
THE UNIVERSITY
*of* EDINBURGH

# Efficient Probabilistic Inversion of Geophysical Data

**Muhammad Atif Nawaz**

Thesis presented for the degree of

**Doctor of Philosophy**

in

**Geology and Geophysics**


THE UNIVERSITY
*of* EDINBURGH

**2019**

School of GeoSciences, The University of Edinburgh
Scotland, United Kingdom

**Ph.D. Thesis**

*School of GeoSciences*
*The University of Edinburgh*

**Muhammad Atif Nawaz**

*M.Sc. Geophysics*
*Quaid-i-Azam University*

**Supervisors:**

**Prof. Andrew Curtis**

*Professor of Mathematical Geoscience*
*The University of Edinburgh*

**Dr. Mark Chapman**

*Reader in Rock Physics*
*The University of Edinburgh*

**Examiners:**

**Prof. Kathy Whaler**

*Professor of Geophysics*
*The University of Edinburgh*

**Prof. Klaus Mosegaard**

*Deputy Head of the Niels Bohr Institute*
*University of Copenhagen*

*To my family*

# Declaration

I declare that this thesis has been composed solely by myself and that it has not been submitted, either in whole or in part, in any previous application for a degree. Except where otherwise acknowledged, the work presented is entirely my own.

Part of this research has been either published, accepted for publication or is submitted in peer-reviewed journals as: Nawaz & Curtis (2017), Nawaz & Curtis (2018), Nawaz & Curtis (2019), and Nawaz *et al*. (2019).

Muhammad Atif Nawaz

10-Feb-2019

# Abstract

Estimation of uncertainties is critical for subsequent decision making in all applications of geosciences such as geological hazard analysis and risk mitigation, management and exploitation of subsurface resources, and environmental waste disposal. More efficient probabilistic inversion methods in geosciences are vital to making rapid and improved predictions of geological hazards and estimation of subsurface resources from geophysical data, and estimation of associated uncertainties. While this thesis focuses on seismic data inversion for the estimation of geological properties, the methods developed may find a wide variety of applications in all fields of research that involve spatial data analysis.

New concepts, models and methods are developed to perform more efficient probabilistic inversion by making use of the latest developments in machine learning and Bayesian inverse theory to solve geophysical inverse problems. The major contribution of this thesis is the development of efficient geostatistical inversion methods for approximate inference for structured inverse problems where probabilistic dependence between unknown model parameters may be expressed as a *Markov random field* (MRF). These methods are many orders of magnitude faster than the corresponding sampling based methods in such types of inverse problems. Further, some of the commonly used but avoidable assumptions in conventional geostatistical inversion methods are progressively relaxed and finally removed in this research. The faster inversion methods allow more complex models to be evaluated for more accurate predictions and improved estimation of uncertainty for given compute power and time.

Most existing geostatistical inversion methods are based on the localized likelihoods assumption, whereby the seismic data at a location are assumed to depend on the geology only at that location. Such an assumption is unrealistic because of imperfect seismic data acquisition and processing, and fundamental limitations of seismic imaging methods. It is also assumed in most such previous research that the data are completely free of any correlated noise or errors. Although these requirements are almost never met in reality, existing methods use these assumptions to make solutions computationally tractable. Both of these assumptions are progressively removed in this thesis while still allowing computationally tractable solutions to be found for suitably structured problems. The class of problems considered here spans a broad range of spatial data analysis and geosciences, where geology

at a location is assumed to depend directly only on the geology within some pre-specified neighbourhood of that location – the so called *Markovian assumption* – which is the core assumption across the entire literature of geostatistics and has been proven to be valid for all practical purposes.

Exact Bayesian inference is intractable in most models of practical interest because it requires normalization of the posterior distribution by integrating model parameters over a very high dimensional space. Therefore, approximate inference is used in practice. Stochastic sampling (e.g., by using *Markov-chain Monte Carlo* – McMC) is the most commonly used approximate inference method but is computationally expensive and detection of its convergence is often based on subjective criteria and hence is unreliable. New Bayesian inversion methods are introduced that estimate the spatial distribution of geological properties from attributes of seismic data, by showing how the usual probabilistic inverse problem can be solved using an optimization framework while still providing full probabilistic results – the so called *variational inference* approach. The intractable posterior distribution is replaced by a tractable approximation in the variational approach. Inference can then be performed using the approximate distribution in an optimization framework, thus circumventing the need for sampling, while still providing probabilistic results.

The methods developed in this thesis infer the post-inversion (posterior) probability density of the unknown model parameters from seismic data and geological prior information. These methods are shown to be robust against weak prior information and correlated noise in the data. The methods are computationally efficient, and are expected to be applicable to 3D models of realistic size on modern computers without incurring any significant computational limitations.

# Lay Summary

Earthquakes bring disasters. But, they are also a “blessing in disguise” since the waves generated by an earthquake, called seismic waves, carry useful information about the medium they pass through (or reflect from). This information satiates our appetite for knowledge about what is inside of our home planet, the Earth. Such waves may also be produced artificially on a much smaller scale to investigate the rocks below the surface of the Earth. Just like an echo reflects back from a cliff, seismic waves reflect back from boundaries between different rock layers; in fact, the type of seismic waves that are most useful to us are the same as the sound waves in air. We can make images (called seismic images) of the underground rock structures by recording and digitally processing these reflected waves.

Seismic images are used for identifying hidden resources such as oil, gas and minerals in the Earth’s crust. The challenge with seismic images is that these resources are often too deep and appear as blurred in these images, just like a distant object in a photographic image appears to be blurred and may not be easily identifiable. So extracting useful information from seismic images involves uncertainty about the exact location and depth or the desired resources, and challenges such as determining whether extraction of the desired subsurface resources would be economically feasible. Such uncertainties are also involved in other uses of seismic waves. For example, seismic waves are also studied to predict earthquakes and to avoid or mitigate associated hazards. Assessment of uncertainty is critical for making decision based on the seismic images, but it is computationally very expensive. As a result, uncertainties are often ignored which costs time, money and risks.

Efficient methods are developed in this thesis for assessment of uncertainties in the analysis of seismic images. These methods are many times faster than the commonly used previous methods. Also, many previous methods commonly use some unrealistic assumptions to limit their computational complexity. Such assumptions are removed in this thesis. So, besides the advantage of computational efficiency, this research may also provide substantial improvement in terms of quality of results for better subsequent decision making regarding exploitation of subsurface resources, and regarding natural hazard prediction and mitigation.

# Acknowledgements

This thesis is the result of my long aspiration to become a scientific researcher and the last four years of intense hard work. First of all, I would like to express my heartiest gratitude to the God, *Allah (SWT)* the Almighty for all of His blessings and for giving me the strength to complete this research successfully.

I am profoundly grateful to my supervisor, *Prof. Andrew Curtis*, for providing me the opportunity to work on this Ph.D. project, and for his invaluable scientific advice and support during the course of this research. I feel honoured to have worked under the supervision of one of the leading scientists in the field of geophysics. His astute scientific approach has left deep imprints on my understanding of science and my passion for research.

I would also like to express my sincere gratitude to my examiners *Prof. Kathy Whaler* and *Prof. Klaus Mosegaard* for their thorough review of this thesis and insightful feedback that has significantly improved this thesis. Most of the work presented in this thesis is either published or is under review in peer-reviewed scientific journals. I would like to express my gratitude to *Prof. Klaus Mosegaard*, *Prof. Malcolm Sambridge*, *Prof. Henning Omre*, and the many anonymous reviewers whose comments have enhanced the quality of this document.

I am also grateful to *TOTAL UK* for initiating and sponsoring this project, and approving publications that resulted from this research. In particular, I would like to thank *Dr. Mohammad Shahraeeni*, *Dr. Constantin Gerea* and *Benoit Paternoster* of TOTAL for providing me the data used in this project, and for their constructive suggestions and feedbacks on this research.

I am sincerely grateful to the staff at the *School of Geosciences* and at the *College of Science and Engineering* in the *University of Edinburgh*. In particular, I would like to mention *Ross Taylor* for helping me regarding digital computing resources, and *Katy Cameron*, *Dawn Sives*, *Lisa Guenther*, and *Lucy Wall* for being supportive throughout my time here and especially during the submission of this thesis.

I am also thankful to my colleagues: *Dr. Carlos da Costa Filho*, *Dr. Satyan Singh*, *Dr. Erica Galetti*, *Dr. Melody Runge*, *Xin Zhang*, *Stephanie Earp*, *Dominic Cummings*, *Angus Lomas*, and my friends: *Asif*, *Qamar*, *Ahmed*, *Ammad*, *Ejaz*, *Waqas*, *Farhan* and many others for a wonderful time we have had together.

I am also grateful to all of my teachers throughout my life for every piece of knowledge I learnt from them. Also, I am very grateful to my managers during my job career who helped me significantly in my career progression and helped me develop my scientific and technical skills. I would like to mention especially *Robert Engelman*, *Ram Sunder*, *William H. Borland*, *Lukas Wihardjo* and *Vivian Pistre*.

I would like to express my heartfelt gratitude to my dear family: my grandparents, parents, brothers, sister, in-laws, uncles, aunts, and cousins who have always made my life so pleasant. Special thanks to my (late) maternal grandfather, *Muhammad Sharif*, who has always been a source of inspiration and guidance for me. I miss him every day of my life. Also, special thanks to my parents, brothers and sister for their support through my life and during my PhD. Also, thanks to my brother *Saad* and his family for visiting us in Edinburgh and making our holidays so much fun.

My wife, *Sadaf*, is undoubtedly a major reason for the existence of this thesis. My heartiest thanks to my better-half for convincing me to pursue my dream of diving into the ocean of scientific endeavours, and for her support throughout my Ph.D.

Last but definitely not the least, I would like to express by deepest gratitude to my lovely children: *Eshaal*, *Hooria*, and *Rayyan* (the latter two were born during my Ph.D.) for being patient with my busy schedule, and for making my life so lively and colourful.

# Table of Contents

# Notation and Abbreviations

Notation and abbreviations that are repeatedly used throughout this thesis are given below for convenience.

### Notation

Some of the commonly used notation is listed in the table below.

General:

| Symbol | Description |
|---|---|
| **i.e.** | reads "*that is*" or "in other words" |
| **e.g.** | reads "for example" |
| ≡ | reads "is identical to" or "is defined to be equal to" |
| ≅ | reads "is approximately equal to" |

A linear index denoted by lower case letters such as $i$ and $j$ to define the locations (or cells) in our model. Sets are represented with italic, regular (non-boldface) capital (English or Greek) letters, e.g., $\mathcal{V}$ and $\mathcal{G}$. Boldface font with lower case (English or Greek) letters is used for vectors, e.g., $\boldsymbol{r}$ or $\boldsymbol{\beta}$, and upper case letters is used for matrices, e.g., $\boldsymbol{R}$. The identity matrix is represented as $\boldsymbol{I}$. A superscript $T$ stands for transpose of a vector or matrix. The notation '$\mathcal{O}$' known as 'big-O' is used to describe computational complexity of an algorithm.

In an iterative algorithm, bracketed superscripts indicate an estimate of a quantity at the iteration number specified in brackets during the course of an iterative update, e.g., $\theta^{(l)}$ represents an estimate of some quantity $\theta$ after $l$ iterations of an iterative algorithm. In the context of supervised machine learning, bracketed superscripts indicate an index over training examples, e.g. $\boldsymbol{m}^{(i)}$ represents the $i^{\text{th}}$ instance of a quantity $\boldsymbol{m}$.

A hat, or caret, over a parameter (or random variable) denotes its estimator, e.g., $\hat{\theta}$ represents an estimator of $\theta$. The left-arrow symbol '←', e.g. in '$x \leftarrow f(x)$' denotes the 'assignment' (or the 'update') operation which means that "the value of $x$ on the left is update from its old value using the expression $f(x)$ on the right-hand-side (RHS) of the left-arrow".

Some of the commonly used notation related to the set theory and probability theory are listed in the tables below.

Set Theory:

| Symbol | Description |
|---|---|
| : | reads “such that” |
| $\subset$ | reads “is a proper subset of” |
| $\subseteq$ | reads “is an improper subset of”, i.e. “is a subset of or is equal to” |
| $\in$ | reads “is a member of” or “belongs to” |
| $\notin$ | reads “is not a member of” or “does not belongs to” |
| $\lvert\cdot\rvert$ | Refers to “cardinality” (the number of elements) of a set |
| $\forall$ | reads “for all” |

Probability Theory:

| Symbol | Description |
|---|---|
| $\sim$ | reads “is distributed as” |
| $\boldsymbol{\mathcal{P}(x)}$ | Probability of $x$ |
| $\boldsymbol{\mathcal{P}(x;\theta)}$ | Probability of $x$ parameterized by fixed parameters $\Theta$ |
| $\boldsymbol{\mathcal{P}(x\lvert y)}$ | Conditional probability of $x$ given $y$ |
| $\boldsymbol{\mathcal{P}(x,y)}$ | Joint probability of $x$ and $y$ |
| $\boldsymbol{\mathcal{P}(x\lvert y;\theta)}$ | Conditional probability of $x$ given $\boldsymbol{y}$ parameterized by fixed parameters $\Theta$ |
| $\boldsymbol{\mathcal{P}(x,y;\theta)}$ | Joint probability of $x$ and $\boldsymbol{y}$ parameterized by fixed parameters $\Theta$ |
| $\boldsymbol{f(\theta;x)}$ | Probability density function (PDF) $f$ for a random variable $x$ as a function of fixed or random parameters $\Theta$. |
| $\boldsymbol{f(\theta;x\lvert y)}$ | Probability density function (PDF) $f$ for a random variable $x$ given another random variable $y$ as a function of fixed or random parameters $\Theta$. |

Note that the notations $f(\Theta;x)$ and $\mathcal{P}(x;\Theta)$ are equivalent if $x\sim f(\Theta)$ and $\Theta$ are fixed parameters (not random). Thus the semicolon in $\mathcal{P}(x;\Theta)$ emphasizes that this should not be confused with the joint probability $\mathcal{P}(x,\Theta)$, or the conditional probability $\mathcal{P}(x|\Theta)$ of $x$ given $\Theta$, when $\Theta$ is a random variable.

## Abbreviations

Below is a list of abbreviations that are commonly used in this thesis:

| Abbreviation | Description |
|---|---|
| **2D** | 2-Dimensional |
| **3D** | 3-Dimensional |
| **AVO** | Amplitude Variation with Offset |
| **BP** | Belief Propagation (algorithm) |
| **CI** | Conditional Independence (assumption on data given geology or any model parameters) |
| **CRF** | Conditional Random Field |
| **EM** | Expectation-Maximization (algorithm) |
| **FWI** | Full Waveform Inversion |
| **GM** | Gaussian Mixture (distribution) |
| **GMM** | Gaussian Mixture Model |
| **HMM** | Hidden Markov Model |
| **HMRF** | Hidden Markov Random Field |
| **IP** | P-Wave Impedance |
| **IS** | S-Wave Impedance |
| **LBP** | Loopy-Belief Propagation (algorithm) |
| **LL** | Localized likelihoods |
| **MAP** | Maximum-a-Posteriori |
| **McMC** | Markov-chain Monte Carlo |
| **MF** | Mean Field (approximation) |
| **MRF** | Markov Random Field |
| **PGM** | Probabilistic Graphical Model |
| **QLL** | Quasi-Localized Likelihoods |
| **Std.** | Standard Deviation |
| **SGMM** | Spatial Gaussian Mixture Model |
| **VB** | Variational Bayes |
| **VP** | P-Wave Velocity |
| **VPVS** | P-Wave to S-Wave Velocity Ratios |
| **VS** | S-Wave Velocity |

***THESIS***

# Chapter 1 Introduction

## 1.1 Motivation

### *1.1.1 The Need for Bayesian Inversion in Geosciences*

Assessment of geological heterogeneity plays a vital role in reservoir characterization and fluid-flow prediction in all subsurface reservoirs, and in the quantification of concomitant reservoir development and economic risk. The degree of heterogeneity in the subsurface is almost always underestimated because the amount of available geophysical data is always limited, and is usually sparse compared to the scale of variation of subsurface geological parameters such as discrete rock types (or facies) and continuous rock properties. For instance, seismic data provides spatially extensive subsurface coverage but is limited in resolution, usually to heterogeneity on length scales greater than tens or hundreds of metres. Borehole data, on the other hand, exhibit far higher resolution along the borehole trajectory but boreholes are usually sparsely distributed and therefore provide poor spatial coverage. These differences in spatial coverage and resolution provide different types and degrees of information, and both reduce and introduce uncertainties about the unknown model parameters. Further, inference of geological parameters from geophysical data is a non-unique problem, which means that different geological models could produce the same data within the data noise tolerance. Thus additional information, commonly referred to as the *geological prior information* (Curtis & Wood, 2004a; Curtis & Wood, 2004b), is required to obtain meaningful and geologically realistic results from geophysical data in the face of uncertainty. An appropriate data inversion or inference method must therefore combine a variety of measurements with different spatial coverage and resolution, together with all available prior information, and must assess the true resultant state of information and uncertainty about the subsurface. Bayesian inversion offers a convenient mathematical framework to achieve this.

Bayesian inversion models the unknown parameters of interest as random variables and describes the degree of uncertainty in these parameters in terms of probability distributions. A probabilistic description reflects a lack of knowledge about the true values of these parameters, or inherent variability of these parameters at scales smaller than the resolution of

geophysical data. The probability distribution that describes uncertainty in model parameters given only the prior information is called the *prior distribution*, while the distribution that describes the state of information given only the observed data is called the *likelihood*. In essence, Bayesian inversion combines the prior distribution and the likelihood to obtain the so-called *posterior distribution* which is the complete solution of the inverse problem. The *posterior distribution* describes the total resultant state of information given all of the observed data and prior information. Thus Bayesian inversion not only allows meaningful predictions about the unknown model parameters from observed data in the light of prior information, it also improves one's confidence in subsequent decision making by allowing assessment of uncertainty in the predictions.

### *1.1.2 Challenges in Stochastic Bayesian Inversion*

Exact Bayesian inference is impractical in practice because it requires normalization of the posterior distribution which is intractable for large models and must be approximated. Also, the computation and digital storage of complete joint posterior probability distributions over a large number of parameters using Bayesian inversion is intractable in most models of practical interest given available computing power and capacity. Probability distributions in high dimensional spaces are therefore generally explored through stochastic sampling, usually using the *Markov-chain Monte Carlo* (McMC) method, (e.g. Mosegaard & Tarantola, 1995; Mosegaard & Sambridge, 2002; Sambridge & Mosegaard, 2002) – a suite of general methods that theoretically produce a set of samples of parameter values which converges in density to that of the true posterior probability distribution as the number of samples tends to infinity. McMC methods therefore obtain a numerical approximation of the true posterior distribution using a finite number of samples, with a theoretical guarantee of asymptotic convergence only as sampling extends to infinity.

McMC based inversion methods are computationally expensive in most models of practical interest because as the number of parameters gets large, one can only expect that the distribution of samples would converge after generating an infeasible number of samples – often referred to as the curse of dimensionality (Curtis & Lomax, 2002). Further, detection of convergence of McMC based methods is usually based on subjective criteria. As a result, any estimate of the posterior that is obtained from any specific, fixed, finite set of Monte Carlo

samples may be biased by that particular set of samples depending on the criteria used for the detection of convergence. Such a bias is usually referred to as *convergence-related bias*. Posterior distributions can sometimes be assumed to take factorizable forms that divide high dimensional problems into lower dimensional problems, alleviating some of the difficulties in McMC sampling (e.g., Mosegaard & Tarantola, 1995; Mosegaard & Sambridge, 2002; Sambridge & Mosegaard, 2002; Gallagher *et al.* 2009). However, even for well-designed formulations of such posterior distributions the curse of dimensionality remains a barrier to rapid and accurate sampling based estimation. Therefore, efficient probabilistic inversion methods are sought that allow reliable detection of convergence. Additionally, efficient inversion methods may also allow a more complex and wider range of possible geological models to be assessed for more accurate predictions about the subsurface and improved estimation of uncertainty in the predictions.

### *1.1.3 The Need for Multi-point Geostatistics Based Prior Information*

The aim of probabilistic geophysical inversion is to produce geological models from geophysical data that represent our true state of knowledge about the subsurface combining the prior geological information with the information extracted from all of the available data. Such models are subsequently used for modelling and monitoring of subsurface fluid flow, and for making operational decisions in the field. Accurate statistical representation of geological heterogeneity in such models is a key requirement for successful use of these models. *Multi-Point Statistics* (MPS) based stochastic simulation methods have been developed (e.g. Guardiano & Srivastava, 1993) in Geostatistics that model higher-order statistics in contrast to the two-point variogram-based methods (such as Kriging: Journel, 1974), allowing realistic representation of heterogeneity and spatial continuity in geological models. Stochastic seismic inversion methods have been developed (e.g. Debeye *et al.* 1996) which generate a large number of realizations of the subsurface using MPS simulation. These methods incorporate geological prior information in terms of spatial geological patterns extracted from *training images* (TI), which are conceptual and graphical depictions of subsurface geological structures. As mentioned earlier, stochastic inversion methods are slow to converge, and convergence is neither guaranteed nor detectable.

More efficient inversion methods have been proposed within the Bayesian framework (e.g. Buland & Omre, 2003) under the linear Gaussian assumption, where both the unknown model parameters and measurement errors are assumed to be distributed as Gaussian and the forward model (relationship between model parameters and data) is assumed to be linear. Computational efficiency stems from the fact that the posterior distribution of desired model parameters can be derived analytically in this case. However these methods mostly rely on two-point statistics based prior information, which does not adequately model complex patterns of geological properties (Tahmasebi, 2018). The use of MPS based prior information therefore becomes inevitable in conditioning inverse problems when data is sparse and uncertain, which is often the case in subsurface modelling. This thesis aims to develop efficient Bayesian inversion methods that allow, or are easily extensible for, incorporation of MPS based prior information. Probabilistic inversion of geophysical data using geostatistical prior information is henceforth referred to as *geostatistical inversion*, and is the focus of this thesis.

### *1.1.4 Common Assumptions in Geostatistical Inversion*

Elastic properties of rocks, such as P-wave and S-wave impedances and Vp/Vs ratios, that may be derived from the seismic waveform data are commonly referred to as seismic attributes. Sensibly chosen seismic attributes corresponding to a given geological facies typically tend to cluster together. Therefore, geological facies may be inferred to some degree of certainty by clustering of different geophysical data or their attributes. Similarly, rock properties of petrophysical interest, such as clay volume, porosity and pore-space water saturation, may also be inferred from suitably chosen attributes of geophysical data based on their expected correlations with these attributes.

In order to appreciate the significance of scientific contribution of this thesis, it is first necessary to understand the set of assumptions that are commonly made in spatial statistical inference problems. To limit the analytical and computational complexity of spatial inverse problems, most previous research in geostatistical inversion makes two common assumptions: the *localized likelihoods* (LL) assumption and *conditional independence* (CI) of data (see e.g., Larsen *et al.* 2006; Caers *et al*. 2006; Hoffman & Caers, 2007; Ulvmoen & Omre, 2010; Shahraeeni & Curtis, 2011; Shahraeeni *et al.* 2012; Walker & Curtis, 2014a; and Grana, 2018). The LL assumption requires the seismic attributes at a location to depend on geological

parameters only at that location. Such an assumption is unrealistic because of imperfect data acquisition and processing procedures, and fundamental limitations of geophysical imaging methods. The CI assumption requires observed data to be independent of each other given the model parameters. In other words, no correlations in data noise across space (or time on temporal grids) are accounted for: the data are assumed to be mutually uncorrelated, apart from their interdependence due to correlated geological parameters. This is also unrealistic.

Although the LL and CI assumptions allow simpler mathematical treatment of the inverse problem and more efficient computation of its solution, they come at the cost of introducing two major limitations in modelling. First, these assumptions under-estimate long-range correlations present in the data. As a consequence, only short range correlations can be captured by the likelihoods: long-range correlations may only be captured through the geological prior information. Second, solutions ignore any correlated noise present in the seismic data which may percolate erroneously into the inversion results, e.g. any residual data acquisition foot-print such as that due to inhomogeneous equipment or ray/wave path distributions, improper focusing in the imaging process due to model errors, or residual multiples and surface wave noise in seismic images. Since such effects commonly impact all seismic surveys, any acquired data may contain long-range correlations due both to the reflected signal from geological layers, and to noise resulting from inaccuracies in data processing, or the acquisition footprint (Chopra & Larsen, 2000). Accounting for long-range correlations is therefore vital for the realistic reconstruction of complex geological patterns and thus for reliable subsurface modelling.

## 1.2 Importance of This Research

Resources stored in the Earth's subsurface are in shortening supply, yet are key to satisfying societal demand for energy (subsurface oil, gas, geothermal heat reservoirs and Uranium), materials (ore and mineral deposits), advanced technology and efficient engines (rare Earth elements) and fresh water. Subsurface reservoirs in rock pore space are also a key storage resource for waste material (nuclear waste and $CO_2$). Such resources are investigated and characterized, and the risks and economics associated with their development and use is mostly assessed from information collected at the Earth's surface; this is a highly uncertain process. Most contributions to this uncertainty are currently ignored because of the human

effort and computational cost required to include their effects. This may lead to suboptimal and poorly informed decisions, expensive errors, and economic inefficiency in a highly costly and national security-critical industrial sector.

Most of the cost (computational and human) derives from nonlinearity in physical relationships between what we can observe from the surface, and what we want to know about the subsurface. This appears to require expensive, Monte Carlo based computational methods to be used to interpret observed data, and also makes it hard for humans to judge what they genuinely do and do not know about the Earth's subsurface. In most cases, investigators therefore resort to methods developed in the 1960's to 1980's that use simplistic, linearized (approximate) physical relationships, which ignore most of what we know about nonlinearity. This is the cause of many of the errors described above. Forty years on, there is a need for modern methods of analysis that account for all known nonlinearity, and use that knowledge to reduce uncertainty to make better decisions.

This thesis aims to develop new methods for estimation of geological properties from geophysical data that account for non-linearity and allow assessment of uncertainties in geophysical data analysis in a computationally efficient manner.

## 1.3 Contribution

The fact that geological heterogeneity is ubiquitous in the subsurface, the need to combine prior geological information and geophysical data at multiple scales and resolution, the limitations of conventional sampling-based inversion methods, and the need for removal or relaxation of the localized likelihoods (LL) and conditional independence (CI) assumptions present significant challenges in Bayesian inversion to solve complex geophysical problems. This thesis aims to address these challenges by introducing novel scientific concepts and models, and by developing efficient methods for inverting geophysical data for spatial distribution of geological properties. The major contributions of this thesis are summarized below:

1. *Novel sampling-free inversion methods are developed to solve geophysical inverse problems based on a class of highly applicable yet structured models of parameter dependencies and a variational approach for approximate Bayesian inference. These are many orders of magnitude faster than the conventional McMC based inversion methods.*

2. *The commonly used localized-likelihoods (LL) and conditional independence (CI) assumptions in geostatistical inversion are progressively relaxed and finally removed – see section 1.4 and chapters 4 & 5.*

3. *A new probabilistic graphical model is introduced for fast approximate inference in 2D (and potentially multi-dimensional) inverse problems – the "2D Hidden Markov Model", and analytical expressions are derived to provide closed-form solutions for posterior marginal distributions of unknown model parameters (see chapter 4). This is also a novel contribution to the fields of mathematics and statistics, besides geosciences.*

4. *The concept of Quasi-Localized Likelihoods (QLL) is introduced as a relaxation of the localized likelihoods (LL) assumption (see chapter 5) in order to account for spatial blurring and consequential loss of resolution in the observed data. The latter is a typical problem in geophysical and many other types of remote-sensing data (e.g. in computer vision and medical imaging). So, this will potentially also be useful in many other fields of research.*

5. *A new machine learning model, the "spatial Gaussian mixture model", is developed for unsupervised clustering of data with limited resolution (such as seismic attributes), which acknowledges the spatial probabilistic dependence between both data and the model parameters. Thus, both prior and the likelihood distributions are defined in terms of spatially dependent parameters (see chapter 5). This is also a novel contribution to machine learning, besides geosciences.*

6. *A framework is introduced for probabilistic inversion of geophysical data, called "discriminative Bayesian inversion", which allows the incorporation of machine learning strategies within the Bayesian paradigm for solving inverse problems in order to address some of the most difficult challenges in spatial Bayesian inversion: to remove the LL and CI assumptions (see chapter 6).*

7. *A new approximate variational inference method, "higher-order mean field inference", is developed for performing efficient probabilistic inference in models with complex spatial dependencies among data and model parameters of interest (see chapter 6).*

8. *An efficient inversion method is developed for inversion of geophysical data for geological properties where parameters of the forward problem are estimated within*

*inversion. Analytical expressions are derived to update each of the unknown parameters in the model for a Gaussian posterior distribution (see chapter 7).*

9. *An efficient inversion method is developed for joint estimation of geological facies and petrophysical rock properties from seismic attributes, while honouring spatial dependencies among these parameters (see chapter 8).*

### *1.3.1 Peer-Reviewed Research Papers*

All of the research work presented in this thesis has been published in, submitted to, or is in preparation for publication in peer-reviewed journals. Few changes to notation and organization are made here compared to the publications in order to maintain consistency and coherency of this thesis.

- *Chapters 2 and 3 are in preparation for publications as review papers.*
- *Chapter 4 is published as:*

  Nawaz, M.A. & Curtis, A., 2017. Bayesian inversion of seismic attributes for geological facies using a hidden Markov model, *Geophysical Journal International*, 208, 1184–1200.

- *Chapter 5 is published as:*

  Nawaz, M.A. & Curtis, A., 2018. Variational Bayesian inversion of seismically derived non-localized rock properties for the spatial distribution of geological facies, *Geophysical Journal International*, 214, 845–875. doi: 10.1093/gji/ggy163.

- *Chapter 6 has been accepted for a publication as:*

  Nawaz, M.A. & Curtis, A., 2019. Rapid Discriminative Variational Bayesian Inversion of Geophysical Data for the Spatial Distribution of Geological Properties, *Journal of Geophysical Research: The Solid Earth*. (Accepted for publication).

- *Chapter 7 is in preparation for a publication.*
- *Chapter 8 is submitted as:*

  Nawaz, M.A., Curtis, A., Shahraeeni, M.S., & Gerea, C., 2019. Variational Bayesian Inversion of Seismic Attributes Jointly for Geological Facies and Petrophysical Rock Properties, *Geophysics*. (Submitted).

### *1.3.2 Conference Proceedings*

The following conference abstracts resulted from this research work.

- *Nawaz, M.A. & Curtis, A., 2016. Fast Bayesian Inversion of Seismic Data for Geological Facies using Localized Likelihoods. Poster presentation at: PETEX 2016, 15-17 November 2016, The Petroleum Exploration Society of Great Britain (PSEGB), London, UK. https://www.petex.info/wp-content/uploads/PETEX-2016-Programme_low-res.pdf.*
- *Nawaz, M.A. & Curtis, A., 2018. Uncertainty quantification and minimization in spatial problems. Poster presentation at: Uncertainty Quantification and Computational Imaging workshop, 23-24 April 2018, International Centre for Mathematical Sciences (ICMS), Edinburgh, UK.* http://www.icms.org.uk/uncertaintyquantification.php*.*
- *Nawaz, M.A. & Curtis, A., 2018. Uncertainty Reduction in Bayesian Inversion of Geophysical Data for Geological Facies using Machine Learning. Poster presentation at: PETEX 2018, 27-29 November 2018, The Petroleum Exploration Society of Great Britain (PSEGB), London, UK.* https://www.petex.info/wp-content/uploads/PETEX-2018-low-res-v2.pdf*.*
- *Nawaz, M.A. & Curtis, A., 2018. Variational Bayesian Inversion of Quasi-Localized Seismic Attributes for the Spatial Distribution of Geological Facies. Poster presentation at: 80th EAGE Conference & Exhibition 2018: Workshop on Seismic Inversion into Lithology/Fluid Classes, 10 June 2018, European Association of Geoscientists and Engineers (EAGE), Copenhagen, Denmark.* https://events.eage.org/en/2018/eage-annual-2018/technical-programme/workshops/workshop-05*.*

## 1.4 Thesis Outline

An overview of this thesis and its structure is given below.

**Chapter 2** introduces Bayesian probability theory and describes its application in the assessment of uncertainty in geophysical data analysis. Challenges in practical applications of Bayesian theory for probabilistic inference in high dimensional problems and possible solutions are highlighted. Specifically, developments in numerical optimization based techniques are

reviewed as viable alternatives to the more commonly used but computationally expensive stochastic methods for approximate Bayesian inference.

**Chapter 3** presents the philosophy behind the use of geological prior information in Bayesian solution of geophysical inverse problems. This chapter describes how to represent prior information about spatial variations in geology in a mathematical and graphical form as a probability distribution that can be directly input to a Bayesian inversion algorithm. Probabilistic graphical models are introduced for this purpose. In particular, hidden Markov model and Markov random fields are described as probabilistic models for spatial variations in geological properties.

**Chapter 4** introduces an efficient method for Bayesian inversion of discrete variables such as geological facies from the attributes of seismic data such as P-wave and S-wave impedances. Similar to most previous research in this field, this chapter makes the LL and CI assumptions. It is shown that the posterior distribution under these assumptions can be estimated analytically using a HMM and is therefore many orders of magnitude faster than sampling based methods for the same problem under the same set of assumptions.

**Chapter 5** introduces a new Bayesian inversion method that estimates the spatial distribution of geological facies from attributes of seismic data, by showing how the usual probabilistic inverse problem can be solved efficiently using an optimization framework while still providing fully probabilistic results. The LL assumption is relaxed in this method to account for spatial blurring of data. A new spatial Gaussian mixture model is introduced to perform classification in spatial problems where data from multiple classes (e.g. facies) shows strong similarities and so the classes are not easily discernible.

**Chapter 6** introduces a new approach to Bayesian inversion that directly models the desired spatial distribution of geological properties using supervised machine learning combined with spatial probabilistic inference, in contrast to the typical generative approach that models data generated from a given set of unknown model parameters (geological properties) as a part of probabilistic inference. The tasks of data modelling and spatial inference are thus separated in this method, which allows removal of LL and CI assumptions without significantly compromising on the computational efficiency of the method. This chapter also introduces a new probabilistic inference method “higher-order mean field inference” that allows multi-point statistics based prior information to be incorporated in

optimization based probabilistic inference. The method is supported by a synthetic and a real data example from New Zealand.

**Chapter 7** presents a hierarchical Bayesian inversion method for estimation of spatial distribution of continuous rock properties from geophysical data. The solution is derived analytically for a Gaussian posterior (post-inversion) distribution of the desired model parameters for both linearized and non-linear forward problems. The Bayesian inference is performed in an optimization framework where the posterior distribution evolves in each iteration as guided by both data and the prior information. Since updates are performed using analytical expressions, the method is computationally efficient and provides fully Bayesian results. The method is supported by a real data example from the North Sea.

**Chapter 8** extends the method developed in chapter 5 for joint inversion of continuous and discrete rock properties from geophysical data. In particular, the method estimates petrophysical rock properties and geological facies simultaneously from elastic attributes of seismic data. The method is supported by a real data example from the North Sea.

**Chapter 9** provides an overview of the strategies used in this research for the development of efficient and practical methods for probabilistic inference. A somewhat rough comparison of computational efficiency of optimization based versus sampling based probabilistic inference methods is provided for geostatistical inversion. Potential applications and future extensions of this research are reviewed.

**Chapter 10** concludes this thesis. Achievements of this body of research are articulated.

# Chapter 2 Bayesian Inversion

## 2.1 Inverse Problems

Inverse problems comprise of an unknown set of model parameters $\boldsymbol{m}$ that describe a physical system, and a set of observed parameters $\boldsymbol{d}$, commonly referred to as data, that consist of measurements obtained using a physical experiment conducted to obtain information about the physical system. The data may thus be considered to have been generated by a physical process that depends on the model parameters. The objective of solving an inverse problem is to infer the model parameters from empirical observations, often under some suitable constraints on the physical system. Solving an inverse problem is in general a challenging task because the observations usually do not depend on the unknown model parameters directly; they are also convoluted by the physical experiment, i.e. the method of obtaining these observations. For example, the observations are often made using a physical field (e.g. electromagnetic or seismic wave field) that is generated by a physical source. The observed data in this case also involves the effects of the source on the observations. The influence of physical experiment on data is generally described by a forward problem, which is expressed in the form of a mathematical expression or a numerical algorithm that defines how the data is generated from a given set of model parameters. The forward problem depends on the geometry of the experiment and structure of the physical system. Uncertainties are further introduced by a limited set of observations when the number of unknown parameters exceeds the number of useful observations, and by the presence of noise in the data. Inference of unknown parameters of interest from a typically limited set of noisy observations is the subject of (probabilistic) inverse problems.

## 2.2 Ill-Posedness of Geophysical Inversion

In the context of geophysical inverse problems, $\boldsymbol{m}$ refers to the geological properties of interest such as discrete litho-fluid classes (also called facies) or the continuous elastic and/or petrophysical properties of rocks (e.g. impedance, density, porosity and permeability), and $\boldsymbol{d}$ refers to any physical or digital observations that carry information about the unknown geological properties $\boldsymbol{m}$.

Various geophysical measurements are used to obtain information about the geological structures and rock properties of the subsurface. The data is often noisy and are corrupted by the data acquisition effects. Thus inversion of geophysical data to estimate geological models may involve significant uncertainties. Uncertainties in geophysical inversion must be assessed to make a proper use of the derived geological models (e.g. for petroleum exploration and production) and to avoid any associated risks.

We often seek information about the physical properties of rocks within a specified volume of earth, through the observations carried out at a boundary (typically the surface of the earth or in a borehole) of that volume. The number of observations is usually much smaller than the number of model parameters to be inferred from these observations. Other complicating factors include limited bandwidth and noise in the data, and assumptions and possible inaccuracies in theoretical relationships between data and model parameters (e.g. errors in forward modelling). Geophysical inverse problems are therefore inherently *ill-posed*, which means that the solution does not exist, is non-unique, or is instable – each of these cases is described below.

Non-existence of a solution refers to the situation that none of the models from a given set of possible solutions, usually referred to as the *model space*, can predict data according to the forward problem. This means that either the set of possible solutions considered is not rich enough or the forward problem does not satisfactorily represent the real physical experiment that generated the observed data. As an example, inversion of multi-component elastic seismic data that contains both P-wave and S-waves using an acoustic forward model cannot yield a solution.

Non-uniqueness of a solution refers to the situation when two very different models can explain the observed data equally well (within numerical tolerance). Non-uniqueness of a solution is typically caused by the observed data that bears insufficient information about the desired model parameters, and therefore more data may be required to discriminate between different models reasonably well. An example is the inversion of anisotropic seismic velocity model from reflection move-outs in a CMP gather where different combinations of vertical seismic velocity and anisotropy parameters may produce very similar seismic reflection move-outs.

Instability of a solution refers to the situation when the solution is very sensitive to the noise level in the data, and a small perturbation in the observed data may result in an arbitrarily large perturbation in the model parameters. Thus two very different models can generate two different predicted responses which differ only within the noise level of the observed data. Obviously, a possible cause of instability of a solution is the presence of high level of noise in the data, e.g. due to imperfect data acquisition, recording and handling methods. Another common reason for instability of a solution is the presence of discontinuities and nonlinearities in the forward problem.

### *2.2.1 Regularization of Inverse Problems*

Several strategies have been suggested in order to cope with the ill-posedness of inverse problems. A very commonly used method to obtain the solution of an ill-posed problem is known as *regularization*. Tikhonov regularization (Tikhonov, 1963) is a commonly used approach. It was developed by a Russian mathematician Andrei Tikhonov in 1943 and was inspired by the physical insight of field geophysicists that led them to the discovery of oil bearing subsurface geological structures using surface electrical measurements. Regularization solves an ill-posed problem by altering it such that it becomes well-posed. Thus, instead of solving the original ill-posed problem, regularization seeks the solution of a nearby well-posed problem. How close the well-posed problem should be to the original ill-posed problem and how the closeness between two problems is quantified, is usually determined by using subjective criteria and is still a topic of active research.

## 2.3 Bayesian Solution of Inverse Problems

Due to the inherent ill-posedness of most inverse problems of geophysical interest, a single solution that is in agreement with the observed data and satisfies any other desired constraints does not completely characterize the complete solution to the inverse problem. A complete characterization of the model space with respect to the given problem involves searching for all possible solutions and assessment of associated uncertainty (or degree of non-uniqueness). For this purpose, a probabilistic solution to the inverse problem is desired which also provides a quantitative estimate of how probable is each model to be a valid solution of the problem in question.

Bayesian framework is a probabilistic paradigm for solving inverse problems that acknowledges uncertainty in the parameters of interest, which is described in terms of probability distributions (Tarantola & Valette, 1982; Mosegaard & Tarantola, 1995). Bayesian inversion regularizes inverse problems using prior information about the expected solution(s) that is independent of the observed data. The prior information is represented by a probability distribution $\mathcal{P}(\boldsymbol{m})$ over all possible solutions for model parameters $\boldsymbol{m}$, called the *prior distribution*, and describes uncertainty in the model $\boldsymbol{m}$ before observing any data. The probability distribution that describes how likely is any given set of model parameters $\boldsymbol{m}$ to have generated the observed data $\boldsymbol{d}$ is called the *likelihood* and is represented by $\mathcal{P}(\boldsymbol{d}|\boldsymbol{m})$. The likelihood is often modelled as a deterministic (linear or nonlinear) function of model parameters and an additive stochastic component representing noise in the data. The likelihood encodes the information in the data $\boldsymbol{d}$ regarding the unknown model parameters $\boldsymbol{m}$.

The Bayesian solution to an inverse problem is also a probability distribution $\mathcal{P}(\boldsymbol{m}|\boldsymbol{d})$, called *posterior distribution*, over all possible solutions which are consistent with the data. The posterior distribution describes residual uncertainty in the model parameters that remains after combining the prior information with the information contained in the data regarding the desired model parameters, and represents the complete solution to an inverse problem. Thus the Bayesian solution also acknowledges the possibility of non-uniqueness of the solution of an inverse problem, and allows assessment of uncertainty in the solution. In comparison, a non-probabilistic solution to an inverse problem only provides a single 'best' solution and does not allow estimation of uncertainty in the solution.

Using the notation defined in the beginning of this thesis, we may generalize the notation for prior, likelihood and posterior distributions as $\mathcal{P}(\boldsymbol{m};\Theta)$, $\mathcal{P}(\boldsymbol{d}|\boldsymbol{m};\Theta)$ and $\mathcal{P}(\boldsymbol{m}|\boldsymbol{d};\Theta)$ parameterized by some nuisance parameters $\Theta$, which may not be of primary interest but are nevertheless important in the analysis of the model parameters of primary interest, $\boldsymbol{m}$. Since $\Theta$ is a set of parameters, it may include any number of parameters that are required to specify a probability distribution and does not specifically require these distributions to have the same functional form. The functional form of a probability distribution should be clear from the context. The parameters $\Theta$ are mostly assumed to be unknown but fixed, except in chapter 7 where we use both random and fixed parameters in the so called *hierarchical Bayesian model*. For now, we assume that $\Theta$ are unknown but fixed parameters.

The posterior distribution may be expressed using *Bayes' theorem* as

$$\mathcal{P}(\boldsymbol{m}|\boldsymbol{d};\Theta) = \frac{\mathcal{P}(\boldsymbol{m},\boldsymbol{d};\Theta)}{\mathcal{P}(\boldsymbol{d};\Theta)} = \frac{\mathcal{P}(\boldsymbol{d}|\boldsymbol{m};\Theta)\mathcal{P}(\boldsymbol{m};\Theta)}{\mathcal{P}(\boldsymbol{d};\Theta)} \quad 2.1$$

where the denominator $\mathcal{P}(\boldsymbol{d})$ represents the marginal likelihood of the observed data $\boldsymbol{d}$ for all possible sets of model parameters $\boldsymbol{m}$, and is therefore also referred to as the *model evidence*, or simply *evidence*, given by

$$\mathcal{P}(\boldsymbol{d};\Theta) = \int_{\boldsymbol{m}} \mathcal{P}(\boldsymbol{m},\boldsymbol{d};\Theta)d\boldsymbol{m} = \int_{\boldsymbol{m}} \mathcal{P}(\boldsymbol{d}|\boldsymbol{m};\Theta)\mathcal{P}(\boldsymbol{m};\Theta)d\boldsymbol{m} \quad 2.2$$

Since the data $\boldsymbol{d}$ is observed as a an instance of the underlying random data variable, the evidence $\mathcal{P}(\boldsymbol{d};\Theta)$ acts as an unknown constant that ensures normalization of the posterior distribution $\mathcal{P}(\boldsymbol{m}|\boldsymbol{d};\Theta)$ to be a valid probability distribution. Estimation of $\mathcal{P}(\boldsymbol{d};\Theta)$ using equation 2.2 requires integration over a possibly high dimensional space, which is intractable for most models of practical interest. It is this intractability that makes exact Bayesian inference impractical for realistic scale problems given the computational limitations of current digital technology. Approximate inference must therefore be performed.

The most widely used method for approximate inference is to explore the probability distributions in high dimensional spaces by stochastic sampling such as using Markov-chain Monte Carlo (McMC) based methods. However, as discussed in section 1.1.2, McMC based methods tend to be computationally intensive and slow to converge in high dimensional problems. Not only that convergence in McMC based methods is not guaranteed in high dimensional problems, the detection of convergence is a challenge in itself and it often involves subjective criteria. Further, McMC generates chains of samples that are distributed according to the posterior distribution as the number of samples tends to infinity. However, in most applications successive samples are highly correlated which severely reduces the information content of any finite sample set compared to a similarly sized set of independent samples. Hence, one seeks alternative methods of probabilistic inference which avoid McMC based sampling. Walker & Curtis (2014a) developed a facies inversion method using exact sampling as an efficient alternative to the McMC sampling. In that method every sample is an independent sample from the posterior probability distribution which is not the case in McMC methods. However, their method is also computationally intensive and requires large

computer memory in high dimensional problems. Therefore, efficient probabilistic inversion methods are required that allow reliable detection of convergence. Such methods may also help improve estimation of uncertainty in the inverse problems by allowing evaluation of a large number of and possibly more complex geological models using the same compute time and power.

## 2.4 Variational Bayesian Inference

Estimating $\mathcal{P}(\boldsymbol{d};\Theta)$ in Bayesian inverse problems is challenging for most problems of practical interest since its evaluation requires summation and/or integration over a very high dimensional space. Rather than trying to estimate $\mathcal{P}(\boldsymbol{d};\Theta)$ as a general function of $\Theta$ which is intractable, we first try to estimate $\Theta$ from the observations $\boldsymbol{d}$; once $\Theta$ has been fixed, estimating $\mathcal{P}(\boldsymbol{d};\Theta)$ is a more tractable problem. The parameters $\Theta$ can be estimated from $\boldsymbol{d}$ using the *maximum-likelihood* (ML) method that aims to find the parameters by setting $\Theta = \hat{\Theta}_{ML}$ that maximizes the joint likelihood, or equivalently the logarithm of joint likelihood $\mathcal{L}(\Theta;\boldsymbol{m},\boldsymbol{d}) \equiv \log \mathcal{P}(\boldsymbol{m},\boldsymbol{d};\Theta)$, of $\boldsymbol{m}$ and $\boldsymbol{d}$ as a function of parameters $\Theta$:

$$\hat{\Theta}_{ML} = \underset{\Theta}{\operatorname{argmax}}\{\log \mathcal{P}(\boldsymbol{m},\boldsymbol{d};\Theta)\} \equiv \underset{\Theta}{\operatorname{argmax}}\{\mathcal{L}(\Theta;\boldsymbol{m},\boldsymbol{d})\} \tag{2.3}$$

If the model $\boldsymbol{m}$ is known, $\mathcal{L}(\Theta;\boldsymbol{m},\boldsymbol{d})$ defines the *joint log-likelihood* as a function of the model parameters $\Theta$. However, since $\boldsymbol{m}$ is unknown, it must be marginalized out resulting in the *marginal log-likelihood* $\mathcal{L}(\Theta;\boldsymbol{d})$ of the observed data $\boldsymbol{d}$, henceforth referred to as *log-evidence*, that can be written as a function of parameters $\Theta$ as

$$\mathcal{L}(\Theta;\boldsymbol{d}) \equiv \log \mathcal{P}(\boldsymbol{d};\Theta) = \log \int_{\boldsymbol{m}} \mathcal{P}(\boldsymbol{m},\boldsymbol{d};\Theta) d\boldsymbol{m} \tag{2.4}$$

Estimation of $\Theta$ is hard in this case since the above integral may not be computed analytically. Further, even if the integral may be approximated numerically, its computational complexity increases exponentially with the dimensionality of model parameters $\boldsymbol{m}$.

In order to address these difficulties while avoiding stochastic sampling, a variational approach to inference – known as *variational Bayes* (VB) (Neal & Hinton, 1998; Beal, 2003; Nawaz & Curtis, 2018), is mostly adopted in this thesis which is a more efficient alternative to

McMC in models where posterior distribution may be approximated by a factorizable form with reasonably accuracy. VB approximates the intractable posterior distribution $\mathcal{P}(\boldsymbol{m}|\boldsymbol{d};\Theta)$ by a simpler, so called *auxiliary* or *variational* distribution $\mathcal{Q}(\boldsymbol{m}|\boldsymbol{d})$ of the unknown model parameters $\boldsymbol{m}$ from a family $\mathbb{Q}$ of distributions which is more amenable to analytical and numerical treatment. Such an approximation is commonly referred to as the *variational approximation*. The term '*variational*' is derived from the field of *calculus of variations* in mathematics that is used in this method to obtain functional approximation of the intractable true posterior.

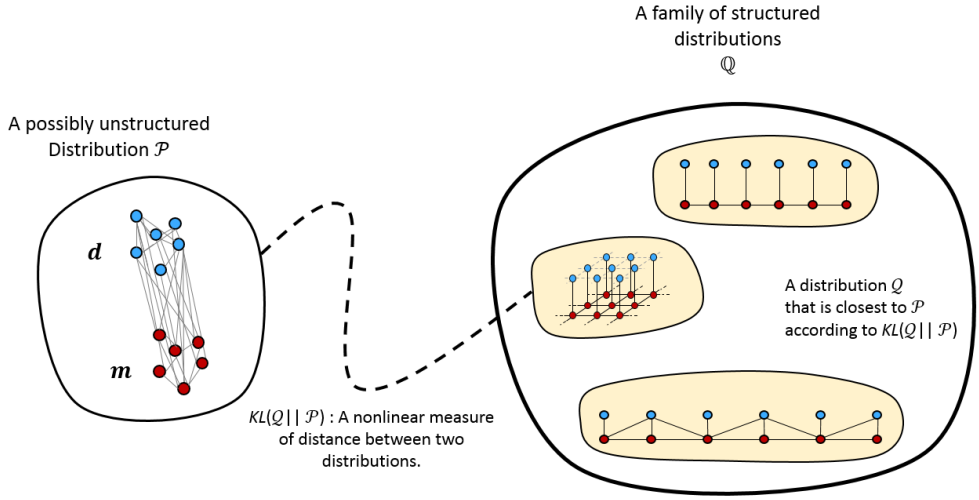



*Figure 2.1 A schematic illustration of variational Bayes method. It searches for a distribution $\mathcal{Q}$, called variational distribution, from a structured family $\mathbb{Q}$ of distributions that minimizes the relative-entropy $KL(\mathcal{Q}||\mathcal{P})$ between $\mathcal{Q}$ and the true unknown distribution $\mathcal{P}$. If $\mathcal{P}$ also belongs to $\mathbb{Q}$, the relative entropy can be minimized to its minimum possible value 0. That is, $\mathcal{Q}$ equals $\mathcal{P}$ in this case.*

The variational distribution is typically chosen to have a factorizable form. Typically the true posterior $\mathcal{P}(\boldsymbol{m}|\boldsymbol{d};\Theta)$ does not belong to $\mathbb{Q}$ (see figure 2.1). We will see in chapter 3 that spatial models that are of common interest in geosciences (or in general spatial data analysis) fall under this category, and therefore, VB is an attractive approach for probabilistic inference in such models. VB inference provides both the posterior point statistics of interest such as the *maximum-a-posteriori* (MAP) solution, as well as estimates of uncertainty in the posterior solution. Bayesian inversion based on the variational approximation is referred to as *variational Bayesian inversion* (VBI) (Kiebel *et al.* 2008; Jin & Zou, 2010; Nawaz & Curtis, 2018). Unlike McMC which estimates the posterior distribution by exploring the model space through

stochastic sampling, VB approximates the true but unknown posterior distribution using a deterministic optimization approach. It is shown below how VB exploits factorization properties of the variational distribution to transform probabilistic inference problem into a constrained optimization problem under the variational approximation.

The expected joint log-likelihood $\mathcal{L}(\Theta; \boldsymbol{m}, \boldsymbol{d})$ with respect to $\mathcal{Q}(\boldsymbol{m}|\boldsymbol{d}) \in \mathbb{Q}$ may then be defined as a function of $\Theta$ as

$$\mathbb{E}_{\mathcal{Q}}[\mathcal{L}(\Theta; \boldsymbol{m}, \boldsymbol{d})] \equiv \int_{\boldsymbol{m}} \mathcal{Q}(\boldsymbol{m}|\boldsymbol{d}) \log \mathcal{P}(\boldsymbol{m}, \boldsymbol{d}; \Theta)\, d\boldsymbol{m} \qquad 2.5$$

which is linear in the joint log-likelihood and is equally factorizable. The notation $\mathbb{E}_{\mathcal{Q}}[\cdot]$ represents expectation of the argument with respect to the variational distribution $\mathcal{Q}(\boldsymbol{m}|\boldsymbol{d})$. As shown below, this allows estimation of posterior marginal distributions and the MAP solution to the Bayesian inverse problem through inference on $\mathcal{Q}(\boldsymbol{m}|\boldsymbol{d})$ rather than the unknown true posterior $\mathcal{P}(\boldsymbol{m}|\boldsymbol{d}; \Theta)$. Since there is no ambiguity in the arguments of $\mathcal{Q}(\boldsymbol{m}|\boldsymbol{d})$ as it does not explicitly depend on $\Theta$, we often denote it just as $\mathcal{Q}$.

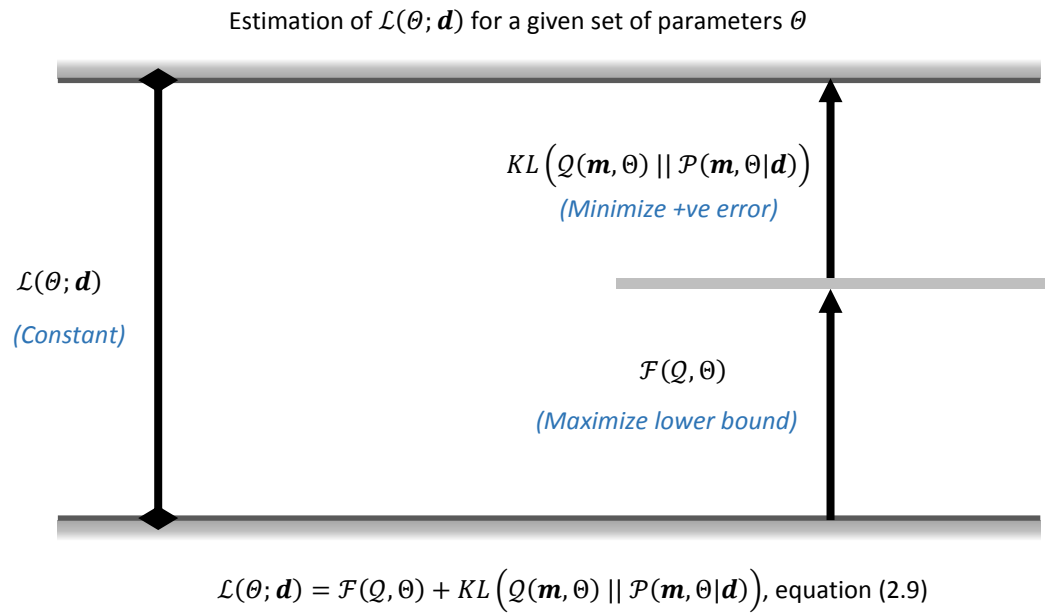



*Figure 2.2 A schematic illustration of minimizing the relative-entropy $KL\big(\mathcal{Q}(\boldsymbol{m}|\boldsymbol{d})||\mathcal{P}(\boldsymbol{m}|\boldsymbol{d}; \Theta)\big)$ between the variational distribution $\mathcal{Q}(\boldsymbol{m}|\boldsymbol{d})$ and the true posterior distribution $\mathcal{P}(\boldsymbol{m}|\boldsymbol{d}; \Theta)$ for a fixed set of parameters $\Theta$. Since the marginal log-likelihood $\mathcal{L}(\Theta; \boldsymbol{d})$ of observed variables $\boldsymbol{d}$ is a constant for fixed $\Theta$, maximizing the variational free energy $\mathcal{F}(\mathcal{Q}, \Theta)$ with respect to $\mathcal{Q}$ corresponds to minimizing the relative-entropy $KL\big(\mathcal{Q}(\boldsymbol{m}|\boldsymbol{d})||\mathcal{P}(\boldsymbol{m}|\boldsymbol{d}; \Theta)\big)$ between $\mathcal{Q}(\boldsymbol{m}|\boldsymbol{d})$ and $\mathcal{P}(\boldsymbol{m}|\boldsymbol{d}; \Theta)$.*

The expected joint log-likelihood $\mathbb{E}_Q[\mathcal{L}(\Theta; \boldsymbol{m}, \boldsymbol{d})]$ acts as a lower bound on the marginal log-likelihood $\mathcal{L}(\Theta; \boldsymbol{d})$ as can be seen by

$$\mathcal{L}(\Theta; \boldsymbol{d}) = \log \int_{\boldsymbol{m}} \mathcal{P}(\boldsymbol{m}, \boldsymbol{d}; \Theta) d\boldsymbol{m}$$

$$= \log \mathbb{E}_Q \left[ \frac{\mathcal{P}(\boldsymbol{m}, \boldsymbol{d}; \Theta)}{\mathcal{Q}(\boldsymbol{m}|\boldsymbol{d})} \right]$$

$$\geq \mathbb{E}_Q[\log \mathcal{P}(\boldsymbol{m}, \boldsymbol{d}; \Theta)] - \mathbb{E}_Q[\log \mathcal{Q}(\boldsymbol{m}|\boldsymbol{d})] \quad 2.6$$

[using Jensen's inequality]

$$= \mathbb{E}_Q[\mathcal{L}(\Theta; \boldsymbol{m}, \boldsymbol{d})] + \mathcal{S}(\mathcal{Q}) \quad 2.7$$

$$\equiv \mathcal{F}(\mathcal{Q}, \Theta) \quad 2.8$$

where $\mathcal{S}(\mathcal{Q}) = -\mathbb{E}_Q[\log \mathcal{Q}(\boldsymbol{m}|\boldsymbol{d})]$ is the *entropy* of the distribution $\mathcal{Q}(\boldsymbol{m}|\boldsymbol{d})$ and the functional $\mathcal{F}(\mathcal{Q}, \Theta)$ is called the *variational free energy* or simply *free energy*. These terms have their origin in statistical physics where $\mathcal{F}(\mathcal{Q}, \Theta)$ corresponds to the negative of *Gibbs free energy* (Feynman, 1972). The first term in equation 2.7, $\mathbb{E}_Q[\mathcal{L}(\Theta; \boldsymbol{m}, \boldsymbol{d})]$, represents the expectation of the joint log-likelihood $\mathcal{L}(\Theta; \boldsymbol{m}, \boldsymbol{d})$ with respect to the variational distribution $\mathcal{Q}(\boldsymbol{m}|\boldsymbol{d})$ as defined in equation 2.5.

Although $\mathcal{L}(\Theta; \boldsymbol{d})$ is intractable in most high-dimensional models, its lower bound $\mathcal{F}(\mathcal{Q}, \Theta)$ (see equation 2.6 and 2.8) may be estimated for a suitably chosen $\mathcal{Q}$. The aim in variational optimization is to estimate the variational distribution $\mathcal{Q}(\boldsymbol{m}|\boldsymbol{d})$ of the unknown model parameters $\boldsymbol{m}$ that maximizes the free energy functional $\mathcal{F}(\mathcal{Q}, \Theta)$ with respect to both $\mathcal{Q}$ and $\Theta$, rather than directly estimating $\mathcal{L}(\Theta; \boldsymbol{d})$. The variational approximation therefore allows us to cast the inference problem into a constrained optimization problem, referred to as *variational optimization*. Also by definition

$$\mathcal{F}(\mathcal{Q}, \Theta) = \mathbb{E}_Q[\log \mathcal{P}(\boldsymbol{m}, \boldsymbol{d}; \Theta)] - \mathbb{E}_Q[\log \mathcal{Q}(\boldsymbol{m}|\boldsymbol{d})]$$

$$= \mathbb{E}_{\mathcal{Q}}[\log \mathcal{P}(\boldsymbol{m}|\boldsymbol{d};\Theta)] + \mathbb{E}_{\mathcal{Q}}[\log \mathcal{P}(\boldsymbol{d};\Theta)] - \mathbb{E}_{\mathcal{Q}}[\log \mathcal{Q}(\boldsymbol{m}|\boldsymbol{d})]$$

$$= \mathbb{E}_{\mathcal{Q}}\left[\log \frac{\mathcal{P}(\boldsymbol{m}|\boldsymbol{d};\Theta)}{\mathcal{Q}(\boldsymbol{m}|\boldsymbol{d})}\right] + log\, \mathcal{P}(\boldsymbol{d};\Theta)$$

[since $\log \mathcal{P}(\boldsymbol{d};\Theta)$ is independent of $\mathcal{Q}(\boldsymbol{m}|\boldsymbol{d})$]

$$= -KL\big(\mathcal{Q}(\boldsymbol{m}|\boldsymbol{d})||\mathcal{P}(\boldsymbol{m}|\boldsymbol{d};\Theta)\big) + \mathcal{L}(\Theta;\boldsymbol{d}) \qquad 2.9$$

where $KL\big(\mathcal{Q}(\boldsymbol{m}|\boldsymbol{d})||\mathcal{P}(\boldsymbol{m}|\boldsymbol{d};\Theta)\big)$ is the *Kullback-Leibler (KL) divergence* (also called *relative-entropy*, Shannon, 1948) between $\mathcal{Q}(\boldsymbol{m}|\boldsymbol{d})$ and $\mathcal{P}(\boldsymbol{m}|\boldsymbol{d};\Theta)$, which is a measure of difference between its two argument distributions, and is given by

$$KL\big(\mathcal{Q}(\boldsymbol{m}|\boldsymbol{d})||\mathcal{P}(\boldsymbol{m}|\boldsymbol{d};\Theta)\big) \equiv \mathbb{E}_{\mathcal{Q}}\left[\log \frac{\mathcal{Q}(\boldsymbol{m}|\boldsymbol{d})}{\mathcal{P}(\boldsymbol{m}|\boldsymbol{d};\Theta)}\right]$$

$$= \int_{\boldsymbol{m}} \mathcal{Q}(\boldsymbol{m}|\boldsymbol{d}) \log \frac{\mathcal{Q}(\boldsymbol{m}|\boldsymbol{d})}{\mathcal{P}(\boldsymbol{m}|\boldsymbol{d};\Theta)}\, d\boldsymbol{m} \geq 0 \qquad 2.10$$

For notational brevity, the relative entropy as given above is henceforth represented as $KL(\mathcal{Q}||\mathcal{P})$. Since $\mathcal{L}(\Theta;\boldsymbol{d})$ is independent of $\mathcal{Q}(\boldsymbol{m}|\boldsymbol{d})$, maximizing $\mathcal{F}(\mathcal{Q},\Theta)$ is equivalent to minimizing the relative-entropy $KL(\mathcal{Q}||\mathcal{P})$ (2.9). The KL divergence takes a minimum value of zero when the two distributions that it compares are identical. Therefore, by maximizing the free energy $\mathcal{F}(\mathcal{Q},\Theta)$ for a given set of parameters $\Theta$ the *variational Bayesian* inference effectively estimates $\mathcal{Q}$ that best approximates the posterior distribution $\mathcal{P}(\boldsymbol{m}|\boldsymbol{d};\Theta)$ (see figure 2.2).

Note that the variational formulation in equations 2.8 and 2.9 above is exact for any arbitrary $\mathcal{Q}$, but the free energy $\mathcal{F}(\mathcal{Q},\Theta)$ is still intractable. Therefore, we need to approximate $\mathcal{F}(\mathcal{Q},\Theta)$ so that it can be maximized in order to estimate $\mathcal{L}(\Theta;\boldsymbol{d})$. This can be achieved either by restricting the functional $\mathcal{F}(\mathcal{Q},\Theta)$ to a specific tractable form, or by restricting $\mathcal{Q}$ to take a specific form (e.g. a factorizable form) that makes $\mathcal{F}(\mathcal{Q},\Theta)$ tractable. The former approach allows approximate $\mathcal{L}(\Theta;\boldsymbol{d})$ in an iterative fashion using a variational form of the *expectation-maximization* (EM) (Dempster *et al.*, 1977; Beal, 2003) algorithm, such that its lower bound

$\mathcal{F}(Q,\Theta)$ is increased while decreasing $KL(Q||\mathcal{P})$ for a given set of parameters $\Theta$ in each iteration. The latter approach allows *mean field (MF) approximation* (Stanley, 1971) to true posterior $\mathcal{P}(\boldsymbol{m}|\boldsymbol{d};\Theta)$, which restricts the variational distribution $Q(\boldsymbol{m}|\boldsymbol{d})$ to have an explicit factorizable form. We will use both of these approaches in this thesis depending on which approach is more suitable to the problem being solved. The EM algorithm is used in chapters 5 and 8, and MF approximation is used in chapters 6 and 7.

VB method for probabilistic inference is inspired from the developments of *mean field (MF) methods* in statistical physics (Feynman, 1972), and has its roots in machine learning (Hinton & Zemel, 1994; Jaakkola, 1997; Jordan *et al*. 1999; Neal & Hinton, 1999; Beal, 2003). It has been applied to solve (linear or weakly nonlinear) inverse problems in various domains of research. Kiebel *et al*. (2008) developed a VBI method for medical image processing. Jin & Zou (2010) discussed regularization and convergence properties of the VB method, and proposed two variational approximations of the posterior distribution which they applied to heat conduction problems. Yanqin & Guoshan (2014) solved blind seismic deconvolution problem using the VB method, and Penz *et al*. (2018) inverted electro-magnetic data using the VB method for geophysical applications. Nawaz & Curtis (2018) and Nawaz & Curtis (2019) (chapters 5 and 6 in this thesis) used variational inference for inversion of geophysical data for geological properties constrained by (multi-point) geostatistical prior information. An elegant generalized treatise on the variational inference methods and their convergence properties can be found in Koller & Friedman, 2009.

# Chapter 3 Probabilistic Representation of Geological Prior Information

Central to most resource and risk assessments is geological heterogeneity (subsurface variations in rock properties). Unknown heterogeneity diminishes our knowledge about the distribution and economics of subsurface resources. High uncertainty about geological heterogeneity is the usual manifestation of the need for geological information to be incorporated in the geophysical inverse problems.

Owing to heterogeneity in the natural world, geology may appear as random at various scales. Nevertheless, geological parameters at nearby locations are more likely to be similar than at distant locations. This is supported by Tobler's first law of geography, which states that "*everything is related to everything else, but near things are more related than distant things*" (Tobler, 1970). This implies that geological properties at any location are strong correlated only within a certain neighbourhood of that location. This is a common observation in geostatistical *variogram analysis* – study of correlations in geological properties as a function of spatial distance between the locations of their observation (Mariethoz & Caers, 2014).

The spatial context in geology induces similar correlations in geophysical data pertaining geological properties of interest within the volume of subsurface that is investigated by geophysical observations. Such spatial probabilistic dependence between geology and geophysical data may be incorporated as geological prior information in order to mitigate ill-posedness of geophysical inverse problems. Geological prior information must have been obtained independent of the data under current analysis, but is nevertheless important in order to assure that inferred subsurface models are geologically realistic (Curtis & Wood, 2004a; Curtis & Wood, 2004b; Mariethoz & Caers, 2014). Such information ultimately derives from previously acquired data, or from prior experience of geoscientists on the local and regional geology. Accordingly, prior information may be obtained through direct expert elicitation (Curtis & Wood, 2004b; Bond *et al*. 2007; Polson & Curtis, 2010, 2015; Curtis, 2012; Arnold & Curtis, 2018), through literature review (Curtis & Wood, 2004b), or may be inferred through an indirect interactive process (Boschetti & Moresi, 2001; Walker & Curtis, 2014b), or

from modelling of geological processes that might have produced the geological structures in a given depositional environment (Hill *et al*. 2009).

## 3.1 Objective Representation of Prior Information

How best to describe and incorporate geological prior information in Bayesian inverse problems is still a subject of active research. The prior information can have a strong influence on final models so it is key that the choice of parameters allows both that it can be combined with information from currently observed data, and that solutions can be updated in the light of either new data or new updates in the prior information (Walker & Curtis, 2014c). The prior information may be parametrized, through field observations (e.g. Hodgetts *et al*. 2004; Jones *et al*. 2004; Verwer *et al*. 2004), as probability distributions (e.g. Hansen *et al*. 2016), variograms and statistics (e.g. Hansen *et al*. 2006; Hansen *et al*. 2008; Lindberg *et al*. 2015; Rezvandehy & Deutsch, 2017), images (e.g. Strebelle, 2001; Arpat, 2005; Journel & Zhang, 2006; González *et al*. 2008; Mariethoz & Caers, 2014), geological process models (Griffiths *et al*. 2001; Burgess & Emery, 2004; Tetzlaff, 2004; Hill *et al*. 2009), logic trees (e.g. Pshenichny, 2004) and a variety of other methods. The Bayesian inversion methods developed in this thesis allow injection of prior information in any of these forms, however for the most part we assume that prior information about spatial distribution of geological facies is available in the form of training images which are described in section 3.2 below, and about the continuous rock properties is available in the form of spatial statistics (such as covariance function, Hansen *et al*. 2006).

## *3.2 Training Images*

Geological phenomena always exhibit some degree of spatial correlation and continuity, but also apparent randomness in space at various scales. Such spatial variability may be described by a geological continuity model that is ultimately governed by geological processes. The spatial variability in geophysically detectable rock properties (e.g., elastic or electromagnetic properties or density) generally follows the spatial distribution of geological facies (distinctly classifiable litho-fluid types) but is often more complex than the spatial variability of the facies themselves. For this reason, geoscientists can provide better a priori constraints on the spatial distribution of discrete geological facies than on the variability in the

continuous rock properties in space. A convenient way to quantitatively embody priori information about the spatial distributions of geological facies in space is through a *training image* (Mariethoz & Caers, 2014).

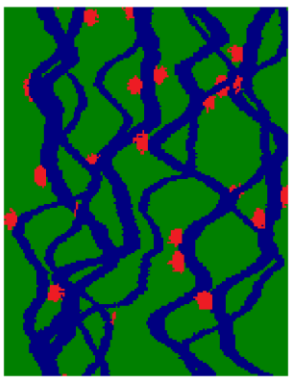


*Figure 3.1: An example of a 2D training image. Green colour represents shale, blue represents channel sands, and red represents over-bank sand deposits.*

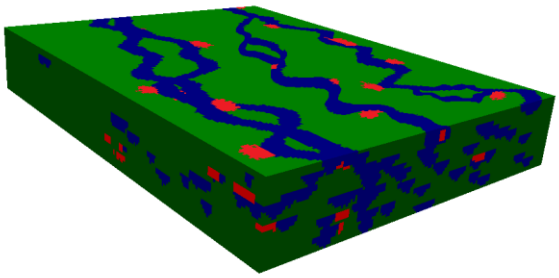


*Figure 3.2: An example of a 3D training image. Colour scheme is same as in figure 3.1.*

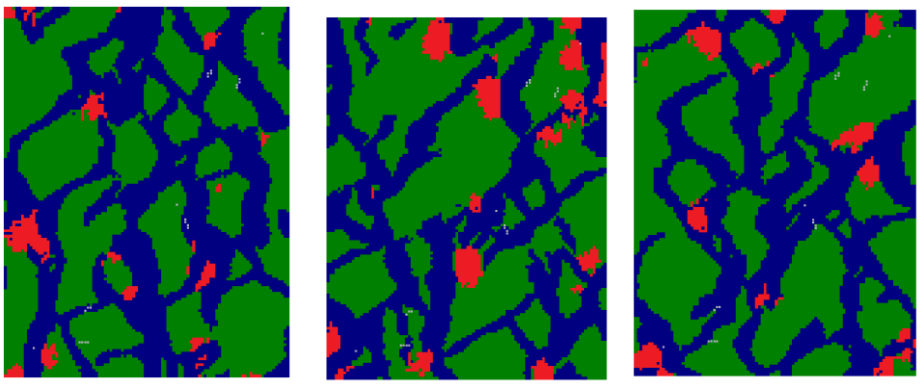


*Figure 3.3: Stochastically simulated facies using the training image in figure 3.1. Colour scheme is same as in figure 3.1.*

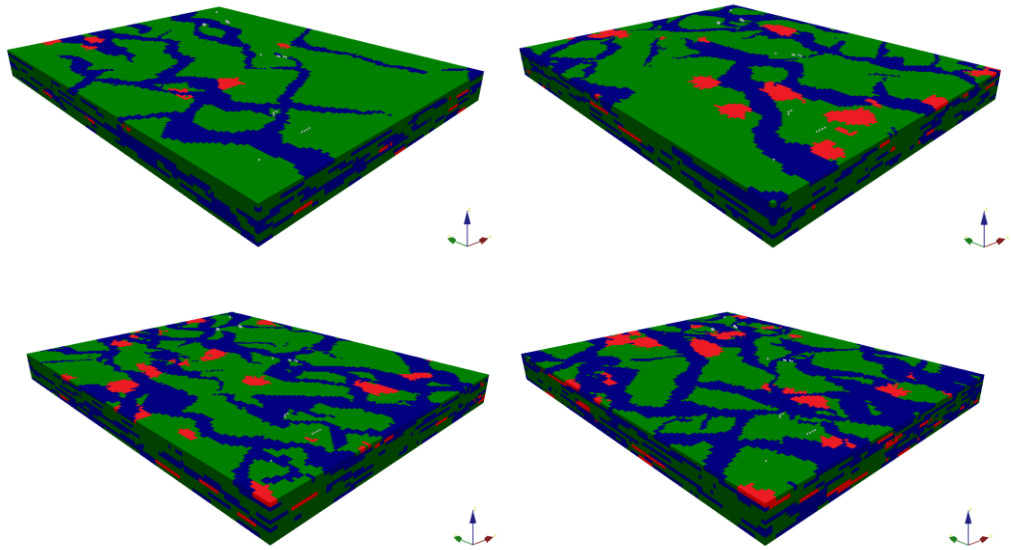


*Figure 3.4: Stochastically simulated facies using the training image in figure 3.2. Colour scheme is same as in figure 3.1.*

A training image is a conceptual depiction of typical patterns of geological features that are expected to exist in the subsurface based on the subjective opinion of geoscientists, or on other objective geological measurements of facies distributions. It is a pictorial manifestation of spatial continuity of facies, and captures the statistics of facies heterogeneity over a lattice of model cells that is consistent with the true geology. These statistics may then be extracted from the training image(s) as and when desired, and may then be injected into Bayesian inversion in the form of prior information. Thus a training image also serves as a compact embodiment of joint and conditional probability distributions over spatial variables which would otherwise require a comparatively large amount of computer memory for their digital storage. Another conceptual advantage of using a training image is that it restricts the expected spatial patterns of facies to a limited set of geologically plausible patterns as depicted in the image. This typically reduces an intractably large number of parameters needed to describe probabilistic dependence to a relatively small and computationally feasible number of parameters in practice. It is, however, important to note that a training image only provides contextual information about the local geology, and not location specific information as is supplied by the data in the form of likelihoods. Figures 3.1 and 3.2 show examples of 2D and 3D training images of three facies: shale, channel sands and over-bank sand deposits

shown in green, blue and red colours, respectively. Figures 3.3 and 3.4 display stochastic realizations from the training images shown in figures 3.1 and 3.2, respectively.

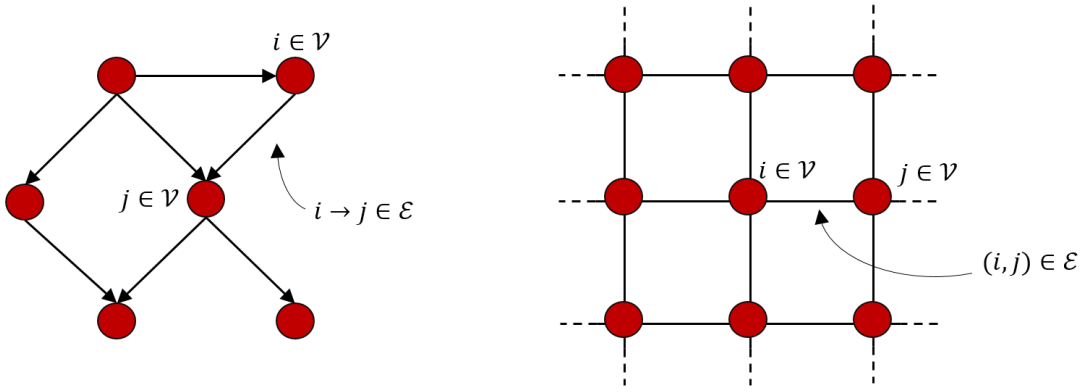
## 3.3 Probabilistic Graphical Model (PGM)

A fundamental requirement in geophysical inversion is to capture the probabilistic spatial distribution and heterogeneity of subsurface properties that is consistent with the true earth, and inject this information into the inversion process in the form of prior information. The aim is to reconstruct the spatial distribution of geological properties (facies and rock properties) in a Bayesian framework by combining the data likelihoods with this *a priori* information. Depending on the inversion algorithm and on the type and complexity of the *a priori* information, different methods exist which mathematically transform *a priori* information into probability distributions. However, when prior information on spatial distribution of geological properties only involves correlations between these properties at neighbouring locations, the strength of such correlations may be encoded in parameters which depend on the relative locations of the neighbours. This can be achieved by parameterizing the spatial distributions of geological properties in the form of a *probabilistic graphical model* (PGM) (Koller & Friedman, 2009) – a structured representation of probabilistic dependence among various parameters of interest, which is described below.

A PGM is a graphical representation of a multivariate probability distribution, typically over a large number of random variables, which decomposes into factors each of which depends only on a smaller subset of variables. Such factorization plays a vital role in probabilistic inference in high dimensions: it connotes a conditional independence structure among some subsets of variables which is crucial for tractable inference in such models. Thus a PGM can accurately represent a joint probability distribution over a large number of variables, while allowing efficient inference by capitalizing on the conditional independence among most of these variables.

A graph $\mathbb{G}(\mathcal{V}, \mathcal{E})$ defines a set of vertices $\mathcal{V}$ (also called nodes) which represent random variables and a set of edges $\mathcal{E}$ where each edge connects exactly one vertex to another in the graph. For brevity, $\mathbb{G}(\mathcal{V}, \mathcal{E})$ is often represented just as $\mathbb{G}$ in the following. The edges represent direct probabilistic dependence between connected vertices. The edges may be undirected

(represented by line segments) or directed (represented by arrows) depending on the directionality of probabilistic influence (see figure 3.5).




*(a)* *(b)*

*Figure 3.5 Examples of probabilistic graphical models (PGM). (a) A directed PGM (also called a Bayes net). (b) An undirected PGM (also called a Markov random field – MRF).*

## *3.3.1 Directed PGM*

A directed PGM encodes causal probabilistic dependence between each of the connected pairs of variables. Causality is induced on the graph by defining ordered relationships that introduce the notions of past and future with respect to a given cell (e.g., see figure 3.5a). The causality also defines the flow of probabilistic influence across all of the cells in the model. Each vertex (random variable) in a directed graph $\mathbb{G}$ is associated with a *probability density function* (PDF). A directed graph $\mathbb{G}$ defines a relationship $\rightarrow$ on its vertices such that for any two vertices $i$ and $j$, the relationship $i \rightarrow j \in \mathcal{E}$ holds when $j$ directly depends on $i$, i.e. not through some other vertex in the graph. The vertex at the head of the arrow is called the *child vertex* and the vertex on its tail is called the *parent vertex*. So $i \rightarrow j \in \mathcal{E}$ implies that $i$ is a parent of $j$, such that the child vertices depend on their parents while the parent vertices do not depend on their children. Cyclic dependence refers to the case when a vertex depends on any of its children (or grand-children), which is not permissible in directed PGMs. For this reason, directed PGMs are sometimes more explicitly referred to as *directed acyclic graphs* (DAG). The PDF associated with each vertex (random variable) may therefore only be defined in terms of its parent variables (and not any other variables in the model). A

directed PGM is commonly known as a *Bayesian network* or *Bayes net* in the machine learning community (Koller & Friedman, 2009).

A path in a directed graph is defined by an ordered sequence of vertices in $\mathcal{V}$ such that a path between any two vertices $i$ and $j$ is said to exist when $j$ depends on $i$, either directly or indirectly through any other vertices. A directed PGM $\mathbb{G}$ defines an order $<$ on its vertices such that for any two vertices $i$ and $j$, $i < j$ when there exists an unblocked path from $i$ to $j$ in $\mathbb{G}$; i.e. when the probabilistic influence may flow from $i$ to $j$. Similarly, $\mathbb{G}$ also defines a partial order $\leq$ which is similar to the order $<$ except that it also allows that $i = j$. The orders $>$ and $\geq$ are similarly defined such that $i > j$ implies that there exists no direct or indirect path from $i$ to $j$ in $\mathbb{G}$; and $i \geq j$ implies that either there is no direct or indirect path from $i$ to $j$ in $\mathbb{G}$, or that $i = j$.

Every vertex $i \in \mathcal{V}$ is associated with a set $\mathcal{N}_{\backslash i} \subset \mathcal{V} \backslash \{i\}$ of neighbouring vertices; these share an edge in $\mathcal{E}$ from the vertex $i \in \mathcal{V}$, and are referred to as the *neighbourhood* of $i$. So $j \in \mathcal{N}_{\backslash i}$ if and only if $i \to j \in \mathcal{E}$ or $j \to i \in \mathcal{E}$. By definition, the neighbouring relationship must satisfy two properties: a vertex cannot be a neighbour of itself, i.e., $i \notin \mathcal{N}_{\backslash i}$ (as is emphasized by the subscript '$\backslash i$'), and the neighbouring relationship is commutative, i.e., $i \in \mathcal{N}_{\backslash j} \Rightarrow j \in \mathcal{N}_{\backslash i}$. Define *neighbourhood cardinality*, denoted as $|\mathcal{N}_i|$, as the number of vertices in the neighbourhood of a given vertex $i$. The neighbourhood cardinality of a cell at the boundary of a model is usually lower than that of a cell further away from the model boundaries. A common type of directed PGMs is the *hidden Markov model* (HMM) which is introduced and described in 1-D in section 3.4 below, and is extended to 2-D in chapter 4.

### *3.3.2 Undirected PGM*

An *undirected PGM* encodes non-causal probabilistic dependence between various random variables and offers natural representation of spatial phenomena. In contrast to the directed PGMs which are expressed by PDFs defined over each vertex, undirected PGMs are usually expressed in terms of non-negative *potential functions*, each of which is defined over a typically small set of variables such that they together encode the full joint probability distribution over all of the variables. If the potential functions are defined over pairs of variables, the associated graph is referred to as a *pairwise PGM*. However, if the potential functions are defined over some larger subsets of variables or clusters of (more than two)

vertices, the corresponding graph is called a *cluster graph* or *higher-order PGM*. A higher-order PGM can model more complex joint distributions over variables than a pairwise PGM over the same set of variables (Koller & Friedman, 2009). An undirected PGM is commonly known as a *Markov network*, or when it models spatial phenomena it is referred to as a *Markov random field* (MRF) (Koller & Friedman, 2009), which is introduced in section 3.5 below. Different variants of a MRF are used in this thesis to inject geological prior information in geophysical inversion.

## 3.4 Hidden Markov Model (HMM)

A stochastic process is a non-deterministic method to generate random variables as a function of an independent variable, such as time or space. A hidden Markov model (HMM) is a directed graphical model that represents a dual stochastic process: a stochastic process representing observations with an underlying stochastic process representing unobserved (also called hidden) states or model parameters. *Hidden Markov-chain*, or *1D-HMM*, is one-dimensional representation of a more general class of HMM's and are used to represent probability distributions over sequences of observations (figure 3.6) – see Stratonovich (1960), Baum et al. (1970), and Baum (1972). The observations are assumed to be produced by underlying unobserved (hidden) states that represent local (in time or space) instances of the underlying stochastic process.

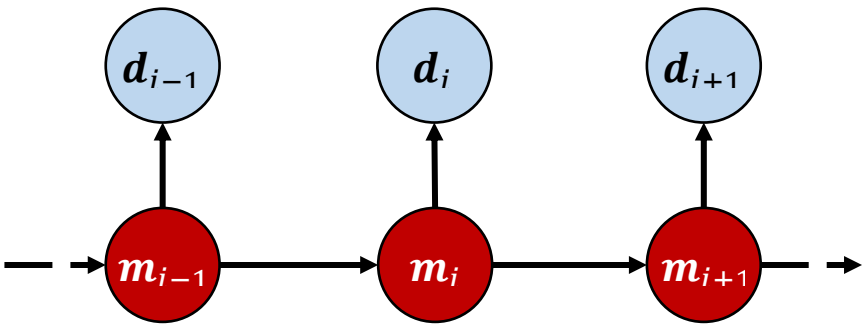



*Figure 3.6 An illustration of a hidden Markov-chain (1D-HMM). Arrow directions represent the directions of probabilistic influence between hidden states* $\boldsymbol{m}$ *and observations* $\boldsymbol{d}$*. Subscripts represent the index (typically time or space) of the corresponding state or observation.*

In a 1D-HMM, the observations and the underlying hidden states are indexed with a parameter $i$, which commonly refers to time but may also refer to some other measurement index such as space. A 1D-HMM assumes that the observation $\boldsymbol{d}_i$ at index $i$ was generated by a stochastic process whose state $\boldsymbol{m}_i$ is hidden from the observer. While data variables are typically assumed to be continuous, we assume that the state variables are also continuous for the sake of generality. It also assumes that the hidden states are sequentially distributed according to an underlying stochastic process that satisfies the (1st-order) *Markov property*: given the hidden state $\boldsymbol{m}_{i-1}$ at index $i-1$, the current state $\boldsymbol{m}_i$ at index $i$ is conditionally independent of all of the previous states $\boldsymbol{m}_1, \dots, \boldsymbol{m}_{i-2}$ at indices prior to $i-1$:

$$\mathcal{P}(\boldsymbol{m}_i|\boldsymbol{m}_1, \dots, \boldsymbol{m}_{i-1}) = \mathcal{P}(\boldsymbol{m}_i|\boldsymbol{m}_{i-1}) \quad 3.1$$

This means that a HMM is a memory-less process: the state $\boldsymbol{m}_{i-1}$ at index $i-1$ is assumed to encapsulate all of the history of the current state $\boldsymbol{m}_i$ at index $i$, and knowing the current state $\boldsymbol{m}_i$ is sufficient to generate the future states at indices $i+1$ and beyond. Another fundamental assumption of a HMM is that for a given state, the observation from that state is conditionally independent of all other observations and hidden states in the model.

$$\mathcal{P}\left(\boldsymbol{d}_i|\boldsymbol{m}, \boldsymbol{d}_{\setminus i}\right) = \mathcal{P}(\boldsymbol{d}_i|\boldsymbol{m}_i) \quad 3.2$$

where $\boldsymbol{m} = (\boldsymbol{m}_1, \dots, \boldsymbol{m}_n)$ and $\boldsymbol{d}_{\setminus i} = (\boldsymbol{d}_1, \dots, \boldsymbol{d}_{i-1}, \boldsymbol{d}_i, \boldsymbol{d}_{i+1}, \dots, \boldsymbol{d}_n)$ are the vectors of all hidden states, and data observed at all times indices $i$. This assumption may be decomposed into the following two assumptions

$$\mathcal{P}\left(\boldsymbol{d}_i|\boldsymbol{m}, \boldsymbol{d}_{\setminus i}\right) = \mathcal{P}(\boldsymbol{d}_i|\boldsymbol{m}) \quad 3.3$$

$$\mathcal{P}(\boldsymbol{d}_i|\boldsymbol{m}) = \mathcal{P}(\boldsymbol{d}_i|\boldsymbol{m}_i) \quad 3.4$$

The assumptions in equations 3.3 and 3.4 correspond to the conditional independence (CI) of data, and the localized likelihoods (LL) assumptions, respectively, as mentioned in section 1.1.4.

## 3.5 Markov Random Field (MRF)

A MRF is a structured set of probabilistic relationships among various parameters of interest at multiple locations, under the assumption that the parameters at any given location are directly dependent only on the parameters in some arbitrary but pre-specified neighbourhood of that location, the so called (1st-order) *Markov assumption*. MRF is widely used in geostatistics as a model for probabilistic dependence among geological properties of interest at multiple locations. The Markovian assumption thus requires that given the geology in the neighbourhood of any location in the model, the geology at that location is conditionally independent of the geology in the rest of the model, i.e. knowledge of geology in the rest of the model has no influence on the geology at the vertex in question. Such a model is simple enough that it allows rigorous and efficient probabilistic inference by leveraging the conditional independence structure of the model, yet sophisticated enough to represent complex spatial patterns of geological properties in the form of prior information. The class of problems considered in this thesis are those that can be represented with sufficient accuracy by this type of model.

The mathematical specification of a MRF requires defining the order of probabilistic dependence among various random variables in a model. For example, second order (or pairwise) dependence refers to the case when the joint distribution over all of the variables can be fully specified using only up to two-point statistics such as mean and covariance of random variables. Similarly, high-order dependence refers to the case when complete specification of the joint distribution requires high-order (or multi-point) statistics. A MRF defined in terms of only pairwise dependence of random variables is referred to as a *pairwise MRF*, while a MRF that involves higher-order dependence is referred to as a *higher-order MRF* or a *factor graph* (Koller & Friedman, 2009).

### 3.5.1 Pairwise MRF

Mathematically, a pairwise MRF is defined using graph theory terminology as an undirected graphical model $\mathbb{G}(\mathcal{V}, \mathcal{E})$ which defines the topology of some physical space (figure 3.7), where $\mathcal{V} = \{1, \dots, n\}$ is a set of vertices (also called nodes), and $\mathcal{E} = \{(i,j) : i,j \in \mathcal{V} \wedge i \neq j\}$ is the set of undirected edges (or connections between vertices) in the graph where $\mathcal{E} \subset$

$\mathcal{V} \times \mathcal{V}$. The edges in an undirected graphical model have no orientation and represent unordered pairs, i.e., an edge $(i, j) \in \mathcal{E}$ is identical to the edge $(j, i) \in \mathcal{E}$. A *path* in the graph is defined by an ordered sequence of vertices in $\mathcal{V}$ such that any two consecutive vertices in this sequence share an edge from $\mathcal{E}$. For any disjoint sets $A, B, C \subset \mathcal{V}$, set $C$ is said to *separate* $A$ and $B$ if every path from any vertex in $A$ to any vertex in $B$ passes through $C$.

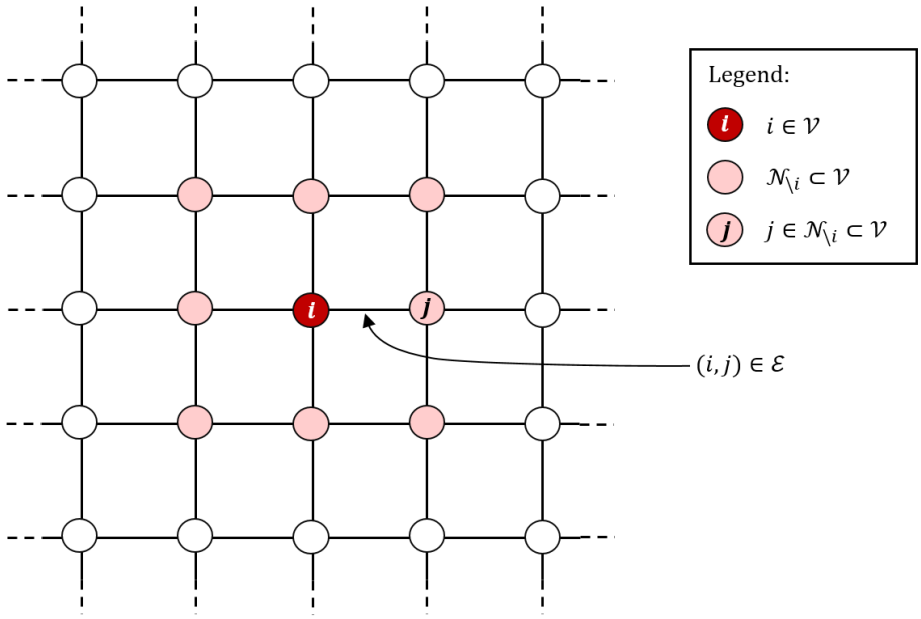


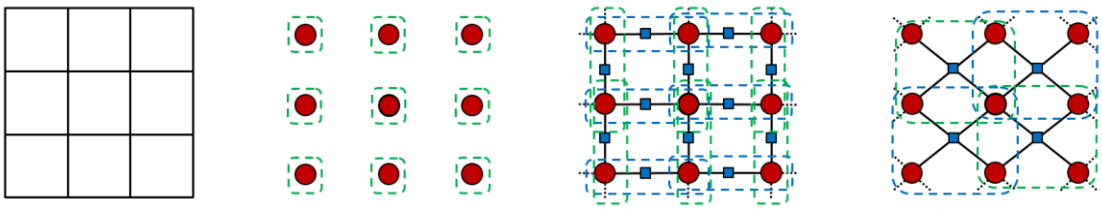

*Figure 3.7 A graphical representation of a Markov random field (MRF) where circles represent vertices $\mathcal{V}$ and the connecting lines represent edges $\mathcal{E}$ in the graph. The central dark-red circle represents any vertex $i$ under consideration and the light-red circles around it form the Markov blanket (neighbourhood) $\mathcal{N}_{\setminus i}$ of $i$. The dotted lines show possible extension of edges and the graph that is not shown in the figure. This graph contains only pairwise cliques, i.e. cliques that contain just two vertices that share an edge (as used in chapters 5 and 8 in this thesis). A more complex MRF may also involve diagonal edges, thus containing cliques of size 3 or more.*

Every vertex $i \in \mathcal{V}$ is associated with a set $\mathcal{N}_{\setminus i} \subset \mathcal{V} \setminus \{i\}$ of neighbouring vertices; these share an edge in $\mathcal{E}$ from the vertex $i \in \mathcal{V}$, and are referred to as the *neighbourhood* of $i$. So $j \in \mathcal{N}_{\setminus i}$ if and only if $(i, j) \in \mathcal{E}$. By definition, the neighbouring relationship must satisfy two properties: a vertex cannot be a neighbour of itself, i.e., $i \notin \mathcal{N}_{\setminus i}$ (as is emphasized by the subscript '$\setminus i$'), and the neighbouring relationship is commutative, i.e., $i \in \mathcal{N}_{\setminus j} \Rightarrow j \in \mathcal{N}_{\setminus i}$. The neighbourhood $\mathcal{N}_{\setminus i}$ of a vertex $i$ in a MRF is also sometimes referred to as the *Markov blanket* of $i$. We often need to consider the set $\mathcal{N}_{\setminus i} \cup \{i\}$ which is used in the rest of this document, so

in order to reduce the notational clutter we denote it with $\mathcal{N}_i$, and also refer to it as the neighbourhood of $i$ while the subscript clearly indicates whether the vertex $i$ is included in the set or not. A *neighbourhood system* $\mathcal{N}$ in graph $\mathbb{G}(\mathcal{V}, \mathcal{E})$ is defined as

$$\mathcal{N} = \left\{\mathcal{N}_{\setminus i} \subset \mathcal{V}\backslash\{i\} : \forall i \in \mathcal{V}\right\} \quad 3.5$$

A *clique* $c \subseteq \mathcal{V}$ of a graph is any subset of its vertices which are fully connected. In other words, for any two vertices $i, j \in c \subseteq \mathcal{V}$, there exists an edge between $i$ and $j$, i.e. $(i, j) \in \mathcal{E}$. A *maximal clique* $\hat{c}$ of a graph is a clique that is not a proper subset of any other clique. So $\hat{c}$ is a maximal clique of $\mathbb{G}$ if it fails to remain a clique when any additional vertex from $\mathcal{V}\backslash\hat{c}$ is added to $\hat{c}$. Thus for every clique $c$ in $\mathbb{G}$ there exists a maximal clique $\hat{c}$ in $\mathbb{G}$ such that $c \subseteq \hat{c}$. The *order* of a clique $c$, represented as $|c|$, refers to the number of vertices in $c$. The *tree width* $\omega$ of a graph is the order of its largest maximal clique, that is $\omega = \max|\hat{c}|$. The set of all of the cliques in $\mathbb{G}$ is represented by $\mathcal{C}$ and the set of all of the maximal cliques in $\mathbb{G}$ is represented by $\hat{\mathcal{C}}$, such that $\hat{\mathcal{C}} \subseteq \mathcal{C}$.



*(a)* *(b)* *(c)* *(d)*

*Figure 3.8 (a) A standard gridded (cellular) model, and (b, c & d) Markov random fields (MRF) where vertices (shown as circles) represent random variables and the edges (links between vertices) indicate probabilistic dependence between the connected vertices (or the associated random variables). In typical applications the vertices in b, c and d might represent the random variables in a gridded cellular model such as in a. Small squares on the edges represent the factors (clique potentials) in the probability distribution corresponding to the connected edges. Rounded rectangles with dashed boundaries enclose cliques in the graph. (b) A MRF with independent variables, represented by cliques/factors defined over individual variables. The neighbourhood of each cell in this case is an empty set. (c) Pairwise MRF with maximum clique size of 2. The neighbourhood of any vertex in this case consists of the four vertices that share an edge (or a pairwise factor) with that vertex. (d) A higher-order MRF (also called a cluster graph) with maximum clique size of 4. The neighbourhood of any vertex in this case consists of the surrounding eight vertices that share a factor with that vertex.*

Let us consider a random vector $\boldsymbol{m} = (\boldsymbol{m}_1, \boldsymbol{m}_2, \dots, \boldsymbol{m}_n)^T$ that represents model parameters of interest such as geological properties at the $n = |\mathcal{V}|$ locations in the model or vertices in the graph. Note that we use a boldface notation for model parameters $\boldsymbol{m}_i$ in a cell $i$, since this may generally be a vector of multiple properties of interest at each location. By definition, for any two vertices $i, j \in \mathcal{V}$ if $j \in \mathcal{N}_{\setminus i}$ then no conditional independence relationships exist or are assumed between the associated random variables $\boldsymbol{m}_i$ and $\boldsymbol{m}_j$. The random vector $\boldsymbol{m}$ forms a MRF over the graph $\mathbb{G}$ if it satisfies two properties: the *positivity property* according to which the joint probability of the random variables $\boldsymbol{m}$ is strictly positive, i.e., $\mathcal{P}(\boldsymbol{m}) > 0$, for all possible configurations of $\boldsymbol{m}$, and the *Markovian property* which requires that given the parameters $\boldsymbol{m}_{\mathcal{N}_{\setminus i}}$ in the neighbourhood $\mathcal{N}_{\setminus i}$ of a vertex $i$, the parameters $\boldsymbol{m}_i$ at $i$ become conditionally independent of the parameters in the rest of the model, i.e., $\mathcal{P}(\boldsymbol{m}_i|\boldsymbol{m}_{\setminus i}) = \mathcal{P}\left(\boldsymbol{m}_i|\boldsymbol{m}_{\mathcal{N}_{\setminus i}}\right)$, where $\boldsymbol{m}_{\setminus i} \equiv \boldsymbol{m} \setminus \{\boldsymbol{m}_i\}$ and $\boldsymbol{m}_{\mathcal{N}_{\setminus i}} \equiv \{\boldsymbol{m}_j : j \in \mathcal{N}_{\setminus i}\}$. The Markovian property implies that for any disjoint subsets $A, B, C \subset \mathcal{V}$ such that $C$ separates $A$ from $B$ in $\mathbb{G}$, we have $\boldsymbol{m}_A$ is conditionally independent of $\boldsymbol{m}_B$ given $\boldsymbol{m}_C$ represented as $(\boldsymbol{m}_A \perp\!\!\!\perp \boldsymbol{m}_B|\boldsymbol{m}_C)$, where $\boldsymbol{m}_X = \{\boldsymbol{m}_i : i \in X\}, X \in \{A, B, C\}$. Therefore, according to the Markovian property, any two unobserved vertices in a MRF are conditionally independent if all paths between them pass through the observed vertices.

### *3.5.2 Higher-order MRF*

A higher-order MRF is an undirected graphical model $\mathbb{G}(\mathcal{V}, \Psi)$, where $\mathcal{V} = \{1, \dots, n\}$ is a set of vertices which defines the topology of some physical space, and $\Psi = \{\psi_c : \mathcal{V}^{|c|} \rightarrow \mathbb{R}^+, c \subseteq \mathcal{V}\}$ is a set of non-negative potential functions $\psi_c$ defined over each clique $c$ in $\mathbb{G}(\mathcal{V}, \Psi)$, where $|c| > 2, \forall c \in \mathcal{V}$ (figure 3.8). Accordingly, a potential function $\psi_c$ is also called a *(high-order) clique potential*. For brevity, $\mathbb{G}(\mathcal{V}, \Psi)$ is represented just as $\mathbb{G}$ in the following. Similar to the pairwise MRF where each vertex $i \in \mathcal{V}$ is associated with a hidden variable $\boldsymbol{m}_i \in \boldsymbol{m}$ (e.g. representing geological properties in our model), each clique $c$ in a higher-order MRF is associated with a subset $\boldsymbol{m}_c$ of $\boldsymbol{m}$. Thus, a higher order graph models probabilistic dependence among more than just pairs of variables at a time. A higher-order potential function $\psi_c(\boldsymbol{m}_c)$ defined over local configurations $\boldsymbol{m}_c \subseteq \boldsymbol{m}$ models mutual affinity or relative compatibility of random variables $\boldsymbol{m}_c$ in $c \subseteq \mathcal{V}$, and need not be a well defined probability. Potential functions

are explained in detail in section 3.5.3 below. Finally, we define the *neighbourhood* $\mathcal{N}_c$ of a clique $c$ as a set of all of the maximal cliques $\hat{c}$ that contain it:

$$\mathcal{N}_c = \{\hat{c} \in \hat{\mathcal{C}} : c \in \hat{c}\} \tag{3.6}$$

and the *neighbourhood cardinality* $|\mathcal{N}_c|$ of a clique $c$ as the number of maximal cliques $\hat{c}$ that contain $c$.

### *3.5.3 Gibbs Distribution*

A mathematically tractable specification of a joint probability distribution over a MRF is provided by the *Hammersley-Clifford* theorem (Hammersley & Clifford, 1971) proved by Besag, 1974. It states that any joint distribution over a MRF may be expressed as a *Gibbs distribution* which takes the form

$$\mathcal{P}(\boldsymbol{m}) = \frac{1}{\mathcal{Z}} \exp\left\{-\frac{1}{T}\sum_{c\in\mathcal{C}} E_c(\boldsymbol{m}_c)\right\} \tag{3.7}$$

where $\mathcal{C}$ represents the set of cliques in the graph, $E_c(\boldsymbol{m}_c)$ represents the *energy function* of the local configurations $\boldsymbol{m}_c \subseteq \boldsymbol{m}$ of each clique $c$ in the graph $\mathbb{G}$ such that low energy states correspond to high probability configurations, $T$ is a parameter called *temperature*, and $\mathcal{Z}$ is a constant known as the *partition function* that ensures normalization of the joint distribution to be a valid probability function and is given by the integral of the numerator over the domain of $\boldsymbol{m}$, i.e.

$$\mathcal{Z} = \int_{\boldsymbol{m}} \exp\left\{-\frac{1}{T}\sum_{c\in\mathcal{C}} E_c(\boldsymbol{m}_c)\right\} d\boldsymbol{m} \tag{3.8}$$

In a MRF, the energy states of a system are conventionally expressed in the form of strictly positive potential functions over cliques, called *clique potentials* $\psi_c(\boldsymbol{m}_c)$, given by

$$\psi_c(\boldsymbol{m}_c) = \exp\left\{-\frac{E_c(\boldsymbol{m}_c)}{T}\right\} \tag{3.9}$$

such that the joint distribution over a MRF may be expressed as a product of clique potentials

$$\mathcal{P}(\boldsymbol{m}) = \frac{1}{\mathcal{Z}} \prod_{c \in \mathcal{C}} \psi_c(\boldsymbol{m}_c) \quad 3.10$$

The clique potentials $\psi_c(\boldsymbol{m}_c)$ are real-valued positive functions of local configurations $\boldsymbol{m}_c \subseteq \boldsymbol{m}$ in each clique $c$ in the graph $\mathbb{G}$.

In a pairwise MRF, the clique potentials are defined over pairwise cliques, i.e., edges from $\mathcal{E}$ in the graph. The pairwise clique potentials are functions of two neighbouring variables expressed as $\psi_{ij}(\boldsymbol{m}_i, \boldsymbol{m}_j)$ such that $(i,j) \in \mathcal{E}$. The pairwise clique potentials are also referred to as *edge potentials* for obvious reasons, and model the affinity or relative compatibility of two neighbouring random variables in a pairwise MRF. Equation 3.10 takes the following form for a pairwise MRF

$$\mathcal{P}(\boldsymbol{m}) = \frac{1}{\mathcal{Z}} \prod_{(i,j) \in \mathcal{E}} \psi_{ij}(\boldsymbol{m}_i, \boldsymbol{m}_j) \quad 3.11$$

If $\boldsymbol{m}$ is discrete, the clique potentials $\psi_c(\boldsymbol{m}_c)$ in a higher-order MRF may be defined by scanning a training image and building histograms for various configurations of $\boldsymbol{m}_c$ over pixels in a clique $c$ with offset distances and direction depending on the clique structure. For a pairwise MRF, the edge potentials $\psi_{ij}(\boldsymbol{m}_i, \boldsymbol{m}_j)$ may be defined in a similar manner by building histograms for various combinations of $\boldsymbol{m}_i$ and $\boldsymbol{m}_j$ over pixels in a training image where offset and direction between locations $i$ and $j$ depend on the size and shape of the neighbourhood structure. For example, a histogram of geological facies (discrete variables) is built by counting the occurrence of any two facies in laterally or vertically adjacent pixels in the training image. These counts are then normalized over all possible combinations of facies within the same configuration of pixels across the training image to give prior probabilities. This assigns zero probability to configurations of facies that are geologically implausible, such as brine directly over gas.

For continuous $\boldsymbol{m}$, the potential functions are defined using an explicit functional form of the PDF of $\boldsymbol{m}$. For example, a Gaussian distribution may be used for continuous variables $\boldsymbol{m}$ a pairwise MRF that is also referred to as a *Gaussian Markov random field* (GMRF, Rue & Held, 2005); or a more general *Gaussian mixture* (GM) distribution in any MRF, also referred to as a *Gaussian mixture Markov random field* (GM-MRF, e.g. see Zhang *et al*. 2016).

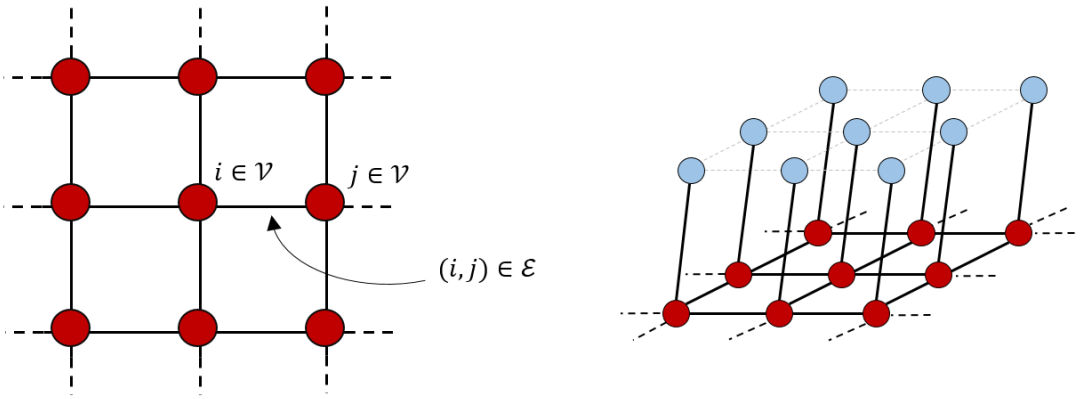
It is important to note here that a MRF that defines clique potentials over higher-order cliques in a graph is more expressive and can model more complex features than a pairwise MRF (Koller & Friedman, 2009). This concept is similar to that of multi-point statistics (MPS) based prior information used in Geostatistics to model complex geological features such as meandering channels in a deltaic environment, compared to two-point statistics (such as covariance) based prior information which may not model such complex features adequately (e.g. Strebelle, 2001; Arpat, 2005; Journel & Zhang, 2006; González *et al*. 2008; Mariethoz & Caers, 2014, Tahmasebi, 2018).

## *3.6 Hidden Markov Random Field (MRF)*

A variant of a MRF known as a *hidden Markov random field* (HMRF) which includes vertices that are fixed and represent observed data, in addition to the unobserved vertices that follow a MRF model (figure 3.9). Thus, each vertex in a HMRF represents either an observed or an unobserved (or hidden) random variable. Similar to the notation used for model parameters $\boldsymbol{m}$ in section 3.5.1, we define a set of observed variables or data $\boldsymbol{d} = (\boldsymbol{d}_1, \boldsymbol{d}_2, \dots, \boldsymbol{d}_n)^T$, where data $\boldsymbol{d}_i$ at a location $i$ is denoted with a boldface font because it may be a vector containing multiple measurements, e.g. multiple seismic attributes such as P-wave and S-wave impedances measured at the same location $i$. A HMRF may be visualized as consisting of two layers, where the upper layer contains the observed variables $\boldsymbol{d}$, and the lower layer contains the hidden variables $\boldsymbol{m}$ (figure 3.9). A HMRF essentially requires the CI assumption (equation 3.3). Additionally, the LL assumption (equation 3.4) is also commonly used in HMRF models (figure 3.9, also see section 1.1.4).

A MRF (or HMRF) is a preferred model of spatial distribution of geological properties due to its desirable properties such as the Markovian property due to two main reasons. First, a MRF allows capturing the patterns of geological parameters typically observed directly in geological outcrops, or indirectly through geophysical measurements taken at the surface or earth or in a borehole. Second, the Markovian property of a MRF limits the amount of computations required for probabilistic inference significantly. The commonly used LL and CI assumptions in a HMRF further limit the computational complexity of solving a spatial inverse problem. However, as discussed in section 1.1.4, these assumptions are unrealistic and may

introduce errors in solutions. The latter is demonstrated with synthetic data examples in sections 5.6.1 and 6.6.2.




*(a)* *(b)*

*Figure 3.9 (a) a Markov random field (MRF), and (b) a typical hidden Markov random field (HMRF), where vertices (shown as circles) represent random variables and the edges (links between vertices) indicate probabilistic dependence between the connected vertices (or the associated random variables). Red circles represent hidden vertices or unobserved variables (model parameters) and the blue circles represent observed vertices (data). A typical HMRF assumes localized likelihoods (LL) where each observed variable depends only on the unobserved variable at the same location. Both (a) and (b) represent pairwise MRFs with a maximum clique size of 2. The neighbourhood of any hidden vertex (red circle) in this case consists of a maximum of four vertices that share an edge with that vertex.*

## 3.7 Structure of a PGM

The structure of a graphical model is designed based on the expected range and density (or comparative sparsity) of probabilistic dependencies or statistical correlations among the variables of interest. The graphical models used in physics are often designed based on prior information derived from scientific theory and models (e.g. *Ising* and *Pott's* models in statistical physics, Baxter, 1989). In information and communication theory as used in statistics and computer science, the graphical models are learnt from the data. In geosciences, both of these avenues are open to us. For instance, in the case of horizontally layered geological strata we expect longer range correlations in rock properties in the lateral dimension than in the vertical dimension. Such geological knowledge can be learnt from experts and parametrized as

discussed in the Introduction section, and injected in Bayesian inversions in the form of prior information. On the other hand, to model the intrinsic variability of geological properties or spatial patterns of facies within a stratum, we could decide to learn the graphical model as part of the method of solution, by maximizing the log-likelihood of observed data with respect to the model parameters.

A MRF can be used to model any distribution that takes the form of a Gibbs distribution (by the Hammersley Clifford theorem). Accurate definition of a Gibbs distribution requires estimation or design of the graphical model which may be combined with parameter estimation during training. The parameterization must then be chosen such that simpler distributions which are sufficiently expressive are preferred over more complex and possibly more expressive distributions. Such an approach aims to satisfy two competing goals: that the chosen distribution is sufficiently structured to capture the desired details in the target distribution, and that it allows tractable inference from the chosen distribution.

In this thesis, it is assumed that the structure of spatial dependence among neighbouring locations, i.e. structure of the MRF, is known *a priori* and is fixed. A more general approach in inversion would be to estimate the structure of the MRF along with the desired parameters, which is proposed as a potential future extension of this work in section 9.5.2.

## 3.8 Synergy between Geology and Statistical Physics

The MRF model has its origin in statistical physics where it was introduced to model the energy states of a large number of mutually interacting particles which exhibit a stochastic behavior, but where their mutual interactions obey some natural rules. For example, a natural system commonly prefers lower energy states and it continuously updates the local energy states of the particles that compose the system until the system attains the lowest energy state. Local energy states of particles depend only on their interactions with neighbouring particles. Such a behavior can be modelled with a MRF called an *Ising* or *Pott's model* as this provides a mathematical specification of any joint distribution over a large number of particles by exploiting the conditional independence among most of the (non-neighbouring) particles. We use our MRF model to parameterize the prior information on geological facies as embodied within a training image, since heterogeneity typically observed in geology may be

assumed to be globally random while the facies in neighbouring locations are more likely to be similar than those in the distant cells.

In the context of inversion of geophysical data for geological properties, a MRF is used to specify the prior information about the spatial distribution of facies. A MRF is a useful model in spatial statistics as it decomposes probabilistic dependence among various random variables into selected subsets (cliques in a graph) which adequately capture natural spatial correlations among these variables by exploiting conditional independence among them due to their relative spatial locations.

# Chapter 4 Bayesian Inversion using a Hidden Markov Model

## 4.1 Summary

An efficient method for Bayesian inversion of categorical variables, such as geological facies, using a hidden Markov model (HMM) is proposed that does not require stochastic sampling. A new 2D HMM is introduced over a grid of cells where observations represent localized data constraining each cell. The data represents seismic attributes such as P-wave and S-wave impedances; categorical variables are the hidden states and represent the geological rock types in each cell – facies of distinct subsets of lithology and fluid combinations such as shale, brine-sand and gas-sand. The observations at each location are assumed to be generated from a random function of the hidden state (facies) at that location, and to be distributed according to a certain probability distribution that is independent of hidden states at other locations – the so called localized likelihoods assumption. The facies at a location cannot be determined solely by the data at that location as it also depends on the spatial distribution of facies elsewhere, which is injected in the form of prior information presented in the form of a training image. The prior information presented in other forms can be used in the method as desired. The method provides direct estimates of posterior marginal probability distributions in each model cell, so these do not need to be estimated from samples such as in McMC. On a 2D test example the method is shown to outperform previous methods significantly, at a fraction of the computational cost of these methods. In many foreseeable applications there are no serious impediments to extending the method to 3D cases.

## 4.2 Introduction

There is always uncertainty in the estimation of geological facies from the observed data at a given location. The uncertainty is either due to uncertainty in the measurement of geophysical data or due to the intrinsic uncertainty in the relationship between facies and the data, or both. This implies that the data inferred at a given location are related to a certain facies at that location according to some probability distribution. Although the actual

observation in a geophysical experiment is the raw data, the inferred attributes (such as seismic attributes) are referred to as 'observed data' or 'observations' herein. This explicitly distinguishes them from the geological facies which are treated as 'hidden' (not observed) variables. The probability of observing (or inferring) a specific set of data at a fixed location given that a particular facies exists at that location, is called the data *likelihood*. Uncertainty in the attributes is accounted for within the likelihood. Since spatial correlation of facies is controlled by the prior probabilities it is commonly assumed that the data likelihood is localized (see e.g., Larsen *et al.* 2006; Ulvmoen & Omre, 2010; Ulvmoen *et al.* 2010; Walker & Curtis, 2014a): that is, given the facies at a location, the data at that location are conditionally independent of both facies and the data at all other locations. This assumption is henceforth referred to as the condition of *localized likelihoods*.

The contextual information expressed as prior probabilities of spatial correlations of facies may be combined with the local information provided by likelihood probabilities in a Bayesian framework. Thus we obtain posterior probabilities of the spatial distribution of geological facies that conform to both the observed data (e.g. seismic attributes) and prior constraints. However, a major problem is that the full posterior distribution is usually analytically intractable for standard high-dimensional models and must be explored through simulation and sampling based inference, e.g., by using Markov-chain Monte Carlo (McMC) methods. As discussed in section 1.1.2, McMC based sampling is computationally demanding as it requires many samples to converge to the true distribution.

Walker & Curtis (2014a) developed a method for Bayesian inversion of two-dimensional spatial data using an exact sampling alternative to McMC. This allows independent samples of the target distribution to be drawn without requiring convergence, thus circumventing convergence related biases. Their algorithm requires large memory and is computationally intensive for real-scale seismic data and geological modelling problems. If distribution functions such as marginals of the posterior distribution in each model cell are required, these must then be calculated from the set of samples generated. A different approach is taken in this research: the marginal posterior distributions of facies in each cell in the model are computed directly. This incorporates prior geological information and the data likelihood in a similar manner to Walker & Curtis (2014a). However, computation of marginal posterior distributions in each cell in the model is computationally more efficient and requires less memory.

In this chapter, some definitions and notation are first introduced which are used in the rest of the chapter. These definitions allow a 1D-like treatment of the 2D-HMM, while fully acknowledging the higher dimensional spatial dependence among cells in the model. Analytical expressions are derived for marginal posterior distributions at each location in the model given the data and the neighbouring geological facies. Then test results of computing marginal probability distributions are presented from an application of this method to a synthetic geological model of siliciclastic-filled river channels in shale, with three geological facies – shale, brine-sand and gas-sand. A brief discussion is finally provided comparing this method with previous research with reference to the test results, before concluding.

## 4.3 Model

Previous work in the field of geostatistical inversion used Markov-chains and hidden Markov models for inversion of seismic data for geological facies (e.g., Larsen et al. 2006; Ulvmoen & Omre 2010; Ulvmoen et al. 2010; Hammer & Tjelmeland 2011; Rimstad & Omre 2013; Lindberg & Omre 2014 & 2015). Larsen et al. (2006) inverted pre-stack seismic data using a 1D Markov-chain prior model of lithology-fluid classes along vertical profiles through a reservoir zone. Ulvmoen & Omre (2010) and Ulvmoen et al. (2010) extended the model of Larsen et al. (2006) by introducing lateral alignments among neighbouring 1D vertical Markov-chains to model lateral coupling of lithology-fluid classes as commonly found in geological strata. Such a graphical structure is called a *profile Markov random field* (see e.g., Eddy 1998). Rimstad & Omre (2013) also used a profile Markov random field based prior but with a different parametrization. Rimstad et al. (2012) inverted seismic AVO data for lithology/fluid classes, elastic properties and porosity using a MRF prior model. Lindberg & Omre (2015) used a convolved two-level 1D-HMM for inversion of categorical variables (such as lithology-fluid classes) represented as the bottom hidden-layer of the model, continuous system response variables (such as reflection coefficients) represented as the middle hidden-layer, and the measured convolved data represented in the observation layer. A common feature among all of these approaches for facies inversion is that they are based on inference from full posterior distribution which must be explored through simulation (sampling) based inference, e.g., using McMC methods.

By contrast, analytic expression for marginal posterior distributions of geological facies conditioned on the seismic attribute data are derived in this chapter using a 2D-HMM (see

section 4.5) which is computationally efficient by many orders of magnitude compared to the previous research on the same problem under the same set of assumptions.

### *4.3.1 2D Hidden Markov Model (2D-HMM)*

Many extensions of hidden Markov-chains to 2D have been proposed in the literature for applications to 2D data such as images in computer vision, but these either convert 2D data into 1D and then apply a pseudo-2D approach (e.g., Abend et al. 1965; Daleno et al. 2010; Ma et al. 2008; Bevilacqua et al. 2007), or attempt to obtain approximate results by introducing assumptions which limit the spatial dependence among neighbouring cells (locations) in the model (e.g., Li et al. 2000; Othman & Aboulnasr 2003; Baumgartner et al. 2013). The main contribution of this research is that it presents analytic, closed-form solutions for approximate marginal posterior distributions of hidden states conditioned on the observed data using a 2D-HMM which incorporates the full 2D coupling of hidden states.

A *2D hidden Markov model (2D-HMM)* may be designed over a rectangular two-dimensional grid where hidden states correspond to the geological facies, and observations correspond to localized data (seismic attributes such as P-wave and S-wave impedances). The hidden states in a 2D-HMM form a special case of a Markov random field (MRF), called a *hidden Markov mesh* or a *causal MRF* (Abend et al. 1965; Cressie & Davidson, 1998). Causality is induced in the grid by directional conditional dependence among the cells in the model, and allows the analytic derivation of marginal posterior distributions.

Herein a 2D-HMM is represented as a directed graph $\mathbb{G}(\mathcal{V}, \mathcal{E})$ where vertices $\mathcal{V}$ are defined over a rectangular grid of cells and edges $\mathcal{E}$ represent horizontal, vertical and diagonal dependence between neighbouring cells in the grid. In a 2D-HMM, we use double indexing to represent vertices such as $(i, j) \in \mathcal{V}$, where $i$ runs vertically (row-wise) and $j$ runs horizontally (column-wise) in the 2D grid. Similarly, $\boldsymbol{d}_{ij}$ represents a vector of data values in cell $(i, j)$; and regular small letters are used for scalar variables, for example, $\kappa_{ij}$ represents geological facies in cell $(i, j)$. However, at a later stage in this chapter, the notion of a partition of the set of cells is introduced that is represented by $P$, and a linear indexing of cells is used within a partition, such that $\kappa_{P,i}$ (note the comma in the subscript) represents the geological facies in the $i^{\text{th}}$ cell within the partition $P$. Ordered relationships between cells, such as $<$ and $>$, are defined as described in the text below.

Since there is a one-to-one mapping between each cell $(i,j) \in \mathcal{V}$ in the model and the corresponding hidden state $\kappa_{ij}$, we may denote the cells in the model with the corresponding hidden state so that we may use the same notation $\boldsymbol{K}$ to denote the set of vertices in the graph as well as the set of geological facies in the model. With 2D indexing, let us denote the neighbourhood of a vertex $(i,j) \in \mathcal{V}$ as $\mathcal{N}(i,j)$. The definition of neighbourhood in a graph implies that given the facies $\boldsymbol{\kappa}_{\mathcal{N}(i,j)}$ in the neighbourhood of a vertex $(i,j) \in \mathcal{V}$ the facies $\kappa_{ij}$ at $(i,j)$ is conditionally independent of all other facies outside of its neighbourhood

$$\mathcal{P}\left(\kappa_{ij}|\boldsymbol{\kappa}_{\setminus ij}\right) = \mathcal{P}\left(\kappa_{ij}|\boldsymbol{\kappa}_{\mathcal{N}(i,j)}\right) \tag{4.1}$$

where $\boldsymbol{\kappa}_{\setminus ij}$ represents the set of geological facies in all cells in the model except $(i,j)$.

Now define a *partition element* or simply a *partition* as a non-empty ordered set of vertices (cells in the model) where each vertex is a neighbour of the next vertex in the set, and the first and last vertices in the set lie at the boundary of the model. A non-empty ordered set of disjoint partitions can be defined such that all of the neighbours of any cell in one partition lie either in the same, the previous or the next partition. Such a family of non-empty, ordered, disjoint partition elements defines a *partition family* over the (graphical) model.

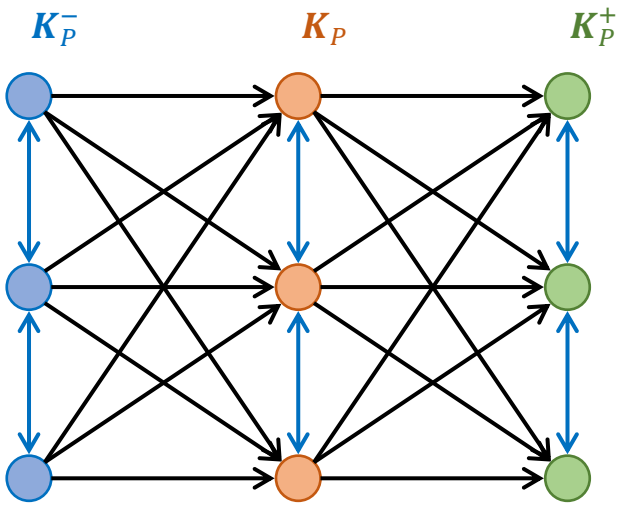



*Figure 4.1: An example of a causal 2D-HMM. The arrow directions represent the directions of probabilistic dependence between various cells (circles) in the model. Circles shown in blue, orange, and green colour represent cells in the past partition $\boldsymbol{K}_P^-$, the current partition $\boldsymbol{K}_P$, and the future partition $\boldsymbol{K}_P^+$, respectively. Bi-directional blue coloured arrows represent acausal dependence between nodes within the same partition, i.e. no directionality holds in this case.*

## 4.4 Marginal Posterior Distribution in a 2D-HMM

In order to derive a recursive formulation of the marginal posterior distribution $\mathcal{P}(\kappa_{ij}|\boldsymbol{D})$ of facies $\kappa_{ij}$ in a cell $(i,j)$ conditioned to data $\boldsymbol{D}$, define a partition $\boldsymbol{K}_P$ as a set of cells ordered from 1 to $n$, such that $\kappa_{ij} \in \boldsymbol{K}_P$; that is, $\exists\, \kappa_{P,k} \in \boldsymbol{K}_P$ such that $\kappa_{P,k} = \kappa_{ij}$, for some $k$, and

$$\boldsymbol{K}_P = \left\{ \kappa_{P,1}, \kappa_{P,2}, \dots, \kappa_{P,k}, \dots, \kappa_{P,n} : \kappa_{P,k} = \kappa_{ij} \text{ for some } k \right\} \quad 4.2$$

Just as the notation with double letters $ij$ in the subscript represents the location of a cell in the model, the notation $\kappa_{P,\cdot}$ (with a $P,\cdot$ in the subscript) is used in equation 4.2 to represent ordering of cells within the partition $\boldsymbol{K}_P$. So by definition of a partition $\kappa_{P,1} \in \mathcal{N}(\kappa_{P,2})$, $\kappa_{P,2} \in \mathcal{N}(\kappa_{P,3})$, … , $\kappa_{P,n-1} \in \mathcal{N}(\kappa_{P,n})$.

A partition may be defined as a row, a column or an arbitrary set of cells that satisfies equation 4.2. The shape of the partition is chosen with consideration of computational convenience, the size and shape of the computational model, and the neighbourhood structure. It is preferable to define the partition along the shorter dimension of the model in order to limit the partition size, as the memory required to store the joint distribution of facies within a partition may grow significantly with the partition size.

Define $\boldsymbol{K}_P^-$ as the set of cells which constitute the immediate past of the partition $\boldsymbol{K}_P$ based on the direction induced by causality (figures 4.1 and 4.2)

$$\boldsymbol{K}_P^- = \left\{ \kappa_{kl} : \exists\, \kappa_{ij} \in \boldsymbol{K}_P, \text{ such that } \kappa_{kl} \rightarrow \kappa_{ij} \in \mathcal{E} \right\} \quad 4.3$$

Similarly, define $\boldsymbol{K}_P^+$ as the set of cells which constitute the immediate future of the partition $\boldsymbol{K}_P$ based on the direction induced by causality (figures 4.1 and 4.2)

$$\boldsymbol{K}_P^+ = \left\{ \kappa_{kl} : \exists\, \kappa_{ij} \in \boldsymbol{K}_P, \text{ such that } \kappa_{ij} \rightarrow \kappa_{kl} \in \mathcal{E} \right\} \quad 4.4$$

By definition $\boldsymbol{K}_P^- \cap \boldsymbol{K}_P = \boldsymbol{K}_P \cap \boldsymbol{K}_P^+ = \boldsymbol{K}_P^- \cap \boldsymbol{K}_P^+ = \emptyset$ and $\boldsymbol{K} = \bigcup_P \boldsymbol{K}_P, \forall\, P$, where $\cap$ represents intersection, $\cup_P$ represents union over all P, and $\emptyset$ is the empty set.

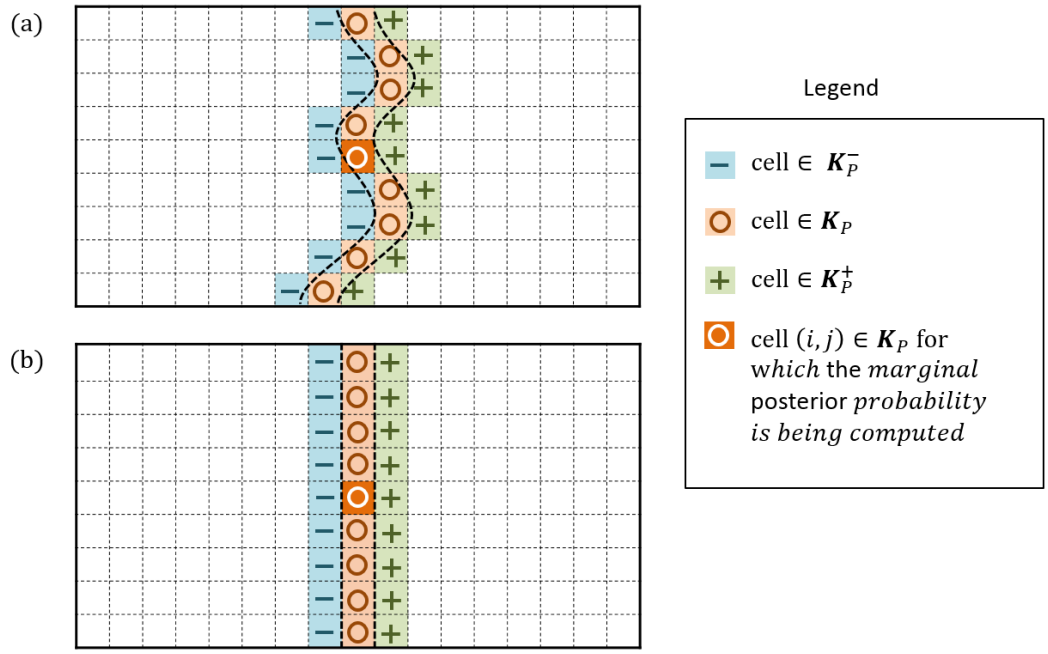



*Figure 4.2: Examples of a partition defined over a graphical model as a set of nodes that divides the model into two non-overlapping parts. An ordered set of such partitions defines a partition family over the graphical model. (a) An arbitrary partition, and (b) a partition (element) defined over a column (or the shorter dimension) in the model. The cells shown with symbols "-", "o" and "+" form the previous (past) partition $\boldsymbol{K}_P^-$, the current partition $\boldsymbol{K}_P$, and the next (future) partition $\boldsymbol{K}_P^+$, respectively. The dark-orange coloured cell with symbol "o" represents the cell $(i,j) \in \boldsymbol{K}_P$ for which marginal posterior probability is being computed.*

Also, define $\boldsymbol{K}_{\leq P}$ as

$$\boldsymbol{K}_{\leq P} = \left\{ \kappa_{kl} : \exists\, \kappa_{ij} \in \boldsymbol{K}_P, \text{such that } \kappa_{kl} \leq \kappa_{ij} \right\} \tag{4.5}$$

It follows that $\boldsymbol{K}_P \subset \boldsymbol{K}_{\leq P}$. Similarly, define $\boldsymbol{K}_{>P}$ as

$$\boldsymbol{K}_{>P} = \left\{ \kappa_{kl} : \exists\, \kappa_{ij} \in \boldsymbol{K}_P, \text{such that } \kappa_{ij} < \kappa_{kl} \right\} \tag{4.6}$$

Figure 4.2 shows examples of partitions $\boldsymbol{K}_P^-$, $\boldsymbol{K}_P$ and $\boldsymbol{K}_P^+$ defined (a) arbitrarily, and (b) as a column of cells in the model. Figure 4.3 shows the regions corresponding to $\boldsymbol{K}_{\leq P}$ and $\boldsymbol{K}_{>P}$ for a partition defined as a column of cells in the model. Similarly define

$$\boldsymbol{D}_P = \{\, \boldsymbol{d}_{kl} : \exists\, \kappa_{kl} \in \boldsymbol{K}_P, \text{ such that } \mathcal{P}(\boldsymbol{d}_{kl}|\boldsymbol{K}) = \mathcal{P}(\boldsymbol{d}_{kl}|\kappa_{kl})\} \quad 4.7$$

$$\boldsymbol{D}_{\leq P} = \{\, \boldsymbol{d}_{kl} : \exists\, \kappa_{kl} \in \boldsymbol{K}_{\leq P}, \text{ such that } \mathcal{P}(\boldsymbol{d}_{kl}|\boldsymbol{K}) = \mathcal{P}(\boldsymbol{d}_{kl}|\kappa_{kl})\} \quad 4.8$$

$$\boldsymbol{D}_{>P} = \{\, \boldsymbol{d}_{kl} : \exists\, \kappa_{kl} \in \boldsymbol{K}_{>P}, \text{ such that } \mathcal{P}(\boldsymbol{d}_{kl}|\boldsymbol{K}) = \mathcal{P}(\boldsymbol{d}_{kl}|\kappa_{kl})\} \quad 4.9$$

A key assumption in computing marginal posterior distributions using a 2D-HMM is that $\boldsymbol{D}_{>P}$ and $\boldsymbol{D}_{\leq P}$ are conditionally independent given the facies $\boldsymbol{K}$.

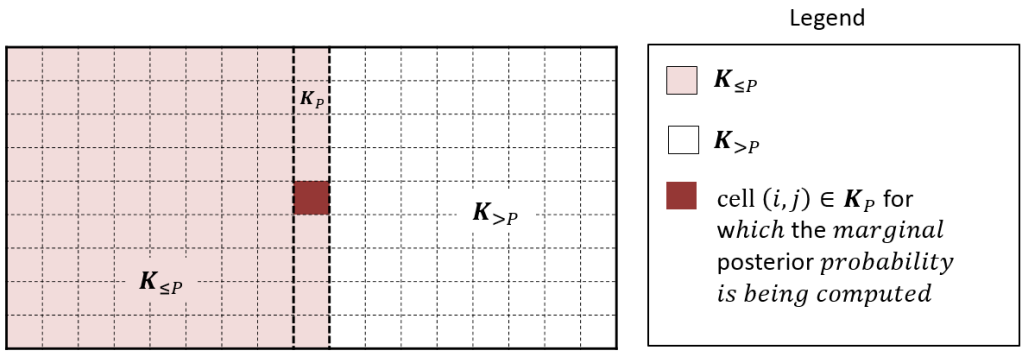



*Figure 4.3: Illustration of partitions $\boldsymbol{K}_{\leq P}$ and $\boldsymbol{K}_{>P}$. The cells with dashed border represent the partition $\boldsymbol{K}_P$, which are also part of partition $\boldsymbol{K}_{\leq P}$.*

## *4.4.1 Conditional Dependence between Partitions*

Due to causality, the probability of a facies being present in a given cell depends on the facies in the previous partition $\boldsymbol{K}_{\bar{P}}$ as well as in the current partition $\boldsymbol{K}_P$. Such a dependence can be computed from the conditional probabilities of facies in the current partition $\boldsymbol{K}_P$ given the facies in the previous partition $\boldsymbol{K}_{\bar{P}}$. From the above definitions it follows that (allowing for different numbers of cells in $\boldsymbol{K}_P$ and $\boldsymbol{K}_{\bar{P}}$ in general)

$$\mathcal{P}(\boldsymbol{K}_P|\boldsymbol{K}_{\bar{P}}) = \mathcal{P}\left(\kappa_{P,1}, \kappa_{P,2}, \ldots, \kappa_{P,n}|\kappa_{\bar{P},1}, \kappa_{\bar{P},2}, \ldots, \kappa_{\bar{P},r}\right) \quad 4.10$$

where $\kappa_{P,\cdot}$ and $\kappa_{\bar{P},\cdot}$ represent the ordering of elements within partitions $\boldsymbol{K}_P$ and $\boldsymbol{K}_{\bar{P}}$ respectively. Then

$$\mathcal{P}(\boldsymbol{K}_P|\boldsymbol{K}_{\bar{P}}^-) = \mathcal{P}\left(\kappa_{P,1} \mid \kappa_{P,2}, \dots, \kappa_{P,n}, \kappa_{\bar{P},1}^-, \kappa_{\bar{P},2}^-, \dots, \kappa_{\bar{P},r}^-\right) \cdot \mathcal{P}\left(\kappa_{P,2}, \dots, \kappa_{P,n} \mid \kappa_{\bar{P},1}^-, \kappa_{\bar{P},2}^-, \dots, \kappa_{\bar{P},r}^-\right)$$

$$= \mathcal{P}\left(\kappa_{P,1}|\boldsymbol{K}_{\mathcal{N}}^-(P,1)\right) \cdot \mathcal{P}\left(\kappa_{P,2}, \dots, \kappa_{P,n} \mid \kappa_{\bar{P},1}^-, \kappa_{\bar{P},2}^-, \dots, \kappa_{\bar{P},r}^-\right)$$

$$= \mathcal{P}\left(\kappa_{P,1}|\boldsymbol{K}_{\mathcal{N}}^-(P,1)\right) \cdot \mathcal{P}\left(\kappa_{P,2}|\kappa_{P,3}, \dots, \kappa_{P,n}, \kappa_{\bar{P},1}^-, \kappa_{\bar{P},2}^-, \dots, \kappa_{\bar{P},r}^-\right) \cdot \mathcal{P}\left(\kappa_{P,3}, \dots, \kappa_{P,n} \mid \kappa_{\bar{P},1}^-, \kappa_{\bar{P},2}^-, \dots, \kappa_{\bar{P},r}^-\right)$$

$$= \mathcal{P}\left(\kappa_{P,1}|\boldsymbol{K}_{\mathcal{N}}^-(P,1)\right) \cdot \mathcal{P}\left(\kappa_{P,2}|\boldsymbol{K}_{\mathcal{N}}^-(P,2)\setminus\{\kappa_{P,1}\}\right) \cdot \mathcal{P}\left(\kappa_{P,3} \mid \kappa_{P,4}, \dots, \kappa_{P,n}, \kappa_{\bar{P},1}^-, \kappa_{\bar{P},2}^-, \dots, \kappa_{\bar{P},r}^-\right)$$

$$= \mathcal{P}\left(\kappa_{P,1}|\boldsymbol{K}_{\mathcal{N}}^-(P,1)\right) \cdot \mathcal{P}\left(\kappa_{P,2}|\boldsymbol{K}_{\mathcal{N}}^-(P,2)\setminus\{\kappa_{P,1}\}\right) \cdot \mathcal{P}\left(\kappa_{P,3}|\boldsymbol{K}_{\mathcal{N}}^-(P,3)\setminus\{\kappa_{P,1},\kappa_{P,2}\}\right) \dots \cdot \mathcal{P}\left(\kappa_{P,n}|\boldsymbol{K}_{\mathcal{N}}^-(P,n)\setminus\{\kappa_{P,1},\ \kappa_{P,2}, \dots, \kappa_{P,n-1}\}\right)$$

$$= \prod_{i=1}^{n} \mathcal{P}\left(\kappa_{P,i}|\boldsymbol{K}_{\mathcal{N}}^-(P,i)\setminus\{\kappa_{P,<i}\}\right) \qquad 4.11$$

where $\{\kappa_{P,<i}\} = \{\kappa_{P,h}: h < i\}$ and $\boldsymbol{K}_{\mathcal{N}}^-(P,i) = (\boldsymbol{K}_{\bar{P}}^- \cup \boldsymbol{K}_P) \cap \mathcal{N}(P,i)$.

The conditional probabilities on the right-hand side of equation 4.11 represent the prior information on the spatial correlation of geological facies. These can be computed directly from the patterns of facies distributions depicted in the training image which correspond to the various shapes and sizes of partitions and the neighbourhood structure. As an example, see Toftaker & Tjelmeland (2013) for a proposed method to build a prior model from a training image using a binary MRF and its partially ordered approximation, and Arnesen & Tjelmeland (2016) for a proposed prior distribution for parameters and structure of a binary MRF. For spatial inversion of geological facies, the spatial correlations read from the training image are assumed to be stationary, i.e., they are assumed to be independent of location within the model.

## 4.5 Derivation of Marginal Posterior Distribution

The idea of a partition $\boldsymbol{K}_P$ thus imposes a natural ordering which (in the following) allows 1D-like treatment of the underlying 2D-HMM while fully acknowledging the 2D structure of probabilistic dependence between cells in the model. Using the above definitions, we can derive the recursive formulation for the marginal posterior distribution conditioned to the data $\boldsymbol{D}$ because $\mathcal{P}(\kappa_{ij}|\boldsymbol{D}) \propto \mathcal{P}(\kappa_{ij}, \boldsymbol{D})$ since the data $\boldsymbol{D}$ is measured and fixed. Setting $\kappa_{ij} = \kappa_{P,q}$,

$$\mathcal{P}(\kappa_{ij}|\boldsymbol{D}) \propto \mathcal{P}(\kappa_{P,q}, \boldsymbol{D})$$

$$= \sum_{\boldsymbol{K}_P \setminus \{\kappa_{P,q}\}} \mathcal{P}(\boldsymbol{K}_P, \boldsymbol{D})$$

[by definition of a marginal distribution over $\kappa_{P,q}$]

$$= \sum_{\boldsymbol{K}_P \setminus \{\kappa_{P,q}\}} \mathcal{P}(\boldsymbol{K}_P, \boldsymbol{D}_{\leq P}) P(\boldsymbol{D}_{>P}|\boldsymbol{K}_P)$$

[since $\boldsymbol{D}_{\leq P}$ is independent of $\boldsymbol{D}_{>P}$]

$$= \sum_{\boldsymbol{K}_P \setminus \{\kappa_{P,q}\}} \alpha(\boldsymbol{K}_P)\, \beta(\boldsymbol{K}_P) \qquad 4.12$$

where $\alpha(\boldsymbol{K}_P) = \mathcal{P}(\boldsymbol{K}_P, \boldsymbol{D}_{\leq P})$ and $\beta(\boldsymbol{K}_P) = \mathcal{P}(\boldsymbol{D}_{>P}|\boldsymbol{K}_P)$ are the equivalent 2D forward and backward probabilities as those used for 1D hidden Markov-chains in the dynamic programming based algorithms of Baum (1972), Baum and Petrie (1966), Baum et al. (1970), Viterbi (1967) and Forney Jr. (1973). Note that the summation in equation 4.12 represents summations over all of the cells $\kappa_{P,\cdot}$ in the partition $\boldsymbol{K}_P$ except the cell $\kappa_{ij} = \kappa_{P,q}$.

Since $\boldsymbol{K}_P$'s, by definition, form a partition over the model space, $\alpha(\boldsymbol{K}_P)$ can be expressed using the recursive formulation of Baum's forward-backward algorithm (*Baum 1972*) for a 1D hidden Markov-chain as

$$\alpha(\boldsymbol{K}_P) = \mathcal{P}(\boldsymbol{K}_P, \boldsymbol{D}_{\leq P})$$

$$= \mathcal{P}(\boldsymbol{D}_P|\boldsymbol{K}_P) \sum_{\boldsymbol{K}_P^-} \mathcal{P}(\boldsymbol{K}_P|\boldsymbol{K}_P^-)\, \alpha(\boldsymbol{K}_P^-) \tag{4.13}$$

where summation is over all of the facies in all of the cells in $\boldsymbol{K}_P^-$. On substitution from equation 4.11 for $\mathcal{P}(\boldsymbol{K}_P|\boldsymbol{K}_P^-)$ and assuming localized likelihoods, equation 4.13 takes the form

$$\alpha(\boldsymbol{K}_P) = \prod_{i=1}^{n} \mathcal{P}\left(\boldsymbol{d}_{P,i}|\kappa_{P,i}\right) \cdot \sum_{\boldsymbol{K}_P^-} \left( \prod_{j=1}^{n} \mathcal{P}\left(\kappa_{P,j}|\boldsymbol{K}_{\mathcal{N}}^-(P,j) \setminus \{\kappa_{P,<j}\}\right) \right) \cdot \alpha(\boldsymbol{K}_P^-) \tag{4.14}$$

The factors $\mathcal{P}(\boldsymbol{D}_P|\boldsymbol{K}_P) = \prod_{i=1}^{n} \mathcal{P}\left(\boldsymbol{d}_{P,i}|\kappa_{P,i}\right)$ in equation 4.14 represent the data likelihood given the geological facies in each cell assuming localized likelihoods. The data likelihood is given by the probabilistic forward model and is explained in the next section.

Similarly, $\beta(\boldsymbol{K}_P)$ can be expressed in a recursive formulation as

$$\beta(\boldsymbol{K}_P) = \mathcal{P}(\boldsymbol{D}_{>P}|\boldsymbol{K}_P)$$

$$= \sum_{\boldsymbol{K}_P^+} \mathcal{P}(\boldsymbol{K}_P^+|\boldsymbol{K}_P)\, \mathcal{P}(\boldsymbol{D}_P^+|\boldsymbol{K}_P^+)\, \beta(\boldsymbol{K}_P^+) \tag{4.15}$$

where the summation is over all possible combinations of facies in all of the cells in $\boldsymbol{K}_P^+$. On substitution from equation 4.11 for $\mathcal{P}(\boldsymbol{K}_P^+|\boldsymbol{K}_P)$ and assuming localized likelihoods, equation 4.15 takes the form

$$\beta(\boldsymbol{K}_P) = \sum_{\boldsymbol{K}_P^+} \left\{ \left( \prod_{i=1}^{n} \mathcal{P}\left(\boldsymbol{d}_{P^+,i}|\kappa_{P^+,i}\right) \right) \cdot \left( \prod_{j=1}^{n} \mathcal{P}\left(\kappa_{P^+,j}|\boldsymbol{K}_{\mathcal{N}}^-(P^+,j) \setminus \{\kappa_{P^+,>j}\}\right) \right) \cdot \beta(\boldsymbol{K}_P^+) \right\} \tag{4.16}$$

Substituting equations 4.14 and 4.16 into equation 4.12 gives a recursive formulation for the marginal posterior distribution in a given cell in the model. $\alpha(\boldsymbol{K}_P)$ in equation 4.14 is

computed in the forward direction (increasing $P$) while $\beta(\boldsymbol{K}_P)$ in equation 4.16 is computed in the backward direction (decreasing $P$). This process is repeated for each cell of interest $(i,j)$ in the model.

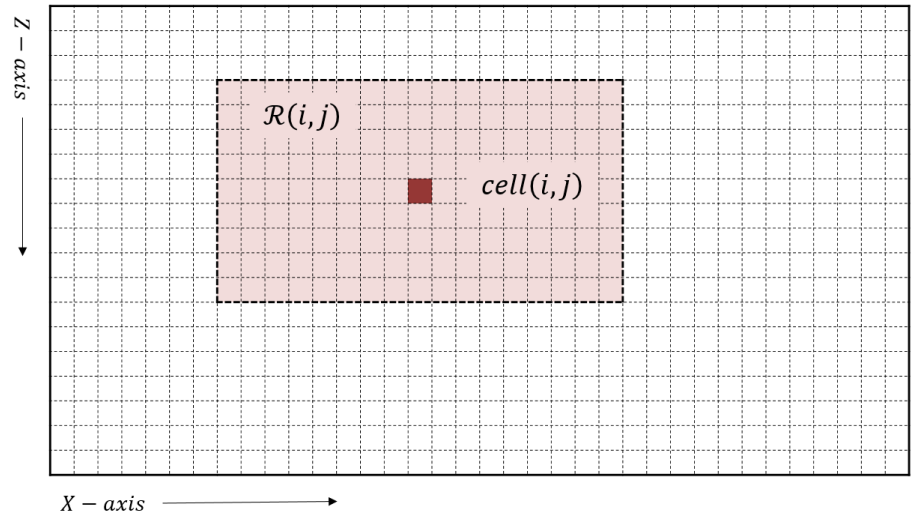



*Figure 4.4: Illustration of the complete model, region of influence $\mathcal{R}(i,j)$ shown as red shaded cells defined over a sub-set of the model around cell $(i,j)$, and the cell $(i,j)$ shown in maroon colour. The marginal posterior distribution of cell $(i,j)$ is computed with the assumption that the facies at cell $(i,j)$ depends on the facies in the neighbouring cells which in turn depend on their neighbours and so on. Thus the facies at cell $(i,j)$ show a spatial correlation with facies across the complete model. The assumption of $\mathcal{R}(i,j)$ in the algorithm, however, removes the conditional dependence of the facies at $(i,j)$ on data observed at locations outside of this region.*

The number of summations in equations 4.12 to 4.16 increases exponentially with the model size. This means that a naïve recursive computation of forward and backward probabilities becomes intractable for models of practical size. In order to limit the computational time and memory, an approximate marginal posterior distribution can be obtained by limiting the grid size considered around each cell. This introduces a practical and fundamental assumption that the facies at a given cell $(i,j)$ is conditionally independent of data observed at locations outside of a certain region of influence $\mathcal{R}(i,j)$ around the cell $(i,j)$, that is

$$\mathcal{P}\left(\kappa_{ij}|\boldsymbol{D}\right) = \mathcal{P}\left(\kappa_{ij}|\boldsymbol{D}_{\mathcal{R}(i,j)}\right) \qquad 4.17$$

where $\boldsymbol{D}_{\mathcal{R}(i,j)}$ represents the set of data within $\mathcal{R}(i,j)$. Figure 4.4 shows an illustration of the concept of a region of influence. Also, the choice of size of a partition allows us to further limit the number of summations required in equations 4.12 to 4.16 by summing over only the plausible geological facies, for example, by summing over only those facies configurations which are found in the training image. This was achieved by directly scanning the training image for the facies configurations in a manner similar to that used in so-called direct sampling (Mariethoz et al. 2010). Tjelmeland & Austad (2012) used a different approach to approximate recursive calculations in a binary MRF by approximating the interaction parameters between neighbouring nodes to zero when they are very small.

## 4.6 Synthetic Test

In order to test the algorithm and to benchmark it against pre-existing algorithms it is applied to the same synthetic inverse problem as was used by Walker & Curtis (2014a). The synthetic example is based on two 2D geological cross-sections extracted from a 3D geological process model of channels with filled and overbank sand deposits emplaced in background shale. Most of the channels are filled with brine. Gas is introduced in some of the channels while obeying gravitational ordering of the two fluids. The sample space of the facies in each cell is therefore given by

$$\mathcal{G} = \{ S\text{hale}, \text{Brine-sand}, \text{Gas-sand} \} \qquad 4.18$$

One of the geological cross-sections (with dimensions of 200 x 200 model cells) defines the training image (figure 4.5a), while the other was used as a target cross-section (with dimensions of 100 x 100 model cells) representing the true Earth (figure 4.5b). The training image was used to define the prior spatial conditional distributions of facies. The size of the region of influence $\mathcal{R}(\mathrm{i},\mathrm{j})$ was arbitrarily taken to be 7 and 9 model cells in each dimension and the partition $G_P$ was defined as a column of 7 cells. The size of the partition was chosen arbitrarily whereas its shape was chosen with computational convenience in mind.

The prior information is extracted from the training image in the form of prior probabilities $\mathcal{P}\left(\kappa_{ij}|\boldsymbol{K}_{\mathcal{N}(i,j)}\right)$ and $\mathcal{P}(\boldsymbol{K}_P|\boldsymbol{K}_{\bar{P}})$. The expression $\mathcal{P}\left(\kappa_{ij}|\boldsymbol{K}_{\mathcal{N}(i,j)}\right)$ represents the probability of existence of a facies $\kappa_{ij}$ in a cell $(i,j) \in \boldsymbol{K}$ given facies configuration $\boldsymbol{K}_{\mathcal{N}(i,j)}$ in the neighbourhood of cell $(i,j)$, and $\mathcal{P}(\boldsymbol{K}_P|\boldsymbol{K}_{\bar{P}})$ represents the spatial correlation of facies

configurations in consecutive partitions $\boldsymbol{K}_{\bar{P}}$ and $\boldsymbol{K}_{P}$. It is assumed that the prior information extracted from the training image is stationary over the model space and the probabilities computed therefrom encapsulate the expected spatial correlations of facies. In order to confirm that, realizations from prior probabilities are generated (see figure 4.7). Given that the prior realizations were generated using a partition of size 7 cells along a column, it is expected that the prior information (and hence these realizations) to preserve small-scale geometrical features and fluid orderings but not the large scale shapes of the channels. In figure 4.7 this is observed to be the case. Where gas exists it is never beneath oil, flat tops of channels are preserved, but the overall semi-circular valley-style channel shape is not. This means that by incorporating prior information we are only constraining the spatial correlations of various facies, and not the shapes of the channels – which ideally should come from the data likelihoods. If so, the prior probabilities combined with the data-derived likelihoods might produce subsurface structures with geologically plausible spatial correlations of facies.

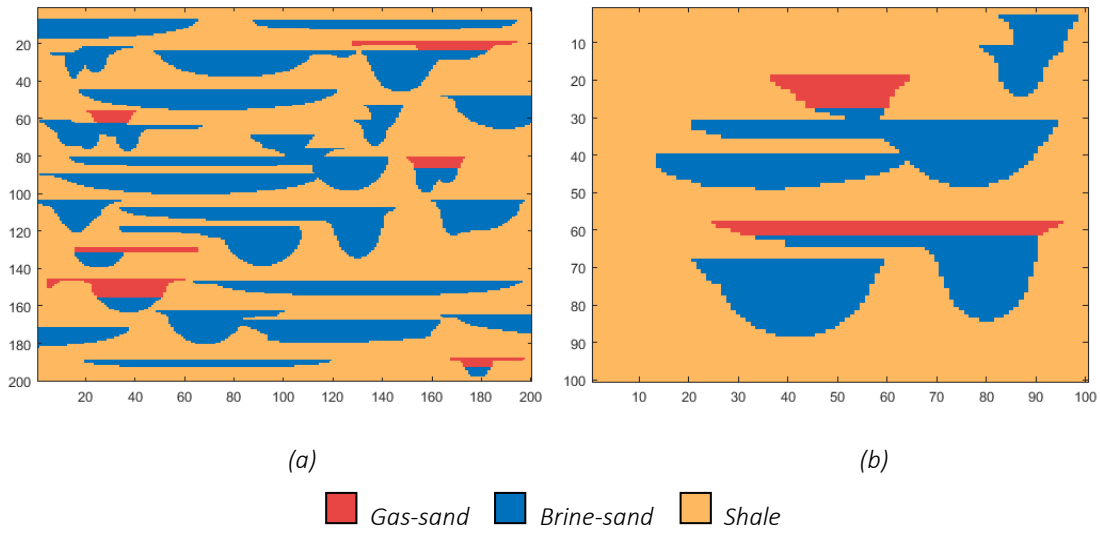
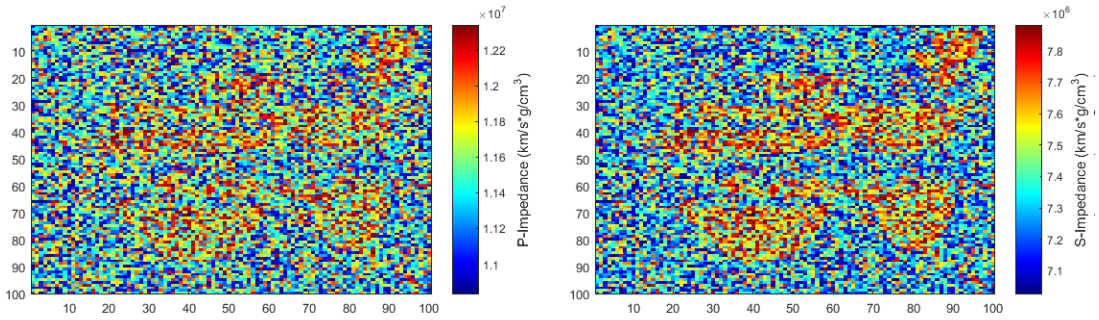



*Figure 4.5: (a) The training image (TI) and (b) the target image extracted as 2D cross-sections from a 3-D geological process model containing channels with filled and overbank sand deposits and shale in the background. The sand is filled with brine or gas, which obey gravitational ordering of the two fluids. The training image in (a) represents a conceptual depiction of typical forms of expected geological structures and spatial distributions of facies. It encodes prior information in the form of spatial conditional distributions of facies. The target image in (b) represents the true geological model which is the target for spatial facies inversion. It is expected to contain statistically similar spatial patterns and conditional distributions of facies as the training image.*




*(a)* *(b)*

*Figure 4.6: (a) P-wave and (b) S-wave impedance attributes generated independently in each cell in the target cross-section using a probabilistic forward model based on the Yin-Marin shaly-sand rock physics model (Marion 1990; Yin et al. 1993; Avseth et al. 2005) with added Gaussian noise.*

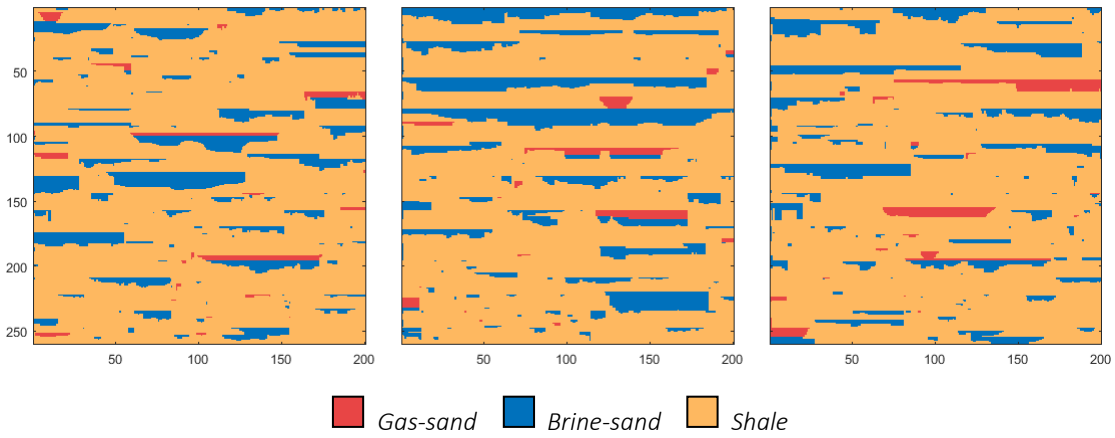



*Figure 4.7: Three realizations generated from prior probabilities computed from the training image shown in figure 4.5a using partition defined as a column of 7 cells.*

The target cross-section (figure 4.5b) was extracted from the same 3D geological process model as the training image, and it therefore contains similar spatial distributions of facies as the training image. The target cross-section was used as a model to generate synthetic seismic attributes which were used to represent real data-derived attributes in our example. These were then inverted for facies using our algorithm with the aim to reproduce the original target cross-section.

Collocated synthetic seismic attributes, P-wave and S-wave impedances $\boldsymbol{d}_{ij}$, were generated independently in each cell $(i,j)$ in the target cross-section using the localized likelihood assumption and a probabilistic forward model $\mathcal{P}(\boldsymbol{d}_{ij}|\kappa_{ij})$. The Yin-Marion shaly-

sand model (Marion 1990; Yin et al. 1993; Avseth et al. 2005) was used to predict P-wave and S-wave impedances from the given geological facies $\kappa_i \in \mathcal{G}$.

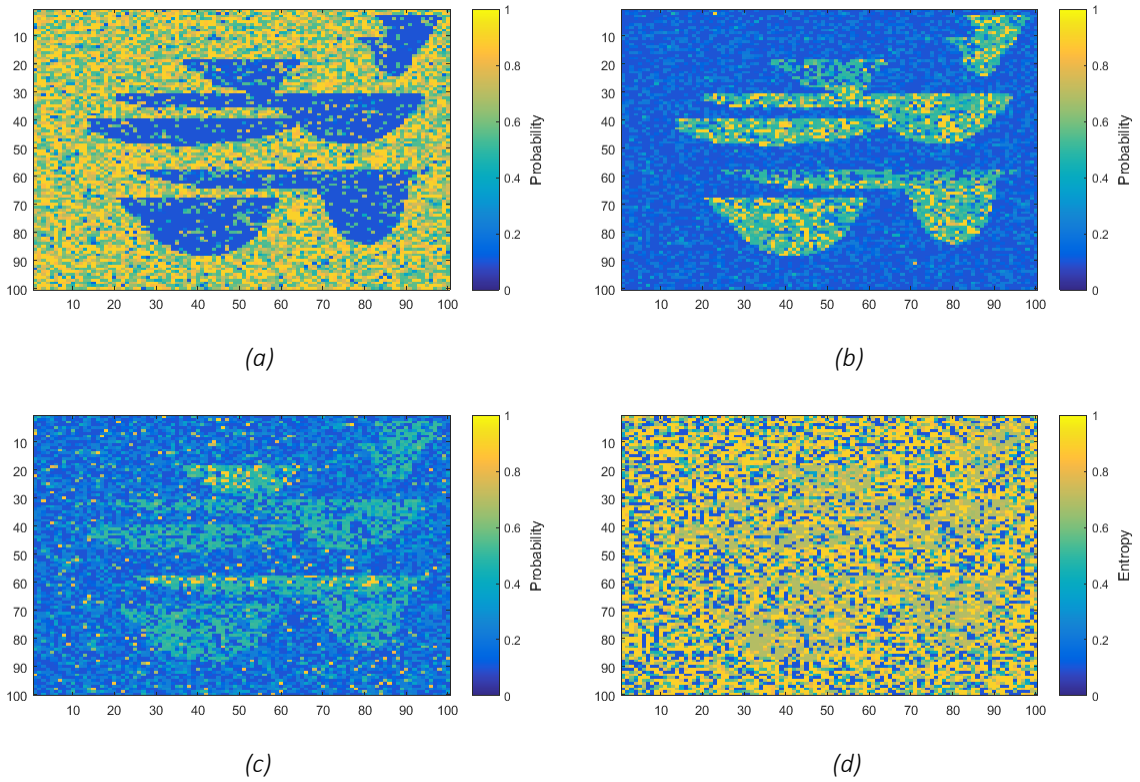
*Table 4.1: Lower and Upper bounds used to define Uniform distributions $\mathcal{P}(\boldsymbol{m}_k|\kappa_{ij})$ over petrophysical parameters $\boldsymbol{m}_k = [V_{clay}, \varphi_{sand}, S_w]_k$.*

| Lithology-Fluid Class | Clay Content by Volume ($V_{clay}$) | Sandstone Matrix Porosity ($\varphi_{sand}$) | Water Saturation ($S_w$) |
|---|---|---|---|
| **Shale** | [ 0.50, 0.90 ] | [ 0.10, 0.40 ] | [ 1.00, 1.00 ] |
| **Brine-sand** | [ 0.00, 0.20 ] | [ 0.20, 0.40 ] | [ 0.40, 1.00 ] |
| **Gas-sand** | [ 0.10, 0.40 ] | [ 0.20, 0.40 ] | [ 0.00, 0.30 ] |

*Table 4.2: Covariance matrices of seismic attributes (P-wave and S-wave impedances) for the three facies considered. The diagonal entries in the above matrices are variances of P-wave and S-wave impedances, whereas the cross-diagonal entries are the covariances of P-wave and S-wave impedances.*

| Lithology-Fluid Class | Covariance matrix for seismic attributes: P-wave and S-wave impedances |
|---|---|
| **Shale** | $\begin{bmatrix} 1.0 & 0.3 \\ 0.3 & 0.5 \end{bmatrix}$ |
| **Brine-sand** | $\begin{bmatrix} 0.8 & 0.3 \\ 0.3 & 0.6 \end{bmatrix}$ |
| **Gas-sand** | $\begin{bmatrix} 0.7 & 0.3 \\ 0.3 & 0.5 \end{bmatrix}$ |

The Yin-Marion model is defined by rock-physics parameters $\boldsymbol{m}_k = [\ V_{clay},\ \phi_{sand},\ S_w\ ]_k$ where $V_{clay}$ is the volume of clay, $\phi_{sand}$ is the matrix porosity of sand, $S_w$ is the water saturation (with gas saturation given by $S_g = 1 - S_w$), and the subscript $k$ refers to each facies. Gaussian random noise (as described below) was then added to the predicted model in order to formulate the model probabilistically as $\mathcal{P}(\boldsymbol{d}_{ij}|\boldsymbol{m}_k)$. The likelihood $\mathcal{P}(\boldsymbol{d}_{ij}|\kappa_{ij})$ is then given in terms of rock-physics parameters $\boldsymbol{m}_k$ by

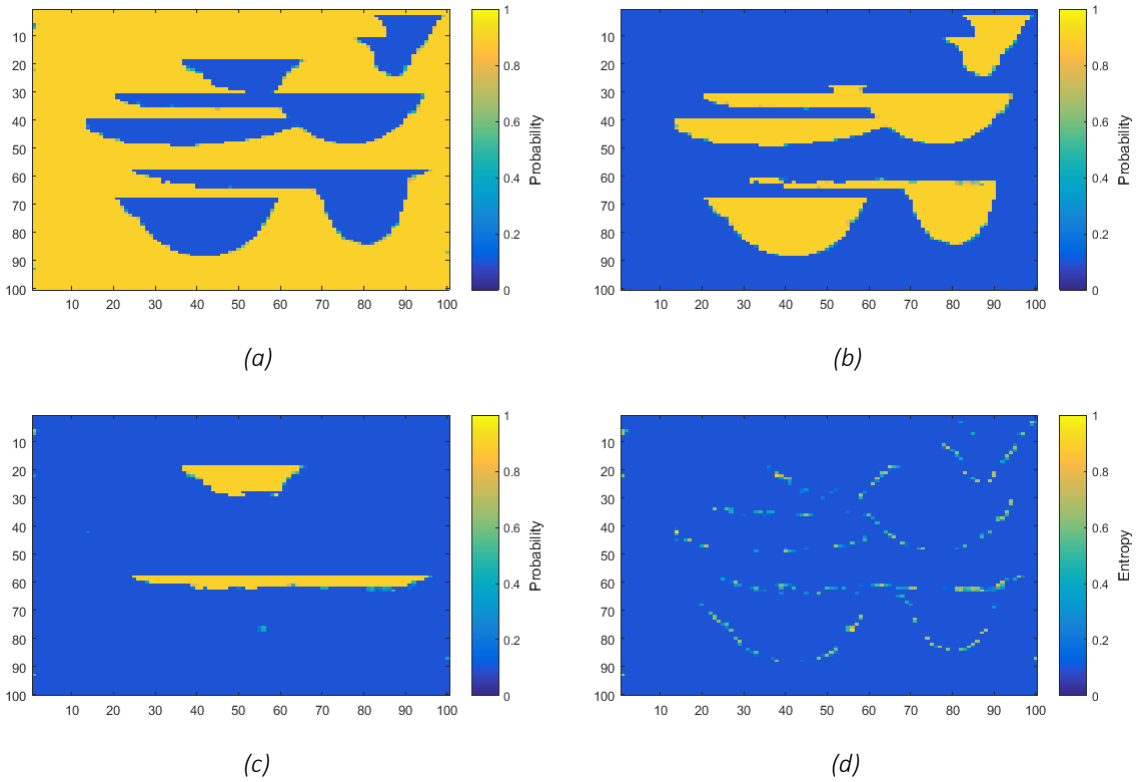



*(a)* *(b)* *(c)* *(d)*

*Figure 4.8: Likelihood functions $P(\boldsymbol{d}_{ij}|\kappa_{ij})$ for each of the geological facies, (a) shale, (b) brine-sand and (c) gas-sand, and (d) entropy (a measure of uncertainty of classification), given the seismic attributes. Results are computed from a Gaussian mixture model using neural networks (Meier et al. 2007a & b; Shahraeeni & Curtis 2011; Shahraeeni et al. 2012). In each plot, bright yellow colour represents high probability (close to 1) and dark blue colour represents low probability (close to 0). The likelihoods are normalized so that the sum of likelihoods for each of the facies in any cell equals 1.*

$$\mathcal{P}(\boldsymbol{d}_{ij}|\kappa_{ij}) = \iiint_{\boldsymbol{L}}^{\boldsymbol{B}} \mathcal{P}(\boldsymbol{d}_{ij}|\boldsymbol{m}_k)\, \mathcal{P}(\boldsymbol{m}_k|\kappa_{ij})\, d\boldsymbol{m}_k \qquad 4.19$$

where **L** and **B** (bold-face letters to represent vector bounds) respectively represent the lower and upper bounds on each parameter in $\boldsymbol{m}_k$. The conditional distribution $\mathcal{P}(\boldsymbol{m}_k|\kappa_{ij})$ describing the probabilistic relationship between rock-physical parameters $\boldsymbol{m}_k$ and the geological facies $\kappa_{ij}$ in each cell of the target cross-section, was set to Uniform within predefined lower and upper bounds [**L**, **B**] on each parameter in $\boldsymbol{m}_k$ given in Table 4.1. The distribution $\mathcal{P}(\boldsymbol{d}_{ij}|\boldsymbol{m}_k)$ is given by the deterministic Yin-Marion shaly-sand model $g(\boldsymbol{m}_k)$ and a stochastic component in the form of Gaussian random noise $\boldsymbol{\epsilon}$ added to the predicted model in order to formulate the model probabilistically.




*(a)* *(b)* *(c)* *(d)*

*Figure 4.9: Cell-wise marginal posterior distributions computed using 2D-HMM model for each of the geological facies, (a) shale, (b) brine-sand (c) gas-sand, and (d) entropy (a measure of uncertainty of classification).*

The data $\boldsymbol{d}_{ij}$ in each model cell $(i,j)$ may then be written as

$$\boldsymbol{d}_{ij} = g(\boldsymbol{m}_k) + \boldsymbol{\epsilon} \qquad 4.20$$

where $\boldsymbol{\epsilon} \sim N(\boldsymbol{e}; 0, \boldsymbol{\Sigma}_e)$, $N$ represents the Gaussian function, and $\boldsymbol{\Sigma}_e$ is the noise covariance matrix given in Table 4.2 for each of the facies. The likelihood $\mathcal{P}(\boldsymbol{d}_{ij}|\kappa_{ij})$ was computed for each of the geological facies (figure 4.8) from a Gaussian mixture model (GMM) using neural networks (Meier et al. 2007a & b; Shahraeeni & Curtis 2011; Shahraeeni et al. 2012). Since the likelihood only uses attributes to discriminate between the facies it allows reasonable discrimination between sand and shale (figure 4.8a), but could hardly discriminate between brine-sand and gas-sand (figure 4.8b & c). Also, the likelihood functions are noisy and do not adhere to the statistical spatial distribution of facies as depicted in the training image. This corroborates the need to introduce prior geological knowledge incorporating the spatial correlation of facies.

The distribution $\mathcal{P}(\boldsymbol{d}_{ij}|\kappa_{ij})$ in equation 4.19 was sampled sequentially: first for $\boldsymbol{m}_k$ from $\mathcal{P}(\boldsymbol{m}_k|\kappa_{ij})$ and then for $\boldsymbol{d}_{ij}$ from $\mathcal{P}(\boldsymbol{d}_{ij}|\boldsymbol{m}_k)$, to obtain synthetic data $\boldsymbol{d}_{ij}$ given the facies $\kappa_{ij}$ in each cell in the model (target cross-section). The data thus obtained are shown in figure *4.6*. The marginal posterior distributions for each of the facies in each cell in the model were computed (figure 4.9) incorporating both the prior geological knowledge elicited from the training image (figure 4.5) and the likelihood functions (figure 4.8).

## 4.7 Computational Complexity

The computational complexity of this algorithm may be expressed mathematically as the maximum number of floating point operations required to compute the posterior marginal distributions in each cell in the model:

$$4 \times r^2 \times (r-1) \times c \times |\mathcal{G}|^r \qquad 4.21$$

where $r$ is the number of rows and $c$ is the number of columns in the region of influence $\mathcal{R}(i,j)$ around the cell $(i,j)$ under consideration, and $|\mathcal{G}|$ represents the size of the sample space of geological facies (i.e., the number of geological facies considered). It is assumed in deriving the above expression that the partition $\boldsymbol{K}_P$ is defined as a column of $r$ cells. The variable $r$ has the maximum order 3 in equation 4.21 and it also appears in the exponent of $|\mathcal{G}|$. This means that it is desirable to define the partition $\boldsymbol{K}_P$ along the shorter dimension of $\mathcal{R}(i,j)$.

The size of the space of geological facies (i.e., the number of discrete facies classes) $|\mathcal{G}|$ is an important factor in the above expression. Because of its exponentiation it must be chosen to be as small as possible.As the size of the partition increases, a naïve approach to storing the joint distribution of facies would require an exponential amount of computer memory. However, since a training image may only depict a finite number of facies configurations, it limits the shapes and scales of the facies configurations that are considered geologically plausible. As a consequence, we only need to compute probabilities and perform sampling for a limited number of configurations. The partition size can, therefore, be taken as large as any one of the dimensions of the training image, thus allowing this method to be easily extensible to 3D without becoming computationally intractable.

## 4.8 Discussion

The computation of a full joint distribution $\mathcal{P}(\boldsymbol{K}|\boldsymbol{D})$ of geological facies conditioned to seismic and well data is computationally intractable even for small synthetic models. Previous research in probabilistic seismic inversion by Walker & Curtis (2014a) relied on the computation of approximate posterior conditional distributions of facies. A different approach is used here: marginal posterior distributions for each of the facies in each cell in the model are computed, conditioned to the data in that cell and to prior facies distributions. Computation of marginal posteriors is orders of magnitude faster than the previous approach, and requires far lower memory. In terms of quality of prior information incorporated into the inversion process, our method clearly outperforms the method of Walker & Curtis (2014a): the realizations from prior distributions (figure 4.7) and the computed marginal posterior distributions (figure 4.9) show flat tops of channels in our example model, whereas the previous method could not produce flat tops in samples from the same example. The main reason for this difference is that this method computes prior probabilities of spatial distribution of facies over partitions (7 cells in a column in our example) as compared to the neighbourhood structure (3x3 cells as was used in previous work). The size of partition is typically larger than the size of neighbourhood structure in the same dimension (7 cells versus 3 cells).

Other previously existing methods that invert seismic data for geological facies using hidden Markov or similar models (e.g., Larsen et al. 2006; Ulvmoen & Omre 2010; Ulvmoen et al. 2010; Hammer & Tjelmeland 2011; Rimstad & Omre 2013; Lindberg & Omre 2014 & 2015) rely on sampling from full posterior distributions using McMC methods. As described earlier in section 4.2, McMC based methods are slow to converge for high dimensional problems and they may suffer from convergence related bias. For this reason, a comparison of the presented method with McMC based methods in terms of computational efficiency would be essentially meaningless. Nonetheless, a comparison can be made in terms of the amount and quality of prior information incorporated in the inversion process. Since this method is based on full 2-dimensional relationships among cells in neighbouring partitions, it incorporates more prior information as compare to the 1D Markov-chain based methods (e.g., Larsen et al. 2006). However, the amount and quality of prior information incorporated is comparable between 2D-HMM based priors used in this research and the profile Markov random field based priors

of Ulvmoen & Omre (2010) and Ulvmoen et al. (2010). The advantage of the presented method over the latter methods remains that it performs inference directly for posterior marginal distributions (in contrast to the full joint distribution) while avoiding the use of sampling.

The localized likelihood assumption is a fundamental assumption in this algorithm as the observations (here seismic attributes) must be conditionally independent given the hidden states (geological facies) in a 2D-HMM. The localized likelihood assumption also allowed us to factorize the likelihood probability $\mathcal{P}(\boldsymbol{D}_P|\boldsymbol{K}_P) = \prod_{i=1}^{n} \mathcal{P}(\boldsymbol{d}_{P,i}|\kappa_{P,i})$ as a product of factors each involving the local likelihood in each cell in the model, as used in equations 4.14 and 4.16. This means that the data are assumed to possess spatial correlations that are only due to the spatial correlations present in the geology (facies and rock properties). However, spatial correlations are also induced in the data by the non-localized nature and limited resolution of seismic data, and by correlated noise that was not accounted for during the process that estimated the attributes. This means that seismic data must be corrected for non-localizing effects of seismic wave propagation such as attenuation, Fresnel zone smearing, etc. as much as possible. This in turn requires that the input data are supplied after proper de-noising (and migration in case of seismic data in which all wave propagation effects have been accounted for).

Spatial correlations in attributes due to the correlations in geology, on the other hand, are exploited in the inference to improve the spatial correlations in the inverted facies. Therefore, although the assumption of localized likelihoods allows us to compute approximate marginal posterior distributions in a closed form solution, it effectively limits the amount of information present in the data that could otherwise be useful in the reconstruction of spatial correlations of facies (specifically, our method ignores correlations in the attributes between cells). As a consequence, this method relies significantly on the prior information, rather than the data, to reconstruct the spatial correlations expected in the geology. The data, therefore, provide the location specific information and the prior knowledge provides information on the spatial correlations to be recovered in the inversion. An advantage of using prior information in this way is that it reduces sensitivity to random noise in the data. However, a more sophisticated approach would exploit the fact that the data at neighbouring locations are spatially correlated, depending on the temporal and spatial resolution of the seismic data. Such an approach is discussed in chapter 5.

The incorporation of prior information from a training image is dependent on the configuration of pixels that are scanned to compute spatial conditional probabilities. Since we scan the training image with a stencil that is the same as the partition $\boldsymbol{K}_P$, it is important to define the shape and size of the partition such that the information gathered can reproduce structures present in the training image up to any desired accuracy. In our synthetic case we defined a partition as a column of 7 cells as we found that the prior information thus gathered from the training image was sufficient to reconstruct the actual marginal distributions with reasonable accuracy. This was because the vertical variations in facies in our training image are correlated over smaller length scales than the lateral variations, and hence 7 cells were sufficient. Because we compute the conditional probabilities of facies patterns by sampling from the training image using a direct sampling approach (Mariethoz et al. 2010), we can easily extend the size of the neighbourhood structure within the memory limits of modern computers. A reasonably large neighbourhood structure increases the computational time but still remains tractable. The shape of the partition can also be chosen with arbitrary complexity to model complex spatial distributions of facies provided the ordering of partitions can still be defined as required by the algorithm. Since the size of partition defines the size of the region of influence along any one of the dimensions, the partition size should be chosen large enough that the region of influence may contain any large scale recoverable features in the training image.

The assumption of the region of influence $\mathcal{R}(i,j)$ around each cell $(i,j)$ in the model is based on the observation that the facies at any location in the subsurface have probabilistic dependence only on the data observed in a certain region around it. This region can be taken reasonably large but finite. In fact it could be as large as the size of the training image. If the region-of-influence is large enough to capture the large scale facies patterns depicted in the training image, this assumption only limits the data correlations outside this region and not the correlation of geological facies. As a consequence, the concept of region-of-influence not only makes the algorithm tractable without limiting the size of the overall model, it also offers a reliable estimation of posterior marginal distributions of facies at the point of interest. Since this assumption is no stricter than the assumption of localized likelihoods, it is therefore valid for all models that are built with the assumption of localized likelihoods. This also applies to models in various other fields of research, such as image and video processing. Other researchers who used two-dimensional extensions of HMM either limited the spatial

interactions of neighbouring cells in the model, or they assumed a 1D underlying graphical model (pseudo 2D HMM). Both of these approaches prohibited incorporation of full two-dimensional interactions of cells (facies correlations in our synthetic example). The assumption of the region of influence allowed us to derive the equations to compute marginal posterior distributions with full two-dimensional spatial interactions among neighbouring cells in the model.

Although the use of a 2D-HMM for spatial inversion of geological facies from seismic data is demonstrated, extension of the method to 3D or higher dimensions is straightforward. Since marginal posterior distributions can be computed in each cell independently, this approach can be parallelized on heterogeneous computer architectures to exploit the maximum efficiency deliverable from the modern day computational and graphical processors.

## 4.9 Conclusions

A new 2D hidden Markov model is introduced to compute marginal posterior probabilities of geological facies from geophysical data. The prior knowledge is incorporated in terms of spatial statistics of facies distributions in space that can be represented in the form of a training image or otherwise. The prior probabilities are independent of data, and only contribute location independent contextual information. Since the data are observed, they are fixed. The observed data represents any type of data (e.g., P-wave and S-wave impedances) that can discriminate between geological facies present at any point in the model to some degree of confidence. The likelihood is assumed to be localized, which implies that given the geological facies at a location, the data observed at that location are assumed to be conditionally independent of the geological facies and data at any other location in the model. The implication of the localized likelihoods assumption is that the seismic data is assumed to be processed and corrected for any non-localized effects of seismic wave propagation.

Previous researchers who used 2D hidden Markov models made assumptions that limit the interaction between neighbouring cells. The presented method makes no such assumptions and models the full 2D interactions between neighbouring cells in the model. However, this method does assume that there lies a region of influence around each cell in the model such that any observations (data) outside of this region have no correlation with the observation in the cell under consideration. Such an assumption does not limit the spatial

correlations among the hidden states, and is therefore valid for any model that is based on the localized likelihoods assumption. This method has been tested on synthetic data and is found to be reliable and many orders of magnitude faster than previous research on the same problem under the same set of assumptions.

# Chapter 5 Variational Bayesian Inversion

## 5.1 Summary

A new Bayesian inversion method is introduced in this chapter that estimates the spatial distribution of geological facies from attributes of seismic data, by showing how the usual probabilistic inverse problem can be solved using an optimization framework still providing full probabilistic results. The method infers the posterior probability of the facies plus some other unknown model parameters, from geophysical data (e.g. seismic attributes) and geological prior information presented as a Markov random field (MRF) (see section 3.5.1). The localized likelihoods (LL, see section 1.1.4) assumption is relaxed in this chapter: probabilistic dependence is allowed between data observed at a location and the geological properties (facies in particular: well-defined and distinct rock and fluid types) in any neighbourhood of that location through a spatial filter. Such likelihoods are henceforth referred to as *quasi-localized*.

The variational Bayes method introduced in section 2.4 is used as a more efficient sampling-free alternative to stochastic inference that offers reliable detection of convergence of the desired posterior distribution. The presented method thus obviates the need for sampling, while still providing probabilistic results. It is shown in a noisy synthetic example that this method recovered the coefficients of the spatial filter with reasonable accuracy, and recovered the correct facies distribution. This method is also shown to be robust against weak prior information and quasi-localized likelihoods (QLL), and that it outperforms previous methods which rely on the LL assumption. This method is computationally efficient, and is expected to be applicable to 3D models of realistic size on modern computers without incurring any significant computational limitations.

## 5.2 Introduction

Geophysical data can be strongly correlated spatially due to mixing of information across different spatial locations (sometimes referred to as blurring or smearing). For example, this occurs in seismic imaging due to errors in the velocity model which cause mislocation of seismic attributes, Fresnel zone smearing, migration errors due to the limited apertures of

seismic arrays, and a number of other factors. Facies inversion methods often ignore such spatial blurring of geophysical data, and rely on the LL assumption for computational and analytical convenience. This assumption was implicit or explicit in most of the previous research (e.g., Larsen *et al.* 2006; Ulvmoen & Omre, 2010; Ulvmoen *et al.* 2010; Shahraeeni & Curtis, 2011; Shahraeeni *et al.* 2012; Grana et al. 2013; Walker & Curtis, 2014a; Nawaz & Curtis, 2017; and Grana, 2018). A more robust inversion method is required that acknowledges the non-localized nature of geophysical data and incorporates spatial correlations present in the data, which is addressed in this chapter.

Bayesian inversion for geological facies typically involves cluster analysis within data such as seismic attributes and/or any other continuous rock properties. Each cluster is considered to represent a particular facies for which we desire to estimate the data *likelihood* – the probability that each data point belongs to a specific cluster or facies. Likelihoods obtained from cluster analysis are often assumed to be spatially localized in the sense that given the facies in any spatial model cell, the data in that cell are assumed to be conditionally independent of the facies and attributes in the rest of the model. Such an assumption is commonly used in previous research (e.g., Larsen *et al.* 2006; Ulvmoen & Omre, 2010; Ulvmoen *et al.* 2010; Walker & Curtis, 2014a; and Nawaz & Curtis, 2017) and is referred to as the *localized likelihoods* assumption. Unfortunately, geophysical data generally contain strong spatial correlations due to inaccurate processing and limited resolution of seismic imaging that results in spatial blurring or smearing which contravenes the localized likelihoods assumption. Another common assumption for the sake of computational efficiency and analytical convenience is that geological facies are spatially independent, i.e. facies in any model cell are assumed to be independent of those in the rest of the model. Such an approach has also been implicitly or explicitly used in the literature (e.g. Shahraeeni & Curtis, 2011; Shahraeeni *et al.* 2012; Grana, 2018) with the hope that the spatial continuity of facies may be recovered from the spatial continuity of seismic data. A typical problem with these methods is that they are more susceptible to noise present in the seismic data, and provide probability estimates with high entropy (uncertainty) for those data points that fall equidistant from cluster centres. Spatial coupling (probabilistic dependence between neighbouring locations) based on prior information may be introduced in the model parameters to reconstruct desired spatial correlations in their posterior distributions.

This prior information is injected using a Markov random field (MRF, see section 3.5.1) to allow for spatial coupling of model parameters (facies in this case). A number of other methods for probabilistic inversion use the Markovian assumption, e.g. Larsen *et al.* (2006), Ulvmoen & Omre (2010), Ulvmoen *et al.* (2010) and Rimstad *et al.* (2012). Since exact Bayesian inference is intractable in real-scale models with spatial coupling between the parameters, approximate inference becomes inevitable. A stochastic approach based on *Markov-chain Monte Carlo* (McMC) simulations is commonly used for this purpose (Doyen *et al.* 1989; Mukerji*et al.* 2001; Grana *et al.* 2012; Wang *et al.* 2016). The variational Bayes method (see section 2.4), which is computationally efficient and is suitable for a MRF model is used in this chapter.

In this research, the likelihood of observing (or estimating) geophysical data at a location given the geological facies in the neighbouring locations is modelled by a new form of a GMM introduced in this chapter – *spatial Gaussian mixture model* (SGMM), which represents a spatial form of the GM distribution. Such likelihoods are referred to as *quasi-localized likelihoods* (QLL). Examples of previous research on 1D Bayesian inversion methods in which likelihoods are not (fully) localized include Lindberg & Omre (2014 & 2015), Grana et al. (2017) and Lindberg *et al.* (2015). The new method presented in this chapter is multi-dimensional, and it allows for joint estimation of SGMM (spatial GM distribution) parameters and the spatial distribution of geological facies. The parameters of the SGMM are spatially constrained through both the prior distribution of facies, and their QLL. As a result, the GM distribution parameters are chosen such that they provide best estimates of the spatial distribution of geological facies that are consistent with the prior information, and which are also constrained by the geophysical data.

In the following, the coordinate system that presents the data with respect to their geographical locations and characterizes the spatial distribution of facies given by the prior information is referred to as the *model space*, and the coordinate system that is used to cross-plot data (e.g. multiple seismic attributes such as P-wave and S-wave impedances) and allows analysis of their mutual correlation irrespective of their spatial locations is referred to as the *attribute space*. The presented method uses a variational form of *expectation-maximization* (EM) (Dempster *et al.,* 1977; Beal, 2003) algorithm which iteratively estimates the posterior marginal distributions of facies in the model space during the E-step, and updates the parameters of the GMM in the attribute space during the M-step.

## 5.3 Model

The probabilistic inverse problem that we solve is to infer the unknown geological facies $\boldsymbol{\kappa}$ from the observed geophysical data or its attributes $\boldsymbol{d}$. The Bayesian solution of the inverse problem is given by the posterior distribution $\mathcal{P}(\boldsymbol{\kappa}|\boldsymbol{d})$ of $\boldsymbol{\kappa}$ given $\boldsymbol{d}$ that can be expressed by the *Bayes' theorem* (equation 2.1) as

$$\mathcal{P}(\boldsymbol{\kappa}|\boldsymbol{d}) = \frac{\mathcal{P}(\boldsymbol{\kappa},\boldsymbol{d})}{\mathcal{P}(\boldsymbol{d})} = \frac{\mathcal{P}(\boldsymbol{d}|\boldsymbol{\kappa})\mathcal{P}(\boldsymbol{\kappa})}{\mathcal{P}(\boldsymbol{d})} \tag{5.1}$$

The denominator $\mathcal{P}(\boldsymbol{d})$ represents the *marginal likelihood* of the observed data $\boldsymbol{d}$ (also called *evidence*). It acts as normalization constant, and is given by

$$\mathcal{P}(\boldsymbol{d}) = \sum_{\boldsymbol{\kappa}} \mathcal{P}(\boldsymbol{d},\boldsymbol{\kappa}) = \sum_{\boldsymbol{\kappa}} \mathcal{P}(\boldsymbol{d}|\boldsymbol{\kappa})\mathcal{P}(\boldsymbol{\kappa}) \tag{5.2}$$

Below, we first describe a model for the prior distribution $\mathcal{P}(\boldsymbol{\kappa})$ of facies in subsection 5.3.1, then we describe a model for the likelihood $\mathcal{P}(\boldsymbol{d}|\boldsymbol{\kappa})$ of data $\boldsymbol{d}$ given facies $\boldsymbol{\kappa}$ in subsection 5.3.2, and then in section 5.3.3 the prior and the likelihood are combined using equation 5.1 to obtain the posterior distribution.

### *5.3.1 Prior Model*

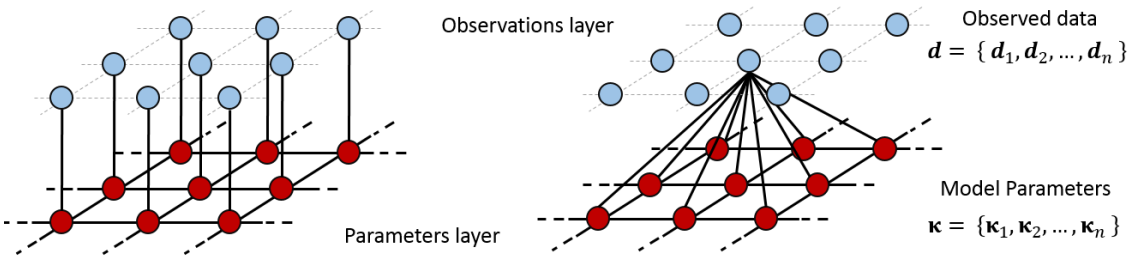
A variant of a typical HMRF model (see figure 3.9) is used in this chapter where observed variables $\boldsymbol{d}_i$ at a location $i$ depend directly on the hidden variables $\boldsymbol{\kappa}_{\mathcal{N}_i}$ within the neighbourhood $\mathcal{N}_i$ of $i$ (with $i$ inclusive). This means that given the facies $\kappa_i$ at $i$, the corresponding data $\boldsymbol{d}_i$ are not assumed to be conditionally independent of rest of the facies $\boldsymbol{\kappa}_{\setminus i}$ in the model. This relaxes the LL assumption (figure 5.1b). This concept is further developed in section 5.3.2. Note that the Markovian property still requires that the conditional independence (CI) assumption is maintained on data, i.e. the observed variables are assumed to be mutually conditionally independent given the hidden variables (entire facies model).

Geological prior information about the spatial distributions of facies is assumed to be available as a joint distribution $\mathcal{P}(\boldsymbol{\kappa})$ of facies in the form of a pairwise MRF (see equation 3.11) which is given by

$$\mathcal{P}(\boldsymbol{\kappa}) = \frac{1}{Z} \prod_{(i,j)\in\mathcal{E}} \psi_{ij}(\kappa_i, \kappa_j) \qquad 5.3$$

The prior probability of occurrence of facies $\kappa_i$ at a location $i$ given the facies $\boldsymbol{\kappa}_{\mathcal{N}_{\setminus i}}$ in its neighbourhood $\mathcal{N}_{\setminus i}$ is therefore given by

$$\mathcal{P}\left(\kappa_i | \boldsymbol{\kappa}_{\mathcal{N}_{\setminus i}}\right) \propto \prod_{j\in\mathcal{N}_{\setminus i}} \psi_{ij}(\kappa_i, \kappa_j) \qquad 5.4$$




*(a) A typical HMRF model* *(b) HMRF model as used in this chapter*

*Figure 5.1: A graphical depiction of a hidden Markov random field (HMRF) with two layers where the upper layer consists of observed variables $\boldsymbol{d}$ represented by light-blue circles and the lower layer consists of hidden variables (model parameters, facies herein) $\boldsymbol{\kappa}$ represented by dark-red circles. The solid black lines represent the edges between connected vertices in the model whereas dotted grey lines in the upper layer are only guidelines included for clarity in order to portray the relative positions of observed vertices in the model grid. The grid is shown in 2 dimensions with a 3x3 square matrix of vertices for illustration purpose only. The actual grid may be higher dimensional and much larger in size. (a) A typical HMRF model where data $\boldsymbol{d}_i$ at a location $i$ depends directly only on the facies $\boldsymbol{\kappa}_i$ at that location. (b) A variant of the HMRF model used in this chapter where each observed variable $\boldsymbol{d}_i$ at location $i$ depends directly on all hidden variables (facies) $\boldsymbol{\kappa}_{\mathcal{N}_i}$ within an arbitrary but pre-specified neighbourhood $\mathcal{N}_i$ of $i$ (with $i$ inclusive). The edges between hidden and observed variables are shown only for one observed variable for clarity, but all observed vertices in the model are assumed to be connected to hidden variables in a similar fashion.*

## *5.3.2 Likelihood*

The likelihood of data observed at a location $i$ given the facies in the neighbourhood $\mathcal{N}_i$ of that location is estimated in order to account for the blurring effect of the band-limited seismic data. This is referred to as *quasi-localized likelihoods* (QLL) since the dependence of

data on facies in neighbouring locations may not be regarded as fully non-localized likelihoods (unless the neighbourhood spans the entire domain). It is important to note there that the QLL assumption is less stringent than the LL assumption.

All facies classification methods assume that the variation in rock properties within a facies is smaller than variations between different facies. Any ambiguity in classification due to overlap of rock properties among multiple facies might be able to be resolved to some extent by introducing the spatial context of each data point. This can be done by conditioning each data point on its spatial neighbours based on the information contained in spatial priors and/or QLL. For example, if a particular facies is more likely to be present in the neighbourhood of a given location, then the same facies is more likely to be present at that location (compared to other facies), provided that the data observed within some neighbourhood of that location also support that. In this manner, QLL reduce the entropy of (the degree of uncertainty in) classification by introducing spatial context of observations and geology, compared to the localized likelihoods which offer no spatial context for the classification task.

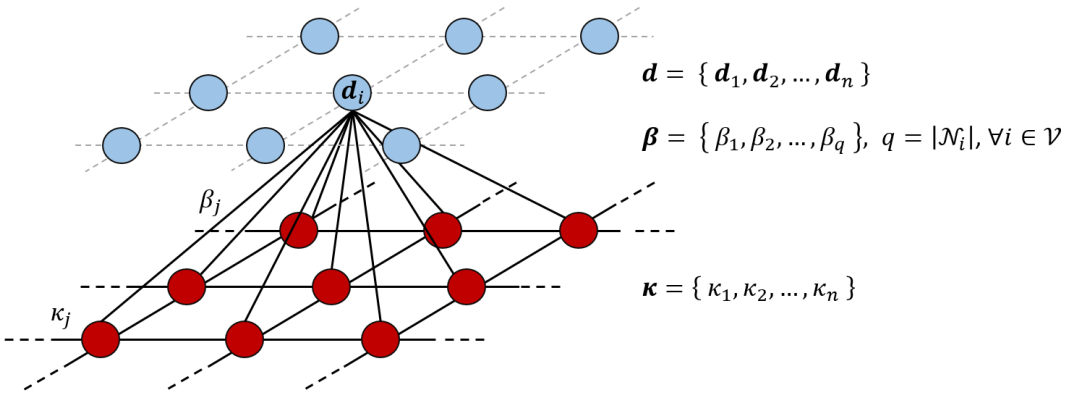



*Figure 5.2: A graphical depiction of a hidden Markov random field (HMRF) as in figure 5.1b where each edge between an observed variable $\boldsymbol{d}_i$ at a location $i$ and the hidden variables $\boldsymbol{\kappa}_j, j \in \mathcal{N}_i$ within the neighbourhood $\mathcal{N}_i$ of $i$ is associated with a weight parameter $\boldsymbol{\beta}_j$ which may be interpreted as the strength of the connection between the two variables in the definition of quasi-localized likelihoods.*

Even though the likelihoods are not assumed to be localized it is still assumed that the data at each location are conditionally independent given the facies model. This implies that any spatial correlations in the observations are assumed to be a direct consequence of spatial

distribution of facies, and not due to correlations that are independent of the geology and are introduced by the measurement process (for example, due to correlated random or systematic noise).

We consider a set $\mathcal{G} = \{1, \dots, K\}$ of discrete variables representing geological facies. Each facies $k \in \mathcal{G}$ is defined in terms of expected attributes $\boldsymbol{\mu}_k$ and the corresponding covariance matrix $\boldsymbol{\Sigma}_k$ that represents intra-facies variations. Let $\boldsymbol{R}_{\mathcal{N}_i} = \left(\boldsymbol{r}_j : j \in \mathcal{N}_i\right)$ be a $p \times q$ matrix of $p$ dimensional data at each of the $q$ locations in the neighbourhood $j \in \mathcal{N}_i$ such that $|\mathcal{N}_i| = q$ is fixed and is independent of location $i$ in the graph, and $\boldsymbol{r}_j$ represents expected local facies responses at each location given by some mapping from the discrete facies $\kappa_j$ to the domain of observed variables $\boldsymbol{d}$. To make this more concrete, define $\boldsymbol{r}_j$ as the expectation of a set of superposed Gaussian distributions $N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),\ k \in \mathcal{G}$ for each facies, weighted by some estimate of the marginal distribution $\hat{\mathcal{P}}_j\left(\kappa_j\right)$ of the facies at each location $j$:

$$\boldsymbol{r}_j = \mathbb{E}\left(\sum_{k \in \mathcal{G}} \hat{\mathcal{P}}_j\left(\kappa_j = k\right) N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right) = \sum_{k \in \mathcal{G}} \hat{\mathcal{P}}_j\left(\kappa_j = k\right)\boldsymbol{\mu}_k \tag{5.5}$$

The data $\boldsymbol{d}_i$ observed at a location $i$ are assumed to be a weighted linear combination of facies responses $\boldsymbol{R}_{\mathcal{N}_i}$ in neighbourhood $\mathcal{N}_i$ such that

$$\boldsymbol{d}_i = \sum_{j \in \mathcal{N}_i} \beta_j \boldsymbol{r}_j + \boldsymbol{\varepsilon}_i = \boldsymbol{R}_{\mathcal{N}_i}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i \tag{5.6}$$

where $\boldsymbol{d}_i$ is a $p \times 1$ vector of $p$ dimensional data, $\boldsymbol{\beta}$ is a $q \times 1$ vector of regression coefficients, and $\boldsymbol{\varepsilon}_i$ is a $p \times 1$ vector of errors which are assumed to be jointly distributed according to a Normal distribution $N(0, \boldsymbol{\Sigma}_\varepsilon)$. The data are assumed to have been pre-standardized to have unit variance, so that the definition of regressors $\boldsymbol{R}_{\mathcal{N}_i}$ allows us to interpret $\boldsymbol{\beta}$ as a weighting kernel over all of the attributes observed at multiple locations in the neighbourhood of $i$ (figure 5.2). The attributes can be de-standardized later to their original means and variances for display and interpretation purposes. Now define the set of parameters as $\Theta \equiv \{\boldsymbol{\beta}, \boldsymbol{\Sigma}_\varepsilon, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}, k \in \mathcal{G}$. So, given the expected facies responses $\boldsymbol{R}_{\mathcal{N}_i}$ in the neighbourhood of $i$, the data $\boldsymbol{d}_i$ are Normally distributed with mean $\boldsymbol{R}_{\mathcal{N}_i}\boldsymbol{\beta}$ and covariance matrix $\boldsymbol{\Sigma}_\varepsilon$. The quasi-localized likelihood of $\boldsymbol{d}_i$ computed at $i$ given the geological facies $\boldsymbol{\kappa}_{\mathcal{N}_i} \equiv \left\{\kappa_j : j \in \mathcal{N}_i\right\} \subseteq \boldsymbol{\kappa}$ in the neighbourhood $\mathcal{N}_i$ of location $i$ is therefore given by

$$\mathcal{P}\left(\boldsymbol{d}_i|\boldsymbol{\kappa}_{\mathcal{N}_i};\Theta\right)=\mathcal{P}\left(\boldsymbol{d}_i|\boldsymbol{R}_{\mathcal{N}_i}\boldsymbol{\beta};\Theta\right)=N\left(\boldsymbol{R}_{\mathcal{N}_i}\boldsymbol{\beta},\boldsymbol{\Sigma}_{\varepsilon}\right) \quad 5.7$$

We can show that the likelihood of observing data $\boldsymbol{d}_i$ at a $i$ given the geological facies $\boldsymbol{\kappa}_{\mathcal{N}_{\setminus i}}$ in the neighbourhood $\mathcal{N}_{\setminus i}$ of $i$ and the parameters $\Theta$ is given by

$$\begin{aligned}\mathcal{P}\left(\boldsymbol{d}_i|\boldsymbol{\kappa}_{\mathcal{N}_{\setminus i}};\Theta\right)&=\sum_{\kappa_i}\mathcal{P}\left(\boldsymbol{d}_i,\kappa_i|\boldsymbol{\kappa}_{\mathcal{N}_{\setminus i}};\Theta\right)\\&=\sum_{\kappa_i}\mathcal{P}\left(\boldsymbol{d}_i|\boldsymbol{\kappa}_{\mathcal{N}_i};\Theta\right)\mathcal{P}\left(\kappa_i|\boldsymbol{\kappa}_{\mathcal{N}_{\setminus i}};\Theta\right)\end{aligned} \quad 5.8$$

which indeed represents a *spatial Gaussian mixture model* (SGMM) with components given by the QLL $\mathcal{P}\left(\boldsymbol{d}_i|\boldsymbol{\kappa}_{\mathcal{N}_i};\Theta\right)$ in equation 5.7, each of which is scaled with the *spatial priors* $\mathcal{P}\left(\kappa_i|\boldsymbol{\kappa}_{\mathcal{N}_{\setminus i}};\Theta\right)$ given by the MRF prior model – equation 5.4.

In a so called *generative model*, the data $\boldsymbol{d}$ are assumed to have been generated by the unobserved facies $\boldsymbol{\kappa}$ according to a probability distribution $\mathcal{P}(\boldsymbol{d}|\boldsymbol{\kappa};\Theta)$, where $\Theta$ is the set of parameters that models the dependencies between the facies and the observed data. Under the assumption of conditional independence of data $\boldsymbol{d}$ given the facies $\boldsymbol{\kappa}$ and the parameters $\Theta$, the likelihood $\mathcal{P}(\boldsymbol{d}|\boldsymbol{\kappa};\Theta)$ of observed data $\boldsymbol{d}$ given a particular facies model $\boldsymbol{\kappa}$ is given by

$$\mathcal{P}(\boldsymbol{d}|\boldsymbol{\kappa};\Theta)=\prod_i\mathcal{P}\left(\boldsymbol{d}_i|\boldsymbol{\kappa}_{\mathcal{N}_i};\Theta\right)=\prod_i N\left(\boldsymbol{R}_{\mathcal{N}_i}\boldsymbol{\beta},\boldsymbol{\Sigma}_{\varepsilon}\right) \quad 5.9$$

It is important to note here that although we use the same notation $\mathcal{N}_i$ for the neighbourhood of $i$ in the expressions for the prior (5.4) and the likelihood (5.6 to 5.9) distributions, these neighbourhood structures need not be the same. That is, we can use a different template for the neighbourhood structure to model each of these distributions.

We can write equation 5.6 for all of the $n$ cells in the model as

$$\boldsymbol{d}=\boldsymbol{R}\boldsymbol{\beta}+\boldsymbol{\varepsilon} \quad 5.10$$

where $\boldsymbol{d}=(\boldsymbol{d}_1,\boldsymbol{d}_2,\dots,\boldsymbol{d}_n)^T$ is a $np\times 1$ vector of $p$ dimensions in each of the $n$ cells, $\boldsymbol{R}=\left(\boldsymbol{R}_{\mathcal{N}_1},\boldsymbol{R}_{\mathcal{N}_2},\dots,\boldsymbol{R}_{\mathcal{N}_n}\right)^T$ is a $np\times q$ matrix of facies responses at $q$ neighbours of each of the $n$ cells, and $\boldsymbol{\varepsilon}$ is a $np\times 1$ vector of errors that are assumed to be uncorrelated with the

covariates $\boldsymbol{R}$ and are jointly distributed according to a Normal distribution $N(0, \boldsymbol{\Sigma_\varepsilon})$. Therefore, given the facies responses $\boldsymbol{R}$, the data are Normally distributed with mean $\boldsymbol{R\beta}$ and covariance matrix $\boldsymbol{\Sigma}_\varepsilon$, that is, $\boldsymbol{d}|\boldsymbol{\kappa} \sim N(\boldsymbol{R\beta}, \boldsymbol{\Sigma}_\varepsilon)$. Thus, the log-likelihood $\mathcal{L}(\Theta; \boldsymbol{d}|\boldsymbol{\kappa})$ as a function of parameters $\Theta$ may be written as

$$\mathcal{L}(\Theta; \boldsymbol{d}|\boldsymbol{\kappa}) \equiv \log \mathcal{P}(\boldsymbol{d}|\boldsymbol{\kappa}; \Theta)$$

$$= \sum_i \log \mathcal{P}\left(\boldsymbol{d}_i|\boldsymbol{\kappa}_{\mathcal{N}_i}; \Theta\right) \qquad \text{[using the conditional independence assumption over } \boldsymbol{d}\text{]}$$

$$= \sum_i \log \mathcal{P}\left(\boldsymbol{d}_i|\boldsymbol{R}_{\mathcal{N}_i}\boldsymbol{\beta}; \Theta\right) \qquad \text{[using equation 5.9]}$$

$$= \sum_i \log\left\{(2\pi)^{-n/2}|\boldsymbol{\Sigma_\varepsilon}|^{-1/2} \exp\left(\frac{-1}{2}\left(\boldsymbol{d}_i - \boldsymbol{R}_{\mathcal{N}_i}\boldsymbol{\beta}\right)^T \boldsymbol{\Sigma}_\varepsilon{}^{-1}\left(\boldsymbol{d}_i - \boldsymbol{R}_{\mathcal{N}_i}\boldsymbol{\beta}\right)\right)\right\}$$

[using mathematical definition of a Gaussian distribution]

$$= -\frac{n^2}{2}\log(2\pi) - \frac{n}{2}\log|\boldsymbol{\Sigma_\varepsilon}| - \frac{1}{2}\sum_i \left(\boldsymbol{d}_i - \boldsymbol{R}_{\mathcal{N}_i}\boldsymbol{\beta}\right)^T \boldsymbol{\Sigma}_\varepsilon{}^{-1}\left(\boldsymbol{d}_i - \boldsymbol{R}_{\mathcal{N}_i}\boldsymbol{\beta}\right)$$

$$= -\frac{n^2}{2}\log(2\pi) - \frac{n}{2}\log|\boldsymbol{\Sigma_\varepsilon}| - \frac{1}{2}(\boldsymbol{d} - \boldsymbol{R\beta})^T \boldsymbol{\Sigma}_\varepsilon{}^{-1}(\boldsymbol{d} - \boldsymbol{R\beta}) \qquad 5.11$$

The expression 5.11 for $\mathcal{L}(\Theta; \boldsymbol{d}|\boldsymbol{\kappa})$ will be used later in section 5.4.1 (The M-Step).

Under the conditional independence (CI) assumption, the QLL $\mathcal{P}(\boldsymbol{d}|\boldsymbol{\kappa}; \Theta)$ given by equation 5.9 can be written as

$$\mathcal{P}(\boldsymbol{d}|\boldsymbol{\kappa}; \Theta) = \prod_{i \in \mathcal{V}} \mathcal{P}\left(\boldsymbol{d}_i|\boldsymbol{\kappa}_{\mathcal{N}_i}; \Theta\right) = \prod_{i \in \mathcal{V}} \varphi_i\left(\boldsymbol{d}_i, \boldsymbol{\kappa}_{\mathcal{N}_i}; \Theta\right) \qquad 5.12$$

where $\varphi_i\left(\boldsymbol{d}_i, \boldsymbol{\kappa}_{\mathcal{N}_i}; \Theta\right) = \mathcal{P}\left(\boldsymbol{d}_i|\boldsymbol{\kappa}_{\mathcal{N}_i}; \Theta\right)$ represents a potential function of $\boldsymbol{d}_i$ and $\boldsymbol{\kappa}_{\mathcal{N}_i}$ that is called the *vertex potential* in a MRF model. It represents the physical dependency between observables and facies in the model, including errors in the data.

### *5.3.3 Posterior Distribution*

The posterior distribution $\mathcal{P}(\boldsymbol{\kappa}|\boldsymbol{d};\Theta)$ of facies $\boldsymbol{\kappa}$ given the observed data $\boldsymbol{d}$ and parameters $\Theta$ is given by the Bayes' theorem (equation 2.1). With the prior $\mathcal{P}(\boldsymbol{\kappa})$ given by equation 5.3 and the likelihood $\mathcal{P}(\boldsymbol{d}|\boldsymbol{\kappa},\Theta)$ given by equation 5.12, the posterior distribution $\mathcal{P}(\boldsymbol{\kappa}|\boldsymbol{d},\Theta)$ may be written as

$$\mathcal{P}(\boldsymbol{\kappa}|\boldsymbol{d};\Theta) = \frac{\mathcal{P}(\boldsymbol{d},\boldsymbol{\kappa};\Theta)}{\mathcal{P}(\boldsymbol{d};\Theta)} = \frac{1}{\mathcal{Z}'}\prod_{i\in\mathcal{V}}\varphi_i\left(\boldsymbol{d}_i,\boldsymbol{\kappa}_{\mathcal{N}_i}\right)\prod_{(i,j)\in\mathcal{E}}\psi_{ij}\left(\kappa_i,\kappa_j\right) \qquad 5.13$$

where constant $\mathcal{P}(\boldsymbol{d};\Theta)$ has been absorbed in $\mathcal{Z}'$. This demonstrates that although we only assumed that the prior distribution $\mathcal{P}(\boldsymbol{\kappa})$ on the facies $\boldsymbol{\kappa}$ is a MRF, the posterior distribution $\mathcal{P}(\boldsymbol{\kappa}|\boldsymbol{d};\Theta)$ and the joint distribution $\mathcal{P}(\boldsymbol{d},\boldsymbol{\kappa};\Theta)$ then also turn out to be MRFs as a consequence of the CI assumption (on $\boldsymbol{d}$). Note that without such an assumption the joint distribution would not be tractable, making inference impossible for models of practical interest. The above formulation is quintessentially the *generative approach* as it models the posterior distribution $\mathcal{P}(\boldsymbol{\kappa}|\boldsymbol{d};\Theta)$ via the joint distribution $\mathcal{P}(\boldsymbol{d},\boldsymbol{\kappa};\Theta)$, as opposed to the *discriminative approach* that directly models the posterior distribution (see chapter 6).

Vertex potentials $\varphi_i\left(\boldsymbol{d}_i,\boldsymbol{\kappa}_{\mathcal{N}_i};\Theta\right)$ are estimated from the data using a rock physics model of the relationship between facies and observed data. The edge potentials $\psi_{ij}\left(\kappa_i,\kappa_j\right)$, on the other hand, are estimated only from the prior information (e.g. expressed in the form of a training image). This means that any spatial correlations in the data are only used in the reconstruction of the spatial distribution of facies through the likelihood function.

The form of the probability distribution in equation 5.13 suggests that this model is an undirected alternative to a 2D-HMM (chapter 4; Nawaz & Curtis, 2017). Although causality in a HMM has no direct physical interpretation in a spatial context, this allows for analytical computation of posterior probabilities. A MRF (or HMRF), on the other hand, is a more natural representation of spatial phenomena but it does not allow analytical computation of posterior probabilities because of the intractable normalizing constant $\mathcal{Z}'$.

## 5.4 Variational Bayesian Inference

We use the variational Bayes (VB) method for probabilistic inference that was introduced in section 2.4. Besides its computational efficiency, the variational Bayes method is a natural choice for probabilistic inference in our current model due to the fact that the posterior distribution $\mathcal{P}(\boldsymbol{\kappa}|\boldsymbol{d};\Theta)$ as given by equation 5.13 is fully factorized as a consequence of the CI assumption (on $\boldsymbol{d}$). We recall from section 2.4 that the VB method approximates the intractable posterior distribution $\mathcal{P}(\boldsymbol{\kappa}|\boldsymbol{d};\Theta)$ by replacing it with a tractable approximation $\mathcal{Q}(\boldsymbol{\kappa}|\boldsymbol{d})$, or simply $\mathcal{Q}$, the so called *variational distribution*, from a family $\mathbb{Q}$ of distributions that are more easily manipulated.

In our current model, equation 2.9 takes the form

$$\mathcal{L}(\Theta;\boldsymbol{d}) = \mathcal{F}(\mathcal{Q},\Theta) + KL\big(\mathcal{Q}(\boldsymbol{\kappa}|\boldsymbol{d})||\mathcal{P}(\boldsymbol{\kappa}|\boldsymbol{d};\Theta)\big) \tag{5.14}$$

where the relative entropy (the second term on RHS of equation 5.14) is given by

$$KL(\mathcal{Q}||\mathcal{P}) = \mathbb{E}_{\mathcal{Q}}\left[log\frac{\mathcal{Q}(\boldsymbol{\kappa}|\boldsymbol{d})}{\mathcal{P}(\boldsymbol{\kappa}|\boldsymbol{d};\Theta)}\right] = \sum_{\boldsymbol{\kappa}} \mathcal{Q}(\boldsymbol{\kappa}|\boldsymbol{d})\, log\frac{\mathcal{Q}(\boldsymbol{\kappa}|\boldsymbol{d})}{\mathcal{P}(\boldsymbol{\kappa}|\boldsymbol{d};\Theta)} \tag{5.15}$$

and $\mathcal{F}(\mathcal{Q},\Theta)$ (the free energy functional) is a lower bound on the log-evidence $\mathcal{L}(\Theta;d)$. $\mathcal{F}(\mathcal{Q},\Theta)$ may be defined in terms of the entropy $\mathcal{S}(\mathcal{Q})$ of $\mathcal{Q}(\boldsymbol{\kappa}|\boldsymbol{d})$, and expected joint log-likelihood $\mathbb{E}_{\mathcal{Q}}[\mathcal{L}(\Theta;\boldsymbol{d},\boldsymbol{\kappa})]$ (see equation 2.8) as

$$\mathcal{F}(\mathcal{Q},\Theta) \equiv \mathbb{E}_{\mathcal{Q}}[\mathcal{L}(\Theta;\boldsymbol{d},\boldsymbol{\kappa})] + \mathcal{S}(\mathcal{Q}) \tag{5.16}$$

Since $\mathcal{P}(\boldsymbol{d},\boldsymbol{\kappa};\Theta)$ factorizes over the cliques in a MRF by definition (due to CI assumption), it follows from equation 5.16 that $\mathbb{E}_{\mathcal{Q}}[\mathcal{L}(\Theta;\boldsymbol{d},\boldsymbol{\kappa})]$ can be computed efficiently, but estimation of $\mathcal{S}(\mathcal{Q})$, and hence $\mathcal{F}(\mathcal{Q},\Theta)$, is still computationally expensive. In order to overcome this difficulty, we use a variational form of the *expectation-maximization* (EM) algorithm (Dempster *et al.*, 1977; Beal, 2003) which approximates $\mathcal{F}(\mathcal{Q},\Theta)$ in an iterative fashion such that the lower-bound $\mathcal{F}(\mathcal{Q},\Theta)$ of $\mathcal{L}(\Theta;\boldsymbol{d})$ is increased while decreasing $KL(\mathcal{Q}||\mathcal{P})$ within each iteration.

### *5.4.1 The Expectation-Maximization (EM) Algorithm*

The EM algorithm involves two steps in each iteration: the so-called E-step and the M-step, which aim to alternately maximize the free-energy $\mathcal{F}(Q,\Theta)$ with respect to $Q$ and $\Theta$, respectively. In concept, the E-step operates in the 'model space' to estimate the posterior distribution $Q$ of facies $\boldsymbol{\kappa}$ (which factorizes in a MRF as shown by equation 3.7) for a given estimate of parameters $\Theta$, whereas the M-step operates in the 'attributes space' to update the current estimate of parameters $\Theta$ by maximizing their likelihood for the current estimate of the posterior distribution $Q$ of facies (figure 5.3).

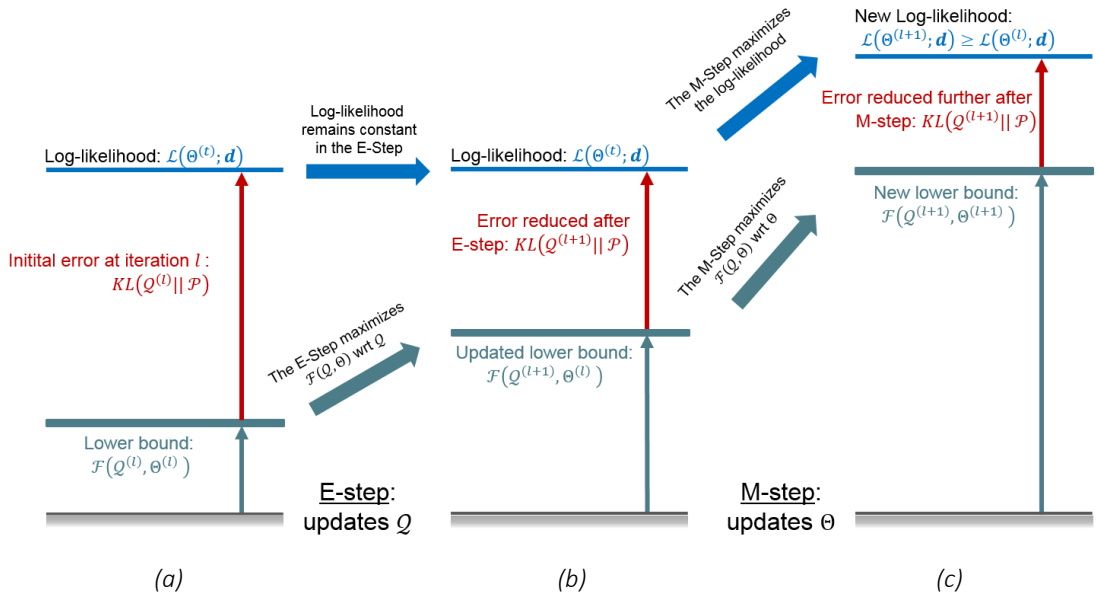



*Figure 5.3: A schematic illustration of the EM algorithm. (a) After $l$ iterations, we have estimates of the variational distribution as $Q^{(l)}$ and the parameters $\Theta^{(l)}$. (b) The E-step of the $l+1^{th}$ iteration maximizes the lower-bound (free energy) $\mathcal{F}(Q^{(l)},\Theta^{(l)})$ with respect to $Q$ which is updated to $Q^{(l+1)}$ and the lower-bound is updated to $\mathcal{F}(Q^{(l+1)},\Theta^{(l)}) \geq \mathcal{F}(Q^{(l)},\Theta^{(l)})$. The log-likelihood $\mathcal{L}(\Theta^{(l)};\boldsymbol{d})$ remains constant during the E-step for fixed $\Theta^{(l)}$, and as a consequence the error term is reduced to $KL\left(Q^{(l+1)}||\mathcal{P}(\boldsymbol{\kappa}|\boldsymbol{d},\Theta^{(l)})\right) \leq KL\left(Q^{(l)}||\mathcal{P}(\boldsymbol{\kappa}|\boldsymbol{d},\Theta^{(l)})\right)$. (c) The M-step updates parameters $\Theta^{(l)}$ to $\Theta^{(l+1)}$ by maximizing the log-likelihood to $\mathcal{L}(\Theta^{(l+1)};\boldsymbol{d}) \geq \mathcal{L}(\Theta^{(l)};\boldsymbol{d})$ thereby maximizing the lower-bound to $\mathcal{F}(Q^{(l+1)},\Theta^{(l+1)}) \geq \mathcal{F}(Q^{(l+1)},\Theta^{(l)})$ and reducing the error further to $KL\left(Q^{(l+1)}||\mathcal{P}(\boldsymbol{\kappa}|\boldsymbol{d},\Theta^{(l+1)})\right)$. In this manner, the error keeps on reducing monotonically in each iteration of the EM algorithm which iterates until convergence.*

Alternate E- and M-steps therefore improve the estimates of $\mathcal{Q}$ and $\Theta$ such that the free energy $\mathcal{F}(\mathcal{Q},\Theta)$ is guaranteed not to decrease in any iteration. With a suitable initialization, the EM algorithm is guaranteed to converge to a local optimum within a relatively small number of iterations (Balakrishnan *et al*. 2017).

## The E-Step

In the E-step of iteration $l$, the variational distribution $\mathcal{Q}(\boldsymbol{\kappa}|\boldsymbol{d})$ over the facies $\boldsymbol{\kappa}$ is estimated from the current estimate of the model parameters $\Theta^{(l)}$ by maximizing the free-energy $\mathcal{F}(\mathcal{Q},\Theta)$ with respect to $\mathcal{Q}$. The E-step may therefore be written as

$$\mathcal{Q}^{(l+1)} = \underset{\mathcal{Q}}{\operatorname{argmax}}\{\mathcal{F}(\mathcal{Q},\Theta^{(l)})\} \tag{5.17}$$

where the bracketed superscripts refer to the iteration number. Since $\mathcal{F}(\mathcal{Q},\Theta)$ is a lower bound of $\mathcal{L}(\Theta;\boldsymbol{d})$ (by equation 5.14), maximizing the lower bound $\mathcal{F}(\mathcal{Q},\Theta^{(l)})$ of the log-evidence $\mathcal{L}(\Theta^{(l)};\boldsymbol{d})$ with respect to $\mathcal{Q}$ results in $\mathcal{Q}^{(l+1)}$ equal to the estimate $\hat{\mathcal{P}}(\boldsymbol{\kappa}|\boldsymbol{d},\Theta^{(l)})$ of the true but unknown posterior distribution $\mathcal{P}(\boldsymbol{\kappa}|\boldsymbol{d},\Theta)$. This can be proved by setting $\mathcal{Q}$ equal to $\mathcal{P}(\boldsymbol{\kappa}|\boldsymbol{d},\Theta^{(l)})$ in the inequality 2.6.

Since exact evaluation of the free energy $\mathcal{F}(\mathcal{Q},\Theta)$ is intractable, we seek more efficient approximate alternatives. The distribution $\mathcal{Q}(\boldsymbol{\kappa}|\boldsymbol{d})$ in a pairwise MRF may be specified by approximate marginal distributions $b_i(\kappa_i)$ over the vertices, and $b_{ij}(\kappa_i,\kappa_j)$ over the edges in the graphical model as defined below. The negative of the free energy $-\mathcal{F}(\mathcal{Q},\Theta)$ can then be approximated for pairwise MRFs by the *Bethe's free energy* $\hat{\mathcal{F}}_B$, also called *Kikuchi free energy* for general MRFs (Yedidia *et al*. 2001a, b), given by

$$\begin{aligned}\hat{\mathcal{F}}_B = &\sum_{(i,j)\epsilon\mathcal{E}}\sum_{(\kappa_i,\kappa_j)} b_{ij}(\kappa_i,\kappa_j)\log\left(\frac{b_{ij}(\kappa_i,\kappa_j)}{\varphi_i(\kappa_i)\,\varphi_j(\kappa_j)\,\psi_{ij}(\kappa_i,\kappa_j)}\right)\\ &-\sum_{i\epsilon\mathcal{V}}(|\mathcal{N}_{\setminus i}|-1)\sum_{\kappa_i} b_i(\kappa_i)\log\left(\frac{b_i(\kappa_i)}{\varphi_i(\kappa_i)}\right)\end{aligned} \tag{5.18}$$

where $|\mathcal{N}_{\setminus i}|$ represents the neighbourhood cardinality of $i$ (excluding $i$), i.e., the number of vertices that are neighbours of $i$. The Bethe's free energy only approximates the entropy term $\mathcal{S}(\mathcal{Q})$ in equation 5.16 which is hard to compute; the expectation term $\mathbb{E}_{\mathcal{Q}}[\mathcal{L}(\Theta;\boldsymbol{d},\boldsymbol{\kappa})]$ remains

exact. The approximate marginal distributions $b_i(\kappa_i)$ and $b_{ij}(\kappa_i, \kappa_j)$ are commonly referred to as *pseudo-marginals* or *beliefs*. The above expression for Bethe's free energy $\hat{\mathcal{F}}_B$ is a direct consequence of a re-parametrization of the posterior distribution from the original parameters in terms of potential functions $\varphi_i(\kappa_i)$ and $\psi_{ij}(\kappa_i, \kappa_j)$, to the new parameters in terms of beliefs $b_i(\kappa_i)$ and $b_{ij}(\kappa_i, \kappa_j)$ under the following so called *admissibility constraints*:

$$\mathcal{P}(\boldsymbol{\kappa}|\boldsymbol{d}) = \frac{1}{\mathcal{Z}} \prod_i \varphi_i(\boldsymbol{d}_i, \boldsymbol{\kappa}_{\mathcal{N}_i}) \prod_{(i,j)\epsilon\mathcal{E}} \psi_{ij}(\kappa_i, \kappa_j) \propto \prod_{(i,j)\epsilon\mathcal{E}} b_{ij}(\kappa_i, \kappa_j) \Big/ \prod_i b_i(\kappa_i)^{|\mathcal{N}_{\setminus i}|-1} \tag{5.19}$$

The Bethe's free energy $\hat{\mathcal{F}}_B$ is exactly equal to the free energy $\mathcal{F}(\mathcal{Q}, \Theta)$ for an acyclic (1D linear or tree-structured) pairwise MRF (Koller & Friedman, 2009). In general MRFs, Yedidia *et al.* (2001a, b) showed that the stationary points of Bethe's free-energy correspond to the fixed points of an iterative message-passing algorithm, the so called *belief propagation* (BP) algorithm introduced by Pearl, 1982.

BP performs approximate inference in graphical models by estimating marginal distributions of unobserved variables conditioned on any observed variables by passing messages over edges in the graph. A message $m_{j\to i}(\kappa_i)$ from the vertex $j$ to the vertex $i$ is a real function with domain $\kappa_i$, the set of values that can be taken by an unobserved vertex $i$, and represents probabilistic influence of a vertex $j$ on the vertex $i$. In other words, a message $m_{j\to i}(\kappa_i)$ encodes 'belief' of a vertex $j$ about the state $\kappa_i$ of an unobserved vertex $i$. The beliefs $b_i(\kappa_i)$ and $b_{ij}(\kappa_i, \kappa_j)$ can be expressed in terms of messages as

$$b_i(\kappa_i) \propto \varphi_i(\kappa_i) \prod_{j\in\mathcal{N}_{\setminus i}} m_{j\to i}(\kappa_i) \tag{5.20}$$

$$b_{ij}(\kappa_i, \kappa_j) \propto \varphi_i(\kappa_i)\varphi_j(\kappa_j)\psi_{ij}(\kappa_i, \kappa_j) \prod_{h\in\mathcal{N}_{\setminus i}\setminus\{j\}} m_{h\to i}(\kappa_i) \prod_{h\in\mathcal{N}_{\setminus j}\setminus\{i\}} m_{h\to j}(\kappa_j) \tag{5.21}$$

Combining these equations yields the BP equation (Pearl, 1982)

$$m_{j\to i}(\kappa_i) \propto \sum_{\kappa_j} \varphi_j(\kappa_j)\psi_{ij}(\kappa_i, \kappa_j) \prod_{h\in\mathcal{N}_{\setminus j}\setminus\{i\}} m_{h\to j}(\kappa_j) \tag{5.22}$$

which forms a schedule for message passing, and shows how a vertex encodes messages that it receives from its neighbours except the target vertex, and passes the encoded messages to its target neighbouring vertex. The schedule starts with a vertex $j$ receiving messages

$m_{h\to j}(\kappa_j)$ from each of its neighbours $h \in \mathcal{N}_{\setminus j}\backslash\{i\}$ except its target vertex $i$. Figure 5.4 shows a schematic illustration of the schedule of messages received by a given vertex from its neighbours except the target vertex, and the message it sends to its target neighbouring vertex. The received messages are multiplied together for each of the possible values of $\kappa_j$ and then scaled with the vertex and edge potentials $\varphi_j(\kappa_j)$ and $\psi_{ij}(\kappa_i, \kappa_j)$ for a given value of the state $\kappa_i$ of $i$. The resulting scaled products of messages are then summed over all of the possible values of $\kappa_j$ and then forwarded by the vertex $j$ to the vertex $i$ encoding the belief of $j$ regarding the state of $i$ being equal to $\kappa_i$. The observed vertices in a HMRF also send messages to their neighbouring hidden vertices, however they cannot receive any messages as their values are fixed.

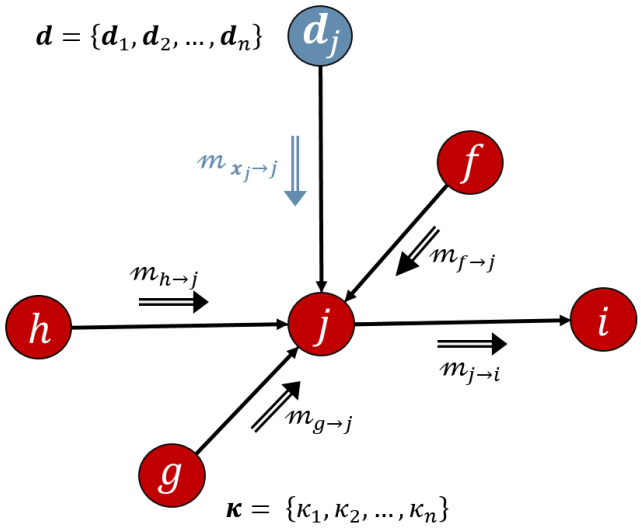



*Figure 5.4: A schematic illustration of message passing. The blue circle represents an observed vertex (or variable), red circles represent hidden vertices, and the solid lines connecting these circles represent edges in the graphical model. Double-lined arrows represent messages flowing between vertices as labelled. The vertex $j$ receives messages $m_{\cdot\to j}$ from all of its neighbours (including the observed vertex $\boldsymbol{d}_j$) except the vertex $i$ which is the current target for a message from $j$. The messages received by $j$ are combined together and encoded into a message $m_{j\to i}$ according to equation 5.22. The encoded message $m_{j\to i}$ is then forwarded by $j$ to $i$. Only hidden vertices can receive messages. Observed vertices can only send messages to their neighbouring hidden vertices, and cannot receive any messages as their values are fixed. Propagation of messages in this manner between all vertices in a graph constitutes what is commonly known as the belief propagation (BP) algorithm.*

*Algorithm 5.1*: Loopy-belief propagation (LBP) over an undirected graphical model $\mathbb{G}(\mathcal{V}, \mathcal{E})$ with accuracy $\varepsilon$ and for a maximum number of iterations $L$. Comments follow the hash signs '#' till the end of each line.

1. *Set sum_product ← true # or false for max-product*
2. *Initialize messages* $m_{j\to i}{}^{(0)}(\kappa_i)$
3. *Set* $l \leftarrow 1$ *# LBP iteration number*
4. *while* $l \leq L$
5. *Set* $\delta \leftarrow 0$
6. *for each* $i \in \mathcal{V}$
7. *for each* $j \in \mathcal{N}_{\backslash i} \subset \mathcal{V}$
8. *if sum_product # for sum-product algorithm*
9. *Compute* $m_{j\to i}{}^{(l)}(\kappa_i)$ *using equation 5.22*
10. *else # for max-product algorithm*
11. *Compute* $m_{j\to i}{}^{(l)}(\kappa_i)$ *using equation 5.23*
12. *end if*
13. *Set* $\delta \leftarrow max\left(\delta,\ m_{j\to i}{}^{(l)}(\kappa_i) - m_{j\to i}{}^{(l-1)}(\kappa_i)\right)$
14. *end for j*
15. *end for i*
16. *if* $\delta < \varepsilon$
17. *Update beliefs using equations 5.20 and 5.21*
18. *print 'Converged!'*
19. *exit*
20. *end if*
21. *Set* $l \leftarrow l + 1$
22. *end while*
23. *print 'Not converged!'*

Equation 5.22 is often referred to as the *sum-product equation* for obvious reasons, and forms the basis of the BP algorithm. The BP algorithm is an exact inference method for tree-structured (or 1D) graphs in which case it can be shown to converge to the true marginal distributions in a number of iterations equal to the diameter of the tree – the maximum number of edges between any two vertices in the graph (Koller & Friedman, 2009; Loïc, 2016). In cyclic graphs (such as used in spatial problems and in this research), a variant of BP known

as the *loopy-belief propagation* (LBP) can be used which is an approximate inference method. LBP is not guaranteed to converge, however, it has been shown empirically to converge in most cases (Pearl, 1982 & 1988; Murphy *et al.* 1999). We discuss this point further in section 5.7. Nevertheless, the LBP algorithm has seen wide applicability and success in various fields of research, for example in statistics (e.g., Pearl 1988; Yasuda, 2015), digital signal and image processing (e.g., Sudderth & Freeman, 2008), artificial intelligence (e.g., Tatikonda & Jordan, 2002) and biology (e.g., Sinoquet & Mourad, 2014). In LBP, the messages are passed iteratively until convergence is detected or until a maximum number of iterations is exceeded. Convergence may be detected if all vertices are updated by an amount less than a predefined tolerance.Messages are generally initialized with unity or with random numbers greater than a positive tolerance, and then updated according to a pre-defined message schedule using equation 5.22. After the messages have converged based on some convergence detection criteria, the vertex beliefs are updated according to equation 5.20 to give approximate marginal posterior distributions. Despite that the vertex potentials $\varphi_i(\kappa_i)$ and the edge potentials $\psi_{ij}(\kappa_i, \kappa_j)$ need not be exact probabilities, their marginalization and normalization ensures numerical stability of the LBP algorithm. Also, since the LBP involves several iterative multiplications of potential functions at each vertex, the LBP algorithm is usually run in the logarithmic domain in order to avoid numerical underflow.

The LBP algorithm may also be used to perform *maximum-a-posteriori* (MAP) inference which computes the most likely configuration, rather than the approximate marginal posterior distributions. MAP inference minimizes the error probability that the most likely configuration, also known as the *MAP estimate*, does not coincide with the true one. This can be achieved by replacing the summation in the sum-product equation 5.22 with the *max* function yielding the corresponding *max-product equation* as

$$m_{j\to i}(\kappa_i) \propto \max_{\kappa_j}\left\{\varphi_j(\kappa_j)\psi_{ij}(\kappa_i,\kappa_j)\prod_{h\in\mathcal{N}_{\backslash j}\backslash\{i\}} m_{h\to j}(\kappa_j)\right\} \tag{5.23}$$

The LBP algorithm on a MRF is summarized in *Algorithm 5.1*. If *Algorithm 5.1* converges, the beliefs $b_i(\kappa_i)$ and $b_{ij}(\kappa_i, \kappa_j)$ are updated using equations 5.20 and 5.21. The variational distribution $Q^{(l+1)}$ at the end of the E-step of $(l+1)^{\text{th}}$ iteration of the EM algorithm is then approximated to $\mathcal{P}(\boldsymbol{\kappa}|\boldsymbol{d})$ using equation 5.19.

## The M-Step

In the M-step, the current estimate of the variational distribution $\mathcal{Q}^{(l+1)}$ obtained from the E-step is used to compute the updated set of parameters $\Theta^{(l+1)}$ that maximize the free-energy $\mathcal{F}(\mathcal{Q},\Theta)$ with respect to $\Theta$. The M-step may therefore be written as

$$\Theta^{(l+1)} = \underset{\Theta}{\operatorname{argmax}}\, \mathcal{F}\left(\mathcal{Q}^{(l+1)},\Theta\right) = \underset{\Theta}{\operatorname{argmax}}\, \mathbb{E}_{\mathcal{Q}^{(l+1)}}[\mathcal{L}(\Theta;\boldsymbol{d},\boldsymbol{\kappa})] \tag{5.24}$$

which follows from the fact that $\mathcal{S}(\mathcal{Q})$ in equation 5.16 is independent of $\Theta$. Thus maximizing $\mathcal{F}(\mathcal{Q},\Theta)$ with respect to $\Theta$ only requires that $\mathbb{E}_{\mathcal{Q}}[\mathcal{L}(\Theta;\boldsymbol{d},\boldsymbol{\kappa})]$ be maximized with respect to $\Theta$. Accordingly, it turns out that the M-step may only require a few statistics of the facies model $\boldsymbol{\kappa}$ computed in the E-step, instead of the full distribution $\mathcal{Q}^{(l+1)}(\boldsymbol{\kappa}|\boldsymbol{d})$. Expanding equation 5.24 in terms of the log-evidence and substituting from equation 5.11 gives

$$\begin{aligned}\Theta^{(l+1)} = \underset{\Theta}{\operatorname{argmax}} \sum_i \sum_{\kappa_i} {b_i}^{(l+1)}(\kappa_i) \Bigg( & -\frac{n^2}{2}\log(2\pi) - \frac{n}{2}\log|\boldsymbol{\Sigma}_\varepsilon| \\ & -\frac{1}{2}\left(\boldsymbol{d}_i - \boldsymbol{R}_{\mathcal{N}_i}\boldsymbol{\beta}\right)^T {\boldsymbol{\Sigma}_\varepsilon}^{-1}\left(\boldsymbol{d}_i - \boldsymbol{R}_{\mathcal{N}_i}\boldsymbol{\beta}\right)\Bigg)\end{aligned} \tag{5.25}$$

The solution to the above equation can be obtained with and without the assumption of *homoscedasticity* whereby the covariance matrix $\boldsymbol{\Sigma}_\varepsilon$ is assumed to be scalar such that $\boldsymbol{\Sigma}_\varepsilon = \sigma^2\boldsymbol{I}$. With this assumption, maximizing log-likelihood under the constraints $\sum_{\kappa_i} b_i(\kappa_i) = 1$ is equivalent to minimizing the residual sum-of-squares

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_i \left(\boldsymbol{d}_i - \boldsymbol{R}_{\mathcal{N}_i}\boldsymbol{\beta}\right)^T \left(\boldsymbol{d}_i - \boldsymbol{R}_{\mathcal{N}_i}\boldsymbol{\beta}\right) \tag{5.26}$$

which gives the *ordinary least-squares* (OLS) solution

$$\widehat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{R}^T\boldsymbol{R})^{-1}\boldsymbol{R}^T\boldsymbol{d} \tag{5.27}$$

The OLS solution is also the unbiased maximum-likelihood solution if $\boldsymbol{R}$ is a full-rank matrix, otherwise one may seek the *regularized least squares* (RLS) solution given by

$$\widehat{\boldsymbol{\beta}}_{RLS} = (\boldsymbol{R}^T\boldsymbol{R} + k\boldsymbol{I})^{-1}\boldsymbol{R}^T\boldsymbol{d} \tag{5.28}$$

where $k$ is the control (regularization) parameter which governs the relative strength of regularization (damping) applied. Similarly, the maximum-likelihood solution of equation 5.25 with respect to $\boldsymbol{\Sigma_\varepsilon} = \sigma^2 \boldsymbol{I}$ is given by

$$\hat{\sigma}^2 = \frac{1}{n}\sum_i \left(\boldsymbol{d}_i - \boldsymbol{R}_{\mathcal{N}_i}\widehat{\boldsymbol{\beta}}\right)^T \left(\boldsymbol{d}_i - \boldsymbol{R}_{\mathcal{N}_i}\widehat{\boldsymbol{\beta}}\right) \tag{5.29}$$

but this is a biased estimator; the bias-corrected estimate (Rencher, 2002) is given by

$$\hat{\sigma}^2 = \frac{1}{n-q-1}\sum_i \left(\boldsymbol{d}_i - \boldsymbol{R}_{\mathcal{N}_i}\widehat{\boldsymbol{\beta}}\right)^T \left(\boldsymbol{d}_i - \boldsymbol{R}_{\mathcal{N}_i}\widehat{\boldsymbol{\beta}}\right) \tag{5.30}$$

where we recall that $q = |\mathcal{N}_i|$ is the neighbourhood cardinality, which is assumed to be a constant for each location $i$ in our graphical model. In the general case of *heteroscedasticity* whereby the covariance matrix is non-scalar, maximizing log-likelihood is equivalent to minimizing the residual weighted sum-of-squares

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmin}} \left\{ n \log|\boldsymbol{\Sigma_\varepsilon}| + \sum_i \left(\boldsymbol{d}_i - \boldsymbol{R}_{\mathcal{N}_i}\boldsymbol{\beta}\right)^T \boldsymbol{\Sigma_\varepsilon}^{-1} \left(\boldsymbol{d}_i - \boldsymbol{R}_{\mathcal{N}_i}\boldsymbol{\beta}\right) \right\} \tag{5.31}$$

With $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon}} = (\hat{\sigma}_{kl}),\ k,l \in \{1,\dots,p\}$ where $\hat{\sigma}_{kl}$ is estimated using equation 5.30, the *generalized least-squares* (GLS) solution is given by

$$\widehat{\boldsymbol{\beta}}_{GLS} = \left(\boldsymbol{R}^T \left(\boldsymbol{I}_n \otimes \widehat{\boldsymbol{\Sigma}}_\varepsilon\right)^{-1} \boldsymbol{R}\right)^{-1} \boldsymbol{R}^T \left(\boldsymbol{I}_n \otimes \widehat{\boldsymbol{\Sigma}}_\varepsilon\right)^{-1} \boldsymbol{d} \tag{5.32}$$

where $\otimes$ represents the *Kronecker product* that multiplies a matrix with each element of the other matrix. Mathematically, it is defined between two matrices $\boldsymbol{A} = [a_{mn}]$ and $\boldsymbol{B} = [b_{pq}]$ as a $(mp \times nq)$ matrix with elements

$$(\boldsymbol{A} \otimes \boldsymbol{B})_{i,j} = a_{\lfloor (i-1)/p \rfloor + 1, \lfloor (j-1)/q \rfloor + 1}\, b_{i - \lfloor (i-1)/p \rfloor p,\, j - \lfloor (j-1)/q \rfloor q}, \tag{5.33}$$

where $\lfloor \cdot \rfloor$ represents the *floor* function which returns the greatest integer less than or equal to its argument.

The parameters $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k,\ k \in \mathcal{G}$ of the GM distribution are iteratively updated by weighted averages of the data $\boldsymbol{d}_i$ at each location $i$ with respect to the current estimates of the posterior marginal distributions $\hat{\mathcal{P}}_i(\kappa_i|\boldsymbol{d};\Theta)$ as estimated in the E-step of the current iteration $l$, as

$$\boldsymbol{\mu}_k{}^{(l+1)} = \frac{\sum_{i=1}^{n} \hat{\mathcal{P}}_i\left(\kappa_i|\boldsymbol{d}, \Theta^{(l)}\right)\boldsymbol{d}_i}{\sum_{i=1}^{n} \hat{\mathcal{P}}_i(\kappa_i|\boldsymbol{d}, \Theta^{(l)})} \tag{5.34}$$

$$\boldsymbol{\Sigma}_k{}^{(l+1)} = \frac{\sum_{i=1}^{n} \hat{\mathcal{P}}_i\left(\kappa_i|\boldsymbol{d}, \Theta^{(l)}\right) \cdot \left(\boldsymbol{d}_i - \boldsymbol{\mu}_k{}^{(l+1)}\right)\left(\boldsymbol{d}_i - \boldsymbol{\mu}_k{}^{(l+1)}\right)^T}{\sum_{i=1}^{n} \hat{\mathcal{P}}_i(\kappa_i|\boldsymbol{d}, \Theta^{(l)})} \tag{5.35}$$

where $\hat{\mathcal{P}}_i\left(\kappa_i|\boldsymbol{d}, \Theta^{(l)}\right)$ is approximated by the vertex beliefs $b_i{}^{(l)}(\kappa_i)$ estimated from the LBP algorithm in the E-step of $l^{\text{th}}$ iteration.

In summary, at the end of $(l+1)^{\text{th}}$ iteration the E-Step of the EM algorithm yields the free energy $\mathcal{F}\left(\mathcal{Q}^{(l+1)}, \Theta^{(l)}\right)$ as an approximation to $\mathcal{L}\left(\Theta^{(l)}; \boldsymbol{d}\right)$ which is the upper bound of $\mathcal{F}\left(\mathcal{Q}, \Theta^{(l)}\right)$, and the M-step maximizes $\mathcal{F}\left(\mathcal{Q}^{(l+1)}, \Theta^{(l)}\right)$ with respect to $\Theta$. Therefore the E-step improves the estimate of the posterior distribution of facies $\hat{\mathcal{P}}(\boldsymbol{\kappa}|\boldsymbol{d}; \Theta)$ in the model space while the M-step improves the estimates of model parameters $\Theta$ in the attribute space, such that the combined E-M steps are guaranteed not to decrease the lower bound $\mathcal{F}(\mathcal{Q}, \Theta)$ of $\mathcal{L}(\Theta; \boldsymbol{d})$ during any iteration of the EM algorithm.

## 5.5 Computational Complexity

The computational complexity of this algorithm is defined by the cost of the LBP algorithm in the E-step and the solution of the linear problem in the M-step. The computational cost of LBP algorithm depends on the number of iterations required for convergence. The cost of E-step is therefore given by

$$C_E \leq n * K^2 * \max|\mathcal{N}| * L \tag{5.36}$$

where $n = |\mathcal{V}|$ is the number of locations (vertices in the graph), $K = |\mathcal{G}|$ is the number of facies considered, $\max|\mathcal{N}|$ represents the maximum neighbourhood cardinality (the maximum number of neighbouring vertices $\mathcal{N}_{\setminus i}$ of any vertex $i \in \mathcal{V}$ in the graph), and $L$ is the total number of iterations in the LBP algorithm.

Although there are cases when LBP does not converge (as in the case of repulsive potential functions, Koller & Friedman, 2009), we consider the number of iterations assuming that the algorithm does converge. If the LBP converges, the required number of iterations depends on the desired accuracy in *Algorithm 5.1*, the initial values of beliefs, model size and

complexity. Initial beliefs close to a local optimum result in a smaller number of iterations. A good choice for initial beliefs are the localized likelihoods as were computed in chapter 4 (also see Walker & Curtis (2014a); and Nawaz & Curtis (2017)). Starting with reasonable initial beliefs, the LBP algorithm requires 10's to 100's of iterations in most cases, depending on the model size and complexity. To limit computational demands in large models, the regions in the graph in which beliefs do not change significantly in some pre-defined number of previous iterations may be eliminated from future iterations, thus effectively reducing the size of the graph as the number of iterations increases.

The computational cost of the M-step is given by

$$C_M = n * p * (\max|\mathcal{N}|)^2 \qquad 5.37$$

where $p$ is the dimensionality of data observed at each location. The total computational cost of the EM algorithm is therefore

$$C_{total} \leq (C_E + C_M) * L_{EM} \qquad 5.38$$

where $L_{EM}$ is the total number of EM iterations.

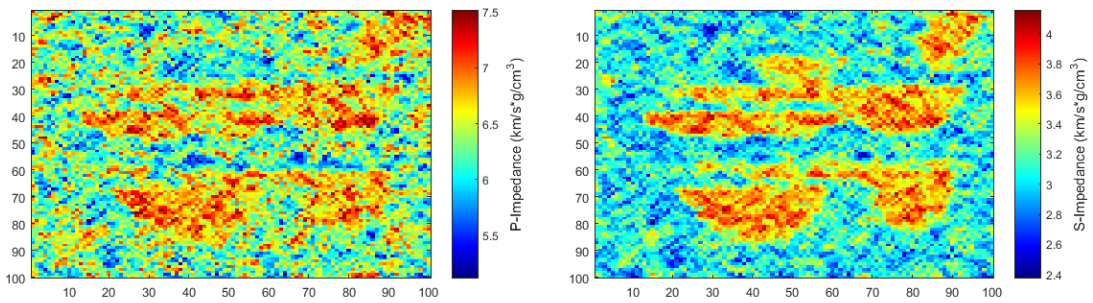
Convergence of the EM algorithm is fast and guaranteed provided that the LBP algorithm in the E-step converges. Expressions 5.36 and 5.37 are quadratic in the number of facies ($K = |\mathcal{G}|$) and the maximum size ($\max|\mathcal{N}|$) of the neighbourhood structure in the graph. The size of the neighbourhood structure defines the extent of spatial correlations in data that is incorporated within the likelihood function. The maximum size of the neighbourhood structure must therefore not be excessively large in order to avoid prohibitive computational costs. For this reason, the likelihoods in this method cannot be solved in fully non-localized form in realistic problems, hence the term *quasi-localized*. Therefore, similar to many other inversion methods in geostatistics, the current method is also based on a trade-off between computational tractability and the extent of spatial correlations incorporated from the data. All other parameters in the expressions 5.36 and 5.37 are linear and therefore do not cause serious computational implications.

For large models which require parallelization of the algorithm in order to improve computational speed, each iteration of the LBP algorithm in the E-step may be parallelized over the vertices of the graph (the *for* loop in line 6 of ***Algorithm 5.1***). A key consideration concerning the convergence and computational performance of the LBP algorithm is message

scheduling. Although synchronous scheduling may be desired where all of the messages are updated at once for higher computation efficiency, an asynchronous schedule is optimal both for convergence and performance. Koller & Friedman (2009) suggested a residual belief propagation schedule which dynamically detects convergence in different parts of the graph and schedules messages in the parts where beliefs disagree most strongly. Also, the solution of the linear problem in the M-step may be parallelized to improve performance (e.g., Koc & Piedra, 1991).

## 5.6 Synthetic Test

In order to test this method, and in particular to benchmark it against previous research, synthetic seismic attribute data are generated similar to and using the same synthetic model as was used in chapter 4, section 4.6. The prior information was extracted by scanning the training image (figure 4.5a) in terms of prior probabilities $\mathcal{P}\left(\kappa_i|\kappa_j \in \boldsymbol{\kappa}_{\mathcal{N}_{\setminus i}}\right)$ constructed from histograms of various facies configurations that occur in the image, under the assumption that they are stationary over the entire model space.




*(a)* *(b)*

*Figure 5.5: Synthetic (a) P-wave and (b) S-wave impedance attributes first sampled independently in each cell of the target cross-section in figure 4.5b using a probabilistic forward model based on a Gaussian distribution per facies (see equation 4.20). The impedance sections thus obtained are then spatially filtered using the 5x5 banana-shaped kernel in equation 5.39 to mimic blurring caused by non-localized effects of seismic data processing.*

Localized synthetic seismic attributes $\boldsymbol{d}_i$ (P-wave and S-wave impedances) in each model cell were first generated using the same rock physics model as described in section 4.6. In order to model the non-localized blurring effect of seismic imaging, collocated synthetic

seismic attributes, $\boldsymbol{d}_i$, were then simulated in each cell $i$ from local facies responses $\boldsymbol{d'}_{\mathcal{N}_i}$ within the neighbourhood $\mathcal{N}_i$ of $i$. This is achieved by using a Gaussian likelihood (see equation 5.12). The spatial filter $\boldsymbol{\beta}$ was chosen as a 5x5 banana-shaped kernel to represent the kind of blurring that may take place during seismic migration due to inaccurate velocity model:

$$\boldsymbol{\beta} = \mathrm{vec}\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.25 & 0 & 0 \\ 0 & 0.125 & 0.125 & 0.125 & 0 \\ 0.0625 & 0.125 & 0 & 0.125 & 0.0625 \end{bmatrix} \tag{5.39}$$

where $\mathrm{vec}[\cdot]$ is any function that transforms its argument matrix to a vector, and the resulting P-wave and S-wave impedances are shown in figure *5.5*. The impedance profiles thus obtained show blurring due to spatial averaging effect of the filter. This results in significant overlap of data representing different facies in the attributes space. Thus, conventional clustering methods that do not account for the spatial nature of the data may not reliably discriminate between multiple facies, and this is what we want to achieve with the current method.

The synthetic seismic attributes were then inverted with the aim to reproduce the target cross-section (figure 4.5b). The initial estimates of parameters $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k, k \in \mathcal{G}$ of the Gaussian mixture distribution were obtained using a *mixture density neural network* (MDN; Meier et al. 2007a,b & 2009; Shahraeeni & Curtis, 2011; Shahraeeni *et al.* 2012) based on clustering of seismic attributes. In a real problem, estimates of these parameters may also be obtained from prior information based on well-logs or other data sources. The localized likelihoods were estimated from a GMM with components $N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), k \in \mathcal{G}$. The spatial filter $\boldsymbol{\beta}$ was initialized to a centered-spike with amplitude equal to 1 at the central element of the kernel while the rest of the elements were all set to 0. The initialization of $\boldsymbol{\beta}$ as a centered-spike effectively results in estimation of localized likelihoods $\mathcal{P}(\boldsymbol{d}_i|\kappa_i)$ as a starting point before the parameters $\Theta$ (and hence $\boldsymbol{\beta}$) are updated during the M-step of the EM algorithm. Since the localized likelihoods are estimated only from the seismic attributes observed at the location of estimation, they are susceptible to noise in the data (figures 5.6 and 5.7) and therefore do not abide by the geological plausibility rules for various facies configurations (such as gravitational ordering of fluids) and the conditional spatial distributions of facies depicted in the training image.

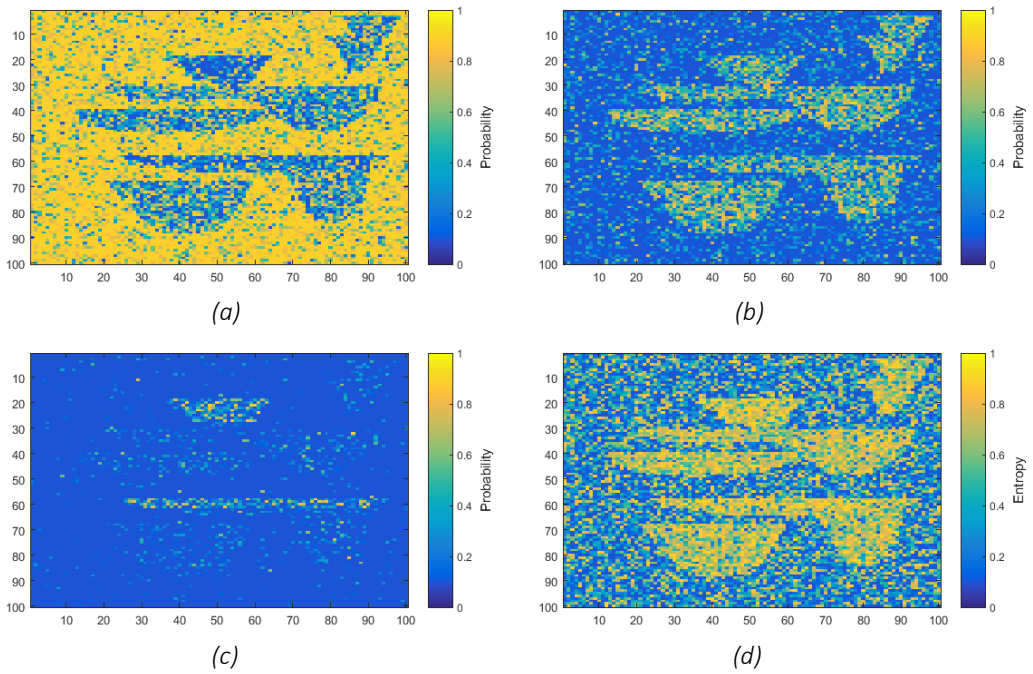



*Figure 5.6: Model space plots of initial cell-wise marginal likelihoods of (a) shale, (b) brine-sand and (c) gas-sand computed from the initial estimates of parameters, and (d) entropy as a measure of uncertainty in the model under the initial likelihoods.*

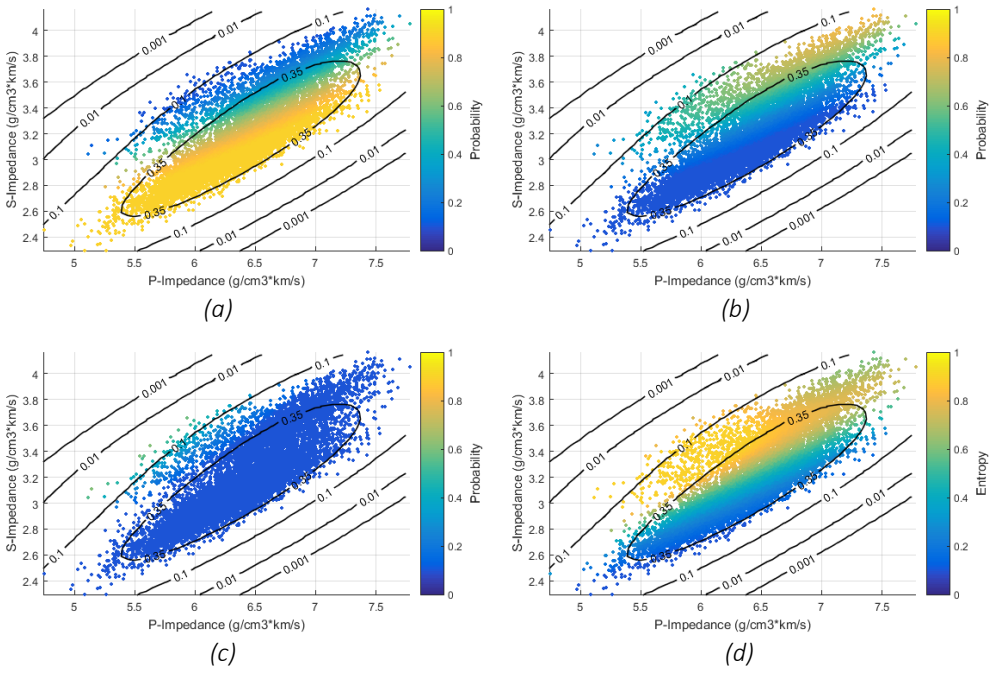



*Figure 5.7: Attribute space plots of initial cell-wise marginal likelihoods of (a) shale, (b) brine-sand and (c) gas-sand computed from the initial estimates of parameters, and (d) entropy as a measure of uncertainty of the model under the initial likelihoods. Equidistant contours represent the initial Gaussian mixture distribution for the three facies.*

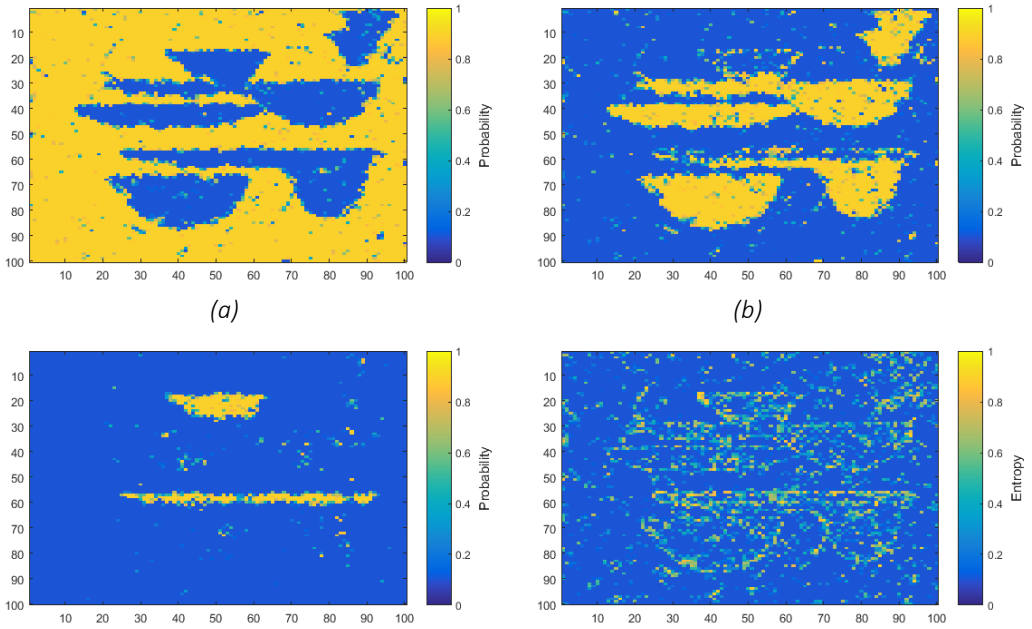



*Figure 5.8: Model space plots of updated cell-wise quasi-localized marginal likelihoods of (a) shale, (b) brine-sand and (c) gas-sand computed from the updated model parameters after running the EM algorithm, and (d) normalized entropy as a measure of model uncertainty under the updated likelihoods.*

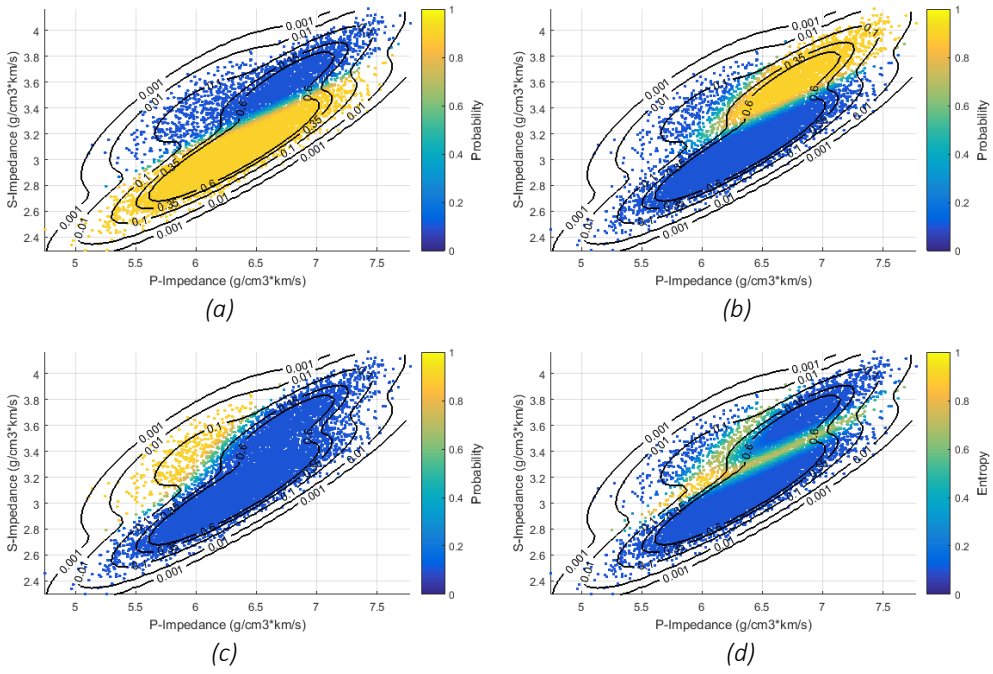



*Figure 5.9: Attribute space plots of updated cell-wise quasi-localized marginal likelihoods of (a) shale, (b) brine-sand and (c) gas-sand computed from the updated model parameters after running the EM algorithm, and (d) entropy as a measure of the model uncertainty under the updated likelihoods. Equidistant contours represent the Gaussian mixture distribution for the three components (facies).*

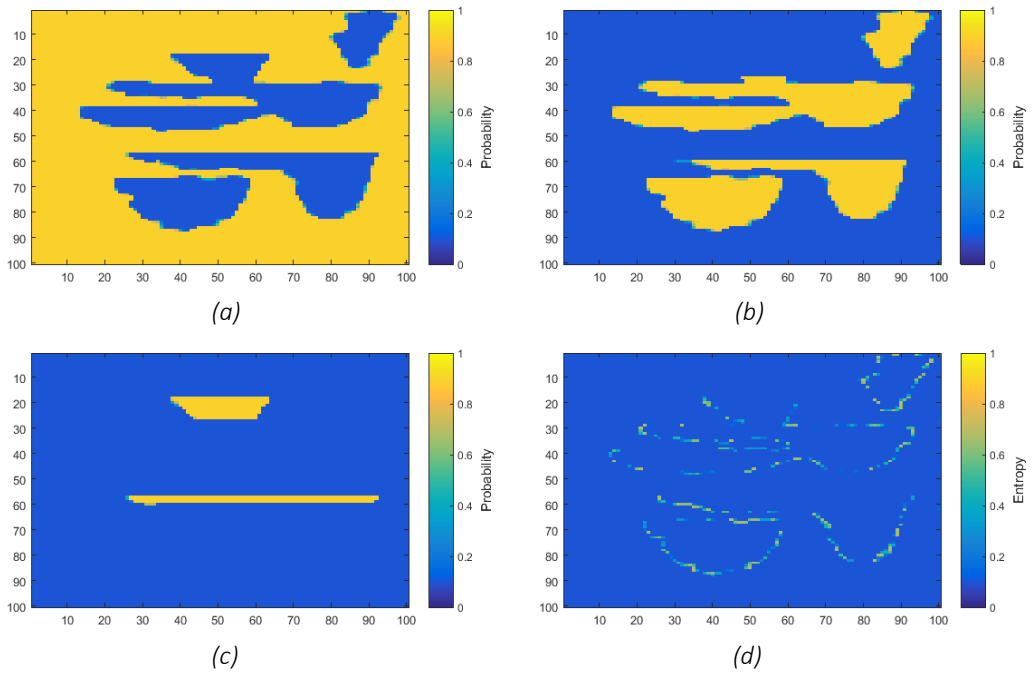



*Figure 5.10: Model space plots of cell-wise marginal posterior distributions of (a) shale, (b) brine-sand and (c) gas-sand, and (d) entropy as a measure of model uncertainty under the marginal posterior distributions.*

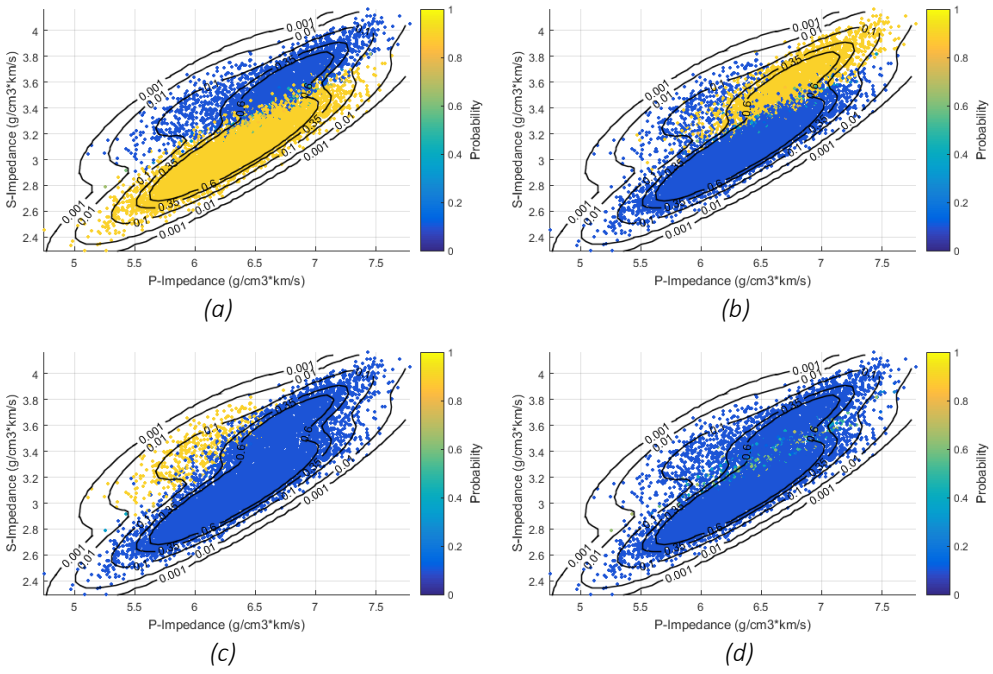



*Figure 5.11: Attribute space plots of cell-wise marginal posterior distributions of (a) shale, (b) brine-sand and (c) gas-sand, and (d) normalized entropy as a measure of model uncertainty under the marginal posterior distributions. Equidistant contours represent probability distributions of individual components of Gaussian mixture.*

The expectation-maximization (EM) algorithm was then used to estimate the marginal posterior distributions and the parameters $\Theta$. Contrary to the general practice of initializing the LBP algorithm with random or constant beliefs, we initialized it with the localized-likelihoods estimated using a MDN. Such initialization of vertex beliefs with the estimated localized likelihoods allowed faster convergence. The parameters $\Theta$ were then updated in the attribute space during the M-step using the current estimate of posterior marginal distributions $\hat{\mathcal{P}}_j(\kappa_j)$ obtained from the E-step, as follows. The filter coefficients $\boldsymbol{\beta}$ were estimated using equation 5.32 with the expected facies responses $\boldsymbol{r}_j$ at each location $j$ computed using equation 5.5. The parameters $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k, k \in \mathcal{G}$ were updated using equations 5.34 and 5.35. The parameters updated during the M-step were then used in the E-step of the subsequent iteration until convergence. On convergence, the EM algorithm resulted in estimates of QLL that show a higher quality of facies discrimination (figures 5.8 and 5.9), and the estimates of marginal posterior distributions $\mathcal{P}(\boldsymbol{\kappa}_i|\boldsymbol{d}, \hat{\Theta})$ for facies $\boldsymbol{\kappa}$ in each model cell $i$ (figures 5.10 and 5.11) given the observed seismic attributes $\boldsymbol{d}$ and the final estimate $\hat{\Theta}$ of parameters $\Theta$, by incorporating both the prior information $\mathcal{P}(\boldsymbol{\kappa})$ elicited from the training image (figure 4.5a) and the final estimates of QLL $\mathcal{P}(\boldsymbol{d}|\boldsymbol{\kappa}, \hat{\Theta})$ (figure 5.8a-c).

Figure 5.12 shows the maximum-a-posteriori (MAP) estimate of the geological facies obtained from the max-product equation 5.23 based LBP using the parameters updated by the EM algorithm. The MAP estimate matches quite reasonably with the 'true' geology (figure 4.5b). Figures 5.6-5.11 (d) show entropy of distributions in each of the corresponding figures (a-c) as a measure of uncertainty under the respective distributions. It is evident from these figures that the entropy reduces significantly starting with the entropy of the localized likelihoods in figure 5.6d & 5.7d to the entropy in the marginal posterior distributions in figures 5.10d & 5.11d.

Although the prior information $\mathcal{P}(\boldsymbol{\kappa})$ was formulated from the training image as spatial distributions between just two neighbouring locations at a time (the so called 2-point statistics, or pairwise cliques), the approximate posterior distributions $\mathcal{P}(\boldsymbol{\kappa}|\boldsymbol{x}, \hat{\Theta})$ estimated by LBP algorithm are reasonably close to the desired target distributions $\mathcal{P}(\boldsymbol{\kappa}|\boldsymbol{x}, \Theta)$. This suggests that Bayesian inversion using QLL requires much less prior information about the conditional spatial distributions of facies to yield reliable estimates of posterior marginal distributions of facies. By contrast, the previous research (Walker & Curtis, 2014a; Nawaz & Curtis, 2017) based on localized likelihoods used prior information extracted using larger templates in the form of

joint distributions of facies over multiple points at a time from the same training image (for geological patterns of the same complexity). This is evident from figures 5.6 and 5.8 since the localized likelihoods used in the first iteration of the EM algorithm are much noisier than the QLL estimated using parameters updated in the M-step. The current algorithm can, however, be modified to incorporate the prior information from cliques of size greater than two. Such a modification is expected to allow the reconstruction of richer features observed in more complex geologies. This is achieved in chapter 6.

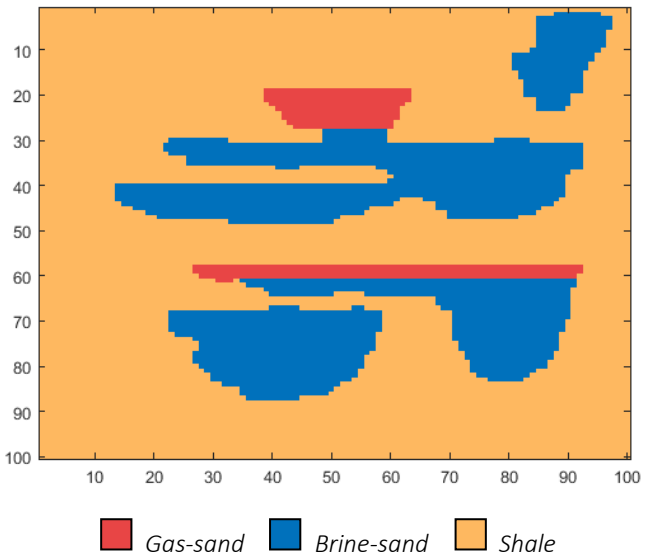



*Figure 5.12: Model space plot of the inverted MAP estimate of facies in each of the model cell using the variational Bayesian inversion (VBI) showing a reasonable reconstruction of the target model (figure 4.5b).*

It is also noteworthy that the marginal posterior distributions are updated in the model space during the E-step of the EM algorithm such that the spatial conditional distributions of various facies comply with those encapsulated in the training image. As a consequence of this, the model parameters are updated in the attribute space to reflect the inter-mixing of attributes (and overlap of their distributions) that are generated by different facies (Gaussian components) – see figure 5.13.

Coefficients of the estimated spatial filter $\widehat{\boldsymbol{\beta}}$ were estimated from the M-step of the last iteration of the EM algorithm under the constraint that the resulting matrix is laterally symmetric (symmetric across columns). The estimated coefficients are shown below in the matrix form

$$\widehat{\boldsymbol{\beta}} = \mathrm{vec}\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.25 & 0 & 0 \\ 0.018 & 0.118 & 0.162 & 0.118 & 0.018 \\ 0.153 & 0.016 & 0.049 & 0.016 & 0.153 \end{bmatrix} \tag{5.40}$$

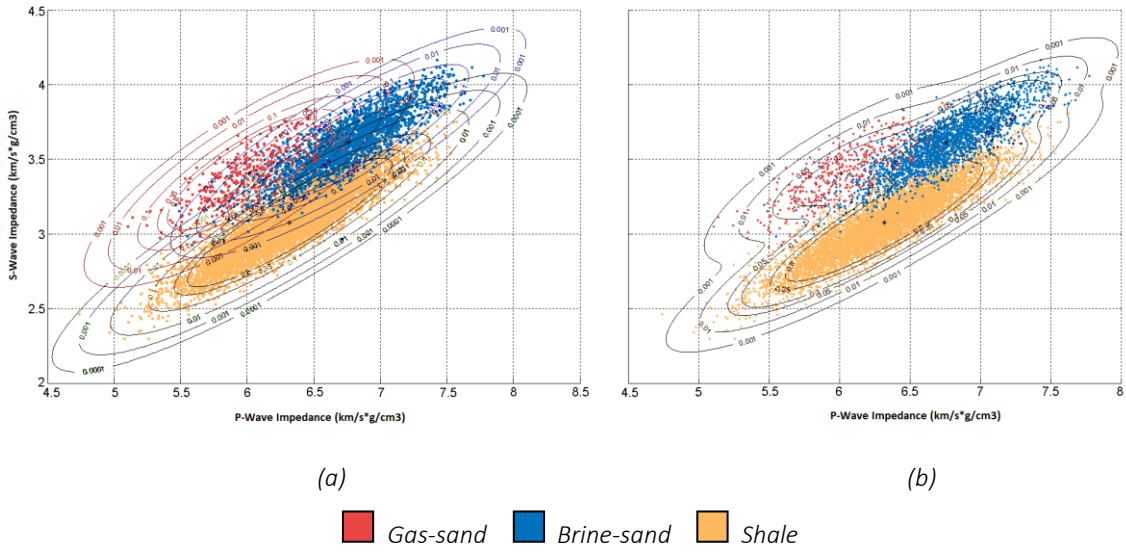



*Figure 5.13: Attribute space plots of (a) components of the Gaussian mixture model, and the seismic attributes colour-coded with the facies of maximum marginal posterior distributions: shale (yellow colour), brine-sand (blue colour) and gas-sand (red colour). (b) The Gaussian mixture distribution obtained from the sum of Gaussian components per facies as displayed in (a).*

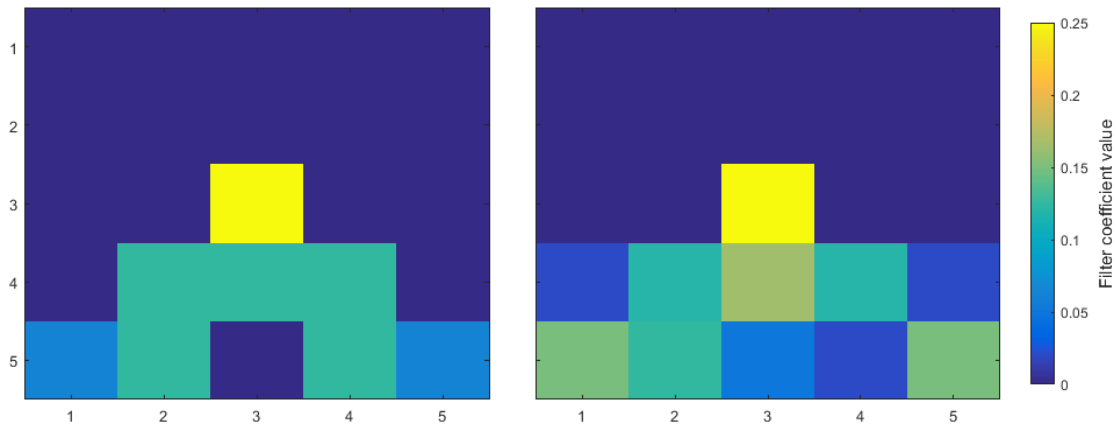



*Figure 5.14: Comparison of (a) the spatial filter $\boldsymbol{\beta}$ used to blur the synthetic attributes and (b) the recovered spatial filter $\widehat{\boldsymbol{\beta}}$. The amplitudes are scaled to a maximum value of 0.25 in both the plots.*

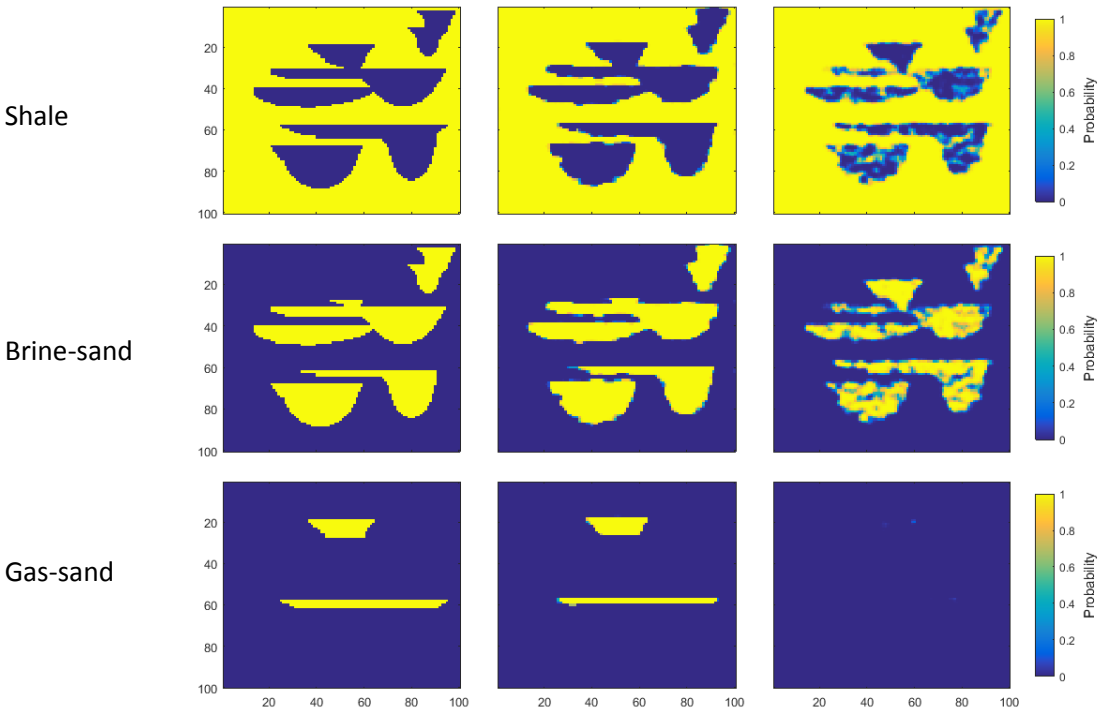
Figure 5.14 shows a comparison of the spatial filter $\boldsymbol{\beta}$ that was used to blur the seismic attributes and the estimated spatial filter $\widehat{\boldsymbol{\beta}}$, both scaled to a maximum amplitude value of 0.25, showing that while not perfect, a reasonable estimate of the spatial blurring can be obtained.

### *5.6.1 Comparison with Localized Likelihoods Based Inversion*

The previously published facies inversion methods that use localized likelihoods (LL, e.g., Larsen *et al.* 2006; Ulvmoen & Omre, 2010; Ulvmoen *et al.* 2010; Walker & Curtis, 2014a; and Nawaz & Curtis, 2017) assume that any spatial correlations present in the data (seismic attributes) are a direct consequence of, and therefore can be completely described by, the spatial distribution of facies as encoded in the prior information. In effect, these methods may not account for any spatial correlations present in the data due to other effects unrelated to the geology, such as those due to spatial blurring caused by processing related artefacts and limited resolution of seismic data. Also, such methods do not make effective use of any spatial correlations in the data that are related to the local geology. We may hypothesize that these methods have been successful to-date mainly because they rely too much on the prior information to reconstruct the spatial distribution of facies. This hypothesis suggests that in the case that the prior information is limited (e.g., using small neighbourhood templates to scan the training image) or is inconsistent with the true geology (e.g., if geological patterns in the training image are not rich enough or are different from those present in the true subsurface), the LL based inversion methods may not be successful. The quasi-localized likelihoods (QLL), on the other hand, complement the prior information by incorporating the spatial correlations present in the data within some neighbourhood of each location in the model. The current QLL based method is therefore expected to be more robust against insufficient or incorrect prior information.

In order to test this hypothesis, the synthetic test data from section 5.6 was used to compare the current 'QLL based method' with the 'LL based method' presented in chapter 4 (also see Nawaz & Curtis; 2017). The comparison is made in terms of the quality of inverted posterior marginal distributions when the data are spatially blurred (i.e., when the seismic attributes for various facies overlap significantly in the attribute space) and the amount of prior information is either limited or inconsistent with the true geology.

Figure 5.15 shows such a comparison with respect to the amount of prior information used. The prior information on the spatial distribution of facies is extracted from the training image using a 3x3 template and then supplied to the LL based method. This corresponds to a clique size of 9, i.e., the prior information is encoded as a joint distribution over neighbouring vertices in a square matrix with 3 rows and 3 columns. In comparison, since the current method uses only pairwise cliques, it uses only 2-point statistics based prior information. Even though the current method uses much less prior information, it reconstructs the marginal posterior distributions quite reasonably.




*(a) "True"* *(b) Using QLL* *(c) Using LL*

*Figure 5.15: Model space plots of the inverted cell-wise marginal distributions per facies – shale, brine-sand and gas-sand in the order from the top to the bottom row: column (a) true marginal distributions in the synthetic model as in figure 4.5b, column (b) that obtained using our current method which is based on the quasi-localized likelihoods, and column (c) that obtained using the method presented in chapter 4 (also see Nawaz & Curtis, 2017) which solves the problem using the localized likelihoods assumption.*

The results using the LL based method (right column in figure 5.15) show that this method could not discriminate between brine-sand and gas-sand and indeed failed to detect any gas-sand. Also the reconstruction of the spatial distribution of shale and brine-sand is not

as good as in the current method (middle column in figure 5.15). In this case, we found that if we increased the size of the prior template to 5x7, the LL based method can reconstruct the posterior marginal distributions just as reasonably as with QLL. This explains the previous success of methods that assumed LL: they can work well with quasi-localized data, but only if the prior information supplied is sufficiently strong to overcome the unrealistic LL assumption.

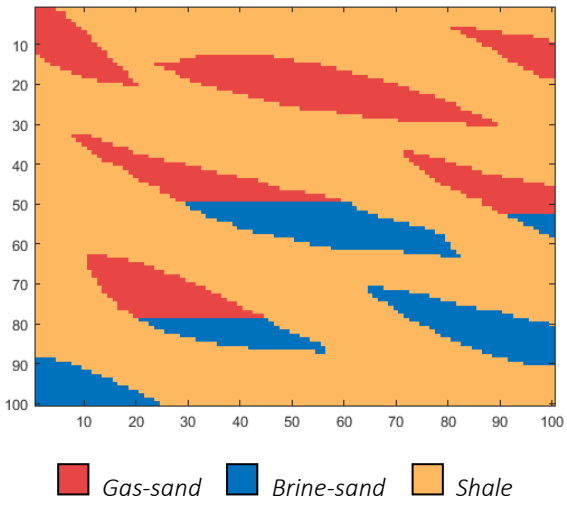
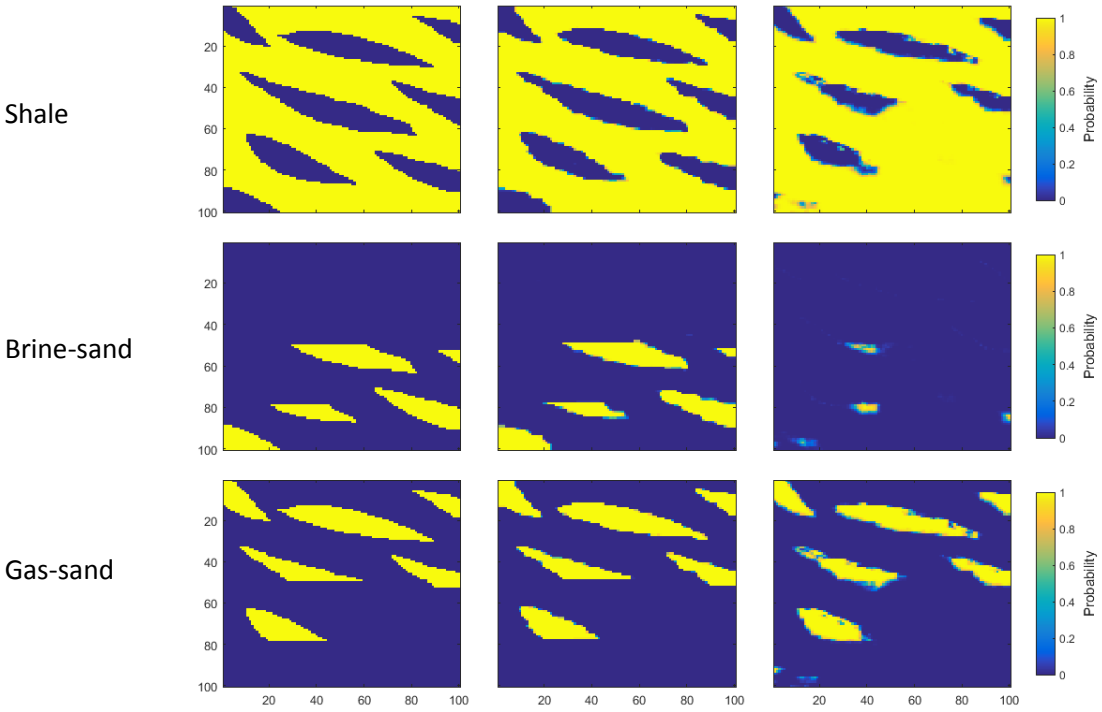



*Figure 5.16: The target image representing the 'true' geological model consisting of dipping sand lenses (with no over-bank deposits), in a hypothetical scenario where the stratum is tilted after lithification. This is the target for spatial facies inversion in the case that the prior information presented in the form of training image in figure 4.5a is inconsistent with this 'true' geological image.*

Next synthetic seismic attributes were generated as described in section 5.6 except that the 'true' geology now contains dipping sand lenses (with no over-bank deposits), in a hypothetical scenario where the stratum is tilted after lithification (figure 5.16). The same training image is used as in figure 4.5a with sand channels and over-bank deposits with a background shale in an assumed horizontal stratum (i.e., without tilting). This allowed making a comparison between the two methods when the prior information supplied in the form of a training image is inconsistent with the true geology (figure 5.17). In this case, the prior information on the spatial distribution of facies is supplied to the LL based method by using a 5x3 template. This corresponds to a clique size of 15. The prior information for QLL still uses only 2-point statistics. In this case the LL based method fails to discriminate between Shale and Brine-sand, though the reconstruction of posterior marginal distributions of Gas-sand is somewhat reasonable (right column in figure 5.17). The current QLL based method, however,

reconstructed the posterior marginal distributions of all of the three facies reasonably well (middle column in figure 5.17) and therefore proves to be significantly more robust against incorrect prior information than LL based methods.




*(a) "True"* *(b) Using QLL* *(c) Using LL*

*Figure 5.17: Model space plots of the inverted cell-wise marginal distributions per facies – shale, brine-sand and gas-sand in the order from the top to the bottom row: (left column) true marginal distributions in the synthetic model as in figure 5.16, (middle column) that obtained using our method as presented in this chapter which is based on the quasi-localized likelihoods, and (right column) that obtained using the method presented in chapter 4 (also see Nawaz & Curtis, 2017) which is based on the localized likelihoods assumption.*

The above comparisons show that the inversion methods based on the LL assumption require well informed priors: that is, the priors must be sufficiently informative to overcome errors due to the incorrect localized assumption, and must be consistent with the true geology. This means that the geological patterns depicted in a training image must be rich – diverse enough to include any possible facies patterns expected to be present in the subsurface. The current QLL based method, on the other hand, is expected to perform better even in the case that we do not have strong prior information, or that our prior information is only partially consistent with the true geology.

## 5.7 Discussion

The localized likelihoods (LL) assumption was used in the previous research by Walker & Curtis (2014a) and Nawaz & Curtis (2017) in order to address the computational intractability of mathematical inference in models with non-localized likelihoods. The current method evades such computational intractability by retaining the conditional independence (CI) assumption on data given the model parameters (facies), and resorting to an iterative optimization based approximation (the EM algorithm) rather than an analytical approach as in chapter 4 for estimation of marginal posterior distributions of facies. This chapter introduces the concept of quasi-localized likelihoods (QLL) as a step towards methods that incorporate fully non-localized likelihoods, which is clearly a topic for future research.

The quality of facies discrimination with QLL (figure 5.8) has shown to be higher than with LL (figure 5.6) since geophysical data contain spatial correlations and are not independent. The prior information further improves the discrimination and the spatial distribution of facies when combined with the QLL, for example ensuring that geologically implausible configurations (e.g., Brine-sand directly overlaying Gas-sand in some areas of figure 5.6) are disregarded in the computation of marginal posterior distributions (figure 5.10). Although this has not been tested explicitly, it is to be expected that the older methods that use LL from other authors cited herein will have similar short-comings to those of the previous work presented in chapter 4 (see also, Nawaz & Curtis; 2017).

A major challenge with any inference or parameter estimation based on the loopy-belief propagation (LBP) algorithm is that there is no theoretical guarantee about its convergence. This contrasts with McMC based methods which are theoretically guaranteed to converge asymptotically. The empirical evidence, on the other hand, is very strong that LBP does converge in most cases. Koller & Friedman (2009) discussed many different possible reasons of non-convergence of LBP and their suggested remedies. In the situations when LBP fails to converge, it is observed that the non-convergence is either local or is due to oscillations in the beliefs. Koller & Friedman (2009) suggested using a dampening of the difference between two subsequent updates of beliefs as a remedy for oscillatory beliefs. If non-convergence is local, most of the beliefs converge except just a few. Averaged beliefs over a number of iterations may be used in case of local non-convergence.

In case of non-convergence of LBP algorithm, a careful examination of the problem is recommend before applying any such remedy. These problems may be caused by conflicts between the vertex and edge potentials which are likely to be caused by the presence of noise in the data, or by problematic parameters learnt during the M-step. In such cases, the input attributes must be properly chosen or conditioned for the problem. Interested readers are recommended to consult Mooij & Kappen (2007) for a detailed account on the sufficient conditions for convergence of the LBP algorithm. Nevertheless, in contrast to McMC based methods where it is impossible to detect convergence objectively, non-convergence in LBP is always detectable (see *Algorithm 5.1*).

The variational form of the EM algorithm as used in this research is expected to offer a significant step towards generalization of the variational Bayesian inversion (VBI) for solving problems which specifically involve a spatial grid of observed data that are collocated with the unknown model parameters. The current method can be extended further to invert for continuous variables (such as rock properties) in spatial inverse problems. This is demonstrated in chapter 8, where this method is extended for joint estimation of geological facies and (continuous) petrophysical rock properties from (elastic) seismic attributes.

Since the current method uses a pairwise MRF as the spatial model for the distribution of facies, it is anticipated that it may not be so capable of reconstructing complex spatial patterns of geological facies (e.g., those found in aerial view of intersecting sand channels in a deltaic environment). Multi-point statistics based simulation (Strebelle 2001; Caers & Zhang, 2004; Arpat, 2005; Journel & Zhang, 2006; Mariethoz & Caers, 2014) and related stochastic inversion methods have been developed for such cases (Haas & Dubrule, 1994; Francis, 2005; Nunes *et al*. 2016). In chapter 6, a general MRF with higher-order cliques is proposed which may be able to reconstruct complex spatial patterns. However, the current method has not been tested for such a case.

## 5.8 Conclusions

A Bayesian method is presented for inversion of geological facies from geophysical data (such as seismic attributes) under the variational approximation as a computationally efficient alternative to the commonly used Markov-chain Monte Carlo (McMC) based methods. In addition, the current method also allows for reliable detection of convergence, in contrast to

the McMC based spatial inversion methods which are known to have difficulties with detection of convergence. The likelihoods are assumed to have a Gaussian distribution with expectations at a location given by a linear combination of local facies responses within the neighbourhood of that location. Such likelihoods are termed quasi-localized likelihoods (QLL) which refer to a relaxation of the assumption of localized likelihoods as was generally used in previous research. The data are assumed to be conditionally independent (CI) given the geological facies and are assumed to be distributed as a Gaussian mixture distribution with number of components given by the number of facies considered. It is also shown that the QLL define a *spatial Gaussian mixture model* (SGMM) for observed data at a location given the model parameters (facies) at the neighbouring locations. The prior spatial distribution of facies is modelled as a Markov random field (MRF), and it was shown that by virtue of the CI assumption on data, the joint and hence the posterior distribution of facies given the observed data also represent a MRF (specifically a hidden MRF).

The current method is compared with the previous LL based methods of facies inversion using a synthetic data example. It shows that the current method requires far less prior information to reconstruct an accurate estimate of the true marginal posterior distributions of facies in the subsurface as compared to a previous LL based inversion method. Also the current method proves to be more robust against prior information that is not consistent with the true geology.

# Chapter 6 Discriminative Variational Bayesian Inversion

## 6.1 Summary

A novel, fully probabilistic and non-linear inversion method is introduced in this chapter to estimate the spatial distribution of geological properties (depositional facies, diagenetic rock types, or other rock properties) from geophysical data (e.g. seismic data). Both localized likelihoods (LL) and conditional independence (CI) assumptions on data are removed in this chapter. Contrary to the conventional generative approach that models solution probabilities via the likelihood of observed data, the current method uses a discriminative approach that directly models the posterior distribution of the geological properties given the data. This reduces the modelling effort significantly, and allows machine learning algorithms such as neural networks to be deployed to solve large scale geophysical inverse problems. The proposed method honours spatial distributions of geological properties supplied as multi-point geostatistical prior information about local geology. This requires spatial probabilistic inference for which a novel and efficient approximate inference method, 'higher-order mean field approximation', is proposed in this chapter within the variational Bayesian framework (see section 6.4.1). The proposed method thus avoids extensive sampling during inference, yet provides fully probabilistic Bayesian results, and is therefore scalable to higher dimensional problems. With the help of a synthetic example it is shown that this method can be trained using supervised learning to be robust against correlated noise (undesired features convolved with the desired signal) present in the data as long as we can provide statistical characteristics of the noise. This method is also applied to a real 3D seismic data from New Zealand to estimate probability of presence of geological faults at any point in the subsurface.

## 6.2 Introduction

Bayesian inversion is usually performed by defining a joint probability distribution over all of the observed as well as the unobserved (or hidden) variables. Modelling the joint distribution over all of the variables is commonly referred to as *generative modelling*, since

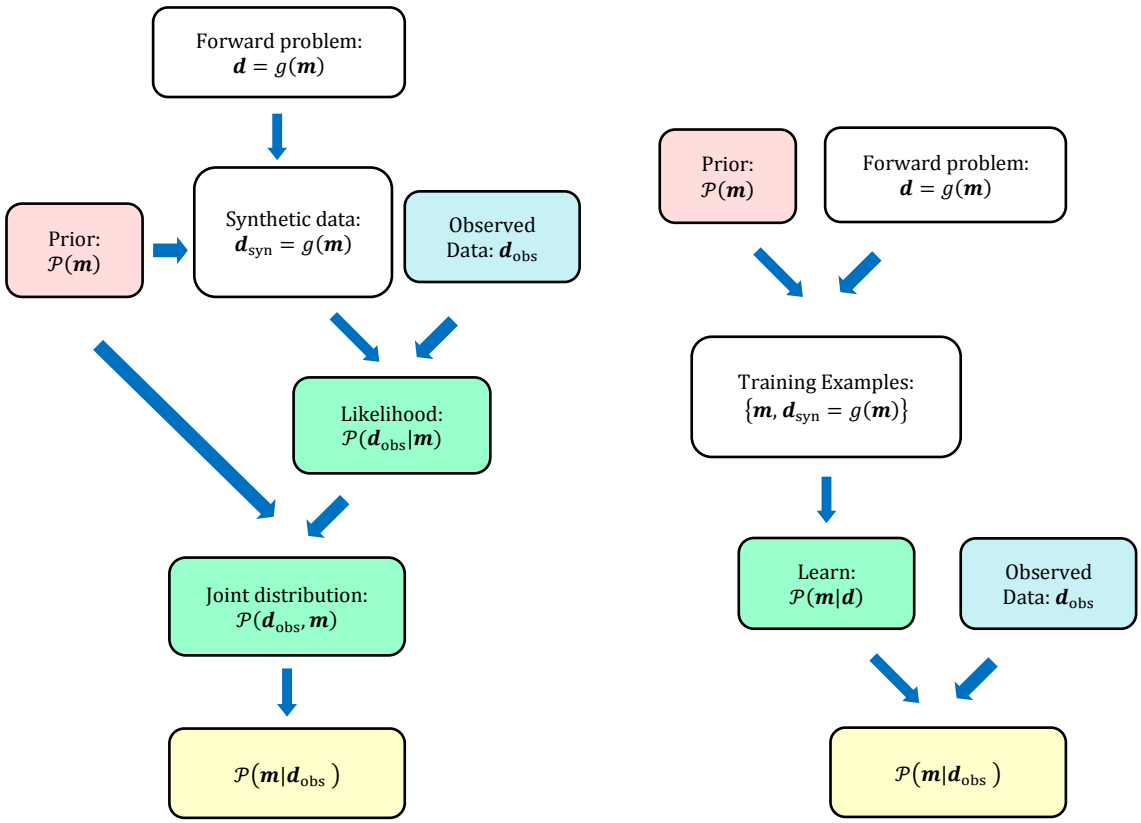
given the joint distribution over all of the variables we can use it to *generate* new synthetic data corresponding to known model parameter values. This is the standard method in most previous geophysical applications of probabilistic inverse theory

The computation of joint posterior probability distributions over a large number of parameters using Bayesian inversion is computationally intractable. To limit the analytical and computational complexity of modelling the joint distribution over all of the variables, previous research in geostatistical inversion (e.g. Larsen *et al.* 2006; Caers *et al*. 2006; Hoffman & Caers, 2007; Ulvmoen & Omre, 2010; Shahraeeni & Curtis, 2011; Shahraeeni *et al.* 2012; Walker & Curtis, 2014a; Nawaz & Curtis, 2017; and Grana, 2018) relied on the assumptions of localized likelihoods (LL) and conditional independence (CI) of data (see section 1.1.4). Another typical assumption in spatial inversion using soft conditioning data (such as seismic) is that seismic data are spatially smooth and therefore smooth spatial patterns of geological parameters may be inferred directly by using such data without the need to perform spatial inference (Caers & Ma, 2002; Shahraeeni & Curtis, 2011; Shahraeeni *et al.* 2012; Grana, 2018). However, this approach is more susceptible to noise present in the data. Examples of previous research in which the localized likelihoods assumption has been relaxed in 1D Bayesian inversion methods are: Lindberg & Omre (2014 & 2015), and Grana et al. (2017). The LL assumption was relaxed in chapter 5 (also see Nawaz & Curtis, 2018) by introducing multi-dimensional quasi-localized likelihoods which relate observed data at a location to the model parameters in any finite neighbourhood of that location.

In this chapter, both LL and CI assumptions are removed. Using *non-localized likelihoods* in solving a spatial inverse problem requires coupling of the model and data spaces such that all of the model parameters may be conditioned (depend) on any of the data, irrespective of the locations of observations. Conversely, the current method also allows data from anywhere in the model to be related to the model parameters at any location if there exists such a logical or conceptual association.

Exact computation of fully non-localized likelihoods is intractable in most models of practical interest. To address this problem while also avoiding McMC, a Bayesian inversion method is proposed that directly models the posterior distribution without requiring that the joint distribution over all of the variables is specified. This approach is called *discriminative modelling*. Although classification of data using a discriminative model is a common choice in the machine learning community, Bayesian inversion using a generative model is the standard

approach in large scale geophysical problems. This chapter therefore introduces a new approach of Bayesian inversion in geophysical problems based on a discriminative model in which prior and likelihoods are implicitly incorporated such that the posterior distribution is modelled without explicit mathematical modelling of the joint distribution over the observed and hidden variables (figure 6.1). Bayesian inversion using a discriminative model is referred to as *discriminative Bayesian inversion* (DBI) or simply *discriminative inversion*.




*(a)* *(b)*

*Figure 6.1. Flow chart comparison of (a) the conventional method of geophysical probabilistic inversion using a generative model, and (b) the discriminative probabilistic inversion method introduced here. Colours match related steps between (a) and (b).*

The proposed method uses supervised learning using training examples of expected spatial distributions of model parameters and the corresponding data. So this method may require some type of Monte Carlo sampling to generate an example database for training purposes, but typically this is a far lower dimensional and more computationally tractable

sampling process than that is required to use general McMC methods to solve the entire inference problem.

The structure of rest of this chapter is as follows. Bayesian inversion is reviewed in section 6.3, where the conventional approach to Bayesian inversion based on a generative model is first discussed in more detail in the subsection 6.3.1 to explain why it is difficult to remove assumptions of localized likelihoods and conditional independence of data. The discriminative inversion approach is then introduced in the subsection 6.3.2 as a tractable alternative to the generative approach in large and complex models. A model for the posterior distribution is proposed in subsection 6.3.3. A mathematical formulation of the variational Bayes method is introduced in section 6.4 to perform inference, and an approximate inference method is derived in section 6.4.1, and an associated method for parameter estimation is presented in section 6.4.2. The computational complexity of this method is discussed in section 6.5. After providing mathematical details of the method, a synthetic test example is provided in section 6.6 where this method is applied to invert multiple seismic attributes for geological facies (shale, brine-sand and gas-sand) in the presence of strongly correlated noise. A real data example from New Zealand is provided in section 6.7 to demonstrate application of this method in probabilistic interpretation of faults in 3D seismic data. Finally the implications of the method are discussed in section 6.8 and conclusions of this research are provided in section 6.9.

## 6.3 Bayesian Inversion

The probabilistic inverse problem that we solve is to infer the unknown geological model parameters $\boldsymbol{m}$ from the observed geophysical data or attributes $\boldsymbol{d}$. The Bayesian solution of the inverse problem is given by the posterior distribution $\mathcal{P}(\boldsymbol{m}|\boldsymbol{d})$ of $\boldsymbol{m}$ given $\boldsymbol{d}$ combines the likelihood $\mathcal{P}(\boldsymbol{d}|\boldsymbol{m})$ of observing $\boldsymbol{d}$ given that $\boldsymbol{m}$ is the true model, and the prior distribution $\mathcal{P}(\boldsymbol{m})$ of model parameters, using *Bayes' theorem* as

$$\mathcal{P}(\boldsymbol{m}|\boldsymbol{d}) = \frac{\mathcal{P}(\boldsymbol{m},\boldsymbol{d})}{\mathcal{P}(\boldsymbol{d})} = \frac{\mathcal{P}(\boldsymbol{d}|\boldsymbol{m})\mathcal{P}(\boldsymbol{m})}{\mathcal{P}(\boldsymbol{d})} \tag{6.1}$$

The denominator $\mathcal{P}(\boldsymbol{d})$ represents the *marginal likelihood* of the observed data $\boldsymbol{d}$ (also called *evidence*). It acts as normalization constant, and is given by

$$\mathcal{P}(\boldsymbol{d}) = \int_{\boldsymbol{m}} \mathcal{P}(\boldsymbol{d}, \boldsymbol{m}) d\boldsymbol{m} = \int_{\boldsymbol{m}} \mathcal{P}(\boldsymbol{d}|\boldsymbol{m}) \mathcal{P}(\boldsymbol{m}) d\boldsymbol{m} \quad 6.2$$

### 6.3.1 Bayesian Inversion using a Generative Model

We see from equations 6.1 and 6.2 that the prior $\mathcal{P}(\boldsymbol{m})$ and the likelihood $\mathcal{P}(\boldsymbol{d}|\boldsymbol{m})$ completely specify the posterior distribution through the joint distribution $\mathcal{P}(\boldsymbol{d}, \boldsymbol{m})$. A model that describes the probability of $\boldsymbol{m}$ given $\boldsymbol{d}$ in terms of their joint distribution is commonly referred to as a *generative model*. Thus, a generative model explicitly expresses the posterior distribution $\mathcal{P}(\boldsymbol{m}|\boldsymbol{d})$ in terms of the data likelihood $\mathcal{P}(\boldsymbol{d}|\boldsymbol{m})$ and the prior model distribution $\mathcal{P}(\boldsymbol{m})$ using equation 6.1.

Explicit specification of priors and likelihoods in Bayes' theorem has an intuitive meaning: the data $\boldsymbol{d}$ are assumed to have been generated by the unknown model $\boldsymbol{m}$ according to a pre-specified probability distribution $\mathcal{P}(\boldsymbol{d}|\boldsymbol{m})$, while the probability $\mathcal{P}(\boldsymbol{m})$ of $\boldsymbol{m}$ is known *a priori*. It is for this reason that explicit modelling of the posterior distribution $\mathcal{P}(\boldsymbol{m}|\boldsymbol{d})$ in terms of the joint distribution $\mathcal{P}(\boldsymbol{d}, \boldsymbol{m}) = \mathcal{P}(\boldsymbol{d}|\boldsymbol{m})\mathcal{P}(\boldsymbol{m})$ is known as *generative modelling*, since given the joint distribution over all of the hidden as well as observed variables we can artificially *generate* more data from it.

Estimation of a joint distribution over all of the variables offers a full description of a probabilistic system; it allows marginalization and conditioning over any subset of variables in order to perform inference, sampling, and prediction. For this reason, generative modelling seems to be an attractive approach. However, the joint distribution over observed and hidden variables is generally too complex to be modelled accurately. In addition, since the generative approach requires modelling a joint distribution over all of the variables that comprise a system, it may turn out to be an inefficient approach in situations where our objective is to solve a specific problem rather than to characterize the entire system. For instance, in geophysics our objective is usually only to compute the conditional distribution of $\boldsymbol{m}$ given $\boldsymbol{d}$; we can achieve this by manipulating the probabilistic relationships among various dependent variables mathematically, without modelling the full joint distribution over both $\boldsymbol{m}$ and $\boldsymbol{d}$. In a dense system where every variable depends on a large number of other variables, this task is practically as daunting as estimating the full joint distribution over all of the variables. However, many problems of practical interest regarding spatial phenomena involve sparse

systems (Besag, 1974). Examples include cases where parameter dependencies can be modelled as a Markov random field (MRF – see chapter 3) in which marginalization can be performed efficiently using dynamic programming (Denardo, 2003) or some approximate methods that do not require estimation of the full joint distribution (Koller & Friedman, 2009). In such a case, estimating the joint distribution over all of the variables may be regarded as a cumbersome and unnecessary intermediate step that requires immense modelling efforts and intense computational power.

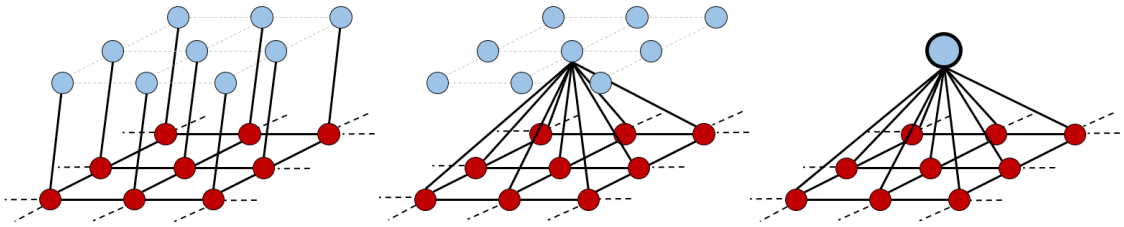
### 6.3.2 Bayesian Inversion using a Discriminative Model

The above concerns regarding generative modelling lead us to explore the alternative *discriminative modelling* approach. This directly estimates the posterior distribution $\mathcal{P}(\boldsymbol{m}|\boldsymbol{d})$ of $\boldsymbol{m}$ given $\boldsymbol{d}$ as a non-linear mathematical function, without modelling their joint distribution $\mathcal{P}(\boldsymbol{m},\boldsymbol{d})$ as an intermediate step. In this manner, discriminative modelling alleviates some of the effort required to model any complex dependencies among variables through their full joint distribution, and proves to be parsimonious in the use of computational resources. The notion of discriminative modelling emerged in the field of machine learning with the introduction of the discriminative *logistic regression* method for classification as an alternative to the generative *Naïve Bayes* classification method. A detailed description of these methods is beyond the scope of this thesis and interested readers are referred to the relevant machine learning literature (e.g. Ng & Jordan, 2002; Jebara, 2002; Kumar & Hebert, 2003).

Since discriminative modelling does not require estimation of the joint distribution over hidden as well as the observed variables, we can deploy the modelling effort and computational resources to incorporate additional sophistication in the model without tremendously increasing the computational cost of the overall method. Based on this notion, a *discriminative Bayesian inversion* (DBI) method is proposed here that uses non-localized likelihoods and accounts for correlations observed in the data, without making any conditional independence assumptions about the observed variables. Inversion methods that are based on a generative model are computationally too demanding to allow for such sophistication in the model. Below we present a discriminative analogue of a MRF which we use as a model for the posterior distribution $\mathcal{P}(\boldsymbol{m}|\boldsymbol{d})$ that implicitly incorporates spatial priors and non-localized likelihoods.

### 6.3.3 Posterior Model

We model the posterior distribution $\mathcal{P}(\boldsymbol{m}|\boldsymbol{d})$ of model parameters $\boldsymbol{m}$ given observed data $\boldsymbol{d}$ as a *conditional random field* (CRF: Lafferty *et al.* 2001), also called a *discriminative random field* (Kumar & Hebert, 2003), which is essentially a hidden Markov random field (HMRF) defined over $\boldsymbol{m}$, and conditioned by the data $\boldsymbol{d}$. A schematic comparison of a HMRF and a CRF is shown in figure 6.2 in the form of a graphical model.



(a) (b) (c)

*Figure 6.2: A schematic representation of (a & b) hidden Markov random fields (HMRF), and (c) a conditional random field (CRF). A dark-red circle represents hidden variables $\boldsymbol{m}_i$ at a location $i$ in the model, light-blue circles represent observed variables $\boldsymbol{d}_i$, larger light-blue circle with a thick border in (c) represents all of the observed data $\boldsymbol{d}$, and solid black lines connecting circles represent direct probabilistic dependence between the connected variables. Dotted lines only represent the location grid and not the probabilistic dependence. The HMRF in (a) assumes localized likelihoods, and in (b) assumes quasi-localized likelihoods (chapter 5, also see Nawaz & Curtis, 2018). Both HMRFs assume conditional independence of data $\boldsymbol{d}$ given model parameters $\boldsymbol{m}$. The CRF in (c) makes no such assumptions. This figure shows only pairwise cliques represented by pairs of connected hidden variables (red circles). In general, cliques may represent higher order dependence among more than two variables, and hence may extend beyond the 3x3 grid of pairwise connected variables shown here in red.*

According to the *Hammersley-Clifford theorem* (Besag, 1974), the posterior distribution $\mathcal{P}(\boldsymbol{m}|\boldsymbol{d})$ may therefore be written as a *Gibbs distribution* in terms of a product of potential functions $\psi(\boldsymbol{m}_{\hat{c}}, \boldsymbol{d})$ defined over the domain of model parameters $\boldsymbol{m}_{\hat{c}}$ within a maximal clique $\hat{c}$ in the model and data $\boldsymbol{d}$. The logarithm of potential functions are typically expressed as a linear combination of some, generally non-linear, pre-specified vector of feature functions $\boldsymbol{f}(\boldsymbol{m}_{\hat{c}}, \boldsymbol{d})$ of $\boldsymbol{m}_{\hat{c}}$ and $\boldsymbol{d}$ with relative weights $\boldsymbol{w}$ such that the posterior distribution $\mathcal{P}(\boldsymbol{m}|\boldsymbol{d}; \boldsymbol{w})$ of $\boldsymbol{m}$ given $\boldsymbol{d}$ parameterized by $\boldsymbol{w}$ may be written as

$$\mathcal{P}(\boldsymbol{m}|\boldsymbol{d};\boldsymbol{w}) = \frac{1}{\mathcal{Z}(\boldsymbol{d};\boldsymbol{w})}\prod_{\hat{c}\in\hat{\mathcal{C}}}\psi(\boldsymbol{m}_{\hat{c}},\boldsymbol{d}) = \frac{1}{\mathcal{Z}(\boldsymbol{d};\boldsymbol{w})}\exp\left(\sum_{\hat{c}\in\hat{\mathcal{C}}}\boldsymbol{w}^T\boldsymbol{f}(\boldsymbol{m}_{\hat{c}},\boldsymbol{d})\right) \quad 6.3$$

where the denominator $\mathcal{Z}(\boldsymbol{d};\boldsymbol{w})$ is the normalization constant given by

$$\mathcal{Z}(\boldsymbol{d};\boldsymbol{w}) = \int_{\boldsymbol{m}}\exp\left(\sum_{\hat{c}\in\hat{\mathcal{C}}}\boldsymbol{w}^T\boldsymbol{f}(\boldsymbol{m}_{\hat{c}},\boldsymbol{d})\right)d\boldsymbol{m} \quad 6.4$$

which is a function of the observed data $\boldsymbol{d}$ parametrized by $\boldsymbol{w}$, and is referred to as the *evidence*.

The feature functions $\boldsymbol{f}(\boldsymbol{m}_{\hat{c}},\boldsymbol{d})$ are assumed to encode sufficient statistics of the desired distribution $\mathcal{P}(\boldsymbol{m}|\boldsymbol{d};\boldsymbol{w})$, and their eloquent specification is therefore crucial for accurate modelling of the true posterior distribution. For example, feature functions may be defined as a measure of how likely are some features in the data given the spatial distributions of geological properties within a maximal clique. The feature functions thus implicitly model the spatial priors over $\boldsymbol{m}_{\hat{c}}$ and the non-localized likelihoods that define the probabilistic relationship between $\boldsymbol{m}_{\hat{c}}$ and $\boldsymbol{d}$. Since conditioning can be performed over the entire set of observed variables $\boldsymbol{d}$, no localization of likelihoods is required in this model. In a discriminative framework, this is what allows the direct modelling of the posterior distribution, which may otherwise be intractable if no conditional independence is assumed over the observed variables (for example in a HMRF).

Once the feature functions have been defined, the next step is to devise efficient methods to estimate parameters $\boldsymbol{w}$ in equations 6.3 and 6.4, and for spatial inference. Spatial inference involves estimating the normalization constant $\mathcal{Z}(\boldsymbol{d};\boldsymbol{w})$, the marginal posterior distributions over cliques and individual variables in the model, and any posterior statistics of interest such as the most likely overall model $\boldsymbol{m}^*$ of $\boldsymbol{m}$ given $\boldsymbol{d}$, such that $\boldsymbol{m}^* = \mathrm{argmax}_{\boldsymbol{m}}\{\mathcal{P}(\boldsymbol{m}|\boldsymbol{d};\boldsymbol{w})\}$. Parameter estimation can be performed in a supervised manner by using training examples of model $\boldsymbol{m}_{\hat{c}}$ and the corresponding data $\boldsymbol{d}$ in order to obtain an estimate of the parameters $\boldsymbol{w}$ that best describe the distribution of $\boldsymbol{m}$ given $\boldsymbol{d}$ under the posterior model in equation 6.3. Within section 6.4, we discuss inference and parameter estimation methods in a CRF model in subsections 6.4.1 and 6.4.2, respectively.

## 6.4 Variational Bayesian Inference

We use the variational Bayes (VB) method (see section 2.4) for spatial inference which is an efficient and prominent method for approximate Bayesian inference in decomposable models such as the CRF used in this research. For a given data $\boldsymbol{d}$ we want to maximize $\mathcal{Z}(\boldsymbol{d};\boldsymbol{w})$ as a function of $\boldsymbol{w}$ which is intractable. VB defines a lower bound on the log-evidence $\mathcal{L}(\boldsymbol{w};\boldsymbol{d}) \equiv \log \mathcal{Z}(\boldsymbol{d};\boldsymbol{w})$ which is maximized with respect to $\boldsymbol{w}$ as a surrogate for maximization of the generally intractable log-evidence. In effect, we use the VB method to approximate a generally intractable joint posterior distribution $\mathcal{P}(\boldsymbol{m}|\boldsymbol{d};\boldsymbol{w})$ with a variational distribution $\mathcal{Q}(\boldsymbol{m}|\boldsymbol{d}) \in \mathbb{Q}$, where $\mathbb{Q}$ is a family of tractable distributions. The variational distribution $\mathcal{Q}$ is chosen from $\mathbb{Q}$ such that it minimizes the *KL-divergence* $KL\big(\mathcal{Q}(\boldsymbol{m}|\boldsymbol{d})||\mathcal{P}(\boldsymbol{m}|\boldsymbol{d};\boldsymbol{w})\big)$, or simply $KL(\mathcal{Q}||\mathcal{P})$, given by

$$KL(\mathcal{Q}||\mathcal{P}) = \mathbb{E}_{\mathcal{Q}}\left[\log\frac{\mathcal{Q}(\boldsymbol{m}|\boldsymbol{d})}{\mathcal{P}(\boldsymbol{m}|\boldsymbol{d};\boldsymbol{w})}\right] = \int_{\boldsymbol{m}} \mathcal{Q}(\boldsymbol{m}|\boldsymbol{d}) \log\frac{\mathcal{Q}(\boldsymbol{m}|\boldsymbol{d})}{\mathcal{P}(\boldsymbol{m}|\boldsymbol{d};\boldsymbol{w})}\, d\boldsymbol{m} \geq 0 \qquad 6.5$$

where $\mathbb{E}_{\mathcal{Q}}$ represents the expectation with respect to distribution $\mathcal{Q}$. Equality to zero holds in equation 6.5 when $\mathcal{Q}(\boldsymbol{m}|\boldsymbol{d}) = \mathcal{P}(\boldsymbol{m}|\boldsymbol{d};\boldsymbol{w})$.

In order to estimate $\mathcal{Q}(\boldsymbol{m}|\boldsymbol{d})$ as an approximation to the desired $\mathcal{P}(\boldsymbol{m}|\boldsymbol{d};\boldsymbol{w})$, we express the log evidence $\mathcal{L}(\boldsymbol{w};\boldsymbol{d})$ in terms of $KL(\mathcal{Q}||\mathcal{P})$ (see equation 2.9) as

$$\mathcal{L}(\boldsymbol{w};\boldsymbol{d}) = \mathcal{F}(\mathcal{Q},\boldsymbol{w}) + KL(\mathcal{Q}||\mathcal{P}) \qquad 6.6$$

where $\mathcal{F}(\mathcal{Q},\boldsymbol{w})$ is the *variational free energy* (or simply *free energy*) and is given by substituting equation 6.3 in equation 2.8, as

$$\mathcal{F}(\mathcal{Q},\boldsymbol{w}) = \mathbb{E}_{\mathcal{Q}}\left(\sum\nolimits_{\hat{c}\in\hat{\mathcal{C}}} \log\psi(\boldsymbol{m}_{\hat{c}},\boldsymbol{d})\right) + \mathcal{S}(\mathcal{Q}) \qquad 6.7$$

where $\mathcal{S}(\mathcal{Q}) = -\int_{\boldsymbol{m}} \mathcal{Q}(\boldsymbol{m}|\boldsymbol{d}) \log\mathcal{Q}(\boldsymbol{m}|\boldsymbol{d})\, d\boldsymbol{m}$ is the entropy of the variational distribution $\mathcal{Q}(\boldsymbol{m}|\boldsymbol{d})$ as a function of data $\boldsymbol{d}$. The expectation term of the free energy in equation 6.7 involves expectations over individual cliques with respect to $\mathcal{Q}$ and is therefore easy to compute for cliques of reasonable size, provided that the choice of the family of possible $\mathcal{Q}$'s allows efficient inference. The entropy term, on the other hand, involves expectations over all

possible realizations of $\boldsymbol{m}$ and does not necessarily factorize. Thus the computational complexity of the entropy term depends on the properties of $Q$. This entails some approximation to overcome the computational complexity of $\mathcal{S}(Q)$.

### 6.4.1 Mean Field Approximation

Within the VB framework, various approximate inference methods have been proposed to address the intractability in large scale probabilistic graphical models or models involving variables with dense dependencies. In chapter 5 (also see Nawaz & Curtis, 2018), we used Bethe's approximation (Bethe, 1935; Yedidia *et al*. 2001a & b) in the *Loopy-Belief Propagation* (LBP) method for a pairwise graphical model to estimate marginal posterior distribution of model parameters under the quasi-localized likelihoods assumption. Here we use the *mean field* (MF) approximation (Opper & Saad, 2001; Koller & Friedman, 2009) as discussed below.

The mean field inference method originated in statistical physics and was inspired by the observations of statistical behaviour of atoms and molecules of various substances such as gases, condensed matter and magnetic materials (Stanley, 1971). It is used in statistical physics to make predictions regarding phase transitions in a substance (i.e. discontinuities in the aggregate properties of a substance as a function of some model parameters). We use this concept to predict spatial distribution of geological facies and discontinuities in the aggregate physical properties of rocks such as porosity, permeability and elasticity, from geophysical data.

In models with no cyclic dependencies among variables, dynamic programming can be used to perform exact inference by exploiting the conditional independence between most variables (Denardo, 2003). In graphs with cycles (or loops), the MF method makes variational inference viable. The MF approximation is based on numerical optimization and assumes some type of independence over the hidden variables $\boldsymbol{m}$. In the context of a CRF, this independence is assumed to be conditioned to the observed variables.

A naïve MF approximation (Jaakkola, 1997; Koller & Friedman, 2009) assumes that all of the hidden variables $\boldsymbol{m}_i, i \in \mathcal{V}$ are independent of each other, i.e.

$$Q(\boldsymbol{m}|\boldsymbol{d}) \cong \prod_{i \in \mathcal{V}} Q_i(\boldsymbol{m}_i|\boldsymbol{d}) \qquad 6.8$$

Such a fully factorized distribution may not capture the information in a general multivariate distribution $\mathcal{Q}(\boldsymbol{m}|\boldsymbol{d})$. However, owing to the Markovian property of a CRF, a factorized distribution with factors of reasonable size and structure may be chosen as a good approximation. We obtain a mean field approximation by taking $\mathbb{Q}$ to be a family of factorizable distributions such that the variational distribution $\mathcal{Q}(\boldsymbol{m}|\boldsymbol{d}) \in \mathbb{Q}$ factorizes into marginal distributions $\mathcal{Q}_c(\boldsymbol{m}_c|\boldsymbol{d})$ over some proper sub-cliques $c$ of the maximal cliques $\hat{c}$ in the model, with some pre-specified order $|c| = q$, such that

$$\mathcal{Q}(\boldsymbol{m}|\boldsymbol{d}) \cong \prod\nolimits_{c \subset \hat{c} \in \hat{\mathcal{C}}} \mathcal{Q}_c(\boldsymbol{m}_c|\boldsymbol{d}) \tag{6.9}$$

We refer to this approximation as the *higher-order mean field approximation*. Note that the above equation degenerates to the naïve MF approximation given by equation 6.8 for $|c| = 1$.

The approximate marginal posterior distributions $\mathcal{Q}_c(\boldsymbol{m}_c|\boldsymbol{d})$ over sub-cliques $c$ may be obtained by maximizing $\mathcal{F}(\mathcal{Q}, \boldsymbol{w})$ as a function of $\mathcal{Q}$ for a given set of parameters $\boldsymbol{w}$ (see Appendix A), which gives

$$\mathcal{Q}_c(\boldsymbol{m}_c|\boldsymbol{d}) \leftarrow \frac{1}{\mathcal{Z}_c(\boldsymbol{d})} \exp\left\{ \sum\nolimits_{\hat{c} \in \hat{\mathcal{C}}: c \subset \hat{c}} \mathbb{E}_{\mathcal{Q}_{\setminus c}}[\boldsymbol{w}^T \boldsymbol{f}(\boldsymbol{m}_{\hat{c}}, \boldsymbol{d}) \mid \boldsymbol{m}_c] \right\} \tag{6.10}$$

where $\mathcal{Q}_{\setminus c}$ represents the per-clique marginals of $\mathcal{Q}$ except for the clique $c$. Thus marginal distribution $\mathcal{Q}_c$ over each approximating clique $c$ is updated by using the expression on the RHS of the of left-arrow. The subscript $\hat{c} \in \hat{\mathcal{C}}: c \subset \hat{c}$ of summation in the above expression reads "for all $\hat{c}$ in $\hat{\mathcal{C}}$ such that $c$ is a proper subset of $\hat{c}$". In simple words, the summation in this expression runs over each maximal clique $\hat{c}$ in the model that contains the approximating clique $c$ that is being updated.

The system of $|\mathcal{C}|$ nonlinear update equations 6.10 collectively represent the *higher-order mean field equations* which may be solved successively in an iterative manner. Since each update has a closed form solution, the free energy $\mathcal{F}(\mathcal{Q}, \boldsymbol{w})$ increases monotonically in each iteration; convergence is therefore guaranteed. However, there are some caveats about convergence that are discussed in section 6.8 which must not be ignored. The factorized distribution $\mathcal{Q}(\boldsymbol{m}|\boldsymbol{d})$ can therefore be evaluated by summation of terms which are defined over a relatively small number of variables (small compared to the exponential number of terms over all of the variables for an un-factorizable distribution). As a consequence, the

computational cost depends mainly on the size of the factors (cliques) and not on the structure of the spatial dependencies. This allows tractable approximate inference in graphs with complex structures where exact inference would require exponential time.

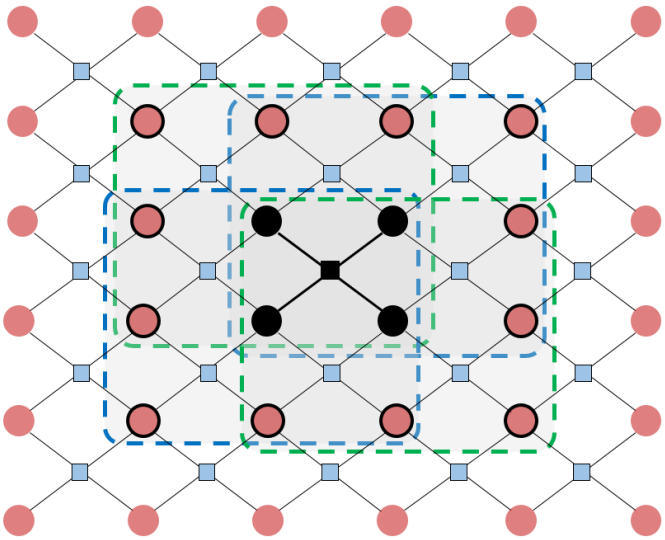


*Figure 6.3: Graphical illustration of the mean field updates. Circles represent vertices in the graph (or the hidden variables $\boldsymbol{m}$), squares which connect vertices through edges (lines) represent clique configurations $\boldsymbol{m}_c$ (also called factors) over approximating cliques $\boldsymbol{c}$ with size 2x2 vertices. Consider an approximating clique $\boldsymbol{c}$ with 2x2 vertices in the centre (shown in black colour) for the mean field update of the approximate marginal distribution $Q_c(\boldsymbol{m}_c|\boldsymbol{d})$. Assume that the maximal cliques $\hat{c}$ in the graph have a size of 3x3 vertices. Four of the (3x3) maximal cliques which share the approximating clique $\boldsymbol{c}$ are shown as rounded rectangles with dashed boundaries. For the model parameters $\boldsymbol{m}_c$ in $\boldsymbol{c}$, the summation in equation 6.10 runs over the set of maximal cliques that share $\boldsymbol{c}$ to compute the conditional expectation over the factors $\boldsymbol{m}_{\hat{c}}$ given $\boldsymbol{m}_c$.*

Although the form of updates is different, the MF update algorithm resembles message passing over a cluster graph, e.g. cluster belief propagation (Koller & Friedman, 2009), where clusters refer to higher-order cliques and messages represent approximate marginal distributions over cliques. Figure 6.3 shows a graphical illustration of the mean field update of $Q_c(\boldsymbol{m}_c|\boldsymbol{d})$ with an example where the approximating clique $c$ has a size of 2x2 vertices, while the maximal cliques $\hat{c}$ in the graph have a size of 3x3 vertices. Unlike Bethe's approximation, the MF approximation does not approximate the objective (the energy functional); it only approximates the restricted optimization space $\mathbb{Q}$ of distributions. The quality of the higher-order mean field approximation depends on the difference in the order of maximal cliques $\hat{c}$

and the approximating cliques $c$: the smaller the difference $|\hat{c}| - |c|$, the better the approximation.

### 6.4.2 Parameter Estimation

The CRF parameters $\boldsymbol{w}$ in equation 6.3 can be estimated by using the *regularized maximum conditional-likelihood* method that searches for the parameters that maximize the conditional log-likelihood of the model for a given training data set (Sutton & McCallum, 2012). In other words, in parameter estimation we aim to find a set of parameters $\boldsymbol{w}$ that makes the approximate posterior distribution $\mathcal{Q}(\boldsymbol{m}|\boldsymbol{d})$ as close to the true distribution $\mathcal{P}(\boldsymbol{m}|\boldsymbol{d};\boldsymbol{w})$ as possible. This method requires computation of the gradient of the log-likelihood $\mathcal{L}(\boldsymbol{w};\boldsymbol{d})$ which is intractable and cannot be computed exactly. For this reason, we also use the mean field approximation to estimate the log-likelihood.

The true joint distribution $\mathcal{P}(\boldsymbol{m},\boldsymbol{d})$ over the entire model is unknown but we assume that we have a training data set that consists of *independent and identically distributed* ($i.i.d.$) samples from the true distribution over maximal cliques – pre-specified subsets of the model. We assume that the training data $D = \{\boldsymbol{m}^{(i)}, \boldsymbol{d}^{(i)}: i = 1, \dots, N\}$ contain $N$ pairs of local configurations of the hidden variables $\boldsymbol{m}^{(i)} = \{\boldsymbol{m}_{\hat{c}}{}^{(i)}: \forall \hat{c} \in \mathcal{C}\}$ over maximal cliques $\hat{c}$ and the corresponding input data $\boldsymbol{d}^{(i)}$, where the bracketed superscript $(i)$ indicates an index over the training instance. The input data $\boldsymbol{d}^{(i)}$ are not required to have the same topology as that of a clique template, however, there should exist some conceptual or logical association between $\boldsymbol{d}^{(i)}$ and $\boldsymbol{m}^{(i)}$. For example, the training data could be prepared from some real data that are manually interpreted and classified by experts, or built from stochastic simulation of geological properties and corresponding data using a variety of earth models (or training images) of expected geology.

The conditional log-likelihood $\mathcal{L}(\boldsymbol{w};\boldsymbol{m}|\boldsymbol{d}) \equiv \log \mathcal{P}(\boldsymbol{m}|\boldsymbol{d};\boldsymbol{w})$ is then given by

$$\mathcal{L}(\boldsymbol{w};\boldsymbol{m}|\boldsymbol{d}) = \sum_{i=1}^{N} \left[\log \mathcal{P}\left(\boldsymbol{m}^{(i)}|\, \boldsymbol{d}^{(i)};\, \boldsymbol{w}\right)\right] - \lambda|\boldsymbol{w}|^{\mathbf{2}}$$

$$= \sum_{i=1}^{N} \left[\sum_{\hat{c}\in\hat{\mathcal{C}}} \boldsymbol{w}^T \boldsymbol{f}\left(\boldsymbol{m}_{\hat{c}}{}^{(i)}, \boldsymbol{d}^{(i)}\right) - \log \mathcal{Z}\left(\boldsymbol{d}^{(i)};\boldsymbol{w}\right)\right] - \lambda\left|\boldsymbol{w}_p\right|^2 \qquad 6.11$$

where we used equation 6.3 in the second equality, and $\lambda > 0$ is a regularization parameter which controls the strength of regularization. The conditional log-likelihood $\mathcal{L}(\boldsymbol{w}; \boldsymbol{m}|\boldsymbol{d})$ in the above equation cannot be maximized analytically. We therefore use gradient based non-linear numerical optimization. Substituting for $\mathcal{Z}(\boldsymbol{d}^{(i)}; \boldsymbol{w})$ from equations 6.4, the gradient of the conditional log-likelihood in equation 6.11 may be written as

$$\nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}; \boldsymbol{m}|\boldsymbol{d}) = \sum_{i=1}^{N}\left[\boldsymbol{f}(\boldsymbol{m}_{\hat{c}}{}^{(i)}, \boldsymbol{d}^{(i)}) - \mathbb{E}_{\boldsymbol{m}\sim\mathcal{P}(\boldsymbol{m}|\boldsymbol{d}^{(i)},\boldsymbol{w})}\left[\boldsymbol{f}(\boldsymbol{m}_{\hat{c}}, \boldsymbol{d}^{(i)})\right]\right] - 2\lambda\boldsymbol{w} \qquad 6.12$$

The zero-gradient conditions thus require that the feature functions $\boldsymbol{f}(\boldsymbol{m}_{\hat{c}}, \boldsymbol{d})$ have the same expectations under the model (the CRF) and the empirical (training) distributions. We therefore approximate the expected features using the mean field inference method as

$$\mathbb{E}_{\boldsymbol{m}_{\hat{c}}\sim\mathcal{P}(\boldsymbol{m}_{\hat{c}}|\boldsymbol{d}^{(i)};\,\boldsymbol{w})}\left[\boldsymbol{f}(\boldsymbol{m}_{\hat{c}}, \boldsymbol{d}^{(i)})\right] \cong \mathbb{E}_{\boldsymbol{m}_c\sim\mathcal{Q}_{\boldsymbol{w}}(\boldsymbol{m}_c|\boldsymbol{d}^{(i)})}\left[\boldsymbol{f}(\boldsymbol{m}_{\hat{c}}, \boldsymbol{d}^{(i)})\right]$$

$$= \sum_{\boldsymbol{m}_c}\mathcal{Q}_{\boldsymbol{w}}(\boldsymbol{m}_c|\boldsymbol{d}^{(i)})\,\boldsymbol{f}(\boldsymbol{m}_{\hat{c}}, \boldsymbol{d}^{(i)}) \qquad 6.13$$

where $\mathcal{Q}_{\boldsymbol{w}}(\boldsymbol{m}_c|\boldsymbol{d}^{(i)})$ refers to the marginals of $\mathcal{Q}_{\boldsymbol{w}}(\boldsymbol{m}|\boldsymbol{d}^{(i)})$ under the approximation $\mathcal{Q}_{\boldsymbol{w}}(\boldsymbol{m}|\boldsymbol{d}^{(i)}) \cong \mathcal{P}(\boldsymbol{m}|\boldsymbol{d}^{(i)}; \boldsymbol{w})$, i.e. for a given set of parameters $\boldsymbol{w}$. Since we assume that all of the variables ($\boldsymbol{m}^{(i)}$ and $\boldsymbol{d}^{(i)}$) are observed in the training data, the log-likelihood $\mathcal{L}(\boldsymbol{w}; \boldsymbol{m}|\boldsymbol{d})$ is a concave function. Therefore, any local maximum is indeed a global maximum. The log-likelihood can therefore be maximized using gradient ascent optimization as long as we can compute the gradient exactly; however this is known to be too slow to converge (Yuan, 2010). Newton or quasi-Newton type methods such as the so called *BFGS* method (Dennis & Schnabel, 1996) use local curvature of the objective function to achieve faster convergence; however these methods require computation and inversion of the Hessian matrix $\mathbf{H}$ given by

$$\boldsymbol{H}\left(\mathcal{L}(\boldsymbol{w}; \boldsymbol{m}|\boldsymbol{d})\right) = -\sum_{i=1}^{N}\left(\mathrm{Cov}_{\boldsymbol{m}\sim\mathcal{P}(\boldsymbol{m}|\boldsymbol{d}^{(i)};\boldsymbol{w})}\left[\boldsymbol{f}(\boldsymbol{m}, \boldsymbol{d}^{(i)})\right]\right) - 2\lambda\boldsymbol{I} \qquad 6.14$$

Computational complexity of evaluating the exact Hessian matrix is quadratic in the number of parameters, i.e. $O(n_{\boldsymbol{w}}^2)$ per iteration, where $n_{\boldsymbol{w}}$ represents the number of parameters. For a small number of parameters, computing the exact Hessian matrix is feasible and the 2nd-order optimization methods offer faster convergence in this case. However, if the

number of parameters is large, computing the Hessian matrix may not be feasible. The limited-memory version of the BFGS method, known as the *L-BFGS* method (Nocedal, 1980; Fletcher, 1987; Nocedal & Wright, 2006), may be used in this case.

A key challenge in performing probabilistic inversion with non-localized likelihoods is that the inverse problem is highly nonlinear because inference for the posterior distribution requires some estimates of the model parameters (CRF weights $\boldsymbol{w}$), whereas estimation of the model parameters requires some estimates of the posterior distribution. This paradox may be solved by first performing inference with randomly initialized parameters $\boldsymbol{w}$ to approximate the marginal posterior distributions, and then updating the parameters by using these approximate posterior distributions. Then inference and parameter estimation are carried out in an iterative fashion until both the model parameters and estimated marginal posterior distributions converge to within a predefined tolerance.

## 6.5 Computational Complexity

The computational complexity of this method can be divided into the three main components of the algorithm: learning the feature functions, mean field inference, and CRF parameter estimation. Feature functions are a rather general concept, and their learning cost depends on the complexity of the task and on the exact method used for learning. For example, the computational complexity of learning feature functions using a *multi-layer perceptron* (MLP) neural network is at most quadratic in the total number of neurons in the network.

The overall cost $C_{MF}$ of the mean field algorithm, expressed in terms of the maximum number of floating point operations required, is given by

$$C_{MF} \leq |\mathcal{C}| * \max|\boldsymbol{m}_c| * \max|\mathcal{N}_c| * L_{MF} \tag{6.15}$$

where $c$ is the clique that defines the mean field approximation in equation 6.10, $|\mathcal{C}|$ is the total number of cliques $c$ in the model, $\max|\boldsymbol{m}_c|$ is the maximum dimensionality of model parameters in a clique $c$, $\max|\mathcal{N}_c|$ is the maximum number of maximal cliques $\hat{c}$ that contain $c$ as a subset in the model which is also referred to as the *neighbourhood cardinality* of the model, and $L_{MF}$ is the total number of MF iterations.

Similarly, the cost $C_{PE}$ of parameter estimation for the CRF model with the L-BFGS method is given by

$$C_{PE} \leq (|\mathcal{C}| * \max|\boldsymbol{m}_c| * \max|\mathcal{N}_c|) * N * n_{\boldsymbol{w}}^2 * L_{PE} \quad 6.16$$

where $N$ is the number of training examples, $n_{\boldsymbol{w}}$ is the number of weights in the CRF model, and $L_{PE}$ is the total number iterations required for the L-BFGS algorithm to converge.

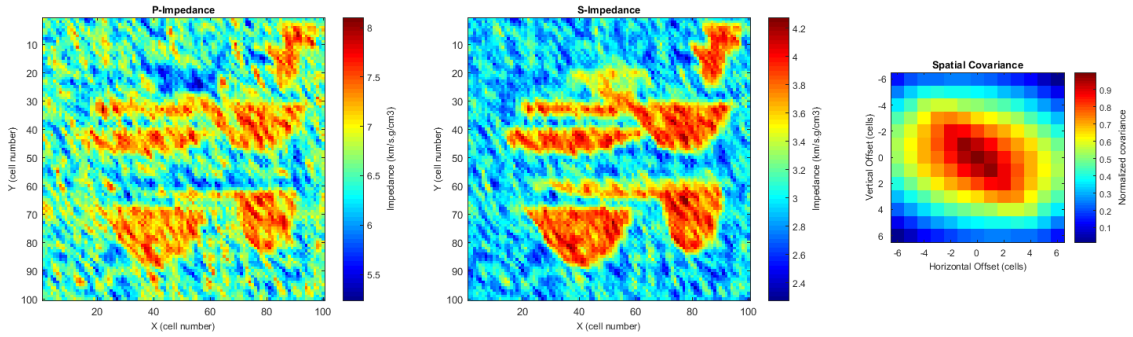
Equations 6.15 and 6.16 show that the factors which control the cost of this method are the number $n_{\boldsymbol{w}}$ of CRF parameters, the dimensionality of model parameters within a clique $|\boldsymbol{m}_c|$, and the size of neighbourhood cardinality $|\mathcal{N}_c|$. The latter two factors themselves depend on the clique size $|c|$ of the approximating distribution $Q(\boldsymbol{m}|\boldsymbol{d})$, and the maximal clique size $|\hat{c}|$ in the graph. If the clique size is too small, it may not be able to capture the expected complexity in the model parameters, and subsequent inference may not be able to model the true spatial distribution of model parameters. If the clique size is too large it may increase the required computational cost significantly. A trade-off is thus required between geological complexity that is to be modelled and the required computational resources. Nevertheless, the above cost is expected to be far lower than would be required to solve the same problem using Monte Carlo methods for the class of problems which involve non-localized likelihoods and which make no conditional independence assumptions on the observed data in high dimensions.

## *6.6 Synthetic Test*

Removing the assumption of localized likelihoods and conditional independence of data means that our method should be able to learn any correlations present in the data due to spatial blurring of data or due to correlated noise, provided that it can be represented within our probabilistic CRF model. This means that this method ought to be robust against correlated noise present in data as long as we can model some salient characteristics (or features), e.g. spatial correlation of noise. In order to test this, and to benchmark the current method against previous research, the same test Earth model (shown in figure 4.5b) is used here as was used in sections 4.6 and 5.6 (also see Walker & Curtis (2014a), and Nawaz & Curtis, 2017 & 2018). Here, for the first time, it is demonstrated that the new method is capable of inverting seismic attributes for facies with reasonable accuracy even in the presence of

strongly correlated systematic and/or random noise. Such noise introduces strong correlations in the data that may be learnt using a neural network and properly accounted for in the inversion process.

Synthetic P-wave and S-wave impedance profiles were generated from the target cross-section (figure 4.5b) to represent the corresponding real-data derived seismic attributes. Synthetic attributes were first generated as described in section 5.6. Then, correlated systematic and random noise was introduced in the simulated seismic attributes in the form of NW-SE oriented random streaks of amplitudes by convolving the noise-free attribute sections with a NW-SE oriented filter, in order to generate collocated synthetic seismic attributes (P-wave and S-wave impedances) as the noisy input $\boldsymbol{d}$ for our method (figure 6.4a and b). Noise in real seismic data due to acquisition foot-print, non-uniform source directivity, or multiple scattering of energy in the subsurface are examples of such a case where noise is convolved with the desired signal, i.e. and is not just additive. The aim is to train our algorithm to disregard the correlated noise and reproduce the true distribution of facies. We refer to the resulting synthetic attributes as the 'true data' as these were then inverted with our method with the aim to reproduce the 'target geology' (figure 4.5b).

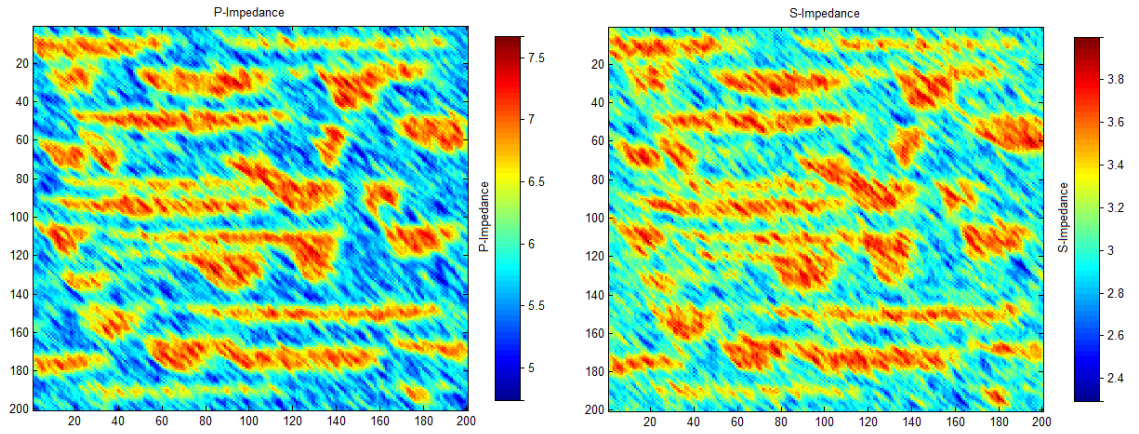



*(a)* *(b)* *(c)*

*Figure 6.4: Synthetic (a) P-wave and (b) S-wave impedance attributes used as input for the synthetic test. (c) Spatial covariance matrix computed from the synthetic attributes (P-wave and S-wave impedances) cross-sections in panels a and b, for a maximum vertical and lateral offset of 6x6 cells.*

The spatial covariance matrix was computed from these synthetic attributes (the input data) which provides an estimate of the spatial variability of impedances in the presence of strongly correlated noise. The computed covariance matrix was then tapered to retain the maximum amplitudes along the main diagonal while the off-diagonal correlations were

suppressed to yield a filter that can introduce similar correlated noise in the simulated examples that we used later for supervised learning. The normalized spatial covariance matrix is shown in figure 6.4c which shows strong correlations in the NW-SE direction similar to the orientation of noise streaks in the data. Such an approach where noise is estimated from the observations under the assumption of stationarity is commonly referred to as *empirical Bayes*.




*(a)* *(b)*

*Figure 6.5: Simulated (a) P-wave and (b) S-wave impedance sections generated by convolving stochastically simulated attributes from the training image (figure 4.5a) with the spatial correlation matrix in figure 6.4c in order to mimic the correlated noise observed in the input attribute sections (figure 6.4a & b). These simulated sections are used to generate stochastic examples for training the neural network in order to learn feature functions.*

Prior information was extracted from the training image (figure 4.5a) in terms of prior probabilities $\mathcal{P}(\boldsymbol{m}_{\hat{c}})$ constructed from histograms of various facies configurations that occur in the image. We chose two clique templates each with a size of 9x9 model cells to relate facies patterns in a clique with the corresponding P-wave and S-wave impedances, respectively. The size of the clique template was chosen based on the size and shape of features observed to be present in the attributes, and it defines a maximal clique in the underlying graphical model. The approximating clique was chosen to have a pairwise structure, such that each cell in the 2D model has four neighbour, two horizontal and two vertical.

Next we prepared examples of seismic attributes and the desired facies patterns. Since the attributes that are used as the data to test our method are synthetically generated, we use the term 'simulated' (rather that 'synthetic') for the attribute sections used to build stochastic examples for training a neural network to learn feature functions. Simulated attributes were

generated using the rock physics model described above from facies patterns present in the training image (figure 4.5a). In order to introduce correlated noise in the simulated attributes, these were cross-correlated with the tapered form of the spatial covariance matrix estimated from the 'true data' shown in figure 6.4(a & b). The resulting noisy sections of P-wave and S-wave impedances simulated from the facies present in the training image are show in figure 6.5.

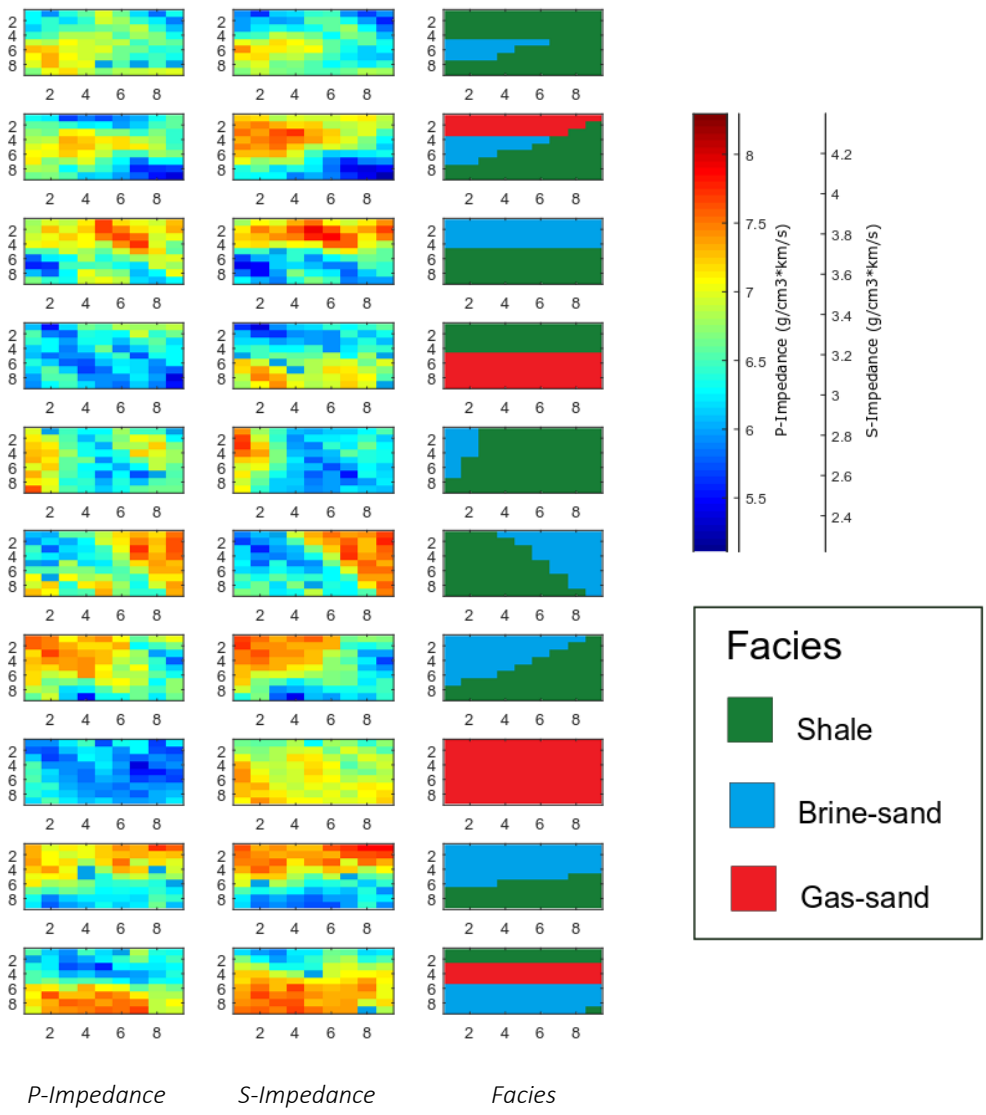



*Figure 6.6: Examples of simulated P-wave and S-wave impedances and corresponding facies patterns in a window of size 9x9 model cells. These examples were used to train a neural network in order to learn feature functions.*

An example database was then prepared for supervised learning in the form of two sets of facies patterns extracted from the training image (figure 4.5a) within the pre-specified clique templates, and the corresponding cells in the simulated attributes sections (figure 6.5). In the context of supervised learning, we refer to the facies patterns in the example database as the 'target facies', and the corresponding simulated attributes as 'simulated features'. The simulated features were extracted from each of the simulated attribute sections (figure 6.5) using windows of the same size as the clique templates (9x9 cells). In this example, the size of training features was chosen to be the same as that of the clique templates (9x9 cells) which adequately captured the salient characteristics of data and correlated noise with respect to the corresponding facies configurations. A total of 5000 examples were stochastically generated with random permutations of facies configurations within the predefined clique templates and the corresponding features (simulated P-wave and S-wave impedances) in the example database. Figure 6.6 shows a few such examples with training features from each of the clique templates and the corresponding target facies.

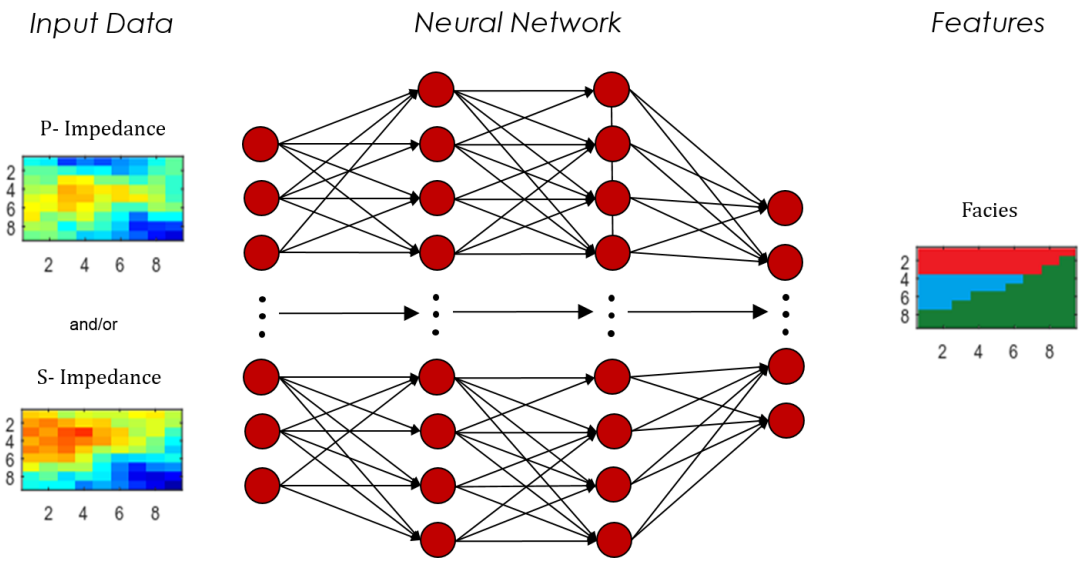



*Figure 6.7: A schematic illustration of learning feature functions from input attributes (P-wave and/or S-wave impedances) using a neural network.*

Feature functions were then defined for each of the clique templates as a vector of indicator variables corresponding to the facies in each cell of the clique template. Each of the indicator variables is set to 1 for the facies present in the target pattern, and 0 for all other

facies patterns. Separate neural networks were then trained with the training features (e.g. the P-Impedance and S-Impedance columns in figure 6.6) as input and the corresponding feature functions (e.g. the indicator representation of the facies columns in figure 6.6) as the desired output for each of the clique templates. In this manner the outputs of a trained neural network may be interpreted as a measure of how likely is a facies configuration for a given input feature (figure 6.7). After training the neural networks on stochastically generated examples, features were extracted from the 'true data' corresponding to each of the clique templates, and the associated feature functions were computed using the trained neural network.

After computing the feature functions, the CRF weights $\boldsymbol{w}$ were initialized randomly and approximate inference was performed using the mean field update equations 6.10 to obtain the variational distribution $\mathcal{Q}(\boldsymbol{m}|\boldsymbol{d})$ as an approximation to the model distribution $\mathcal{P}(\boldsymbol{m}|\boldsymbol{d};\boldsymbol{w})$. These posterior distributions were then used to update the CRF weights using the quasi-Newton optimization method L-BFGS. Since estimation of both the posterior distributions and the CRF weights requires the other to be known, each of these were alternately updated in an iterative fashion until both converged within a pre-specified tolerance. The final estimates of the marginal posterior distributions in each cell are shown in figure 6.8(a-c) and the entropy (a measure of uncertainty) is shown in figure 6.8d. The map of the facies that has the maximum of the marginals in each cell, shown in figure 6.8e, shows reasonable reconstruction of the target geology (figure 4.5b) given that the input attributes contain strongly correlated noise. The quality of prediction is quantified in terms of success rate computed as a percentage of cells with predicted facies for each of the three facies in the model. This is shown by the *confusion matrix* in figure 6.8f.

The quality of prediction is very good as the overall accuracy rate is 97%. The major errors lie in false prediction of shale when the true facies was brine-sand, and false prediction of brine-sand when the true facies was gas-sand. Errors are mostly found at the transitions between different facies where entropy is at its highest, see figure 6.8d. The high accuracy of prediction resulted from the fact that the noise follows a linear (NW-SE) trend (figure 6.4c) that is different from the trend of geological correlations, and that the prior information extracted from the training image is a good representation of the 'true' geology. Either of these may not be guaranteed in real data problems. Therefore, the accuracy rate may not be as good in practical situations and it depends on the quality of geological prior information and

our ability to discriminate noise correlations from expected geological correlations. Nevertheless, high prediction accuracy in this synthetic example does show that the method is reliable provided the required inputs are available with reasonable accuracy.

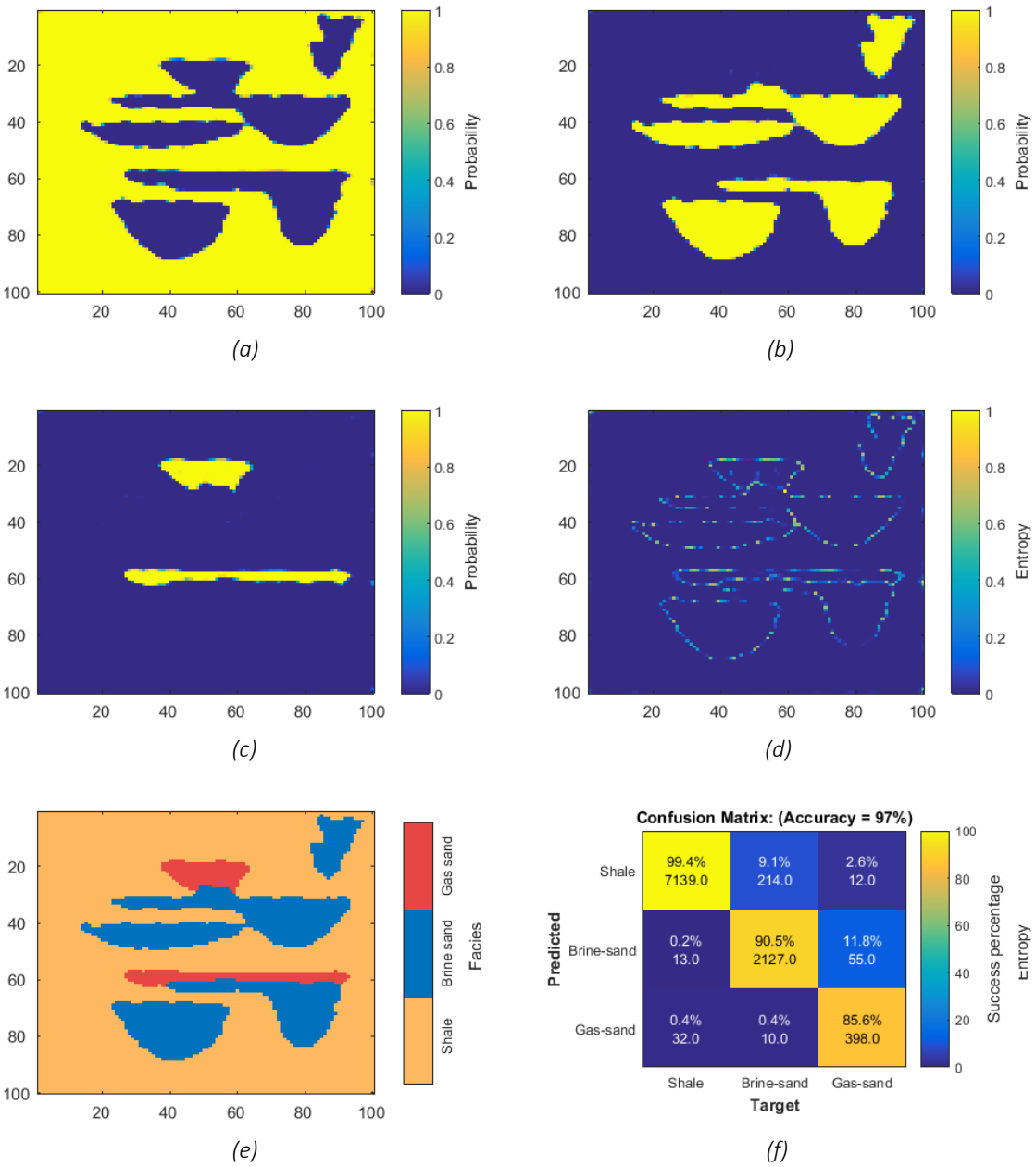



*Figure 6.8: (a-c) Approximate marginal posterior distributions for the three facies: shale, brine-sand and gas-sand, obtained after mean field approximation with optimized CRF parameters. (d) Entropy of the approximate marginal posterior distributions. (e) Facies with maximum marginal distribution. Note that this is not a maximum-a-posteriori (MAP) estimate (i.e. it is not a realization). (f) Confusion matrix showing the success rate of predictions versus targets for the three facies.*

### 6.6.1 Summary of the Method as Applied Above

The following is a step-wise summary of the overall method used in this synthetic example:

1. *Define the graphical model (maximal clique size) and extract facies patterns from the training image to construct the prior distribution.*
2. *Identify features of the data and collect data correlation statistics.*
3. *Perform forward simulation of data that corresponds to the training image incorporating the correlation statistics.*
4. *Define clique templates, and feature functions that relate data features in a clique template to facies patterns in a maximal clique.*
5. *Train a machine learning model (e.g. a neural network) on training examples extracted from the training image and its associated simulated data, to learn feature functions from the data.*
6. *Define a CRF model using equation 6.3 with feature functions as the basis functions and initialize CRF weights $\boldsymbol{w}$ randomly.*
7. *Perform mean field inference using equations 6.10 to estimate approximate posterior distribution $Q(\boldsymbol{m}|\boldsymbol{d})$ from the current estimate of CRF weights $\boldsymbol{w}$.*
8. *Update CRF weights $\boldsymbol{w}$ using a non-linear optimization method (e.g. L-BFGS) with the gradient of the conditional log-likelihood in equation 6.12 computed from the current estimate of approximate posterior distribution $Q(\boldsymbol{m}|\boldsymbol{d})$.*
9. *Repeat steps 8 and 9 until the approximate posterior distribution $Q(\boldsymbol{m}|\boldsymbol{d})$ and the CRF weights $\boldsymbol{w}$ converge to within a predefined tolerance.*

### 6.6.2 Comparison with Quasi-Localized Likelihoods Based Inversion

For a comparison we applied our previous method of facies inversion using quasi-localized likelihoods (chapter 5, and Nawaz & Curtis, 2018) to the data with strongly correlated noise as shown in figure 6.4. It was shown in section 5.6 that the quasi-localized method performs significantly better than localized methods in this problem. In order to make a fair comparison between the two methods, the QLL based method presented in chapter 5 was

modified to use higher-order cliques of size 9x9 instead of just pairwise cliques. The results from the QLL based method are shown in figure 6.9: these exhibit good discrimination between shale and sand (figure 6.9a), while the discrimination between brine-sand and gas-sand is poor (figure 6.9b & c). The latter occurs because although spatial inference is performed in order to reproduce geologically plausible patterns of facies (as depicted in the training image in figure 4.5a), the method could not handle correlated noise in the data. Here we recall that most previously existing methods of facies inversion assume that any correlations present in the data are a direct consequence of correlations in the geology – the so called conditional independence (CI) assumption on data. The current method, on the other hand, provides a new mathematical framework for probabilistic inference that incorporates complex features in the data that should be regarded or discarded during the inversion process, and is capable of providing reliable results (figure 6.8) even in the presence of strongly correlated noise.

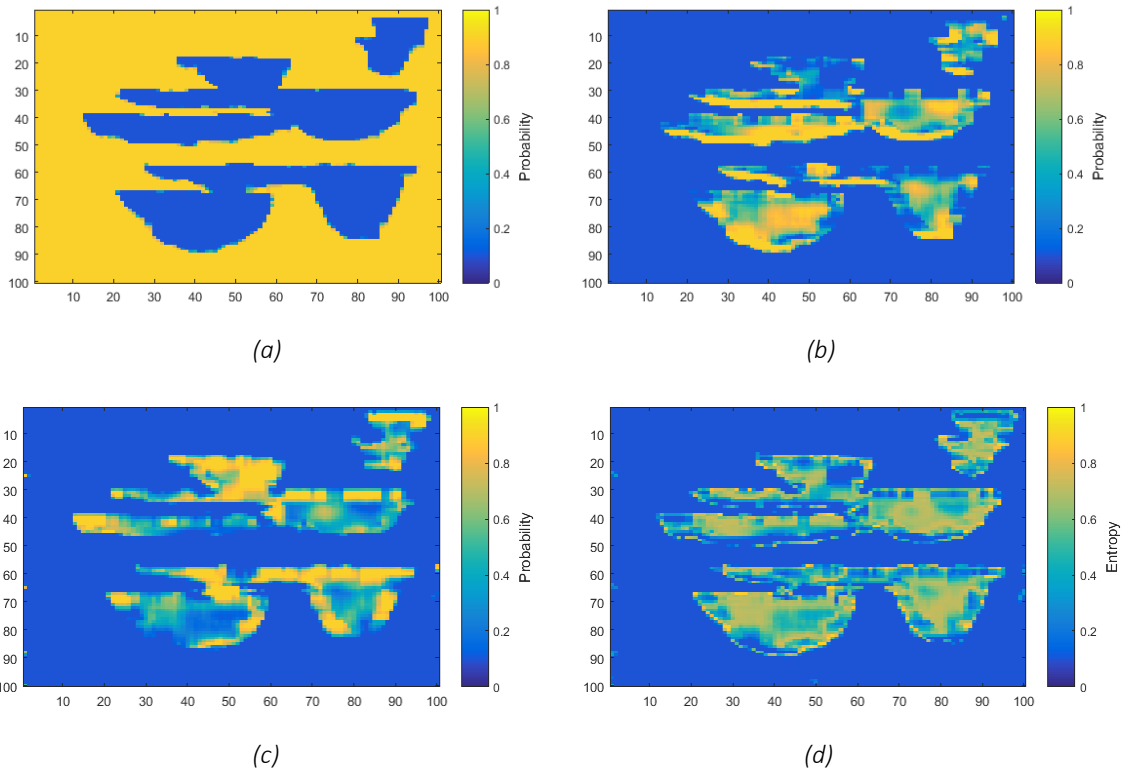
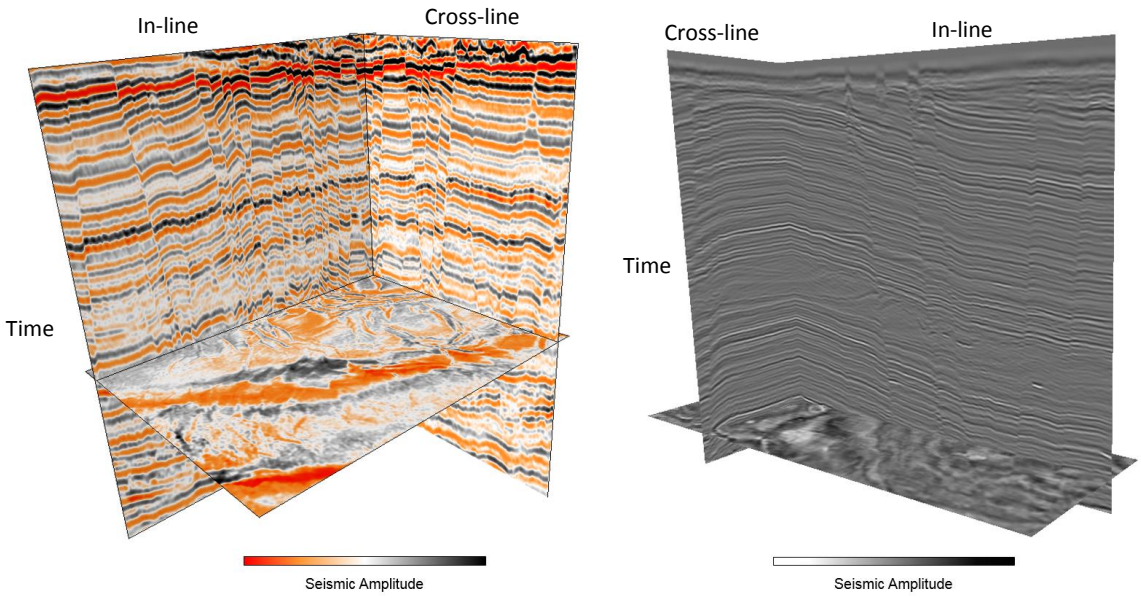



*Figure 6.9: Approximate posterior marginal distributions for the three facies obtained using the quasi-localized likelihoods based facies inversion method of Nawaz & Curtis, 2018 in the presence of strongly correlated noise, (a) shale, (b) brine-sand, and (c) gas-sand.*

## 6.7 Application: Fault Interpretation in 3D Seismic Data

The concept of feature functions is general and is widely used in a wide range of machine learning applications. As a consequence of that, discriminative Bayesian inversion (DBI) method developed in this research is also general and may be used to solve a variety of problems in geosciences, and in other fields of research. This is demonstrated here by a real data example where we aim to compute probability of presence of a geological fault at each sample in a 3D seismic image.




*(a)* *(b)*

*Figure 6.10: Two different 3D seismic images used to demonstrate application of the current method for fault interpretation using supervised learning, (a) Seismic-1 – used to generate training dataset, and (b) Seismic-2 – used for validation. Different colour schemes are used for these images to clearly identify these from each other in the figures below.*

A number of seismic attributes indicate presence of faults (Randen & Sønneland, 2005), either directly by detecting lateral discontinuity (e.g. coherency dip and variance) or indirectly by enhancing amplitude variations in reflection events (e.g. instantaneous gradient and phase, and similarity). However, not all discontinuities correspond to the presence of faults; they may correspond to stratigraphic breaks such as pinch-outs. Therefore, identification of faults requires spatial context and cognitive interpretation of discontinuities. Supervised learning,

e.g. using artificial neural networks (ANN) may resolve ambiguities in fault interpretation by analyzing a certain volume of the 3D seismic image (and its attributes) around each sample. Training examples can be generated from manually interpreted seismic data, or from synthetic data computed from expected models of geology (Wu *et al.* 2018).

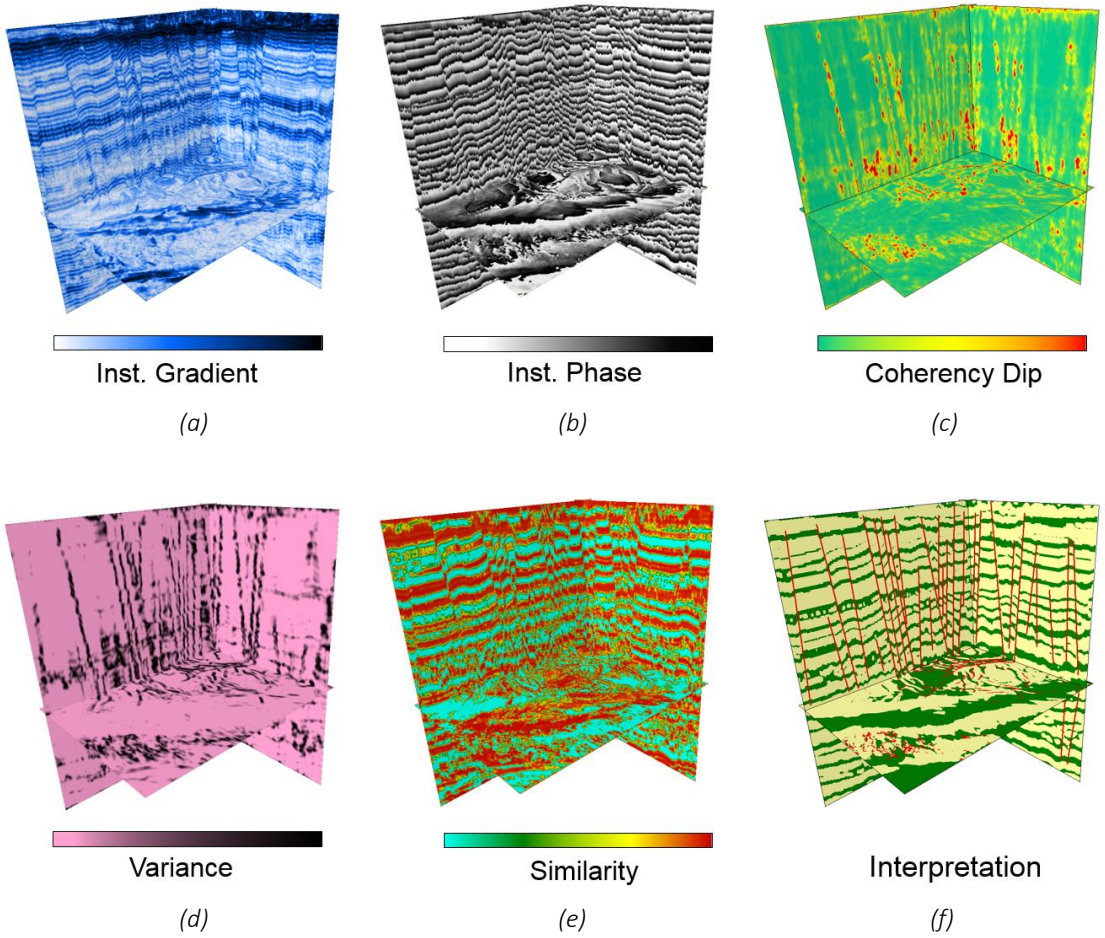



*Figure 6.11: Various seismic attributes computed from the 3D seismic image shown in figure 6.10a, and manual fault interpretation. (a) Instantaneous gradient, (b) Instantaneous phase, (c) coherency dip, (d) variance, (e) similarity, and (f) manually interpreted faults (shown in red colour) superimposed on interfaces (shown in green colour) enhanced using edge detection with a yellow background. Amplitudes in a-e were normalized between 0 and 1.*

Two 3D seismic images from different basins in New Zealand were used to demonstrate an application of the current method for automatic fault interpretation from 3D seismic data. We refer to these images as Seismic-1 and Seismic-2 (figure 6.10), where Seismic-1 was used to generate training examples by manually interpreting faults, and Seismic-2 was used for

validation. First, the following seismic attributes (shown in figure 6.11a-e) were computed from Seismic-1:

1. *Instantaneous Gradient: It provides direction of the normal to the surface of maximum coherency at each point in the seismic data.*
2. *Instantaneous Phase: It enhances both the continuity and discontinuity at each point in the seismic image.*
3. *Coherency Dip: It measures the dip direction of coherency at each sample, where the latter is sensitive to lateral variations in the seismic amplitudes (e.g. as caused by the presence of faults).*
4. *Variance: It is a measure of signal unconformity in terms of trace to trace variability and enhances discontinuities.*
5. *Similarity: A measure of how similar two or more trace segments are. It is computed at each sample by analysing a number of traces around that sample within a short time window.*

Further, a new 3D image of faults was generated by interpreting faults manually on Seismic-1, such that a value of 1 was assigned to each sample where a fault is present, and a value 0 otherwise. Continuous lateral reflections (horizons) were then auto-tracked using edge detection, and interpreted faults were superimposed on it for visualization (figure 6.11f).

A 3D window of 16 samples along each of the dimensions was chosen to represent a maximal clique in our graphical model. The approximating clique was chosen to have a pairwise structure, such that each cell in the 3D model has six neighbours, two along each dimension. Although not required by definition, we chose the window to extract attribute data to be the same as the dimensions of a maximal clique. A total of 80,000 collocated windows were selected at randomly chosen coordinates (inline and crossline numbers, and time) from each of the 3D volumes of interpreted faults as desired targets, and computed seismic attributes as input features (figure 6.12). Presence of faults at the edges of the target windows was manually suppressed to avoid spurious results, and the dimensionality of target windows along the time direction was reduced from 16x16x16 (along inline, crossline, and time dimensions) samples to 16x16x1 samples using principal component analysis. This retains most of the information in the target windows since faults are typically oriented nearly vertically.

Such a dimensionality reduction may not be acceptable for low angle faults, but it was not a problem for either of the 3D seismic images used in this example.

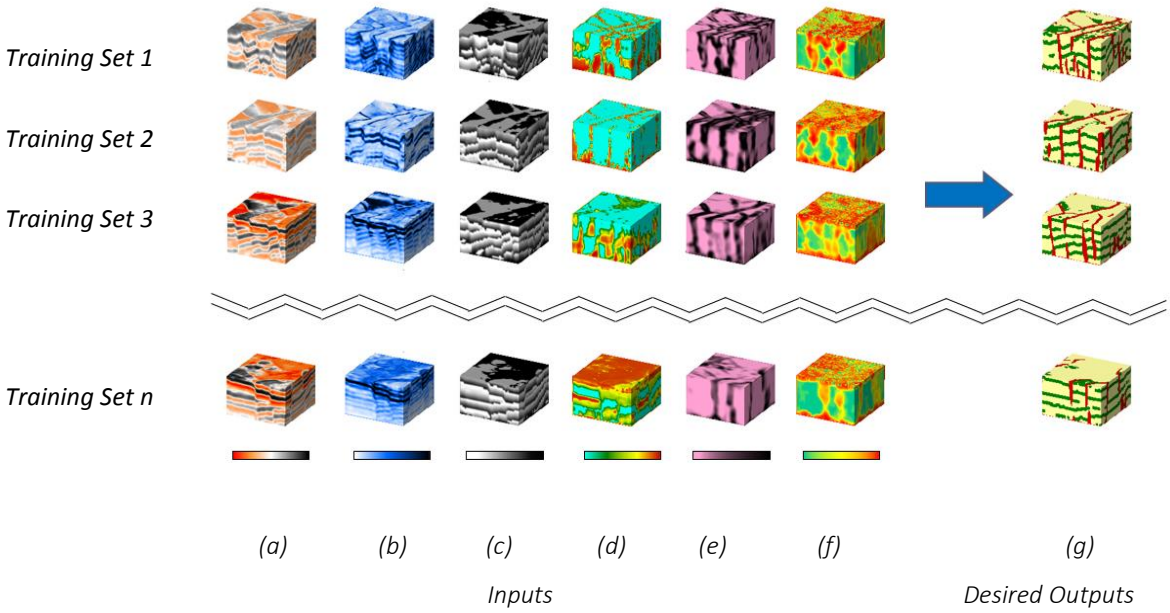



*Figure 6.12: Training examples generated from the input (a) seismic amplitude data in the form of attributes: (b) instantaneous gradient, (c) instantaneous phase, (d) coherency dip, (e) variance, and (f) similarity, extracted within a short 3D window of 16 samples along each dimension, and (g) the corresponding fault interpretation used as the desired output from neural network. The interpretation shows manually interpreted faults in red colour, and auto-tracked horizons in green colour with a yellow background. The horizons are only used for display purpose. These examples were used to train a neural network in order to learn feature functions.*

A convolutional neural network (CNN) with three convolutional layers and one fully connected (dense) layer (figure 6.13) was used to learn the likelihood of presence of fault at each sample from the input seismic attributes. Note that no pooling layer was used since use of filter stride is more favourable to reduce the size of the network (Springenberg *et al.* 2014). The inner-product of the output vector from the CNN and the desired output (after dimensionality reduction) was interpreted as a feature function $f(\boldsymbol{m}_{\hat{c}}, \boldsymbol{d})$. Since all of the attributes were used in learning the CNN, relative weights $\boldsymbol{w}$ of different features (attributes) were implicitly defined in the dense layer. This resulted in a single combined feature function which relates data and the model parameters (presence or otherwise of fault in each cell) within a clique. The feature functions were transformed into clique-wise probabilities using

equation 6.3, and spatial inference was then performed using equation 6.10 to recompose these probabilities into a full scale 3D image of fault probability at each sample.

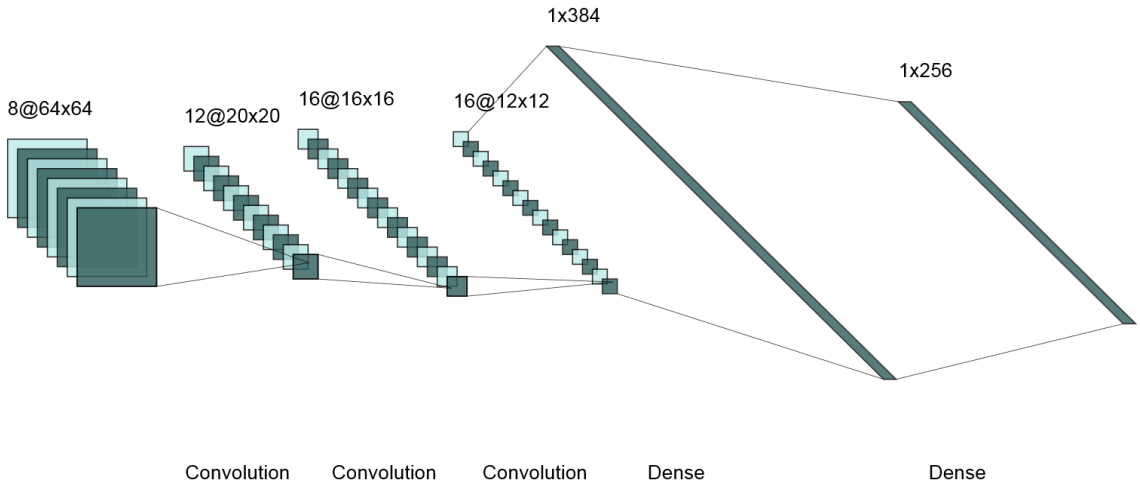
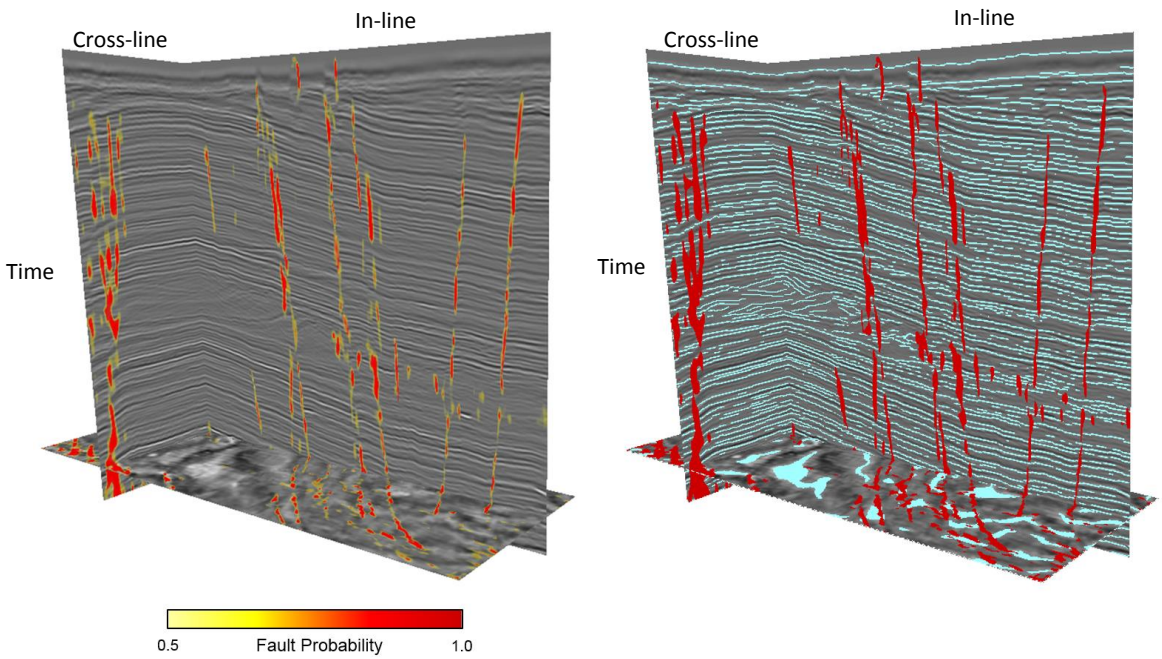



*Figure 6.13: Architecture of the convolutional neural network (CNN) used to learn feature functions for fault interpretation. Number of filters are shown in each layer. The filter size was 7x7 in the first layer and 5x5 in the second and third layers. Filter stride was 3 in the first layer and 1 in the second and third layers.*




*(a)* *(b)*

*Figure 6.14: 3D seismic image shown in figure 6.10b, (a) with fault probability at each sample overlaid, (b) with faults (shown in red, where fault probability is greater than 0.5) and auto-tracked horizons (shown in blue) overlaid.*

Figure 6.14a shows the fault probability at each sample in the 3D image overlaid on the original seismic image, and figure 6.14b shows the fault indicator where a value of 1 is assigned to a sample where fault probability is greater than 0.5, and a value of 0 is assigned otherwise. Figure 6.14b also shows auto-tracked horizons (in blue) for visual comparison, which shows a number of breaks in the horizons which do not actually correspond to the presence of a fault. Such breaks are either caused by stratigraphic features or due to noise in the data. Such features are not identified as faults by the method as it was guided in a supervised manner. Major faults and most of the minor faults have been interpreted with reasonable accuracy.

## *6.8 Discussion*

Both generative and discriminative modelling require reasonable knowledge of the underlying relationship between model parameters and the data. This relationship is often presented in the form of mathematical or computational functions in generative modelling, and is presented as (often simulated) training examples from which mathematical functions, here referred to as feature functions, may be derived in the discriminative approach. The advantage of the discriminative approach is that it learns the inverse of the underlying forward model, and the inverse may be arbitrarily complex and non-linear, may represent non-uniqueness in that inverse relation, and may represent the true model-data relationship (given suitable training examples from the real relationship) rather than a synthetic approximation to that relationship. Consequently, discriminative modelling may learn more complexity in a problem with less effort than is required to produce an accurate generative model for the same problem.

As an example, we showed with a synthetic example in section 6.6 that we only needed to model and learn some statistical characteristics of correlated noise present in the data in order to discard it during inversion of the noisy data. Applying the generative modelling approach to such an example requires reliable prediction of the correlated noise. Formulating the joint distribution over noisy data and the desired model parameters in a generative approach can be hard as it would require reliable prediction of the noise along with the signal for any given set of model parameters. The discriminative approach simplifies it by not attempting to model the noise; only statistical characteristics of noise are needed in order to discriminate between signal and noise.

Generating and learning from training examples may be a tedious task, however, the effort spent preparing training examples and learning the inverse mapping (from data to model parameters) often depends mostly on the complexity of the problem, and not so much on the size of the problem in cases where the problem can be decomposed (factorized) into smaller sub-components. This means that the same training examples that are prepared for inversion of a small seismic section may be used to invert a large 3D seismic volume provided that the assumption of stationarity (that the same training examples are equally appropriate everywhere in the volume) is valid. In other words, the expensive part of our method (the learning stage) operates at a scale that is greatly reduced compared to the full problem, allowing the method to scale to far larger problems.

The feature functions must be defined such that they effectively capture complex relationships between the geological model and the data. Various machine learning methods have been proposed to achieve this task, for example random forests (Ho, 1995), support vector machine (SVM, Cortes & Vapnik, 1995) and deep neural networks (DNN, e.g. Hinton *et al*. 2006) such as convolutional neural networks (CNN, Zhang, 1988). The decision about which method is used to learn feature functions depends mainly on the type and complexity of the features that are to be modelled, and requires an interpretive approach. The general approach presented here allows any such method, or a combination thereof to be employed under the assumption that the training examples represent the data-model relationship reasonably well, and that the accuracy of feature functions learnt from the training examples is acceptable.

Training examples can be created in at least two ways: feature vectors could be extracted from real data and manually classified by experts to provide the corresponding geological parameters, or pairs of feature vectors and their classes could be created by stochastically generating synthetic data for a variety of earth models (or training images) of expected geological features. The former approach is a type of expert elicitation in which statistical information is elicited from experts based on their subjective opinion about the extracted data features (Polson & Curtis, 2010, 2015; Walker & Curtis, 2014b; Macrae *et al.* 2016). The latter approach uses a generative framework where data are modelled from the spatial distribution of geological properties obtained, for example, by using geological process modelling. Task specific features in the data must be captured in the training examples to define feature functions. Although the overall inversion still uses a discriminative framework for learning the posterior distribution of facies across the entire model given all of the data, it

may thus be decomposed into smaller generative models, each of which only models the facies distribution within a maximal clique (or a clique template) and a specific associated data feature.

Feature functions do not require the data to be defined in the same domain as the geological model, so geological properties in each clique may potentially be related to features in all of the data. For example, the geology may be spatial and the seismic data may be in space-time domain. If the desired data features are prohibitively large to be stored in computer memory and subsequently analyzed, their size may be reduced by using dimensionality reduction techniques such as principal component analysis (PCA, Pearson, 1901). Complex feature functions may be learnt, and the definition of the posterior distribution in equation 6.3 shows that any number of feature functions can be included in the design. Thus our method is reasonably general and may be applied to a variety of problems and many types of data.

A principal motivation of the current research was to remove two commonly used assumptions in probabilistic inversion: the localized likelihoods assumptions and the conditional independence assumption on data. This is achieved in the posterior probability model since the feature functions implicitly encode the prior distribution and the non-localized likelihoods. Our method does not require the data to be defined on a spatial grid that is the same as the geological model. Therefore we may hope to extend this method to seismic tomography and FWI type problems in future.

The proposed inversion method combines supervised machine learning with spatial inference to solve the spatial inverse problem. Spatial inference corrects inaccuracies and reduces uncertainties in the feature functions by constraining the spatial distribution of model parameters at neighbouring locations to be consistent with both the spatial priors and the non-localized likelihoods. The dimensionality of model parameters in a large clique template may be too high. This is addressed by the mean field (MF) approximation. The naive MF method is quite limited as it assumes independence of individual vertices; the quality of such an approximation is governed by the density (as opposed to sparsity), scale and strength of neglected interactions among various variables of interest. The higher-order mean field approximation defined in this chapter attempts to ameliorate the loss due to neglecting significant interactions among variables as it assumes independence of non-maximal cliques in the graph: if the size of such cliques is sufficiently large to capture the expected spatial

distributions of geological properties, mean field inference proves to be an efficient and reliable approximation in models where the posterior distribution is factorizable (e.g. in a MRF).

Any solution of the MF equations is a stationary point and is not guaranteed to be an optimum. However, in practice a MF solution is empirically known to converge to local optima in most scenarios because it is highly unlikely for a solution to get stuck at an unstable stationary point (e.g. a saddle point, Koller & Friedman, 2009). Also, it is important to note that the locally optimal solution obtained from the MF updates is not guaranteed to be the same as the globally best factorized approximation $Q$. This is because the solution depends on the initial CRF weights $\boldsymbol{w}$ and on the ordering of MF updates, both of which should usually be chosen randomly. In our experience, as long as the approximating cliques are large enough to capture the expected spatial patterns of facies, the MF algorithm converges to a consistent solution. In principle the MF equations may be solved within a global optimization framework such as simulated annealing for global optimization of the free energy functional $\mathcal{F}(Q, \boldsymbol{w})$ in equation 6.7, although we found that there was no need to do so in examples that we have tested.

In the light of above discussion, the quality of solutions from the proposed method is determined by the choice of feature functions and their accuracy, i.e. how well they relate the data and corresponding model parameters, the amount of prior information injected (defined by the maximal clique size), and how close the size of approximating cliques is to the maximal clique size. The latter factor mainly governs the computational cost of the method, and essentially defines a trade-off between accuracy and computational efficiency. The MF inference that we deploy offers a more computationally efficient method compared to McMC, however, it is worthwhile to note that McMC is a general method that is in principle applicable to any inverse problem, while mean field inference offers a reasonable approximation only in models where the true posterior distribution is factorizable (e.g. a MRF). A MRF model is used in this research because it is the most widely used model in spatial statistics (in particular geostatistics), even in most McMC-based geostatistical inversion methods (e.g. Ulvmoen & Omre, 2010; Rimstad & Omre, 2010; Luo & Tjelmeland, 2017). A fair comparison of accuracy and computational cost of McMC versus mean field inference requires such comparison to be made with respect to a given problem, i.e. under the same set of assumptions. We leave such a comparison as a topic for future research.

## *6.9 Conclusions*

We introduced a discriminative approach to Bayesian inversion of geophysical data for geological facies. This method models the posterior distribution of facies given the observed data directly using a conditional random field (CRF), as opposed to the commonly used generative-modelling based Bayesian inversion that models the posterior distribution through the joint distribution of facies and data. For problems that are decomposable into interlinked sub-problems as described herein, the presented discriminative approach thus circumvents the prohibitive amount of computational time and digital storage commonly required by the joint distribution, and allows tractable inversion in complex problems for which the conventional generative approach becomes intractable. This allowed us to add more sophistication to our model and remove the commonly used assumptions of localized likelihoods and conditional independence of data, without incurring significant computational limitations. Our proposed method incorporates spatial prior information and non-localized likelihoods, and is therefore capable of modelling complex correlations in both data and geology.

We avoided the use of stochastic sampling and introduced a higher-order mean field method for approximate inference within the variational Bayesian framework. Convergence to a local (and potentially global) optimum is guaranteed in this method. The mean field inference may be performed within a global optimization framework such as simulated annealing or genetic algorithms to encourage global convergence. In a synthetic example, we demonstrated that this method is capable of inverting seismic attributes for facies with reasonable accuracy even in the presence of strongly correlated noise.

# Chapter 7 Linearized Variational Bayesian Inversion Using a Hierarchical Model

## 7.1 Summary

This chapter introduces an efficient Bayesian inversion method based on variational inference using a hierarchical model. The hierarchical approach treats the parameters of the inverse problem as random variables that are estimated as part of the solution to the inverse problem. The variational Bayesian framework is first formulated in general terms as an iterative optimization algorithm, and then an analytical solution is derived for each update in terms of a Gaussian posterior distribution of the desired model parameters. The proposed method jointly estimates the parameters of the forward model and the noise level in the data along with the solution of the inverse problem, while providing a quantitative assessment of uncertainties in these estimates. The forward model is initially assumed to be linearized, and later a non-linear extension of this method is proposed. Since the probabilistic inference problem is performed within an optimization framework, the proposed method avoids stochastic sampling of the solution space, yet provides fully probabilistic Bayesian results more efficiently.

## 7.2 Introduction

In this chapter, a variational Bayesian inversion (VBI) method is introduced to estimate the posterior distribution of spatially coupled (probabilistically dependent) geological properties of subsurface rocks from geophysical data using a hierarchical Bayesian framework with a linearized forward model. Introducing spatial probabilistic dependence among the model parameters ensures that the solution of the inverse problem captures expected spatial correlations in the model parameters that are consistent with both the data and available geostatistical prior information. The hierarchical Bayesian framework regards all the parameters of an inverse problem including the coefficients of the forward model as random variables, all of which are estimated as a part of the solution. Thus, in contrast to the more conventional non-hierarchical approach, the hierarchical model accounts for all or most of the

uncertainty that is present in the problem. The computational efficiency of VBI comes from the fact that its solution can be given in terms of a set of *fixed point equations* (Koller & Friedman, 2009) that can be solved iteratively such that each iteration updates the parameters of the distributions in a closed form (see section 7.5.2).

The proposed method is first formulated in general terms so that it is applicable to any (geophysical) inverse problem where the forward model is linear or may be linearized, and the errors are Normally distributed. This covers a wide range of problems in geosciences and other fields of research. The solution is then derived analytically assuming that both the prior and the likelihood distributions are Gaussian. Again, this is a commonly used assumption in geophysical literature. For example, Buland & Omre (2003a) performed *amplitude variation with offset* (AVO, also called *amplitude variation with angle* – AVA) inversion of seismic data using an explicit analytical form of the posterior distribution under the Gaussian model assumption. The analytical form allows computationally efficient solution, however, their method does not allow spatial coupling of model parameters (elastic rock properties). Thus elastic properties along a vertical trace/bin location are assumed to be independent of model parameters at all other locations. Further, the parameters of the forward model (seismic wavelet in this case) and the noise covariance was also assumed to be known *a priori*. Buland & Omre (2003b) removed such assumptions and performed AVO inversion with spatially coupled parameters using a hierarchical Bayesian model. However, their method requires stochastic sampling using McMC and is therefore computationally expensive.

The method proposed in this chapter defines a more general hierarchical Bayes model and is computationally more efficient. This method is applied to an AVO inversion problem on real 2D seismic data from the North Sea. The AVO inversion from raw seismic data is a highly non-linear problem. Nonlinearity is mainly caused by the factors such as arrival time move-outs, multiple scattering and amplitude decay of seismic energy with travel distance. If such effects are removed from the seismic data during processing while preserving true reflection amplitudes, the residual nonlinearity may only be due to the intrinsic nonlinearity of seismic wave reflection that can be modelled using the Zoeppritz equations (Zoeppritz, 1919). These equations may be linearized, e.g. see Aki & Richards (Aki & Richards, 1980). This means that the processed seismic images presented in the form of partial reflection angle stacks (stacked data for only a range of reflection angles) may be suitable for AVO inversion with a linearized approximation (Buland & Omre, 2003a & 2003b). This is what we will explore in section 7.7.

The rest of this chapter is organized as follows. The Bayesian framework for probabilistic inversion is reviewed in explicit terms of parameters of the prior and likelihood distributions in section 7.3. Then the hierarchical Bayesian model and the associated hyper-prior distributions for Gaussian prior and likelihoods are introduced in section 7.4. Then, in section 7.5, the variational Bayes (VB) inference method and the associated mean field (MF) approximation is introduced for the current hierarchical Bayesian model. Until this point the discussion is kept in general terms. Then in the subsection 7.5.2, the mean field update equations are derived in a closed form for the specific case of Gaussian prior and likelihood distributions. In section 7.6, the computational cost of the proposed method is discussed. The application of this method on real data is presented in section 7.7. Finally, we discuss the implications of the method and conclude in sections 7.8 and 7.9, respectively.

## 7.3 Model

We use the so called *generative model* whereby the data $\boldsymbol{d}$ are assumed to have been generated by the model $\boldsymbol{m}$ according to the likelihood distribution $\mathcal{P}(\boldsymbol{d}|\boldsymbol{m};\Theta_{d|m})$, which is defined in terms of a set of parameters $\Theta_{d|m}$ that model the relationship between $\boldsymbol{m}$ and $\boldsymbol{d}$. Similarly, we may express the prior distribution as $\mathcal{P}(\boldsymbol{m};\Theta_m)$, where $\Theta_m$ are its parameters. The posterior distribution $\mathcal{P}(\boldsymbol{m}|\boldsymbol{d};\Theta)$ of the model $\boldsymbol{m}$ given the data $\boldsymbol{d}$ may then be expressed in terms of the parameters $\Theta \equiv \Theta_m \cup \Theta_{d|m}$, which is given by the Bayes' theorem (see equation 2.1):

$$\mathcal{P}(\boldsymbol{m}|\boldsymbol{d};\Theta) = \frac{\mathcal{P}(\boldsymbol{m},\boldsymbol{d};\Theta)}{\mathcal{P}(\boldsymbol{d};\Theta)} = \frac{\mathcal{P}(\boldsymbol{d}|\boldsymbol{m};\Theta_{d|m})\mathcal{P}(\boldsymbol{m};\Theta_m)}{\mathcal{P}(\boldsymbol{d};\Theta)} \tag{7.1}$$

where the denominator $\mathcal{P}(\boldsymbol{d};\Theta)$ is the marginal likelihood of the observed data $\boldsymbol{d}$, and acts as a normalization constant which is given by

$$\mathcal{P}(\boldsymbol{d};\Theta) = \int_{\boldsymbol{m}} \mathcal{P}(\boldsymbol{d}|\boldsymbol{m};\Theta_{d|m})\mathcal{P}(\boldsymbol{m};\Theta_m)\, d\boldsymbol{m} \tag{7.2}$$

Below, we first describe a model for the prior distribution $\mathcal{P}(\boldsymbol{m})$ of model parameters in subsection 7.3.1, then we describe a model for the likelihood $\mathcal{P}(\boldsymbol{d}|\boldsymbol{m})$ of data $\boldsymbol{d}$ given $\boldsymbol{m}$ in subsection 7.3.2, and then in section 7.4 the prior and the likelihood are extended in a hierarchical Bayesian model.

### *7.3.1 Prior Model*

A Normal distribution is commonly used in a wide range of inverse problems in geophysics (Tarantola & Valette, 1982; Buland & Omre, 2003a; Hansen *et al.* 2006; Lang & Grana, 2018) as the prior distribution over the model parameters $\boldsymbol{m}$. The correlations among various elements of $\boldsymbol{m}$ are expressed in the form of a covariance matrix $\boldsymbol{\Sigma}_m$. We use a spatial (location dependent) form of a Normal distribution, known as the *Gaussian Markov random field* (GMRF) (Rue & Held, 2005) as the prior distribution over $\boldsymbol{m}$ for its analytical attractiveness and computational advantages. A GMRF is a Normal (or Gaussian) distribution which assumes that the model parameters $\boldsymbol{m}$ satisfy the *Markovian* property whereby the model parameters $\boldsymbol{m}_i$ at a location $i$ in the model are assumed to be conditionally independent given the parameters $\boldsymbol{m}_{\mathcal{N}_{\setminus i}}$ at the neighbouring locations $\mathcal{N}_{\setminus i}$ of $i$. The conditional distribution of $\boldsymbol{m}_i$ given $\boldsymbol{m}_{\setminus i}$ (model parameters at all locations except $i$) can then be expressed as

$$\mathcal{P}\left(\boldsymbol{m}_i \mid \boldsymbol{m}_{\setminus i}\right) = \mathcal{P}\left(\boldsymbol{m}_i \mid \boldsymbol{m}_{\mathcal{N}_{\setminus i}}\right) = N\left(\sum_{j \in \mathcal{N}_{\setminus i}} \beta_{ij}\boldsymbol{m}_j\,, \lambda_i^{-1}\right) \tag{7.3}$$

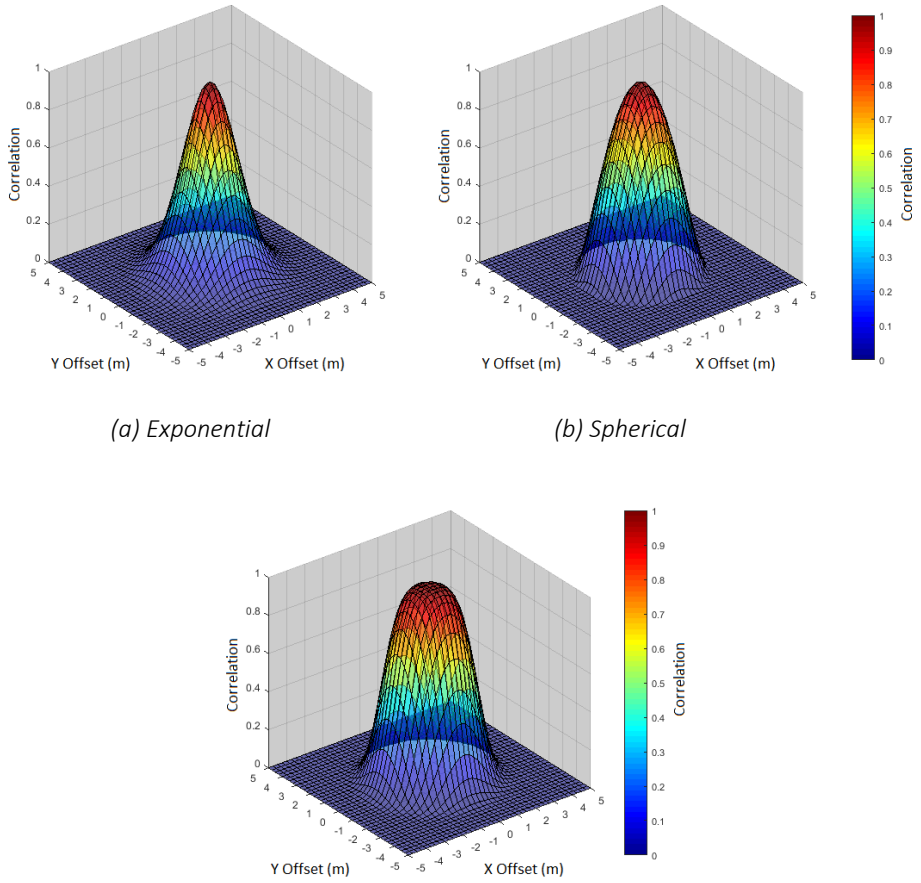
where $\lambda_i > 0, \forall i$ and $\lambda_i\beta_{ij} = \lambda_j\beta_{ji}, \forall i \neq j$. The parameter $\lambda_i$ defines the conditional precision (inverse of variance) of $\boldsymbol{m}_i$ given $\boldsymbol{m}_j, j \in \mathcal{N}_{\setminus i}$. The parameters $\beta_{ij}$ introduce spatial context in a GMRF and are generally defined as monotonically decreasing functions of displacement (or lag) between the locations $i$ and $j$. For example, $\beta_{ij}$ may be expressed in terms of a spatial covariance function $\nu(h)$ of lag $h$ between $i$ and $j$ (Buland & Omre, 2003a). Commonly used spatial covariance functions are the exponential $\nu_{\text{exp}}(h)$, spherical $\nu_{\text{sph}}(h)$ and Gaussian $\nu_{\text{gauss}}(h)$ covariance functions given by:

$$\nu_{\text{exp}}(h) = \exp\left(-\frac{3h}{d}\right) \tag{7.4}$$

$$\nu_{\text{sph}}(h) = \begin{cases} 1 - 3h/2d + h^3/2d^3 & h \leq d \\ 0 & h > d \end{cases} \tag{7.5}$$

$$\nu_{\text{gauss}}(h) = \exp\left(-\frac{3h^2}{d^2}\right) \tag{7.6}$$

where $d$ is the range parameter that defines the maximum correlation length (see figure 7.1).




*(a) Exponential* *(b) Spherical*

*(c) Gaussian*

*Figure 7.1: Illustration of the three spatial correlation functions given in equations 7.4 to 7.6.*

A customary and analytically convenient expression for a GMRF is in terms of its *precision* matrix (or the inverse of covariance matrix) $\boldsymbol{\Lambda} = (\Lambda_{ij})$ given by:

$$\Lambda_{ij} = \begin{cases} \lambda_i & i = j \\ -\lambda_i \beta_{ij} & i \neq j \end{cases} \qquad 7.7$$

The precision matrix $\boldsymbol{\Lambda}$ is a symmetric and positive definite matrix where the symmetry is ensured by the condition: $\lambda_i \beta_{ij} = \lambda_j \beta_{ji}, \forall i \neq j$, while a necessary condition for positive definiteness is that $\lambda_i > 0, \forall i$. However, positive definiteness further requires additional (and often complicated) constraints on the $\beta_{ij}$'s. A sufficient condition for positive definiteness requires that $\boldsymbol{\Lambda}$ is a diagonal dominant matrix, i.e. each diagonal entry is larger than the sum of the absolute off-diagonal entries. $\Lambda_{ij} = 0$ (or $\beta_{ij} = 0$) implies that the model parameters $\boldsymbol{m}_i$ and $\boldsymbol{m}_j$ at locations $i$ and $j \neq i$ are conditionally independent given the rest.

The model parameters $\boldsymbol{m}_i$ at $i$ are represented in vector form as it may be composed of multiple quantities (measurements) at each location. For example, the model parameters may include P-wave and S-wave velocities and density at each location in the seismic AVO inversion problem. In such a case, a stationary covariance matrix $\boldsymbol{\Sigma}_0$ may be defined that captures correlations among various components of $\boldsymbol{m}_i, \forall i$. The GMRF precision matrix $\boldsymbol{\Lambda}$ and the inverse stationary covariance matrix ${\boldsymbol{\Sigma}_0}^{-1}$ may be composed together through the Kronecker product $\otimes$ to give a $m \times m$ model precision matrix $\boldsymbol{\Lambda}_m$ (or the inverse of model covariance matrix $\boldsymbol{\Sigma}_m$) of the overall $m \times 1$ model vector $\boldsymbol{m}$ as

$$\boldsymbol{\Lambda}_m = {\boldsymbol{\Sigma}_m}^{-1} = \boldsymbol{\Lambda} \otimes \boldsymbol{\Sigma}_0^{-1} \tag{7.8}$$

which captures the spatial correlations among all elements of $\boldsymbol{m}$. In the case when the model parameters $\boldsymbol{m}$ contain a single quantity (measurement) at each location, $\boldsymbol{\Lambda}_m$ equals $\boldsymbol{\Lambda}$ scaled by the variance of $\boldsymbol{m}$. In general, since $\beta_{ij}$ is a monotonically decreasing function of lag between any two locations, model correlations may be ignored beyond a certain neighbourhood around each location. This renders $\boldsymbol{\Lambda}_m$ as a sparse symmetric block matrix.

The GMRF prior probability distribution over the model parameters $\boldsymbol{m}$ can then be expressed as a Normal distribution $\boldsymbol{m} \sim \mathrm{N}\left(\boldsymbol{\mu}_m, {\boldsymbol{\Lambda}_m}^{-1}\right)$ with mean $\boldsymbol{\mu}_m$ and precision matrix $\boldsymbol{\Lambda}_m$, which is given by the following *probability density function* (PDF):

$$\mathcal{P}(\boldsymbol{m}; \Theta_{\boldsymbol{m}}) = N\left(\boldsymbol{m}; \boldsymbol{\mu}_m, {\boldsymbol{\Lambda}_m}^{-1}\right)$$

$$= \frac{1}{(2\pi)^{n_m/2}} |\boldsymbol{\Lambda}_m|^{1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{m} - \boldsymbol{\mu}_m)^T \boldsymbol{\Lambda}_m (\boldsymbol{m} - \boldsymbol{\mu}_m)\right\} \tag{7.9}$$

where $n_m$ is the dimensionality of $\boldsymbol{m}$, and

$$\Theta_{\boldsymbol{m}} \equiv \{\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m\} \tag{7.10}$$

Note here that the prior probability of model parameters $\boldsymbol{m}$ is simply a Gaussian distribution where the spatial context is encoded within the precision matrix $\boldsymbol{\Lambda}_m$ through equation 7.8.

### *7.3.2 Likelihood Model*

The (generally nonlinear) relationship between $\boldsymbol{m} \in \mathcal{M}$ and $\boldsymbol{d} \in D$ may be expressed as a deterministic forward function $g: \mathcal{M} \rightarrow D$ that maps the model space $\mathcal{M}$ to the data space $D$. The stochasticity (i.e. the uncertainty in this relationship) may be introduced by adding stochastic errors $\boldsymbol{\epsilon}$ in the data that are not modelled by $g$. Thus we may write

$$\boldsymbol{d} = g(\boldsymbol{m}) + \boldsymbol{\epsilon} \tag{7.11}$$

The forward function $g$ captures various effects related to the data acquisition process such as the blurring effect of seismic wavelet in seismic data, and the effect of recording geometry (e.g. source-receiver offsets on seismic amplitudes). In the case when the above equation can be represented or at least approximated by a discrete convolution operation, it may be written as a system of linear equations in matrix form as

$$\boldsymbol{d} = \boldsymbol{G}\boldsymbol{m} + \boldsymbol{\epsilon} \tag{7.12}$$

where $\boldsymbol{d}$ and $\boldsymbol{m}$ are $n \times 1$ and $m \times 1$ vectors, $\boldsymbol{G}$ is a $n \times m$ *block circulant Toeplitz matrix* constructed from the coefficients of the forward function $g$ that relates $\boldsymbol{d}$ and $\boldsymbol{m}$, and $\boldsymbol{\epsilon}$ represents a $n \times 1$ vector of stochastic errors. The set of $n$ equations 7.12 represents the forward problem of predicting the data $\boldsymbol{d}$ from a given set of model parameters $\boldsymbol{m}$. The inverse problem is then to search for all possible sets of model parameters that are admissible under some required prior constraints, and minimize the residuals $\boldsymbol{\epsilon}$ to within an acceptable tolerance. Using the commutative property of a linear operation over circulant matrices, the above equation may also be represented as

$$\boldsymbol{d} = \boldsymbol{M}\boldsymbol{g} + \boldsymbol{\epsilon} \tag{7.13}$$

where $\boldsymbol{M}$ is a $n \times m$ block circulant Toeplitz matrix constructed from the model parameters $\boldsymbol{m}$, and $\boldsymbol{g}$ is a $m \times 1$ vector of coefficients of $g$. Equations 7.12 and 7.13 are both exactly equivalent as the block Toeplitz matrix $\boldsymbol{G}$ is constructed by circular displacements of the vector $\boldsymbol{g}$. We consider $\boldsymbol{g}$ as a random vector so that prior uncertainty in the coefficients of the forward model may also be acknowledged in the solution of the inverse problem. We model $\boldsymbol{g}$

using a Normal distribution $\mathrm{N}(\boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g{}^{-1})$ with mean vector $\boldsymbol{\mu}_g$ and precision matrix $\boldsymbol{\Lambda}_g$ (or covariance matrix $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g{}^{-1}$), which is given by the following PDF:

$$\mathcal{P}(\boldsymbol{g}) = N(\boldsymbol{g}; \boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g{}^{-1})$$

$$= \frac{1}{(2\pi)^{n_m/2}} |\boldsymbol{\Lambda}_g|^{1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{g} - \boldsymbol{\mu}_g)^T \boldsymbol{\Lambda}_g (\boldsymbol{g} - \boldsymbol{\mu}_g)\right\} \quad 7.14$$

The stochastic errors $\boldsymbol{\epsilon}$ in equations 7.12 and 7.13 are commonly assumed to be *independent and identically distributed* $(i.i.d.)$ according to a Normal distribution with zero mean and constant variance $\sigma^2$. Then the variance $\sigma^2$ of stochastic errors also represents the variance of the data $\boldsymbol{d}$ around their mean value $\boldsymbol{Gm}$, and describes the *signal to noise ratio* (SNR) of $\boldsymbol{d}$. The data likelihood $\mathcal{P}(\boldsymbol{d}|\boldsymbol{m})$ is therefore given by the following PDF:

$$\mathcal{P}(\boldsymbol{d}|\boldsymbol{m}; \Theta_{d|m}) = N(\boldsymbol{d}; \boldsymbol{Gm}, \boldsymbol{\Sigma}_\epsilon)$$

$$= \frac{1}{(2\pi)^{n_d/2} |\boldsymbol{\Sigma}_\epsilon|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{d} - \boldsymbol{Gm})^T \boldsymbol{\Sigma}_\epsilon^{-1} (\boldsymbol{d} - \boldsymbol{Gm})\right\} \quad 7.15$$

where $n_d$ is the dimensionality of $\boldsymbol{d}$, $\boldsymbol{\Sigma}_\epsilon = \sigma^2 \boldsymbol{I}_{n\times n}$ is the covariance matrix of errors $\boldsymbol{\epsilon}$ and $\Theta_{d|m} = \{\boldsymbol{g}, \boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g, \boldsymbol{\Sigma}_\epsilon\}$ are the parameters of the likelihood function.

In many geophysical applications, the data $\boldsymbol{d}$ may be composed of multiple number, say $n_a$, of data vectors $\boldsymbol{d}_i$ each of size $n_s \times 1$ such that $\boldsymbol{d} = [\boldsymbol{d}_1, \boldsymbol{d}_2, \dots, \boldsymbol{d}_{n_a}]^T$. Examples of such case are seismic AVO inversion of partial-angle stacks (Buland & Omre, 2003a; Lang & Grana, 2018), where $n_a$ represents the number of AVO angles to be inverted and $n_s$ represents the number of samples in each trace. For clarity, we refer to the vectors $\boldsymbol{d}_i$ as the 'components' of $\boldsymbol{d}$, and each scalar element $d_i$ of $\boldsymbol{d}$ as the 'elements' of $\boldsymbol{d}$. For the sake of generality, we allow each component $\boldsymbol{d}_i$ to have a different variance ${\sigma_i}^2$. This allows capturing differences in SNR across various components of $\boldsymbol{d}$. Thus, the data covariance matrix $\boldsymbol{\Sigma}_\epsilon$ takes a block diagonal form $\boldsymbol{\Sigma}_\epsilon = \mathrm{diag}({\sigma_1}^2, \dots, {\sigma_{n_a}}^2) \otimes \boldsymbol{I}_{n_s}$, where ${\sigma_i}^2$ represents the variance of $\boldsymbol{d}_i$. Note that the constant variance assumption is a special case of this block diagonal form where ${\sigma_i}^2 = \sigma^2, \forall i$.

For analytical convenience, it is customary to work with the precision matrix $\boldsymbol{\Lambda}_\epsilon$ which is the inverse of the covariance matrix $\boldsymbol{\Sigma}_\epsilon$, i.e. $\boldsymbol{\Lambda}_\epsilon = \boldsymbol{\Sigma}_\epsilon^{-1}$. Since $\boldsymbol{\Sigma}_\epsilon$ is a diagonal matrix, therefore $\boldsymbol{\Lambda}_\epsilon$ is also a diagonal matrix given by $\boldsymbol{\Lambda}_\epsilon = \mathrm{diag}(\lambda_1, \dots, \lambda_{n_a}) \otimes \boldsymbol{I}_{n_s}$, where $\lambda_i = 1/\sigma_i^2$. So we can write:

$$\mathcal{P}(\boldsymbol{d}|\boldsymbol{m}; \Theta_{d|m}) = N(\boldsymbol{d}; \boldsymbol{Gm}, \boldsymbol{\Lambda}_\epsilon^{-1})$$

$$= \frac{1}{(2\pi)^{n_d/2}} |\boldsymbol{\Lambda}_\epsilon|^{1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{d} - \boldsymbol{Gm})^T \boldsymbol{\Lambda}_\epsilon (\boldsymbol{d} - \boldsymbol{Gm})\right\} \quad 7.16$$

Thus, the likelihood parameters may be expressed as:

$$\Theta_{d|m} = \{\boldsymbol{g}, \boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g, \boldsymbol{\Lambda}_\epsilon\} \quad 7.17$$

## 7.4 Hierarchical Bayesian Model

Analytical solutions of the posterior distribution are commonly sought by assuming that the parameters $\Theta$ are fixed and are known *a priori*. Tarantola & Valette (1982) and Mosegaard & Tarantola (2002) showed that the solution for a linear inverse problem with Gaussian prior and likelihood yields a Gaussian posterior distribution with mean $\boldsymbol{\mu}_{m|d}$ and covariance matrix $\boldsymbol{\Sigma}_{m|d}$ that can be computed in a closed form as

$$\boldsymbol{\mu}_{m|d} = \widetilde{\boldsymbol{\mu}}_m + \widetilde{\boldsymbol{\Sigma}}_m \widetilde{\boldsymbol{G}}^T \left(\widetilde{\boldsymbol{G}} \widetilde{\boldsymbol{\Sigma}}_m \widetilde{\boldsymbol{G}}^T + \widetilde{\boldsymbol{\Sigma}}_\epsilon\right)^{-1} \left(\boldsymbol{d} - \widetilde{\boldsymbol{G}} \widetilde{\boldsymbol{\mu}}_m\right) \quad 7.18$$

$$\boldsymbol{\Sigma}_{m|d} = \widetilde{\boldsymbol{\Sigma}}_m - \widetilde{\boldsymbol{\Sigma}}_m \widetilde{\boldsymbol{G}}^T \left(\widetilde{\boldsymbol{G}} \widetilde{\boldsymbol{\Sigma}}_m \widetilde{\boldsymbol{G}}^T + \widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}\right)^{-1} \widetilde{\boldsymbol{G}} \widetilde{\boldsymbol{\Sigma}}_m \quad 7.19$$

where the forward linear operator $\widetilde{\boldsymbol{G}}$, prior mean $\widetilde{\boldsymbol{\mu}}_m$, prior covariance matrix $\widetilde{\boldsymbol{\Sigma}}_m$, and the covariance matrix of data errors $\widetilde{\boldsymbol{\Sigma}}_\epsilon$ are assumed to be fixed and known *a priori*. These equations have found wide applicability in solving linear Gaussian inverse problems in geophysics while avoiding Monte Carlo sampling (e.g. Buland *et al.* 2003; Buland & Omre, 2003a; Hansen *et al.* 2006; Lang & Grana, 2018).

In this chapter, we use the fully Bayesian approach that regards the parameters $\Theta$ as random variables with prior distributions expressed in terms of hyper-priors – the distribution of the parameters of the prior distributions. The parameters $\Theta$ are therefore not assumed to be fixed and known with certainty *a priori*; they are estimated within the solution of the inverse problem. The prior information about the unknown parameters $\boldsymbol{m}$ and $\Theta$ incorporated through their prior distributions is expected to alleviate the ill-posedness of the inverse problem and account for uncertainty in all of the parameters involved. Thus, a hierarchical approach allows a more data-adaptive estimation of the model parameters, which are otherwise assumed to be accurately known *a priori* in a non-hierarchical formulation. Bayes' theorem (equation 7.1) can then be written as

$$\mathcal{P}(\boldsymbol{m}, \Theta|\boldsymbol{d}) = \frac{\mathcal{P}(\boldsymbol{m}, \boldsymbol{d}, \Theta)}{\mathcal{P}(\boldsymbol{d})} = \frac{\mathcal{P}(\boldsymbol{d}|\boldsymbol{m}, \Theta_{d|m})\mathcal{P}(\boldsymbol{m}|\Theta_m)\mathcal{P}(\Theta_{\boldsymbol{m}}, \Theta_{d|m})}{\mathcal{P}(\boldsymbol{d})} \tag{7.20}$$

where the denominator $\mathcal{P}(\boldsymbol{d})$ represents the marginal likelihood of the observed data $\boldsymbol{d}$, and acts as a normalization constant. It is now given by

$$\mathcal{P}(\boldsymbol{d}) = \iiint \mathcal{P}(\boldsymbol{d}|\boldsymbol{m}, \Theta_{d|m})\mathcal{P}(\boldsymbol{m}|\Theta_m)\mathcal{P}(\Theta_m, \Theta_{d|m})\, d\boldsymbol{m}\, d\Theta_m\, d\Theta_{d|m} \tag{7.21}$$

Note that the parameters $\Theta$ are also regarded here as random variables. This is what differentiates the approach used in the rest of this thesis to the hierarchical Bayesian approach used in this chapter. However, in comparison to the model parameters $\boldsymbol{m}$, parameters $\Theta$ are regarded as nuisance parameters which are not of primary interest, but these must still be accounted for in the analysis of $\boldsymbol{m}$.

The posterior distribution $\mathcal{P}(\boldsymbol{m}, \Theta|\boldsymbol{d})$ constitutes the complete solution of an inverse problem and describes the associated uncertainties.

### *7.4.1 Hyper-priors*

Within a hierarchical modelling framework the parameters $\Theta_m = \{\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m\}$ and $\Theta_{d|m} = \{\boldsymbol{g}, \boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g, \boldsymbol{\Lambda}_{\boldsymbol{\epsilon}}\}$ of the prior and likelihood PDFs are themselves defined in terms of hyper-priors. For the joint prior distribution $\mathcal{P}(\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m)$ over the model expectation $\boldsymbol{\mu}_m$ and the model precision matrix $\boldsymbol{\Lambda}_m$ we use a *Normal-Wishart* ($NW$) distribution (Bishop, 2006), such that

$$\mathcal{P}(\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m) = \mathcal{P}(\boldsymbol{\mu}_m|\boldsymbol{\Lambda}_m)\mathcal{P}(\boldsymbol{\Lambda}_m) = N(\boldsymbol{\mu}_m|\boldsymbol{\Lambda}_m; \boldsymbol{\pi}_m, (\tau_m\boldsymbol{\Lambda}_m)^{-1})\, W(\boldsymbol{\Lambda}_m; \boldsymbol{W}_\Lambda, v_\Lambda) \tag{7.22}$$

Here, $\boldsymbol{\mu}_m|\boldsymbol{\Lambda}_m \sim N(\boldsymbol{\pi}_m, (\tau_m\boldsymbol{\Lambda}_m)^{-1})$ represents a Normal distribution with ($m$-vector) mean $\boldsymbol{\pi}_m$ and $m \times m$ covariance matrix $(\tau_m\boldsymbol{\Lambda}_m)^{-1}$ with $\tau_m$ being the scale of the precision matrix $\boldsymbol{\Lambda}_m$, given by the following PDF

$$N(\boldsymbol{\mu}_m|\boldsymbol{\Lambda}_m; \boldsymbol{\pi}_m, (\tau_m\boldsymbol{\Lambda}_m)^{-1}) = (2\pi)^{-n_d/2}|\tau_m\boldsymbol{\Lambda}_m|^{1/2} \exp\left\{-\frac{\tau_m}{2}(\boldsymbol{\mu}_m - \boldsymbol{\pi}_m)^T\boldsymbol{\Lambda}_m(\boldsymbol{\mu}_m - \boldsymbol{\pi}_m)\right\} \tag{7.23}$$

and $\boldsymbol{\Lambda}_m \sim W(\boldsymbol{W}_\Lambda, v_\Lambda)$ represents a Wishart distribution with $n_d \times n_d$ scale matrix $\boldsymbol{W}_\Lambda$ and the number of degrees of freedom $v_\Lambda$, given by the following PDF

$$W(\boldsymbol{\Lambda}_m; \boldsymbol{W}_\Lambda, v_\Lambda) = B(\boldsymbol{W}_\Lambda, v_\Lambda)|\boldsymbol{\Lambda}_m|^{(v_\Lambda - m - 1)/2} \exp\left\{-\frac{1}{2}Tr\left[\boldsymbol{W}_m{}^{-1}\boldsymbol{\Lambda}_m\right]\right\} \tag{7.24}$$

where $\mathrm{Tr}[.]$ represents the trace operator of a matrix that represents the sum of the diagonal elements of its argument matrix, and

$$B(\boldsymbol{W}_\Lambda, v_\Lambda) = |\boldsymbol{W}_\Lambda|^{-v_\Lambda/2}\left(2^{(mv_\Lambda)/2}\,\pi^{m(m-1)/4}\prod_{i=1}^{m}\Gamma\left(\frac{v_\Lambda + 1 - i}{2}\right)\right)^{-1} \tag{7.25}$$

where $\Gamma$ is the *gamma function* (Davis, 1959) , defined as

$$\Gamma(x) = \int_0^\infty x^{t-1}\, e^{-x} dx \tag{7.26}$$

The Normal-Wishart distribution is the conjugate prior of the multivariate Normal distribution with unknown mean and precision matrix (Bishop, 2006). Therefore, its choice for the joint prior distribution $\mathcal{P}(\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m)$ allows convenient analytical derivation of the posterior distribution in a hierarchical Bayesian formulation and results in a multivariate Normal posterior distribution of model parameters $\boldsymbol{m}$ (see section 7.5.2). Besides its analytical convenience, the Normal-Wishart distribution is also an intuitive model for the joint prior distribution of the mean and precision matrix. This is because for a given precision matrix $\boldsymbol{\Lambda}_m$, the Normal distribution penalizes a value of model expectation $\boldsymbol{\mu}_m$ that is too different from its expectation $\boldsymbol{\pi}_m$ and therefore ensures that any updates in $\boldsymbol{\mu}_m$ during the course of

inversion are regularized. Similarly, the Wishart distribution ensures that the elements of the $\boldsymbol{\Lambda}_m$ remain within their expected scales as imposed by the prior information during inversion.

The joint prior distribution $\mathcal{P}(\boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g)$ over the expectation $\boldsymbol{\mu}_g$ and the precision matrix $\boldsymbol{\Lambda}_g$ of the linearized coefficients $\boldsymbol{g}$ of the forward model $g$ is also modelled using a Normal-Wishart distribution, given by

$$\mathcal{P}(\boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g) = \mathcal{P}(\boldsymbol{\mu}_g|\boldsymbol{\Lambda}_g)\mathcal{P}(\boldsymbol{\Lambda}_g) = N\left(\boldsymbol{\mu}_g|\boldsymbol{\Lambda}_g; \boldsymbol{\pi}_g, (\tau_g \boldsymbol{\Lambda}_g)^{-1}\right) W(\boldsymbol{\Lambda}_g; \boldsymbol{W}_g, v_g) \quad 7.27$$

which has the same form as for $\mathcal{P}(\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m)$ given in equation 7.22, and the PDFs and hyper-parameters involved are defined accordingly as in equations 7.23 to 7.25 above.

We use a *Gamma distribution* $G$ as the hyper-prior over each component $\lambda_i$ of the data precision matrix $\boldsymbol{\Lambda}_\epsilon$, such that $\lambda_i \sim G(a_i, b_i)$ where $a_i$ and $b_i$ are the so called *shape* and *rate* parameters. The PDF of the Gamma distribution is given by

$$\lambda_i \sim G(a_i, b_i) = \frac{{b_i}^{a_i} \, {\lambda_i}^{a_i - 1} \exp\{-b_i \lambda_i\}}{\Gamma(a_i)} \quad 7.28$$

The expected value of a Gamma distribution is given by $a/b$. Each component $\lambda_i$ of $\boldsymbol{\Lambda}_\epsilon$ is assumed to be independent, so their joint distribution may be expressed as

$$\mathcal{P}(\boldsymbol{\Lambda}_\epsilon) = \prod_i \mathcal{G}(\lambda_i; a_i, b_i) = \prod_i \frac{{b_i}^{a_i} \, {\lambda_i}^{a_i - 1} \exp\{-b_i \lambda_i\}}{\Gamma(a_i)} \quad 7.29$$

### *7.4.2 Graphical Representation of Hierarchical Bayesian Model*

The probabilistic dependence among all of the random variables in the hierarchical model may be depicted in the form of a *directed acyclic graph* (DAG), also known as a *Bayesian network*, (see e.g. Koller & Friedman, 2009) as shown in figure 7.2.

The circular nodes in the graph represent random variables, and square nodes represent fixed parameters which are assumed to be known *a priori*. The directed edges (arrows) represent causal probabilistic dependence between the connected nodes such that the node at the head of the arrow (called the *child node*) depends on the node on its tail (called the *parent node*). Each circular node (random variable) is associated with a PDF which is

defined in terms of the nodes' parent variables. The node corresponding to the data $\boldsymbol{d}$ is coloured to represent that it is observed while all other random variables are unobserved. Since only the variables shown as square nodes are assumed to be known *a priori*, and the data only depends indirectly on these variables through the unknown random parameters of the model, the inversion is more data adaptive and is less influenced by inaccuracies in the prior information.

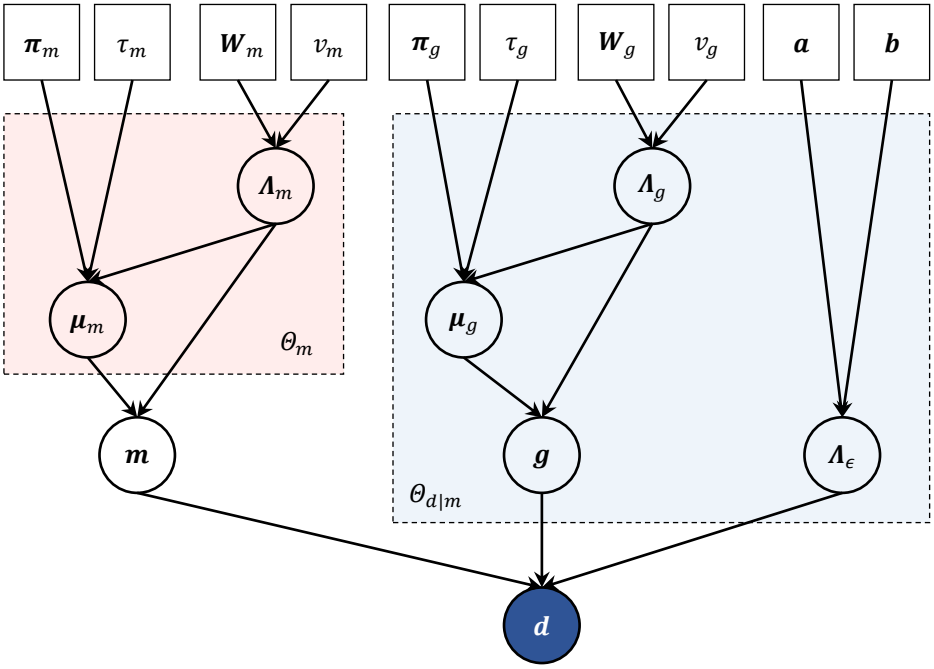



*Figure 7.2: Probabilistic dependence among various variables in the hierarchical model is expressed in the form of a directed acyclic graph (DAG). Circular nodes represent random variables and square nodes represent fixed parameters that are assumed to be known a priori. The data node represented by the variable* $\boldsymbol{d}$ *is coloured to reflect that it is observed.*

## 7.5 Variational Bayesian Inference

We use the variational Bayesian (VB) method for approximate inference, where the true posterior distribution $\mathcal{P}(\boldsymbol{m}, \Theta|\boldsymbol{d})$ is approximated by a variational distribution $\mathcal{Q}(\boldsymbol{m}, \Theta) \in \mathbb{Q}$ chosen from a family $\mathbb{Q}$ of distributions, by minimizing the KL divergence $KL\big(\mathcal{Q}(\boldsymbol{m}, \Theta)||\mathcal{P}(\boldsymbol{m}, \Theta|\boldsymbol{d})\big)$, or simply $KL(\mathcal{Q}||\mathcal{P})$ (see equation 2.10). Minimizing $KL(\mathcal{Q}||\mathcal{P})$ is equivalent to maximizing the variational free energy $\mathcal{F}(\mathcal{Q})$ as given by equation 2.9.

As we saw in section 5.4, the variational free energy $\mathcal{F}(\mathcal{Q})$ may be maximized by using the *expectation-maximization* algorithm (Dempster *et al.* 1977; Neal & Hinton, 1999; Beal, 2003; Nawaz & Curtis, 2018) for a given set of parameters $\Theta$. However, since $\Theta$ is considered unknown in a hierarchical model and must be estimated within the solution to the inverse problem, we use an alternative *mean field (MF) approximation* (Feynman, 1972; Jaakkola, 1997; Opper & Saad, 2001; Nawaz & Curtis, 2019), which is similar in concept but different in formulation to the MF approximation used in section 6.4.1. Here, the variational distribution $\mathcal{Q}(\boldsymbol{m}, \Theta)$ is chosen from a factorized family $\mathbb{Q}$ of distributions under the MF approximation. This corresponds to conditional independence assumptions over at least some of the parameters of interest given the observed data. This is described in the subsection below.

## *7.5.1 Mean Field (MF) Approximation*

We are interested in obtaining the joint posterior distribution $\mathcal{P}(\boldsymbol{m}, \Theta|\boldsymbol{d})$ over $\boldsymbol{m}$ and $\Theta$ after having observed the data $\boldsymbol{d}$, which are given by equation 7.20. Since the joint estimation of $\boldsymbol{m}$ and $\Theta$ is a nonlinear inverse problem, a closed form solution for the posterior distribution is not possible under the minimum set of assumptions we make below. Nevertheless, a set of update equations may be derived in a closed form under the variational approximation, which can be solved iteratively to estimate the desired posterior distribution. We assume factorization of the posterior distribution as below

$$\mathcal{P}(\boldsymbol{m}, \Theta|\boldsymbol{d}) = \mathcal{P}(\boldsymbol{m}|\Theta, \boldsymbol{d})\mathcal{P}(\Theta|\boldsymbol{d})$$

$$\cong \mathcal{Q}(\boldsymbol{m}, \Theta)$$

$$\equiv \mathcal{Q}(\boldsymbol{m})\mathcal{Q}(\Theta_m)\mathcal{Q}\left(\Theta_{d|m}\right) \qquad 7.30$$

where $\mathcal{P}(\boldsymbol{m}, \Theta|\boldsymbol{d})$ is approximated by $\mathcal{Q}(\boldsymbol{m}, \Theta)$, which factorizes as in equation 7.3. Such an approximation may be obtained in a closed form in terms of fixed point equations that minimize $KL(\mathcal{Q} \,||\, \mathcal{P})$ by forming the Lagrangian $L(\boldsymbol{m}, \Theta, \gamma_1, \gamma_2)$ subject to the normalization constraints for $\mathcal{Q}(\boldsymbol{m})$ and $\mathcal{Q}(\Theta)$ as

$$L(\boldsymbol{m},\Theta,\gamma_1,\gamma_2) = KL(\mathcal{Q}\,||\,\mathcal{P}) + \gamma_1\left(\int \mathcal{Q}(\boldsymbol{m})\,d\boldsymbol{m} - 1\right) + \gamma_2\left(\int \mathcal{Q}(\Theta)\,d\Theta - 1\right)$$

$$= -\iint \mathcal{Q}(\boldsymbol{m})\mathcal{Q}(\Theta)\log\frac{\mathcal{P}(\boldsymbol{m},\Theta|\boldsymbol{d})}{\mathcal{Q}(\boldsymbol{m})\mathcal{Q}(\Theta)}\,d\boldsymbol{m}\,d\Theta + \gamma_1\left(\int \mathcal{Q}(\boldsymbol{m})\,d\boldsymbol{m} - 1\right) + \gamma_2\left(\int \mathcal{Q}(\Theta)\,d\Theta - 1\right) \quad 7.31$$

where $\gamma_1$ and $\gamma_2$ are the Lagrange parameters. Taking partial derivatives of $L(\boldsymbol{m},\Theta,\gamma_1,\gamma_2)$ with respect to $\mathcal{Q}(\boldsymbol{m})$ and $\mathcal{Q}(\Theta)$ and setting the results equal to zero gives the update equations for each of these distributions as

$$\mathcal{Q}(\boldsymbol{m}) = k(\Theta)\exp\{\,\mathbb{E}_{\mathcal{Q}(\Theta)}[\log\mathcal{P}(\boldsymbol{m},\Theta,\boldsymbol{d})]\,\} \quad 7.32$$

$$\mathcal{Q}(\Theta) = k(\boldsymbol{m})\exp\{\,\mathbb{E}_{\mathcal{Q}(\boldsymbol{m})}[\log\mathcal{P}(\boldsymbol{m},\Theta,\boldsymbol{d})]\,\} \quad 7.33$$

where $k(\Theta)$ and $k(\boldsymbol{m})$ are factors constant in the variable being updated, and $\mathbb{E}_{\mathcal{Q}}[\cdot]$ represents expectation of its argument with respect to $\mathcal{Q}$. For example, $\mathbb{E}_{\mathcal{Q}(\Theta)}[\cdot]$ may be expressed as

$$\mathbb{E}_{\mathcal{Q}(\Theta)}[\log\mathcal{P}(\boldsymbol{m},\Theta,\boldsymbol{d})] = \int \mathcal{Q}(\Theta)\log\mathcal{P}(\boldsymbol{m},\Theta,\boldsymbol{d})\,d\Theta \quad 7.34$$

For a complete analytical derivation of equation 7.32 see Appendix B. Equation 7.33 may be obtained in a similar manner.

Equations 7.32 and 7.33 form the essence of the MF inference and show complex inter-dependence of the random variables involved. For example, the approximate posterior distribution $\mathcal{Q}(\boldsymbol{m})$ of $\boldsymbol{m}$ requires expectation of the joint distribution $\mathcal{P}(\boldsymbol{m},\Theta,\boldsymbol{d})$ with respect to approximate distribution $\mathcal{Q}(\Theta)$ of the rest of the parameters, while the approximate posterior distribution of $\mathcal{Q}(\Theta)$ requires expectation of $\mathcal{P}(\boldsymbol{m},\Theta,\boldsymbol{d})$ with respect to $\mathcal{Q}(\boldsymbol{m})$. Such interdependence of the unknown approximate distributions makes this system of equations nonlinear and therefore an iterative scheme is required to update these distributions in order to maximize the variational free energy $\mathcal{F}(\mathcal{Q})$, or to minimize $KL(\mathcal{Q}||\mathcal{P})$, until convergence within some predefined tolerance is achieved. The MF update equations 7.32 and 7.33 are

solved to achieve this in a coordinate ascent manner as shown by a schematic illustration in figure 7.3.

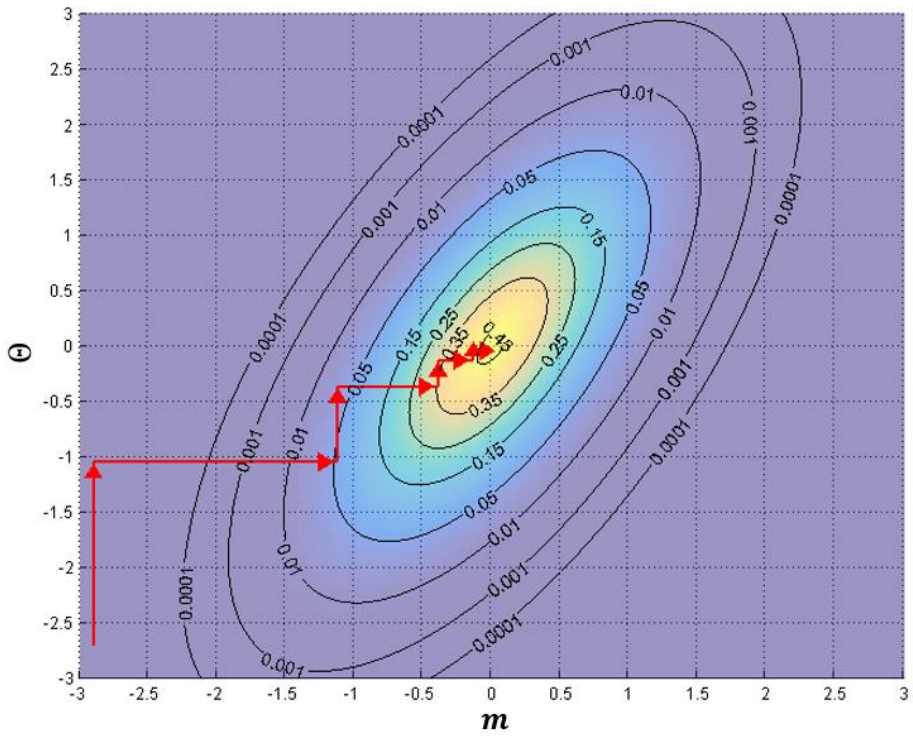



*Figure 7.3: A schematic illustration of mean field (MF) update equations for two variables* $\boldsymbol{m}$ *and* $\Theta$*, which are updated in a coordinate ascent manner in order to maximize the variational free energy* $\mathcal{F}(Q)$ *which is a functional of the approximate distribution* $Q(\boldsymbol{m}, \Theta)$*. Each update involves two steps in this case corresponding to the two variables: in the first step* $\mathcal{F}(Q)$ *is maximized with respect to* $\Theta$ *only (shown as vertical red arrows) and in the second step* $\mathcal{F}(Q)$ *is maximized with respect to* $\boldsymbol{m}$ *only (shown as horizontal red arrows). So one iteration of the maximization algorithm is represented by consecutive vertical and horizontal arrows. At the point of convergence,* $\mathcal{F}(Q)$ *is maximized, or equivalently the relative entropy* $KL(Q \,||\, \mathcal{P})$ *between the approximate posterior distribution* $Q(\boldsymbol{m}, \Theta)$ *and the true posterior distribution* $\mathcal{P}(\boldsymbol{m}, \Theta|\boldsymbol{d})$ *is minimized, within some predefined tolerance.*

Depending on the functional form of $\mathcal{P}(\boldsymbol{m}, \Theta, \boldsymbol{d})$ and the degree of non-linearity of the forward model $\boldsymbol{d} = g(\boldsymbol{m}, \Theta)$, analytical equations or numerical algorithms may be devised for minimization of $\mathrm{KL}(Q \,||\, \mathcal{P})$. Up until this point the MF formulation is quite general and is applicable to any probability distributions of interest. For further discussion and general treatment of MF update equations see e.g. Opper & Saad (2001). In the subsection below, we focus on a linear Gaussian inverse problem and derive closed form solutions of the MF update

equations 7.32 and 7.33 analytically using the functional form of the full joint distribution $\mathcal{P}(\boldsymbol{m}, \Theta, \boldsymbol{d})$ for Normally distributed prior $\mathcal{P}(\boldsymbol{m}|\Theta_m)$ and likelihood $\mathcal{P}\big(\boldsymbol{d}|\boldsymbol{m}, \Theta_{d|m}\big)$ as defined in equations 7.9 and 7.16, respectively (except that $\Theta_m$ and $\Theta_{d|m}$ are now regarded as random variables).

### *7.5.2 Analytical Derivation of MF Equations for Gaussian Distributions*

We first cast the above mean field (MF) formulation explicitly in terms of the parameters $\Theta = \Theta_m \cup \Theta_{d|m}$, where $\Theta_m = \{\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m\}$ define the prior distribution in equation 7.9, and $\Theta_{d|m} = \big\{\boldsymbol{g}, \boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g, \boldsymbol{\Lambda}_\epsilon\big\}$ define the likelihood in equation 7.16, respectively. The desired posterior distribution in equation 7.30 may then be written as

$$\mathcal{P}(\boldsymbol{m}, \Theta|\boldsymbol{d}) \cong \mathcal{Q}(\boldsymbol{m})\mathcal{Q}(\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m)\mathcal{Q}(\boldsymbol{g})\mathcal{Q}\big(\boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g\big)\mathcal{Q}(\boldsymbol{\Lambda}_\epsilon) \tag{7.35}$$

The MF update equations 7.32 and 7.33 may then be expressed for each of the random variables involved as

$$\mathcal{Q}(\boldsymbol{m}) = k(\Theta)\exp\big\{\,\mathbb{E}_{\mathcal{Q}(\Theta)}[\log\mathcal{P}(\boldsymbol{m}, \Theta, \boldsymbol{d})]\,\big\} \tag{7.36}$$

$$\mathcal{Q}(\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m) = k\big(\boldsymbol{m}, \Theta_{\setminus\boldsymbol{\mu}_m, \setminus\boldsymbol{\Lambda}_m}\big)\exp\Big\{\,\mathbb{E}_{\mathcal{Q}(\boldsymbol{m}, \Theta_{\setminus\boldsymbol{\mu}_m, \setminus\boldsymbol{\Lambda}_m})}[\log\mathcal{P}(\boldsymbol{m}, \Theta, \boldsymbol{d})]\,\Big\} \tag{7.37}$$

$$\mathcal{Q}(\boldsymbol{g}) = k\big(\boldsymbol{m}, \Theta_{\setminus g}\big)\exp\Big\{\,\mathbb{E}_{\mathcal{Q}(\boldsymbol{m}, \Theta_{\setminus g})}[\log\mathcal{P}(\boldsymbol{m}, \Theta, \boldsymbol{d})]\,\Big\} \tag{7.38}$$

$$\mathcal{Q}\big(\boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g\big) = k\left(\boldsymbol{m}, \Theta_{\setminus\boldsymbol{\mu}_g, \setminus\boldsymbol{\Lambda}_g}\right)\exp\Big\{\,\mathbb{E}_{\mathcal{Q}\left(\boldsymbol{m}, \Theta_{\setminus\boldsymbol{\mu}_g, \setminus\boldsymbol{\Lambda}_g}\right)}[\log\mathcal{P}(\boldsymbol{m}, \Theta, \boldsymbol{d})]\,\Big\} \tag{7.39}$$

$$\mathcal{Q}(\boldsymbol{\Lambda}_\epsilon) = k\big(\boldsymbol{m}, \Theta_{\setminus\boldsymbol{\Lambda}_\epsilon}\big)\exp\Big\{\,\mathbb{E}_{\mathcal{Q}(\boldsymbol{m}, \Theta_{\setminus\boldsymbol{\Lambda}_\epsilon})}[\log\mathcal{P}(\boldsymbol{m}, \Theta, \boldsymbol{d})]\,\Big\} \tag{7.40}$$

where $k$'s are functions of the unknown random variables except the variable being updated, and expectations $\mathbb{E}_{\mathcal{Q}(\boldsymbol{x})}[\cdot]$ are defined similar to equation 7.34 obtained by integrating all of the

unknown random variables out of $\log \mathcal{P}(\boldsymbol{m}, \Theta, \boldsymbol{d})$ except the variable being updated. For example $\mathbb{E}_{\mathcal{Q}(\Theta)}[\log \mathcal{P}(\boldsymbol{m}, \Theta, \boldsymbol{d})]$ in equation 7.36 is given by

$$\mathbb{E}_{\mathcal{Q}(\boldsymbol{m})}\left[\log \mathcal{P}\left(\boldsymbol{m}, \boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m, \boldsymbol{g}, \boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g, \boldsymbol{\Lambda}_\epsilon, \boldsymbol{d}\right)\right]$$

$$= \int\int\int\int\int\int \mathcal{Q}(\boldsymbol{m})\mathcal{Q}(\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m)\mathcal{Q}(\boldsymbol{g})\mathcal{Q}\left(\boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g\right)\mathcal{Q}(\boldsymbol{\Lambda}_\epsilon) \qquad 7.41$$

$$\log \mathcal{P}\left(\boldsymbol{m}, \boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m, \boldsymbol{g}, \boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g, \boldsymbol{\Lambda}_\epsilon, \boldsymbol{d}\right) d\boldsymbol{\mu}_m \, d\boldsymbol{\Lambda}_m \, d\boldsymbol{g} \, d\boldsymbol{\mu}_g d\boldsymbol{\Lambda}_g d\boldsymbol{\Lambda}_\epsilon$$

We seek a closed form solution for each of the MF update equations 7.36 to 7.40 by solving these analytically. This requires a functional form of the full joint distribution $\mathcal{P}(\boldsymbol{m}, \Theta, \boldsymbol{d})$ that may be expressed in terms of the Normal PDFs for the priors $\mathcal{P}(\boldsymbol{m}|\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m)$ and $\mathcal{P}\left(\boldsymbol{g}|\boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g\right)$ and the likelihood $\mathcal{P}(\boldsymbol{d}|\boldsymbol{m}, \boldsymbol{\Lambda}_\epsilon)$ (see sections 7.3.1 and 7.3.2), and the hyper-prior distributions as Normal-Wishart PDF for $\mathcal{P}(\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m)$ and $\mathcal{P}\left(\boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g\right)$, and Gamma PDF for $\mathcal{P}(\boldsymbol{\Lambda}_\epsilon)$ (see section 7.4.1), as follows:

$$\log \mathcal{P}(\boldsymbol{m}, \Theta, \boldsymbol{d}) = \log \mathcal{P}\left(\boldsymbol{m}, \boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m, \boldsymbol{g}, \boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g, \boldsymbol{\Lambda}_\epsilon, \boldsymbol{d}\right)$$

$$= \log \mathcal{P}(\boldsymbol{d}|\boldsymbol{m}, \boldsymbol{\Lambda}_\epsilon) + \log \mathcal{P}(\boldsymbol{m}|\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m) + \log \mathcal{P}(\boldsymbol{\mu}_m|\boldsymbol{\Lambda}_m)$$

$$+ \log \mathcal{P}(\boldsymbol{\Lambda}_m) + \log \mathcal{P}(\boldsymbol{g}) + \log \mathcal{P}(\boldsymbol{\Lambda}_\epsilon)$$

$$= \log N\left(\boldsymbol{d}; \boldsymbol{G}\boldsymbol{m}, {\boldsymbol{\Lambda}_\epsilon}^{-1}\right) + \log N\left(\boldsymbol{m}; \boldsymbol{\mu}_m, {\boldsymbol{\Lambda}_m}^{-1}\right) + \log N(\boldsymbol{\mu}_m|\boldsymbol{\Lambda}_m; \boldsymbol{\pi}_m, (\tau_m \boldsymbol{\Lambda}_m)^{-1})$$

$$+ \log W(\boldsymbol{\Lambda}_m; \boldsymbol{W}_m, v_m) + \log N\left(\boldsymbol{g}; \boldsymbol{\mu}_g, {\boldsymbol{\Lambda}_g}^{-1}\right)$$

$$+ \log N\left(\boldsymbol{\mu}_g|\boldsymbol{\Lambda}_g; \boldsymbol{\pi}_g, \left(\tau_g \boldsymbol{\Lambda}_g\right)^{-1}\right) + \log W\left(\boldsymbol{\Lambda}_g; \boldsymbol{W}_g, v_g\right)$$

$$+ \sum_{i=1}^{n_a} \log G(\lambda_i; a_i, b_i) \qquad 7.42$$

[using equations 7.9, 7.14, 7.16, 7.22, 7.27 and 7.29]

$$= \frac{1}{2}\log|\boldsymbol{\Lambda}_\epsilon| - \frac{1}{2}(\boldsymbol{d}-\boldsymbol{G}\boldsymbol{m})^T\boldsymbol{\Lambda}_\epsilon(\boldsymbol{d}-\boldsymbol{G}\boldsymbol{m})$$

$$+\frac{1}{2}\log|\boldsymbol{\Lambda}_m| - \frac{1}{2}(\boldsymbol{m}-\boldsymbol{\mu}_m)^T\boldsymbol{\Lambda}_m(\boldsymbol{m}-\boldsymbol{\mu}_m)$$

$$+\frac{m}{2}\log\tau_m + \frac{1}{2}\log|\boldsymbol{\Lambda}_m| - \frac{\tau_m}{2}(\boldsymbol{\mu}_m-\boldsymbol{\pi}_m)^T\boldsymbol{\Lambda}_m(\boldsymbol{\mu}_m-\boldsymbol{\pi}_m)$$

$$+\frac{(v_m-m-1)}{2}\log|\boldsymbol{\Lambda}_m| - \frac{1}{2}Tr\left[\boldsymbol{W}_\Lambda{}^{-1}\boldsymbol{\Lambda}_m\right]$$

$$+\frac{1}{2}\log|\boldsymbol{\Lambda}_g| - \frac{1}{2}\left(\boldsymbol{g}-\boldsymbol{\mu}_g\right)^T\boldsymbol{\Lambda}_g\left(\boldsymbol{g}-\boldsymbol{\mu}_g\right)$$

$$+\frac{m}{2}\log\tau_g + \frac{1}{2}\log|\boldsymbol{\Lambda}_g| - \frac{\tau_g}{2}\left(\boldsymbol{\mu}_g-\boldsymbol{\pi}_g\right)^T\boldsymbol{\Lambda}_g\left(\boldsymbol{\mu}_g-\boldsymbol{\pi}_g\right)$$

$$+\frac{\left(v_g-m-1\right)}{2}\log|\boldsymbol{\Lambda}_g| - \frac{1}{2}Tr\left[\boldsymbol{W}_g{}^{-1}\boldsymbol{\Lambda}_g\right]$$

$$+\sum_{i=1}^{n_a}\{(a_i-1)\log\lambda_i - b_i\lambda_i\} + \text{constant} \qquad 7.43$$

Substituting for $\log\mathcal{P}\left(\boldsymbol{m},\boldsymbol{\mu}_m,\boldsymbol{\Lambda}_m,\boldsymbol{g},\boldsymbol{\mu}_g,\boldsymbol{\Lambda}_g,\boldsymbol{\Lambda}_\epsilon,\boldsymbol{d}\right)$ from equation 7.43 into the MF update equations 7.36 to 7.40, followed by some algebraic manipulation, we get these update equations in closed form for each of the unknown variables $\boldsymbol{m},\boldsymbol{\mu}_m,\boldsymbol{\Lambda}_m,\boldsymbol{g},\boldsymbol{\mu}_g,\boldsymbol{\Lambda}_g$ and $\boldsymbol{\Lambda}_\epsilon$ which are given below.

The update equation for $\mathcal{Q}(\boldsymbol{m})$ shows that it is a Normal distribution given by

$$\mathcal{Q}(\boldsymbol{m}) = N(\boldsymbol{\mu}_m{}^*,\boldsymbol{\Sigma}_m{}^*) = N(\boldsymbol{\mu}_m{}^*,(\boldsymbol{\Lambda}_m{}^*)^{-1}) \qquad 7.44$$

where the mean $\boldsymbol{\mu}_m{}^*$ and covariance matrix $\boldsymbol{\Sigma}_m{}^*$ (or precision matrix $\boldsymbol{\Lambda}_m{}^*$) may be computed from the current estimate of rest of the random variables as

$$\boldsymbol{\Sigma}_m{}^* = (\boldsymbol{\Lambda}_m{}^*)^{-1} = \left(\widehat{\boldsymbol{G}}^T\widehat{\boldsymbol{\Lambda}}_\epsilon\widehat{\boldsymbol{G}} + \widehat{\boldsymbol{\Lambda}}_m\right)^{-1} = \left(\widehat{\boldsymbol{G}}^T\widehat{\boldsymbol{\Sigma}}_\epsilon{}^{-1}\widehat{\boldsymbol{G}} + \widehat{\boldsymbol{\Sigma}}_m{}^{-1}\right)^{-1} \qquad 7.45$$

$$\boldsymbol{\mu}_m{}^* = (\boldsymbol{\Lambda}_m{}^*)^{-1}\left(\widehat{\boldsymbol{G}}^T\widehat{\boldsymbol{\Lambda}}_\epsilon\boldsymbol{d} + \widehat{\boldsymbol{\Lambda}}_m\widehat{\boldsymbol{\mu}}_m\right) = (\boldsymbol{\Sigma}_m{}^*)\left(\widehat{\boldsymbol{G}}^T\widehat{\boldsymbol{\Sigma}}_\epsilon{}^{-1}\boldsymbol{d} + \widehat{\boldsymbol{\Sigma}}_m{}^{-1}\widehat{\boldsymbol{\mu}}_m\right) \qquad 7.46$$

The complete derivation of these equations is given in Appendix C. Although not apparent from its current form, the above result is interestingly similar to the MAP estimate of the model parameters $\boldsymbol{m}$ for a Gaussian prior distribution with fixed parameters in equations 7.18 and 7.19 (Tarantola & Valette, 1982; and Mosegaard & Tarantola, 2002). However, with some algebraic manipulation equations 7.45 and 7.46 can be shown to have a similar form to equations 7.18 and 7.19, respectively, except that the fixed parameters $\tilde{\boldsymbol{\mu}}_m$ and $\tilde{\boldsymbol{\Sigma}}_m$, $\tilde{\boldsymbol{\Sigma}}_\epsilon$, and $\tilde{\boldsymbol{G}}$ in equations equations 7.18 and 7.19 are now replaced with their respective counterparts in the hierarchical model, i.e. the current estimate of expected values of the corresponding random variables $\hat{\boldsymbol{\mu}}_m$ and $\hat{\boldsymbol{\Sigma}}_m$, $\hat{\boldsymbol{\Sigma}}_\epsilon$, and $\hat{\boldsymbol{G}}$, respectively.

The update equation for $Q(\boldsymbol{g})$ shows that it is also a Normal distribution that is given by

$$Q(\boldsymbol{g}) = N(\boldsymbol{\mu}_g{}^*, \boldsymbol{\Sigma}_g{}^*) = N\left(\boldsymbol{\mu}_g{}^*, (\boldsymbol{\Lambda}_g{}^*)^{-1}\right) \tag{7.47}$$

where the mean $\boldsymbol{\mu}_g{}^*$ and covariance matrix $\boldsymbol{\Sigma}_g{}^*$ (or precision matrix $\boldsymbol{\Lambda}_g{}^*$) may be computed from the current estimate of rest of the random variables as

$$\boldsymbol{\Sigma}_g{}^* = (\boldsymbol{\Lambda}_g{}^*)^{-1} = \left(\hat{\boldsymbol{M}}^T \hat{\boldsymbol{\Lambda}}_\epsilon \hat{\boldsymbol{M}} + \hat{\boldsymbol{\Lambda}}_g\right)^{-1} = \left(\hat{\boldsymbol{M}}^T \hat{\boldsymbol{\Sigma}}_\epsilon{}^{-1} \hat{\boldsymbol{M}} + \hat{\boldsymbol{\Sigma}}_g{}^{-1}\right)^{-1} \tag{7.48}$$

$$\boldsymbol{\mu}_g{}^* = (\boldsymbol{\Lambda}_g{}^*)^{-1}\left(\hat{\boldsymbol{M}}^T \hat{\boldsymbol{\Lambda}}_\epsilon \boldsymbol{d} + \hat{\boldsymbol{\Lambda}}_g \hat{\boldsymbol{\mu}}_g\right) = \boldsymbol{\Sigma}_g{}^*\left(\hat{\boldsymbol{M}}^T \hat{\boldsymbol{\Sigma}}_\epsilon{}^{-1} \boldsymbol{d} + \hat{\boldsymbol{\Sigma}}_g{}^{-1} \hat{\boldsymbol{\mu}}_g\right) \tag{7.49}$$

where $\boldsymbol{M}$ is defined in equation 7.13. The complete derivation of these equations is exactly similar to the derivation for the approximate posterior distribution $Q(\boldsymbol{m})$ given in Appendix C. Also, these have exactly the same form as that of equations 7.18 and 7.19, except that the fixed parameters are replaced with the current estimates of the mean value of the corresponding random variables.

The update equations for the joint distribution $Q(\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m)$ are derived in Appendix D, which show that it is a Normal-Wishart distribution given by

$$Q(\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m) = N(\boldsymbol{\mu}_m | \boldsymbol{\Lambda}_m; \boldsymbol{\pi}_m{}^*, \tau_m{}^* \boldsymbol{\Lambda}_m)\, W(\boldsymbol{\Lambda}_m; \boldsymbol{W}_m{}^*, v_m{}^*) \tag{7.50}$$

where the updated current estimates of the mean $\boldsymbol{\pi}_m{}^*$, precision matrix scale $\tau_m{}^*$, scale matrix $\boldsymbol{W}_m{}^*$ and degrees of freedom $v_m{}^*$ may be computed from the current estimate of rest of the random variables as

$$\tau_m{}^* = 1 + \hat{\tau}_m \tag{7.51}$$

$$\boldsymbol{\pi}_m{}^* = (\tau_m{}^*)^{-1}(\widehat{\boldsymbol{m}} + \tau_m \widehat{\boldsymbol{\pi}}_m) \tag{7.52}$$

$$\boldsymbol{W}_m{}^* = \left(\widehat{\boldsymbol{W}}_m{}^{-1} + \hat{\tau}_m(\tau_m{}^*)^{-1}(\widehat{\boldsymbol{m}} - \widehat{\boldsymbol{\pi}}_m)(\widehat{\boldsymbol{m}} - \widehat{\boldsymbol{\pi}}_m)^T\right)\widehat{\boldsymbol{\Lambda}}_m \tag{7.53}$$

$$v_m{}^* = \hat{v}_m + 2 \tag{7.54}$$

The update equations for the joint distribution $\mathcal{Q}(\boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g)$ may be obtained analogously to the those for $(\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m)$ (see Appendix D), which show that it is also a Normal-Wishart distribution that is given by

$$\mathcal{Q}(\boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g) = N(\boldsymbol{\mu}_g|\boldsymbol{\Lambda}_g; \boldsymbol{\pi}_g{}^*, \tau_g{}^*\boldsymbol{\Lambda}_g)\, W(\boldsymbol{\Lambda}_g; \boldsymbol{W}_g{}^*, v_g{}^*) \tag{7.55}$$

where the updated current estimates of the mean $\boldsymbol{\pi}_g{}^*$, precision matrix scale $\tau_g{}^*$, scale matrix $\boldsymbol{W}_g{}^*$ and degrees of freedom $v_g{}^*$ may be computed from the current estimate of rest of the random variables as

$$\tau_g{}^* = 1 + \hat{\tau}_g \tag{7.56}$$

$$\boldsymbol{\pi}_g{}^* = \left(\tau_g{}^*\right)^{-1}\left(\widehat{\boldsymbol{g}} + \hat{\tau}_g \widehat{\boldsymbol{\pi}}_g\right) \tag{7.57}$$

$$\boldsymbol{W}_g{}^* = \left(\widehat{\boldsymbol{W}}_g{}^{-1} + \hat{\tau}_g\left(\tau_g{}^*\right)^{-1}\left(\widehat{\boldsymbol{g}} - \widehat{\boldsymbol{\pi}}_g\right)\left(\widehat{\boldsymbol{g}} - \widehat{\boldsymbol{\pi}}_g\right)^T\right)\widehat{\boldsymbol{\Lambda}}_g \tag{7.58}$$

$$v_g{}^* = \hat{v}_g + 2 \tag{7.59}$$

The update equations for $\mathcal{Q}(\boldsymbol{\Lambda}_\epsilon)$ are derived in Appendix E, which show that it is a Gamma distribution given by

$$\mathcal{Q}(\boldsymbol{\Lambda}_\epsilon) = \prod_{i=1}^{n_a} \mathcal{Q}(\lambda_i) = \prod_{i=1}^{n_a} G(\lambda_i \mid a_i{}^*, b_i{}^*) \tag{7.60}$$

with scale $a_i{}^*$ and rate $b_i{}^*$ parameters updated according to

$$a_i{}^* = \hat{a}_i + \frac{n_s}{2} \tag{7.61}$$

$$b_i{}^* = \hat{b}_i + \frac{\hat{S}_i}{2} \tag{7.62}$$

where $\hat{S}_i$ is defined as

$$\hat{S}_i = \left\| \boldsymbol{d}_i - \widehat{\boldsymbol{G}}_i \widehat{\boldsymbol{m}} \right\|^2 = \|\boldsymbol{d}_i\|^2 + \left\| \widehat{\boldsymbol{G}}_i \widehat{\boldsymbol{m}} \right\|^2 - 2\, \boldsymbol{d}_i{}^T \widehat{\boldsymbol{G}}_i \widehat{\boldsymbol{m}} \tag{7.63}$$

Since $\hat{S}_i \geq 0, \forall i$ according to its definition in equation 7.63 and $\hat{S}_i \to 0, \forall i$ as the inversion proceeds, the above update equations 7.61 and 7.62 show that both $a^*{}_i$ and $b^*{}_i$ increase monotonically. However, $a^*{}_i$ increases at a constant rate across iterations while the rate of increase of $b^*{}_i$ dcreases as the inversion proceeds. This means that as inversion proceeds towards an optimum solution, the expected value of the precision $\lambda_i$ of each component $\boldsymbol{d}_i$ of $\boldsymbol{d}$, which is given by $a^*{}_i / b^*{}_i$, increases. Equivalently, the expected value of variance $\sigma_i{}^2 = 1/\lambda_i$ decreases. It is therefore recommended to start with a high initial variance $\sigma_i{}^2$, or low initial precision $\lambda_i$, of the data noise.

After a complete iteration, the updated parameters (with an asterisk '*' superscript) replace the current estimate of corresponding parameters (marked with a caret accent '^'). Iterative updates continues in this manner until convergence. Since each update monotonically increases the free energy $\mathcal{F}(\mathcal{Q})$, or monotonically decreases the relative entropy $\mathrm{KL}(\mathcal{Q}||\mathcal{P})$, convergence to a fixed point on the surface of $\mathcal{F}(\mathcal{Q})$ is guaranteed in this method. The convergence point is the global optimum for a Gaussian posterior distribution as presented here. For further discussion on convergence, see section 7.8.

## 7.6 Computational Complexity

The computational complexity of this method mainly depends on the cost of each of the MF updates (using equations 7.36 to 7.40), which in turn is mainly determined by the cost of matrix multiplication and inversion operations. In a densely coupled model (where model parameters at any location are correlated with model parameters in the rest of the model), the cost of matrix inversion can be as high as $\mathcal{O}(n^3)$, where $n$ is the size of one dimension of a matrix. However, since all of the matrices involved ($\boldsymbol{G}$, $\boldsymbol{M}$, $\boldsymbol{W}_m$, $\boldsymbol{W}_g$, $\boldsymbol{\Lambda}_m$, $\boldsymbol{\Lambda}_g$ and $\boldsymbol{\Lambda}_\epsilon$) are sparse, where $\boldsymbol{G}$ and $\boldsymbol{M}$ are also block Toeplitz matrices and $\boldsymbol{\Lambda}_\epsilon$ is also a block diagonal matrix, these can be inverted using efficient numerical algorithms, e.g. using (band-) Cholesky decomposition (Asif & Moura, 2005; Martinsson *et al.* 2005; Rue & Held, 2005; Lin *et al.* 2011). The computational advantage of dealing with sparse matrices comes from the fact that we neither need to store, nor do we need to compute, the elements that are known to be zero. Further, the presented method may be implemented in the frequency domain using Fast Fourier Transform (FFT) to further improve its computational efficiency (Buland *et al.* 2003). The computational cost of this method is $\mathcal{O}(n \log n)$ in the frequency domain, and is at most $\mathcal{O}(n^2)$ in the time domain.

Depending on the degree of sparsity of matrices involved, the computational cost of this method may be much less than $\mathcal{O}(n^2)$ in the time domain. In order to maximize the use of sparse structure of these matrices we need to minimize the number of non-zero terms. This may require permutations of the indices of the prior GMRF model in space such that the prior precision matrix $\widehat{\boldsymbol{\Lambda}}_m$ and its scale matrix $\boldsymbol{W}_m$ become a narrow band matrices (i.e. most non-zero entries get concentrated at or close to the main diagonal of the matrix). For details on such operations see Rue & Held (2005). The size of the dimension $n$ in the above expressions may be limited under the assumption of conditional independence of each bin gather (multiple partial-angle traces corresponding to one CDP), which significantly reduces the computational cost of this method.

## 7.7 Application: Seismic AVO Inversion

A real 2D seismic data example from the North Sea is presented here to demonstrate a practical application of the proposed variational Bayesian inversion (VBI) method. The input

data consists of six *pre-stack time migrated* (PSTM) partial-angle seismic stacks at the following angles: 04-12, 12-20, 20-28, 28-36, 36-44, and 40-48 degrees, with middle angles: 8, 16, 24, 32, 40 and 44 degrees, respectively (figure 7.4), and well logs from a borehole (figure 7.5) that is located on the available 2D seismic section.

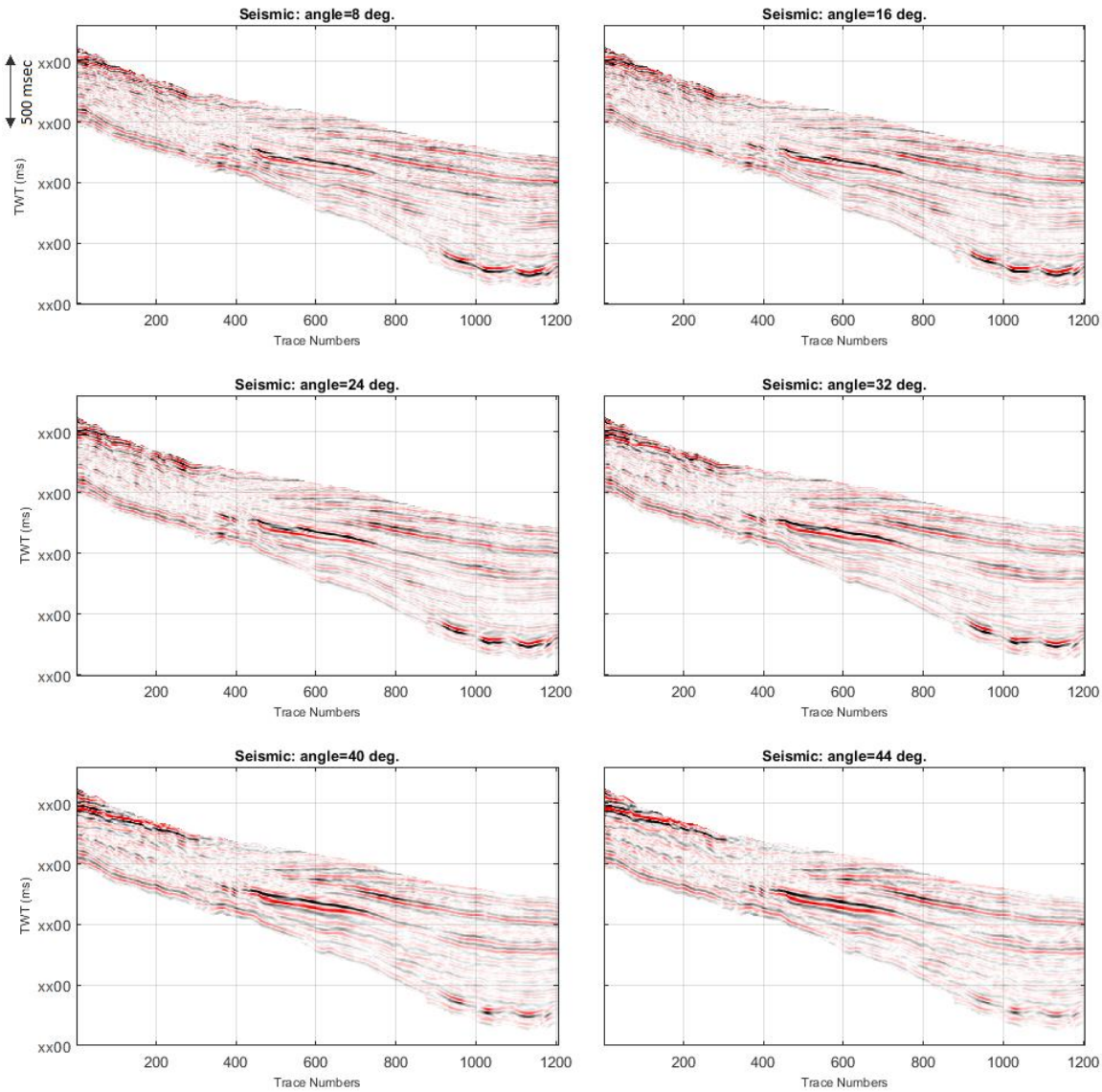



*Figure 7.4: Partial-angle seismic stacks used in AVO inversion for elastic seismic attributes $V_p$, $V_s$, and $\rho$.*

Seismic AVO data may be expressed in terms of contrasts in the elastic parameters across different rock and/or fluid interfaces in the subsurface using non-linear Zoeppritz equations (Aki & Richards, 1980). Linear approximations to the reflection amplitudes as a

function of reflection angle and reflectivity contrasts in elastic parameters across an interface are proposed by Aki & Richards (1980), which for PP (P-to-P wave) reflection are given by:

$$R_{PP}(\theta) = \frac{1}{2} sec^2\,\theta\, R_p \; - 4\gamma^2\, sin^2\,\theta\, R_s + \frac{1}{2}(1 - 4\gamma^2\, sin^2\,\theta) R_d \qquad 7.64$$

where $R_p$, $R_s$ and $R_d$ are respectively the P-wave, S-wave and density reflectivity contrasts across the interface, $\gamma$ is the average S-wave velocity to P-wave velocity ratio ($\bar{V}_s/\bar{V}_p$), and $\theta$ is the average of PP reflection and transmission angles at the interface. The reflectivity contrasts in the above equation may be approximated by the first order derivatives of material properties with respect to the independent variable of recording (e.g. time), such that $R_p = \partial/\partial t\ \log V_p(\boldsymbol{x}, t)$, $R_s = \partial/\partial t\ \log V_s(\boldsymbol{x}, t)$ and $R_d = \partial/\partial t\ \log \rho(\boldsymbol{x}, t)$, where $V_p(\boldsymbol{x}, t)$, $V_s(\boldsymbol{x}, t)$ and $\rho(\boldsymbol{x}, t)$ are the P-wave & S-wave velocities and density, respectively, as a function of spatial coordinates $\boldsymbol{x}$ and time $t$.

The forward model in equation 7.12 may then be written as

$$\boldsymbol{d} = \boldsymbol{G}\boldsymbol{m} + \boldsymbol{\varepsilon} = \boldsymbol{W}\boldsymbol{A}\boldsymbol{D}\boldsymbol{m} + \boldsymbol{\varepsilon} \qquad 7.65$$

where $\boldsymbol{m}$ is a vector of logarithms of discretized elastic attributes, i.e. $\boldsymbol{m} = \log\left[\boldsymbol{V}_p, \boldsymbol{V}_s, \boldsymbol{\rho}\right]^T$ in each model cell that corresponds to one sample of the seismic data, $\boldsymbol{W}$ is a block Toeplitz matrix of seismic wavelet(s), $\boldsymbol{A}$ is a matrix of coefficients of reflectivity contrasts in equation 7.64, and $\boldsymbol{D}$ is a matrix of first-order difference operators, and $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{\Sigma}_\varepsilon)$ is a matrix of independent and Normally distributed stochastic errors (Buland & Omre, 2003a).

The elastic properties of rocks generally contain wide-band variations in space. However, seismic data are typically band-limited, and therefore do not contain all of the required information to reconstruct the elastic properties. Also the coefficients of reflectivity contrasts in equation 7.64 depend on the average (low frequency) S-wave to P-wave velocity ratio $\gamma \equiv \bar{V}_s/\bar{V}_p$. This suggests that additional low frequencies constraints must be introduced for reliable and accurate modelling of the trend in elastic parameters. Low frequency models (LFM) were built using cokriging of low-pass filtered well-log data and seismic migration velocities (figure 7.6). These LFMs were used to define the initial expectations $\boldsymbol{\pi}_m$ of model parameters.

We intend to invert the partial-angle stacks for the following subsurface elastic properties (also called attributes): P-wave velocity $V_p$, S-wave velocity $V_s$, and density $\rho$. Initial data analysis and quality check (QC) was performed before inversion which include the following steps: well data analysis to verify the Gaussian assumption and to estimate spatial correlation range, AVO attributes analysis, wavelet extraction to analyze frequency and phase of the seismic data, and to generate synthetic seismograms with multiple signal-to-noise ratios (SNR), and inversion test on these synthetic seismograms before applying inversion to the seismic data. The parameters required for inversion were also computed during this analysis phase. These steps are described below.

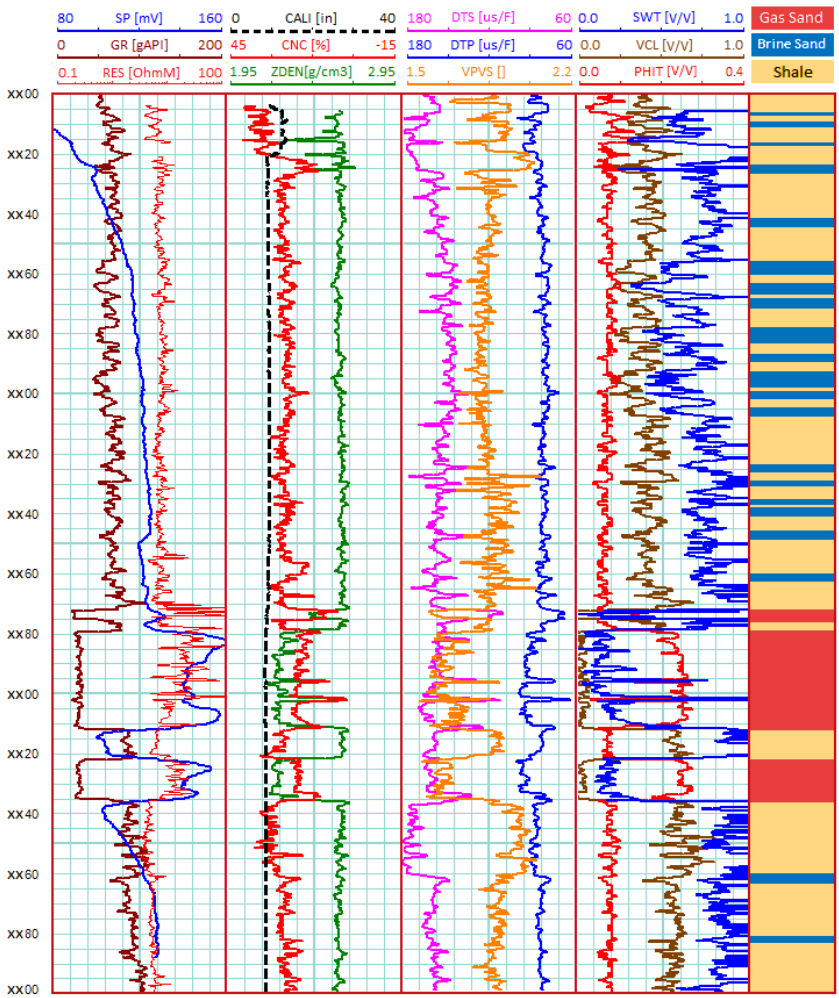



*Figure 7.5: Well-log data and interpreted facies profile from well W1. Standard well-log pneumonics are used for the well log curves as shown in the headers above the display tracks. The well data were used to calibrate low frequency model (LFM) and to construct prior precision (inverse covariance) matrix $\boldsymbol{\Lambda_m}$ of the model parameters (see equation 7.8).*

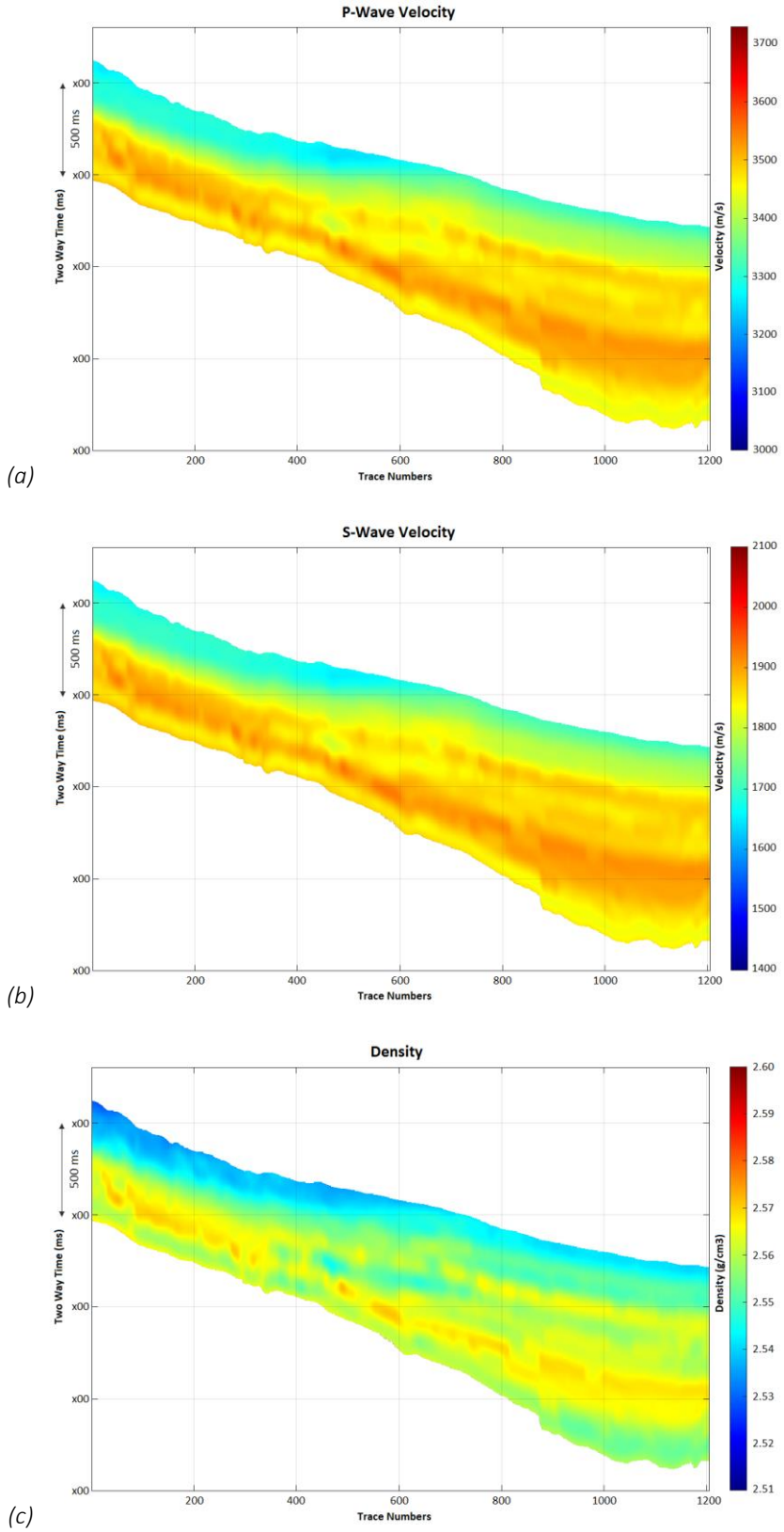



*Figure 7.6: Low frequency models of elastic rock properties (a) $V_p$, (b) $V_s$, and (c) $\rho$.*

### *7.7.1 Well Data Analysis*

As a critical QC step, well log data (figure 7.5) were used to verify if the Gaussian assumption is acceptable for this geology. For this purpose, the logarithms of elastic properties $V_p$, $V_s$, $\rho$ obtained from well logs were visualized on a Normal probability plot (figure 7.7). For perfectly Normal distributed parameters, the top row plots should be linear. From figure 7.7 we see that distributions of the logarithm of velocities are close to Normal with some departure. The logarithm of density shows an almost linear trend for most of the observations except some observations which deviate significantly from the linear trend. On inspection, it was identified that the low density within the reservoir interval due to presence of gas caused this deviation. In one respect it shows a clear anomaly that acts as a direct hydrocarbon indicator (DHI) that we have seen in figure 7.7. Nevertheless, for now we assume that the Gaussian assumption is valid.

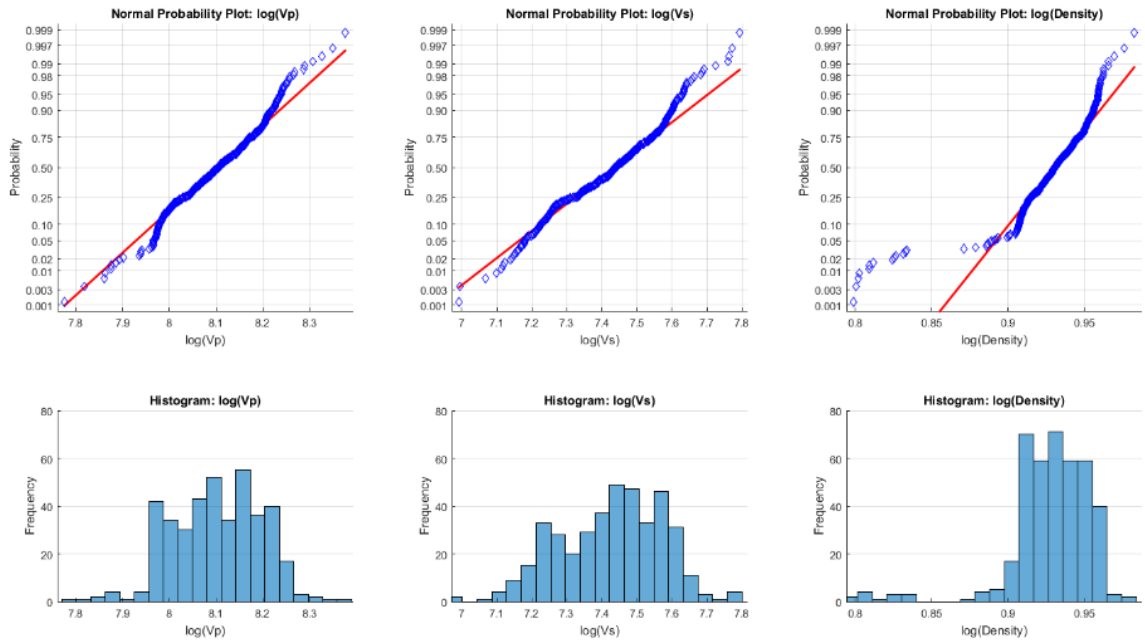



*Figure 7.7: (Top row) Normal probability plots and (bottom row) histograms of logarithms of elastic properties $V_p$, $V_s$, and $\rho$ obtained from well-logs.*

The initial expectation $\boldsymbol{\pi}_m$ of the model parameters were obtained by applying logarithm to the LFMs. The scale matrix $\boldsymbol{W}_m$ was defined as the initial estimate of the prior precision matrix $\boldsymbol{\Lambda}_m$ that was defined in terms of a spatial precision matrix $\boldsymbol{\Lambda}$ and a stationary

covariance matrix $\boldsymbol{\Sigma}_0$ of the three elastic properties. $\boldsymbol{\Sigma}_0$ was computed from these well logs as below:

$$\boldsymbol{\Sigma}_0 = \begin{bmatrix} 1.0 & 0.89 & 0.75 \\ 0.89 & 1.0 & 0.53 \\ 0.75 & 0.53 & 1.0 \end{bmatrix} \quad 7.66$$

where the three columns (or rows) are sequenced as $V_p$, $V_s$ and $\rho$.

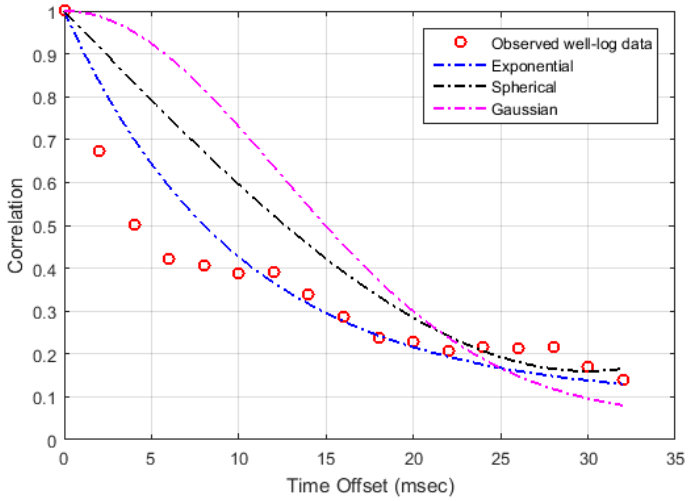



*Figure 7.8: Temporal correlation function estimated from the well-log data (red circles), and different analytical functions (equations 7.4 to 7.6) plotted for comparison. Exponential correlation function was used to define the prior distribution.*

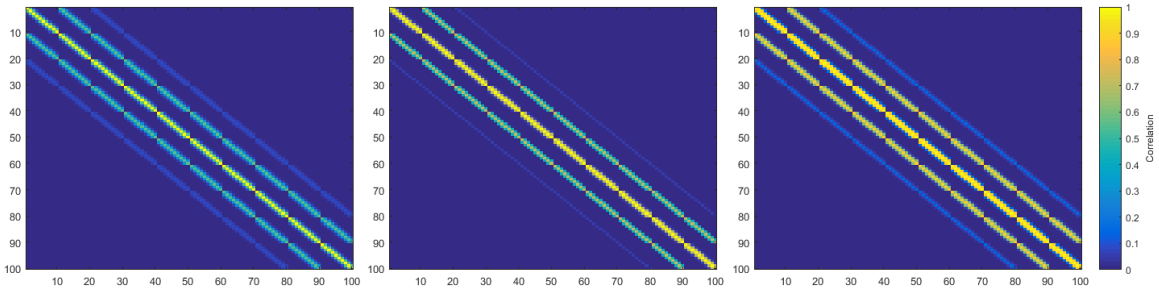


*Figure 7.9: Examples of spatial precision matrices computed using the three correlation functions given in equations 7.4 to 7.6, for a hypothetical model of size 10x10 cells, where spatial correlation length is 5 cells.*

The spatial precision matrix $\boldsymbol{\Lambda}$ was defined using a GMRF with horizontal and vertical coupling of parameters. The vertical correlation length was chosen as 20 ms that was estimated from spatial correlation of log curves (figure 7.8). The horizontal correlation length

was chosen as 6 traces across. This corresponds to twice the vertical correlation length as layer properties are typically expected to be highly correlated in the lateral direction. If data from closely spaced wells or nearly horizontal wells was available, a better estimate of horizontal correlation length could be obtained. The exponential correlation function (equation 7.4) was used to model spatial correlations as it appears to have a better match in the vertical direction compared to the other two correlation functions. The matrix $\boldsymbol{\Lambda}$ is sparse and non-zero values are mostly concentrated close its main diagonal. As an example, spatial precision matrices computed using the three correlation functions given in equations 7.4 to 7.6 are shown in figure 7.9 for a hypothetical model of size 10x10 cells, where spatial correlation length is 5 cells. Equation 7.8 was used to compose $\boldsymbol{\Lambda}$ and $\boldsymbol{\Sigma_0^{-1}}$ together to obtain $\boldsymbol{\Lambda}_m$ as an initial estimate of $\boldsymbol{W}_m$. The initial degrees of freedom $v_m$ of the prior Wishart distribution of $\boldsymbol{\Lambda}_m$ was set to one more than the total number of samples in a trace obtained by combining logarithms of the three elastic properties, which corresponds to non-informative prior for a Wishart distribution.

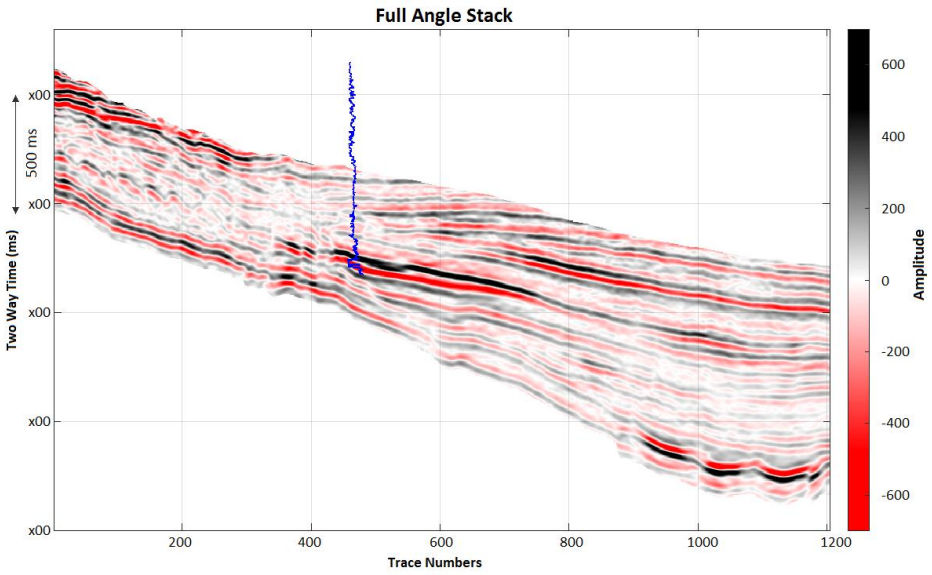


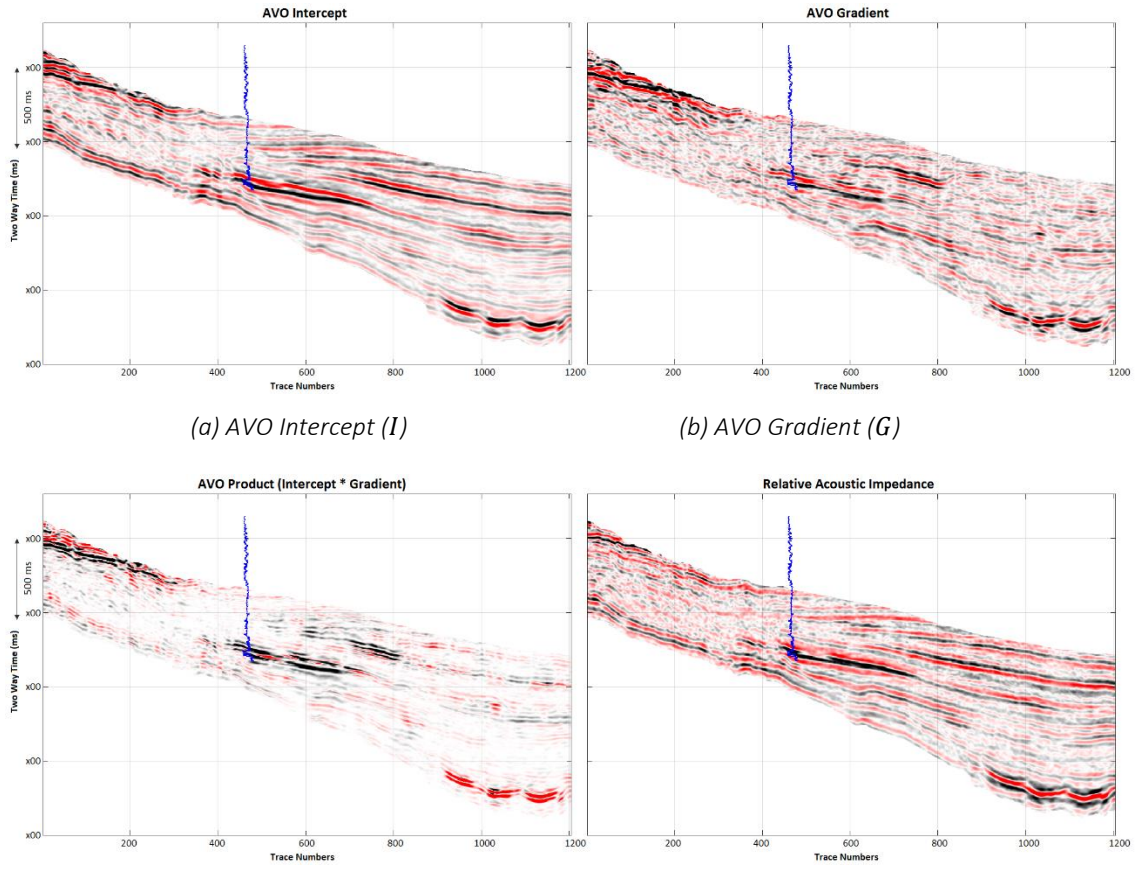

*Figure 7.10: Full-angle stack of seismic data with Gamma Ray ($GR$) log displayed at the location of a well in the study area.*

## *7.7.2 AVO Attributes Analysis*

Full-angle stack of seismic data was computed as a sum of all of the partial-angle stacks (figure 7.10), and the following AVO attributes were computed from the partial-angle stacks:

*AVO intercept* ($I$), *gradient* ($G$), *product* (intercept times gradient, $I^*G$) (Foster *et al*. 2010), and *relative acoustic impedance* ($RAI$) (figure 7.11). AVO attributes are represent properties of interface between rock layers and are often helpful in identifying hydrocarbons. The top of the gas reservoir can be easily identified by a *Class-3 AVO anomaly* (Foster *et al*. 2010) identified in the seismic data using the AVO attributes (figure 7.12).

The $RAI$ is a layer property unlike seismic data which measures reflectivity contrast across interfaces. However, $RAI$ is low in resolution as the effect of seismic wavelet has not been removed in its computation. In this manner, it gives a quick idea about minimum details that can be resolved in a subsequent full inversion workflow.




*(a) AVO Intercept ($I$)* *(b) AVO Gradient ($G$)*

*(c) AVO Product ($I * G$)* *(d) Relative Acoustic Impedance ($RAI$)*

*Figure 7.11: Seismic waveform attributes. (a) AVO intercept ($I$), (b) AVO gradient ($G$), (c) AVO product (intercept times gradient, $I * G$). The attributes (a-c) were computed using the first four partial-angle stacks (with mid-angles 8, 16, 24, and 32 degrees) shown in figure 7.4(a-d). (d) Relative acoustic impedance (RAI) attribute computed from the full-stack seismic data shown in figure 7.1.*

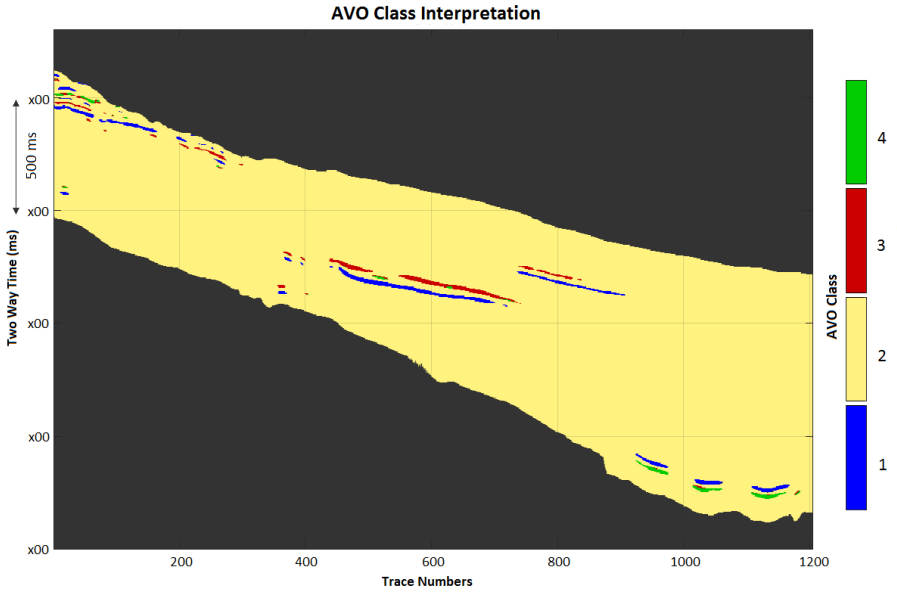



*Figure 7.12: AVO class attribute computed from the AVO intercept and gradient attributes (Foster et al. 2010).*

## *7.7.3 Seismic Wavelet Analysis*

The forward model in equation 7.65 comprises of wavelet $\boldsymbol{W}$ and the Aki-Richards coefficients $\boldsymbol{A}$. The wavelet was assumed to have a Gaussian prior distribution. The coefficients $\boldsymbol{A}$ were assumed to be fixed given the low frequency models. A number of initial wavelets were extracted from each of the partial-angle seismic stacks at the well location (figure 7.13) using a frequency domain approach (Walden & White, 1998).

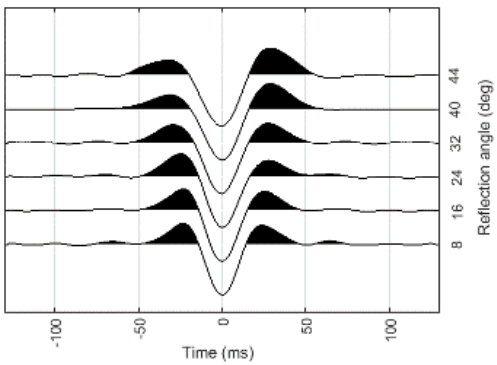



*Figure 7.13: Initial estimate of seismic wavelets extracted from each of the partial-angle stacks at the well location.*

The central frequency of the wavelet was Uniformly sampled between 15 and 25 Hz, and phase was Uniformly sampled from -10 to 15 degrees. These ranges were chosen based on correlation coefficients between seismic traces in the vicinity of borehole and the synthetic seismograms obtained using these wavelets. The initial expectation $\boldsymbol{\pi}_g$ of the wavelet was defined by combining the mean of the wavelets extracted from each of the partial-angle stacks into a single trace, referred to as the wavelet trace. The scale matrix $\boldsymbol{W}_g$ was defined to be the precision matrix of the wavelet trace, and the initial degrees of freedom $v_g$ of the prior Wishart distribution of $\boldsymbol{\Lambda}_g$ was set in a similar manner as for $v_m$.

### *7.7.4 Inversion of Noisy Synthetic Seismograms*

The mean extracted wavelets were then used to generate angle-dependent synthetics seismograms using both the nonlinear Zoeppritz equations and their linearized Aki-Richards approximation (figure 7.14). The difference between the nonlinear and linearized synthetics was also computed which is shown in the right-most track of figure 7.14. The difference shows noticeable differences at higher angles due to linearization errors.

Synthetic seismograms were then generated using linearized forward model 7.64 for signal-to-noise ratios (SNR) of 1, 5, and 15 (figure 7.15), and were inverted using the current method in order to analyze sensitivity of the inversion process to the data noise (figure 7.16). Each of the figures 7.16 (a-c) shows the measured log curves of $V_p$, $V_s$, and $\rho$ in black color in tracks 1-3 from left, and the inverted log curves with maximum-a-posterior (MAP, which is also equal to the mean due to Gaussian assumption) shown in red and the 2$^{nd}$ standard deviation (Std.) shown as shaded yellow regions bounded by dashed red curves. Tracks 4-6 show the computed and reconstructed AVO synthetics, and their differences, respectively.

The standard deviation (Std.) provides quantification of uncertainty in the predicted log curves. For precise inversion, exactly 95.4% of the actual measured log samples should fall within the 2$^{nd}$ Std. of the posterior distribution. We define the percentage of measured log samples contained within the 2$^{nd}$ Std. of the predicted distributions to the ideal value of 95.4% as the *confidence ratio* (CR). An ideal CR is therefore 1.0 which refers to perfect prediction of uncertainty for a Gaussian distribution. A CR value greater than 1.0 represents over-estimation of uncertainty, and vice versa.

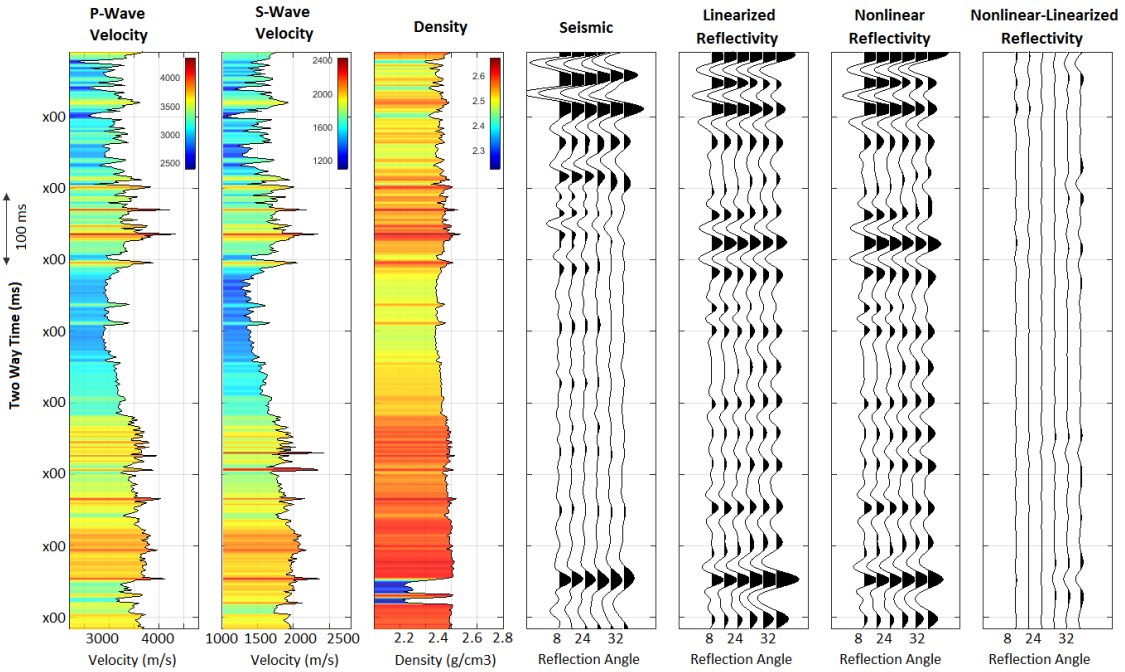



*Figure 7.14: Track numbers from left to right: elastic property logs (1) $V_p$, (2) $V_s$, and (3) $\rho$, (4) seismic AVO traces at the well location, synthetic seismogram (5) using linearized Aki-Richards' approximation and (6) using nonlinear Zoeppritz equations, and (7) the difference between nonlinear and linearized reflectivity.*

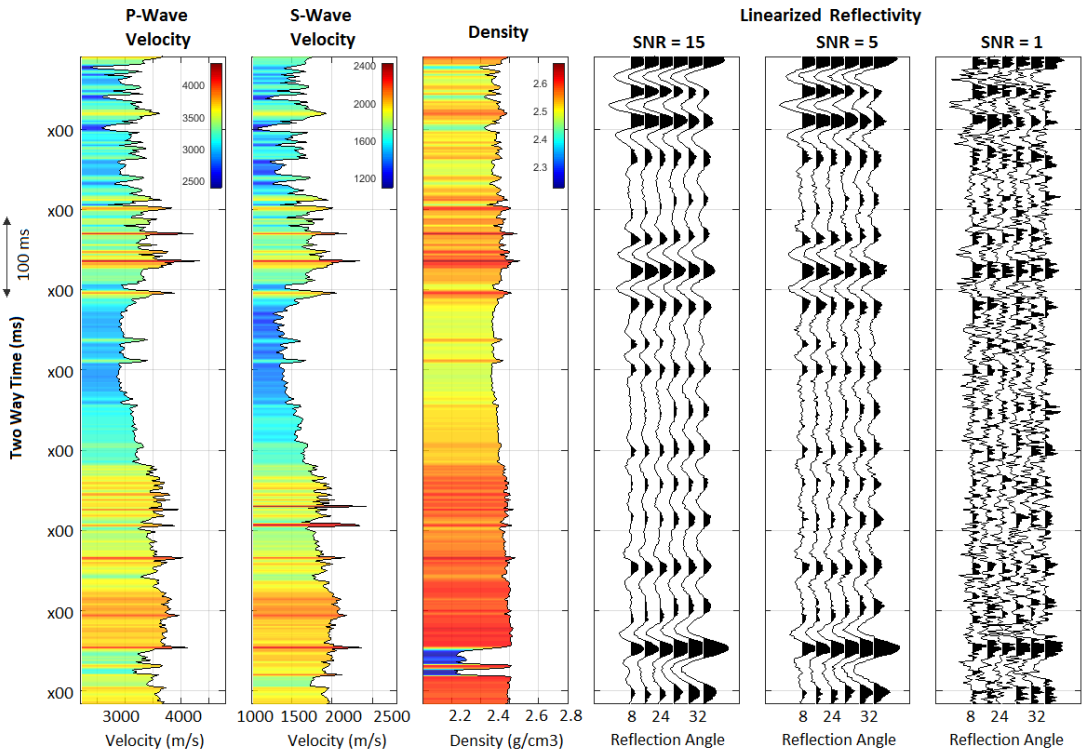



*Figure 7.15: Track numbers from left to right: elastic property logs (1) $V_p$, (2) $V_s$, and (3) $\rho$, synthetic seismogram computed with signal-to-noise ratio (SNR) values of (4) 15, (5) 5, and (6) 1.*

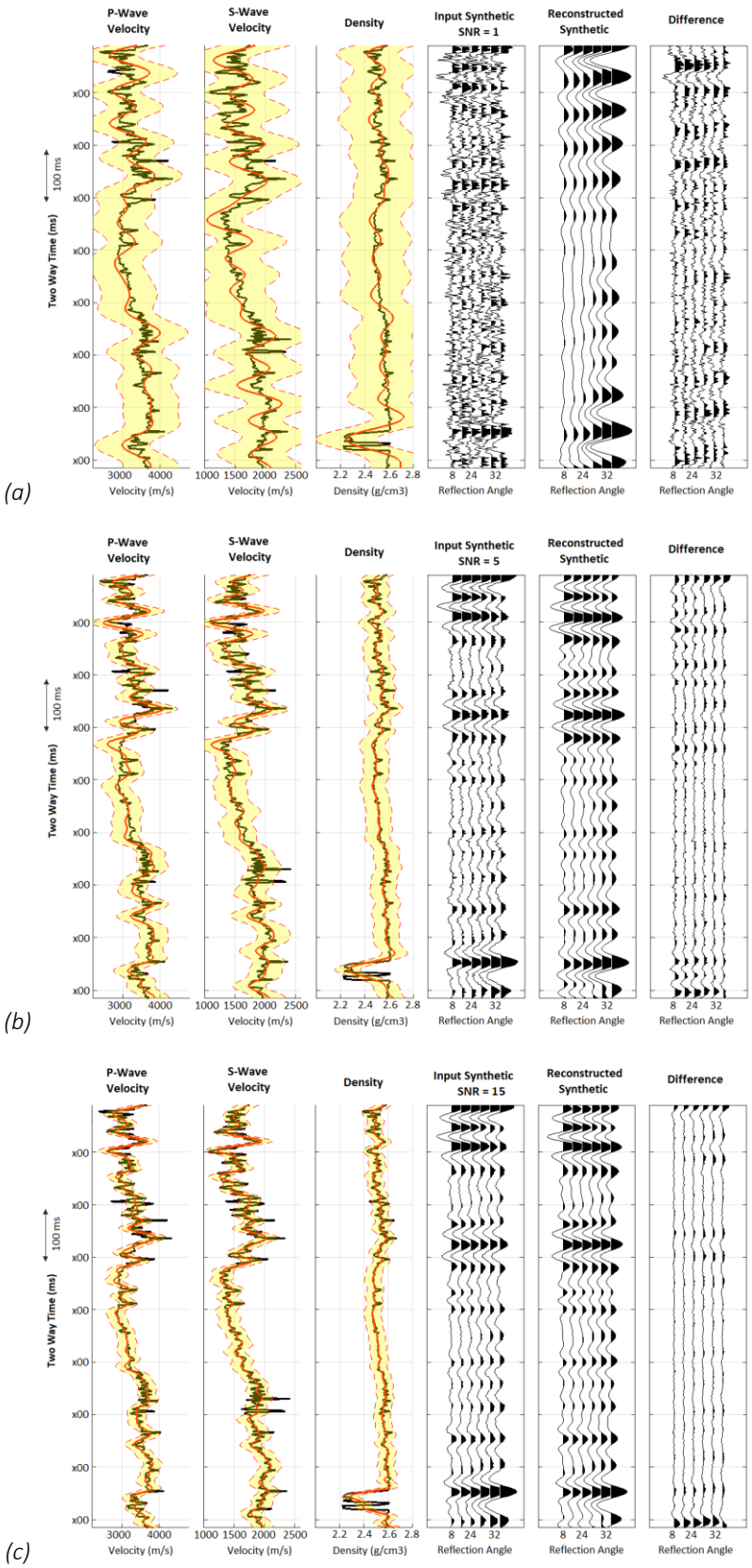



*Figure 7.16: Inversion of synthetic seismograms with different signal-to-noise ratio (SNR). (a) SNR=1, (b) SNR=5, and (c) SNR=15. See text for details.*

*Table 7.1: Prior and posterior standard deviations (Std.) and confidence ratios (CR) for the inverted elastic properties at well location for various signal-to-noise ratios (SNR) computed with respect to the measured (reference) log curves. Confidence ratio is defined in the text.*

| Signal to Noise Ratio (SNR) | Elastic property | Prior Std. | Posterior Std. | Confidence Ratio (CR) |
|---|---|---|---|---|
| **1** | P-wave velocity, $V_p$ (m/s): | 400 | 382.12 | 1.036 |
| **1** | S-wave velocity, $V_s$ (m/s): | 300 | 311.03 | 1.046 |
| **1** | Density, $\rho$ (g/cm3): | 0.130 | 0.126 | 1.048 |
| **5** | P-wave velocity, $V_p$ (m/s): | 400 | 215.34 | 1.005 |
| **5** | S-wave velocity, $V_s$ (m/s): | 300 | 156.09 | 1.004 |
| **5** | Density, $\rho$ (g/cm3): | 0.130 | 0.066 | 1.007 |
| **15** | P-wave velocity, $V_p$ (m/s): | 400 | 124.25 | 1.001 |
| **15** | S-wave velocity, $V_s$ (m/s): | 300 | 90.751 | 1.001 |
| **15** | Density, $\rho$ (g/cm3): | 0.130 | 0.041 | 0.986 |

Table 7.1 shows the prior and posterior Std. of elastic properties and values of their CR for various SNR values. The prior Std. were estimated from the single well data and were increased by almost 20% to round-off figures in order to account for possible underestimation of uncertainty due to availability of well data from only one well. The results show that for SNR equal to 1, the posterior uncertainty remains almost the same as the prior since data are contributing little to no information. Also the MAP estimate shows some instability in the form of ringing (figure 7.16a). The posterior uncertainty reduces significantly for SNR values of 5 and 15. The uncertainty is significantly over-estimated for SNR=1, and is close to the perfect value of 1.0 for SNR values of 5 and 15 except for density for which it is a bit underestimated for SNR=15. We know from figure 7.7 that density at the reservoir level significantly deviates from the Normal trend, and that explains slight underestimation of uncertainty in density as SNR improves.

With the synthetics example, an SNR of 5 showed an accuracy of prediction that is very close to perfect. This means that if we assume that there are no other factors that may cause inaccuracies in the inversion results, and if the real seismic data have an SNR value of at least 5, then this inversion method is expected to produce reasonable results for these data. Though

in practice a higher SNR will be needed to obtain reliable inversion results. This is because when inverting the synthetic seismograms, the forward model and the prior correlations in time and among different elastic properties are known with high precision. In reality the linearized forward model is not expected to be accurate, and the correlations in model properties away from the well location may differ significantly from the known correlations at the well location(s).

An initial estimate of noise variance ${\sigma_i}^2 = 1/\lambda_i$ in the seismic data was obtained using spherical coherence analysis (White, 1984) between the noise-free synthetic computed at the well location and 20 seismic bins (partial-angle stacks) on each side of the well. The correlated energy from trace to trace is regarded as signal and the uncorrelated energy is regarded as noise. The maximum estimated noise variance across all of the partial-angle stacks was found to be 0.026 which corresponds to a minimum SNR value of 37.8 (assuming the signal is normalized to have variance equal to 1.0). The parameters of the gamma distribution for $\lambda_i, \forall i$ were initialized as $a_i = 0.001$ and $b_i = 0.001$, which correspond to weak prior on the noise precision (inverse variance) and SNR value of 1.

### *7.7.5 Inversion of Seismic AVO Data*

Once all the required input parameters (fixed parameters in DAG shown in figure 7.2) were set, partial-angle seismic stack data (figure 7.4) were inverted under the assumption that each bin gather is conditionally independent given the model parameters. In this manner, each bin gather may be inverted in parallel within each iteration of the MF update algorithm (equations 7.36 to 7.40). Multiple iterations still need to run in sequence, but this does not have a significant impact on the computational cost since MF algorithm generally requires just a few iterations to converge, usually less than 10. In this example, the method converged within minutes on a desktop computer for a total of 6 iterations (figure 7.17). The MAP (which is also equal to the mean) estimates of elastic properties and their standard deviations after convergence are shown in figures 7.18 and 7.19. For comparison, up-scaled well logs are overlaid on the MAP estimates of respective elastic properties in figure 7.18. Pearson's correlation coefficient was computed between the inverted elastic properties at the well location and the up-scaled well logs which showed acceptable correlation of 0.724, 0.683 and 0.716 for $V_p$, $V_s$ and $\rho$, respectively.

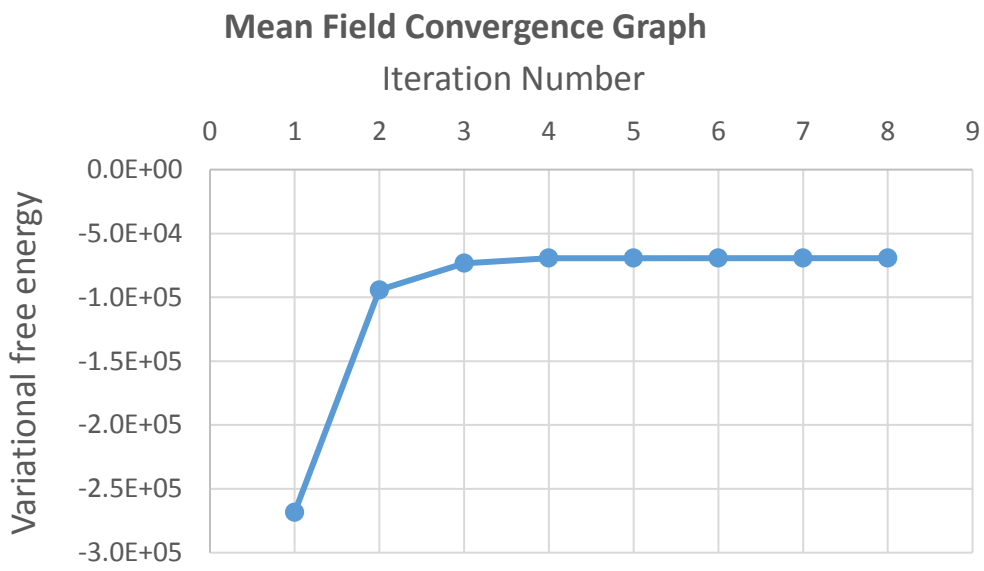
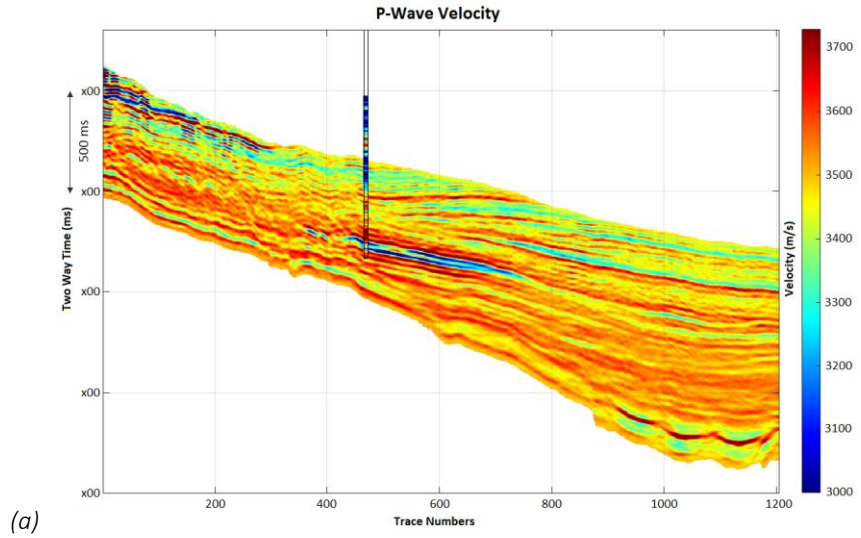



*Figure 7.17: Convergence of the mean field (MF) algorithm showing maximization of variational free energy in 6 iterations to within a tolerance of 0.01.*

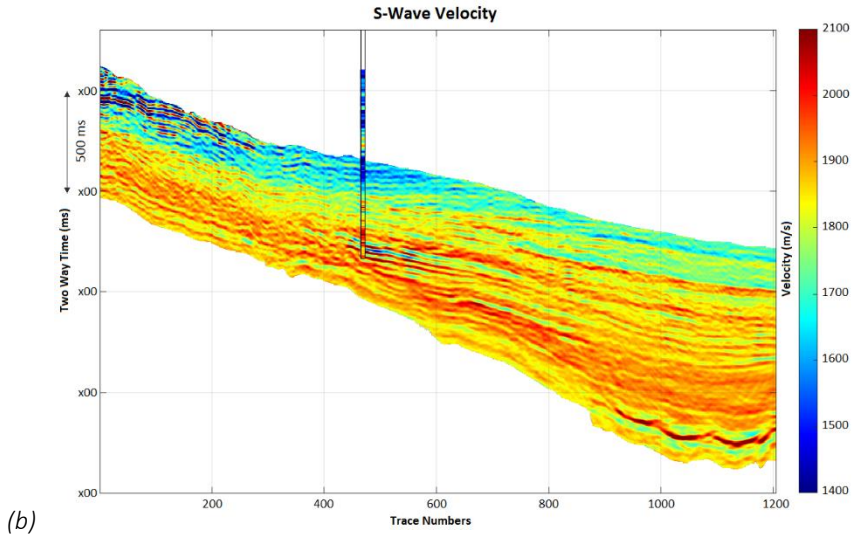
Also, the final MAP estimates and the 2nd standard deviations of wavelets are plotted in figure 7.20. The shaded pink regions represent posterior uncertainty in the wavelet estimates. These wavelets show some differences in frequency and phase compared to the initial estimates shown in figure 7.13. The reason for this difference is that the initial wavelets were estimated only from a few traces in the vicinity of the well, while the posterior wavelets are updated by the inversion process that involved all of the traces.

Synthetic partial-angle sections were computed from the inverted MAP estimates of elastic properties using the linearized forward model. The synthetic sections and their differences from the input seismic are shown in figures 7.21 and 7.22, respectively. The noise is found to be mostly low except at few traces (bin locations) particularly to the left of the plot. We can see from the noise sections (figure 7.23) that this is caused by high dips that were not accounted for by our spatial model. However, some of these dips also appear to be caused by migration smiles. Resolving such discrepancies requires further investigations which we leave for future work.

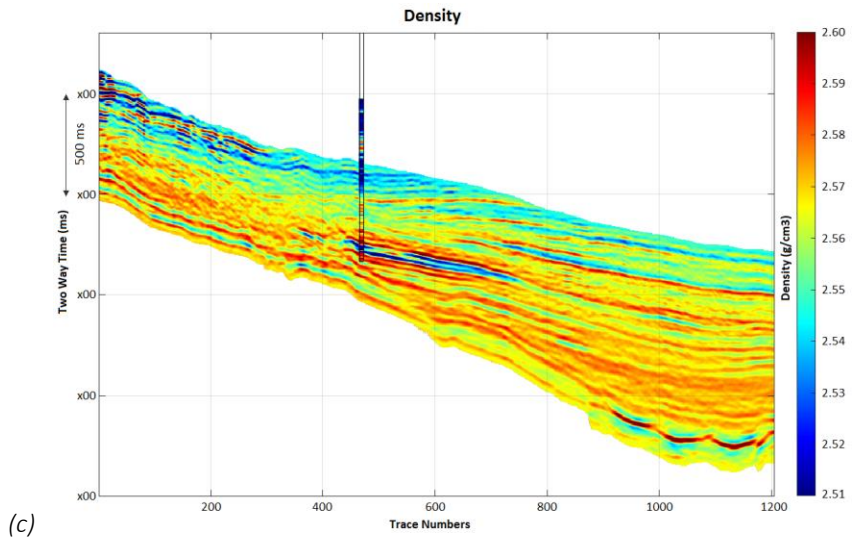
The objective of this study was to evaluate the performance of the proposed VBI method on a real problem in terms of the quality of results and computational performance, both of which are found to be satisfactory.




(a)

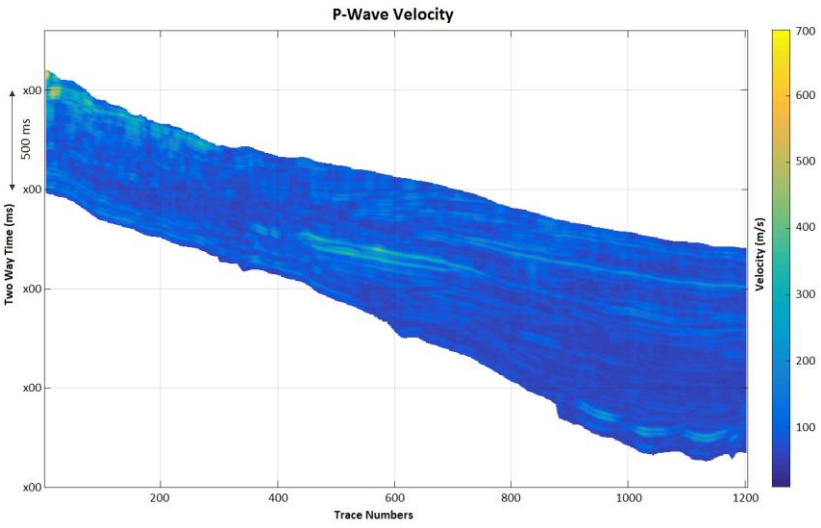



(b)

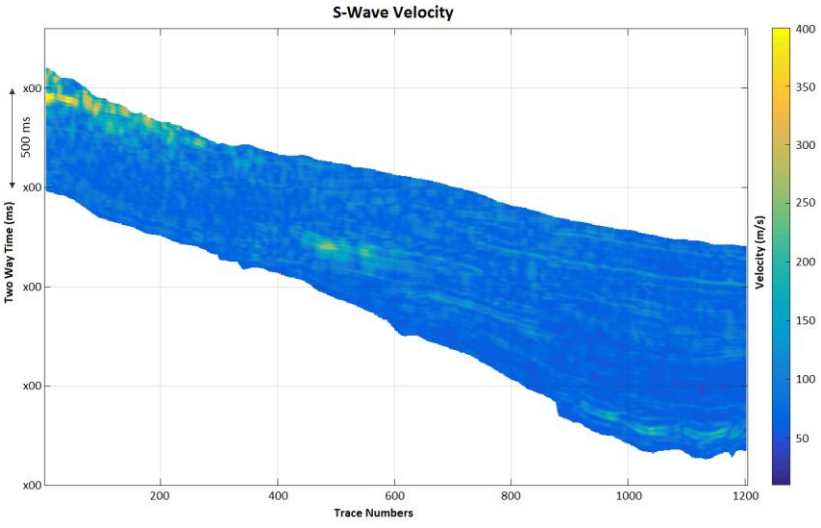



(c)

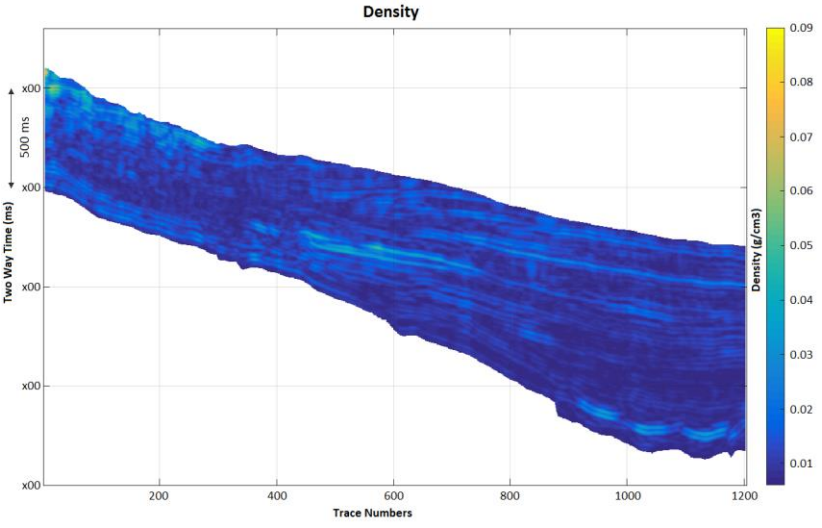
*Figure 7.18: The maximum-a-posteriori (MAP) estimates of elastic properties (a) $V_p$, (b) $V_s$, and (c) $\rho$.*




*(a)*




*(b)*




*(c)*

*Figure 7.19: Posterior standard deviation of elastic properties (a) $V_p$, (b) $V_s$, and (c) $\rho$.*

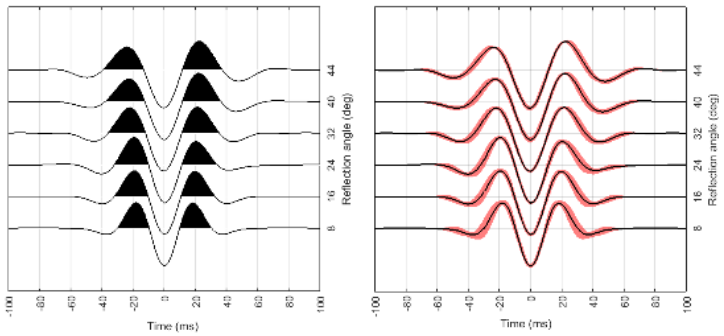



*Figure 7.20: Posterior MAP estimates and standard deviation of seismic wavelets corresponding to each of the input partial-angle seismic stacks in figure 7.4.*

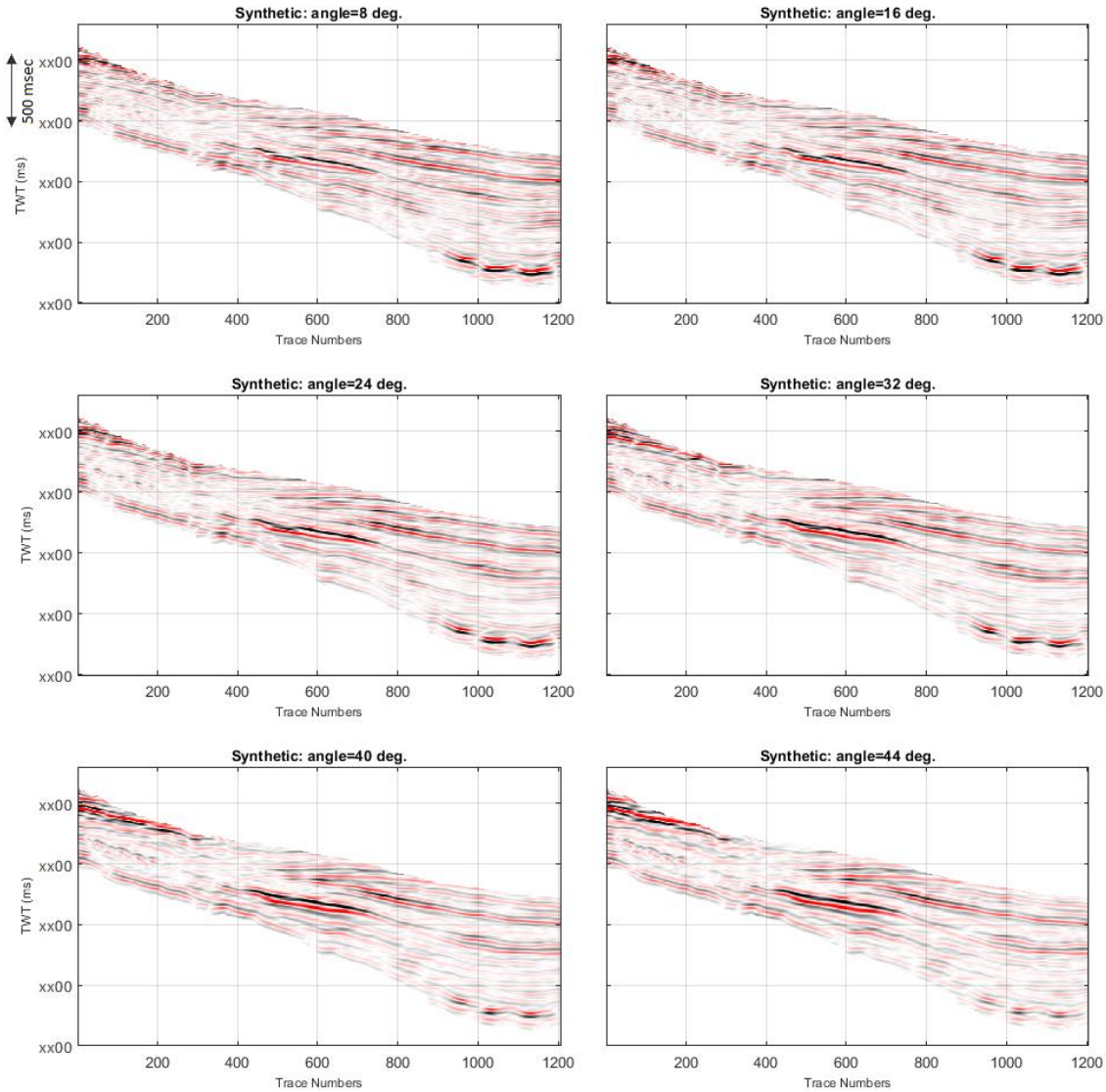



*Figure 7.21: Simulated seismic sections computed from the posterior MAP estimates of elastic properties corresponding to each of the input partial-angle seismic stacks in figure 7.4.*

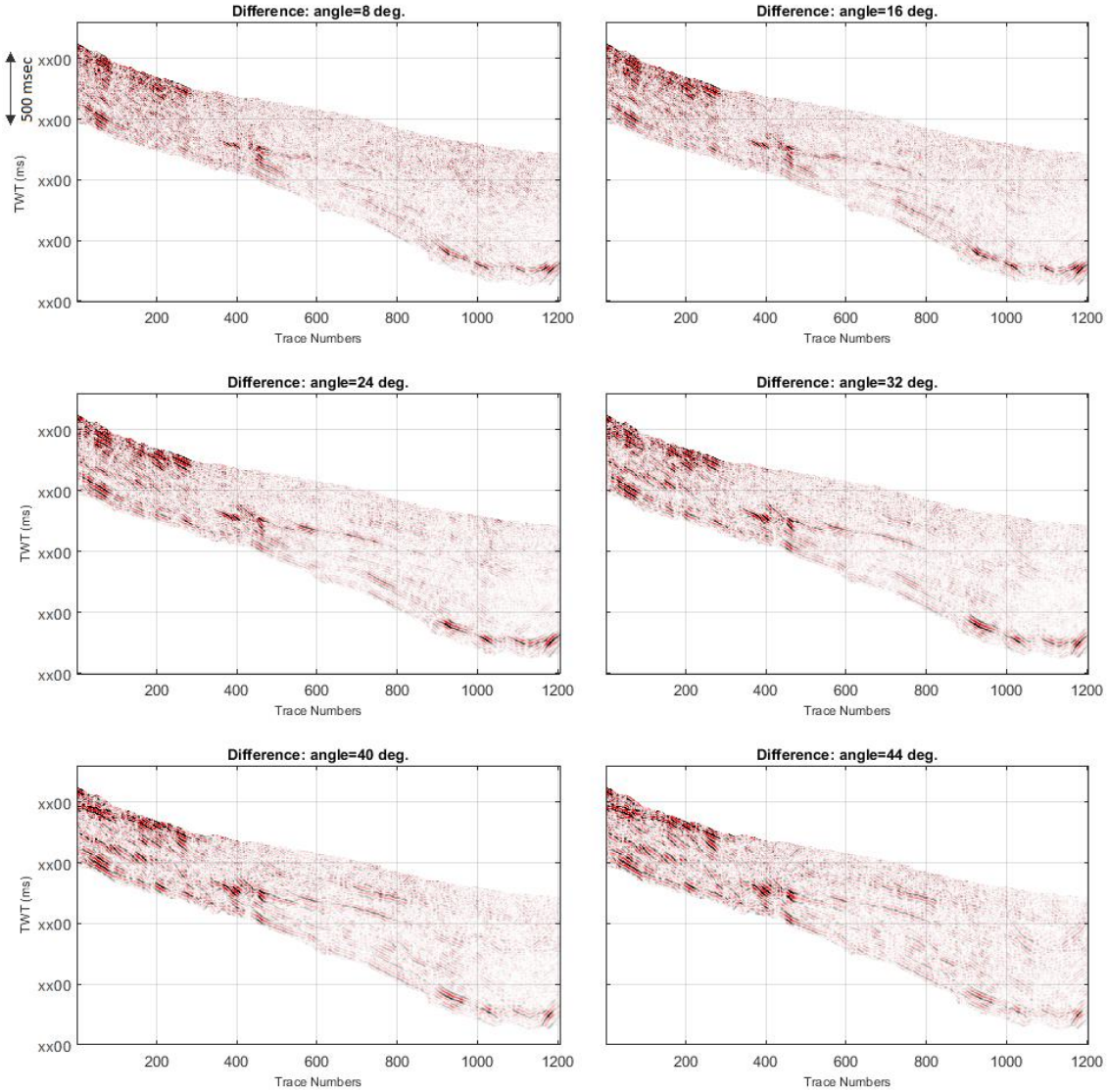



*Figure 7.22: Difference between observed (figure 7.4) and simulated seismic sections (figure 7.21). These differences represent errors in the seismic data that are not explained by the forward model.*

## 7.8 Discussion

*Markov-chain Monte Carlo* (McMC) method for stochastic sampling is general in its application and requires little modification to remain applicable even if the prior and/or likelihood distributions in an inverse problem need to be changed. This is not the case with variational Bayes (VB) method. If any of the distributions (e.g. prior or likelihood) are to be changed in a VBI method, the fixed point equation pertaining to the new distribution needs to be re-derived. If other distributions share some parameters with the distribution that is to be

changed, the entire system of fixed point equations may need to be re-derived analytically. Thus, a given VBI method is only applicable to a given set of distributions involved. In this chapter, we present the VBI method for a commonly used distribution for both prior and likelihood in geophysical applications – the *Normal distribution* (also known as the *Gaussian distribution*).

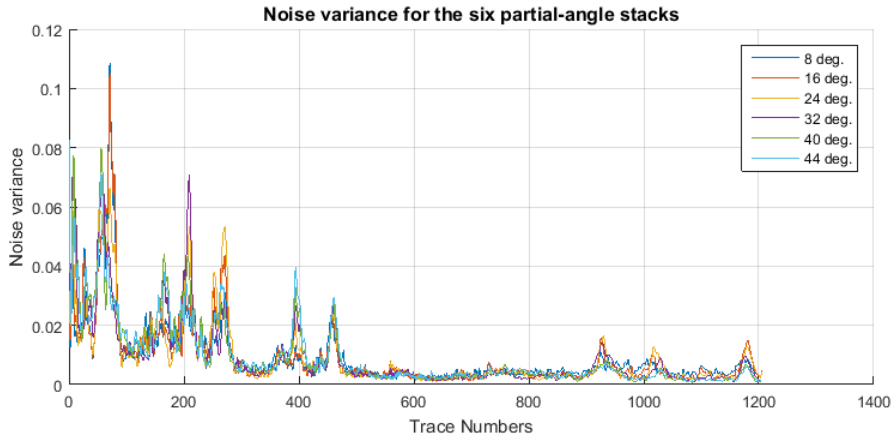



*Figure 7.23: Per-trace variance of noise (figure 7.22) for each of the partial-angle stack. This plot shows that the noise is mostly very low except at few locations where a significant amount of coherent energy could not be explained by the model.*

A factorized variational approximation $\mathcal{Q}(\boldsymbol{m}, \Theta)$ is proposed for the true but unknown posterior distribution $\mathcal{P}(\boldsymbol{m}, \Theta|\boldsymbol{d})$ of geological properties $\boldsymbol{m}$ and model parameters $\Theta$ given the observed geophysical data $\boldsymbol{d}$ using a hierarchical model. The variational approximation assumes conditional independence among the random variables $\boldsymbol{m}$ and $\Theta$ given the data $\boldsymbol{d}$. Such conditional independence assumptions allowed analytical derivation of the mean field (MF) update equations 7.32 and 7.33 in closed form. The computational efficiency of the current method mainly stems from these closed form solutions of the MF update equations. Additionally, computational advantage is also achieved by using a Gaussian Markov random field (GMRF) prior model which induces further conditional independence assumptions (the Markovian assumption) among geological properties at multiple locations. In fact, such assumptions are widely used in almost the entire literature on the solution of linearized spatial inverse problems, including the McMC based solutions of inverse problems with a non-linear forward problem (Rabben *et al.* 2008). Thus, the current method offers a more efficient solution within a set of assumptions that are commonly used in most (linearized) inverse problems in literature.

Non-existence of an inverse problem refers to the case when no valid solution lies within the restricted space of admissible solutions. It is particularly important for variational methods since these methods achieve computational advantage by restricting the space of admissible posterior distributions. In any application of the current method the form of prior and likelihood distributions used must be validated to ensure that these represent the prior information and the stochastic relationships between data and model parameters adequately. The use of Gaussian prior and likelihood is based on the repertoire of examples in geophysical literature where these distributions have been used successfully. Nevertheless, it is vital to ensure that these are applicable to the problem in hand. For example, the distribution of some of the reservoir properties may be skewed and/or multimodal, and may not therefore be modelled with a Gaussian prior. In such cases, Gamma distribution may be used to model a skewed distribution with a single mode, or any mixture distribution such as a *Gaussian mixture* (GM) distribution (Nawaz & Curtis, 2018; 2019) may be used to model a multimodal distribution. In such a case, the analytical solution given in section 7.5.2 may not be applicable, and one must re-derive the solution starting from the general MF update equations 7.32 and 7.33, or update the equations as appropriate. Such a lack of analytical generality of a variational Bayesian method is its major limitation compared to the more general McMC based methods. However, once the MF update equations are solved to give closed-form updates of the parameters involved, the effort spent in analytical treatment of the problem pays off in terms of computational efficiency of the VB methods.

Another important consideration in the application of this method is that it requires the forward problem to be linearizable. It is this linearization that allows probabilistic independence of the parameters $\Theta_m$ and $\Theta_{d|m}$ of the prior and likelihood distributions, respectively. Let us consider the travel-time tomography problem for example. The observed travel-times represent the data in this case while the velocities of the media represent the model parameters in this case. We know that the travel-times are function of velocities and the ray paths, where ray paths are themselves functions of velocities. In this case, the likelihood parameters $\Theta_{d|m}$ define the ray paths while the prior parameters define $\Theta_m$ velocities, both of which are strongly coupled. This makes the problem non-linear and therefore cannot be solved with the presented method in its current form. Another example of such non-linear problems is the full-waveform inversion (FWI) of geophysical data. An efficient

probabilistic solution of non-linear problems therefore requires further developments which we leave for future research.

In comparison to the McMC based inference that has a theoretical guarantee of asymptotic convergence as the number of samples tend to infinity, MF inference is guaranteed to converge to a local optimum in the coordinate space of random variables involved (Xing *et al*. 2003, Koller & Friedman, 2009). This is because each MF parameter update maximizes the lower bound $\mathcal{F}(Q)$ with respect to that coordinate given the mean values of the rest of the parameters. The overall iterative MF update thus works in a coordinate ascent manner (see figure 7.3) with a local convergence guarantee. However, global convergence is not guaranteed as the solution depends on the initial conditions and the order in which MF update equations are solved. To obtain a global optimum solution, the MF equations may be solved within a global optimization framework such as simulated annealing with multiple initializations and ordering of MF updates, both chosen randomly. However, this is typically not required for simple unimodal distributions such as a Gaussian as we used in this research.

## 7.9 Conclusions

A method for probabilistic inversion of geophysical data is introduced using variational Bayesian inference in a hierarchical model as an efficient alternative to the McMC based stochastic inversion methods. Besides the desired model properties, the hierarchical Bayesian inversion estimates the parameters of the prior and likelihood distributions as a part of the solution to the inverse problem. The variational Bayesian approach casts the probabilistic inference problem in an optimization framework, which is solved by the MF approximation in a coordinate ascent manner. The presented method jointly estimates the parameters of the forward model and the noise level in the data along with the solution of the inverse problem, while providing a quantitative assessment of posterior (post-inversion) uncertainties in these estimates. Our method avoids sampling of the solution space, yet provides fully probabilistic Bayesian results.

# Chapter 8 Joint Variational Bayesian Inversion for Facies and Rock Properties

## 8.1 Summary

In this chapter, an efficient probabilistic inversion method is introduced for joint estimation of geological facies (discrete litho-fluid classes) and petrophysical rock properties such as porosity, clay volume and water saturation, from seismic data attributes (derived quantities) such as P-wave and S-wave impedances and Vp/Vs ratios. Similar to the previous facies inversion methods presented in this thesis, the current method also honours spatial correlations in geological facies that are supplied as prior information. Additionally, mutual probabilistic dependence among various seismic attributes and petrophysical rock properties are also honoured through spatial correlations in facies. Seismic attributes and petrophysical properties are jointly modelled using a *Gaussian mixture* (GM) distribution whose parameters are initialized by unsupervised learning using well-log data. Rock physical models may be used to augment the training data if the existing well data are limited, however this is not required if sufficient well data are available. The joint posterior distribution of petrophysical rock properties and geological facies given the observed seismic attributes is updated in an iterative fashion. The variational Bayesian inversion method introduced in chapter 5 is extended here to circumvent the need for stochastic sampling, while still providing full probabilistic results. The application of this method is demonstrated on a real data example from the North Sea.

## 8.2 Introduction

3D Seismic data offer an extensive coverage of the subsurface and provide essential information required to build models of subsurface fluid reservoirs. Such models are used for estimation of reserves and for making decisions regarding development of subsurface resources. At the very least, the structural architecture of a reservoir may be defined based on geological interpretation of 3D seismic data. Additional information in the form of spatial distribution of geological facies (discrete litho-fluid types) and petrophysical rock properties (continuous physical properties of rocks such as porosity and permeability) is also required for

quantitative reservoir characterization, and in establishing a meaningful link between various features of seismic data, and the static and dynamic characteristics of subsurface fluid reservoirs. However, such information cannot be inferred from seismic data directly, and must be obtained from other sources of information such as well data. Since well data are usually limited and sparse, we need to perform mapping of these properties over the entire reservoir. Such a mapping is usually performed by inversion of seismic data to ensure that the mapped properties are consistent with the seismic data.

For a given geological facies, petrophysical rock properties are often well correlated with seismic attributes, such as P-wave and S-wave impedances. Therefore, seismic waveform data and their attributes provide useful constraints on the spatial distribution of both geological facies and petrophysical rock properties. Examples of seismic attributes are P-wave and S-wave velocities ($V_p$ and $V_s$) and impedances ($I_p$ and $I_s$), the ratio of P-wave to S-wave velocity ($\gamma \equiv V_p/V_s$), Poisson's ratio ($\sigma$), density ($\rho$), Lame's coefficients ($\lambda$ and μ), and amplitude variation with offset (AVO) attributes such as intercept ($A$), gradient ($B$) and their product ($A * B$). Examples of petrophysical properties are porosity ($\varphi$), volume of clay ($V_{cl}$) in siliciclastic reservoirs, and pore space water saturations ($S_w$). Although seismic attributes are generally estimated from the observed seismic waveform data, we refer to them as the *observed data* since these are considered as fixed inputs to our method. The elastic rock properties (or seismic attributes) and the petrophysical rock properties are together referred to as rock properties. Petrophysical rock properties and geological facies are henceforth together referred to as *model parameters* of interest.

Estimation of petrophysical rock properties from seismic attributes is a non-unique inverse problem, but it can be regularized in a meaningful way if the solution can be constrained by the distribution of geological facies. Further, discrimination of geological facies from the seismic attributes may be improved if petrophysical rock properties are estimated and as such can be regarded as (uncertain) data along with the seismic attributes. Thus knowledge of either facies or petrophysical properties helps in the discrimination or estimation of the other. Since both of these are unknown, their inference from seismic attributes is a joint, usually nonlinear problem. In this chapter, we solve this nonlinear problem in an iterative fashion, by alternately estimating one of these unknowns from the current estimate of the other in each iteration, with the objective of improving the overall joint model. The method presented in chapter 5 (also see Nawaz & Curtis, 2018) is extended here to estimate the spatial

distribution of both petrophysical rock properties and facies from seismic attributes jointly, by using *variational Bayesian inversion* (VBI). This avoids extensive sampling during inference, yet provides fully probabilistic Bayesian results.

The rest of this chapter is organized as follows. The Bayesian inversion framework is first formulated for the current problem in section 8.3. The Markov random field (MRF, see sections 3.5 and 5.3.1) model for prior distribution of spatially coupled (probabilistically dependent) facies is reviewed in section 8.3.1. The quasi-localized likelihoods model that was proposed in section 5.3.2 is extended in section 8.3.2 to include petrophysical rock properties. Then the variational Bayes (VB) method is presented in section 8.4 for joint estimation of spatial distributions of geological facies and petrophysical rock properties from seismic attributes. After providing the mathematical details of this method, a real data example is provided from the North Sea in section 8.6, where the inversion results are first shown for a gas reservoir on well-log data and then across a 2D seismic attributes section. The data example is followed by a discussion on the method in section 8.7, and finally the conclusions in section 8.8.

## 8.3 Model

We want to infer petrophysical rock properties $\boldsymbol{r}$ and facies $\boldsymbol{\kappa}$ jointly from the seismic attributes $\boldsymbol{d}$ along with their associated uncertainty of prediction. In terms of probability theory, we seek the so called *posterior distribution* $\mathcal{P}(\boldsymbol{r}, \boldsymbol{\kappa}|\boldsymbol{d})$ of unknown model parameters $\boldsymbol{r}$ and $\boldsymbol{\kappa}$ conditioned on the realized data $\boldsymbol{d}$. For this purpose, we use the *generative modelling* approach as was used in chapter 5. The forward model is usually a deterministic or stochastic relationship that can be used to express the likelihood $\mathcal{P}(\boldsymbol{d}|\boldsymbol{r}, \boldsymbol{\kappa})$ of data given the unknown model parameters. For the observed data, this conditional distribution is called the *data likelihood*. The posterior distribution $\mathcal{P}(\boldsymbol{r}, \boldsymbol{\kappa}|\boldsymbol{d})$ and the data likelihood $\mathcal{P}(\boldsymbol{d}|\boldsymbol{r}, \boldsymbol{\kappa})$ are related according to Bayes' theorem as

$$\mathcal{P}(\boldsymbol{r}, \boldsymbol{\kappa}|\boldsymbol{d}) = \frac{\mathcal{P}(\boldsymbol{d}|\boldsymbol{r}, \boldsymbol{\kappa})\mathcal{P}(\boldsymbol{r}|\boldsymbol{\kappa})\mathcal{P}(\boldsymbol{\kappa})}{\mathcal{P}(\boldsymbol{d})} \tag{8.1}$$

where $\mathcal{P}(\boldsymbol{\kappa})$ represents the prior distribution of facies, $\mathcal{P}(\boldsymbol{r}|\boldsymbol{\kappa})$ represents the conditional prior distribution of the petrophysical properties $\boldsymbol{r}$ given a particular facies model $\boldsymbol{\kappa}$, and $\mathcal{P}(\boldsymbol{d})$ represents the marginal probability of data $\boldsymbol{d}$ – the *evidence* which is given by

$$\mathcal{P}(\boldsymbol{d}) = \sum_{\boldsymbol{\kappa}} \int \mathcal{P}(\boldsymbol{d}|\boldsymbol{r},\boldsymbol{\kappa})\mathcal{P}(\boldsymbol{r}|\boldsymbol{\kappa})\mathcal{P}(\boldsymbol{\kappa})d\boldsymbol{r} \quad 8.2$$

Below, we first describe a model for the prior distribution $\mathcal{P}(\boldsymbol{\kappa})$ of facies in subsection 8.3.1, and then we merge the conditional prior distribution $\mathcal{P}(\boldsymbol{r}|\boldsymbol{\kappa})$ and the likelihood $\mathcal{P}(\boldsymbol{d}|\boldsymbol{r},\boldsymbol{\kappa})$ to form the joint distribution $\mathcal{P}(\boldsymbol{d},\boldsymbol{r}|\boldsymbol{\kappa})$ of rock properties $\boldsymbol{d}$ and $\boldsymbol{r}$ given facies $\boldsymbol{\kappa}$ in subsection 8.3.2.

## *8.3.1 Facies Prior Model*

Here, we use the same pairwise Markov random field (MRF) model to encode prior information about spatial distribution of geological facies as was used in chapter 5. A pairwise MRF factorizes (according to *Hammerseley-Clifford* theorem: Besag, 1974) into pairwise potential functions $\psi_{ij}(\kappa_i,\kappa_j)$ called *edge potentials*, such that the prior distribution $\mathcal{P}(\boldsymbol{\kappa})$ of facies $\boldsymbol{\kappa}$ may be expressed as

$$\mathcal{P}(\boldsymbol{\kappa}) = \frac{1}{Z}\prod_{(i,j)\in\mathcal{E}} \psi_{ij}(\kappa_i,\kappa_j) \quad 8.3$$

which is same as equations 3.11 and 5.3. The prior conditional probability of occurrence of facies $\kappa_i$ at a location $i$ in the model given the facies $\boldsymbol{\kappa}_{\mathcal{N}_{\setminus i}}$ in its neighbourhood $\mathcal{N}_{\setminus i}$ is therefore given by equation 5.4 as

$$\mathcal{P}\left(\kappa_i|\boldsymbol{\kappa}_{\mathcal{N}_{\setminus i}}\right) \propto \prod_{j\in\mathcal{N}_{\setminus i}} \psi_{ij}(\kappa_i,\kappa_j) \quad 8.4$$

which defines the spatial coupling of facies in terms of the pairwise clique potential functions $\psi_{ij}(\kappa_i,\kappa_j)$.

## *8.3.2 Likelihood Model*

Two main approaches are used for modelling the relationship between data and model parameters: physics based modelling and the data driven modelling. Physics based models define a mapping from the model parameters to the observed data based on the physics of the problem. Such models are always semi-empirical in that they contain free parameters that are tuned such that the derived model matches observed examples of model parameter values

and corresponding data. Examples are the parameterized empirical Gardner relationship between density and seismic velocity (Gardner *et al.* 1974), and the soft-sand and stiff-sand rock physics models (Dvorkin & Nur, 1996) with Gaussian distributed noise. Such models typically require a small number (often 3 or 4) of parameters to be calibrated to fit petro-elastic data (e.g. $V_p$ and $\varphi$) from siliciclastic rocks. On the other hand, the data driven approach defines and fits a non-parametric model to the observed samples – a model which cannot be defined in terms of a finite number of parameters. An example of a data driven model is non-parametric *kernel mixture density* (Grana, 2018) that fits a pre-specified base function (the kernel function) at each data point to approximate any complex probability distribution.

The physics based approach may allow intuitive interpretation of the observed data, for example, fitting the soft-sand and stiff-sand models to petro-elastic data (e.g. $V_p$ and $\varphi$) may provide information about the compactness of the rocks under investigation. However, for this to be possible the models need to be simple, and consequently they may not capture salient features of any particular dataset. This may lead to inaccurate estimation of posterior (post-inference) uncertainties of the model parameters conditioned to the observed data. The data driven models incorporate little or no physical intuition about the relationship between model parameters and observed data, however they are flexible in the level of detail that they can capture. Also, in contrast to physics based models which are often valid only for a particular type of geology, data driven models may be applied to any geology. However, data driven models may easily over-fit the data and consequently result in biased posterior estimates of the model parameters.

We use a middle ground: a *Gaussian mixture model* (GMM), which is a semi-parametric way of representing an arbitrarily complex and possibly multimodal distribution. A GMM defines a *Gaussian mixture* (GM) distribution as a linear combination (weighted sum) of Gaussian probability density functions (PDF). It is similar to the kernel mixture density with Gaussian kernels but it typically requires a much smaller number of kernels than the number of data points to be fit. For a random variable $\boldsymbol{x}$, a GM distribution with $T$ components may be expressed by the following PDF:

$$\mathcal{P}(\boldsymbol{x}) = \sum_{t=1}^{T} \alpha_t \, g_t(\boldsymbol{x}) = \sum_{t=1}^{T} \alpha_t \, N(\boldsymbol{x} \mid \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t) \qquad 8.5$$

where $g_t(\boldsymbol{x}) = N(\boldsymbol{x} \mid \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$ represents a Gaussian PDF with mean $\boldsymbol{\mu}_t$ and covariance matrix $\boldsymbol{\Sigma}_t$, and $\alpha_t$ is the weight of the $t^{\text{th}}$ component of the mixture. A GM distribution is a universal approximator of PDFs: given a sufficient number of Gaussian kernels with appropriate parameters, it can approximate any complex PDF to any desired non-zero accuracy (McLachlan & Peel, 2000).

GM distributions have been widely used to model the distribution of rock properties in geophysical literature (e.g. Meier *et al*. 2007a, b, and 2009; Grana & Della Rossa, 2010; Shahraeeni & Curtis, 2011; Grana *et al*. 2017; Nawaz & Curtis, 2017 & 2018). Shahraeeni & Curtis (2011) used a *mixture density network* (MDN) (Bishop, 2995) which is a type of neural network that can be trained to emulate a desired conditional distribution with a GM distribution. They used it to compute cell-wise posterior distributions of petrophysical properties conditioned on the observed seismic attributes in each model cell after the network is trained on well data. In the current work, we use a variant of the *expectation-maximization* (EM) algorithm (Dempster *et al*. 1977, Nawaz & Curtis, 2018) to model the joint distribution of all rock properties (elastic and petrophysical) as a GM distribution. The posterior distribution of petrophysical properties conditioned on the observed seismic attributes may then be obtained analytically; by marginalizing or by conditioning on the joint distribution depending on whether the data (seismic attributes or elastic rock properties) uncertainties are included in the model or not or not, respectively. As opposed to the MDN approach that uses supervised learning from training examples, the presented method is based on unsupervised learning and is computationally more efficient as it avoids the computational cost of generating and learning from training examples.

A rock physics model is usually used to relate elastic properties and corresponding petrophysical properties. However, if sufficient well coverage is available the joint distribution of rock properties may be estimated directly from the well data, i.e. without requiring a rock physics model. This allows the correlation between any combination of rock properties, as well as the variance of each of the rock properties to be captured. The conditional prior distribution $\mathcal{P}(\boldsymbol{r}|\boldsymbol{\kappa})$ of petrophysical rock properties $\boldsymbol{r}$ given geological facies $\boldsymbol{\kappa}$ is usually modelled using well logs that have been up-scaled at the dominant seismic wavelength relative to seismic attributes $\boldsymbol{d}$ (Grana & Della Rossa, 2010), and the likelihood $\mathcal{P}(\boldsymbol{d}|\boldsymbol{r}, \boldsymbol{\kappa})$ is usually modelled using rock physics models (Bosch *et al*. 2010; Grana & Della Rossa, 2010; Lang & Grana, 2018; Grana, 2018) calibrated with the well data and local geological information. We adopt a

different approach: we model both of the conditional prior $\mathcal{P}(\boldsymbol{r}|\boldsymbol{\kappa})$ and the likelihood $\mathcal{P}(\boldsymbol{d}|\boldsymbol{r},\boldsymbol{\kappa})$ jointly using up-scaled well-logs in the form of a joint distribution $\mathcal{P}(\boldsymbol{d},\boldsymbol{r}|\boldsymbol{\kappa};\Theta)$ of elastic attributes $\boldsymbol{d}$ and petrophysical properties $\boldsymbol{r}$ given the facies $\boldsymbol{\kappa}$, defined in terms of a set of parameters $\Theta$ which we will define and estimate below. Therefore, the current method does not require a rock physics model to be used. However, if well coverage is limited, available well data may be augmented by using an appropriate rock physics model prior to the estimation of the joint PDF of rock properties.

We adopt the quasi-localized likelihoods model of Nawaz & Curtis (2018) where rock properties $\boldsymbol{d}_i$ and $\boldsymbol{r}_i$ in each cell $i$ are conditioned on the facies $\boldsymbol{\kappa}_{\mathcal{N}_i}$ in some pre-specified neighbourhood $\mathcal{N}_i$ of $i$. The quasi localized likelihoods defined in this manner, $\mathcal{P}\left(\boldsymbol{d}_i,\boldsymbol{r}_i|\boldsymbol{\kappa}_{\mathcal{N}_i};\Theta\right)$, may be very high dimensional depending on the size of the neighbourhood structure $\mathcal{N}_i$. This may increase the computational cost of the method significantly. However, since facies in the neighbouring locations tend to be similar in a MRF model, there is a high probability that any one facies dominates other facies within any neighbourhood. This suggests that we can reduce the dimensionality of quasi-localized likelihoods by defining the most probable facies $\hat{\kappa}_i$ in cell $i$ as the one that maximizes the sum of some estimate of marginal probabilities $\hat{\mathcal{P}}\left(\kappa_j\right)$ of facies $\kappa_j$ at locations $j \in \mathcal{N}_i$, i.e.

$$\hat{\kappa}_i = \underset{\kappa}{argmax} \sum_{j\in\mathcal{N}_i} \hat{\mathcal{P}}\left(\kappa_j\right) = \underset{\kappa}{argmax} \sum_{j\in\mathcal{N}_i} \sum_{\hat{\boldsymbol{\kappa}}_{\mathcal{N}_{\setminus j}}} \mathcal{P}\left(\kappa_j|\hat{\boldsymbol{\kappa}}_{\mathcal{N}_{\setminus j}}\right) \tag{8.6}$$

where $\mathcal{P}\left(\kappa_j|\hat{\boldsymbol{\kappa}}_{\mathcal{N}_{\setminus j}}\right)$ is the prior probability of facies $\kappa_j$ at a location $j$ given some estimate $\hat{\boldsymbol{\kappa}}_{\mathcal{N}_{\setminus j}}$ of the facies $\boldsymbol{\kappa}_{\mathcal{N}_{\setminus j}}$ in the neighbourhood $\mathcal{N}_{\setminus j}$ of $j$ given by equation 8.4.

Since the prior distribution $\mathcal{P}(\boldsymbol{\kappa})$ of facies is expressed as a Gibbs distribution, it factorizes over cliques in the model according to equation 8.3. A similar factorization of $\mathcal{P}(\boldsymbol{d},\boldsymbol{r}|\boldsymbol{\kappa};\Theta)$ can be achieved by assuming conditional independence of rock properties ($\boldsymbol{d}$ and $\boldsymbol{r}$) given the facies $\boldsymbol{\kappa}$ such that

$$\mathcal{P}(\boldsymbol{d},\boldsymbol{r}|\boldsymbol{\kappa};\Theta) = \prod_{i\in\mathcal{V}} \mathcal{P}\left(\boldsymbol{d}_i,\boldsymbol{r}_i|\boldsymbol{\kappa}_{\mathcal{N}_i};\Theta\right) \cong \prod_{i\in\mathcal{V}} \mathcal{P}(\boldsymbol{d}_i,\boldsymbol{r}_i|\hat{\kappa}_i;\Theta) \tag{8.7}$$

The probability of $\boldsymbol{d}$ given $\boldsymbol{\kappa}$ may then be expressed as

$$\mathcal{P}(\boldsymbol{d}|\boldsymbol{\kappa};\Theta) = \prod_{i\in\mathcal{V}} \int \mathcal{P}\big(\boldsymbol{d}_i, \boldsymbol{r}_i|\boldsymbol{\kappa}_{\mathcal{N}_i};\Theta\big) d\boldsymbol{r}_i \cong \prod_{i\in\mathcal{V}} \int \mathcal{P}(\boldsymbol{d}_i, \boldsymbol{r}_i|\hat{\kappa}_i;\Theta)\, d\boldsymbol{r}_i \equiv \prod_{i\in\mathcal{V}} \varphi_i(\hat{\kappa}_i) \qquad 8.8$$

where $\varphi_i(\kappa_i) \equiv \int \mathcal{P}(\boldsymbol{d}_i, \boldsymbol{r}_i|\hat{\kappa}_i;\Theta)\, d\boldsymbol{r}_i$ is a potential function of $\kappa_i$ referred to as the *vertex potential* in a MRF model. It models the likelihood of observing seismic attributes $\boldsymbol{d}_i$ and current estimate of petrophysical properties $\boldsymbol{r}_i$ at a location $i$ which may be regarded as the up-scaled response of facies $\boldsymbol{\kappa}_{\mathcal{N}_i}$ within the neighbourhood of $i$ (Nawaz & Curtis, 2018). If the estimate of marginal probability $\hat{\mathcal{P}}(\kappa_j)$ in equation 8.6 is obtained from the current estimate of posterior marginal distribution of facies in cell $i$, the approximations 8.7 and 8.8 correspond to the notion of *empirical Bayes*.

Petrophysical rock properties are usually obtained from well log data, and are therefore much higher in resolution compared to the seismic attributes. In order to account for the difference in resolution, the rock properties $\boldsymbol{x}_i = [\boldsymbol{d}_i, \boldsymbol{r}_i]$ at a location $i$ are assumed to be a weighted linear combination of the corresponding high resolution rock properties $\boldsymbol{h}_j$ at the neighbouring locations $j \in \mathcal{N}_i$ such that

$$\boldsymbol{x}_i = \sum_{j\in\mathcal{N}_i} \beta_j \boldsymbol{h}_j + \boldsymbol{\varepsilon}_i \qquad 8.9$$

where $\boldsymbol{x}_i$ is a $p \times 1$ vector of rock properties (seismic attributes $\boldsymbol{d}_i$ and the petrophysical properties $\boldsymbol{r}_i$), $\beta_j$ are the regression coefficients, and $\boldsymbol{\varepsilon}_i$ is a vector of errors which are assumed to be jointly distributed according to a *Normal distribution* $N(0, \boldsymbol{\Sigma}_\varepsilon)$. The regression coefficients $\beta_j$ in this expression act as coefficients of a spatial averaging filter, and may be estimated within the inversion process (Nawaz & Curtis, 2018), or may be fixed *a priori* based on vertical averaging of well-logs at the seismic wavelengths.

We use a Gaussian mixture (GM) distribution to model $\mathcal{P}(\boldsymbol{d}_i, \boldsymbol{r}_i|\hat{\kappa}_i;\Theta)$ that is defined as a linear combination of a given number of Gaussian kernels, usually referred to as the components of the mixture distribution. Defining $\boldsymbol{x}_i \equiv [\boldsymbol{d}_i, \boldsymbol{r}_i]^T$, i.e. a vector of rock properties in cell $i$, the GM distribution is expressed as

$$\mathcal{P}(\boldsymbol{x}_i|\hat{\kappa}_i = k;\Theta) = \sum_{t=1}^{T_k} \alpha_{t,k}\, g_{t,k}(\boldsymbol{x}_i)\,, \qquad \forall i \in \mathcal{V} \qquad 8.10$$

where $T_k$ is the number of mixture components (which may be different for each facies $k$), $\alpha_{t,k}$ is the component weight and is included in $\Theta$, and $g_{t,k}(\boldsymbol{x}_i)$ is the Gaussian kernel for the $t^{\text{th}}$ component and facies $\hat{\kappa}_i = k$. The Gaussian kernels $g_{t,k}(\boldsymbol{x}_i)$ are given by

$$g_{t,k}(\boldsymbol{x}_i) = g_{t,k}\left(\begin{bmatrix}\boldsymbol{d}_i\\ \boldsymbol{r}_i\end{bmatrix}\right) = N\left(\begin{bmatrix}\boldsymbol{\mu}_d\\ \boldsymbol{\mu}_r\end{bmatrix}_{t,k}, \begin{bmatrix}\boldsymbol{\Sigma}_{d,d} & \boldsymbol{\Sigma}_{d,r}\\ \boldsymbol{\Sigma}_{r,d} & \boldsymbol{\Sigma}_{r,r}\end{bmatrix}_{t,k}\right), \qquad \forall i \in \mathcal{V} \tag{8.11}$$

where $N$ represents the probability density function (PDF) of the Normal distribution, $\boldsymbol{\mu}$'s and $\boldsymbol{\Sigma}$'s are means and block covariance matrices of the kernel (and are also included in $\Theta$) with subscripts indicating the data $\boldsymbol{d}$ or the petrophysical properties $\boldsymbol{r}$ components of $\boldsymbol{x}_i$. The expression for a Gaussian kernel may also be expressed explicitly as

$$g_{t,k}(\boldsymbol{x}_i) = (2\pi)^{-p/2}\left|\boldsymbol{\Sigma}_{t,k}\right|^{-1/2} \exp\left\{\frac{-1}{2}\left(\boldsymbol{x}_i - \boldsymbol{\mu}_{t,k}\right)^T {\boldsymbol{\Sigma}_{t,k}}^{-1}\left(\boldsymbol{x}_i - \boldsymbol{\mu}_{t,k}\right)\right\}, \qquad \forall i \in \mathcal{V} \tag{8.12}$$

where $p$ is the dimensionality of $\boldsymbol{x}_i$, and $\boldsymbol{\mu}_{t,k}$ and $\boldsymbol{\Sigma}_{t,k}$ are mean and covariance matrix of the kernel $g_{t,k}(\boldsymbol{x}_i)$ given by

$$\boldsymbol{\mu}_{t,k} = \begin{bmatrix}\boldsymbol{\mu}_d\\ \boldsymbol{\mu}_r\end{bmatrix}_{t,k} \tag{8.13}$$

and

$$\boldsymbol{\Sigma}_{t,k} = \begin{bmatrix}\boldsymbol{\Sigma}_{d,d} & \boldsymbol{\Sigma}_{d,r}\\ \boldsymbol{\Sigma}_{r,d} & \boldsymbol{\Sigma}_{r,r}\end{bmatrix}_{t,k} \tag{8.14}$$

Since the joint conditional distribution $\mathcal{P}(\boldsymbol{d}, \boldsymbol{r}|\boldsymbol{\kappa}; \Theta)$ of seismic attributes $\boldsymbol{d}$ and rock properties $\boldsymbol{r}$ given facies $\boldsymbol{\kappa}$ (and the distribution parameters $\Theta$) is modelled as a GM distribution, and the prior distribution of facies $\mathcal{P}(\boldsymbol{\kappa})$ is modelled as a MRF, the overall model of joint distribution $\mathcal{P}(\boldsymbol{d}, \boldsymbol{r}, \boldsymbol{\kappa}; \Theta)$ of the data $\boldsymbol{d}$ and model parameters $\boldsymbol{r}$ and $\boldsymbol{\kappa}$ represents a *Gaussian mixture - Markov random field* (GM-MRF). The parameters $\Theta$ may be defined as $\Theta \equiv \{\alpha_{t,k}, \boldsymbol{\mu}_{t,k}, \boldsymbol{\Sigma}_{t,k}\}, \forall t, k$. We may initialize $\Theta$ using some training data (e.g. up-scaled well logs) and, as we show in section 8.4, $\Theta$ may be updated as a part of the inversion process.

### *8.3.3 Posterior Model*

The posterior distribution in equation 8.1 may be written as

$$\mathcal{P}(\boldsymbol{r},\boldsymbol{\kappa}|\boldsymbol{d};\Theta)=\frac{\mathcal{P}(\boldsymbol{d},\boldsymbol{r}|\boldsymbol{\kappa};\Theta)\mathcal{P}(\boldsymbol{\kappa})}{\mathcal{P}(\boldsymbol{d};\Theta)} \tag{8.15}$$

Substituting equations 8.3 and 8.7 into equation 8.15 we get

$$\mathcal{P}(\boldsymbol{r},\boldsymbol{\kappa}|\boldsymbol{d};\Theta)=\frac{\mathcal{P}(\boldsymbol{d},\boldsymbol{r},\boldsymbol{\kappa};\Theta)}{\mathcal{P}(\boldsymbol{d};\Theta)}\cong\frac{1}{\mathcal{Z}'}\prod_{i\in\mathcal{V}}\mathcal{P}(\boldsymbol{d}_i,\boldsymbol{r}_i|\hat{\kappa}_i;\Theta)\prod_{(i,j)\in\mathcal{E}}\psi_{ij}(\kappa_i,\kappa_j) \tag{8.16}$$

where $\mathcal{P}(\boldsymbol{d};\Theta)$ has been absorbed in the normalization constant $\mathcal{Z}'$ on the right hand side. This demonstrates that although we only assumed that the prior distribution $\mathcal{P}(\boldsymbol{\kappa})$ on facies $\boldsymbol{\kappa}$ is a MRF, the posterior distribution $\mathcal{P}(\boldsymbol{r},\boldsymbol{\kappa}|\boldsymbol{d};\Theta)$ and the joint distribution $\mathcal{P}(\boldsymbol{d},\boldsymbol{r},\boldsymbol{\kappa};\Theta)$ then also turn out to be MRFs. This is a consequence of the spatial conditional independence assumption on rock properties $\boldsymbol{d}$ and $\boldsymbol{r}$, and we show in section 8.4 that such a factorization of the posterior distribution is crucial for making inference tractable for real-scale models.

## 8.4 Variational Bayesian (VB) Inference

We use the variational Bayes (VB) method to approximate the intractable posterior distribution $\mathcal{P}(\boldsymbol{r},\boldsymbol{\kappa}|\boldsymbol{d};\Theta)$ by a tractable variational distribution $\mathcal{Q}(\boldsymbol{r},\boldsymbol{\kappa}|\boldsymbol{d})$, or simply $\mathcal{Q}$. Probabilistic inference can then be performed in an optimization framework, as discussed in section 2.4. Any choice of the variational distribution $\mathcal{Q}$ can be used to define a lower bound $\mathcal{F}(\mathcal{Q},\Theta)$ on the log-evidence $\mathcal{L}(\Theta;\boldsymbol{d})$ (Neal & Hinton, 1998; Beal, 2003; Nawaz & Curtis, 2018), such that

$$\mathcal{L}(\Theta;\boldsymbol{d})=\mathcal{F}(\mathcal{Q},\Theta)+KL\big(\mathcal{Q}||\mathcal{P}(\boldsymbol{r},\boldsymbol{\kappa}|\boldsymbol{d};\Theta)\big) \tag{8.17}$$

where the lower bound $\mathcal{F}(\mathcal{Q},\Theta)$ is called the variational free energy or simply free energy. The term $KL\big(\mathcal{Q}||\mathcal{P}(\boldsymbol{r},\boldsymbol{\kappa}|\boldsymbol{d};\Theta)\big)\geq 0$ is the *Kullback-Liebler (KL) divergence* between $\mathcal{Q}$ and $\mathcal{P}(\boldsymbol{r},\boldsymbol{\kappa}|\boldsymbol{d};\Theta)$, which is given by

$$KL\big(\mathcal{Q}||\mathcal{P}(\boldsymbol{r},\boldsymbol{\kappa}|\boldsymbol{d};\Theta)\big)=\mathbb{E}_{\mathcal{Q}}\left[\log\frac{\mathcal{Q}(\boldsymbol{r},\boldsymbol{\kappa}|\boldsymbol{d})}{\mathcal{P}(\boldsymbol{r},\boldsymbol{\kappa}|\boldsymbol{d};\Theta)}\right]=\sum_{\boldsymbol{\kappa}}\int\mathcal{Q}(\boldsymbol{r},\boldsymbol{\kappa}|\boldsymbol{d})\log\frac{\mathcal{Q}(\boldsymbol{r},\boldsymbol{\kappa}|\boldsymbol{d})}{\mathcal{P}(\boldsymbol{r},\boldsymbol{\kappa}|\boldsymbol{d};\Theta)}d\boldsymbol{r} \tag{8.18}$$

Although $\mathcal{L}(\Theta;\boldsymbol{d})$ is intractable, its lower bound $\mathcal{F}(\mathcal{Q},\Theta)$ may be estimated for a suitably chosen $\mathcal{Q}$. An iterative scheme may be devised to estimate $\mathcal{L}(\Theta;\boldsymbol{d})$ by successively updating $\mathcal{Q}$ and $\Theta$ in each iteration. For example, a variational form of the *expectation-maximization* (EM)

algorithm (Dempster *et al.* 1977) may be used to approximate $\mathcal{L}(\Theta;\boldsymbol{x})$ in an iterative fashion such that its lower bound $\mathcal{F}(\mathcal{Q},\Theta)$ is increased while decreasing $KL\big(\mathcal{Q}||\mathcal{P}(\boldsymbol{r},\boldsymbol{\kappa}|\boldsymbol{d};\Theta)\big)$ for a given set of parameters $\Theta$ in each iteration (see sections 2.4 and 5.4.1).

We use the same VBI method that was used in section 5.4, except that here we model the joint distribution of seismic attributes and petrophysical rock properties as a Gaussian mixture distribution, instead of just the seismic attributes. Also, in contrast to chapter 5 where we modeled the GM distribution of seismic attributes with a single Gaussian component per facies, here we use multiple Gaussian components per facies to model complex multimodal distributions of rock properties within the same facies.

### *8.4.1 The Expectation-Maximization (EM) Algorithm*

The EM algorithm alternately maximize the free-energy $\mathcal{F}(\mathcal{Q},\Theta)$ with respect to $\mathcal{Q}$ and $\Theta$ in the so-called *E-step* and *M-step*, respectively. This improves the estimates of $\mathcal{Q}$ and $\Theta$ such that the log-evidence $\mathcal{L}(\Theta;\boldsymbol{x})$ is guaranteed not to decrease in any iteration (figure 5.3). This strategy effectively estimates $\mathcal{Q}$ that best approximates the posterior distribution $\mathcal{P}(\boldsymbol{r},\boldsymbol{\kappa}|\boldsymbol{d};\Theta)$ on convergence. With a suitable initialization, the EM algorithm is guaranteed to converge to a local optimum within a reasonably small number of iterations (Balakrishnan *et al.* 2017).

#### The E-Step

The E-step of the EM algorithm at any iteration $l$ updates the variational distribution $\mathcal{Q}(\boldsymbol{r},\boldsymbol{\kappa}|\boldsymbol{d})$ by maximizing the free-energy $\mathcal{F}(\mathcal{Q},\Theta)$ with respect to $\mathcal{Q}$ while keeping the parameters $\Theta^{(l)}$ fixed such that

$$\mathcal{Q}^{(l+1)} = \underset{\mathcal{Q}}{\operatorname{argmax}}\left\{\mathcal{F}\left(\mathcal{Q},\Theta^{(l)}\right)\right\} \tag{8.19}$$

where the bracketed superscripts refer to the iteration number. It was shown in chapter 5 that the E-step of the EM algorithm can be solved using the loopy-belief propagation (LBP) (Murphy *et al.* 1999; Yedidia *et al.* 2001a, b; Koller & Friedman, 2009) algorithm as discussed in chapter 5, which performs approximate inference and is applicable in any general graphical model (e.g. a graphical model with cyclic dependencies among variables or loops, as used in this research).

The marginal conditional distribution of $\boldsymbol{r}_i$ given $\boldsymbol{d}_i$ and $\hat{\kappa}_i$ may be obtained by conditioning on the joint GM distribution of $\boldsymbol{d}$ and $\boldsymbol{r}$ using the current estimate of parameters $\Theta^{(l)}$ at any iteration $l$ (equation 8.10), which may be expressed as another GM distribution as

$$\mathcal{P}\left(\boldsymbol{r}_i|\boldsymbol{d}_i, \hat{\kappa}_i = k, \Theta^{(l)}\right) = \sum_{t=1}^{T_k} \alpha_{t,k}^{(l)}\, g_{t,k}(\boldsymbol{r}_i|\boldsymbol{d}_i)\,, \qquad \forall i \in \mathcal{V} \tag{8.20}$$

where the bracketed superscript refers to the iteration number, and the Gaussian kernel $g_{t,k}(\boldsymbol{r}_i|\boldsymbol{d}_i)$ for the $t^{\text{th}}$ mixture component and facies $\hat{\kappa}_i = k$ is given by

$$g_{t,k}(\boldsymbol{r}_i|\boldsymbol{d}_i) = N\left(\left[\,\boldsymbol{\mu}_{r|d}^{(l)}\,\right]_{t,k}, \left[\,\boldsymbol{\Sigma}_{r|d}^{(l)}\,\right]_{t,k}\right), \qquad \forall i \in \mathcal{V} \tag{8.21}$$

with mean $\boldsymbol{\mu}_{r|d}$ and covariance matrix $\boldsymbol{\Sigma}_{r|d}$ estimated from the current estimate $\Theta^{(l)}$ of the parameters $\Theta$ of the joint distribution of $\boldsymbol{d}$ and $\boldsymbol{r}$ (equation 8.10) by

$$\boldsymbol{\mu}_{r|d}^{(l)} = \boldsymbol{\mu}_r^{(l)} + \boldsymbol{\Sigma}_{r,d}^{(l)}\, \boldsymbol{\Sigma}_{d,d}^{(l)\,-1}\left(\boldsymbol{d} - \boldsymbol{\mu}_d^{(l)}\right) \tag{8.22}$$

and

$$\boldsymbol{\Sigma}_{r|d}^{(l)} = \boldsymbol{\Sigma}_{r,r}^{(l)} - \boldsymbol{\Sigma}_{r,d}^{(l)}\, \boldsymbol{\Sigma}_{d,d}^{(l)\,-1}\, \boldsymbol{\Sigma}_{d,r}^{(l)} \tag{8.23}$$

Since petrophysical properties $\boldsymbol{r}$ are assumed to be conditionally independent given facies $\boldsymbol{\kappa}$, their joint posterior distribution $\mathcal{P}\left(\boldsymbol{r}|\boldsymbol{\kappa}, \boldsymbol{d}, \Theta^{(l)}\right)$ given $\boldsymbol{d}$ and $\boldsymbol{\kappa}$ over the entire graphical model $\mathbb{G}$ at any iteration $l$ may be expressed as

$$\mathcal{P}\left(\boldsymbol{r}|\boldsymbol{\kappa}, \boldsymbol{d}, \Theta^{(l)}\right) = \prod\nolimits_{i\in\mathcal{V}} \mathcal{P}\left(\boldsymbol{r}_i|\boldsymbol{d}_i, \hat{\kappa}_i = k, \Theta^{(l)}\right)$$

$$= \prod\nolimits_{i\in\mathcal{V}} \sum_{t=1}^{T_k} \alpha_{t,k}^{(l)}\, g_{t,k}(\boldsymbol{r}_i|\boldsymbol{d}_i)\,, \forall i \in \mathcal{V} \tag{8.24}$$

### The M-Step

The M-step of the EM algorithm at any iteration $l$ computes an updated set of parameters $\Theta^{(l+1)}$ by maximizing the free-energy $\mathcal{F}(\mathcal{Q},\Theta)$ with respect to $\Theta$ while keeping the variational distribution $\mathcal{Q}$ fixed at its value $\mathcal{Q}^{(l+1)}$ estimated during the E-step, such that

$$\Theta^{(l+1)} = \underset{\Theta}{\operatorname{argmax}}\, \mathcal{F}\left(\mathcal{Q}^{(l+1)},\Theta\right) \tag{8.25}$$

The above expression can be used to show that parameters $\Theta_{t,k} \equiv \left\{\alpha_{t,k}, \boldsymbol{\mu}_{t,k}, \boldsymbol{\Sigma}_{t,k}\right\}$ of the joint GM distribution of $\boldsymbol{x} \equiv [\boldsymbol{d},\boldsymbol{r}]$ to be updated for all of the facies $(k \in \mathcal{G})$ and mixture components $(t = 1,\dots,T_k)$ as follows:

$$\alpha_{t,k}{}^{(l+1)} = \frac{1}{n}\sum_{i=1}^{n} \hat{\mathcal{P}}\left(\kappa_i = k|\boldsymbol{d},\Theta_{t,k}^{(l)}\right) \tag{8.26}$$

$$\boldsymbol{\mu}_{t,k}{}^{(l+1)} = \frac{\sum_{i=1}^{n} \hat{\mathcal{P}}\left(\kappa_i = k|\boldsymbol{d},\Theta_{t,k}^{(l)}\right)\boldsymbol{x}_i}{\sum_{i=1}^{n} \hat{\mathcal{P}}\left(\kappa_i = k|\boldsymbol{d},\Theta_{t,k}^{(l)}\right)} \tag{8.27}$$

$$\boldsymbol{\Sigma}_{t,k}{}^{(l+1)} = \frac{\sum_{i=1}^{n} \hat{\mathcal{P}}\left(\kappa_i = k|\boldsymbol{d},\Theta_{t,k}^{(l)}\right)\cdot\left(\boldsymbol{x}_i - \boldsymbol{\mu}_{t,k}{}^{(l+1)}\right)\left(\boldsymbol{x}_i - \boldsymbol{\mu}_{t,k}{}^{(l+1)}\right)^T}{\sum_{i=1}^{n} \hat{\mathcal{P}}\left(\kappa_i = k\,|\boldsymbol{d},\Theta_{t,k}^{(l)}\right)} \tag{8.28}$$

where $\hat{\mathcal{P}}\left(\hat{\kappa}_i = k\,|\boldsymbol{d},\Theta_{t,k}^{(l)}\right)$ is the current estimate (at iteration $l$) of the marginal distribution of facies $\hat{\kappa}_i = k$ at location $i$ estimated in the E-step, and acts as weight for averaging the rock properties $\boldsymbol{x}_i \equiv [\boldsymbol{d}_i,\boldsymbol{r}_i]$ at a location $i$ in order to honour the spatial dependence among facies $\boldsymbol{\kappa}$.

## 8.5 The Approximate Posterior Distribution

On convergence of the EM algorithm, $\mathcal{Q}(\boldsymbol{\kappa}|\boldsymbol{d})$ approximates the true posterior distribution $\mathcal{P}(\boldsymbol{\kappa}|\boldsymbol{d})$ of facies $\boldsymbol{\kappa}$ given seismic attributes $\boldsymbol{d}$, such that the desired joint posterior distribution $\mathcal{P}(\boldsymbol{r},\boldsymbol{\kappa}|\boldsymbol{d})$ may be approximated as

$$\mathcal{P}(\boldsymbol{r},\boldsymbol{\kappa}|\boldsymbol{d}) = \mathcal{P}(\boldsymbol{r}|\boldsymbol{\kappa},\boldsymbol{d})\mathcal{P}(\boldsymbol{\kappa}|\boldsymbol{d}) \cong \mathcal{Q}(\boldsymbol{r},\boldsymbol{\kappa}|\boldsymbol{d}) = \hat{\mathcal{P}}\left(\boldsymbol{r}|\boldsymbol{\kappa},\boldsymbol{d},\hat{\Theta}\right)\mathcal{Q}(\boldsymbol{\kappa}|\boldsymbol{d}) \tag{8.29}$$

where $\hat{\theta}$ is the final estimate of parameters $\theta$. Note that in the above expression the variational approximation $\mathcal{P}(\boldsymbol{\kappa}|\boldsymbol{d}) \cong \mathcal{Q}(\boldsymbol{\kappa}|\boldsymbol{d})$ on the form of posterior distribution is used only for the posterior distribution of facies, and no approximation on the form of the posterior distribution $\mathcal{P}(\boldsymbol{r}|\boldsymbol{\kappa}, \boldsymbol{d})$ of petrophysical properties $\boldsymbol{r}$ is assumed; only the value $\hat{\mathcal{P}}\left(\boldsymbol{r}|\boldsymbol{\kappa}, \boldsymbol{d}, \hat{\theta}\right)$ of $\mathcal{P}(\boldsymbol{r}|\boldsymbol{\kappa}, \boldsymbol{d}; \theta)$ is approximated by the use of estimated parameters $\hat{\theta}$.

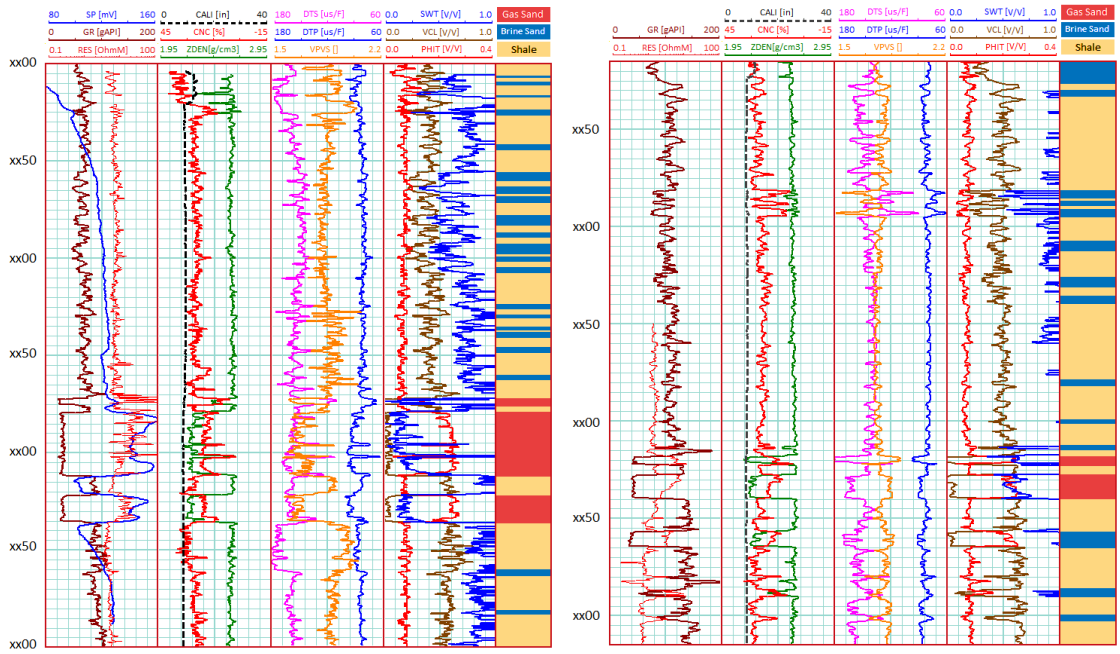
For a discussion on computational complexity of this variational method, see section 5.5 since the current method is an extension of the method presented in chapter 5 and so the computational efficiency of these methods is similar.

## 8.6 Field Example: North Sea

We apply the joint inversion method to estimate the spatial distribution of petrophysical rock properties and geological facies from well data and seismic attributes from the North Sea. The data available for testing this method include well logs from two wells, W1 and W2 (figure 8.1), and vertical 2D sections of seismic attributes, P-wave impedance ($\mathrm{I_p}$), S-wave impedance ($\mathrm{I_s}$), and Vp/Vs ratios ($\mathrm{V_p/V_s}$) (figure 8.2), that are located on the available 2D seismic section. The seismic attributes were available from a previous inversion of seismic waveform data. We are interested in classifying the seismic attribute data into three geological facies: shale, brine-sand and gas-sand, which are jointly estimated together with petrophysical properties of interest: clay volume ($\mathrm{V_{cl}}$), water saturation ($\mathrm{S_w}$) and porosity ($\varphi$). The well log data were first analyzed and the three facies of interest (shale, brine-sand and gas-sand) were interpreted from the log data. Cross-plots of pairs of elastic properties are shown in figure 8.3 with the colour scales set to (a) the facies interpreted from the well-log data and (b) the volume of clay. The gas-sand points are well separated while the brine-sand and shale points show a significant overlap.

The prior spatial distribution of facies was modelled as a MRF using a training image (TI) that represents a conceptual depiction of typical forms of expected geological structures and spatial distributions of facies in the subsurface (figure 8.4). The TI encodes the spatial conditional distributions of facies graphically. The prior information was extracted from the training image in terms of prior probabilities $\mathcal{P}\left(\kappa_i|\boldsymbol{\kappa}_{\mathcal{N}\setminus i}\right)$ constructed from histograms of various facies configurations in the image using equation 8.4. The prior probabilities encapsulate the spatial conditional distributions of facies under the assumption that they are

stationary over the entire model space. Since our input seismic attributes span a small 2D vertical section, stationarity is an acceptable assumption in this case. If, however, the aim is to invert a large region (or volume) of space or depth/time interval, the priors must be conditioned to the location using zonation or depth trends that capture the expected variability of facies patterns in space.




*(a) W1* *(b) W2*

*Figure 8.1: Well-log data and facies profiles in two wells: W1 and W2. Standard well-log pneumonics are used for the well log curves as shown in the headers above the display tracks. The colour codes for three facies, i.e. yellow for shale, blue for brine-sand and red for gas-sand, are used as standard in all of the subsequent figures in this chapter. The data from W1 is used as input for modelling the facies dependent prior joint distribution of elastic (seismic attributes) and petrophysical rock properties. The data from W2 was used only for cross-validation (testing) of the inversion results.*

The initial distribution of facies-dependent rock properties for seismic inversion was built from well log data. The well logs from W1 were used to model the prior distribution of rock properties. W1 encountered only dry gas in the reservoir formations (Sands-A, B and C), while W2 encountered brine in C-sand. For this reason, log data from W2 within the C-Sand interval was used for calibration of the prior distribution. Apart from the C-sand interval, W2 data was only used for validation (testing) of the inversion results. In order to reliably build the probability distribution of rock properties within a subsurface section (or volume), a significant

amount of well data is typically required. However, wells are often sparsely located and the well data are usually limited. In such a case, rock physics modelling and Monte Carlo (MC) simulation must be performed to augment the existing well data in order to build the prior distribution. If we construct a prior distribution using log data only from one well, it would not contain sufficient information to represent the entire model that is to be inverted. Thus, we first build a probabilistic rock physics model of the reservoir formations and then simulate rock properties from it to augment the existing well data.

We performed fluid substitution by synthetically replacing gas with brine in the reservoir sands to simulate the reservoir scenarios that are not actually encountered in W1. This requires a suitable rock physics model to be calibrated with the well data (Bosch *et al.* 2010). We investigated two related rock physics models: the *soft-sand* and *stiff-sand* models (Dvorkin & Nur, 1996). The soft-sand model assumes that the sand is unconsolidated and the cement is deposited away from the grain contacts, while the stiff-sand model assumes that the sand is strongly consolidated due to the deposition of cement material at the grain contacts. The parameters of these models are the *coordination number* $C_n$, the *critical porosity* $\varphi_c$, and the *hydrostatic pressure* $P$. $C_n$ refers to the average number of contacts that each grain has with its surrounding grains, and $\varphi_c$ refers to the initial porosity at the time of deposition (before the implacement of cement). Figure 8.5 shows the $\varphi - V_p$ crossplot overlaid on the two models using different values for $C_n$ and $\varphi_c$. Higher values of $C_n$ and $\varphi_c$ show a better fit of the well data with the soft-sand model than with the stiff-sand model. This suggests that the compaction of reservoir sands can be described by the *intermediate stiff-sand* model (Mavko *et al.* 2009).

The rock physics modelling involves a number of intermediate parameters, such as mineral and fluid properties, that introduce uncertainties in the desired elastic properties of brine-saturated rock. Such intermediate parameters are regarded as confounding variables and are assigned *Uniform* prior distributions listed in Table 8.1. MC simulation was then performed to sample these confounding variables, followed by upscaling of well logs and fluid substitution using Gassmann's equations (Berryman, 1999) to model brine and gas saturated rock with prior probabilities of brine-sand and gas-sand taken from the training image. The simulated data were then combined with the existing well data to obtain augmented data that are expected a priori to represent the elastic properties of rocks in the entire model.

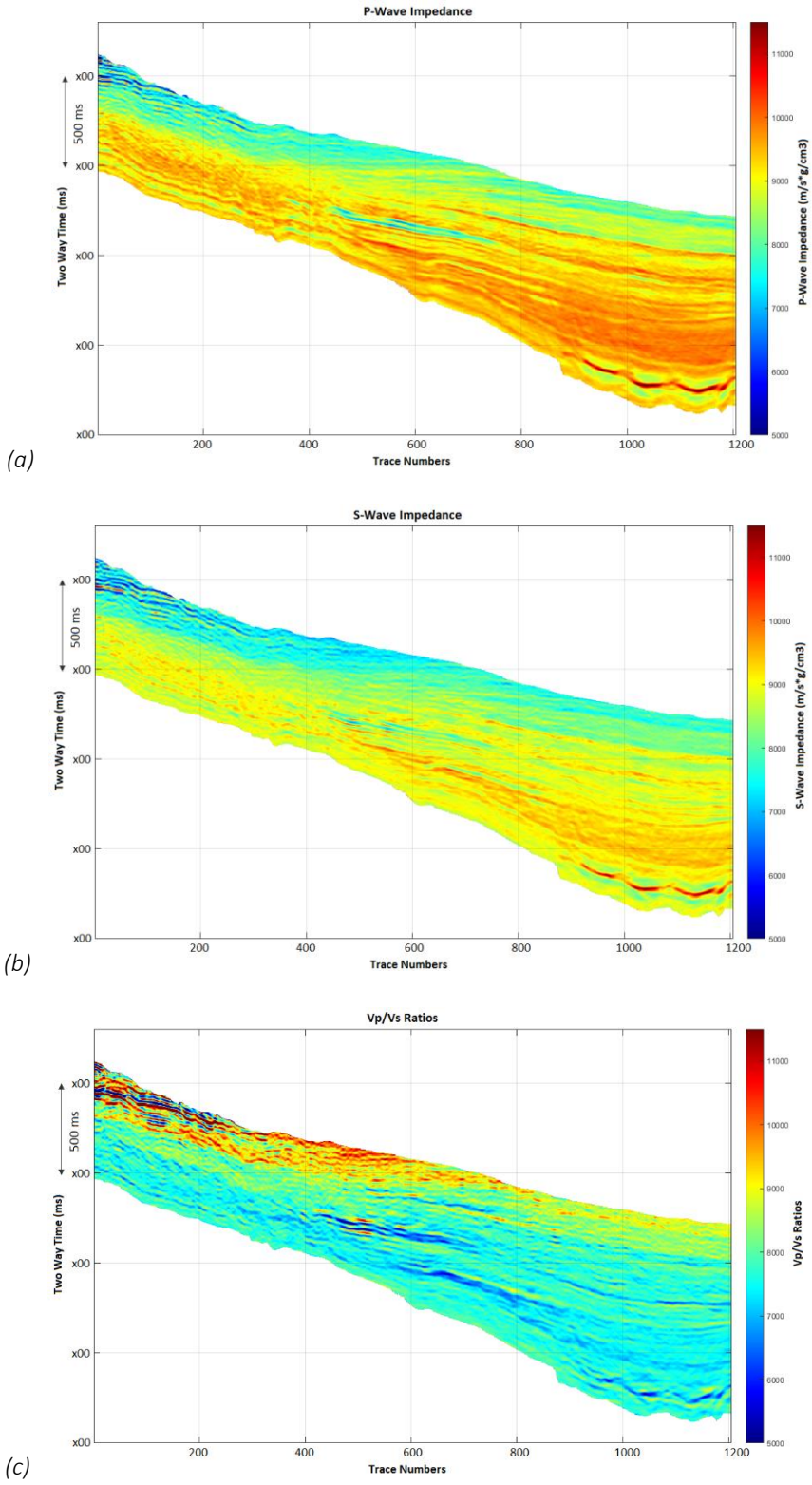
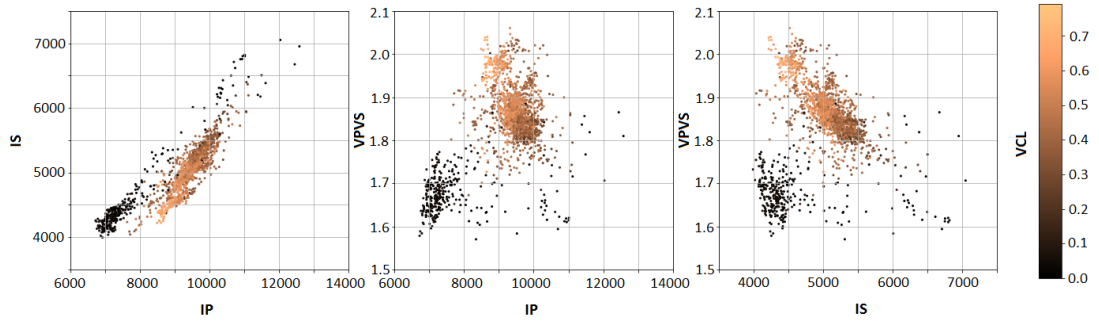
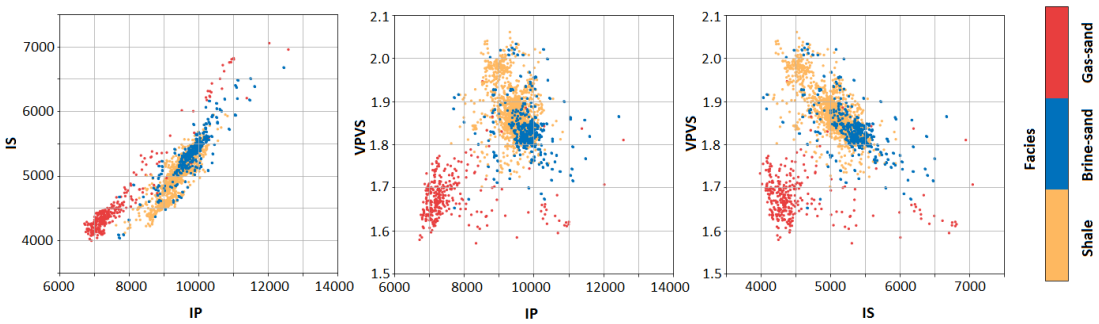



*Figure 8.2: Seismic attributes (a) P-wave impedance, (b) S-wave impedance, and (c) Vp/Vs ratios, derived from a selected 2D section of waveform seismic data. These attributes are used as inputs to our method for the joint inversion of geological facies and petrophysical rock properties.*




*(a) $I_p$ versus $I_s$* *(b) $I_p$ versus $V_p/V_s$* *(c) $I_s$ versus $V_p/V_s$*




*(d) $I_p$ versus $I_s$* *(e) $I_p$ versus $V_p/V_s$* *(f) $I_s$ versus $V_p/V_s$*

*Figure 8.3: Cross-plots between various combinations of P-wave and S-wave impedances and the Vp/Vs ratios observed in the well log data. The cross-plots are colour coded with respect to the volume of clay ($V_{cl}$) in (a)-(c) and with respect to the interpreted facies (d)-(f). The gas-sand points are well separated from the other facies, while the brine-sand and shale points have a significant overlap.*

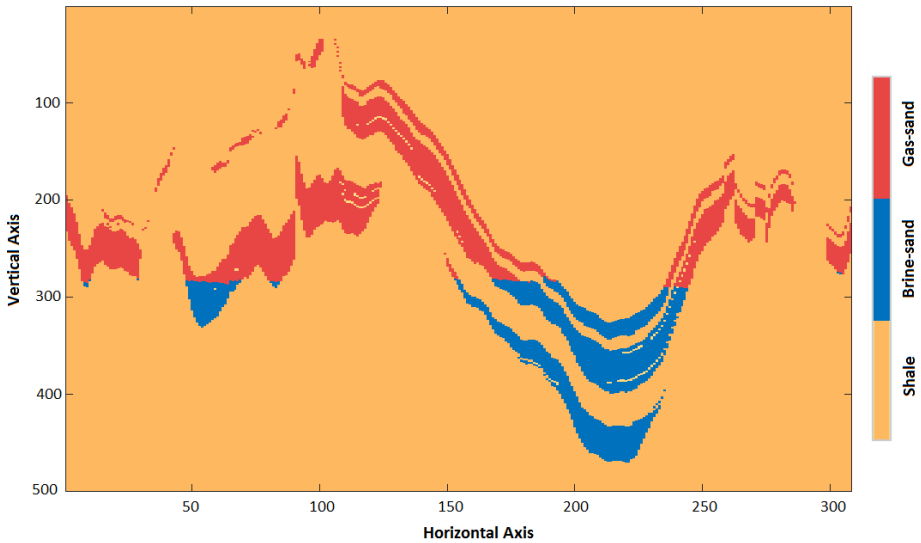



*Figure 8.4: The training image used to model the spatial prior distribution of facies that is constructed from histograms of various facies configurations found in this image.*

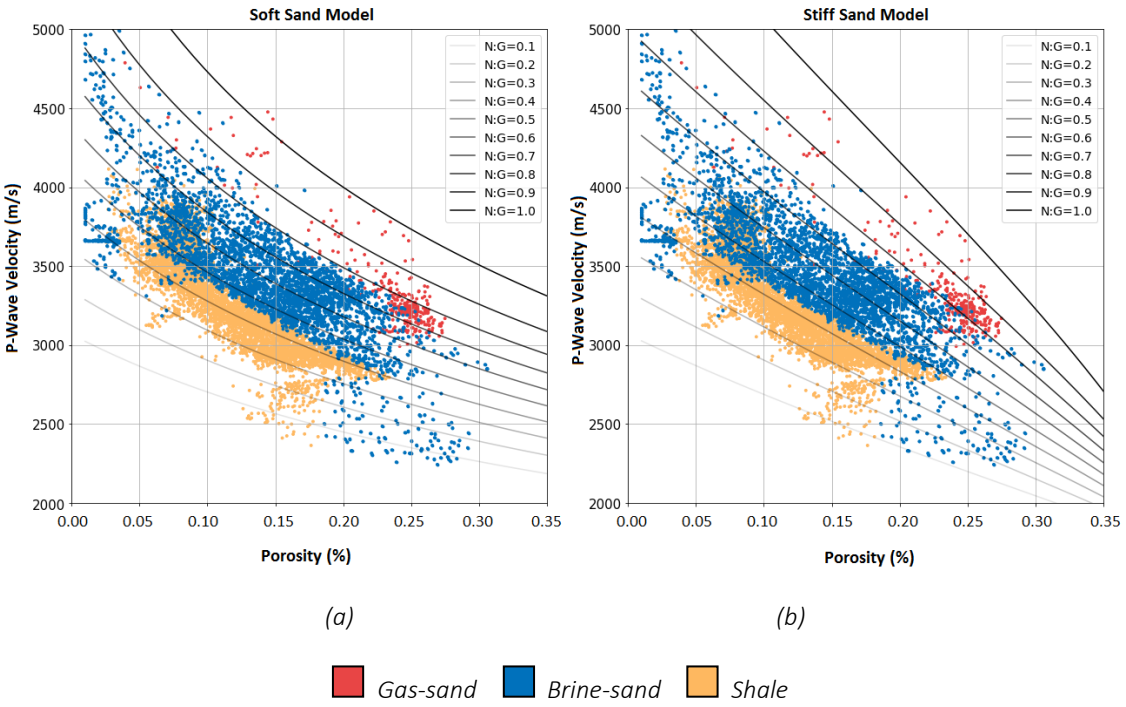



*Figure 8.5: Porosity ($\varphi$) vs. P-wave velocity ($V_p$) cross-plots with colour codes based on the facies interpreted from the well data. The overlaid rock physics template (lines with different shades of grey) correspond to trends for different net-to-gross ($N:G$) ratios predicted using (a) the soft-sand and (b) the stiff-sand model. Each of the two rock physics models are calibrated using different set of parameters: the coordination number $C_n = 13$ and the critical porosity $\varphi_c = 0.5$ for the soft-sand model, and $C_n = 5$ and the critical porosity $\varphi_c = 0.4$ for the stiff-sand model. This shows that the reservoir can be modelled using the Intermediate stiff-sand model (Mavko et al. 2009), i.e. either by a stiffer soft-sand model or a softer stiff-sand model.*

Figure 8.6 shows $I_p$ versus $V_p/V_s$, and $I_p$ versus $I_s$ cross-plots for a comparison between the original well data, the data after fluid substitution (brine replacing gas in the reservoir) using mean values of the confounding parameters, and the augmented data using MC simulations. The prior facies dependent joint distribution of the petrophysical and elastic rock properties (figure 8.7) was modelled as a GM distribution using the augmented data. Each of these facies dependent GM distributions was modelled as a mixture of two Gaussian components in order to capture possible multimodal behaviour of rock properties within each facies.

Before applying our method to invert elastic seismic attributes for petrophysical properties and facies, we first test the method by inverting the elastic logs from W2 for

petrophysical properties and facies. This also validates the consistency of the prior distribution built using rock physics modelling against the log data from W2. Recall that the W2 data were not used in building the prior distribution. The joint inversion for petrophysical rock properties and facies was performed by updating the prior distribution of rock properties by conditioning on the seismic attributes (figure 8.8) using the EM algorithm as discussed in section 8.4.1. The E-step of the EM algorithm approximates the posterior marginal distributions of facies by using the LBP algorithm, while the M-step updates the parameters of the joint distribution of rock properties given facies estimated in the E-step using equations 8.26-8.28. The marginal conditional distribution of petrophysical properties given the observed elastic properties (elastic well logs in this case) may be computed for each facies at any iteration by conditioning on the joint distribution of rock properties given facies using equation 8.20. However, this is typically required only after convergence of the EM algorithm.

*Table 8.1: Prior Uniform distribution ranges used for the intermediate rock physics parameters.*

| Rock Physics Parameter | Range |
|---|---|
| **Coordination number, $C_n$** | 5 – 13 |
| **Critical porosity, $\varphi_c$** | 0.4 – 0.5 |
| **Hydrostatic pressure, $P$** | 40 – 55 |
| **Mineral density, $\rho_m$** | 2.5 – 2.8 g/cm$^3$ |
| **Mineral bulk modulus, $K_m$** | 15 – 38 GPa |
| **Mineral shear modulus, $\mu_m$** | 5 – 44 GPa |
| **Brine density, $\rho_b$** | 1.0 – 1.1 g/cm$^3$ |
| **Brine bulk modulus, $K_b$** | 2.2 – 2.8 GPa |
| **Gas density, $\rho_g$** | 0.15 – 0.25 g/cm$^3$ |
| **Gas bulk modulus, $K_g$** | 0.04 – 0.06 GPa |
| **Error in volume of clay, $\Delta V_{cl}$** | 0.0 – 0.2 |
| **Error in water saturation, $\Delta S_w$** | 0.0 – 0.1 |
| **Error in porosity, $\Delta\varphi$** | 0.0 – 0.1 |


*Figure 8.6: (a)-(c) $I_p$ versus $V_p/V_s$, (d)-(f) $I_s$ versus $V_p/V_s$, and (g)-(i) $I_p$ versus $I_s$ cross-plots. The first column (a, d & g) displays the cross-plots using log data from W1. The second column (b, e & h) displays the cross-plots using the original well data together with the well data after replacing gas with brine in the sand layers using Gassmann fluid substitution modelling to show the effect of brine on the elastic properties of reservoir sands. The third column (c, f & i) displays the cross-plots using Monte Carlo (MC) simulated data using the soft-sand model with intermediate rock physics parameters as shown in Table 8.1 to simulate a wide range of possible values that might not have been sampled in the well data.*

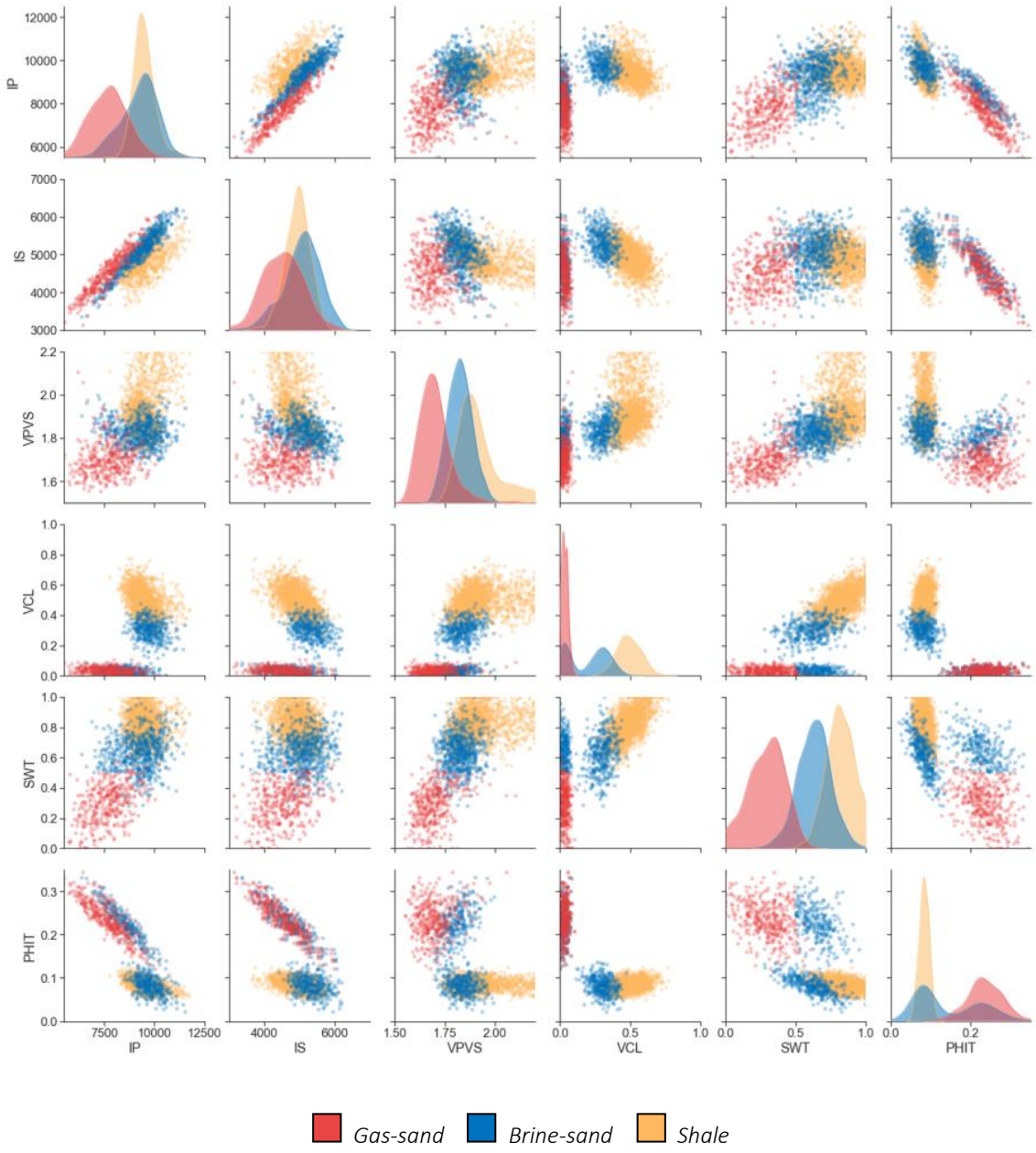



*Figure 8.7: Matrix-plot of samples from components of the prior joint distribution of elastic and petrophysical rock properties. The first three components are the elastic properties: P-wave impedance $I_p$ (IP log), S-wave impedance $I_s$ (IS log) and the P-wave to S-wave velocity ratios $V_p/V_s$ (VPVS log), and the last three components are the petrophysical properties: clay volume $V_{cl}$ (VCL log), water saturation $S_w$ (SWT log) and porosity $\varphi$ (PHIT log). The diagonal plots represent smoothed histograms of each of the components, and the off-diagonal plots show facies dependent correlations between the respective components. Yellow points represent shale, blue represent brine-sand, and red points represent gas-sand.*

Testing the inversion method on the well log data provides a best case scenario for our method since the BP algorithm performs exact inference in the 1D case. Therefore any inaccuracies in the inversion results in this case are not a result of any approximation used in probabilistic inference, but may be attributed to the approximations used in rock physics modelling. The inversion results are shown in figure 8.8. The input to inversion are the measured elastic well logs (P-wave and S-wave impedances and Vp/Vs ratios) that are shown as solid-black curves in the tracks 1-3.

The output of inversion is the joint posterior GM distribution of the elastic and petrophysical rock properties and facies. The joint posterior distribution was conditioned on the observed elastic well logs using equations 8.20 and marginalized to obtain the posterior distribution of inverted petrophysical logs (VCL, SWT and PHIT). Each of the marginal posterior GM distributions of petrophysical properties were approximated with univariate Gaussian distributions for display and interpretation purposes. The solid-red curves in tracks 4-6 are means of posterior distribution of petrophysical properties. The yellow shaded regions bounded by the dashed-red curves in tracks 1-6 are the $2^{nd}$ standard deviation of the posterior distribution of corresponding rock properties. The actually observed petrophysical logs are shown as solid-black curves in tracks 4-6 for comparison.

The standard deviation (Std.) of rock properties quantifies the natural variability of these properties, and also provides quantification of uncertainty of the predicted petrophysical properties. For precise inversion results, exactly 95.4% of the actual observed log samples should fall within the $2^{nd}$ standard deviation of the posterior distribution. Let us define the percentage of actual petrophysical log samples contained within the $2^{nd}$ standard deviation of the predicted distributions to the ideal value of 95.4% as the *confidence ratio* (CR). An ideal CR is therefore 1.0 which refers to perfect prediction of uncertainty for a Gaussian distribution. A CR value greater than 1.0 represents over-estimation of uncertainty, and vice versa. The CR for well data inversion of the petrophysical properties are shown in Table 8.2. The uncertainty is slightly under-estimated for the inverted petrophysical properties (with CR ranging between 0.93 and 0.98). It is interesting to note that since our method estimates the posterior conditional distributions of petrophysical properties from the joint distribution of elastic and petrophysical rock properties, it yields uncertainty in the input elastic properties under the joint distribution as well (as shown by the yellow shaded regions in tracks 1-3).

The similarity between the mean inversion results and the corresponding reference log curves is estimated in terms of Pearson's correlation coefficient, herein referred to simply as correlation. Excellent correlation of 0.91 and 0.93 is obtained for inverted $V_{cl}$ and $\varphi$ (compared to the measured reference log curves VCL and PHIT, respectively), while a relatively lower correlation of 0.81 is obtained between the inverted $S_w$ and the measured SWT log curve. It shows that the elastic properties have a higher correlation with clay volume and porosity than with water saturation, which is also evident from figure 8.7.

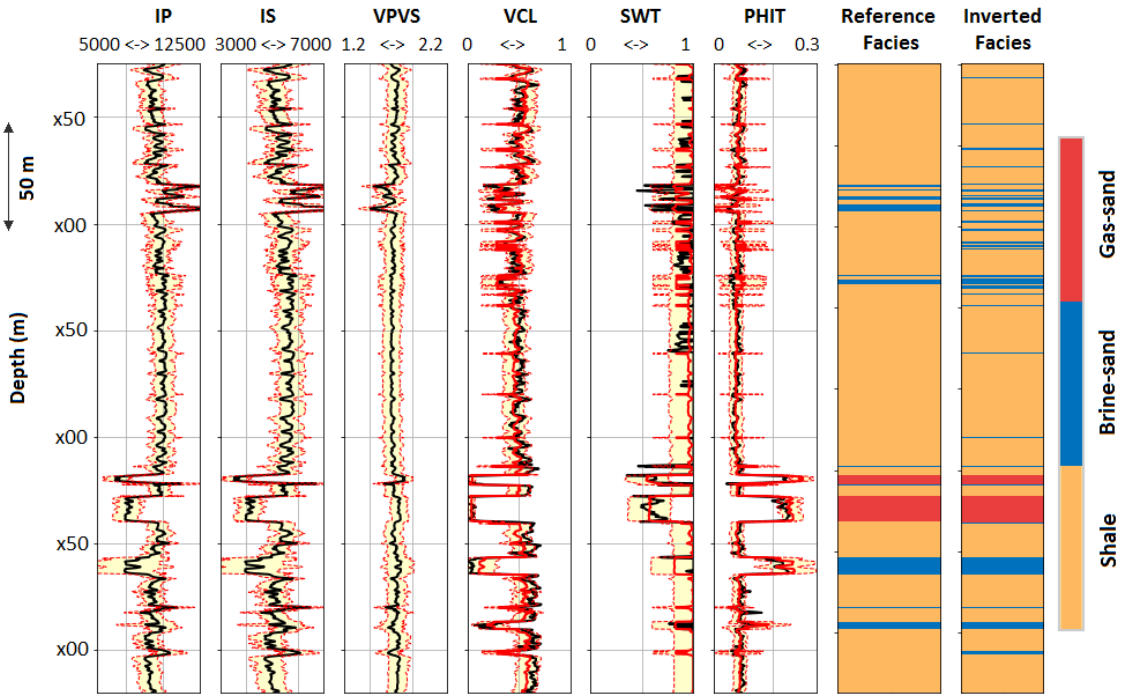



*Figure 8.8: Well logs inversion results. The first three tracks display the input elastic rock properties: P-wave impedance $I_p$ (IP log), S-wave impedance $I_s$ (IS log) and the P-wave to S-wave velocity ratios $V_p/V_s$ (VPVS log), shown in the solid-black lines estimated from the sonic (DTP and DTS) and density (ZDEN) logs shown in figure 8.2. The solid-black curves in tracks 4-6 are the reference petrophysical well logs, and solid-red curves the mean inverted petrophysical properties: clay volume $V_{cl}$ (VCL log), water saturation $S_w$ (SWT log) and porosity $\varphi$ (PHIT log). Track-7 displays the reference facies interpreted from the well data and track-8 shows the inverted facies. The yellow shaded regions bounded by the dashed-red curves represent the 2nd standard deviation of the posterior marginal distributions of the petrophysical rock properties in tracks 4-6, and the 2nd standard deviation of the conditional marginals of the joint distribution of rock properties obtained by conditioning on the estimated posterior mean petrophysical properties and integrating out the elastic properties other than the one that is plotted in tracks 1-3.*

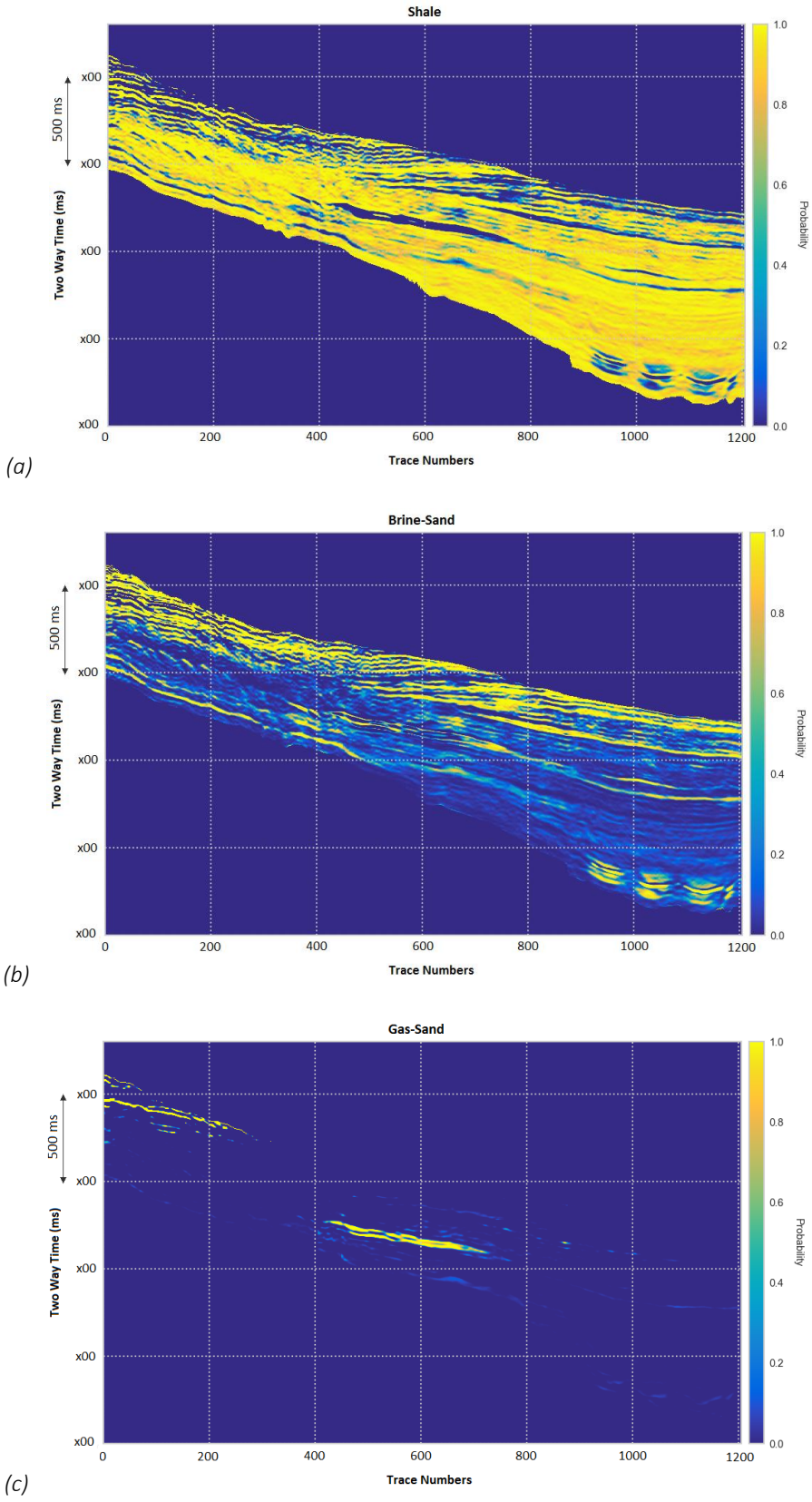



*Figure 8.9: Cell-wise posterior marginal distributions of (a) shale, (b) brine-sand, and (c) gas-sand. Yellow colour represents high probability (value=1.0) and dark blue colour represents low probability (value=0.0).*

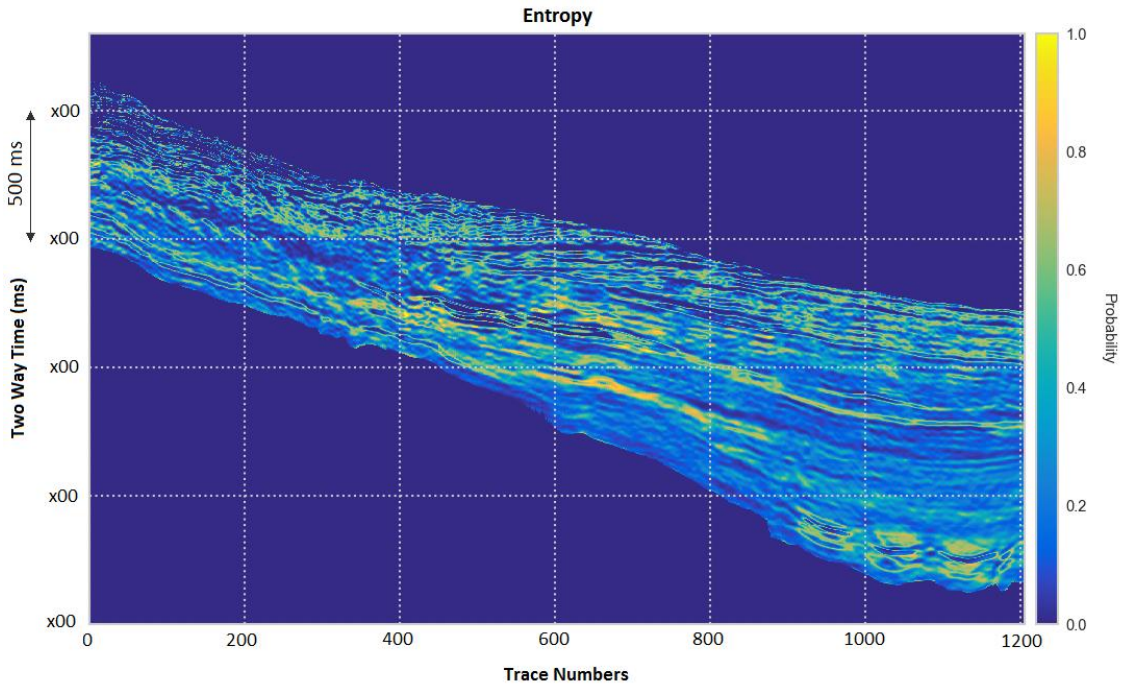
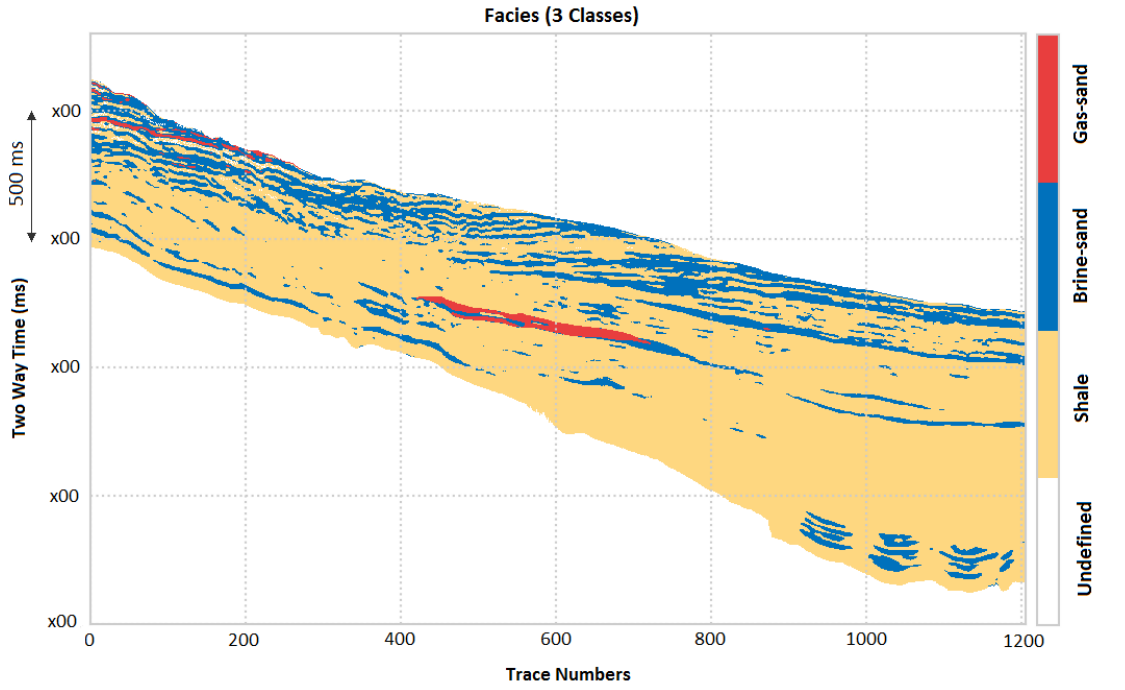


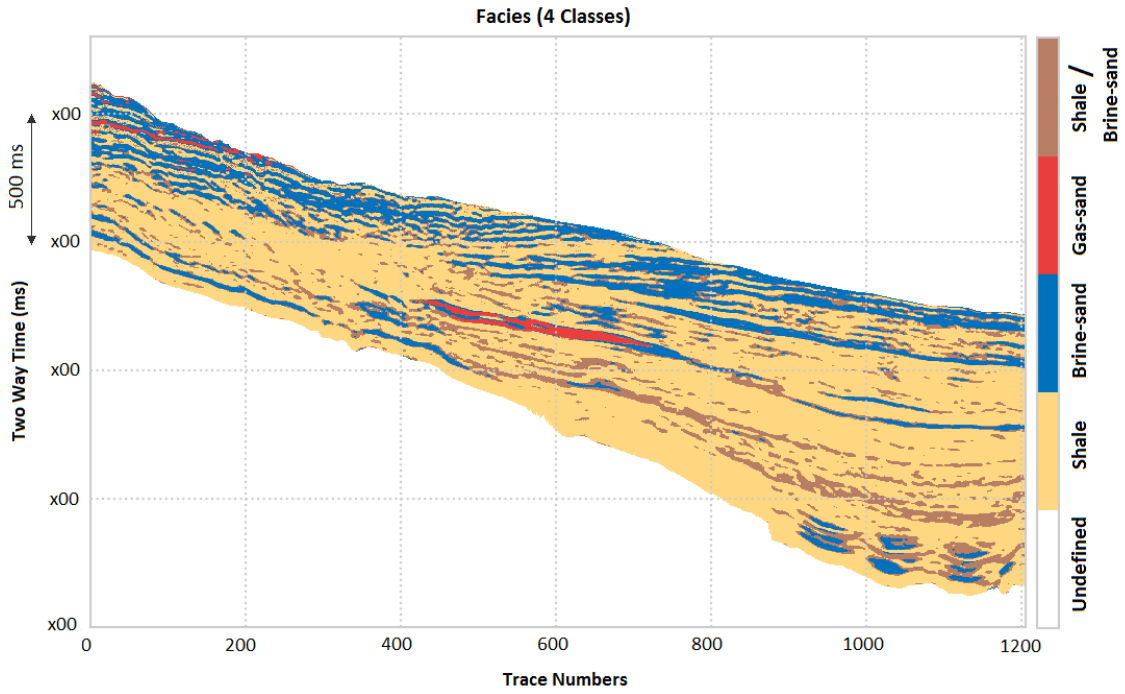

*Figure 8.10: Cell-wise posterior marginal entropy of facies classification shown in figure 8.9 scaled between 0.0 and 1.0. Yellow colour represents high entropy (value=1.0) and dark blue colour represents low entropy (value=0.0).*

*Table 8.2: Accuracy measures for the petrophysical properties and facies inverted at well locations computed with respect to the actually measured (reference) log curves and facies interpreted from well data. Confidence ratio and success rate are defined in the text.*

| Property & Accuracy measure | Well-log inversion (W2) | Seismic inversion (W1) | Seismic inversion (W2) |
|---|---|---|---|
| **Volume of clay, $V_{cl}$: Confidence ratio** | 0.93 | 0.82 | 0.73 |
| **Volume of clay, $V_{cl}$: Correlation** | 0.91 | 0.59 | 0.72 |
| **Water saturation, $S_w$: Confidence ratio** | 0.96 | 0.82 | 0.91 |
| **Water saturation, $S_w$: Correlation** | 0.81 | 0.68 | 0.61 |
| **Porosity, $\varphi$: Confidence ratio** | 0.98 | 0.77 | 0.89 |
| **Porosity, $\varphi$: Correlation** | 0.93 | 0.60 | 0.81 |
| **Shale prediction: Success rate** | 0.94 | 0.83 | 0.82 |
| **Brine-sand prediction: Success rate** | 0.76 | 0.60 | 0.66 |
| **Gas-sand prediction: Success rate** | 0.98 | 0.80 | 0.96 |
| **Overall facies prediction: Success rate** | 0.90 | 0.74 | 0.81 |

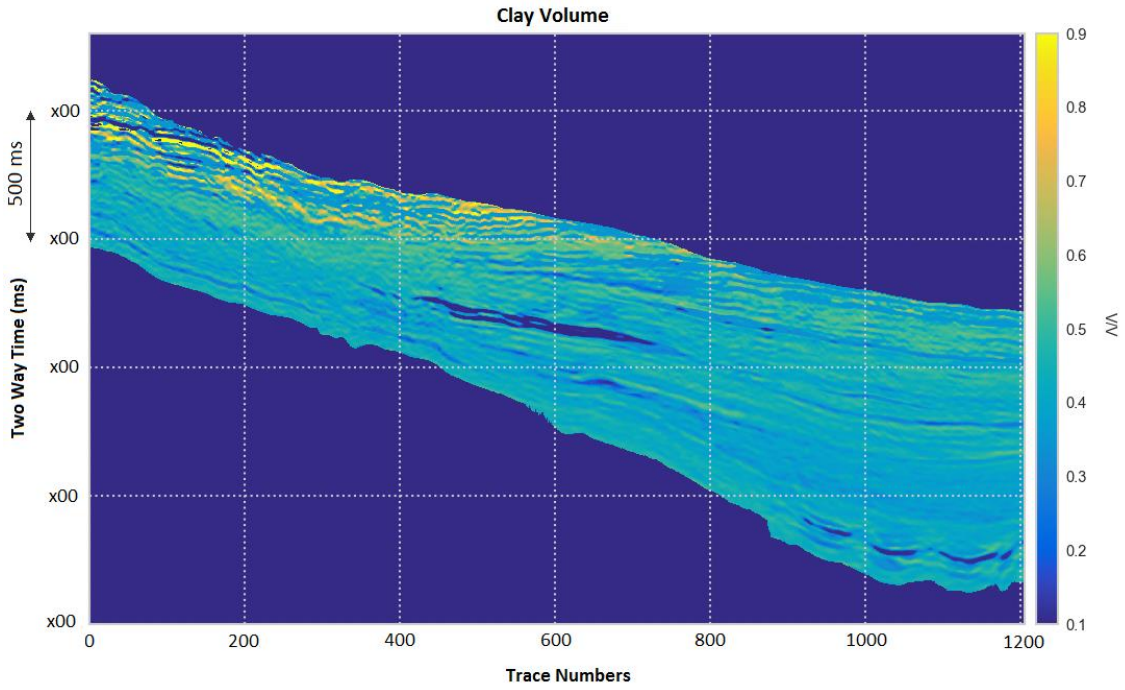



*(a)*

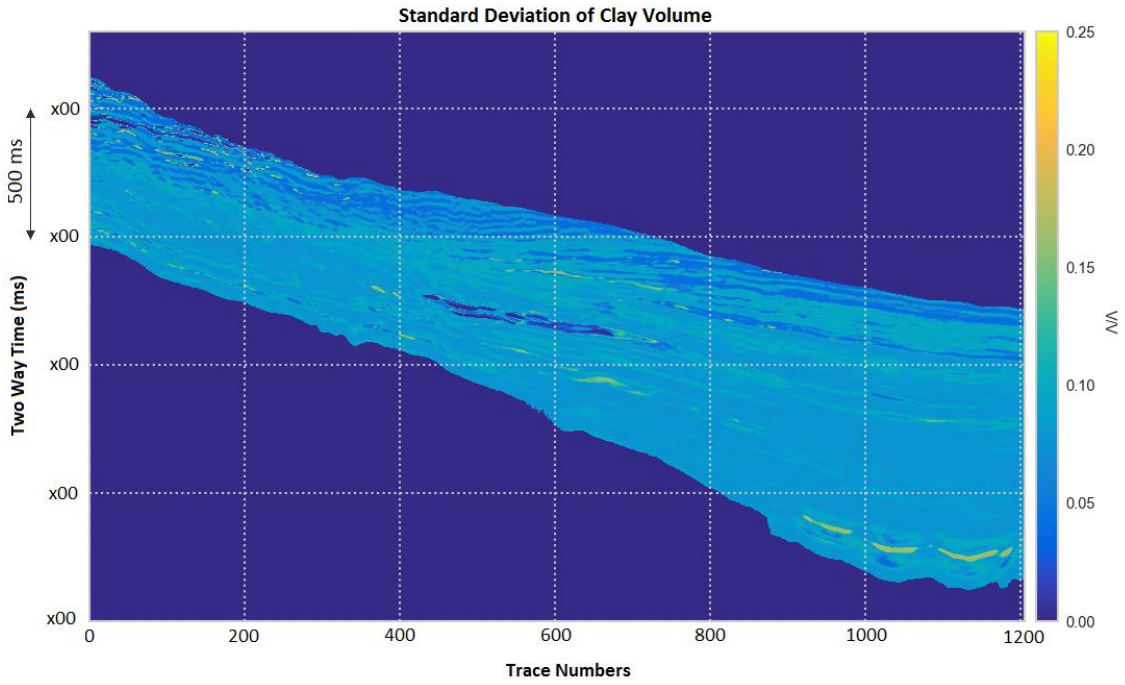



*(b)*

*Figure 8.11: Maps of facies with maximum marginal distribution in each cell. (a) Map of the three inverted facies: Shale (SH: shown in yellow), brine-sand (BS: blue) and gas-sand (GS: red). (b) Map with an additional facie "Shale/Sand" (SS: brown) identified from high entropy layers in figure 8.10.*

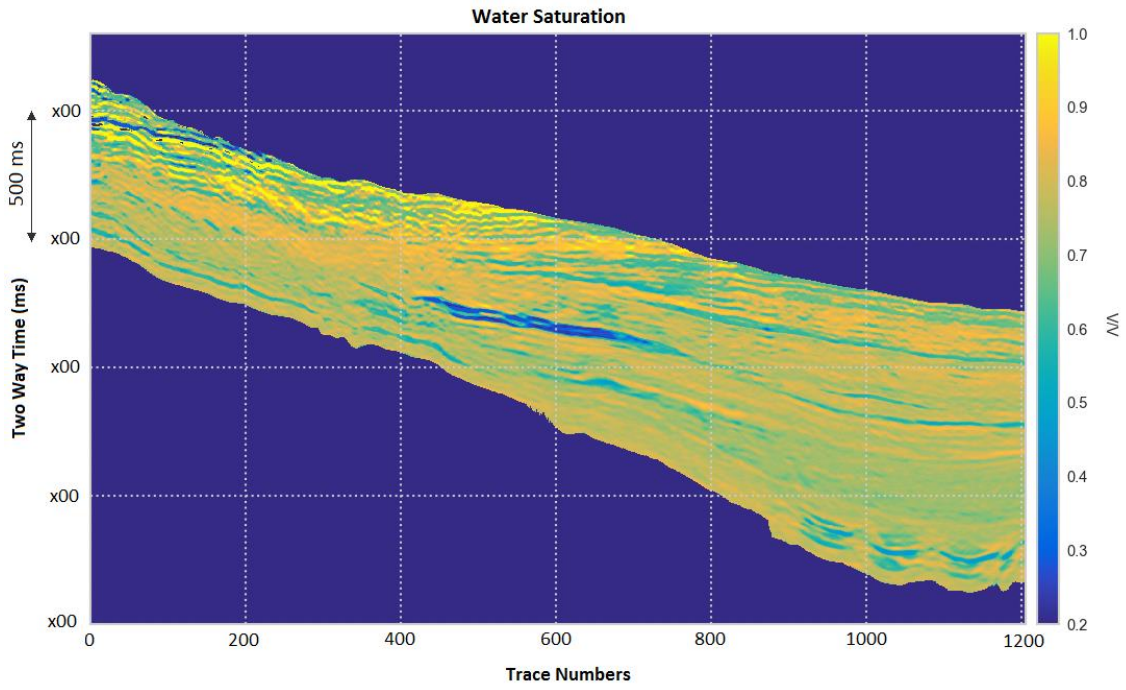



*(a)*

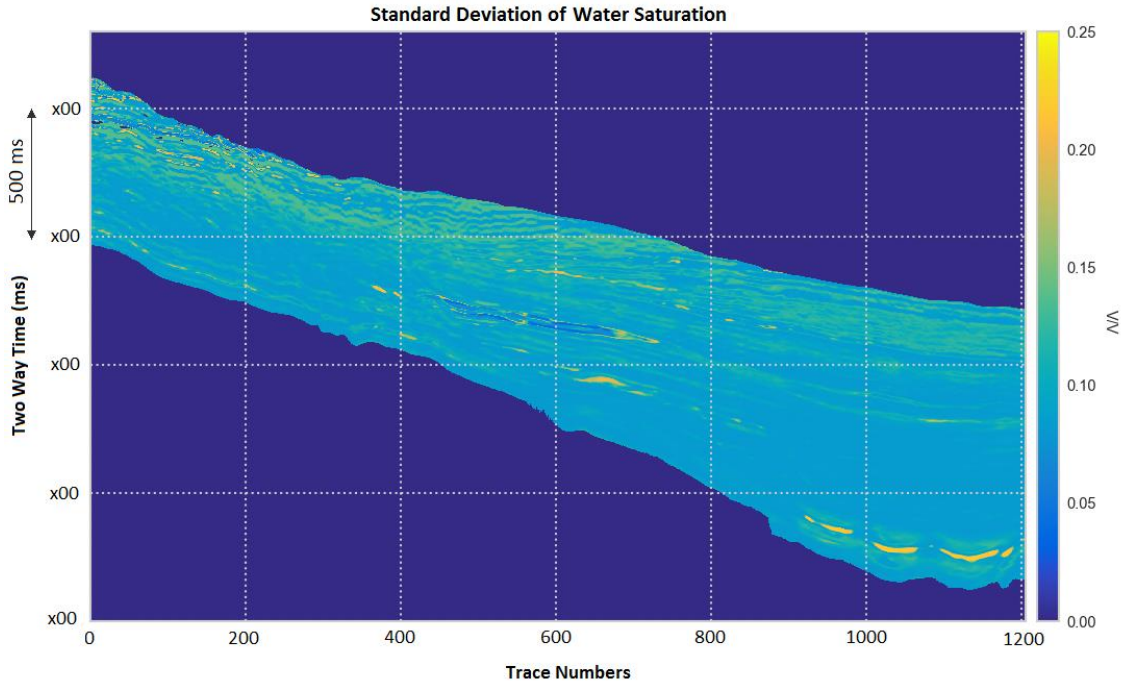



*(b)*

*Figure 8.12: Cell-wise map of (a) clay volume ($V_{cl}$) and (b) its standard deviations (Std.). Yellow colour represents high values and dark blue colour represents low values of the respective properties.*

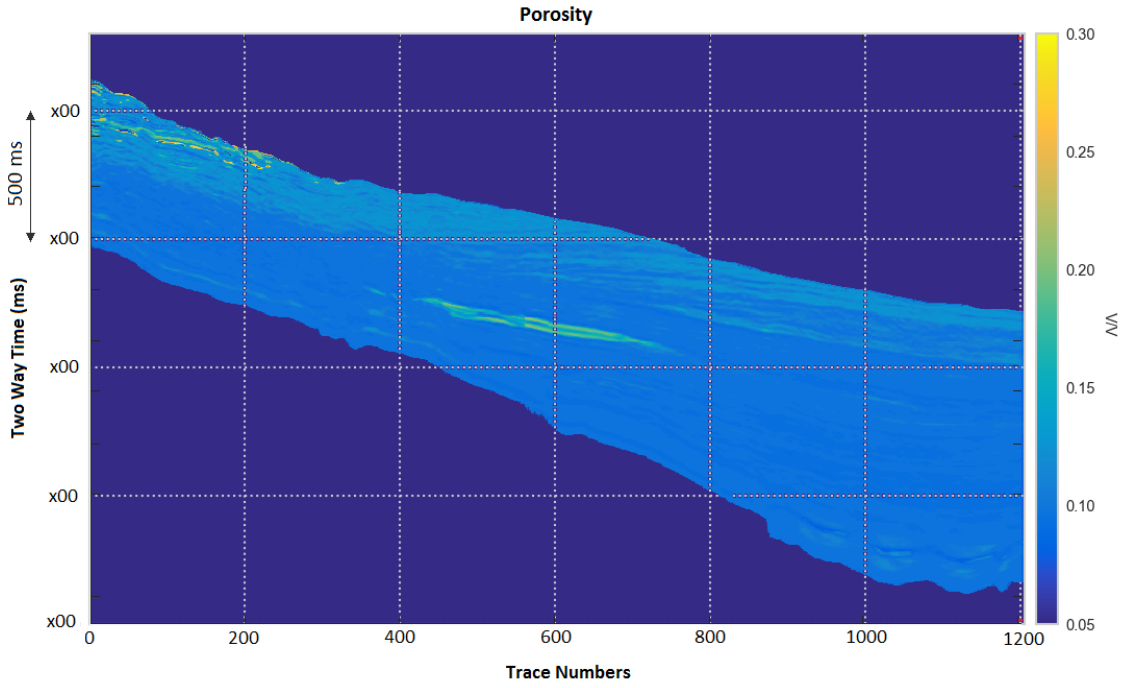



*(a)*

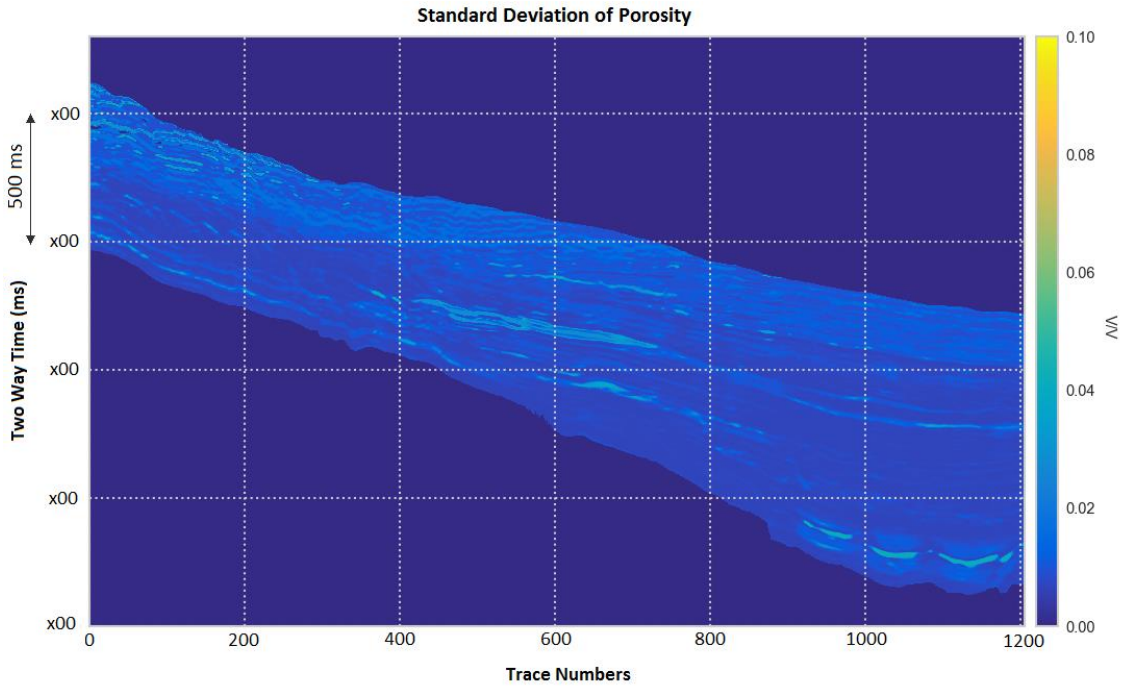



*(b)*

*Figure 8.13: Cell-wise map (a) water saturation ($S_w$) and (b) its standard deviations (Std.). Yellow colour represents high values and dark blue colour represents low values of the respective properties.*




*(a)*




*(b)*

*Figure 8.14: Cell-wise map (a) porosity (**φ**) and (b) its standard deviations (Std.). Yellow colour represents high values and dark blue colour represents low values of the respective properties.*

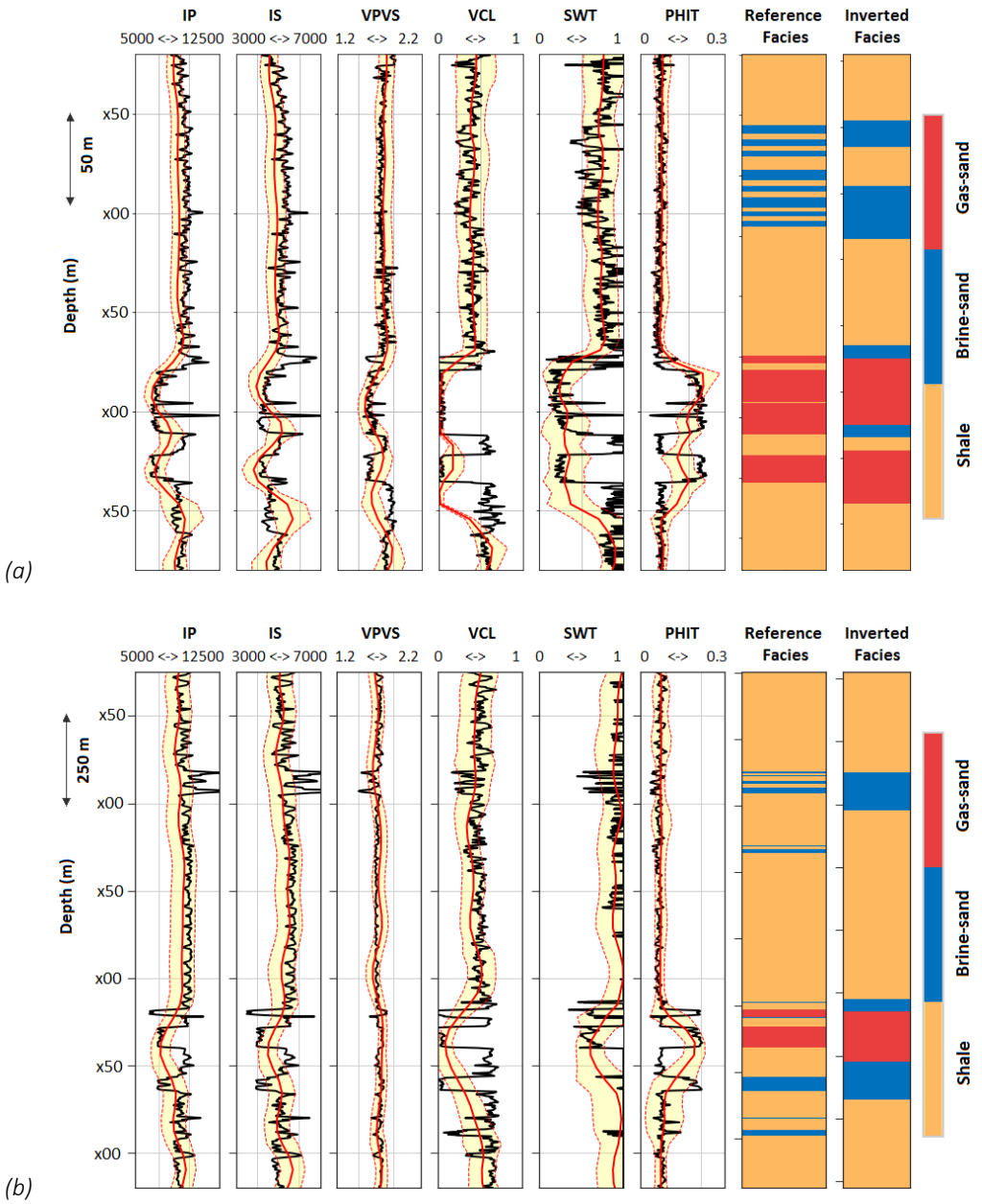


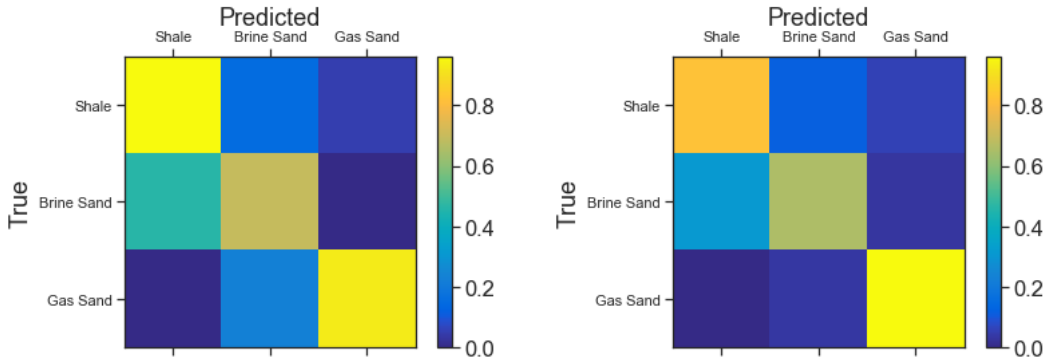

*Figure 8.15: Seismic attributes inversion results at the (a) W1 and (b) W2 well locations. The first three tracks display the elastic rock properties: P-wave impedance $I_p$ (IP log), S-wave impedance $I_s$ (IS log) and the P-wave to S-wave velocity ratios $V_p/V_s$ (VPVS log), where the reference elastic well logs are shown in solid-black lines and the seismic attributes used as input to the inversion are shown in solid-red lines. The solid-black curves in tracks 4-6 are the reference petrophysical well logs, and solid-red curves are the mean inverted petrophysical properties: clay volume $V_{cl}$ (VCL log), water saturation $S_w$ (SWT log) and porosity $\varphi$ (PHIT log). Track-7 displays the reference facies interpreted from the well data and track-8 shows the inverted facies. The yellow shaded regions bounded by the dashed-red curves in tracks 1-6 represent the 2nd standard deviation of the posterior marginal distributions of the respective rock properties.*

The success rate refers to the percentage of facies correctly predicted at the well location. The success rate is very good for shale (94%) and a bit low for brine-sand (76%), whereas the gas-sand has an excellent 98% predicted rate as the gas-sand properties are well discriminated from the rest of the two facies (figure *8.6*). As mentioned earlier, a 1D inversion with our method provides the best case results since the probabilistic inference is exact in this case, and minor discrepancies between predicted and actual properties are due to the approximations used in rock physics modelling. Since the two wells are located quite close together (about 2.0 km apart), the reservoir properties are not expected to be too different and the assumption of stationarity appears to be valid.

After verifying the inversion results at the well log scale, the inversion method was applied to invert the available elastic seismic attributes jointly for the spatial distributions of facies and petrophysical rock properties. The limited resolution of the seismic attributes is accounted for within the inversion framework using a boxcar averaging kernel (the regression coefficients in equation 8.9) whose length is determined by the dominant seismic wavelength. Figure 8.9 shows the marginal posterior distributions of the three facies and the entropy (a measure of uncertainty) of these distributions scaled between 0.0 and 1.0. Figure 8.10 shows the entropy (a measure of uncertainty) of the marginal distributions shown in figure 8.9 scaled between 0.0 and 1.0. The entropy is mostly low except at the transitions between different facies, but it appears to be high within some layers too. Since gas-sand has well discriminated properties as seen in the log data, high entropy within some layers indicates presence of mix brine-sand and shale lithology that is not well discriminated. Figure 8.11(a) shows the facies map with maximum marginal distributions in each model cell for the three inverted facies: shale, brine-sand, and gas-sand. Figure 8.11(b) shows the facies map with an additional facies defined as a combination of non-discriminated shale-sand identified to exist in the cells where entropy is greater than a cutoff value of 0.5 (i.e. 50% of the scaled entropy range from 0.0 to 1.0). Even though we inverted for 3 facies, the entropy of the marginal posterior distributions identifies that an additional facies may also be interpreted as shaly-sand or sandy-shale shown in brown colour in figure 8.11(b).

The inverted petrophysical properties along with their standard deviations are shown in figures 8.12 to 8.14. The gas reservoir consists of three sand layers (A, B and C), while only two layers are well identified which appear to be merging towards the right in the inversion results, possibly due to limited resolution of the input seismic attributes. The seismic attribute

inversion results at the well locations are shown in figure 8.15. The measured well logs are shown in solid-black curves for reference. The solid-red curves in tracks 1-3 are the input seismic attributes along the boreholes in tracks 1-3, and means of the posterior distribution of petrophysical properties in tracks 4-6. The yellow shaded regions bounded by the dashed-red curves in tracks 1-6 are the 2nd standard deviation of the posterior distribution of corresponding rock properties.




*(a) Well: W1* *(b) Well: W2*

*Figure 8.16: Confusion matrix plots for facies prediction from seismic attributes at the locations of wells (a) W1 and (b) W2.*

The quantitative analysis of seismic attributes inversion results is summarized in Table 8.2. The uncertainty is under-estimated for the inverted petrophysical properties in both of the wells (with CR ranging between 0.73 and 0.89). Acceptable correlations (ranging between 0.59 and 0.81) are found between the inverted petrophysical properties and the respective observed well logs. Lower correlations and coverage ratios are mainly due to a significantly lower resolution of input seismic attributes compared to the well logs. Facies prediction rates are very good for gas-sand and shale (between 80% and 96%) and are a bit low (60% and 66% in W1 and W2, respectively) for brine-sand because brine-sand exists mostly in the form of thin layers (figure 8.2) which are below seismic resolution. Figure 8.16 shows the confusion matrix plot of facies predictions at the well locations of W1 and W2. The confusion matrix displays the percentage of predicted facies along columns with respect to the true facies along rows. For example, the element at index $[1,1]$, i.e. top left square, represents the percentage of facies predicted as shale when the true facies is shale. Similarly, the element at index $[1,2]$ (2nd box from left on the top row), represents the percentage of facies predicted as brine-sand when the true facies is shale, and so on. For a good prediction, the diagonal elements must

have high values (shown as a colour closer to yellow), and the off-diagonal elements must have a low value (shown as a colour closer to dark blue).

## 8.7 Discussion

A major contribution of this research is the development of a computationally efficient inversion method for spatially correlated continuous (petrophysical) rock properties jointly with discrete rock properties (facies), using a sampling-free (i.e. without using McMC) yet fully probabilistic approach. The spatial correlations in continuous rock properties are governed by the spatial continuity of geological facies such that the inversion results honour both the data and the spatial prior information following the Bayesian philosophy.

The presented method avoids the common approach of petrophysical inversion that is based on an explicit use of a forward rock physics model (e.g. Bosch *et al*. 2009; Lang & Grana, 2018) that defines the relationship between data and model parameters. Contrary to that previous work, a pure data driven approach does not require any models; the relationship between the data and model parameters is expressed in the form of a probability distribution. Both approaches have their merits and demerits. For example, forward modelling always requires some simplistic assumptions about rock composition and structures which govern their properties. Such assumptions are undesirable when sufficient well data are available, in which case a data driven approach may perform better. On the other hand, rock physics models are more helpful in interpreting the inversion results.

Our method is primarily data driven; it builds facies dependent joint distributions of all of the continuous rock properties (elastic as well as petrophysical properties) and thus implicitly involves correlations between rock properties without requiring any forward model. However, a forward rock physics model may be used to augment the existing well data by generating samples of potential reservoir scenarios that are not encountered in the existing wells, or in case of limited availability of well data. Augmenting the existing well data in this manner also ensures that the prior distribution does not over-fit the existing well data, which refers to the case when inversion might perfectly predict model parameters close to the well location but may fail at other locations. Explicit use of forward modelling for solving an inverse problem often requires further assumptions such as linearity of the relationship between data

and model parameters (e.g. Grana *et al*. 2017) for computational efficiency. The presented method makes no such assumptions; it is fully nonlinear and is still computationally efficient.

An additional advantage of the presented method is that the prior joint distribution of elastic and petrophysical properties implicitly introduces prior information on the petrophysical properties. Only the prior information on the facies is separately required which can be provided in the form of training images. A training image depicts the expected spatial continuity of geological facies which can be modelled using geological process modelling (Griffiths *et al*. 2001; Hill *et al*. 2009) or other methods (e.g. Lindberg *et al*. 2015; Mariethoz & Caers, 2014). Prior information on both facies and the petrophysical properties helps to regularize the nonlinear joint inversion problem.

Mixture density estimation has been widely used in the rock physics or petrophysical inversion literature. Grana (2018) used a data dependent non-parametric kernel density estimation (KDE) method. This approach may be computationally expensive in the case of a large dataset since it requires the fitting of a predefined kernel at each data point. Also, like any other data driven method, KDE is highly susceptible to over-fitting. Parametric distributions (e.g. Gaussian), on the other hand, are often too simple to reliably model a complex probability density function (PDF). In this chapter, a semi-parametric Gaussian mixture (GM) distribution is used. A GM distribution is robust enough to capture any level of detail in any complex PDF provided a sufficient number of kernels are used, but it typically requires a much smaller number of parameters compared to a non-parametric distribution, and is therefore less prone to over-fitting.

Shahraeeni & Curtis (2011) used a GM distribution within a mixture density network (MDN) based inversion method for estimation of petrophysical parameters. They used a GM distribution with diagonal covariance matrices. A large number of kernels are required in such a case in order to reasonably represent a distribution with significantly nonlinearly correlated components. For example, P-wave and S-wave impedances are generally strongly correlated. In this work Gaussian components with full covariance matrices are used which capture any correlations among various variables. Such correlations are useful in regularizing an inverse problem in order to mitigate non-uniqueness of the solution. Although a GM distribution with full covariance involves more parameters per kernel, it requires a much smaller number of components to accurately model a given distribution.

A common approach in geophysical literature is to use a GM distribution with one component per facies to be inverted. This approach is generalized in this research by using multiple mixture components per facies. This allows the modelling of multimodal distributions caused by the intrinsic variability of rock properties within the same facies, e.g. due to patchy saturation, multiple types of porosity (pores, vugs, and fractures in carbonates), etc.

An application of the method is demonstrated on a real dataset from the North Sea. Attributes estimated from a 2D seismic section were inverted with restricted depth range under the assumption of stationarity, i.e. the statistical relationship between the rock properties do not vary with location. If, however, a larger subsurface volume is to be inverted, non-stationarity may be a challenge which can be addressed by the introduction of spatial and depth trends in the rock properties, and zonation to account for changing patterns of facies (Mariethoz & Caers, 2014). In spite of such strategies, sufficient sampling of rock properties in the subsurface still remains a critical requirement for reliable inversion in any possible scenario.

In the real data example, the input seismic attributes (P-wave and S-wave impedances and Vp/Vs ratios) were obtained deterministically from the seismic waveform data which does not provide an estimation of uncertainty in the estimated attributes. Thus, the uncertainty in input attributes due to errors in their estimation process were not incorporated; only the uncertainty due to intrinsic variability of rock properties within each facies were incorporated. This resulted in under-estimation of the posterior uncertainty in petrophysical properties. This suggests that the ignored uncertainties should also be acknowledged for an improved estimation of posterior uncertainties in the petrophysical properties.

The presented method requires a predefined structure of the Markov random field (MRF) which means that the size of the neighbourhood is fixed. This approach is similar to sequential simulation methods in Geostatistics that use a predefined template for spatial conditioning of neighbouring variables (Strebelle 2001, Mariethoz & Caers, 2014). A more general approach would invert the neighbourhood structure and size along with the model parameters using a hierarchical Bayes approach (Luo & Tjelmeland, 2018). This is left as a topic of future research.

## 8.8 Conclusions

A Bayesian inversion method is presented for joint estimation of geological facies and petrophysical rock properties and their associated uncertainties from seismic attributes. The presented method is based on a variational optimization approach which is a computationally efficient alternative to the commonly used *Markov-chain Monte Carlo* (McMC) based methods. The McMC based inversion methods do not offer objective criteria for detection of convergence of the posterior distribution in high dimensional problems, whereas our method allows reliable detection of convergence and remains computationally efficient in high dimensions (when inverting 3D seismic data, for example).

The presented method honours expected spatial distribution of facies from both data and the prior geological information that may be presented in the form of a training image. The prior spatial distribution of facies is modelled as a Markov random field (MRF). The input seismic attributes and the unknown petrophysical rock properties are jointly modelled using a Gaussian mixture (GM) distribution, and are assumed to be conditionally independent at each location given the geological facies in the neighbourhood of that location – the so called quasi-localized likelihoods (QLL) assumption. The prior joint GM distribution is updated using the expectation-maximization (EM) algorithm in an iterative fashion. The EM algorithm alternately updates an approximation to the posterior distribution of petrophysical properties and the geological facies in the so called E-step, and the GM distribution parameters in the so called M-step in each iteration. Efficient inference on the spatially correlated facies is performed using the loopy-belief propagation (LBP) algorithm within the E-step of the EM algorithm. Both LBP and the EM algorithm are computationally efficient and therefore the presented method is applicable to real-scale 3D problems.

Application of this method is demonstrated on a real dataset from the North Sea. The application shows reasonable accuracy of inversion results. However, like most other inversion methods, limited resolution of seismic data and lack of sufficient well data to provide prior information remain potential challenges for this method to produce reliable results.

# Chapter 9 Discussion

This chapter discusses the overall contribution and potential applications of this body of research as below.

## 9.1 Promoting Uncertainty Assessment in Upstream Geophysical Data Analysis

The recorded geophysical data are typically massive in size. Petabytes ($10^{15}$ bytes = 1000 terabytes) of acquired seismic data is not astounding anymore. With this "big-data" revolution, new hardware and software technologies are needed to analyze such massive amount of data. While hardware technology is growing day by day, there is still a gap in the development of software technology which is limited by the speed of developments in computational sciences. Just as developments in science lead to new developments in technology, technological advancements stimulate new developments in science. The research presented herein is stimulated by the growing need for big data analysis in geosciences in the presence of uncertainty, with the aim to develop more efficient methods that are expected to encourage uncertainty assessment in up-stream geophysical data analysis.

Geophysical data are usually processed with an interpretive approach to obtain a single image of the subsurface that is 'best' in the view of interpreter(s). Such an approach often involves parameter selection in an ad-hoc manner. As a result, assessment of uncertainty in the obtained image set aside, the presence of uncertainty is not even acknowledged. The main reason for using a non-probabilistic approach is the overwhelming computational cost of stochastic sampling for uncertainty assessment. As the geophysical data is processed, its size reduces significantly. For example, raw seismic gathers are usually 100's to 1000's of terabytes in size, whereas the processed seismic images are usually 10's to 100's of gigabytes in size. Probabilistic approach becomes practically applicable only after the raw seismic gathers are reduced to an image of the earth. Significant amount of uncertainty in the subsurface parameters (e.g. velocities) is not accounted for in the preparation of earth images.

The efficient probabilistic geophysical inversion methods developed in this thesis are expected to promulgate and disseminate the probabilistic approach and uncertainty

assessment in up-stream geophysical data analysis. Although the focus of this thesis was not on pre-stack seismic data analysis, it is expected to stimulate further developments in this field (e.g. tomography and full waveform inversion).

## 9.2 Deterministic Approach to Probabilistic Inversion

The main achievement of this thesis is the development of efficient methods for probabilistic inversion of geophysical data while honouring the geological prior information. Mathematical representation of geological parameters in the form of a Markov random field (MRF), or its variations such as hidden Markov model (HMM) and conditional random field (CRF), is a common feature of the new methods developed. New strategies are explored to achieve computational efficiency that use deterministic approach in contrast to the sampling based approach, e.g. using the Markov-chain Monte Carlo (McMC) method. The new methods developed in this thesis use deterministic approach and are applicable to structured models where posterior distribution may be represented in a factorizable form, e.g. a MRF.

A general perception in the geosciences community regarding deterministic inversion is that it is computationally efficient but it only provides point statistics of the solution such as the maximum-a-posteriori (MAP) or the mean solution. This thesis together with some other related research (e.g. Yanqin & Guoshan, 2014; Penz *et al*. 2018) is expected to change this perception since the structured models for which these methods are applicable span a wide range of problems in geosciences, and in fact in many other disciplines that involve space and/or time dependent variables. These new methods offer the computational efficiency of the deterministic approach while still providing the full joint posterior distribution in terms of marginal distributions – the factors that constitute the joint distribution over all of the desired model parameters.

## 9.3 Review of Strategies Used for Efficient Probabilistic Inversion

As discussed in section 1.1.2, sampling based stochastic inference, e.g. using the Markov-chain Monte Carlo (McMC) method, is computationally slow and this thesis aims at developing new methods for solving geophysical inverse problems using efficient approximate

probabilistic inference methods. The main strategy was to avoid sampling and explore deterministic methods for mathematical treatment of probabilistic dependence among various random variables involved in the inverse problem. A key feature of geophysical inverse problems (or in general any spatial and/or temporal data analysis problem) is that these often involve structured set of probabilistic dependence among various parameters of interest. Structured dependence refers to the case when each random variable in a model is strongly correlated with just a few other variables, and is only weakly correlated with the rest of the variables in the model. This induces 'indirect' probabilistic dependence among a large number of variables. When indirect probabilistic dependence exists between any two variables, it connotes with conditional independence assumption between these variables given rest of the variables. The assumption of conditional independence among various model parameters is commonly referred to as the *Markovian assumption*. With reference to geological models, the Markovian assumption requires that geology at a location depends directly only on geology within some pre-specified neighbourhood of that location. In other words, given the geological properties in some pre-specified neighbourhood of a particular location in the model, properties in the rest of the model provide no additional information about the properties at that particular location.

The conditional independence (CI) assumption of data given the model parameters, as discussed in section 1.1.4, is different from the Markovian assumption. The former refers to the assumption that any correlations present in the observed data are direct consequence of correlations in the model parameters, while the latter is a characteristic of models that involve structured probabilistic dependencies. The Markovian assumption is a characteristic feature of two commonly used PGMs in spatial and/or temporal data analysis methods: the *hidden Markov model* (HMM) and the *Markov random field* (MRF). According to the *Hammerley-Clifford theorem* (Besag, 1974), the Markovian assumption induces factorization of any probability distribution defined over model parameters. This means that the joint posterior distribution over all of the model parameters decomposes into factors (called *Gibbs factors*) each of which is typically much lower in dimensions than the full joint posterior distribution. Although this makes probabilistic inference more feasible, exact Bayesian inference still remains intractable even with the Markovian assumption in most models of realistic scales since it requires normalization of the posterior distribution which must be performed over the entire high dimensional space. Thus approximate inference is inevitable in high dimensions.

The Markovian assumption is ubiquitous in the entire geostatistical literature, and has been proven to be valid for all practical purposes. It has also been used widely in the geophysical literature to formulate McMC based stochastic sampling algorithms. However, inverse problems that involve MRF as a model of probabilistic dependence among parameters of interest, lend themselves to the variational formulation naturally. The methods developed in this thesis exploit factorization of the posterior distribution under the Markovian assumption in order to devise more efficient probabilistic inversion methods. Thus instead of sampling the solution space stochastically, the solution was obtained using analytical (chapter 4), numerical (chapters 5 and 6), and combined analytical and numerical, i.e. semi-analytical (chapters 7 and 8) approaches while making use of machine learning where appropriate.

Another strategy used (chapter 6) for efficient probabilistic inversion is to decompose large scale inverse problems into interlinked sub-problems that can be solved efficiently using machine learning. The solutions of the sub-problems can then be recomposed using numerical optimization based Bayesian inference as discussed in section 2.4. Nonlinearities in model-data relationships cause ill-posedness of geophysical inverse problems. Linearization based simplistic approximations ignore such known nonlinearities, and therefore cause errors in the solution. For this reason, despite the fact that linearization generally allows solving an inverse problem efficiently, it is avoided in this thesis to a large extent (except in chapter 7, where nonlinear solution is proposed as a potential future extension of this the presented method). The model-data relationships are learnt using machine learning methods instead.

As a comparison with McMC, it is important to point out that McMC is a general method, whereas the deterministic alternatives such as variational Bayes (VB) are more objective oriented methods. This is one of the reasons these methods are computationally more efficient than McMC. For example, probabilistic inference generally requires some sort of marginalization over at least some of the variables. HMM based inference and VB focus on the estimation of the desired marginal distributions, while McMC must estimate the full joint posterior distribution first, which may be marginalized subsequently to obtain the desired marginal distributions.

## 9.4 Gain in Computational Efficiency

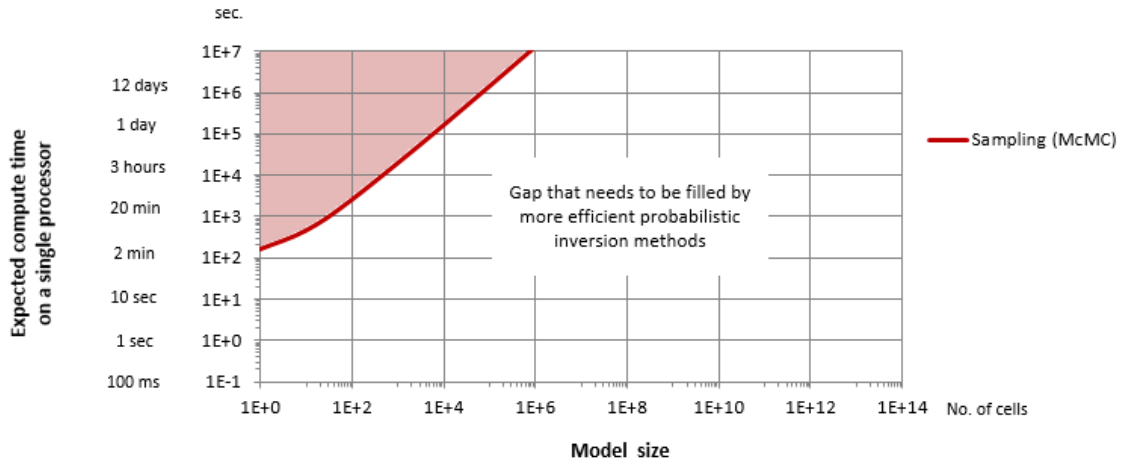
Stochastic inversion methods explore the solution space in order to obtain an estimate of the posterior distribution in the form of randomly generated realizations of the model that are consistent with data and prior information. Each such realization is a possible instantiation of the unknown reality. These methods are computationally very expensive and their computational cost increases significantly with the number of dimensions of the solution space. A quantitative comparison of the computational cost of variational Bayesian (VB) inference and McMC is quite difficult in general. Also, McMC is a global search method while VB is a local optimization method. Thus a direct comparison of the computational performance in general is meaningless. A fair comparison between the two methods for any general problem requires extension of the methods developed in this thesis to solve highly nonlinear problems such as tomography and full waveform inversion (FWI) (e.g. using global optimization strategies), and then such a comparison be made. Nevertheless, we may still compare the two approaches to probabilistic inversion for the models to which the presented methods are applicable, i.e. models with the Markovian assumption.

Assessing the computational cost of McMC based methods is quite difficult because of many factors that are unknown beforehand, e.g. acceptance ratio of the generated samples, total number of samples to be generated, rate and detection of convergence etc. Perhaps it is for this reason that published literature on geostatistical inversion methods that use McMC sampling rarely assess the computational cost of these methods quantitatively. Nevertheless, McMC based methods are known to take hours for even very small problems, whereas methods developed in this thesis take from a few seconds to minutes to solve most of the similar problems. An advantage of increased computational efficiency is that these methods may be employed to solve more complex problems without incurring significant computational costs. This has been demonstrated with examples in chapter 5 where the localized likelihoods (LL) assumption has been relaxed and in chapter 6 where this assumption has been removed. In chapter 6 the conditional independence (CI) assumption on data has also been removed.

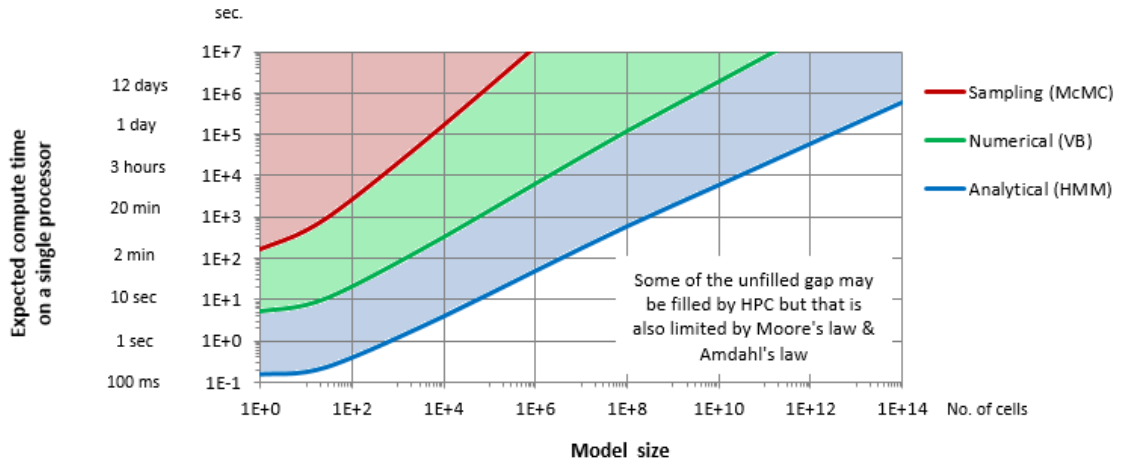
A fair comparison of the computational cost of McMC and the deterministic methods developed in this thesis requires such comparison to be made with respect to a given problem, i.e. under the same set of assumptions. For example, Walker & Curtis (2014a) performed Gibbs sampling on the same synthetic problem that was presented in chapter 4 under the localized

likelihoods assumption. They performed $10^9$ iterations which took approximately 24 hours whereas the method based on a HMM presented in chapter 4 took just 0.2 seconds on the exact same problem with a grid size of 100 x 100 cells on a single processor. This shows that the HMM based inference is about 5 to 6 orders of magnitude faster than McMC for this type of problems. Mohammad-Djafari & Ayasso (2009) showed a comparison of computational cost of VB inference versus McMC in an unsupervised learning problem that is similar to the variational Bayesian inversion (VBI) method (chapters 5 and 8). They showed that VB is about 4 orders of magnitude faster than McMC in their problem. The discriminative Bayesian inversion (DBI) method (chapter 6) which is also based on VB inference uses the so called discriminative modeling approach which typically requires supervised learning that may be a tedious task for some problems and requires interpretive approach to generate and learn from the training examples. Its computational cost depends on both the cost of learning the inverse mapping from data to the desired model parameters, and the cost of spatial inference that ensures that the inferred model honours spatial correlations supplied in the form of geological prior information. The additional cost of supervised learning is justified because this method is only expected to be used where its counterpart, the so called generative modeling approach, is computationally lot more expensive. Various methods have been developed by the machine learning community, such as *stochastic gradient descent* (SGD), that allow fast supervised learning. Apart from the manual time and effort involved in supervised learning, the computational cost of inference in DBI is similar to that of VBI.

The cost of generating one stochastic sample of the complete model and one iteration of the VB inference methods such as loopy-belief propagation (LBP) and mean field (MF) approximation is quite similar for most problems irrespective of the model size and the expected range of spatial correlations within the model space. While VB inference mostly requires 10's to 100's of iterations for most problems, MCMC requires hundreds of millions of iterations in moderate sized models (Eidsvik *et al*. 2004; Walker & Curtis, 2014a). This again suggests that according to a rough estimate deterministic inversion methods developed in this thesis are expected to be at least 2 to 4 orders of magnitude faster than the corresponding McMC based methods for most (linear, or weakly non-linear) geophysical inverse problems. Further developments are needed to improve computational efficiency for highly nonlinear problems. This is further discussed in section 9.5.1.




*(a)*




*(b)*

*Figure 9.1: A rough comparison of computation cost of McMC based geostatistical inversion methods with the efficient probabilistic methods developed in this thesis. (a) The white colour represents the gap in computational efficiency that needs to be filled. (b) Some the gap is filled by the methods developed in this thesis. Some of the remaining gap may be filled by using high performance computing (HPC). Some gap will still remain as governed by the Moore's law and Amdahl's law even with the used of HPC.*

McMC is expected to yield a global solution in highly non-linear problems provided that sufficient number of iterations are performed. The VB inference, on the other hand, yields a local solution in the probability-parameters ($Q - \Theta$) space, which may or may not correspond to the global solution in the space of model parameters $\boldsymbol{m}$. This suggests that VB inference must be performed within a global optimization framework such as using simulated annealing (SA) or genetic algorithm (GA) in order to fully explore the solution space in case of highly non-

linear problems, such as tomography and full-waveform inversion. This is expected to increase the cost of VB inference by 1 or 2 orders of magnitude for most realistic scale problems, corresponding to 10's or 100's of independent runs of inference each with different initial conditions. Nevertheless, the deterministic inference methods developed in this thesis are still expected to provide a significant computational gain compared to McMC in highly non-linear problems. Verification and quantification of this is clearly a direction for future research.

In the light of above discussion, a rough quantitative comparison of computational efficiency of sampling based inference methods (e.g. McMC) and the deterministic inference methods developed in this thesis may be provided for MRF/HMM based models. Figure 9.1a shows the compute time required on an average single processor as a function of model size (number of cells) for McMC based methods. The graph has been taken from previous literature on McMC based inversion methods that use the Markovian assumption. The region in red shows the feasible region for McMC based methods, and the region in white shows the gap that needs to be filled with more efficient probabilistic methods. This thesis aims to fill some of this gap.

Figure 9.1b shows a rough quantitative comparison of the three methods: sampling based inference using McMC, numerical estimation of the posterior distribution using VB (chapters 5 to 8), and analytical solution for the posterior distribution using HMM (chapter 4). Exact computational time required by any of these methods depends upon a number of other factors such as the correlation range in model or data space, and the time required for forward model computation. Such factors are ignored in this comparison as these are likely to affect the computational efficiency of all of these methods in a similar manner. Only size of the model in terms of number of cells is considered variable.

The area above a curve in this graph represents the feasible region for the method corresponding to that curve. Thus, the region shown in red is feasible for all of the three methods. The green region is feasible for VB and HMM based solutions and not for McMC. Similarly, the region shown in blue colour is only feasible for HMM based solution. The green and blue regions thus represent the gap that is filled by this thesis. The white region represents the gap that still remains unfilled. Some of this gap may be filled by high performance computing (HPC) using multiple processors and by harnessing the parallel computing ability of graphical processing units (GPU). It is worthwhile to note here that McMC is parallelizable only to a certain degree, whereas significant computational gain may further be achieved using VB

and HMM based inference since these are easily parallelizable to a higher degree. The HMM based method in chapter 4 is embarrassingly parallelizable, whereas the inference part of each iteration of the variational methods presented in chapters 5 to 8 are also easily parallelizable. However, even using HPC some of the unfilled gap (white region in figure 9.1b) is still inevitable as governed by the *Moore's law* and *Amdahl's law* which define the current technological limits of maximum achievable compute power.

## 9.5 Directions for Future Research

The directions for further advancements in this research are proposed below:

### *9.5.1 Efficient Probabilistic Inversion Using Global Optimization*

The solution of a highly non-linear problem (e.g. FWI) is typically non-unique; a large number of solutions may produce the same data within some acceptable tolerance. Variational inference as used in this thesis is not guaranteed to provide the global solution (Saddiki *et al.* 2017). In order to address this limitation, it is therefore proposed to use these methods within a global optimization framework such as simulated annealing (SA) and genetic algorithm (GA). Further, since stochastic sampling based methods (e.g. McMC) generally provide a global solution provided that sufficient number of samples are generated, a fair comparison of optimization based inference with stochastic methods is only possible when global optimization is performed. Even with a global optimization approach, variational inference typically requires fewer number of iterations compared to stochastic inference (Gultekin et al. 2018). The computational performance of methods developed in this thesis is therefore still expected to be better than corresponding stochastic methods in most problems.

### *9.5.2 Model Selection*

The computational challenges of exact Bayesian inference originate due to the intractable denominator in Bayes theorem (equation 2.1), called evidence or the marginal data-likelihood, since its evaluation requires integration of the product of prior and likelihood over a potentially high dimensional space. Besides acting as the normalization constant, this term also carries further significance. By marginalizing out the model parameters, this term determines how well the mathematical model itself fits the observed data for all possible

combinations of model parameters. The mathematical model here refers to the functional form of prior and likelihood distributions, which in turn depends on the prior representation such as the structure of the HMM or MRF, and the relationship between data and desired model parameters.

The approximate Bayesian inference methods used in this thesis estimate the evidence term using numerical optimization. In particular, variational inference defines a lower bound on the logarithm of the evidence term, which is referred to as log-likelihood. The lower-bound is maximized to obtain an estimate of the evidence term. Since and estimate of the evidence term is obtained in these methods, it can help in model selection which is computationally hard using the stochastic methods.

The methods developed in this thesis assume a predefined structure of the Markov random field (MRF) which means that the size of the neighbourhood and structure of probabilistic dependence among various variables is fixed. This approach is similar to sequential simulation methods in Geostatistics that use a predefined template for spatial conditioning of neighbouring variables (Strebelle 2001, Mariethoz & Caers, 2014). A more general approach is proposed to invert the neighbourhood structure and size along with the model parameters using a hierarchical Bayes approach. Such an approach may also include the parameters of the forward problem to be estimated as a part of inversion process, where necessary.

### *9.5.3 Hierarchical Geological Modelling Within Geophysical Inversion*

Geological prior information is often derived from *geological process modelling* (GPM) which simulates the expected spatial distribution of geological properties (e.g. Hill *et al*. 2009). Such an approach requires a good knowledge of the parameters such as deposition rate of sediments and the available accommodation space in the basin to be known with reasonable accuracy, which is often not the case. As a result, the prior information derived from GPM results may not be well representative of the true geology. The hierarchical Bayes approach of model selection may be taken to a new level where geophysical inversion may be combined with GPM to obtain the result by solving these problems simultaneously in an iterative manner. This is expected to be computationally intensive. However, since the geophysical inversion methods presented in this thesis are computationally efficient, these methods are not expected to impede such a development.

### *9.5.4 Comparison with McMC*

The new methods developed in this thesis using the deterministic approach require that the posterior distribution is factorizable over some subsets of the set of model parameters, called cliques in a graph. Since these methods exploit factorization of the posterior distribution under the Markovian assumption, the quality of inversion is expected to be good as long as MRF is a valid model. This has also been verified through synthetic tests where the inversion results could be compared against the 'true' model known *a priori*. The quality of results have also been verified through application of these methods on real data from North sea, where inversion results are validated against observed borehole data at the well locations. However, there still remains a need for a detailed comparison of these methods with the corresponding McMC based methods in terms of quality. Such a comparison is proposed for future research.

### *9.5.5 Probabilistic Approach to Imaging*

The forward problem in geophysics models the relationship between data and the model parameters of interest (such as geological properties). Many geophysical inverse problems, such as tomography and full waveform inversion, are highly nonlinear because the forward problem depends on the unknown model parameters themselves. For example, seismic travel times represent the observed data in travel-time tomography and are used to infer subsurface velocities. Travel times depend on the ray paths in the subsurface which are themselves defined by the unknown velocities. So forward modelling requires the solution to be known a priori. Such non-linearity is handled by solving the problem in an iterative manner via optimization based methods or by repeated modelling of ray paths with randomly sampled velocity models using McMC. The optimization based methods are efficient but may not fully characterize the solution space.

The forward problem is typically assumed to be deterministic in geophysical inverse problems. Given that the forward problem depends on model parameters (such as seismic velocities) in highly non-linear geophysical problems, and that the model parameters are themselves regarded as random variables in probabilistic inversion, this suggests that the forward problem should also be regarded as stochastic. For example, seismic wave propagation through media with uncertain parameters may be regarded as a stochastic problem where both the source and the model properties may be treated as random variables.

Such a stochastic forward problem may be solved by using *stochastic partial differential equations* (SPDEs) that propagate initial parameter uncertainties to uncertainties in the observed data. Such a stochastic forward problem can be solved within a Bayesian inference framework to yield probabilities of the desired subsurface parameters. This approach will allow fast, more accurate and fully probabilistic analysis of geophysical data (such as seismic images) and geological properties without using MC sampling.

### *9.5.6 Addressing Subjectivity Bias in Inverse Problems*

The prior information injected into geophysical inversion may be biased to the prior knowledge and experience of the individual geoscientists involved. In order to reduce the subjectivity bias, use of machine learning models is proposed to build prior information (figure 9.2). Artificial neural networks (ANNs) may be trained on outcrops and borehole data (well logs and cores) to learn geological patterns which may subsequently be used as prior information. ANNs trained on a large variety of data from different geological environments may encode such information in terms of probability distributions. This is expected to reduce the subjectivity bias significantly.

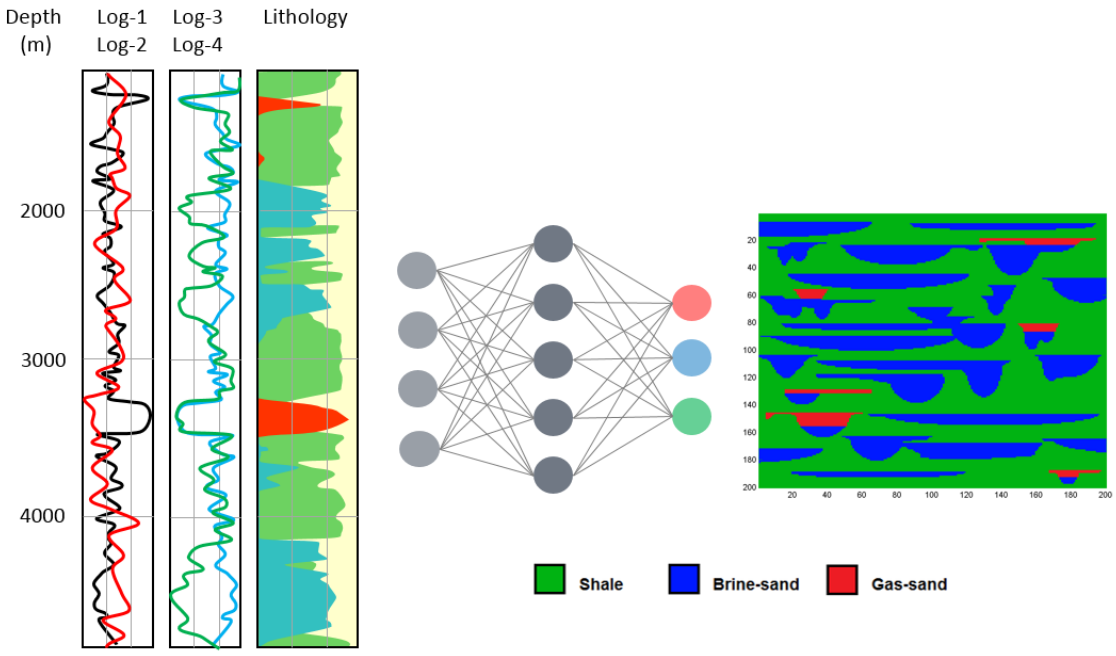



*Figure 9.2: Geological facies patterns may be built from well data to represent prior information in a given geological environment using neural networks.*

### *9.5.7 Real-Time Reservoir Monitoring and Earthquake Early Warning*

Extension of the methods developed in this thesis to seismic travel-time tomography may find useful applications in real-time monitoring of subsurface reservoirs for resource production and for waste storage for climate change mitigation. As a result, real-time decisions could be made to ensure safe and productive field operations. Another application is the development of more accurate earthquake early warning (EEW) systems (figure 9.3). Since MC sampling is too slow for real-time applications, the existing EEW systems (e.g. Burkett *et al.* 2014) use ad-hoc and subjective criteria for identification of possible earthquakes. Consequently, these systems generate a large number of false-positive warnings (Finazzi et al. 2016). The development of fast, fully probabilistic travel-time estimation will eradicate this limitation of the existing EEW systems and will allow more reliable early warnings for earthquakes. In the second step, full wavefield subsurface imaging methods will be developed to provide detailed probabilistic results with applications in characterization and time-lapse monitoring of subsurface reservoirs, for development planning, reserves estimation and risk and economic forecasting with probabilistic reservoir simulation models, while accounting for uncertainties in the results. These methods will also be useful in fast and reliable uncertainty assessment in non-geophysical applications such as medical imaging (e.g. CT and MRI scans).

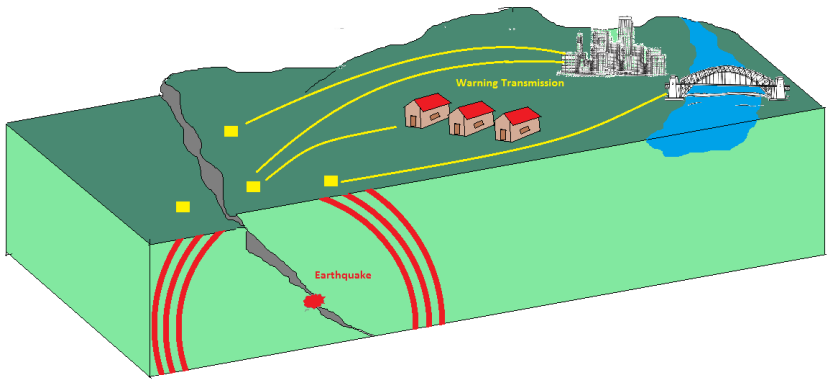



*Figure 9.3: Earthquake early warning system (EEW). Detection of waves at seismometers triggers warnings sent to cities at the speed of light with messages containing expected seismic waves intensities and arrival times.*

# Chapter 10 Conclusions

New concepts, models and methods were developed in this thesis to perform more efficient probabilistic inversion by making use of the latest developments in machine learning and Bayesian inverse theory to solve geophysical inverse problems. The major contribution of this thesis is the development of efficient geostatistical inversion methods for approximate inference for structured inverse problems where probabilistic dependence between unknown model parameters may be expressed as a *Markov random field* (MRF). The methods developed in this thesis perform inference within an optimization framework, thus circumventing the need for stochastic sampling, while still providing probabilistic results. These methods are many orders of magnitude faster than the corresponding sampling based methods in such types of inverse problems.

Further, the assumptions of localized likelihoods and conditional independence of data that are commonly used in conventional geostatistical inversion methods are relaxed and/or removed in this research while still allowing computationally tractable solutions to be found for suitably structured problems. The class of problems considered here spans a broad range of spatial data analysis and geosciences. The methods developed in this thesis infer the post-inversion (posterior) probability density of the unknown model parameters from geophysical data and geological prior information. These methods are shown to be robust against weak prior information and correlated noise in the data.

# References

Aki, K., and Richards, P. G., 1980, *Quantitative seismology*: W. H. Freeman & Co.

Arpat G. B., 2005. Sequential simulation with patterns, PhD thesis, Stanford University.

Asif, A., and Moura, J. M. F., 2005, Block matrices with l-block-banded inverse: Inversion algorithms: *IEEE Transactions on Signal Processing*, **53**(2).

Avseth, P., Mukerji, T. & Mavko, G., 2005. *Quantitative Seismic Interpretation*, Vol. 1, Cambridge Univ. Press, Cambridge, U.K., ISBN: 9780521151351.

Bachrach, R., 2006. Joint estimation of porosity and saturation using stochastic rock-physics modelling, *Geophysics*, **71**(5), O53-O63.

Balakrishnan, S., Wainwright, M.J., & Yu, B., 2017. Statistical guarantees for the EM algorithm: From population to sample-based analysis. The Annals of Statistics, 45(1), 77–120, doi: 10.1214/16-AOS1435.

Beal, M.J. 2003. Variational Algorithms for Approximate Bayesian Inference, PhD. Thesis, Gatsby Computational Neuroscience Unit, University College London.

Besag, J., 1974. Spatial interaction and the statistical analysis of lattice systems, *Journal of the Royal Statistical Society, Series B (Methodological)*, **36**(2), 192-236.

Bethe, H. 1935.Statistical Theory of Superlattices, Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences. **150**(871), 552-575.

Bishop, C. 1995. *Neural networks for pattern recognition*. Oxford University Press.

Bishop, C. M., 2006, *Pattern Recognition and Machine Learning*: Springer Science+Business Media. p.690.

Bosch, M., Carvajal, C., Rodrigues, J., Torres, A., Aldana, M., & Sierra, J. 2009. Petrophysical seismic inversion conditioned to well-log data: Methods and application to a gas reservoir. *Geophysics*, **74**(2), O1-O15. https://doi.org/10.1190/1.3043796.

Bosch, M., Mukerji, T., & Gonzalez, E. F., 2010. Seismic inversion for reservoir properties combining statistical rock physics and geostatistics: A review, *Geophysics*, **75**(5), 75A165-75A176

Buland, A. & Omre, H., 2003a. Bayesian linearized AVO inversion, *Geophysics*, **68**(1), 185-198.

Buland, A., and Omre, H., 2003b, Joint AVO inversion, wavelet estimation and noise-level estimation using a spatially coupled hierarchical Bayesian model: *Geophysical Prospecting*, **51**, 531-550. doi:10.1046/j.1365-2478.2003.00390.x

Burkett, E.R., Give, D.D. & Jones, L.M. 2014. ShakeAlert—An earthquake early warning system for the United States west coast, Fact Sheet 2014-3083, https://doi.org/10.3133/fs20143083.

Caers, J. & Ma, X., 2002. Modeling conditional distributions of facies from seismic using neural nets. *Mathematical Geology*, 34, 143-167. https://doi.org/10.1023/A:1014460101588.

Caers, J., & Zhang, T., 2004, Multiple-point geostatistics: a quantitative vehicle for integrating geologic analogs into multiple reservoir models, in Grammer, G. M., Harris, P. M., and Eberli, G. P., eds., *Integration of outcrop and modern analogs in reservoir modeling*. Am. Assoc. Petrol. Geol. Memoir p. 384–394.

Caers, J., Hoffman, T., Strebelle, S. & Wen, X.H., 2006. Probabilistic integration of geologic scenarios, seismic, and production data—a West Africa turbidite reservoir case study. *The Leading Edge*, 25(3), 240-244. https://doi.org/10.1190/1.2184087.

Cordua, K., Hansen, T. & Mosegaard, K. 2012. Monte Carlo full-waveform inversion of crosshole GPR data using multiple-point geostatistical a priori information. *Geophysics*. 77. 19-. 10.1190/geo2011-0170.1.

Chopra, S. & Larsen, G., 2000. Acquisition footprint – Its detection and removal. Recorder, Canadian Society of Exploration Geophysicists, 25(8), October, 2000.

Chopra, S., Castagna, J. & Xu, Y. 2009. Thin-bed reflectivity inversion and some applications. First Break. 31. 27-34. 10.3997/1365-2397.2009009.

Cortes, C. & Vapnik, V.N., 1995. Support-vector networks. Machine Learning, 20(3):273-297.

Cover, T. & Thomas, J., 1991. Elements of information theory, John Wiley & Sons.

Curtis, A. & Lomax, A., 2001. Prior information, sampling distributions and the curse of dimensionality, *Geophysics*, **66**, 372-378, 2001.

Curtis, A. & Wood, R., (editors), 2004a. Geological Prior Information; Informing Science and Engineering. Geological Society of London Special Publication, 239.

Curtis, A. & Wood, R., 2004b. Optional elicitation of probabilistic information from experts. In: *Geological Prior Information: Informing Science and Engineering*. Geological Society of London. Special Publication, 239.

Curtis, A., 2012. The science of subjectivity. *Geology*. 40, pp. 95-96. doi:10.1130/focus012012.1.

Davis, P.J., 1959. Leonhard Euler's Integral: A Historical Profile of the Gamma Function. American Mathematical Monthly. 66 (10): 849–869. doi:10.2307/2309786. JSTOR 2309786.

Debeye, H.W.J., Sabbah, E., and van der Made, P. M., 1996. Stochastic inversion. SEG Technical Program Expanded Abstracts 1996: pp. 1212-1215.

Dempster, A.P., Laird, N.M. and Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society*, Series B (Methodological), **39**: 1–38. 2.2.2, 2.4.3.

Denardo, E.V., 2003. Dynamic programming: Models and applications, Mineola, NY: Dover Publications, ISBN 978-0-486-42810-9.

Dennis, J.E. & Schnabel, R.B., 1996. Numerical methods for unconstrained optimization and nonlinear equations. SIAM, Mathematics.

Doyen, P.M., Guidish, T.M., & de Buyl, M.H., 1989. Monte Carlo Simulation of Lithology from Seismic Data in a Channel-Sand Reservoir, *SPE paper* # 19588.

Dvorkin, J. & Nur, A., 1996. Elasticity of high-porosity sandstones: Theory for two North Sea data sets. *Geophysics*, **61**(5), 1363-1370. https://doi.org/10.1190/1.1444059.

Eddy, S.R., 1998. Profile hidden Markov models, *Bioinformatics*. **14**, 755–763.

Eidsvik, J., Avseth, P., Omre, H., Mukerji, T. & Mavko, G., 2004. Stochastic reservoir characterization using prestack seismic data. *Geophysics*. 69. 10.1190/1.1778241.

Feynman. R.P., 1972. *Statistical Mechanics: A Set of Lectures*, Perseus, Reading, MA, 1972. 2.2.1, 2.3.2.

Finazzi, F., 2016. The Earthquake Network Project: Toward a Crowdsourced Smartphone-Based Earthquake Early Warning System. *Bulletin of the Seismological Society of America*; 106 (3): 1088–1099. doi: https://doi.org/10.1785/0120150354.

Fletcher, R., 1987. *Practical methods of optimization*. 2nd Edition, New York, John Wiley & Sons, ISBN 978-0-471-91547-8.

Foster, D.J., & Mosher, C.C., 1992. Suppression of multiple reflections using the Radon transform. Geophysics, 57(3), 386-395. doi:10.1190/1.1443253.

Foster, D.J., Keys, R.G. & Lane, F.D., 2010. Interpretation of AVO anomalies. Geophysics, 75(5), 75A3-75A13. doi:10.1190/1.3467825.

Francis, A., 2005. Limitations of deterministic and advantages of stochastic inversion, *CSEG Recorder*.

Gallagher, K., Charvin, K., Nielsen, S., Sambridge, M. & Stephenson, J., 2009. Markov chain Monte Carlo (McMC) sampling methods to determine optimal models, model resolution and model choice for Earth Science problems, *Mar. Pet. Geol.*, **26**(4), 525–535.

Gardner, G.H.F, Gardner, L.W., and Gregory A.R., 1974. Formation velocity and density – the diagnostic basics for stratigraphic traps. Geophysics, 39, 770-780. Doi:10.1190/1.1440465.

Grana D. & Della Rossa E., 2010. Probabilistic petrophysical-properties estimation integrating statistical rock physics with seismic inversion, *Geophysics*, **75**(3), O21-O37.

Grana D., Mukerji T., Dvorkin J., & Mavko G., 2012. Stochastic inversion of facies from seismic data based on sequential simulations and probability perturbation method, *Geophysics*, **77**(4), M53-M72.

Grana, D., Paparozzi, E., Mancini, S. & Tarchiani, C., 2013. Seismic driven probabilistic classification of reservoir facies for static reservoir modelling: A case history in the Barents Sea. Geophysical Prospecting. 61. 613-629. 10.1111/j.1365-2478.2012.01115.x.

Grana D. & Mukerji T., 2015. Bayesian inversion of time-lapse seismic data for the estimation of static reservoir properties and dynamic property changes, *Geophysical Prospecting*, **63**(3), 637- 655.

Grana, D., Lang, X., & Wu, W., 2016. Statistical facies classification from multiple seismic attributes: comparison between Bayesian classification and expectation–maximization method and application in petrophysical inversion, *Geophysical Prospecting*, **65**(2), 544-562.

Grana, D., Fjeldstad, T. & Omre, H., 2017. Bayesian Gaussian Mixture Linear Inversion for Geophysical Inverse Problems. *Math Geoscience*, **49**, 493, doi: 10.1007/s11004-016-9671-9.

Grana, D. 2018. Joint facies and reservoir properties inversion. *Geophysics*, **83**(3), M15-M24. https://doi.org/10.1190/geo2017-0670.1.

Griffiths, C., Dyt, C., Paraschivioiu, E., & Liu, K., 2001. SEDSIM in hydrocarbon exploration. In: Merrian, D.F., Davis, J.C. (Eds.), *Geologic Modeling and Simulation*. Kluwer Academic/Plenum Publishers, New York, pp. 71–99.

Guardiano F. B., Srivastava R. M., 1993. Multivariate geostatistics: beyond bivariate moments. In: *Geostatistics Troia'92. Springer*, pp 133–144.

Gultekin, S., Zhang, A. & Paisley, J. 2018. Asymptotic Simulated Annealing for Variational Inference. arXiv:1505.06723v1 [stat.ML].

Haas, A., & Dubrule, O., 1994. Geostatistical inversion — A sequential method of stochastic reservoir modeling constrained by seismic data, *First Break*, **12**, 561–569.

Hammer, H.L. & Tjelmeland, H., 2011. Approximate forward–backward algorithm for a switching linear Gaussian model. *Computational Statistics & Data Analysis*. vol. 55 (1).

Hammersley, J. M., Clifford, P., 1971. Markov fields on finite graphs and lattices, *unpublished work*.

Hansen, T.M., Journel, A.G., Tarantola, A., Mosegaard, K., 2006. Linear inverse Gaussian theory and geostatistics. *Geophysics*, 71(6), R101–R111.

Hansen, T.M., Mosegaard, K., Cordua, K.C., 2008. Using geostatistics to describe complex a priori information for inverse problems. In: Ortiz, J.M., Emery, X. (eds.) VIII *International Geostatistics Congress*, vol. 1, pp. 329–338. Mining Engineering Department, University of Chile, Santiago.

Hansen, T.M., Cordua, K.S., Zunino, A., & Mosegaard, K., 2016. Probabilistic integration of geo-information. In: Integrated imaging of the earth: theory and applications, pp.93-116. Wiley. doi: 10.1002/9781118929063.ch6.

Hill, J., Tetzlaff, D., Curtis, A. & Wood, R. 2009. Modeling shallow marine carbonate depositional systems. *Computers & Geosciences*, **35**, pp. 1862–1874.

Hinton, G.E., and Zemel, R.S., 1994, Autoencoders, Minimum Description Length, and Helmholtz Free Energy: *Advances in Neural Information Processing Systems 6*. J. D. Cowan, G. Tesauro and J. Alspector (Eds.), Morgan Kaufmann: San Mateo, CA.

Ho, T.K., 1995. Random decision forests. Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14-16 August 1995. Pp. 278-282.

Hoffman, B.T. and Caers, J,. 2007. History matching by jointly perturbing local facies proportions and their spatial distribution: Application to a North Sea Reservoir. Journal of Petroleum Sciences and Technology, 57, 3-4, 257-272

Jaakkola, T.S., 1997. Variational methods for inference and learning in graphical models. PhD thesis, Massachusetts Institute of Technology (MIT).

Jebara, T., 2002. *Discriminative, generative and imitative learning*. PhD thesis, Massachusetts Institute of Technology (MIT).

Jin, B., and Zou, J., 2010, Hierarchical Bayesian inference for ill-posed problems via variational method: *Journal of Computational Physics*, **229**(19), 7317-7343.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K., 1999, An introduction to variational methods for graphical models: *Mach. Learn.*, **37**, 183–233.

Journel A. G., 1974. Geostatistics for conditional simulation of orebodies. Economic Geol **69**(5):673–687

Journel, A. & Zhang, T., 2006. The necessity of a multiple-point prior model, *Mathematical Geology*, **38**(5), pp 591–610.

Kiebel, S. J., Daunizeau, J., Phillips, C., and Friston, K. J., 2008, Variational Bayesian inversion of the equivalent current dipole model in EEG/MEG: Neuroimage. 2008 Jan 15, **39**(2):728-41.

Koc, C.K. & Piedra, R.M., 1991. A parallel algorithm for exact solution of linear equations, In *Proceedings of International Conference on Parallel Processing*, Volume III, pages 1-8, St. Charles, IL, August 12-16, 1991. Boca Raton, FL: CR Press.

Koller, D. & Friedman, N., 2009. *Probabilistic Graphical Models*: Principles and Techniques. *MIT Press*.

Kumar, S. & Hebert, M., 2003. Discriminative fields for modeling spatial dependencies in natural images. In: *Advances in Neural Information Processing Systems*, 16. MIT Press, Cambridge, MA.

Lafferty, J., McCallum, A. & Pereira, F.C.N., 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the 18th ICML 2001, 282-289.

Lang, X. & Grana, D. 2018. Bayesian linearized petrophysical AVO inversion. *Geophysics*, **83**(3), M1-M13. https://doi.org/10.1190/geo2017-0364.1

Larsen, A.L., Ulvmoen, M., Omre, H., & Buland, A., 2006. Bayesian lithology/fluid prediction and simulation on the basis of a Markov-chain prior model, *Geophysics*, 71(5), R69–R78.

Lin, L., Yang, C., Meza, J. C., Lu, J., Ying, L., and Weinan, E., 2011, SelInv – An algorithm for selected inversion of a sparse symmetric matrix: *ACM Transactions on Mathematical Software*, **37**(4), p. 40.

Loïc S., 2016. Exact Bayesian Inference in Graphical Models : Tree-structured Network Inference and Segmentation. *Statistics [math.ST]*. *Université Paris-Saclay*, 2016. NNT:2016SACLS210.

Lindberg, D. & Omre, H., 2014. Blind categorical deconvolution in two level hidden Markov models, *IEEE Trans. Geosci. Remote Sens.*, **52**(11), 7435-7447.

Lindberg, D. & Omre, H., 2015. Inference of the transition matrix in convolved hidden Markov models and the generalized Baum-Welch algorithm, *IEEE Trans. Geosci. Remote Sens.*, **53**(12), 6443-6456.

Lindberg, D.V., Rimstad, E., & Omre, H., 2015. Inversion of well logs into facies accounting for spatial dependencies and convolution effects, *Journal of Petroleum Science and Engineering*, **134**, 237–246.

Luo, X. & Tjelmeland, H., 2017.. Prior specification for binary Markov mesh models. Statistics and Computing. doi:10.1007/s11222-018-9813-7.

Macrae, E.J., Bond, C.E., Shipton, Z.K., & Lunn, R.J., 2016. Increasing the quality of seismic interpretation. Interpretation, 4(3), T395-T402. doi: 10.1190/INT-2015-0218.1.

Marion, D.P., 1990. Acoustical, mechanical, and transport properties of sediments and granular materials, PhD thesis, Stanford University, Department of Geophysics.

Mariethoz, G. & Caers, J., 2014. Multiple-point Geostatistics: Stochastic Modeling with Training Images, Wiley-Blackwell.

Martinsson, P.G., Rokhlin, V., and Tygert, M., 2005, A fast algorithm for the inversion of general Toeplitz matrices. *Computers & Mathematics with Applications*, 50(5):741-752. doi: 10.1016/j.camwa.2005.03.011.

Mavko G., Mukerji T., Dvorkin J. 2009. *The rock physics handbook: tools for seismic analysis of porous media*. Cambridge University Press, London.

McLachlan, G. & Peel, D. 2000. *Finite Mixture Models*. Wiley Interscience.

Meier, U., Curtis, A. & Trampert, J., 2007a. A global crustal model constrained by non-linearised inversion of fundamental mode surface waves. *Geophys. Res. Lett.*, **34**, L16304.

Meier, U., Curtis, A. & Trampert, J., 2007b. Global crustal thickness from neural network inversion of surface wave data. *Geophys. J. Int.*, **169**, 706-722.

Meier, U., Trampert, J. & Curtis, A., 2009. Global variations of temperature and water content in the mantle transition zone from higher mode surface waves. *Earth and Planetary Science Letters*, **282**, 91–101.

Mohammad-Djafari, A., & Ayasso, H., 2009. Variational Bayes and mean field approximations for Markov field unsupervised estimation. In: *Proceedings of IEEE International Workshop on Machine Learning for Signal Processing*, Grenoble, France, 2–4 September 2009; pp. 1–6.

Mooij, J.M. & Kappen, J.H., 2007. Sufficient conditions for convergence of the sum-product algorithm, *IEEE Transactions on Information Theory*, **53**(12), 4422-4437.

Mosegaard, K. & Tarantola, A., 1995. Monte Carlo sampling of solutions to inverse problems. Journal of Geophysical Research: Solid Earth, 100(B7), 12431-12447.

Mosegaard, K., and Sambridge, M., 2002. Monte Carlo analysis of inverse problems, Inverse Problems, 18, R29-R54.

Mosegaard, K. & Tarantola, A., 2002. Probabilistic Approach to Inverse Problems. In "*International Handbook of Earthquake and Engineering Seismology*". Academic Press, pp. 237-265.

Mukerji, T., Jørstad, A., Avseth, P., Mavko, G., & Granli, J.R., 2001. Mapping lithofacies and pore-fluid probabilities in a North Sea reservoir: Seismic inversions and statistical rock physics, *Geophysics*, **66**, SPECIAL SECTION, 988-1001.

Murphy, K. P., Weiss, Y., & Jordan, M. I., 1999. Loopy belief propagation for approximate inference: An empirical study, *In Proceedings of Uncertainty in AI*, **9**, 467–475.

Nawaz, M.A. & Curtis, A., 2017. Bayesian inversion of seismic attributes for geological facies using a hidden Markov model, *Geophys. J. Int*.**208**, 1184–1200.

Nawaz, M.A. & Curtis, A., 2018. Variational Bayesian inversion of seismically derived non-localized rock properties for the spatial distribution of geological facies, Geophys. J. Int. 214, 845–875. doi: 10.1093/gji/ggy163.

Nawaz, M.A. & Curtis, A., 2019. Rapid discriminative variational Bayesian inversion of geophysical data for the spatial distribution of geological properties, *JGR – Solid Earth*, (Accepted for publication).

Nawaz, M.A., Curtis, A., Shahraeeni, M.S., & Gerea, C., 2019. Variational Bayesian Inversion of Seismic Attributes Jointly for Geological Facies and Petrophysical Rock Properties, *Geophysics*, (Submitted).

Neal R.M., Hinton G.E. 1998. A view of the EM Algorithm that justifies incremental, sparse, and other variants. In: Jordan M.I. (eds) Learning in Graphical Models. *NATO ASI Series* (Series D: Behavioural and Social Sciences), vol 89. Springer, Dordrecht.

Neal, R. M., and Hinton, G. E., 1999, A view of the EM algorithm that justifies incremental, sparse, and other variants: *Learning in graphical models*, MIT Press, Cambridge, MA.

Ng, A.Y. & Jordan, M.I., 2002. On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. *Advances in neural information processing systems*, 841-848.

Nocedal, J. 1980. Updating quasi-Newton matrices with limited storage. *Mathematics of Computation*. 35(151): 773-782. doi:10.1090/S0025-5718-1980-0572855-7.

Nocedal, J. & Wright, S., 2006. *Numerical Optimization*. 2nd Edition. Springer Series in Operations Research and Financial Engineering, Springer-Verlag New York.

Nunes, R., Soares, A., Azevedo, L. & Pereira, P., 2016. Geostatistical seismic inversion with direct sequential simulation and co-simulation with multi-local distribution functions, *Mathematical Geoscience*, 1–19.

Opper, M. & Saad, D. (Eds.), 2001. Advanced Mean Field Methods: Theory and Practice. The MIT Press, Cambridge, Massachusetts, London, England, 2001, Neural Information Processing Series, 273 pages, hardbound, ISBN 0-262-15054-9.

Pearl, J., 1982. Reverend Bayes on inference engines: A distributed hierarchical approach, *Proceedings of the Second National Conference on Artificial Intelligence*, AAAI-82: Pittsburgh, PA. Menlo Park, California. AAAI Press. 133–136.

Pearl, J., 1988. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference,2nd ed.,*Morgan Kaufmann Publishers Inc*., San Francisco, CA, USA. ISBN 1-55860-479-0.

Penz, S., Duchêne, B., and Mohammad-Djafari, A., 2018. Variational Bayesian inversion of synthetic 3D controlled-source electromagnetic geophysical data. *Geophysics*, **83**(1), E25-E36. Doi:10.1190/geo2016-0682.1.

Polson, D. & Curtis, A., 2010. Dynamics of uncertainty in geological interpretation. Journal of the Geological Society, 167:5-10.

Polson, D. & Curtis, A., 2015. Assessing individual influence on group decisions in geological carbon capture and storage problems. Chapter in, "Collaborative knowledge in scientific research networks", ed. P Diviacco, P. Fox, C. Pshenichny and A. Leadbetter, pub. IGI Books. DOI: 10.4018/978-1-4666-6567-5.ch004.

Rabben, T. E., Tjelmeland, H., and Ursin, B., 2008, Non-linear Bayesian joint inversion of seismic reflection coefficients. *Geophysical Journal International*, **173**(1), p:265–280.

Randen, T. and Sønneland, L., 2005. Atlas of 3D Seismic Attributes. Mathematical Methods and Modelling in Hydrocarbon Exploration and Production, Springer, pp 23-46.

Rencher, A.C., 2002. *Methods of Multivariate Analysis*. 2nd Edition, Wiley-Interscience.

Rimstad, K. & Omre, H., 2010. Impact of rock-physics depth trends and Markov random fields on hierarchical Bayesian lithology/fluid prediction, *Geophysics*, **75**(4), R93–R108.

Rimstad, K., Avseth, P. & Omre, H., 2012. Hierarchical Bayesian lithology/fluid prediction: a North Sea case study, *Geophysics*, **77**(2), B69-B85.

Rimstad, K. & Omre, H., 2010. Impact of rock-physics depth trends and Markov random fields on hierarchical Bayesian lithology/fluid prediction, *Geophysics*, **75**(4), R93–R108.

Rimstad, K. & Omre, H., 2013. Approximate posterior distributions for convolutional two-level hidden Markov models, *Computational Statistics & Data Analysis*. **58**.

Rezvandehy, M. & Deutsch C.V., 2017. Horizontal variogram inference in the presence of widely spaced well data. *Petroleum Geoscience*, 24, 219-235, doi:10.1144/petgeo2016-161.

Rue, H., & Held, L., 2005, *Gaussian Markov Random Fields: Theory and Applications*: Monographs on Statistics and Applied Probability 104, Chapman & Hall/CRC.

Saddiki, H., Trapp, A.C. & Flaherty, P. 2017. A Deterministic Global Optimization Method for Variational Inference. arXiv:1703.07169v1 [stat.ME].

Sambridge, M. & Mosegaard, K., 2002. Monte Carlo Methods in Geophysical Inverse Problems, *Rev. Geophys.*, **40**(3), 3-1–3-29.

Sambridge, M., Bodin, T., Gallagher, K., and Tkalčić, H., 2013, Transdimensional inference in the geosciences, *Phil. Trans. R. Soc. A*, 371, 20110547. doi: 10.1098/rsta.2011.0547

Shahraeeni, M.S. & Curtis, A., 2011. Fast probabilistic nonlinear petrophysical inversion, *Geophysics*, **76**(2), E45-E58.

Shahraeeni, M.S., Curtis, A. & Chao, G., 2012. Fast probabilistic petrophysical mapping of reservoirs from 3D seismic data, *Geophysics*, **77**(3), O1-O19.

Shannon, C.E., 1948. A mathematical theory of communication. *Bell System Technical Journal*, **27**(3), 379-423.

Sinoquet, C. & Mourad, R., 2014. Probabilistic graphical models for genetics, genomics and postgenomics, *Keith Mansfield. Oxford University Press*, 480, 978-0-19-870902-2.

Springenberg, J.T., Dosovitskiy, A., Brox, T., & Riedmiller, M.A., 2014. Striving for Simplicity: The All Convolutional Net. *CoRR*, abs/1412.6806.

Stanley, H.E., 1971. Mean field theory of magnetic phase transitions. Introduction to phase transitions and critical phenomena. Oxford University Press. ISBN 0-19-505316-8.

Strebelle S., 2001.Conditional simulation of complex geological structures using multiple-point statistics. *Mathematical Geology*, **34**(1), 1-21.

Sudderth, E. & Freeman, W., 2008. Signal and image processing with belief propagation, *IEEE Signal Processing Magazine*, **25**(2), 114-141.

Sutton, C. & McCallum, A., 2012. An introduction to conditional random fields. Foundations and Trends® in Machine Learning: 4(4), 267-373.

Tahmasebi P., 2018. Multiple Point Statistics: A Review. In: Daya Sagar B., Cheng Q., Agterberg F. (eds) Handbook of Mathematical Geosciences. Springer, Cham, SN:978-3-319-78999-6, DOI: 10.1007/978-3-319-78999-6_30.

Tarantola, A. & Valette, B., 1982. Inverse problems = quest for information. J. Geophys, 50(3), 150-170.

Tatikonda, S.C. & Jordan, M.I., 2002. Loopy belief propagation and Gibbs measures. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence (UAI'02)*, Adnan Darwiche and Nir Friedman (Eds.). *Morgan Kaufmann Publishers Inc.*, San Francisco, CA, USA, 493-500.

Tarantola, A., 2005, *Inverse problem theory and methods for model parameter estimation*. Society for Industrial and Applied Mathematics

Tikhonov, A.N., 1963. "О решении некорректно поставленных задач и методе регуляризации". Doklady Akademii Nauk SSSR. 151: 501–504.. Translated in "Solution of incorrectly formulated problems and the regularization method". Soviet Mathematics. 4: 1035–1038.

Tobler, W., 1970. A computer movie simulating urban growth in the detroit region. *Economic Geography*, **46**, 234-240.

Ulvmoen, M., Omre, H., 2010. Improved resolution in Bayesian lithology/fluid inversion from prestack seismic data and well observations, Part 1 — Methodology. *Geophysics*, **75**(2), R21-R35.

Ulvmoen, M., Omre, H. & Buland, A., 2010. Improved resolution in Bayesian lithology/fluid inversion from prestack seismic data and well observations: Part 2 — Real case study. *Geophysics*, **75**(2), B73-B82.

Wainwright, M.J. & Jordan, M.I., 2008. Graphical Models, Exponential Families, and Variational Inference, *Foundations and Trends® in Machine Learning*: **1**(1–2), doi: 10.1561/2200000001.

Walden, A.T. and White, R.E., 1998, Seismic wavelet estimation: a frequency domain solution to a geophysical noisy input-output problem: *IEEE Transactions on Geoscience and Remote Sensing*, 36, 287–297.

Walker, M. & Curtis, A., 2014a. Spatial Bayesian inversion with localized likelihoods: an exact sampling alternative to McMC. *J. Geophys. Res. Solid Earth*, **119**, 5741-5761.

Walker, M. & Curtis, A., 2014b. Expert elicitation of geological spatial statistics using genetic algorithms. Geophys. J. Int., 198, pp.342–356, doi: 10.1093/gji/ggu132.

Wang, H., Wellmann, J.F., Li, Z., Wang, X., & Liang, R.Y., 2016. A segmentation approach for stochastic geological modeling using hidden Markov random fields. *Mathematical Geosciences*, **49**(2), 145–177.

White, R.E. 1984. Signal and Noise Estimation from Seismic Reflection Data Using Spectral Coherence Methods. *Proceedings of the IEEE*. 72. 1340 - 1356. 10.1109/PROC.1984.13022.

Wu, X., Shi, Y., Fomel, S., & Liang, L. 2018. Convolutional neural networks for fault interpretation in seismic images. *SEG Technical Program Expanded Abstracts 2018*: 1946-1950.

Xing, E.P., Jordan, M. I., and Russell, S., 2003, A generalized mean field algorithm for variational inference in exponential families. In *Uncertainty in Artificial Intelligence*, **19**, 2003.

Yanqin, L. & Zhang, G. 2014. Blind Seismic Deconvolution Using Variational Bayesian Method. *Journal of Applied Geophysics*. 110. 10.1016/j.jappgeo.2014.09.002.

Yasuda, M., Kataoka, S. & Tanaka, K., 2015. Statistical analysis of loopy belief propagation in random fields, *Phys. Rev*., **92**, 042120, doi: 10.1103/PhysRevE.92.042120.

Yedidia, J.S., Freeman, W.T., and Weiss, Y., 2001a. Bethe free energy, Kikuchi approximations and belief propagation algorithms. Technical report, Mitsubishi Electric Res. Labs. TR-2001-16.

Yedidia, J.S., Freeman, W.T., & Weiss, Y., 2001b.Understanding belief propagation and its generalizations, *Technical report, Mitsubishi Electric Res. Labs*. **TR-2001-15**.

Yin, H., Nur, A., & Mavko, G., 1993. Critical porosity a physical boundary in poroelasticity, *Int. J. Rock Mech. Min. Sci. Geomech. Abstr.*, **30**(7), 805–808.

Yuan, Y., 2010. Gradient Methods for Large Scale Convex Quadratic Functions. In: Optimization and Regularization for Computational Inverse Problems and Applications, Editors: Yanfei Wang, Anatoly G. Yagola, and Changchun Yang, Higher Education Press, Beijing and Springer-Verlag Berlin Heidelberg.

Zhang, W., 1988. Shift-invariant pattern recognition neural network and its optical architecture. Proceedings of annual conference of the Japan Society of Applied Physics.

Zhang, R., Ye, D. H., Pal, D., Thibault, J.-B., Sauer, K. D., & Bouman, C. A., 2016. A Gaussian mixture MRF for model-based iterative reconstruction with applications to low-dose X-ray CT, *IEEE Trans. Comput. Imag.*, **2**(3), 359-374.

Zhao, B., Zhong, Y., Ma, A., & Zhang, L., 2016. A spatial Gaussian mixture model for optical remote sensing image clustering. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 2016*, **9**, 5748–5759.

Zoeppritz, K. 1919. "VIIb. Über Reflexion und Durchgang seismischer Wellen durch Unstetigkeitsflächen." [VIIb. On reflection and transmission of seismic waves by surfaces of discontinuity], Nachrichten von der Königlichen Gesellschaft der Wissenschaften zu Göttingen, *Mathematisch-physikalische Klasse*, 66–84.

## Appendix A: Mathematical Derivation of Equation 6.10

To derive equation 6.10 from equation 6.9, we substitute equation 6.9 into equation 6.7 in the main text which gives

$$\mathcal{F}(\mathcal{Q}, \boldsymbol{w}) = \sum\nolimits_{\hat{c}\in\hat{\mathcal{C}}} \mathbb{E}_{\mathcal{Q}}[\boldsymbol{w}^T \boldsymbol{f}(\boldsymbol{m}_{\hat{c}}, \boldsymbol{d})] + \sum\nolimits_{c} \mathcal{S}(\mathcal{Q}_c)$$

$$= \sum\nolimits_{\hat{c}\in\hat{\mathcal{C}}} \int_{\boldsymbol{m}_{\hat{c}}} \left( \prod\nolimits_{c\subset\hat{c}} \mathcal{Q}_c(\boldsymbol{m}_c|\boldsymbol{d}) \right) \boldsymbol{w}^T \boldsymbol{f}(\boldsymbol{m}_{\hat{c}}, \boldsymbol{d}) d\boldsymbol{m}_{\hat{c}}$$

$$- \sum\nolimits_{c} \int_{\boldsymbol{m}_c} \mathcal{Q}_c(\boldsymbol{m}_c|\boldsymbol{d}) \log \mathcal{Q}_c(\boldsymbol{m}_c|\boldsymbol{d}) \, d\, \boldsymbol{m}_c \qquad \text{A1}$$

The maximum computational complexity of the expectation term (the 1st term in equation A1) is $O\left(\left|\hat{\mathcal{C}}\right| * |\boldsymbol{m}_{\hat{c}}|\right)$ and of the entropy term (the 2nd term in equation A1) is $O(|\mathcal{C}| * |\boldsymbol{m}_c|)$, where $\left|\hat{\mathcal{C}}\right|$ and $|\mathcal{C}|$ are respectively the number of maximal cliques $\hat{c}$ and the approximating cliques $c$ in the graph, and $|\boldsymbol{m}_{\hat{c}}|$ and $|\boldsymbol{m}_c|$ are the maximum dimensionality of model parameters in the maximal and the approximating cliques in the graph, respectively. Thus evaluation of the free energy functional in the above equation can be performed in time that is linear in the maximum dimensionality of factors in $\mathcal{Q}(\boldsymbol{m}|\boldsymbol{d})$ (or cliques in the graph).

In order to optimize the marginal distributions $\mathcal{Q}_c(\boldsymbol{m}_c|\boldsymbol{d})$ of any restricted variational distribution $\mathcal{Q}(\boldsymbol{m}|\boldsymbol{d}) \in \mathbb{Q}$, we iteratively maximize the variational free energy $\mathcal{F}(\mathcal{Q}, \boldsymbol{w})$ within the family $\mathbb{Q}$ of factorizable distributions by successively optimizing each of the marginal distributions at a time. The factorizable form of the mean field (MF) variational distribution allows successive optimization of each factor (marginal distribution) while keeping others fixed in an iterative fashion. We may characterize stationary points of the marginal distribution $\mathcal{Q}_c(\boldsymbol{m}_c|\boldsymbol{d})$ in terms of the rest of the marginals $\mathcal{Q}_{\setminus c}\left(\boldsymbol{m}_{\setminus c}|\boldsymbol{d}\right)$ by restricting the energy functional $\mathcal{F}(\mathcal{Q}, \boldsymbol{w})$ to the terms involving $\mathcal{Q}_c(\boldsymbol{m}_c|\boldsymbol{d})$, which gives

$$\mathcal{F}(\mathcal{Q}_c, \boldsymbol{w}) = \sum\nolimits_{\hat{c}\in\hat{\mathcal{C}}} \mathbb{E}_{\mathcal{Q}}[\boldsymbol{w}^T \boldsymbol{f}(\boldsymbol{m}_{\hat{c}}, \boldsymbol{d})] + \mathcal{S}(\mathcal{Q}_c) + \text{ constant} \qquad \text{A2}$$

We seek to derive update equations for a higher-order MF approximation by characterizing the stationary points of the free energy functional using Lagrange multipliers. Since $\mathcal{F}(\mathcal{Q}_c, \boldsymbol{w})$ is concave in $\mathcal{Q}_c$, we can maximize it by forming a Lagrangian as

$$L_c(\mathcal{Q}) = \sum_{\hat{c}\in\hat{\mathcal{C}}} \mathbb{E}_{\mathcal{Q}}[\boldsymbol{w}^T \boldsymbol{f}(\boldsymbol{m}_{\hat{c}}, \boldsymbol{d})] + \mathcal{S}(\mathcal{Q}_c) + \gamma\left(1 - \int_{\boldsymbol{m}_c} \mathcal{Q}_c(\boldsymbol{m}_c|\boldsymbol{d})\, d\boldsymbol{m}_c\right) \quad \text{A3}$$

where the Lagrange multiplier $\gamma$ enforces the constraint that the marginal $\mathcal{Q}_c(\boldsymbol{m}_c|\boldsymbol{d})$ is a proper distribution. Differentiating $L_c(\mathcal{Q})$ with respect to $\mathcal{Q}_c(\boldsymbol{m}_c|\boldsymbol{d})$ and setting it equal to zero gives:

$$\mathcal{Q}_c(\boldsymbol{m}_c|\boldsymbol{d}) = \frac{1}{\mathcal{Z}_c(\boldsymbol{d})} \exp\left\{\sum_{\hat{c}\in\hat{\mathcal{C}}} \mathbb{E}_{\mathcal{Q}_{\setminus c}}[\boldsymbol{w}^T \boldsymbol{f}(\boldsymbol{m}_{\hat{c}}, \boldsymbol{d})|\, \boldsymbol{m}_c]\right\} \quad \text{A4}$$

as the necessary and sufficient condition for $\mathcal{Q}_c(\boldsymbol{m}_c|\boldsymbol{d})$ to be a local maximum of $\mathcal{F}(\mathcal{Q}_c, \boldsymbol{w})$ given the rest of the marginal distributions $\mathcal{Q}_{\setminus c}(\boldsymbol{m}_{\setminus c}|\boldsymbol{d})$. In this equation $\mathcal{Z}_c(\boldsymbol{d})$ is the local normalization constant that ensures that $\mathcal{Q}_c(\boldsymbol{m}_c|\boldsymbol{d})$ is a valid distribution, and the conditional expectation in the argument of the exponential function is given by

$$\mathbb{E}_{\mathcal{Q}_{\setminus c}}[\boldsymbol{w}^T \boldsymbol{f}(\boldsymbol{m}_{\hat{c}}, \boldsymbol{d})|\, \boldsymbol{m}_c] = \int_{\boldsymbol{m}_{c'}} \left(\prod\nolimits_{c'} \mathcal{Q}_{c'}(\boldsymbol{m}_{c'}|\boldsymbol{d})\right) \boldsymbol{w}^T \boldsymbol{f}(\boldsymbol{m}_{\hat{c}}, \boldsymbol{d})\, d\boldsymbol{m}_{c'} \quad \text{A5}$$

where $c' \in \hat{c}\backslash\{c\} \wedge |c'| = q$. The above conditional expectation in equation A5 is independent of the variational marginal distribution $\mathcal{Q}_c(\boldsymbol{m}_c|\boldsymbol{d})$, but is a function of $\boldsymbol{m}_c$ as a conditioning variable. It may be restricted to terms that involve $c$ by exploiting the conditional independence of $\boldsymbol{m}_c$ given $\boldsymbol{d}$ under the MF approximation, resulting in a closed-form update for the marginal distribution $\mathcal{Q}_c(\boldsymbol{m}_c|\boldsymbol{d})$:

$$\mathcal{Q}_c(\boldsymbol{m}_c|\boldsymbol{d}) \leftarrow \frac{1}{\mathcal{Z}_c(\boldsymbol{d})} \exp\left\{\sum_{\hat{c}\in\hat{\mathcal{C}}: c\subset\hat{c}} \mathbb{E}_{\mathcal{Q}_{\setminus c}}[\boldsymbol{w}^T \boldsymbol{f}(\boldsymbol{m}_{\hat{c}}, \boldsymbol{d}) \mid \boldsymbol{m}_c]\right\} \quad \text{A6}$$

which is same as equation 6.10 in the main text.

## Appendix B: Mathematical Derivation of Equation 7.32

For the fixed point characterization of $KL(Q \,||\, \mathcal{P})$ with respect to $Q(\boldsymbol{m})$, we take the partial derivative of $L(\boldsymbol{m}, \Theta, \gamma_1, \gamma_2)$ with respect to $Q(\boldsymbol{m})$ which gives

$$\frac{\partial}{\partial Q(\boldsymbol{m})} L(\boldsymbol{m}, \Theta, \gamma_1, \gamma_2)$$

$$= -\int \left( \frac{\partial}{\partial Q(\boldsymbol{m})} \int Q(\boldsymbol{m}) \log \frac{\mathcal{P}(\boldsymbol{m}, \Theta, \boldsymbol{d})}{Q(\boldsymbol{m}) Q(\Theta)} d\boldsymbol{m} \right) Q(\Theta) d\Theta + \gamma_1$$

$$= -\iint \left( \frac{\partial}{\partial Q(\boldsymbol{m})} \int Q(\boldsymbol{m}) (\log \mathcal{P}(\boldsymbol{m}, \Theta, \boldsymbol{d}) - \log Q(\boldsymbol{m}) - \log Q(\Theta)) d\boldsymbol{m} \right) Q(\Theta) d\Theta + \gamma_1$$

$$= -\iint \left( \frac{\partial}{\partial Q(\boldsymbol{m})} \int Q(\boldsymbol{m}) (\log \mathcal{P}(\boldsymbol{m}, \Theta, \boldsymbol{d}) - \log Q(\boldsymbol{m})) d\boldsymbol{m} - \log Q(\Theta) \right) Q(\Theta) d\Theta + \gamma_1$$

$$= -\int (\log \mathcal{P}(\boldsymbol{m}, \Theta, \boldsymbol{d}) - \log Q(\boldsymbol{m}) - 1 - \log Q(\Theta)) \, Q(\Theta) d\Theta + \gamma_1 \qquad \text{B1}$$

Setting the above expression for the first derivative of $\mathrm{L}(\boldsymbol{m}, \Theta, \gamma_1, \gamma_2)$ with respect to $Q(\boldsymbol{m})$ equal to zero gives

$$-\int Q(\Theta) \log \mathcal{P}(\boldsymbol{m}, \Theta, \boldsymbol{d}) \, d\Theta + \log Q(\boldsymbol{m}) + 1 + \int Q(\Theta) \log Q(\Theta) \, d\Theta + \gamma_1 = 0$$

$$\log Q(\boldsymbol{m}) = \int Q(\Theta) \log \mathcal{P}(\boldsymbol{m}, \Theta, \boldsymbol{d}) \, d\Theta - \log Z_{\boldsymbol{m}}(\Theta, \gamma_1)$$

$$= \mathbb{E}_{Q(\Theta)}[\log \mathcal{P}(\boldsymbol{m}, \Theta, \boldsymbol{d})] - \log Z_m(\Theta, \gamma_1) \qquad \text{B2}$$

where $\log \mathrm{Z}_m(\Theta, \gamma_1) \equiv -\int Q(\Theta) \log Q(\Theta) \, \mathrm{d}\Theta - \gamma_1 - 1$ is a normalization constant independent of $\boldsymbol{m}$, and $\mathbb{E}_{Q(\Theta)}[\cdot]$ represents expectation with respect to $Q(\Theta)$. Exponentiating equation B2 proves equation 7.32. Similar derivations of these equations may be found in e.g. Koller & Friedman, 2009 and Jin & Zou, 2010.

## Appendix C: Mathematical Derivation of Equations 7.44 to 7.46

For analytical derivation of the MF update equations of $Q(\boldsymbol{m})$ in a closed form, we re-write equation 7.43 by treating all of the terms independent of $\boldsymbol{m}$ as a constant $k_{\backslash \boldsymbol{m}}$ as

$$\log \mathcal{P}\left(\boldsymbol{m}, \boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m, \boldsymbol{g}, \boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g, \boldsymbol{\Lambda}_\epsilon, \boldsymbol{d}\right)$$

$$= -\frac{1}{2}(\boldsymbol{d}-\boldsymbol{G}\boldsymbol{m})^T \boldsymbol{\Lambda}_\epsilon (\boldsymbol{d}-\boldsymbol{G}\boldsymbol{m}) - \frac{1}{2}(\boldsymbol{m}-\boldsymbol{\mu}_m)^T \boldsymbol{\Lambda}_m (\boldsymbol{m}-\boldsymbol{\mu}_m) + k1_{\backslash \boldsymbol{m}}$$

$$= -\frac{1}{2}(\boldsymbol{d}^T \boldsymbol{\Lambda}_\epsilon \boldsymbol{d} - \boldsymbol{d}^T \boldsymbol{\Lambda}_\epsilon \boldsymbol{G}\boldsymbol{m} - \boldsymbol{m}^T \boldsymbol{G}^T \boldsymbol{\Lambda}_\epsilon \boldsymbol{d} + \boldsymbol{m}^T \boldsymbol{G}^T \boldsymbol{\Lambda}_\epsilon \boldsymbol{G}\boldsymbol{m})$$

$$-\frac{1}{2}(\boldsymbol{m}^T \boldsymbol{\Lambda}_m \boldsymbol{m} - \boldsymbol{m}^T \boldsymbol{\Lambda}_m \boldsymbol{\mu}_m - {\boldsymbol{\mu}_m}^T \boldsymbol{\Lambda}_m \boldsymbol{m} + {\boldsymbol{\mu}_m}^T \boldsymbol{\Lambda}_m \boldsymbol{\mu}_m) + k1_{\backslash \boldsymbol{m}}$$

$$= -\frac{1}{2}(\boldsymbol{d}^T \boldsymbol{\Lambda}_\epsilon \boldsymbol{d} - 2\boldsymbol{m}^T \boldsymbol{G}^T \boldsymbol{\Lambda}_\epsilon \boldsymbol{d} + \boldsymbol{m}^T \boldsymbol{G}^T \boldsymbol{\Lambda}_\epsilon \boldsymbol{G}\boldsymbol{m})$$

$$-\frac{1}{2}(\boldsymbol{m}^T \boldsymbol{\Lambda}_m \boldsymbol{m} - 2\boldsymbol{m}^T \boldsymbol{\Lambda}_m \boldsymbol{\mu}_m + {\boldsymbol{\mu}_m}^T \boldsymbol{\Lambda}_m \boldsymbol{\mu}_m) + k1_{\backslash \boldsymbol{m}}$$

as $\boldsymbol{m}^T \boldsymbol{G}^T \boldsymbol{\Lambda}_\epsilon \boldsymbol{d}$ is a scalar value and $\boldsymbol{\Lambda}_\epsilon$ is symmetric

$$= -\frac{1}{2}\{\boldsymbol{m}^T (\boldsymbol{G}^T \boldsymbol{\Lambda}_\epsilon \boldsymbol{G} + \boldsymbol{\Lambda}_m)\boldsymbol{m} - 2\boldsymbol{m}^T (\boldsymbol{G}^T \boldsymbol{\Lambda}_\epsilon \boldsymbol{d} + \boldsymbol{\Lambda}_m \boldsymbol{\mu}_m)\} + k_{\backslash \boldsymbol{m}} \quad \text{C1}$$

Substituting the above expression for $\log \mathcal{P}\left(\boldsymbol{m}, \boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m, \boldsymbol{g}, \boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g, \boldsymbol{\Lambda}_\epsilon, \boldsymbol{d}\right)$ into equation 7.36 gives

$$\log Q(\boldsymbol{m}) = \mathbb{E}_{Q(\Theta)}[\log \mathcal{P}(\boldsymbol{m}, \Theta, \boldsymbol{d})] + \text{constant}$$

$$= \mathbb{E}_{Q(\Theta_{\backslash m})}\left[-\frac{1}{2}\{\boldsymbol{m}^T (\boldsymbol{G}^T \boldsymbol{\Lambda}_\epsilon \boldsymbol{G} + \boldsymbol{\Lambda}_m)\boldsymbol{m} - 2\boldsymbol{m}^T (\boldsymbol{G}^T \boldsymbol{\Lambda}_\epsilon \boldsymbol{d} + \boldsymbol{\Lambda}_m \boldsymbol{\mu}_m)\}\right] + {k'}_{\backslash \boldsymbol{m}}$$

where ${k'}_{\backslash \boldsymbol{m}}$ is another set of terms independent of $\boldsymbol{m}$.

$$= -\frac{1}{2}\{\boldsymbol{m}^T\left(\widehat{\boldsymbol{G}}^T\widehat{\boldsymbol{\Lambda}}_\epsilon\widehat{\boldsymbol{G}} + \widehat{\boldsymbol{\Lambda}}_m\right)\boldsymbol{m} - 2\boldsymbol{m}^T\left(\widehat{\boldsymbol{G}}^T\widehat{\boldsymbol{\Lambda}}_\epsilon\boldsymbol{d} + \widehat{\boldsymbol{\Lambda}}_m\widehat{\boldsymbol{\mu}}_m\right)\} + k'_{\backslash \boldsymbol{m}} \quad \text{C2}$$

where $\widehat{\boldsymbol{x}}$ represents the (current estimate of the) expected value of $\boldsymbol{x}$. The last expression follows from the linearity of the expectation operator. This is the canonical form of a Normal distribution with mean and precision matrix as given in equations 7.45 and 7.46.

## Appendix D: Mathematical Derivation of Equations 7.50 to 7.54

For analytical derivation of the MF update equations of $\mathcal{Q}(\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m)$ in a closed form, we re-write equation 7.43 by treating all of the terms independent of $\boldsymbol{\mu}_m$ and $\boldsymbol{\Lambda}_m$ as a constant $k_{\backslash\boldsymbol{\mu}_m,\backslash\boldsymbol{\Lambda}_m}$ as

$$\log \mathcal{P}\left(\boldsymbol{m}, \boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m, \boldsymbol{g}, \boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g, \boldsymbol{\Lambda}_\epsilon, \boldsymbol{d}\right)$$

$$= \frac{1}{2}\Big[ (v_m - m + 1)\log|\boldsymbol{\Lambda}_m| - (\boldsymbol{m} - \boldsymbol{\mu}_m)^T \boldsymbol{\Lambda}_m (\boldsymbol{m} - \boldsymbol{\mu}_m)$$

$$-\tau_m (\boldsymbol{\mu}_m - \boldsymbol{\pi}_m)^T \boldsymbol{\Lambda}_m (\boldsymbol{\mu}_m - \boldsymbol{\pi}_m) - Tr\left[{\boldsymbol{W}_m}^{-1} \boldsymbol{\Lambda}_m\right] \Big] + k_{\backslash\boldsymbol{\mu}_m,\backslash\boldsymbol{\Lambda}_m}$$

$$= \frac{1}{2}\Big[ (v_m - m + 1)\log|\boldsymbol{\Lambda}_m| - (\boldsymbol{m}^T \boldsymbol{\Lambda}_m \boldsymbol{m} - \boldsymbol{m}^T \boldsymbol{\Lambda}_m \boldsymbol{\mu}_m - {\boldsymbol{\mu}_m}^T \boldsymbol{\Lambda}_m \boldsymbol{m} + {\boldsymbol{\mu}_m}^T \boldsymbol{\Lambda}_m \boldsymbol{\mu}_m)$$

$$-\tau_m ({\boldsymbol{\mu}_m}^T \boldsymbol{\Lambda}_m \boldsymbol{\mu}_m - {\boldsymbol{\mu}_m}^T \boldsymbol{\Lambda}_m \boldsymbol{\pi}_m - {\boldsymbol{\pi}_m}^T \boldsymbol{\Lambda}_m \boldsymbol{\mu}_m + {\boldsymbol{\pi}_m}^T \boldsymbol{\Lambda}_m \boldsymbol{\pi}_m)$$

$$- Tr\left[{\boldsymbol{W}_m}^{-1} \boldsymbol{\Lambda}_m\right] \Big] + k_{\backslash\boldsymbol{\mu}_m,\backslash\boldsymbol{\Lambda}_m}$$

$$= \frac{1}{2}\Big[ (v_m - m + 1)\log|\boldsymbol{\Lambda}_m| - (1 + \tau_m){\boldsymbol{\mu}_m}^T \boldsymbol{\Lambda}_m \boldsymbol{\mu}_m + {\boldsymbol{\mu}_m}^T \boldsymbol{\Lambda}_m (\boldsymbol{m} + \tau_m \boldsymbol{\pi}_m)$$

$$+ (\boldsymbol{m}^T + \tau_m {\boldsymbol{\pi}_m}^T) \boldsymbol{\Lambda}_m \boldsymbol{\mu}_m - \boldsymbol{m}^T \boldsymbol{\Lambda}_m \boldsymbol{m} - \tau_m {\boldsymbol{\pi}_m}^T \boldsymbol{\Lambda}_m \boldsymbol{\pi}_m$$

$$- Tr\left[{\boldsymbol{W}_m}^{-1} \boldsymbol{\Lambda}_m\right] \Big] + k_{\backslash\boldsymbol{\mu}_m,\backslash\boldsymbol{\Lambda}_m}$$

$$= \frac{1}{2}\Big[ (v_m - m + 1)\log|\boldsymbol{\Lambda}_m| - (1 + \tau_m){\boldsymbol{\mu}_m}^T \boldsymbol{\Lambda}_m \boldsymbol{\mu}_m + {\boldsymbol{\mu}_m}^T \boldsymbol{\Lambda}_m (\boldsymbol{m} + \tau_m \boldsymbol{\pi}_m)$$

$$+(\boldsymbol{m}+\tau_m\boldsymbol{\pi}_m)^T\boldsymbol{\Lambda}_m\boldsymbol{\mu}_m-\boldsymbol{m}^T\boldsymbol{\Lambda}_m\boldsymbol{m}-\tau_m{\boldsymbol{\pi}_m}^T\boldsymbol{\Lambda}_m\boldsymbol{\pi}_m-Tr\left[{\boldsymbol{W}_m}^{-1}\boldsymbol{\Lambda}_m\right]$$

$$-\frac{1}{(1+\tau_m)}(\boldsymbol{m}+\tau_m\boldsymbol{\pi}_m)^T\boldsymbol{\Lambda}_m(\boldsymbol{m}+\tau_m\boldsymbol{\pi}_m)$$

$$+\frac{1}{(1+\tau_m)}(\boldsymbol{m}+\tau_m\boldsymbol{\pi}_m)^T\boldsymbol{\Lambda}_m(\boldsymbol{m}+\tau_m\boldsymbol{\pi}_m)\ \Big]+k_{\backslash\boldsymbol{\mu}_m,\backslash\boldsymbol{\Lambda}_m}$$

Re-arranging terms:

$$=\frac{1}{2}\Big[\ (\upsilon_m-m+1)\log|\boldsymbol{\Lambda}_m|-\Big\{\ (1+\tau_m){\boldsymbol{\mu}_m}^T\boldsymbol{\Lambda}_m\boldsymbol{\mu}_m-\ {\boldsymbol{\mu}_m}^T\boldsymbol{\Lambda}_m(\boldsymbol{m}+\tau_m\boldsymbol{\pi}_m)$$

$$-(\boldsymbol{m}+\tau_m\boldsymbol{\pi}_m)^T\boldsymbol{\Lambda}_m\boldsymbol{\mu}_m+\frac{1}{(1+\tau_m)}(\boldsymbol{m}+\tau_m\boldsymbol{\pi}_m)^T\boldsymbol{\Lambda}_m(\boldsymbol{m}+\tau_m\boldsymbol{\pi}_m)\ \Big\}$$

$$-\Big\{\ \boldsymbol{m}^T\boldsymbol{\Lambda}_m\boldsymbol{m}+\tau_m{\boldsymbol{\pi}_m}^T\boldsymbol{\Lambda}_m\boldsymbol{\pi}_m+Tr\left[{\boldsymbol{W}_m}^{-1}\boldsymbol{\Lambda}_m\right]$$

$$-\frac{1}{(1+\tau_m)}(\boldsymbol{m}+\tau_m\boldsymbol{\pi}_m)^T\boldsymbol{\Lambda}_m(\boldsymbol{m}+\tau_m\boldsymbol{\pi}_m)\ \Big\}\ \Big]+k_{\backslash\boldsymbol{\mu}_m,\backslash\boldsymbol{\Lambda}_m}$$

$$=\frac{1}{2}\Big[\ (\upsilon_m-m+1)\log|\boldsymbol{\Lambda}_m|$$

$$-\Big\{\ (1+\tau_m)\left(\boldsymbol{\mu}_m-\frac{\boldsymbol{m}+\tau_m\boldsymbol{\pi}_m}{1+\tau_m}\right)^T\boldsymbol{\Lambda}_m\left(\boldsymbol{\mu}_m-\frac{\boldsymbol{m}+\tau_m\boldsymbol{\pi}_m}{1+\tau_m}\right)\Big\}$$

$$-\Big\{\ \boldsymbol{m}^T\boldsymbol{\Lambda}_m\boldsymbol{m}+\tau_m{\boldsymbol{\pi}_m}^T\boldsymbol{\Lambda}_m\boldsymbol{\pi}_m+Tr\left[{\boldsymbol{W}_m}^{-1}\boldsymbol{\Lambda}_m\right]$$

$$-\frac{1}{(1+\tau_m)}(\ \boldsymbol{m}^T\boldsymbol{\Lambda}_m\boldsymbol{m}+\tau_m\boldsymbol{m}^T\boldsymbol{\Lambda}_m\boldsymbol{\pi}_m+\tau_m{\boldsymbol{\pi}_m}^T\boldsymbol{\Lambda}_m\boldsymbol{m}$$

$$+{\tau_m}^2{\boldsymbol{\pi}_m}^T\boldsymbol{\Lambda}_m\boldsymbol{\pi}_m\ )\ \Big\}\ \Big]+k_{\backslash\boldsymbol{\mu}_m,\backslash\boldsymbol{\Lambda}_m}$$

$$= \frac{1}{2}\Big[ (v_m - m + 1)\log|\boldsymbol{\Lambda}_m|$$

$$-\left\{ (1+\tau_m)\left(\boldsymbol{\mu}_m - \frac{\boldsymbol{m}+\tau_m\boldsymbol{\pi}_m}{1+\tau_m}\right)^T \boldsymbol{\Lambda}_m \left(\boldsymbol{\mu}_m - \frac{\boldsymbol{m}+\tau_m\boldsymbol{\pi}_m}{1+\tau_m}\right)\right\}$$

$$-\Big\{ Tr\left[\boldsymbol{W}_m{}^{-1}\boldsymbol{\Lambda}_m\right] + \frac{1}{(1+\tau_m)}\big( (1+\tau_m)\boldsymbol{m}^T\boldsymbol{\Lambda}_m\boldsymbol{m}$$

$$+(1+\tau_m)\tau_m\boldsymbol{\pi}_m{}^T\boldsymbol{\Lambda}_m\boldsymbol{\pi}_m - \boldsymbol{m}^T\boldsymbol{\Lambda}_m\boldsymbol{m} - \tau_m\boldsymbol{m}^T\boldsymbol{\Lambda}_m\boldsymbol{\pi}_m$$

$$-\tau_m\boldsymbol{\pi}_m{}^T\boldsymbol{\Lambda}_m\boldsymbol{m} - \tau_m{}^2\boldsymbol{\pi}_m{}^T\boldsymbol{\Lambda}_m\boldsymbol{\pi}_m \big)\Big\}\Big] + k_{\backslash\boldsymbol{\mu}_m,\backslash\boldsymbol{\Lambda}_m}$$

Again expanding the terms that may be cancelled out, we get

$$= \frac{1}{2}\Big[ (v_m - m + 1)\log|\boldsymbol{\Lambda}_m|$$

$$-\left\{ (1+\tau_m)\left(\boldsymbol{\mu}_m - \frac{\boldsymbol{m}+\tau_m\boldsymbol{\pi}_m}{1+\tau_m}\right)^T \boldsymbol{\Lambda}_m \left(\boldsymbol{\mu}_m - \frac{\boldsymbol{m}+\tau_m\boldsymbol{\pi}_m}{1+\tau_m}\right)\right\}$$

$$-\Big\{ Tr\left[\boldsymbol{W}_m{}^{-1}\boldsymbol{\Lambda}_m\right] + \frac{1}{(1+\tau_m)}\big( \boldsymbol{m}^T\boldsymbol{\Lambda}_m\boldsymbol{m} + \tau_m\boldsymbol{m}^T\boldsymbol{\Lambda}_m\boldsymbol{m}$$

$$+\tau_m\boldsymbol{\pi}_m{}^T\boldsymbol{\Lambda}_m\boldsymbol{\pi}_m + \tau_m{}^2\boldsymbol{\pi}_m{}^T\boldsymbol{\Lambda}_m\boldsymbol{\pi}_m - \boldsymbol{m}^T\boldsymbol{\Lambda}_m\boldsymbol{m} - \tau_m\boldsymbol{m}^T\boldsymbol{\Lambda}_m\boldsymbol{\pi}_m$$

$$-\tau_m\boldsymbol{\pi}_m{}^T\boldsymbol{\Lambda}_m\boldsymbol{m} - \tau_m{}^2\boldsymbol{\pi}_m{}^T\boldsymbol{\Lambda}_m\boldsymbol{\pi}_m \big)\Big\}\Big] + k_{\backslash\boldsymbol{\mu}_m,\backslash\boldsymbol{\Lambda}_m}$$

$$= \frac{1}{2}\Big[ (v_m - m + 1)\log|\boldsymbol{\Lambda}_m|$$

$$-\left\{ (1+\tau_m)\left(\boldsymbol{\mu}_m - \frac{\boldsymbol{m}+\tau_m\boldsymbol{\pi}_m}{1+\tau_m}\right)^T \boldsymbol{\Lambda}_m \left(\boldsymbol{\mu}_m - \frac{\boldsymbol{m}+\tau_m\boldsymbol{\pi}_m}{1+\tau_m}\right) \right\}$$

$$- \left\{ Tr\left[\boldsymbol{W}_m{}^{-1}\boldsymbol{\Lambda}_m\right] + \frac{\tau_m}{(1+\tau_m)} \left( \boldsymbol{m}^T\boldsymbol{\Lambda}_m\boldsymbol{m} + \boldsymbol{\pi}_m{}^T\boldsymbol{\Lambda}_m\boldsymbol{\pi}_m \right.\right.$$

$$\left.\left. - \boldsymbol{m}^T\boldsymbol{\Lambda}_m\boldsymbol{\pi}_m - \boldsymbol{\pi}_m{}^T\boldsymbol{\Lambda}_m\boldsymbol{m} \right) \right\} \Bigg] + k_{\backslash\boldsymbol{\mu}_m,\backslash\boldsymbol{\Lambda}_m}$$

$$= \frac{1}{2}\Bigg[ (v_m - m + 1)\log|\boldsymbol{\Lambda}_m|$$

$$-\left\{ (1+\tau_m)\left(\boldsymbol{\mu}_m - \frac{\boldsymbol{m}+\tau_m\boldsymbol{\pi}_m}{1+\tau_m}\right)^T \boldsymbol{\Lambda}_m \left(\boldsymbol{\mu}_m - \frac{\boldsymbol{m}+\tau_m\boldsymbol{\pi}_m}{1+\tau_m}\right) \right\}$$

$$- \left\{ Tr\left[\boldsymbol{W}_m{}^{-1}\boldsymbol{\Lambda}_m\right] + \frac{\tau_m}{(1+\tau_m)} (\boldsymbol{m}-\boldsymbol{\pi}_m)^T\boldsymbol{\Lambda}_m(\boldsymbol{m}-\boldsymbol{\pi}_m) \right\} \Bigg] + k_{\backslash\boldsymbol{\mu}_m,\backslash\boldsymbol{\Lambda}_m}$$

$$= \frac{1}{2}\Bigg[ (v_m - m + 1)\log|\boldsymbol{\Lambda}_m|$$

$$-\left\{ (1+\tau_m)\left(\boldsymbol{\mu}_m - \frac{\boldsymbol{m}+\tau_m\boldsymbol{\pi}_m}{1+\tau_m}\right)^T \boldsymbol{\Lambda}_m \left(\boldsymbol{\mu}_m - \frac{\boldsymbol{m}+\tau_m\boldsymbol{\pi}_m}{1+\tau_m}\right) \right\}$$

$$- \left\{ Tr\left[\boldsymbol{W}_m{}^{-1}\boldsymbol{\Lambda}_m\right] + Tr\left[\frac{\tau_m}{(1+\tau_m)} (\boldsymbol{m}-\boldsymbol{\pi}_m)^T\boldsymbol{\Lambda}_m(\boldsymbol{m}-\boldsymbol{\pi}_m)\right] \right\} \Bigg] + k_{\backslash\boldsymbol{\mu}_m,\backslash\boldsymbol{\Lambda}_m}$$

Now using the 'linearity' and '*invariance under cyclic permutation*' property of the trace operator, i.e. $\mathrm{Tr}[ABC] = \mathrm{Tr}[CAB]$:

$$= \frac{1}{2}\Bigg[ (v_m - m + 1)\log|\boldsymbol{\Lambda}_m|$$

$$-\left\{ (1+\tau_m)\left(\boldsymbol{\mu}_m - \frac{\boldsymbol{m}+\tau_m\boldsymbol{\pi}_m}{1+\tau_m}\right)^T \boldsymbol{\Lambda}_m \left(\boldsymbol{\mu}_m - \frac{\boldsymbol{m}+\tau_m\boldsymbol{\pi}_m}{1+\tau_m}\right)\right\}$$

$$-\left\{ Tr\left[\boldsymbol{W}_m{}^{-1}\boldsymbol{\Lambda}_m\right] + Tr\left[\frac{\tau_m}{(1+\tau_m)}(\boldsymbol{m}-\boldsymbol{\pi}_m)(\boldsymbol{m}-\boldsymbol{\pi}_m)^T\boldsymbol{\Lambda}_m\right]\right\}\Bigg] + k_{\backslash\boldsymbol{\mu}_m,\backslash\boldsymbol{\Lambda}_m}$$

$$= \frac{1}{2}\Bigg[ (v_m - m + 1)\log|\boldsymbol{\Lambda}_m|$$

$$-\left\{ (1+\tau_m)\left(\boldsymbol{\mu}_m - \frac{\boldsymbol{m}+\tau_m\boldsymbol{\pi}_m}{1+\tau_m}\right)^T \boldsymbol{\Lambda}_m \left(\boldsymbol{\mu}_m - \frac{\boldsymbol{m}+\tau_m\boldsymbol{\pi}_m}{1+\tau_m}\right)\right\}$$

$$-\left\{ Tr\left[\left(\boldsymbol{W}_m{}^{-1} + \frac{\tau_m}{(1+\tau_m)}(\boldsymbol{m}-\boldsymbol{\pi}_m)(\boldsymbol{m}-\boldsymbol{\pi}_m)^T\right)\boldsymbol{\Lambda}_m\right]\right\}\Bigg] + k_{\backslash\boldsymbol{\mu}_m,\backslash\boldsymbol{\Lambda}_m} \qquad \text{D1}$$

Substituting the expression D1 in equation 7.37 shows that it is a Normal-Wishart distribution with parameters as given in equations 7.51 to 7.54.

## Appendix E: Mathematical Derivation of Equations 7.60 to 7.62

For analytical derivation of the MF update equations of $\mathcal{Q}(\boldsymbol{\Lambda}_\epsilon)$ in a closed form, we re-write equation 7.43 by treating all of the terms independent of $\boldsymbol{\Lambda}_\epsilon$ as a constant $k_{\setminus \boldsymbol{\Lambda}_\epsilon}$ as

$$\log \mathcal{P}\left(\boldsymbol{m}, \boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m, \boldsymbol{g}, \boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g, \boldsymbol{\Lambda}_\epsilon, \boldsymbol{d}\right)$$

$$= \frac{1}{2}\log|\boldsymbol{\Lambda}_\epsilon| - \frac{1}{2}(\boldsymbol{d} - \boldsymbol{G}\boldsymbol{m})^T \boldsymbol{\Lambda}_\epsilon (\boldsymbol{d} - \boldsymbol{G}\boldsymbol{m}) + \sum_{i=1}^{n_a}\{(a_i - 1)\log\lambda_i - b_i\lambda_i\} + k_{\setminus \boldsymbol{\Lambda}_\epsilon} \quad \text{E1}$$

Since $\boldsymbol{\Lambda}_\epsilon$ is a block diagonal matrix

$$\boldsymbol{\Lambda}_\epsilon = \begin{bmatrix} \boldsymbol{\lambda}_1 & \cdots & \boldsymbol{0} \\ \vdots & \ddots & \vdots \\ \boldsymbol{0} & \cdots & \boldsymbol{\lambda}_{n_a} \end{bmatrix} \quad \text{E2}$$

with diagonal matrices $\boldsymbol{\lambda}_i = \lambda_i \boldsymbol{I}_{n_s \times n_s} : i = 1 \ldots n_a$, its determinant $|\boldsymbol{\Lambda}_\epsilon|$ is given by the product of determinants of $\boldsymbol{\lambda}_i$, i.e. $|\boldsymbol{\Lambda}_\epsilon| = \prod_{i=1}^{n_a}|\boldsymbol{\lambda}_i|$. Similarly, since each of $\boldsymbol{\lambda}_i$ is itself a $n_s \times n_s$ scalar matrix with all diagonal elements equal to $\lambda_i$, we have $|\boldsymbol{\lambda}_i| = (\lambda_i)^{n_s}$. This gives $\log|\boldsymbol{\Lambda}_\epsilon| = \log\prod_{i=1}^{n_a}|\boldsymbol{\lambda}_i| = \sum_{i=1}^{n_a}(\log|\boldsymbol{\lambda}_i|) = \sum_{i=1}^{n_a}(\log(\lambda_i)^{n_s}) = \sum_{i=1}^{n_a}(n_s \log\lambda_i) = n_s \sum_{i=1}^{n_a}\log\lambda_i$ .

$$\log|\boldsymbol{\Lambda}_\epsilon| = \log\prod_{i=1}^{n_a}|\boldsymbol{\lambda}_i| = \sum_{i=1}^{n_a}(\log|\boldsymbol{\lambda}_i|) = \sum_{i=1}^{n_a}(\log(\lambda_i)^{n_s}) = n_s\sum_{i=1}^{n_a}\log\lambda_i \quad \text{E3}$$

Also, using the identity $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} = \sum_i A_{ii}x_i^2$, for a diagonal matrix $\boldsymbol{A}$ and vector $\boldsymbol{x}$:

$$(\boldsymbol{d} - \boldsymbol{G}\boldsymbol{m})^T \boldsymbol{\Lambda}_\epsilon (\boldsymbol{d} - \boldsymbol{G}\boldsymbol{m}) = \sum_{i=1}^{n_a}\|\boldsymbol{d}_i - \boldsymbol{G}_i\boldsymbol{m}\|^2\lambda_i \quad \text{E4}$$

Substituting equations E3 and E4 in equation E1, we get

$$\log \mathcal{P}(\boldsymbol{m}, \boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m, \boldsymbol{\Lambda}_\epsilon, \boldsymbol{d})$$

$$= \frac{n_s}{2}\sum_{i=1}^{n_a}\log\lambda_i - \frac{1}{2}\sum_{i=1}^{n_a}\|\boldsymbol{d}_i - \boldsymbol{G}_i\boldsymbol{m}\|^2\lambda_i + \sum_{i=1}^{n_a}(a_i-1)\log\lambda_i - \sum_{i=1}^{n_a} b_i\lambda_i + k_{\backslash\boldsymbol{\Lambda}_\epsilon}$$

$$= \sum_{i=1}^{n_a}\left\{\left(a_i - 1 + \frac{n_s}{2}\right)\log\lambda_i - \frac{1}{2}\|\boldsymbol{d}_i - \boldsymbol{G}_i\boldsymbol{m}\|^2\lambda_i - b_i\lambda_i\right\} + k_{\backslash\boldsymbol{\Lambda}_\epsilon} \quad \text{E5}$$

Substituting equation E5 in equation 7.40 we get:

$$\log\mathcal{Q}(\boldsymbol{\Lambda}_\epsilon) \propto \mathbb{E}_{\mathcal{Q}(\boldsymbol{m},\boldsymbol{\mu}_m,\boldsymbol{\Lambda})}[\log\mathcal{P}(\boldsymbol{m},\Theta,\boldsymbol{d})]$$

$$= \mathbb{E}_{\mathcal{Q}(\boldsymbol{m},\boldsymbol{\mu}_m,\boldsymbol{\Lambda}_m)}\left[\sum_{i=1}^{n_a}\left\{\left(a_i - 1 + \frac{n_s}{2}\right)\log\lambda_i - \frac{1}{2}\|\boldsymbol{d}_i - \boldsymbol{G}_i\boldsymbol{m}\|^2\lambda_i - b_i\lambda_i\right\}\right] + k'_{\backslash\boldsymbol{\Lambda}_\epsilon}$$

Where $k'_{\backslash\boldsymbol{\Lambda}_\epsilon}$ is another set of terms independent of $\boldsymbol{\Lambda}_\epsilon$.

$$= \sum_{i=1}^{n_a}\left\{\left(a_i + \frac{n_s}{2} - 1\right)\log\lambda_i - \frac{1}{2}\,\mathbb{E}_{\mathcal{Q}(\boldsymbol{m},\boldsymbol{\mu}_m,\boldsymbol{\Lambda}_m)}[\|\boldsymbol{d}_i - \boldsymbol{G}_i\boldsymbol{m}\|^2]\lambda_i - b_i\lambda_i\right\} + k'_{\backslash\boldsymbol{\Lambda}_\epsilon} \quad \text{E6}$$

Now using the definition of variance: $\mathbb{E}[\|X\|^2] = \mathrm{var}(X) + \|\mathbb{E}[X]\|^2$, we have

$$\mathbb{E}_{\mathcal{Q}(\boldsymbol{m},\boldsymbol{\mu}_m,\boldsymbol{\Lambda}_m)}[\|\boldsymbol{d}_i - \boldsymbol{G}_i\boldsymbol{m}\|^2] = \mathrm{var}\left(\boldsymbol{d}_i - \widehat{\boldsymbol{G}}_i\widehat{\boldsymbol{m}}\right) + \left\|\mathbb{E}_{\mathcal{Q}(\boldsymbol{m})}\left[\boldsymbol{d}_i - \widehat{\boldsymbol{G}}_i\widehat{\boldsymbol{m}}\right]\right\|^2$$

$$= \frac{1}{\lambda_i} + \left\|\boldsymbol{d}_i - \widehat{\boldsymbol{G}}_i\widehat{\boldsymbol{m}}\right\|^2$$

$$= \frac{1}{\lambda_i} + \hat{S} \quad \text{E7}$$

where

$$\hat{S}_i = \left\|\boldsymbol{d}_i - \widehat{\boldsymbol{G}}_i\widehat{\boldsymbol{m}}\right\|^2$$

$$= \|\boldsymbol{d}_i\|^2 + \left\|\widehat{\boldsymbol{G}}_i\widehat{\boldsymbol{m}}\right\|^2 - 2\,\boldsymbol{d}_i{}^T\widehat{\boldsymbol{G}}_i\widehat{\boldsymbol{m}} \quad \text{E8}$$

where $\widehat{\boldsymbol{m}}$ represents the current estimate of expected value of $\boldsymbol{m}$ at any iteration during the mean field updates. Note that $\hat{S} \geq 0$ and as inversion proceeds and the approximation $\widehat{\boldsymbol{G}}_i\widehat{\boldsymbol{m}}$ of

$\boldsymbol{d}_i$ improves, $\hat{S} \to 0$. Substituting for $\mathbb{E}_{\mathcal{Q}(\boldsymbol{m},\boldsymbol{\mu}_m,\boldsymbol{\Lambda}_m)}[\|\boldsymbol{d}_i - \boldsymbol{G}_i\boldsymbol{m}\|^2]$ from equation E7 in equation E6, we get:

$$\log \mathcal{Q}(\boldsymbol{\Lambda}_{\boldsymbol{\epsilon}}) \propto \mathbb{E}_{\mathcal{Q}(\boldsymbol{m},\boldsymbol{\mu}_m,\boldsymbol{\Lambda}_m)}[\log \mathcal{P}(\boldsymbol{m}, \Theta, \boldsymbol{d})]$$

$$= \sum_{i=1}^{n_a} \left\{ \left(a_i + \frac{n_s}{2} - 1\right) \log \lambda_i - \frac{1}{2}\left(\frac{1}{\lambda_i} + \hat{S}_i\right)\lambda_i - b_i\lambda_i \right\} + k'_{\backslash \boldsymbol{\Lambda}_{\boldsymbol{\epsilon}}}$$

$$= \sum_{i=1}^{n_a} \left\{ \left(a_i + \frac{n_s}{2} - 1\right) \log \lambda_i - \frac{\hat{S}_i\lambda_i}{2} - b_i\lambda_i \right\} + k''_{\backslash \boldsymbol{\Lambda}_{\boldsymbol{\epsilon}}}$$

where $k''_{\backslash \boldsymbol{\Lambda}_{\boldsymbol{\epsilon}}}$ is another set of terms independent of $\boldsymbol{\Lambda}_{\boldsymbol{\epsilon}}$.

$$= \sum_{i=1}^{n_a} \left\{ \left(a_i + \frac{n_s}{2} - 1\right) \log \lambda_i - \left(\frac{\hat{S}_i}{2} + b_i\right)\lambda_i \right\} + k''_{\backslash \boldsymbol{\Lambda}_{\boldsymbol{\epsilon}}} \quad \text{E9}$$

which shows that $\mathcal{Q}(\boldsymbol{\Lambda}_{\boldsymbol{\epsilon}})$ is a product of Gamma distributions with parameters as given in equations 7.61 and 7.62.