).

In order to match the expressiveness of the target model $p_{\boldsymbol{\theta}}$ we parametrise the univariate conditional distributions using flow-like element-wise distributions. We use the standard normal distribution as the base distribution and R sets of linear and rational-quadratic spline transformations, where the parameters of the spline transformations (knot coordinates and derivatives) are given by a partially-shared residual network. The transformations are defined over the same interval as the target flow model, and we use 4 bins and $R = 3$ element-wise transformations.

As before, we use the extended conditionals with a partially-shared model which consists of a shared network and independent per-conditional networks. The shared network takes a completed data-point and computes a shared representation. This representation is then used to compute the element-wise transformation parameters using independent networks for each of the $R{*}D$ transformations. The shared network consists of one residual block with 256 hidden features, and outputs a 128-dimensional shared representation. Each element-wise transformation network consists of 2 residual blocks with 32 hidden features.

Like in the experiments in the previous sections, we optimise the variational parameters $\boldsymbol{\phi}$ using AMSGrad (Reddi et al., 2018). The other VGI hyper-parameters are as follows: $K = 5/10/20$ imputation chains for corresponding missingness of 16.6%/50%/83.3%, $G = 5$ Gibbs update steps, and $M = 1$ sampled missing dimensions in (11).

6.4 Comparison Methods

We compare VGI against Monte Carlo expectation-maximisation (MCEM, Wei and Tanner, 1990) with the flow-specific MCMC sampler of Cannella et al. (2021). MCEM maximises the observed data log-likelihood by iteratively maximising the ELBO

$$\begin{aligned}
\hat{\boldsymbol{\theta}}^t &= \arg\max_{\boldsymbol{\theta}} \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{p_{\hat{\boldsymbol{\theta}}^{t-1}}(\boldsymbol{x}_{\text{mis}}|\boldsymbol{x}^i_{\text{obs}})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}}, \boldsymbol{x}^i_{\text{obs}}) - \log p_{\hat{\boldsymbol{\theta}}^{t-1}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}^i_{\text{obs}})\right] \\
&= \arg\max_{\boldsymbol{\theta}} \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{p_{\hat{\boldsymbol{\theta}}^{t-1}}(\boldsymbol{x}_{\text{mis}}|\boldsymbol{x}^i_{\text{obs}})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}}, \boldsymbol{x}^i_{\text{obs}})\right] \\
&\approx \arg\max_{\boldsymbol{\theta}} \frac{1}{N}\sum_{i=1}^{N} \frac{1}{K}\sum_{k=1}^{K} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i,k)}_{\text{mis}}, \boldsymbol{x}^i_{\text{obs}}), \quad \text{where} \quad \boldsymbol{x}^{(i,k)}_{\text{mis}} \sim p_{\hat{\boldsymbol{\theta}}^{t-1}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}^i_{\text{obs}}).
\end{aligned}$$

We obtain the samples from the conditional distributions of missing data $p_{\hat{\boldsymbol{\theta}}^{t-1}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}^i_{\text{obs}})$ using projected latent Markov chain Monte Carlo (PLMCMC, Cannella et al., 2021). Rather than generating proposals in the data space, PLMCMC generates Metropolis-Hastings proposals (e.g. Bishop, 2006, Chapter 11.2.2) in the better-behaved space of the base distribution of a normalising flow using simple proposal distributions, and hence provides a more efficient way to sample from any conditional distribution under a joint normalising flow model.

The use of PLMCMC with MCEM (in the next subsections denoted as PLMCMC for simplicity) for the estimation of normalising flows from incomplete data is not new to this paper and has been demonstrated by Cannella et al. (2021), and in our experiments we follow their experimental configuration. During the first set of iterations the imputations are sampled from a standard Gaussian distribution, and afterwards the conditional imputations

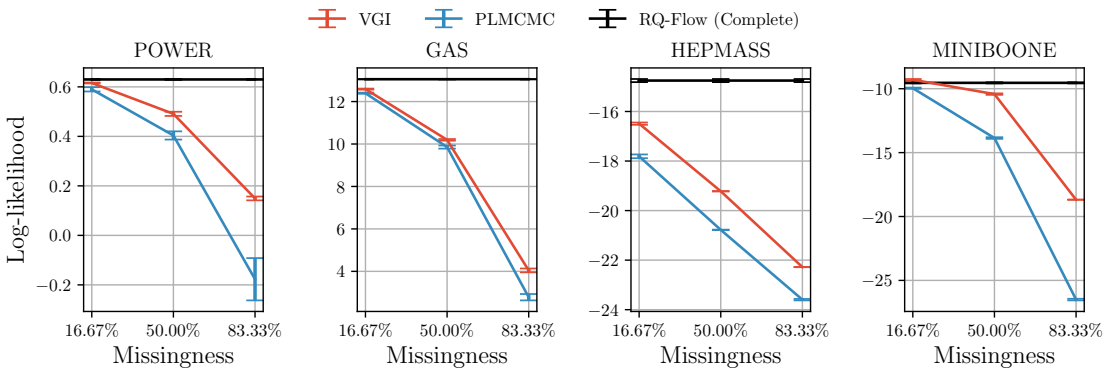


Figure 10: Log-likelihood on complete test UCI data. The error bars show standard error on the mean of five experiment repetitions with different missingness masks and model initialisation. VGI consistently yields a more accurate model $p_{\boldsymbol{\theta}}$ than MCEM with the PLMCMC sampler.

are re-sampled using PLMCMC at regular intervals using chains of increasing length (from 200 to 1000 steps), except for the MINIBOONE data for which we could afford running longer chains of up to 3000 steps due to its relatively small size.

After a limited hyper-parameter search we have found that the same proposal distribution as used in the model estimation experiments by Cannella et al. (2021) also performed best in our experiments. The proposal distribution used is a uniform mixture of a conditional normal distribution with mean as the previous state and a standard deviation of 0.01, and an unconditional normal distribution with mean 0 and standard deviation of 1.0.

We observed that the PLMCMC sampler can experience numerical instability due to the inversion of the flow on the GAS and POWER data sets. To stabilise the sampler on the GAS data set we clipped the accepted MCMC proposals to the observed data hypercube, defined by the minimum and maximum observed values for each dimension. A similar fix has been used in some experiments in the original work by Cannella et al. (2021). However, this fix did not help on the POWER data set, and the results in the following sections partly reflect the instability of the sampler.

6.5 Accuracy of the Fitted Normalising Flow Models

We did not observe any over-fitting in the flow model experiments, hence the model parameters $\hat{\boldsymbol{\theta}}$ from the final training iteration were used in the following evaluation for both methods.

We evaluate the accuracy of the fitted models using the log-likelihood on complete test data, shown in Figure 10. Uniformly across the data sets, VGI produces a more accurate fit than MCEM with the PLMCMC sampler. In principle, MCEM should be able to produce model fits that are equally good or better than any variational method due to sampling from the true conditional distributions instead of using an approximation. However, in

practice the performance of MCEM is limited by the performance of the MCMC sampler, which, due to computational constraints, often does not generate exact samples from the true conditional distribution. Even when using PLMCMC as the sampler, which takes advantage of the characteristics of a normalising flow to improve the performance of the sampler, the adverse effect of using finite-length Markov chains is significant. In an attempt to mitigate this pitfall of MCEM, we have empirically investigated the use of persistent imputations chains, similar to VGI, but found that this drastically reduced the acceptance rate of the proposed transitions in the later training epochs and therefore also impaired the accuracy of the estimated statistical model.

In this evaluation we have trained the models for a fixed number of iterations with both methods—we investigate the accuracy-compute trade-off in the following subsection. We thus note that the results could be further improved by running the methods for more iterations (see Appendix M Figure 25 for possible gains when using larger compute budgets), or via other means in stochastic optimisation.

6.6 Computational Aspects

To investigate the computational performance, we plot the test log-likelihood against the training iteration in Figure 11 and further show the average duration of a training iteration for each method. First, we note that the log-likelihood curves of VGI (red) are consistently ahead of PLMCMC (blue). Hence, even without taking the iteration cost into account, VGI improves the statistical model $p_{\boldsymbol{\theta}}$ significantly faster than PLMCMC. Moreover, the average per-iteration cost of VGI is also smaller than PLMCMC as shown in the bar plots, making VGI overall more efficient than PLMCMC.

The figure further shows that PLMCMC spends a large fraction of its per-iteration cost on performing MCMC (blue hatched bars). Since the PLMCMC sampler requires the inverse transformation to be efficiently computable, the experiments up to this point used a flow model with coupling layers whose inverse can be computed at a similar cost as the forward transformations. However, many normalising flow models cannot be inverted efficiently. One example of such a model class are the autoregressive flows, where the forward transformation can be computed in constant time with the dimensionality of the model, but the inverse requires d sequential transformations (e.g. Papamakarios et al., 2021, Section 3.1). In Figure 12, we compare the training iteration cost in seconds on a coupling-based (darker colour) and autoregressive (lighter colour) flow models on the MINIBOONE data.[25] It can be seen that the cost of PLMCMC on the autoregressive model is significantly larger than for the coupling-based model, which prohibits its application to larger data sets, whilst VGI handles such models much more gracefully.

The results in this section contrast with Section 5.6, where the VGI-based methods were computationally more expensive than the other VAE-specific estimation methods: here we have found that the general-purpose VGI method is more efficient than the existing method for flow model estimation, whilst also producing a more accurate fit.

25. For the autoregressive flow we use the autoregressive rational-quadratic neural spline flow model with the setup of Durkan et al. (2019, Appendix B.1).

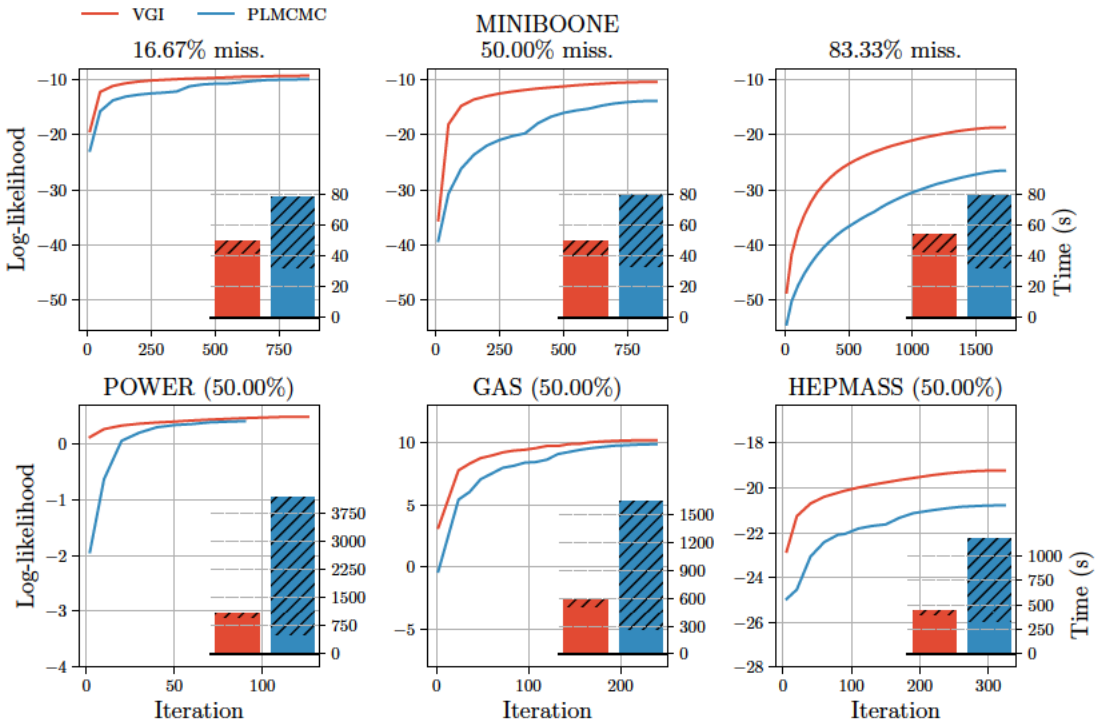


Figure 11: Estimated test log-likelihood on complete data against training iteration and average duration of one training iteration in seconds. The hatched part of the bars indicates the time spent on pseudo-Gibbs sampling in VGI and PLMCMC during MCEM. Top row: results on the MINIBOONE data set for all fractions of missingness. Bottom row: results on the other data sets (POWER, GAS, HEPMASS) for 50% missingness (for the other fractions, see Appendix M Figure 26).

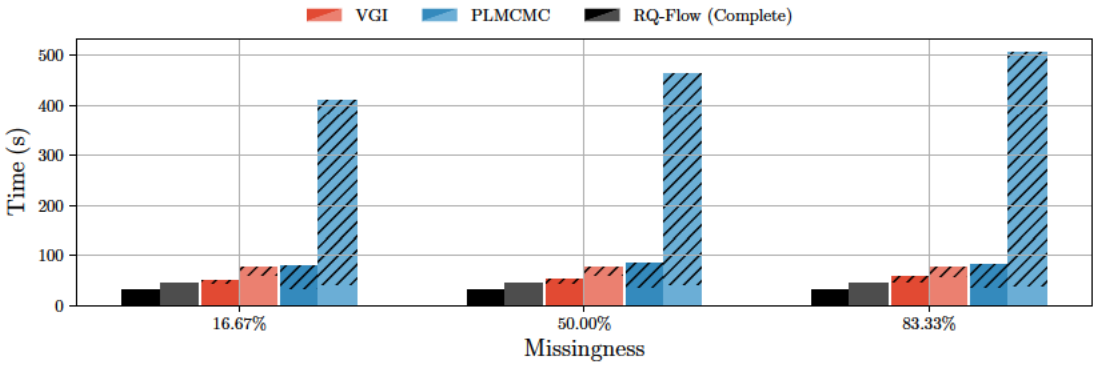


Figure 12: The average duration of a training iteration on the MINIBOONE data set when using coupling-based (darker) and autoregressive (lighter) flow models. VGI easily handles autoregressive models without a significant overhead while PLMCMC scales poorly to such models due to a computationally inefficient inverse.

7. Discussion

In this work we have presented a novel approach to (approximate) maximum-likelihood estimation of statistical models from incomplete data. Our method is applicable to general statistical models and is based on variational inference and Gibbs sampling, both of which have been extensively researched. Hence, the proposed method, termed variational Gibbs inference (VGI), can adopt the diversity of techniques developed for them. We have further introduced a version of VGI for problems where missingness is known to occur in blocks (Section 3.4), for example, in latent-variable models with incomplete data (as demonstrated on a VAE in Section 5).

VGI addresses limitations of (amortised) variational inference (VI) that has strongly limited its applicability to model estimation from incomplete data (see Sections 2.1-2.2): (i) VGI only requires d amortised variational conditionals and thus scales linearly in the dimensionality of the data, gracefully handling the exponential explosion in the number of possible missingness patterns and subsequently the number of conditional variational distributions, and (ii) it enables the specification of a flexible variational family for the conditional distribution of the missing variables using general probabilistic models without restrictions. The parameters of the statistical model $p_{\boldsymbol{\theta}}$ and the variational conditionals $q_{\boldsymbol{\phi}_j}$ are fitted using stochastic gradient ascent, hence it can be used in large-scale data settings that are common in machine learning. The proposed method thus enables efficient learning by drawing on the strengths of amortised variational inference, thanks to property (i), and allows us to well-optimise the ELBO and obtain a good estimate of the target statistical model, thanks to property (ii).

We have validated our method on a toy factor analysis (FA) model against the EM algorithm, which corresponds to the best solution that can be achieved by any variational method, and against a multiple-imputation approach using a popular MICE imputation method in order to evaluate against the two-stage impute-then-fit approaches to statistical model estimation from incomplete data. Our method achieved comparable performance to the classical EM algorithm in a low-dimensional setting and performed similarly to the Monte Carlo EM in the higher-dimensional setting. This suggests that given sufficient compute resources, VGI can achieve good estimates of the model. Moreover, VGI outperformed the impute-then-fit approaches in a higher-dimensional setting even when the (MICE) imputer used the correct family of imputation models relative to the target (FA) model. This suggests that optimising a single variational lower-bound might be a more efficient approach to estimating statistical models from incomplete data than the two-stage impute-then-fit approaches.

We have demonstrated that VGI is particularly well-suited for modern models by successfully fitting important modern statistical models in machine learning, namely VAEs and normalising flows in Sections 5 and 6, respectively.[26] In our experiments, VGI achieved better performance than the existing state-of-the-art model-specific estimation methods.

7.1 Future Work

We see a number of interesting future directions following this work.

26. The evaluated models are for continuous random variables, and we focused on them due to their popularity, but the theory presented in Section 3 is general and also holds for discrete distributions.

Due to the generality of the method, it is a candidate to be implemented in probabilistic modelling platforms as a general-purpose inference engine for incomplete data, or as a library in machine learning frameworks. A readily available framework may help improve the current practice of handling missing data, where simple fixes are often preferred over the principled probabilistic methods like VGI.

We also believe that there is area for improvement in the computational efficiency of the method. It is widely known that the mixing of the Markov chain may be inefficient in Gibbs sampling if the variables are highly correlated, which may slow down the rate of improvement in VGI. However, we can use the rich literature on Gibbs sampling to improve VGI: To improve the mixing rate of the Gibbs imputation chains one could use ordered over-relaxation (Neal, 1995) or apply adaptive-scan Gibbs methods to find Gibbs selection probabilities $\pi(j)$ that are better than the uniform ones that we used in this paper (Łatuszyński et al., 2013; Chimisov et al., 2018; Grathwohl et al., 2021), thereby increasing the overall performance of VGI. Alternatively, one could use the (near-)conditional-independencies of the variational conditionals, if such exist, to update some of the dimensions simultaneously, thus breaking the sequential Gibbs update requirement but potentially increasing the mixing rate of the imputation chains (Angelino et al., 2016; Terenin et al., 2020). However, the effect of these changes on the performance of VGI is yet unknown.

Another area of potential improvement is the computational efficiency of evaluating and sampling the variational model. In this work we have implemented the variational conditionals moderately efficiently by using the extended-Gibbs kernel with partially-shared parameters and then computing all conditional distributions simultaneously using optimised matrix operations. However, we usually do not require all conditional distributions for each data-point; only a subset of sampled missing dimensions in the VGI objective in (11) is required. We believe this could be addressed by leveraging the efficient sparse matrix computation libraries built for graph and compressed neural network processing (Bell and Garland, 2008; Han et al., 2016; Huillcen-Baca and Palomino-Valdivia, 2019).

Finally, while our investigation considered statistical model estimation under the ignorable missingness assumption, which is a standard assumption in the missing data field, in practice, the missingness mechanisms are often non-ignorable, and making the ignorability assumption can result in incorrect inferences. Some recent work on VAEs has shown promising results for non-ignorable missingness (Ipsen et al., 2020; Collier et al., 2021; Ma and Zhang, 2021), however no such work has yet been done on normalising flows. Considering the importance of non-ignorable missingness in practice, extending VGI to statistical model estimation from incomplete data that is subject to non-ignorable missingness is an important future direction towards a general-purpose estimation engine for incomplete data.

Acknowledgments

Vaidotas Simkus gratefully acknowledges the support from the University of Edinburgh and the financial support from Huawei through their grant for PhD Studentships in Dialogue Systems and Data Systems. Benjamin Rhodes was supported in part by the EPSRC Centre for Doctoral Training in Data Science, funded by the UK Engineering and Physical Sciences Research Council (grant EP/L016427/1) and the University of Edinburgh. We would also

like to thank the anonymous reviewers for their constructive feedback and valuable suggestions that have helped us improve the paper.

Appendix A. Ignorable Missingness Assumption

In this section we summarise the three classes of missingness as proposed by Rubin (1976) and their effects on statistical model estimation. An important aspect of the incomplete data model in (1) is that the missingness mask $\boldsymbol{m}$ depends on both observed $\boldsymbol{x}_{\text{obs}}$ and missing $\boldsymbol{x}_{\text{mis}}$ values of the data. This generally means that the estimation of the statistical model $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ is coupled with the missingness model $p(\boldsymbol{m} \mid \boldsymbol{x})$. Rubin (1976) proposed a taxonomy of missing data mechanisms that consists of three levels of simplifying assumptions:

1. Missing not at random (MNAR): no independence assumptions,
2. Missing at random (MAR): $p(\boldsymbol{m} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}) = p(\boldsymbol{m} \mid \boldsymbol{x}_{\text{obs}})$,
3. Missing completely at random (MCAR): $p(\boldsymbol{m} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}) = p(\boldsymbol{m})$.

MAR and MCAR assumptions have been widely used in many incomplete data methods since they decouple the model $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ estimation task from the missingness model $p(\boldsymbol{m} \mid \boldsymbol{x})$

$$\begin{aligned}
\hat{\boldsymbol{\theta}} &= \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{N} \int p_{\boldsymbol{\theta}}(\boldsymbol{x}^i_{\text{obs}}, \boldsymbol{x}^i_{\text{mis}}) p(\boldsymbol{m}^i \mid \boldsymbol{x}^i_{\text{obs}}) \, \mathrm{d}\boldsymbol{x}^i_{\text{mis}} \\
&= \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{N} p(\boldsymbol{m}^i \mid \boldsymbol{x}^i_{\text{obs}}) \prod_{i=1}^{N} \int p_{\boldsymbol{\theta}}(\boldsymbol{x}^i_{\text{obs}}, \boldsymbol{x}^i_{\text{mis}}) \, \mathrm{d}\boldsymbol{x}^i_{\text{mis}} \\
&= \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{N} \int p_{\boldsymbol{\theta}}(\boldsymbol{x}^i_{\text{obs}}, \boldsymbol{x}^i_{\text{mis}}) \, \mathrm{d}\boldsymbol{x}^i_{\text{mis}},
\end{aligned}$$

where the product is due to computing the likelihood for N i.i.d. observed data-points. These assumptions are also known as *ignorable missingness mechanism* because the missingness model can be simply ignored when the only goal is to estimate the model $p_{\boldsymbol{\theta}}(\boldsymbol{x})$.[27]

Appendix B. Gibbs Sampling

Gibbs sampling (Geman and Geman, 1984) is a Markov chain Monte Carlo (MCMC, e.g. Barber, 2017, Chapter 27.3) method for (approximately) sampling a joint model $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ where directly sampling the joint distribution is intractable. The method factorises the distribution using the chain rule $p_{\boldsymbol{\theta}}(\boldsymbol{x}) = p_{\boldsymbol{\theta}}(x_j \mid \boldsymbol{x}_{\setminus j}) p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\setminus j})$, and assumes that the conditional distribution $p_{\boldsymbol{\theta}}(x_j \mid \boldsymbol{x}_{\setminus j})$ is easy to sample from for any index j. Then, the sampler starts with any random initial sample $\boldsymbol{x}^0$, selects one variable index j to sample, and updates the sample $\boldsymbol{x}^{t+1} = \{x_j^{t+1}, \boldsymbol{x}^t_{\setminus j}\}$ with $x_j^{t+1} \sim p_{\boldsymbol{\theta}}(x_j \mid \boldsymbol{x}^t_{\setminus j})$. By continuing this procedure, the Gibbs sampler asymptotically in the number of iterations t samples the joint distribution of the model $p_{\boldsymbol{\theta}}(\boldsymbol{x})$.

27. Assuming the model $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ is flexible enough, otherwise the estimate can be biased even with MAR misingness (Marlin, 2008, Section 3.4).

The order in which each dimension is updated by the Gibbs sampler is called the "scan". The two most common scans are: random, where the dimension index j at every iteration t is chosen uniformly at random, and systematic, in which a fixed order of the variables is selected and the sampler repeatedly iterates over the variables in that order. Our VGI method uses the randomised scan as presented in Section 3. However, we feel that it is important to note that neither of these scans are necessarily optimal, and as highlighted in the discussion there are ways to adapt the scan of the Gibbs sampler (Łatuszyński et al., 2013; Chimisov et al., 2018; Grathwohl et al., 2021).

Gibbs sampling methods have been crucial for sampling and learning some important classes of models in physics, such as the Ising model (e.g. Barber, 2017, Section 4.2.5). However, the required conditionals $p_{\boldsymbol{\theta}}(x_j \mid \boldsymbol{x}_{\setminus j})$ are typically intractable for most modern statistical models, such as VAEs and normalising flows, and hence, efficient sampling is not tractable either. In VGI, we approximate the Gibbs conditionals with an amortised variational model, which allows us to overcome this issue.

Appendix C. Multiple Imputation by Chained Equations (MICE)

The MICE algorithm starts with random imputations of missing data and then alternates between fitting the univariate conditional distributions $f(x_j \mid \boldsymbol{x}_{\setminus j}; \boldsymbol{\phi}_j)$ using the data-points where x_j is observed and imputing the data-points where x_j is missing for $\forall j \in 1 \dots d$. For example, the t-th iteration of the method proceeds as follows

$$
\begin{aligned}
\hat{\boldsymbol{\phi}}_1^{(t)} &= \arg\max_{\boldsymbol{\phi}_1} \sum_{i \in \mathrm{obs}(\mathcal{D},1)} \log f(x_1^{(i,t-1)} \mid \boldsymbol{x}_{\setminus 1}^{(i,t-1)}; \hat{\boldsymbol{\phi}}_1^{(t-1)}) \\
x_1^{(i,t)} &\sim f(x_1^{(i,t)} \mid \boldsymbol{x}_{\setminus 1}^{(i,t-1)}; \hat{\boldsymbol{\phi}}_1^{(t)}) \text{ for } \forall i \in \mathrm{mis}(\mathcal{D}, 1) \\
&\vdots \\
\hat{\boldsymbol{\phi}}_d^{(t)} &= \arg\max_{\boldsymbol{\phi}_d} \sum_{i \in \mathrm{obs}(\mathcal{D},d)} \log f(x_d^{(i,t-1)} \mid \boldsymbol{x}_{\setminus d}^{(i,t)}; \hat{\boldsymbol{\phi}}_d^{(t-1)}) \\
x_d^{(i,t)} &\sim f(x_d^{(i,t)} \mid \boldsymbol{x}_{\setminus d}^{(i,t)}; \hat{\boldsymbol{\phi}}_d^{(t)}) \text{ for } \forall i \in \mathrm{mis}(\mathcal{D}, d),
\end{aligned}
$$

where $\mathrm{obs}(\mathcal{D}, j)$ and $\mathrm{mis}(\mathcal{D}, j)$ are respectively the sets of indices of data-points in the data set that are observed and missing in the j-th feature dimension. The procedure is usually repeated several times over all the missing variables to converge to reasonable imputations. Alternatively, if the conditional models $f(x_j \mid \boldsymbol{x}_{\setminus j}, \phi_j)$ admit it, the approach can be made Bayesian by specifying a prior over the parameters $\boldsymbol{\phi}_j$ and then sampling the parameters $\boldsymbol{\phi}_j^{(t)}$ from the posterior distribution $f(\boldsymbol{\phi}_j \mid \{\boldsymbol{x}_{<j}^{(t)}, \boldsymbol{x}_{\geq j}^{(t-1)}\}_{\mathrm{obs}(\mathcal{D},j)})$ instead of the $\arg\max$ in the above procedure.

Moreover, as dictated by the multiple-imputation framework, the overall imputation procedure is usually repeated K times, each time starting with random baseline imputations, to obtain K independent imputations of the missing data, which can then be used for many downstream tasks.

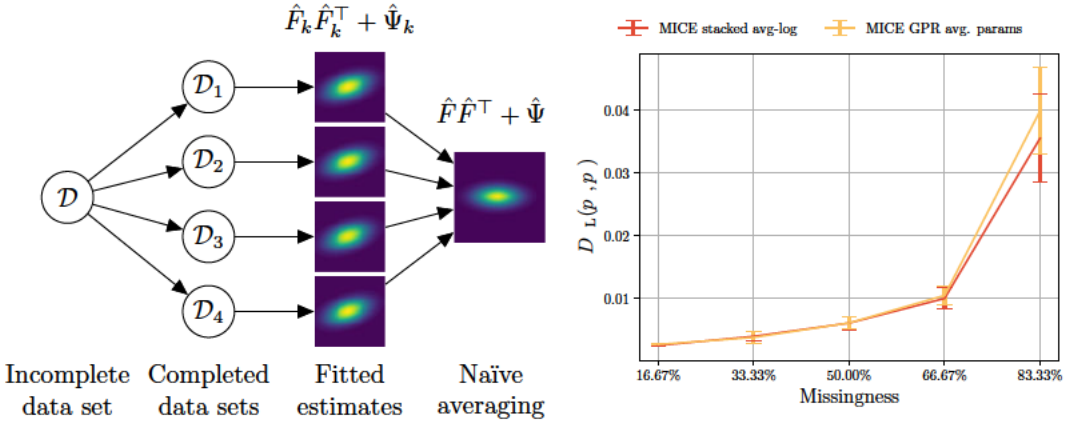


Figure 13: The effect of parameter averaging on a FA model when fitted on imputations from MICE. Left: naïve averaging of parameter estimates from different imputations removes the correlations in the averaged model. Right: comparison of model accuracy after parameter averaging of models trained on independent completed data set (yellow), where generalised Procrustes rotation was applied to the factor loading matrices prior the averaging, and a model trained on several completed data sets stacked together using average log-likelihood (red). Accuracy is measured using KL divergence of the factor analysis model to the ground truth model. Fitting on a stacked completed data set (red) yields equally good parameter estimates as the GPR-based averaging approach.

Appendix D. Combining Multiple Imputations for Statistical Model Estimation

Multiple imputation (MI, Rubin, 1987b) has been the primary tool for the analysis of incomplete data for more than two decades. Many books have covered the standard workflow of principled statistical analysis with multiple imputation and the risks of following invalid procedures (Schafer, 1997; Gelman et al., 2013; van Buuren, 2018).

In this section we first describe the standard analysis workflow with MI and then discuss why it is problematic or cannot be used for statistical models that have non-identifiable parameters. We demonstrate the issue on a classical factor analysis (FA) model, where, unlike for modern deep models (for example, VAEs), there are known methods to resolve the issue. We then explain the procedure we followed in our baseline experiments that use MI-based methods (for example, MICE) and verify on the FA model that it obtains equally good estimates of the statistical model as the standard analysis procedure.

In standard analysis with MI, initially K sets of imputed data are generated using a preferred multiple-imputation method such as the FCS described in Section 2.3. Then, for each imputed data set an independent analysis is performed treating the imputed data as complete and producing K estimates of a desired analysis quantity $\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^K$, namely one from each set of imputed data. Following Rubin's rules (Rubin, 1987b, Result 3.2) the

estimated quantities are then generally combined by averaging

$$\hat{\boldsymbol{\theta}} = \frac{1}{K}\sum_{k=1}^{K}\boldsymbol{\theta}^k.$$

The estimated quantities $\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^K$ can further be used to perform statistical tests and provide confidence intervals of the estimate $\hat{\boldsymbol{\theta}}$.

However, such averaging requires that $\boldsymbol{\theta}$ is identifiable. If identifiability is not possible this procedure may result in meaningless estimates. Parametric non-identifiability is a common property of many statistical models starting from classical models, such as mixtures of Gaussians and factor analysis, to modern deep models, such as variational autoencoders and normalising flows, where the non-identifiability stems from the use of neural networks.

We illustrate the issues arising in MI analysis from non-identifiability on a simple example, the factor analysis model, where the factor loading matrix is generally only identifiable up to a multiplication with an orthonormal matrix. Hence, estimating the factor matrix using MI may produce estimates $W^1, \ldots, W^K$ that could be written as the product of the true factor matrix F and an orthonormal matrix R^k, $W^k = FR^k$. Naïvely averaging such estimates would give

$$\hat{W} = \frac{1}{K}\sum_{k=1}^{K}W^k = \frac{1}{K}\sum_{k=1}^{K}FR^k = F\frac{1}{K}\sum_{k=1}^{K}R^k,$$

where the average over orthonormal matrices is not a valid orthonormal matrix, and hence the average result is not a valid estimate of the parameters. On the left side of Figure 13 we demonstrate that such naïve averaging approach produces a meaningless model estimate.

For classical models, such as mixtures of Gaussians and factor analysis it is often possible to resolve the indeterminacies of the K estimates before averaging. One solution for the FA model is to transform the factor loading matrices before averaging using the generalised (orthogonal) Procrustes rotation (GPR, Ten Berge, 1977; van Ginkel and Kroonenberg, 2014; Lorenzo-Seva and van Ginkel, 2016). However, no obvious solution exists to resolve the indeterminacy of deep models, which are the main interest of this paper, and hence the standard MI analysis methodology via parameter averaging described above is generally not applicable.

Therefore to avoid averaging, an alternative approach was used in our experiments where MI methods (such as MICE) were used to estimate baseline statistical models. Specifically, we have stacked the imputed data into a single K-times imputed data set and then estimate the model $p_{\boldsymbol{\theta}}$ using weighted maximum-likelihood by weighting each imputation by $1/K$

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \frac{1}{N}\sum_{i=1}^{N}\frac{1}{K}\sum_{k=1}^{K}\log p_{\boldsymbol{\theta}}(\tilde{\boldsymbol{x}}_{\text{mis}}^{(i,k)}, \boldsymbol{x}_{\text{obs}}^{i}),$$

where $\tilde{\boldsymbol{x}}_{\text{mis}}^{(i,k)}$ denotes the k-th imputation of the i-th sample.

On the right side of Figure 13 we validate that the stacking approach (red) described above produces equally good parameter estimates as the GPR-based parameter averaging (yellow) for all fractions of missingness. While standard statistical tests may provide invalid

test statistics or confidence intervals if (incorrectly) applied to the stacked data (e.g. van Buuren, 2018, Chapter 5.1) here we are primarily interested in obtaining point estimates of the parameters and hence use the stacked approach since no method (such as the GPR for FA) exists to resolve the parametric indeterminacy of neural networks.

Appendix E. VGI Derivation with Extended-Gibbs Kernel

We here generalise the VGI lower-bound (8) to the case of the extended-Gibbs kernel where the variational conditionals can depend on all variables imputed in the previous iteration. We use this extension to achieve a computationally more efficient method by enabling computation sharing among all variational conditionals (see Section 3.3).

We define the extended-Gibbs kernel by letting the variational conditionals depend on the previous imputation value $\tilde{x}_j$ in dimension j

$$\tilde{\tau}_{\boldsymbol{\phi}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{x}}_{\text{mis}}) = \sum_{j \in \text{idx}(\boldsymbol{m})} \boldsymbol{\pi}(j) q_{\boldsymbol{\phi}_j}(x_j \mid \boldsymbol{x}_{\text{mis} \setminus j}, \tilde{x}_j, \boldsymbol{x}_{\text{obs}}) \delta(\boldsymbol{x}_{\text{mis} \setminus j} - \tilde{\boldsymbol{x}}_{\text{mis} \setminus j}). \tag{15}$$

Marginalising out $\tilde{\boldsymbol{x}}_{\text{mis}}$ in the above equation with respect to the previous imputation distribution f^{t-1} yields the imputation distribution after a single Gibbs update

$$\begin{aligned}
f^t_{\boldsymbol{\phi}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}) &= \int \tilde{\tau}_{\boldsymbol{\phi}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, \tilde{\boldsymbol{x}}_{\text{mis}}) f^{t-1}(\tilde{\boldsymbol{x}}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}) \, \mathrm{d}\tilde{\boldsymbol{x}}_{\text{mis}} \\
&= \sum_{j \in \text{idx}(\boldsymbol{m})} \boldsymbol{\pi}(j) \int f^{t-1}(\tilde{x}_j \mid \boldsymbol{x}_{\text{obs}}) \\
&\qquad\qquad f^{t-1}(\boldsymbol{x}_{\text{mis} \setminus j} \mid \boldsymbol{x}_{\text{obs}}, \tilde{x}_j) q_{\boldsymbol{\phi}_j}(x_j \mid \boldsymbol{x}_{\text{mis} \setminus j}, \tilde{x}_j, \boldsymbol{x}_{\text{obs}}) \, \mathrm{d}\tilde{x}_j \\
&= \sum_{j \in \text{idx}(\boldsymbol{m})} \boldsymbol{\pi}(j) \int f^{t-1}(\tilde{x}_j \mid \boldsymbol{x}_{\text{obs}}) f^t_{\boldsymbol{\phi}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, \tilde{x}_j, j) \, \mathrm{d}\tilde{x}_j \\
&= \mathbb{E}_{\boldsymbol{\pi}(j) f^{t-1}(\tilde{x}_j \mid \boldsymbol{x}_{\text{obs}})} \left[f^t_{\boldsymbol{\phi}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, \tilde{x}_j, j) \right] \qquad (16) \\
f^t_{\boldsymbol{\phi}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, \tilde{x}_j, j) &= q_{\boldsymbol{\phi}_j}(x_j \mid \boldsymbol{x}_{\text{mis} \setminus j}, \tilde{x}_j, \boldsymbol{x}_{\text{obs}}) f^{t-1}(\boldsymbol{x}_{\text{mis} \setminus j} \mid \boldsymbol{x}_{\text{obs}}, \tilde{x}_j). \qquad (17)
\end{aligned}$$

Now we continue from the standard ELBO in (3) and use (16) and (17)

$$\begin{aligned}
\mathcal{L}^t(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}_{\text{obs}}) &\overset{(16)}{=} \mathbb{E}_{\boldsymbol{\pi}(j) f^{t-1}(\tilde{x}_j \mid \boldsymbol{x}_{\text{obs}}) f^t_{\boldsymbol{\phi}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, \tilde{x}_j, j)} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})}{f^t_{\boldsymbol{\phi}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})} \right] \\
&= \mathbb{E}_{\boldsymbol{\pi}(j) f^{t-1}(\tilde{x}_j \mid \boldsymbol{x}_{\text{obs}}) f^t_{\boldsymbol{\phi}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, \tilde{x}_j, j)} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})}{f^t_{\boldsymbol{\phi}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, \tilde{x}_j, j)} \right. \\
&\qquad\qquad \left. + \log \frac{f^t_{\boldsymbol{\phi}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, \tilde{x}_j, j)}{f^t_{\boldsymbol{\phi}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})} \right] \\
&= \mathbb{E}_{\boldsymbol{\pi}(j) f^{t-1}(\tilde{x}_j \mid \boldsymbol{x}_{\text{obs}}) f^t_{\boldsymbol{\phi}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, \tilde{x}_j, j)} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})}{f^t_{\boldsymbol{\phi}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, \tilde{x}_j, j)} \right]
\end{aligned}$$

VARIATIONAL GIBBS INFERENCE

$$
\begin{aligned}
&+ \mathbb{E}_{\boldsymbol{\pi}(j) f^{t-1}(\tilde{x}_j|\boldsymbol{x}_{\text{obs}})} D_{\text{KL}}(f^t_{\boldsymbol{\phi}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, \tilde{x}_j, j) \mid\mid f^t_{\boldsymbol{\phi}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})) \\
&\geq \mathbb{E}_{\boldsymbol{\pi}(j) f^{t-1}(\tilde{x}_j|\boldsymbol{x}_{\text{obs}}) f^t_{\boldsymbol{\phi}}(\boldsymbol{x}_{\text{mis}}|\boldsymbol{x}_{\text{obs}},\tilde{x}_j,j)} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})}{f^t_{\boldsymbol{\phi}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, \tilde{x}_j, j)} \right] \\
&\overset{(17)}{=} \mathbb{E}_{\boldsymbol{\pi}(j) f^{t-1}(\tilde{x}_j|\boldsymbol{x}_{\text{obs}}) f^{t-1}(\boldsymbol{x}_{\text{mis}\setminus j}|\boldsymbol{x}_{\text{obs}},\tilde{x}_j) q_{\boldsymbol{\phi}_j}(x_j|\boldsymbol{x}_{\text{mis}\setminus j},\tilde{x}_j,\boldsymbol{x}_{\text{obs}})} \Bigg[\\
&\qquad \log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})}{q_{\boldsymbol{\phi}_j}(x_j \mid \boldsymbol{x}_{\text{mis}\setminus j}, \tilde{x}_j, \boldsymbol{x}_{\text{obs}})} \Bigg] \\
&\quad - \mathbb{E}_{\boldsymbol{\pi}(j) f^{t-1}(\tilde{x}_j|\boldsymbol{x}_{\text{obs}}) f^{t-1}(\boldsymbol{x}_{\text{mis}\setminus j}|\boldsymbol{x}_{\text{obs}},\tilde{x}_j)} \left[\log f^{t-1}(\boldsymbol{x}_{\text{mis}\setminus j} \mid \boldsymbol{x}_{\text{obs}}, \tilde{x}_j) \right] \\
&= \mathbb{E}_{\boldsymbol{\pi}(j) f^{t-1}(\tilde{x}_j,\boldsymbol{x}_{\text{mis}\setminus j}|\boldsymbol{x}_{\text{obs}}) q_{\boldsymbol{\phi}_j}(x_j|\boldsymbol{x}_{\text{mis}\setminus j},\tilde{x}_j,\boldsymbol{x}_{\text{obs}})} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})}{q_{\boldsymbol{\phi}_j}(x_j \mid \boldsymbol{x}_{\text{mis}\setminus j}, \tilde{x}_j, \boldsymbol{x}_{\text{obs}})} \right] \\
&\quad - \mathbb{E}_{\boldsymbol{\pi}(j) f^{t-1}(\tilde{x}_j,\boldsymbol{x}_{\text{mis}\setminus j}|\boldsymbol{x}_{\text{obs}})} \left[\log f^{t-1}(\boldsymbol{x}_{\text{mis}\setminus j} \mid \boldsymbol{x}_{\text{obs}}, \tilde{x}_j) \right], \qquad (18)
\end{aligned}
$$

where the inequality follows from the non-negativity of KL divergence. Dropping the entropy term in the second line of (18) yields the VGI objective with extended conditionals. We note that the VGI ELBO in (18) obtained using the extended-Gibbs kernel is analogous to (8), only the coloured terms are different. To sum up:

- It uses the imputation value $\tilde{x}_j$ from the previous step in the Markov chain, highlighted in red.
- The variational conditionals $q_{\boldsymbol{\phi}_j}$ depend on the previous step in the Markov chain, highlighted in blue.
- The entropy term of the imputation distribution f^{t-1} from (8) became a conditional entropy, highlighted in magenta. Since the term does not include the parameters of interest $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, it does not need to be computed during optimisation.

The extended-Gibbs conditionals improve the efficiency of the main stage of the VGI algorithm. We note that the extended-Gibbs conditionals need to be separately defined for the variational warm-up stage (line 2 of Algorithm 1 and Algorithm 2), since x_j has no "previous imputation" to condition on. Including the true x_j in the conditioning set for the j-th dimension would cause the variational model to approximate the identity function, which may not be a desirable initialisation. To enable variational model warm-up with the extended conditionals we mask the j-th input to the conditional $q_{\boldsymbol{\phi}_j}$ with a zero.

Furthermore, in the VBGI objective for latent-variable models in (13), we can similarly let the univariate conditionals depend on the previous imputation, which results in the following VBGI objective with extended conditionals

$$
\begin{aligned}
\mathcal{J}^t_{\text{VBGI}}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}_{\text{obs}}) = \mathbb{E}_{f^{t-1}(\tilde{\boldsymbol{x}}_{\text{mis}},\tilde{\boldsymbol{x}}_z|\boldsymbol{x}_{\text{obs}})} \Bigg(& \\
\sum_{j \in \text{idx}(\boldsymbol{m})} \boldsymbol{\pi}(j) \mathbb{E}_{q_{\boldsymbol{\phi}_j}(x_j|\tilde{\boldsymbol{x}}_{\text{mis}},\tilde{\boldsymbol{x}}_z,\boldsymbol{x}_{\text{obs}})} & \left[\log \frac{p_{\boldsymbol{\theta}}(x_j, \tilde{\boldsymbol{x}}_{\text{mis}\setminus j}, \boldsymbol{x}_{\text{obs}} \mid \tilde{\boldsymbol{x}}_z) p_{\boldsymbol{\theta}}(\tilde{\boldsymbol{x}}_z)}{q_{\boldsymbol{\phi}_j}(x_j \mid \tilde{\boldsymbol{x}}_{\text{mis}}, \tilde{\boldsymbol{x}}_z, \boldsymbol{x}_{\text{obs}})} \right]
\end{aligned}
$$

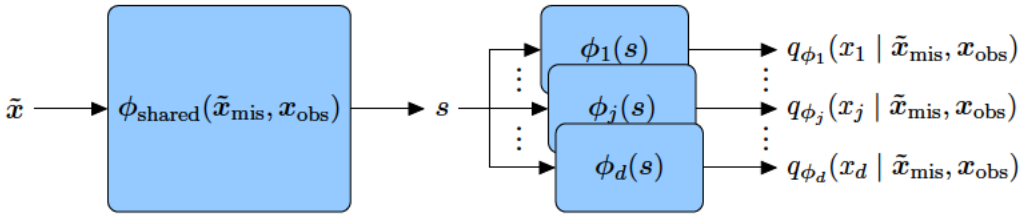


Figure 14: Partially-shared model of the extended variational conditionals.

$$+ \pi(j = z)\mathbb{E}_{q_{\phi_z}(x_z \mid \tilde{x}_{\text{mis}}, \tilde{x}_z, x_{\text{obs}})}\left[\log \frac{p_\theta(\tilde{x}_{\text{mis}}, x_{\text{obs}} \mid x_z) p_\theta(x_z)}{q_{\phi_z}(x_z \mid \tilde{x}_{\text{mis}}, \tilde{x}_z, x_{\text{obs}})}\right]\Bigg).$$

It allows us to achieve a similar computational improvement as in the standard VGI case with the extended conditionals.

Appendix F. Properties of the Gibbs Kernel in the Stationary Regime

We show that in the stationary regime the distribution obtained after a Gibbs transition is independent of the updated dimension j. We use a modified kernel from (4), which updates a fixed dimension j

$$\tau_\phi(x_{\text{mis}} \mid x_{\text{obs}}, \tilde{x}_{\text{mis}}, j) = q_{\phi_j}(x_j \mid x_{\text{mis}\setminus j}, x_{\text{obs}})\delta(x_{\text{mis}\setminus j} - \tilde{x}_{\text{mis}\setminus j}).$$

We now marginalise the kernel with respect to the stationary distribution $q^*(x_{\text{mis}} \mid x_{\text{obs}})$ with conditionals $q_{\phi_j}(x_j \mid x_{\text{mis}\setminus j}, x_{\text{obs}})$

$$\begin{aligned} q^*(x_{\text{mis}} \mid x_{\text{obs}}, j) &= \int \tau_\phi(x_{\text{mis}} \mid x_{\text{obs}}, \tilde{x}_{\text{mis}}, j) q^*(\tilde{x}_{\text{mis}} \mid x_{\text{obs}}) \, \mathrm{d}\tilde{x}_{\text{mis}} \\ &= q_{\phi_j}(x_j \mid x_{\text{mis}\setminus j}, x_{\text{obs}}) q^*(x_{\text{mis}\setminus j} \mid x_{\text{obs}}) \\ &= q^*(x_{\text{mis}} \mid x_{\text{obs}}). \end{aligned}$$

Hence, the stationary distribution is invariant to the updated dimension j.

Appendix G. The Variational Inference Network

Using VGI to fit a statistical model p_θ first requires us to specify and model d variational distributions. One way to model them is to use d inference networks with parameters ϕ_j, one for each variational distribution. Thus, the neural network for the j-th variational distribution takes an (imputed) data-point $\tilde{x}_{\setminus j}$ without the target dimension j and outputs the parameters of the distribution $q_{\phi_j}(x_j \mid \tilde{x}_{\setminus j})$. Note that all inference networks that parametrise the d Gibbs conditionals have the same input-output dimensionality, hence the computation of all variational distribution parameters can be efficiently parallelised using optimised matrix operations.

Considering that the variational distributions are approximating the conditionals of a joint target distribution we motivate that some parameter sharing in the variational model can be beneficial. One way to incorporate partial parameter sharing is to use a two-part

inference network, where the first part of the network is shared among all conditionals and the second part is independent for each conditional. The shared network can then be used for the different conditionals by fixing the j-th input dimension to zero when computing the intermediate representation for the j-th conditional. In our evaluations we compare an independent model with a partially-shared model and find that they both perform similarly well (see Section 4.7). Importantly, a partially-shared model allows us to scale VGI to higher-dimensional data, by increasing the parameter-efficiency of the variational model.

The VGI objective in (11) requires computing M conditional distributions for each datapoint, and the dimensions j^m of those conditionals can differ for each example. One way to compute the parameters of the conditional distributions is to evaluate the inference network for each dimension separately, however, this can require up to d evaluations and hence can be slow in higher dimensions. Alternatively, we could compute all d conditionals in parallel and then select only the required conditionals, however, this would not always be possible, since it would require d-times more memory. To mitigate this, for high-dimensional data we suggest working with the extended-Gibbs variational conditionals where the j-th conditional distribution is allowed to depend on the current imputation in the j-th dimension, thus giving variational distributions of the form $q_{\boldsymbol{\phi}_j}(x_j^t \mid \boldsymbol{x}_{\text{obs}}, \tilde{x}_j^{t-1}, \tilde{\boldsymbol{x}}_{\text{mis}\setminus j}^{t-1})$. Figure 14 depicts such an extended model. We now need a *single* pass through the shared layers, rather than d passes, which allows for a significant amount of the computation to be shared and hence gives a significant gain in computational efficiency. We empirically investigate the effect of the extended variational conditionals in Section 4.7, where we find no significant detrimental effects of the extended model when estimating higher-dimensional statistical models $p_{\boldsymbol{\theta}}$.

Appendix H. VGI Fine-Tuning Algorithm

In this section we describe the VGI fine-tuning algorithm (Algorithm 6) that is used to evaluate the VGI objective in (9) on held-out data. It is similar to the VGI training algorithm (Algorithm 1) when only the parameters $\boldsymbol{\phi}$ of the variational model are updated. In our experiments, we use this procedure to select model parameters $\hat{\boldsymbol{\theta}}$ that perform best on the validation data, which is not otherwise possible with amortised variational inference without fine-tuning due to the inference generalisation gap discussed in Section 3.5.

It starts with an incomplete held-out data set and finds the minimum and maximum values in the observed data that defines an imputation acceptance region for the warm-up stage. Then, the incomplete data are imputed using the initial imputation distribution f_0. In the first iteration of the algorithm, it first "warms up" the incomplete data imputations by performing G_{W} Gibbs updates, rejecting any imputations outside of the acceptance hypercube. Rejecting imputations that are far from observed data distribution mitigates some of the adverse effects of the inference generalisation gap, but other potentially better rejection strategies, such as the use of a Metropolis-Hastings acceptance step, may also be used. Then, the algorithm continues in the same way as Algorithm 1, fine-tuning the variational parameters $\boldsymbol{\phi}$ and updating imputations with G Gibbs updates in each iteration, accepting any imputations that are produced by the Gibbs sampler.

Algorithm 6 VGI fine-tuning algorithm

Input: $p_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x})$, a fitted probabilistic model with parameters $\hat{\boldsymbol{\theta}}$ (kept fixed)
$q_{\hat{\boldsymbol{\phi}}_j}(x_j \mid \boldsymbol{x}_{\setminus j})$ for $j \in \{1 \dots d\}$, var. conditionals fitted on train data with params $\hat{\boldsymbol{\phi}}$
$\mathcal{D}^{\text{val}}$, incomplete validation data set
K, number of imputations of each incomplete data-point
$f^0(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$, initial imputation distribution
G_{W}, number of Gibbs steps to warm up imputations
G, number of Gibbs imputation update steps in each epoch
max_epochs, number of epochs
Output: Validation loss $\hat{\mathcal{J}}_{\text{VGI}}$ and K-times imputed validation data $\mathcal{D}_K^{\text{val}}$

Find the maximum $\boldsymbol{x}_{\max}$ and minimum $\boldsymbol{x}_{\min}$ values for each dimension in $\mathcal{D}^{\text{val}}$
Create K-times imputed data $\mathcal{D}_K^{\text{val}}$ using f^0
$\boldsymbol{\phi}^0 \leftarrow \hat{\boldsymbol{\phi}}$ ▷ Set the initial parameters to the parameters of the fitted variational model
for t in $[1, \textit{max\_epochs}]$ **do**
  for mini-batch $\mathcal{B}_K$ in $\mathcal{D}_K^{\text{val}}$ **do**
    if $t = 1$ **then**
      Update $\mathcal{B}_K$ with G_{W} steps of (pseudo-)Gibbs sampler using the variational conditionals, rejecting any Gibbs update values x_j outside $[\boldsymbol{x}_{\min}(j), \boldsymbol{x}_{\max}(j)]$
    else
      Update $\mathcal{B}_K$ with G steps of (pseudo-)Gibbs sampler using Algorithm 5
    end if
    Store the updated imputations in $\mathcal{B}_K$ for use in the next epoch
    Compute $\hat{\mathcal{J}}_{\text{VGI}}^t$ in (11) using Algorithm 3
    $\boldsymbol{\phi}^t = \boldsymbol{\phi}^{t-1} + \alpha_{\boldsymbol{\phi}} \nabla_{\boldsymbol{\phi}} \hat{\mathcal{J}}_{\text{VGI}}^t$ ▷ Update params of $q_{\boldsymbol{\phi}}$ with a stochastic gradient step
  end for
end for

Appendix I. More on Markov Chain Variational Inference

We review the lower-bound from Markov chain variational inference (MCVI, Salimans et al., 2015), and show that replacing the reverse transition model r with the true reverse operator restores the tighter lower-bound using the marginal of a Markov chain in (3).

The MCVI lower-bound is defined in equation 4 in Salimans et al. (2015) as follows

$$\mathcal{L}_{\text{MCVI}}(\boldsymbol{x}) = \mathbb{E}_{q(\boldsymbol{z}_0,\dots,\boldsymbol{z}^T \mid \boldsymbol{x})} \left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z}^T)}{q_0(\boldsymbol{z}^0 \mid \boldsymbol{x})} + \sum_{t=1}^{T} \log \frac{r(\boldsymbol{z}^{t-1} \mid \boldsymbol{x}, \boldsymbol{z}^t)}{q(\boldsymbol{z}^t \mid \boldsymbol{x}, \boldsymbol{z}^{t-1})} \right]$$

where T is the total number of transitions in a Markov chain, q and r are transition and reverse transition functions, and $\boldsymbol{z}$ denotes the latent variable. The joint variational distribution q and the reverse distribution r are assumed to follow a Markov structure $q(\boldsymbol{z}^0, \dots, \boldsymbol{z}^T \mid \boldsymbol{x}) = q_0(\boldsymbol{z}^0 \mid \boldsymbol{x}) \prod_{t=1}^{T} q(\boldsymbol{z}^t \mid \boldsymbol{x}, \boldsymbol{z}^{t-1})$ and $r(\boldsymbol{z}^0, \dots, \boldsymbol{z}^{T-1} \mid \boldsymbol{x}, \boldsymbol{z}^T) = \prod_{t=1}^{T} r(\boldsymbol{z}^{t-1} \mid \boldsymbol{x}, \boldsymbol{z}^t)$.

We now show that $\mathcal{L}_{\text{MCVI}}$ is a looser lower-bound than the variational ELBO using the marginal of a Markov chain in (3), denoting the marginal distribution $q_T(\boldsymbol{z}^T \mid \boldsymbol{x}) = \int q_0(\boldsymbol{z}^0 \mid$

VARIATIONAL GIBBS INFERENCE

$\boldsymbol{x}) \prod_{t=1}^{T} q(\boldsymbol{z}^t \mid \boldsymbol{x}, \boldsymbol{z}^{t-1}) \, \mathrm{d}\boldsymbol{z}^{t-1}$.

$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}) \geq \mathcal{L}_T(\boldsymbol{x})$

$$
\begin{aligned}
\mathcal{L}_T(\boldsymbol{x}) &= \mathbb{E}_{q_T(\boldsymbol{z}^T|\boldsymbol{x})} \left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z}^T)}{q_T(\boldsymbol{z}^T \mid \boldsymbol{x})} \right] \\
&= \mathbb{E}_{q_T(\boldsymbol{z}^T|\boldsymbol{x})} \left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z}^T)}{q_T(\boldsymbol{z}^T \mid \boldsymbol{x})} + \mathbb{E}_{q(\boldsymbol{z}^{0\ldots T-1}|\boldsymbol{x}, \boldsymbol{z}^T)} \left[\log \frac{r(\boldsymbol{z}^{0\ldots T-1} \mid \boldsymbol{x}, \boldsymbol{z}^T) q(\boldsymbol{z}^{0\ldots T-1} \mid \boldsymbol{x}, \boldsymbol{z}^T)}{q(\boldsymbol{z}^{0\ldots T-1} \mid \boldsymbol{x}, \boldsymbol{z}^T) r(\boldsymbol{z}^{0\ldots T-1} \mid \boldsymbol{x}, \boldsymbol{z}^T)} \right] \right] \\
&= \mathbb{E}_{q(\boldsymbol{z}^0, \ldots, \boldsymbol{z}^T|\boldsymbol{x})} \left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z}^T)}{q_T(\boldsymbol{z}^T \mid \boldsymbol{x})} + \log \frac{r(\boldsymbol{z}^{0\ldots T-1} \mid \boldsymbol{x}, \boldsymbol{z}^T)}{q(\boldsymbol{z}^{0\ldots T-1} \mid \boldsymbol{x}, \boldsymbol{z}^T)} \right] \\
&\quad + \mathbb{E}_{q_T(\boldsymbol{z}^T|\boldsymbol{x})} \left[D_{\mathrm{KL}}(q(\boldsymbol{z}^{0\ldots T-1} \mid \boldsymbol{x}, \boldsymbol{z}^T) \mid\mid r(\boldsymbol{z}^{0\ldots T-1} \mid \boldsymbol{x}, \boldsymbol{z}^T)) \right] \\
&= \mathbb{E}_{q(\boldsymbol{z}^0, \ldots, \boldsymbol{z}^T|\boldsymbol{x})} \left[\log p(\boldsymbol{x}, \boldsymbol{z}^T) + \log \frac{r(\boldsymbol{z}^{0\ldots T-1} \mid \boldsymbol{x}, \boldsymbol{z}^T)}{q(\boldsymbol{z}^{0\ldots T} \mid \boldsymbol{x})} \right] \\
&\quad + \mathbb{E}_{q_T(\boldsymbol{z}^T|\boldsymbol{x})} \left[D_{\mathrm{KL}}(q(\boldsymbol{z}^{0\ldots T-1} \mid \boldsymbol{x}, \boldsymbol{z}^T) \mid\mid r(\boldsymbol{z}^{0\ldots T-1} \mid \boldsymbol{x}, \boldsymbol{z}^T)) \right] \\
&= \mathbb{E}_{q(\boldsymbol{z}^0, \ldots, \boldsymbol{z}^T|\boldsymbol{x})} \left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z}^T)}{q_0(\boldsymbol{z}^0 \mid \boldsymbol{x})} + \log \frac{\prod_{t=1}^{T} r(\boldsymbol{z}^{t-1} \mid \boldsymbol{x}, \boldsymbol{z}^t)}{\prod_{t=1}^{T} q_t(\boldsymbol{z}^t \mid \boldsymbol{x}, \boldsymbol{z}^{t-1})} \right] \\
&\quad + \mathbb{E}_{q_T(\boldsymbol{z}^T|\boldsymbol{x})} \left[D_{\mathrm{KL}}(q(\boldsymbol{z}^{0\ldots T-1} \mid \boldsymbol{x}, \boldsymbol{z}^T) \mid\mid r(\boldsymbol{z}^{0\ldots T-1} \mid \boldsymbol{x}, \boldsymbol{z}^T)) \right] \\
&= \mathbb{E}_{q(\boldsymbol{z}^0, \ldots, \boldsymbol{z}^T|\boldsymbol{x})} \left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z}^T)}{q_0(\boldsymbol{z}^0 \mid \boldsymbol{x})} + \sum_{t=1}^{T} \log \frac{r(\boldsymbol{z}^{t-1} \mid \boldsymbol{x}, \boldsymbol{z}^t)}{q_t(\boldsymbol{z}^t \mid \boldsymbol{x}, \boldsymbol{z}^{t-1})} \right] \\
&\quad + \mathbb{E}_{q_T(\boldsymbol{z}^T|\boldsymbol{x})} \left[D_{\mathrm{KL}}(q(\boldsymbol{z}^{0\ldots T-1} \mid \boldsymbol{x}, \boldsymbol{z}^T) \mid\mid r(\boldsymbol{z}^{0\ldots T-1} \mid \boldsymbol{x}, \boldsymbol{z}^T)) \right] \\
&= \mathcal{L}_{\mathrm{MCVI}}(\boldsymbol{x}) + \mathbb{E}_{q_T(\boldsymbol{z}^T|\boldsymbol{x})} \left[D_{\mathrm{KL}}(q(\boldsymbol{z}^{0\ldots T-1} \mid \boldsymbol{x}, \boldsymbol{z}^T) \mid\mid r(\boldsymbol{z}^{0\ldots T-1} \mid \boldsymbol{x}, \boldsymbol{z}^T)) \right]
\end{aligned}
$$

$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}) \geq \mathcal{L}_T(\boldsymbol{x}) \geq \mathcal{L}_{\mathrm{MCVI}}(\boldsymbol{x})$

Moreover, if we choose the reverse transition function r in MCVI to be the true reverse operator, such that it produces the distribution of the chain at the previous step (e.g. Murray, 2007, Chapter 1.4)

$$
r(\boldsymbol{z}^{t-1} \mid \boldsymbol{x}, \boldsymbol{z}^t) = \frac{q(\boldsymbol{z}^t \mid \boldsymbol{x}, \boldsymbol{z}^{t-1}) q_{t-1}(\boldsymbol{z}^{t-1} \mid \boldsymbol{x})}{q_t(\boldsymbol{z}^t \mid \boldsymbol{x})},
$$

then from MCVI we can recover the tighter lower-bound in (3)

$$
\begin{aligned}
\mathcal{L}_{\mathrm{MCVI}}(\boldsymbol{x}) &= \mathbb{E}_{q(\boldsymbol{z}_0, \ldots, \boldsymbol{z}^T|\boldsymbol{x})} \left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z}^T)}{q(\boldsymbol{z}^0 \mid \boldsymbol{x})} + \sum_{t=1}^{T} \log \frac{r(\boldsymbol{z}^{t-1} \mid \boldsymbol{x}, \boldsymbol{z}^t)}{q(\boldsymbol{z}^t \mid \boldsymbol{x}, \boldsymbol{z}^{t-1})} \right] \\
&= \mathbb{E}_{q(\boldsymbol{z}_0, \ldots, \boldsymbol{z}^T|\boldsymbol{x})} \left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z}^T)}{q(\boldsymbol{z}^0 \mid \boldsymbol{x})} + \sum_{t=1}^{T} \log \frac{q(\boldsymbol{z}^t \mid \boldsymbol{x}, \boldsymbol{z}^{t-1}) q_{t-1}(\boldsymbol{z}^{t-1} \mid \boldsymbol{x})}{q_t(\boldsymbol{z}^t \mid \boldsymbol{x}) q(\boldsymbol{z}^t \mid \boldsymbol{x}, \boldsymbol{z}^{t-1})} \right]
\end{aligned}
$$

$$
\begin{aligned}
&= \mathbb{E}_{q(\boldsymbol{z}_0,\dots,\boldsymbol{z}^T|\boldsymbol{x})}\left[\log\frac{p(\boldsymbol{x},\boldsymbol{z}^T)}{q(\boldsymbol{z}^0 \mid \boldsymbol{x})} + \sum_{t=1}^{T}\log\frac{q(\boldsymbol{z}^t \mid \boldsymbol{x},\boldsymbol{z}^{t-1})q_{t-1}(\boldsymbol{z}^{t-1} \mid \boldsymbol{x})}{q_t(\boldsymbol{z}^t \mid \boldsymbol{x})q(\boldsymbol{z}^t \mid \boldsymbol{x},\boldsymbol{z}^{t-1})}\right] \\
&= \mathbb{E}_{q_T(\boldsymbol{z}^T|\boldsymbol{x})}\left[\log\frac{p(\boldsymbol{x},\boldsymbol{z}^T)}{q_T(\boldsymbol{z}^T \mid \boldsymbol{x})}\right] \\
&= \mathcal{L}_T(\boldsymbol{x})
\end{aligned}
$$

Hence, having to learn a reverse model r loosens the lower-bound on $\log p_{\boldsymbol{\theta}}(\boldsymbol{x})$, which makes maximising the likelihood more difficult. The VGI method, on the other hand, does not require learning a model r to reverse the Markov sampling path, and optimises a lower-bound that asymptotically converges to $\mathcal{L}_T(\boldsymbol{x})$.

Appendix J. Expectation-Maximisation for Factor Analysis with Incomplete Data

We here derive the EM solution for the FA model with incomplete data. From Section 4.1 we know that the FA model is a Gaussian model, with zero-mean standard Gaussian latent variables $\boldsymbol{z}$ and a linear Gaussian generative model $p(\boldsymbol{x} \mid \boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{F}\boldsymbol{z} + \boldsymbol{\mu}, \boldsymbol{\Psi})$. We assume that $\boldsymbol{x}$ has observed $\boldsymbol{x}_{\text{obs}}$ and missing $\boldsymbol{x}_{\text{mis}}$ components, and write down the joint distribution

$$
p\left(\begin{pmatrix}\boldsymbol{x}_{\text{obs}}\\ \boldsymbol{x}_{\text{mis}}\\ \boldsymbol{z}\end{pmatrix}\right) = \mathcal{N}\left(\begin{pmatrix}\boldsymbol{x}_{\text{obs}}\\ \boldsymbol{x}_{\text{mis}}\\ \boldsymbol{z}\end{pmatrix}; \begin{pmatrix}\boldsymbol{\mu}_{\text{obs}}\\ \boldsymbol{\mu}_{\text{mis}}\\ \boldsymbol{0}\end{pmatrix}, \begin{bmatrix}\boldsymbol{F}_{\text{obs}}\boldsymbol{F}_{\text{obs}}^\top + \boldsymbol{\Psi}_{\text{obs}} & \boldsymbol{F}_{\text{obs}}\boldsymbol{F}_{\text{mis}}^\top & \boldsymbol{F}_{\text{obs}}\\ \boldsymbol{F}_{\text{mis}}\boldsymbol{F}_{\text{obs}}^\top & \boldsymbol{F}_{\text{mis}}\boldsymbol{F}_{\text{mis}}^\top + \boldsymbol{\Psi}_{\text{mis}} & \boldsymbol{F}_{\text{mis}}\\ \boldsymbol{F}_{\text{obs}}^\top & \boldsymbol{F}_{\text{mis}}^\top & \boldsymbol{I}\end{bmatrix}\right). \tag{19}
$$

Following the standard EM procedure (e.g. Barber, 2017, Chapter 11) we write down the energy, where we remove a constant term due to the prior on $\boldsymbol{z}$

$$
\begin{aligned}
E(\boldsymbol{\theta};\boldsymbol{\theta}^{(t-1)}) &= \sum_{i=1}^{N}\mathbb{E}_{p_{\boldsymbol{\theta}^{(t-1)}}(\boldsymbol{x}_{\text{mis}},\boldsymbol{z}|\boldsymbol{x}^i_{\text{obs}})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}^i_{\text{obs}},\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{z})\right] \\
&= -\frac{1}{2}\sum_{i=1}^{N}\mathbb{E}\left[d\log(2\pi) + \log\det|\boldsymbol{\Psi}| + (\boldsymbol{x}^i - \boldsymbol{\mu} - \boldsymbol{F}\boldsymbol{z})^\top\boldsymbol{\Psi}^{-1}(\boldsymbol{x}^i - \boldsymbol{\mu} - \boldsymbol{F}\boldsymbol{z})\right] \\
&\propto -\frac{N}{2}\log\det|\boldsymbol{\Psi}| - \frac{1}{2}\sum_{i=1}^{N}\mathbb{E}\left[(\boldsymbol{x}^i - \boldsymbol{\mu} - \boldsymbol{F}\boldsymbol{z})^\top\boldsymbol{\Psi}^{-1}(\boldsymbol{x}^i - \boldsymbol{\mu} - \boldsymbol{F}\boldsymbol{z})\right],
\end{aligned} \tag{20}
$$

where $\boldsymbol{\theta} = (\boldsymbol{F}, \boldsymbol{\mu}, \boldsymbol{\Psi})$ and $\boldsymbol{\theta}^{(t-1)}$ is the current estimate of $\boldsymbol{\theta}$. The expectation operator $\mathbb{E}[\cdot]$ in (20) means expectation with respect to $p_{\boldsymbol{\theta}^{(t-1)}}(\boldsymbol{x}_{\text{mis}}, \boldsymbol{z} \mid \boldsymbol{x}^i_{\text{obs}})$, which is

$$
p_{\boldsymbol{\theta}^{(t-1)}}\left(\begin{pmatrix}\boldsymbol{x}_{\text{mis}}\\ \boldsymbol{z}\end{pmatrix} \,\middle|\, \boldsymbol{x}^i_{\text{obs}}\right) = \mathcal{N}\left(\begin{pmatrix}\boldsymbol{x}_{\text{mis}}\\ \boldsymbol{z}\end{pmatrix}; \begin{pmatrix}\boldsymbol{\mu}_{\text{mis|obs}}\\ \boldsymbol{\mu}_{\text{z|obs}}\end{pmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{\text{mis|obs}} & \boldsymbol{C}_{\text{mis},z|\text{obs}}\\ \boldsymbol{C}^\top_{\text{mis},z|\text{obs}} & \boldsymbol{\Sigma}_{z|\text{obs}}\end{bmatrix}\right). \tag{21}
$$

In the E-step of the algorithm we compute the statistics required in the above from (19) using the rules for conditionals of multivariate Gaussian distributions (Petersen and Pedersen, 2012).

$$\begin{aligned}
\boldsymbol{\Sigma}_{z|\text{obs}} &= \boldsymbol{I} - \boldsymbol{F}_{\text{obs}}^\top(\boldsymbol{F}_{\text{obs}}\boldsymbol{F}_{\text{obs}}^\top + \boldsymbol{\Psi}_{\text{obs}})^{-1}\boldsymbol{F}_{\text{obs}} = (\boldsymbol{I} + \boldsymbol{F}_{\text{obs}}^\top\boldsymbol{\Psi}_{\text{obs}}^{-1}\boldsymbol{F}_{\text{obs}})^{-1} && (22)\\
\boldsymbol{\mu}_{z|\text{obs}} &= \boldsymbol{F}_{\text{obs}}^\top(\boldsymbol{F}_{\text{obs}}\boldsymbol{F}_{\text{obs}}^\top + \boldsymbol{\Psi}_{\text{obs}})^{-1}(\boldsymbol{x}_{\text{obs}} - \boldsymbol{\mu}_{\text{obs}}) = \boldsymbol{\Sigma}_{z|\text{obs}}\boldsymbol{F}_{\text{obs}}^\top\boldsymbol{\Psi}_{\text{obs}}^{-1}(\boldsymbol{x}_{\text{obs}} - \boldsymbol{\mu}_{\text{obs}}) && (23)\\
\boldsymbol{\Sigma}_{\text{mis}|\text{obs}} &= \boldsymbol{F}_{\text{mis}}\boldsymbol{F}_{\text{mis}}^\top + \boldsymbol{\Psi}_{\text{mis}} - \boldsymbol{F}_{\text{mis}}\boldsymbol{F}_{\text{obs}}^\top(\boldsymbol{F}_{\text{obs}}\boldsymbol{F}_{\text{obs}}^\top + \boldsymbol{\Psi}_{\text{obs}})^{-1}\boldsymbol{F}_{\text{obs}}\boldsymbol{F}_{\text{mis}}^\top\\
&= \boldsymbol{F}_{\text{mis}}\boldsymbol{\Sigma}_{z|\text{obs}}\boldsymbol{F}_{\text{mis}}^\top + \boldsymbol{\Psi}_{\text{mis}}\\
\boldsymbol{\mu}_{\text{mis}|\text{obs}} &= \boldsymbol{\mu}_{\text{mis}} + \boldsymbol{F}_{\text{mis}}\boldsymbol{F}_{\text{obs}}^\top(\boldsymbol{F}_{\text{obs}}\boldsymbol{F}_{\text{obs}}^\top + \boldsymbol{\Psi}_{\text{obs}})^{-1}(\boldsymbol{x}_{\text{obs}} - \boldsymbol{\mu}_{\text{obs}}) = \boldsymbol{\mu}_{\text{mis}} + \boldsymbol{F}_{\text{mis}}\boldsymbol{\mu}_{z|\text{obs}}\\
\boldsymbol{C}_{\text{mis},z|\text{obs}} &= \boldsymbol{F}_{\text{mis}} - \boldsymbol{F}_{\text{mis}}\boldsymbol{F}_{\text{obs}}^\top(\boldsymbol{F}_{\text{obs}}\boldsymbol{F}_{\text{obs}}^\top + \boldsymbol{\Psi}_{\text{obs}})^{-1}\boldsymbol{F}_{\text{obs}} = \boldsymbol{F}_{\text{mis}}\boldsymbol{\Sigma}_{z|\text{obs}},
\end{aligned}$$

where in (22) we used Woodbury's identity (e.g. Petersen and Pedersen, 2012) and in (23) we used the push-through identity (Henderson and Searle, 1981). Also note that these statistics depend on the observed data $\boldsymbol{x}_{\text{obs}}^i$ or the observed dimensions $\boldsymbol{m}$, we suppressed the index i in the above notation.

In the M-step of the algorithm we maximise the energy with respect to $\boldsymbol{\theta}$. First, to derive the solution for $\boldsymbol{\mu}$ we write down a simplified energy term where $\boldsymbol{z}$ are marginalised out

$$\begin{aligned}
E_{\boldsymbol{x}}(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t-1)}) &= \sum_{i=1}^{N} \mathbb{E}_{p_{\boldsymbol{\theta}^{(t-1)}}(\boldsymbol{x}_{\text{mis}}|\boldsymbol{x}_{\text{obs}}^i)}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}}^i, \boldsymbol{x}_{\text{mis}})\right]\\
&= -\frac{1}{2}\sum_{i=1}^{N} \mathbb{E}\left[d\log(2\pi) + \log\det|\boldsymbol{\Sigma}| + (\boldsymbol{x}^i - \boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}^i - \boldsymbol{\mu})\right]. && (24)
\end{aligned}$$

The expectation operator $\mathbb{E}[\cdot]$ in the above equation depends on $p_{\boldsymbol{\theta}^{(t-1)}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}^i)$, which can be directly read from (21) as the marginal of a Gaussian (e.g. Petersen and Pedersen, 2012, Section 8.1.2). Then, taking the partial derivative and setting it to zero we get $\hat{\boldsymbol{\mu}}$

$$\begin{aligned}
\frac{\partial E_{\boldsymbol{x}}(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t-1)})}{\partial \boldsymbol{\mu}} &= -\frac{1}{2}\sum_{i=1}^{N} \mathbb{E}\left[\frac{\partial}{\partial \boldsymbol{\mu}}\left((\boldsymbol{x}^i - \boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}^i - \boldsymbol{\mu})\right)\right]\\
&= -\sum_{i=1}^{N} \mathbb{E}\left[(\boldsymbol{x}^i - \boldsymbol{\mu})^\top\right]\boldsymbol{\Sigma}^{-1}\\
&= N\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1} - \sum_{i=1}^{N} \mathbb{E}\left[\boldsymbol{x}^i\right]^\top\boldsymbol{\Sigma}^{-1} = 0\\
\implies \hat{\boldsymbol{\mu}} &= \frac{1}{N}\sum_{i=1}^{N} \hat{\boldsymbol{x}}^i, && (25)
\end{aligned}$$

where $\hat{\boldsymbol{x}}^i = (\boldsymbol{x}_{\text{obs}}^i, \boldsymbol{\mu}_{\text{mis}|\text{obs}}^i)$.

Now, from (20) we derive the updated $\hat{\boldsymbol{F}}$

$$\frac{\partial E(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t-1)})}{\partial \boldsymbol{F}} = -\frac{1}{2}\sum_{i=1}^{N} \mathbb{E}\left[\frac{\partial}{\partial \boldsymbol{F}}\left((\boldsymbol{x}^i - \boldsymbol{\mu} - \boldsymbol{F}\boldsymbol{z})^\top\boldsymbol{\Psi}^{-1}(\boldsymbol{x}^i - \boldsymbol{\mu} - \boldsymbol{F}\boldsymbol{z})\right)\right]$$

$$
\begin{aligned}
&= -\sum_{i=1}^{N} \mathbb{E}\left[(\boldsymbol{x}^i - \boldsymbol{\mu} - \boldsymbol{F}\boldsymbol{z})\boldsymbol{z}^\top \boldsymbol{\Psi}^{-1}\right] \\
&= -\sum_{i=1}^{N} \mathbb{E}\left[(\boldsymbol{x}^i - \boldsymbol{\mu})\boldsymbol{z}^\top\right] \boldsymbol{\Psi}^{-1} + \boldsymbol{F}\sum_{i=1}^{N} \mathbb{E}\left[\boldsymbol{z}\boldsymbol{z}^\top\right] \boldsymbol{\Psi}^{-1} = 0 \\
\implies \hat{\boldsymbol{F}}\boldsymbol{H} &= \boldsymbol{A} \\
\implies \hat{\boldsymbol{F}} &= \boldsymbol{A}\boldsymbol{H}^{-1}
\end{aligned}
$$

where

$$
\begin{aligned}
\boldsymbol{H} &= \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[\boldsymbol{z}\boldsymbol{z}^\top\right] = \frac{1}{N}\sum_{i=1}^{N} \left(\boldsymbol{\Sigma}^i_{z|\text{obs}} + \boldsymbol{\mu}^i_{z|\text{obs}}\boldsymbol{\mu}^{i\top}_{z|\text{obs}}\right) \\
\boldsymbol{A} &= \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[(\boldsymbol{x}^i - \hat{\boldsymbol{\mu}})\boldsymbol{z}^\top\right] = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[\boldsymbol{x}^i\boldsymbol{z}^\top\right] - \hat{\boldsymbol{\mu}}\boldsymbol{\mu}^{i\top}_{z|\text{obs}} \\
&= \frac{1}{N}\sum_{i=1}^{N} \bar{\boldsymbol{C}}^i_{\text{mis},z|\text{obs}} + \hat{\boldsymbol{x}}^i\boldsymbol{\mu}^{i\top}_{z|\text{obs}} - \hat{\boldsymbol{\mu}}\boldsymbol{\mu}^{i\top}_{z|\text{obs}} \\
&= \frac{1}{N}\sum_{i=1}^{N} \bar{\boldsymbol{C}}^i_{\text{mis},z|\text{obs}} + (\hat{\boldsymbol{x}}^i - \hat{\boldsymbol{\mu}})\boldsymbol{\mu}^{i\top}_{z|\text{obs}},
\end{aligned}
$$

and $\bar{\boldsymbol{C}}_{\text{mis},z|\text{obs}}$ is $\boldsymbol{C}_{\text{mis},z|\text{obs}}$ with 0-rows for the observed dimensions.

Now we derive solution for $\hat{\boldsymbol{\Psi}}$, using diag() to denote a function that sets off-diagonal elements to zero

$$
\begin{aligned}
\frac{\partial E(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t-1)})}{\partial \boldsymbol{\Psi}} &= -\frac{\partial}{\partial \boldsymbol{\Psi}}\frac{N}{2}\log\det|\boldsymbol{\Psi}| \\
&\quad - \text{diag}\left(\sum_{i=1}^{N} \mathbb{E}\left[\frac{\partial}{\partial \boldsymbol{\Psi}}\left\{\frac{1}{2}(\boldsymbol{x}^i - \boldsymbol{\mu} - \boldsymbol{F}\boldsymbol{z})^\top \boldsymbol{\Psi}^{-1}(\boldsymbol{x}^i - \boldsymbol{\mu} - \boldsymbol{F}\boldsymbol{z})\right\}\right]\right) \\
&= -\frac{N}{2}\boldsymbol{\Psi}^{-1} + \text{diag}\left(\sum_{i=1}^{N} \mathbb{E}\left[\frac{1}{2}\boldsymbol{\Psi}^{-1}(\boldsymbol{x}^i - \boldsymbol{\mu} - \boldsymbol{F}\boldsymbol{z})(\boldsymbol{x}^i - \boldsymbol{\mu} - \boldsymbol{F}\boldsymbol{z})^\top \boldsymbol{\Psi}^{-1}\right]\right) = 0 \\
\implies \hat{\boldsymbol{\Psi}} &= \text{diag}\left(\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[(\boldsymbol{x}^i - \hat{\boldsymbol{\mu}} - \boldsymbol{F}\boldsymbol{z})(\boldsymbol{x}^i - \hat{\boldsymbol{\mu}} - \boldsymbol{F}\boldsymbol{z})^\top\right]\right) \\
&= \text{diag}\left(\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[(\boldsymbol{x}^i - \hat{\boldsymbol{\mu}})(\boldsymbol{x}^i - \hat{\boldsymbol{\mu}})^\top - 2\boldsymbol{F}\boldsymbol{z}(\boldsymbol{x}^i - \hat{\boldsymbol{\mu}})^\top + \boldsymbol{F}\boldsymbol{z}\boldsymbol{z}^\top\boldsymbol{F}^\top\right]\right) \\
&= \text{diag}\left(\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[(\boldsymbol{x}^i - \hat{\boldsymbol{\mu}})(\boldsymbol{x}^i - \hat{\boldsymbol{\mu}})^\top\right] - 2\boldsymbol{F}\boldsymbol{A}^\top + \boldsymbol{F}\boldsymbol{H}\boldsymbol{F}^\top\right) \\
&= \text{diag}\left(\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[\boldsymbol{x}^i\boldsymbol{x}^{i\top} - 2\boldsymbol{x}^i\hat{\boldsymbol{\mu}}^\top + \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^\top\right] - 2\boldsymbol{F}\boldsymbol{A}^\top + \boldsymbol{F}\boldsymbol{H}\boldsymbol{F}^\top\right)
\end{aligned}
$$

$$= \operatorname{diag}\left(\boldsymbol{V} - 2\boldsymbol{F}\boldsymbol{A}^\top + \boldsymbol{F}\boldsymbol{H}\boldsymbol{F}^\top\right),$$

with

$$\begin{aligned}\boldsymbol{V} &= \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[\boldsymbol{x}^i \boldsymbol{x}^{i\top}\right] - 2\hat{\boldsymbol{x}}^i \hat{\boldsymbol{\mu}}^\top + \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^\top \\ &= \frac{1}{N}\sum_{i=1}^{N} \bar{\boldsymbol{\Sigma}}^i_{\text{mis}|\text{obs}} + \hat{\boldsymbol{x}}^i \hat{\boldsymbol{x}}^{i\top} - 2\hat{\boldsymbol{x}}^i \hat{\boldsymbol{\mu}}^\top + \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^\top \\ &= \frac{1}{N}\sum_{i=1}^{N} \bar{\boldsymbol{\Sigma}}^i_{\text{mis}|\text{obs}} + (\hat{\boldsymbol{x}}^i - \hat{\boldsymbol{\mu}})(\hat{\boldsymbol{x}}^i - \hat{\boldsymbol{\mu}})^\top,\end{aligned}$$

and where $\bar{\boldsymbol{\Sigma}}^i_{\text{mis}|\text{obs}}$ is $\boldsymbol{\Sigma}^i_{\text{mis}|\text{obs}}$ with zero rows and columns for the observed dimensions in $\boldsymbol{x}^i$. The solutions from the current iteration are then set to $\boldsymbol{\theta}^{(t)} = (\hat{\boldsymbol{F}}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Psi}})$ and used in the next iteration.

Similar EM update rules for FA with incomplete data have been derived by Marlin (2008, Chapter 4.3.2), however our update of the mean vector in (25) uses a tighter lower-bound, by marginalising the latents from the energy in (24), while theirs uses a looser lower-bound that results in the use of the energy in (20). Thus, we can expect that using our update rule should converge faster. Finally, we note that our derivation closely follows the derivation of EM for FA with complete data by Barber (2017, Chapter 21.2.2) with extra terms arising due to incomplete data.

Appendix K. Toy Data Ground Truth

The toy data was generated using a FA model with the following parameters as the ground truth:

$$\boldsymbol{F} = \begin{bmatrix} -5 & -2 \\ 4 & 0 \\ -3 & -1 \\ -3 & -3 \\ 1 & 5 \\ -1 & 2 \end{bmatrix} \quad \boldsymbol{\mu} = \begin{bmatrix} 3 \\ -1 \\ 0 \\ 2 \\ -1 \\ 0 \end{bmatrix} \quad \operatorname{diag}(\boldsymbol{\Psi}) = \begin{bmatrix} 50.4794 \\ 30.0988 \\ 6.766 \\ 17.3357 \\ 40.9839 \\ 25.1122 \end{bmatrix}.$$

The toy data intentionally has a low signal-to-noise ratio, which made the estimation problem harder.

Appendix L. VAE Model Selection on Incomplete Held-Out Data

The primary goal of our investigation in Section 5 is to evaluate the accuracy of the VAE model $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ to understand how well it estimates the distribution of the data. Like all other amortised variational methods, the evaluation on held-out data requires fine-tuning of the variational models to the target data due to the inference generalisation gap discussed in Section 3.5. Hence, to perform model selection via validation on held-out data, we

store checkpoints of all parameters (including the generator and the variational models) at predefined intervals during training. Then, we retrospectively fine-tune the variational models to the validation data and select, out of all checkpointed states, the generator parameters that performed best on the fine-tuned validation loss. For VGI we use the validation loss obtained via the fine-tuning algorithm in Appendix H and for the other methods we use their respective loss with the fine-tuned encoder networks.

In Figure 20 of Appendix M we show the corresponding objective curves during fine-tuning of the VGI methods and MVAE,[28] where we observe that the fine-tuning of VGI-based methods converge significantly faster than MVAE. Moreover, in Figure 21 of Appendix M we show the validation (dashed) and fine-tuned validation (dash-dotted) learning curves. We can see that the inference generalisation gap affects all evaluated methods (note the significant gap between dashed and dash-dotted curves), hence confirming that fine-tuning of the variational distributions is necessary to perform model selection.

Appendix M. Additional Figures

In this section we show additional figures from the experiments in Sections 4-6.

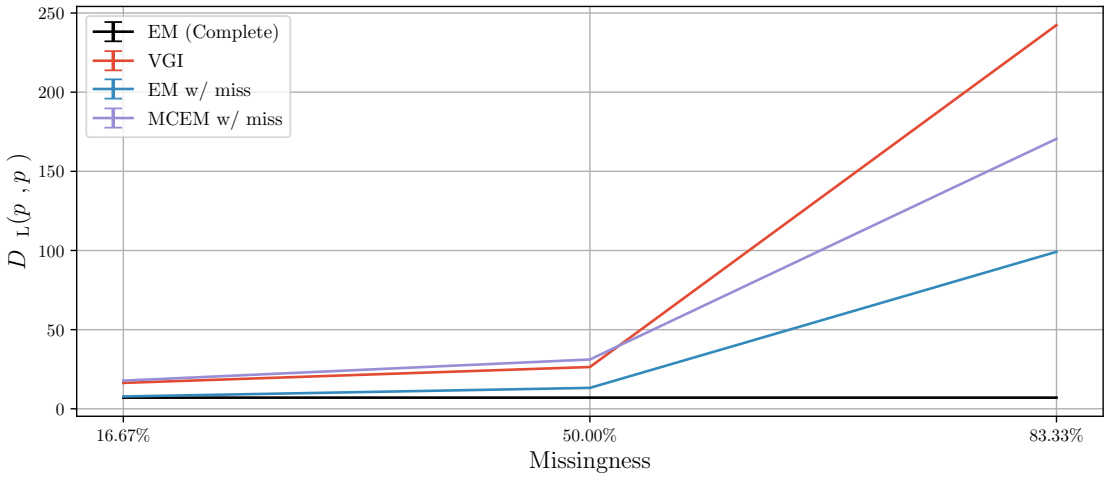


Figure 15: The accuracy of the fitted statistical models as $D_{\mathrm{KL}}(p_* \,||\, p_{\boldsymbol{\theta}})$ on FA-Frey data. MCEM samples imputations from the true conditional $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathrm{mis}} \mid \boldsymbol{x}_{\mathrm{obs}})$ using the learnt model $p_{\boldsymbol{\theta}}$ and using SGA fits the model on the imputed data. Note that the curves for VGI (red) and MCEM (purple) are close, hence we attribute the performance gap between the EM and VGI to the stochastic optimisation.

28. The fine-tuning loss curves of the other VAE-specific methods were similar to MVAE and hence are not shown.

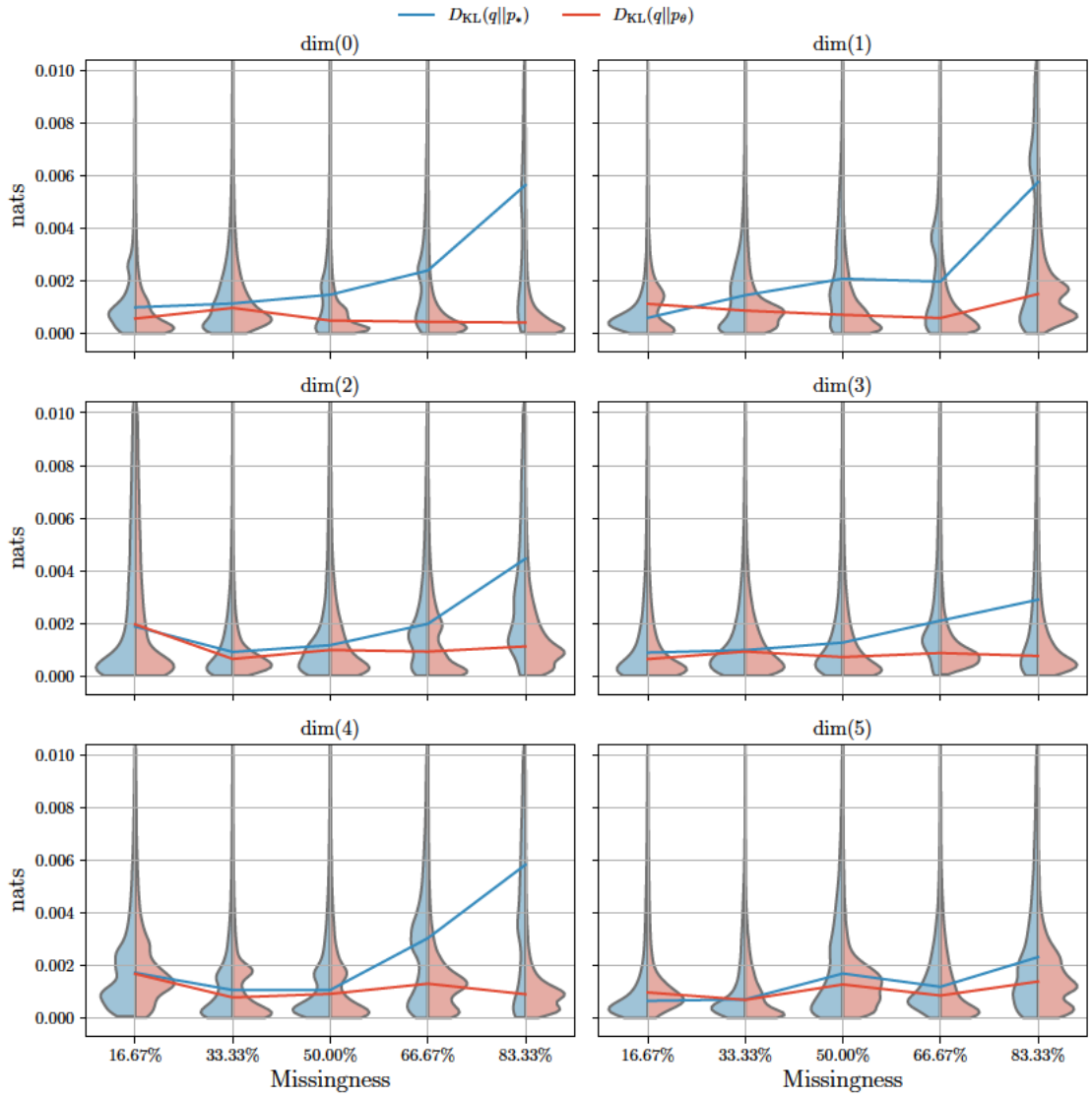


Figure 16: KL divergence, conditioned on the toy test set, between the univariate variational conditional distributions and the ground truth distribution for each feature dimension $D_{\mathrm{KL}}(q_{\phi_j}(x_j \mid \boldsymbol{x}_{\setminus j}) \mid\mid p_*(x_j \mid \boldsymbol{x}_{\setminus j}))$ (blue), and the posterior under the learnt model $D_{\mathrm{KL}}(q_{\phi_j}(x_j \mid \boldsymbol{x}_{\setminus j}) \mid\mid p_{\boldsymbol{\theta}}(x_j \mid \boldsymbol{x}_{\setminus j}))$ (red). The lines show the median conditional KL divergence on the test set.

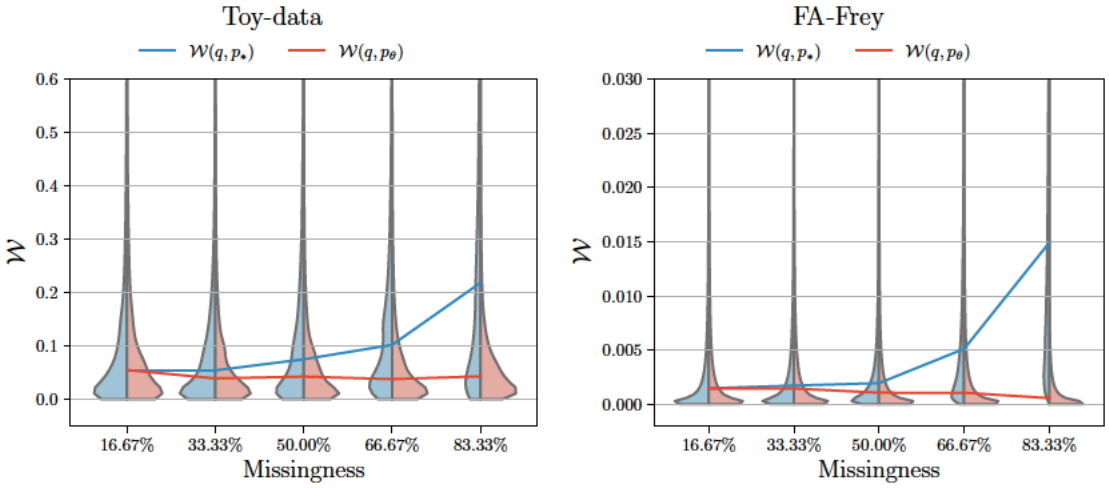


Figure 17: Wasserstein distance, conditioned on the test set (left: toy data, right: FA-Frey), between the univariate variational conditional distributions and the ground truth distribution $\mathcal{W}(q_{\phi_j}(x_j \mid \boldsymbol{x}_{\setminus j}), p_*(x_j \mid \boldsymbol{x}_{\setminus j}))$ (blue), and the posterior under the learnt model $\mathcal{W}(q_{\phi_j}(x_j \mid \boldsymbol{x}_{\setminus j}), p_{\boldsymbol{\theta}}(x_j \mid \boldsymbol{x}_{\setminus j}))$ (red). The lines show the median Wasserstein distance conditioned on the test set.

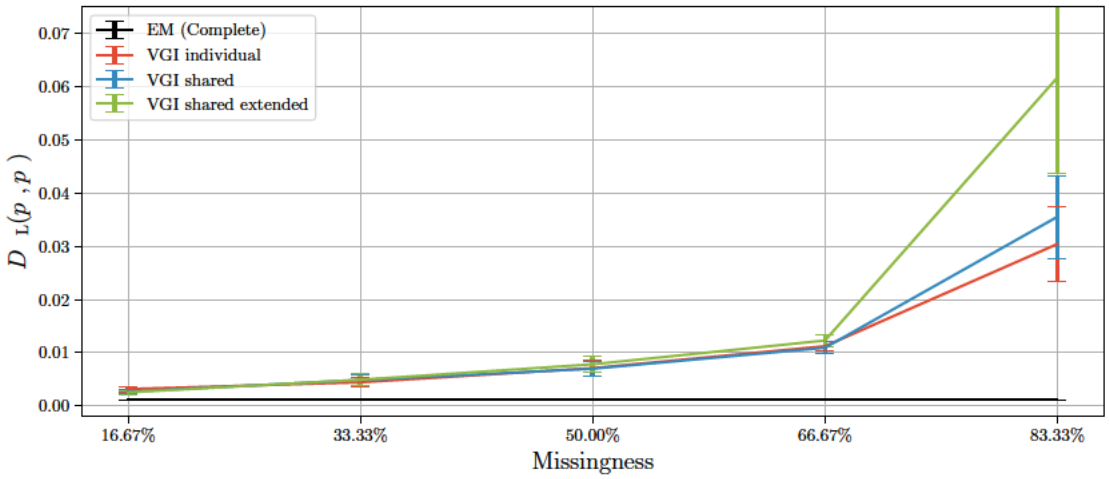


Figure 18: The accuracy of the fitted models on toy data, as measured by $D_{\mathrm{KL}}(p_* \mid\mid p_{\boldsymbol{\theta}})$, comparing independent-weight variational models against shared-weight variational models.

Variational Gibbs Inference

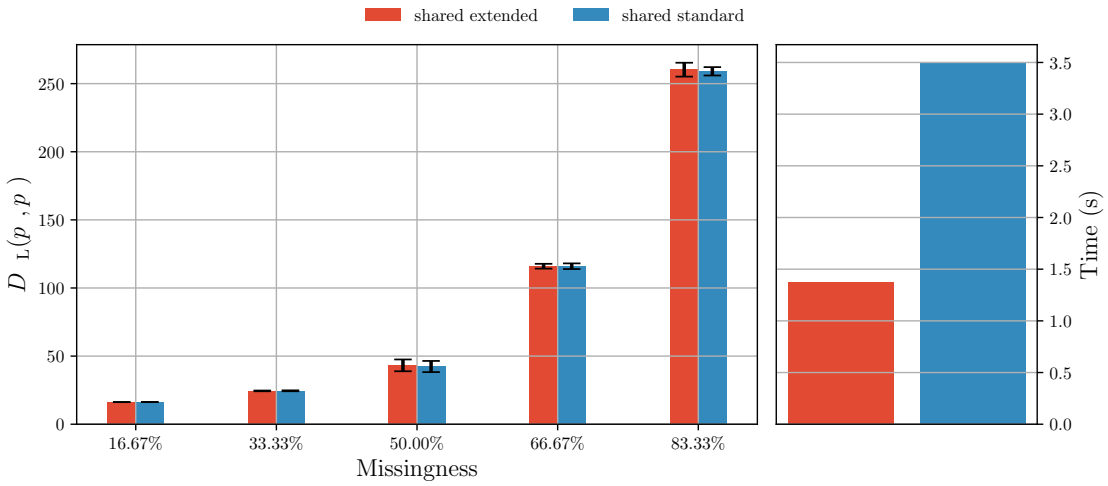


Figure 19: Comparison of VGI estimation accuracy on a FA model using a shared standard variational Gibbs model against a shared extended model, as discussed in Section 3.3. Left: estimated KL divergence between the fitted and the ground truth model. Right: average training time per iteration.

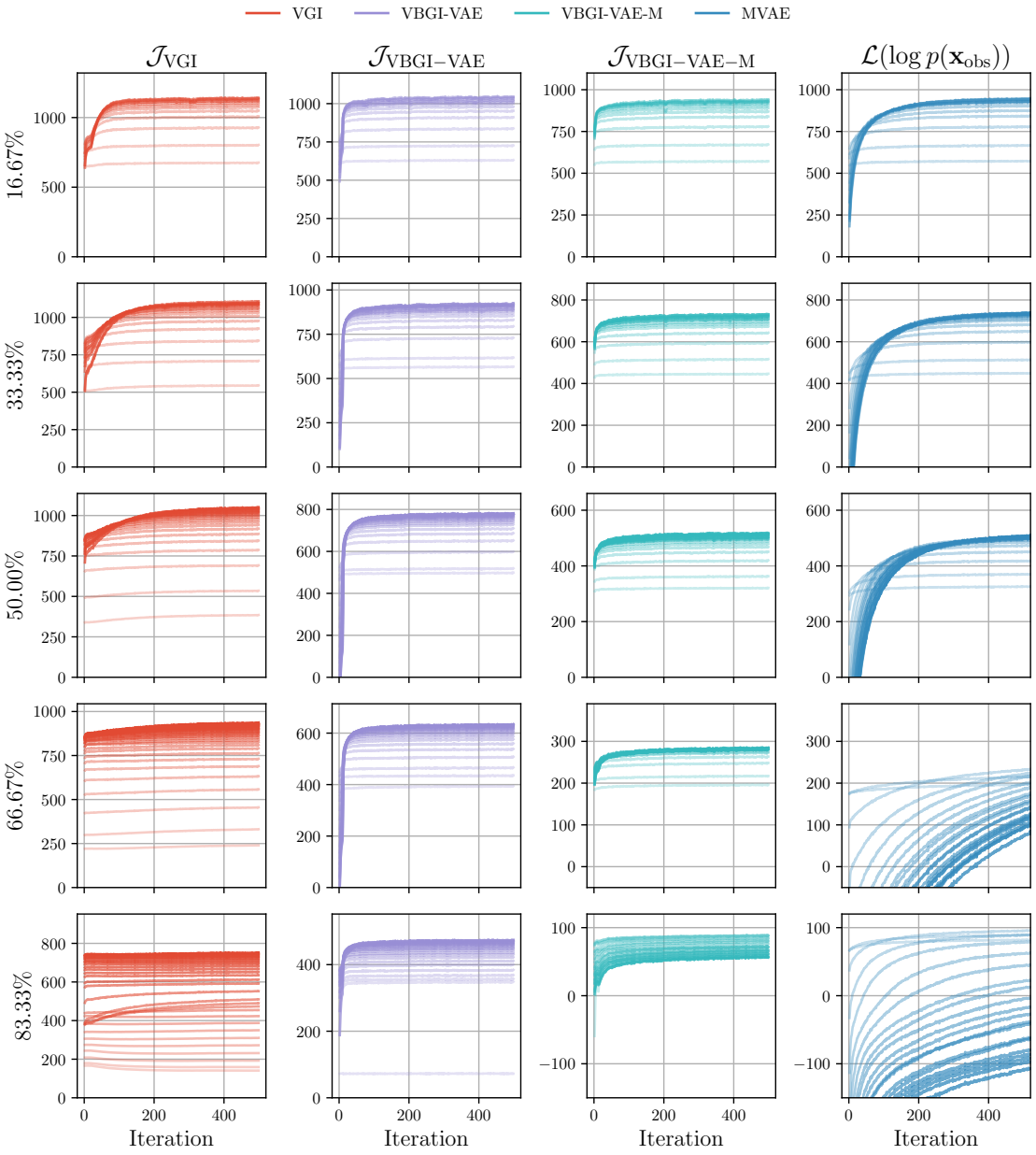


Figure 20: The validation fine-tuning loss curves on VAE-Frey data, where the intensity of the curve represents the training iteration at which the weight snapshot was taken. From an early iteration (less intense) to the last iteration (more intense). The fine-tuning of VGI-based experiments (columns 1-3) is comparatively faster then the VAE-specific methods (right-hand column, the curves of all VAE-specific methods were similar).

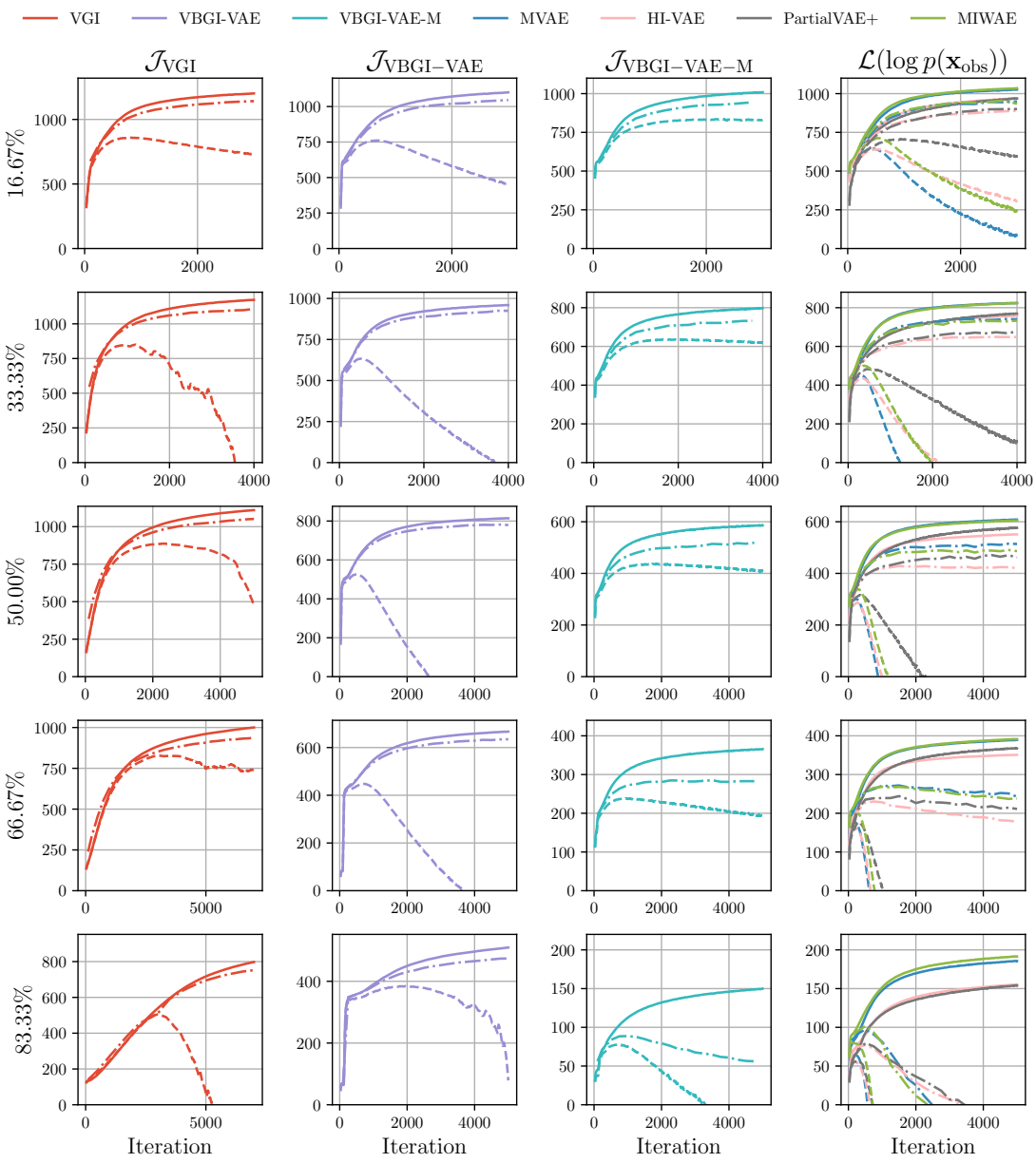


Figure 21: The training (solid), validation (dashed), and fine-tuned validation (dash-dotted) loss curves on VAE-Frey data. The gap between the validation and fine-tuned validation curves shows the inference generalisation gap. Note the difference in scale of the objectives, thus they should not be compared directly. Because of the generalisation gap, the validation curves before fine-tuning (dashed red) diverged for VGI when the number of variational models was large. For visualisation purposes, we thus only accepted Gibbs updates in the hypercube defined by the minimum and maximum values in the observed data.

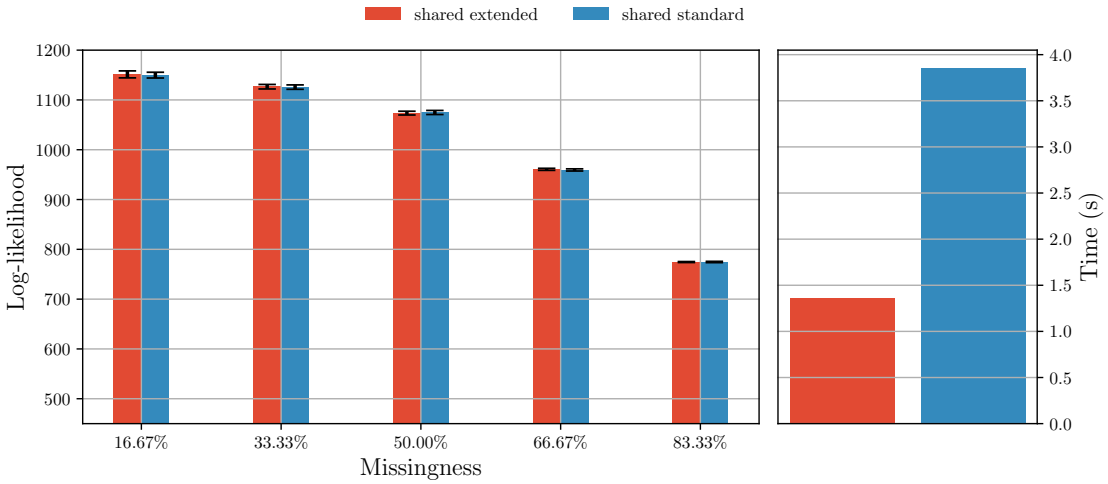


Figure 22: Comparison of VGI estimation on a VAE model using shared standard-Gibbs variational conditionals against shared extended-Gibbs model, as discussed in Section 3.3. Left: estimated test log-likelihood of the fitted model. Right: average training time per iteration.

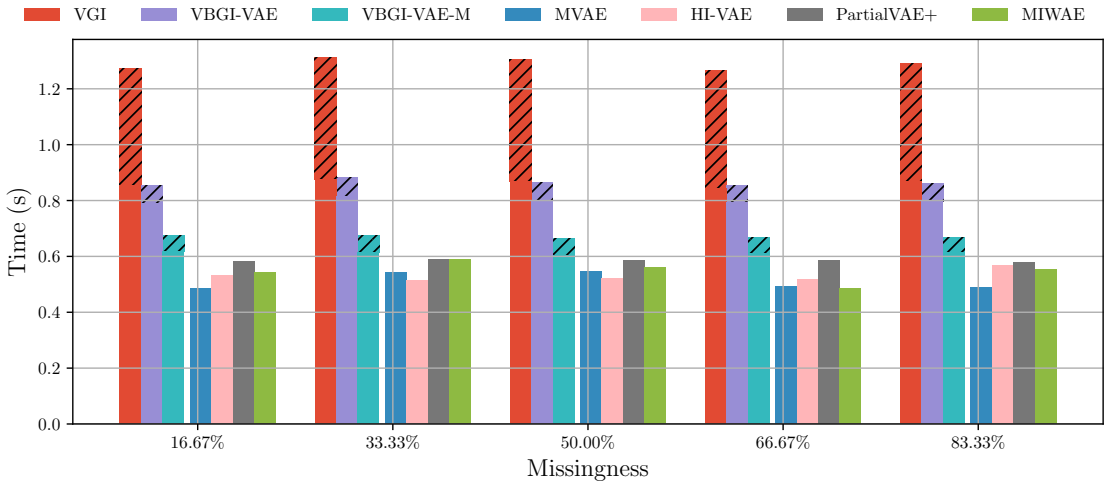


Figure 23: Average time of one training iteration in seconds on VAE-Frey data. The hatched part of the bars indicates the time spent on updating the imputations via pseudo-Gibbs sampling.

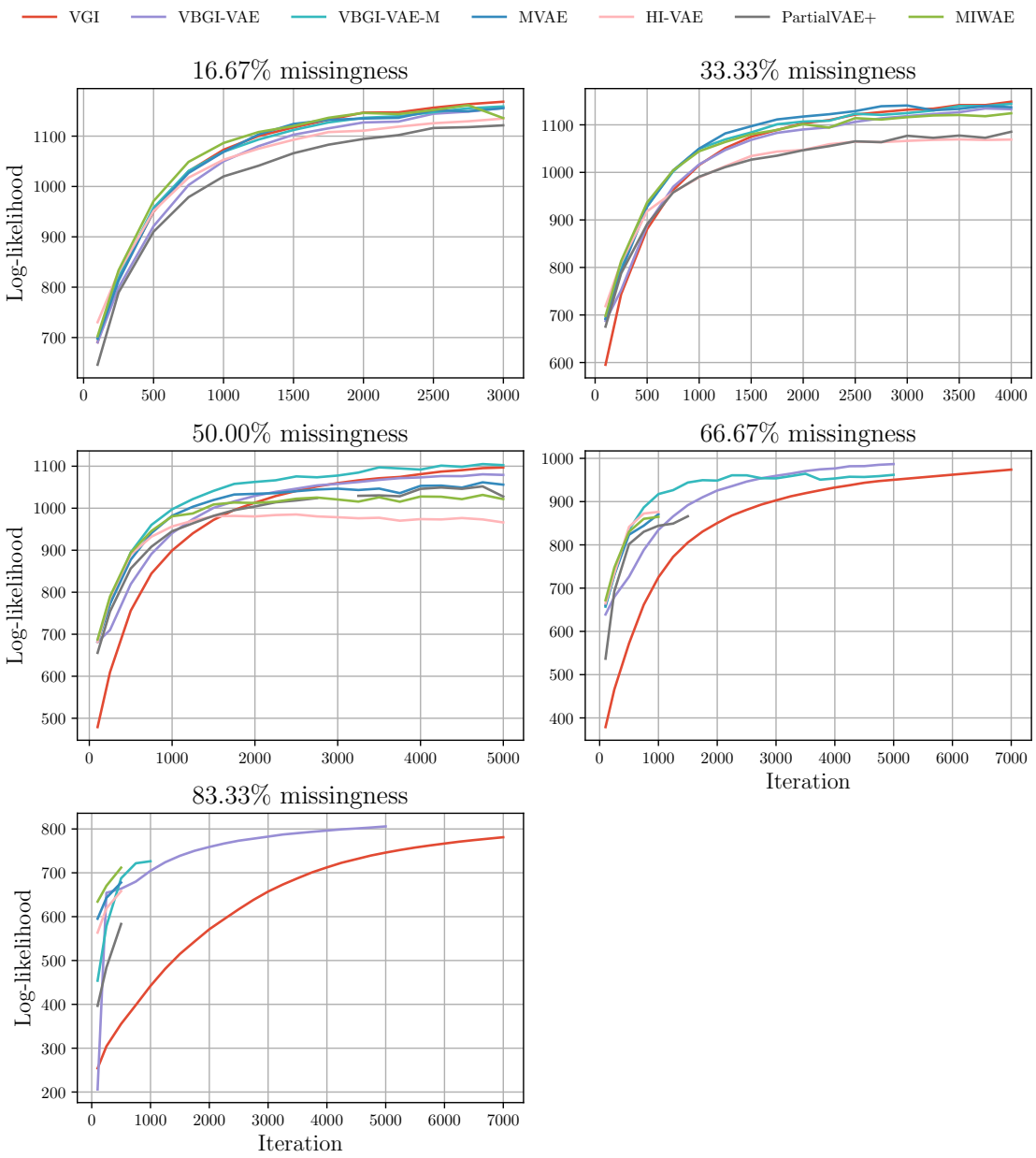


Figure 24: Estimated test log-likelihood against the training iteration on complete VAE-Frey data. Estimated using importance sampling as described in Section 5.5, where the variational encoder was refitted to the test data to mitigate the inference generalisation gap.

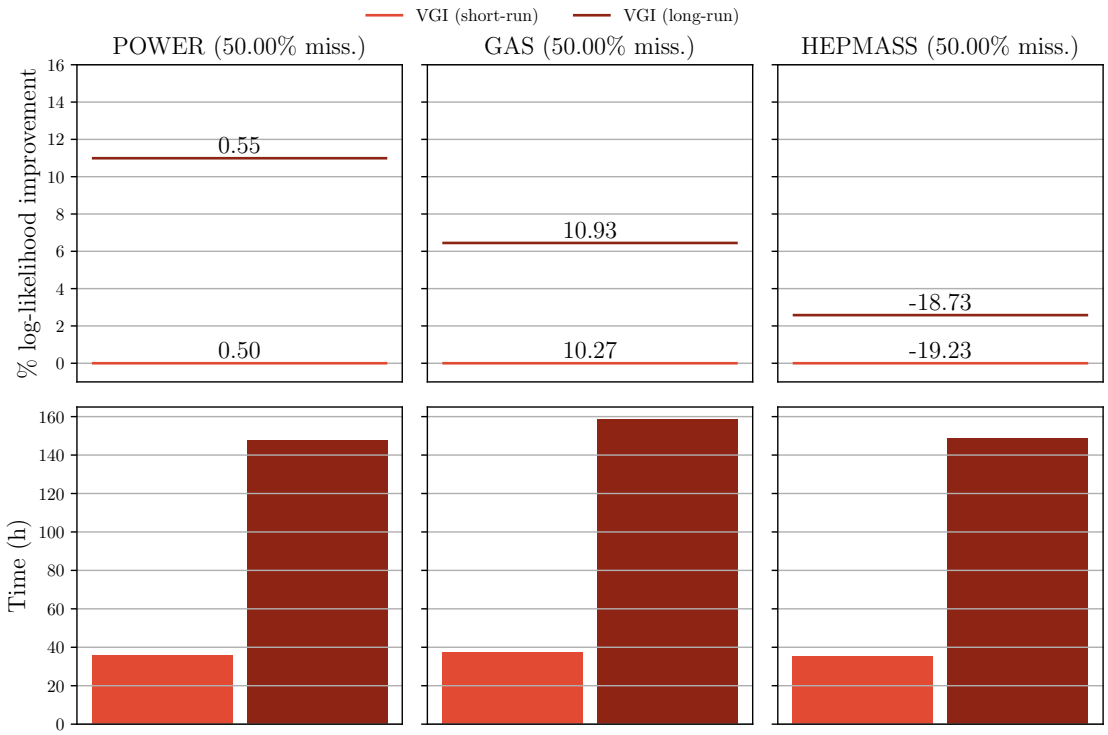


Figure 25: The percentage gain in test log-likelihood by running the VGI significantly longer. The numbers above the horizontal lines show the test log-likelihood, and the bars show the total training time. Other than the number of maximum iterations, the experimental setup was identical to the main experiments. The results show that the model fits in the main evaluation (Section 6.5) can be further improved by using a larger computational budget.

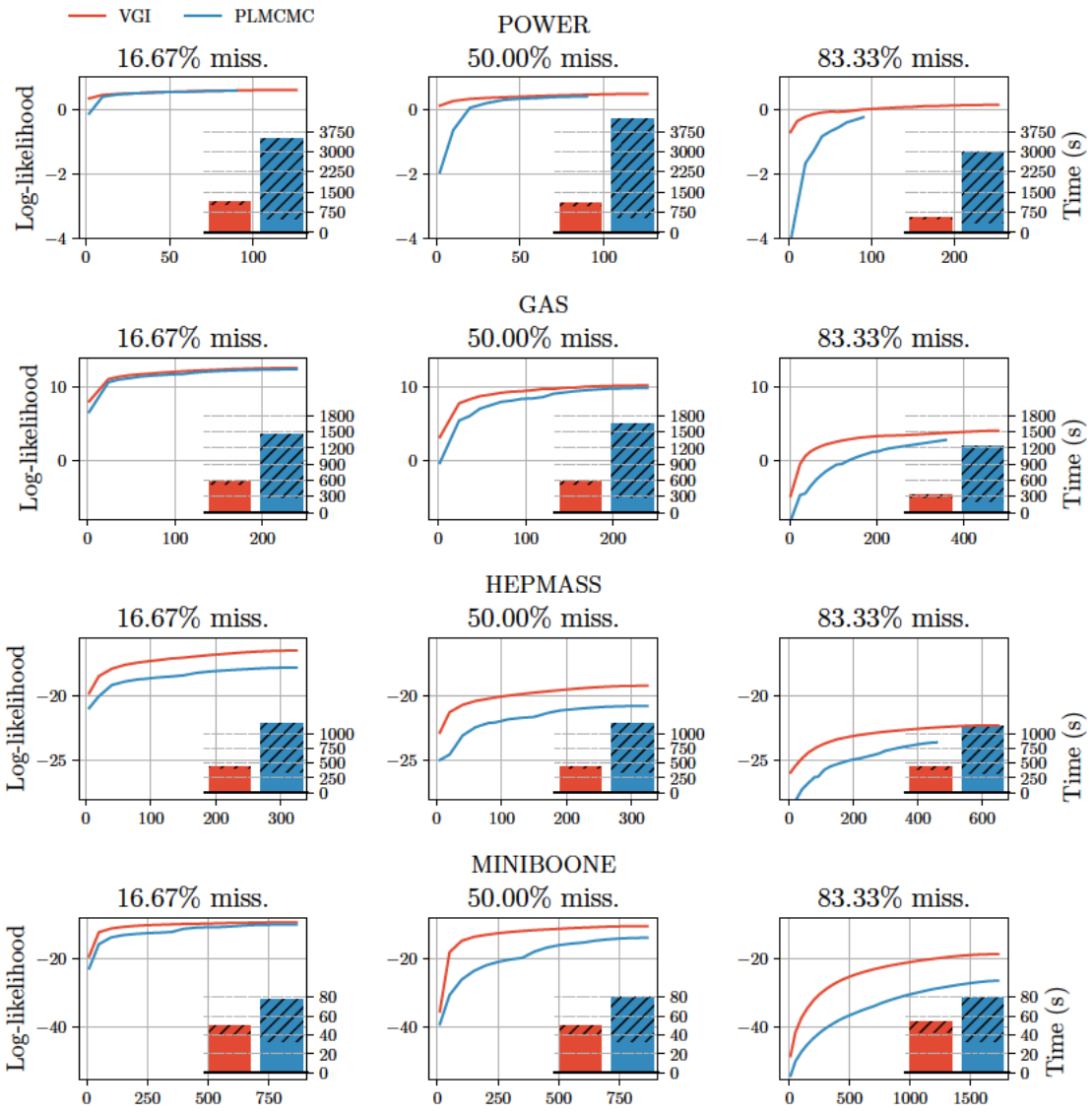


Figure 26: Estimated test log-likelihood on complete data against training iteration and average duration of one training iteration in seconds. The hatched part of the bars indicates the time spent on pseudo-Gibbs sampling in VGI and performing PLMCMC during MCEM. In all of the top row and the third column the PLMCMC log-likelihood curves stop earlier than VGI due to the computational budget, since the average iteration of PLMCMC was significantly more expensive than VGI.

References

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021. ISSN 1566-2535. doi: 10.1016/j.inffus.2021.05.008.

Jaan Altosaar, Rajesh Ranganath, and David M. Blei. Proximity Variational Inference. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*. arXiv, May 2017. doi: 10.48550/arXiv.1705.08931.

Elaine Angelino, Matthew James Johnson, and Ryan P. Adams. Patterns of Scalable Bayesian Inference. *Foundations and Trends in Machine Learning*, 9(2-3):119–247, February 2016. doi: 10.1561/2200000052.

Barry C. Arnold, Enrique Castillo, and Jose Maria Sarabia. *Conditional Specification of Statistical Models*. Springer-Verlag, New York, 1999. ISBN 978-0-387-77500-5. doi: 10.1007/978-0-387-98135-2.

David Barber. *Bayesian Reasoning and Machine Learning*. Cambridge University Press, 2017. ISBN 978-0-511-80477-9. doi: 10.1017/CBO9780511804779.

Jonathan W. Bartlett, Shaun R. Seaman, Ian R. White, and James R. Carpenter. Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Statistical Methods in Medical Research*, 24(4):462–487, August 2015. ISSN 0962-2802. doi: 10.1177/0962280214521348.

Melanie L. Bell, Mallorie Fiero, Nicholas J. Horton, and Chiu-Hsieh Hsu. Handling missing data in RCTs; a review of the top medical journals. *BMC Medical Research Methodology*, 14(1):118, November 2014. ISSN 1471-2288. doi: 10.1186/1471-2288-14-118.

Nathan Bell and Michael Garland. Efficient Sparse Matrix-Vector Multiplication on CUDA. Technical Report VR-2008-004, NVIDIA Corporation, 2008.

Christopher M. Bishop. *Machine Learning and Pattern Recognition*. Springer, March 2006. ISBN 0-387-31073-8.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017. ISSN 1537274X. doi: 10.1080/01621459.2017.1285773.

Jaap P. L. Brand. *Development, Implementation and Evaluation of Multiple Imputation Strategies for the Statistical Analysis of Incomplete Data Sets*. PhD thesis, Erasmus University Rotterdam, 1999.

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance Weighted Autoencoders. In *International Conference on Learning Representations (ICLR)*, San Juan, Puerto Rico, September 2015.

Chris Cannella, Mohammadreza Soltani, and Vahid Tarokh. Projected Latent Markov Chain Monte Carlo: Conditional Sampling of Normalizing Flows. In *International Conference on Learning Representations (ICLR)*, Austria, June 2021.

Cyril Chimisov, Krzysztof Latuszynski, and Gareth Roberts. Adapting The Gibbs Sampler. *arXiv preprints*, 2018.

Mark Collier, Alfredo Nazabal, and Christopher K. I. Williams. VAEs in the Presence of Missing Data. In *Workshop on the Art of Learning with Missing Values (Artemiss) at International Conference on Machine Learning (ICML)*, March 2021.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*, volume 39. Wiley-Interscience, 2006. ISBN 0-471-24195-4. doi: 10.1109/tit.1993.1603955.

Chris Cremer, Xuechen Li, and David Duvenaud. Inference Suboptimality in Variational Autoencoders. In *International Conference on Machine Learning (ICML)*, May 2018.

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977. doi: 10.1111/j.2517-6161.1977.tb01600.x.

Dheeru Dua and Casey Graff. UCI Machine Learning Repository, 2017.

Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural Spline Flows. In *Advances in Neural Information Processing Systems (NeurIPS)*, December 2019.

Yarin Gal. *Uncertainty in Deep Learning*. PhD thesis, University of Cambridge, 2016.

Andrew Gelman, John Carlin, Hal Stern, David Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis, 3rd Edition*. Chapman & Hall/CRC, 2013.

Stuart Geman and Donald Geman. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741, November 1984. doi: 10.1109/TPAMI.1984.4767596.

Samuel J. Gershman and Noah D. Goodman. Amortized Inference in Probabilistic Reasoning. In *Annual Meeting of the Cognitive Science Society*, volume 36, 2014.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, Cambridge, MA, USA, 2016.

Will Grathwohl, Kevin Swersky, Milad Hashemi, David Duvenaud, and Chris J. Maddison. Oops I Took A Gradient: Scalable Sampling for Discrete Distributions. In *International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 3831–3841. PMLR, February 2021.

Song Han, Xingyu Liu, Huizi Mao, Jing Pu, Ardavan Pedram, Mark A. Horowitz, and William J. Dally. EIE: Efficient inference engine on compressed deep neural network. *ACM SIGARCH Computer Architecture News*, 44(3):243–254, June 2016. ISSN 0163-5964. doi: 10.1145/3007787.3001163.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034, 2015. ISBN 9781467383912. doi: 10.1109/ICCV.2015.123.

David Heckerman, David Maxwell Chickering, Christopher Meek, Robert Rounthwaite, and Carl Kadie. Dependency Networks for Inference, Collaborative Filtering, and Data Visualization. *Journal of Machine Learning Research*, 1(Oct):49–75, 2000.

Harold V. Henderson and Shayle R. Searle. On Deriving the Inverse of a Sum of Matrices. *SIAM Review*, 23(1):53–60, 1981. ISSN 0036-1445. doi: 10.1137/1023004.

Herwin Alayn Huillcen-Baca and Flor de Luz Palomino-Valdivia. Efficient Sparse Matrix-Vector Multiplication on GPUs using the CSR Format, Pinned Memory and Overlap Data Transfer. In *International Conference on Electronics, Electrical Engineering and Computing (INTERCON)*, pages 1–4, August 2019. doi: 10.1109/INTERCON.2019.8853624.

Niels Bruun Ipsen, Pierre-Alexandre Mattei, and Jes Frellsen. Not-MIWAE: Deep Generative Modelling with Missing not at Random Data. In *International Conference on Learning Representations (ICLR)*, June 2020.

Pavel Izmailov, Sharad Vikram, Matthew D. Hoffman, and Andrew Gordon Wilson. What Are Bayesian Neural Network Posteriors Really Like? In *International Conference on Machine Learning (ICML)*, pages 4629–4640. PMLR, July 2021.

Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An Introduction to Variational Methods for Graphical Models. *Machine Learning*, 37(2): 183–233, 1999. doi: 10.1023/A:1007665907178.

Diederik P. Kingma and Jimmy Lei Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*, December 2014.

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations (ICLR)*, December 2013.

Krzysztof Łatuszyński, Gareth O. Roberts, and Jeffrey S. Rosenthal. Adaptive Gibbs samplers and related MCMC methods. *Annals of Applied Probability*, 23(1):66–98, 2013. ISSN 10505164. doi: 10.1214/11-AAP806.

Yang Li, Shoaib Akbar, and Junier B. Oliva. Flow Models for Arbitrary Conditional Likelihoods. In *International Conference on Machine Learning (ICML)*, 2020.

Roderick J. A. Little. Missing-Data Adjustments in Large Surveys. *Journal of Business & Economic Statistics*, 6(3):287–296, 1988. ISSN 0735-0015. doi: 10.2307/1391878.

Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data: Second Edition*. Wiley-Interscience, 2002. ISBN 0-471-18386-5.

Urbano Lorenzo-Seva and Joost R. van Ginkel. Multiple Imputation of missing values in exploratory factor analysis of multidimensional scales: Estimating latent trait scores. *Anales de Psicología*, 32(2):596–608, May 2016. ISSN 0212-9728. doi: 10.6018/analesps.32.2.215161.

Chao Ma and Cheng Zhang. Identifiable Generative Models for Missing Not at Random Data Imputation. In *Neural Information Processing Systems (NeurIPS)*, October 2021.

Chao Ma, Sebastian Tschiatschek, Konstantina Palla, José Miguel Hernández-Lobato, Sebastian Nowozin, and Cheng Zhang. EDDI: Efficient dynamic discovery of high-value information with partial VAE. In *International Conference on Machine Learning (ICML)*, pages 7483–7504, 2019. ISBN 9781510886988.

Wesley Maddox, Timur Garipov, Pavel Izmailov, Dmitry Vetrov, and Andrew Gordon Wilson. A Simple Baseline for Bayesian Uncertainty in Deep Learning. In *Neural Information Processing Systems (NeurIPS)*, pages 1–25, Canada, 2019.

Benjamin M. Marlin. *Missing Data Problems in Machine Learning*. PhD thesis, University of Toronto, 2008.

Benjamin M. Marlin, Richard S. Zemel, Sam T. Roweis, and Malcolm Slaney. Recommender systems: Missing data and statistical model estimation. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2686–2691, 2011. ISBN 9781577355120. doi: 10.5591/978-1-57735-516-8/IJCAI11-447.

Pierre-Alexandre Mattei and Jes Frellsen. Refit your Encoder when New Data Comes by. In *Workshop on Bayesian Deep Learning at Neural Information Processing Systems (NeurIPS)*, page 4, Montreal, Canada, 2018.

Pierre-Alexandre Mattei and Jes Frellsen. MIWAE: Deep Generative Modelling and Imputation of Incomplete Data Sets. In *International Conference on Machine Learning (ICML)*, 2019.

Kevin P. Murphy. *Probabilistic Machine Learning: An Introduction*. MIT Press, 2021.

Iain Murray. *Advances in Markov Chain Monte Carlo Methods*. PhD thesis, University College London, 2007.

Alfredo Nazábal, Pablo M. Olmos, Zoubin Ghahramani, and Isabel Valera. Handling Incomplete Heterogeneous Data using VAEs. *Pattern Recognition*, 107, 2020. ISSN 0031-3203. doi: 10.1016/j.patcog.2020.107501.

Radford M. Neal. Suppressing Random Walks in Markov Chain Monte Carlo Using Ordered Overrelaxation. Technical Report 9508, Department of Statistics, University of Toronto, Toronto, Ontario, Canada, June 1995.

Kenneth H. Norwich. *Information, Sensation and Perception*. San Diego: Academic Press, 1993. ISBN 0-12-521890-7.

Art B. Owen. *Monte Carlo Theory, Methods and Examples.* https://artowen.su.domains/mc/, 2013.

Topi Paananen, Juho Piironen, Paul-Christian Bürkner, and Aki Vehtari. Implicitly adaptive importance sampling. *Statistics and Computing*, 31(2):16, March 2021. ISSN 0960-3174, 1573-1375. doi: 10.1007/s11222-020-09982-2.

George Papamakarios, Theo Pavlakou, and Iain Murray. Masked Autoregressive Flow for Density Estimation. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing Flows for Probabilistic Modeling and Inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron J. Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. ISSN 2375-0529. doi: 10.1145/2786984.2786995.

Kaare Brandt Petersen and Michael Syskind Pedersen. The Matrix Cookbook. *Technical University of Denmark*, 16(4):1–16, 2012. ISSN 1611-020X. doi: 10.1017/CBO9780511470943.008.

Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of Adam and beyond. In *International Conference on Learning Representations (ICLR)*, pages 1–23, 2018.

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning (ICML)*, 2015. ISBN 978-1-5108-1058-7.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic Backpropagation and Approximate Inference. In *International Conference on Machine Learning (ICML)*, Beijing, China, 2014.

Donald B. Rubin. Inference and Missing Data. *Biometrika*, 63(3):581, December 1976. doi: 10.2307/2335739.

Donald B. Rubin. A Noniterative Sampling/Importance Resampling Alternative to the Data Augmentation Algorithm for Creating a Few Imputations When Fractions of Missing Information are Modest: The SIR Algorithm. *Journal of the American Statistical Association*, 82(398):543–546, 1987a.

Donald B. Rubin. *Multiple Imputation for Nonresponse in Surveys.* John Wiley & Sons, New York, 1987b. ISBN 0-471-08705-X. doi: 10.2307/3172772.

Donald B. Rubin. Nested multiple imputation of NMES via partially incompatible MCMC. *Statistica Neerlandica*, 57(1):3–18, 2003. ISSN 00390402. doi: 10.1111/1467-9574.00217.

Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv:1609.04747 [cs]*, June 2017.

Tim Salimans, Diederik P. Kingma, and Max Welling. Markov Chain Monte Carlo and Variational Inference: Bridging the Gap. In *International Conference on Machine Learning (ICML)*, pages 1218–1226, 2015.

Joseph L. Schafer. *Analysis of Incomplete Multivariate Data*. Chapman and Hall, New York, 1997. ISBN 0-412-04061-1.

James C. Spall. *Introduction to Stochastic Search and Optimization*. Series in Discrete Mathematics and Optimization. Wiley-Interscience, 2003. doi: 10.1007/978-1-4939-1323-7_1.

Daniel J. Stekhoven and Peter Bühlmann. MissForest - nonparametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, January 2012. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btr597.

Martin Tanner and Wing Hung Wong. The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association*, 82(398):528–540, 1987.

Jos M. F. Ten Berge. Orthogonal procrustes rotation for two or more matrices. *Psychometrika*, 42(2):267–276, June 1977. ISSN 1860-0980. doi: 10.1007/BF02294053.

Alexander Terenin, Daniel Simpson, and David Draper. Asynchronous Gibbs Sampling. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

Tijmen Tieleman. Training restricted Boltzmann machines using approximations to the likelihood gradient. In *International Conference on Machine Learning (ICML)*, pages 1064–1071, 2008. ISBN 9781605582054. doi: 10.1145/1390156.1390290.

Stef van Buuren. *Flexible Imputation of Missing Data*. CRC Press LLC, second edition, 2018. ISBN 978-1-138-58831-8.

Stef van Buuren and Catharina G. M. Oudshoorn. Multivariate Imputation by Chained Equations. Technical report, TNO Prevention and Health, 2000.

Stef van Buuren, Jaap P. L. Brand, Catharina G. M. Groothuis-Oudshoorn, and Donald B. Rubin. Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, 76(12):1049–1064, 2006. ISSN 00949655. doi: 10.1080/10629360600810434.

Joost van Ginkel and Pieter Kroonenberg. Using Generalized Procrustes Analysis for Multiple Imputation in Principal Component Analysis. *Journal of Classification*, 31(2):242–269, 2014.

Larry Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer Texts in Statistics. Springer, 2005. ISBN 978-1-4419-2322-6.

Greg C. G. Wei and Martin A. Tanner. A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms. *Journal of the American Statistical Association*, 85(411):699–704, September 1990. doi: 10.1080/01621459.1990.10474930.

Christopher K. I. Williams, Charlie Nash, and Alfredo Nazábal. Autoencoders and Probabilistic Inference with Missing Data: An Exact Solution for The Factor Analysis Case. *arXiv preprint*, 1801.03851, January 2018.

Laurent Younes. On the convergence of markovian stochastic algorithms with rapidly decreasing ergodicity rates. *Stochastics and Stochastic Reports*, 65(3-4):177–228, February 1999. ISSN 1045-1129. doi: 10.1080/17442509908834179.

Mingtian Zhang, Peter Hayes, and David Barber. Generalization Gap in Amortized Inference. In *Workshop on Bayesian Deep Learning at Neural Information Processing Systems (NeurIPS)*, page 6, 2021.

6.2 Additional discussion

Wake-sleep alternative to the EM. In this chapter, we have proposed a novel method based on the variational EM algorithm for model estimation from incomplete data. The key objective in the paper corresponds to a case of the variational lower-bound in eq. (2.25). It is well known, that optimising this bound corresponds to maximising the marginal log-likelihood and minimising the reverse KL between the variational distribution and the imputation distribution:

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}}) \geq \mathbb{E}_{f(\boldsymbol{x}_{\text{mis}}|\boldsymbol{x}_{\text{obs}})}\left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}}, \boldsymbol{x}_{\text{obs}})}{f(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})}\right] \tag{6.1}$$

$$= \log p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}}) + \mathbb{E}_{f(\boldsymbol{x}_{\text{mis}}|\boldsymbol{x}_{\text{obs}})}\left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})}{f(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})}\right] \tag{6.2}$$

$$= \log p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}}) - D_{\text{KL}}(f(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}) \parallel p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})). \tag{6.3}$$

If the variational distribution is flexible enough then the KL divergence can be minimised to 0 and optimisation of the lower-bound would correspond to directly maximising the marginal log-likelihood $\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}})$. However, the reverse KL objective has a "mode-seeking" behaviour (*e.g.* Murphy, 2021, Section 6.2.6), which could mean that some of the imputation distribution modes may be omitted in the variational approximation, as a result making the KL term non-zero and thus biasing the model estimate by reducing its entropy.

An alternative to the variational EM, originally proposed for latent variable models, is the wake-sleep algorithm (Hinton et al., 1995). The algorithm uses the forward KL divergence $D_{\text{KL}}(p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}) \parallel f(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}))$, instead of the reverse, when optimising the variational distribution, while keeping the standard variational lower-bound in eq. (6.1) for the model parameters $\boldsymbol{\theta}$. The potential advantage of this objective for the variational distribution is that it is "mass-covering" (*e.g.* Murphy, 2021, Section 6.2.6), which means that it attempts to have positive probability mass wherever $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ is non-zero, thus avoiding the "mode-seeking" issues of the reverse KL. Similar to the variational EM case, if the variational distribution is flexible enough, then the KL would be 0 and model estimation would correspond exactly to maximum-likelihood. However, a caveat of the mass-covering objective is that the variational distribution $f(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ may overestimate the support of the true conditional distribution $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$, which in turn would also cause bias to the model estimate by increasing its entropy.

The similarity between the two methods lies in their shared objective of minimising KL divergences between the variational distribution and the model. Therefore, when using the wake-sleep algorithm, similar considerations as those previously discussed for the

variational EM algorithm in this chapter must be taken into account when specifying the variational distribution. Namely, we need to work with flexible variational distributions that can handle arbitrary patterns of missingness (see section 2.2 of our publication), otherwise the KL terms will be high, introducing bias. One way to address this issue is to use the full-conditional specification of the variational distribution in the wake-sleep algorithm, similar to the approach we used in VGI. Or alternatively, we could use the importance-weighted objective (Burda et al., 2015; Bornschein and Bengio, 2015) instead of the standard lower-bound for parameters $\boldsymbol{\theta}$, which is able to transform samples from a simpler variational distribution into the (approximate) samples from the model conditional (Cremer et al., 2017) and thus would potentially allow using less flexible variational distributions and still be able to accurately estimate the model, as long as the computational budget is large.

A further complication with the wake-sleep objective for the variational distribution is that sampling from $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$, needed to approximate the forward KL objective $D_{\text{KL}}(p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}) \parallel f(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}))$, is generally intractable. The standard approach in wake-sleep literature for latent variable models is to jointly sample both $\boldsymbol{x}_{\text{mis}}$ and $\boldsymbol{x}_{\text{obs}}$ from the model, which is generally assumed tractable. This makes optimising the objective feasible but introduces a shift: the variational distribution f is conditioned on samples from the model rather than the true training distribution. This shift can lead to issues like the amortisation generalisation gap (Cremer et al., 2018; Zhang et al., 2021), where the variational distribution may not generalise well to the true training data set after being fitted on model-generated samples. Additionally, in the incomplete data setting, where the split into observed $\boldsymbol{x}_{\text{obs}}$ and missing $\boldsymbol{x}_{\text{mis}}$ parts depends on the unknown missingness mechanism (see section 2.2.1), this approach requires estimating the missingness distribution $p(\boldsymbol{m} \mid \boldsymbol{x})$. Then, to sample $\boldsymbol{x}_{\text{obs}}$ and $\boldsymbol{x}_{\text{mis}}$ one would first sample jointly from $p_{\boldsymbol{\theta}}(\boldsymbol{x})$, followed by sampling a missingness mask from $p(\boldsymbol{m} \mid \boldsymbol{x})$, allowing the split into observed $\boldsymbol{x}_{\text{obs}} = \boldsymbol{x}[\boldsymbol{m}]$ and missing $\boldsymbol{x}_{\text{mis}} = \boldsymbol{x}[1 - \boldsymbol{m}]$ components.

Overall, the wake-sleep algorithm and its variants, such as reweighted wake-sleep (Bornschein and Bengio, 2015), present interesting opportunities for model estimation from incomplete data, particularly as they can help address the entropy underestimation issues associated with variational EM. However, multiple technical challenges may need to be understood and resolved before these methods can be effectively applied to incomplete data.

Arbitrarily-conditional posteriors. As discussed in section 2.2 of our publication, in order to effectively use amortised variational inference for model estimation from incomplete data we need to work with flexible variational distributions that can handle

arbitrary patterns of missingness that may occur in the data set. A recent work on posterior inference in simulator-based models uses a diffusion model with transformer architecture as the conditional posterior which enables flexible modelling of arbitrarily-conditional distributions (Gloeckler et al., 2024). However, unlike the models considered in this chapter, simulator-based models generally do not have a tractable likelihood, and hence the approximate posterior cannot be trained using the variational lower-bound. Instead, the authors of this paper choose to sample the data from the model and then fit the approximate posterior using maximum-likelihood, which is similar to the approach used in the wake-sleep algorithm (Hinton et al., 1995). Hence, applying this method for model estimation from incomplete data encounters similar technical challenges as those discussed above for the wake-sleep algorithm.

Another recent work aims to perform arbitrarily-conditional probabilistic inference on probabilistic programs by modelling the inference task with a language model that takes the probabilistic program along with the observed values as input (Wu and Goodman, 2022). While the method primarily assumes a fully-observed data set that is masked during training, the authors also introduce a fine-tuning procedure based on the variational ELBO, making it applicable to incomplete data sets. To achieve a flexible posterior family, the authors suggest using auto-regressive distributions similar to (Germain et al., 2015). However, this approach is computationally expensive when training with the ELBO, due to the need for autoregressive sampling and backpropagation through each step, making it infeasible to use in high-dimensional high-missingness settings.

In contrast to the approaches discussed above, VGI aims to both fit a target statistical model and perform probabilistic inference on the missing variables, whereas the other methods focus solely on probabilistic inference without model fitting. This makes the VGI setting significantly more challenging, highlighting the importance of computational efficiency. To address this problem effectively, VGI uses flexible variational distributions specified via the variational kernel, needs to amortise only D distributions instead of 2^D, making amortisation more efficient and accurate, and uses a simple, computationally efficient objective to train both the model and the variational distribution. Furthermore, the method does not require learning the missingness mechanism (assuming ignorable missingness), which would be necessary if adapting the approach from (Gloeckler et al., 2024). Yet, incorporating features from these works, such as neural network architectures, into VGI could yield further improvements and thus require an investigation.

Chapter 7

Discussion

The remarkable flexibility of deep models offers promising opportunities to improve data-driven applications across many practical domains. But, many important domains are affected by missing data, where these models face new challenges that must be understood and addressed before their full potential can be unlocked.

Our investigation centred around two popular classes of deep models: variational autoencoders (VAEs) and normalising flows. The focus was on two key statistical tasks essential for the principled handling of incomplete data: missing data imputation and model estimation with missing values in the data. By exploring these tasks within the context of deep models, this thesis aims to further our understanding of the challenges involved and propose mitigating strategies, thereby paving the way for wider adoption of deep models in domains afflicted by missing data.

In the following sections, we summarise the contributions of this thesis and discuss potential future directions.

7.1 Summary of contributions

In chapter 4, we investigated the use of pre-trained VAEs for multiple imputation of missing data, addressing the first research question of this thesis:

Research Question 1: What are the challenges when using pre-trained VAEs for missing data imputation, and how might these challenges be overcome?

We focused this investigation on existing methods based on Gibbs-like procedures that aim to sample the conditional distribution of missing variables from a standard, jointly-specified, VAE model. Using an archetypical scenario, we have shown that their practical efficiency may be limited and identified three pitfalls as the sources of these limitations. These pitfalls stem from the VAEs' inherent tendency, and a common objective, to learn a structured latent space. Drawing upon these pitfalls, we introduced two new sampling methods that improve empirical sampling efficiency for missing data imputation using pre-trained VAEs.

In chapter 5, we proceeded to the task of VAE estimation from incomplete data, focusing on settings where simplifying assumptions enable efficient marginalisation of the missing variables. Hence, tackling the second research question:

Research Question 2: What are the factors that contribute to the increased complexity of VAE estimation from incomplete data, and what strategies can be used to effectively address this additional complexity due to missing data?

To this end, we have shown that marginalisation of the missing variables, while theoretically recommended (section 2.2.6), made the subsequent variational inference task harder compared to the fully-observed data case. The complexity arises from the model's posterior distribution over the latent variables, which tends to be more complicated in the presence of incomplete-data compared to fully-observed data scenarios. Consequently, fitting VAEs from incomplete data may require more flexible variational distributions in order to well-match the complicated incomplete-data posteriors.

However, the choice of the variational distribution for a VAE model is often motivated by certain inductive biases that induce desired properties into the model. This means that by modifying the variational distribution to handle the complexities due to missingness we may lose these desired inductive biases. This finding highlights a potential trade-off between the need for more flexible variational distributions to accommodate incomplete data and the importance of preserving the intended inductive biases of the VAE model. Addressing this, we showed that the model posterior over the latents with incomplete-data is a mixture of fully-observed data posteriors. Using this observation we introduced two new estimation methods based on variational-mixtures that improve VAE estimation from incomplete data while enabling the re-use of the variational families from the fully-observed data case.

In chapter 6, we turned to the more challenging task of estimating general deep statistical models from incomplete data where simplifying assumptions such as efficient marginalisation of the missing variables may not be applicable. Thus, addressing the final research question:

Research Question 3: How do we efficiently estimate general statistical models for which simplifying assumptions, like marginalisation of the missing variables, may not be applicable?

As discussed in section 2.2.6, the expectation-maximisation (EM) algorithm and its variants provide a principled framework for model estimation from incomplete data when

such simplifying assumptions do not hold. But the application of this family of methods has remained relatively under-explored in the context of deep models with missing data. To address this gap, we proposed a new general-purpose method for model estimation from incomplete data, called variational Gibbs inference (VGI). VGI extends the variational EM framework by introducing a novel variational approximation of the missing data distribution. Specifically, VGI approximates the missing data distribution using a Markov chain with a learnable variational kernel that closely resembles the kernel of a Gibbs sampler. In doing so, VGI reduces an exponential growth (with respect to data dimensionality) of variational distributions for missing data into a linear growth, requiring only one conditional variational distribution per data dimension, while also enabling the use of flexible variational families thus ensuring that the model can be optimised effectively.

7.2 Outlook

A common thread throughout this thesis has been the task of conditional sampling from deep statistical models. This task was explored either as a primary goal for missing data imputation, as in chapter 4, or as an integral component of the model estimation procedures, as seen in chapters 5 and 6. For conditional sampling methods to be effective in the context of missing data imputation, they must be able to effectively adapt to the target distribution of each incomplete data point. Additionally, to be effective for model estimation, these conditional sampling techniques must be able to adapt to the evolving target distribution as the model undergoes optimisation. Considering these requirements, to what extent do our proposed methods achieve this adaptability, and what future directions are promising to further improve this area?

The two proposed methods in chapter 4 are per-data-point adaptive, as they iteratively adapt their proposal distributions to the target distribution for each data-point. However, their efficiency may depend on manually tuned *global* hyperparameters balancing exploration and exploitation, suggesting that these methods may be further improved by automatically selecting the hyperparameters on a *per-data-point* basis. Moreover, similar to existing approaches, these proposed methods re-use the encoder distribution from VAE training process allowing them to be (partially) adaptive to the changing target distribution when used for model estimation. On the other hand, the imputation distribution in chapter 6 is directly adapted to fit the evolving target distribution via a learnable variational kernel. However, the variational Gibbs kernel uses a fixed randomised scan strategy, which may limit the mixing of the Markov chains, as different strategies may potentially work better for different data-points. This suggests, that VGI can be further improved us-

ing techniques from the adaptive-Gibbs sampling literature (*e.g.* Neal, 1995; Łatuszyński et al., 2013; Chimisov et al., 2018; Grathwohl et al., 2021), which aim to adapt the scan strategy, to boost the method's adaptation to individual data-points. Overall, in chapters 4 and 6 we proposed promising methods for conditional sampling of jointly-specified models in various settings that partially fulfil the two key criteria of adaptability and we believe that they may further benefit from investigation of the discussed directions.

A further interesting research direction in the context of imputation, that we did not explore in this thesis, are gradient-based samplers, such as Hamiltonian Monte Carlo (HMC, Duane et al., 1987; Neal, 2011) or Metropolis-adjusted Langevin algorithm (MALA, Roberts and Tweedie, 1996; Dwivedi et al., 2019). These gradient-based methods may be able to adapt to both individual data-points and the changing target distributions as they use the local (gradient) information of the (changing) model to construct an implicit proposal for any target distribution.

Another core topic of this thesis was the estimation of deep statistical models from incomplete data. We investigated this topic in the context of VAEs (chapter 5) and in the context of more general deep statistical models (chapter 6). While the proposed methods in these chapters show promising results, there is still room for improvement to facilitate wider adoption of deep models in domains affected by missing data. Potential areas for improvement include increasing the estimation accuracy, improving computational efficiency, and simplifying the ease-of-use of these methods.

Both chapters 5 and 6 used tools from the variational inference literature to efficiently and accurately estimate deep models from incomplete data. In chapter 5, variational inference was used to approximate the incomplete data posterior of the latent variables using amortised mixture distributions. Meanwhile, in chapter 6, it was used to approximate the conditional distributions of the missing variables using a Gibbs Markov chain. Yet, the variational inference literature offers a diverse array of tools, each with their own advantages and disadvantages (Zhang et al., 2018), and this thesis has only explored a small subset of these tools. Hence, exploring alternative tools from the variational inference literature is an important future direction. For example, exploring alternative variational objectives may improve posterior mode coverage (Bornschein and Bengio, 2015; Zhang et al., 2018; Ambrogioni et al., 2018; Wan et al., 2020), compared to the typically-used reverse KL divergence, as a result potentially further improving the accuracy of model estimation. By expanding the exploration of variational inference techniques beyond those covered in this thesis, there is an opportunity to leverage the strengths of different approaches, potentially yielding further improvements in accuracy, efficiency, and flexibility for estimating deep statistical models from incomplete data.

Another potential avenue for improvement may involve extending the VGI method introduced in chapter 6 by incorporating a Metropolis–Hastings (MH) correction step (Metropolis et al., 1953; Hastings, 1970; Gelman and Rubin, 1992). This would ensure that the imputation Gibbs chains asymptotically converge to the correct conditional distribution of missing variables, even in the presence of variational approximation errors, potentially improving the method's estimation accuracy and ease-of-use by preventing Markov chain divergence. To ensure efficient MH-corrected sampling, the variational objective in VGI may be modified to additionally (implicitly) maximise the probability of proposal acceptance in the MH step (*e.g.* Thin et al., 2020). By incorporating such an MH step and optimising for high proposal acceptance rates, we could potentially further improve VGI's robustness, accuracy, and usability for general deep statistical model estimation from incomplete data.

Furthermore, probabilistic programming tools (Salvatier et al., 2016; Carpenter et al., 2017; Bingham et al., 2019) are being actively developed to provide ease-of-use of probabilistic inference and estimation techniques, eliminating the need for often complicated implementation processes. The generality of the VGI method proposed in chapter 6 positions it well for implementation in such probabilistic programming tools. We believe that by abstracting away implementation complexities, these tools can make principled techniques like VGI more accessible and user-friendly, potentially accelerating their real-world impact and driving wider adoption in fields dealing with missing data challenges.

Finally, while our contributions focused primarily on two classes of deep models, namely variational autoencoders and normalising flows, recently diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021) and flow-matching models (Lipman et al., 2023; Tong et al., 2023) have gathered substantial attention due to their impressive modelling capabilities in various domains. Yet, applying these models in domains with missing data is challenging. Conditional sampling is generally intractable and requires working in very high-dimensional latent spaces. And, while the missing variables may be tractably marginalised, the effects of the marginalisation on the estimated model and the estimation procedures are not thoroughly studied. We hence believe that further analysis of diffusion and flow-matching models is needed, similar to the analysis we conducted in this thesis for variational autoencoders and normalising flows, in order to better understand their behaviour in the presence of incomplete data sets and further expand their applicability to real-world scenarios with missing data challenges.

Altogether, we believe that deep statistical models hold a significant promise across many domains. However, a key prerequisite to unlocking their full potential is addressing the challenges that arise when working with incomplete data. While further research is nec-

essary to fully overcome these capabilities, this thesis provides important insights into the core challenges faced by deep models in the presence of missing data. The proposed methods demonstrate how using these insights can lead to improved performance when applying deep statistical models in incomplete data settings. Therefore, the findings and techniques presented in this thesis represent a step towards enabling a broader adoption of deep models across domains affected by data missingness, unlocking powerful data-driven capabilities.

References

Luca Ambrogioni, Umut Güçlü, Yağmur Güçlütürk, Max Hinne, Marcel A. J. van Gerven, and Eric Maris. Wasserstein Variational Inference. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. Cited on page 183.

Barry C. Arnold, Enrique Castillo, and Jose Maria Sarabia. *Conditional Specification of Statistical Models*. Springer-Verlag, New York, 1999. Cited on page 26.

David Barber. *Bayesian Reasoning and Machine Learning*. Cambridge University Press, 2017. arXiv: 1011.1669v3. Cited on pages 24, 27.

Matthew J. Beal and Zoubin Ghahramani. The Variational Bayesian EM Algorithm for Incomplete Data: With Application to Scoring Graphical Model Structures. In *Bayesian Statistics 7: Proceedings of the 7th Valencia International Meeting, June 2-6, 2002*, 2003. Cited on pages 24, 33.

Edgar A. Bernal. Training Deep Normalizing Flow Models in Highly Incomplete Data Scenarios with Prior Regularization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Apr. 2021. arXiv: 2104.01482 [cs]. Cited on page 33.

Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, and Noah D. Goodman. Pyro: Deep Universal Probabilistic Programming. *Journal of Machine Learning Research*, 2019. Cited on page 184.

Jörg Bornschein and Yoshua Bengio. Reweighted Wake-Sleep. In *International Conference on Learning Representations (ICLR)*, Apr. 2015. arXiv: 1406.2751 [cs]. Cited on pages 178, 183.

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance Weighted Autoencoders. In *International Conference on Learning Representations (ICLR)*, San Juan, Puerto Rico, Sept. 2015. arXiv: 1509.00519. Cited on pages 14, 32, 178.

Chris Cannella, Mohammadreza Soltani, and Vahid Tarokh. Projected Latent Markov Chain Monte Carlo: Conditional Sampling of Normalizing Flows. In *International Conference on Learning Representations (ICLR)*, Austria, June 2021. arXiv: 2007.06140. Cited on pages 28, 33.

Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus A. Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A

Probabilistic Programming Language. *Journal of Statistical Software*, 2017. Cited on page 184.

Cyril Chimisov, Krzysztof Latuszynski, and Gareth Roberts. Adapting The Gibbs Sampler. *arXiv preprints*, 2018. arXiv: 1801.09299. Cited on page 183.

Nicolas Chopin and Omiros Papaspiliopoulos. *An Introduction to Sequential Monte Carlo*. Springer Series in Statistics. Springer, 2020. Cited on page 27.

Mark Collier, Alfredo Nazabal, and Christopher K. I. Williams. VAEs in the Presence of Missing Data. In *Workshop on the Art of Learning with Missing Values (Artemiss) at International Conference on Machine Learning (ICML)*, Mar. 2021. arXiv: 2006.05301. Cited on page 32.

Chris Cremer, Xuechen Li, and David Duvenaud. Inference Suboptimality in Variational Autoencoders. In *International Conference on Machine Learning (ICML)*, May 2018. arXiv: 1801.03558. Cited on page 178.

Chris Cremer, Quaid Morris, and David Duvenaud. Reinterpreting Importance-Weighted Autoencoders. In *ICLR Workshop*, Feb. 2017. Cited on pages 14, 178.

Biwei Dai and Uros Seljak. Translation and Rotation Equivariant Normalizing Flow (TRENF) for Optimal Cosmological Analysis. *Monthly Notices of the Royal Astronomical Society*, Sept. 2022. arXiv: 2202.05282 [astro-ph]. Cited on page 16.

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 1977. Cited on pages 11, 23, 196.

Laurent Dinh, David Krueger, and Yoshua Bengio. NICE: Non-linear Independent Components Estimation. In *ICLR Workshop*, Apr. 2015. arXiv: 1410.8516. Cited on page 15.

Simon Duane, Anthony D. Kennedy, Brian J. Pendleton, and Duncan Roweth. Hybrid Monte Carlo. *Physics Letters B*, Sept. 1987. Cited on page 183.

Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural Spline Flows. In *Advances in Neural Information Processing Systems (NeurIPS)*, Dec. 2019. arXiv: 1906.04032. Cited on page 15.

Raaz Dwivedi, Yuansi Chen, Martin J. Wainwright, and Bin Yu. Log-Concave Sampling: Metropolis-Hastings Algorithms Are Fast. *Journal of Machine Learning Research*, 2019. Cited on page 183.

Garrett Fitzmaurice, Marie Davidian, Geert Verbeke, and Geert Molenberghs, editors. *Longitudinal Data Analysis*. Chapman and Hall/CRC, New York, Aug. 2008. Cited on page 17.

Andrew Gelman and Donald B. Rubin. Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, Nov. 1992. Cited on page 184.

Stuart Geman and Donald Geman. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Nov. 1984. Cited on page 28.

Mathieu Germain, Karol Gregor, Iain Murray, and Hugo Larochelle. MADE: Masked Autoencoder for Distribution Estimation. In *International Conference on Machine Learning (ICML)*, 2015. arXiv: 1502.03509. Cited on page 179.

Samuel J. Gershman and Noah D. Goodman. Amortized Inference in Probabilistic Reasoning. In *Annual Meeting of the Cognitive Science Society*, 2014. Cited on page 15.

Manuel Gloeckler, Michael Deistler, Christian Weilbach, Frank Wood, and Jakob H. Macke. All-in-One Simulation-Based Inference. In *International Conference on Machine Learning (ICML)*, Apr. 2024. arXiv: 2404.09636 [cs, stat]. Cited on page 179.

Robert J. Glynn, Nan M. Laird, and Donald B. Rubin. Selection Modeling Versus Mixture Modeling with Nonignorable Nonresponse. In Howard Wainer, editor, *Drawing Inferences from Self-Selected Samples*. Springer, New York, NY, 1986. Cited on page 17.

Rafael Gómez-Bombarelli, Jennifer N. Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D. Hirzel, Ryan P. Adams, and Alán Aspuru-Guzik. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Central Science*, Feb. 2018. Cited on page 15.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, Cambridge, MA, USA, 2016. Cited on page 13.

Will Grathwohl, Kevin Swersky, Milad Hashemi, David Duvenaud, and Chris J. Maddison. Oops I Took A Gradient: Scalable Sampling for Discrete Distributions. In *International Conference on Machine Learning (ICML)*, Feb. 2021. Cited on page 183.

Ryan-Rhys Griffiths and José Miguel Hernández-Lobato. Constrained Bayesian Optimization for Automatic Chemical Design Using Variational Autoencoders. *Chemical Science*, Jan. 2020. Cited on page 25.

Ishaan Gulrajani, Kundan Kumar, Faruk Ahmed, Adrien Ali Taiga, Francesco Visin, David Vazquez, and Aaron Courville. PixelVAE: A Latent Variable Model for Natural

Images. In *International Conference on Learning Representations (ICLR)*, Nov. 2016. arXiv: 1611.05013. Cited on page 34.

Wilfred Keith Hastings. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, 1970. JSTOR: 2334940. Cited on pages 28, 184.

James Heckman. The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models. NBER Chapters, National Bureau of Economic Research, Inc, 1976. Cited on page 17.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohammed, and Alexander Lerchner. β-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *International Conference on Learning Representations (ICLR)*, 2017. Cited on page 15.

Geoffrey E. Hinton, Peter Dayan, Brendan J. Frey, and Radford M. Neal. The "Wake-Sleep" Algorithm for Unsupervised Neural Networks. *Science (New York, N.Y.)*, May 1995. PMID: 7761831. Cited on pages 177, 179.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *Neural Information Processing Systems (NeurIPS)*, Dec. 2020. arXiv: 2006.11239 [cs, stat]. Cited on page 184.

Oleg Ivanov, Michael Figurnov, and Dmitry Vetrov. Variational Autoencoder with Arbitrary Conditioning. In *International Conference on Learning Representations (ICLR)*, 2019. arXiv: 1806.02382. Cited on page 27.

Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An Introduction to Variational Methods for Graphical Models. *Machine Learning*, 1999. Cited on pages 14, 103.

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations (ICLR)*, Dec. 2013. arXiv: 1312.6114. Cited on page 14.

Krzysztof Łatuszyński, Gareth O. Roberts, and Jeffrey S. Rosenthal. Adaptive Gibbs Samplers and Related MCMC Methods. *Annals of Applied Probability*, 2013. arXiv: 1101.5838. Cited on page 183.

Steven Cheng-Xian Li, Benjamin M. Marlin, and Bo Jiang. MisGAN: Learning from Incomplete Data with Generative Adversarial Networks. In *International Conference on Learning Representations (ICLR)*, 2019. arXiv: 1902.09599v1. Cited on page 29.

Yang Li, Shoaib Akbar, and Junier B. Oliva. Flow Models for Arbitrary Conditional Likelihoods. In *International Conference on Machine Learning (ICML)*, 2020. arXiv: 1909.06319. Cited on page 27.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow Matching for Generative Modeling. In *International Conference on Learning Representations (ICLR)*, Feb. 2023. arXiv: 2210.02747 [cs, stat]. Cited on page 184.

Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data: Third Edition*. Wiley Series in Probability and Statistics. Wiley, Hoboken, NJ, 3rd edition edition, 2020. Cited on pages 11, 20, 21.

James Lucas, George Tucker, Roger Grosse, and Mohammad Norouzi. Understanding Posterior Collapse in Generative Latent Variable Models. In *Workshop on Deep Generative Models for Highly Structured Data*, International Conference on Learning Representations (ICLR), Mar. 2019. Cited on page 72.

Chao Ma, Sebastian Tschiatschek, Konstantina Palla, José Miguel Hernández-Lobato, Sebastian Nowozin, and Cheng Zhang. EDDI: Efficient Dynamic Discovery of High-Value Information with Partial VAE. In *International Conference on Machine Learning (ICML)*, 2019. arXiv: 1809.11142. Cited on page 31.

Chao Ma, Sebastian Tschiatschek, Richard Turner, José Miguel Hernández-Lobato, and Cheng Zhang. VAEM: A Deep Generative Model for Heterogeneous Mixed Type Data. In *Neural Information Processing Systems (NeurIPS)*, 2020. Cited on page 32.

Chao Ma and Cheng Zhang. Identifiable Generative Models for Missing Not at Random Data Imputation. In *Advances in Neural Information Processing Systems (NeurIPS)*, Oct. 2021. arXiv: 2110.14708. Cited on page 19.

Pierre-Alexandre Mattei and Jes Frellsen. Leveraging the Exact Likelihood of Deep Latent Variable Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, Feb. 2018. arXiv: 1802.04826. Cited on page 28.

Pierre-Alexandre Mattei and Jes Frellsen. MIWAE: Deep Generative Modelling and Imputation of Incomplete Data Sets. In *International Conference on Machine Learning (ICML)*, 2019. Cited on page 32.

Geoffrey J. McLachlan and Thriyambakam Krishnan. *The EM Algorithm and Extensions*. Wiley Series in Probability and Statistics. 2nd edition, 2007. Cited on page 24.

Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, June 1953. Cited on pages 28, 184.

Ning Miao, Emile Mathieu, N. Siddharth, Yee Whye Teh, and Tom Rainforth. On Incorporating Inductive Biases into VAEs. In *International Conference on Learning Representations (ICLR)*, Feb. 2022. arXiv: 2106.13746 [cs, stat]. Cited on page 15.

Karthika Mohan, Judea Pearl, and Jin Tian. Graphical Models for Inference with Missing Data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2013. Cited on page 19.

Kevin P. Murphy. *Probabilistic Machine Learning: Advanced Topics*. Adaptive Computation and Machine Learning Series. The MIT Press, Cambridge, Massachusetts, 2023. Cited on pages 13, 16.

Kevin P. Murphy. *Probabilistic Machine Learning: An Introduction*. MIT Press, 2021. Cited on pages 15, 22, 177.

Razieh Nabi, Rohit Bhattacharya, and Ilya Shpitser. Full Law Identification In Graphical Models Of Missing Data: Completeness Results. In *International Conference on Machine Learning (ICML)*, 2020. arXiv: 2004.04872. Cited on page 19.

Alfredo Nazábal, Pablo M. Olmos, Zoubin Ghahramani, and Isabel Valera. Handling Incomplete Heterogeneous Data Using VAEs. *Pattern Recognition*, 2020. arXiv: 1807.03653. Cited on page 32.

Radford M. Neal. MCMC Using Hamiltonian Dynamics. In *Handbook of Markov Chain Monte Carlo*. Chapman & Hall / CRC Press, May 2011. arXiv: 1206.1901 [physics, stat]. Cited on page 183.

Radford M. Neal. Suppressing Random Walks in Markov Chain Monte Carlo Using Ordered Overrelaxation. Technical Report, Department of Statistics, University of Toronto, Toronto, Ontario, Canada, June 1995. arXiv: bayes-an/9506004. Cited on page 183.

Radford M. Neal and Geoffrey E. Hinton. A View of the EM Algorithm That Justifies Incremental, Sparse, and Other Variants. In Michael I. Jordan, editor, *Learning in Graphical Models*. Kluwer, 1998. Cited on page 23.

Art B. Owen. *Monte Carlo Theory, Methods and Examples*. https://artowen.su.domains/mc/, 2013. Cited on page 28.

Topi Paananen, Juho Piironen, Paul-Christian Bürkner, and Aki Vehtari. Implicitly Adaptive Importance Sampling. *Statistics and Computing*, Mar. 2021. Cited on page 28.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing Flows for Probabilistic Modeling and Inference. *Journal of Machine Learning Research*, 2021. arXiv: 1912.02762. Cited on page 15.

Ignacio Peis, Chao Ma, and José Miguel Hernández-Lobato. Missing Data Imputation and Acquisition with Deep Hierarchical Models and Hamiltonian Monte Carlo. In *Neural Information Processing Systems (NeurIPS)*, Sept. 2022. arXiv: 2202.04599 [cs, stat]. Cited on page 32.

Nikša Praljak, Xinran Lian, Rama Ranganathan, and Andrew L. Ferguson. ProtWave-VAE: Integrating Autoregressive Sampling with Latent-Based Inference for Data-Driven Protein Design. *ACS Synthetic Biology*, Dec. 2023. Cited on page 34.

Tom Rainforth, Adam R. Kosiorek, Tuan Anh Le, Chris J. Maddison, Maximilian Igl, Frank Wood, and Yee Whye Teh. Tighter Variational Bounds Are Not Necessarily Better. In *International Conference on Machine Learning (ICML)*, Mar. 2019. arXiv: 1802.04537. Cited on page 14.

Danilo Jimenez Rezende and Shakir Mohamed. Variational Inference with Normalizing Flows. In *International Conference on Machine Learning (ICML)*, 2015. arXiv: 1505.05770. Cited on page 15.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic Backpropagation and Approximate Inference. In *International Conference on Machine Learning (ICML)*, Beijing, China, 2014. arXiv: 1401.4082v3. Cited on pages 14, 28.

Trevor W. Richardson, Wencheng Wu, Lei Lin, Beilei Xu, and Edgar A. Bernal. MCFlow: Monte Carlo Flow Models for Data Imputation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, Mar. 2020. arXiv: 2003.12628. Cited on page 33.

Gareth O. Roberts and Richard L. Tweedie. Exponential Convergence of Langevin Distributions and Their Discrete Approximations. *Bernoulli*, 1996. JSTOR: 3318418. Cited on page 183.

Donald B. Rubin. Inference and Missing Data. *Biometrika*, Dec. 1976. JSTOR: 2335739. Cited on page 16.

Donald B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, New York, 1987. Cited on pages 11, 21.

John Salvatier, Thomas V. Wiecki, and Christopher Fonnesbeck. Probabilistic Programming in Python Using PyMC3. *PeerJ Computer Science*, Apr. 2016. Cited on page 184.

Victor Garcia Satorras, Emiel Hoogeboom, Fabian B. Fuchs, Ingmar Posner, and Max Welling. E(n) Equivariant Normalizing Flows. In *Neural Information Processing Systems (NeurIPS)*, June 2021. arXiv: 2105.09016. Cited on page 16.

Shaun Seaman, John Galati, Dan Jackson, and John Carlin. What Is Meant by "Missing at Random"? *Statistical Science*, May 2013. Cited on pages 17–20.

Vaidotas Simkus and Michael U. Gutmann. Conditional Sampling of Variational Autoencoders via Iterated Approximate Ancestral Sampling. *Transactions on Machine Learning Research*, 2023. Cited on pages 12, 36.

Vaidotas Simkus and Michael U. Gutmann. Improving Variational Autoencoder Estimation from Incomplete Data with Mixture Variational Families. *Transactions on Machine Learning Research*, 2024. Cited on pages 12, 73.

Vaidotas Simkus, Benjamin Rhodes, and Michael U. Gutmann. Variational Gibbs Inference for Statistical Model Estimation from Incomplete Data. *Journal of Machine Learning Research*, 2023. Cited on pages 12, 104.

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep Unsupervised Learning Using Nonequilibrium Thermodynamics. In *International Conference on Machine Learning (ICML)*, Nov. 2015. arXiv: 1503.03585 [cond-mat, q-bio, stat]. Cited on page 184.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations (ICLR)*, Feb. 2021. arXiv: 2011.13456. Cited on page 184.

James C. Spall. *Introduction to Stochastic Search and Optimization*. Series in Discrete Mathematics and Optimization. Wiley-Interscience, 2003. Cited on page 22.

Ryan R. Strauss and Junier B. Oliva. Posterior Matching for Arbitrary Conditioning. In *Neural Information Processing Systems (NeurIPS)*, Nov. 2022. arXiv: 2201.12414 [cs, stat]. Cited on page 29.

Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. CSDI: Conditional Score-based Diffusion Models for Probabilistic Time Series Imputation. In *Neural Information Processing Systems (NeurIPS)*, Oct. 2021. arXiv: 2107.03502 [cs, stat]. Cited on pages 26, 27.

Achille Thin, Nikita Kotelevskii, Jean-Stanislas Denain, Leo Grinsztajn, Alain Durmus, Maxim Panov, and Eric Moulines. MetFlow: A New Efficient Method for Bridging the Gap between Markov Chain Monte Carlo and Variational Inference, 2020. arXiv: 2002.12253. Cited on page 184.

Alexander Tong, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Kilian Fatras, Guy Wolf, and Yoshua Bengio. Improving and Generalizing Flow-Based Generative Models with Minibatch Optimal Transport. *Transactions on Machine Learning Research*, Oct. 2023. arXiv: 2302.00482 [cs]. Cited on page 184.

Stef van Buuren. *Flexible Imputation of Missing Data*. CRC Press LLC, 2nd edition, 2018. Cited on page 20.

Ramakrishna Vedantam, Ian Fischer, Jonathan Huang, and Kevin P. Murphy. Generative Models of Visually Grounded Imagination. *International Conference on Learning Representations (ICLR)*, May 2017. arXiv: 1705.10762. Cited on page 31.

Neng Wan, Dapeng Li, and Naira Hovakimyan. F-Divergence Variational Inference. In *Neural Information Processing Systems (NeurIPS)*, 2020. Cited on page 183.

Yixin Wang, David Blei, and John P. Cunningham. Posterior Collapse and Latent Variable Non-identifiability. In *Neural Information Processing Systems (NeurIPS)*, 2021. Cited on page 72.

Greg C. G. Wei and Martin A. Tanner. A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms. *Journal of the American Statistical Association*, Sept. 1990. Cited on pages 24, 33, 34, 103.

Jay Whang, Erik Lindgren, and Alex Dimakis. Composing Normalizing Flows for Inverse Problems. In *International Conference on Machine Learning (ICML)*, July 2021. Cited on page 29.

Christopher K. I. Williams, Charlie Nash, and Alfredo Nazábal. Autoencoders and Probabilistic Inference with Missing Data: An Exact Solution for The Factor Analysis Case. *arXiv preprint*, Jan. 2018. arXiv: 1801.03851. Cited on page 31.

Michael J. Williams, John Veitch, and Chris Messenger. Nested Sampling with Normalising Flows for Gravitational-Wave Inference. *Physical Review D*, May 2021. arXiv: 2102.11056. Cited on page 16.

Luhuan Wu, Brian L. Trippe, Christian A. Naesseth, David M. Blei, and John P. Cunningham. Practical and Asymptotically Exact Conditional Sampling in Diffusion Models. In *Neural Information Processing Systems (NeurIPS)*, June 2023. arXiv: 2306.17775 [cs, q-bio, stat]. Cited on page 28.

Mike Wu and Noah Goodman. Foundation Posteriors for Approximate Probabilistic Inference. In *Neural Information Processing Systems (NeurIPS)*, May 2022. Cited on page 179.

Mike Wu and Noah D. Goodman. Multimodal Generative Models for Scalable Weakly-Supervised Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, Feb. 2018. arXiv: 1802.05335. Cited on page 31.

Jinsung Yoon, James Jordon, and Mihaela van der Schaar. GAIN: Missing Data Imputation Using Generative Adversarial Nets. In *International Conference of Machine Learning (ICML)*, June 2018. arXiv: 1806.02920. Cited on pages 27, 29.

Cheng Zhang, Judith Butepage, Hedvig Kjellstrom, and Stephan Mandt. Advances in Variational Inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. arXiv: 1711.05597. Cited on page 183.

Mingtian Zhang, Peter Hayes, and David Barber. Generalization Gap in Amortized Inference. In *Workshop on Bayesian Deep Learning at Neural Information Processing Systems (NeurIPS)*, 2021. Cited on page 178.

Shengjia Zhao, Jiaming Song, and Stefano Ermon. InfoVAE: Information Maximizing Variational Autoencoders. In *Neural Information Processing Systems (NeurIPS)*, May 2018. arXiv: 1706.02262. Cited on page 72.

Appendices

A Link between EM with minimal M-step and SGA

Here, we show that the expectation-maximisation algorithm (EM, Dempster et al., 1977) with a minimal M-step, also known as generalised EM, performing a single stochastic gradient ascent (SGA) step on the evidence lower bound (ELBO), is equivalent to SGA on the marginal log-likelihood $\ell(\boldsymbol{\theta})$ in eq. (2.22). This equivalence provides insight into the relationship between EM-based approaches and direct optimisation of the marginal log-likelihood in the context of incomplete data.

We define the ELBO for incomplete data using eq. (2.25):

$$\mathcal{L}(\boldsymbol{\theta}, \{f^i\}_{i=1}^N) = \frac{1}{N}\sum_{i=1}^N \mathbb{E}_{f^i(\boldsymbol{x}^i_{\text{mis}}|\boldsymbol{x}^i_{\text{obs}})}\left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}^i_{\text{obs}}, \boldsymbol{x}^i_{\text{mis}})}{f^i(\boldsymbol{x}^i_{\text{mis}} \mid \boldsymbol{x}^i_{\text{obs}})}\right] \leq \ell(\boldsymbol{\theta}), \tag{A.1}$$

where f^i in the l.h.s. is a short-hand for $f^i(\boldsymbol{x}^i_{\text{mis}} \mid \boldsymbol{x}^i_{\text{obs}})$. Starting with some initial parameters $\boldsymbol{\theta}^t$ at iteration $t = 0$, the algorithm iterates between the E- and M-steps (see section 2.2.6), producing a sequence of parameters $\boldsymbol{\theta}^0, \ldots, \boldsymbol{\theta}^T$, until it reaches a (local) optimum of the ELBO.

The E-step at iteration t obtains $f^i(\boldsymbol{x}^i_{\text{mis}} \mid \boldsymbol{x}^i_{\text{obs}}) = p_{\boldsymbol{\theta}^t}(\boldsymbol{x}^i_{\text{mis}} \mid \boldsymbol{x}^i_{\text{obs}})$ for all i. Then, the M-step seeks to maximise the ELBO w.r.t. $\boldsymbol{\theta}$, keeping the $\{f^i\}_{i=1}^N$ fixed. Instead of fully maximising the ELBO, which can often be intractable, we here take one step of gradient ascent on the ELBO, with the gradient defined as:

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}}\, \mathcal{L}(\boldsymbol{\theta}, \{f^i\}_{i=1}^N)\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^t} &= \nabla_{\boldsymbol{\theta}}\, \frac{1}{N}\sum_{i=1}^N \mathbb{E}_{f^i(\boldsymbol{x}^i_{\text{mis}}|\boldsymbol{x}^i_{\text{obs}})}\left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}^i_{\text{obs}}, \boldsymbol{x}^i_{\text{mis}})}{f^i(\boldsymbol{x}^i_{\text{mis}} \mid \boldsymbol{x}^i_{\text{obs}})}\right]\Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^t} && \text{(A.2)}\\
&= \nabla_{\boldsymbol{\theta}}\, \frac{1}{N}\sum_{i=1}^N \mathbb{E}_{f^i(\boldsymbol{x}^i_{\text{mis}}|\boldsymbol{x}^i_{\text{obs}})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}^i_{\text{obs}})\right]\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^t} && \text{(A.3)}\\
&= \nabla_{\boldsymbol{\theta}}\, \frac{1}{N}\sum_{i=1}^N \log p_{\boldsymbol{\theta}}(\boldsymbol{x}^i_{\text{obs}})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^t} && \text{(A.4)}\\
&= \nabla_{\boldsymbol{\theta}}\, \ell(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^t}\,, && \text{(A.5)}
\end{aligned}$$

where we obtain the second line by inserting $f^i(\boldsymbol{x}^i_{\text{mis}} \mid \boldsymbol{x}^i_{\text{obs}}) = p_{\boldsymbol{\theta}^t}(\boldsymbol{x}^i_{\text{mis}} \mid \boldsymbol{x}^i_{\text{obs}})$ from the E-step. This means that taking the gradient of the ELBO $\mathcal{L}(\boldsymbol{\theta}, \{f^i\}_{i=1}^N)$ w.r.t. $\boldsymbol{\theta}$ at the current parameter value $\boldsymbol{\theta}^t$ corresponds to taking the gradient of the marginal log-likelihood $\ell(\boldsymbol{\theta})$. This is due to the EM performing a "local" marginalisation of the

missing variables at the current parameter value $\boldsymbol{\theta}^t$. Hence, starting from the same initial parameters $\boldsymbol{\theta}^0$, the above generalised EM algorithm and gradient ascent on the marginal log-likelihood $\ell(\boldsymbol{\theta})$ would yield the same parameter iterates.

Importantly, to fit the parameters $\boldsymbol{\theta}$, the EM algorithm does not require directly evaluating the integral in eq. (2.24) to marginalise the model $p_{\boldsymbol{\theta}}(\boldsymbol{x})$, which enables the construction of many practical methods for parameter estimation from incomplete data.

Moreover, the above equivalence also trivially extends to the SGA case for the optimisation of the marginal log-likelihood $\ell(\boldsymbol{\theta})$ and the ELBO, where the sum over all N data-points would be replaced by a sum over a mini-batch of samples from the data set at any iteration t.