THE UNIVERSITY
of EDINBURGH

Bayesian methods for biomarker evaluation and disease diagnosis

Javier Edgardo Garrido Guillén

Doctor of Philosophy

School of Mathematics

University of Edinburgh

United Kingdom

February, 2022

Declaration

I declare that this thesis was composed by myself and that the work contained therein is my own, except where explicitly stated otherwise in the text. The work has not been submitted for any other degree or profession qualification.

Javier Edgardo Garrido Guillén

Edinburgh, UK, 2021

Abstract

Accurate diagnosis of disease is of fundamental importance in medical research and clinical practice. For such reason, the role that diagnostic testing and screening play is undeniable. The major goal of a diagnostic test is to distinguish between individuals with a well-defined condition, referred as diseased, and individuals with the absence of such condition, known as nondiseased. Before a test is widely used in practice, its discriminatory ability must be rigorously assessed through statistical analysis. The overlap coefficient, which is defined as the proportion of overlap area between two density functions, has gained unarguably popularity as a summary measure of diagnostic accuracy and, in this thesis, it is our main object of study.

In the first chapter of this thesis, we introduce different concepts related to diagnostic tests and how its accuracy might be measured. A brief description of the receiver operating characteristic (ROC) curve, one the most popular existing statistical methods to evaluate the discriminatory ability of a test, is provided as well. We then define the coefficient of overlap and we discuss its advantages and disadvantages over usual summary measures, namely, the area under the ROC curve and the Youden index. At the end of the chapter, we recognize that, as it has been acknowledged in the literature, the performance of a diagnostic test may depend on covariates (e.g., age and/or sex) and failure to incorporate this information may result in misleading or oversimplified conclusions about the accuracy of the test.

In the second chapter we provide a brief introduction to Bayesian inference and Bayesian nonparametric methods, as we have adopted the Bayesian paradigm throughout this thesis.

In the third chapter of this thesis, we develop Bayesian inferential methods for the coefficient of

overlap. Accurate estimation of the coefficient of overlap requires accurately estimating the density functions of test outcomes in both the diseased and nondiseased populations. For such end, we employ a Dirichlet process mixture of normal distributions to model such density functions. Once estimates of the density functions of test results have been obtained, two estimators for the coefficient of overlap are then proposed: one based on numerical integration and another one that further uses the Bayesian bootstrap. Our integrated framework relaxes restrictive distributional assumptions (e.g., normality of test outcomes in each population) of existing approaches. The performance of our methods is assessed through a simulation study and we also provide an application concerned with the search for ovarian cancer biomarkers.

In the fourth chapter of this thesis, we extend our flexible modelling approach for the coefficient of overlap to the covariates' context. We follow a joint approach based on Dirichlet process mixtures, where both test outcomes and covariates are modelled jointly through a multivariate kernel. We use different simulated examples to evaluate the performance of our modelling approach and we provide an application to two real datasets as well. The first application concerns the assessment of the accuracy of the glucose levels as a marker for diabetes changes with age. In turn, in the second application the goal is to study the effect of age and sex on the discriminatory ability of different biomarkers for the Alzheimer's disease.

In the fifth chapter, we include vignettes and examples showing the usage of the R package `OverlapCoefficient`, which implements our methods.

Finally, we discuss future working directions, such as possible generalizations of the coefficient of overlap to handle two or more biomarkers.

Lay summary

Accurate disease diagnosis is crucial for clinical decision making and patients' health care and well-being. Medical diagnostic tests are commonly used to determine whether an individual has a particular disease or not. Therefore, it is essential to rigorously evaluate the ability of a test to successfully discriminate whether a condition is or is not present before it can be widely used in general practice. In real life, there are no "perfect" diagnostic tests that completely discriminate between people with and without a disease. Nevertheless, "imperfect", non-invasive and economically accessible diagnostic tests can accurately identify the presence of a specific medical condition and thus, they can be widely used to screen larger populations.

For example, Alzheimer's disease is the most common cause of dementia, which is one of the leading causes of death among elderly people. To date, there is no cure and it can be definitively diagnosed only once the patient has died (brain tissue autopsy). There is vast research aimed at performing an accurate diagnosis on the early stages of Alzheimer's disease based on potential biomarkers (e.g., genes, proteins, etc.). An accurate diagnosis would help to design efficient treatments and to understand the disease mechanism to generate preventive actions or reduce the risk of suffering from it. Statistical analyses are used to validate these biomarkers. Therefore, it is crucial that potential biomarkers undergo a thorough statistical analysis to evaluate their ability to discriminate patients with the disease.

Different statistical methodologies have been proposed to evaluate the discriminatory ability of diagnostic tests. One approach evaluates how similar or different (closer or far apart) are the probability distributions of the test outcomes. Popular methods under this approach are usually based on the *receiver operating characteristic* (ROC) curve (a graphical tool that displays the probability of correctly

classifying an individual and the probability of misclassifying them across all possible thresholds of a continuous test or biomarker). ROC analysis has well-known limitations (e.g., multiple intersections between the distributions may lead to a misrepresentation of the true discriminatory ability of the test) and several alternatives have been proposed to overcome them. However, in many clinical scenarios their implementation is difficult or cumbersome.

Recently, the *coefficient of overlap*, defined as the area enclosed between two probability distributions, has been proposed as an alternative summary measure of diagnostic accuracy. The coefficient of overlap preserves some of the advantages of the usual summary measures (e.g., the area under the ROC curve and the Youden index) while being non-directional, i.e., regardless of the order of the distributions, the value of the coefficient of overlap will remain unchanged. In this thesis, we focus on the development of flexible methods to estimate and conduct inference about the coefficient of overlap.

Biomedical data (e.g., gene expression, glucose levels in blood, etc.) usually presents complex structures that are difficult to capture by common methods, for example, parametric methods which require assumptions about the probability distributions of the test outcomes that are too restrictive. Bayesian nonparametric methods overcome these limitations and have gained special interest due to their versatility for a wide range of applications.

Bayesian nonparametric methods provide a robust and data-driven framework to conduct inference about the coefficient of overlap. Furthermore, in many clinical applications, both information regarding the test outcomes (e.g., biomarkers in the Alzheimer example) and about the individuals being tested (e.g., their age or sex) is available. It is widely known that individual characteristics might have a great effect on the test accuracy. Thus, it is important to consider this information within the inference process. Ignoring this individuals' characteristics information can lead to oversimplified, biased, or misleading conclusions about the true discrimination of a test. Here, we propose the *covariate-specific coefficient of overlap*, an extension of the coefficient of overlap that considers such characteristics, also known as covariates. Therefore, for each value of the covariates (e.g., age, sex, etc.), we might get a different value of the coefficient of overlap. This helps to determine the optimal and suboptimal populations where to perform the tests on.

Acknowledgements

First, I would like to thank my father, Francisco, and my aunts, Graciela and Guadalupe, who without their support I would not have been able to get to where I am.

I also want to thank my supervisor, Dr Vanda Inácio, for her dedication, even before starting this journey. Thank you for your help and commitment throughout this project, which without your support and guidance I would not have been able to complete.

I am indebted to all the Mexican people, thanks to their work and effort I had the opportunity to be awarded a scholarship through the Consejo Nacional de Ciencia y Tecnología (Conacyt) - Beca al Extranjero 2017.

Thanks to my sister, Gaby, for your endless help both academic and personal. Thank you for your advice and your help in correcting my written pieces, including this thesis.

Finally, I would like to dedicate this thesis to my mother, Rosalba, and my wife, Karina. I could write a second thesis just to thank you both for your infinite help, support, trust and much more that I am not even able to describe. Mom, thank you for being the best, for taking care of me and for always being an inspiration to me, I love you. Morcín, thank you for always being by my side and for encouraging me to be a better version of myself day by day. Thanks to you this thesis has beautiful graphics. The credit is yours on the stick-breaking process diagram. I know that throughout these four years, we have lived through very difficult times. However, these have made us grow and be stronger, as individuals and as the wonderful team that we have forged. I love you.

To Rosalba and Karina

Contents

Chapter 1

Introduction

This thesis concerns with the assessment of the discriminatory ability of continuous biomarkers through the so-called coefficient of overlap. Throughout this thesis we will use the terms "biomarker", "marker" and "diagnostic test" (or simply referred as "test") interchangeably, although they do not necessarily have the same meaning always. First, we introduce several concepts related to diagnostic testing and measures of diagnostic accuracy, including the popular receiver operating characteristic (ROC) curve. We continue our discussion with a brief review of existing summary indices based on the ROC curve which are widely used to evaluate a test discriminatory ability. We then present the coefficient of overlap as an alternative to common summary measures of diagnostic accuracy, highlighting its advantages and disadvantages over them. Finally, we introduce the covariate-specific coefficient of overlap, intended to determine the optimal and suboptimal populations where to perform the tests on based on individuals' characteristics.

1.1 Diagnostic testing

A diagnostic test is a medical procedure performed on an individual to identify the presence or absence of a well-defined condition referred, throughout this thesis, as disease. An individual without such condition is known as nondiseased. Diagnostic testing has several purposes, for example, it provides

reliable information about an individual's condition, it may influence health care policies or it can be used to understand disease mechanism (Zhou et al., 2011, p. 3). It is an important tool to monitor or determine the course of treatment as well. In other words, diagnostic tests are used to determine the presence or absence of a disease on symptomatic individuals and, usually, they are expensive and/or invasive (Ruf et al., 2017). In contrast, screening tests are used to detect potential disease indicators in a large number of apparently healthy individuals. They can be cheaper and less invasive than diagnostic tests and, essentially, they only indicate the suspicion of a disease, which needs to be confirmed (Ruf et al., 2017). A test's accuracy is the ability of a test to discriminate among different states of health (Zweig and Campbell, 1993). If the test outcomes do not differ among states, we say that the test has negligible accuracy, making it useless from a diagnostic viewpoint. However, if the test outcomes do not overlap at all for the different health states, then the test has a perfect accuracy (Zhou et al., 2011, p. 4). Intermediate degrees of overlap correspond to different degrees of accuracy.

It is important to note that test outcomes may not be a true representation of an individual's condition. For such reason, we assume that the true status is measured without error using a definitive gold standard test. A gold standard is a source of information completely different from the test under consideration, which tell us about the true condition of the individual (Zhou et al., 2011, p. 4). An ideal test (perfect accuracy) would correctly classify all individuals and would be inexpensive, i.e., a cheap gold standard. However, such tests are rare or non-existent, for example, gold standards may involve the autopsy report or invasive procedures such as a biopsy or a surgery. In diagnostic test accuracy studies, we are interested in how well a test performs compared to the truth. To determine how useful the test under consideration might be, a sample of individuals who have been tested is taken. In this thesis, we assume that we have two samples, one from nondiseased individuals and the other from diseased individuals, who have been classified using a gold standard or any other sort of reference.

1.2 Measures of diagnostic accuracy

Test outcomes can be dichotomous, ordinal or continuous. Dichotomous tests include, among others, home pregnancy tests or rapid lateral flow tests (RFT) for Covid-19. Ordinal scales are often used in medical imaging to perform a diagnosis (e.g., X-rays, magnetic resonance imaging, etc.). Blood pressure or glucose levels in blood are typical examples of continuous markers. In this thesis, we will focus on continuous test outcomes, as they are the most common in practice. Let $Y \in \mathbb{R}$ be a continuous random variable representing the test outcomes. The diagnosis is made by comparing the test result to some threshold or cut-off, denoted by $c \in \mathbb{R}$. Under the assumption that larger test outcomes are more indicative of the presence of the disease, without loss of generality, we can assume that the decision rule is as follows

- if $Y \geq c$, then the individual is classified as diseased and
- if $Y < c$, then the individual is classified as nondiseased.

If such assumption is not met, one can simply reverse the decision rule. Further, let $D \in \{0, 1\}$ be the binary variable representing the true health status of the individual, with $D = 1$ denoting the presence of the condition and $D = 0$ its absence. The accuracy of a test can be measured by its sensitivity and specificity. Let $Y_{\bar{D}}$ and Y_D be two independent continuous random variables representing the test outcomes from the nondiseased and diseased population, with corresponding cumulative distribution functions (cdf) and probability density functions (pdf) given by $F_{\bar{D}}(y) = \mathbb{P}(Y_{\bar{D}} \leq y)$ and $f_{\bar{D}}(y)$; and $F_D(y) = \mathbb{P}(Y_D \leq y)$ and $f_D(y)$, respectively. The sensitivity, Se, also known as the true positive fraction (TPF), is the test's ability to identify the condition when it is actually present, that is,

$$\mathrm{Se}(c) = \mathrm{TPF}(c) = \mathbb{P}\left(Y \geq c \mid D = 1\right) = \mathbb{P}\left(Y_D \geq c\right) = 1 - F_D(c).$$

Note that, for each threshold value, we might get a different TPF. The specificity, Sp, also known as the true negative fraction (TNF), is the test's ability to rule out the condition when it is truly absent, i.e.,

$$\mathrm{Sp}(c) = \mathrm{TNF}(c) = \mathbb{P}\left(Y < c \mid D = 0\right) = \mathbb{P}\left(Y_{\bar{D}} < c\right) = F_{\bar{D}}(c).$$

There are two types of error when measuring a marker, namely,

- a diseased individual may test negative, this is called a false negative or
- a nondiseased individual may test positive, this is known as false positive.

These classification errors are measured by the false negative fraction (FNF) and the false positive fraction (FPF), respectively,

$$\mathrm{FNF}\,(c) = \mathbb{P}\,(Y < c \mid D = 1) = \mathbb{P}\,(Y_D < c) = F_D(c),$$

and

$$\mathrm{FPF}\,(c) = \mathbb{P}\,(Y \geq c \mid D = 0) = \mathbb{P}\,(Y_{\bar{D}} \geq c) = 1 - F_{\bar{D}}(c).$$

Usually, once a diagnostic test has been demonstrated to be accurate enough, a threshold or cut-off to screen subjects in practice must be determined. This threshold value can vary depending on different circumstances, for example, a life threatening disease will require high sensitivity, but high risk procedures to treat a likely non life threatening disease may require a higher specificity.

1.3 ROC analysis

The ROC curve is a widely used graphical tool to assess a test's accuracy. It graphically displays the test's false positive fraction versus the true positive fraction across all possible decision thresholds. Thus, the ROC curve is the set of points

$$\{(\mathrm{FPF}(c), \mathrm{TPF}(c)) \mid c \in \mathbb{R}\} \equiv \{(1 - F_{\bar{D}}(c), 1 - F_D(c)) \mid c \in \mathbb{R}\}\,. \tag{1.1}$$

In Figure 1.1, we can observe that the ROC curve measures the separation between the distribution of the test outcomes in the nondiseased and diseased population. It also illustrates the trade-off between the TPF and the FPF as the cut-off changes. Alternatively, we can rewrite (1.1) as

$$\{(p, \mathrm{ROC}(p)) \mid 0 \leq p \leq 1\}\,,$$

with $\text{ROC}(p) = 1 - F_D\left\{F_{\bar{D}}^{-1}(1-p)\right\}$ and where $F_{\bar{D}}^{-1}(1-p) = \inf\{z \mid F_{\bar{D}}(z) \geq 1-p\}$ stands for the quantile function of the nondiseased group. ROC curves lie in the unit square. The segment line comprised between the vertices $(0,0)$ and $(1,1)$ is called the chance diagonal. An ROC curve falling along this diagonal means that the test classifies no better than chance, thus making it useless from a diagnostic viewpoint. In contrast, the farther the curve from the diagonal is, the better the discriminatory ability of the test is. It is also possible to have an ROC curve below the chance diagonal. However, this can be fixed by simply reversing the decision rule, that is, classify as diseased an individual when $Y < c$, and as nondiseased when $Y \geq c$.

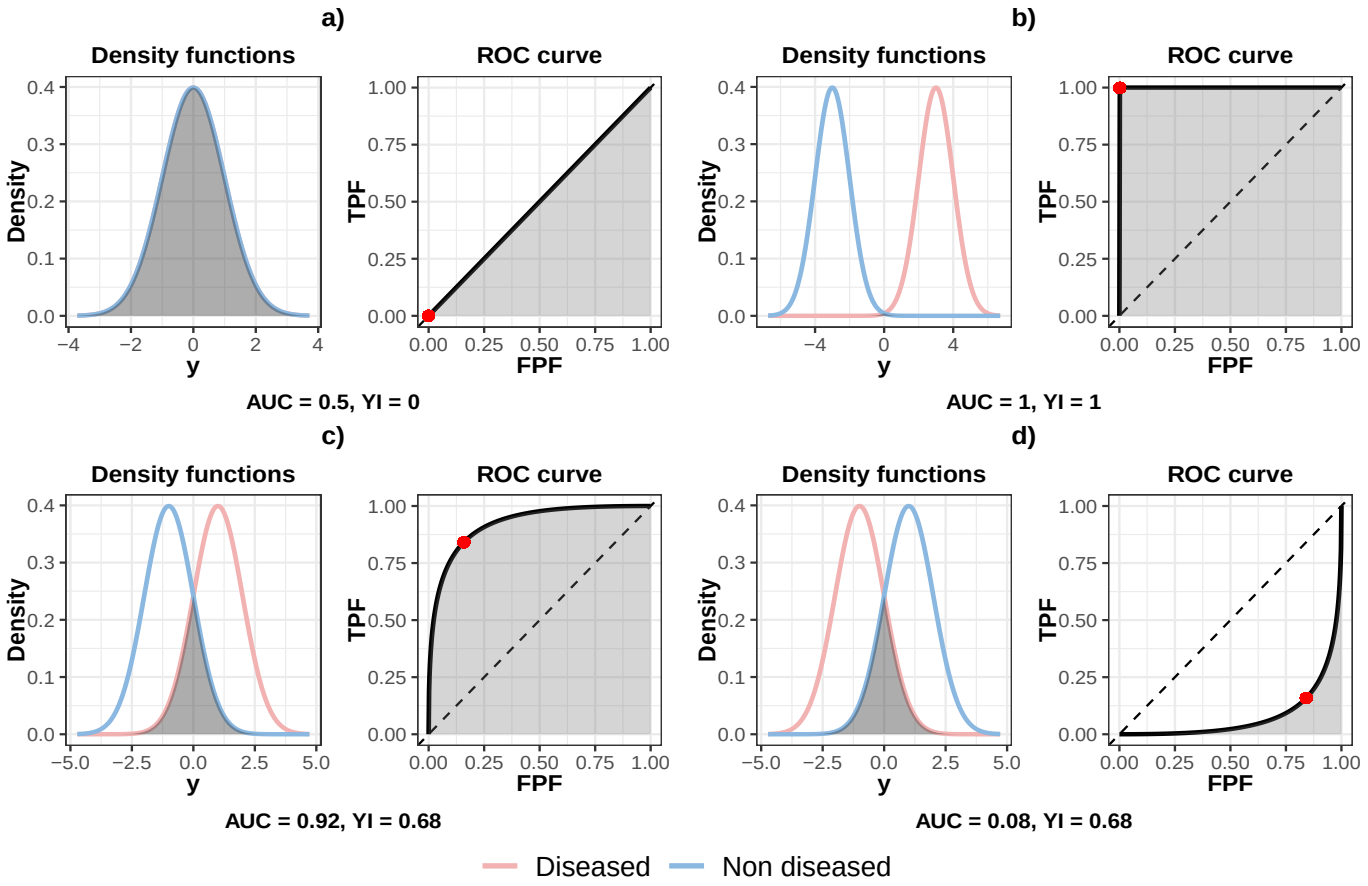

Figure 1.1: Different hypothetical densities of test outcomes in the nondiseased (blue line) and diseased (red line) groups along with their corresponding ROC curve. (a) Uninformative test; (b) perfect test; (c) and (d) intermediate situations. In the ROC curve plots, the dotted line represents the chance diagonal, the shaded area is the AUC and the red point is the YI.

ROC curves can also be used to visually compare the accuracy of different tests. However, it turns

out difficult to determine how better is a test compared to another, because a change in the threshold may affect both tests differently (Turner, 1978). For such reason, it is often useful to summarize the test accuracy by a single index or value. The most popular one is the area under the ROC curve (AUC). It is defined as

$$\text{AUC} = \int_0^1 \text{ROC}(p)dp.$$

A test with AUC = 1 would have a perfect discriminatory ability. Whereas, a test with AUC = 0.5 would have no discrimination power. If AUC < 0.5, then it can be fixed by reversing the decision rule. It is common to interpret the AUC as the probability that a randomly selected diseased individual's test outcome is greater than that of a randomly chosen nondiseased individual (Zhou et al., 2011, p. 28), i.e., $\text{AUC} = \mathbb{P}(Y_D > Y_{\bar{D}})$.

Another popular summary measure used in practice is the Youden index (YI) (Youden, 1950), defined as

$$\text{YI} = \max_{c\in\mathbb{R}} \{\text{TPF}(c) + \text{TNF}(c) - 1\} = \max_{c\in\mathbb{R}} \{F_{\bar{D}}(c) - F_D(c)\} = \max_{0\leq p\leq 1} |\text{ROC}(p) - p|.$$

When the distributions of the test outcomes are completely separated, YI = 1. This means a perfectly accurate test. In contrast, when the distributions completely overlap, YI = 0, thus meaning an useless test. An appealing feature of the YI over the AUC, is that it provides a criterion for choosing the optimal cut-off c^*, that is

$$c^* = \underset{c\in\mathbb{R}}{\text{argmax}} \{F_{\bar{D}}(c) - F_D(c)\}.$$

The YI can be interpreted as the point on the ROC curve that maximizes the distance between the ROC curve and the chance diagonal. This can be clearly observed in Figure 1.1, where the YI is denoted by a red dot over the ROC curve. Similar to the AUC, if the ROC curve lies below the chance diagonal, we might obtain YI < 0. However, again we can fix it reversing the decision rule.

1.4 The coefficient of overlap

Despite their popularity, the AUC and the YI have some limitations. The AUC can only distinguish location differences between the underlying test outcomes distributions. In contrast, the YI does detect differences in both location and shape (Samawi et al., 2017). However, within the ROC setting, the mean of the diseased population is usually assumed to be larger than that of the nondiseased population. This may lead to erroneous values of the AUC and YI if the means of the distributions are swapped, as it occurs in Figure 1.2 panel (10), although the classification rule can always be reversed.

Recently, the coefficient of overlap (OVL), defined as the proportion of overlap area between two density functions, has been proposed as an alternative summary measure of diagnostic accuracy (Samawi et al., 2017, Wang and Tian, 2017). It can be thought as a measure of agreement between two probability distributions (Inman and Bradley Jr, 1989) and it has been successfully applied in different fields such as medicine (Samawi et al., 2017), ecology (Ridout and Linkie, 2009), demography (Clemons and Bradley Jr., 2000), economics (Chan et al., 2017) and genetics (Wang and Tian, 2017, Silva-Fortes et al., 2012).

According to Weitzman (1970), the coefficient of overlap is defined as

$$\mathrm{OVL}\left(Y_{\bar{D}}, Y_D\right) = \int_{-\infty}^{\infty} \min\{f_{\bar{D}}(y), f_D(y)\} dy. \tag{1.2}$$

Schmid and Schmidt (2006) provided another representation of the OVL as a sum of error probabilities, that is,

$$\begin{aligned}
\mathrm{OVL}\left(Y_{\bar{D}}, Y_D\right) &= \int_{-\infty}^{\infty} \min\left\{f_{\bar{D}}(y), f_D(y)\right\} dy \\
&= \int_{-\infty}^{\infty} \mathbb{1}_{\{f_{\bar{D}}(y) < f_D(y)\}} f_{\bar{D}}(y) dy + \int_{-\infty}^{\infty} \mathbb{1}_{\{f_{\bar{D}}(y) \geq f_D(y)\}} f_D(y) dy \\
&= \mathbb{E}_{Y_{\bar{D}}}\left(\mathbb{1}_{\{f_{\bar{D}}(Y_{\bar{D}}) < f_D(Y_{\bar{D}})\}}\right) + \mathbb{E}_{Y_D}\left(\mathbb{1}_{\{f_{\bar{D}}(Y_D) \geq f_D(Y_D)\}}\right) \\
&= \mathbb{P}(f_{\bar{D}}(Y_{\bar{D}}) < f_D(Y_{\bar{D}})) + \mathbb{P}(f_{\bar{D}}(Y_D) \geq f_D(Y_D)).
\end{aligned} \tag{1.3}$$

Indeed, $\mathbb{P}(f_{\bar{D}}(Y_{\bar{D}}) < f_D(Y_{\bar{D}}))$ is the probability of choosing f_D when $f_{\bar{D}}$ is the true density and $\mathbb{P}(f_{\bar{D}}(Y_D) \geq f_D(Y_D))$ is the probability of choosing $f_{\bar{D}}$ when f_D is the true density.

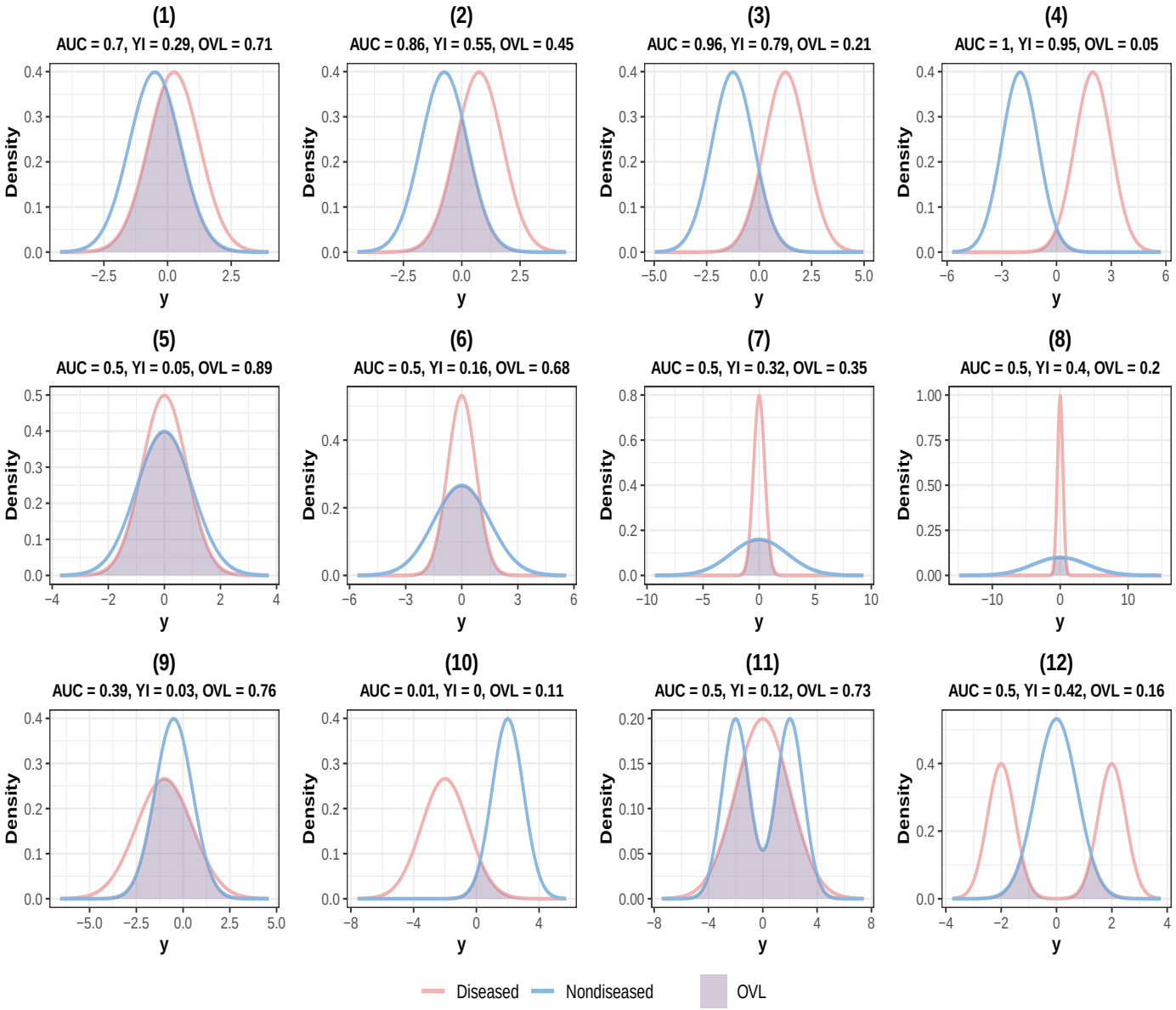

Figure 1.2: Examples of the OVL under different situations. In Examples (1)–(4) the means differ between groups, but the scales are equal. Examples (5)–(8) consider different scales, but equal means. Examples (9)–(12) consider distinct means and scales. Examples (9) and (10) highlight the non-directional property of the OVL. In cases when the mean of the non-diseased group is larger, the AUC and YI fail. Finally, Examples (11) and (12) show that, in cases where the mean is equal, the AUC may not detect shape differences and although the YI does, it fails to take into account multiple crossings of the density functions, thus underestimating the discriminatory ability of the test.

The OVL ranges from 0 to 1. When the distributions do not overlap at all (perfect diagnostic accuracy), OVL = 0. Whereas OVL = 1 means that the distributions are identical and thus the test is useless from a diagnostic viewpoint. To make the OVL values comparable with the other summary measures, we can take its complement, i.e., $1 - \text{OVL}$. Therefore, a larger value corresponds to a better discriminatory ability.

An advantage of the OVL over the AUC and YI is that, in addition of taking into account both the location and shape of the underlying distributions, it is non-directional (Schmid and Schmidt, 2006). Being non-directional means that regardless which distribution (nondiseased or diseased) has the larger mean, the OVL value remains unchanged. This is illustrated in Figure 1.3, albeit the mean of the nondiseased group is greater than that of the diseased group (right panel), the value of the OVL is the same. A disadvantage of the OVL is that an optimal cut-off value cannot be easily obtained, as for example, with the YI.

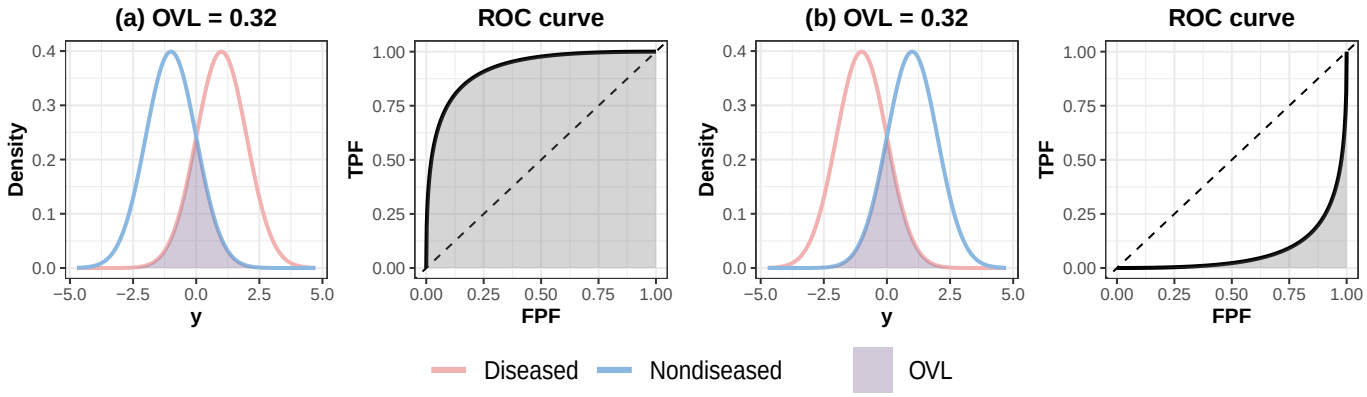

Figure 1.3: The OVL non-directional property illustrated. Panel (a): $Y_{\bar{D}} \sim \text{N}(-1, 1)$, $Y_D \sim \text{N}(1, 1)$. Panel (b): $Y_{\bar{D}} \sim \text{N}(1, 1)$, $Y_D \sim \text{N}(-1, 1)$.

Samawi et al. (2017) derived the relationship between the YI and the OVL under different situations. Specifically, they discussed that when multiple intersections occur between the distributions, as it may be the case of multimodal distributions, the equivalence between the YI and the OVL no longer holds. Furthermore, in this situation, the YI is always less or equal than $1 - \text{OVL}$. For instance, consider the scenario shown in Figure 1.4, both distributions have multiple modes and they intersect at multiple locations as well. In Figure 1.4 left, we clearly observe that both groups are well differentiated. This is

fully captured by the OVL (OVL = 0.04). Conversely, the AUC and the YI consider the discrimination of this marker no better than chance. Even when the decision rule is reversed, both measures regard this marker as uninformative albeit it is highly informative. At this point, it is fair to mention the work by Martinez-Camblor et al. (2017), Martínez-Camblor and Pardo-Fernández (2019) and Pérez-Fernández et al. (2021), who discuss generalisations of the ROC curve to handle these situations.

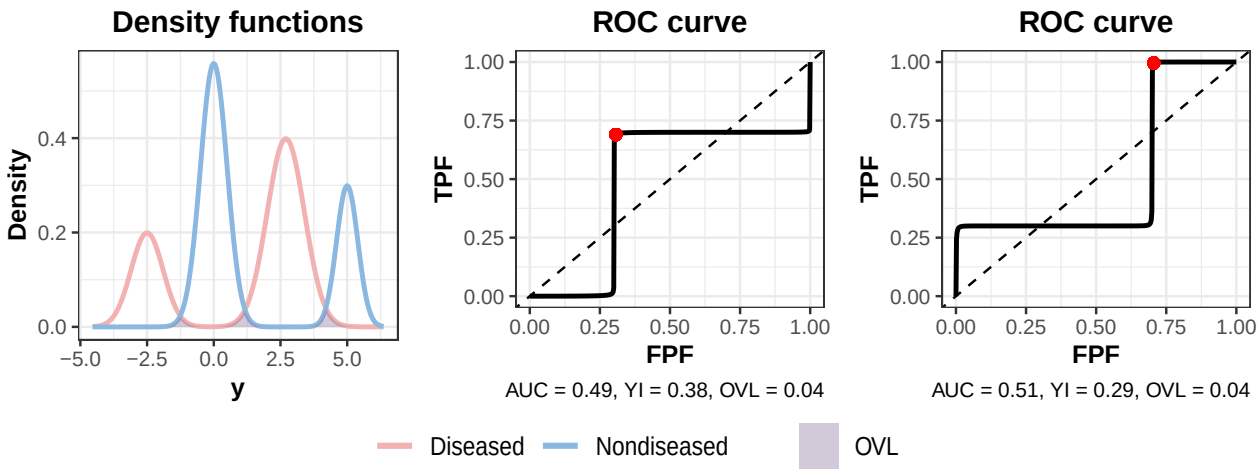

Figure 1.4: Left: the distributions of the test outcomes are given by $Y_{\bar{D}} \sim 0.7\mathrm{N}(0, 0.5^2) + 0.3\mathrm{N}(5, 0.4^2)$, $Y_D \sim 0.3\mathrm{N}(-2.5, 0.6^2) + 0.7\mathrm{N}(2.7, 0.7^2)$. Middle: the resulting ROC curve. Right: the resulting ROC curve if the decision rule is reversed. The YI is denoted by a red point over the ROC.

One interesting feature that the OVL shares with the AUC and the YI is the invariance property, described as follows

Result 1.4.1 *Let g be a strictly increasing and differentiable transformation defined on the supports of $Y_{\bar{D}}$ and Y_D. Then,*

$$\mathrm{OVL}\left(g(Y_{\bar{D}}), g(Y_D)\right) = \mathrm{OVL}\left(Y_{\bar{D}}, Y_D\right).$$

Proof. The corresponding density functions of $g(Y_{\bar{D}})$ and $g(Y_D)$ are given by $\frac{dg^{-1}(u)}{du} f_{\bar{D}}(g^{-1}(u))$ and

$\frac{dg^{-1}(u)}{du} f_D(g^{-1}(u))$, respectively. Then, the OVL is

$$\begin{aligned}
\mathrm{OVL}\left(g(Y_{\bar{D}}), g(Y_D)\right) &= \int_{-\infty}^{\infty} \min\left\{ \frac{dg^{-1}(u)}{du} f_{\bar{D}}(g^{-1}(u)), \frac{dg^{-1}(u)}{du} f_D(g^{-1}(u)) \right\} du \\
&= \int_{-\infty}^{\infty} \min\left\{ f_{\bar{D}}(g^{-1}(u)), f_D(g^{-1}(u)) \right\} \frac{dg^{-1}(u)}{du} du \\
&= \int_{-\infty}^{\infty} \min\left\{ f_{\bar{D}}(y), f_D(y) \right\} dy \quad \text{(change of variable)} \\
&= \mathrm{OVL}\left(Y_{\bar{D}}, Y_D\right).
\end{aligned}$$

We must emphasize that it is not a good practice to rely blindly on any of these measures. The OVL must be complementary to the ROC curve and associated measures. For example, in Figure 1.5 we illustrate three hypothetical situations taken from Pepe et al. (2003). The ideal scenario is shown in panel (a), where both distributions are clearly distinguished. However, the scenarios shown in panels (b) and (c) may be a little tricky. Consider a practical situation where, for instance, it is important to maintain the false positive fractions as low as possible. This can be due to highly expensive or invasive procedures implied by a positive outcome (Pepe et al., 2003). Recall from Section 1.2, that given a threshold c, we can compute the false positive fraction (FPF) as $\mathrm{FPF} = \mathbb{P}(Y_{\bar{D}} \geq c)$ and fix it, for instance, at 0.02 (for the particular distributions used in this example, we set $c = 4$). This means that approximately 2% of nondiseased individuals screen positive. Despite that the OVL for scenario (c) (OVL = 0.48) is much lower than that from scenario (b) (OVL = 0.71), the true positive fraction ($\mathrm{TPF} = \mathbb{P}(Y_D \geq c)$) is about 0.0345. Whereas, for scenario (b) TPF = 0.3091. This means that, under scenario (c), the test detects roughly 3% of the diseased, compared to the 30% identified under scenario (b), even if the OVL is higher in the latter. As we can observe, it is important to not only look at the ROC curve or a single diagnostic summary measure, but to look at the underlying distributions as well. In our case, we get these for free, because the OVL depends entirely on the density functions.

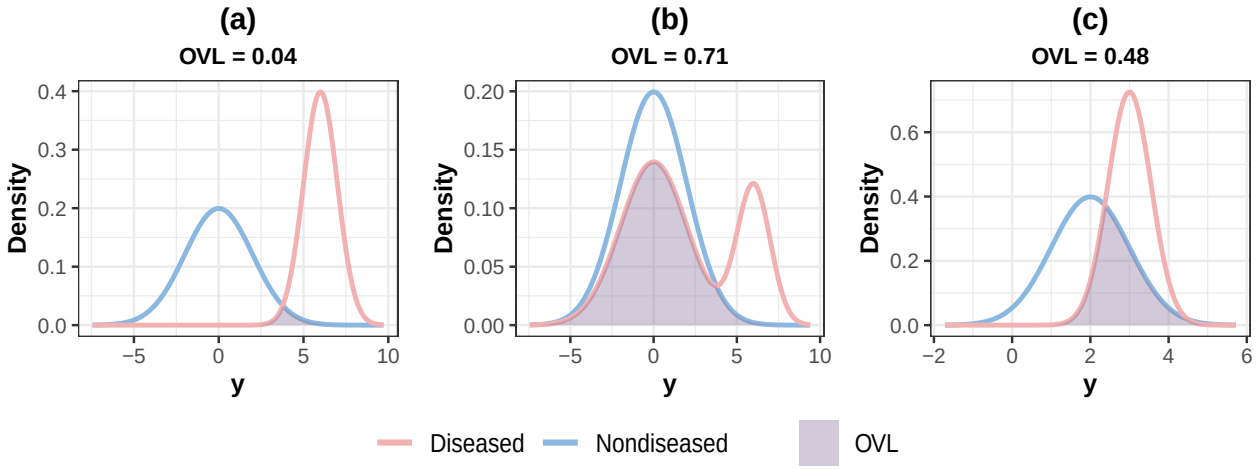

Figure 1.5: Panel (a): $Y_{\bar{D}} \sim \mathrm{N}(0, 2^2)$, $Y_D \sim \mathrm{N}(6, 1)$. The ideal case where both distributions are well-differentiated. Panel (b): $Y_{\bar{D}} \sim \mathrm{N}(0, 2^2)$, $Y_D \sim 0.7\mathrm{N}(0, 2^2) + 0.3\mathrm{N}(6, 1)$. There is a subset of the diseased population which is clearly distinguishable and this might be of more practical interest. Panel (c): $Y_{\bar{D}} \sim \mathrm{N}(2, 1)$, $Y_D \sim \mathrm{N}(3, 0.55^2)$. Even though the OVL is lower than in panel (b), the range of the diseased distribution falls within that of the nondiseased group. This implies that the true positive fraction is lower than in scenario (b).

1.5 Covariate information

It is well-known that individual's characteristics may influence the accuracy of a diagnostic test (Pepe, 2003, Chapter 6). For such reason, it is of particular interest to determine if and how the covariates affect the test accuracy. This is important because identifying sub-populations where the test performs satisfactorily (or poorly) can lead to a more efficient, and accurate, screening and diagnosis or prompt the search for new markers. Ignoring this information may yield to biased or oversimplified inferences. Several methods have been proposed to study the covariates effect under the ROC curve framework. The three main approaches are discussed in Pepe (1998) and Pepe (2003, Chapter 6). One of these approaches is known as the induced approach, where the distributions of both groups are independently modelled and then the ROC curve is induced from them. In this thesis, we follow this induced approach for the coefficient of overlap.

Let $\boldsymbol{X}_{\bar{D}}$ and $\boldsymbol{X}_D$ be two p–dimensional covariate vectors from the nondiseased and diseased population with the corresponding conditional density functions of the test outcomes given by $f_{\bar{D}}(\cdot \mid \boldsymbol{X}_{\bar{D}} = \boldsymbol{x}_{\bar{D}})$

and $f_D(\cdot \mid \boldsymbol{X}_D = \boldsymbol{x}_D)$, respectively. For simplicity and to ease notation, suppose that the same covariates are measured in both groups. Given a covariate value $\boldsymbol{x}$, the covariate-specific ROC curve (Pepe, 1998) is defined as

$$\text{ROC}(p \mid \boldsymbol{x}) = 1 - F_D\left\{F_{\bar{D}}^{-1}(1-p \mid \boldsymbol{X}_{\bar{D}} = \boldsymbol{x}) \mid \boldsymbol{X}_D = \boldsymbol{x}\right\}, \quad 0 \leq p \leq 1,$$

where $F_D(y \mid \boldsymbol{X}_D = \boldsymbol{x}) = \mathbb{P}(Y_D \leq y \mid \boldsymbol{X}_D = \boldsymbol{x})$ is the conditional cdf of Y_D given $\boldsymbol{X}_D = \boldsymbol{x}$ and $F_{\bar{D}}(y \mid \boldsymbol{X}_{\bar{D}} = \boldsymbol{x})$ is similarly defined, with $F_{\bar{D}}^{-1}(1-p \mid \boldsymbol{X}_{\bar{D}} = \boldsymbol{x})$ the corresponding conditional quantile function. Note that for every possible value $\boldsymbol{x}$, we might obtain a different ROC curve. To summarize the information of these collection of ROC curves, the covariate-specific area under the ROC curve (AUC) and the covariate-specific Youden index (Faraggi, 2003) have been proposed. They are, respectively, defined as

$$\text{AUC}(\boldsymbol{x}) = \int_0^1 \text{ROC}(p \mid \boldsymbol{x})dp \quad \text{and} \quad \text{YI}(\boldsymbol{x}) = \max_{c \in \mathbb{R}} \left|F_{\bar{D}}(c \mid \boldsymbol{X}_D = \boldsymbol{x}) - F_D(c \mid \boldsymbol{X}_{\bar{D}} = \boldsymbol{x})\right|.$$

Both the AUC and the YI involve computationally expensive procedures. Whereas the estimation of the AUC needs to find the quantile function of the nondiseased group numerically (unless a parametric model is assumed), the YI requires solving a maximization problem and, possibly, getting stuck at a local maxima. Although, a grid search algorithm can be used, in high dimensional settings (three or more continuous covariates) this may be infeasible. As mentioned before, both measures usually assume that the mean of the diseased population is larger than that of the nondiseased population, meaning that extra care must be taken and further arrangements have to be performed if the mean of the diseased population is smaller than that of the nondiseased population. For such reasons, we propose the covariate-specific coefficient of overlap (OVL) as an alternative diagnostic summary measure. The covariate-specific OVL is defined as

$$\text{OVL}(\boldsymbol{x}) = \int_{-\infty}^{\infty} \min\left\{f_{\bar{D}}(y \mid \boldsymbol{X}_{\bar{D}} = \boldsymbol{x}), f_D(y \mid \boldsymbol{X}_D = \boldsymbol{x})\right\} dy. \tag{1.4}$$

Like its unconditional counterpart (with no covariates), the covariate-specific OVL ranges from zero to one and has the same interpretation described previously. In Figure 1.6 is illustrated a simple example with a single continuous covariate x, where x lies in a uniform grid in $(0, 1)$. Panels a), b)

and c) depict the conditional density function of the nondiseased and diseased groups, for $x = 0.2$, $x = 0.5$ and $x = 0.8$, respectively. Here, it is possible to observe how the overlap value changes as a function of x (Figure 1.6 bottom row right). The covariate-specific OVL inherits all the properties of the unconditional case, that is, it is invariant under strictly increasing transformations, it takes into account both location and shape of the underlying distributions and it is non-directional. The results are straightforward as they follow from the no covariates case, conditioning on a fixed value of the covariates (for proofs of the unconditional case see, for example, Schmid and Schmidt (2006)). Being non-directional in the covariates setting is particularly advantageous, because ROC type measures require knowing in advance the conditional distributions.

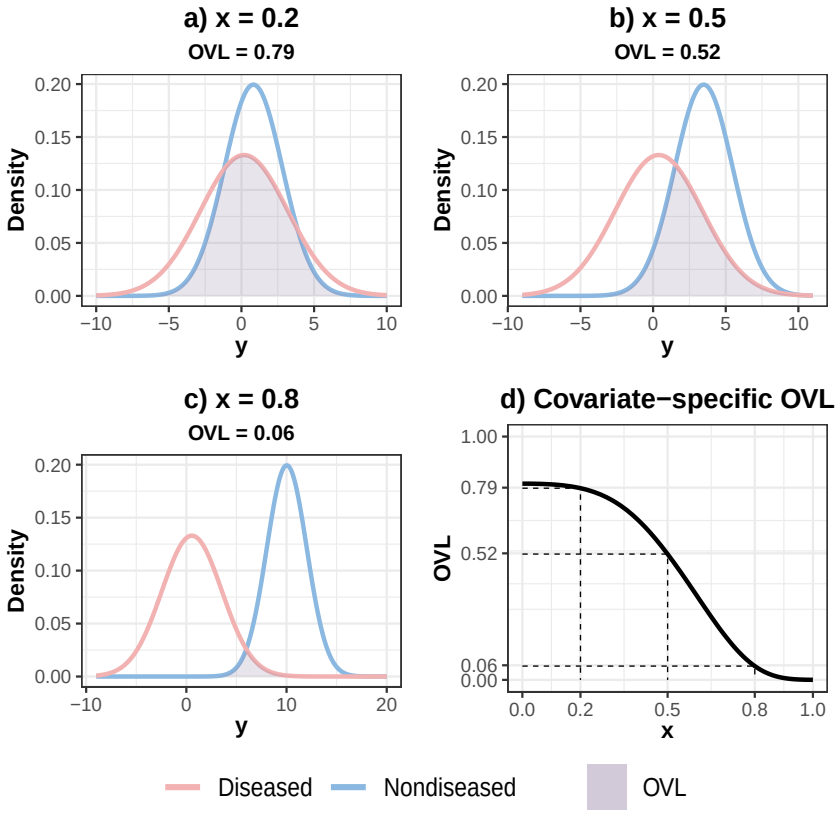

Figure 1.6: Conditional distributions for the test outcomes given three particular values of the covariate ($x = 0.2$, $x = 0.5$ and $x = 0.8$). Distributions for test outcomes: $Y_{\bar{D}} \sim \mathrm{N}(-\exp\{-x\} + 1, 3^2)$, $Y_D \sim \mathrm{N}(\exp\{3x\} - 1, 2^2)$. Plots a), b) and c) depict the conditional density distributions for the nondiseased (blue) and diseased (red) populations, where the gray area represents the overlap coefficient. The corresponding covariate-specific OVL is shown in the bottom right plot.

We must emphasize that the effect of the covariates on test outcomes does not necessarily mean a change in the OVL. For example, Figure 1.7 shows the case of a biomarker with a single binary covariate, $x \in \{0, 1\}$, say, e.g., sex. The underlying distributions of test outcomes are shifted, but the OVL values for both sub-groups are the same. This is because the amount of separation between the two distributions, regardless of the shift in the test outcomes, remains the same.

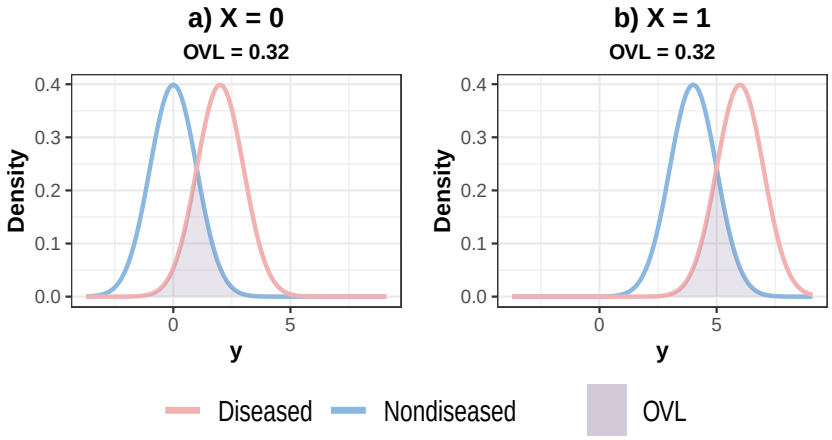

Figure 1.7: Hypothetical example where a single binary covariate affects the test outcomes distributions, but the OVL remains unchanged.

1.6 Thesis outline

The main object of study throughout this thesis is the coefficient of overlap (OVL). Our focus in this work is to develop a flexible framework to estimate and conduct inference about the OVL in both the unconditional and conditional case. For such end, we make use of Bayesian nonparametric methods, which have been shown to be flexible enough to model the complex structures observed in biomedical data (see, for example, Inácio de Carvalho et al., 2013, Inácio et al., 2020 and Johnson and de Carvalho, 2015). The remainder of the thesis is organized as follows. In Chapter 2, we introduce briefly the Bayesian paradigm and some Bayesian nonparametric methods. We focus our discussion around the Dirichlet process mixture (DPM) models, which are the core of our modelling approaches. Different model comparison criteria are provided as well. We use them to compare our proposed models to other modelling approaches.

In Chapter 3, we present our proposed modelling framework, based on Dirichlet process mixtures and the Bayesian bootstrap, for conducting inference about the overlap coefficient when there are no covariates. We propose two OVL estimators based on (1.2) and (1.3). The practical performance of our proposed estimators was assessed through a simulation study, which revealed that our estimators are able to recover the true value of the overlap coefficient. We found that the empirical coverage probability of the corresponding 95% credible intervals is close to the nominal value under a variety of conceivable test outcomes distributions. We illustrate our approach through an application concerned with the search for biomarkers of ovarian cancer. We have submitted an article based on the content of this chapter to *Statistics in Medicine*.

In Chapter 4, we propose the covariate-specific coefficient of overlap, intended to determine the optimal and suboptimal populations where to perform the tests on. We follow a joint approach where the test outcomes and the covariates are both treated as random and are modelled jointly using a DPM model. Generally, under the conditional approach, specific functional forms of how the covariates influence, for example, the mean of the conditional distribution of the test outcomes are assumed. In this sense, our modelling approach relaxes such assumptions and any non-linearity or heterogeneity is automatically accommodated. Further, other covariate-dependent functions of interest are available, such as the conditional mean, quantile or variance function. A simulation study demonstrated that our proposed estimator works well under a variety of scenarios, being able to recover the functional form of the coefficient of overlap. We also found that the empirical coverage probability of the 95% credible intervals, on average, is below the nominal value in complex scenarios. However, the distance between the real curve and the 95% credibility band is minimal. We illustrate our modelling approach through two different real datasets. For one of the datasets, we examine how the discriminatory ability of glucose levels changes with age. The other dataset is concerned with the search of biomarkers for the Alzheimer's disease. The material in this chapter also serves as a basis for a second article that will be submitted soon.

In Chapter 5, we include vignettes of our developed R (R Core Team, 2020) package, called `OverlapCoefficient`. R is a free-source programming language widely used among statisticians and non-statisticians for data analysis. In this package, we have implemented all our methods into stan-

dalone functions, prioritizing an user-friendly experience. The package is currently in a private Github repository (`https://github.com/javier-gg/OverlapCoefficient`). However, we aim to make the package publicly available once we submit the second article.

Finally, further directions of research are discussed. These include possible generalizations of the OVL to handle two or more biomarkers.

Chapter 2

Bayesian analysis

Throughout this thesis, we use Bayesian methods to conduct inference about the distribution of the test outcomes and ultimately, about the coefficient of overlap and its covariate-specific counterpart. For such reason, we start our discussion giving a brief introduction to the Bayesian paradigm and the Bayesian nonparametric methods. Excellent references dedicated to Bayesian inference can be found, among others, in Hoff (2009), Bolstad and Curran (2016), Gelman et al. (2013), Christensen et al. (2011) and McElreath (2018). There is already vast research on Bayesian nonparametric methods as well. A general review can be found in Walker et al. (1999) and Müller and Quintana (2004). Other excellent and concise discussions include, among others, Hjort (2003), Hjort et al. (2010), Lijoi et al. (2010), Rodriguez and Müller (2013) and Walker (2013).

2.1 The Bayesian paradigm

Bayesian statistics has its origins in the posthumous paper *An Essay Towards Solving a Problem in the Doctrine of Chance* written by the Reverend Thomas Bayes (1701-1761) and published by his friend Richard Price in 1793. However, Bayesian statistics did not become popular until the 80's, early 90's, because of the complex computations involved. Thus, its implementation into real scientific problems resulted impractical and fell into abandonment. Pierre–Simon Laplace was the first to adopt Bayes'

ideas and to translate them into a formal theorem in the 19th century. However, it was not until the middle of the 20th century, that interest was renewed and de Finetti, Jeffreys, Savage and Lindley, among others, developed a complete methodology based on the Bayes' theorem (Bolstad and Curran, 2016, p. 6).

Bayesian inference essentially works as follows: we wish to perform inference about the characteristics of a population, represented mathematically as a random variable Y. Typically, these characteristics are expressed in terms of a vector of parameters $\boldsymbol{\theta}$ from a probability model $f(\cdot \mid \boldsymbol{\theta})$, which corresponds to the probability density function (pdf) of Y and for which we observe the data $\boldsymbol{y} = (y_1, \ldots, y_n)$. Then, the Bayes' rule states that

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{p(\boldsymbol{\theta}) f(\boldsymbol{y} \mid \boldsymbol{\theta})}{f(\boldsymbol{y})},$$

where $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ is referred as the posterior distribution and $p(\boldsymbol{\theta})$ as the prior distribution. The posterior distribution contains all the information about the parameters after observing the data. Whereas the prior represents the prior beliefs about the parameters before observing any data. The probability model or sampling distribution, $f(\boldsymbol{y} \mid \boldsymbol{\theta})$, is usually referred as the likelihood. Often, it is denoted as $L(\boldsymbol{\theta}; \boldsymbol{y})$ to emphasize that it is a function of $\boldsymbol{\theta}$ and the observed data, $\boldsymbol{y}$, is regarded as fixed or constant. The term in the denominator, $f(\boldsymbol{y})$, is called the marginal distribution of the data, for example, in the continuous case, this corresponds to

$$f(\boldsymbol{y}) = \int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}) f(\boldsymbol{y} \mid \boldsymbol{\theta}) d\boldsymbol{\theta}.$$

Note that this last term does not depend on $\boldsymbol{\theta}$ and can be considered as a constant. Thus, it is common to rewrite the Bayes' rule as

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto p(\boldsymbol{\theta}) p(\boldsymbol{y} \mid \boldsymbol{\theta}),$$

yielding to the unnormalized posterior distribution. Hence, the Bayesian approach gives a natural mechanism to accommodate all the information available and the uncertainty into a single framework. Then, in light of new information, by using the Bayes' rule the initial beliefs or hypotheses are updated. Therefore, the Bayesian paradigm is thus consistent with how science works.

Unlike the frequentist approach, the Bayesian paradigm models all the uncertainty with probability distributions and always obey to the laws of probability. Further, considering the parameters as random allows to make probability statements rather than confidence statements based on what might have occurred in the long run (Bolstad and Curran, 2016, p. 7). Prediction is straightforward for Bayesian methods and they do not depend solely on large sample theory. Finally, Bayesian methods for complex models are usually more straightforward than non Bayesian methods. Despite these advantages, the Bayesian framework has not been exempt of criticism. Partly, because of its subjective nature and the underlying complexity of its calculations. Opposition consider that methods, as scientific research, must be completely objective and that the choice of a prior based on individual beliefs makes the Bayesian approach unreliable, as different priors may lead to different posteriors and thus to different conclusions. However, obtaining new knowledge based on previous experience is what makes science, science. The scientific method uses information available to build up hypotheses, then, after conducting an experiment and analysing the resulting data, it updates the knowledge about the study object and confirms or rejects such hypotheses.

Setting the prior distribution can be, indeed, challenging. In practice, previous knowledge usually comes in form of previous studies, expert opinion or experience. However, translating this into a proper mathematical object (the prior distribution) can be a non trivial task. This is known as the elicitation process and there is extensive research on this. However, it falls out of the scope of this thesis. We refer the reader, for example, to Spiegelhalter et al. (2004); Berry and Stangl (2018); Moyé (2016) or Zou et al. (2011), for an overview. In absence of prior knowledge, one can use non informative or relative vague priors that reflect the degree of uncertainty. Conjugate priors are typically used to aid the computation of the posterior distribution. This is because conjugate priors are distribution families, such that the corresponding posterior distribution belongs to the same family as the prior distribution, e.g., a normal prior for the mean of a normal sampling distribution is conjugate, because the posterior distribution for the mean is normal as well. Under certain conditions, the Bayesian central limit theorem assures that whatever the choice of the prior is, it becomes less and less important as the sample size increases (Gelman et al., 2013, pp. 83–84).

Computational power narrowed the spread of Bayesian methods. For almost two centuries, appli-

cations were confined to cases where either conjugacy allowed for analytical posteriors or where low dimension of the parameters vector permitted numerical integration. Nowadays, thanks to the advance of technology, the computation of the posterior has become feasible. Monte Carlo methods arose from the necessity of giving practical, rather than abstract solutions to a wide range of problems, spanning from casino bets to nuclear physics (Eckhardt, 1987). Monte Carlo integration was developed along these methods back in the 1940's. It is a widely used technique, which approximates complex or infeasible integrals. Basically, the idea is to generate a sample of random values, $\theta_1, \ldots, \theta_S$, which are then evaluated in an integrand f and finally, approximate the integral by the sample mean, that is,

$$\int f(\theta) d\theta \approx \frac{1}{S} \sum_{s=1}^{S} f(\theta_s).$$

This is crucial in the Bayesian setting, because if we could sample, say S, random samples from the posterior distribution, $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(S)}$, then we could approximate the posterior distribution. Further, any other summary measure of interest would be available, because by the Law of Large Numbers,

$$\frac{1}{S} \sum_{s=1}^{S} g\left(\boldsymbol{\theta}^{(s)}\right) \overset{s \to \infty}{\longrightarrow} \mathbb{E}\left(g(\boldsymbol{\theta}) \mid \boldsymbol{y}\right),$$

where $g(\cdot)$ is any arbitrary function. A natural question may arise now, how to simulate from the posterior distribution? Several methods have been proposed for such end, for example, the rejection sampling, importance sampling or sampling importance resampling (Gelman et al., 2013, pp. 264–265). However, Monte Carlo Markov chain (MCMC) methods have become more popular due to their flexibility and applicability to almost any statistical problem (Gelfand and Smith, 1990).

MCMC methods combine the use of Markov chains to generate a sequence of random numbers from the posterior and Monte Carlo integration to summarize these simulations. A Markov chain is a stochastic process whose draws only depend on the previous state of the chain, i.e., if $\boldsymbol{\theta}^{(i+1)}$ is the current realization of the chain, then its probability only depends on $\boldsymbol{\theta}^{(i)}$. A well-constructed Markov chain, will eventually converge to the stationary distribution, in our case, the posterior distribution (Gelman et al., 2013, pp. 275–276). Ergodic theory ensures that convergence is achieved after some realisations. However, it is still an active area of research to assess when the chain has converged. Some authors have proposed different convergence diagnostics to help identifying when the chain has not converged,

for example, trace plots, the Brooks-Gelman-Rubin method (Gelman et al., 1992; Brooks and Gelman, 1998) or the Geweke diagnostic (Geweke, 1992). All these diagnostics are available in the R package `coda` (Plummer et al., 2006) and we discuss them briefly below.

Trace plots

A trace plot is graphical tool which displays the posterior simulated value at each MCMC iteration. Trace plots can help to determine an appropriate burn-in period. If the simulated values are within a region of the posterior sample space and they do not show any periodicity or specific trend (i.e., the plot only shows randomness), then there is no evidence of lack of convergence.

Geweke convergence diagnostic

The Geweke convergence diagnostic is based on testing the posterior sample for equality of means between the first and last part of a Markov chain (usually, the first 10% and the last 50%). Under the null hypothesis, the Geweke's statistic has an asymptotically standard normal distribution, thus if the samples are drawn from the stationary distribution, both means should be equal. This leads to use bands between -2 and 2 (roughly the quantiles 0.025 and 0.975 of a standard normal distribution) to visually evaluate this statistic.

Additionally, a common problem, albeit inherent to the methodology related to MCMC methods, is the autocorrelation of the simulated posterior sample. The effective sample size represents the posterior sample size adjusted for autocorrelation. For instance, if the effective sample size of a posterior sample is close to the number of MCMC iterations, then the observations are less autocorrelated and we would obtain an approximately independent sample. For such reason, it is also important to look at the effective sample size to assess the validity of our inferences.

Different approaches have been proposed to construct such Markov chains. The first algorithm was proposed by Metropolis et al. (1953) and later generalised by Hastings (1970) yielding to the Metropolis–Hastings (MH) algorithm. The MH method is based on an acceptance-rejection scheme and has multiple applications due to its versatility, because it allows for models without conjugate distributions. However, appropriate proposal distributions and tuning parameters are needed to perform an efficient sampling.

A particular case of the MH algorithm is known as the Gibbs sampler. It was proposed by Geman and Geman (1984) and it exploits conjugacy to obtain full conditional distributions to sample from. A full conditional is the conditional distribution of a parameter given everything else, for example, if $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$, then the full conditional for θ_i is given by $p(\theta_i \mid \theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_p, \boldsymbol{y})$. In this thesis, we will use the Gibbs sampler algorithm because the full conditional distributions of all parameters from our proposed models are available in closed form.

We have reduced an apparently infeasible problem into a more manageable one, sampling from univariate distributions rather than difficult multivariate distributions and using these simulations in conjunction with Monte Carlo integration to obtain quantities of interest. Along with the increasing computational power, there are plenty softwares available dedicated to the implementation of Bayesian methods, for example, WinBUGS (Lunn et al., 2000), JAGS (Plummer, 2003), Stan (Stan Development Team, 2021), INLA (Rue et al., 2009), etc. In our case, we have implemented our methods in an R package `OverlapCoefficient`, which is described in Chapter 5.

2.2 Bayesian nonparametric methods

As we have seen in Section 1.4, estimating the OVL, mainly, consists in estimating accurately the underlying density functions of the test outcomes in each group. One possible approach is to assume a parametric model for each of them. A popular choice is the so-called binormal model (Pepe, 2003, p. 81), which assumes a normal distribution for both populations, possibly, after transforming the Y scale (e.g., using a Box-Cox transformation). In many practical situations, the test outcomes may present a complex structure (e.g., multimodality, skewness, excess of kutosis, etc.), which can be difficult to capture with a single parametric model. This motivates the use of a broader class of models that allows for flexibility and robustness against mis-specification of a parametric model (Müller et al., 2015, p. 1).

Parametric models are described by a finite-dimensional vector of parameters $\boldsymbol{\theta}$. Assuming that data come from an underlying probability distribution G, a parametric model arises from a family of distributions $\{G_{\boldsymbol{\theta}} \mid \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$, indexed by a set of parameters $\boldsymbol{\theta}$ of a finite-dimensional set $\boldsymbol{\Theta}$. Often,

this assumption is too restrictive and limits the scope and type of inferences that could be drawn (Müller et al., 2015, p. 1). Conversely, Bayesian nonparametrics (BNP) broadens the class of models for G, such that, attention is now moved to $\{G \mid G \in \mathcal{G}\}$, with $\mathcal{G}$ being an infinite-dimensional space of distributions. Considering this type of families requires priors for $\mathcal{G}$ over the space of probability measures. Such priors are known as Bayesian nonparametric priors (Müller and Mitra, 2013; Müller et al., 2015). Usually, infinite dimensional parameters of interest come in form of functions, e.g., probability density functions or conditional mean functions (regression functions) (Müller et al., 2015, p. 4). We must note that the term nonparametric in BNP may be misleading. It does not mean that there are no parameters in BNP methods. They are, in fact, massively parametric. In contrast, it refers to the fact that these methods are free from restrictive or mis-specification of particular parametric models (Inácio, 2012). Our discussion will be around one of the most popular BNP priors, the Dirichlet process (DP). Later, we will introduce the Dirichlet process mixture (DPM) model, which will be the core of our modelling approaches.

2.2.1 Mixtures of normals

Before introducing the DP and the DPM, we will begin our discussion around finite mixture models because, to a certain extent, they provide the background to understand DPMs. In many practical situations, the outcomes of a diagnostic test may present a complex structure. For example, within the diseased population there might be sub-populations which differentiate several stages of the same disease. Multimodality is not the sole characteristic usually observed in biomedical data, skewness or excess of kurtosis may be present as well. This motivates the use of a broader class of models that allows for more flexibility, such as mixture models. In a mixture model we assume that the data arise from one or many sub-populations or clusters and each group can be represented by a single parametric model. Thus, the complete data can be characterized by a weighted sum of distributions.

In what follows, we will focus on mixtures of normal distributions, which are known to approximate any smooth continuous distribution on the real line, provided one uses an appropriate number of components (Lo 1984, Rossi 2014, p. 5). The density function of a mixture of univariate normals is

given by

$$f(y \mid \boldsymbol{\omega}, \boldsymbol{\theta}) = \sum_{k=1}^{K} \omega_k \phi(y \mid \mu_k, \sigma_k^2), \quad \omega_k \geq 0, \quad \sum_{k=1}^{K} \omega_k = 1, \tag{2.1}$$

where $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_K)$ is the vector of weights, $\boldsymbol{\theta} = (\mu_1, \ldots, \mu_K, \sigma_1^2, \ldots, \sigma_K^2)$ and $\phi(\cdot \mid \mu_k, \sigma_k^2)$ denotes the density function of a normal random variable with mean μ_k and variance σ_k^2. The K mass points in (2.1) are often called the components of the mixture. Equivalently, we can rewrite the model in (2.1) in a general form as

$$f(y \mid \boldsymbol{\omega}, \boldsymbol{\theta}) = \int \phi(y \mid \mu, \sigma^2) dG(\mu, \sigma^2).$$

Here $G(\cdot)$ is known as the mixing distribution and in this case, it is a discrete distribution defined as

$$G(\cdot) = \sum_{k=1}^{K} \omega_k \delta_{(\mu_k, \sigma_k^2)}(\cdot),$$

where $\delta_a(\cdot)$ denotes a point mass at a. The success of normal mixture models relies in great part due to the ease of inference methods (Rossi, 2014, p. 5). Although the likelihood is not tractable for the model in (2.1), because it is a product of sums, we can overcome this issue by introducing latent random variables $z_i \in \{1, \ldots, K\}$, for $i = 1, \ldots, n$. These represent the component to which the i–th observation was allocated. In this sense, if $z_i = k$, then the observation y_i arises from a normal distribution with mean μ_k and variance σ_k^2. Under the Bayesian approach, we must set prior distributions for $\boldsymbol{\omega}$ and $\boldsymbol{\theta}$. Often, for conjugacy reasons, a Dirichlet prior is placed over the vector of weights, whereas a normal-inverse-gamma prior is used for the mean and variance of the mixture components.

Mixture models are widely used in practice because of their flexibility. However, they have some limitations. The most important one is, possibly, choosing the number of components, which is not a trivial task. For example, a prior on K can be placed, but this is difficult to implement efficiently in practice (e.g., using reversible jump type of algorithms), because it is computationally expensive. Another approach would be to assess the goodness of fit of a set of K values. For example, a mixture model using $K = 2, 3, 4, 5$ could be fitted and then, some kind of model comparison criterion (e.g., the

LPML, the WAIC and/or the DIC) could be used to decide which model yields the best fit. A detailed description of model comparison criteria is given in Section 2.3. Recent approaches have emerged to overcome common problems encountered in mixture models, for example, the work by Van Havre et al. (2015) and Frühwirth-Schnatter and Malsiner-Walli (2019). The former propose solutions to the non identifiability of the mixture components, poorly mixing of the MCMC algorithms and the label switching problem. In the latter, the authors proposed the use of sparse priors on the weights to overcome overfitting. This approach poses a direct link to Dirichlet process mixtures.

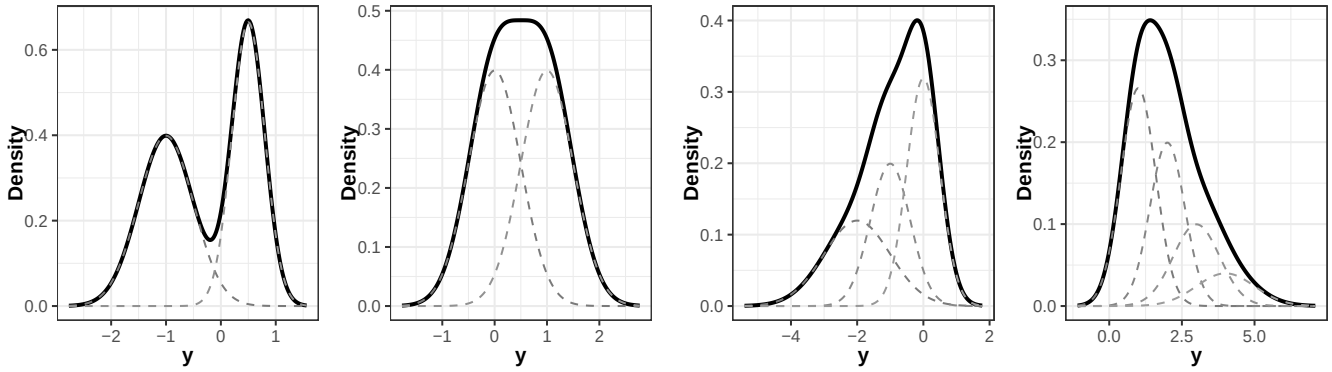

Figure 2.1: Examples of mixtures of univariate normal distributions.

2.2.2 Dirichlet processes

A Dirichlet process (DP) (Ferguson, 1973, 1974) is a stochastic process whose realisations are probability measures. Formally, we say that G follows a DP with parameters α and G^*, denoted by $G \sim \text{DP}(\alpha, G^*)$, if and only if for any finite and measurable partition, $A_1, \ldots, A_K$, of the sample space $\mathscr{X}$, the random vector

$$(G(A_1), \ldots, G(A_K)) \sim \text{Dirichlet}\left(\alpha G^*(A_1), \ldots, \alpha G^*(A_K)\right).$$

The parameter $\alpha > 0$ is called the precision or concentration parameter and G^* is called the baseline distribution. For any subset A of $\mathscr{X}$, the mean and the variance are given by

$$\begin{aligned} \mathbb{E}\left(G(A)\right) &= G^*(A), \\ \text{Var}\left(G(A)\right) &= \frac{G^*(A)(1 - G^*(A))}{\alpha + 1}. \end{aligned} \tag{2.2}$$

This comes from the fact that $G(A) \sim \text{Beta}\left[\alpha G^*(A), \alpha(1 - G^*(A))\right]$. Thus, the designation of G^* as baseline distribution can be regarded as our best guess for G, with α controlling how certain we are about such guess. As can be seen from (2.2), larger values of α imply realisations of G that are closer to G^*.

One important property of the DP is that it is conjugate with respect to i.i.d. sampling (Ferguson, 1973, Theorem 1), that is, consider the following

$$\begin{aligned} y_1, \ldots, y_n \mid G &\overset{\text{iid}}{\sim} G \\ G &\sim \text{DP}(\alpha, G^*). \end{aligned}$$

Then the posterior distribution of G is given by

$$G \mid y_1, \ldots, y_n \sim \text{DP}\left(\alpha + n, \frac{\alpha}{\alpha + n} G^* + \frac{1}{\alpha + n} \sum_{i=1}^{n} \delta_{y_i}\right).$$

A posterior point estimate for $G(t)$ is

$$\mathbb{E}(G(t) \mid y_1, \ldots, y_n) = \frac{\alpha}{\alpha + n} G^*(t) + \frac{n}{\alpha + n} G_n(t),$$

where $G_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{y_i \leq t\}$ is the empirical distribution function of the data. Interestingly, the DP has a connection with the Bayesian bootstrap (BB) (Rubin, 1981). Gasparini (1995) showed that the BB can be regarded as a non-informative version of the DP, as the precision parameter α approaches zero. Unlike the frequentist bootstrap (Efron, 1979), where the proportion of times each observation appears in a bootstrap replication follows a discrete uniform distribution, the Bayesian bootstrap generates a posterior probability for each observation from a Dirichlet distribution. This means that the data is regarded as fixed, we do not resample from it. Hence, the weights in the BB are smoother than those from the classical bootstrap.

Undoubtedly, the most useful representation of the DP was given by Sethuraman (1994). This is the so-called stick-breaking representation, according to which G has an almost sure representation of the form

$$G(\cdot) = \sum_{k=1}^{\infty} \omega_k \delta_{\boldsymbol{\theta}_k}(\cdot), \quad \omega_k = \begin{cases} v_k, & \text{if} \quad k = 1 \\ v_k \prod_{t<k} (1 - v_t), & \text{if} \quad k \geq 2 \end{cases}, \tag{2.3}$$

where the atoms $\boldsymbol{\theta}_k \overset{\text{iid}}{\sim} G^*$ and $v_k \overset{\text{iid}}{\sim} \text{Beta}(1, \alpha)$ are mutually independent for $k \geq 1$. Basically, the idea consists of starting with a stick of one unit length and breaking it at random according to a beta distribution. The first break happens at $\omega_1 = v_1$, which is the proportion assigned to $\boldsymbol{\theta}_1$, ω_2 is the proportion of the remaining $1 - v_1$ length stick allocated to $\boldsymbol{\theta}_2$, and so on. Thus, the remaining stick will be shorter and shorter. Figure 2.2 illustrates the stick-breaking process. Note that this construction implies that G is a discrete distribution with probability one.

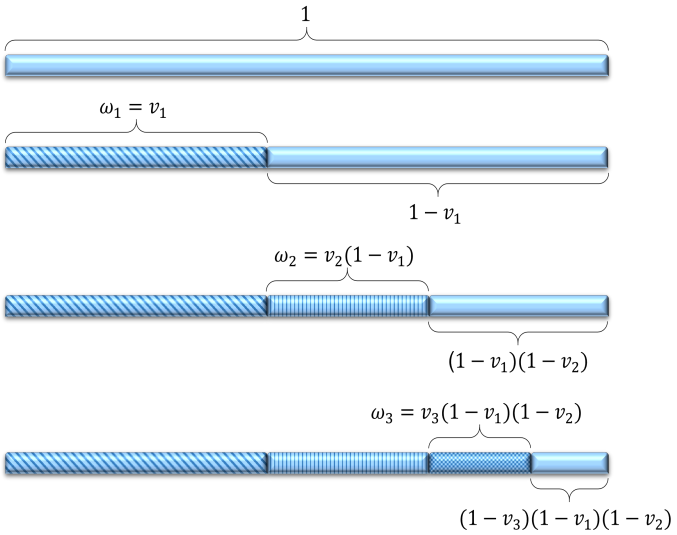

Figure 2.2: The stick-breaking process.

The weights of the stick-breaking process decrease geometrically in k (Rossi, 2014, p. 69). Since $\mathbb{E}(v_k) = \frac{1}{1+\alpha}$, small values of α result in a fast decline of the weights. This benefits a higher mass to be placed on the first atoms, see Figure 2.3 top left. Whereas, larger values of α allow more mass to be spread on more atoms as shown in Figure 2.3 top right. In the bottom row of Figure 2.3 are depicted 30 corresponding cdfs trajectories of the DP and, as we might observe, it is more evident the effect of the α parameter. As α increases the smoother and closer to the baseline distribution the trajectories are. This versatility of the stick-breaking construction has permitted to build up efficient MCMC algorithms (Ishwaran and James, 2001).

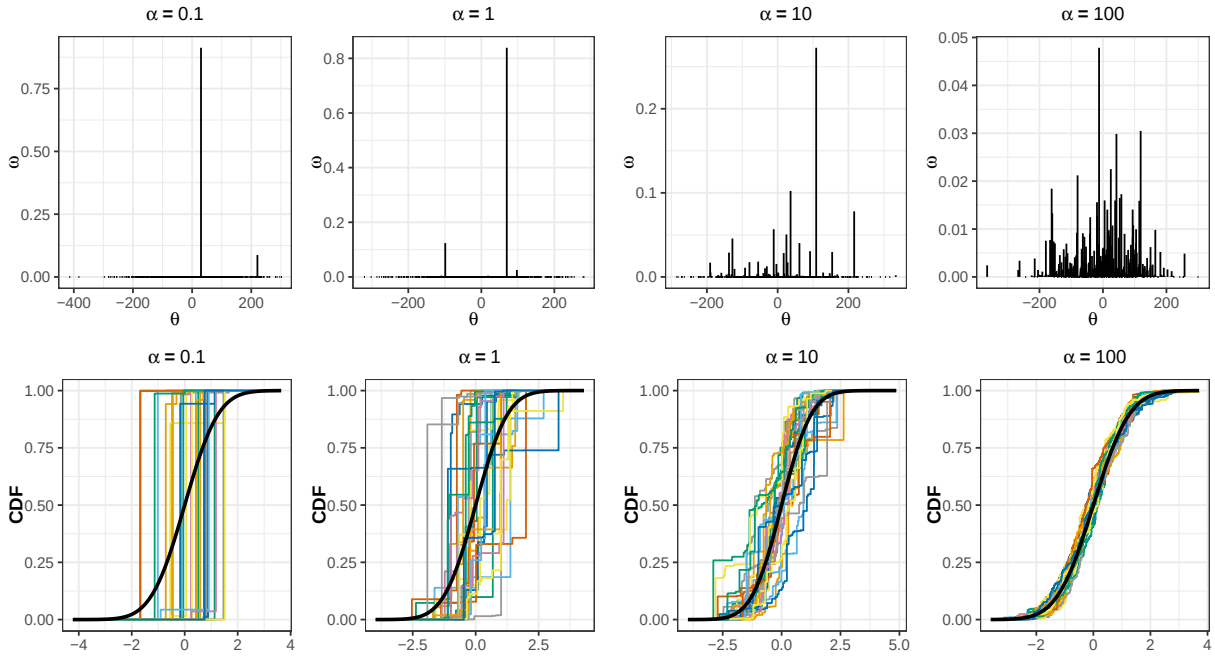

Figure 2.3: Top: Draws from a Dirichlet process with $G^* = \mathrm{N}(0, 1)$ and different values of the concentration parameter α. The spiked lines are placed at 1000 sampled values of G^*. Their heights are determined by the weights ω_k using a truncated version of the stick-breaking representation, so the weights sum up to one. Bottom: 30 associated trajectories based on 1000 draws. The solid line is the cdf of the standard normal distribution.

2.2.3 Dirichlet process mixtures

Despite the usefulness and flexibility of the DP formulation, we have observed that it generates almost surely discrete probability measures. For such reason, it would not be suitable for approximating continuous distributions as in our case. A possible solution to this limitation is found by convolving the trajectories of the DP with some continuous kernel (Müller et al., 2015, p. 11). This approach was introduced by Ferguson (1983), Lo (1984), Escobar (1988), Escobar (1994) and Escobar and West (1995). Let G be a distribution defined on a finite dimensional parameter space $\boldsymbol{\Theta}$ of a continuous pdf $f(\cdot \mid \boldsymbol{\theta})$ with parameter vector $\boldsymbol{\theta}$, then the pdf of a mixture of f with respect to G is given by

$$f(\cdot \mid G) = \int f(\cdot \mid \boldsymbol{\theta}) dG(\boldsymbol{\theta}). \tag{2.4}$$

Further, if a DP prior is placed on the mixing distribution G, then this results in the so-called

Dirichlet process mixture (DPM) model, which may be specified as follows

$$
\begin{aligned}
y_1, \ldots, y_n \mid f &\overset{\text{iid}}{\sim} f, \\
f(\cdot \mid G) &\sim \int f(\cdot \mid \boldsymbol{\theta}) dG(\boldsymbol{\theta}), \\
G \mid \alpha, \boldsymbol{\psi} &\sim \text{DP}(\alpha, G^*), \quad G^* = G^*(\cdot \mid \boldsymbol{\psi}), \\
\alpha, \boldsymbol{\psi} &\sim p(\alpha) p(\boldsymbol{\psi}).
\end{aligned}
$$

It is worth to say, that in many applications, a common choice is to set $\alpha = 1$. However, as in our case, a gamma prior can be placed instead (Escobar and West, 1995). Also, we can use the Sethuraman's stick-breaking representation to express f as an infinite mixture of parametric distributions as follows

$$
f(\cdot \mid G) = \sum_{k=1}^{\infty} \omega_k f(\cdot \mid \boldsymbol{\theta}_k).
$$

Note that the DPM generates either discrete or continuous distributions depending on the nature of the selected kernel. For example, if the kernel is a Poisson distribution, as shown in Figure 2.4 top right, the resulting distribution will be discrete. In contrast, Figure 2.4 bottom right illustrates the resulting distribution for a Gaussian kernel.

Samplers based on the Pòlya urn representation of the DP (Blackwell and MacQueen, 1973), blocked Gibbs (Ishwaran and James, 2001) or "reversible jump" (Richardson and Green, 1997) samplers are popular approaches used to draw simulations from the posterior distribution in DPM models. Throughout this thesis, we will use the blocked Gibbs sampler, which is based on approximating G by truncation of the stick-breaking representation. As we mentioned earlier, the stick-breaking weights decrease geometrically as k increases. Thus, it turns out reasonable to replace the infinite sum in (2.3) by a sum of the first, say K terms, with K sufficiently large. Hence, we can rewrite (2.3) as

$$
G^K(\cdot) = \sum_{k=1}^{K} \omega_k \delta_{\boldsymbol{\theta}_k}(\cdot), \quad \omega_k = \begin{cases} v_k, & \text{if} \quad k = 1 \\ v_k \prod_{t<k} (1 - v_t), & \text{if} \quad 2 \leq k \leq K \end{cases}. \tag{2.5}
$$

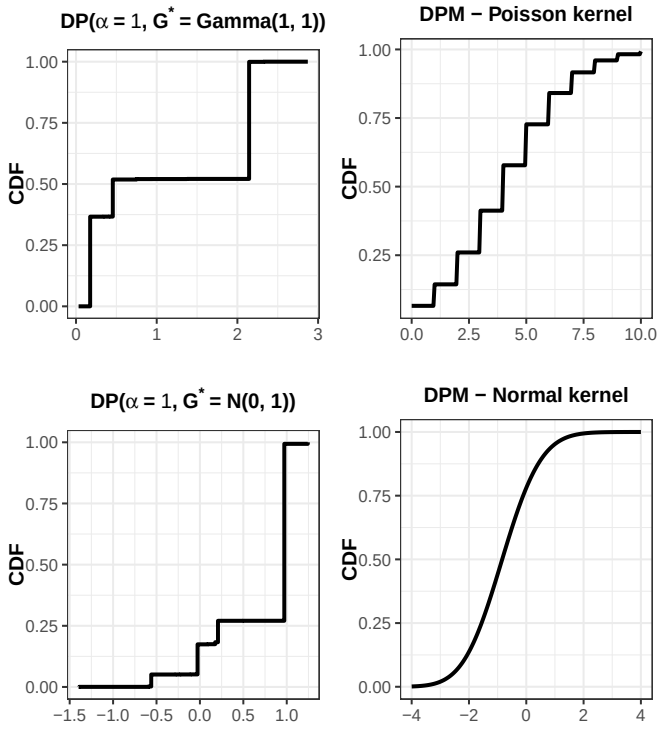

Figure 2.4: Top: Draw from a Dirichlet process with $G^* = \Gamma(1, 1)$ and $\alpha = 1$ (left) and the corresponding DPM distribution using a Poisson kernel (right). Bottom: Draw from a Dirichlet process with $G^* = \mathrm{N}(0, 1)$ and $\alpha = 1$ (left) and the corresponding DPM distribution mixing only over μ using a normal kernel with $\sigma = 0.5$.

The weights ω_k, now follow a truncated stick-breaking construction. The inputs of the weights are distributed according to a beta distribution, i.e., $v_1, \ldots, v_{K-1} \overset{\text{iid}}{\sim} \text{Beta}(1, \alpha)$, and to ensure they add up to one, we set $v_K = 1$. Ishwaran and Zarepour (2000) showed that for higher order weights in the stick breaking representation $\mathbb{E}\left(\sum_{k=K}^{\infty} \omega_k\right) = \left(\frac{\alpha}{\alpha+1}\right)^{K-1}$. For instance, if we set $K = 20$ and $\alpha = 1$, then $\mathbb{E}\left(\sum_{k=K}^{\infty} \omega_k\right) < 10^{-6}$. Further, Ishwaran and James (2001) proved that G^K converges to a DP, when $K \to \infty$.

We shall note that K is not the exact number of components expected to be observed, but instead an upper bound on it, because some of the components may be unoccupied. This representation provides a direct practical implementation of the DPM model, because it reduces an infinite dimensional problem to a finite one. Thus, traditional sampling methods used for mixture models can be used.

2.2.4 The induced conditional density approach

In many practical applications, interest relies in investigating the effect of a set of predictors or covariates on the test outcomes. In other words, we are interested in studying if and how the distribution of the test outcomes changes as a function of the predictors. Thus, our main goal is to model the distribution of the test outcomes given the covariates. For this end, several approaches have been proposed in the literature both frequentist and Bayesian as well as parametric and nonparametric. However, we will restrict ourselves to fully Bayesian nonparametric regression methods.

The induced conditional density approach was firstly described by Müller et al. (1996). The authors modelled jointly a continuous response and continuous covariates using a multivariate normal distribution as the kernel of the DPM and then, they induced the conditional distribution of the response by simply using the properties of the multivariate normal distribution. Although they looked only at a particular function of interest, namely, the mean function (regression function), $m(\boldsymbol{x}) = \mathbb{E}(y \mid \boldsymbol{x})$, other numerical features, such as, the quantile or variance function are available for free. Another advantage of this approach is that all the heterogeneity and functional forms are automatically accommodated in a single framework (Rossi, 2014, p. 93). Let $\boldsymbol{u}_i = (y_i, \boldsymbol{x}_i)$ be the complete data. The model is defined as

$$\begin{aligned} \boldsymbol{u}_i \mid G &\sim \int \phi(\boldsymbol{u}_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) dG(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad i = 1, \ldots, n, \\ G \mid \alpha, G^* &\sim \mathrm{DP}(\alpha, G^*). \end{aligned}$$

Note that this model induces a mixture of normal distributions, where the weights depend on the covariates, that is,

$$\begin{aligned} f\left(y \mid \boldsymbol{x}\right) &= \frac{f(y, \boldsymbol{x})}{f(\boldsymbol{x})} \\ &= \sum_{k=1}^{\infty} q_k(\boldsymbol{x}) \phi\left(y \mid \tilde{\mu}_k, \tilde{\sigma}_k^2\right), \end{aligned}$$

where

$$q_k(\boldsymbol{x}) = \frac{\omega_k \phi\left(\boldsymbol{x} \mid \boldsymbol{\mu}_k^x, \boldsymbol{\Sigma}_k^{xx}\right)}{\sum_{\ell=1}^{\infty} \omega_\ell \phi\left(\boldsymbol{x} \mid \boldsymbol{\mu}_\ell^x, \boldsymbol{\Sigma}_\ell^{xx}\right)},$$

$\tilde{\mu}_k = \mu_k^y + \boldsymbol{\Sigma}_k^{yx}(\boldsymbol{\Sigma}_k^{xx})^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k^x)$ and $\tilde{\sigma}_k^2 = \Sigma_k^{yy} - \boldsymbol{\Sigma}_k^{yx}(\boldsymbol{\Sigma}_k^{xx})^{-1}\boldsymbol{\Sigma}_k^{xy}$, for $k \geq 1$, with $\boldsymbol{\mu}_k^x$ and $\boldsymbol{\Sigma}_k^{xx}$ the mean vector and covariance matrix induced by the joint $\mathrm{N}(y, \boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ distribution, where

$$\boldsymbol{\mu}_k = \begin{pmatrix} \mu_k^y \\ \boldsymbol{\mu}_k^x \end{pmatrix} \text{ and } \boldsymbol{\Sigma}_k = \begin{pmatrix} \Sigma_k^{yy} & \boldsymbol{\Sigma}_k^{yx} \\ \boldsymbol{\Sigma}_k^{xy} & \boldsymbol{\Sigma}_k^{xx} \end{pmatrix}.$$

Possible extensions are available to handle both continuous, $\boldsymbol{x}^c$, and categorical covariates, $\boldsymbol{x}^d$, considering appropriate kernels and assuming a multiplicative structure of the joint distribution. Thus, one possible model is

$$\boldsymbol{u}_i \mid G \overset{\text{iid}}{\sim} \int k(\boldsymbol{x}_i^d \mid \boldsymbol{\theta}_1) k(y_i, \boldsymbol{x}_i^c \mid \boldsymbol{\theta}_2) dG(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2),$$

where $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are the corresponding parameters of the discrete and continuous kernel, respectively. A discrete kernel might be, for example, a Bernoulli kernel. Another possibility for priors over collections of conditional density functions is the dependent Dirichlet process introduced below.

2.2.5 Dependent Dirichlet processes

Under the nonparametric framework, the parameter of interest in a regression model is now a set of random probability measures at location (covariate value) $\boldsymbol{x}$, say $\mathcal{G} = \{G_{\boldsymbol{x}} \mid \boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^p\}$, where $G_{\boldsymbol{x}}$ is defined on the response sample space $\mathcal{Y}$. From a Bayesian viewpoint, we shall define a prior for $\mathcal{G}$. Common BNP priors for such end are based on extensions of (2.4), by replacing G with $G_{\boldsymbol{x}}$ (Quintana et al., 2020).

One of these first extensions was the dependent Dirichlet process (DDP) introduced by MacEachern (1999, 2000). The key idea is based on replacing the weights and/or locations of the Sethuraman's stick-breaking representation by an appropriate stochastic process on $\mathcal{X}$, such that the random measures are marginally DP distributed. It has the following representation

$$G_{\boldsymbol{x}}(\cdot) = \sum_{k=1}^{\infty} \omega_k(\boldsymbol{x}) \delta_{\boldsymbol{\theta}_k(\boldsymbol{x})}(\cdot), \quad \omega_k(\boldsymbol{x}) = \begin{cases} v_k(\boldsymbol{x}), & \text{if} \quad k = 1 \\ v_k(\boldsymbol{x}) \prod_{t<k} (1 - v_t(\boldsymbol{x})), & \text{if} \quad k \geq 2 \end{cases},$$

where $v_k(\boldsymbol{x})$ are stochastic processes defined in $[0,1]$ with index set $\mathcal{X}$ whose marginal distributions are $\text{Beta}(1,\alpha_{\boldsymbol{x}})$ and $\boldsymbol{\theta}_k(\boldsymbol{x})$ are mutually independent stochastic processes with index set $\mathcal{X}$ and marginal distributions $G^*_{\boldsymbol{x}}$, for $k \geq 1$.

A canonical DDP construction was provided by MacEachern (1999, 2000). It is based on transforming two independent stochastic processes, via the inverse transformation method (Devroye, 1986), to induce the desired marginal distributions for $v_k(\boldsymbol{x})$ and $\boldsymbol{\theta}_k(\boldsymbol{x})$. However, its practical implementation is difficult, because it requires the specification of its different components. This has motivated variations where the dependence is introduced only through the weights or the atoms. In the latter case, MacEachern considered the so-called single-weights DDP model, defined as

$$G_{\boldsymbol{x}}(\cdot) = \sum_{k=1}^{\infty} \omega_k \delta_{\boldsymbol{\theta}(\boldsymbol{x})}(\cdot), \quad \omega_k = v_k \prod_{t<k}(1-v_t),$$

where $v_k \overset{\text{iid}}{\sim} \text{Beta}(1,\alpha)$ and $\boldsymbol{\theta}_k(\boldsymbol{x})$ are independent stochastic processes with index set $\mathcal{X}$ and marginal distribution $G^*_{\boldsymbol{x}}$. Due to the fact that posterior inference for this model is straightforward using sampling algorithms similar to those for the DP, this is one of the most popular forms of DDPs.

The ANOVA-DDP (De Iorio et al., 2004) and the linear DDP (LDDP) models are popular examples of single-weights DDP models. As its name suggests, the ANOVA-DDP is based on an ANOVA-type regression model, where the atoms depend on a vector of categorical covariates. Specifically, $\boldsymbol{\theta}_k(\boldsymbol{x}) = \lambda_k' d_{\boldsymbol{x}}$, where $\lambda_k \overset{\text{iid}}{\sim} G^*$ and $d_{\boldsymbol{x}}$ is the design vector of the corresponding covariates. On the other hand, the LDDP involves a linear combination of the covariates. The dependence is explicitly introduced through the atoms, such that, the induced marginal distribution is given by

$$y \sim \sum_{k=1}^{\infty} \omega_k \text{N}(y \mid \lambda_k' \boldsymbol{x}, \sigma^2),$$

the weights follow a stick-breaking construction and $\boldsymbol{x}$ is the usual vector of covariates (continuous and/or categorical). An advantage of the joint modelling approach over the DDP is that the former allows the weights to vary smoothly over the covariates values and there is no need to specify how the covariates affect the test outcomes.

As we have seen so far, we only have highlighted some of the several extensions of the DPM model. However, an excellent and comprehensive review can be found in Quintana et al. (2020).

2.3 Model comparison criteria

In many practical situations, a single model is not solely available to fit the data, usually there is a set of candidate models. Then, a natural question that arises is how to know which of them provide the best fit. However, this does not mean that the model is good per se. This may be only for the sake of comparison or to seek further improvement directions, such as a different prior specification. For instance, we may want to compare the fit of a DPM to the fit of a simpler modelling approach, such as a normal linear model. For this end, we will employ three different model comparison criteria. We will begin our discussion introducing the log pseudomarginal likelihood (LPML) (Geisser and Eddy, 1979; Gelfand and Dey, 1994). Then, we will describe the widely applicable information criterion (WAIC) (Gelman et al., 2014; Watanabe, 2010) and a modified version of the popular deviance information criterion (DIC) (Spiegelhalter et al., 2002), specially tailored to mixture models that was proposed by Celeux et al. (2006). Finally, we will present a recent alternative proposed by Cheng et al. (2019) to measure the uncertainty around the comparison of multiple models, based on the Bayesian bootstrap and the conditional predictive ordinates.

2.3.1 Log pseudomarginal likelihood (LPML)

One way to assess the performance of a model is through the accuracy of its predictions (Gelman et al., 2013, p. 166). The idea is to replace the marginal distribution by some predictive version of it, namely the pseudomarginal likelihood. The log pseudomarginal likelihood (LPML) (Geisser and Eddy, 1979) is based on the conditional predictive ordinates (CPO). For a set of observations $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$ of a response y with covariates $\boldsymbol{x}$, conditional density function $f(\cdot \mid \boldsymbol{x})$ and parameter vector $\boldsymbol{\theta}$, the i–th CPO is given by

$$\text{CPO}_i = f\left(y_i \mid \boldsymbol{y}^{(-i)}, \boldsymbol{x}^{(-i)}\right) = \int_{\boldsymbol{\Theta}} f(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}) p\left(\boldsymbol{\theta} \mid \boldsymbol{y}^{(-i)}, \boldsymbol{x}^{(-i)}\right) d\boldsymbol{\theta},$$

where $p\left(\boldsymbol{\theta} \mid \boldsymbol{y}^{(-i)}, \boldsymbol{x}^{(-i)}\right)$ is the posterior density of $\boldsymbol{\theta}$, both based on all the observations except the i–th. This can also be defined in the unconditional context simply dropping the covariates from the notation. Gelfand and Dey (1994) showed that, the CPO_i can be easily computed from an MCMC

sample of size S, as follows

$$\text{CPO}_i^{-1} \approx \frac{1}{S}\sum_{s=1}^{S}\frac{1}{f^{(s)}\left(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}^{(s)}\right)}, \tag{2.6}$$

where $f^{(s)}\left(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}^{(s)}\right)$ is a single iteration of the conditional density. Given the above, the log pseudomarginal likelihood is defined as

$$\text{LPML} = \sum_{i=1}^{n}\log(\text{CPO}_i). \tag{2.7}$$

In this sense, we would choose the model that maximizes the LPML, i.e., the higher the LPML, the better the model is. In some cases, the weights $w_{is} = \frac{1}{f^{(s)}\left(y_i|\boldsymbol{x}_i,\boldsymbol{\theta}^{(s)}\right)}$ may have infinity variance (Epifani et al., 2008). As an alternative, Gelman et al. (2014) suggest replacing such weights with $\widetilde{w}_{is} = \min\{w_{is}, \bar{w}_i\sqrt{S}\}$, where $\bar{w}_i = \frac{1}{S}\sum_{s=1}^{S} w_{is}$. Hence, the stabilized version of the CPO_i is given by

$$\text{CPO}_i^{-1} = \frac{\sum_{s=1}^{S}\widetilde{w}_{is}f^{(s)}\left(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}^{(s)}\right)}{\sum_{s=1}^{S}\widetilde{w}_{is}}. \tag{2.8}$$

2.3.2 Widely applicable information criterion (WAIC)

The widely applicable information criterion (WAIC) is a fully Bayesian approach comprised of the model fit and complexity. The idea is to use the expected log predictive density (elpd) as a measure of the overall fit. However, since we do not know the parameters $\boldsymbol{\theta}$, we usually estimate the log predictive density through the log pointwise predictive density (lppd) which is known to be a biased estimator (Gelman et al., 2013, p. 169). Different approaches have been proposed to correct this bias, we will follow one based upon the posterior variance, that is,

$$\begin{aligned}
\text{elpd} &= \text{lppd} - p_{\text{WAIC}},\\
\text{lppd} &= \sum_{i=1}^{n}\log\left[\frac{1}{S}\sum_{s=1}^{S}f^{(s)}\left(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}^{(s)}\right)\right], \qquad (2.9)\\
p_{\text{WAIC}} &= \sum_{i=1}^{n}\frac{1}{S-1}\sum_{s=1}^{S}\left\{\log\left[f^{(s)}\left(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}^{(s)}\right)\right] - \frac{1}{S}\sum_{s=1}^{S}\log\left[f^{(s)}\left(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}^{(s)}\right)\right]\right\}^2. \qquad (2.10)
\end{aligned}$$

Finally, from the expressions described above, Gelman et al. (2014) propose to use the deviance scale to make WAIC comparable with other measures such as the DIC, i.e.,

$$\text{WAIC} = -2\text{elpd}.$$

The smaller the WAIC is, the better the fit of the model under consideration.

2.3.3 Deviance information criterion (DIC)

The deviance information criterion (DIC) is a popular alternative for model assessment and model comparison. However, it has not been exempt of criticism. As Celeux et al. (2006) mentioned, there have been some authors pointing out possible inconsistencies in the definition of the DIC, particularly, for mixture models (DeIorio and Robert, 2002). Under the model comparison framework, the deviance is defined as

$$D(\boldsymbol{\theta}) = -2\log\left[f(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})\right].$$

Spiegelhalter et al. (2002) define the DIC as follows

$$\begin{aligned}
\text{DIC} &= \overline{D(\boldsymbol{\theta})} + p_D \\
&= 2\overline{D(\boldsymbol{\theta})} - D(\tilde{\boldsymbol{\theta}}) \\
&= -4\mathbb{E}_{\boldsymbol{\theta}}\left\{\log\left[f(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})\right] \mid \boldsymbol{y}\right\} + 2\log\left[f(\boldsymbol{y} \mid \boldsymbol{x}, \tilde{\boldsymbol{\theta}})\right],
\end{aligned}$$

where $\overline{D(\theta)}$ is the posterior mean of the deviance, p_D is the effective dimension, and $\tilde{\boldsymbol{\theta}}$ is an estimate of $\boldsymbol{\theta}$, often $\tilde{\boldsymbol{\theta}} = \mathbb{E}(\boldsymbol{\theta} \mid \boldsymbol{y})$. As it is well-known, mixture models present non-identifiability, thus the estimate of $\boldsymbol{\theta}$ makes no sense. However, in most cases (as in ours) the inferential focus is on the density itself. For this reason, Celeux et al. (2006) mention that a more natural choice for $D(\tilde{\boldsymbol{\theta}})$ is to select an estimator, $\hat{f}(\boldsymbol{y})$, of the density, $f(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})$, since it is invariant under permutations of the component labels. One possible choice might be $\mathbb{E}_{\boldsymbol{\theta}}[f(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}) \mid \boldsymbol{y}]$. Hence, they define the DIC_3 as

$$\text{DIC}_3 = -4\mathbb{E}_{\boldsymbol{\theta}}\left\{\log\left[f(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})\right] \mid \boldsymbol{y}\right\} + 2\log\left[\hat{f}(\boldsymbol{y})\right],$$

where $\hat{f}(\boldsymbol{y}) = \prod_{i=1}^{n} \hat{f}(y_i) \approx \prod_{i=1}^{n} \mathbb{E}_{\boldsymbol{\theta}}[f(y_i \mid \boldsymbol{x}, \boldsymbol{\theta}) \mid \boldsymbol{y}]$. The DIC_3 can be easily approximated using the MCMC iterations, that is,

$$\text{DIC}_3 \approx -4\frac{1}{S}\sum_{s=1}^{S}\sum_{i=1}^{n} \log f^{(s)}\left(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}^{(s)}\right) + 2\sum_{i=1}^{n} \log\left\{\frac{1}{S}\sum_{s=1}^{S} f^{(s)}\left(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}^{(s)}\right)\right\}. \tag{2.11}$$

Therefore, the model that provides the best fit is the one with the smallest DIC_3 value.

2.3.4 Posterior rank probability

None of the model comparison criteria described thus far provides a measure about the uncertainty in the choice of one model over another. Recently, Cheng et al. (2019) have proposed an approach combining the CPOs with the Bayesian bootstrap to measure such uncertainty. Let M_1 and M_2 be two competing models with $f(y_i \mid \boldsymbol{y}^{(-i)}, \boldsymbol{x}^{(-i)}, M_1)$ and $f(y_i \mid \boldsymbol{y}^{(-i)}, \boldsymbol{x}^{(-i)}, M_2)$ the corresponding CPO for the i–th observation under each model, respectively. Then,

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\log\left[f(y_i \mid \boldsymbol{y}^{(-i)}, \boldsymbol{x}^{(-i)}, M_1)\right] - \log\left[f(y_i \mid \boldsymbol{y}^{(-i)}, \boldsymbol{x}^{(-i)}, M_2)\right]\right\},$$

measures the difference of the average prediction of both models. For example, if this difference is negative, then M_2 is preferred, otherwise M_1 is better than M_2. However, this difference does not provide any probabilistic quantification about how much better is one model compared to the other. Thus, Cheng et al. (2019) proposed to approximate the probability of a model being better than another using the BB. The idea is to generate, say B, weights $w_{i,b} \sim \text{Dirichlet}(n; 1, \ldots, 1)$ for each observation and then summarize them to obtain the probability of M_1 being better than M_2, i.e.,

$$\mathbb{P}(M_1 \text{ better than } M_2) = \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}\left\{\frac{1}{n}\sum_{j=1}^{n} w_{i,b} \log\left(\frac{f(y_i \mid \boldsymbol{y}^{(-i)}, \boldsymbol{x}^{(-i)}, M_1)}{f(y_i \mid \boldsymbol{y}^{(-i)}, \boldsymbol{x}^{(-i)}, M_2)}\right) > 0\right\}.$$

Further, one can compare multiple models choosing the one with the highest rank, on each of the B bootstrap samples, by sorting the models with

$$\frac{1}{n}\sum_{i=1}^{n} w_{i,b} \log f(y_i \mid \boldsymbol{y}^{(-i)}, \boldsymbol{x}^{(-i)}, M_m). \tag{2.12}$$

Then, counting the number of times N_m, each model was the best and computing its probability as N_m/B, this is the so-called posterior rank probability.

Chapter 3

Bayesian nonparametric inference for the coefficient of overlap

Biomedical data often exhibit nonstandard features, such as multimodality, skewness, excess of kurtosis, among others. Usual parametric approaches are too restrictive to represent accurately these features and can lead to misleading results about the discriminatory capacity of a test. The overlap coefficient has been recognized as an alternative summary measure of diagnostic accuracy (Samawi et al. 2017, Wang and Tian 2017). We present a Bayesian nonparametric approach for conducting inference about the coefficient of overlap based on Dirichlet process mixtures (DPM) of normal distributions. Mixtures of normals are our preferred choice because they are well-known to approximate any smooth density in the real line (Lo, 1984; Rossi, 2014, p. 5). We propose two estimators to conduct robust inference about the coefficient of overlap. The first estimator is based on numerical integration and the second estimator makes use of the Bayesian bootstrap, both use DPMs as well. We assess the performance of our proposed methods through multiple simulation studies. An application to real data concerning the search for biomarkers for ovarian cancer is provided as well.

3.1 Introduction

Investigating the discriminatory ability of a diagnostic test is crucial for an accurate clinical diagnosis. Statistical methods based on the receiver operating characteristic (ROC) curve are popular and are the most widely used tools to evaluate the accuracy of a continuous diagnostic test. Pepe (2003, Chapter 4) and Zhou et al. (2011, Chapter 2, section 3) are excellent resources for a frequentist overview of ROC curve estimation. However, existing work from a Bayesian perspective is limited. One of the earliest works on Bayesian methods for estimating the ROC curve was introduced by Erkanli et al. (2006), who proposed a semi-parametric Bayesian approach, based on Dirichlet process mixtures (DPM). Also, it is fair to mention the work by Gu et al. (2008), who proposed an estimator for the ROC curve and associated measures based on the Bayesian bootstrap and Branscum et al. (2008), whose work is based on Polya trees. An excellent review on ROC curve estimation based on DPM models and Bayesian bootstrap is also provided by Inácio de Carvalho et al. (2015).

Although flexible and robust approaches for the ROC curve have been proposed, the traditional associated diagnostic summary measures, such as the area under the ROC curve (AUC) and the Youden index (YI) are not flawless and present some drawbacks. Perhaps, one of the most notable is the assumption that larger test outcome values are more indicative of disease. If this assumption is not met, the decision rule can always be reversed. However, this process may be cumbersome at times.

Recently, the coefficient of overlap (OVL) has been proposed as an alternative measure of diagnostic accuracy because of its advantages over these summary measures. Some of the existing approaches to conduct inference about the coefficient of overlap are based on parametric alternatives which require either transform the data, the assumption of normality or both (Wang and Tian, 2017; Reiser and Faraggi, 1999). However, in many practical situations, the test outcomes may present a complex structure, such as asymmetry, multiple modes or other nonstandard features, which can be difficult to capture with a single parametric model. For instance, within the diseased population there might be subpopulations which differentiate several stages of the same disease. Conversely, in distribution-free approaches discussed, among others, in Schmid and Schmidt (2006), Ridout and Linkie (2009), Clemons and Bradley Jr. (2000) and Pastore and Calcagnì (2019), parametric assumptions are relaxed

by using kernel techniques. However, only approximate confidence intervals are available. In general, this is achieved resorting to other methods, such as the bootstrap (Efron, 1979). This motivates us to use a class of models that allows to represent a wide variety of density shapes, such as mixture models. In addition, working from a Bayesian perspective point and interval estimates are obtained into a single integrated framework.

Bayesian research on the coefficient of overlap is scarce. Recently, Núñez-Antonio et al. (2018) proposed a Bayesian nonparametric framework to estimate the OVL. However, their approach concerns the study of interactions among species in a particular region. Given the nature of their application, circular distributions are be used. For this end, the authors proposed a Dirichlet process mixture model based on projected normals. In this chapter, we propose a flexible framework to estimate the coefficient of overlap based on Dirichlet process mixtures of normal distributions (Escobar and West, 1995; Richardson and Green, 1997). In a mixture model, we assume that the data come from one or many subpopulations or clusters and each group can be represented by a single parametric model. Hence, the complete data can be characterized as a weighted sum of distributions. We propose two estimators based on (1.2) and (1.3). For the first estimator (the one in (1.2)), the integral is replaced by a numerical integration method. More precisely, we use the trapezoidal rule for this end. We call this estimator the "plain vanilla" estimator. In the second estimator (the one in (1.3)), the outer probabilities are computed assuming a Dirichlet process (DP) and letting the concentration parameter of the DP approaches zero ($\alpha \to 0$). This limiting case is also known as the Bayesian bootstrap (Rubin, 1981). For this reason, we call this estimator the DPM-BB estimator.

3.2 Methods

The expressions in (1.2) and (1.3) suggest a two-step modelling procedure. The first step involves estimating the density functions of the test outcomes in each population. In the second step, for the estimator based on (1.2), the integral can be computed via numerical integration. While the estimator based on (1.3) requires estimating the outer probabilities.

3.2.1 Density estimation

Let $Y_{\bar{D}}$ and Y_D be two continuous random variables representing the test outcomes in the nondiseased and diseased population, with density functions given by $f_{\bar{D}}(\cdot)$ and $f_D(\cdot)$, respectively. In what follows, let $\{y_{\bar{D}i}\}_{i=1}^{n_{\bar{D}}}$ and $\{y_{Dj}\}_{j=1}^{n_D}$ be two independent random samples of test outcomes of size $n_{\bar{D}}$ and n_D from the nondiseased and diseased populations, respectively, with

$$y_{\bar{D}1}, \ldots, y_{\bar{D}n_{\bar{D}}} \mid f_{\bar{D}} \overset{\text{iid}}{\sim} f_{\bar{D}}, \quad \text{and} \quad y_{D1}, \ldots, y_{Dn_D} \mid f_D \overset{\text{iid}}{\sim} f_D.$$

Traditional approaches usually assume a normal distribution for both f_D and $f_{\bar{D}}$, possibly after some transformation of Y_D and $Y_{\bar{D}}$ (e.g., the logarithm or, more generally, a Box-Cox transformation). However, such model would be unsuitable for test outcomes data exhibiting asymmetry, multiple modes or excess of kurtosis. Mixtures of normal distributions can be used to represent a wide variety of density shapes and therefore they are our preferred choice. Although mixture models are widely used in practice, special care must be taken in order to overcome overfitting (a detailed discussion can be found in Section 2.2.1). An equally flexible alternative is to use Dirichlet process mixtures of normal distributions, which have shown to accurately approximate any smooth density on the real line. In the remainder, we will describe our modelling approach for the diseased group (the one for the nondiseased group follows analogously). The density function of the diseased population is given by

$$f_D(y_{Dj}) = \int \phi(y_{Dj} \mid \mu_D, \sigma_D^2) \mathrm{d}G_D(\mu_D, \sigma_D^2), \quad G_D \sim \mathrm{DP}(\alpha_D, G_D^*), \tag{3.1}$$

where $\phi(\cdot \mid \mu_D, \sigma_D^2)$ denotes the probability density function of a normal distribution with mean μ_D and variance σ_D^2. The mixing distribution G_D follows a Dirichlet process (Ferguson, 1973) with precision parameter $\alpha_D > 0$ and baseline distribution G_D^*. Since $\mathbb{E}\left[G_D(\mu_D, \sigma_D^2)\right] = G_D^*(\mu_D, \sigma_D^2)$, the designation of G_D^* as baseline distribution can be regarded as our best guess for G_D, with α_D controlling how certain we are about such guess. Larger values of α_D imply realizations of G_D that are closer to G_D^*.

The constructive definition of the DP given by Sethuraman (1994), is unarguably its most popular

representation and it allows to write G_D as an infinite sum of weighted point masses as follows

$$G_D(\cdot) = \sum_{l=1}^{\infty} \omega_{Dl} \delta_{(}\mu_{Dl}, \sigma_{Dl}^2)(\cdot), \quad \omega_{Dl} = \begin{cases} v_{Dl}, & \text{if} \quad l = 1 \\ v_{Dl} \prod_{t<l} (1 - v_{Dt}), & \text{if} \quad l \geq 2 \end{cases}, \tag{3.2}$$

where the atoms $(\mu_{Dl}, \sigma_{Dl}^2) \overset{\text{iid}}{\sim} G_D^*(\mu_D, \sigma_D^2)$ and $v_{Dl} \overset{\text{iid}}{\sim} \text{Beta}(1, \alpha_D)$ are mutually independent for $l \geq 1$. This allows us to write the density in (3.1) as an infinite mixture of normal distributions, that is,

$$f_D(y_{Dj}) = \sum_{l=1}^{\infty} \omega_{Dl} \phi(y_{Dj} \mid \mu_{Dl}, \sigma_{Dl}^2). \tag{3.3}$$

The probability mass on the atoms $(\mu_{Dl}, \sigma_{Dl}^2)$ decreases geometrically as the index becomes larger, thus the number of mixture components is determined automatically. Leveraging conjugacy properties, we set $G_D^*(\mu_D, \sigma_D^2) \equiv \text{N}(\mu_D \mid a_{\mu_D}, b_{\mu_D}^2)\text{IG}(\sigma_D^2 \mid a_{\sigma_D^2}, b_{\sigma_D^2})$, where $\text{IG}(a, b)$ denotes an inverse-gamma distribution with shape parameter a and scale parameter b. Despite we can complete our model by placing a gamma prior on the DP precision parameter, $\alpha_D \sim \Gamma(a_{\alpha_D}, b_{\alpha_D})$, we have fixed it ($\alpha_D = 1$) in the simulation study and in the application. In what follows, we present the model in its full generality where α_D is allowed to be random. The precision parameter has a direct relationship with the number of occupied mixture components, say L_D^*. For moderate to large sample sizes, Liu (1996) has shown that the conditional prior mean and variance of the number of occupied components, given a fixed α_D value and sample size n_D, are

$$\mathbb{E}(L_D^* \mid \alpha_D) = \alpha_D \log\left(\frac{\alpha_D + n_D}{\alpha_D}\right), \quad \text{Var}(L_D^* \mid \alpha_D) = \alpha_D \left\{\log\left(\frac{\alpha_D + n_D}{\alpha_D}\right) - 1\right\}. \tag{3.4}$$

These expressions can be averaged over the $\Gamma(a_{\alpha_D}, b_{\alpha_D})$ prior for α_D to obtain $\mathbb{E}(L_D^*)$ and $\text{Var}(L_D^*)$, thus selecting the values of a_{α_D} and b_{α_D} to agree with the prior guesses. Further, to allow us to easily simulate from the posterior distribution, we employ a truncation of the stick-breaking construction proposed by Ishwaran and Zarepour (2000), where the infinite sum in (3.2) is replaced by a finite sum of the first, say L_D, terms. Additionally, the authors showed that $\mathbb{E}\left(\sum_{l=L_D}^{\infty} \omega_{Dl}\right) = \left(\frac{\alpha_D}{\alpha_D+1}\right)^{L_D-1}$. For instance, if we set $L_D = 20$ and $\alpha_D = 1$, then $\mathbb{E}\left(\sum_{l=L_D}^{\infty} \omega_{Dl}\right) < 10^{-6}$. Hence, we can rewrite (3.2) as

$$G_D(\cdot) = \sum_{l=1}^{L_D} \omega_{Dl} \delta_{(\mu_{Dl}, \sigma_{Dl}^2)}(\cdot),$$

where the weights ω_{Dl}, now follow a truncated stick-breaking construction, that is, $\omega_{D1} = v_{D1}$, $\omega_{Dl} = v_{Dl}\prod_{t<l}(1 - v_{Dt})$, for $l = 2, \ldots, L_D$. The inputs of the weights are distributed according to a beta distribution, i.e., $v_{D1}, \ldots, v_{DL_D-1} \overset{\text{iid}}{\sim} \text{Beta}(1, \alpha_D)$ and $v_{DL_D} = 1$ to ensure the weights add up to one. Finally, the density function in (3.3) can be expressed as follows

$$f_D(\cdot) = \sum_{l=1}^{L_D} \omega_{Dl}\phi(\cdot \mid \mu_{Dl}, \sigma^2_{Dl}). \tag{3.5}$$

Note that in this case L_D does not represent the number of components but an upper bound on such number instead, since some of the components may be unoccupied. Also, note that the randomness of $\boldsymbol{\omega}_D = (\omega_{D1}, \ldots, \omega_{DL_D})$ relies on the fact that $\boldsymbol{v}_D = (v_{D1}, \ldots, v_{DL_D})$ is random (beta variates).

3.2.2 Posterior inference

We introduce latent variables z_{Dj} to identify the mixture component to which the j–th diseased individual is allocated to, for $j = 1, \ldots, n_D$. Therefore, $z_{Dj} = l$ means that the individual is allocated to the normal component with mean μ_{Dl} and variance σ^2_{Dl}. This data augmentation allow us to rewrite the model hierarchically as

$$y_{Dj} \mid \boldsymbol{\mu}_D, \boldsymbol{\sigma}^2_D, z_{Dj} \overset{\text{ind.}}{\sim} \text{N}(\mu_{Dz_{Dj}}, \sigma^2_{Dz_{Dj}}), \quad j = 1, \ldots, n_D, \tag{3.6}$$

$$\mathbb{P}(z_{Dj} = l \mid \boldsymbol{v}_D) = \omega_{Dl}, \quad l = 1, \ldots, L_D, \tag{3.7}$$

$$v_{Dl} \mid \alpha_D \overset{\text{iid}}{\sim} \text{Beta}(1, \alpha_D), \quad l = 1, \ldots, L_D - 1, \tag{3.8}$$

$$(\mu_{Dl}, \sigma^2_{Dl}) \overset{\text{iid}}{\sim} \text{N}(a_{\mu_D}, b^2_{\mu_D})\text{IG}(a_{\sigma^2_D}, b_{\sigma^2_D}), \quad l = 1, \ldots, L_D, \tag{3.9}$$

$$\alpha_D \mid a_{\alpha_D}, b_{\alpha_D} \sim \Gamma(a_{\alpha_D}, b_{\alpha_D}). \tag{3.10}$$

where $\boldsymbol{\mu}_D = (\mu_{D1}, \ldots, \mu_{DL_D})$, $\boldsymbol{\sigma}^2_D = (\sigma^2_{D1}, \ldots, \sigma^2_{DL_D})$. The full conditional distributions for all parameters are available in closed form, thus allowing for direct posterior simulation through a simple Gibbs sampler.

Full conditional distributions

Using Equations (3.6)–(3.10) and denoting by $\boldsymbol{\theta}_D = (\boldsymbol{v}_D, \boldsymbol{\mu}_D, \boldsymbol{\sigma}^2_D)$ we can express the likelihood as

follows

$$
\begin{aligned}
L(\boldsymbol{\theta}_D; \boldsymbol{y}_D, \boldsymbol{z}_D) &= \prod_{j=1}^{n_D} \omega_{Dz_{Dj}} \phi\left(y_{Dj} \mid \mu_{Dz_{Dj}}, \sigma^2_{Dz_{Dj}}\right) \\
&= \prod_{l=1}^{L_D} \prod_{j:z_{Dj}=l} \omega_{Dl} \phi\left(y_{Dj} \mid \mu_{Dl}, \sigma^2_{Dl}\right) \\
&= \prod_{l=1}^{L_D} \omega_{Dl}^{n_{Dl}} \prod_{j:z_{Dj}=l} \phi\left(y_{Dj} \mid \mu_{Dl}, \sigma^2_{Dl}\right),
\end{aligned}
$$

where $n_{Dl} = \sum_{j=1}^{n_D} \mathbb{I}\{z_{Dj} = l\}$. Thus, the joint posterior distribution is given by

$$
\begin{aligned}
p(\boldsymbol{z}_D, \boldsymbol{\theta}_D, \alpha_D \mid \boldsymbol{y}_D) &\propto L(\boldsymbol{\theta}_D; \boldsymbol{y}_D, \boldsymbol{z}_D) p(\boldsymbol{\theta}_D) p(\alpha_D) \\
&= L(\boldsymbol{\theta}_D; \boldsymbol{y}_D, \boldsymbol{z}_D) \prod_{l=1}^{L_D} p(\mu_{Dl}) p(\sigma^2_{Dl}) \prod_{l=1}^{L_D-1} p(v_{Dl} \mid \alpha_D) p(\alpha_D) \\
&\propto \prod_{l=1}^{L_D} \left\{ v_{Dl} \prod_{t<l} (1 - v_{Dt}) \right\}^{n_{Dl}} \prod_{j:z_{Dj}=l} \phi\left(y_{Dj} \mid \mu_{Dl}, \sigma^2_{Dl}\right) \\
&\quad \times \prod_{l=1}^{L_D} \exp\left\{ -\frac{1}{2b^2_{\mu_D}} \left(\mu_{Dl} - a_{\mu_D}\right)^2 \right\} \\
&\quad \times \prod_{l=1}^{L_D} \left(\sigma^2_{Dl}\right)^{-a_{\sigma^2_D}-1} \exp\left\{ -\frac{b_{\sigma^2_D}}{\sigma^2_{Dl}} \right\} \\
&\quad \times \prod_{l=1}^{L_D-1} \alpha_D (1 - v_{Dl})^{\alpha_D - 1} \\
&\quad \times \alpha_D^{a_{\alpha_D}-1} \exp\left\{ -b_{\alpha_D} \alpha_D \right\}.
\end{aligned}
$$

First, we derive the full conditional for the mean of each component, that is,

$$
\begin{aligned}
p(\mu_{Dl} \mid \text{else}) &\propto L(\boldsymbol{\theta}_D; \boldsymbol{y}_D, \boldsymbol{z}_D) p(\mu_{Dl}) \\
&\propto \exp\left\{ -\frac{1}{2\sigma^2_{Dl}} \sum_{j:z_{Dj}=l} \left(y_{Dj} - \mu_{Dl}\right)^2 \right\} \times \exp\left\{ -\frac{1}{2b^2_{\mu_D}} \mu^2_{Dl} \right\}.
\end{aligned}
$$

Adding the terms in the exponents and ignoring the $-1/2$ for the moment, we can work within the

exponential function

$$\frac{1}{\sigma^2_{Dl}}\left\{\sum_{j:z_{Dj}=l} y^2_{Dj} - 2\sum_{j:z_{Dj}=l} y_{Dj}\mu_{Dl} + n_{Dl}\mu^2_{Dl}\right\} + \frac{1}{b^2_\mu}\mu^2_{Dl} \propto \left(\frac{n_{Dl}}{\sigma^2_{Dl}} + \frac{1}{b^2_\mu}\right)\mu^2_{Dl} - 2\left\{\frac{1}{\sigma^2_{Dl}}\sum_{j:z_{Dj}=l} y_{Dj}\right\}\mu_{Dl}.$$

Then, completing the square

$$p(\mu_{Dl} \mid \text{else}) \propto \exp\left\{-\frac{1}{2}\left(\frac{n_{Dl}}{\sigma^2_{Dl}} + \frac{1}{b^2_\mu}\right)\left[\mu_{Dl} - \left(\frac{n_{Dl}}{\sigma^2_{Dl}} + \frac{1}{b^2_\mu}\right)^{-1}\left(\frac{1}{\sigma^2_{Dl}}\sum_{j:z_{Dj}=l} y_{Dj}\right)\right]^2\right\},$$

which we recognise as the kernel of a normal distribution with mean $\left(\frac{n_{Dl}}{\sigma^2_{Dl}} + \frac{1}{b^2_\mu}\right)^{-1}\left(\frac{1}{\sigma^2_{Dl}}\sum_{j:z_{Dj}=l} y_{Dj}\right)$ and variance $\left(\frac{n_{Dl}}{\sigma^2_{Dl}} + \frac{1}{b^2_\mu}\right)^{-1}$, that is,

$$\mu_{Dl} \mid \text{else} \sim \text{N}\left(\frac{\frac{1}{\sigma^2_{Dl}}\sum_{j:z_{Dj}=l} y_{Dj}}{\frac{n_{Dl}}{\sigma^2_{Dl}} + \frac{1}{b^2_\mu}}, \frac{1}{\frac{n_{Dl}}{\sigma^2_{Dl}} + \frac{1}{b^2_\mu}}\right).$$

The full conditional distributions for the variance of the mixture components are given by

$$\begin{aligned}
p(\sigma^2_{Dl} \mid \text{else}) &\propto L(\boldsymbol{\theta}_D; \boldsymbol{y}_D, \boldsymbol{z}_D)p(\sigma^2_{Dl}) \\
&\propto (\sigma^2_{Dl})^{-\frac{n_{Dl}}{2}} \exp\left\{-\frac{1}{2\sigma^2_{Dl}}\sum_{j:z_{Dj}=l}(y_{Dj} - \mu_{Dl})^2\right\} \times (\sigma^2_{Dl})^{-a_{\sigma^2_D}-1}\exp\left\{-\frac{b_{\sigma^2_D}}{\sigma^2_{Dl}}\right\} \\
&= (\sigma^2_{Dl})^{-a_{\sigma^2_D}-\frac{n_{Dl}}{2}-1}\exp\left\{-\frac{1}{\sigma^2_{Dl}}\left[b_{\sigma^2_D} + \frac{1}{2}\sum_{j:z_{Dj}=l}(y_{Dj} - \mu_{Dl})^2\right]\right\},
\end{aligned}$$

which is the kernel of an inverse-gamma distribution, thus

$$\sigma^2_{Dl} \mid \text{else} \sim \text{IG}\left(a_{\sigma^2_D} - \frac{n_{Dl}}{2}, b_{\sigma^2_D} + \frac{1}{2}\sum_{j:z_{Dj}=l}(y_{Dj} - \mu_{Dl})^2\right).$$

Now, for the stick-breaking weights, let us first look at the full conditional of v_{D1}

$$\begin{aligned}
p(v_{D1} \mid \text{else}) &\propto \prod_{l=1}^{L_D}\left\{v_{Dl}\prod_{t<l}(1 - v_{Dt})\right\}^{n_{Dl}}(1 - v_{D1})^{\alpha_D - 1} \\
&\propto v^{n_{D1}}_{D1}(1 - v_{D1})^{n_{D2}} \ldots (1 - v_{D1})^{n_{DL_D}}(1 - v_{D1})^{\alpha_D - 1} \\
&= v^{n_{D1}}_{D1}(1 - v_{D1})^{\sum_{l=2}^{L_D} n_{Dl} + \alpha_D - 1},
\end{aligned}$$

which is the kernel of a beta distribution with parameters $1 + n_{D1}$ and $\alpha_D + \sum_{l=2}^{L_D} n_{Dl}$. Given this, we can generalize and assert that

$$v_{Dl} \mid \text{else} \sim \text{Beta}\left(1 + n_{Dl}, \alpha_D + \sum_{h=l+1}^{L_D} n_{Dh}\right).$$

Next, we will derive the full conditional distribution for the concentration parameter of the Dirichlet process, that is

$$\begin{aligned}
p(\alpha_D \mid \text{else}) &\propto \prod_{l=1}^{L_D-1} p(v_{Dl} \mid \alpha_D) p(\alpha_D) \\
&\propto \prod_{l=1}^{L_D-1} \alpha_D (1 - v_{Dl})^{\alpha_D - 1} \times \alpha_D^{a_{\alpha_D}-1} \exp\{-b_{\alpha_D}\alpha_D\} \\
&= \alpha_D^{a_{\alpha_D}+L_D-2} \exp\left\{\log\left[\prod_{l=1}^{L_D-1}(1 - v_{Dl})\right]^{\alpha_D - 1} - b_{\alpha_D}\alpha_D\right\} \\
&\propto \alpha_D^{a_{\alpha_D}+L_D-2} \exp\left\{-\left[b_{\alpha_D} - \sum_{l=1}^{L_D-1} \log(1 - v_{Dl})\right]\alpha_D\right\},
\end{aligned}$$

which is clearly the kernel of a gamma distribution with shape $a_{\alpha_D} + L_D - 1$ and rate given by $b_{\alpha_D} - \sum_{l=1}^{L_D-1} \log(1 - v_{Dl})$, i.e.,

$$\alpha_D \mid \text{else} \sim \Gamma\left(a_{\alpha_D} + L_D - 1, b_{\alpha_D} - \sum_{l=1}^{L_D-1} \log(1 - v_{Dl})\right).$$

Finally, for the latent variables we have

$$\begin{aligned}
\mathbb{P}(z_{Dj} = h \mid \text{else}) &= \frac{L(\boldsymbol{\theta}_D; y_{Dj}, z_{Dj} = h)\mathbb{P}(z_{Dj} = h)}{\sum_{l=1}^{L_D} L(\boldsymbol{\theta}_D; y_{Dj}, z_{Dj} = l)\mathbb{P}(z_{Dj} = l)} \\
&= \frac{\omega_{Dh}\phi\left(y_{Dj} \mid \mu_{Dh}, \sigma^2_{Dh}\right)}{\sum_{l=1}^{L_D} \omega_{Dl}\phi\left(y_{Dj} \mid \mu_{Dl}, \sigma^2_{Dl}\right)}, \quad j = 1, \ldots, n_D, \;\; h = 1, \ldots, L_D.
\end{aligned}$$

These full conditional distributions allow us to implement a simple Gibbs sampler to simulate, say S, draws from the posterior distribution after burn-in. Thus, at iteration s, the estimated conditional density is given by

$$f_D^{(s)}(y_D) = \sum_{l=1}^{L_D} \omega_{Dl}^{(s)} \phi\left(y_D \mid \mu_{Dl}^{(s)}, \sigma^{2(s)}_{Dl}\right).$$

3.2.3 Our estimators

The estimators for the OVL that we consider are those based on Equations (1.2) and (1.3). For the first estimator (the one in (1.2)), we replaced the integral by a numerical integration method over all the posterior draws. More precisely, we use the trapezoidal rule for this end. Denoting by $g^{(s)}(y) = \min\left\{f_{\bar{D}}^{(s)}(y), f_{D}^{(s)}(y)\right\}$, a single Gibbs sampler iteration of our "plain-vanilla" estimator is given by

$$\begin{aligned}\text{OVL}^{(s)} &= \frac{\Delta y}{2}\sum_{i=1}^{N}\left\{g^{(s)}(y_{i-1}) + g^{(s)}(y_i)\right\},\\ &= \frac{\Delta y}{2}\left\{g^{(s)}(y_0) + 2\sum_{i=1}^{N-1} g^{(s)}(y_i) + g^{(s)}(y_N)\right\}, \quad s = 1,\ldots,S,\end{aligned} \tag{3.11}$$

where $\min\{y_{\bar{D}}, y_D\} = y_0 < \cdots < y_N = \max\{y_{\bar{D}}, y_D\}$ is an equally spaced grid and Δy is the length of each sub-interval.

For the second estimator (the one in (1.3)) we model the outer probabilities, $\mathbb{P}(f_{\bar{D}}(Y_{\bar{D}}) < f_D(Y_{\bar{D}}))$ and $\mathbb{P}(f_{\bar{D}}(Y_D) \geq f_D(Y_D))$, by using a DP and considering the limiting case, where the concentration parameter approaches zero. This limiting case is known as the Bayesian bootstrap (BB) and it can be regarded as a non-informative version of the DP (Gasparini, 1995, Theorem 2). The BB was proposed by Rubin (1981), and unlike the frequentist bootstrap (Efron, 1979), where the proportion of times each observation is drawn in a bootstrap replication follows a discrete uniform distribution, the Bayesian bootstrap samples correspond to discrete distributions supported at the observed data points with weights distributed according to a Dirichlet distribution (Gelman et al., 2013, p. 548). Hence, the difference between these two bootstraps relies on how the probabilities attached to each observation were generated.

Let $U_{\bar{D}} = f_{\bar{D}}(Y_{\bar{D}}) - f_D(Y_{\bar{D}})$ and $U_D = f_{\bar{D}}(Y_D) - f_D(Y_D)$ be two random variables with cumulative distribution functions given by $F_{U_{\bar{D}}}$ and F_{U_D}, respectively. Then, the outer probabilities can be written as follows

$$\begin{aligned}\mathbb{P}(f_{\bar{D}}(Y_{\bar{D}}) < f_D(Y_{\bar{D}})) &= \mathbb{P}(U_{\bar{D}} < 0) = F_{U_{\bar{D}}}(0),\\ \mathbb{P}(f_{\bar{D}}(Y_D) \geq f_D(Y_D)) &= \mathbb{P}(U_D \geq 0) = 1 - F_{U_D}(0).\end{aligned}$$

Thus, the OVL in expression (1.3) can be computed as follows

$$\text{OVL} = F_{U_{\bar{D}}}(0) + 1 - F_{U_D}(0).$$

Now, we assume a DP for $F_{U_{\bar{D}}}$ and F_{U_D}, i.e.,

$$\begin{aligned} u_{\bar{D}1}, \ldots, u_{\bar{D}n_{\bar{D}}} \mid F_{U_{\bar{D}}} &\overset{\text{ind.}}{\sim} F_{U_{\bar{D}}}, \quad F_{U_{\bar{D}}} \mid \alpha_{U_{\bar{D}}}, F^*_{U_{\bar{D}}} \sim \text{DP}(\alpha_{U_{\bar{D}}}, F^*_{U_{\bar{D}}}), \\ u_{D1}, \ldots, u_{Dn_D} \mid F_{U_D} &\overset{\text{ind.}}{\sim} F_{U_D}, \quad F_{U_D} \mid \alpha_{U_D}, F^*_{U_D} \sim \text{DP}(\alpha_{U_D}, F^*_{U_D}). \end{aligned}$$

Due to the conjugacy property of the DP (Ferguson, 1973, Theorem 1), the resulting posterior distributions of $F_{U_{\bar{D}}}$ and F_{U_D} are DP as well, that is,

$$\begin{aligned} F_{U_{\bar{D}}} \mid u_{\bar{D}1}, \ldots, u_{\bar{D}n_{\bar{D}}} &\sim \text{DP}\left(\alpha_{U_{\bar{D}}} + n_{\bar{D}}, \frac{\alpha_{U_{\bar{D}}}}{\alpha_{U_{\bar{D}}} + n_{\bar{D}}} F^*_{U_{\bar{D}}} + \frac{n_{\bar{D}}}{\alpha_{U_{\bar{D}}} + n_{\bar{D}}} \hat{F}^{\text{emp}}_{U_{\bar{D}}}\right), \\ F_{U_D} \mid u_{D1}, \ldots, u_{Dn_D} &\sim \text{DP}\left(\alpha_{U_D} + n_D, \frac{\alpha_{U_D}}{\alpha_{U_D} + n_D} F^*_{U_D} + \frac{n_D}{\alpha_{U_D} + n_D} \hat{F}^{\text{emp}}_{U_D}\right), \end{aligned}$$

where $\hat{F}^{\text{emp}}_{U_{\bar{D}}}(u) = \frac{1}{n_{\bar{D}}} \sum_{i=1}^{n_{\bar{D}}} \mathbb{I}\left\{u_{\bar{D}i} \leq u\right\}$ is the empirical distribution function of $U_{\bar{D}}$, $\tilde{F}^{\text{emp}}_{U_D}$ is similarly defined. For the sake of computational simplicity, we consider the limiting case where the concentration parameters $\alpha_{U_{\bar{D}}}$ and α_{U_D} approach zero, in some sense this means that an uniformative prior is imposed. Note that we do not even need to specify $F^*_{U_{\bar{D}}}$ or $F^*_{U_D}$, as the baseline distribution vanishes in the limiting case. Hence, the posterior distributions are simplified to

$$\begin{aligned} F_{U_{\bar{D}}} \mid u_{\bar{D}1}, \ldots, u_{\bar{D}n_{\bar{D}}} &\sim \text{DP}\left(n_{\bar{D}}, \hat{F}^{\text{emp}}_{U_{\bar{D}}}\right), \\ F_{U_D} \mid u_{D1}, \ldots, u_{Dn_D} &\sim \text{DP}\left(n_D, \hat{F}^{\text{emp}}_{U_D}\right). \end{aligned}$$

We can use the truncated stick-breaking representation of the DP to draw posterior realisations of F_{U_D} (the procedure for $F_{U_{\bar{D}}}$ follows analogously) as follows

$$F_{U_D}(u) = \sum_{k=1}^{K_D} \omega_{Dk} \mathbb{I}\left\{\theta_{Dk} \leq u\right\}, \quad \omega_{Dk} = \begin{cases} h_{Dk}, & \text{if} \quad k = 1 \\ h_{Dk} \prod_{t<k} (1 - h_{Dt}), & \text{if} \quad 2 \leq k \leq K_D \end{cases},$$

where $h_{D1}, \ldots, h_{DK_D - 1} \overset{\text{iid}}{\sim} \text{Beta}(1, n_D)$, $h_{DK_D} = 1$ and $\theta_{Dk} \overset{\text{iid}}{\sim} \hat{F}^{\text{emp}}_{U_D}$, for $k = 1, \ldots, K_D$. Note that in this case, K_D ($K_{\bar{D}}$) must be much larger that n_D ($n_{\bar{D}}$) to ensure a good coverage of all possible

values of U_D ($U_{\bar{D}}$). According to Gasparini (1995, Theorem 2), this construction is equivalent to the BB. Therefore, instead of generating a large number of weights and draws from the empirical cdf for each posterior realisation of F_{U_D} ($F_{U_{\bar{D}}}$), we choose to generate the weights directly from a Dirichlet distribution as in the BB.

Once the densities have been estimated, we can compute posterior realisations of $U_{\bar{D}}$ and U_D as follows

$$u_{\bar{D}i}^{(s)} = f_{\bar{D}}^{(s)}(y_{\bar{D}i}) - f_D^{(s)}(y_{\bar{D}i}) = \sum_{\ell=1}^{L_{\bar{D}}} \omega_{\bar{D}\ell}^{(s)} \phi\left(y_{\bar{D}i} \mid \mu_{\bar{D}\ell}^{(s)}, (\sigma_{\bar{D}\ell}^{(s)})^2\right) - \sum_{l=1}^{L_D} \omega_{Dl}^{(s)} \phi\left(y_{\bar{D}i} \mid \mu_{Dl}^{(s)}, (\sigma_{Dl}^{(s)})^2\right), \quad i = 1, \ldots, n_{\bar{D}},$$

$$u_{Dj}^{(s)} = f_{\bar{D}}^{(s)}(y_{Dj}) - f_D^{(s)}(y_{Dj}) = \sum_{\ell=1}^{L_{\bar{D}}} \omega_{\bar{D}\ell}^{(s)} \phi\left(y_{Dj} \mid \mu_{\bar{D}\ell}^{(s)}, (\sigma_{\bar{D}\ell}^{(s)})^2\right) - \sum_{l=1}^{L_D} \omega_{Dl}^{(s)} \phi\left(y_{Dj} \mid \mu_{Dl}^{(s)}, (\sigma_{Dl}^{(s)})^2\right), \quad j = 1, \ldots, n_D.$$

We can now compute the probabilities involved in Equation (1.3) using the previous expressions combined with the Bayesian bootstrap. Hence, the corresponding realisation of the OVL is given by

$$\begin{aligned} \mathrm{OVL}_{BB}^{(s)} &= F_{U_{\bar{D}}}^{(s)}(0) + 1 - F_{U_D}^{(s)}(0) && (3.12) \\ &= \sum_{i=1}^{n_{\bar{D}}} q_{\bar{D}i}^{(s)} \mathbb{I}\left\{u_{\bar{D}i}^{(s)} < 0\right\} + \sum_{j=1}^{n_D} q_{Dj}^{(s)} \mathbb{I}\left\{u_{Dj}^{(s)} \geq 0\right\}, \quad s = 1, \ldots, S, && (3.13) \end{aligned}$$

where $(q_{\bar{D}1}^{(s)}, \ldots, q_{\bar{D}n_{\bar{D}}}^{(s)}) \sim \mathrm{Dirichlet}(n_{\bar{D}}; 1, \ldots, 1)$ and $(q_{D1}^{(s)}, \ldots, q_{Dn_D}^{(s)}) \sim \mathrm{Dirichlet}(n_D; 1, \ldots, 1)$. Although we are considering continuous test outcomes, measurements are made with finite precision, thus meaning that ties in the test outcomes within each population can occur and this also implies ties in $U_{\bar{D}}$ and/or U_D. In such case, the parameter vector of the Dirichlet distribution must be adjusted according to the number of ties, that is, adding together the ones for each repeated test outcome. We have chosen to model $F_{U_{\bar{D}}}$ and F_{U_D} using a Bayesian bootstrap instead of a DPM for merely computational convenience. Despite that we can use a DPM with an appropriate kernel for modelling purposes, this would imply that for each set $\{u_{\bar{D}i}\}_{i=1}^{n_{\bar{D}}}$ and $\{u_{Dj}\}_{j=1}^{n_D}$ we would need to generate another S' realisations of the resulting models to compute the required cumulative distribution functions.

Finally, point estimates for both estimators of the OVL can be obtained by averaging over the posterior ensembles $\left\{\mathrm{OVL}^{(1)}, \ldots, \mathrm{OVL}^{(S)}\right\}$ and $\left\{\mathrm{OVL}_{BB}^{(1)}, \ldots, \mathrm{OVL}_{BB}^{(S)}\right\}$. For example, for the DPM-BB estimator $\widehat{\mathrm{OVL}}_{BB} = \sum_{s=1}^{S} \mathrm{OVL}_{BB}^{(s)}$. A 95% credible interval can also be obtained using the 2.5%

and 97.5% percentiles of the corresponding ensemble. We have observed that for very large overlap between the densities of the test outcomes, some realisations of (3.12) can be slightly above one and therefore, in such case, we can consider $\min\left\{1, \mathrm{OVL}_{BB}^{(s)}\right\}$.

3.3 Simulation study

We carried out a simulation study to evaluate the performance of our OVL estimators across three different scenarios. For each simulation scenario, three different configurations, that lead to small, intermediate and large overlaps, were used and 100 datasets were generated using sample sizes of $(n_{\bar{D}}, n_D) = (50, 50)$, $(n_{\bar{D}}, n_D) = (100, 100)$, $(n_{\bar{D}}, n_D) = (200, 200)$ and $(n_{\bar{D}}, n_D) = (500, 500)$. Although we assumed equal sample sizes for both groups to simplify computations, in practice it is common to find imbalanced data, e.g., a larger number of nondiseased compared to diseased individuals. However, unless the difference in sample sizes is very large, the results should not differ too much from those presented here.

3.3.1 Simulation scenarios

The simulation scenarios considered are listed in Table 3.1. Scenario I corresponds to the case where test outcomes in the two populations follow normal distributions. In Scenario II, test outcomes in both groups arise from different and non-normal distributions, namely, a gamma distribution and a skew normal distribution. Lastly, Scenario III considers mixtures of normal distributions in each of the two populations.

3.3.2 Models

We standardised the data (i.e., subtracted the mean and divided by the standard deviation) to facilitate prior specification and avoid computational issues, but we transformed back to the original scale when presenting the results. According to Hanson (2006), in absence of prior information on the number of components needed to adequately describe the density functions, we set $\alpha_d = 1$, for $d \in \{\bar{D}, D\}$. This

Table 3.1: Different distributional assumptions for $Y_{\bar{D}}$ and Y_D under Scenario I, II, and III. Note that $\text{SN}(\nu, \eta, \lambda)$ denotes a skew normal distribution with location ν, scale η, and skewness parameter λ.

Scenario	$Y_{\bar{D}}$	Y_D	**OVL**
I	$\text{N}(-0.75, 1^2)$	$\text{N}(2.5, 1^2)$	0.104
	$\text{N}(1.1, 1^2)$	$\text{N}(2.5, 1^2)$	0.484
	$\text{N}(2.2, 1^2)$	$\text{N}(2.5, 1^2)$	0.880
II	$\Gamma(3, 1)$	$\text{SN}(5, 2, 5)$	0.172
	$\Gamma(3, 1)$	$\text{SN}(3, 2, 5)$	0.482
	$\Gamma(3, 1)$	$\text{SN}(1.25, 2, 5)$	0.860
III	$0.5\text{N}(-2.5, 1^2) + 0.5\text{N}(0.5, 1^2)$	$0.5\text{N}(2.5, 1^2) + 0.5\text{N}(5, 1^2)$	0.159
	$0.5\text{N}(-1.15, 1^2) + 0.5\text{N}(1.5, 1^2)$	$0.5\text{N}(1.5, 1^2) + 0.5\text{N}(3.5, 1^2)$	0.510
	$0.5\text{N}(0, 1^2) + 0.5\text{N}(3, 1^2)$	$0.5\text{N}(0.5, 1^2) + 0.5\text{N}(3.25, 1^2)$	0.896

prior specification favours a small number of occupied mixture components relative to the sample size (Gelman et al., 2013, p. 553). We can compute the conditional prior mean and standard deviation of the number of occupied components using the expressions in (3.4) derived by Liu (1996). For instance, for a sample size of 50, such choice leads to a prior expected (standard deviation) number of occupied components of approximately 4(2), of approximately 5(2) and 6(2) for sample sizes of 100 and 200, respectively, and for a sample size of 500, of approximately 7(2). For the mean and variance of the components, we used as hyperparameters $a_{\mu_d} = 0, b^2_{\mu_d} = 10, a_{\sigma^2_d} = 2$ and $b_{\sigma^2_d} = 0.5$ leading to a relatively vague prior distributions. This is due the fact that test outcomes are standardised, thus we expect the means of the mixture components to be around zero, therefore $a_{\mu_d} = 0$. The variance $b^2_{\mu_d} = 10$ implies that about 95% of the drawn values of μ_{dl} roughly lie within -6 and 6. Whereas $a_{\sigma^2_d} = 2$ leads to a prior on σ^2_{dl} with infinite variance (hence, in some sense, vague) centred around a finite mean ($b_{\sigma^2_d} = 0.5$). Note that this prior on σ^2_{dl} favours variances less than one. The standardised data have marginal unit variance, so the within-component variance σ^2_{dl} is expected to be smaller than the marginal variance. We capped the stick-breaking construction up to $L_d = 10$, meaning that a maximum of 10 normal distributions were used to approximate the densities in each group. Posterior inference was based on

5,000 iterations after discarding 2,000 iterations of the Gibbs sampler as the burn-in. Finally, for the trapezoidal rule we used a grid of 1,001 equally spaced points.

We compared our Bayesian nonparametric estimators against their frequentist counterparts. We estimated the densities using one of the most popular nonparametric methods, the Gaussian kernel density estimator (see e.g., Clemons and Bradley Jr., 2000 and Schmid and Schmidt, 2006). Then, the density of the diseased group can be written as

$$\hat{f}_D(y) = \frac{1}{n_D h_D} \sum_{j=1}^{n_D} \phi\left(\frac{y - y_{Dj}}{h_D}\right),$$

where $\phi(\cdot)$ is the standard normal density function and $h_D > 0$ is a smoothing parameter known as the bandwidth, selected using the rule of thumb derived by Silverman (1986, Chapter 3) and computed as follows

$$h_D = 0.9 n_D^{-0.2} \min\{\hat{\sigma}_D, \text{IQR}_D/1.34\},$$

where $\hat{\sigma}_D$ is the estimated standard deviation of the diseased group and IQR_D is the inter-quartile range. A similar expression holds for the density function of the test outcomes in the nondiseased group. The estimated densities are then plugged in on Equation (1.2) to obtain the corresponding "plain vanilla" estimator. The integral was computed numerically through the trapezoidal rule on a grid of 1,001 equally spaced points. For the estimator of (1.3), the frequentist counterpart was proposed by Schmid and Schmidt (2006) and it is given by

$$\widehat{\text{OVL}} = \frac{1}{n_{\bar{D}}} \sum_{i=1}^{n_{\bar{D}}} \mathbb{I}\left\{\hat{f}_{\bar{D}}(y_{\bar{D}i}) - \hat{f}_D(y_{\bar{D}i}) < 0\right\} + \frac{1}{n_D} \sum_{j=1}^{n_D} \mathbb{I}\left\{\hat{f}_{\bar{D}}(y_{Dj}) - \hat{f}_D(y_{Dj} \geq 0)\right\}.$$

3.3.3 Results

The boxplots displayed in Figures 3.1–3.3 summarize the estimates of the OVL for each method and for the three parameters configurations and sample sizes considered. In general, we observe that our Bayesian nonparametric estimators perform well in most scenarios, yielding values close to the true OVL. We note that all the estimators present a slight bias when there is a large overlap between the two density functions of the test outcomes, something already discussed in Ridout and Linkie (2009).

However, the DPM-BB estimator outperforms its competitors in terms of bias, only in few cases, the frequentist counterparts are on par. Further, the DPM-BB estimator presents higher variability. Conversely, the "plain vanilla" estimator presents larger bias in such cases, being more evident in small samples $(n_{\bar{D}}, n_D) = (50, 50)$. However, as the sample size increases, the variability of the estimators and the bias decrease.

For our proposed estimators, we investigated the empirical coverage probability of the 95% credible intervals as well as their width. The results are shown in Tables 3.2–3.4 and in Figures 3.4–3.6, respectively. The "plain vanilla" estimator presents, in most cases, a coverage close to the nominal value of 0.95. However, we have observed that this estimator presented larger bias for higher values of the OVL, especially when $(n_{\bar{D}}, n_D) = (50, 50)$, thus its coverage is expected to be below the nominal value. On the other hand, the DPM-BB estimator presented a coverage greater than the nominal value in almost all the scenarios, even when $(n_{\bar{D}}, n_D) = (50, 50)$. At this point, it is worth to mention that the width of the credible intervals of this estimator is larger than that of the numerical integration-based estimator; in particular, for the cases in which the overlap between the two densities is high. Therefore, higher coverage is to be expected. However, similar to the previous results, the width of the credible intervals decreases as the sample size increases.

In conclusion, our Bayesian nonparametric estimators perform well in the simulated scenarios considered. We observe that there is a slight bias when the true OVL is large. However, the DPM-BB estimator still provides good results, in terms of bias. Further, our results suggest that our estimators may be consistent, because as the sample size increases, the estimated OVL values become more concentrated around the true OVL. We study the empirical coverage probability of the 95% credible intervals of our proposed estimators. We found that, in most cases, the nominal value of 0.95 was reached. When the overlapping area between the two densities of test outcomes is large, we note that the coverage is below the nominal value, even for the DPM-BB estimator and when $(n_{\bar{D}}, n_D) = (500, 500)$. Finally, we observe that the width of the credible intervals of the DPM-BB estimator is larger compared to that of the "plain vanilla" estimator. This means that higher coverage is achieved, but at the expense of higher uncertainty.

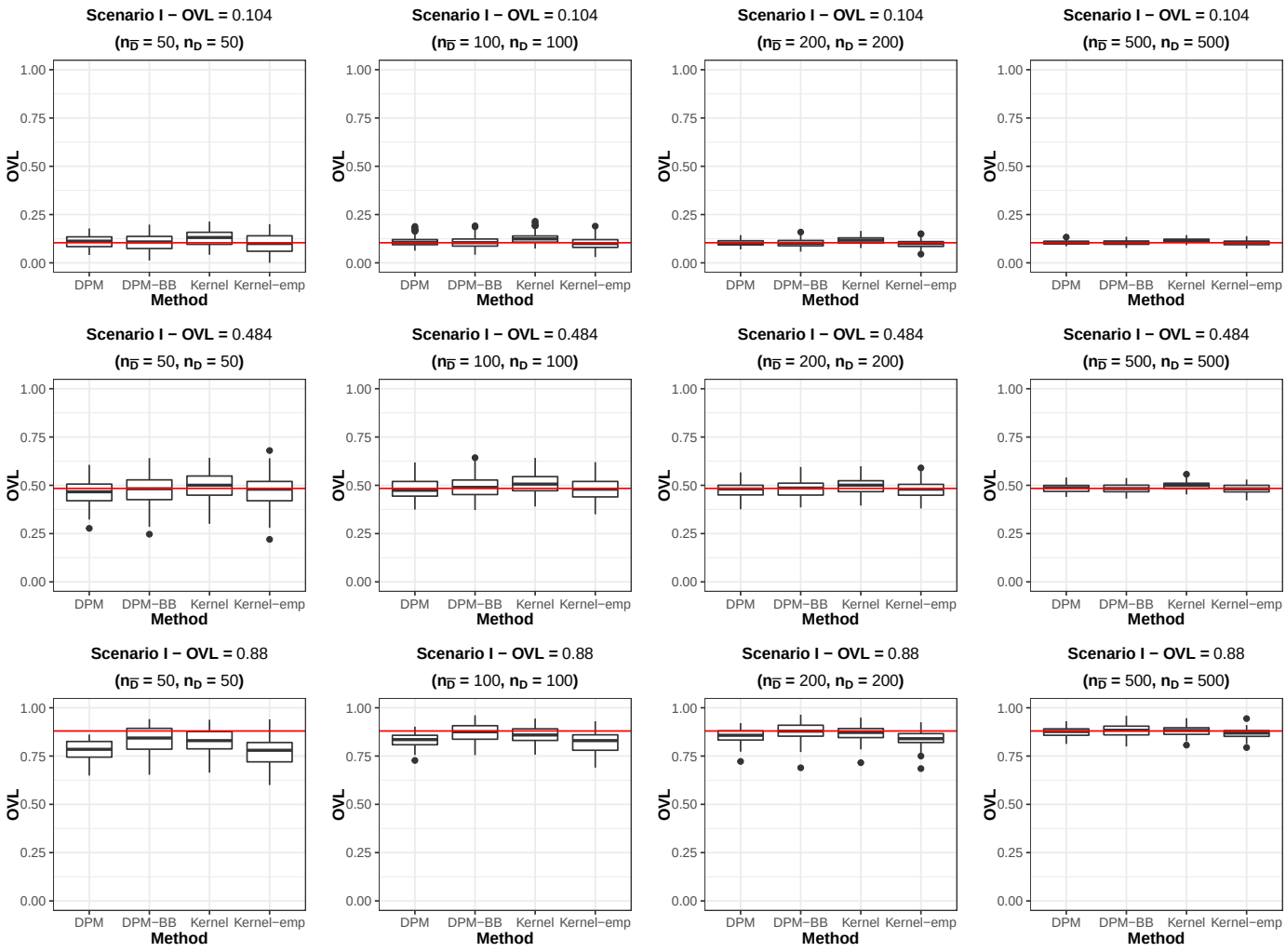

Figure 3.1: Results for Scenario I. Boxplots of the OVL estimators across the 100 simulated datasets and for the three different parameter configurations and sample sizes considered. The solid red line represents the true OVL. DPM: posterior mean, for each dataset, under the DPM plain vanilla estimator; DPM-BB: posterior mean of the Bayesian bootstrap-based estimator; Kernel: Gaussian kernel estimator based in (1.2); Kernel-emp: Gaussian kernel estimator based in (1.3).

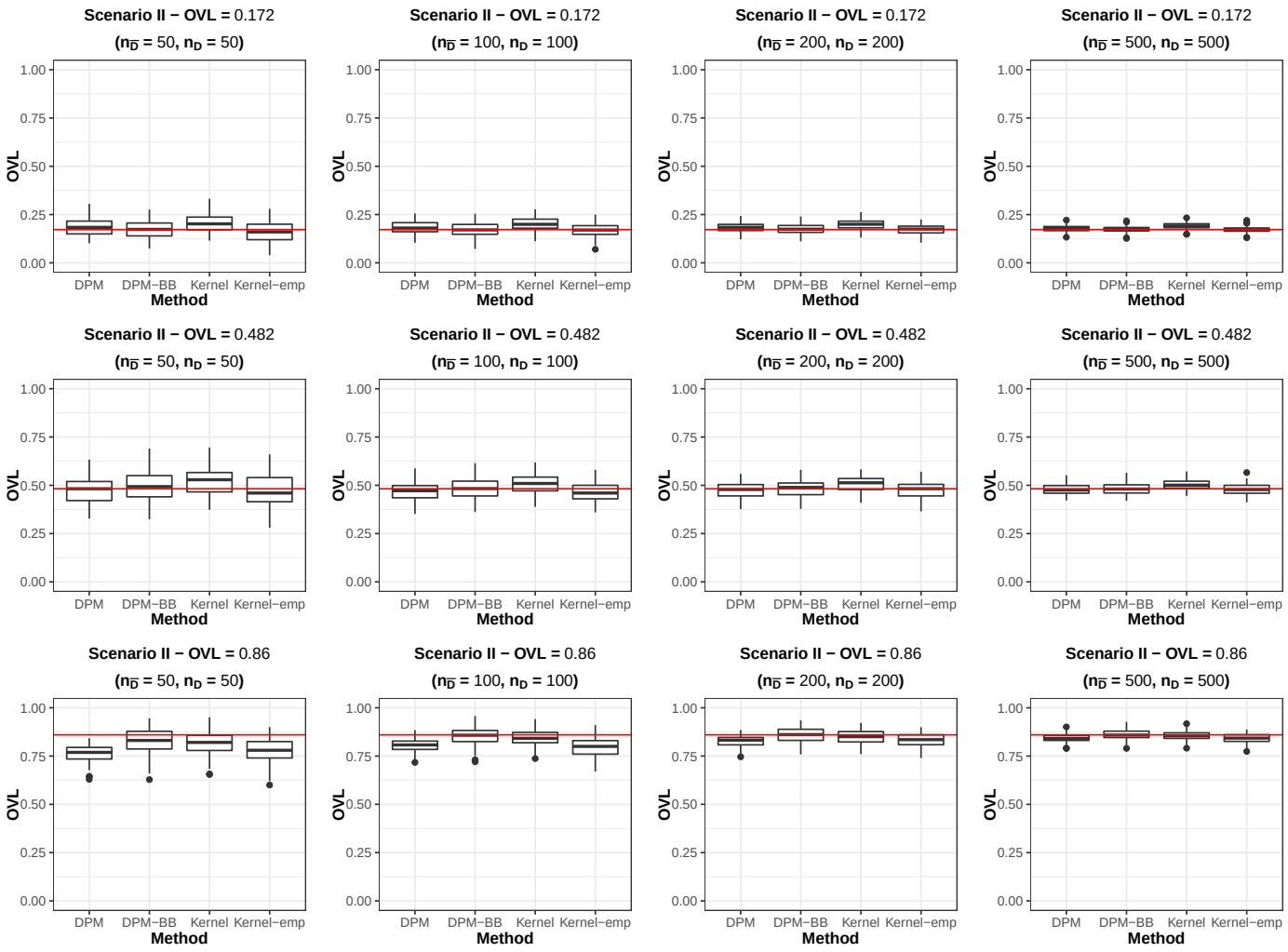

Figure 3.2: Results for Scenario II. Boxplots of the OVL estimators across the 100 simulated datasets and for the three different parameter configurations and sample sizes considered. The solid red line represents the true OVL. DPM: posterior mean, for each dataset, under the DPM plain vanilla estimator; DPM-BB: posterior mean of the Bayesian bootstrap-based estimator; Kernel: Gaussian kernel estimator based in (1.2); Kernel-emp: Gaussian kernel estimator based in (1.3).

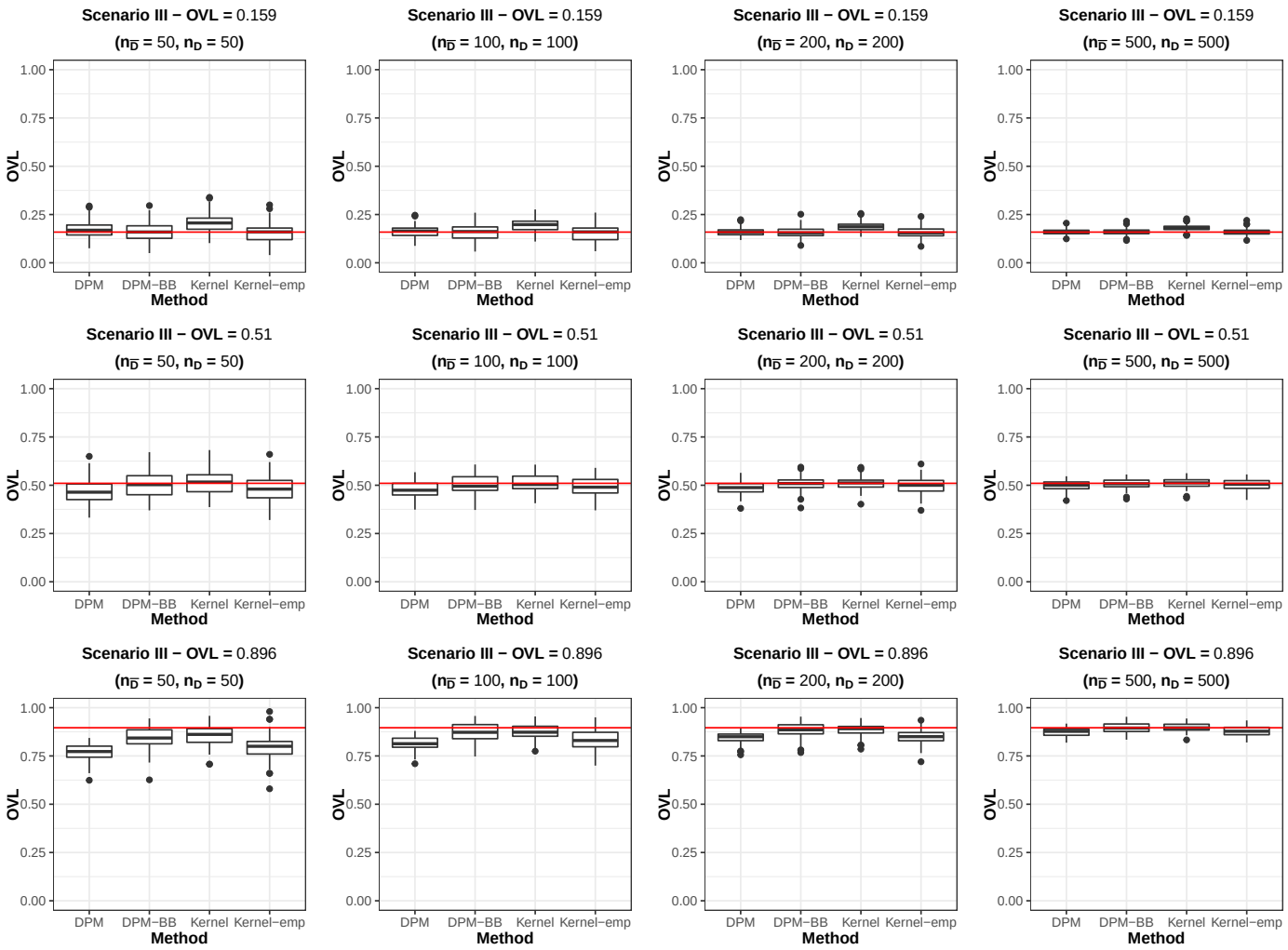

Figure 3.3: Results for Scenario III. Boxplots of the OVL estimators across the 100 simulated datasets and for the three different parameter configurations and sample sizes considered. The solid red line represents the true OVL. DPM: posterior mean, for each dataset, under the DPM plain vanilla estimator; DPM-BB: posterior mean of the Bayesian bootstrap-based estimator; Kernel: Gaussian kernel estimator based in (1.2); Kernel-emp: Gaussian kernel estimator based in (1.3).

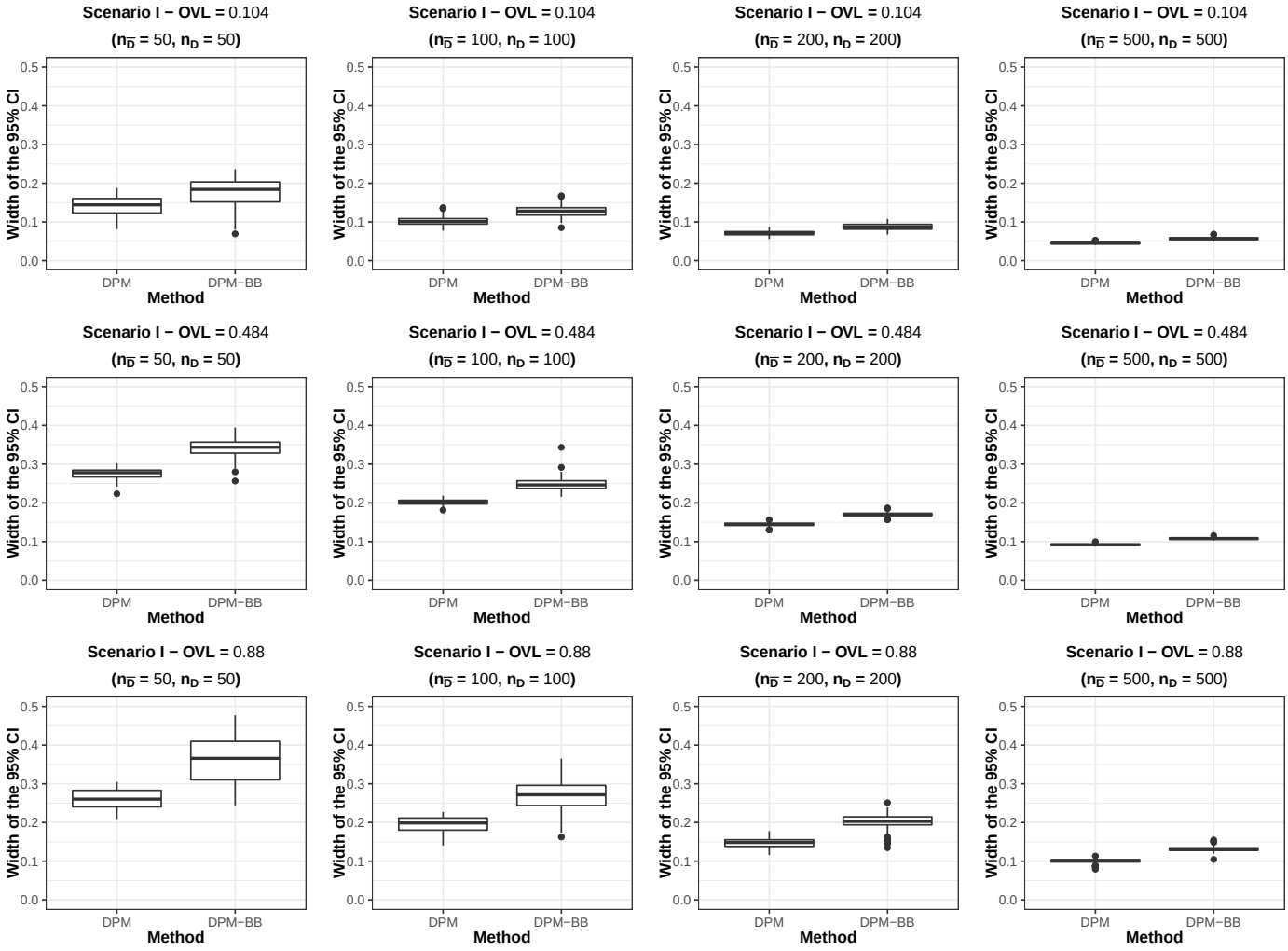

Figure 3.4: Results for Scenario I. Boxplots representing the width of the 95% credible intervals of our Bayesian nonparametric OVL estimators across the 100 simulated datasets and for the three different parameter configurations and sample sizes considered. DPM: plain vanilla estimator; DPM-BB: Bayesian bootstrap-based estimator.

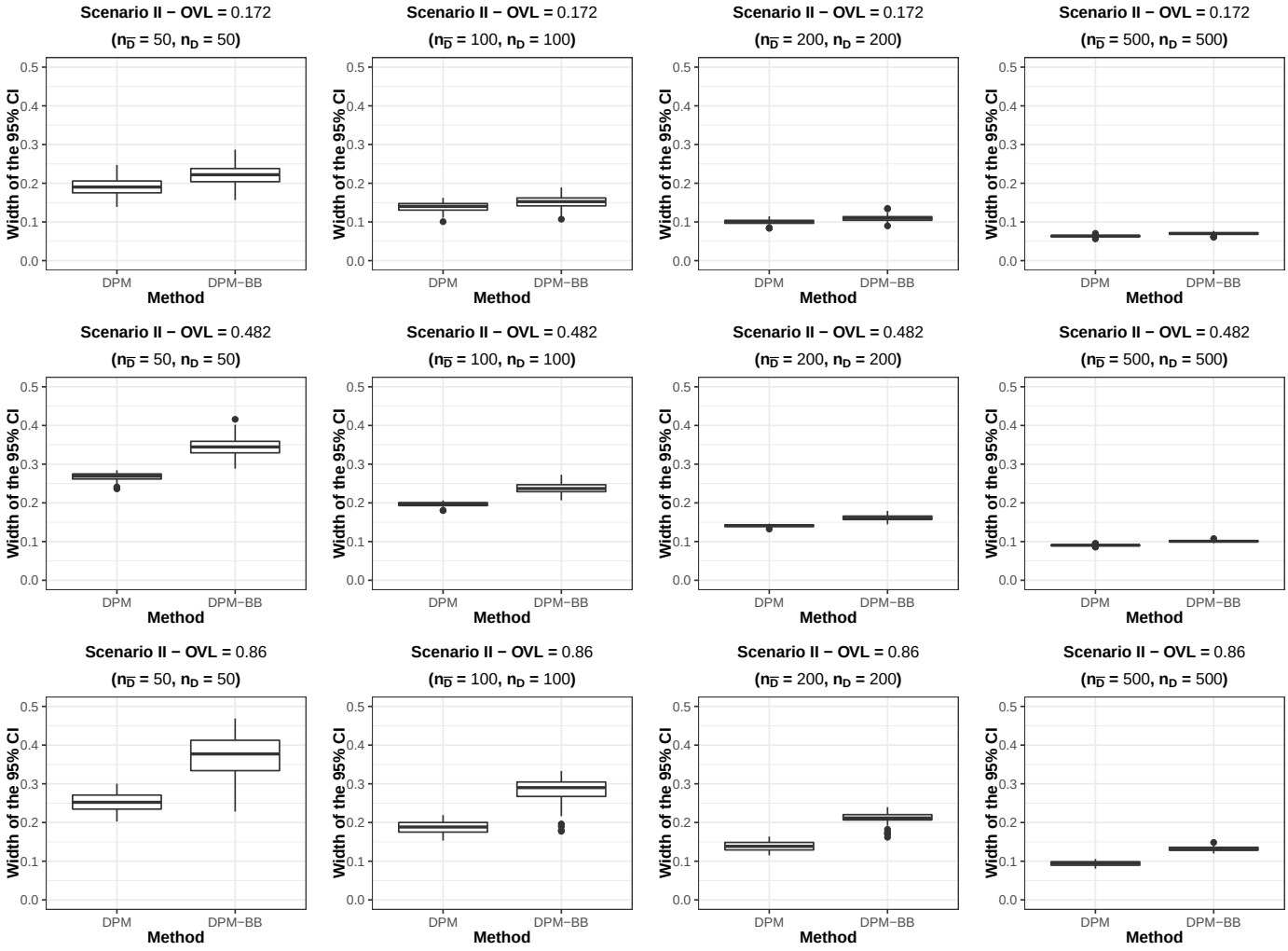

Figure 3.5: Results for Scenario II. Boxplots representing the width of the 95% credible intervals of our Bayesian nonparametric OVL estimators across the 100 simulated datasets and for the three different parameter configurations and sample sizes considered. DPM: plain vanilla estimator; DPM-BB: Bayesian bootstrap-based estimator.

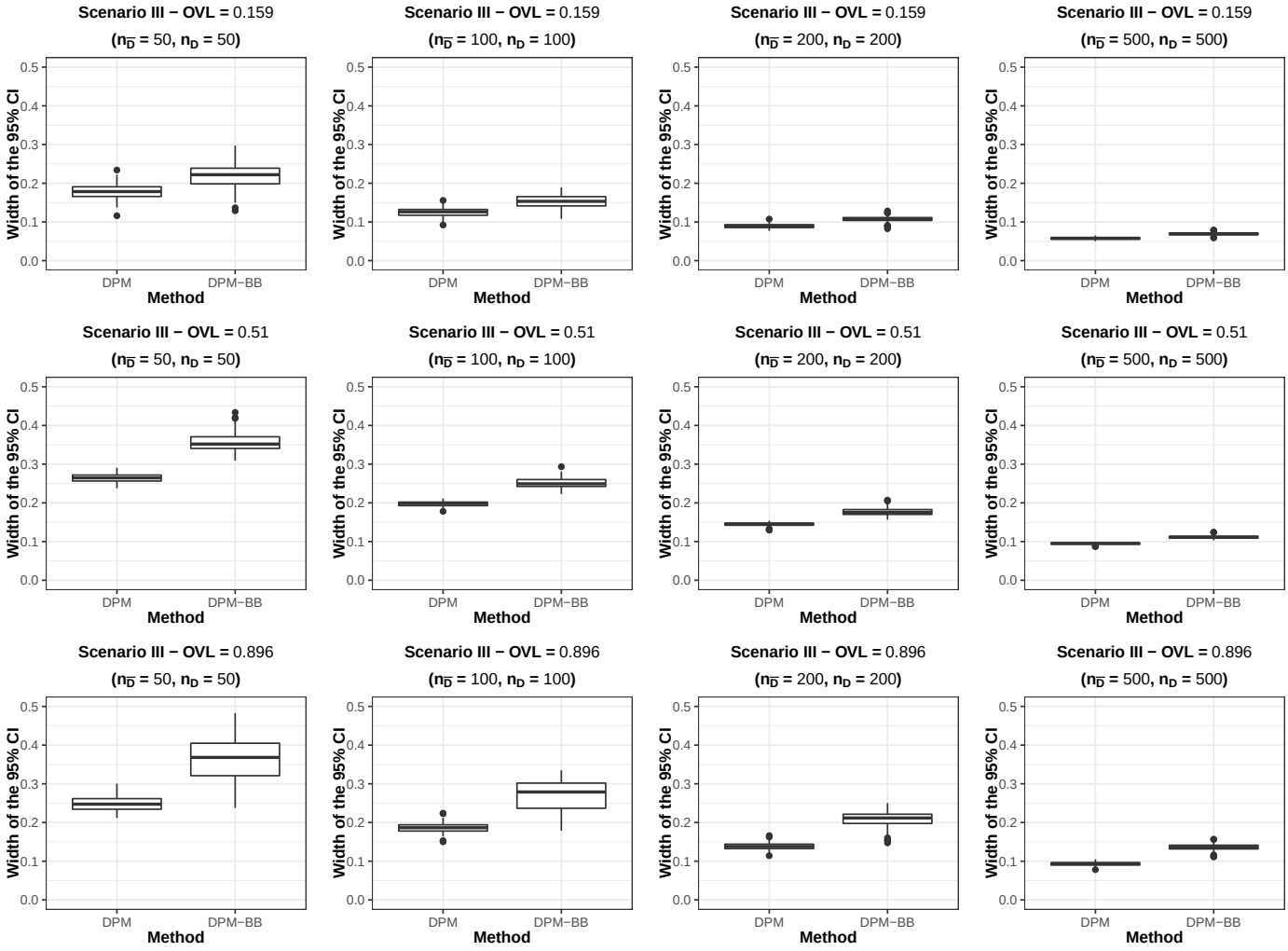

Figure 3.6: Results for Scenario III. Boxplots representing the width of the 95% credible intervals of our Bayesian nonparametric OVL estimators across the 100 simulated datasets and for the three different parameter configurations and sample sizes considered. DPM: plain vanilla estimator; DPM-BB: Bayesian bootstrap-based estimator.

Table 3.2: Scenario I. Empirical coverage probability of the 95% credible intervals

OVL	$(n_{\bar{D}}, n_D)$	**DPM**	**DPM-BB**
0.104	(50, 50)	0.98	0.97
	(100, 100)	0.94	0.95
	(200, 200)	0.95	0.96
	(500, 500)	0.98	0.98
0.484	(50, 50)	0.94	0.96
	(100, 100)	0.97	0.98
	(200, 200)	0.93	0.95
	(500, 500)	0.99	0.99
0.880	(50, 50)	0.76	0.98
	(100, 100)	0.97	1.0
	(200, 200)	0.94	0.97
	(500, 500)	0.97	0.97

Table 3.3: Scenario II. Empirical coverage probability of the 95% credible intervals

OVL	$(n_{\bar{D}}, n_D)$	**DPM**	**DPM-BB**
0.172	(50, 50)	0.97	1.0
	(100, 100)	0.96	0.98
	(200, 200)	0.93	0.96
	(500, 500)	0.91	0.96
0.482	(50, 50)	0.95	0.97
	(100, 100)	0.95	0.95
	(200, 200)	0.96	0.96
	(500, 500)	0.91	0.95
0.860	(50, 50)	0.76	0.99
	(100, 100)	0.84	1.0
	(200, 200)	0.85	1.0
	(500, 500)	0.94	0.99

Table 3.4: Scenario III. Empirical coverage probability of the 95% credible intervals

OVL	$(n_{\bar{D}}, n_D)$	**DPM**	**DPM-BB**
0.159	(50, 50)	0.95	0.96
	(100, 100)	0.95	0.95
	(200, 200)	0.98	0.94
	(500, 500)	0.94	0.95
0.510	(50, 50)	0.95	1.0
	(100, 100)	0.92	0.99
	(200, 200)	0.89	0.99
	(500, 500)	0.95	0.98
0.896	(50, 50)	0.43	0.99
	(100, 100)	0.65	1.0
	(200, 200)	0.77	0.97
	(500, 500)	0.87	1.0

3.3.4 Induced prior on OVL

In Section 3.3.3, we observed that the DPM estimator yields biased results when the overlap between the density functions of the nondiseased and diseased groups is very large. Therefore, we investigate whether the prior specification we used was a cause for such bias. To this end, we simulate from the prior predictive distribution, that is, the prior predictive distribution of the diseased group is defined as follows

$$f_D(y_D) = \int p(\boldsymbol{\theta}_D) f(y_D \mid \boldsymbol{\theta}_D) d\boldsymbol{\theta}_D,$$

where $\boldsymbol{\theta}_D = (\boldsymbol{\mu}_D, \boldsymbol{\sigma}_D^2, \boldsymbol{\omega}_D)$. We can obtain posterior samples, say S, from the prior predictive distribution sampling from each of the prior distributions of the parameters defined in Equations (3.6)–(3.10), thus a single draw from the prior predictive distribution in the diseased group is given by

$$f_D(\cdot) = \sum_{l=1}^{L_D} \omega_{Dl}^{(s)} \phi(\cdot \mid \mu_{Dl}^{(s)}, (\sigma_{Dl}^{(s)})^2).$$

The prior predictive distribution in the nondiseased group is similarly obtained. Then, we can simply plug in the densities into Equation (3.11) and obtain the induced prior on the OVL. We compared the induced OVL under the same prior specification as in the simulation study detailed in Section 3.3.2 to an informative prior based on the data. More precisely, the hyperparameters that we used for the mean of the components are $a_{\mu_d} = \bar{y}_d$ and $b_{\mu_d}^2 = \frac{100}{n_d}\hat{\sigma}_d^2$, for $d \in \{\bar{D}, D\}$; and for the hyperparameters of the variance, we used $a_{\sigma_d^2} = 2$ and $b_{\sigma_d^2} = \hat{\sigma}_d^2/2$, where $\bar{y}_d$ and $\hat{\sigma}_d^2$ are the corresponding sample mean and variance, respectively. We used the same values of α_d and L_d, and we used 301 points of a grid from -15 to 15 considering 10,000 simulations. It is worth to mention that under the uniformative prior, we centered and scaled the data, but for the informative prior we do not.

The distribution of the OVL before observing any data under both priors is depicted in Figure 3.7. We notice that by using the uniformative prior, the two prior predictive distributions (Figure 3.8 left) can be quite different, meaning that the induced prior on the overlap is concentrated on low values. This can be clearly observed in the violin plot (Figure 3.7 right plot of the first panel), where under such prior specification, the distribution concentrates on low values of the OVL. In turn,

under an informative (data-based) prior, the induced OVL values follow an almost perfectly symmetric distribution. It is possible to observe the same behaviour in the corresponding histograms.

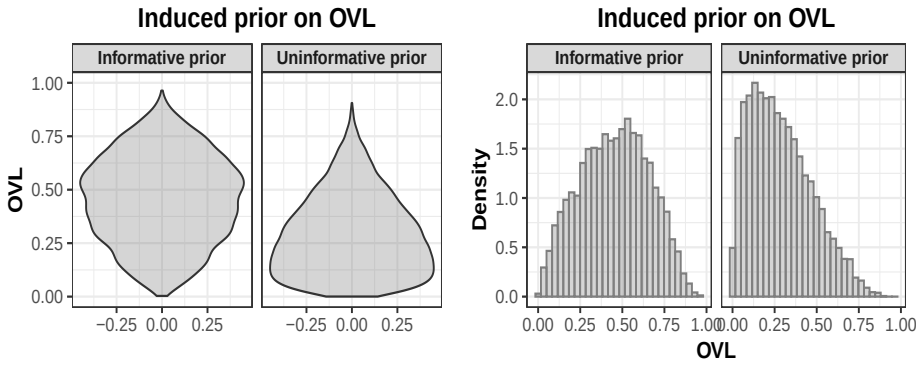

Figure 3.7: Violin plot and histograms of the induced prior on the OVL under different priors.

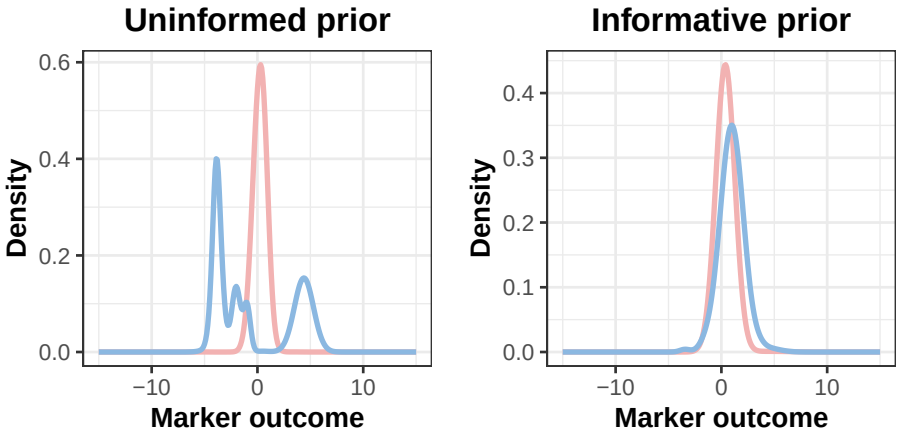

Figure 3.8: A single draw of the prior predictive distributions in the nondiseased (blue) and diseased (red) group under the two considered priors.

We ran a simulation study to investigate whether there is a difference on the resulting OVL under both priors. We follow the same simulation settings described in Section 3.3.2 regarding the number of simulated datasets, iterations, burn-in and grid points. We used the case 3 of Scenario I, where the true data generating process is given by

$$y_{\bar{D}i} \overset{\text{ind}}{\sim} \text{N}(2.2, 1), \quad i = 1, \ldots, n_{\bar{D}} = 50,$$
$$y_{Dj} \overset{\text{ind}}{\sim} \text{N}(2.5, 1), \quad j = 1, \ldots, n_{D} = 50.$$

Results are shown in Figure 3.9. It is clear to observe no evidence that an informed prior performs better in terms of bias, even in the case where the sample size is small ($(n_{\bar{D}}, n_D) = (50, 50)$) in each group. Results for case 3 of Scenario II and III are similar.

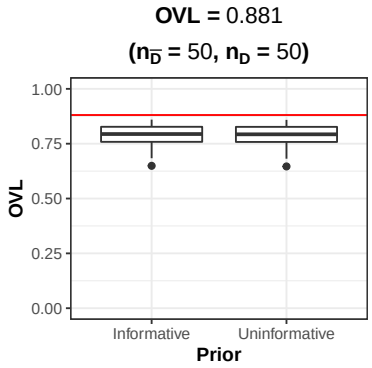

Figure 3.9: Boxplots of the DPM estimator across the 100 simulated datasets. The solid red line represents the true OVL. Informative: prior based on the data; Uninformative: the prior specification used in Section 3.3.2.

3.4 Application to ovarian cancer diagnosis

One of the most important and challenging objectives in medical research is the evaluation of potential biomarkers that could distinguish between cancerous and normal organ tissues. It is of particular interest to identify genes that are differentially expressed in ovarian cancerous tissue compared with a normal ovarian tissue. According to Pepe et al. (2003) if a gene is expressed differently between a patient with cancer and a healthy one, then this gene could be used as a basis for population screening as long as it may be detected in blood or urine.

We analyse the mRNA expression of 1,536 clones of genes (Pepe et al., 2003). The dataset comprises ovarian tissue from 30 subjects with cancer and 23 control subjects. A gene would be an ideal candidate as cancer marker if its values are clearly differentiable in cancer tissue from those in normal tissue (over-expressed gene). Our aim is to seek for over-expressed genes. As a standard practice in the literature (see, e.g., Quackenbush, 2002), we used the logarithm of base 2 of the relative gene expression intensities.

For illustration purposes, we selected genes 9, 93 and 1033 which represent different types of sit-

uations commonly found in practice. Gene 9 is an example of a biomarker exhibiting larger values associated to the nondiseased group, therefore the classification rule behind the AUC and YI would need to be adjusted. In turn, gene 93 can be regarded as an ideal scenario. Finally, gene 1033 illustrates a bimodal situation where a subgroup of the group with ovarian cancer (diseased group) is well-differentiated from the group without ovarian cancer (nondiseased group). In Section 1.4, we discussed that if a marker is able to distinguish clearly a subset of diseased individuals, then it might be more of practical interest even if the overlap is large. However, this is not the case, because the diseased density falls entirely within the range of the nondiseased density. Posterior inference was obtained using 10,000 MCMC iterations after discarding 2,000 iterations as burn-in period. The same prior information described in Section 3.3.2 was used here. We monitor the MCMC chains convergence of the induced density in each group for three randomly selected relative gene expression intensities through their corresponding trace plots and Geweke diagnostics. Results are shown, respectively, in Figures 3.10–3.12 and Figure 3.13 and they do not suggest any lack of convergence. We also observed that the effective sample size was reasonably high enough (Figure 3.14). Finally, we assessed the goodness of fit of our models visually using QQ-plots of the model residuals (Dunn and Smyth, 1996) shown in Figure 3.15. Here, we observe no visible signs of misfit, which can be corroborated in the histograms shown in Figure 3.16.

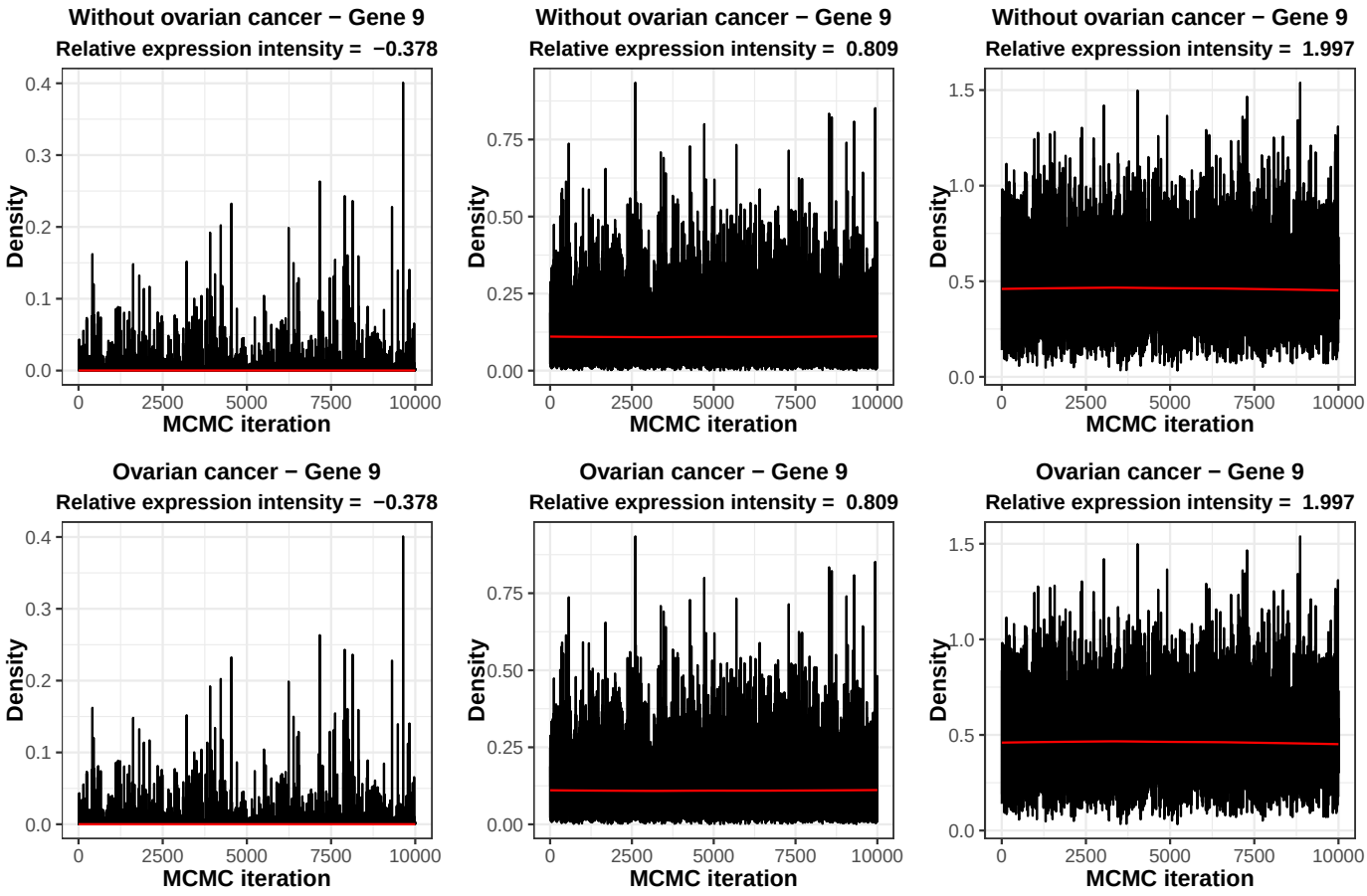

Figure 3.10: Gene 9. Trace plots after burn-in for three particular values of relative expression intensity for both groups (with and without ovarian cancer). The red line represents the mean regression line of a LOWESS model. We use it to check if there is any sort of trend in the chain.

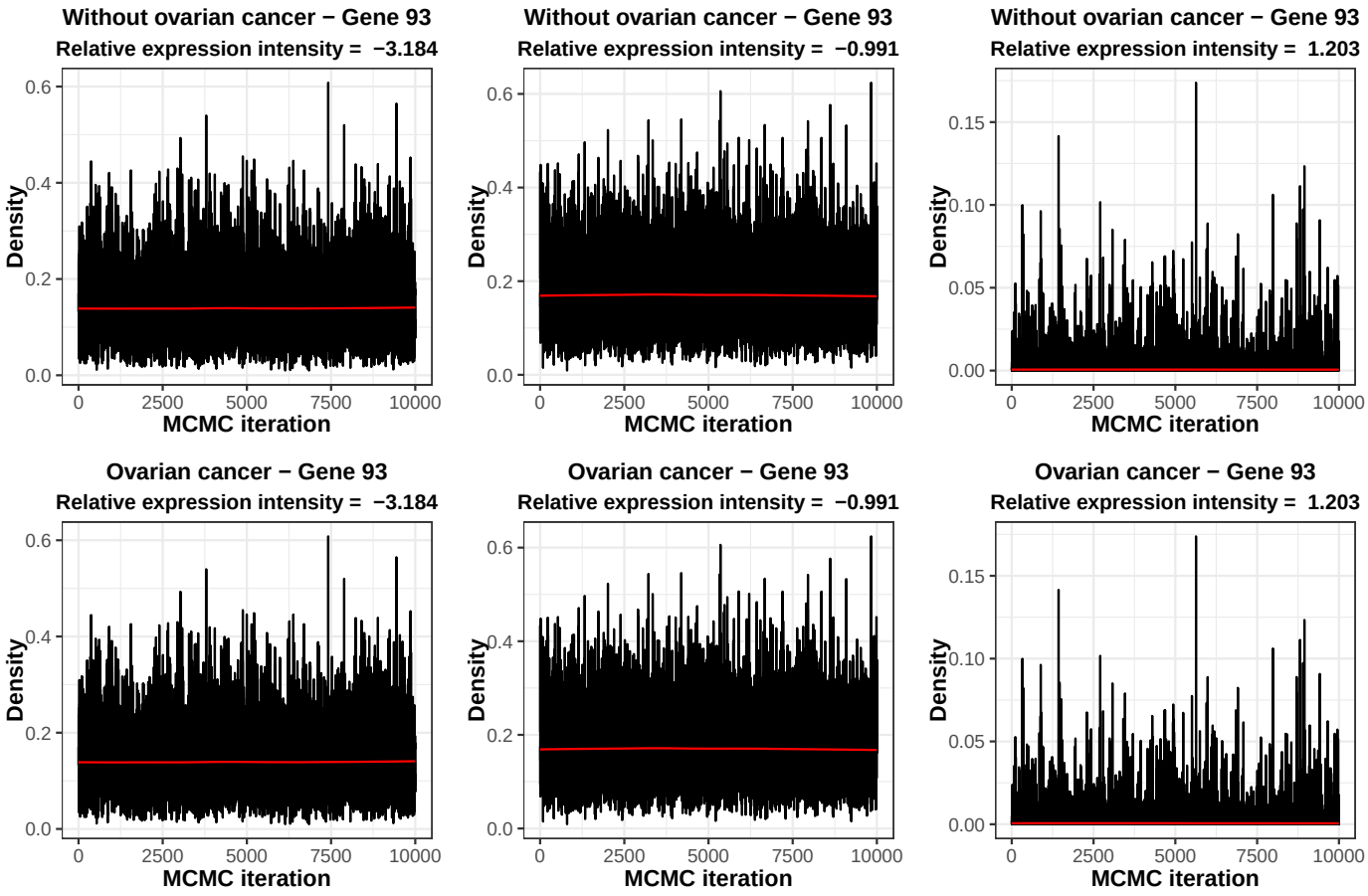

Figure 3.11: Gene 93. Trace plots after burn-in for three particular values of relative expression intensity for both groups (with and without ovarian cancer). The red line represents the mean regression line of a LOWESS model. We use it to check if there is any sort of trend in the chain.

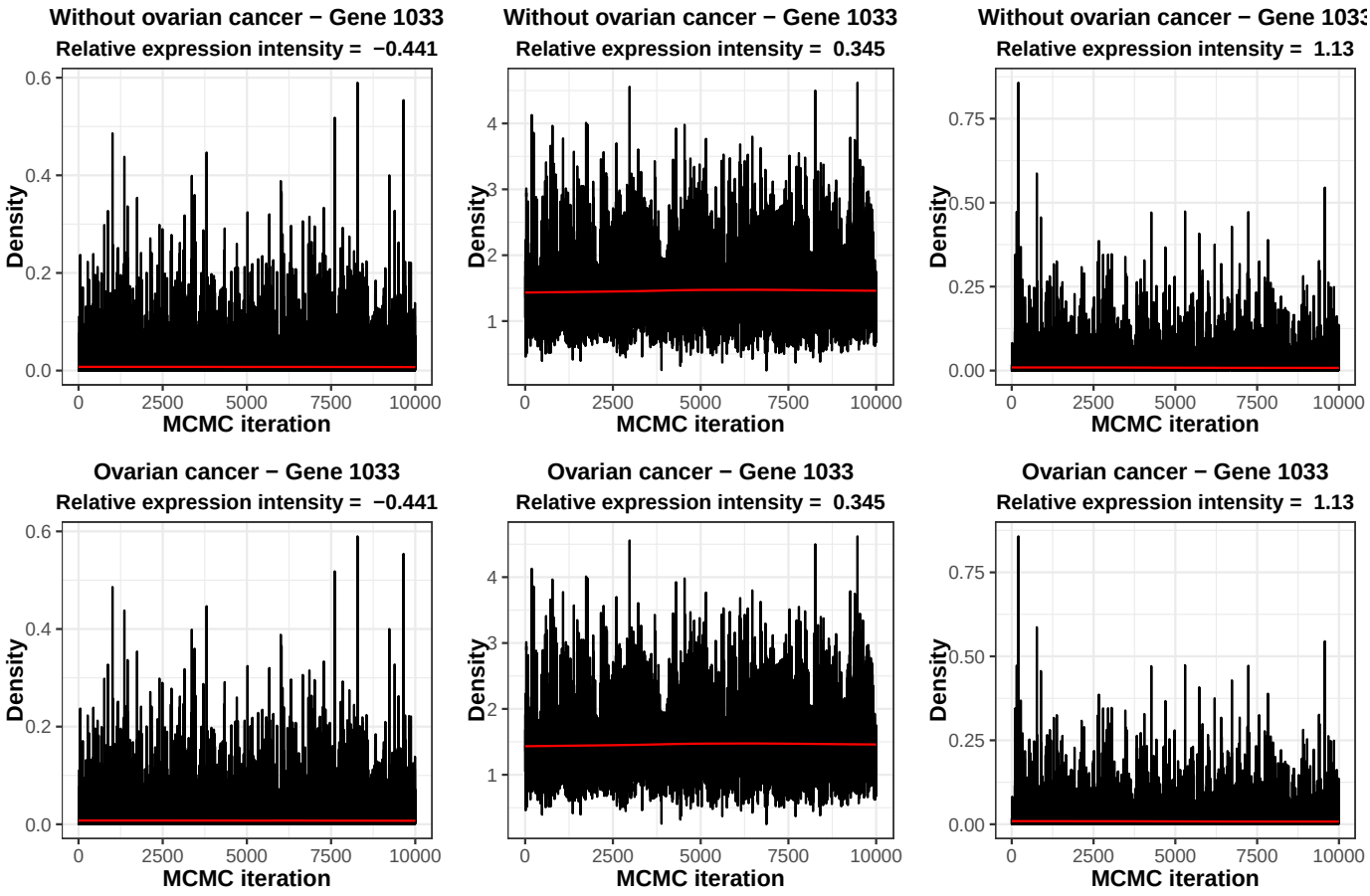

Figure 3.12: Gene 1033. Trace plots after burn-in for three particular values of relative expression intensity for both groups (with and without ovarian cancer). The red line represents the mean regression line of a LOWESS model. We use it to check if there is any sort of trend in the chain.

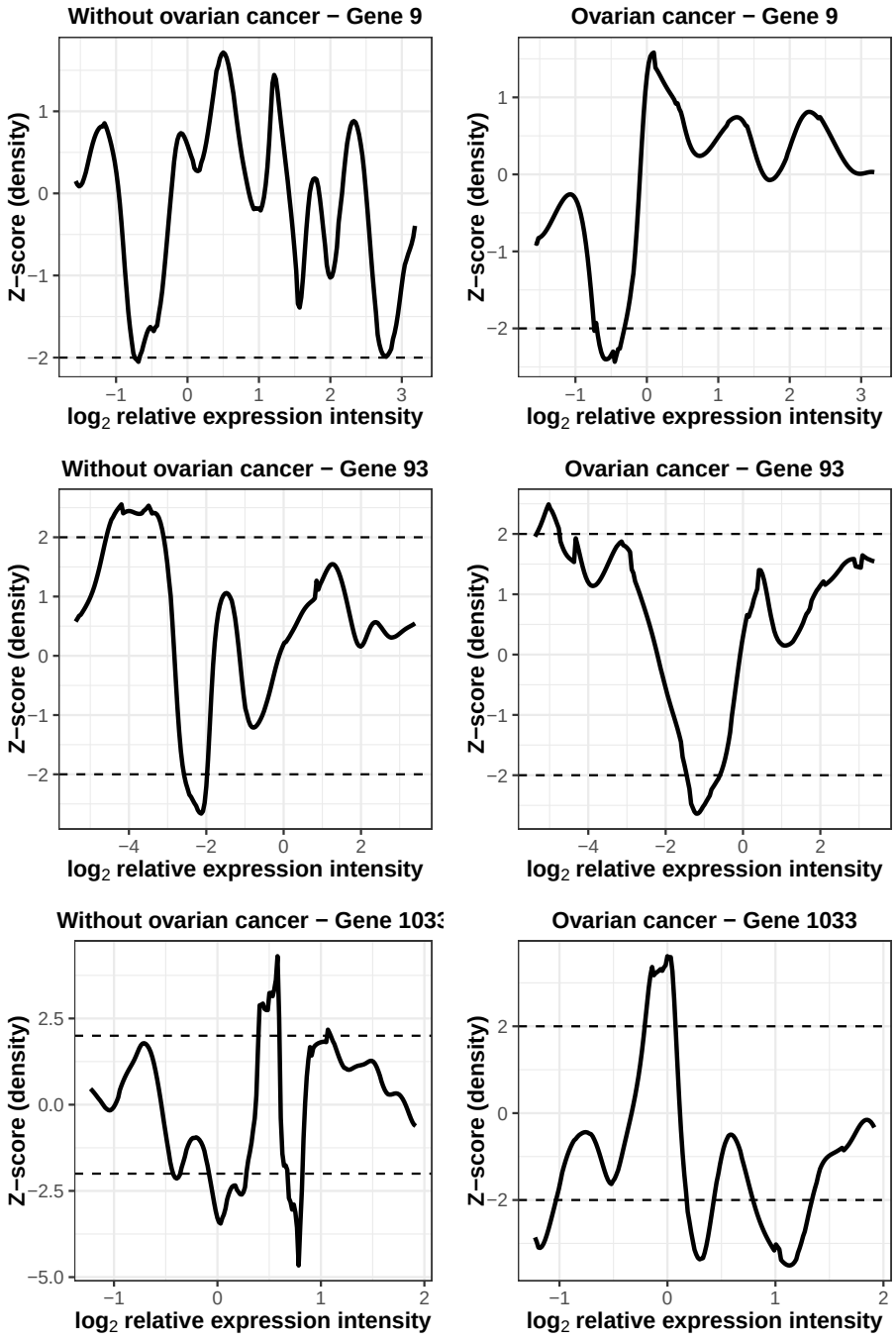

Figure 3.13: Geweke diagnostic plots for the estimated densities of gene 9, 93 and 1033 for the groups without cancer (left) and with cancer (right).

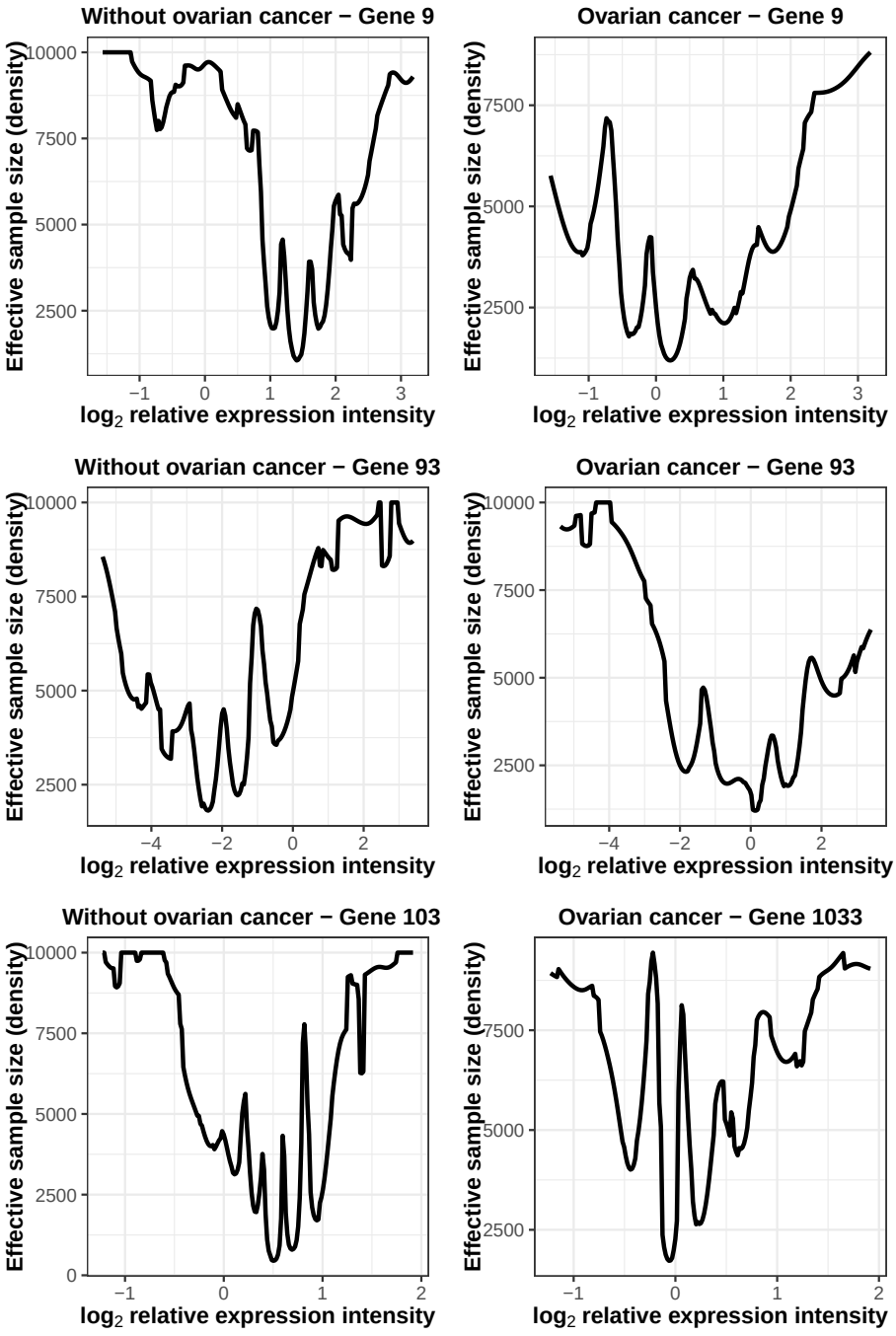

Figure 3.14: Effective sample size for the estimated densities of gene 9, 93 and 1033 for the groups without cancer (left) and with cancer (right).

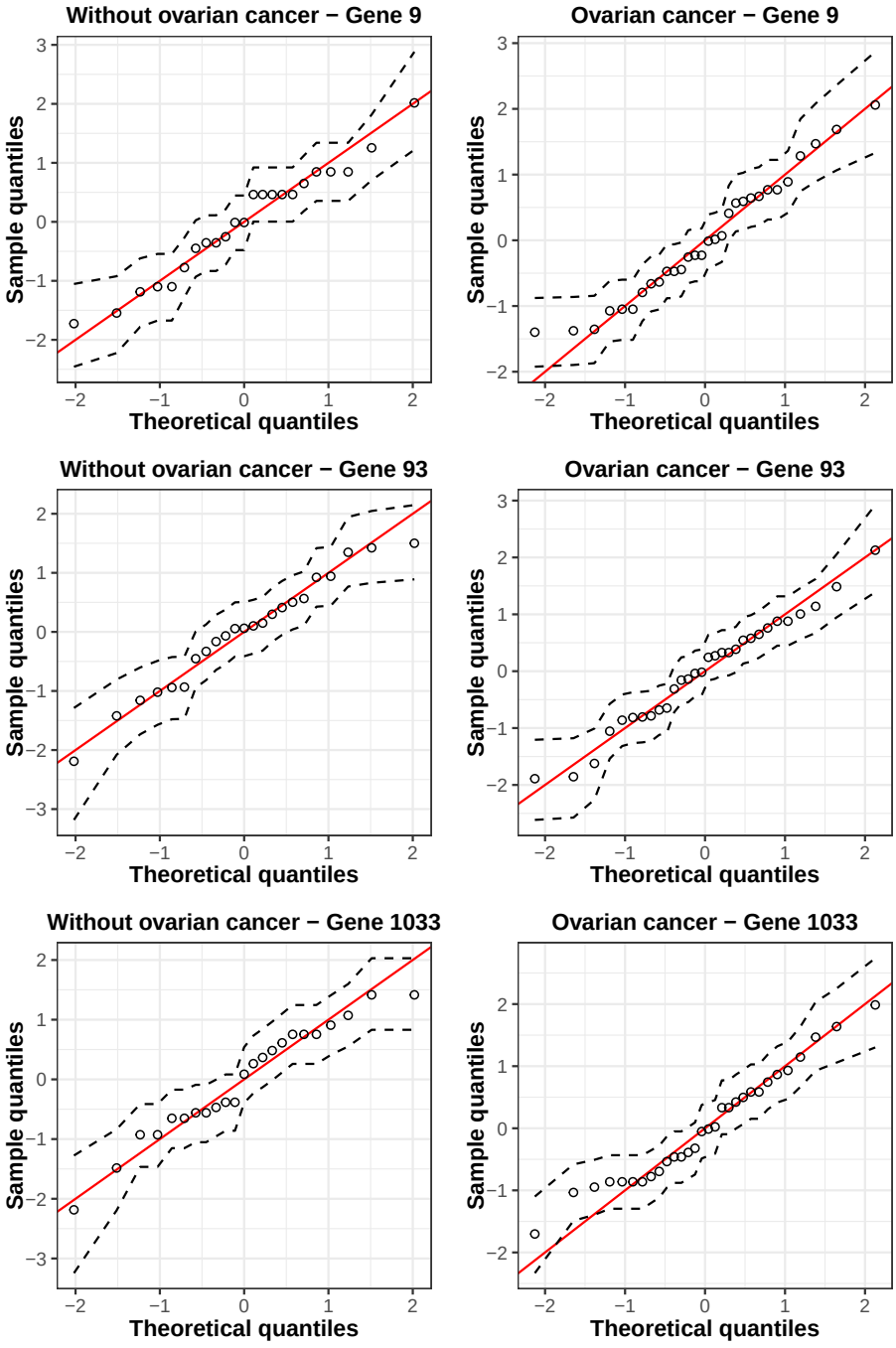

Figure 3.15: QQ-plots of the relative expression intensities. Theoretical quantiles correspond to a standard normal distribution. The circles denote the posterior mean quantiles and the dashed lines represent the corresponding 95% credible bands.

The estimated densities under the DPM model for each of the genes considered are depicted in Figure 3.16. Table 3.5 displays the estimated OVL based on (3.11) and (3.12). For comparison purposes we included the estimates using the kernel approach described in Section 3.3.2, the confidence intervals shown are based on 1,000 bootstrap replicates. Genes 9 and 93, depicted in the top and middle rows of Figure 3.16, respectively, show a similar behaviour and a low OVL, therefore they are clear examples of differentially expressed genes. Thus, both might be used as basis for screening purposes. In turn, gene 1033 presents a larger OVL and, as mentioned above, the densities are within the same range, making this gene less appropriate to be considered as a candidate.

Table 3.5: Estimated coefficient of overlap for each method. For the Bayesian methods, the posterior means (95% credible intervals) are reported. For kernel estimates, bootstrap 95% confidence intervals are reported.

Gene	**DPM**	**DPM-BB**	**Kernel**	**Kernel-emp**
Gene 9	0.185 (0.0813, 0.323)	0.164 (0.046, 0.359)	0.194 (0.063, 0.324)	0.133 (0.033, 0.277)
Gene 93	0.159 (0.064, 0.292)	0.123 (0.025, 0.303)	0.170 (0.048, 0.272)	0.100 (0.000, 0.233)
Gene 1033	0.483 (0.316, 0.666)	0.485 (0.272, 0.757)	0.527 (0.323, 0.691)	0.487 (0.277, 0.674)

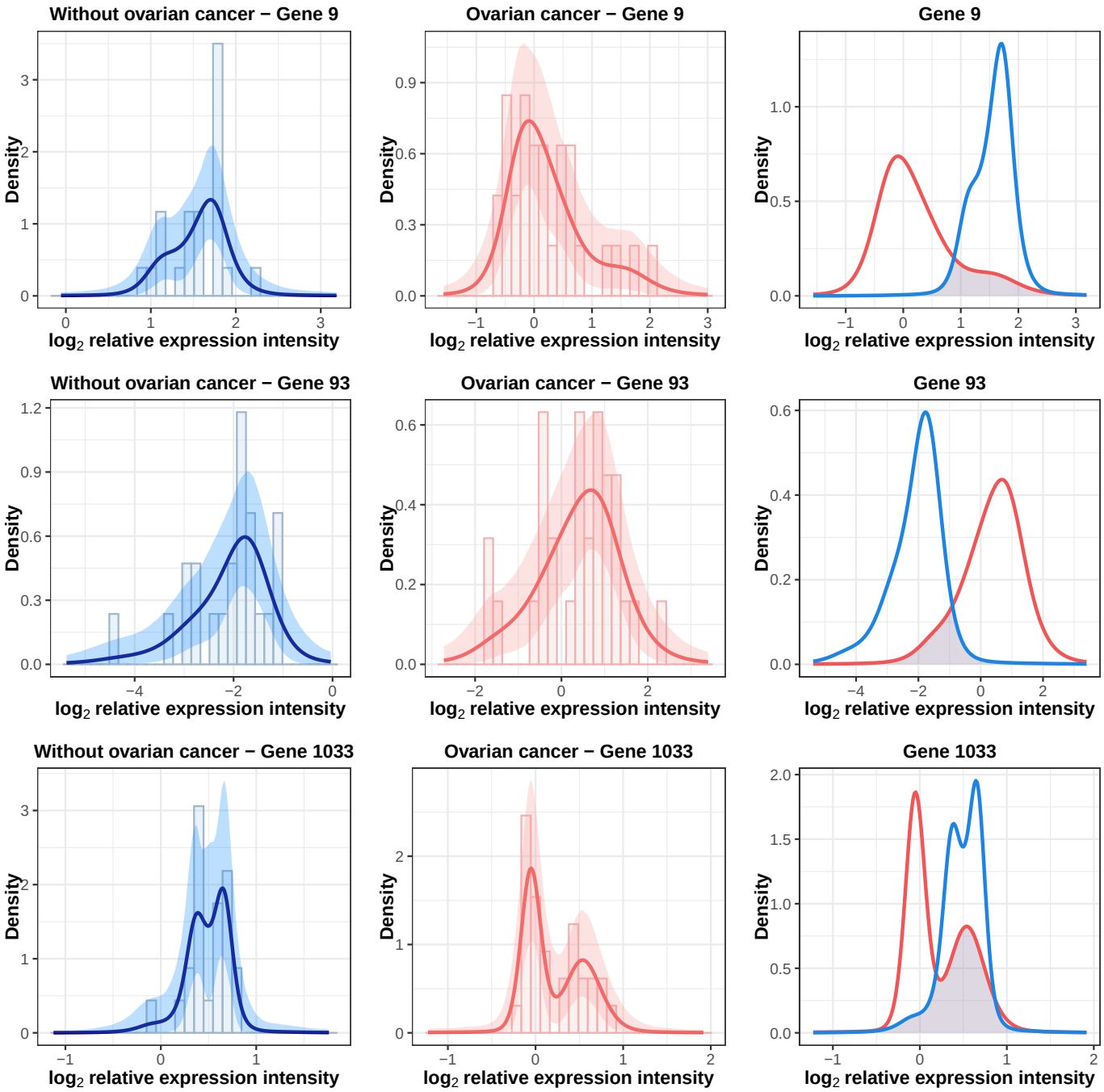

Figure 3.16: Histograms and estimated densities (posterior mean and pointwise 95% credible bands) of the $\log_2$ relative gene expression intensities in both groups (with and without ovarian cancer) for gene 9 (top row), gene 93 (middle row) and gene 1033 (bottom row).

In conclusion, genes 9 and 93 appear to be good candidates as biomarkers for ovarian cancer. While gene 1033 might not be a suitable choice, because it presents a larger OVL and the densities of both groups are within the same range.

3.5 Discussion

We have developed a flexible framework to estimate and conduct inference about the coefficient of overlap. We followed a Bayesian nonparametric approach that combines Dirichlet process mixtures and the Bayesian bootstrap, thus relaxing restrictive assumptions commonly used in practice. Further, point and interval estimates are available into a single integrated framework. The simulation study revealed that our estimators are able to produce accurate estimates of the true overlap coefficient under a variety of scenarios and sample sizes commonly encountered in diagnostic studies. These results also suggests that our estimators might be consistent and, particularly, that the DPM-BB estimator is on par with or exceeds current nonparametric estimators of the coefficient of overlap. We investigated the empirical coverage probability of the 95% credible intervals of our proposed estimators. We found that the "plain vanilla" estimator leads to lower coverages, especially, for situations where the true overlap is close to one. In turn, the DPM-BB estimator showed higher coverage for almost all cases. However, the width of its credible intervals was larger as well, meaning higher uncertainty. Finally, our methods were illustrated through an application concerning the search for biomarkers for ovarian cancer. This application allowed us to demonstrate that the coefficient of overlap as a summary measure of diagnostic accuracy has undeniable advantages over traditional measures, such as the AUC or YI: i) it is non-directional and ii) has an appealing graphical interpretation.

It is worth noting that in cases where the main interest is to rank different diagnostic tests (as in our application of gene expression data), the coefficient of overlap serves as a very appealing alternative to traditional measures associated to the receiver operating characteristic curve. However, the OVL is not suitable to determine a cut-off value(s) to screen subjects in practice. In such case, measures based on the ROC curve, or its generalisation, are possibly better suited for this task. Hence, our aim is not for the overlap coefficient to replace such measures, but rather to serve as a complement.

Chapter 4

Bayesian nonparametric inference for the covariate-specific coefficient of overlap

As it was already mentioned in the previous chapters, the coefficient of overlap has gained unarguably popularity as a summary measure of the discriminatory performance of a continuous diagnostic test. In many practical situations, covariates such as age and/or sex, may influence the behaviour of the test and ultimately, its accuracy. Therefore, it is essential to take this covariate information into account. We propose a Bayesian nonparametric modelling framework, based on Dirichlet process mixtures, to estimate and conduct inference about the covariate-specific coefficient of overlap, thus extending the coefficient of overlap to the regression setting. We follow a joint approach where the test outcomes and the covariates are both treated as random and are modelled jointly, therefore allowing to accommodate any type of functional form of how the covariates influence the test outcomes distributions. Our proposed estimator performs well on a variety of simulated scenarios and it is able to recover the true covariate-specific coefficient of overlap curves. We illustrate our proposed methodology through its application to two different real datasets. The first application investigates if and how the accuracy of glucose levels as a marker for diabetes changes with age. The second application concerns the search for biomarkers for Alzheimer's disease and how these may be affected by the sex and age of the patients.

4.1 Introduction

In many situations, the behaviour of a diagnostic test may be influenced by particular characteristics of the subjects. For example, a diagnostic test might have a good discriminatory ability in women, but it might perform poorly in men. Therefore, failure to incorporate this information may result in misleading or oversimplified conclusions about the accuracy of the test. In order to identify the optimal and suboptimal populations where to perform the tests on, it is crucial to take covariate information into account.

Popular methods to evaluate the discriminatory ability of continuous tests in presence of covariates are based on the ROC curve. Pepe (2003, Chapter 6) and Zhou et al. (2011, Chapter 8) are excellent resources for a frequentist overview of ROC curve regression. However, existing work under the Bayesian perspective is rather limited, for instance, the work by Inácio de Carvalho et al. (2013) and Rodríguez and Martínez (2014), who developed a Bayesian nonparametric framework for ROC regression based on dependent Dirichlet processes and Gaussian processes, respectively. Recently, de Carvalho et al. (2020) proposed an affinity measure based on the so-called Hellinger affinity as an alternative summary measure of diagnostic accuracy and defined its covariate-specific version as follows

$$\kappa(\boldsymbol{x}) = \int_{-\infty}^{\infty} \sqrt{f_{\bar{D}}(y \mid \boldsymbol{x})}\sqrt{f_D(y \mid \boldsymbol{x})}dy,$$

where $\boldsymbol{x}$ is a vector of covariates. Their proposed summary measure has among its main advantages being able to accommodate a situation in which both low and high test outcomes are associated with the absence of the disease and the tests results in the diseased group lie in between the two modes of the nondiseased test outcomes distribution, such that, there is a perfect separation between the distributions of both groups. This implies that the AUC or the YI would falsely indicate that the test classification is no better than chance. Additionally, their measure does not require to know in advance if larger values of the test are more indicative of the disease and it avoids computing conditional quantiles over a grid of covariate values, which is computationally demanding.

All these discussed advantages are also shared by the coefficient of overlap and both range from 0 to 1, with lower values indicating higher discrimination power and larger values suggesting useless tests. Further, both summary measures are invariant to monotone increasing transformations. Figure

4.1 shows a comparison between both measures under the same covariate-specific scenario taken from de Carvalho et al. (2020). We observe that both curves present a similar behaviour. However, the covariate-specific OVL values are lower than those from the affinity measure. Further, the OVL has an appealing graphical interpretation, connected directly to the resulting OVL value, the larger it is, the higher the overlap between the densities is. In turn, the affinity measure scale does not have provide such information about the densities. It is difficult to know what a resulting value of this measure means in terms of overlap between the densities.

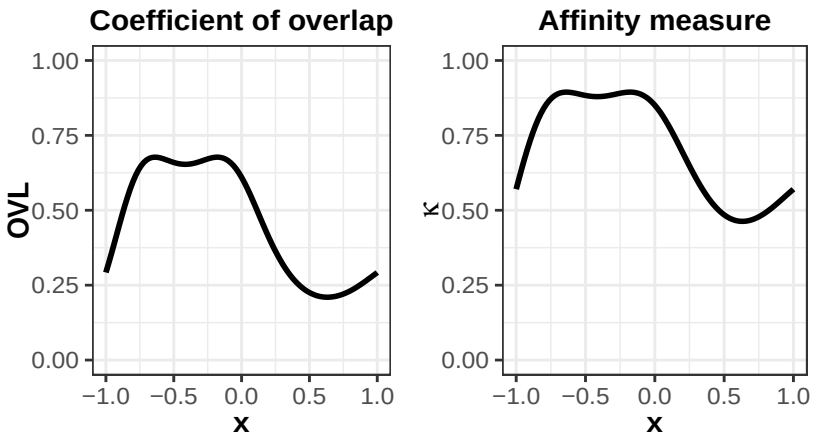

Figure 4.1: Covariate-specific OVL (left) and affinity measure (right) comparison. The true data generation process is given by: $Y_{\bar{D}} \sim \mathrm{N}\left(\sin\left[\pi(x_{\bar{D}}+1)\right], 0.5^2\right)$, $Y_D \sim \mathrm{N}\left(0.5+x_D^2, 1\right)$.

The no covariates case for the OVL was already considered in Chapter 3. However, to the best of our knowledge, no methodology has been proposed yet for conditional OVL estimation. The outcomes of a medical diagnostic test usually present a complex structure (e.g., multimodality, skewness, excess of kutosis, etc.) which can be difficult to capture with a single parametric model. Furthermore, the effect of the covariates on the test outcomes might be non-linear. This motivated us to use a class of models that is both flexible enough to represent a wide variety of density shapes and that allows to handle any functional form of the covariates. The methodology followed by de Carvalho et al. (2020) for the covariate-specific affinity is similar to the one we present in this chapter for the covariate-specific OVL. However, unlike our proposed approach, the authors directly modelled the conditional densities with a single-weights DDP and an additive model on the mean of each distribution.

Our focus in this chapter is to develop a method to flexibly estimate and conduct inference about the

covariate-specific OVL. Given that the covariate-specific OVL depends on the underlying conditional density functions, the problem of estimating it, reduces to the problem of accurately estimating the underlying conditional densities. For this end, we follow a joint approach based on Dirichlet process mixtures (Müller et al., 1996). Under this approach, both test outcomes and covariates are considered as random and are modelled jointly. To allow our model to handle binary covariates, we replace the multivariate normal kernel proposed by Müller et al. (1996) by a normal/Bernoulli one. It is worth to mention that our approach can be easily extended to handle categorical covariates, replacing the Bernoulli kernel by a Multinomial kernel. Then, by using the properties of the normal distribution, we can easily compute the conditional distributions. This framework has some advantages compared to traditional conditional approaches, such as linear regression models. It does not require to specify any particular functional form of how the covariates affect the test outcomes, because when modelling both variables jointly, any non-linearity or heterogeneity is automatically accommodated (Rossi, 2014, p. 93). This means that other functions of interest, such as the conditional mean (regression function), quantile or variance functions (available for free as well) are fully data driven because they depend implicitly on the covariates. Whereas, for example, in linear regression models there are usually more restrictive assumptions that inhibit capturing other features of data, such as heteroscedasticity.

4.2 Methods

The estimation of the covariate-specific OVL can be regarded as a two-step procedure. In the first step, we must estimate the conditional density functions in the nondiseased and diseased group. The second step involves computing the integral in (1.4), and for this end, we resort to a numerical integration method, specifically, the trapezoidal rule.

4.2.1 Conditional density estimation

Let $\left\{\left(y_{\bar{D}i}, \boldsymbol{x}^c_{\bar{D}i}, \boldsymbol{x}^d_{\bar{D}i}\right)\right\}_{i=1}^{n_{\bar{D}}}$ and $\left\{\left(y_{Dj}, \boldsymbol{x}^c_{Dj}, \boldsymbol{x}^d_{Dj}\right)\right\}_{j=1}^{n_D}$ be independent random samples of size $n_{\bar{D}}$ and n_D from the nondiseased and diseased population, respectively. Here, $y_{\bar{D}i}$ and y_{Dj} represent the test outcomes of the nondiseased and diseased groups, for $i = 1, \ldots, n_{\bar{D}}$ and $j = 1, \ldots, n_D$; $\boldsymbol{x}^c_{\bar{D}i} =$

$(x^c_{\bar{D}i,1}, \ldots, x^c_{\bar{D}i,p-1})'$ and $\boldsymbol{x}^c_{Dj} = (x^c_{Dj,1}, \ldots, x^c_{Dj,p-1})'$ are two $(p-1)$–dimensional vectors of continuous covariates; and $\boldsymbol{x}^d_{\bar{D}i} = (x^d_{\bar{D}i,1}, \ldots, x^d_{\bar{D}i,q})'$ and $\boldsymbol{x}^d_{Dj} = (x^d_{Dj,1}, \ldots, x^d_{Dj,q})'$ are two q–dimensional vectors of binary covariates.

For the sake of simplicity and to ease notation, we will assume that the covariates are the same in both groups. However, our modelling approach can easily deal with different covariates in each group. For example, consider the covariates sex and severity of disease. Sex (female and male) can be measured in both groups (nondiseased and diseased), but severity of disease can only be measured across the diseased group. First, we would compare the nondiseased female subjects against females with mild disease and then nondiseased females against females with severe disease. Then, we would repeat the same procedure for male subjects. Alternatively, we could consider three disease groups: nondiseased, mild and severely diseased. However, in this case we would need to seek for a further extension to the multi-class case of the OVL.

In what follows, we will describe our modelling procedure only for the diseased population as a similar one is applicable to the nondiseased group. We will model the joint density as a DPM using a normal/Bernoulli kernel, that is, a multivariate normal kernel is used for the test outcomes and the continuous covariates and independent Bernoulli distributions are used for the binary covariates. Thus, the joint density for the diseased group can be expressed as follows

$$
\begin{aligned}
f_D(y_D, \boldsymbol{x}^c_D, \boldsymbol{x}^d_D) &= \int \phi\left(y_D, \boldsymbol{x}^c_D \mid \boldsymbol{\mu}_D, \boldsymbol{\Sigma}_D\right) \prod_{k=1}^{q} \mathrm{Bern}(x^d_{D,k} \mid \pi_{Dk}) \mathrm{d}G_D(\boldsymbol{\mu}_D, \boldsymbol{\Sigma}_D, \boldsymbol{\pi}_D), \\
G_D &\sim \mathrm{DP}(\alpha_D, G^*_D(\boldsymbol{\mu}_D, \boldsymbol{\Sigma}_D, \boldsymbol{\pi}_D)),
\end{aligned} \tag{4.1}
$$

where $\phi\left(\cdot \mid \boldsymbol{\mu}_D, \boldsymbol{\Sigma}_D\right)$ is the density function of a p–variate normal distribution with mean vector $\boldsymbol{\mu}_D$ and covariance matrix $\boldsymbol{\Sigma}_D$, $\mathrm{Bern}(\cdot \mid \pi_{Dk})$ stands for a Bernoulli density function with probability of success π_{Dk}, for $k = 1, \ldots, q$, and $\boldsymbol{\pi}_D = (\pi_{D1}, \ldots, \pi_{Dq})$. Note that we can easily incorporate categorical covariates into our model by simply replacing the Bernoulli kernel by a Multinomial one and choosing an appropriate prior, such as a Dirichlet distribution. The mixing distribution G_D follows a Dirichlet process with concentration parameter $\alpha_D > 0$ and baseline distribution G^*_D. The baseline distribution G^*_D encapsulates any prior knowledge that might be known about G_D, while the parameter α_D governs the concentration of G_D around G^*_D, with smaller (larger) values implying higher (lower)

uncertainty. The constructive definition of the DP (Sethuraman, 1994), is undeniably its most popular representation, under which G_D can be written as an infinite sum of point masses as follows

$$G_D(\cdot) = \sum_{l=1}^{\infty} \omega_{Dl} \delta_{\boldsymbol{\theta}_{Dl}}(\cdot), \quad \omega_{Dl} = \begin{cases} v_{D1}, & \text{if} \quad l = 1 \\ v_{Dl} \prod_{t<l} (1 - v_{Dt}), & \text{if} \quad l \geq 2 \end{cases}, \tag{4.2}$$

where δ_a denotes a point mass at a, $\boldsymbol{\theta}_{Dl} = (\boldsymbol{\mu}_{Dl}, \boldsymbol{\Sigma}_{Dl}, \boldsymbol{\pi}_{Dl}) \overset{\text{iid}}{\sim} G_D^*(\boldsymbol{\mu}_D, \boldsymbol{\Sigma}_D, \boldsymbol{\pi}_D)$, $\boldsymbol{\pi}_{Dl} = (\pi_{Dl,1}, \ldots, \pi_{Dl,q})$ and $v_{Dl} \overset{\text{iid}}{\sim} \text{Beta}(1, \alpha_D)$ are mutually independent for $l \geq 1$. For conjugacy reasons, we take

$$\begin{aligned} G_D^* (\boldsymbol{\mu}_D, \boldsymbol{\Sigma}_D, \boldsymbol{\pi}_D \mid \boldsymbol{\mu}_{D0}, \boldsymbol{V}_{D0}, \nu_{D0}, \boldsymbol{\Psi}_{D0}, \boldsymbol{a}_{\boldsymbol{\pi}_D}, \boldsymbol{b}_{\boldsymbol{\pi}_D}) &= \text{N}_p(\boldsymbol{\mu}_D \mid \boldsymbol{\mu}_{D0}, \boldsymbol{V}_{D0}) \text{IW}_p(\boldsymbol{\Sigma}_D \mid \nu_{D0}, \boldsymbol{\Psi}_{D0}) \\ &\quad \times \prod_{k=1}^{q} \text{Beta}(\pi_{Dk} \mid a_{\pi_{Dk}}, b_{\pi_{Dk}}), \end{aligned}$$

where $\text{IW}_p(\boldsymbol{\Sigma}_D \mid \nu_{D0}, \boldsymbol{\Psi}_{D0})$ denotes that the positive-definite matrix $\boldsymbol{\Sigma}_D$ follows an inverse Wishart distribution. We are using the parametrization where the expected value is equal to $\frac{1}{\nu_{D0}-p-1}\boldsymbol{\Psi}_{D0}$ and ν_{D0} degrees of freedom. The hyperparameters $\boldsymbol{a}_{\boldsymbol{\pi}_D} = (a_{\pi_{D1}}, \ldots, a_{\pi_{Dq}})$ and $\boldsymbol{b}_{\boldsymbol{\pi}_D} = (b_{\pi_{D1}}, \ldots, b_{\pi_{Dq}})$ are regarded as fixed. We complete our model by placing priors on $\alpha_D, \boldsymbol{\mu}_{D0}, \boldsymbol{V}_{D0}$ and $\boldsymbol{\Psi}_{D0}$. Specifically, we let $\alpha_D \sim \Gamma(a_{\alpha_D}, b_{\alpha_D})$, where $\Gamma(a_{\alpha_D}, b_{\alpha_D})$ denotes a gamma distribution with shape a_{α_D} and rate b_{α_D}. For the hyperpriors, we set $\boldsymbol{\mu}_{D0} \sim \text{N}_p(\boldsymbol{m}_D, \boldsymbol{S}_D)$, $\boldsymbol{V}_{D0} \sim \text{IW}_p(r_D, \boldsymbol{Q}_D)$ and $\boldsymbol{\Psi}_{D0} \sim \text{W}_p(\nu_{D1}, \boldsymbol{\Psi}_{D1})$, where $\text{W}_p(\nu_{D1}, \boldsymbol{\Psi}_{D1})$ denotes a Wishart distribution with mean $\nu_{D1}\boldsymbol{\Psi}_{D1}$ and ν_{D1} degrees of freedom. All the hyperparameters are regarded as fixed.

To facilitate posterior simulation, we will employ a truncated version of the DPM model proposed by Ishwaran and Zarepour (2000), which caps the infinite representation in (4.2) to a finite number, say L_D. They showed that for higher order weights in the stick breaking representation $\mathbb{E}\left(\sum_{l=L_D}^{\infty} \omega_{Dl}\right) = \left(\frac{\alpha_D}{\alpha_D+1}\right)^{L_D-1}$. For instance, if we set $L_D = 20$ and $\alpha_D = 1$, then $\mathbb{E}\left(\sum_{l=L_D}^{\infty} \omega_{Dl}\right) < 10^{-6}$. Hence, we can rewrite (4.2) as

$$G_D(\cdot) = \sum_{l=1}^{L_D} \omega_{Dl} \delta_{\boldsymbol{\theta}_{Dl}}(\cdot).$$

The weights ω_{Dl}, now follow a truncated stick-breaking construction, that is, $\omega_{D1} = v_{D1}$, $\omega_{Dl} = v_{Dl} \prod_{t<l} (1 - v_{Dt})$, for $l = 2, \ldots, L_D$. The inputs of the weights are distributed according to a beta

distribution, i.e., $v_{D1}, \ldots, v_{DL_D-1} \overset{\text{iid}}{\sim} \text{Beta}(1, \alpha_D)$, and to ensure they add up to one, we set $v_{DL_D} = 1$. Finally, we can express the joint density in (4.1) as

$$f_D(y_D, \boldsymbol{x}_D^c, \boldsymbol{x}_D^d) = \sum_{l=1}^{L_D} \omega_{Dl} \phi\left(y_D, \boldsymbol{x}_D^c \mid \boldsymbol{\mu}_{Dl}, \boldsymbol{\Sigma}_{Dl}\right) \prod_{k=1}^{q} \text{Bern}(x_{D,k}^d \mid \pi_{Dl,k}). \tag{4.3}$$

Note that (4.3) resembles a finite normal/Bernoulli mixture. However, in this case L_D do not represent the number of components, but an upper bound on such number, instead. We can use the expressions in (3.4) derived by Liu (1996) to aid the selection of L_D.

Although not of direct interest in our setting, the marginal density of the covariates can be recovered as follows

$$\begin{aligned}
f_D(\boldsymbol{x}_D^c, \boldsymbol{x}_D^d) &= \int_{-\infty}^{\infty} f_D(y_D, \boldsymbol{x}_D^c, \boldsymbol{x}_D^d) dy_D \\
&= \int_{-\infty}^{\infty} \sum_{l=1}^{L_D} \omega_{Dl} \phi\left(y_D, \boldsymbol{x}_D^c \mid \boldsymbol{\mu}_{Dl}, \boldsymbol{\Sigma}_{Dl}\right) \prod_{k=1}^{q} \text{Bern}(x_{Dj}^d \mid \pi_{Dl,k}) dy_D \\
&= \sum_{l=1}^{L_D} \omega_{Dl} \prod_{k=1}^{q} \text{Bern}(x_{Dk}^d \mid \pi_{Dl,k}) \int_{-\infty}^{\infty} \phi\left(y_D, \boldsymbol{x}_D^c \mid \boldsymbol{\mu}_{Dl}, \boldsymbol{\Sigma}_{Dl}\right) dy_D \\
&= \sum_{l=1}^{L_D} \omega_{Dl} \prod_{k=1}^{q} \text{Bern}(x_{Dk}^d \mid \pi_{Dl,k}) \phi\left(\boldsymbol{x}_D^c \mid \boldsymbol{\mu}_{Dl}^x, \boldsymbol{\Sigma}_{Dl}^{xx}\right).
\end{aligned} \tag{4.4}$$

Equation (4.4) follows from the properties of the multivariate normal distribution. Recall that marginally, the distribution of a subset of a multivariate normal random vector is normal with mean vector $\boldsymbol{\mu}_{Dl}^x$ and covariance matrix $\boldsymbol{\Sigma}_{Dl}^{xx}$, where $\boldsymbol{\mu}_{Dl}^x$ and $\boldsymbol{\Sigma}_{Dl}^{xx}$ are the mean vector and covariance matrix induced by the joint $\text{N}_p(y_D, \boldsymbol{x}_D^c \mid \boldsymbol{\mu}_{Dl}, \boldsymbol{\Sigma}_{Dl})$ distribution, that is

$$\boldsymbol{\mu}_{Dl} = \begin{pmatrix} \mu_{Dl,1} \\ \mu_{Dl,2} \\ \vdots \\ \mu_{Dl,p} \end{pmatrix} = \begin{pmatrix} \mu_{Dl}^y \\ \boldsymbol{\mu}_{Dl}^x \end{pmatrix} \text{ and } \boldsymbol{\Sigma}_{Dl} = \begin{pmatrix} \Sigma_{Dl,11} & \Sigma_{Dl,12} & \ldots & \Sigma_{Dl,1p} \\ \Sigma_{Dl,21} & \Sigma_{Dl,22} & \ldots & \Sigma_{Dl,2p} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{Dl,p1} & \Sigma_{Dl,p2} & \ldots & \Sigma_{Dl,pp} \end{pmatrix} = \begin{pmatrix} \Sigma_{Dl}^{yy} & \boldsymbol{\Sigma}_{Dl}^{yx} \\ \boldsymbol{\Sigma}_{Dl}^{xy} & \boldsymbol{\Sigma}_{Dl}^{xx} \end{pmatrix}.$$

Now, the conditional density function is given by

$$\begin{aligned} f_D(y_D \mid \boldsymbol{x}_D^c, \boldsymbol{x}_D^d) &= \frac{f_D(y_D, \boldsymbol{x}_D^c, \boldsymbol{x}_D^d)}{f_D(\boldsymbol{x}_D^c, \boldsymbol{x}_D^d)} \\ &= \frac{\sum_{l=1}^{L_D} \omega_{Dl} \phi\left(y_D, \boldsymbol{x}_D^c \mid \boldsymbol{\mu}_{Dl}, \boldsymbol{\Sigma}_{Dl}\right) \prod_{k=1}^{q} \text{Bern}(x_{Dk}^d \mid \pi_{Dl,k})}{\sum_{l=1}^{L_D} \omega_{Dl} \phi\left(\boldsymbol{x}_D^c \mid \boldsymbol{\mu}_{Dl}^x, \boldsymbol{\Sigma}_{Dl}^{xx}\right) \prod_{k=1}^{q} \text{Bern}(x_{Dk}^d \mid \pi_{Dl,k})} \\ &= \frac{\sum_{l=1}^{L_D} \omega_{Dl} \phi\left(\boldsymbol{x}_D^c \mid \boldsymbol{\mu}_{Dl}^x, \boldsymbol{\Sigma}_{Dl}^{xx}\right) \phi\left(y_D \mid \tilde{\mu}_{Dl}, \tilde{\sigma}_{Dl}^2\right) \prod_{k=1}^{q} \text{Bern}(x_{Dk}^d \mid \pi_{Dl,k})}{\sum_{l=1}^{L_D} \omega_{Dl} \phi\left(\boldsymbol{x}_D^c \mid \boldsymbol{\mu}_{Dl}^x, \boldsymbol{\Sigma}_{Dl}^{xx}\right) \prod_{k=1}^{q} \text{Bern}(x_{Dk}^d \mid \pi_{Dl,k})}. \end{aligned} \tag{4.5}$$

Equation (4.5) follows from the chain rule and the properties of the multivariate normal distribution. Dropping the parameters from the notation, we know that $p(y_D, \boldsymbol{x}_D^c) = p(y_D \mid \boldsymbol{x}_D^c)p(\boldsymbol{x}_D^c)$, where $p(y_D \mid \boldsymbol{x}_D^c)$ is the density function of a normal distribution with mean $\tilde{\mu}_{Dl} = \mu_{Dl}^y + \boldsymbol{\Sigma}_{Dl}^{yx}(\boldsymbol{\Sigma}_{Dl}^{xx})^{-1}(\boldsymbol{x}_D^c - \boldsymbol{\mu}_{Dl}^x)$ and variance $\tilde{\sigma}_{Dl}^2 = \Sigma_{Dl}^{yy} - \boldsymbol{\Sigma}_{Dl}^{yx}(\boldsymbol{\Sigma}_{Dl}^{xx})^{-1}\boldsymbol{\Sigma}_{Dl}^{xy}$, and $p(\boldsymbol{x}_D^c)$ is the density of a multivariate normal distribution with mean vector $\boldsymbol{\mu}_{Dl}^{xx}$ and covariance matrix $\boldsymbol{\Sigma}_{Dl}^{xx}$. Further, note that Equation (4.5) can be rewritten as

$$f_D\left(y_D \mid \boldsymbol{X}_D^c = \boldsymbol{x}_D^c, \boldsymbol{X}_D^d = \boldsymbol{x}_D^d\right) = \sum_{l=1}^{L_D} q_{Dl}(\boldsymbol{x}_D^c, \boldsymbol{x}_D^d) \phi\left(y_D \mid \tilde{\mu}_{Dl}, \tilde{\sigma}_{Dl}^2\right) \tag{4.6}$$

where

$$q_{Dl}(\boldsymbol{x}_D^c, \boldsymbol{x}_D^d) = \frac{\omega_{Dl} \phi\left(\boldsymbol{x}_D^c \mid \boldsymbol{\mu}_{Dl}^x, \boldsymbol{\Sigma}_{Dl}^{xx}\right) \prod_{k=1}^{q} \text{Bern}(x_{D,k}^d \mid \pi_{Dl,k})}{f_D(\boldsymbol{x}_D^c, \boldsymbol{x}_D^d)}.$$

We notice that the conditional distribution resembles again a mixture of distributions, where the weights, $q_{Dl}(\boldsymbol{x}_D^c, \boldsymbol{x}_D^d)$, now depend on the covariates. This is an advantage over the single-weights DDP (covered in Section 2.2.5 and used by de Carvalho et al., 2020), where the weights are constant across the values of the covariates. Under this modelling framework, other features are also available, such as the conditional mean and variance function, $\mathbb{E}(y_D \mid \boldsymbol{x}_D^c, \boldsymbol{x}_D^d)$ and $\text{Var}(y_D \mid \boldsymbol{x}_D^c, \boldsymbol{x}_D^d)$, respectively. The conditional mean function is given by

$$\begin{aligned} \mathbb{E}(Y_D \mid \boldsymbol{X}_D^c = \boldsymbol{x}_D^c, \boldsymbol{X}_D^d = \boldsymbol{x}_D^d) &= \int_{-\infty}^{\infty} y_D f_D\left(y_D \mid \boldsymbol{X}_D^c = \boldsymbol{x}_D^c, \boldsymbol{X}_D^d = \boldsymbol{x}_D^d\right) dy_D \\ &= \int_{-\infty}^{\infty} y_D \sum_{l=1}^{L_D} q_{Dl}\left(\boldsymbol{x}_D^c, \boldsymbol{x}_D^d\right) \phi\left(y_D \mid \tilde{\mu}_{Dl}, \tilde{\sigma}_{Dl}^2\right) dy_D \\ &= \sum_{l=1}^{L_D} q_{Dl}\left(\boldsymbol{x}_D^c, \boldsymbol{x}_D^d\right) \tilde{\mu}_{Dl}, \end{aligned}$$

and the conditional variance function is

$$\begin{aligned}
\mathrm{Var}(Y_D \mid \boldsymbol{X}_D^c = \boldsymbol{x}_D^c, \boldsymbol{X}_D^d = \boldsymbol{x}_D^d) &= \mathbb{E}(Y_D^2 \mid \boldsymbol{x}_D^c, \boldsymbol{x}_D^d) - \mathbb{E}^2(Y_D \mid \boldsymbol{x}_D^c, \boldsymbol{x}_D^d) \\
&= \int_{-\infty}^{\infty} y_D^2 \sum_{l=1}^{L_D} q_{Dl}\left(\boldsymbol{x}_D^c, \boldsymbol{x}_D^d\right) \phi\left(y_D \mid \tilde{\mu}_{Dl}, \tilde{\sigma}_{Dl}^2\right) dy_D - \mathbb{E}^2(Y_D \mid \boldsymbol{x}_D^c, \boldsymbol{x}_D^d) \\
&= \sum_{l=1}^{L_D} q_{Dl}\left(\boldsymbol{x}_D^c, \boldsymbol{x}_D^d\right) \left\{\tilde{\sigma}_{Dl}^2 + \tilde{\mu}_{Dl}^2\right\} - \mathbb{E}^2(Y_D \mid \boldsymbol{x}_D^c, \boldsymbol{x}_D^d) \\
&= \sum_{l=1}^{L_D} q_{Dl}\left(\boldsymbol{x}_D^c, \boldsymbol{x}_D^d\right) \tilde{\sigma}_{Dl}^2 + \sum_{l=1}^{L_D} q_{Dl}\left(\boldsymbol{x}_D^c, \boldsymbol{x}_D^d\right) \left\{\tilde{\mu}_{Dl}^2 - \mathbb{E}^2(Y_D \mid \boldsymbol{x}_D^c, \boldsymbol{x}_D^d)\right\}.
\end{aligned}$$

4.2.2 Posterior inference

Posterior inference is available through explicit full conditional distributions via a data augmentation approach. We introduce (latent) configuration variables z_{Dj} to identify the label of the mixture components to which the j–th individual is allocated to, meaning that if $z_{Dj} = l$, then the j–th individual is allocated to component l, for $l = 1, \ldots, L_D$ and $j = 1, \ldots, n_D$. Denoting by $\boldsymbol{u}_{Dj} = (y_{Dj}, \boldsymbol{x}_{Dj}^c)$, this data augmentation allow us to rewrite our model hierarchically as

$$\boldsymbol{u}_{Dj}, \boldsymbol{x}_{Dj}^d \mid z_{Dj}, \boldsymbol{\theta}_D \overset{\text{ind.}}{\sim} \mathrm{N}_p\left(\boldsymbol{\mu}_{Dz_{Dj}}, \boldsymbol{\Sigma}_{Dz_{Dj}}\right) \prod_{k=1}^{q} \mathrm{Bernoulli}(\pi_{Dz_{Dj},k}), \quad j = 1, \ldots, n_D, \tag{4.7}$$

$$\mathbb{P}\left(z_{Dj} = l \mid \boldsymbol{v}_D\right) = \omega_{Dl}, \quad j = 1, \ldots, n_D, \quad l = 1, \ldots, L_D, \tag{4.8}$$

$$(\boldsymbol{\mu}_{Dl}, \boldsymbol{\Sigma}_{Dl}) \mid \boldsymbol{\mu}_{D0}, \boldsymbol{V}_{D0}, \nu_{D0}, \boldsymbol{\Psi}_{D0} \overset{\text{iid}}{\sim} \mathrm{N}_p\left(\boldsymbol{\mu}_{Dl} \mid \boldsymbol{\mu}_{D0}, \boldsymbol{V}_{D0}\right) \mathrm{IW}_p\left(\boldsymbol{\Sigma}_{Dl} \mid \nu_{D0}, \boldsymbol{\Psi}_{D0}\right), \quad l = 1, \ldots, L_D, \tag{4.9}$$

$$\pi_{Dl,k} \mid a_{\pi_{Dk}}, b_{\pi_{Dk}} \overset{\text{iid}}{\sim} \mathrm{Beta}\left(a_{\pi_{Dk}}, b_{\pi_{Dk}}\right), \quad k = 1, \ldots, q, \quad l = 1, \ldots, L_D, \tag{4.10}$$

$$v_{Dl} \mid \alpha_D \overset{\text{iid}}{\sim} \mathrm{Beta}(1, \alpha_D), \quad l = 1, \ldots, L_D - 1, \tag{4.11}$$

$$\boldsymbol{\mu}_{D0} \mid \boldsymbol{m}_D, \boldsymbol{S}_D \sim \mathrm{N}_p(\boldsymbol{m}_D, \boldsymbol{S}_D), \tag{4.12}$$

$$\boldsymbol{V}_{D0} \mid r_D, \boldsymbol{Q}_D \sim \mathrm{IW}_p(r_D, \boldsymbol{Q}_D), \tag{4.13}$$

$$\boldsymbol{\Psi}_{D0} \mid \nu_{D1}, \boldsymbol{\Psi}_{D1} \sim \mathrm{W}_p(\nu_{D1}, \boldsymbol{\Psi}_{D1}), \tag{4.14}$$

$$\alpha_D \mid a_{\alpha_D}, b_{\alpha_D} \sim \Gamma(a_{\alpha_D}, b_{\alpha_D}), \tag{4.15}$$

where $\boldsymbol{\theta}_D = (\boldsymbol{v}_D, \boldsymbol{\mu}_{D1}, \ldots, \boldsymbol{\mu}_{DL_D}, \boldsymbol{\Sigma}_{D1}, \ldots, \boldsymbol{\Sigma}_{DL_D}, \boldsymbol{\pi}_{D1}, \ldots, \boldsymbol{\pi}_{DL_D})$, $\boldsymbol{v}_D = (v_{D1}, \ldots, v_{DL_D})$ and $\boldsymbol{\omega}_D = (\omega_{D1}, \ldots, \omega_{DL_D})$.

Full conditional distributions

Under model described in (4.7)–(4.15), we can express the likelihood as follows

$$\begin{aligned}
L(\boldsymbol{\theta}_D; \boldsymbol{u}_D, \boldsymbol{x}_D^d, \boldsymbol{z}_D) &= \prod_{j=1}^{n_D} \omega_{Dz_{Dj}} \phi\left(\boldsymbol{u}_{Dj} \mid \boldsymbol{\mu}_{Dz_{Dj}}, \boldsymbol{\Sigma}_{Dz_{Dj}}\right) \prod_{k=1}^{q} \text{Bern}(x_{Dj,k}^d \mid \pi_{Dz_{Dj},k}) \\
&= \prod_{l=1}^{L_D} \prod_{j:z_{Dj}=l} \omega_{Dl} \phi\left(\boldsymbol{u}_{Dj} \mid \boldsymbol{\mu}_{Dl}, \boldsymbol{\Sigma}_{Dl}\right) \prod_{k=1}^{q} \text{Bern}(x_{Dj,k}^d \mid \pi_{Dl,k}) \\
&= \prod_{l=1}^{L_D} \omega_{Dl}^{n_{Dl}} \prod_{j:z_{Dj}=l} \phi\left(\boldsymbol{u}_{Dj} \mid \boldsymbol{\mu}_{Dl}, \boldsymbol{\Sigma}_{Dl}\right) \prod_{k=1}^{q} \text{Bern}(x_{Dj,k}^d \mid \pi_{Dl,k}),
\end{aligned}$$

where $n_{Dl} = \sum_{j=1}^{n_D} \mathbb{I}\{z_{Dj} = l\}$. The joint posterior distribution is given by

$$\begin{aligned}
p(\boldsymbol{z}_D, \boldsymbol{\theta}_D, \boldsymbol{\mu}_{D0}, \boldsymbol{V}_{D0}, \boldsymbol{\Psi}_{D0}, \alpha_D \mid \boldsymbol{u}_D, \boldsymbol{x}_D^d) &\propto L(\boldsymbol{\theta}_D; \boldsymbol{u}_D, \boldsymbol{x}_D^d, \boldsymbol{z}_D) p(\boldsymbol{\theta}_D) p(\boldsymbol{\mu}_{D0}) p(\boldsymbol{V}_{D0}) p(\boldsymbol{\Psi}_{D0}) p(\alpha_D) \\
&= L(\boldsymbol{\theta}_D; \boldsymbol{u}_D, \boldsymbol{x}_D^d, \boldsymbol{z}_D) \prod_{l=1}^{L_D} \left\{ p(\boldsymbol{\mu}_{Dl} \mid \boldsymbol{\mu}_{D0}, \boldsymbol{V}_{D0}) p(\boldsymbol{\Sigma}_{Dl} \mid \nu_{D0}, \boldsymbol{\Psi}_{D0}) \right. \\
&\quad \left. \times \prod_{k=1}^{q} p(\pi_{Dl,k} \mid a_{\pi_{Dk}}, b_{\pi_{Dk}}) \right\} \prod_{l=1}^{L_D-1} p(v_{Dl} \mid \alpha_D) p(\boldsymbol{\mu}_{D0}) p(\boldsymbol{V}_{D0}) p(\boldsymbol{\Psi}_{D0}) p(\alpha_D) \\
&\propto \prod_{l=1}^{L_D} \left\{ v_{Dl} \prod_{t<l} (1 - v_{Dt}) \right\}^{n_{Dl}} \prod_{j:z_{Dj}=l} \left\{ \phi\left(\boldsymbol{u}_{Dj} \mid \boldsymbol{\mu}_{Dl}, \boldsymbol{\Sigma}_{Dl}\right) \right. \\
&\quad \left. \times \prod_{k=1}^{q} \text{Bern}(x_{Dj,k}^d \mid \pi_{Dl,k}) \right\} \\
&\quad \times \prod_{l=1}^{L_D} |\boldsymbol{V}_{D0}|^{-1/2} \exp\left\{ -\frac{1}{2} \left(\boldsymbol{\mu}_{Dl} - \boldsymbol{\mu}_{D0}\right)' \boldsymbol{V}_{D0}^{-1} \left(\boldsymbol{\mu}_{Dl} - \boldsymbol{\mu}_{D0}\right) \right\} \\
&\quad \times \prod_{l=1}^{L_D} |\boldsymbol{\Psi}_{D0}|^{-\nu_{D0}/2} |\boldsymbol{\Sigma}_{Dl}|^{-(\nu_{D0}+p+1)/2} \exp\left\{ -\frac{1}{2} \text{tr}\left(\boldsymbol{\Psi}_{D0} \boldsymbol{\Sigma}_{Dl}^{-1}\right) \right\} \\
&\quad \times \prod_{l=1}^{L_D} \prod_{k=1}^{q} \pi_{Dl,k}^{a_{\pi_{Dk}}-1} (1 - \pi_{Dl,k})^{b_{\pi_{Dk}}-1} \\
&\quad \times \prod_{l=1}^{L_D-1} \alpha_D (1 - v_{Dl})^{\alpha_D - 1} \\
&\quad \times \exp\left\{ -\frac{1}{2} \left(\boldsymbol{\mu}_{D0} - \boldsymbol{m}_D\right)' \boldsymbol{S}_D^{-1} \left(\boldsymbol{\mu}_{D0} - \boldsymbol{m}_D\right) \right\} \\
&\quad \times |\boldsymbol{V}_{D0}|^{-(r_D+p+1)/2} \exp\left\{ -\frac{1}{2} \text{tr}\left(\boldsymbol{Q}_D \boldsymbol{V}_{D0}^{-1}\right) \right\} \\
&\quad \times |\boldsymbol{\Psi}_{D0}|^{(\nu_{D1}-p-1)/2} \exp\left\{ -\frac{1}{2} \text{tr}\left(\boldsymbol{\Psi}_{D1}^{-1} \boldsymbol{\Psi}_{D0}\right) \right\} \\
&\quad \times \alpha_D^{a_{\alpha_D}-1} \exp\left\{ -b_{\alpha_D} \alpha_D \right\}.
\end{aligned}$$

The joint posterior distribution does not have a recognisable form, but the full conditional distributions of each parameter does. Firstly, we will derive the full conditional for $\boldsymbol{\mu}_{Dl}$ as follows

$$\begin{aligned}
p(\boldsymbol{\mu}_{Dl} \mid \text{else}) &\propto L(\boldsymbol{\theta}_D; \boldsymbol{u}_D, \boldsymbol{x}_D^d, \boldsymbol{z}_D) p(\boldsymbol{\mu}_{Dl} \mid \boldsymbol{\mu}_{D0}, \boldsymbol{V}_{D0}) \\
&\propto \prod_{j:z_{Dj}=l} \exp\left\{-\frac{1}{2}(\boldsymbol{u}_{Dj} - \boldsymbol{\mu}_{Dl})' \boldsymbol{\Sigma}_{Dl}^{-1} (\boldsymbol{u}_{Dj} - \boldsymbol{\mu}_{Dl})\right\} \times \exp\left\{-\frac{1}{2}(\boldsymbol{\mu}_{Dl} - \boldsymbol{\mu}_{D0})' \boldsymbol{V}_{D0}^{-1} (\boldsymbol{\mu}_{Dl} - \boldsymbol{\mu}_{D0})\right\} \\
&\propto \exp\left\{-\frac{1}{2}\boldsymbol{\mu}_{Dl}' \left(\boldsymbol{V}_{D0}^{-1} + n_{Dl}\boldsymbol{\Sigma}_{Dl}^{-1}\right) \boldsymbol{\mu}_{Dl} + \boldsymbol{\mu}_{Dl}' \left(\boldsymbol{V}_{D0}^{-1}\boldsymbol{\mu}_{D0} + \boldsymbol{\Sigma}_{Dl}^{-1} \sum_{j:z_{Dj}=l} \boldsymbol{u}_{Dj}\right)\right\}.
\end{aligned}$$

Hoff (2009, p. 107–108) showed that if a random vector $\boldsymbol{\theta}$ has a density on $\mathbb{R}^p$ proportional to $\exp\left\{-\frac{1}{2}\boldsymbol{\theta}'\boldsymbol{A}\boldsymbol{\theta} + \boldsymbol{\theta}'\boldsymbol{b}\right\}$ for some matrix $\boldsymbol{A}$ and vector $\boldsymbol{b}$, then $\boldsymbol{\theta}$ must have a multivariate normal distribution with covariance matrix given by $\boldsymbol{A}^{-1}$ and mean vector $\boldsymbol{A}^{-1}\boldsymbol{b}$. Therefore,

$$\boldsymbol{\mu}_{Dl} \mid \text{else} \sim \mathrm{N}_p\left(\left[\boldsymbol{V}_{D0}^{-1} + n_{Dl}\boldsymbol{\Sigma}_{Dl}^{-1}\right]^{-1} \left(\boldsymbol{V}_{D0}^{-1}\boldsymbol{\mu}_{D0} + \boldsymbol{\Sigma}_{Dl}^{-1} \sum_{j:z_{Dj}=l} \boldsymbol{u}_{Dj}\right), \left[\boldsymbol{V}_{D0}^{-1} + n_{Dl}\boldsymbol{\Sigma}_{Dl}^{-1}\right]^{-1}\right).$$

Now, we will derive the full conditional for the covariance matrix of the mixture components $\boldsymbol{\Sigma}_{Dl}$ as follows

$$\begin{aligned}
p(\boldsymbol{\Sigma}_{Dl} \mid \text{else}) &\propto L(\boldsymbol{\theta}_D; \boldsymbol{u}_D, \boldsymbol{x}_D^d, \boldsymbol{z}_D) p(\boldsymbol{\Sigma}_{Dl} \mid \nu_{D0}, \boldsymbol{\Psi}_{D0}) \\
&\propto |\boldsymbol{\Sigma}_{Dl}|^{-n_{Dl}/2} \exp\left\{-\frac{1}{2} \sum_{j:z_{Dj}=l} (\boldsymbol{u}_{Dj} - \boldsymbol{\mu}_{Dl})' \boldsymbol{\Sigma}_{Dl}^{-1} (\boldsymbol{u}_{Dj} - \boldsymbol{\mu}_{Dl})\right\} \\
&\quad \times |\boldsymbol{\Sigma}_{Dl}|^{-(\nu_{D0}+p+1)/2} \exp\left\{-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Psi}_{D0}\boldsymbol{\Sigma}_{Dl}^{-1}\right)\right\} \\
&= |\boldsymbol{\Sigma}_{Dl}|^{-(n_{Dl}+\nu_{D0}+p+1)/2} \exp\left\{-\frac{1}{2}\left[\mathrm{tr}\left(\sum_{j:z_{Dj}=l} (\boldsymbol{u}_{Dj} - \boldsymbol{\mu}_{Dl})' \boldsymbol{\Sigma}_{Dl}^{-1} (\boldsymbol{u}_{Dj} - \boldsymbol{\mu}_{Dl})\right) + \mathrm{tr}\left(\boldsymbol{\Psi}_{D0}\boldsymbol{\Sigma}_{Dl}^{-1}\right)\right]\right\} \\
&= |\boldsymbol{\Sigma}_{Dl}|^{-(n_{Dl}+\nu_{D0}+p+1)/2} \exp\left\{-\frac{1}{2}\left[\mathrm{tr}\left(\sum_{j:z_{Dj}=l} (\boldsymbol{u}_{Dj} - \boldsymbol{\mu}_{Dl})(\boldsymbol{u}_{Dj} - \boldsymbol{\mu}_{Dl})' \boldsymbol{\Sigma}_{Dl}^{-1}\right) + \mathrm{tr}\left(\boldsymbol{\Psi}_{D0}\boldsymbol{\Sigma}_{Dl}^{-1}\right)\right]\right\} \\
&= |\boldsymbol{\Sigma}_{Dl}|^{-(n_{Dl}+\nu_{D0}+p+1)/2} \exp\left\{-\frac{1}{2}\mathrm{tr}\left(\left[\sum_{j:z_{Dj}=l} (\boldsymbol{u}_{Dj} - \boldsymbol{\mu}_{Dl})(\boldsymbol{u}_{Dj} - \boldsymbol{\mu}_{Dl})' + \boldsymbol{\Psi}_{D0}\right] \boldsymbol{\Sigma}_{Dl}^{-1}\right)\right\}.
\end{aligned}$$

The first equality is due to the fact that $\sum_{j:z_{Dj}=l} (\boldsymbol{u}_{Dj} - \boldsymbol{\mu}_{Dl})' \boldsymbol{\Sigma}_{Dl}^{-1} (\boldsymbol{u}_{Dj} - \boldsymbol{\mu}_{Dl})$ is a scalar, so it equals its trace, and that the trace is a linear mapping, i.e., if $\boldsymbol{A}$ and $\boldsymbol{B}$ are two $(n \times n)$ matrices,

then $\text{tr}(\boldsymbol{A}+\boldsymbol{B}) = \text{tr}(\boldsymbol{A}) + \text{tr}(\boldsymbol{B})$. Second equality obeys the cyclic property of the trace, that is, if $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ are matrices of dimension $(n \times p)$, $(p \times p)$ and $(p \times n)$, respectively, then $\text{tr}(\boldsymbol{ABC}) = \text{tr}(\boldsymbol{BCA}) = \text{tr}(\boldsymbol{CAB})$. In the third equality we have used the linear mapping property again. Finally, we recognize this as the kernel of an inverse Wishart distribution, thus

$$\boldsymbol{\Sigma}_{Dl} \mid \text{else} \sim \text{IW}_p \left(\nu_{D0} + n_{Dl}, \boldsymbol{\Psi}_{D0} + \sum_{j:z_{Dj}=l} (\boldsymbol{u}_{Dj} - \boldsymbol{\mu}_{Dl})(\boldsymbol{u}_{Dj} - \boldsymbol{\mu}_{Dl})' \right).$$

The full conditional distribution for the probability of success $\pi_{Dl,k}$ is given by

$$\begin{aligned} p(\pi_{Dl,k} \mid \text{else}) &\propto L(\boldsymbol{\theta}_D; \boldsymbol{u}_D, \boldsymbol{x}_D^d, \boldsymbol{z}_D) p(\pi_{Dl,k} \mid a_{\pi_{Dk}}, b_{\pi_{Dk}}) \\ &\propto \prod_{j:z_{Dj}=l} \pi_{Dl,k}^{x_{Dj,k}^d} (1-\pi_{Dl,k})^{1-x_{Dj,k}^d} \times \pi_{Dl,k}^{a_{\pi_{Dk}}-1} (1-\pi_{Dl,k})^{b_{\pi_{Dk}}-1} \\ &= \pi_{Dl,k}^{a_{\pi_{Dk}} + \sum_{j:z_{Dj}=l} x_{Dj,k}^d - 1} (1-\pi_{Dl,k})^{b_{\pi_{Dk}} + n_{Dl} - \sum_{j:z_{Dj}=l} x_{Dj,k}^d - 1}, \end{aligned}$$

which is the kernel of a beta distribution with parameters $a_{\pi_{Dk}} + \sum_{j:z_{Dj}=l} x_{Dj,k}^d$ and $b_{\pi_{Dk}} + n_{Dl} - \sum_{j:z_{Dk}=l} x_{Dj,k}^d$, i.e.,

$$\pi_{Dl,k} \mid \text{else} \sim \text{Beta} \left(a_{\pi_{Dk}} + \sum_{j:z_{Dj}=l} x_{Dj,k}^d, b_{\pi_{Dk}} + n_{Dl} - \sum_{j:z_{Dj}=l} x_{Dj,k}^d \right).$$

Since the stick-breaking weights are similarly defined as in the unconditional case (see Section 3.2.2), we know that

$$v_{Dl} \mid \text{else} \sim \text{Beta} \left(1 + n_{Dl}, \alpha_D + \sum_{h=l+1}^{L_D} n_{Dh} \right).$$

The full conditional of the prior mean of the mean components is given by

$$\begin{aligned} p(\boldsymbol{\mu}_{D0} \mid \text{else}) &\propto \prod_{l=1}^{L_D} p(\boldsymbol{\mu}_{Dl} \mid \boldsymbol{\mu}_{D0}, \boldsymbol{V}_{D0}) p(\boldsymbol{\mu}_{D0}) \\ &\propto \exp\left\{ -\frac{1}{2} \sum_{l=1}^{L_D} (\boldsymbol{\mu}_{Dl} - \boldsymbol{\mu}_{D0})' \boldsymbol{V}_{D0}^{-1} (\boldsymbol{\mu}_{Dl} - \boldsymbol{\mu}_{D0}) \right\} \times \exp\left\{ -\frac{1}{2} (\boldsymbol{\mu}_{D0} - \boldsymbol{m}_D)' \boldsymbol{S}_D^{-1} (\boldsymbol{\mu}_{D0} - \boldsymbol{m}_D) \right\} \\ &\propto \exp\left\{ -\frac{1}{2} \boldsymbol{\mu}_{D0}' \left(\boldsymbol{S}_D^{-1} + L_D \boldsymbol{V}_{D0}^{-1} \right) \boldsymbol{\mu}_{D0} + \boldsymbol{\mu}_{D0}' \left(\boldsymbol{S}_D^{-1} \boldsymbol{m}_D + \boldsymbol{V}_{D0}^{-1} \sum_{l=1}^{L_D} \boldsymbol{\mu}_{Dl} \right) \right\}, \end{aligned}$$

which is the kernel of a multivariate normal distribution. Therefore,

$$\boldsymbol{\mu}_{D0} \mid \text{else} \sim \mathrm{N}_p \left(\left[\boldsymbol{S}_D^{-1} + L_D \boldsymbol{V}_{D0}^{-1} \right]^{-1} \left(\boldsymbol{S}_D^{-1} \boldsymbol{m}_D + \boldsymbol{V}_{D0}^{-1} \sum_{l=1}^{L_D} \boldsymbol{\mu}_{Dl} \right), \left[\boldsymbol{S}_D^{-1} + L_D \boldsymbol{V}_{D0}^{-1} \right]^{-1} \right).$$

In turn, the full conditional distribution of the prior on $\boldsymbol{\mu}_{Dl}$ covariance matrix is

$$\begin{aligned} p(\boldsymbol{V}_{D0} \mid \text{else}) &\propto \prod_{l=1}^{L_D} p(\boldsymbol{\mu}_{Dl} \mid \boldsymbol{\mu}_{D0}, \boldsymbol{V}_{D0}) p(\boldsymbol{V}_{D0}) \\ &= |\boldsymbol{V}_{D0}|^{-L_D/2} \exp\left\{ -\frac{1}{2} \sum_{l=1}^{L_D} (\boldsymbol{\mu}_{Dl} - \boldsymbol{\mu}_{D0})' \boldsymbol{V}_{D0}^{-1} (\boldsymbol{\mu}_{Dl} - \boldsymbol{\mu}_{D0}) \right\} \\ &\quad \times |\boldsymbol{V}_{D0}|^{-(r_D+p+1)/2} \exp\left\{ -\frac{1}{2} \mathrm{tr}\left(\boldsymbol{Q}_D \boldsymbol{V}_{D0}^{-1} \right) \right\}, \end{aligned} \tag{4.16}$$

and using similar arguments to those presented for the full conditional distribution of $\boldsymbol{\Sigma}_{Dl}$, we can express Equation (4.16) as follows

$$p(\boldsymbol{V}_{D0} \mid \text{else}) \propto |\boldsymbol{V}_{D0}|^{-(L_D+r_D+p+1)/2} \exp\left\{ -\frac{1}{2} \mathrm{tr}\left(\left[\sum_{l=1}^{L_D} (\boldsymbol{\mu}_{Dl} - \boldsymbol{\mu}_{D0})(\boldsymbol{\mu}_{Dl} - \boldsymbol{\mu}_{D0})' + \boldsymbol{Q}_D \right] \boldsymbol{V}_{D0}^{-1} \right) \right\}.$$

Thus, it is the kernel of an inverse Wishart distribution, and hence

$$\boldsymbol{V}_{D0} \mid \text{else} \sim \mathrm{IW}_p \left(L_D + r_D, \sum_{l=1}^{L_D} (\boldsymbol{\mu}_{Dl} - \boldsymbol{\mu}_{D0})(\boldsymbol{\mu}_{Dl} - \boldsymbol{\mu}_{D0})' + \boldsymbol{Q}_D \right).$$

Similarly, we can derive the full conditional distribution of $\boldsymbol{\Psi}_{D0}$, as

$$\begin{aligned} p(\boldsymbol{\Psi}_{D0} \mid \text{else}) &\propto \prod_{l=1}^{L_D} p(\boldsymbol{\Sigma}_{Dl} \mid \nu_{D0}, \boldsymbol{\Psi}_{D0}) p(\boldsymbol{\Psi}_{D0}) \\ &= |\boldsymbol{\Psi}_{D0}|^{L_D \nu_{D0}/2} \exp\left\{ -\frac{1}{2} \mathrm{tr}\left(\sum_{l=1}^{L_D} \boldsymbol{\Psi}_{D0} \boldsymbol{\Sigma}_{Dl}^{-1} \right) \right\} \times |\boldsymbol{\Psi}_{D0}|^{(\nu_{D1}-p-1)/2} \exp\left\{ -\frac{1}{2} \mathrm{tr}\left(\boldsymbol{\Psi}_{D1}^{-1} \boldsymbol{\Psi}_{D0} \right) \right\} \\ &= |\boldsymbol{\Psi}_{D0}|^{(L_D \nu_{D0} + \nu_{D1} - p - 1)/2} \exp\left\{ -\frac{1}{2} \mathrm{tr}\left(\left[\sum_{l=1}^{L_D} \boldsymbol{\Sigma}_{Dl}^{-1} + \boldsymbol{\Psi}_{D1}^{-1} \right] \boldsymbol{\Psi}_{D0} \right) \right\}, \end{aligned}$$

which is the kernel of a Wishart distribution with $L_D \nu_{D0} + \nu_{D1}$ degrees of freedom and scale matrix $\left[\sum_{l=1}^{L_D} \boldsymbol{\Sigma}_{Dl}^{-1} + \boldsymbol{\Psi}_{D1}^{-1} \right]^{-1}$, i.e.,

$$\boldsymbol{\Psi}_{D0} \mid \text{else} \sim \mathrm{W}_p \left(L_D \nu_{D0} + \nu_{D1}, \left[\sum_{l=1}^{L_D} \boldsymbol{\Sigma}_{Dl}^{-1} + \boldsymbol{\Psi}_{D1}^{-1} \right]^{-1} \right).$$

We have already derived the full conditional distribution of the concentration parameter of the Dirichlet process (see Section 3.2.2), thus

$$\alpha_D \mid \text{else} \sim \Gamma\left(a_{\alpha_D} + L_D - 1, b_{\alpha_D} - \sum_{l=1}^{L_D-1} \log(1 - v_{Dl})\right).$$

Finally, for the latent variables we have

$$\begin{aligned}\mathbb{P}(z_{Dj} = h \mid \text{else}) &= \frac{L(\boldsymbol{\theta}_D; \boldsymbol{u}_{Dj}, \boldsymbol{x}^d_{Dj}, z_{Dj} = h)\mathbb{P}(z_{Dj} = h)}{\sum_{l=1}^{L_D} L(\boldsymbol{\theta}_D; \boldsymbol{u}_{Dj}, \boldsymbol{x}^d_{Dj}, z_{Dj} = l)\mathbb{P}(z_{Dj} = l)} \\ &= \frac{\omega_{Dh}\phi\left(\boldsymbol{u}_{Dj} \mid \boldsymbol{\mu}_{Dh}, \boldsymbol{\Sigma}_{Dh}\right)\prod_{k=1}^{q}\text{Bern}\left(x^d_{Dj,k} \mid \pi_{Dh,k}\right)}{\sum_{l=1}^{L_D}\omega_{Dl}\phi\left(\boldsymbol{u}_{Dj} \mid \boldsymbol{\mu}_{Dl}, \boldsymbol{\Sigma}_{Dl}\right)\prod_{k=1}^{q}\text{Bern}\left(x^d_{Dj,k} \mid \pi_{Dl,k}\right)}, \quad j = 1, \ldots, n_D, \;\; h = 1, \ldots, L_D.\end{aligned}$$

These full conditional distributions allow us to implement a simple Gibbs sampler to simulate, say S, draws from the posterior distribution. Thus, at iteration s, for $s = 1, \ldots, S$, the estimated conditional density is given by

$$f_D^{(s)}\left(y_D \mid \boldsymbol{X}^c_D = \boldsymbol{x}^c_D, \boldsymbol{X}^d_D = \boldsymbol{x}^d_D\right) = \sum_{l=1}^{L_D} q^{(s)}_{Dl}(\boldsymbol{x}^c_D, \boldsymbol{x}^d_D)\phi\left(y_D \mid \tilde{\mu}^{(s)}_{Dl}, \tilde{\sigma}^{2(s)}_{Dl}\right),$$

where

$$q^{(s)}_{Dl}(\boldsymbol{x}^c_D, \boldsymbol{x}^d_D) = \frac{\omega^{(s)}_{Dl}\phi\left(\boldsymbol{x}^c_D \mid \boldsymbol{\mu}^{x(s)}_{Dl}, \boldsymbol{\Sigma}^{xx(s)}_{Dl}\right)\prod_{k=1}^{q}\text{Bern}(x^d_{D,k} \mid \pi^{(s)}_{Dl,k})}{f^{(s)}_D(\boldsymbol{x}^c_D, \boldsymbol{x}^d_D)},$$

$\tilde{\mu}^{(s)}_{Dl} = \mu^{y(s)}_{Dl} + \boldsymbol{\Sigma}^{yx(s)}_{Dl}\left(\boldsymbol{\Sigma}^{xx(s)}_{Dl}\right)^{-1}\left(\boldsymbol{x}^c_D - \boldsymbol{\mu}^{x(s)}_{Dl}\right)$ and $\tilde{\sigma}^{2(s)}_{Dl} = \Sigma^{yy(s)}_{Dl} - \boldsymbol{\Sigma}^{yx(s)}_{Dl}\left(\boldsymbol{\Sigma}^{xx(s)}_{Dl}\right)^{-1}\boldsymbol{\Sigma}^{xy(s)}_{Dl}$. The marginal density of the covariates can be expressed in a similar way as follows

$$f^{(s)}_D(\boldsymbol{x}^c_D, \boldsymbol{x}^d_D) = \sum_{l=1}^{L_D}\omega^{(s)}_{Dl}\phi\left(\boldsymbol{x}^c_D \mid \boldsymbol{\mu}^{x(s)}_{Dl}, \boldsymbol{\Sigma}^{xx(s)}_{Dl}\right)\prod_{k=1}^{q}\text{Bern}(x^d_{D,k} \mid \pi^{(s)}_{Dl,k}).$$

4.2.3 Inference for the covariate-specific OVL

Our proposed estimator for the covariate-specific OVL is based on (1.4). The integral is replaced by a numerical integration method over all the posterior draws, more precisely, we use the trapezoidal rule. Denoting by $g^{(s)}(y \mid \boldsymbol{x}) = \min\left\{f^{(s)}_{\bar{D}}(y \mid \boldsymbol{x}), f^{(s)}_D(y \mid \boldsymbol{x})\right\}$, for any fixed value $\boldsymbol{x} = (\boldsymbol{x}^c, \boldsymbol{x}^d)$ of the

covariates, a single Gibbs sampler iteration of our proposed estimator is given by

$$\begin{aligned}\mathrm{OVL}^{(s)}(\boldsymbol{x}) &= \frac{\Delta y}{2}\sum_{i=1}^{N}\left\{g^{(s)}(y_{i-1}\mid \boldsymbol{x}) + g^{(s)}(y_i \mid \boldsymbol{x})\right\},\\ &= \frac{\Delta y}{2}\left\{g^{(s)}(y_0\mid \boldsymbol{x}) + 2\sum_{i=1}^{N-1} g^{(s)}(y_i\mid \boldsymbol{x}) + g^{(s)}(y_N\mid \boldsymbol{x})\right\}, \quad s=1,\ldots,S, \end{aligned} \tag{4.17}$$

where $\min\{y_{\bar{D}}, y_D\} = y_0 < \cdots < y_N = \max\{y_{\bar{D}}, y_D\}$ is an equally spaced grid and Δy is the length of each sub-interval. Finally, we can take $\widehat{\mathrm{OVL}}(\boldsymbol{x}) = \frac{1}{S}\sum_{s=1}^{S}\mathrm{OVL}^{(s)}(\boldsymbol{x})$ as a point estimate of the covariate-specific OVL at $\boldsymbol{x}$ and we can obtain symmetric $100(1-\alpha)\%$ pointwise credible bands using the $\alpha/2$ and $1-\alpha/2$ quantiles of the posterior ensemble $\left\{\mathrm{OVL}^{(1)}(\boldsymbol{x}),\ldots,\mathrm{OVL}^{(S)}(\boldsymbol{x})\right\}$.

4.3 Simulation study

In this section we illustrate the performance of our method through different simulated examples. We have simulated 100 datasets of sample sizes $(n_{\bar{D}}, n_D) = (100, 100)$, $(n_{\bar{D}}, n_D) = (200, 200)$ and $(n_{\bar{D}}, n_D) = (500, 500)$ each. Similar to the unconditional case, even though we assumed equal sample sizes for both groups, in practice it is common to find imbalanced data, e.g., a larger number of nondiseased compared to diseased individuals. However, unless the difference in sample sizes is very large, the results should not differ too much from those presented here. For all the scenarios, we have simulated continuous covariates independently from a uniform distribution, that is,

$$\begin{aligned} x_{\bar{D}i,1} &\overset{\text{iid}}{\sim} \mathrm{U}(0,1), \quad i = 1,\ldots,n_{\bar{D}},\\ x_{Dj,1}, &\overset{\text{iid}}{\sim} \mathrm{U}(0,1), \quad j = 1,\ldots,n_D. \end{aligned}$$

Further, we generated binary covariates from a Bernoulli distribution, namely,

$$x_{\bar{D}i,2}, x_{Dj,2} \overset{\text{iid}}{\sim} \text{Bernoulli}(0.5).$$

For each simulated dataset, we simulated 20,000 MCMC iterations of the Gibbs sampler after discarding 20,000 iterations as burn-in period. For the trapezoidal rule, we used 201 equally spaced points and we computed the covariate-specific OVL over 11 values of the continuous covariate.

4.3.1 Simulation scenarios

- Scenario I

In Scenario I, we consider a homoscedastic linear mean regression model for both nondiseased and diseased group. Specifically, the true data generating process is given by

$$y_{\bar{D}i} \mid x_{\bar{D}i,1} \overset{\text{ind}}{\sim} \text{N}\left(2 + 4x_{\bar{D}i,1}, 0.8^2\right), \quad i = 1, \ldots, n_{\bar{D}},$$
$$y_{Dj} \mid x_{Dj,1} \overset{\text{ind}}{\sim} \text{N}\left(3 + 4x_{Dj,1}, 0.5^2\right), \quad j = 1, \ldots, n_D.$$

Our main purpose is to investigate the performance of the covariate-specific OVL when the covariates do not affect the discriminatory capacity of the test, i.e., it is constant across x_1. Although, they do affect the test outcomes.

- Scenario II

In Scenario II we investigate the performance of our method when the test outcomes arise from different and non normal parametric families, namely, a gamma distribution and a mixture of normal distributions for the nondiseased and diseased group, respectively. The effect of the covariates is non-linear. We used the following data generating process

$$y_{\bar{D}i} \mid x_{\bar{D}i,1} \overset{\text{ind}}{\sim} \Gamma\left(10, 10\left[(2x_{\bar{D}i,1} - 1)^2 + 0.5\right]^{-1}\right), \quad i = 1, \ldots, n_{\bar{D}},$$
$$y_{Dj} \mid x_{Dj,1} \overset{\text{ind}}{\sim} 0.2\text{N}\left(-\{2x_{Dj,1} - 1\}^2 - 1.5, 0.5^2\right) + 0.8\text{N}\left(-\{2x_{Dj,1} - 1\}^2 + 2.5, 1\right), \quad j = 1, \ldots, n_D.$$

Note that this setting involves a heteroscedastic situation in both groups, because the variance of the gamma distribution is

$$\text{Var}\left(y_{\bar{D}i} \mid x_{\bar{D}i,1}\right) = \frac{\left[(2x_{\bar{D}i,1} - 1)^2 + 0.5\right]^2}{10},$$

and the variance of the mixture is given by

$$\text{Var}\left(y_{Dj} \mid x_{Dj,1}\right) = 4.6 + (2x_{Dj,1} - 1)^2\left[(2x_{Dj,1} - 1)^2 - 2.4\right].$$

- Scenario III

Finally, in Scenario III we consider Gaussian heteroscedastic non-linear regression models for both the nondiseased and diseased groups. Further, we include an interaction between the continuous and binary covariate. We simulated data as follows

$$y_{\bar{D}i} \mid x_{\bar{D}i,1}, x_{\bar{D}i,1} \overset{\text{ind}}{\sim} \mathrm{N}\left(\mu_{\bar{D}}(x_{\bar{D}i,1}, x_{\bar{D}i,2}), \sigma^2_{\bar{D}}(x_{\bar{D}i,1})\right), \quad i = 1, \ldots, n_{\bar{D}},$$
$$y_{Dj} \mid x_{Dj,1}, x_{Dj,2} \overset{\text{ind}}{\sim} \mathrm{N}\left(\mu_{D}(x_{Dj,1}, x_{Dj,2}), \sigma^2_{D}(x_{Dj,1})\right), \quad j = 1, \ldots, n_{D},$$

where

$$\mu_{\bar{D}}(x_{\bar{D}i,1}, x_{\bar{D}i,2}) = \left(0.7\sin(x_{\bar{D}i,1})\cos(6x_{\bar{D}i,1}) + 0.4\right)(1 - x_{\bar{D}i,2}) + 0.3\exp\left\{-2x_{\bar{D}i,1} + 1\right\}x_{\bar{D}i,2},$$
$$\mu_{D}(x_{Dj,1}, x_{Dj,2}) = \left(0.5\sin(x_{Dj,1})\cos(8x_{Dj,1}) - 0.2\right)(1 - x_{Dj,2}) + 0.4\exp\left\{-x^2_{Dj,1} + 0.3x_{Dj,1} + 1\right\}x_{Dj,2},$$
$$\sigma_{\bar{D}}(x_{\bar{D}i,1}) = 0.6(1 - x_{\bar{D}i,1})^3 - 0.2x^2_{\bar{D}i,1} + 0.5x_{\bar{D}i,1},$$
$$\sigma_{D}(x_{Di,1}) = \phi(4x_{Dj,1}) + 0.2.$$

4.3.2 Models

Again, for notational convenience we present the prior specification for the diseased group, the same one was used for the nondiseased group. To avoid numerical difficulties within the MCMC algorithm and to facilitate prior specification, we have centred and scaled the data, i.e., we modelled $\boldsymbol{\Sigma}_{D0}^{-1/2}(\boldsymbol{u}_D - \bar{\boldsymbol{u}}_D)$, where $\boldsymbol{\Sigma}_{D0}$ is the sample covariance matrix and $\bar{\boldsymbol{u}}_D$ is the mean vector, instead of the original observations $\boldsymbol{u}_D$. To compute $\boldsymbol{\Sigma}_{D0}^{-1/2}$, we used the spectral decomposition of $\boldsymbol{\Sigma}_{D0}$, that is

$$\boldsymbol{\Sigma}_{D0} = \boldsymbol{\Lambda}\boldsymbol{D}\boldsymbol{\Lambda}^{-1},$$

where $\boldsymbol{\Lambda}$ is a $(p \times p)$ matrix whose j–th column is the j–th eigenvector of $\boldsymbol{\Sigma}_{D0}$ and $\boldsymbol{D}$ is a diagonal matrix whose elements are the corresponding eigenvalues. It follows that, $\boldsymbol{\Sigma}_{D0}^{1/2} = \boldsymbol{\Lambda}\boldsymbol{D}^{1/2}\boldsymbol{\Lambda}^{-1}$.

We use relatively vague prior distributions. We take $\boldsymbol{m}_D = \mathbf{0}_p$ and $\boldsymbol{S}_D = 0.25\boldsymbol{I}_p$ for the prior mean $\boldsymbol{\mu}_{D0}$, where $\mathbf{0}_p$ and $\boldsymbol{I}_p$ denote the p–dimensional zero vector and identity matrix, respectively. For the

prior covariance matrix $\boldsymbol{V}_{D0}$, we take $r_D = p+2$ and $\boldsymbol{Q}_D = 0.25\boldsymbol{I}_p$. This hyperparameters choice leads to a distribution for the mean vector components centred around $\boldsymbol{0}_p$ with covariance matrix given by $0.5\boldsymbol{I}_p$. Whereas taking $\nu_{D0} = p+2$, $\nu_{D1} = p$ and $\boldsymbol{\Psi}_{D1} = 0.5p^{-1}\boldsymbol{I}_p$ favours covariance matrices around $\boldsymbol{I}_p$ because $\mathbb{E}(\boldsymbol{\Sigma}_{Dl}) = \boldsymbol{I}_p$. Since the data have unit variance (and zero covariance), we expect that marginally the covariance matrix of the components is close to the identity matrix. We use $a_{\alpha_D} = 2$ and $b_{\alpha_D} = 2$ as hyperparameters for the concentration parameter of the DP, and under this parameter specification $\mathbb{E}(\alpha_D) = 1$ and $\text{Var}(\alpha_D) = 0.5$. For the binary covariates, we choose $a_{\pi_D} = 0.5$ and $b_{\pi_D} = 0.5$, which leads to the uninformative Jeffrey's prior. Finally, we capped the stick-breaking construction up to $L_D = 20$, this means that a maximum of 20 multivariate normal distributions were used to approximate the joint density. Note that with these choices, if $\boldsymbol{u}_D$ is drawn from the prior then

$$\begin{aligned}
\mathbb{E}(\boldsymbol{u}_D) &= \mathbb{E}(\mathbb{E}(\boldsymbol{u}_D \mid \boldsymbol{\mu}_D, \boldsymbol{\Sigma}_D)) \\
&= \mathbb{E}(\boldsymbol{\mu}_D) \\
&= \boldsymbol{0}_p \\
\text{Var}(\boldsymbol{u}_D) &= \mathbb{E}(\text{Var}(\boldsymbol{u}_D \mid \boldsymbol{\mu}_D, \boldsymbol{\Sigma}_D)) + \text{Var}(\mathbb{E}(\boldsymbol{u}_D \mid \boldsymbol{\mu}_D, \boldsymbol{\Sigma}_D)) \\
&= \mathbb{E}(\boldsymbol{\Sigma}_D) + \text{Var}(\boldsymbol{\mu}_D) \\
&= 0.5\boldsymbol{I}_p + 0.5\boldsymbol{I}_p \\
&= \boldsymbol{I}_p.
\end{aligned}$$

The performance of our proposed model is compared to a linear mean regression model, where the test outcomes are assumed Gaussian with constant variance σ_D^2 and the mean is a linear function of the covariates, i.e.,

$$\eta_{Dj} = \beta_{D0} + \beta_{D1}x^c_{Dj,1} + \cdots + \beta_{Dp-1}x^c_{Dj,p-1} + \gamma_{D1}x^d_{Dj,1} + \cdots + \gamma_{Dq}x^d_{Dj,q}, \quad j = 1, \ldots, n_D.$$

Thus, the corresponding conditional density function is given by

$$f(y_{Dj} \mid \boldsymbol{X}^c_D = \boldsymbol{x}^c_{Dj}, \boldsymbol{X}^d_D = \boldsymbol{x}^d_{Dj}) = \phi\left(y_{Dj} \mid \eta_{Dj}, \sigma^2_D\right),$$

Finally, the estimated densities are plugged in (1.4) to obtain the covariate-specific OVL.

The fit of the models is visually assessed by simply looking at the posterior realisations of our proposed covariate-specific OVL estimator compared to the true one. The performance of our model when compared to the LM was evaluated through different criteria, namely, the log pseudomarginal likelihood, the widely applicable criterion, the modified deviance information criterion and the posterior rank probability (details on these criteria can be found in Section 2.3). We plug Equation (4.6) in (2.6), (2.8) and (2.11) to obtain the CPO's, the stabilised CPO's and the DIC_3, respectively. Once the CPO's are computed, we can compute the LPML and the posterior rank probability described in (2.7) and (2.12), respectively. The densities in Equations (2.9) and (2.10) are replaced by (4.6) to obtain the WAIC.

4.3.3 Results

Figures 4.2–4.19 depict the mean of the posterior means across the 100 replications and the pointwise bands are based on the 2.5% and 97.5% percentiles of such 100 posterior means. Tables 4.4–4.6 show the model comparison criteria results, the values shown are averages based on all the replications. We computed an empirical probability of our model being better than the LM for each criterion, for example, for the WAIC we computed

$$\mathbb{P}(\text{WAIC}_{\text{DPM}} < \text{WAIC}_{\text{LM}}) = \frac{1}{100}\sum_{i=1}^{100}\mathbb{I}\left\{\text{WAIC}_{\text{DPM},i} < \text{WAIC}_{\text{LM},i}\right\},$$

where $\text{WAIC}_{\text{DPM},i}$ is the WAIC obtained for the i–th replication under our modelling approach. Similarly, $\text{WAIC}_{\text{LM},i}$ is the WAIC under the LM.

Although the LM outperforms, as expected, the DPM when the effect of the covariate is linear, our modelling approach is able to recover the true covariate-specific OVL, mean functions and conditional densities even in situation when $(n_{\bar{D}}, n_D) = (100, 100)$ (Figure 4.2). When the test outcomes arise from different parametric families (Scenario II), the DPM does not accurately estimate the covariate-specific OVL, specially at the boundaries and for the smaller sample sizes of $(n_{\bar{D}}, n_D) = (100, 100)$ and $(n_{\bar{D}}, n_D) = (200, 200)$. This is depicted in Figures 4.8 and 4.10 (top left). One possible reason is that the DPM tends to sub-estimate the mode of the nondiseased distributions (Figures 4.9 and 4.11). However, as the sample size increases, the estimation of the covariate-specific OVL and the

mean functions becomes more accurate. When $(n_{\bar{D}}, n_D) = (500, 500)$, our modelling approach recovers completely the conditional densities (Figure 4.13) and thus, the true covariate-specific OVL (Figure 4.12 top left).

Scenario III is the most challenging one. In Figure 4.14 top left, we observe that the estimation of the OVL is rather poor when $(n_{\bar{D}}, n_D) = (100, 100)$. In this case, our modelling approach is not able to fully capture the curvatures induced by the interaction between the continuous and the binary covariate. In Figure 4.15, we may confirm this by observing that the DPM tends to poorly estimate the variance of the conditional distributions. Nevertheless, similar to the previous scenarios, as the sample size increases, these shortcomings start to disappear and the DPM starts recovering accurately the covariate-specific OVL, the mean and variance functions and the underlying conditional distributions (Figures 4.18 and 4.19).

We explore the empirical coverage probability of the pointwise 95% credible intervals. As a summary measure, we averaged this coverage across the grid of the continuous covariate. The results are summarised in Tables 4.1–4.3. For Scenario I, the coverage of our estimator is close to the nominal value of 0.95. However, this was not the case for Scenarios II and III. In the former, only for larger sample sizes we get a coverage close to the nominal value, but still below. Whereas, for Scenario III, the coverage is much lower, even when the sample size is high. However, the coverage is close to the nominal value when $x_2 = 1$ and $(n_{\bar{D}}, n_D) = (500, 500)$.

Finally, Tables 4.4–4.6 display the results of different model comparison criteria discussed in Section 2.3. We observe that our modelling approach outperforms clearly the LM in Scenarios II and III for all sample sizes. For Scenario I, the LM is preferable as expected.

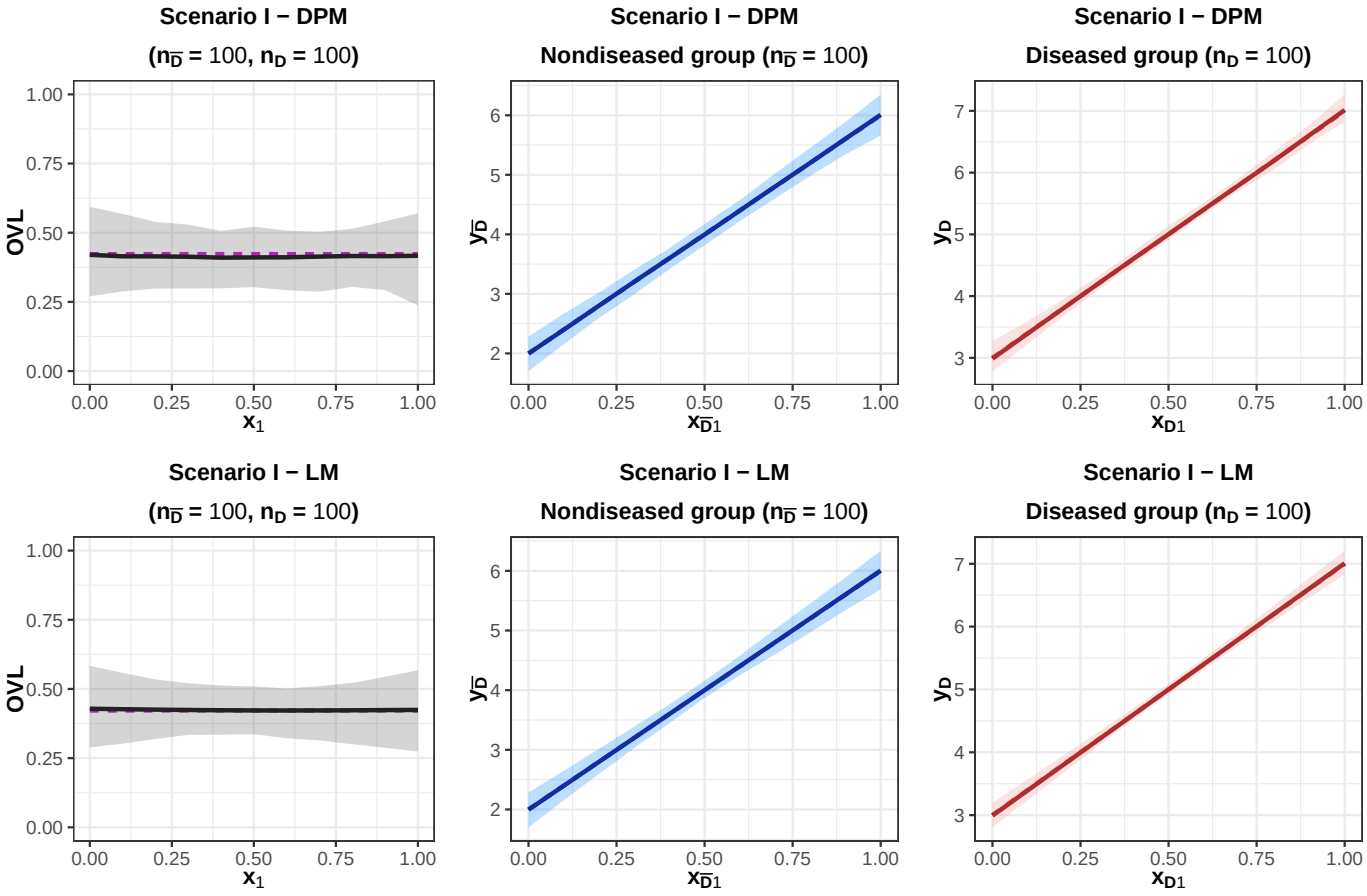

Figure 4.2: Results for the DPM model and the LM under Scenario I and $(n_{\bar{D}}, n_D) = (100, 100)$ are displayed in the top and bottom panels, respectively. True (dotted line) and mean across simulations (solid line) of the posterior mean of the covariate-specific OVL (left) and mean function of the nondiseased (middle) and diseased (right) group. A band constructed using the pointwise 2.5% and 97.5% percentiles of the posterior means across simulations is presented as the shaded area.

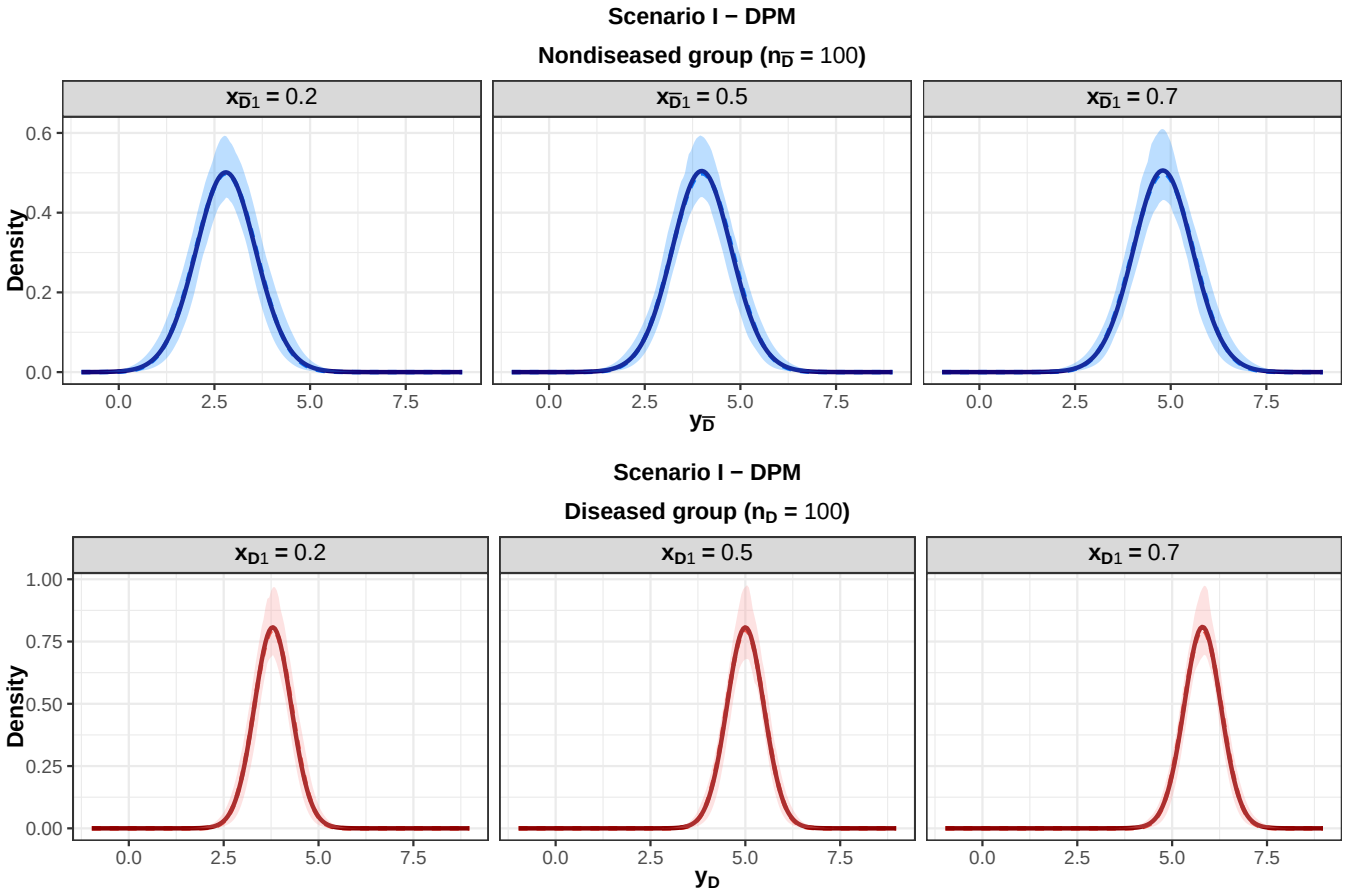

Figure 4.3: Conditional density functions for three particular values of the covariate under the DPM model for Scenario I and $(n_{\bar{D}}, n_D) = (100, 100)$. True (dotted line) conditional distribution and mean across simulations (solid line) of the posterior mean of the conditional density for the nondiseased (top) and diseased (bottom) group. A band constructed using the pointwise 2.5% and 97.5% percentiles of the posterior means across simulations is presented as the shaded area.

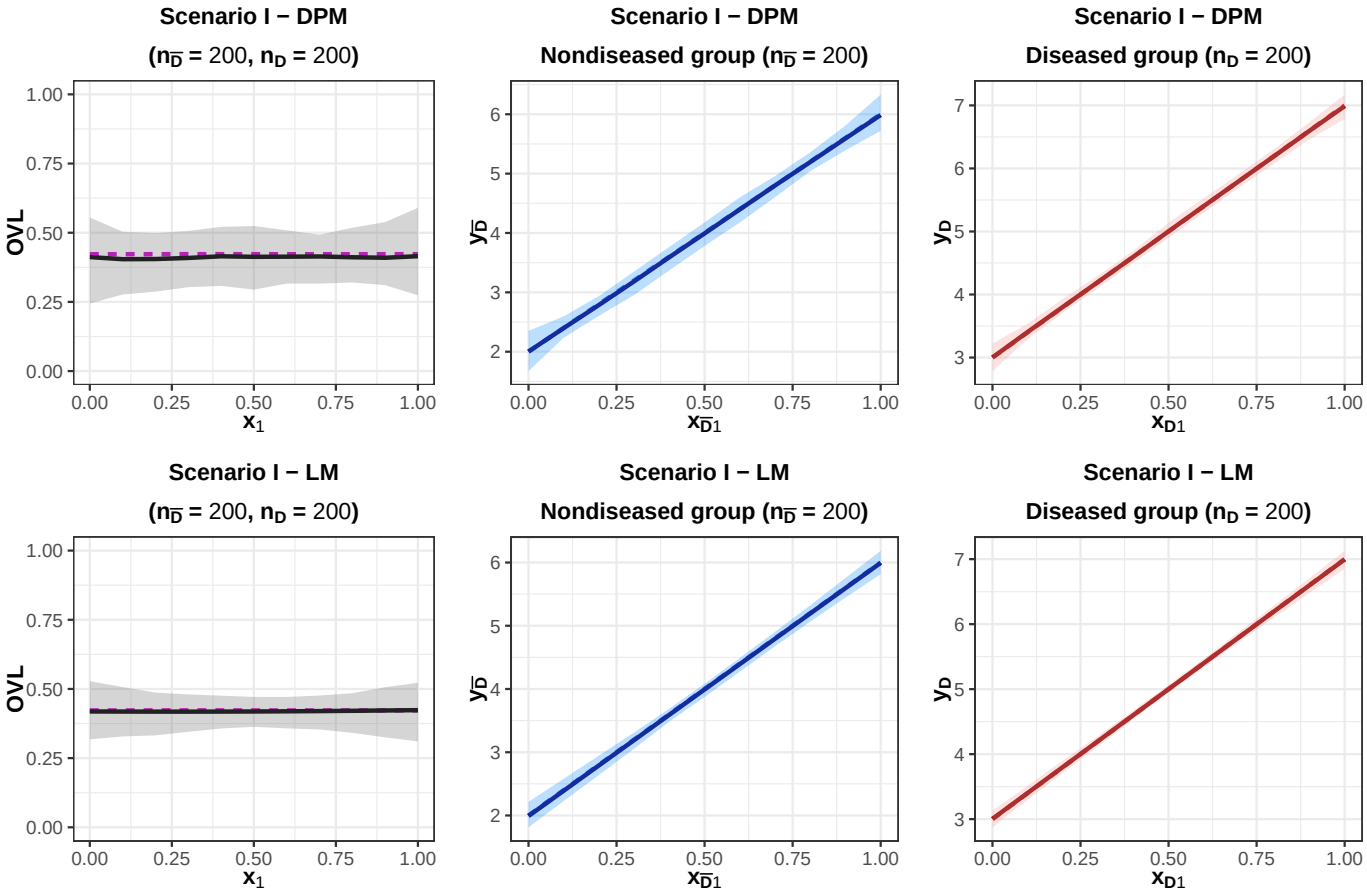

Figure 4.4: Results for the DPM model and the LM under Scenario I and $(n_{\bar{D}}, n_D) = (200, 200)$ are displayed in the top and bottom panels, respectively. True (dotted line) and mean across simulations (solid line) of the posterior mean of the covariate-specific OVL (left) and mean function of the nondiseased (middle) and diseased (right) group. A band constructed using the pointwise 2.5% and 97.5% percentiles of the posterior means across simulations is presented as the shaded area.

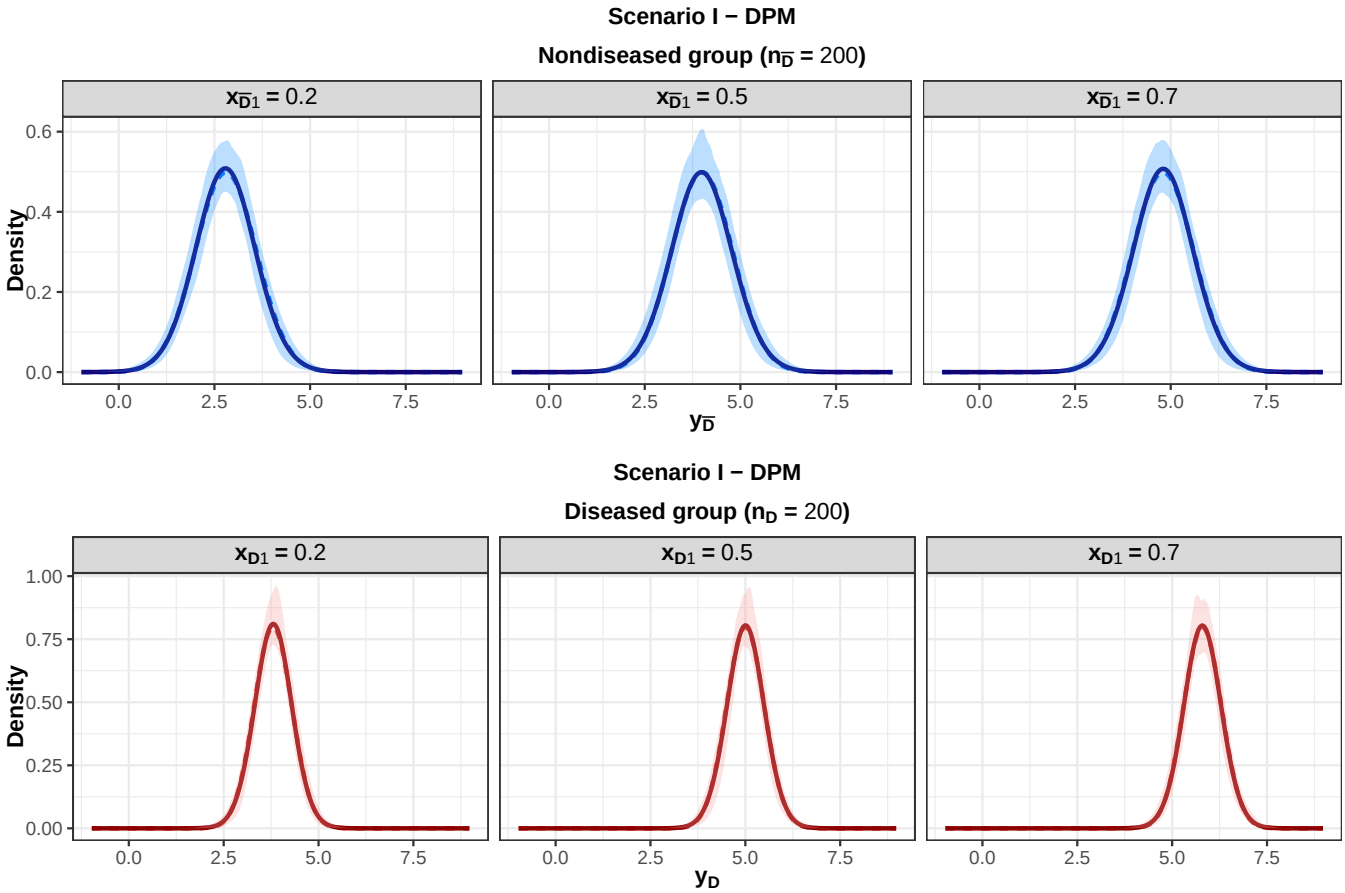

Figure 4.5: Conditional density functions for three particular values of the covariate under the DPM model for Scenario I and $(n_{\bar{D}}, n_D) = (200, 200)$. True (dotted line) conditional distribution and mean across simulations (solid line) of the posterior mean of the conditional density for the nondiseased (top) and diseased (bottom) group. A band constructed using the pointwise 2.5% and 97.5% percentiles of the posterior means across simulations is presented as the shaded area.

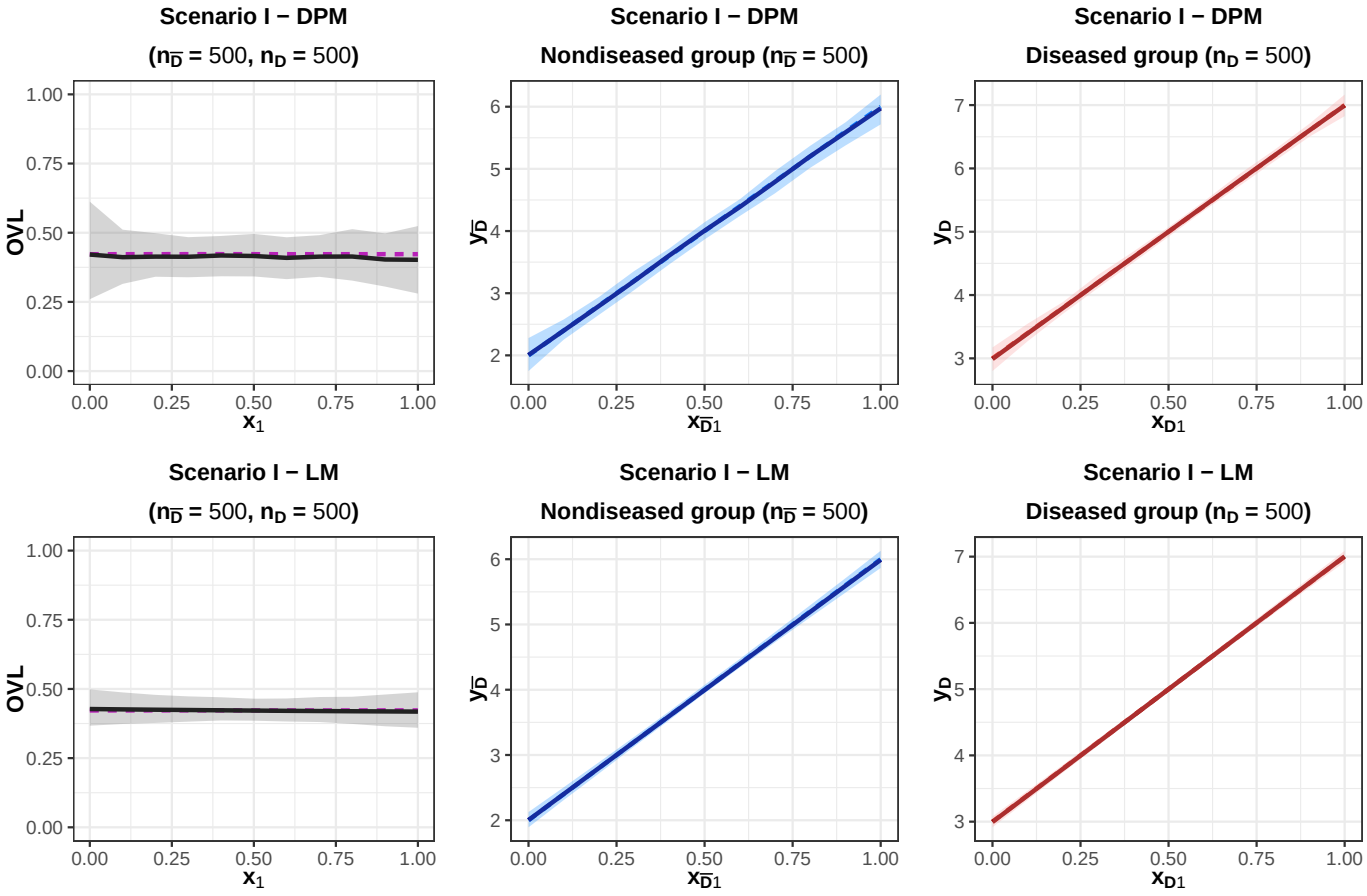

Figure 4.6: Results for the DPM model and the LM under Scenario I and $(n_{\bar{D}}, n_D) = (500, 500)$ are displayed in the top and bottom panels, respectively. True (dotted line) and mean across simulations (solid line) of the posterior mean of the covariate-specific OVL (left) and mean function of the nondiseased (middle) and diseased (right) group. A band constructed using the pointwise 2.5% and 97.5% percentiles of the posterior means across simulations is presented as the shaded area.

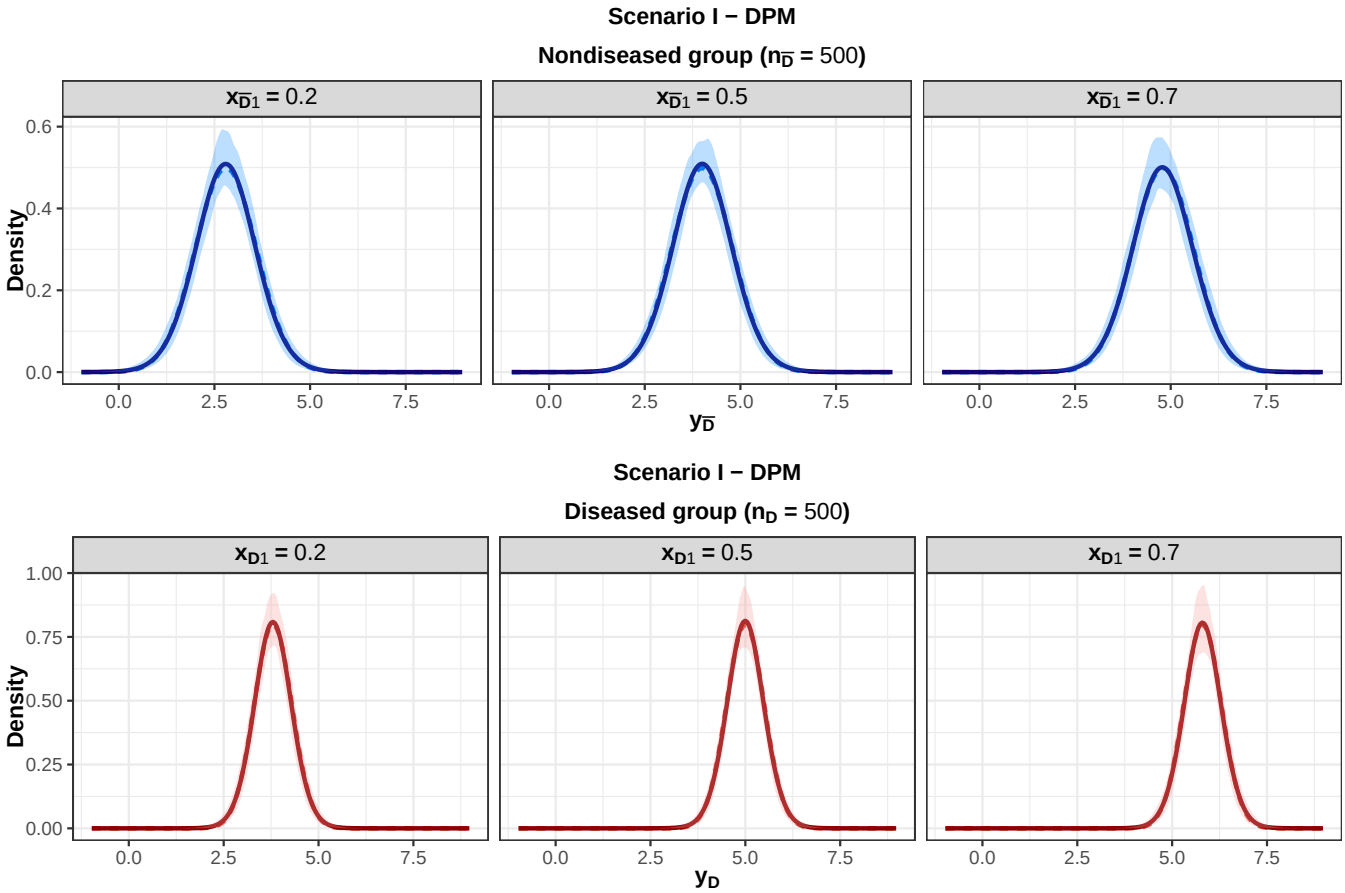

Figure 4.7: Conditional density functions for three particular values of the covariate under the DPM model for Scenario I and $(n_{\bar{D}}, n_D) = (500, 500)$. True (dotted line) conditional distribution and mean across simulations (solid line) of the posterior mean of the conditional density for the nondiseased (top) and diseased (bottom) group. A band constructed using the pointwise 2.5% and 97.5% percentiles of the posterior means across simulations is presented as the shaded area.

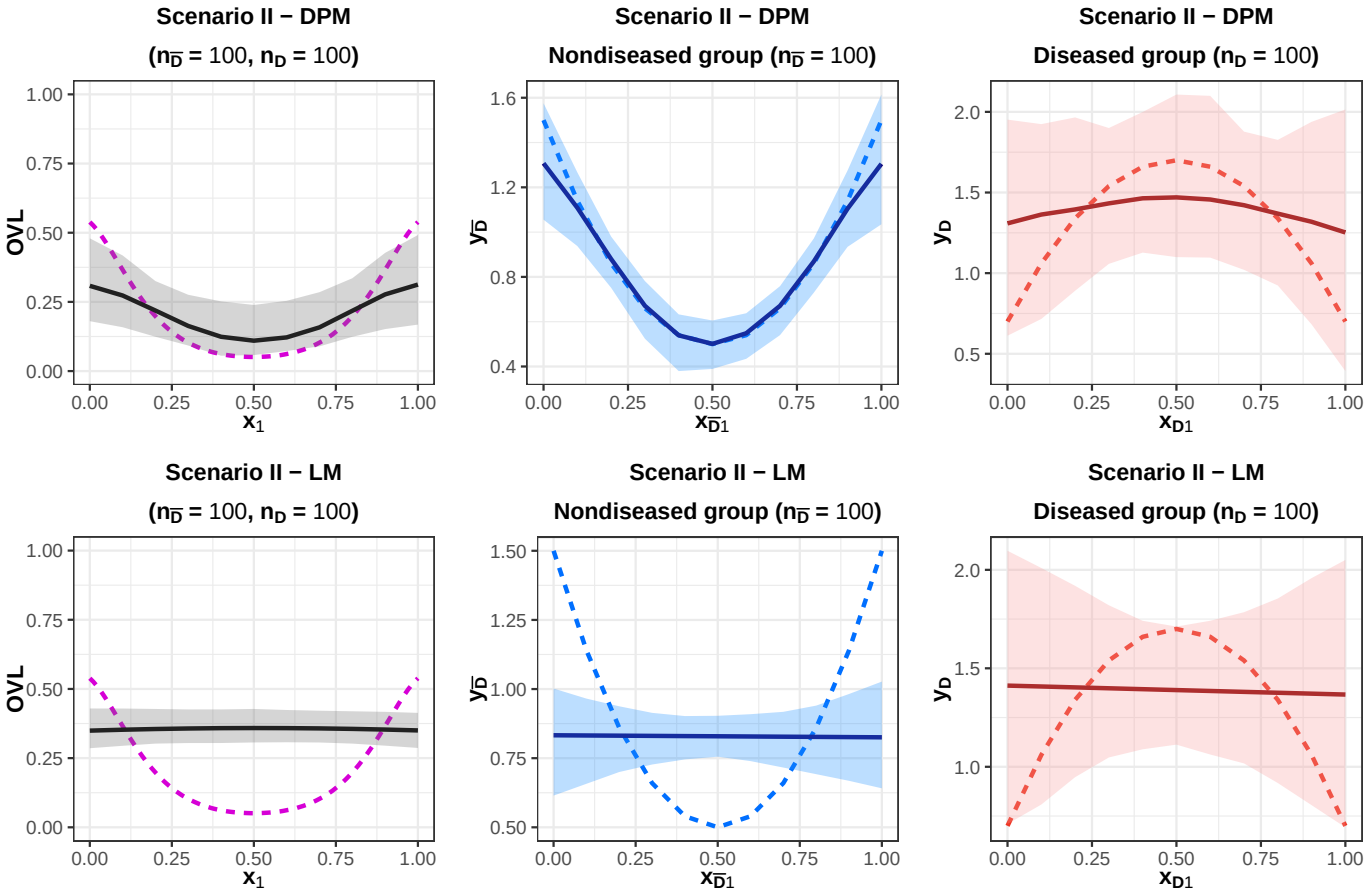

Figure 4.8: Results for the DPM model and the LM under Scenario II and $(n_{\bar{D}}, n_D) = (100, 100)$ are displayed in the top and bottom panels, respectively. True (dotted line) and mean across simulations (solid line) of the posterior mean of the covariate-specific OVL (left) and mean function of the nondiseased (middle) and diseased (right) group. A band constructed using the pointwise 2.5% and 97.5% percentiles of the posterior means across simulations is presented as the shaded area.

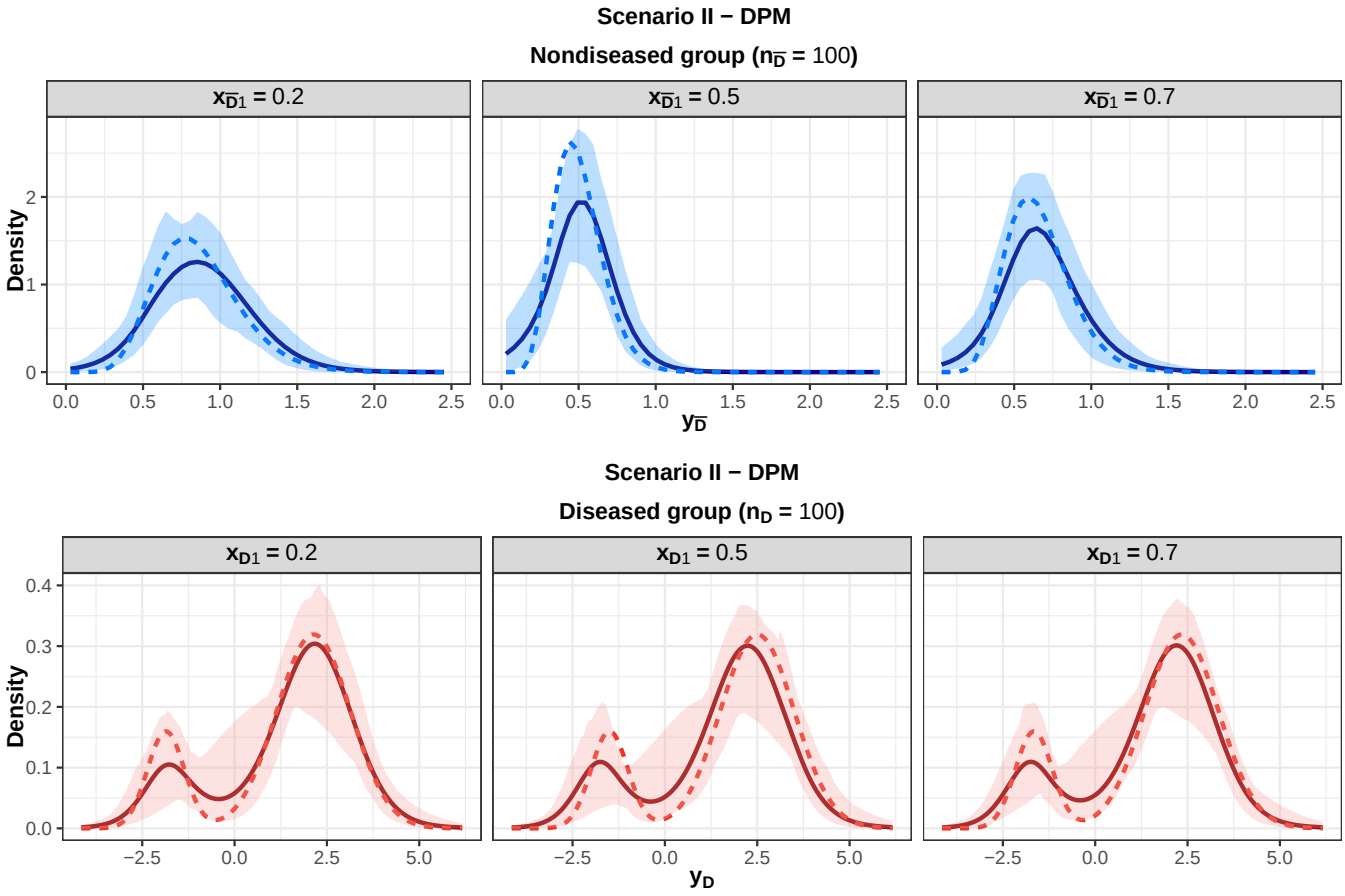

Figure 4.9: Conditional density functions for three particular values of the covariate under the DPM model for Scenario II and $(n_{\bar{D}}, n_D) = (100, 100)$. True (dotted line) conditional distribution and mean across simulations (solid line) of the posterior mean of the conditional density for the nondiseased (top) and diseased (bottom) group. A band constructed using the pointwise 2.5% and 97.5% percentiles of the posterior means across simulations is presented as the shaded area.

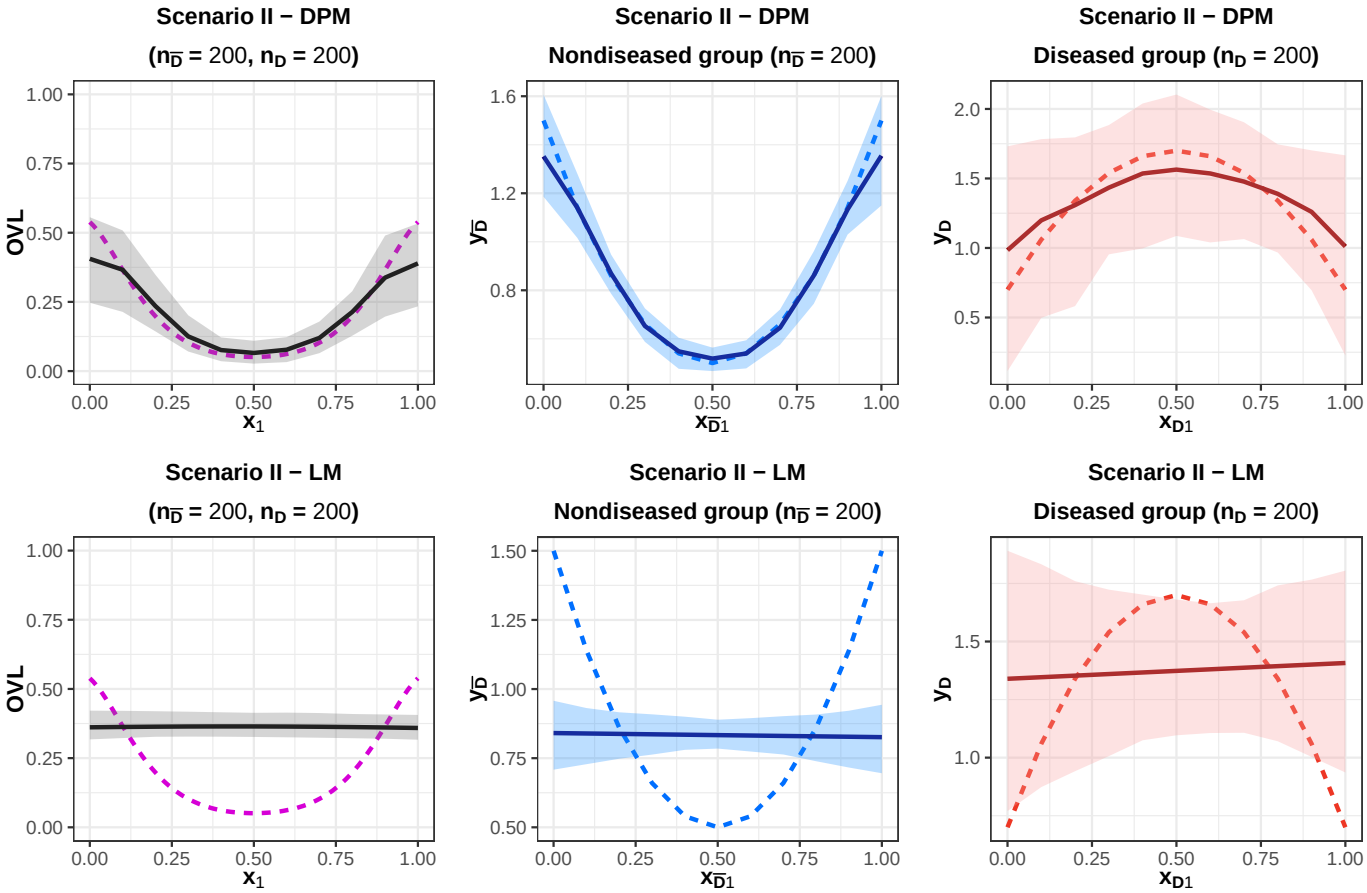

Figure 4.10: Results for the DPM model and the LM under Scenario II and $(n_{\bar{D}}, n_D) = (200, 200)$ are displayed in the top and bottom panels, respectively. True (dotted line) and mean across simulations (solid line) of the posterior mean of the covariate-specific OVL (left) and mean function of the nondiseased (middle) and diseased (right) group. A band constructed using the pointwise 2.5% and 97.5% percentiles of the posterior means across simulations is presented as the shaded area.

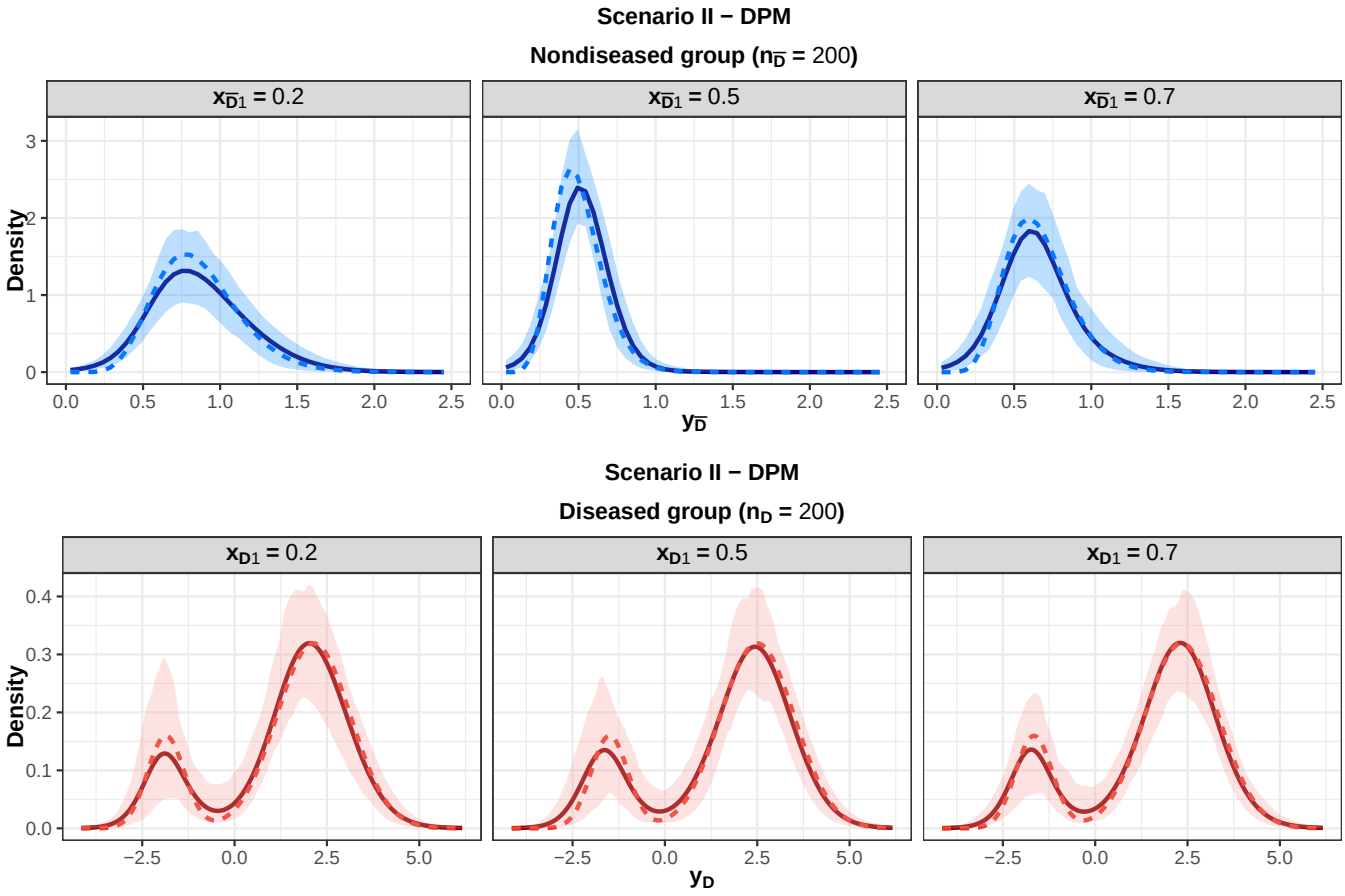

Figure 4.11: Conditional density functions for three particular values of the covariate under the DPM model for Scenario II and $(n_{\bar{D}}, n_D) = (200, 200)$. True (dotted line) conditional distribution and mean across simulations (solid line) of the posterior mean of the conditional density for the nondiseased (top) and diseased (bottom) group. A band constructed using the pointwise 2.5% and 97.5% percentiles of the posterior means across simulations is presented as the shaded area.

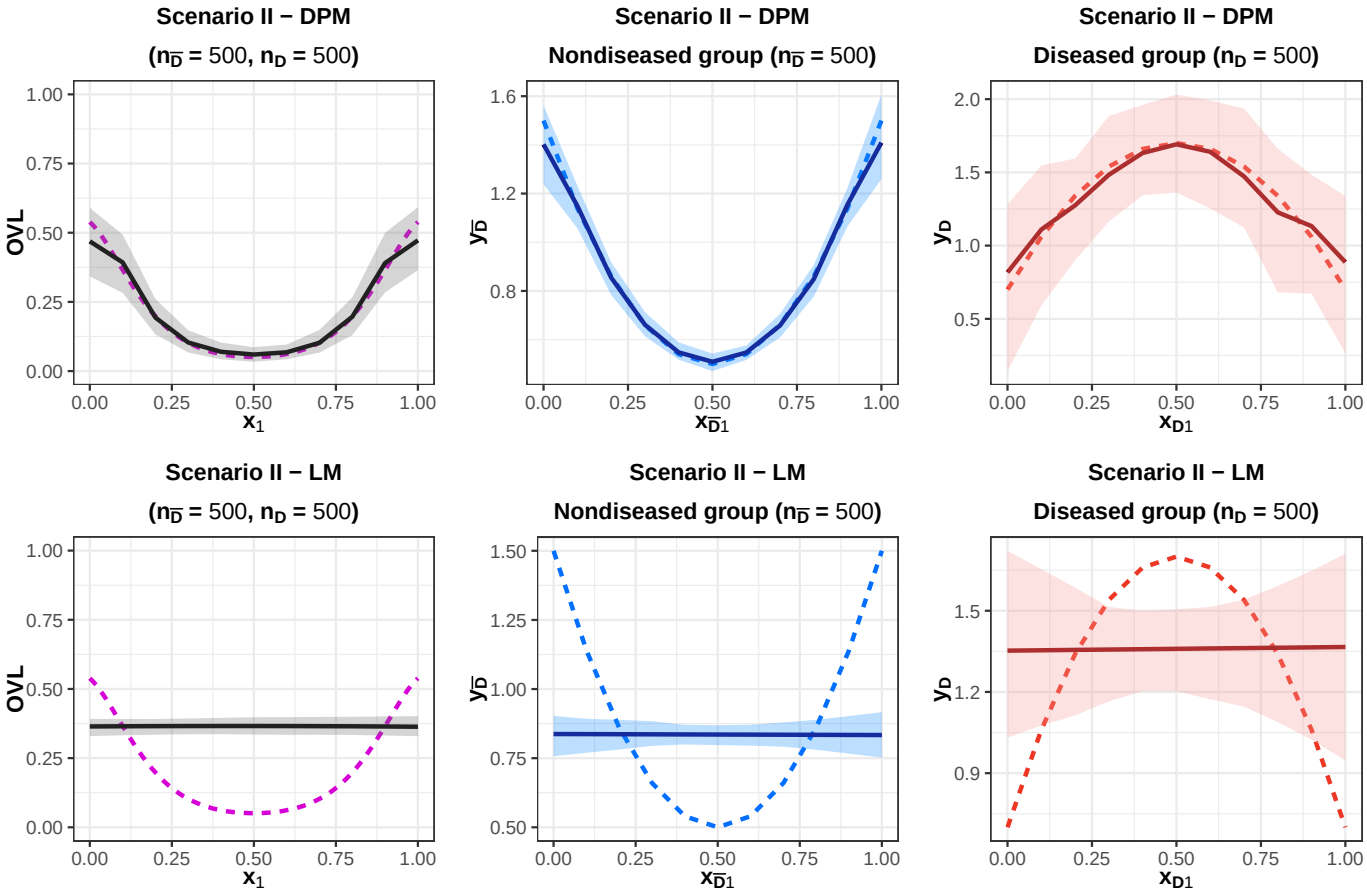

Figure 4.12: Results for the DPM model and the LM under Scenario II and $(n_{\bar{D}}, n_D) = (500, 500)$ are displayed in the top and bottom panels, respectively. True (dotted line) and mean across simulations (solid line) of the posterior mean of the covariate-specific OVL (left) and mean function of the nondiseased (middle) and diseased (right) group. A band constructed using the pointwise 2.5% and 97.5% percentiles of the posterior means across simulations is presented as the shaded area.

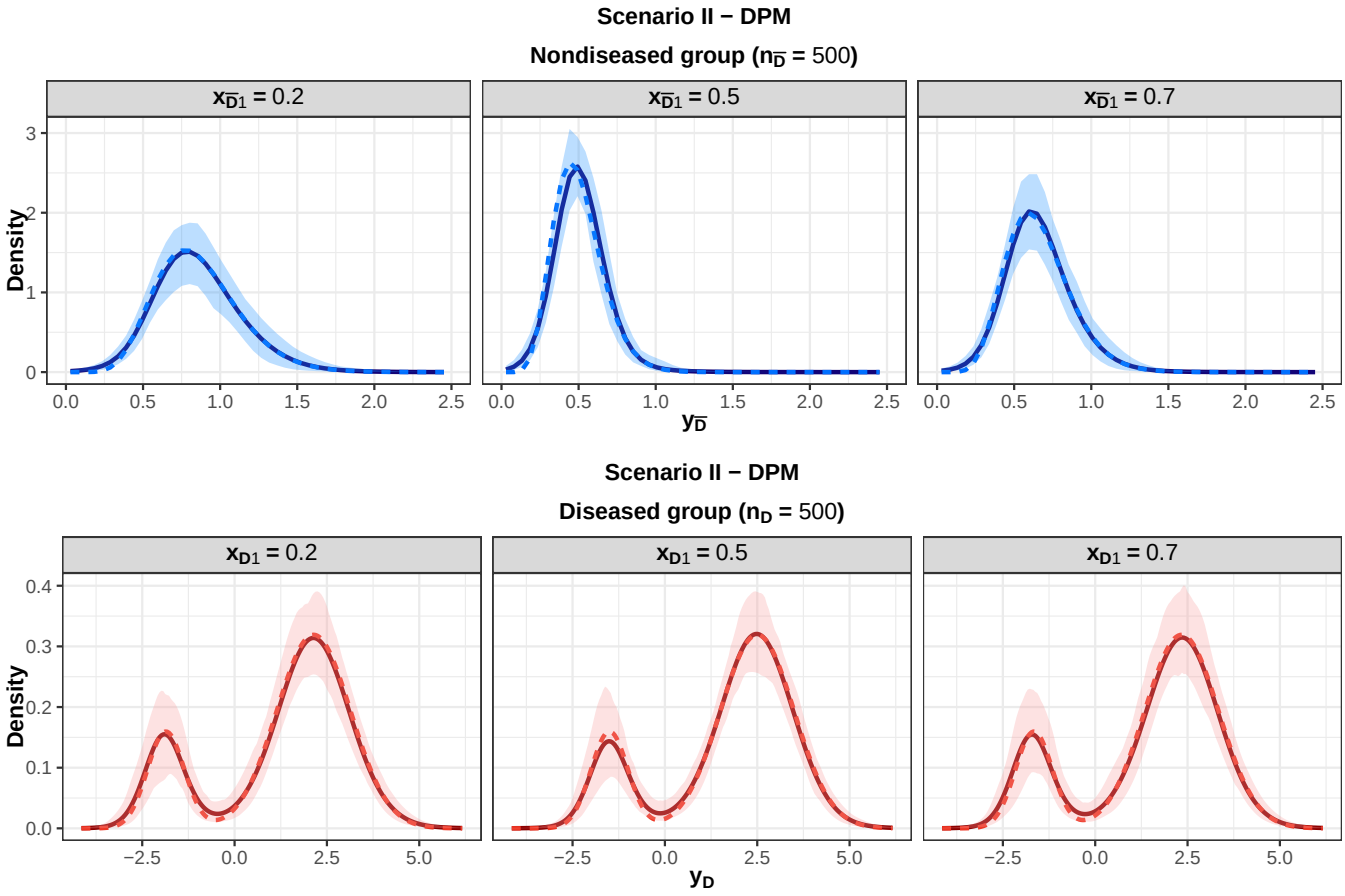

Figure 4.13: Conditional density functions for three particular values of the covariate under the DPM model for Scenario II and $(n_{\bar{D}}, n_D) = (500, 500)$. True (dotted line) conditional distribution and mean across simulations (solid line) of the posterior mean of the conditional density for the nondiseased (top) and diseased (bottom) group. A band constructed using the pointwise 2.5% and 97.5% percentiles of the posterior means across simulations is presented as the shaded area.

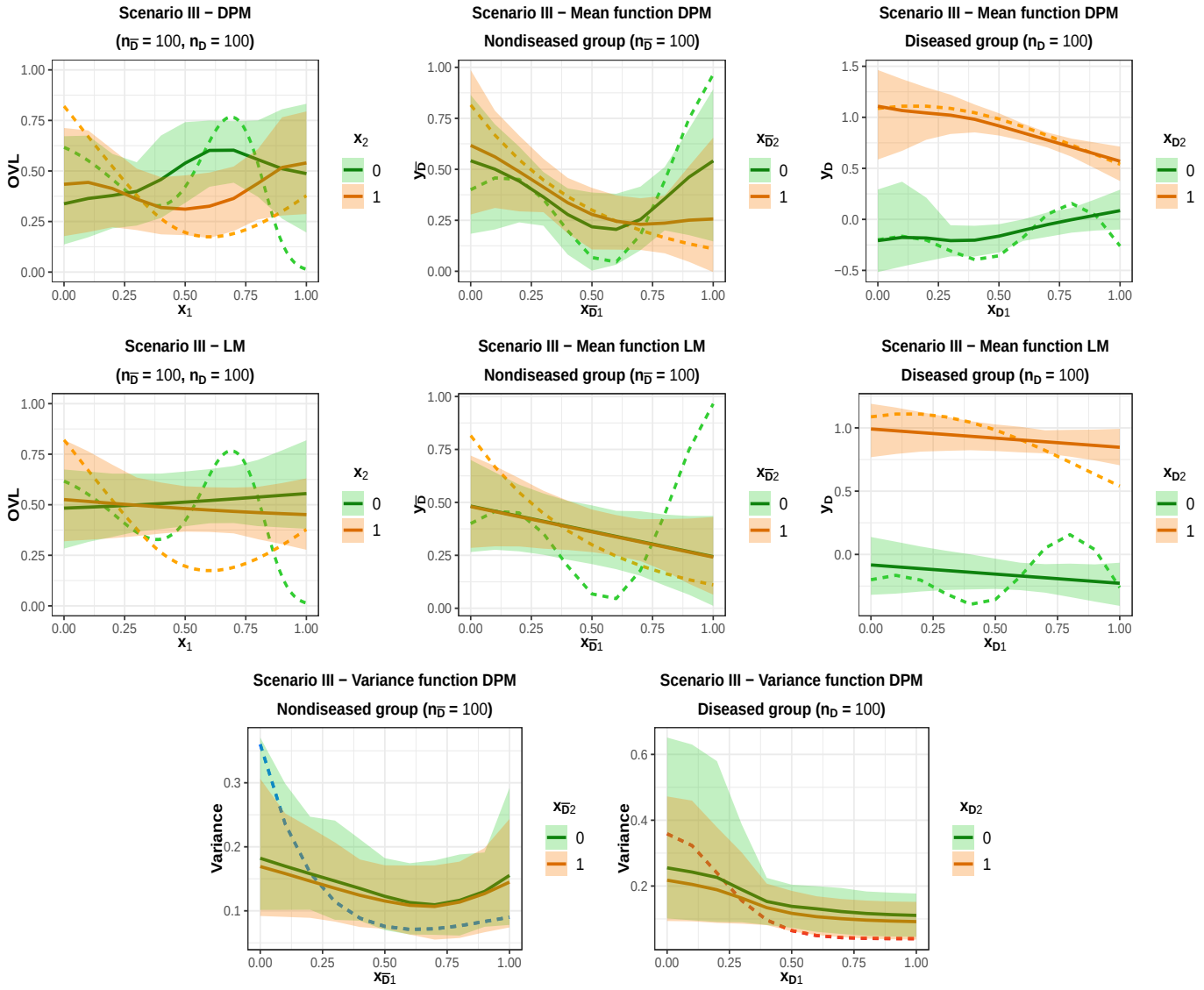

Figure 4.14: Results for the DPM model (top and bottom row) and the LM (middle row) under Scenario III and $(n_{\bar{D}}, n_D) = (100, 100)$. True (dotted line) and mean across simulations (solid line) of the posterior mean of the covariate-specific OVL (left) and mean/variance function of the nondiseased (middle) and diseased (right) group. A band constructed using the pointwise 2.5% and 97.5% percentiles of the posterior means across simulations is presented as the shaded area.

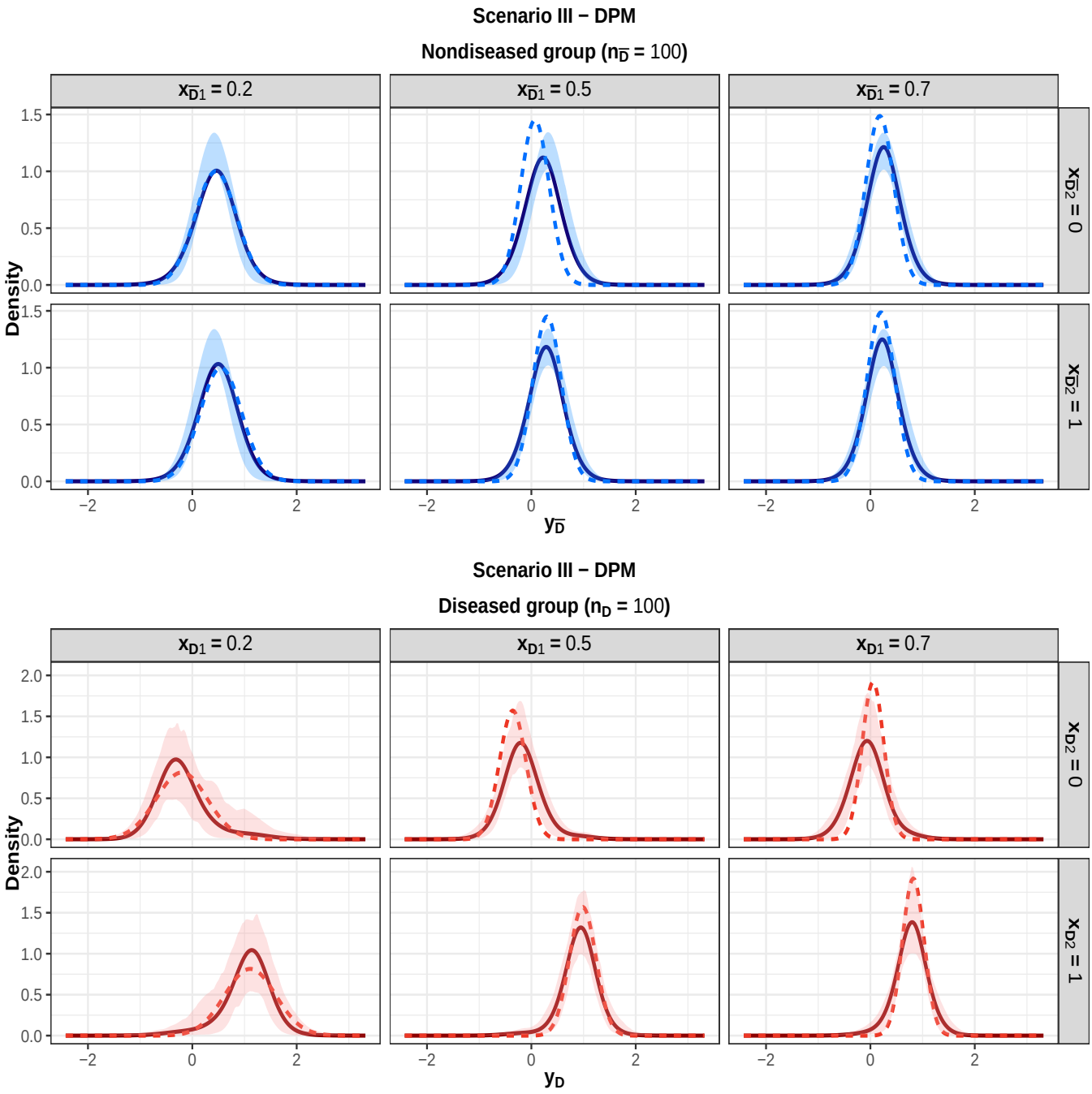

Figure 4.15: Conditional density functions for three particular values of the continuous covariate and for the two values of the binary covariate under the DPM model for Scenario III and $(n_{\bar{D}}, n_D) = (100, 100)$. True (dotted line) conditional distribution and mean across simulations (solid line) of the posterior mean of the conditional density for the nondiseased (top) and diseased (bottom) group. A band constructed using the pointwise 2.5% and 97.5% percentiles of the posterior means across simulations is presented as the shaded area.

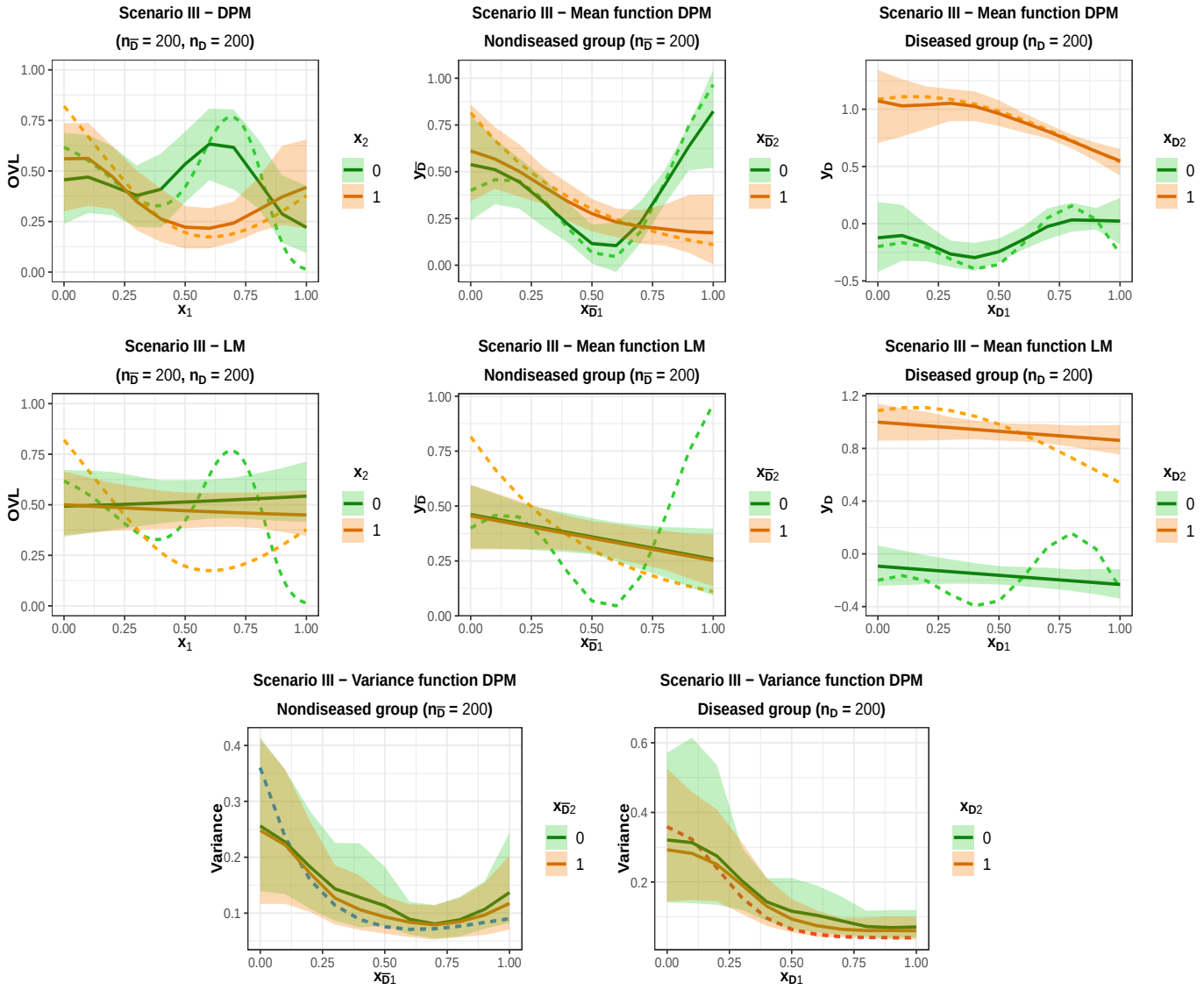

Figure 4.16: Results for the DPM model (top and bottom row) and the LM (middle row) under Scenario III and $(n_{\bar{D}}, n_D) = (200, 200)$. True (dotted line) and mean across simulations (solid line) of the posterior mean of the covariate-specific OVL (left) and mean/variance function of the nondiseased (middle) and diseased (right) group. A band constructed using the pointwise 2.5% and 97.5% percentiles of the posterior means across simulations is presented as the shaded area.

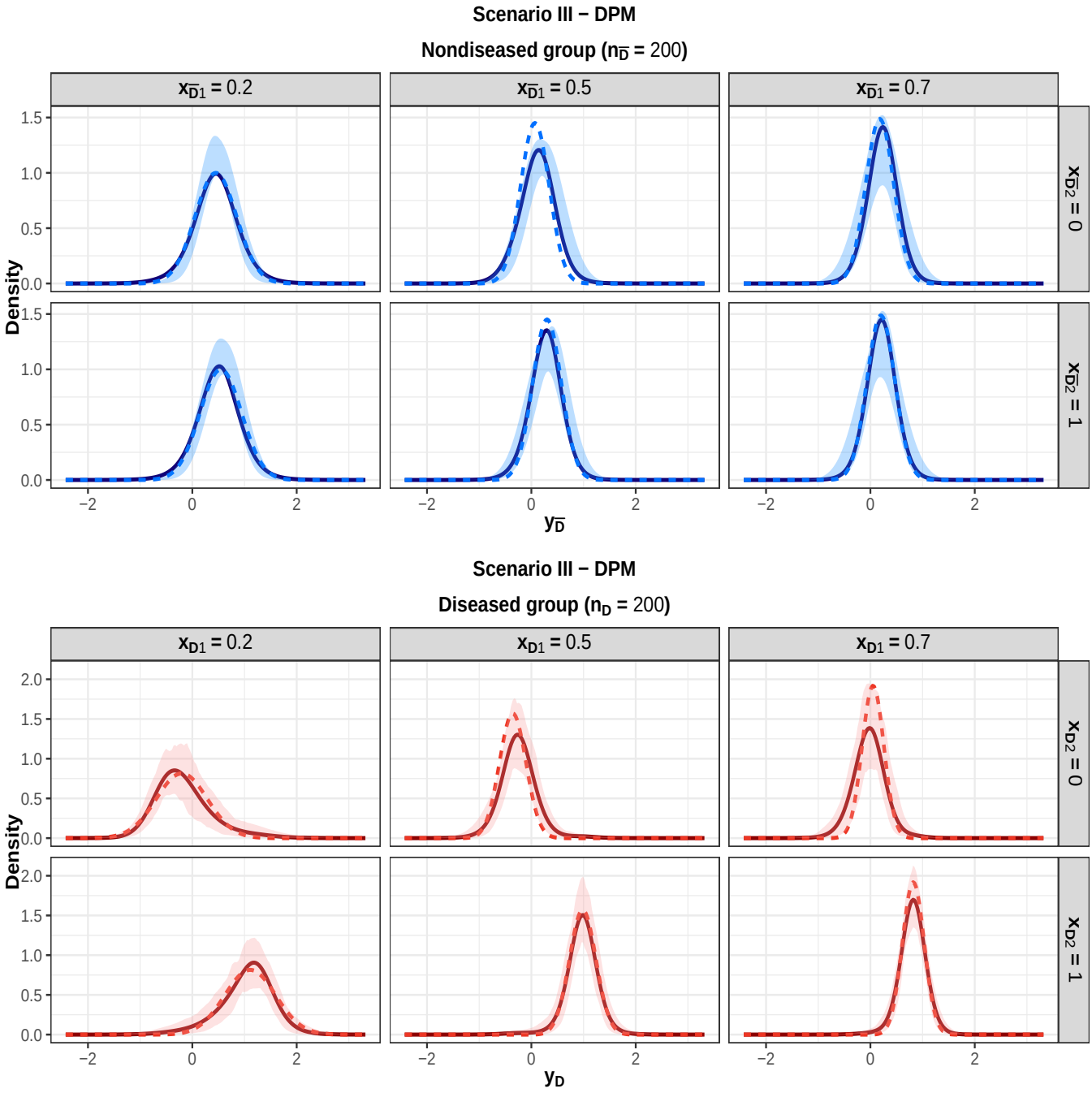

Figure 4.17: Conditional density functions for three particular values of the continuous covariate and for the two values of the binary covariate under the DPM model for Scenario III and $(n_{\bar{D}}, n_D) = (200, 200)$. True (dotted line) conditional distribution and mean across simulations (solid line) of the posterior mean of the conditional density for the nondiseased (top) and diseased (bottom) group. A band constructed using the pointwise 2.5% and 97.5% percentiles of the posterior means across simulations is presented as the shaded area.

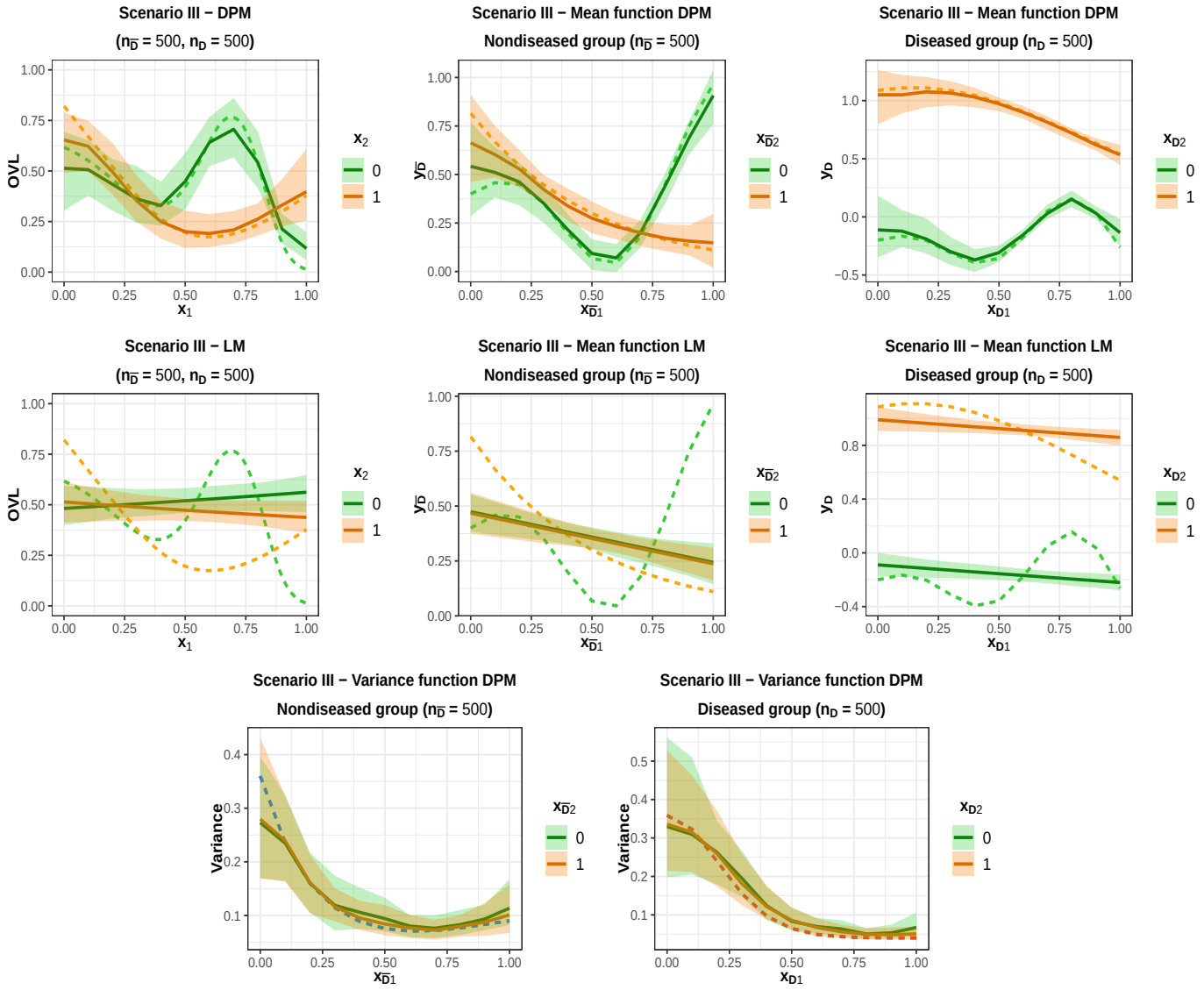

Figure 4.18: Results for the DPM model (top and bottom row) and the LM (middle row) under Scenario III and $(n_{\bar{D}}, n_D) = (500, 500)$. True (dotted line) and mean across simulations (solid line) of the posterior mean of the covariate-specific OVL (left) and mean/variance function of the nondiseased (middle) and diseased (right) group. A band constructed using the pointwise 2.5% and 97.5% percentiles of the posterior means across simulations is presented as the shaded area.

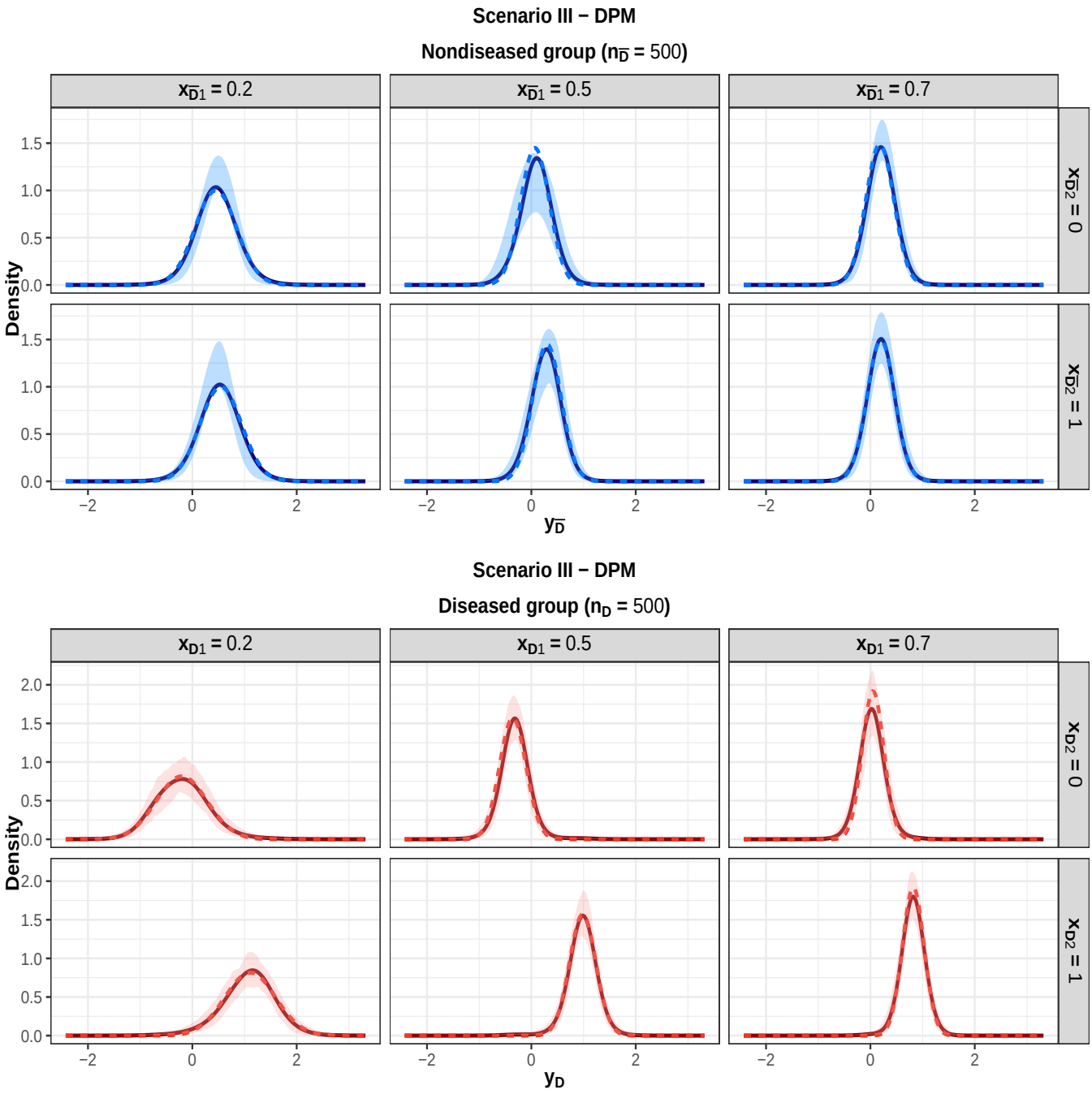

Figure 4.19: Conditional density functions for three particular values of the continuous covariate and for the two values of the binary covariate under the DPM model for Scenario III and $(n_{\bar{D}}, n_D) = (500, 500)$. True (dotted line) conditional distribution and mean across simulations (solid line) of the posterior mean of the conditional density for the nondiseased (top) and diseased (bottom) group. A band constructed using the pointwise 2.5% and 97.5% percentiles of the posterior means across simulations is presented as the shaded area.

Table 4.1: Scenario I. Average of the empirical coverage probability of the 95% credible intervals

$(n_{\bar{D}}, n_D)$	**OVL**	**Mean function**	
		Nondiseased	**Diseased**
(100, 100)	0.94	0.95	0.95
(200, 200)	0.94	0.96	0.96
(500, 500)	0.97	0.98	0.97

Table 4.2: Scenario II. Average of the empirical coverage probability of the 95% credible intervals

$(n_{\bar{D}}, n_D)$	**OVL**	**Mean function**	
		Nondiseased	**Diseased**
(100, 100)	0.58	0.85	0.84
(200, 200)	0.82	0.89	0.87
(500, 500)	0.94	0.93	0.95

Table 4.3: Scenario III. Average of the empirical coverage probability of the 95% credible intervals

$\boldsymbol{X}_2$	$(n_{\bar{D}}, n_D)$	**OVL**	**Mean function**		**Variance function**	
			Nondiseased	**Diseased**	**Nondiseased**	**Diseased**
	(100, 100)	0.6	0.65	0.61	0.61	0.43
0	(200, 200)	0.75	0.88	0.72	0.87	0.64
	(500, 500)	0.84	0.89	0.87	0.92	0.84
	(100, 100)	0.57	0.78	0.84	0.61	0.52
1	(200, 200)	0.89	0.91	0.94	0.9	0.8
	(500, 500)	0.93	0.91	0.95	0.94	0.88

Table 4.4: Scenario I. Model comparison criteria (Average over 100 simulated data sets)

$(n_{\bar{D}}, n_D)$	**Group**	**Criterion**	**DPM**	**LM**	**Empirical prob.**
(100, 100)	Nondiseased	Adjusted-LPML	-121.67	**-121.48**	0.56
		WAIC	243.28	**242.95**	0.56
		DIC_3	242.91	**242.72**	0.59
		Posterior rank probability	**0.5027**	0.4973	0.65
	Diseased	Adjusted-LPML	-74.15	**-73.91**	0.51
		WAIC	148.24	**147.8**	0.51
		DIC_3	147.84	**147.57**	0.57
		Posterior rank probability	0.4908	**0.5092**	0.61
(200, 200)	Nondiseased	Adjusted-LPML	-240.72	**-239.72**	0.26
		WAIC	481.35	**479.43**	0.27
		DIC_3	480.8	**479.31**	0.33
		Posterior rank probability	0.3106	**0.6894**	0.29
	Diseased	Adjusted-LPML	-147.5	**-146.58**	0.26
		WAIC	294.91	**293.16**	0.26
		DIC_3	294.38	**293.04**	0.28
		Posterior rank probability	0.3198	**0.6802**	0.26
(500, 500)	Nondiseased	Adjusted-LPML	-601.65	**-597.69**	0.03
		WAIC	1203.18	**1195.38**	0.03
		DIC_3	1202.28	**1195.33**	0.06
		Posterior rank probability	0.1182	**0.8818**	0.03
	Diseased	Adjusted-LPML	-370.45	**-366.46**	0.06
		WAIC	740.78	**732.92**	0.06
		DIC_3	739.85	**732.87**	0.06
		Posterior rank probability	0.1234	**0.8766**	0.06

Table 4.5: Scenario II. Model comparison criteria (Average over 100 simulated data sets)

$(n_{\bar{D}}, n_D)$	**Group**	**Criterion**	**DPM**	**LM**	**Empirical prob.**
(100, 100)	Nondiseased	Adjusted-LPML	**-12.92**	-53.39	1.0
		WAIC	**25.33**	106.74	1.0
		DIC_3	**23.07**	106.19	1.0
		Posterior rank probability	**1.0**	0.0	1.0
	Diseased	Adjusted-LPML	**-189.05**	-206.28	0.98
		WAIC	**377.79**	412.54	0.98
		DIC_3	**376.38**	412.34	0.98
		Posterior rank probability	**0.9556**	0.0444	0.98
(200, 200)	Nondiseased	Adjusted-LPML	**-13.37**	-106.85	1.0
		WAIC	**26.37**	213.7	1.0
		DIC_3	**24.07**	213.37	1.0
		Posterior rank probability	**1.0**	0.0	1.0
	Diseased	Adjusted-LPML	**-365.22**	-409.95	1.0
		WAIC	**730**	819.9	1.0
		DIC_3	**727.75**	819.79	1.0
		Posterior rank probability	**0.9999**	0.0001	1.0
(500, 500)	Nondiseased	Adjusted-LPML	**-15.2**	-266.05	1.0
		WAIC	**30.06**	532.1	1.0
		DIC_3	**27.61**	531.95	1.0
		Posterior rank probability	**1.0**	0.0	1.0
	Diseased	Adjusted-LPML	**-904.26**	-1025.37	1.0
		WAIC	**1808.18**	2050.73	1.0
		DIC_3	**1805.65**	2050.69	1.0
		Posterior rank probability	**1.0**	0.0	1.0

Table 4.6: Scenario III. Model comparison criteria (Average over 100 simulated data sets)

$(n_{\bar{D}}, n_D)$	**Group**	**Criterion**	**DPM**	**LM**	**Empirical prob.**
(100, 100)	Nondiseased	Adjusted-LPML	**-42.57**	-52.53	0.95
		WAIC	**84.75**	105.02	0.95
		DIC_3	**83.07**	104.53	0.94
		Posterior rank probability	**0.918**	0.082	0.95
	Diseased	Adjusted-LPML	**-39.01**	-49.99	0.95
		WAIC	**77.6**	99.97	0.96
		DIC_3	**75.66**	99.44	0.97
		Posterior rank probability	**0.9141**	0.0859	0.95
(200, 200)	Nondiseased	Adjusted-LPML	**-71.18**	-106.97	1.0
		WAIC	**141.84**	213.94	1.0
		DIC_3	**139.29**	213.64	1.0
		Posterior rank probability	**0.9996**	0.0004	1.0
	Diseased	Adjusted-LPML	**-62.07**	-97.87	1.0
		WAIC	**123.79**	195.72	1.0
		DIC_3	**121.1**	195.42	1.0
		Posterior rank probability	**0.9996**	0.0004	1.0
(500, 500)	Nondiseased	Adjusted-LPML	**-157.65**	-256.74	1.0
		WAIC	**314.97**	513.48	1.0
		DIC_3	**312.52**	513.36	1.0
		Posterior rank probability	**1.0**	0.0	1.0
	Diseased	Adjusted-LPML	**-126.63**	-247.26	1.0
		WAIC	**252.88**	494.53	1.0
		DIC_3	**250.07**	494.39	1.0
		Posterior rank probability	**1.0**	0.0	1.0

4.4 Applications

We applied our methods to two different real datasets. The first application is aimed to investigate the effect of age on the glucose levels as a biomarker for diabetes. The second application is intended to assess the effect of age and sex on different biomarkers for Alzheimer's disease.

4.4.1 Glucose level as biomarker for diabetes

Diabetes mellitus is a condition in which glucose levels in blood are too high and cannot be controlled by the lack of a hormone called insulin. This may be due to the immune system (Type 1) or because the body is unable to produce it in sufficient quantity (Type 2). In recent years, its prevalence in the world population has increased. According to estimates by the World Health Organization in 2016, the number of adults with diabetes doubled since 2014, from 4.7% to 8.5% of the adult population (World Health Organization, 2016). Moreover, in the same year, diabetes caused over 1.6 million of deaths (4%) (World Health Organization, 2018, p. 7). Among other risk factors, it is closely related to being overweight, a condition also responsible for other diseases that are among the leading causes of death worldwide (e.g., cardiovascular diseases). Unfortunately there is no cure yet, but an early diagnosis can help to improve patients' lives. Conversely, if the patients do not have enough care or proper treatment, they might develop other serious medical complications, such as blindness or leg amputations. For all the reasons mentioned above, diabetes is considered as a serious public health problem. Numerous institutions around the world have gathered efforts to combat this condition (e.g., Diabetes UK in the UK, Federación Mexicana de Diabetes in Mexico, among others). However, more work is needed along with better health policies on each government.

We have applied our method to data from a population based survey of diabetes in Cairo, Egypt (Smith and Thompson, 1996). The data comprises measurements on postprandial (after a meal) blood glucose obtained from a finger stick on 88 subjects with diabetes and 198 non diabetic. Our primary goal is to evaluate the effect of age in the accuracy of glucose as a biomarker of diabetes. We centred and scaled the data and we used the same prior information described in Section 4.3.2. Posterior inference was based on 30,000 MCMC iterations after discarding 30,000 iterations as burn-in period. Further, we

used trace plots and the Geweke diagnostic to assess the convergence of our MCMC algorithm. Note that our object of interest are not the parameters of the mixture model, but the conditional density itself. Results are presented in Figures 4.20–4.22 and 4.23, respectively, and they do not suggest lack of convergence. Figure 4.24 shows that the effective sample size is high enough as well. Also, we have fitted a linear model (LM) for comparing purposes.

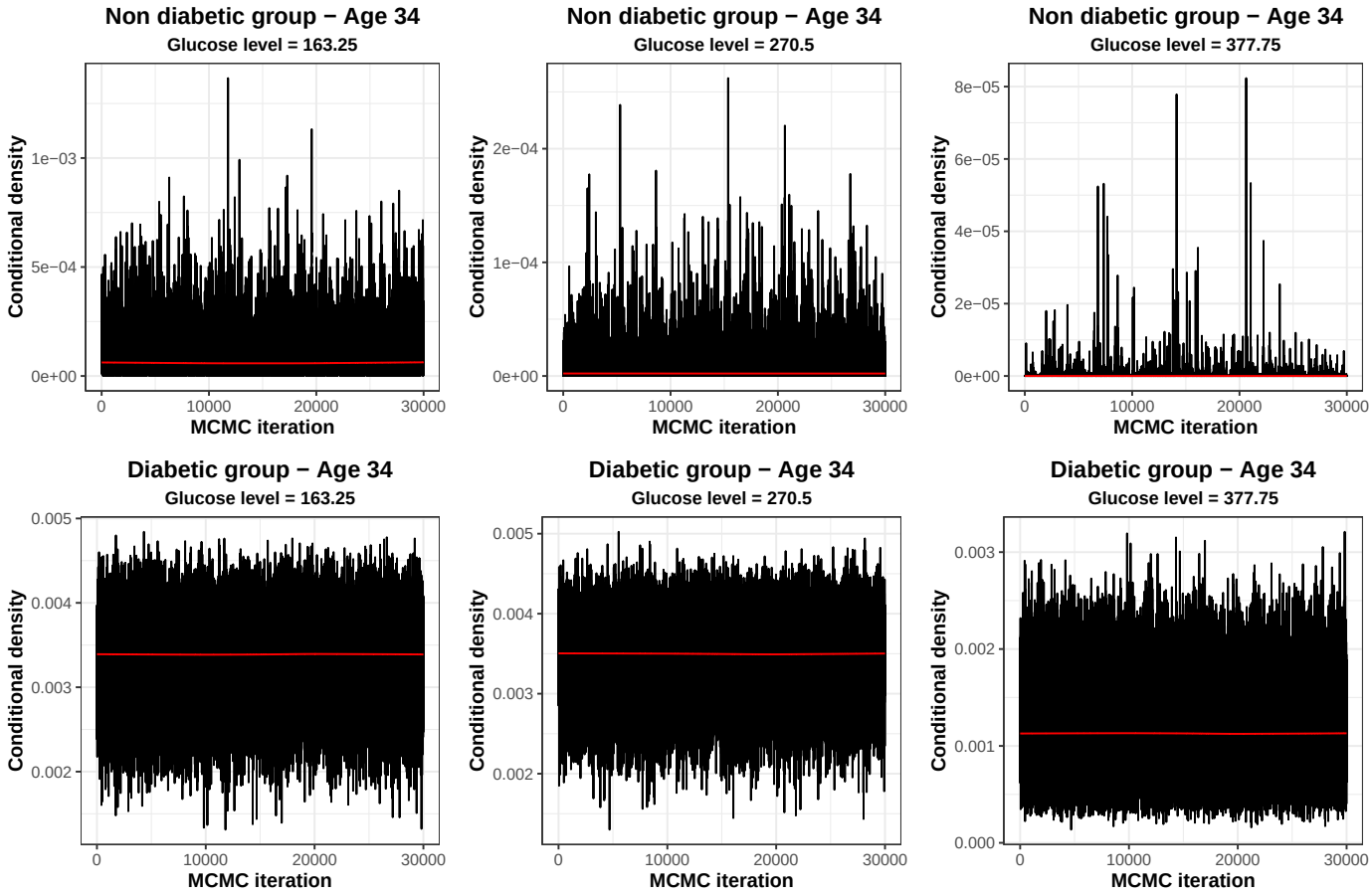

Figure 4.20: Trace plots of three particular values of the glucose levels density for a specific age in the nondiseased (top) and diseased (bottom) group. The red line represents the mean regression line of a LOWESS model. We use it to check if there is any sort of trend in the chain.

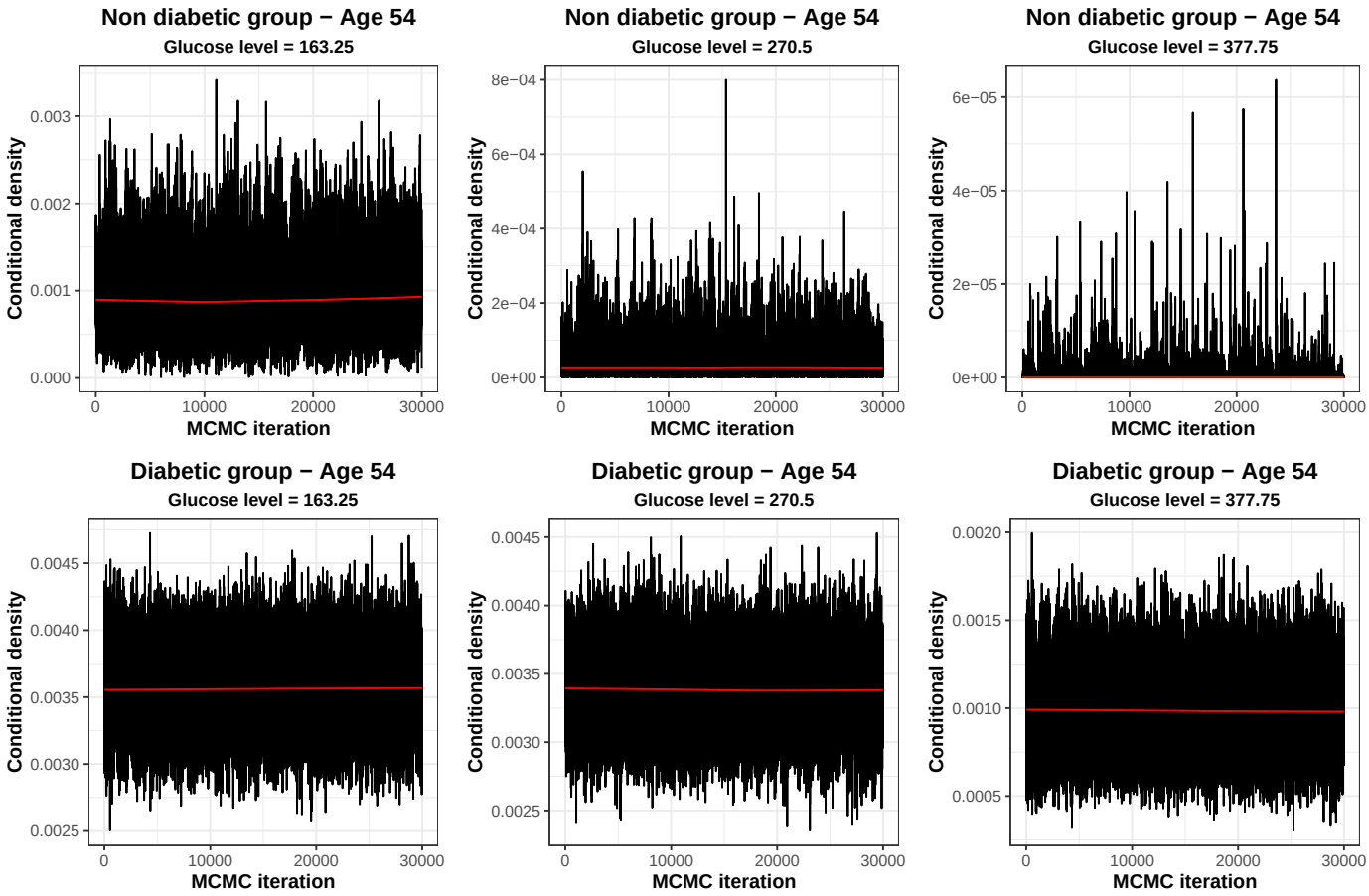

Figure 4.21: Trace plots of three particular values of the glucose levels density for a specific age in the nondiseased (top) and diseased (bottom) group. The red line represents the mean regression line of a LOWESS model. We use it to check if there is any sort of trend in the chain.

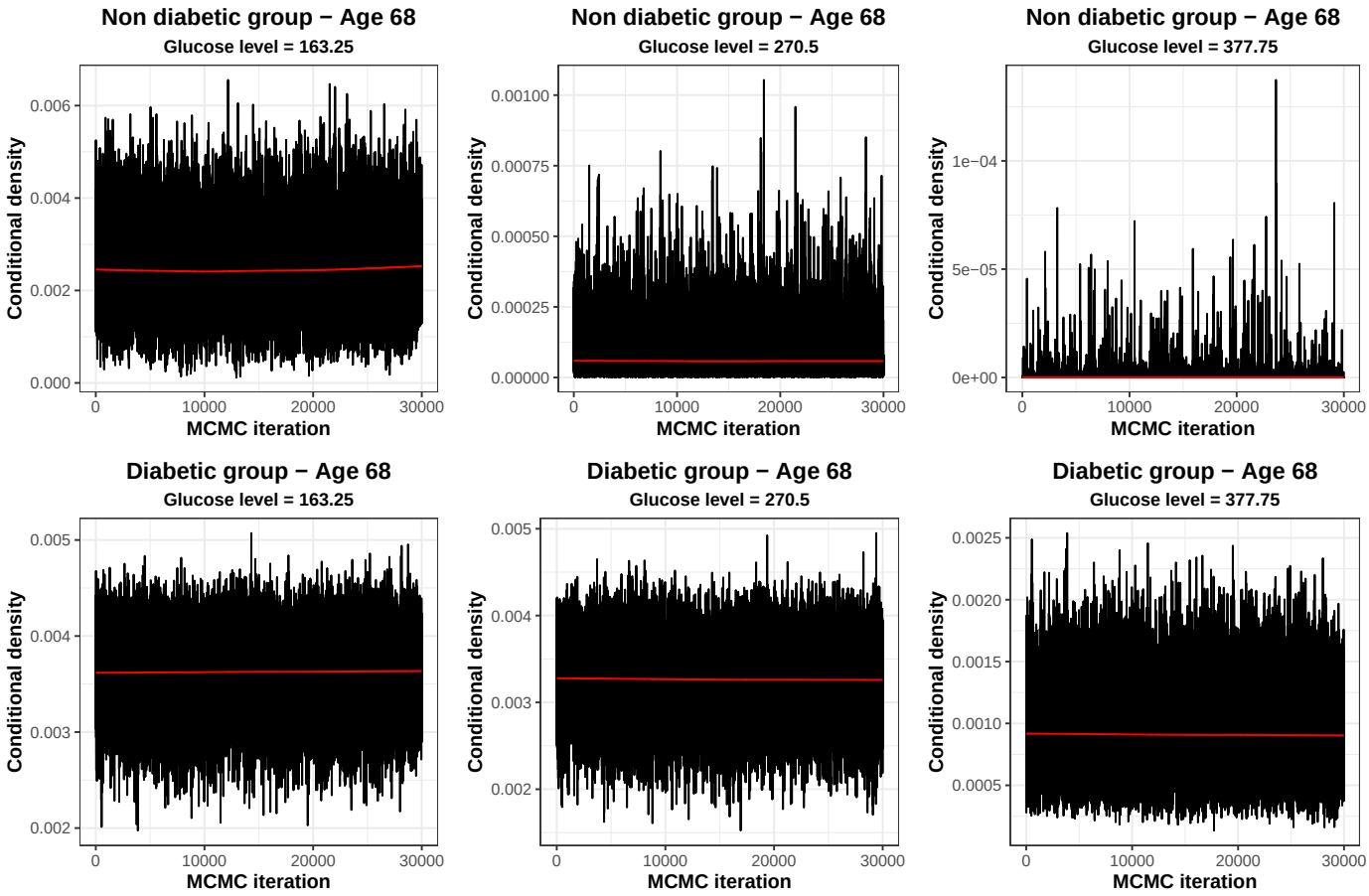

Figure 4.22: Trace plots of three particular values of the glucose levels density for a specific age in the nondiseased (top) and diseased (bottom) group. The red line represents the mean regression line of a LOWESS model. We use it to check if there is any sort of trend in the chain.

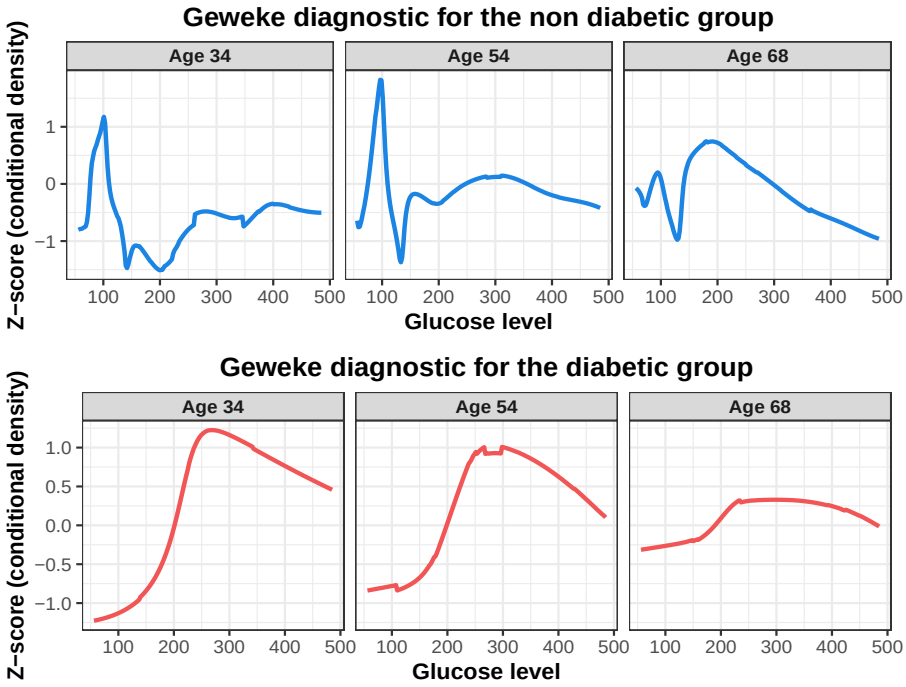

Figure 4.23: Geweke diagnostic for the conditional density across a grid of the glucose levels for the nondiseased (top) and diseased (bottom) group.

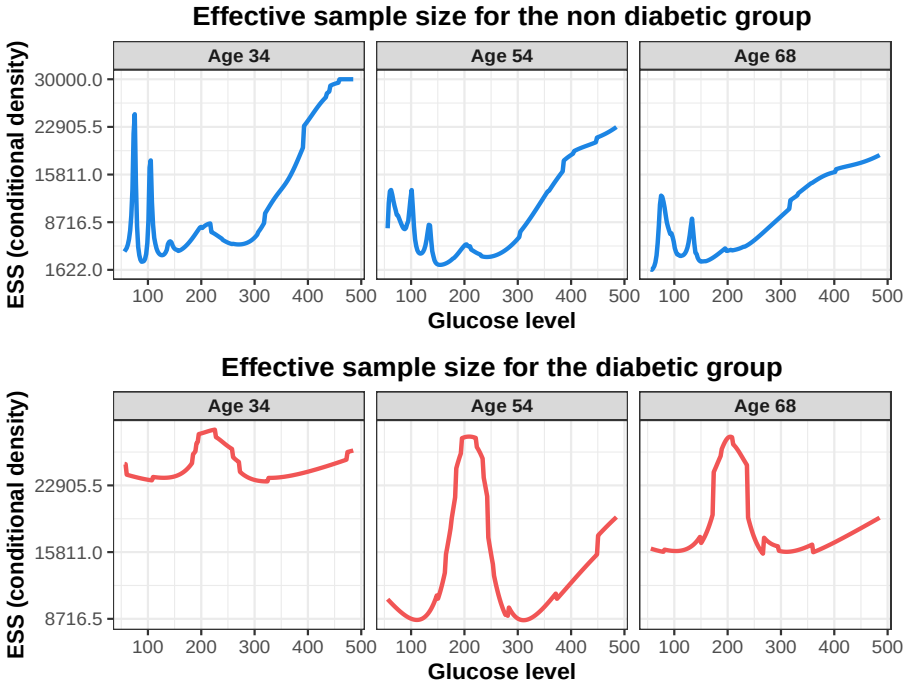

Figure 4.24: Effective simple size for the conditional density across a grid of the glucose levels for the nondiseased (top) and diseased (bottom) group.

The results from the DPM model are presented in the top row of Figure 4.25. We observe that the estimated covariate-specific OVL (top left) increases steadily with age, meaning less diagnosis accuracy. Credible bands widen after 80 years old. This is due to the fact that there are no data on diabetic subjects for this age range, meaning higher uncertainty. Figure 4.25 top middle shows the mean function of the non diabetic group, we observe a slight increase as age increases. Whereas for the diseased group (top right), we observe an almost imperceptible linear decrease. Although the LM manages to capture the increasing trend of the OVL, the credible band is much narrower (Figure 4.25 bottom left), which implies less uncertainty even for subjects over 80 years old.

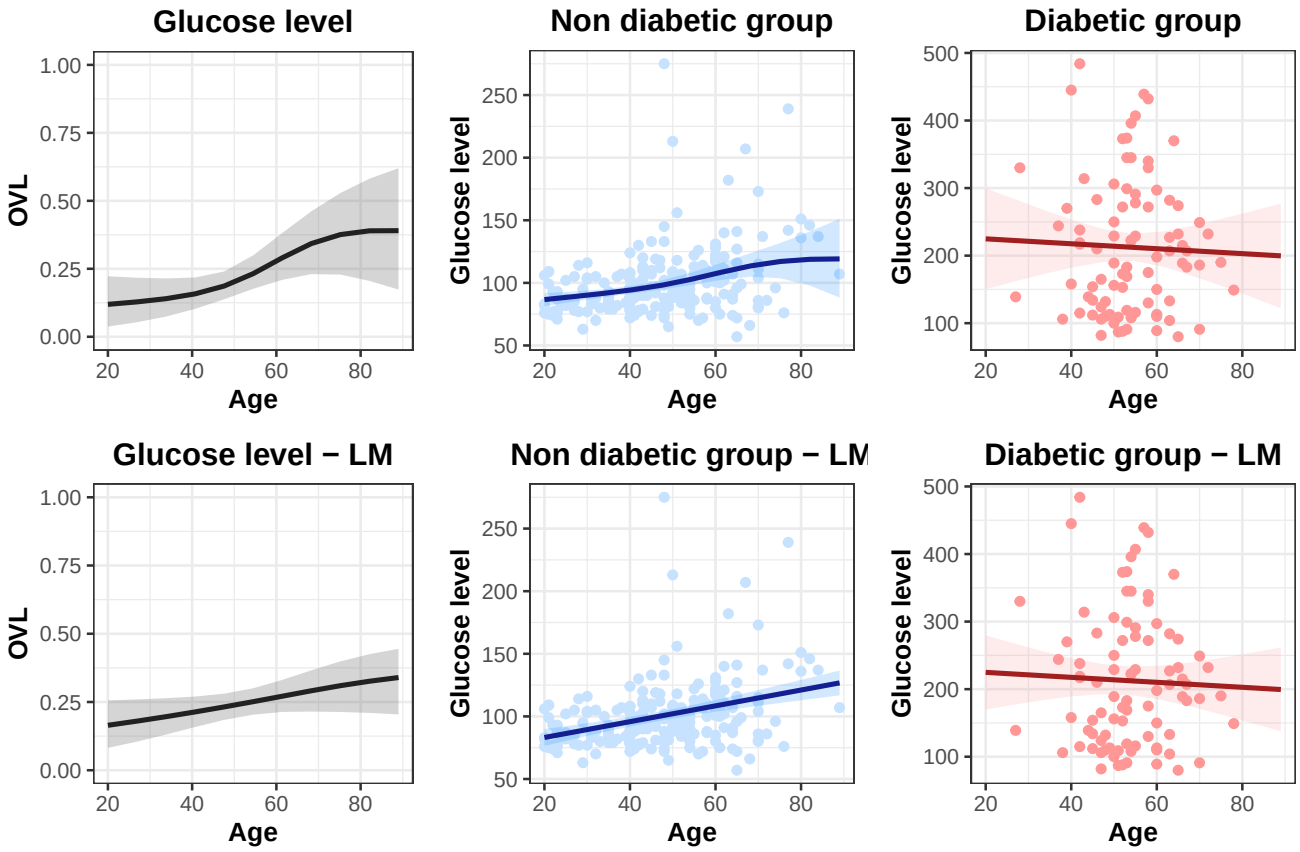

Figure 4.25: Results from the DPM model (top row) and LM (bottom row). Posterior mean (solid line) and 95% pointwise credible bands (shaded area) of the covariate-specific OVL (left), mean function of the non diabetic (middle) and diabetic (right) group.

In the top row of Figure 4.26, there are depicted examples of conditional densities of both groups for three particular ages (quartiles). In these plots, it is possible to note that, the conditional densities

of the diabetic group have heavier tails, as expected. To check the goodness of fit of our model, we performed posterior predictive checks in both groups. The idea is very simple, replicated data using the posterior predictive distribution should look similar to the observed data. More precisely, we can draw from the posterior predictive distribution using the posterior samples. For example, using the same notation introduced in Section 4.2.2, let $\tilde{\boldsymbol{u}}_D = (\tilde{y}_D, \tilde{x}_D)$ be a new observation of the glucose level and age of an individual in the diabetic group, respectively. Then, new observations can be generated as follows

$$\tilde{\boldsymbol{u}}_D^{(s)} \overset{\text{iid}}{\sim} \sum_{l=1}^{L_D} \omega_{Dl}^{(s)} \mathrm{N}\left(\boldsymbol{u}_D \mid \boldsymbol{\mu}_{Dl}^{(s)}, \boldsymbol{\Sigma}_{Dl}^{(s)}\right), \quad s = 1, \ldots, S.$$

Note that we can drop the simulated ages, $\tilde{x}_D^{(s)}$, to obtain samples from the marginal posterior predictive distribution of the glucose level, our variable of interest. Then, we can compute some statistics based on those samples to help us know if there are discrepancies that indicate potential failings of our model. Bottom row of Figure 4.26 shows the posterior predictive checks for both models. Here, we observe that our model seems to produce less skewed distributions for the diseased group (bottom right). However, there are no clear signs of misfit of our model. It is important to bear in mind that when modelling strictly positive test outcomes, like in this case, there might be replications where the minimum is negative (statistic not shown). The reason is that we are using a normal distribution as part of the kernel of the DPM model. Hence, it might assign positive probability mass to negative values.

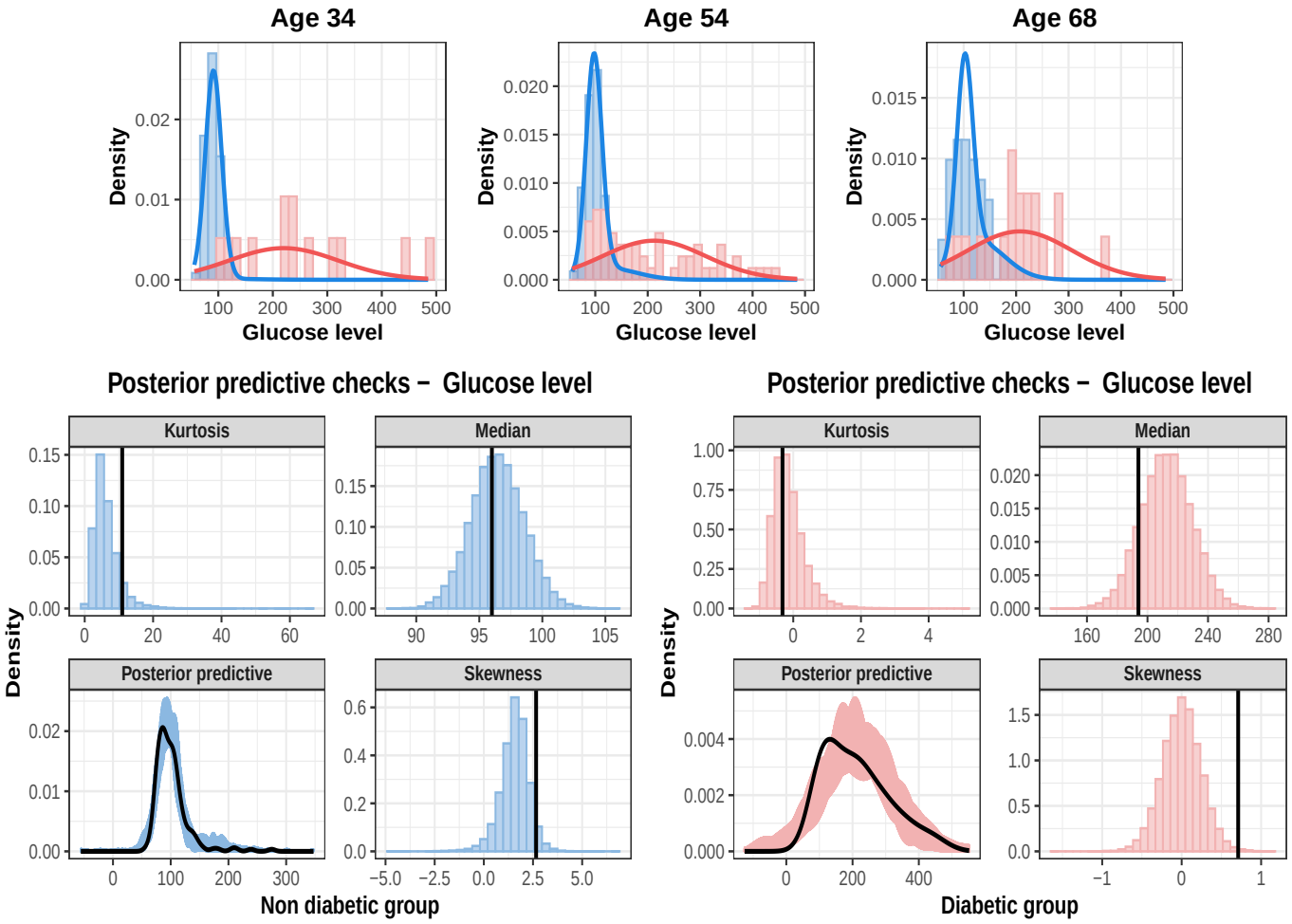

Figure 4.26: Top row: conditional histograms along with the estimated conditional densities of nondiseased (blue) and diseased (red) group (posterior mean and 95% credible bands) for three specific ages. Bottom row: Posterior predictive checks for the nondiseased (left) and diseased (right) group.

Finally, Table 4.7 shows the model comparison criteria results. We note that the DPM model provides a better fit for the non diabetic group. Although, the LPML, WAIC and DIC_3 are all very similar for both models in the diabetic goup, these criteria suggest that the LM may be a better choice. In conclusion, there seems to be evidence of less diagnostic accuracy of glucose levels as age increases.

Table 4.7: Diabetes model comparison criteria

Group	Criterion	DPM	LM
Non diabetic	Adjusted LPML	**-879.14**	-934.32
	WAIC	**1757.74**	1868.87
	DIC_3	**1755.98**	1867.37
	Posterior rank probability	**1.0**	0.0
Diabetic	Adjusted LPML	-532.08	**-531.66**
	WAIC	1064.14	**1063.31**
	DIC_3	1063.91	**1063.15**
	Posterior rank probability	0.0412	**0.9588**

4.4.2 Biomarkers for Alzheimer's disease

Alzheimer's disease (AD) is a chronic neurodegenerative disease, i.e., it is a progressive condition that affects multiple brain functions (loss of the structure or function of neurons, including their death) and in which symptoms gradually develop over the years and eventually become more severe. It is commonly associated with gradual memory loss and other cognitive abilities, i.e., the ability to process thought. Around 60-70% of dementia cases are due to AD (World Health Organization, 2019). The exact cause of AD is not yet fully understood, although there are certain factors that may increase the risk of developing the condition, such as increasing age, a family history, untreated depression or other lifestyle factors and conditions associated with cardiovascular disease (Reitz and Mayeux, 2014). Usually, AD occurs in three stages: early (cognitively normal), middle (mild cognitive impairment) and late (AD) stage (World Health Organization, 2019). There is currently no cure for AD. However, numerous ongoing clinical studies and research are adding up all their efforts to find new treatments, as well as possible biomarkers that help the early detection of this disease.

One of these major efforts is the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Mueller et al., 2005), which is a longitudinal multicenter study designed to develop clinical, imaging, genetic, and biochemical biomarkers for the early detection and tracking of AD. It has made major contributions

to AD research, enabling the sharing of data between researchers around the world (Toga and Crawford, 2015). According to the World Health Organization (2019), one of the main goals is the early diagnosis. In this way, the ADNI study aims to detect AD at the earliest possible stage (pre-dementia) and identify ways to track the disease's progression with certain biomarkers. Many authors have pointed out the cerebrospinal fluid (CSF) amyloid-β (Aβ) and the CSF tau as potential biomarkers for AD (Jack et al., 2016; Welge et al., 2009; Blennow and Zetterberg, 2018). Aβ is a measure of Aβ42 in cerebrospinal fluid biospecimen and reflects the accumulation of the proteins amyloid-β in the brain. Whereas, the CSF tau is a measure of tau protein in cerebrospinal fluid biospecimen, an increase is associated to neuronal injury. Other recognised biomarkers include the hippocampal volume (Rahman et al., 2016) and the hypometabolic convergence index (HCI) (Chen et al., 2011). The hippocampus is responsible of multiple brain functions associated with memory, thus a change in its volume would suggest signs of dementia, and the HCI is a summary measure of different scores produced by the cerebral hypometabolism of a subject and it proved to be effective in identifying the disease (Chen et al., 2011).

ADNI participants undergo a series of tests that are repeated over time. These include, among others, clinical evaluations and neuropsychological tests. Participants who are cognitively normal (CN) are control subjects that do not show signs of depression, mild cognitive impairment or dementia. In contrast, mild cognitive impairment (MCI) subjects present memory complaints, but show an absence of significant levels of impairment in other cognitive domains, including preserving activities of daily living. Finally, AD subjects are considered as probable dementia patients according to AD-specific criteria, such as the Clinical Dementia Rating (CDR). Subjects can only be diagnosed with AD after death (e.g., through a brain biopsy) and therefore labels of the groups may contain potential noise.

We used data from the ADNI study to investigate the ability of Aβ, Tau, hippocampal volume and HCI to discriminate early and middle stages (CN and MCI subjects collapsed) from late stage participants. CN and MCI subjects were grouped together for two reasons: i) the coefficient of overlap is currently only defined for two classes, and ii) the OVL values are not largely different between CN and MCI groups. In addition, preliminary analyses suggested similarities in the conditional distributions of CN and MCI groups. Therefore, we refer to this group as the "nondiseased" and to late stage patients

as the "diseased" group. The way we collapsed the data is solely for illustration purposes and it is data-driven rather than application-based. In practice, however, it is important to differentiate between cognitively normal and mild cognitive impairment individuals because merging these two classes might vanish or hide important features of one or both groups. Data comprise 894 nondiseased subjects (421 females and 473 males) and 138 diseased individuals (54 female and 84 male). Further, since it is well-known that the risk of AD increases with age, we included this covariate as well as the sex of the patients in our analyses. Histograms of the nondiseased and diseased groups by sex are presented in Figure 4.27. In these, we can observe similarities between female and male subjects within each group (nondiseased and diseased).

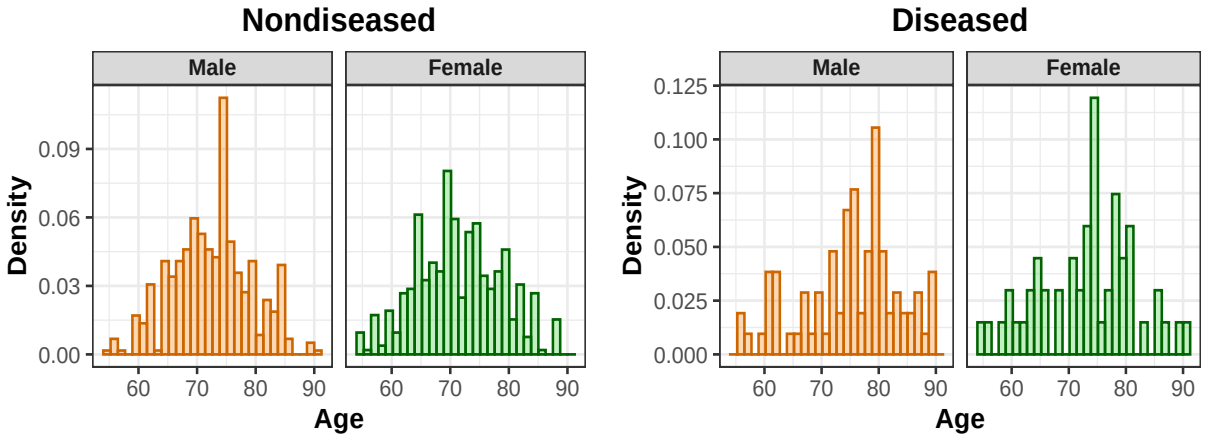

Figure 4.27: Histograms of age and sex of the nondiseased (left) and diseased (right) groups.

We centred and scaled the data to avoid computational issues. We performed 100,000 MCMC iterations after a burn-in period of 50,000. The prior information used was the same described in Section 4.3.2. We assessed convergence of our MCMC algorithm through trace plots (not shown) and the Geweke diagnostic of the conditional densities for three particular ages across a grid of the biomarkers, which are depicted in Figures 4.28–4.31. We have also computed the effective sample size of the conditional density, results are shown in Figures 4.32–4.35.

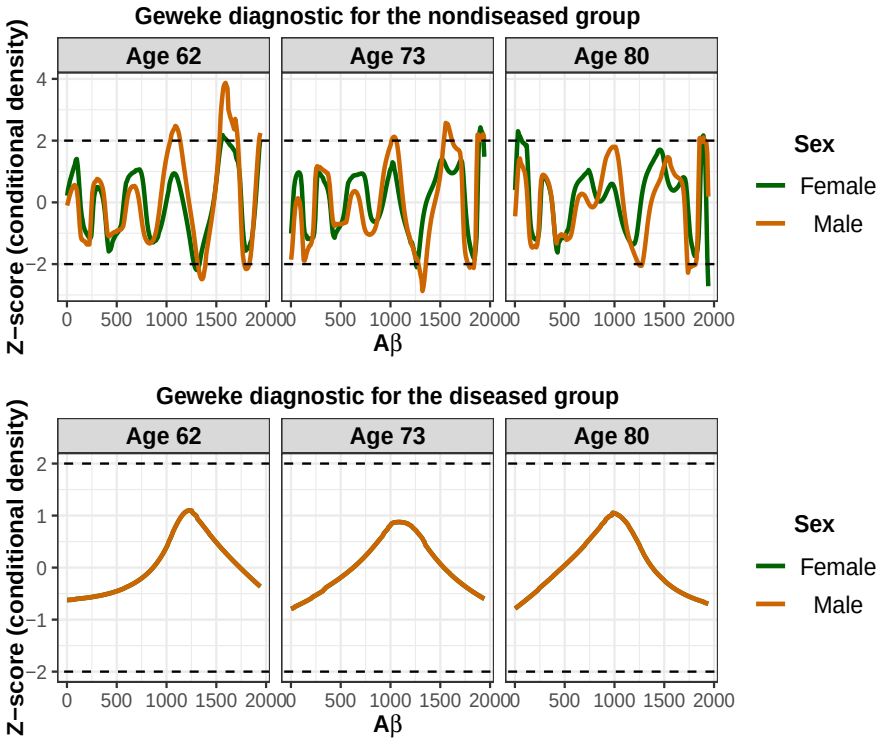

Figure 4.28: Geweke diagnostic for the conditional density across a grid of the Aβ levels for the nondiseased (top) and diseased (bottom) group.

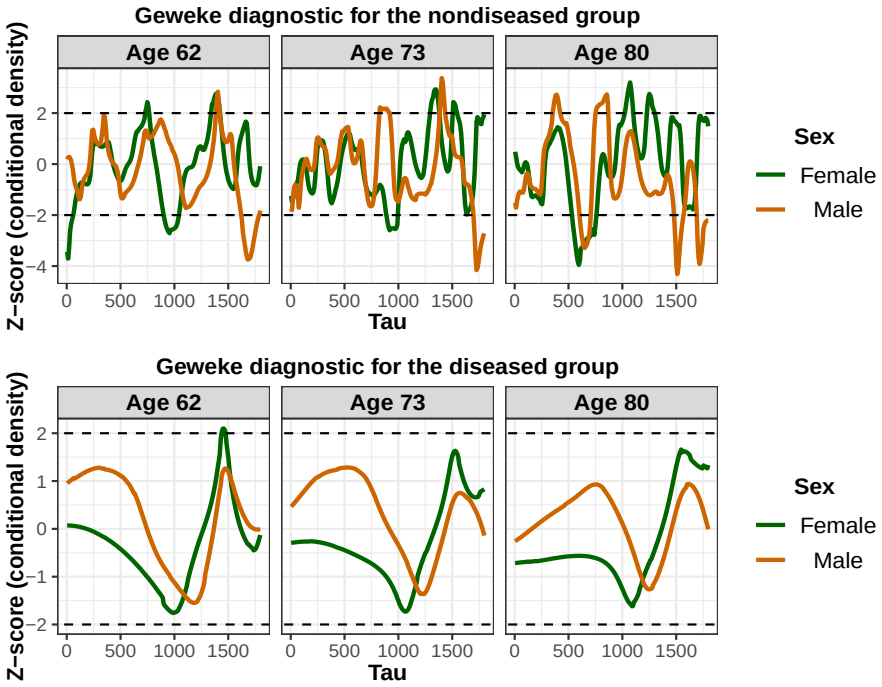

Figure 4.29: Geweke diagnostic for the conditional density across a grid of the Tau levels for the nondiseased (top) and diseased (bottom) group.

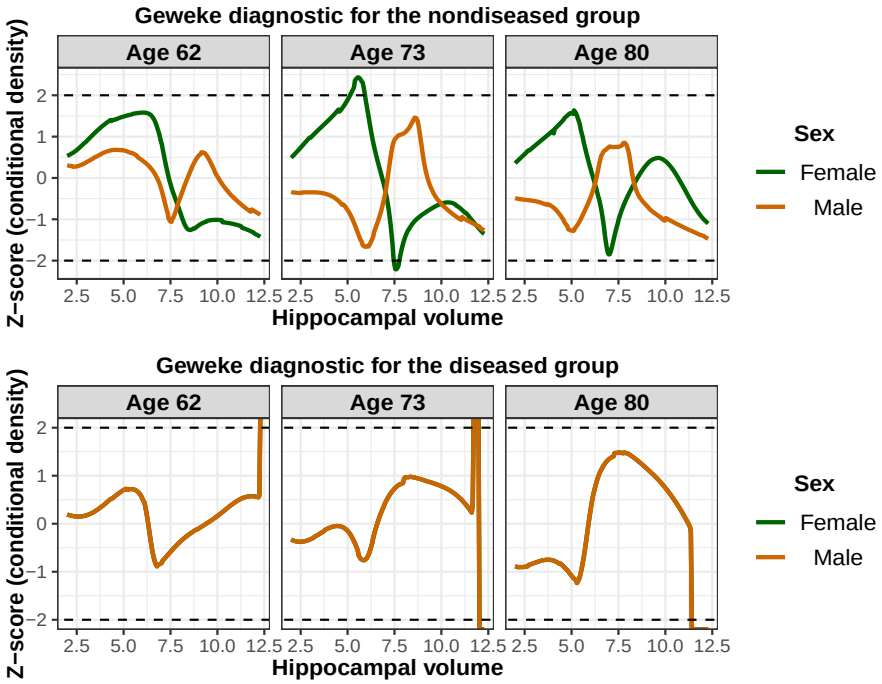

Figure 4.30: Geweke diagnostic for the conditional density across a grid of the hippocampal volume for the nondiseased (top) and diseased (bottom) group.

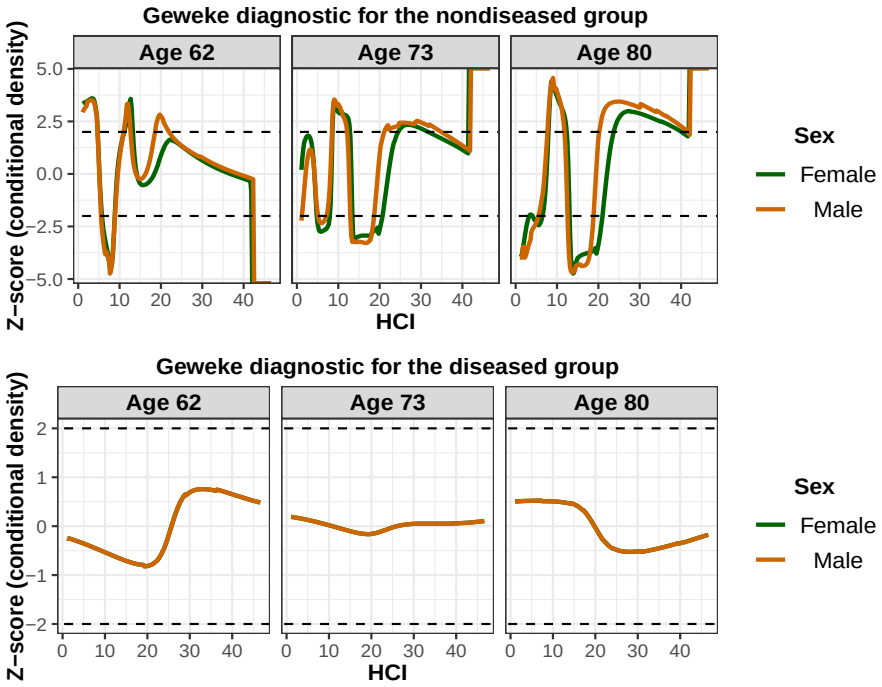

Figure 4.31: Geweke diagnostic for the conditional density across a grid of the HCI for the nondiseased (top) and diseased (bottom) group.

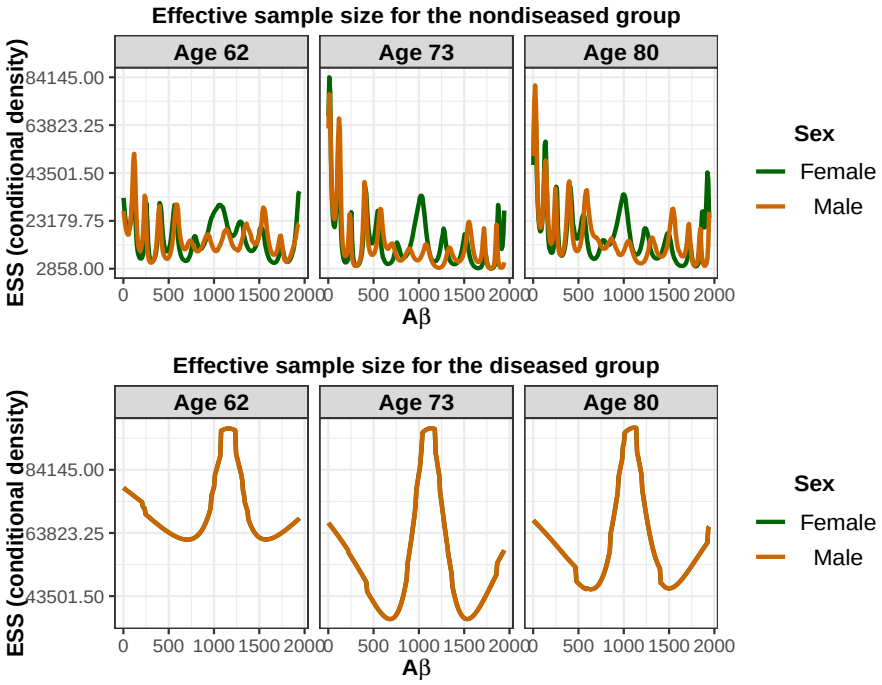

Figure 4.32: Effective simple size for the conditional density across a grid of the Aβ levels for the nondiseased (top) and diseased (bottom) group.

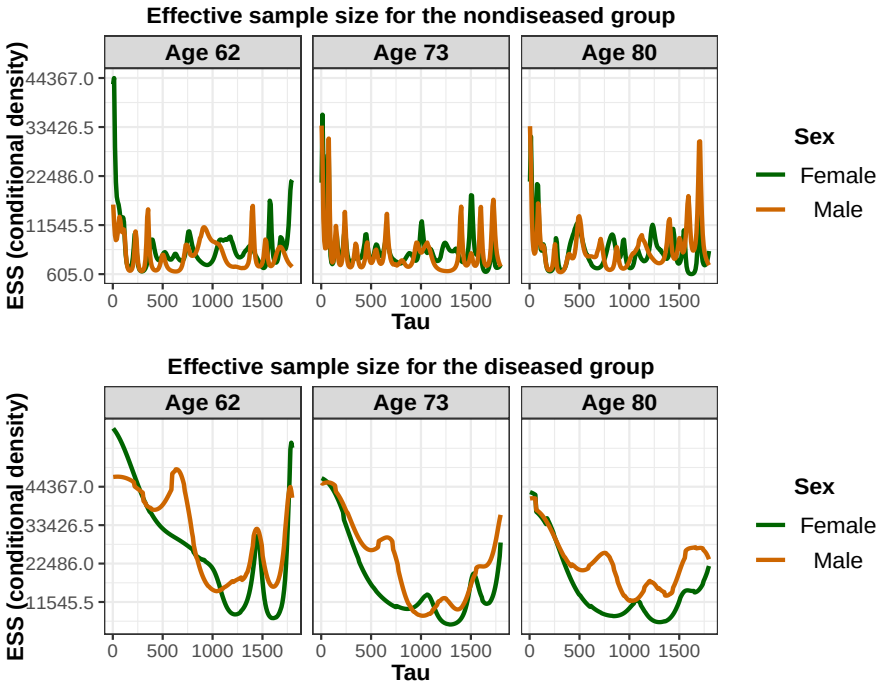

Figure 4.33: Effective simple size for the conditional density across a grid of the Tau levels for the nondiseased (top) and diseased (bottom) group.

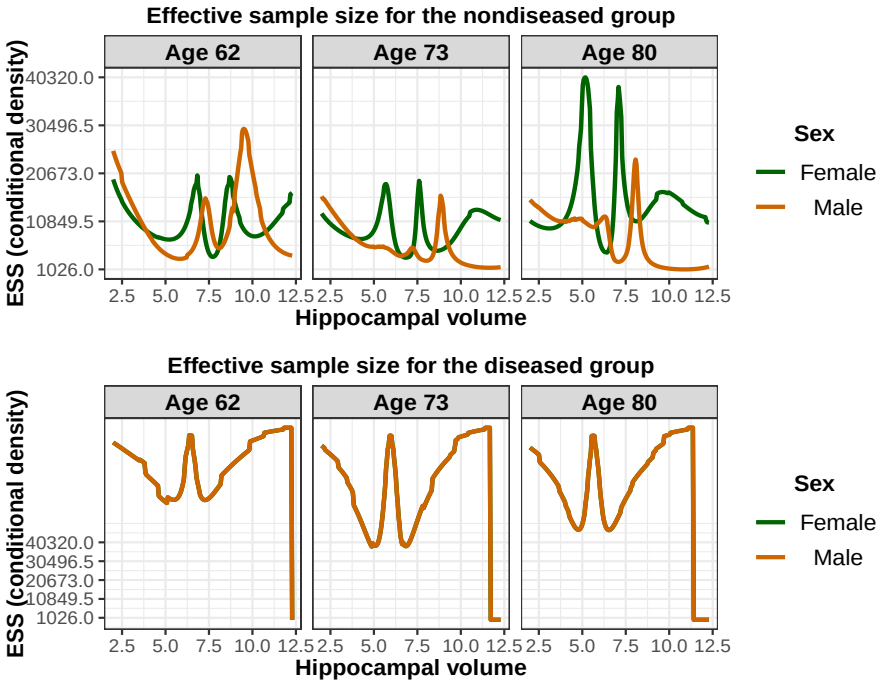

Figure 4.34: Effective simple size for the conditional density across a grid of the hippocampal volume for the nondiseased (top) and diseased (bottom) group.

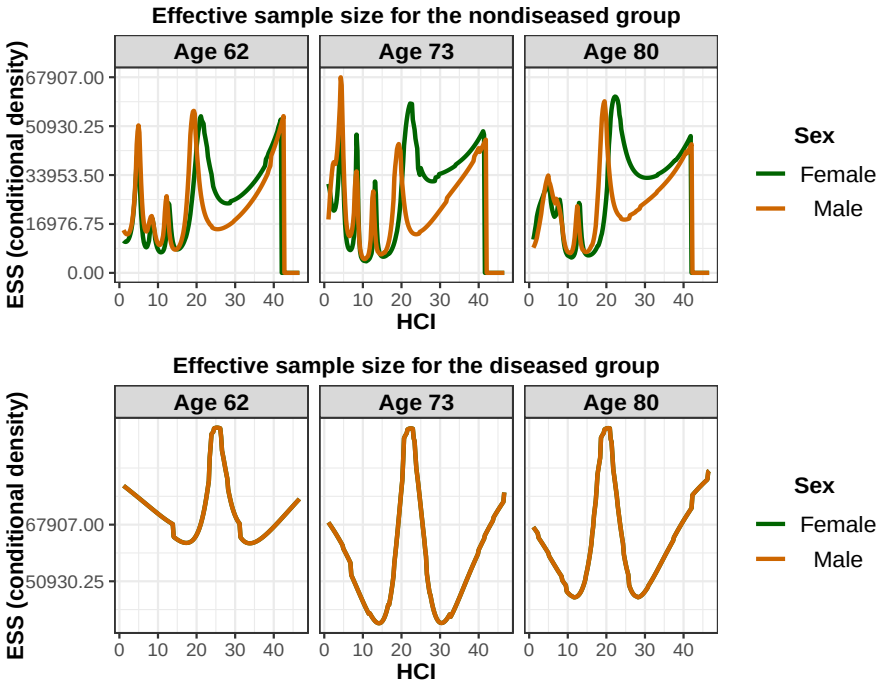

Figure 4.35: Effective simple size for the conditional density across a grid of the HCI for the nondiseased (top) and diseased (bottom) group.

The results for the Aβ levels are displayed in Figure 4.36. We observe that the covariate-specific OVL (top left) increases as age increases and that the curve from female patients is always below the curve from males. This means that the Aβ levels are less accurate for older male individuals. The mean functions, shown in top middle and right, revealed little differences, on average, between female and male individuals. Posterior predictive checks (Figure 4.37) suggest no visible signs of misfit.

Results for Tau are presented in Figure 4.38. In this case, we observe a similar behaviour for the covariate-specific OVL, that is, an increasing trend and lower values for female subjects. However, there is much more uncertainty and discrimination seems indistinguishable between females and males for those over 80 years. The mean functions show differences, on average, between healthy females and males. Figure 4.39 displays the posterior predictive checks, which suggest a good fit. They show that our model recovers other features of data such as kurtosis or skewness very well.

Conversely, for the hippocampal volume, discrimination is better in males (Figure 4.40 top left). The mean functions are almost linear (top middle and right). For such reason, we observed a better fit of the LM (see 4.8). In Figure 4.41, the posterior predictive checks show a some difficulties capturing certain characteristics of the nondiseased group (e.g., kurtosis).

The results for the HCI are depicted in Figure 4.42. Probably, the HCI provides the best discrimination among all the evaluated biomarkers. Similar to previous cases, the HCI suggests a higher accuracy in early years and in female subjects. There are no apparent differences between females and males shown in the mean functions. Again, no visible signs of misfit are shown in the posterior predictive checks (Figure 4.43).

Finally, Table 4.8 shows the results from the model comparison criteria. For almost all cases, our modelling approach is preferable. As noted previously, the LM seems to be a reasonable choice to model hippocampal volume values, which is corroborated by all the criteria in Table 4.8.

In conclusion, we found that all the analysed biomarkers are less accurate as age increases. The majority of them seems to discriminate better female patients. The results suggest that HCI best discriminates CN and MCI from AD subjects. In contrast, both tau levels and hippocampal volume showed higher OVL values, thus making them less reliable biomarkers to differentiate between these

categories.

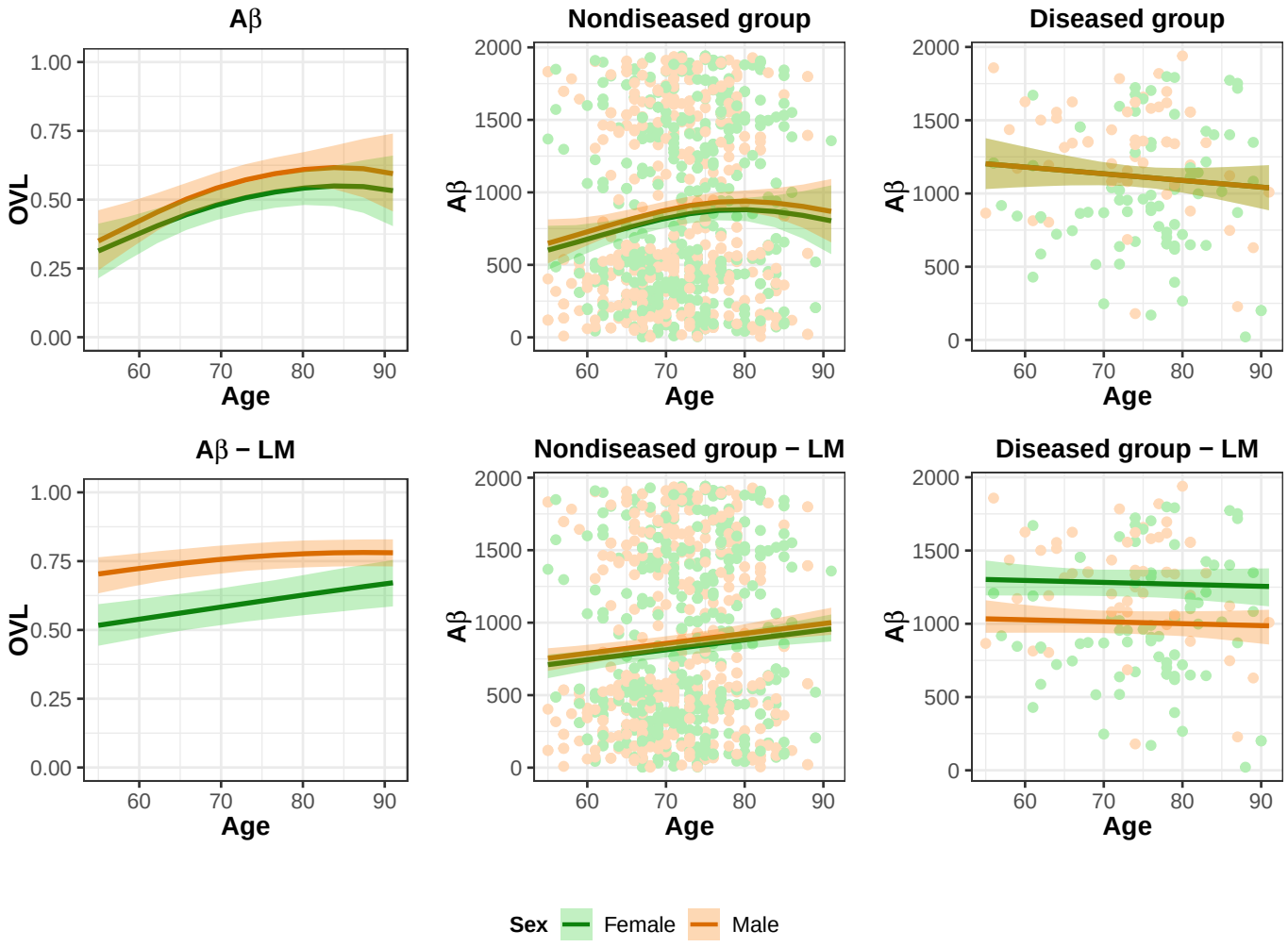

Figure 4.36: Results for Aβ under the DPM model (top row) and LM (bottom row). Posterior mean (solid line) and 95% pointwise credible bands (shaded area) of the covariate-specific OVL (left) and the mean functions of the nondiseased (middle) and diseased (right) group.

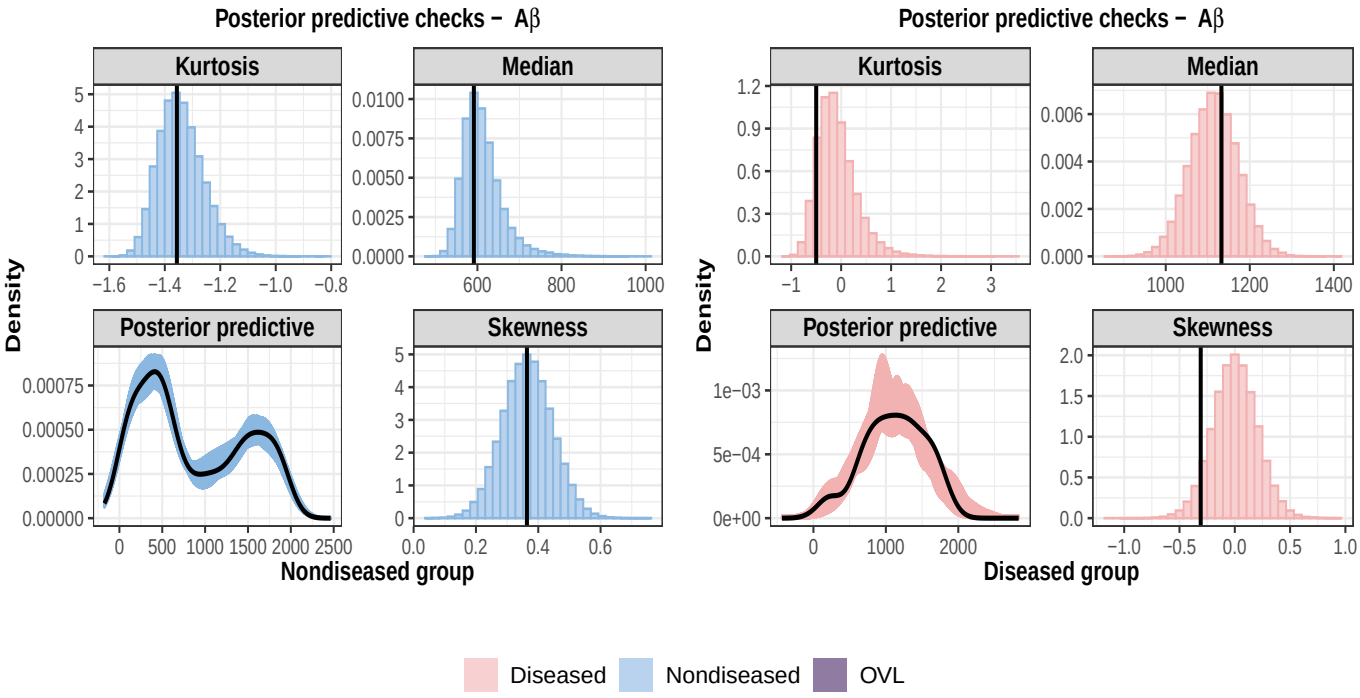

Figure 4.37: Results for Aβ. Posterior predictive checks for the nondiseased (left) and diseased (right) group.

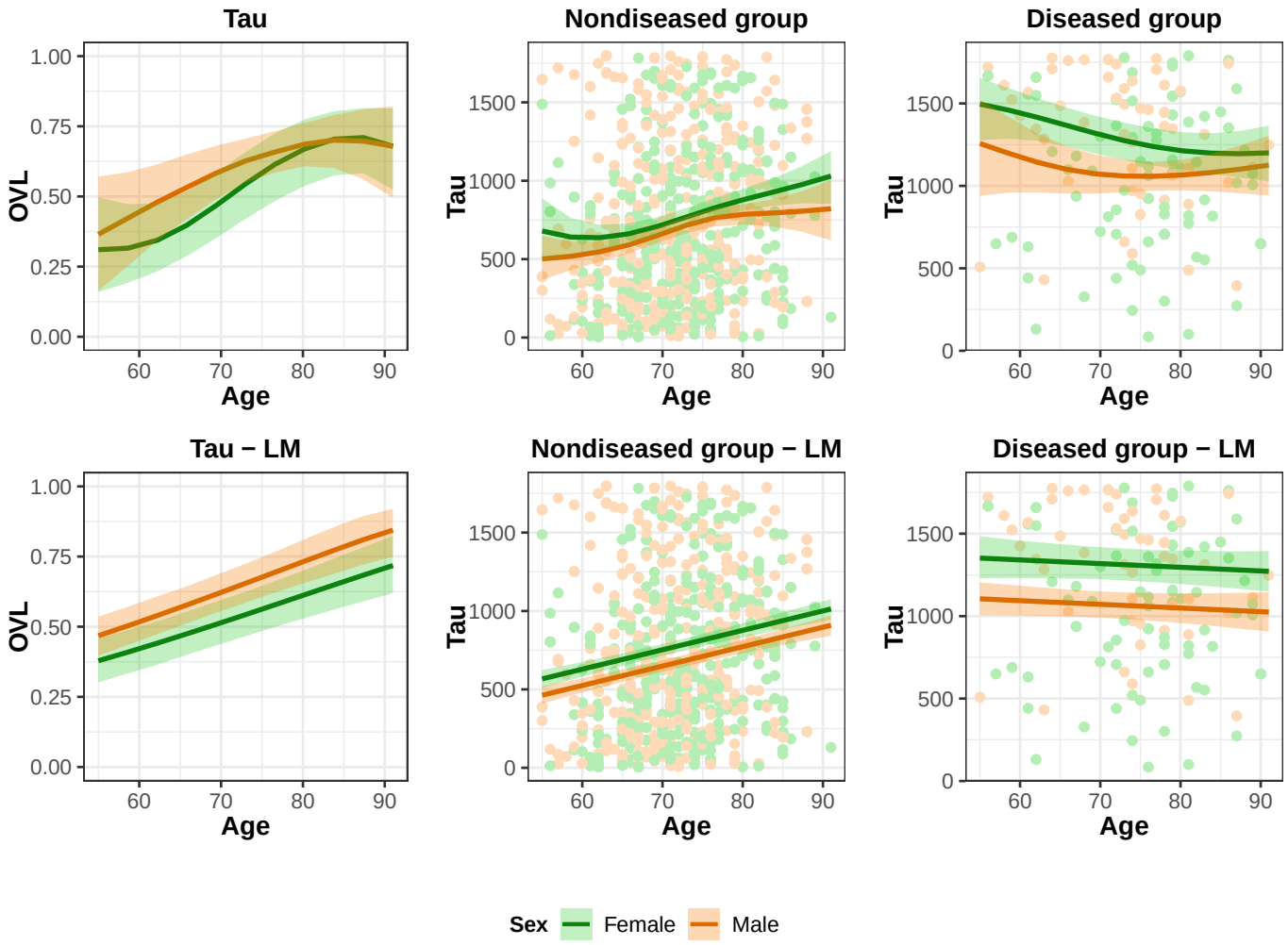

Figure 4.38: Results for Tau under the DPM model (top row) and LM (bottom row). Posterior mean (solid line) and 95% pointwise credible bands (shaded area) of the covariate-specific OVL (left) and the mean functions of the nondiseased (middle) and diseased (right) group.

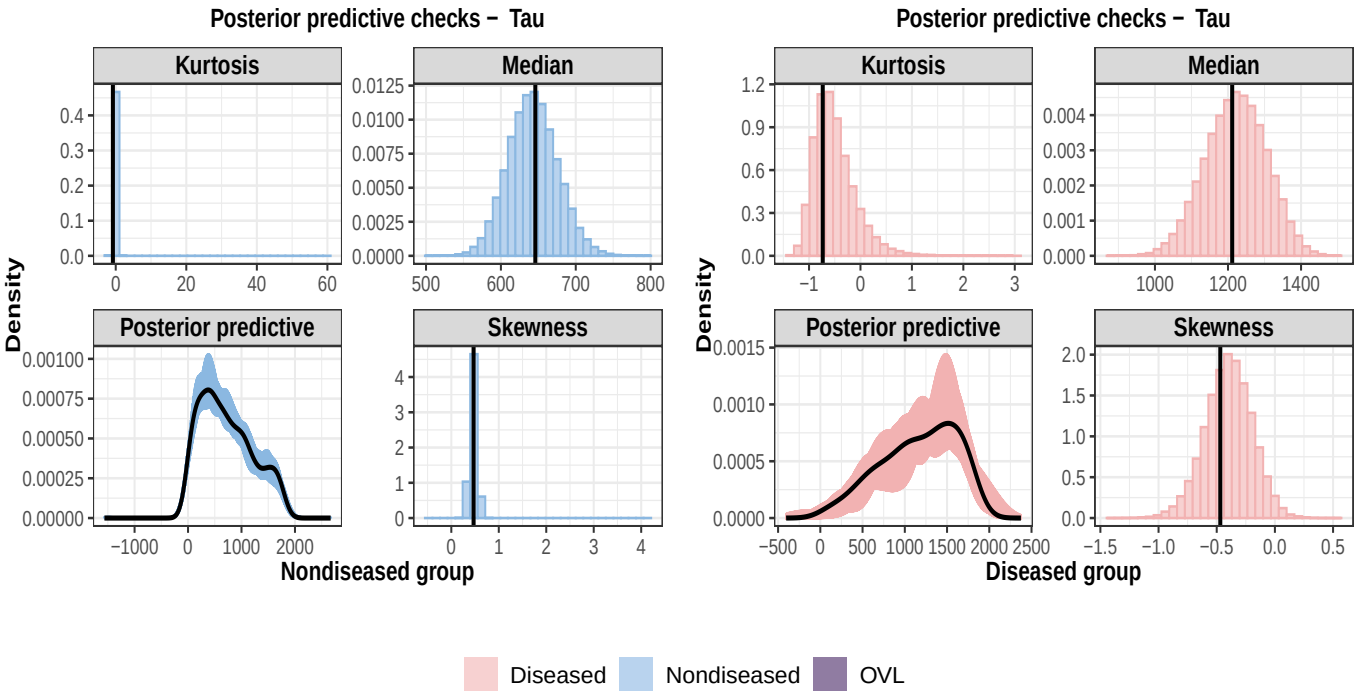

Figure 4.39: Results for Tau. Posterior predictive checks for the nondiseased (left) and diseased (right) group.

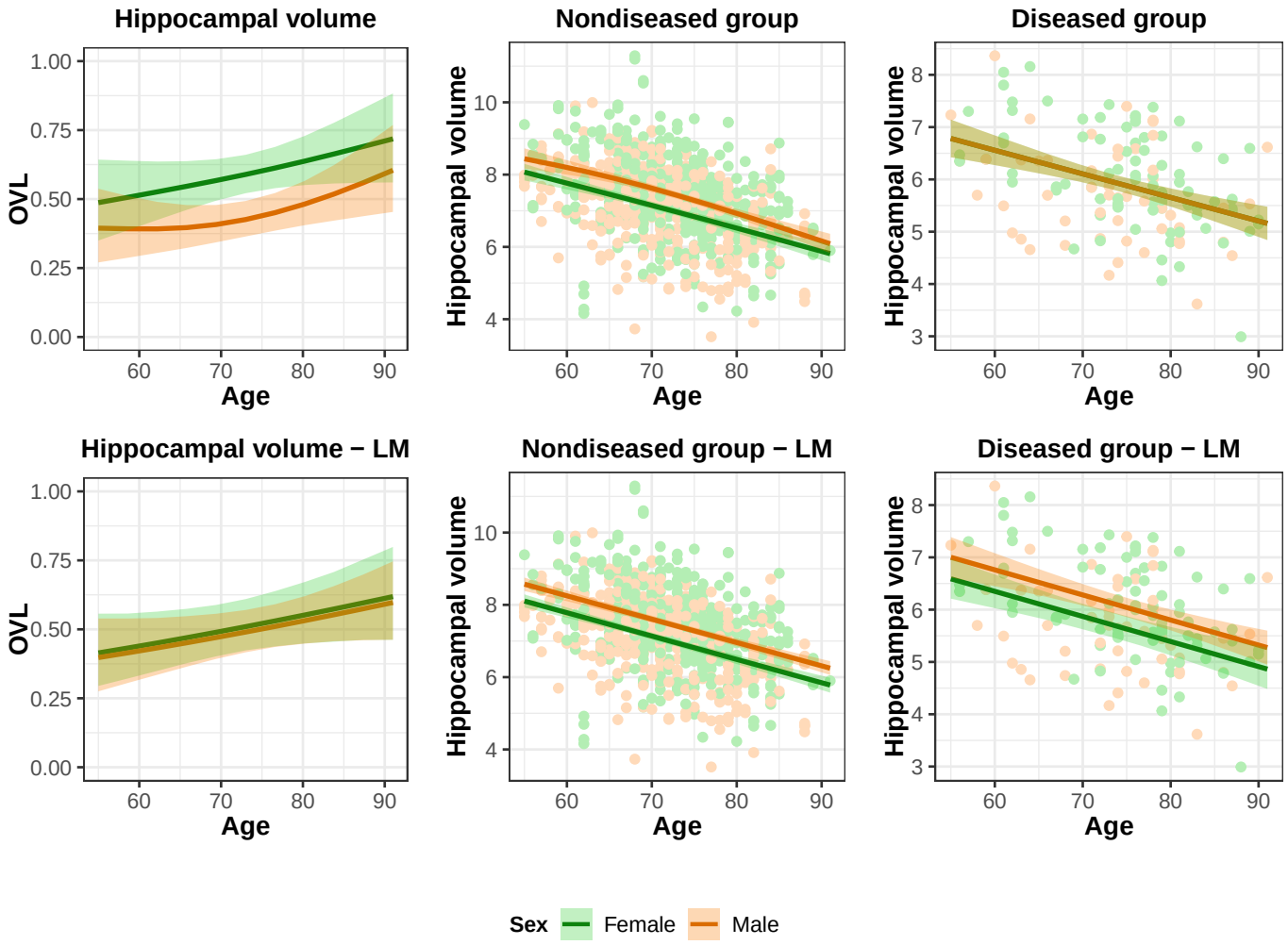

Figure 4.40: Results for the hippocampal volume under the DPM model (top row) and LM (bottom row). Posterior mean (solid line) and 95% pointwise credible bands (shaded area) of the covariate-specific OVL (left) and the mean functions of the nondiseased (middle) and diseased (right) group.

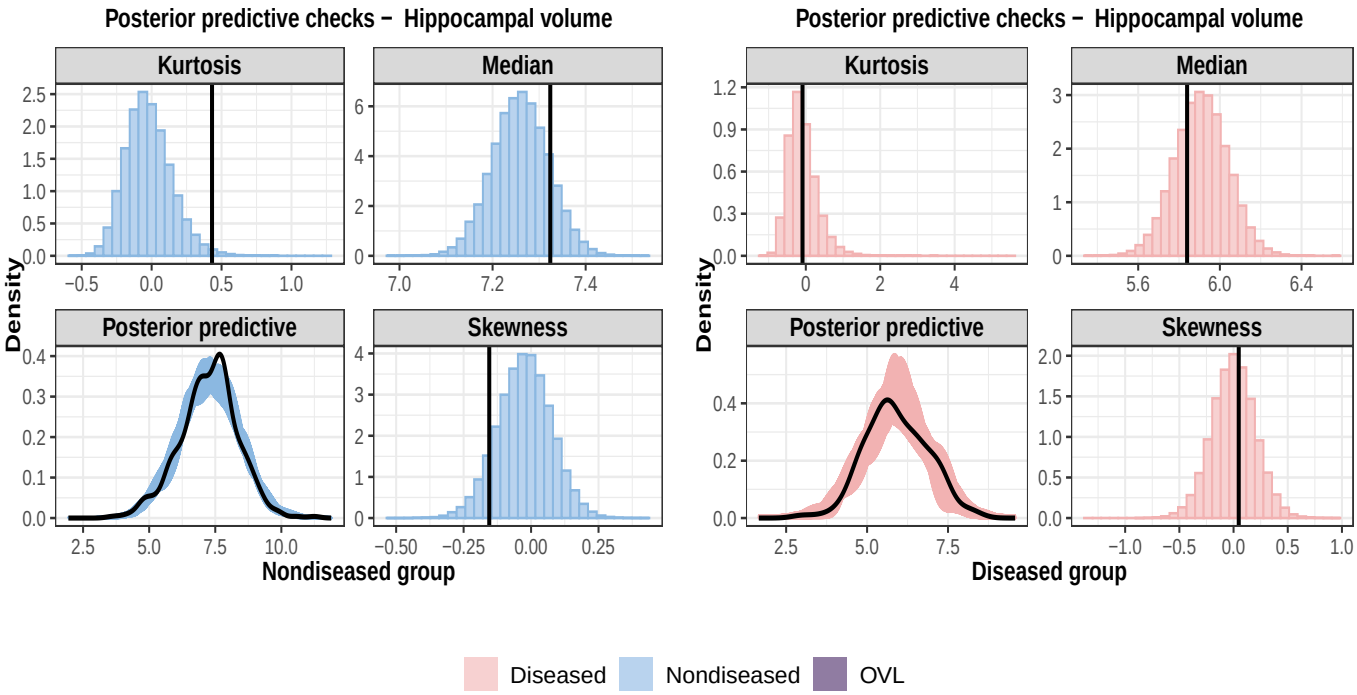

Figure 4.41: Results for the hippocampal volume. Posterior predictive checks for the nondiseased (left) and diseased (right) group.

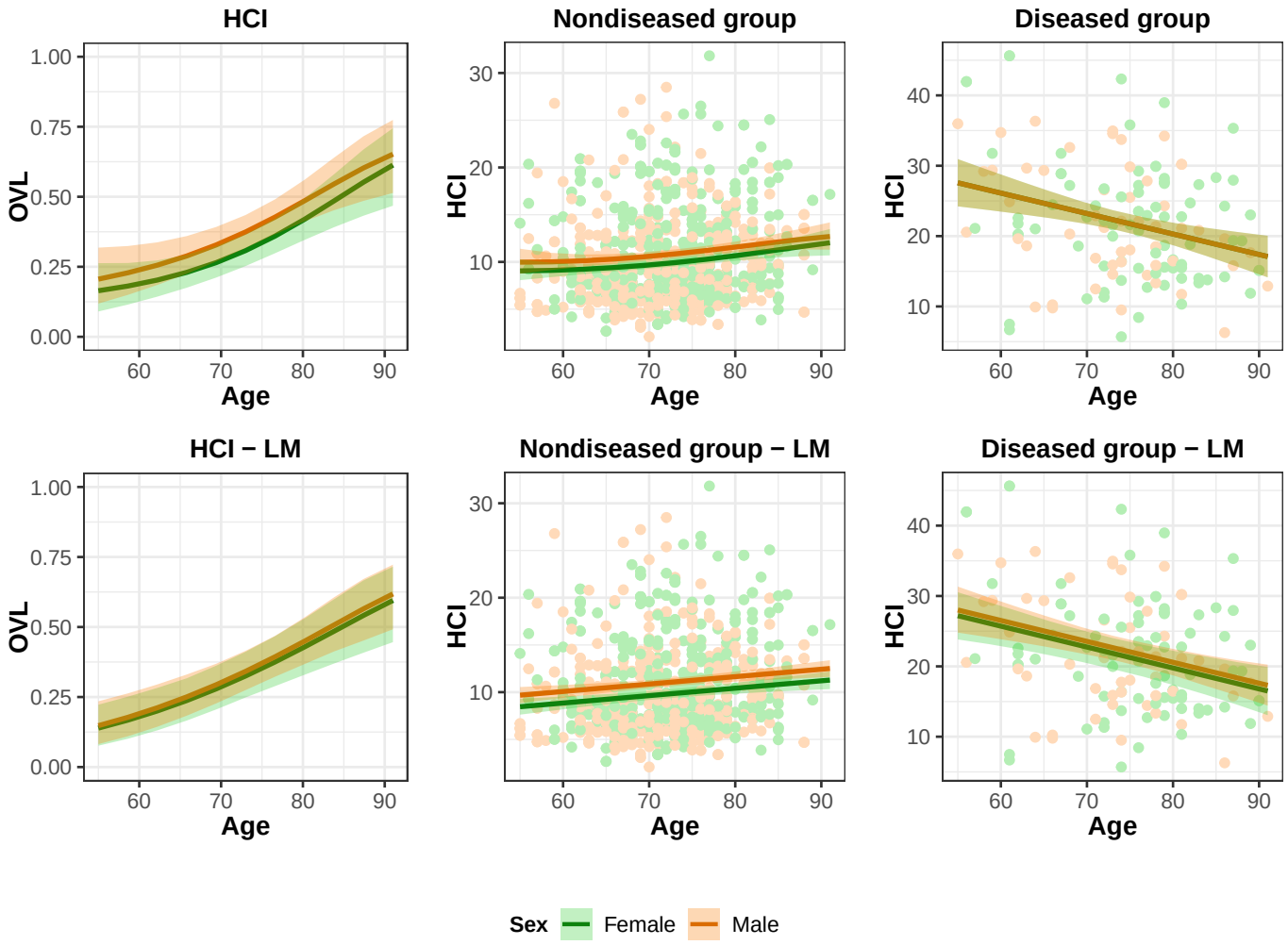

Figure 4.42: Results for the HCI under the DPM model (top row) and LM (bottom row). Posterior mean (solid line) and 95% pointwise credible bands (shaded area) of the covariate-specific OVL (left) and the mean functions of the nondiseased (middle) and diseased (right) group.

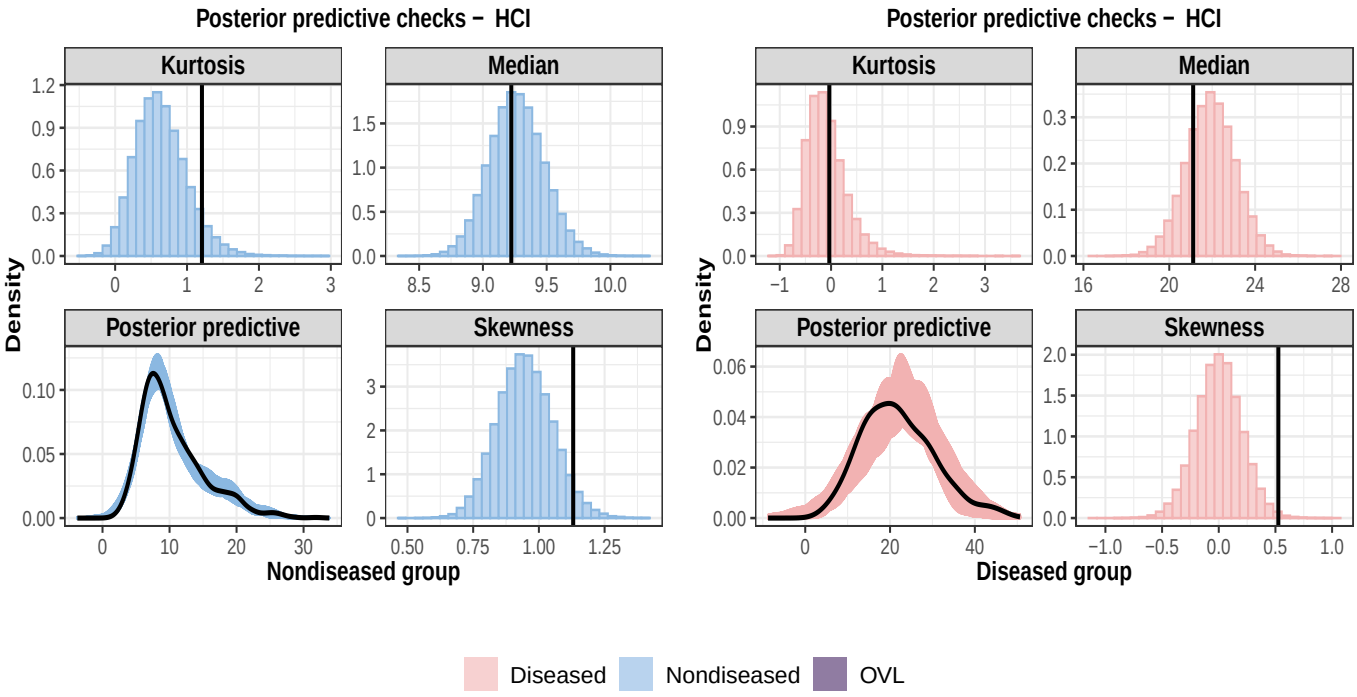

Figure 4.43: Results for the HCI. Posterior predictive checks for the nondiseased (left) and diseased (right) group.

Table 4.8: Model comparison criteria. Alzheimer's disease application

Marker	Group	Criterion	DPM	LM
Aβ	Healthy	Adjusted LPML	**-6688.92**	-7007.45
		WAIC	**13377.82**	14014.9
		DIC_3	**13377.16**	14014.88
		Posterior rank probability	**1.00**	0.00
	Diseased	Adjusted LPML	-1032.93	**-1026.39**
		WAIC	2065.86	**2052.77**
		DIC_3	2065.71	**2052.62**
		Posterior rank probability	0.0395	**0.9605**
Tau	Healthy	Adjusted LPML	**-6626.31**	-6780.78
		WAIC	**13252.29**	13561.57
		DIC_3	**13249.95**	13561.55
		Posterior rank probability	**1.00**	0.00
	Diseased	Adjusted LPML	**-1023.81**	-1033.1
		WAIC	**2047.56**	2066.19
		DIC_3	**2046.84**	2066.08
		Posterior rank probability	**0.9891**	0.0109
Hippocampal volume	Healthy	Adjusted LPML	-1259.62	**-1259.06**
		WAIC	2519.27	**2518.13**
		DIC_3	2518.89	**2518.08**
		Posterior rank probability	0.4113	**0.5887**
	Diseased	Adjusted LPML	-179.79	**-177.28**
		WAIC	359.58	**354.55**
		DIC_3	359.42	**354.3**
		Posterior rank probability	0.1857	**0.8143**

Continuation of Table 4.8

Marker	Group	Criterion	DPM	LM
HCI	Healthy	Adjusted LPML	**-2531.16**	-2644.76
		WAIC	**5062.31**	5289.52
		DIC_3	**5062.11**	5289.45
		Posterior rank probability	**1.0**	0.0
	Diseased	Adjusted LPML	**-486.66**	-487.21
		WAIC	**973.32**	974.41
		DIC_3	**973.11**	974.17
		Posterior rank probability	**0.7982**	0.2018

4.5 Discussion

We presented a flexible and robust approach to conduct inference about the covariate-specific coefficient of overlap. We modelled the joint distribution of the test outcomes and the covariates by using a DPM with a normal/Bernoulli kernel, thus allowing continuous and binary covariates. Further, by modelling jointly the test outcomes and the covariates, we do not need to specify any particular functional forms of how the covariates affect the test outcomes, thus any non-linearity or heterogeneity is automatically accommodated. Other functions of interest are available as well, such as the conditional mean, quantile and variance function. Also, we detailed a Gibbs scheme to simulate from the posterior, making inference straightforward.

The simulation study revealed that our estimator is able to recover the true covariate-specific OVL in a number of conceivable situations as the sample size increases. Specially, for more complex situations, such as Scenario III, where a heteroscedastic situation with an interaction between a binary and a continuous covariate was considered, it requires a larger sample size to accurately estimate the covariate-specific OVL and other functions of interest. These results suggest that our proposed estimator may be consistent. Further, our modelling approach captures well other features of data, such as the conditional mean and variance function. Empirical coverage probability of the 95% credible

intervals was investigated as well. While it is true that as sample size increases, so does the coverage, in complex scenarios, the coverage of our estimator remains below the nominal value. This is largely due to the boundaries, where fewer observations are available. The model comparison criteria results suggest that, even in linear scenarios, our modelling approach with relatively small sample sizes, may provide a good fit. When the test outcomes arise from different non Gaussian parametric families or present complex structures, the DPM clearly outperforms the LM, meaning that more reliable conclusions about a test's accuracy can be drawn.

We applied our methods to two real datasets. The first application was intended to investigate if and how the accuracy of glucose levels as a marker for diabetes changes with age. We observed that there seems to be evidence that glucose levels are less accurate as age increases. Our modelling approach is preferable for the non diabetic group. Whereas, the LM provides a better fit for the diabetic group. However, the results from our model for the diabetic group were consistent with those from the LM albeit higher uncertainty (broader credible bands). Therefore, we can state that conclusions drawn for the covariate-specific OVL under our modelling framework are still reliable. The second application concerned the search for biomarkers for Alzheimer's disease and how these may be affected by the sex and age of the patients. We found that age is clearly a key variable and the accuracy in diagnosis decreases as age increases. There seems to be evidence that the majority of the analysed biomarkers perform better in female patients. However, this difference is slight. The results suggest that HCI best discriminates individuals with normal cognition and mild impairment from AD subjects.

Chapter 5

OverlapCoefficient: An R package for Bayesian nonparametric inference about the coefficient of overlap

In this chapter we include vignettes and examples of the `OverlapCoefficient` package, an R package in which we have implemented all the methods described in the previous chapters.

The `OverlapCoefficient` R package (`https://github.com/javier-gg/OverlapCoefficient`) implements the Bayesian nonparametric modelling frameworks described in Chapter 3 and 4 to estimate the coefficient of overlap and its covariate-specific counterpart. The main function is called `ovl`, which internally fits a Dirichlet process mixture model to each of the groups (nondiseased and diseased) and computes the (conditional) density function on a given grid to obtain the (covariate-specific) OVL using a numerical integration method (the default method is the trapezoidal rule). The Bayesian bootstrap-based estimator for the OVL described in Section 3.2.3, can be computed using the function `ovlBB`.

For the sake of comparison, we have included kernel based estimators of the OVL as well as implementations of the popular binormal model, respectively, in the `ovlfreq`, `ovlprob` and `ovlNorm` functions, respectively. Also, we include different utility functions to, for example, estimate the (con-

ditional) density functions and the conditional mean/variance functions for the fitted models.

The vignettes demonstrate plenty examples of how to use each of the functions. Further examples can be found in the Github repository at `https://github.com/javier-gg/OverlapCoefficient`. On the date this thesis has been submitted, this repository is still private for security reasons. Nevertheless, we aim to make both the repository and the package publicly available once the article based on Chapter 4 is submitted.

A website for the package has also been created using the `pkgdown` package. This will be available once the repository is public as well. The site contains the help files as shown in the vignettes of the package as well as additional worked examples. This has been intended to maintain the package for a future submission to the CRAN.

Package 'OverlapCoefficient'

August 11, 2021

Type Package

Title Bayesian nonparametric inference for the coefficient of overlap

Date 2020-10-20

Version 0.0.2

Author Javier E. Garrido Guillén, Vanda Inácio

Maintainer Javier E. Garrido Guillén <javier.garridog@ed.ac.uk>

Description Implements Bayesian nonparametric methods for estimating the coefficient of overlap and its covariate-specific version.

License GPL-3 + file LICENSE

Encoding UTF-8

LazyData true

Imports Rcpp (>= 1.0.5), MASS, mvnfast, CholWishart, pbapply, statmod, ggplot2

LinkingTo Rcpp, RcppArmadillo

URL https://github.com/javier-gg/OverlapCoefficient

BugReports https://github.com/javier-gg/OverlapCoefficient/issues

Suggests knitr, rmarkdown

VignetteBuilder knitr

Archs x64

R topics documented:

`blm` *Bayesian linear model*

Description

Fits a Gaussian linear regression model under the Bayesian framework.

Usage

```
blm(formula, data, prior, mcmc, silent = FALSE)
```

Arguments

`formula`	An object of class `formula`, similar to the one passed to `lm`.
`data`	An optional data frame containing the data. If data is missing, then the variables are taken from the environment.
`prior`	An optional named list specifying the prior information. See "Details".
`mcmc`	An optional named list containing the values for the MCMC algorithm. See "Details".
`silent`	An optional logical value specifying whether a progress bar should be shown or not. The default value is `FALSE`, meaning that a progress bar is shown.

Details

The function implements the Bayesian counterpart of the `lm` function. For the moment, explicit arguments are needed instead of a formula, i.e., data from the response variable must be passed to the `y` argument and a suitable design matrix must be passed to the `X` argument. The hierarchical representation of the model is as follows

$$y \mid \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2 \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2),$$

$$\boldsymbol{\beta} \mid \boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0 \sim \mathrm{N}(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0),$$

$$\sigma^2 \mid a_{\sigma^2}, b_{\sigma^2} \sim \Gamma^{-1}(a_{\sigma^2}, b_{\sigma^2}),$$

where $\Gamma^{-1}(a, b)$ stands for an inverse-gamma distribution with shape a and scale b. Note that $\boldsymbol{\Sigma}_0$ represents the covariance matrix. Hence, the hyperparameters $\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0, a_{\sigma^2}, b_{\sigma^2}$ must be included in

a named list object and passed to the `prior` argument as `list(beta_0,Sigma_0,a_sigma2,b_sigma2)`. If `prior` is missing, then relatively vague hyperparameter values are used, specifically `list(beta_0 = rep(0,p),Sigma_0 = 100*diag(p),a_sigma2 = 0.001,b_sigma2 = 0.001)`, where p is the number of parameters.

The argument `mcmc` specify the number of saved iterations, burn-in period and thinning of the chain. It must be a named list containing these values. Default values are `list(nsave = 3000,nburn = 1000,nthin = 1)`.

Value

An list of class `blm` containing:

`y`	The original response data.
`X`	The original design matrix.
`prior`	The prior information used.
`beta`	A numeric matrix containing the posterior sample of the regression coefficients.
`sigma2`	A numeric vector containing a posterior sample of the variance.

Author(s)

Javier E. Garrido Guillén

References

Hoff, P. D. (2009). A first course in Bayesian statistical methods (Vol. 580). New York: Springer.

See Also

`lm`, `summary.blm`, `meanblm`

Examples

```
## Not run:
##############################
## Example using mtcars data ##
##############################
## Load data
data(mtcars)

## Fit a linear regression model
## using all default settings
fitlm<-blm(mpg ~ wt + cyl, data = mtcars)

## Posterior summary
summary(fitlm)


##########################################
## Another example, using simulated data ##
##########################################
## Set a seed for reproducibility
set.seed(123)

## Simulate 500 observations from
```

```
## a mixture of normals, where the
## weights depend on a continuous
## covariate (example taken from
## Dunson et al., 2007)
n<-500
x<-runif(n)
y1<-x + rnorm(n, 0, sqrt(0.01))
y2<-x^4 + rnorm(n, 0, sqrt(0.04))
u<-runif(n)
prob<-exp(-2*x)
y<-ifelse(u < prob, y1, y2)

## Fit a linear regression model,
## changing prior and number of
## MCMC iterations
myprior<-list(beta_0 = c(0, 0), Sigma_0 = 100*diag(2),
a_sigma2 = 0.5, b_sigma2 = 0.5)
mymcmc<-list(nsave = 10000, nburn = 5000, nthin = 1)
fitlm<-blm(y ~ x, prior = myprior, mcmc = mymcmc)

## Posterior summary
summary(fitlm)

## Visualize conditional mean function
## Create a data frame with covariate values
newdata<-data.frame(x = seq(0, 1, len = 101))
meanfunc<-meanblm(fitlm, x = newdata)

## Load ggplot2
require(ggplot2)

## Plot the (posterior mean)
## mean function and the data
plot(meanfunc) +
geom_point(data = data.frame(x = x, y = y),
mapping = aes(x = x, y = y),
alpha = 0.5, color = "slategray2")

## End(Not run)
```

`criteria` *Model comparison criteria*

Description

Computes the LPML, WAIC and DIC for a given model.

Usage

```
criteria(fit, ncores = 1, silent = FALSE)

## S3 method for class 'blm'
criteria(fit, ncores = 1, silent = FALSE)
```

```
## S3 method for class 'dpm'
criteria(fit, ncores = 1, silent = FALSE)
```

Arguments

`fit`	A resulting object from a call to `dpm` or `blm`.
`ncores`	An optional integer specifying the number of cores that should be used. The default value is 1.
`silent`	An optional logical argument specifying whether a progress bar should be printed or not. The default value is `FALSE`, then a progress bar is printed.

Details

This function computes the log–pseudomarginal likelihood (LPML) based on the result showed by Gelfand & Dey (1994).

$$\text{LPML} = \sum_{i=1}^{n} \log(\text{CPO}_i),$$

where $\text{CPO}_i^{-1} = \frac{1}{S}\sum_{s=1}^{S} \frac{1}{f(y_i|\boldsymbol{\theta}^{(s)})}$. Here, S represents the number of MCMC iterations.

A weighted alternative suggested by Gelman et al. (2014) is also implemented.

$$\text{CPO}_i^{-1} = \frac{\sum_{s=1}^{S} w_{is} f(y_i \mid \boldsymbol{\theta}^{(s)})}{\sum_{s=1}^{S} w_{is}}$$

where $w_{i,s} = \min\left\{\frac{1}{f(y_i|\boldsymbol{\theta}^{(s)})}, \sqrt{S}\frac{1}{S}\sum_{s=1}^{S}\frac{1}{f(y_i|\boldsymbol{\theta}^{(s)})}\right\}$.

Addionally, the function computes the Watanabe-Akaike or widely applicable information criterion (WAIC) using the posterior variance to correct the bias.

$$\text{WAIC} = -2\text{elpd},$$

where elpd stands for the expected log–predictive density.

And a modification of the popular deviance information criterion (Spiegelhalter et al., 2002) proposed by Celeux et al. (2006).

$$\text{DIC}_3 = -4\frac{1}{S}\sum_{s=1}^{S}\sum_{i=1}^{n}\log[f(y_i \mid \boldsymbol{\theta}^{(s)})] + 2\sum_{i=1}^{n}\log\left[\frac{1}{S}\sum_{s=1}^{S} f(y_i \mid \boldsymbol{\theta}^{(s)})\right].$$

Value

A list containing the following:

`values`	A numeric vector containing the values of the four criteria.
`CPO`	A numeric vector containing the corresponding CPOs.

Author(s)

Javier E. Garrido Guillén

References

Celeux G., Forbes F., Robert C.P. & Titterington D.M. (2006), "Deviance Information Criteria for Missing Data Models", Bayesian analysis, **1**(4), 651–673.
Gelfand, A. E. & Dey, D. K. (1994), "Bayesian model choice: asymptotics and exact calculations", Journal of the Royal Statistical Society, Series B **56**(3), 501–514.
Gelman, A., Hwang, J. & Vehtari, A. (2014), "Understanding predictive information criteria for Bayesian models", Statistics and Computing **24**(6), 997–1016.

See Also

ddpm, dblm

Examples

```
## Not run:
############################
## Compare a DPM and a LM ##
############################
## Set a seed for reproducibility
set.seed(123)

## Simulate 500 observations from
## a mixture of normals, where the
## weights depend on a continuous
## covariate (example taken from
## Dunson et al., 2007)
n<-500
x<-runif(n)
y1<-x + rnorm(n, 0, sqrt(0.01))
y2<-x^4 + rnorm(n, 0, sqrt(0.04))
u<-runif(n)
prob<-exp(-2*x)
y<-ifelse(u < prob, y1, y2)

## Fit a Dirichlet process mixture model
fitdpm<-dpm(y ~ x,
mcmc = list(nsave = 10000, nburn = 5000, nthin = 1))

## Fit a linear regression model
fitlm<-blm(y ~ x,
mcmc = list(nsave = 10000, nburn = 5000, nthin = 1))

## Compute model comparison criteria
criteriadpm<-criteria(fitdpm)
criterialm<-criteria(fitlm)

## Show results
res<-cbind(criteriadpm$values, criterialm$values)
colnames(res)<-c("DPM", "LM")
res

## End(Not run)
```

dbern *Product of independent Bernoulli distributions*

Description

Density function of a product of independent Bernoulli distributions with probability of success prob.

Usage

```
dbern(x, prob)
```

Arguments

x	A numeric matrix of quantiles. See "Details".
prob	A numeric vector of probabilities of dimension equal to the number of columns in x.

Details

This function computes the density of a product of q independent Bernoulli distributions, that is

$$f(\boldsymbol{x}) = \Pi_{k=1}^{q} \pi_k^{x_k} \left(1 - \pi_k\right)^{1-x_k},$$

where $\boldsymbol{x} = (x_1, \ldots, x_q)$.

Value

A numeric vector containing the density.

Author(s)

Javier E. Garrido Guillén

Examples

```
## Not run:
#####################
## Univariate case ##
#####################
## Observed matrix
x<-matrix(c(0, 1, 0, 0), ncol = 1)

## Probability of success
prob<-0.3

## Density function
dbern(x, prob)


######################
## Multivariate case ##
```

```
######################
## Observed matrix
x<-matrix(rbinom(15, 1, 0.5), ncol = 3, nrow = 5)

## Probabilities of success
prob<-c(0.3, 0.1, 0.6)

## Density function
dbern(x, prob)

## End(Not run)
```

---

dblm *Linear model conditional density*

Description

Conditional density of a Gaussian linear regression model.

Usage

```
dblm(fit, grid, x = NULL, silent = FALSE)
```

Arguments

fit	A resulting object from a call to blm.
grid	An optional numeric vector of quantiles of the response variable. If missing, a uniform grid from min(y) -1 to max(y) + 1 of length 201 is used.
x	An optional data frame containing the values of the covariates in fit used for computing the conditional density.
silent	An optional logical argument specifying whether a progress bar should be printed or not. The default value is FALSE, then a progress bar is printed.

Details

The conditional density is given by

$$f(y \mid \boldsymbol{x}) = \phi(y \mid \boldsymbol{x}\boldsymbol{\beta}, \sigma^2),$$

where $\phi(\cdot \mid \mu, \sigma^2)$ denotes the density function of a normal distribution with mean μ and variance σ^2.

y must be a numeric $(N \times p)$ matrix, where N is the total number of evaluation points of the conditional density function and p the number of covariates (including the intercept). By default, the first column will be assumed to be the response variable. The rest of the columns will be assumed to be the covariates in the same order as in fit.

Value

A list containing:

grid	The grid used to compute the density.
density	A numeric matrix containing the posterior density sample.

Author(s)

Javier E. Garrido Guillén.

See Also

blm

Examples

```
## Not run:
########################
## Simulated example ##
########################
## True mean function
mu<-function(xc, xd) (0.7*sin(xc)*cos(6*xc) + 0.4)*
(1 - xd) + 0.3*exp(-2*xc + 1)*xd

## True standard deviation function
sigma<-function(xc) 0.6*(1 - xc)^3 - 0.2*xc^2 + 0.5*xc

## Simulate data
## Set seed for reproducibility,
## simulate 200 observations of
## a normal distribution where
## mean and variance depend on
## continuous/binary covariate
set.seed(123)
n<-200
xc<-runif(n)
xd<-rbinom(n, 1, 0.5)
y<-rnorm(n, mu(xc, xd), sigma(xc))

## Create data frame with data
## NOTE: binary covariates must
## be defined as factors
mydata<-data.frame(y = y,
xc = xc,
xd = factor(xd))

## Fit a linear regression model
fitlm<-blm(y ~ xc + xd, mydata)

## For illustration purposes, we
## choose three random values of
## the continuous covariate xc
xcpred<-runif(3)
xdpred<-factor(c(0, 1))

## New data for conditional densities
newdata<-expand.grid(xc = xcpred, xd = xdpred)

## Compute the density function
dens<-dblm(fitlm, x = newdata)

## Compute posterior mean and
## 95% pointwise credible bands
```

```
meandens<-apply(dens$density, 1, mean)
lowerdens<-apply(dens$density, 1, quantile, 0.025)
upperdens<-apply(dens$density, 1, quantile, 0.975)

## Auxiliary variables
grid<-dens$grid
ngrid<-length(grid)

## Compute the true conditional densities
denst<-vector()
for(j in c(0, 1)){
for(i in 1:length(xcpred)){
denst<-c(denst, dnorm(grid, mu(xcpred[i], j), sigma(xcpred[i])))
}
}

## Data frame for plotting
dtf<-data.frame(grid = dens$grid,
xc = rep(round(xcpred, 3), each = ngrid),
xd = rep(c(0, 1), each = ngrid*3),
density = meandens,
lower = lowerdens,
upper = upperdens)

## True conditional densities data frame
truedf<-data.frame(x = grid,
y = denst,
xc = rep(round(xcpred, 3), each = ngrid),
xd = rep(c(0, 1), each = ngrid*3))

## Load ggplot2
require(ggplot2)

## Conditional densities plot
p_dens<-ggplot(dtf, aes(x = grid, y = density)) +
labs(x = "y", y = "Density") +
geom_line(lwd = 1) +
geom_ribbon(aes(ymin = lower, ymax = upper),
alpha = 0.3) +
geom_line(data = truedf,
mapping = aes(x = x, y = y),
lty = 2, color = "red") +
facet_grid(xd ~ xc,
labeller = label_bquote(
cols = bold(x^c ~ "=" ~ .(xc)),
rows = bold(x^d ~ "=" ~ .(xd)))) +
theme_custom()
p_dens

## End(Not run)
```

---

ddpm *Dirichlet process mixture conditional density*

Description

Conditional density of a Dirichlet process mixture (DPM) model using a normal/Bernoulli kernel.

Usage

```
ddpm(fit, grid, x = NULL, ncores = 1, nthin = 1, silent = FALSE)
```

Arguments

`fit`	A resulting object from a call to `dpm`.
`grid`	An optional numeric vector of quantiles of the response variable. If missing, a uniform grid from `min(y) -1` to `max(y) + 1` of length 201 is used.
`x`	An optional data frame containing the values of the covariates in `fit` used for computing the conditional density.
`ncores`	An optional integer specifying the number of cores that should be used. The default value is 1. See "Note".
`nthin`	An optional integer specifying whether to perform thinning of the chain or not. Default value is 1, meaning that no thinning is performed. This is useful to speed up computations once convergence was assessed.
`silent`	An optional logical argument specifying whether a progress bar should be printed or not. The default value is `FALSE`, then a progress bar is printed.

Details

The conditional density is given by

$$f(y \mid \boldsymbol{x}^c, \boldsymbol{x}^d) = \sum_{l=1}^{L} q_l(\boldsymbol{x}) \phi(y \mid \mu_l, \sigma_l^2),$$

where $\boldsymbol{x} = (\boldsymbol{x}^c, \boldsymbol{x}^d)$, $q_l(\boldsymbol{x}) \propto \omega_l \phi(\boldsymbol{x}^c \mid \boldsymbol{\mu}_l^x, \boldsymbol{\Sigma}_l^{xx}) \prod_{k=1}^{q} \text{Bern}(x_k^d \mid \pi_{lk})$ are the components weights with $\boldsymbol{\mu}_l^x$ and $\boldsymbol{\Sigma}_l^{xx}$ the corresponding mean vector and covariance matrix induced by the joint multivariate normal distribution and $\text{Bern}(x_k^d \mid \pi_{lk})$ denoting the density function of a Bernoulli distribution with success probability π_{lk}, $\phi(\cdot \mid \mu_l, \sigma_l^2)$ denotes the density function of a normal distribution with mean μ_l and variance σ_l^2, again with μ_l and σ_l^2 the induced mean and variance, respectively.

Value

A list containing:

`grid`	The grid used to compute the density.
`density`	A numeric matrix containing the posterior density sample.

Note

Changing the default value of the number of cores might slow down the process in some cases. This function can be really slow, specially if binary covariates.

Author(s)

Javier E. Garrido Guillén.

References

Fronczyk, K., Kottas, A. & Munch, S. (2012), "Flexible modeling for stock-recruitment relationships using Bayesian nonparametric mixtures", Environmental and Ecological Statistics **19**(2), 183–204.
Müller, P., Erkanli, A. & West, M. (1996), "Bayesian curve fitting using multivariate normal mixtures", Biometrika **83**(1), 67–79.

See Also

dpm

Examples

```
## Not run:
########################
## No covariates case ##
########################
## Simulate data
## Set seed for reproducibility
set.seed(123)
n<-200
y<-rnorm(n, 0, 1)

## Fit a Dirichlet process mixture model
fitdpm<-dpm(y ~ 1)

## Compute the density function
dens<-ddpm(fitdpm)

## Compute posterior mean and
## 95% pointwise credible bands
meandens<-apply(dens$density, 1, mean)
lowerdens<-apply(dens$density, 1, quantile, 0.025)
upperdens<-apply(dens$density, 1, quantile, 0.975)

## Use a different grid
mygrid<-seq(mean(y) - 3*sd(y), mean(y) + 3*sd(y), len = 512)
densa<-ddpm(fitdpm, grid = mygrid)

## Compute posterior mean and
## 95% pointwise credible bands
meandensa<-apply(densa$density, 1, mean)
lowerdensa<-apply(densa$density, 1, quantile, 0.025)
upperdensa<-apply(densa$density, 1, quantile, 0.975)

## Performing thinning
densb<-ddpm(fitdpm, nthin = 10)
dim(dens$density)
dim(densb$density)

## Compute posterior mean and
## 95% pointwise credible bands
meandensb<-apply(densb$density, 1, mean)
lowerdensb<-apply(densb$density, 1, quantile, 0.025)
upperdensb<-apply(densb$density, 1, quantile, 0.975)
```

```
## Data frames for plotting
dtf<-data.frame(grid = dens$grid,
density = meandens,
lower = lowerdens,
upper = upperdens)
dtfa<-data.frame(grid = densa$grid,
density = meandensa,
lower = lowerdensa,
upper = upperdensa)
dtfb<-data.frame(grid = densb$grid,
density = meandensb,
lower = lowerdensb,
upper = upperdensb)

## Load ggplot2 and gridExtra
require(ggplot2)
require(gridExtra)

## Density plots
p_dens<-ggplot(dtf, aes(x = grid, y = density)) +
labs(x = "y",
y = "Density") +
geom_line(lwd = 1) +
geom_ribbon(aes(ymin = lower, ymax = upper),
alpha = 0.3) +
theme_custom()
p_densa<-ggplot(dtfa, aes(x = grid, y = density)) +
labs(x = "y",
y = "Density") +
geom_line(lwd = 1) +
geom_ribbon(aes(ymin = lower, ymax = upper),
alpha = 0.3) +
theme_custom()
p_densb<-ggplot(dtfb, aes(x = grid, y = density)) +
labs(x = "y",
y = "Density") +
geom_line(lwd = 1) +
geom_ribbon(aes(ymin = lower, ymax = upper),
alpha = 0.3) +
theme_custom()

## Arrange plots
grid.arrange(p_dens, p_densa, p_densb, ncol = 2)


######################
## Covariates case ##
######################
## True mean function
mu<-function(xc, xd) (0.7*sin(xc)*cos(6*xc) + 0.4)*
(1 - xd) + 0.3*exp(-2*xc + 1)*xd

## True standard deviation function
sigma<-function(xc) 0.6*(1 - xc)^3 - 0.2*xc^2 + 0.5*xc

## Simulate data
## Set seed for reproducibility,
```

```
## simulate 200 observations of
## a normal distribution where
## mean and variance depend on
## continuous/binary covariate
n<-200
xc<-runif(n)
xd<-rbinom(n, 1, 0.5)
y<-rnorm(n, mu(xc, xd), sigma(xc))

## Create data frame with data
## NOTE: binary covariates must
## be defined as factors
mydata<-data.frame(y = y,
xc = xc,
xd = factor(xd))

## Fit a Dirichlet process mixture model
fitdpm<-dpm(y ~ xc + xd, mydata)

## For illustration purposes, we
## choose three random values of
## the continuous covariate xc
xcpred<-runif(3)
xdpred<-factor(c(0, 1))

## New data for conditional densities
newdata<-expand.grid(xc = xcpred, xd = xdpred)

## Compute the density function
dens<-ddpm(fitdpm, x = newdata)

## Compute posterior mean and
## 95% pointwise credible bands
meandens<-apply(dens$density, 1, mean)
lowerdens<-apply(dens$density, 1, quantile, 0.025)
upperdens<-apply(dens$density, 1, quantile, 0.975)

## Auxiliary variables for the plot
grid<-dens$grid
ngrid<-length(grid)

## Compute the true conditional densities
denst<-vector()
for(j in c(0, 1)){
for(i in 1:length(xcpred)){
denst<-c(denst, dnorm(grid, mu(xcpred[i], j), sigma(xcpred[i])))
}
}

## Changing the number of cores: Slower!
densa<-ddpm(fitdpm, x = newdata, ncores = 2)

## Compute posterior mean and
## 95% pointwise credible bands
meandensa<-apply(densa$density, 1, mean)
lowerdensa<-apply(densa$density, 1, quantile, 0.025)
upperdensa<-apply(densa$density, 1, quantile, 0.975)
```

```
## Data frames for plotting
dtf<-data.frame(grid = dens$grid,
xc = rep(round(xcpred, 3), each = ngrid),
xd = rep(c(0, 1), each = ngrid*3),
density = meandens,
lower = lowerdens,
upper = upperdens)
dtfa<-data.frame(grid = densa$grid,
xc = rep(round(xcpred, 3), each = ngrid),
xd = rep(c(0, 1), each = ngrid*3),
density = meandensa,
lower = lowerdensa,
upper = upperdensa)
truedf<-data.frame(x = grid, y = denst,
xc = rep(round(xcpred, 3), each = ngrid),
xd = rep(c(0, 1), each = ngrid*3))

## Load ggplot2
require(ggplot2)

## Conditional densities plot
p_dens<-ggplot(dtf, aes(x = grid, y = density)) +
labs(x = "y", y = "Density") +
geom_line(lwd = 1) +
geom_ribbon(aes(ymin = lower, ymax = upper),
alpha = 0.3) +
geom_line(data = truedf,
mapping = aes(x = x, y = y), lty = 2,
color = "red") +
facet_grid(xd ~ xc,
labeller = label_bquote(
cols = bold(x^c ~ "=" ~ .(xc)),
rows = bold(x^d ~ "=" ~ .(xd)))) +
theme_custom()
p_dens

p_densa<-ggplot(dtfa, aes(x = grid, y = density)) +
labs(x = "y", y = "Density") +
geom_line(lwd = 1) +
geom_ribbon(aes(ymin = lower, ymax = upper),
alpha = 0.3) +
geom_line(data = truedf,
mapping = aes(x = x, y = y), lty = 2,
color = "red") +
facet_grid(xd ~ xc,
labeller = label_bquote(
cols = bold(x^c ~ "=" ~ .(xc)),
rows = bold(x^d ~ "=" ~ .(xd)))) +
theme_custom()
p_densa

## End(Not run)
```

---

dpm *Dirichlet process mixture model*

Description

Fits a Dirichlet process mixture (DPM) model using a normal/Bernoulli kernel.

Usage

```
dpm(formula, data, prior, mcmc, start = NULL,
scale = TRUE, silent = FALSE)
```

Arguments

formula	An object of class formula, similar to the one passed to lm.
data	An optional data frame containing the data. If data is missing, then the variables are taken from the environment.
prior	An optional named list specifying the prior information. See "Details".
mcmc	An optional named list containing the values for the MCMC algorithm. See "Details".
start	An optional named list specifying the starting values for the parameters. Default value is NULL, meaning that starting values are based on the data. See inits for an example of how it should be passed.
scale	An optional logical argument specifying whether the data should be centered and standardrized or not. The default value is TRUE.
silent	An optional logical argument specifying whether a progress bar should be printed or not. The default value is FALSE, then a progress bar is printed.

Details

Let Y be a continuous random variable with $\boldsymbol{X}^c$ and $\boldsymbol{X}^d$ two $p-1$ and q–dimensional vectors of continuous and binary covariates, respectively. The joint density of $(Y, \boldsymbol{X}^c, \boldsymbol{X}^d)$ is modelled as a Dirichlet process mixture (DPM) using a normal/Bernoulli kernel, that is, the joint density can be expressed as follows

$$f(y, \boldsymbol{x}^c, \boldsymbol{x}^d) = \int \phi(\boldsymbol{u}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \prod_{k=1}^{q} \text{Bern}(x_k^d \mid \pi_k) \text{d}G(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}), \quad G \sim \text{DP}(\alpha, G^*),$$

where $\boldsymbol{u} = (y, \boldsymbol{x}^c)$, $\phi(\cdot \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the density function of a p–variate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, $\text{Bern}(\cdot \mid \pi_k)$ stands for a Bernoulli density function with probability of success π_k. The mixing distribution G follows a Dirichlet process (DP) with concentration parameter α and baseline distribution G^*.

Using a latent random variable z for the components of the mixture, the model can be represented hierarchically as follows

$$\boldsymbol{u}, \boldsymbol{x}^d \mid z, \boldsymbol{\theta} \sim \text{N}(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z) \prod_{k=1}^{q} \text{Bernoulli}(\pi_{zk}),$$

$$z \mid \boldsymbol{v} \sim \text{Mult}(1, \boldsymbol{\omega}),$$

$$(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l) \mid \boldsymbol{\mu}_0, \boldsymbol{V}_0, \nu_0, \boldsymbol{\Psi}_0 \sim \text{N}(\boldsymbol{\mu}_l \mid \boldsymbol{\mu}_0, \boldsymbol{V}_0)\text{IW}(\boldsymbol{\Sigma}_l \mid \nu_0, \boldsymbol{\Psi}_0),$$

$$\pi_{lk} \mid a_{\pi_k}, b_{\pi_k} \sim \text{Beta}(a_{\pi_k}, b_{\pi_k}),$$

$$v_l \mid \alpha \sim \text{Beta}(1, \alpha),$$

where $\boldsymbol{\theta} = (v_1, \ldots, v_L, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_L, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_L, \boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_L)$, $\boldsymbol{\omega}$ are the stick-breaking weights, $\text{IW}(\nu, \boldsymbol{\Psi})$ denotes an inverse-Wishart distribution with mean $\frac{1}{(\nu-p-1)}\boldsymbol{\Psi}$ and ν degrees of freedom.

Hence, the hyperparameters $\boldsymbol{\mu}_0, \boldsymbol{V}_0, \nu_0, \boldsymbol{\Psi}_0, a_\pi, b_\pi, \alpha$ must be included along with the maximum number of distributions used in the mixture, i.e., L, in a named list object and passed to the `prior` argument as `list(L,mu_0,V_0,nu_0,Psi_0,a_pi,b_pi,alpha)`. Further specification is possible through

$$\begin{aligned}
\boldsymbol{\mu}_0 \mid \boldsymbol{m}, \boldsymbol{S} &\sim \text{N}(\boldsymbol{m}, \boldsymbol{S}),\\
\boldsymbol{V}_0 \mid r, \boldsymbol{Q} &\sim \text{IW}(r, \boldsymbol{Q}),\\
\boldsymbol{\Psi}_0 \mid \nu_1, \boldsymbol{\Psi}_1 &\sim \text{W}(\nu_1, \boldsymbol{\Psi}_1),\\
\alpha \mid a_\alpha, b_\alpha &\; \Gamma(a_\alpha, b_\alpha),
\end{aligned}$$

where $\text{W}(\nu, \boldsymbol{\Psi})$ denotes a Wishart distribution with mean $\nu\boldsymbol{\Psi}$ and ν degrees of freedom. In this case, if `mu_0` is not specified, then `m` and `S` must be supplied. Similar, if `V_0` is not specified, `r` and `Q` must be supplied. If `Psi_0` is not supplied, then `nu_1` and `Psi_1` must be provided. Finally, if one wants to allow the data to inform about the appropriate value of α, then `a_alpha` and `b_alpha` must be specified.

If `prior` is not supplied and `scale = TRUE`, then relatively vague hyperparameters are used based on the centered and scaled data, namely `prior = list(L = 20,m = rep(0,p),S = 0.25*diag(p),r = p + 2,Q = 0.25*diag(p),nu_0 = p + 2,nu_1 = p,Psi_1 = 0.5/p*diag(p),a_alpha = 2,b_alpha = 2,a_pi = rep(0.5,q),b_pi = rep(0.5,q))`, where `p` and `q` are the dimension of $\boldsymbol{u}$ and $\boldsymbol{x}^d$, respectively. Otherwise, an empirical Bayes approach is used to set the hyperparameters, i.e., the sample mean and sample covariance matrix are used to specify the corresponding hyperparameters.

The argument `mcmc` specify the number of saved iterations, burn-in period and thinning of the chain. It must be a named list containing these values. Default values are `list(nsave = 3000,nburn = 1000,nthin = 1)`.

The starting values for $\boldsymbol{\omega}, \boldsymbol{\mu}, \boldsymbol{\Sigma}$ and $\boldsymbol{\pi}$ are set by default considering the scaled and centered data. If different starting values are needed, e.g., to run diagnostic checks, we recommend to use the utility function `inits`. Otherwise, a named list object must be specified with the following elements:

1. `omega`: a numeric vector of `L` elements summing up to one, representing the weigths of the mixture.
2. `mu`: a $(p \times L)$ matrix representing the starting values of the components mean, where p is the dimension of $\boldsymbol{u}$, i.e., the number of continuous variables including the response.
3. `sigma`: a list of length `L` which each element is a $(p \times p)$ positive-definite matrix representing the starting values of the components covariance matrix.
4. `pi`: a $(q \times L)$ matrix representing the starting values of the components probability of success, where q is the dimension of $\boldsymbol{x}^d$, i.e., the number of binary covariates.

Value

An list of class `dpm` containing:

`y`	The original continuous variables.
`xd`	The original binary covariates.
`prior`	The prior information used.
`params`	A list containing a posterior sample of the parameters $\boldsymbol{\omega}, \boldsymbol{\mu}, \boldsymbol{\Sigma}$ and $\boldsymbol{\pi}$.

The structure of the `params` object is

`params[[s]][[l]]`: are the posterior parameters at the s–th MCMC iteration of the l–th component.

Warning

Currently, the function does not check for the dimensions of the starting values. We encourage the use of inits.

Author(s)

Javier E. Garrido Guillén

References

Escobar, M. D. & West, M. (1995), "Bayesian density estimation and inference using mixtures", Journal of the American Statistical Association **90**(430), 577–588.
Ferguson, T. S. (1974), "Prior distributions on spaces of probability measures", The Annals of Statistics **2**(4), 615–629.
Ishwaran, H. & Zarepour, M. (2000), "Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models", Biometrika **87**(2), 371–390.
Richardson, S. & Green, P. J. (1997), "On Bayesian analysis of mixtures with an unknown number of components (with discussion)", Journal of the Royal Statistical Society, Series B **59**(4), 731–792.

Examples

```
## Not run:
########################
## No covariates case ##
########################
## Simulate data
## Set seed for reproducibility
set.seed(123)
n<-200
y<-rnorm(n, 0, 1)

## Fit a Dirichlet process mixture model
fitdpm<-dpm(y ~ 1)

## Plot method: Only available when no covariates
plot(fitdpm)

## Fit a Dirichlet process mixture model
## without standardizing the data
fitdpma<-dpm(y ~ 1, scale = FALSE)

## User-defined prior
myprior<-list(L = 10,
mu_0 = 0, V_0 = 10,
nu_0 = 4, Psi_0 = 1,
a_alpha = 2, b_alpha = 2)
fitdpmb<-dpm(y ~ 1, prior = myprior)

## Changing MCMC settings
mymcmc<-list(nsave = 10000, nburn = 5000, nthin = 1)
fitdpmc<-dpm(y ~ 1, mcmc = mymcmc)

## Changing starting values and silent
## NOTE: see help(inits) for details
fitdpmd<-dpm(y ~ 1, start = inits(20, 1), silent = TRUE)
```

```
## Create a data frame containing
## the true density function
grid<-seq(-4, 4, len = 201)
dtf<-data.frame(x = grid, y = dnorm(grid, 0, 1))

## Load ggplot2 and gridExtra
require(ggplot2)
require(gridExtra)

## Plot the densities along with the true one
p<-plot(fitdpm, title = "Default settings") +
geom_line(data = dtf,
mapping = aes(x = x, y = y),
lwd = 1, lty = 2,
col = "red")
pa<-plot(fitdpma, title = "Unstandardized data") +
geom_line(data = dtf,
mapping = aes(x = x, y = y),
lwd = 1, lty = 2,
col = "red")
pb<-plot(fitdpmb, title = "User-defined prior") +
geom_line(data = dtf,
mapping = aes(x = x, y = y),
lwd = 1, lty = 2,
col = "red")
pc<-plot(fitdpmc, title = "User-defined MCMC settings") +
geom_line(data = dtf,
mapping = aes(x = x, y = y),
lwd = 1, lty = 2,
col = "red")
pd<-plot(fitdpmd, title = "Random starting values") +
geom_line(data = dtf,
mapping = aes(x = x, y = y),
lwd = 1, lty = 2,
col = "red")

## Arrange plots
grid.arrange(p, pa, pb, pc, pd, ncol = 3)


######################
## Covariates case ##
######################
## True mean function
mu<-function(xc, xd) (0.7*sin(xc)*cos(6*xc) + 0.4)*
(1 - xd) + 0.3*exp(-2*xc + 1)*xd

## True standard deviation function
sigma<-function(xc) 0.6*(1 - xc)^3 - 0.2*xc^2 + 0.5*xc

## Simulate data
## Set seed for reproducibility,
## simulate 200 observations of
## a normal distribution where
## mean and variance depend on
## continuous/binary covariate
```

```
n<-200
xc<-runif(n)
xd<-rbinom(n, 1, 0.5)
y<-rnorm(n, mu(xc, xd), sigma(xc))

## Create data frame with data
## NOTE: binary covariates must
## be defined as factors
mydata<-data.frame(y = y,
xc = xc,
xd = factor(xd))

## Fit a Dirichlet process mixture model
fitdpm<-dpm(y ~ xc + xd, mydata)

## Changing prior and starting values
myprior<-list(L = 10,
m = c(0, 0), S = 100*diag(2),
r = 4, Q = 0.5*diag(2),
nu_0 = 4, nu_1 = 2, Psi_1 = 0.1*diag(2),
a_alpha = 2, b_alpha = 2,
a_pi = 0.5, b_pi = 0.5)

## NOTE: see help(inits) for a description
fitdpma<-dpm(y ~ xc + xd, mydata,
prior = myprior,
start = inits(10, 2, 1))

## Plot methods: pops an error!
## Visualisation: either conditional densities
## check help(ddpm) or mean/variance functions
## check help(meandpm) or help(vardpm)
plot(fitdpm)
plot(fitdpma)

## End(Not run)
```

`gq` *Numerical integration using a Gaussian quadrature*

Description

Performs a numerical integration using a Gaussian quadrature by a direct call to the `gauss.quad` function in the `statmod` package.

Usage

```
gq(f, grid, N = 500)
```

Arguments

`f`	A numeric vector containing the function evaluated at every single point of the supplied grid.

grid	A numeric vector containing a grid of the interval where the integral will be computed on.
N	An optional integer specifying the number of nodes and weights of the Gaussian quadrature. The default value is 500.

Details

See gauss.quad for further details.

Value

The function returns the value of the integral in the supplied interval.

Author(s)

Javier E. Garrido Guillén

See Also

gauss.quad

Examples

```
## Set a grid of the interval
grid<-seq(-4, 4, len = 201)

## Evaluate the function in such grid
f<-dnorm(grid)

## Compute the integral
gq(f, grid)

## Change the number of nodes
gq(f, grid, 2000)
```

---

inits *Random starting values for DPM models*

Description

A random generator of starting values for DPM models.

Usage

```
inits(L, p, q = NULL)
```

Arguments

L	Maximum number of densities used in the mixture.
p	Dimension of the continuous variables (response and covariates).
q	Dimension of the binary covariates.

Details

This is a utility function helpful when different chains started at different values are needed. It will generate random starting values for ω, μ, Σ and π.

Value

A list containing:

omega	A numeric vector of L elements summing up to one, representing the weigths of the mixture.
mu	A p x L matrix representing the starting values of the components mean, where p is the dimension of u, i.e., the number of continuous variables including the response.
sigma	A list of length L which each element is a (p x p) positive-definite matrix representing the starting values of the components covariance matrix.
pi	a (q x L) matrix representing the starting values of the components probability of success, where q represents the number of binary covariates.

Author(s)

Javier E. Garrido Guillén

See Also

dpm

Examples

```
## Not run:
########################
## No covariates case ##
########################
## Generate randomly the initial
## values of the parameters for
## a Dirichlet process mixture
## without covariates and default
## settings
start<-inits(20, 1)

## User-defined prior
myprior<-list(L = 10,
mu_0 = 0, V_0 = 10,
nu_0 = 4, Psi_0 = 1,
a_alpha = 2, b_alpha = 2)

## Corresponding starting values
start<-inits(10, 1)


####################
## Covariates case ##
####################
## Generate randomly the initial
## values of the parameters for
```

```
## a Dirichlet process mixture
## with 1 continuous covariate
## and 3 binary covariates using
## default settings
start<-inits(20, 2, 3)

## User-defined prior
myprior<-list(L = 10,
m = c(0, 0), S = 100*diag(2),
r = 4, Q = 0.5*diag(2),
nu_0 = 4, nu_1 = 2, Psi_1 = 0.1*diag(2),
a_alpha = 2, b_alpha = 2,
a_pi = 0.5, b_pi = 0.5)

## Corresponding starting values
start<-inits(10, 2, 1)

## End(Not run)
```

---

mbb *Model comparison based on Bayesian bootstrap (Posterior rank probability)*

Description

Computes the posterior rank probability for a set of models.

Usage

```
mbb(cpo1, cpo2, ..., B = 10000)
```

Arguments

cpo1	A numeric vector containing the conditional predictive ordinates (CPO) of the first model.
cpo2	A numeric vector containing the CPO of the second model.
...	Further numeric vectors containing the CPO of other competing models.
B	Number of Bayesian bootstrap replications. Default value is 10,000.

Details

The posterior rank probability was proposed by Cheng et al. (2019). It combines the conditional predictive ordinates (CPO) with the Bayesian bootstrap (BB) to provide a measure of uncentainty in the choice of one model over another. Thus, one can compare multiple models choosing the one with the highest rank, on each of B bootstrap samples, by sorting the models with

$$\frac{1}{n}\sum_{i=1}^{n} w_{i,b}\log(\text{CPO}_{i,M}),$$

where n is the sample size, $w_{i,b}$ are the correspoinding weights of the BB and $\text{CPO}_{i,M}$ are the corresponding CPOs of model M. Then, counting across all bootstrap samples the number of times N_m, each model was the best and computing its probability as N_m/B, gives the so-called posterior rank probability.

Value

A vector containing the posterior rank probabilities for each model.

Author(s)

Javier E. Garrido Guillén

References

Cheng, L., Ramchandran, S., Vatanen, T., Lietzén, N., Lahesmaa, R., Vehtari, A. & Lähdesmäki, H. (2019), "An additive Gaussian process regression model for interpretable non-parametric analysis of longitudinal data", Nature communications **10**(1), 1–11.

Examples

```
## Not run:
############################################
## Compare different prior specifications ##
############################################
## Simulate data
## Set seed for reproducibility
n<-200
y<-rnorm(n, 0, 1)

## First model: default settings
fitdpm<-dpm(y ~ 1)

## Second model: fixed alpha
priora<-list(L = 10,
mu_0 = 0, V_0 = 10,
nu_0 = 4, Psi_0 = 1,
alpha = 1)
fitdpma<-dpm(y ~ 1, prior = priora)

## Third model: all random - option 1
priorb<-list(L = 10,
m = 0, S = 100,
r = 4, Q = 1,
nu_0 = 4, nu_1 = 1, Psi_1 = 0.5,
a_alpha = 4, b_alpha = 4)
fitdpmb<-dpm(y ~ 1, prior = priorb)

## Fourth model: all random - option 2
priorc<-list(L = 20,
m = 0, S = 0.25,
r = 4, Q = 0.25,
nu_0 = 4, nu_1 = 1, Psi_1 = 0.5,
a_alpha = 2, b_alpha = 2)
fitdpmc<-dpm(y ~ 1, prior = priorc)

## Compute the CPOs
cpo<-criteria(fitdpm)$cpo
cpoa<-criteria(fitdpma)$cpo
cpob<-criteria(fitdpmb)$cpo
cpoc<-criteria(fitdpmc)$cpo
```

```
## Compute the posterior rank probability
mbb(cpo, cpoa, cpob, cpoc, B = 100000)


############################
## Compare a DPM and a LM ##
############################
## Set a seed for reproducibility
set.seed(123)

## Simulate 500 observations from
## a mixture of normals, where the
## weights depend on a continuous
## covariate (example taken from
## Dunson et al., 2007)
n<-500
x<-runif(n)
y1<-x + rnorm(n, 0, sqrt(0.01))
y2<-x^4 + rnorm(n, 0, sqrt(0.04))
u<-runif(n)
prob<-exp(-2*x)
y<-ifelse(u < prob, y1, y2)

## Fit a Dirichlet process mixture model
fitdpm<-dpm(y ~ x,
mcmc = list(nsave = 10000, nburn = 5000, nthin = 1))

## Fit a linear regression model
fitlm<-blm(y ~ x,
mcmc = list(nsave = 10000, nburn = 5000, nthin = 1))

## Compute the CPOs
cpodpm<-criteria(fitdpm)$cpo
cpolm<-criteria(fitlm)$cpo

## Compute the posterior rank probability
mbb(cpodpm, cpolm)

## End(Not run)
```

`meanblm` *Mean regression function for linear models*

Description

Returns the mean function of a Gaussian linear regression model.

Usage

```
meanblm(fit, x)
```

Arguments

`fit`	An output object from the `blm` function.

x	A data frame containing the values of all the covariates included in the fit model.

Details

This function works in a similar fashion to the predict function for lm objects.

Value

A list object of class blmx containing

x	The original data.
posterior	A numeric matrix containing the posterior mean function sample.

Author(s)

Javier E. Garrido Guillén

See Also

blm

Examples

```
## Not run:
#######################
## Simulated example ##
#######################
## True mean function
mu<-function(xc, xd) (0.7*sin(xc)*cos(6*xc) + 0.4)*
(1 - xd) + 0.3*exp(-2*xc + 1)*xd

## True standard deviation function
sigma<-function(xc) 0.6*(1 - xc)^3 - 0.2*xc^2 + 0.5*xc

## Simulate data
## Set seed for reproducibility,
## simulate 200 observations of
## a normal distribution where
## mean and variance depend on
## continuous/binary covariate
n<-200
xc<-runif(n)
xd<-rbinom(n, 1, 0.5)
y<-rnorm(n, mu(xc, xd), sigma(xc))

## Create data frame with data
## NOTE: binary covariates must
## be defined as factors
mydata<-data.frame(y = y,
xc = xc,
xd = factor(xd))

## Fit a linear regression model
fitlm<-blm(y ~ xc + xd, mydata)
```

```
## New data for the conditional mean function
xcpred<-seq(0, 1, len = 11)
xdpred<-factor(c(0, 1))
newdata<-expand.grid(xc = xcpred, xd = xdpred)

## Data frame of the true mean function
truedf<-data.frame(xc = xcpred,
truemean = c(mu(xcpred, 0), mu(xcpred, 1)),
xd = rep(c(0, 1), each = length(xcpred)))

## Compute the mean function
meanfunc<-meanblm(fitlm, newdata)

## Load ggplot2
require(ggplot2)

## Plot the (posterior mean) mean function,
## the true function and the data
plot(meanfunc) +
geom_point(data = mydata,
mapping = aes(x = xc,  y = y),
alpha = 0.5, col = "slategray2") +
geom_line(data = truedf,
mapping = aes(x = xc, y = truemean),
lwd = 1, color = "red", lty = 2) +
facet_wrap(~ xd)

## End(Not run)
```

`meandpm` *Mean function for DPM models*

Description

Returns the mean function of a DPM model.

Usage

```
meandpm(fit, x, silent = FALSE)
```

Arguments

`fit`	A resulting object from a call to `dpm`.
`x`	A data frame containing the values of all the covariates included in the `fit` model.
`silent`	An optional logical argument specifying whether a progress bar should be printed or not. The default value is `FALSE`, then a progress bar is printed.

Details

This function works in a similar fashion to the `predict` function for `lm` objects.

The `x` argument must contain the covariates included in the `fit` object with the same classes.

Value

A list object of class dpmx containing

x	The original data.
posterior	A numeric matrix containing the posterior mean function sample.
xc	The continuous covariates.
xd	The binary covariates.

Author(s)

Javier E. Garrido Guillén

See Also

dpm

Examples

```
## Not run:
#######################
## Simulated example ##
#######################
## True mean function
mu<-function(xc, xd) (0.7*sin(xc)*cos(6*xc) + 0.4)*
(1 - xd) + 0.3*exp(-2*xc + 1)*xd

## True standard deviation function
sigma<-function(xc) 0.6*(1 - xc)^3 - 0.2*xc^2 + 0.5*xc

## Simulate data
## Set seed for reproducibility,
## simulate 200 observations of
## a normal distribution where
## mean and variance depend on
## continuous/binary covariate
n<-200
xc<-runif(n)
xd<-rbinom(n, 1, 0.5)
y<-rnorm(n, mu(xc, xd), sigma(xc))

## Create data frame with data
## NOTE: binary covariates must
## be defined as factors
mydata<-data.frame(y = y,
xc = xc,
xd = factor(xd))

## Fit a Dirichlet process mixture model
fitdpm<-dpm(y ~ xc + xd, mydata)

## New data for the conditional mean/variance function
xcpred<-seq(0, 1, len = 21)
xdpred<-factor(c(0, 1))
newdata<-expand.grid(xc = xcpred, xd = xdpred)
```

```
## Data frame of the true functions
truedf<-data.frame(xc = xcpred,
truemean = c(mu(xcpred, 0), mu(xcpred, 1)),
truevar = sigma(xcpred)^2,
xd = rep(c(0, 1), each = length(xcpred)))

## Compute the mean function
meanfunc<-meandpm(fitdpm, newdata)

## Load ggplot2
require(ggplot2)

## Plot the (posterior mean) mean function,
## the true function and the data
plot(meanfunc) +
geom_point(data = mydata,
mapping = aes(x = xc,  y = y),
alpha = 0.5, col = "slategray2") +
geom_line(data = truedf,
mapping = aes(x = xc, y = truemean),
lwd = 1, color = "red", lty = 2) +
facet_wrap(~ xd)


##########################################
## Another example, using simulated data ##
##########################################
## Set a seed for reproducibility
set.seed(123)

## Simulate 500 observations from
## a mixture of normals, where the
## weights depend on a continuous
## covariate (example taken from
## Dunson et al., 2007)
n<-500
x<-runif(n)
y1<-x + rnorm(n, 0, sqrt(0.01))
y2<-x^4 + rnorm(n, 0, sqrt(0.04))
u<-runif(n)
prob<-exp(-2*x)
y<-ifelse(u < prob, y1, y2)

## Fit a Dricihlet process mixture model
fitdpm<-dpm(y ~ x,
mcmc = list(nsave = 10000, nburn = 5000, nthin = 1))

## New data for the conditional mean/variance function
xpred<-seq(0, 1, len = 21)
newdata<-data.frame(x = xpred)

## Data frame of the true functions
truedf<-data.frame(x = xpred,
truemean = exp(-2*xpred)*xpred + (1 - exp(-2*xpred))*xpred^4)

## Compute the mean function
meanfunc<-meandpm(fitdpm, newdata)
```

```
## Load ggplot2
require(ggplot2)

## Plot the (posterior mean) mean function,
## the true function and the data
plot(meanfunc) +
geom_point(data = data.frame(x = x, y = y),
mapping = aes(x = x,  y = y),
alpha = 0.5, col = "slategray2") +
geom_line(data = truedf,
mapping = aes(x = x, y = truemean),
lwd = 1, color = "red", lty = 2)

## End(Not run)
```

---

ovl *Bayesian nonparametric overlap coefficient estimator*

Description

Estimates the (covariate-specific) overlap coefficient using a Dirichlet process mixture model.

Usage

```
ovl(formula.h, formula.d,
data, group, tag.h,
prior.h, prior.d,
mcmc,
start.h = NULL, start.d = NULL,
scale = TRUE, grid,
newdata = NULL,
ncores = 1, nthin = 1,
integral = trapezoidal, ...,
criteria = TRUE, silent = FALSE)
```

Arguments

formula.h	An object of class formula specifying the model for the healthy group.
formula.d	An object of class formula specifying the model for the diseased group.
data	An optional data frame containing the data. If data is missing, then the variables are taken from the environment.
group	The name of the grouping variable contained in the data.
tag.h	The value used to label the healthy individuals.
prior.h	An optional named list specifying the prior information for the healthy group. See dpm for details.
prior.d	An optional named list specifying the prior information for the diseased group. See dpm for details.
mcmc	An optional named list containing the values for the MCMC algorithm. See dpm for details.

`start.h`	An optional named list specifying the starting values for the parameters of the healthy group. See `dpm` for details.
`start.d`	An optional named list specifying the starting values for the parameters of the diseased group. See `dpm` for details.
`scale`	An optional logical argument specifying whether the data should be centered and standardrized or not. The default value is `TRUE`.
`grid`	A numeric vector specifying the grid where the densities must be evaluated and where the integral will be computed on. If missing, a uniform grid of length 201 on the range of the response variable.
`newdata`	An optional data frame containing the values of the covariates used for computing the conditional density and thus, the covariate-specific overlap coefficient.
`ncores`	An optional integer specifying the number of cores that should be used. The default value is 1. See `ddpm` for details.
`nthin`	An optional integer specifying whether to perform thinning of the chain or not for the computation of the densities. See `ddpm` for details.
`integral`	A numerical integration function used to compute the overlap coefficient. Possible values include `trapezoidal`, which is the default, `simpson` and `gq`. Other user-defined functions are possible, but the same format is required as in the aforementioned ones.
`...`	Additional arguments passed to `integral`.
`criteria`	A logical value indicating whether to compute model comparison criteria or not. Default value is `TRUE`.
`silent`	An optional logical argument specifying whether a progress bar should be printed or not. The default value is `FALSE`, then a progress bar is printed.

Details

When no covariates are considered, the function computes the overlap coefficient (OVL), defined by Weitzman (1970) as

$$\int_{-\infty}^{\infty} \min\left\{f_0(y), f_1(y)\right\} dy,$$

where $f_0(y)$ and $f_1(y)$ are the corresponding density functions estimated using a Dirichlet process mixture model as in `dpm`.

In the case of included covariates, the function computes the covariate-specific OVL, defined as

$$\text{OVL}(\boldsymbol{x}) = \int_{-\infty}^{\infty} \min\left\{f_0(y \mid \boldsymbol{X} = \boldsymbol{x}), f_1(y \mid \boldsymbol{X} = \boldsymbol{x})\right\} dy.$$

In this implementation the integral is replaced by a numerical integration method specified in the `integral` function.

Value

A list object of class `ovl` containing:

`fit0`	The fitted object for the healthy group.
`fit1`	The fitted object for the diseased group.

dens0	The density object for the healthy group.
dens1	The density object for the diseased group.
ovl	The posterior sample of the (covariate-specific) overlap coefficient.
newdata	The data used for computing the covariate-specific overlap coefficient.
criteria0	The model comparison criteria for the healthy group.
criteria1	The model comparison criteria for the diseased group.

Author(s)

Javier E. Garrido Guillén

References

Samawi, H. M., Yin, J., Rochani, H. & Panchal, V. (2017), "Notes on the overlap measure as an alternative to the Youden index: How are they related?", Statistics in Medicine **36**(26), 4230–4240.
Schmid, F. & Schmidt, A. (2006), "Nonparametric estimation of the coefficient of overlapping–theory and empirical application", Computational Statistics & Data Analysis **50**(6), 1583–1596.
Weitzman, M. S. (1970), "Measures of overlap of income distributions of white and Negro families in the United States", Vol. 22, US Bureau of the Census.

See Also

dpm, ddpm, criteria

Examples

```
## Not run:
########################
## No covariates case ##
########################
## Simulate data
## Set seed for reproducibility
set.seed(123)
n<-200
y0<-rnorm(n, 0, 1)
y1<-rnorm(n, 1, 1)

## Compute the true OVL
lower<-min(qnorm(0.0001, 0, 1),
qnorm(0.0001, 1, 1))
upper<-max(qnorm(0.9999, 0, 1),
qnorm(0.9999, 1, 1))
grid<-seq(lower, upper, len = 512)
d0<-dnorm(grid, 0, 1)
d1<-dnorm(grid, 1, 1)
covlt<-trapezoidal(pmin(d0, d1), grid)

## Compute the coefficient of overlap
(res<-ovl(y0 ~ 1, y1 ~ 1))

## Posterior summary
summary(res)
covlt
```

```
## Using the Simpson's rule
(res_simp<-ovl(y0 ~ 1, y1 ~ 1, integral = simpson))

## Posterior summary
summary(res_simp)
covlt


#####################
## Covariates case ##
#####################
## True mean functions
mu1<-function(xc, xd) (0.7*sin(xc)*cos(6*xc) + 0.4)*
(1 - xd) + 0.3*exp(-2*xc + 1)*xd
mu2<-function(xc, xd) (0.5*sin(xc)*cos(8*xc) - 0.2)*
(1 - xd) + 0.4*exp(-xc^2 + 0.3*xc + 1)*xd

## True standard deviation functions
sigma1<-function(xc) 0.6*(1 - xc)^3 - 0.2*xc^2 + 0.5*xc
sigma2<-function(xc) dnorm(4*xc) + 0.2

## Create a grid to compute the true OVL
x<-seq(0, 1, len = 101)
ngrid<-201
lower0<-min(qnorm(0.0001, mu1(x, 0), sigma1(x)),
qnorm(0.0001, mu2(x, 0), sigma2(x)))
upper0<-max(qnorm(0.9999, mu1(x, 0), sigma1(x)),
qnorm(0.9999, mu2(x, 0), sigma2(x)))
lower1<-min(qnorm(0.0001, mu1(x, 1), sigma1(x)),
qnorm(0.0001, mu2(x, 1), sigma2(x)))
upper1<-max(qnorm(0.9999, mu1(x, 1), sigma1(x)),
qnorm(0.9999, mu2(x, 1), sigma2(x)))
grid<-seq(min(lower0, lower1), max(upper0, upper1),
len = ngrid)

## Compute the true OVL
covlt0<-covlt1<-vector()
for(i in 1:length(x)){
d00<-dnorm(grid, mu1(x[i], 0), sigma1(x[i]))
d10<-dnorm(grid, mu2(x[i], 0), sigma2(x[i]))
covlt0[i]<-trapezoidal(pmin(d00, d10), grid)

d01<-dnorm(grid, mu1(x[i], 1), sigma1(x[i]))
d11<-dnorm(grid, mu2(x[i], 1), sigma2(x[i]))
covlt1[i]<-trapezoidal(pmin(d01, d11), grid)
}

## True OVL data frame
truedf<-data.frame(xc = x, covlt = c(covlt0, covlt1),
xd = factor(rep(c(0, 1), each = length(x))))

## Simulate data
## Set seed for reproducibility,
## simulate 200 observations of
## a normal distribution where
## mean and variance depend on
## continuous/binary covariate
```

```
set.seed(123)
n<-200
xc0<-runif(n)
xd0<-rbinom(n, 1, 0.5)
xc1<-runif(n)
xd1<-rbinom(n, 1, 0.5)
y0<-rnorm(n, mu1(xc0, xd0), sigma1(xc0))
y1<-rnorm(n, mu2(xc1, xd1), sigma2(xc1))
mydata<-data.frame(y = c(y0, y1),
xc = c(xc0, xc1),
xd = as.factor(c(xd0, xd1)),
group = rep(c(0, 1), each = n))

## New data for the conditional densities
xcpred<-seq(0, 1, len = 11)
newdata<-expand.grid(xc = xcpred, xd = factor(c(0, 1)))

## Compute the covariate-specific OVL
(res<-ovl(y ~ xc + xd,
y ~ xc + xd,
mydata, "group", 0,
newdata = newdata))

## Posterior summary
summary(res)

## End(Not run)
```

---

ovlBB *Bayesian bootstrap-based overlap coefficient estimator*

Description

Estimates the overlap coefficient using the sum of error probabilities representation combined with the Bayesian bootstrap.

Usage

```
ovlBB(biomarker, data, group, tag.h,
prior.h, prior.d, mcmc,
start.h = NULL, start.d = NULL,
scale = TRUE, grid, ncores = 1, nthin = 1,
criteria = TRUE, silent = FALSE)
```

Arguments

biomarker	A character string specifying the name of the biomarker.
data	A data frame containing the data.
group	The character string specifying the name of the grouping variable.
tag.h	The value used to label the healthy individuals.
prior.h	An optional named list specifying the prior information for the healthy group. See dpm for details.

prior.d	An optional named list specifying the prior information for the diseased group. See dpm for details.
mcmc	An optional named list containing the values for the MCMC algorithm. See dpm for details.
start.h	An optional named list specifying the starting values for the parameters of the healthy group. See dpm for details.
start.d	An optional named list specifying the starting values for the parameters of the diseased group. See dpm for details.
scale	An optional logical argument specifying whether the data should be centered and standardrized or not. The default value is TRUE.
grid	A numeric vector specifying the grid where the densities must be evaluated. If missing, a uniform grid of length 201 on the range of the biomarker.
ncores	An optional integer specifying the number of cores that should be used. The default value is 1. See ddpm for details.
nthin	An optional integer specifying whether to perform thinning of the chain or not for the computation of the densities. See ddpm for details.
criteria	A logical value indicating whether to compute model comparison criteria or not. Default value is TRUE.
silent	An optional logical argument specifying whether a progress bar should be printed or not. The default value is FALSE, then a progress bar is printed.

Details

From the alternative definition of the coefficient of overlap as a sum of error probabilities derived by Schmid & Schmidt (2006). The Bayesian bootstrap-based estimator is given by

$$\text{OVL}_{\text{BB}} = \min\left\{1, \sum_{i=1}^{n_0} q_{0i} 1_{\{f_0(y_{0i}) < f_1(y_{0i})\}} + \sum_{j=1}^{n_1} q_{1j} 1_{\{f_0(y_{1j}) \geq f_1(y_{1j})\}}\right\},$$

where $\boldsymbol{q}_0 = (q_{01}, \ldots, q_{0n_0})$ and $\boldsymbol{q}_1 = (q_{11}, \ldots, q_{1n_1})$ are vectors of weights from the Bayesian bootstrap which are drawn from a Dirichlet $(1, \ldots, 1)$ distribution.

Value

A list object of class ovl containing:

fit0	The fitted object for the healthy group.
fit1	The fitted object for the diseased group.
dens0	The density object for the healthy group.
dens1	The density object for the diseased group.
ovl	The posterior sample of the overlap coefficient.
criteria0	The model comparison criteria for the healthy group.
criteria1	The model comparison criteria for the diseased group.

Author(s)

Vanda Inácio, Javier E. Garrido Guillén

References

Schmid, F. & Schmidt, A. (2006), "Nonparametric estimation of the coefficient of overlapping–theory and empirical application", Computational Statistics & Data Analysis **50**(6), 1583–1596.
Rubin, D. B. (1981), "The Bayesian bootstrap", The Annals of Statistics **9**(1), 130–134.

See Also

dpm, ddpm, criteria, ovl

Examples

```
## Not run:
########################
## Simulated example ##
########################
## Simulate data
## Set seed for reproducibility
set.seed(123)
n<-200
y0<-rnorm(n, 0, 1)
y1<-rnorm(n, 1, 1)

## Compute the true OVL
lower<-min(qnorm(0.0001, 0, 1),
qnorm(0.0001, 1, 1))
upper<-max(qnorm(0.9999, 0, 1),
qnorm(0.9999, 1, 1))
grid<-seq(lower, upper, len = 512)
d0<-dnorm(grid, 0, 1)
d1<-dnorm(grid, 1, 1)
covlt<-trapezoidal(pmin(d0, d1), grid)

## Create a data frame containing the data
mydata<-data.frame(y = c(y0, y1),
group = rep(c(0, 1), each = n))

## Compute the coefficient of overlap
(res_BB<-ovlBB("y", mydata, "group", 0))
summary(res_BB)
covlt

## End(Not run)
```

---

ovlfreq *Frenquentist estimator of the overlap coefficient*

Description

Estimates the overlap coefficient using a Gaussian kernel density estimator.

Usage

```
ovlfreq(y0, y1, grid, bw0 = "nrd0", bw1 = "nrd0",
B = 1000, alpha = 0.05,
integral = trapezoidal, ...)
```

Arguments

y0	A numeric vector containing the data from the nondiseased group.
y1	A numeric vector containing the data from the diseased group.
grid	A numeric vector specifying the grid where the overlap coefficient will be computed. If not supplied, a uniform grid ranging from min(y0,y1) -1 to max(y0,y1) + 1 will be used.
bw0	The smoothing bandwidth for the nondiseased group to be used. See density for details.
bw1	The smoothing bandwidth for the diseased group to be used. See density for details.
B	An integer specifying the number of bootstrap replications. The default value is set to 1,000.
alpha	A numeric value between 0 and 1 specifying the significance level used to construct a confidence interval for the overlap coefficient.
integral	The numerical integration function that should be used to compute the overlap coefficient. Possible values include trapezoidal, which is the default, simpson and gq. Other user-defined functions are possible, but the same format is required as in the aforementioned ones.
...	Additional arguments passed to the integral function.

Details

The overlap coefficient is defined by Weitzman (1970) as

$$\int_{-\infty}^{\infty} \min\left\{f_0(y), f_1(y)\right\} dy,$$

where $f_0(y)$ and $f_1(y)$ are the corresponding density functions. In this implementation a Gaussian kernel density estimator is used to estimate the density functions and the integral is replaced by a numerical integration method specified in the integral function.

Value

A list object of class ovlfreq containing the following:

OVL	A data frame containing the estimated overlap coefficient along with a (1 - alpha)*100% confidence interval.
dens0	An object of the class density resulting of the call to density for the nondiseased group.
dens1	An object of the class density resulting of the call to density for the diseased group.
var.names	A numeric vector of length 2 containing the deparsed name of the y0 and y1 arguments.

Author(s)

Javier E. Garrido Guillén

References

Samawi, H. M., Yin, J., Rochani, H. & Panchal, V. (2017), "Notes on the overlap measure as an alternative to the Youden index: How are they related?", Statistics in Medicine **36**(26), 4230–4240.
Schmid, F. & Schmidt, A. (2006), "Nonparametric estimation of the coefficient of overlapping–theory and empirical application", Computational Statistics & Data Analysis **50**(6), 1583–1596.
Weitzman, M. S. (1970), "Measures of overlap of income distributions of white and Negro families in the United States", Vol. 22, US Bureau of the Census.

See Also

density, trapezoidal, simpson, gq, ovlprob,

Examples

```
## Not run:
#######################
## Simulated example ##
#######################
## Simulate data
## Set seed for reproducibility
set.seed(123)
n<-200
y0<-rnorm(n, 0, 1)
y1<-rnorm(n, 1, 1)

## Compute the true OVL
lower<-min(qnorm(0.0001, 0, 1),
qnorm(0.0001, 1, 1))
upper<-max(qnorm(0.9999, 0, 1),
qnorm(0.9999, 1, 1))
grid<-seq(lower, upper, len = 512)
d0<-dnorm(grid, 0, 1)
d1<-dnorm(grid, 1, 1)
covlt<-trapezoidal(pmin(d0, d1), grid)

## Compute the coefficient of overlap
(resfreq<-ovlfreq(y0, y1))
covlt

## End(Not run)
```

---

ovlNorm *Bayesian parametric overlap coefficient estimator*

Description

Estimates the (covariate-specific) overlap coefficient using a normal-normal model.

Usage

```
ovlNorm(formula.h, formula.d,
data, group, tag.h,
prior.h, prior.d,
```

```
mcmc,
grid, newdata = NULL,
integral = trapezoidal, ...,
criteria = TRUE, silent = FALSE)
```

Arguments

formula.h	An object of class formula specifying the model for the healthy group.
formula.d	An object of class formula specifying the model for the diseased group.
data	An optional data frame containing the data. If data is missing, then the variables are taken from the environment.
group	The name of the grouping variable contained in the data.
tag.h	The value used to label the healthy individuals.
prior.h	An optional named list specifying the prior information for the healthy group. See blm for details.
prior.d	An optional named list specifying the prior information for the diseased group. See blm for details.
mcmc	An optional named list containing the values for the MCMC algorithm. See blm for details.
grid	A numeric vector specifying the grid where the densities must be evaluated and where the integral will be computed on. If missing, a uniform grid of length 201 on the range of the response variable.
newdata	An optional data frame containing the values of the covariates used for computing the conditional density and thus, the covariate-specific overlap coefficient.
integral	A numerical integration function used to compute the overlap coefficient. Possible values include trapezoidal, which is the default, simpson and gq. Other user-defined functions are possible, but the same format is required as in the aforementioned ones.
...	Additional arguments passed to integral.
criteria	A logical value indicating whether to compute model comparison criteria or not. Default value is TRUE.
silent	An optional logical argument specifying whether a progress bar should be printed or not. The default value is FALSE, then a progress bar is printed.

Details

When no covariates are considered, the function computes the overlap coefficient (OVL), defined by Weitzman (1970) as

$$\int_{-\infty}^{\infty} \min\left\{f_0(y), f_1(y)\right\} dy,$$

where $f_0(y)$ and $f_1(y)$ are the corresponding density functions estimated assuming a normal model as in blm.

In the case of included covariates, the function computes the covariate-specific OVL, defined as

$$\text{OVL}(\boldsymbol{x}) = \int_{-\infty}^{\infty} \min\left\{f_0(y \mid \boldsymbol{X} = \boldsymbol{x}), f_1(y \mid \boldsymbol{X} = \boldsymbol{x})\right\} dy.$$

In this implementation the integral is replaced by a numerical integration method specified in the integral function.

Value

A list object of class `ovln` containing:

`fit0`	The fitted object for the healthy group.
`fit1`	The fitted object for the diseased group.
`dens0`	The density object for the healthy group.
`dens1`	The density object for the diseased group.
`ovl`	The posterior sample of the (covariate-specific) overlap coefficient.
`newdata`	The data used for computing the covariate-specific overlap coefficient.
`criteria0`	The model comparison criteria for the healthy group.
`criteria1`	The model comparison criteria for the diseased group.

Author(s)

Javier E. Garrido Guillén

References

Samawi, H. M., Yin, J., Rochani, H. & Panchal, V. (2017), "Notes on the overlap measure as an alternative to the Youden index: How are they related?", Statistics in Medicine **36**(26), 4230–4240.
Schmid, F. & Schmidt, A. (2006), "Nonparametric estimation of the coefficient of overlapping–theory and empirical application", Computational Statistics & Data Analysis **50**(6), 1583–1596.
Weitzman, M. S. (1970), "Measures of overlap of income distributions of white and Negro families in the United States", Vol. 22, US Bureau of the Census.

See Also

`blm`, `dblm`, `criteria`

Examples

```
## Not run:
########################
## No covariates case ##
########################
## Simulate data
## Set seed for reproducibility
set.seed(123)
n<-200
y0<-rnorm(n, 0, 1)
y1<-rnorm(n, 1, 1)

## Compute the true OVL
lower<-min(qnorm(0.0001, 0, 1),
qnorm(0.0001, 1, 1))
upper<-max(qnorm(0.9999, 0, 1),
qnorm(0.9999, 1, 1))
grid<-seq(lower, upper, len = 512)
d0<-dnorm(grid, 0, 1)
d1<-dnorm(grid, 1, 1)
covlt<-trapezoidal(pmin(d0, d1), grid)

## Compute the coefficient of overlap
```

```
(res<-ovlNorm(y0 ~ 1, y1 ~ 1))

## Posterior summary
summary(res)
covlt

## Using the Simpson's rule
(res_simp<-ovlNorm(y0 ~ 1, y1 ~ 1, integral = simpson))

## Posterior summary
summary(res_simp)
covlt


######################
## Covariates case ##
######################
## True mean functions
mu1<-function(xc, xd) (0.7*sin(xc)*cos(6*xc) + 0.4)*
(1 - xd) + 0.3*exp(-2*xc + 1)*xd
mu2<-function(xc, xd) (0.5*sin(xc)*cos(8*xc) - 0.2)*
(1 - xd) + 0.4*exp(-xc^2 + 0.3*xc + 1)*xd

## True standard deviation functions
sigma1<-function(xc) 0.6*(1 - xc)^3 - 0.2*xc^2 + 0.5*xc
sigma2<-function(xc) dnorm(4*xc) + 0.2

## Create a grid to compute the true OVL
x<-seq(0, 1, len = 101)
ngrid<-201
lower0<-min(qnorm(0.0001, mu1(x, 0), sigma1(x)),
qnorm(0.0001, mu2(x, 0), sigma2(x)))
upper0<-max(qnorm(0.9999, mu1(x, 0), sigma1(x)),
qnorm(0.9999, mu2(x, 0), sigma2(x)))
lower1<-min(qnorm(0.0001, mu1(x, 1), sigma1(x)),
qnorm(0.0001, mu2(x, 1), sigma2(x)))
upper1<-max(qnorm(0.9999, mu1(x, 1), sigma1(x)),
qnorm(0.9999, mu2(x, 1), sigma2(x)))
grid<-seq(min(lower0, lower1), max(upper0, upper1),
len = ngrid)

## Compute the true OVL
covlt0<-covlt1<-vector()
for(i in 1:length(x)){
d00<-dnorm(grid, mu1(x[i], 0), sigma1(x[i]))
d10<-dnorm(grid, mu2(x[i], 0), sigma2(x[i]))
covlt0[i]<-trapezoidal(pmin(d00, d10), grid)

d01<-dnorm(grid, mu1(x[i], 1), sigma1(x[i]))
d11<-dnorm(grid, mu2(x[i], 1), sigma2(x[i]))
covlt1[i]<-trapezoidal(pmin(d01, d11), grid)
}

## True OVL data frame
truedf<-data.frame(xc = x, covlt = c(covlt0, covlt1),
xd = factor(rep(c(0, 1), each = length(x))))
```

```
## Simulate data
## Set seed for reproducibility,
## simulate 200 observations of
## a normal distribution where
## mean and variance depend on
## continuous/binary covariate
set.seed(123)
n<-200
xc0<-runif(n)
xd0<-rbinom(n, 1, 0.5)
xc1<-runif(n)
xd1<-rbinom(n, 1, 0.5)
y0<-rnorm(n, mu1(xc0, xd0), sigma1(xc0))
y1<-rnorm(n, mu2(xc1, xd1), sigma2(xc1))
mydata<-data.frame(y = c(y0, y1),
xc = c(xc0, xc1),
xd = as.factor(c(xd0, xd1)),
group = rep(c(0, 1), each = n))

## New data for the conditional densities
xcpred<-seq(0, 1, len = 11)
newdata<-expand.grid(xc = xcpred, xd = factor(c(0, 1)))

## Compute the covariate-specific OVL
(res<-ovlNorm(y ~ xc + xd,
y ~ xc + xd,
mydata, "group", 0,
newdata = newdata))

## Posterior summary
summary(res)

## End(Not run)
```

`ovlprob` *Frequentist estimator of the overlap coefficient based on the sum of error probabilities*

Description

Estimates the overlap coefficient using the sum of error probabilities representation and a Gaussian kernel density estimator.

Usage

```
ovlprob(y0, y1, grid, bw0 = "nrd0", bw1 = "nrd0",
B = 1000, alpha = 0.05)
```

Arguments

`y0`	A numeric vector containing the data from the nondiseased group.
`y1`	A numeric vector containing the data from the diseased group.

grid	A numeric vector specifying the grid where the overlap coefficient will be computed. If not supplied, a uniform grid ranging from min(y0,y1) -1 to max(y0,y1) + 1 will be used.
bw0	The smoothing bandwidth for the nondiseased group to be used. See density for details.
bw1	The smoothing bandwidth for the diseased group to be used. See density for details.
B	An integer specifying the number of bootstrap replications. The default value is set to 1,000.
alpha	A numeric value between 0 and 1 specifying the significance level used to construct a confidence interval for the overlap coefficient.

Details

Another way of thinking the coefficient of overlap is as a sum of error probabilities (Schmid & Schmidt, 2006), that is

$$\text{OVL}(Y_0, Y_1) = \Pr\left(f_0(Y_0) < f_1(Y_0)\right) + \Pr\left(f_0(Y_1) \geq f_1(Y_1)\right),$$

where f_0 and f_1 are the density functions of the nondiseased and diseased group, respectively.

Value

A list object of class ovlfreq containing the following:

OVL	A data frame containing the estimated overlap coefficient along with a (1 - alpha)*100% confidence interval.
dens0	An object of the class density resulting of the call to density for the nondiseased group.
dens1	An object of the class density resulting of the call to density for the diseased group.
var.names	A numeric vector of length 2 containing the deparsed name of the y0 and y1 arguments.

Author(s)

Javier E. Garrido Guillén

References

Schmid, F. & Schmidt, A. (2006), "Nonparametric estimation of the coefficient of overlapping–theory and empirical application", Computational Statistics & Data Analysis **50**(6), 1583-1596.

See Also

density, ovlfreq

Examples

```
## Not run:
#######################
## Simulated example ##
#######################
## Simulate data
## Set seed for reproducibility
set.seed(123)
n<-200
y0<-rnorm(n, 0, 1)
y1<-rnorm(n, 1, 1)

## Compute the true OVL
lower<-min(qnorm(0.0001, 0, 1),
qnorm(0.0001, 1, 1))
upper<-max(qnorm(0.9999, 0, 1),
qnorm(0.9999, 1, 1))
grid<-seq(lower, upper, len = 512)
d0<-dnorm(grid, 0, 1)
d1<-dnorm(grid, 1, 1)
covlt<-trapezoidal(pmin(d0, d1), grid)

## Compute the coefficient of overlap
(res<-ovlprob(y0, y1))
covlt

## End(Not run)
```

---

plot.dpm *Plot DPM density*

Description

Plots the density function of an object of class "dpm".

Usage

```
## S3 method for class 'dpm'
plot(x, ...)

## S3 method for class 'dpm'
autoplot(x, grid, title = "Density estimation",
ylab = "Density", xlab = NULL,
hist_color = "slategray3",
hist_fill = "slategray2", hist_alpha = 0.3,
lwd = 1, lty = "solid", col = "#120377",
band_color = NA, band_fill = "dodgerblue",
band_alpha = 0.3)
```

Arguments

x	A resulting object from a call to dpm.

grid	An optional numeric vector of quantiles of the response variable. If missing, a uniform grid from min(y) -1 to max(y) + 1 of length 201 is used.
title	The title of the plot.
xlab	x-axis label.
ylab	y-axis label.
hist_color	Histogram contour color.
hist_fill	Histogram fill color.
hist_alpha	Histogram opacity level.
lwd	The line width.
lty	The line type.
col	The color of the line.
band_color	Credible bands contour color.
band_fill	Credible bands fill color.
band_alpha	Credible bands opacity level.
...	Other arguments passed to ggplot.

Details

This method is only available for the no covariates case. It will plot the estimated density function (posterior mean) and 95% credible bands, superimposed on the histogram of data.

Value

A ggplot.

Author(s)

Javier E. Garrido Guillén

See Also

dpm

Examples

```
## Not run:
########################
## Simulated example ##
########################
## Simulate data
## Set seed for reproducibility
set.seed(123)
n<-200
y<-rnorm(n, 0, 1)

## Fit a Dirichlet process mixture model
fitdpm<-dpm(y ~ 1)

## Plot
plot(fitdpm)

## End(Not run)
```

plot.dpmx *Conditional mean/variance function plot*

Description

Plots the conditional mean/variance function of a DPM model or a LM.

Usage

```
## S3 method for class 'dpmx'
plot(x, ...)

## S3 method for class 'blmx'
plot(x, ...)

## S3 method for class 'dpmx'
autoplot(x, ylab = "E(y | x)",
title = "Posterior mean function", lwd = 1,
lty = "solid", col = "#120377",
band_alpha = 0.2, band_color = NA,
band_fill = "dodgerblue", size = 3,
width = 0.2, bar_color = "black")

## S3 method for class 'blmx'
autoplot(x, ylab = "E(y | x)",
title = "Posterior mean function", lwd = 1,
lty = "solid", col = "#120377",
band_alpha = 0.2, band_color = NA,
band_fill = "dodgerblue", size = 3,
width = 0.2, bar_color = "black")
```

Arguments

x	A resulting object from a call to either meandpm, vardpm or meanblm.
ylab	y-axis label.
title	The title of the plot.
lwd	The line width.
lty	The line type.
col	The color of the line.
band_alpha	Credible bands opacity level.
band_color	Credible bands contour color.
band_fill	Credible bands fill color.
size	Size of the points.
width	Error bar width.
bar_color	Error bar color.
...	Other plotting arguments.

Details

This method is only available for a single continuous covariate and/or a single binary covariate. It will plot the conditional mean ("dpmx" or "blmx" object) or variance ("dpmx" object) function (posterior mean) and 95% credible bands.

Value

A ggplot.

Author(s)

Javier E. Garrido Guillén

See Also

meandpm, vardpm, meanblm

Examples

```
## Not run:
####################################
## Only a single binary covariate ##
####################################
## True mean function
funcmean<-function(x) 0.6 - 0.8*x

## True standard deviation function
funcsd<-function(x) 0.3*(1 - x) + 0.6

## True functions data frame
x<-c(0, 1)
dfmt<-data.frame(x = factor(x),
truemean = funcmean(x),
truevar = funcsd(x)^2)

## Simulate data
## Set seed for reproducibility
## Simulate 200 observartions of
## a normal distribution, where
## the mean depends on a binary
## covariate (ANOVA-type model)
set.seed(123)
n<-200
x<-rbinom(n, 1, 0.5)
y<-rnorm(n, funcmean(x), funcsd(x))

## Create a data frame with the data
mydata<-data.frame(y = y, x = factor(x))

## Create a data frame to compute the
## conditional mean/variance function
newdata<-data.frame(x = factor(c(0, 1)))

## Fit a Dirichlet process mixture model
fitdpm<-dpm(y ~ x, mydata)
```

```
## Compute the mean function
meanfunc<-meandpm(fitdpm, newdata)
plot(meanfunc)

## Load ggplot2
require(ggplot2)

## Add the true mean
plot(meanfunc) +
geom_point(aes(x = x, y = truemean),
data = dfmt,
size = 3, color = "dodgerblue")

## Compute the variance function
varfunc<-vardpm(fitdpm, newdata)
plot(varfunc,
title =  "Posterior variance function",
ylab = "Variance")

## Add the true variance
plot(varfunc,
title =  "Posterior variance function",
ylab = "Variance") +
geom_point(aes(x = x, y = truevar),
data = dfmt, size = 3,
color = "dodgerblue")


#####################################
## Continuous and binary covariate ##
#####################################
## True mean function
mu<-function(xc, xd) (0.7*sin(xc)*cos(6*xc) + 0.4)*
(1 - xd) + 0.3*exp(-2*xc + 1)*xd

## True standard deviation function
sigma<-function(xc) 0.6*(1 - xc)^3 - 0.2*xc^2 + 0.5*xc

## Simulate data
## Set seed for reproducibility,
## simulate 200 observations of
## a normal distribution where
## mean and variance depend on
## continuous/binary covariate
n<-200
xc<-runif(n)
xd<-rbinom(n, 1, 0.5)
y<-rnorm(n, mu(xc, xd), sigma(xc))

## Create data frame with data
## NOTE: binary covariates must
## be defined as factors
mydata<-data.frame(y = y,
xc = xc,
xd = factor(xd))

## Fit a Dirichlet process mixture model
```

```
fitdpm<-dpm(y ~ xc + xd, mydata)

## New data for the conditional mean/variance function
xcpred<-seq(0, 1, len = 21)
xdpred<-factor(c(0, 1))
newdata<-expand.grid(xc = xcpred, xd = xdpred)

## Data frame of the true functions
truedf<-data.frame(xc = xcpred,
truemean = c(mu(xcpred, 0), mu(xcpred, 1)),
truevar = sigma(xcpred)^2,
xd = rep(c(0, 1), each = length(xcpred)))

## Compute the mean function
meanfunc<-meandpm(fitdpm, newdata)

## Load ggplot2
require(ggplot2)

## Plot the (posterior mean) mean function,
## the true function and the data
plot(meanfunc) +
geom_point(data = mydata,
mapping = aes(x = xc,  y = y),
alpha = 0.5, col = "slategray2") +
geom_line(data = truedf,
mapping = aes(x = xc, y = truemean),
lwd = 1, color = "red", lty = 2) +
facet_wrap(~ xd)

## Compute the variance function
varfunc<-vardpm(fitdpm, newdata)

## Load ggplot2
require(ggplot2)

## Plot the (posterior mean) variance
## function and the true function
plot(varfunc,
title = "Posterior variance function",
ylab = "Variance") +
geom_line(data = truedf,
mapping = aes(x = xc, y = truevar),
lwd = 1, color = "red", lty = 2)


## Using a linear regression model
fitlm<-blm(y ~ xc + xd, mydata)

## Compute the mean function
meanfunc<-meanblm(fitlm, newdata)

## Load ggplot2
require(ggplot2)

## Plot the (posterior mean) mean function,
## the true function and the data
```

```
plot(meanfunc) +
geom_point(data = mydata,
mapping = aes(x = xc,  y = y),
alpha = 0.5, col = "slategray2") +
geom_line(data = truedf,
mapping = aes(x = xc, y = truemean),
lwd = 1, color = "red", lty = 2) +
facet_wrap(~ xd)

## End(Not run)
```

`plot.ovl` *Overlap coefficient plot*

Description

Plots the (covariate-specific) overlap coefficient.

Usage

```
## S3 method for class 'ovl'
plot(x, ...)

## S3 method for class 'ovln'
plot(x, ...)

## S3 method for class 'ovl'
autoplot(x, xlab = NULL, ylab = "Density",
title = "Overlap coefficient", lwd = 1,
lty = "solid", legend_title = "Group",
group_color = c("red4", "#120377"),
group_labs = c("Diseased", "Healthy"),
band_color = NA, band_fill = "gray45",
band_alpha = 0.3, ovl_color = "black",
size = 3, width = 0.2, bar_color = "black")

## S3 method for class 'ovln'
autoplot(x, xlab = NULL, ylab = "Density",
title = "Overlap coefficient", lwd = 1,
lty = "solid", legend_title = "Group",
group_color = c("red4", "#120377"),
group_labs = c("Diseased", "Healthy"),
band_color = NA, band_fill = "gray45",
band_alpha = 0.3, ovl_color = "black",
size = 3, width = 0.2, bar_color = "black")
```

Arguments

x	A resulting object from a call to ovl, ovlBB or ovlNorm.
xlab	x-axis label.
ylab	y-axis label.

title	The title of the plot.
lwd	The line width.
lty	The line type.
legend_title	The title of the legend.
group_color	A vector of length two representing the line colors.
group_labs	A vector of length two representing the group labels.
band_color	Credible bands contour color.
band_fill	Credible bands fill color.
band_alpha	Credible bands opacity level.
ovl_color	Overlap coefficient line color.
size	Size of the points.
width	Error bar width.
bar_color	Error bar color.
...	Other plotting arguments.

Details

This method is only available for a single continuous covariate and/or a single binary covariate. In the case without covariates, it will plot the density functions along the overlap coefficient. In the covariates case, it will plot the covariate-specific overlap coefficient (posterior mean) and 95% pointwise credible bands.

Value

A ggplot.

Author(s)

Javier E. Garrido Guillén

See Also

ovl, ovlBB, ovlNorm

Examples

```
## Not run:
########################
## No covariates case ##
########################
## Simulate data
## Set seed for reproducibility
set.seed(123)
n<-200
y0<-rnorm(n, 0, 1)
y1<-rnorm(n, 1, 1)

## Compute the true OVL
lower<-min(qnorm(0.0001, 0, 1),
qnorm(0.0001, 1, 1))
upper<-max(qnorm(0.9999, 0, 1),
```

```
qnorm(0.9999, 1, 1))
grid<-seq(lower, upper, len = 512)
d0<-dnorm(grid, 0, 1)
d1<-dnorm(grid, 1, 1)
covlt<-trapezoidal(pmin(d0, d1), grid)

## Compute the coefficient of overlap
res<-ovl(y0 ~ 1, y1 ~ 1)

## Plot method
plot(res, xlab = "y")

## Using the BB-estimator
mydata<-data.frame(y = c(y0, y1),
group = rep(c(0, 1), each = n))
resBB<-ovlBB("y", mydata, "group", 0)

## Plot method
plot(resBB)

## Using a LM
reslm<-ovlNorm(y0 ~ 1, y1 ~ 1)

## Plot method
plot(reslm, xlab = "y")


################################
## A single binary covariate ##
################################
## True means
mu1<-function(x) 0.6 - 0.8*x
mu2<-1.3

## True standard deviations
sigma1<-function(x) 0.3*(1 - x) + 0.6
sigma2<-0.5

## Compute the true OVL
x<-c(0, 1)
lower<-min(qnorm(0.0001, mu1(x), sigma1(x)),
qnorm(0.0001, mu2, sigma2))
upper<-max(qnorm(0.9999, mu1(x), sigma1(x)),
qnorm(0.9999, mu2, sigma2))
grid<-seq(lower, upper, len = 512)
covlt<-vector()
for(i in 1:length(x)){
d0<-dnorm(grid, mu1(x[i]), sigma1(x[i]))
d1<-dnorm(grid, mu2, sigma2)
covlt[i]<-trapezoidal(pmin(d0, d1), grid)
}

## True OVL data frame
dft<-data.frame(x = x, covlt = covlt)

## Simulate data
## Set seed for reproducibility
```

```
## Simulate 200 observartions of
## a normal distribution, where
## the mean depends on a binary
## covariate (ANOVA-type model)
set.seed(123)
n<-200
x0<-rbinom(n, 1, 0.5)
y0<-rnorm(n, mu1(x0), sigma1(x0))
y1<-rnorm(n, mu2, sigma2)
mydata<-data.frame(y = c(y0, y1),
x = factor(c(x0, rep(NA, n))),
group = rep(c(0, 1), each = n))

## New data for conditional densities
newdata<-data.frame(x = factor(c(0, 1)))

## Compute the OVL
res<-ovl(y ~ x, y ~ 1, mydata, "group", 0,
newdata = newdata)

## Load ggplot2
require(ggplot2)

## Plot the posterior mean estimate
## of the OVL and true OVL
plot(res) +
geom_point(aes(x = x, y = covlt),
size = 3,
color = "#D600D6")


#####################################
## Continuous and binary covariate ##
#####################################
## True mean functions
mu1<-function(xc, xd) (0.7*sin(xc)*cos(6*xc) + 0.4)*
(1 - xd) + 0.3*exp(-2*xc + 1)*xd
mu2<-function(xc, xd) (0.5*sin(xc)*cos(8*xc) - 0.2)*
(1 - xd) + 0.4*exp(-xc^2 + 0.3*xc + 1)*xd

## True standard deviation functions
sigma1<-function(xc) 0.6*(1 - xc)^3 - 0.2*xc^2 + 0.5*xc
sigma2<-function(xc) dnorm(4*xc) + 0.2

## Create a grid to compute the true OVL
x<-seq(0, 1, len = 101)
ngrid<-201
lower0<-min(qnorm(0.0001, mu1(x, 0), sigma1(x)),
qnorm(0.0001, mu2(x, 0), sigma2(x)))
upper0<-max(qnorm(0.9999, mu1(x, 0), sigma1(x)),
qnorm(0.9999, mu2(x, 0), sigma2(x)))
lower1<-min(qnorm(0.0001, mu1(x, 1), sigma1(x)),
qnorm(0.0001, mu2(x, 1), sigma2(x)))
upper1<-max(qnorm(0.9999, mu1(x, 1), sigma1(x)),
qnorm(0.9999, mu2(x, 1), sigma2(x)))
grid<-seq(min(lower0, lower1), max(upper0, upper1),
len = ngrid)
```

```
## Compute the true OVL
covlt0<-covlt1<-vector()
for(i in 1:length(x)){
d00<-dnorm(grid, mu1(x[i], 0), sigma1(x[i]))
d10<-dnorm(grid, mu2(x[i], 0), sigma2(x[i]))
covlt0[i]<-trapezoidal(pmin(d00, d10), grid)

d01<-dnorm(grid, mu1(x[i], 1), sigma1(x[i]))
d11<-dnorm(grid, mu2(x[i], 1), sigma2(x[i]))
covlt1[i]<-trapezoidal(pmin(d01, d11), grid)
}

## True OVL data frame
truedf<-data.frame(xc = x, covlt = c(covlt0, covlt1),
xd = factor(rep(c(0, 1), each = length(x))))

## Simulate data
## Set seed for reproducibility,
## simulate 200 observations of
## a normal distribution where
## mean and variance depend on
## continuous/binary covariate
set.seed(123)
n<-200
xc0<-runif(n)
xd0<-rbinom(n, 1, 0.5)
xc1<-runif(n)
xd1<-rbinom(n, 1, 0.5)
y0<-rnorm(n, mu1(xc0, xd0), sigma1(xc0))
y1<-rnorm(n, mu2(xc1, xd1), sigma2(xc1))
mydata<-data.frame(y = c(y0, y1),
xc = c(xc0, xc1),
xd = as.factor(c(xd0, xd1)),
group = rep(c(0, 1), each = n))

## New data for the conditional densities
xcpred<-seq(0, 1, len = 11)
newdata<-expand.grid(xc = xcpred, xd = factor(c(0, 1)))

## Compute the covariate-specific OVL
res<-ovl(y ~ xc + xd,
y ~ xc + xd,
mydata, "group", 0,
newdata = newdata)

## Load ggplot2
require(ggplot2)

## Plot the posterior estimate
## of the covariate-specific OVL
## and the true OVL curve
plot(res) +
geom_line(aes(x = xc, y = covlt),
data = truedf,
lwd = 1, linetype = "22",
color = "#D600D6") +
```

```
facet_wrap(~ xd)


## Using a LM to compute the covariate-specific OVL
reslm<-ovlNorm(y ~ xc + xd,
y ~ xc + xd,
mydata, "group", 0,
newdata = newdata)

## Plot the posterior estimate
## of the covariate-specific OVL
## and the true OVL curve
plot(reslm) +
geom_line(aes(x = xc, y = covlt),
data = truedf,
lwd = 1, linetype = "22",
color = "#D600D6") +
facet_wrap(~ xd)

## End(Not run)
```

`print.dpm` *Print methods*

Description

Prints different values depending on the class of the object ("dpm", "blm", "ovl", "ovln" or "ovlfreq").

Usage

```
## S3 method for class 'dpm'
print(x, ...)

## S3 method for class 'blm'
print(x, ...)

## S3 method for class 'ovl'
print(x, ...)

## S3 method for class 'ovln'
print(x, ...)

## S3 method for class 'ovlfreq'
print(x, ...)
```

Arguments

x	A resulting object of a call to either dpm, blm, ovl, ovlBB, ovlNorm, ovlfreq or ovlprob.
...	Not other arguments used.

Details

This function returns a summary of the call to one of the functions mentioned before.

Author(s)

Javier E. Garrido Guillén

See Also

dpm, blm, ovl, ovlBB, ovlNorm, ovlfreq, ovlprob

`simpson` *Numerical integration using the Simpson's rule*

Description

Performs a numerical integration using the Simpson's rule.

Usage

```
simpson(f, grid)
```

Arguments

`f`	A numeric vector containing the function evaluated at every single point of the supplied grid.
`grid`	A numeric vector containing a grid of the interval where the integral will be computed on.

Value

The function returns the value of the integral in the supplied interval.

Author(s)

Vanda Inácio, Javier E. Garrido Guillén

Examples

```
## Set a grid of the interval
grid<-seq(-4, 4, len = 201)

## Evaluate the function in such grid
f<-dnorm(grid)

## Compute the integral
simpson(f, grid)
```

summary.ovl *Summary methods*

Description

Summary methods for objects of classes "blm", "ovl" and "ovln".

Usage

```
## S3 method for class 'ovl'
summary(object, ...)

## S3 method for class 'blm'
summary(object, ...)

## S3 method for class 'ovln'
summary(object, ...)
```

Arguments

object	A resulting object from a call to ovl, blm or ovlNorm.
...	Not other arguments used.

Details

This function returns the posterior mean of the (covariate-specific) overlap coefficient (for objects of class "ovl" or "ovln") or the posterior mean of all parameters of the LM (for objects of class "blm") along with symmetric credible intervals using the 0.25 and 0.975 quantiles.

Value

A list containing:

posterior	A data frame containing the posterior mean of the (covariate-specific) overlap coefficient along with 95% credible (pointwise) bands.

Author(s)

Javier E. Garrido Guillén

See Also

ovl, blm, ovlNorm

trapezoidal *Numerical integration using the trapezoidal rule*

Description

Performs a numerical integration using the trapezoidal rule.

Usage

```
trapezoidal(f, grid)
```

Arguments

f	A numeric vector containing the function evaluated at every single point of the supplied grid.
grid	A numeric vector containing a grid of the interval where the integral will be computed on.

Value

The function returns the value of the integral in the supplied interval.

Author(s)

Javier E. Garrido Guillén

Examples

```
## Set a grid of the interval
grid<-seq(-4, 4, len = 201)

## Evaluate the function in such grid
f<-dnorm(grid)

## Compute the integral
trapezoidal(f, grid)
```

vardpm *Variance function for DPM models*

Description

Returns the variance function of a DPM model.

Usage

```
vardpm(fit, x, silent = FALSE)
```

Arguments

fit	A resulting object from a call to dpm.
x	A data frame containing the values of all the covariates included in the fit model.
silent	An optional logical argument specifying whether a progress bar should be printed or not. The default value is FALSE, then a progress bar is printed.

Details

This function computes the conditional variance function.

Value

A list object of class dpmx containing

x	The original data.
posterior	A numeric matrix containing the posterior variance function sample.
xc	The continuous covariates.
xd	The binary covariates.

Author(s)

Javier E. Garrido Guillén

See Also

dpm, meandpm

Examples

```
## Not run:
########################
## Simulated example ##
########################
## True mean function
mu<-function(xc, xd) (0.7*sin(xc)*cos(6*xc) + 0.4)*
(1 - xd) + 0.3*exp(-2*xc + 1)*xd

## True standard deviation function
sigma<-function(xc) 0.6*(1 - xc)^3 - 0.2*xc^2 + 0.5*xc

## Simulate data
## Set seed for reproducibility,
## simulate 200 observations of
## a normal distribution where
## mean and variance depend on
## continuous/binary covariate
n<-200
xc<-runif(n)
xd<-rbinom(n, 1, 0.5)
y<-rnorm(n, mu(xc, xd), sigma(xc))

## Create data frame with data
## NOTE: binary covariates must
```

```
## be defined as factors
mydata<-data.frame(y = y,
xc = xc,
xd = factor(xd))

## Fit a Dirichlet process mixture model
fitdpm<-dpm(y ~ xc + xd, mydata)

## New data for the conditional mean/variance function
xcpred<-seq(0, 1, len = 21)
xdpred<-factor(c(0, 1))
newdata<-expand.grid(xc = xcpred, xd = xdpred)

## Data frame of the true functions
truedf<-data.frame(xc = xcpred,
truemean = c(mu(xcpred, 0), mu(xcpred, 1)),
truevar = sigma(xcpred)^2,
xd = rep(c(0, 1), each = length(xcpred)))

## Compute the variance function
varfunc<-vardpm(fitdpm, newdata)

## Load ggplot2
require(ggplot2)

## Plot the (posterior mean) mean function,
## the true function and the data
plot(varfunc, title = "Posterior variance function",
ylab = "Variance") +
geom_line(data = truedf,
mapping = aes(x = xc, y = truevar),
lwd = 1, color = "red", lty = 2) +
facet_wrap(~ xd)


##########################################
## Another example, using simulated data ##
##########################################
## Set a seed for reproducibility
set.seed(123)

## Simulate 500 observations from
## a mixture of normals, where the
## weights depend on a continuous
## covariate (example taken from
## Dunson et al., 2007)
n<-500
x<-runif(n)
y1<-x + rnorm(n, 0, sqrt(0.01))
y2<-x^4 + rnorm(n, 0, sqrt(0.04))
u<-runif(n)
prob<-exp(-2*x)
y<-ifelse(u < prob, y1, y2)

## Fit a linear regression model,
## changing prior and number of
## MCMC iterations
```

```
fitdpm<-dpm(y ~ x,
mcmc = list(nsave = 10000, nburn = 5000, nthin = 1))

## New data for the conditional mean/variance function
xpred<-seq(0, 1, len = 21)
newdata<-data.frame(x = xpred)

## True mean and variance
truemean<-exp(-2*xpred)*xpred +
(1 - exp(-2*xpred))*xpred^4
truevar<-exp(-2*xpred)*(0.01 + xpred^2) +
(1 - exp(-2*xpred))*(0.04 + xpred^8) - truemean^2

## Data frame of the true functions
truedf<-data.frame(x = xpred,
truemean = truemean,
truevar = truevar)

## Compute the variance function
varfunc<-vardpm(fitdpm, newdata)

## Load ggplot2
require(ggplot2)

## Plot the (posterior mean) variance
## function and the true function
plot(varfunc, title = "Posterior variance function",
ylab = "Variance") +
geom_line(data = truedf,
mapping = aes(x = x, y = truevar),
lwd = 1, color = "red", lty = 2)

## End(Not run)
```

Index

Chapter 6

Future work

An interesting question that arose from the recent application of our methods to Alzheimer's disease data was to consider the discriminatory ability of multiple biomarkers. This means that we could extend the definition of the covariate-specific OVL to study the joint discriminatory ability of various diagnostic tests or biomarkers. We discuss it in detail below.

In many practical applications, there are two (or more) biomarkers available for a particular disease (as in the Alzheimer's disease example, studied by Welge et al. (2009) as well) and we wonder about their joint discriminatory ability.

Let $\boldsymbol{Y}_{\bar{D}} = (Y_{\bar{D}1}, Y_{\bar{D}2})$ and $\boldsymbol{Y}_D = (Y_{D1}, Y_{D2})$ two random vectors that represent the test outcomes in the nondiseased and diseased population with joint density functions given by $f_{\bar{D}}(\cdot)$ and $f_D(\cdot)$, respectively. Then, we can define the bivariate overlap coefficient as follows

$$\text{OVL}(\boldsymbol{y}_{\bar{D}}, \boldsymbol{y}_D) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \min \left\{ f_{\bar{D}}(y_1, y_2), f_D(y_1, y_2) \right\} dy_1 dy_2. \tag{6.1}$$

We can model the joint density functions using the approach described in Chapter 4. The double integral may be replaced by a numerical method, such as the trapezoidal rule, leading to the following

estimator

$$
\begin{aligned}
\text{OVL} &\approx \int_{-\infty}^{\infty} \left\{ \boldsymbol{\Delta} y_1 \left(\sum_{k=1}^{N-1} g\left(\tilde{y}_{1k}, y_2\right) + \frac{g\left(\tilde{y}_{1N}, y_2\right) + g\left(\tilde{y}_{10}, y_2\right)}{2} \right) \right\} dy_2 \\
&= \boldsymbol{\Delta} y_1 \left(\sum_{k=1}^{N-1} \int_{-\infty}^{\infty} g\left(\tilde{y}_{1k}, y_2\right) dy_2 + \int_{-\infty}^{\infty} \frac{g\left(\tilde{y}_{1N}, y_2\right) + g\left(\tilde{y}_{10}, y_2\right)}{2} dy_2 \right) \\
&\approx \boldsymbol{\Delta} y_1 \Bigg(\sum_{k=1}^{N-1} \boldsymbol{\Delta} y_2 \left[\sum_{m=1}^{M-1} g\left(\tilde{y}_{1k}, \tilde{y}_{2m}\right) + \frac{g\left(\tilde{y}_{1k}, \tilde{y}_{2M}\right) + g\left(\tilde{y}_{1k}, \tilde{y}_{20}\right)}{2} \right] \\
&\quad + \frac{\boldsymbol{\Delta} y_2}{2} \Bigg[\sum_{m=1}^{M-1} \left\{ g\left(\tilde{y}_{1N}, \tilde{y}_{2m}\right) + g\left(\tilde{y}_{10}, \tilde{y}_{2m}\right) \right\} + \frac{g\left(\tilde{y}_{1N}, \tilde{y}_{2M}\right) + g\left(\tilde{y}_{1N}, \tilde{y}_{20}\right)}{2} \\
&\quad + \frac{g\left(\tilde{y}_{10}, \tilde{y}_{2M}\right) + g\left(\tilde{y}_{10}, \tilde{y}_{20}\right)}{2} \Bigg] \Bigg) \\
&= \boldsymbol{\Delta} y_1 \Bigg(\sum_{k=1}^{N-1} \boldsymbol{\Delta} y_2 \Bigg[\sum_{m=1}^{M-1} g\left(\tilde{y}_{1k}, \tilde{y}_{2m}\right) + \frac{g\left(\tilde{y}_{1k}, \tilde{y}_{2M}\right) + g\left(\tilde{y}_{1k}, \tilde{y}_{20}\right)}{2} \\
&\quad + \sum_{m=1}^{M-1} \frac{g\left(\tilde{y}_{1N}, \tilde{y}_{2m}\right) + g\left(\tilde{y}_{10}, \tilde{y}_{2m}\right)}{2} + \frac{g\left(\tilde{y}_{1N}, \tilde{y}_{2M}\right) + g\left(\tilde{y}_{1N}, \tilde{y}_{20}\right)}{4} \\
&\quad + \frac{g\left(\tilde{y}_{10}, \tilde{y}_{2M}\right) + g\left(\tilde{y}_{10}, \tilde{y}_{20}\right)}{4} \Bigg] \Bigg) \\
&= \boldsymbol{\Delta} y_1 \boldsymbol{\Delta} y_2 \Bigg(\sum_{k=1}^{N-1} \sum_{m=1}^{M-1} g\left(\tilde{y}_{1k}, \tilde{y}_{2m}\right) + \frac{g\left(\tilde{y}_{1k}, \tilde{y}_{2M}\right) + g\left(\tilde{y}_{1k}, \tilde{y}_{20}\right)}{2} + \frac{g\left(\tilde{y}_{1N}, \tilde{y}_{2m}\right) + g\left(\tilde{y}_{10}, \tilde{y}_{2m}\right)}{2} \\
&\quad + \frac{g\left(\tilde{y}_{1N}, \tilde{y}_{2M}\right) + g\left(\tilde{y}_{1N}, \tilde{y}_{20}\right)}{4} + \frac{g\left(\tilde{y}_{10}, \tilde{y}_{2M}\right) + g\left(\tilde{y}_{10}, \tilde{y}_{20}\right)}{4} \Bigg),
\end{aligned}
$$

where $\min\left\{y_{\bar{D}1}, y_{D1}\right\} = \tilde{y}_{10} < \cdots < \tilde{y}_{1N} = \max\left\{y_{\bar{D}1}, y_{D1}\right\}$, $\min\left\{y_{\bar{D}2}, y_{D2}\right\} = \tilde{y}_{20} < \cdots < \tilde{y}_{2M} = \max\left\{y_{\bar{D}2}, y_{D2}\right\}$, $g(\cdot) = \min\left\{f_{\bar{D}}(\cdot), f_D(\cdot)\right\}$, $\boldsymbol{\Delta} y_1 = \frac{\tilde{y}_{1N} - \tilde{y}_{10}}{N}$ and $\boldsymbol{\Delta} y_2 = \frac{\tilde{y}_{2M} - \tilde{y}_{20}}{M}$.

A natural extension of (6.1) to the conditional setting is given by

$$
\text{OVL}(\boldsymbol{x}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \min\left\{ f_{\bar{D}}(y_1, y_2 \mid \boldsymbol{X}_{\bar{D}} = \boldsymbol{x}), f_D(y_1, y_2 \mid \boldsymbol{X}_D = \boldsymbol{x}) \right\} dy_1 dy_2,
$$

where $\boldsymbol{x}$ is a vector of covariates (continuous and/or binary). Note that the estimation of the densities is straightforward even in the covariate-specific case, as the methodology described in Chapter 4 can be easily adapted.

Appendices

Appendix A

Inference for the coefficient of overlap based on P-splines and Dirichlet process mixtures

In this section we provide our first approach to conduct inference for the covariate-specific coefficient of overlap. We proposed a conditional density approach based on an additive formulation for the mean of the test outcomes. For this aim, we assume a regression model for the test outcomes where the effect of the continuous covariates is modelled through a linear combination of unknown smooth functions and further covariates may be included in the model in form of an usual linear predictor. Additionally, to allow for more flexibility, we assume that the error follows a Dirichlet process mixture (DPM) of normal distributions. We must emphasize that this methodology was not included in the main part of the thesis, because we experienced issues with the identifiability of the model when interactions between categorical and continuous covariates were included, which we could not fully understand. For this reason, the results presented throughout this chapter must be treated as preliminary. Further research is needed.

A.1 Introduction

In many situations the behaviour of a diagnostic test may depend on individuals' covariates. For example, a diagnostic test might have a good discriminatory ability in women, but it might perform poorly in men. Therefore, failure to incorporate this information may result in misleading or oversimplified conclusions about the accuracy of the test.

Similarly to the no covariates case, several statical methods have been developed to help determining more precisely the differences between the underlying distributions of the test outcomes in the diseased and nondiseased population. A detailed discussion can be found in Section 1.5.

Recently, the overlap coefficient (OVL), defined as the proportion of overlap area between two density functions, has been proposed as an alternative summary measure of diagnostic accuracy. An OVL value of zero means that the distributions do not overlap at all (perfect diagnostic accuracy), whereas a value of one means that the distributions are identical and thus, the test is useless from a diagnostic viewpoint. Let $Y_{\bar{D}}$ and Y_D be two independent continuous random variables representing the test outcomes in the nondiseased and diseased population, with p–dimensional covariate vectors $\boldsymbol{X}_{\bar{D}}$ and $\boldsymbol{X}_D$ and conditional density functions given by $f_{\bar{D}}(\cdot \mid \boldsymbol{X}_{\bar{D}} = \boldsymbol{x}_{\bar{D}})$ and $f_D(\cdot \mid \boldsymbol{X}_D = \boldsymbol{x}_D)$, respectively. For simplicity, we assume that the covariates are the same in both groups. Thus, given a covariate value $\boldsymbol{x}$, the covariate-specific overlap coefficient is defined as

$$\mathrm{OVL}(\boldsymbol{x}) = \int_{-\infty}^{\infty} \min\left\{f_{\bar{D}}(y \mid \boldsymbol{X}_{\bar{D}} = \boldsymbol{x}), f_D(y \mid \boldsymbol{X}_D = \boldsymbol{x})\right\} dy. \tag{A.1}$$

Note that for each possible value of the covariates $\boldsymbol{x}$, we might get a different OVL. The covariate-specific OVL inherits all the properties of the the unconditional case, that is, it is invariant under strictly increasing transformations, it takes into account both location and shape of the underlying distributions and it is non-directional. The results are straightforward since they follow from the no covariates case, conditioning on a fixed value of the covariates (see Schmid and Schmidt, 2006 for proofs on the unconditional case).

Our focus in this work is to develop methods to flexibly estimate the overlap coefficient in the covariate-specific context. Since the OVL depends on the underlying conditional density functions, the

problem of estimating it, reduces to the problem of (accurately) estimating the underlying conditional densities. In many practical situations, the outcomes of a medical diagnostic test may present a complex structure which can be difficult to capture with a single parametric model (e.g., multimodality, skewness, excess of kutosis, etc.). Furthermore, this behaviour might be affected by the presence of covariates whose effect might be non-linear. For those reasons, we propose to model the conditional density of the test outcomes in each group as a DPM of normal distributions, where the mean of each component is modelled as a linear combination of smooth (unknown) functions modelled via P-splines (Eilers and Marx, 1996; Lang and Brezger, 2004). Further, by working under a Bayesian context, point and interval estimates are obtained into a single integrated framework.

A.2 Methods

Let $\{(y_{\bar{D}i}, \boldsymbol{x}'_{\bar{D}i}, \boldsymbol{v}'_{\bar{D}i})\}_{i=1}^{n_{\bar{D}}}$ and $\{(y_{Dj}, \boldsymbol{x}'_{Dj}, \boldsymbol{v}'_{Dj})\}_{j=1}^{n_D}$ be independent random samples of size $n_{\bar{D}}$ and n_D from the nondiseased and diseased population, respectively. We assume that the continuous covariates $\boldsymbol{x}_{\bar{D}i} = \left(x_{\bar{D}i,1}, \ldots, x_{\bar{D}i,p}\right)'$ and $\boldsymbol{x}_{Dj} = \left(x_{Dj,1}, \ldots, x_{Dj,p}\right)'$ have a non-linear effect on the test outcomes; and $\boldsymbol{v}_{\bar{D}i} = \left(v_{\bar{D}i,1}, \ldots, v_{\bar{D}i,q}\right)'$ and $\boldsymbol{v}_{Dj} = \left(v_{Dj,1}, \ldots, v_{Dj,q}\right)'$ are two q–dimensional vectors of further covariates (categorical/continuous), including the intercept, whose effect is assumed to be linear. For the sake of simplicity and to ease notation, we will assume that the covariates are the same in both groups. However, our modelling approach can easily deal with different covariates between groups. For instance, consider the severity of the disease, it is a specific variable that could be measured within the diseased group, but which is senseless within the nondiseased. In what follows, we will describe our modelling procedure only for the diseased population as a similar one is applicable to the nondiseased group. We assume the following regression model for the test outcomes in the diseased group

$$y_{Dj} = h_{D1}(x_{Dj,1}) + \cdots + h_{Dp}(x_{Dj,p}) + \boldsymbol{v}'_{Dj}\boldsymbol{\gamma}_D + \epsilon_{Dj}, \quad j = 1, \ldots, n_D,$$

where $h_{D1}, \ldots, h_{Dp}$ are unknown smooth functions of the covariates $\boldsymbol{x}_{Dj}$, $\boldsymbol{\gamma}_D = (\gamma_{D1}, \ldots, \gamma_{Dq})'$ is the vector of coefficients of the parametric part of the predictor and the error ϵ_{Dj} follows a DPM of normal distributions (Escobar and West, 1995; Richardson and Green, 1997). This setting induces the

following conditional density for the test outcomes

$$f(y_{Dj} \mid \boldsymbol{X}_D = \boldsymbol{x}_{Dj}, \boldsymbol{V}_D = \boldsymbol{v}_{Dj}) = \int \phi(y_{Dj} \mid \eta_{Dj} + \mu, \sigma^2) \mathrm{d}G_D(\mu, \sigma^2), \quad G_D \sim \mathrm{DP}(\alpha_D, G_D^*), \tag{A.2}$$

where $\phi(\cdot \mid \mu, \sigma^2)$ is the density function of a normal distribution with mean μ and variance σ^2, $\eta_{Dj} = h_{D1}(x_{Dj,1}) + \cdots + h_{Dp}(x_{Dj,p}) + \boldsymbol{v}_{Dj}^{'}\boldsymbol{\gamma}_D$ is the predictor, and the mixing distribution G_D follows a Dirichlet process (DP) (Ferguson, 1973) with concentration parameter $\alpha_D > 0$ and baseline distribution G_D^*. Since $\mathbb{E}(G_D(\mu, \sigma^2)) = G_D^*(\mu, \sigma^2)$, the designation of G_D^* as baseline distribution can be regarded as our best guess for G_D, with α_D controlling how certain we are about such guess. Larger values of α_D imply realizations of G_D that are closer to G_D^*. Then, using the stick-breaking construction given by Sethuraman (1994), we can express G_D as follows

$$G_D(\cdot) = \sum_{l=1}^{\infty} \omega_{Dl} \delta_{(\mu_{Dl}, \sigma_{Dl}^2)}(\cdot), \quad \omega_{Dl} = V_{Dl} \prod_{t<l} (1 - V_{Dt}),$$

where δ_a denotes a point mass at a, $(\mu_{Dl}, \sigma_{Dl}^2) \overset{\text{iid}}{\sim} G_D^*$ and $V_{Dl} \overset{\text{idd}}{\sim} \mathrm{Beta}(1, \alpha_D)$ are mutually independent, for $l \geq 1$. For conjugacy reasons, we set a normal-inverse gamma distribution on G_D^*, that is, $G_D^*(\mu, \sigma^2) \equiv \mathrm{N}(\mu \mid 0, b_{\mu_D}^2)\mathrm{IG}(\sigma^2 \mid a_{\sigma_D^2}, b_{\sigma_D^2})$. Note that we need to set the prior mean of μ to zero since we have included an intercept in the parametric part of the model. Citing Chib and Greenberg (2010), μ can be viewed as a random effect and thus, setting the prior mean of the random effect to zero is sufficient to identify the intercept. Thus, we can rewrite the model in (A.2) as

$$f(y_{Dj} \mid \boldsymbol{X}_D = \boldsymbol{x}_{Dj}, \boldsymbol{V}_D = \boldsymbol{v}_{Dj}) = \sum_{l=1}^{\infty} \omega_{Dl} \phi(y_{Dj} \mid \eta_{Dj} + \mu_{Dl}, \sigma_{Dl}^2).$$

Regarding the specification of the predictor η_{Dj}, we propose to approximate each smooth function h_{Dk}, by a cubic B-splines basis with equally spaced knots $x_{Dk,\min} = \xi_{Dk,0} < \cdots < \xi_{Dk,m} = x_{Dk,\max}$. Thus, we can write each of them as a linear combination of $R = m + 3$ B-splines basis functions $B_{Dk,r}$, that is,

$$h_{Dk}(x_{Dj,k}) = \sum_{r=1}^{R} \beta_{Dk,r} B_{Dk,r}(x_{Dj,k}),$$

where $\boldsymbol{\beta}_{Dk} = (\beta_{Dk,1}, \ldots, \beta_{Dk,R})^{'}$ is the corresponding vector of coefficients. Let us denote by $\boldsymbol{B}_{Dk}$ the $(n_D \times R)$ design matrix for each covariate, where the element in row j and column r is given by

$\boldsymbol{B}_{Dk,jr} = B_{Dk,r}(x_{Dj,k})$, for $k = 1, \ldots, p$. For simplicity of notation, we will assume the same number of knots and hence, the same number of basis functions for every function h_{Dk}. Further, note that the mean of every function is not identifiable, because the mean level of each of them is not identified. Hence, we must impose constraints to ensure identifiability. We will use sum-to-zero constraints, which according to (Wood, 2017, p. 175) are the best constraints, more precisely

$$\sum_{j=1}^{n_D} h_{Dk}(x_{Dj,k}) = 0, \quad k = 1, \ldots, p.$$

For this end, we will subtract from each column of the design matrix $\boldsymbol{B}_{Dk}$, its mean. Thus, the column centred matrix is given by

$$\widetilde{\boldsymbol{B}_{Dk}} = \boldsymbol{B}_{Dk} - \mathbf{1}\mathbf{1}'\boldsymbol{B}_{Dk}/n_D,$$

where $\mathbf{1}$ is a n_D–dimensional column vector of ones. In the remainder, we will assume that $\boldsymbol{B}_{Dk}$ is the constrained version. Hence, we can express the predictor in matrix notation

$$\begin{aligned}\boldsymbol{\eta} &= \boldsymbol{B}_{D1}\boldsymbol{\beta}_{D1} + \cdots + \boldsymbol{B}_{Dp}\boldsymbol{\beta}_{Dp} + \boldsymbol{V}_D\boldsymbol{\gamma}_D \\ &= \boldsymbol{B}_D\boldsymbol{\beta}_D + \boldsymbol{V}_D\boldsymbol{\gamma}_D,\end{aligned}$$

where $\boldsymbol{B}_D = [\boldsymbol{B}_{D1} : \cdots : \boldsymbol{B}_{Dp}]$, $\boldsymbol{\beta}_D = (\boldsymbol{\beta}_{D1}, \ldots, \boldsymbol{\beta}_{Dp})'$ and $\boldsymbol{V}_D$ is the design matrix for the linear effects, where for convenience $v_{Dj,1} = 1$, for all $j = 1, \ldots, n_D$. It is well-known that the position and number of knots can have a large influence on the fitted functions (Wood, 2017, p. 169). To overcome this problem we will use P-splines, basically the idea is not to control the smoothness through the number of basis functions, but to use a roughness penalty instead. For this end, we will use a penalty based on differences of adjacent B-splines coefficients as described in Eilers and Marx (1996), with the Bayesian counterpart developed in Lang and Brezger (2004). Usually, under this setting between 20 and 40 knots are enough to guarantee sufficient flexibility and smoothness. Following the construction given by Lang and Brezger (2004), we assume a second-order random walk for the P-splines coefficients prior, that is,

$$\beta_{Dk,r} = 2\beta_{Dk,r-1} - \beta_{Dk,r-2} + u_{Dk,r},$$

where $u_{Dk,r} \overset{\text{idd}}{\sim} \text{N}(0, \tau^2_{Dk})$, for $r = 3, \ldots, R$, and $(\beta_{Dk,1}, \beta_{Dk,2}) \propto const$. This prior specification is equivalent to

$$\boldsymbol{\beta}_{Dk} \mid \tau^2_{Dk} \propto \exp\left\{-\frac{1}{2\tau^2_{Dk}}\boldsymbol{\beta}'_{Dk}\boldsymbol{K}_{Dk}\boldsymbol{\beta}_{Dk}\right\},$$

where the penalty matrix $\boldsymbol{K}_{Dk}$ is constructed from the second-order differences matrix,

$$\boldsymbol{D} = \begin{pmatrix} 1 & -2 & 1 & 0 & \ldots & 0 \\ 0 & 1 & -2 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 & -2 & 1 \end{pmatrix},$$

as $\boldsymbol{K}_{Dk} = \boldsymbol{D}'\boldsymbol{D}$, for $k = 1, \ldots, p$. Note that $\boldsymbol{K}_{Dk}$ is rank deficient with $\text{rank}(\boldsymbol{K}_{Dk}) = R - 2$. For later notational convenience, it is very useful to express the individual penalty matrices as a single block-diagonal matrix along with the additional variance parameters τ^2_{Dk}. These parameters control the amount of smoothness and correspond to the inverse of the smoothing parameters in the classical approach (Lang and Brezger, 2004).

$$\boldsymbol{K}_D = \begin{pmatrix} \frac{1}{\tau^2_{D1}}\boldsymbol{K}_{D1} & \boldsymbol{0} & \ldots & \boldsymbol{0} \\ \boldsymbol{0} & \frac{1}{\tau^2_{D2}}\boldsymbol{K}_{D2} & \ldots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \ldots & \ldots & \frac{1}{\tau^2_{Dp}}\boldsymbol{K}_{Dp} \end{pmatrix}.$$

For fully Bayesian inference, priors are set on the smoothing parameters $\boldsymbol{\tau}^2_D = (\tau^2_{D1}, \ldots, \tau^2_{Dp})$ as well. The usual choice is to set mutually independent inverse-gamma priors, $\tau^2_{Dk} \sim \text{IG}(a_{\tau^2_{Dk}}, b_{\tau^2_{Dk}})$, with small values for the hyperparameters, such as $a_{\tau^2_{Dk}} = 0.0005$ and $b_{\tau^2_{Dk}} = 0.0005$, leading to almost diffuse priors (Lang and Brezger, 2004).

Concerning the prior specification for the linear effects coefficients, we assume diffuse priors, i.e., $\gamma_{Dk} \propto const$, for $k = 1, \ldots, q$. Finally, to allow us to easily simulate from the posterior distribution, we will employ a truncation of the stick-breaking construction. Ishwaran and Zarepour (2000) showed that $\mathbb{E}(\sum_{l=L_D+1}^{\infty} \omega_{Dl}) = \alpha_D^{L_D}(\alpha_D + 1)^{-L_D}$. For instance, setting $\alpha_D = 1$ and $L_D = 20$, $\mathbb{E}(\sum_{l=L_D+1}^{\infty} \omega_{Dl}) <$

10^{-6}. Hence, the conditional density function can be represented as follows

$$f(y_{Dj} \mid \boldsymbol{X}_D = \boldsymbol{x}_{Dj}, \boldsymbol{V}_D = \boldsymbol{v}_{Dj}) = \sum_{l=1}^{L_D} \omega_{Dl} \phi \left(y_{Dj} \mid \boldsymbol{B}_{Dj}\boldsymbol{\beta}_D + \boldsymbol{v}_{Dj}^{'}\boldsymbol{\gamma}_D + \mu_{Dl}, \sigma_{Dl}^2 \right),$$

where $\boldsymbol{B}_{Dj}$ is the j-th row of the design matrix $\boldsymbol{B}_D$ and the weights ω_{Dl}, now follow a truncated stick-breaking construction: $\omega_{D1} = V_{D1}$, $\omega_{Dl} = V_{Dl}\prod_{t<l}(1 - V_{Dt})$, for $l = 2, \ldots, L_D$, and where the inputs of the weights are distributed according to a beta distribution, i.e., $V_{D1}, \ldots, V_{D,L_D-1} \overset{\text{iid}}{\sim} \text{Beta}(1, \alpha_D)$, and $V_{DL_D} = 1$ to ensure they add up to one. Further, to allow the data to inform about the appropriate value of α_D, we place a prior on α_D, specifically, we let $\alpha_D \sim \Gamma(a_\alpha, b_\alpha)$, i.e., a gamma distribution with shape a_α and rate b_α.

Posterior inference is addressed through explicit full conditionals. Let z_{Dj} be a latent random variable representing the component to which the j–th subject is allocated. This data augmentation allow us to rewrite our model hierarchically

$$y_{Dj} \mid z_{Dj}, \boldsymbol{x}_{Dj}, \boldsymbol{v}_{Dj}, \boldsymbol{\theta}_D \overset{\text{ind}}{\sim} \text{N}(\boldsymbol{B}_{Dj}\boldsymbol{\beta}_D + \boldsymbol{v}_{Dj}^{'}\boldsymbol{\gamma}_D + \mu_{Dz_{Dj}}, \sigma_{Dz_{Dj}}^2), \quad j = 1, \ldots, n_D, \tag{A.3}$$

$$z_{Dj} \mid \boldsymbol{V}_{D1}, \ldots, \boldsymbol{V}_{DL_D} \overset{\text{iid}}{\sim} \text{Mult}(1, \boldsymbol{\omega}_D), \tag{A.4}$$

$$\boldsymbol{\beta}_{Dk} \mid \tau_{Dk}^2 \propto \exp\left\{ -\frac{1}{2\tau_{Dk}^2} \boldsymbol{\beta}_{Dk}^{'} \boldsymbol{K}_{Dk} \boldsymbol{\beta}_{Dk} \right\}, \quad k = 1, \ldots, p, \tag{A.5}$$

$$\boldsymbol{\gamma}_D \propto const, \tag{A.6}$$

$$(\mu_{Dl}, \sigma_{Dl}^2) \overset{\text{iid}}{\sim} \text{N}(0, b_{\mu_D}^2)\text{IG}(a_{\sigma_D^2}, b_{\sigma_D^2}), \quad l = 1, \ldots, L_D, \tag{A.7}$$

$$V_{Dl} \mid \alpha_D \overset{\text{iid}}{\sim} \text{Beta}(1, \alpha_D), \quad l = 1, \ldots, L_D - 1, \tag{A.8}$$

$$\alpha_D \sim \Gamma(a_{\alpha_D}, b_{\alpha_D}), \tag{A.9}$$

$$\tau_{Dk}^2 \overset{\text{iid}}{\sim} \text{IG}(a_{\tau_{Dk}^2}, b_{\tau_{Dk}^2}), \tag{A.10}$$

where $\boldsymbol{\omega}_D = (\omega_{D1}, \ldots, \omega_{DL})$ and $\boldsymbol{\theta}_D = (\boldsymbol{\beta}_D, \boldsymbol{\gamma}_D, v_{D1}, \ldots, v_{DL}, \mu_{D1}, \ldots, \mu_{DL_D}, \sigma_{D1}^2, \ldots, \sigma_{DL_D}^2)$. With

the equations described in (A.3)–(A.10), we obtain the following full conditional distributions

$$\boldsymbol{\beta}_D \mid \text{else} \sim \text{N}\left((\boldsymbol{B}_D^{'}\boldsymbol{\Omega}_D^{-1}\boldsymbol{B}_D + \boldsymbol{K}_D)^{-1}\boldsymbol{B}_D^{'}\boldsymbol{\Omega}_D^{-1}(\boldsymbol{y}_D^{*} - \boldsymbol{V}_D\boldsymbol{\gamma}_D), (\boldsymbol{B}_D^{'}\boldsymbol{\Omega}_D^{-1}\boldsymbol{B}_D + \boldsymbol{K}_D)^{-1}\right),$$

$$\boldsymbol{\gamma}_D \mid \text{else} \sim \text{N}\left((\boldsymbol{V}_D^{'}\boldsymbol{\Omega}_D^{-1}\boldsymbol{V}_D)^{-1}\boldsymbol{V}_D^{'}\boldsymbol{\Omega}_D^{-1}(\boldsymbol{y}_D^{*} - \boldsymbol{B}_D\boldsymbol{\beta}_D), (\boldsymbol{V}_D^{'}\boldsymbol{\Omega}_D^{-1}\boldsymbol{V}_D)^{-1}\right),$$

$$\mu_{Dl} \mid \text{else} \sim \text{N}\left(\frac{\frac{1}{\sigma_{Dl}^2}\sum_{j:z_{Dj}=l}\varepsilon_{Dj}}{n_{Dl}/\sigma_{Dl}^2 + 1/b_{\mu_D}^2}, \frac{1}{n_{Dl}/\sigma_{Dl}^2 + 1/b_{\mu_D}^2}\right), \quad l = 1, \ldots, L_D, \text{ where } n_{Dl} = \sum_{j=1}^{n_D}\mathbb{1}\{z_{Dj} = l\},$$

$$\sigma_{Dl}^2 \mid \text{else} \sim \text{IG}\left(a_{\sigma_D^2} + \frac{n_{Dl}}{2}, b_{\sigma_D^2} + \frac{1}{2}\sum_{j:z_{Dj}=l}(\varepsilon_{Dj} - \mu_{Dl})^2\right),$$

$$z_{Dj} \mid \text{else} \sim \text{Mult}(1, \boldsymbol{\pi}_{Dj}), \text{ with } \pi_{Dj,l} = \frac{\omega_{Dl}\phi(y_{Di} \mid \boldsymbol{B}_{Dj}\boldsymbol{\beta}_D - \boldsymbol{v}_{Dj}^{'}\boldsymbol{\gamma}_D + \mu_{Dl}, \sigma_{Dl}^2)}{\sum_{\ell=1}^{L_D}\omega_{D\ell}\phi(y_{Dj} \mid \boldsymbol{B}_{Dj}\boldsymbol{\beta}_D - \boldsymbol{v}_{Dj}^{'}\boldsymbol{\gamma}_D + \mu_{D\ell}, \sigma_{D\ell}^2)}, \ j = 1, \ldots, n_D,$$

$$V_{Dl} \mid \text{else} \sim \text{Beta}\left(n_{Dl} + 1, \sum_{\ell=l+1}^{L_D} n_{D\ell} + \alpha_D\right),$$

$$\alpha_D \mid \text{else} \sim \Gamma\left(a_\alpha + L_D - 1, b_\alpha - \sum_{l=1}^{L_D-1}\log(1 - V_{Dl})\right),$$

$$\tau_{Dk}^2 \mid \text{else} \sim \text{IG}\left(a_{\tau_{Dk}^2} + \frac{\text{rank}(\boldsymbol{K}_{Dk})}{2}, b_{\tau_{Dk}^2} + \frac{1}{2}\boldsymbol{\beta}_{Dk}^{'}\boldsymbol{K}_{Dk}\boldsymbol{\beta}_{Dk}\right), \quad k = 1, \ldots, p,$$

where $\boldsymbol{y}_D^{*} = (y_1 - \mu_{Dz_1}, \ldots, y_{n_D} - \mu_{Dz_{n_D}})$, $\boldsymbol{\Omega}_D = \text{diag}(\sigma_{Dz_1}^2, \ldots, \sigma_{Dz_{n_D}}^2)$ and $\varepsilon_{Dj} = y_{Dj} - \boldsymbol{B}_{Dj}\boldsymbol{\beta}_D - \boldsymbol{v}_{Dj}^{'}\boldsymbol{\gamma}_D$. These full conditional distributions allow us to implement a simple Gibbs sampler to simulate draws, say S, from the posterior distribution. Thus, at iteration s, the estimated conditional density is given by

$$f^{(s)}\left(y_{Dj} \mid \boldsymbol{X}_D = \boldsymbol{x}_{Dj}, \boldsymbol{V}_D = \boldsymbol{v}_{Dj}\right) = \sum_{l=1}^{L_D}\omega_{Dl}^{(s)}\phi\left(y_{Dj} \mid \boldsymbol{B}_{Dj}\boldsymbol{\beta}_D^{(s)} + \boldsymbol{v}_{Dj}^{'}\boldsymbol{\gamma}_D^{(s)} + \mu_{Dl}^{(s)}, (\sigma_{Dl}^{(s)})^2\right).$$

Our proposed estimator for the covariate-specific OVL is based on equation (A.1). The integral will be replaced by a numerical integration method over all the posterior draws, more precisely, we will use the trapezoidal rule for this end. Denoting by $g^{(s)}(y \mid \boldsymbol{x}) = \min\left\{f_{\bar{D}}^{(s)}(y \mid \boldsymbol{x}), f_D^{(s)}(y \mid \boldsymbol{x})\right\}$, for any fixed value $\boldsymbol{x}$ of the covariates, a single Gibbs sampler iteration of our proposed estimator is given by

$$\begin{aligned}\text{OVL}^{(s)}(\boldsymbol{x}) &= \frac{\Delta y}{2}\sum_{i=1}^{N}\left\{g^{(s)}(y_{i-1} \mid \boldsymbol{x}) + g^{(s)}(y_i \mid \boldsymbol{x})\right\}, \\ &= \frac{\Delta y}{2}\left\{g^{(s)}(y_0 \mid \boldsymbol{x}) + 2\sum_{i=1}^{N-1}g^{(s)}(y_i \mid \boldsymbol{x}) + g^{(s)}(y_N \mid \boldsymbol{x})\right\}, \quad s = 1, \ldots, S, \end{aligned} \tag{A.11}$$

where $\min\{y_{\bar{D}}, y_D\} = y_0 < \cdots < y_N = \max\{y_{\bar{D}}, y_D\}$ is an equally spaced grid and Δy is the length of each sub-interval. Note that at this stage, one should compute the corresponding basis functions at $\boldsymbol{x}$. More precisely, $\boldsymbol{B}_D = (\boldsymbol{B}_{D1}, \ldots, \boldsymbol{B}_{Dp})$, where $\boldsymbol{B}_{Dk} = [B_{Dk,1}(x_k), \ldots, B_{Dk,R}(x_k)]$, for $k = 1, \ldots, p$, and $\boldsymbol{B}_{\bar{D}}$ is constructed in a similar way. Finally, we can take $\widehat{\mathrm{OVL}}(\boldsymbol{x}) = \frac{1}{S}\sum_{s=1}^{S} \mathrm{OVL}^{(s)}(\boldsymbol{x})$ as a point estimate of the covariate-specific OVL at $\boldsymbol{x}$ and obtain symmetric $100(1-\alpha)\%$ pointwise credible bands using the $\alpha/2$ and $1-\alpha/2$ quantiles of the posterior ensemble $\left\{\mathrm{OVL}^{(1)}(\boldsymbol{x}), \ldots, \mathrm{OVL}^{(S)}(\boldsymbol{x})\right\}$.

A.3 Simulated examples

In this section we illustrate our methods through five different examples. Since our purpose is only to show our methodology, we have performed just one replication. For the first three scenarios, we have simulated a single continuous covariate for each group, $x_{\bar{D},1} \sim \mathrm{U}(0,1)$ and $x_{D,1} \sim \mathrm{U}(0,1)$. And for the fourth and fifth scenario, we have added a second covariate, $x_{\bar{D},2}, x_{D,2} \sim \mathrm{U}(0,1)$ and $x_{\bar{D},3}, x_{D,3} \sim$ Bernoulli(0.5), respectively. In all cases, we have considered sample sizes $n_{\bar{D}} = 200$ and $n_D = 200$. Regarding the MCMC algorithms, we have saved 2,000 simulations and discarded 500 as the burn-in of the chains.

A.3.1 Toy illustrative examples

The five scenarios considered are listed in Table A.1. These scenarios cover the situation where the relation between the covariates and the test outcomes is i) linear (Scenario I) and ii) non-linear (Scenario II–V). Regarding the distributional assumptions, we have considered that the test outcomes follow: i) a normal distribution for both populations (Scenario I, II and IV), ii) a gamma distribution for the nondiseased group (we are using the shape-scale parametrization, i.e., if $X \sim \Gamma(a, b)$, a is the shape and b the scale parameter of the distribution), and a mixture of normal distributions for the diseased group (Scenario III) and iii) a scaled Student's t-distribution for both populations (Scenario V).

Table A.1: Scenarios considered for the toy examples

Scenario	Model for $Y_{\bar{D}}$	Model for Y_D
1	$y_{\bar{D}i} \overset{\text{ind}}{\sim} \text{N}(0.4 + 0.8x_{\bar{D}i,1}, 2^2)$	$y_{Dj} \overset{\text{ind}}{\sim} \text{N}(0.5 + 0.65x_{Dj,1}, 0.4^2)$
2	$y_{\bar{D}i} \overset{\text{ind}}{\sim} \text{N}\left(\frac{\exp\{-2x_{\bar{D}i,1}\}}{2}, 0.7^2\right)$	$y_{Dj} \overset{\text{ind}}{\sim} \text{N}\left(\frac{\exp\{-2(x_{Dj,1}-1)\}}{2} - 0.4, 0.5^2\right)$
3	$y_{\bar{D}i} \overset{\text{ind}}{\sim} \Gamma\left(10, \frac{10}{(2x_{\bar{D}i,1}-1)^2+0.5}\right)$	$y_{Dj} \overset{\text{ind}}{\sim} 0.2\text{N}(-\{2x_{Dj,1}-1\}^2 - 1.5, 0.5^2)$ $+ 0.8\text{N}(-\{2x_{Dj,1}-1\}^2 + 2.5, 1)$
4	$y_{\bar{D}i} \overset{\text{ind}}{\sim} \text{N}\left(x_{\bar{D}i,1}^3 - 3x_{\bar{D}i,1} + x_{\bar{D}i,2}^2, 1\right)$	$y_{Dj} \overset{\text{ind}}{\sim} \text{N}\big(4\{x_{Dj,1}-0.5\}^2$ $+ 6\{x_{Dj,2}-0.5\}^2 - 0.5, 2^2\big)$
5	$y_{\bar{D}i} \overset{\text{ind}}{\sim} t_3\big(\{x_{\bar{D}i,1}+0.8\}^{4\cos(6x_{\bar{D}i,1}+4.8)-1}$ $- 0.16x_{\bar{D}i,3} + 0.2, 0.5^2\big)$	$y_{Dj} \overset{\text{ind}}{\sim} t_2\big(\{x_{Dj,1}+0.9\}^{4\cos(x_{Dj,1}+0.9)\sin(2x_{Dj,1}+0.2)}$ $- \exp\{-0.2\}x_{Dj,3}, 0.2^2\big)$

A.3.2 Models

To avoid numerical difficulties within the MCMC simulations and to facilitate prior specification, we have scaled the responses (i.e., divided by the standard deviation), but we transformed back to the original scale when presenting the results. For the P-splines we have used $m = 20$ interior knots and a highly disperse prior for the smoothing parameter ($a_{\tau^2_{dk}} = 0.005$, $b_{\tau^2_{dk}} = 0.005$), for $d \in \{\bar{D}, D\}$ and $k = 1, 2$. We used $a_{\alpha_d} = 2$ and $b_{\alpha_d} = 2$ as hyperparameters for the concentration parameter of the DP, under this prior $\mathbb{E}(\alpha_d) = 1$ and $\text{Var}(\alpha_d) = 0.5$, . For the means and variances of the components, we used relatively vague prior distributions, that is, the variance of the mean ($b^2_{\mu_d} = 100$) is large, whereas $a_{\sigma^2_d} = 2$ leads to a prior with infinite variance (hence, in some sense, vague) centred around a finite mean ($b_{\sigma^2_d} = 0.5$). Note that this prior on σ^2_{dl} favours variances less than one, since $\mathbb{E}(\sigma^2_{dl}) = 0.5$. The scaled data have marginal variance of one, so it is expected that the within-component variance σ^2_{dl} to be smaller than the marginal variance. Finally, we capped the stick-breaking construction up to $L_d = 20$, this means that a maximum of 20 normal distributions were used to approximate the conditional densities.

For the first three scenarios, we compared the performance of our proposed model against two popular alternatives, a Bayesian additive model with normally distributed error terms (see e.g., Fahrmeir

et al. 2013, chapter 8 and Lang and Brezger 2004 for an overview), which is a particular case of our model when $L_d = 1$, and a Bayesian normal linear model, also a particular case of our model when $L_d = 1$ and every function h_{Dk} is linear. In the first one, the test outcomes are assumed Gaussian with mean

$$\eta_{Dj} = g_{D1}(x_{Dj,1}) + \cdots + g_{Dp}(x_{Dj,p}), \quad j = 1, \ldots, n_D, \tag{A.12}$$

and constant variance σ_D^2. The smooth functions g_{Dk}, for $k = 1, \ldots, p$, are also approximated by P-splines (see Lang and Brezger (2004) for a detailed description). In the second one, the test outcomes are also assumed Gaussian with constant variance. However, it does not allow for non-linear effects, rather it assumed that the mean is a (more restrictive) linear function of the covariates, i.e.,

$$\eta_{Dj} = \beta_{D0} + \beta_{D1} x_{Dj,1} + \cdots + \beta_{Dp} x_{Dj,p}, \quad j = 1, \ldots, n_D. \tag{A.13}$$

Thus, the corresponding density functions are given by

$$f(y_{Dj} \mid \boldsymbol{X}_{Dj} = \boldsymbol{x}_{Dj}) = \phi(y_{Dj} \mid \eta_{Dj}, \sigma_D^2),$$

where η_{Dj} is defined as in equations (A.12) and (A.13). Finally, the estimated densities are plugged in on equation (A.1) to obtain the covariate-specific OVL.

A.3.3 Results

Figures A.1–A.3 show the results for Scenario I, II and III, respectively. Top row corresponds to our proposed model (PS-DPM). Whereas, the center and bottom rows to the additive model (AM) and the linear model (LM), respectively. Left and center columns show the estimated (posterior mean) mean function along with 95% pointwise credible bands and the true mean function for the nondiseased and diseased group, respectively. Finally, right column depicts the corresponding OVL under each model.

In Figure A.1, we observe that, when normality holds and the effect of the covariate is linear, all the models provide estimates close to the true values. Although the nondiseased group presents greater variability, on average, the test outcomes from both groups are very similar. Therefore, it is expected

that the covariate does not have a significant impact on the discriminatory ability of the test. This is consistent with the observed OVL, which remains almost constant for all covariate values under the three models.

In turn, when the effect of the covariate is non-linear, it is clear that the LM fails to estimate accurately the mean function, as can be observed on the bottom row of Figure A.2. In the case of the nondiseased group, although the true mean function is non-linear, it can be well-approximated by a straight line. Hence, all the models produce good estimates. Moreover, note that, on average, there is a larger difference between the nondiseased and diseased test outcomes for small values of the covariate. However, as it increases, this difference decreases. Therefore, the diagnostic test is less accurate as the covariate increases. This behaviour seems to be well described under our model and the AM (top and center right rows, respectively). Conversely, under the LM the estimation of the covariate-specific OVL presents greater difficulties.

In Figure A.3 we can see a situation in which the test presents a better discrimination for intermediate values of the covariate. In this case, low or high values of the covariate greatly affect the accuracy of the test. Similar to Scenario II, the LM is cannot capture the non-linear effect of the covariate. In addition, when the test outcomes arise from different non normal distributions, even the AM is unable to accurately describe the mean function (center plot). Hence, the estimated covariate-specific OVL under both models the LM and AM is very poor (center and bottom right rows). In turn, the PS-DPM produces substantially better results. It captures the mean functions of both groups and it is able to describe more accurately the behaviour of the covariate-specific OVL, except at the boundaries.

We also computed different model comparison criteria (detailed in Section 2.3) for these three models under consideration. The results are listed in Table A.2. In Scenario I, despite that all models showed a similar performance, it seems that most of the criteria support the LM as the best model for both the nondiseased and diseased group. This was expected, since this scenario considers that the test outcomes are Gaussian with a linear mean function. In this case, the posterior rank probability is the only one that chooses the AM as a better option. However, recall that our purpose is only to illustrate our methodology, thus we are largely subject to Monte Carlo error. From Figure A.2, we observed that although the true mean function of the nondiseased group is non-linear, it can be approximated well

by a straight line. This can be corroborated by all the criteria, therefore the LM provides a better fit. However, a clearer non-linear trend, such as that observed in the diseased group, cannot be captured by the LM. In this case, our model provides a better fit. Finally, when the covariate effect is non-linear and the test outcomes arise from different parametric distributions, the PS-DPM clearly outperforms the other two models.

Figures A.4–A.6 show the results for Scenario IV. However, it is difficult to see clearly whether the fit of our model is good or not. For this reason, we resort to contours plots shown in Figure A.7 and profile plots depicted in Figure A.8, where we have fixed x_2 at three particular values, namely, $x_2 = 0.2$, 0.5 and 0.8. From the latter, we may observe that our model correctly recovers the mean functions in both groups, as well as the corresponding covariate-specific OVL.

Finally, the results for Scenario V are shown in Figure A.9. Top row depicts the case when the binary covariate $x_3 = 0$ and bottom row when $x_3 = 1$. This scenario represents a situation where the test outcomes arise from different non normal parametric families and where the test discriminatory ability, in addition to being affected by a continuous covariate x_1, changes dramatically in presence of an additional binary covariate. When $x_3 = 0$ the test is less accurate for values of x_1 less than 0.5. However, for values between 0.6 and 0.8, its performance improves and then, it decreases again when x_1 is greater than 0.8. When $x_3 = 1$, we may observe an inverse behaviour, that is, the accuracy of the test decreases as x_1 increases, reaching a peak around 0.8, whereas the test becomes more accurate for values greater than 0.8. It should be noted that, in this last sub-interval, our estimate coincides exactly with the true covariate-specific OVL. This is an example of a diagnostic test that shows better performance for some subgroup of the populations.

In conclusion, from these illustrative examples, we can preliminary conclude that when the covariates effect is non-linear and/or the test outcomes arise from different non normal parametric families, our modelling approach outperforms the AM and the LM.

Table A.2: Model comparison criteria

Scenario	Group	Criterion	PS-DPM	AM	LM
I	Nondiseased	Adjusted LPML	-424.19	-424.24	**-423.45**
		WAIC	848.36	848.46	**846.9**
		DIC_3	848.1	848.22	**846.78**
		Posterior rank probability	0.0963	**0.716**	0.1877
	Diseased	Adjusted LPML	-95.44	-95.52	**-95.05**
		WAIC	190.86	191.03	**190.11**
		DIC_3	190.61	190.78	**189.98**
		Posterior rank probability	0.3919	**0.5798**	0.0283
II	Nondiseased	Adjusted LPML	-214.11	-214.1	**-213.21**
		WAIC	428.21	428.19	**426.42**
		DIC_3	427.97	427.96	**426.3**
		Posterior rank probability	0.0001	0.4031	**0.5968**
	Diseased	Adjusted LPML	**-140.84**	-141.06	-151.23
		WAIC	**281.67**	282.07	302.45
		DIC_3	**281.29**	281.62	302.31
		Posterior rank probability	**0.9701**	0.0287	0.0012
III	Nondiseased	Adjusted LPML	**-8.45**	-17.34	-89.33
		WAIC	**16.72**	34.61	178.68
		DIC_3	**15.61**	33.78	178.5
		Posterior rank probability	**0.9617**	0.0383	0.0
	Diseased	Adjusted LPML	**-364.86**	-402.07	-404.4
		WAIC	**729.59**	804.13	808.81
		DIC_3	**728.85**	803.83	808.71
		Posterior rank probability	**1.0**	0.0	0.0

A.4 Application

Diabetes mellitus is a condition in which glucose levels in blood are too high and cannot be controlled by the lack of a hormone called insulin. This may be due to the immune system (Type 1) or because the body is unable to produce it in sufficient quantity (Type 2). In recent years, its prevalence in the world population has increased. According to estimates by the World Health Organization (WHO) in 2016, the number of adults with diabetes doubled since 2014, from 4.7% to 8.5% of the adult population (World Health Organization, 2016). Moreover, in the same year, diabetes caused over 1.6 million of deaths (4%) (World Health Organization, 2018, p. 7). Among other risk factors, it is closely related to overweight, a condition also responsible for other diseases that are among the leading causes of death worldwide (e.g., cardiovascular diseases). Unfortunately, there is no cure currently, further research is necessary. However, an early diagnosis can help to improve patients' lives. Conversely, if the patients do not have enough care or proper treatment, they might develop other serious medical complications, such as blindness or leg amputations. For all the reasons mentioned above, diabetes is considered as a serious public health problem. Numerous institutions around the world have gathered efforts to combat this condition (e.g., Diabetes UK in the UK, Federación Mexicana de Diabetes in Mexico, among others). However, more work is needed along with better health policies on each government.

We have applied our method to data from a population based survey of diabetes in Cairo, Egypt (Smith and Thompson, 1996). The data comprises measurements on postprandial blood glucose obtained from a finger stick on 88 subjects with diabetes and 198 non diabetic. Our primary goal is to evaluate the effect of age in the accuracy of glucose as a biomarker of diabetes. In the top row of Figure A.10, for the non diabetic group (left) we can see that glucose levels tend to increase almost linearly as age increases. Whereas for the diseased group, we can observe an opposite behaviour. Therefore, the estimated age-specific OVL (bottom left) increases with age (less diagnosis accuracy) reaching a peak roughly at 70 years, where it starts to decrease. Finally, in the bottom right of Figure A.10, there are depicted examples of conditional densities of both groups for three particular ages (quartiles). In these plots, it is possible to note that the conditional densities of the diseased group have heavier tails, as expected.

In conclusion, there seems to be evidence that glucose levels are less accurate as a marker for diabetes as age increases.

It is always advisable to fit different models to the data when this is possible. For this reason, we have fitted an AM and an LM. Table A.3 shows the model comparison criteria results. We note that, the PS-DPM seems to be the best option for the non diabetic population. While for the diabetic group, according to the LPML, WAIC and DIC_3, the PS-DPM is the most reasonable option as well. However, according to the posterior rank probability, the AM is almost equally likely to provide a better fit. Despite this, we can conclude that inferences derived from the covariate-specific OVL are more reliable under our proposed framework.

Table A.3: Diabetes model comparison criteria

Group	Criterion	PS-DPM	AM	LM
Non diabetic	Adjusted LPML	**-884.38**	-935.56	-938.96
	WAIC	**1768.09**	1871.41	1878.17
	DIC_3	**1765.48**	1869.03	1876.8
	Posterior rank probability	**0.9999**	0.0	0.0001
Diabetic	Adjusted LPML	**-532.51**	-532.54	-538.18
	WAIC	**1064.59**	1064.66	1076.35
	DIC_3	**1063.38**	1063.56	1076.2
	Posterior rank probability	0.4314	**0.5295**	0.0391

Finally, to check the goodness of fit of the PS-DPM models, we performed posterior predictive checks for both populations. The idea is very simple, replicated data under our model, using the posterior predictive distribution, should look similar to the observed data. More precisely, some statistics could help us to determine if there are discrepancies that indicate potential failings of our model. Figures A.11 and A.12 do not show any visible sign of misfit of our models, since in almost cases, most of the density is close to the observed statistic. Note that in both non diabetic and diabetic group, there were replications where the minimum is negative. However, this is not possible, because glucose levels are always positive. The reason is that we are using a normal distribution as the kernel of the DPM. Hence, it might assign positive probability mass to negative values. The last plot (bottom right) on

each figure, illustrates the observed density compared to the replicated densities. In both cases, it is clear that the observed densities (thick line) seem like another draw from the model. Therefore, we can state that our models fit well to the data.

A.5 Full conditional distributions

In this section, we will describe in detail the derivation of the full conditional distributions of our proposed model. Recall that the conditional density function for the diseased population (again, we will discuss only the case for the diseased group, the procedure is the same for the nondiseased) is given by

$$f(y_{Dj} \mid \boldsymbol{X}_D = \boldsymbol{x}_{Dj}, \boldsymbol{V}_D = \boldsymbol{v}_{Dj}) = \sum_{l=1}^{L_D} \omega_{Dl} \phi \left(y_{Dj} \mid \boldsymbol{B}_{Dj}\boldsymbol{\beta}_D + \boldsymbol{v}_{Dj}^{'}\boldsymbol{\gamma}_D + \mu_{Dl}, \sigma_{Dl}^2 \right). \tag{A.14}$$

Note that the equation described in (A.14) resembles a finite mixture of normal distributions. Denoting by $\boldsymbol{\theta}_D = (V_{D1}, \ldots, V_{DL_D-1}, \boldsymbol{\beta}_D, \boldsymbol{\gamma}_D, \mu_{D1}, \ldots, \mu_{DL_D}, \sigma_{D1}^2, \ldots, \sigma_{DL_D}^2)$, the likelihood of the data is given by

$$L(\boldsymbol{\theta}_D; \boldsymbol{y}_D) = \prod_{j=1}^{n_D} \left\{ \sum_{l=1}^{L_D} \omega_{Dl} \phi \left(y_{Dj} \mid \boldsymbol{B}_{Dj}\boldsymbol{\beta}_D + \boldsymbol{v}_{Dj}^{'}\boldsymbol{\gamma}_D + \mu_{Dl}, \sigma_{Dl}^2 \right) \right\},$$

which is analytically intractable. However, we can simply use data augmentation to overcome this problem. Let z_{Dj} be a latent random variable representing the mixture component to which the j-th observation is allocated. For instance, $z_{Dj} = \ell$ denotes that observation y_{Dj} comes from component ℓ and thus, from a normal distribution with mean $\boldsymbol{B}_{Dj}\boldsymbol{\beta}_D + \boldsymbol{v}_{Di}^{'}\boldsymbol{\gamma}_D + \mu_{D\ell}$ and variance $\sigma_{D\ell}^2$. Under this approach, we can re-express the likelihood as follows

$$\begin{aligned}
L(\boldsymbol{\theta}_D; \boldsymbol{y}_D, \boldsymbol{z}_D) &= \prod_{j=1}^{n_D} \omega_{Dz_{Dj}} \phi \left(y_{Dj} \mid \boldsymbol{B}_{Dj}\boldsymbol{\beta}_D + \boldsymbol{v}_{Dj}^{'}\boldsymbol{\gamma}_D + \mu_{Dz_{Dj}}, \sigma_{Dz_{Dj}}^2 \right) \\
&= \prod_{l=1}^{L_D} \prod_{j:z_{Dj}=l} \omega_{Dl} \phi \left(y_{Dj} \mid \boldsymbol{B}_{Dj}\boldsymbol{\beta}_D + \boldsymbol{v}_{Dj}^{'}\boldsymbol{\gamma}_D + \mu_{Dl}, \sigma_{Dl}^2 \right) \\
&= \prod_{l=1}^{L_D} \omega_{Dl}^{n_{Dl}} \prod_{j:z_{Dj}=l} \phi \left(y_{Dj} \mid \boldsymbol{B}_{Dj}\boldsymbol{\beta}_D + \boldsymbol{v}_{Dj}^{'}\boldsymbol{\gamma}_D + \mu_{Dl}, \sigma_{Dl}^2 \right),
\end{aligned}$$

where $n_{Dl} = \sum_{j=1}^{n_D} \mathbb{I}\{z_{Dj} = l\}$. The joint posterior distribution is given by

$$
\begin{aligned}
p(\boldsymbol{\theta}_D, \boldsymbol{z}_D, \alpha_D, \boldsymbol{\tau}_D^2 \mid \boldsymbol{y}_D) &\propto L(\boldsymbol{\theta}_D; \boldsymbol{y}_D, \boldsymbol{z}_D) p(\boldsymbol{\theta}_D) p(\alpha_D) p(\boldsymbol{\tau}_D^2) \\
&= L(\boldsymbol{\theta}_D; \boldsymbol{y}_D, \boldsymbol{z}_D) p(\boldsymbol{\beta}_D \mid \boldsymbol{\tau}_D^2) p(\boldsymbol{\gamma}_D) \prod_{l=1}^{L_D} p(\mu_{Dl}) p(\sigma_{Dl}^2) \prod_{l=1}^{L_D - 1} p(V_{Dl} \mid \alpha_D) p(\alpha_D) p(\boldsymbol{\tau}_D^2) \\
&\propto \prod_{l=1}^{L_D} \left\{ V_{Dl} \prod_{t<l} (1 - V_{Dt}) \right\}^{n_{Dl}} \prod_{j:z_{Dj}=l} \phi\left(y_{Dj} \mid \boldsymbol{B}_{Dj}\boldsymbol{\beta}_D + \boldsymbol{v}_{Dj}^{'}\boldsymbol{\gamma}_D + \mu_{Dl}, \sigma_{Dl}^2 \right) \\
&\times \prod_{k=1}^{p} (\tau_{Dj}^2)^{-\frac{\text{rank}(\boldsymbol{K}_{Dk})}{2}} \exp\left\{ -\frac{1}{2\tau_{Dk}^2} \boldsymbol{\beta}_{Dk}^{'} \boldsymbol{K}_{Dk} \boldsymbol{\beta}_{Dk} \right\} \\
&\times \prod_{l=1}^{L_D} \exp\left\{ -\frac{1}{2b_{\mu_D}^2} \mu_{Dl}^2 \right\} (\sigma_{Dl}^2)^{-a_{\sigma_D^2} - 1} \exp\left\{ -\frac{b_{\sigma_D^2}}{\sigma_{Dl}^2} \right\} \\
&\times \prod_{l=1}^{L_D - 1} \frac{\Gamma(\alpha_D + 1)}{\Gamma(\alpha_D)\Gamma(1)} (1 - V_{Dl})^{\alpha_D - 1} \\
&\times \alpha_D^{a_{\alpha_D} - 1} \exp\{-b_{\alpha_D}\alpha_D\} \\
&\times \prod_{k=1}^{p} (\tau_{Dk}^2)^{-a_{\tau_k^2} - 1} \exp\left\{ -\frac{b_{\tau_k^2}}{\tau_{Dk}^2} \right\}.
\end{aligned}
$$

Which does not have a recognisable form, but the full conditional distributions do. First, we will derive the full conditional for $\boldsymbol{\beta}_D$. Define $\boldsymbol{\mu}_D = (\mu_{z_{D1}}, \ldots, \mu_{z_{Dn_D}})$ and $\boldsymbol{\Omega}_D = \text{diag}(\sigma_{z_{D1}}^2, \ldots, \sigma_{z_{Dn_D}}^2)$ and recall that in Section A.2 we expressed the individual penalty matrices as a single block-diagonal matrix $\boldsymbol{K}_D$. This allows us to express the full conditional as

$$
\begin{aligned}
p(\boldsymbol{\beta}_D \mid \text{else}) &\propto \exp\left\{ -\frac{1}{2} \left(\boldsymbol{y}_D - \boldsymbol{B}_D\boldsymbol{\beta}_D - \boldsymbol{V}_D\boldsymbol{\gamma}_D - \boldsymbol{\mu}_D \right)^{'} \boldsymbol{\Omega}_D^{-1} \left(\boldsymbol{y}_D - \boldsymbol{B}_D\boldsymbol{\beta}_D - \boldsymbol{V}_D\boldsymbol{\gamma}_D - \boldsymbol{\mu}_D \right) \right\} \\
&\times \exp\left\{ -\frac{1}{2} \boldsymbol{\beta}_D^{'} \boldsymbol{K}_D \boldsymbol{\beta}_D \right\} \\
&\propto \exp\left\{ \boldsymbol{\beta}_D^{'} \boldsymbol{B}_D^{'} \boldsymbol{\Omega}_D^{-1} \boldsymbol{y}_D - \boldsymbol{\beta}_D^{'} \boldsymbol{B}_D^{'} \boldsymbol{\Omega}_D^{-1} \boldsymbol{V}_D \boldsymbol{\gamma}_D - \boldsymbol{\beta}_D^{'} \boldsymbol{B}_D^{'} \boldsymbol{\Omega}_D^{-1} \boldsymbol{\mu}_D - \frac{1}{2} \boldsymbol{\beta}_D^{'} (\boldsymbol{B}_D^{'} \boldsymbol{\Omega}_D^{-1} \boldsymbol{B}_D + \boldsymbol{K}_D) \boldsymbol{\beta}_D \right\} \\
&= \exp\left\{ \boldsymbol{\beta}_D^{'} \left[\boldsymbol{B}_D^{'} \boldsymbol{\Omega}_D^{-1} (\boldsymbol{y}_D - \boldsymbol{V}_D\boldsymbol{\gamma}_D - \boldsymbol{\mu}_D) \right] - \frac{1}{2} \boldsymbol{\beta}_D^{'} (\boldsymbol{B}_D^{'} \boldsymbol{\Omega}_D^{-1} \boldsymbol{B}_D + \boldsymbol{K}_D) \boldsymbol{\beta}_D \right\}.
\end{aligned}
$$

Hoff (2009, p. 107-108) showed that if a random vector $\boldsymbol{\theta}$ has a density on $\mathbb{R}^p$ proportional to $\exp\left\{ -\frac{1}{2}\boldsymbol{\theta}^{'}\boldsymbol{A}\boldsymbol{\theta} + \boldsymbol{\theta}^{'}\boldsymbol{b} \right\}$ for some matrix $\boldsymbol{A}$ and vector $\boldsymbol{b}$, then $\boldsymbol{\theta}$ must have a multivariate normal distribution with covariance matrix given by $\boldsymbol{A}^{-1}$ and mean vector $\boldsymbol{A}^{-1}\boldsymbol{b}$. Therefore, $\boldsymbol{\beta}_D$ must have

a multivariate normal distribution with covariance matrix $\boldsymbol{\Lambda}^{-1} = (\boldsymbol{B}_D^{'}\boldsymbol{\Omega}_D^{-1}\boldsymbol{B}_D + \boldsymbol{K}_D)^{-1}$ and mean $\boldsymbol{\Lambda}^{-1}\boldsymbol{B}_D^{'}\boldsymbol{\Omega}_D^{-1}(\boldsymbol{y}_D - \boldsymbol{V}_D\boldsymbol{\gamma}_D - \boldsymbol{\mu}_D)$, that is,

$$\boldsymbol{\beta}_D \mid \text{else} \sim \text{N}\left((\boldsymbol{B}_D^{'}\boldsymbol{\Omega}_D^{-1}\boldsymbol{B}_D + \boldsymbol{K}_D)^{-1}\boldsymbol{B}_D^{'}\boldsymbol{\Omega}_D^{-1}(\boldsymbol{y}_D - \boldsymbol{V}_D\boldsymbol{\gamma}_D - \boldsymbol{\mu}_D), (\boldsymbol{B}_D^{'}\boldsymbol{\Omega}_D^{-1}\boldsymbol{B}_D + \boldsymbol{K}_D)^{-1}\right).$$

Similarly, the full conditional for $\boldsymbol{\gamma}_D$ is

$$\begin{aligned} p(\boldsymbol{\gamma}_D \mid \text{else}) &\propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{y}_D - \boldsymbol{B}_D\boldsymbol{\beta}_D - \boldsymbol{V}_D\boldsymbol{\gamma}_D - \boldsymbol{\mu}_D\right)^{'}\boldsymbol{\Omega}_D^{-1}\left(\boldsymbol{y}_D - \boldsymbol{B}_D\boldsymbol{\beta}_D - \boldsymbol{V}_D\boldsymbol{\gamma}_D - \boldsymbol{\mu}_D\right)\right\} \\ &\propto \exp\left\{\boldsymbol{\gamma}_D^{'}\left[\boldsymbol{V}_D^{'}\boldsymbol{\Omega}_D^{-1}(\boldsymbol{y}_D - \boldsymbol{B}_D\boldsymbol{\beta}_D - \boldsymbol{\mu}_D)\right] - \frac{1}{2}\boldsymbol{\gamma}_D^{'}(\boldsymbol{V}_D^{'}\boldsymbol{\Omega}_D^{-1}\boldsymbol{V}_D)\boldsymbol{\gamma}_D\right\}. \end{aligned}$$

It follows that, $\boldsymbol{\gamma}_D$ must have a multivariate normal distribution with covariance matrix given by $\boldsymbol{\kappa}^{-1} = (\boldsymbol{V}_D^{'}\boldsymbol{\Omega}_D^{-1}\boldsymbol{V}_D)^{-1}$ and mean $\boldsymbol{\kappa}^{-1}\boldsymbol{V}_D^{'}\boldsymbol{\Omega}_D^{-1}(\boldsymbol{y}_D - \boldsymbol{B}_D\boldsymbol{\beta}_D - \boldsymbol{\mu}_D)$, i.e.,

$$\boldsymbol{\gamma}_D \mid \text{else} \sim \text{N}\left((\boldsymbol{V}_D^{'}\boldsymbol{\Omega}_D^{-1}\boldsymbol{V}_D)^{-1}\boldsymbol{V}_D^{'}\boldsymbol{\Omega}_D^{-1}(\boldsymbol{y}_D - \boldsymbol{B}_D\boldsymbol{\beta}_D - \boldsymbol{\mu}_D), (\boldsymbol{V}_D^{'}\boldsymbol{\Omega}_D^{-1}\boldsymbol{V}_D)^{-1}\right).$$

Now, we will derive the full conditionals for μ_{Dl} and σ_{Dl}^2. First, we have

$$\begin{aligned} p(\mu_{Dl} \mid \text{else}) &\propto \exp\left\{-\frac{1}{2\sigma_{Dl}^2}\sum_{j:z_{Dj}=l}\left(y_{Dj} - \boldsymbol{B}_{Dj}\boldsymbol{\beta}_D - \boldsymbol{v}_{Dj}^{'}\boldsymbol{\gamma}_D - \mu_{Dl}\right)^2\right\} \\ &\times \exp\left\{-\frac{1}{2b_{\mu_D}^2}\mu_{Dl}^2\right\}. \end{aligned}$$

Adding the terms in the exponents and ignoring the $-1/2$ for the moment, we can work within the exponential function. Further, defining $\varepsilon_{Dj} = y_{Dj} - \boldsymbol{B}_{Dj}\boldsymbol{\beta}_D - \boldsymbol{v}_{Dj}^{'}\boldsymbol{\gamma}_D$ it follows that

$$\frac{1}{\sigma_{Dl}^2}\left\{\sum_{j:z_{Dj}=l}\varepsilon_{Dj}^2 - 2\sum_{j:z_{Dj}=l}\varepsilon_{Dj}\mu_{Dl} + n_{Dl}\mu_{Dl}^2\right\} + \frac{1}{b_{\mu_D}^2}\mu_{Dl}^2 \propto \left(\frac{n_{Dl}}{\sigma_{Dl}^2} + \frac{1}{b_{\mu_D}^2}\right)\mu_{Dl}^2 - 2\left\{\frac{1}{\sigma_{Dl}^2}\sum_{j:z_{Dj}=l}\varepsilon_{Dj}\right\}\mu_{Dl}.$$

Then, completing the square

$$p(\mu_{Dl} \mid \text{else}) \propto \exp\left\{-\frac{1}{2}\left(\frac{n_{Dl}}{\sigma_{Dl}^2} + \frac{1}{b_{\mu_D}^2}\right)\left[\mu_{Dl} - \left(\frac{n_{Dl}}{\sigma_{Dl}^2} + \frac{1}{b_{\mu_D}^2}\right)^{-1}\left(\frac{1}{\sigma_{Dl}^2}\sum_{j:z_{Dj}=l}\varepsilon_{Dj}\right)\right]^2\right\},$$

which we can recognise as the kernel of a normal distribution with mean $\left(\frac{n_{Dl}}{\sigma^2_{Dl}} + \frac{1}{b^2_{\mu_D}}\right)^{-1} \left(\frac{1}{\sigma^2_{Dl}} \sum_{j:z_{Dj}=l} \varepsilon_{Dj}\right)$ and variance $\left(\frac{n_{Dl}}{\sigma^2_{Dl}} + \frac{1}{b^2_{\mu_D}}\right)^{-1}$, that is,

$$\mu_{Dl} \mid \text{else} \sim \text{N}\left(\frac{\frac{1}{\sigma^2_{Dl}} \sum_{j:z_{Dj}=l} \varepsilon_{Dj}}{\frac{n_{Dl}}{\sigma^2_{Dl}} + \frac{1}{b^2_{\mu_D}}}, \frac{1}{\frac{n_{Dl}}{\sigma^2_{Dl}} + \frac{1}{b^2_{\mu_D}}}\right), \quad \text{for } l = 1, \ldots, L_D.$$

On the other hand, using the same notation, for the variance we have

$$\begin{aligned} p(\sigma^2_{Dl} \mid \text{else}) &\propto (\sigma^2_{Dl})^{-\frac{n_{Dl}}{2}} \exp\left\{-\frac{1}{2\sigma^2_{Dl}} \sum_{j:z_{Dj}=l} (\varepsilon_{Dj} - \mu_{Dl})^2\right\} \\ &\quad \times (\sigma^2_{Dl})^{-a_{\sigma^2_D}-1} \exp\left\{-\frac{b_{\sigma^2_D}}{\sigma^2_{Dl}}\right\} \\ &= (\sigma^2_{Dl})^{-a_{\sigma^2_D}-\frac{n_{Dl}}{2}-1} \exp\left\{-\frac{1}{\sigma^2_{Dl}}\left[b_{\sigma^2_D} + \frac{1}{2} \sum_{j:z_{Dj}=l} (\varepsilon_{Dj} - \mu_{Dl})^2\right]\right\}, \end{aligned}$$

which is the kernel of an inverse-gamma distribution with shape parameter $a_{\sigma^2_D} - \frac{n_{Dl}}{2}$ and scale parameter $b_{\sigma^2_D} + \frac{1}{2} \sum_{j:z_{Dj}=l} (\varepsilon_{Dj} - \mu_{Dl})^2$, i.e.,

$$\sigma^2_{Dl} \mid \text{else} \sim \text{IG}\left(a_{\sigma^2_D} - \frac{n_{Dl}}{2}, b_{\sigma^2_D} + \frac{1}{2} \sum_{j:z_{Dj}=l} (\varepsilon_{Dj} - \mu_{Dl})^2\right), \quad \text{for } l = 1, \ldots, L_D.$$

Now, for the stick-breaking weights, let us first look at the full conditional of V_{D1}

$$\begin{aligned} p(V_{D1} \mid \text{else}) &\propto \prod_{l=1}^{L_D} \left\{V_{Dl} \prod_{t<l} (1 - V_{Dt})\right\}^{n_{Dl}} (1 - V_{D1})^{\alpha_D - 1} \\ &\propto V_{D1}^{n_{D1}} (1 - V_{D1})^{n_{D2}} \ldots (1 - V_{D1})^{n_{DL}} (1 - V_{D1})^{\alpha_D - 1} \\ &= V_{D1}^{n_{D1}} (1 - V_{D1})^{\sum_{l=2}^{L_D} n_{Dl} + \alpha_D - 1}, \end{aligned}$$

which is the kernel of a beta distribution with parameters $n_{D1} + 1$ and $\sum_{l=2}^{L_D} n_{Dl} + \alpha_D$. Given this, we can generalize and assert that

$$V_{Dl} \mid \text{else} \sim \text{Beta}\left(n_{Dl} + 1, \sum_{k=l+1}^{L} n_{Dk} + \alpha_D\right), \quad \text{for } l = 1, \ldots, L_D - 1.$$

Next, we will derive the full conditional for the concentration parameter of the Dirichlet process, that is

$$\begin{aligned} p(\alpha_D \mid \text{else}) &\propto \prod_{l=1}^{L_D-1} \frac{\Gamma(\alpha_D+1)}{\Gamma(\alpha_D)\Gamma(1)}(1-V_{Dl})^{\alpha_D-1}\alpha_D^{a_\alpha-1}\exp\{-b_\alpha\alpha_D\} \\ &= \alpha_D^{a_\alpha+L_D-2}\exp\left\{\log\left[\prod_{l=1}^{L_D-1}(1-V_{Dl})\right]^{\alpha_D-1} - b_\alpha\alpha_D\right\} \\ &\propto \alpha_D^{a_\alpha+L_D-2}\exp\left\{-\left[b_\alpha - \sum_{l=1}^{L_D-1}\log(1-V_{Dl})\right]\alpha_D\right\}, \end{aligned}$$

which is clearly the kernel of a gamma distribution with shape $a_\alpha + L_D - 1$ and rate given by $b_\alpha - \sum_{l=1}^{L_D-1}\log(1-V_{Dl})$, i.e.,

$$\alpha_D \mid \text{else} \sim \Gamma\left(a_\alpha + L_D - 1, b_\alpha - \sum_{l=1}^{L_D-1}\log(1-V_{Dl})\right).$$

For the additional variance parameter, which controls the amount of smoothness of the fitted functions in our model, the full conditional is given by

$$\begin{aligned} p(\tau^2_{Dk} \mid \text{else}) &\propto (\tau^2_{Dk})^{-\frac{\text{rank}(\boldsymbol{K}_{Dk})}{2}}\exp\left\{-\frac{1}{2\tau^2_{Dk}}\boldsymbol{\beta}'_{Dk}\boldsymbol{K}_{Dk}\boldsymbol{\beta}_{Dk}\right\}(\tau^2_{Dk})^{-a_{\tau^2_k}-1}\exp\left\{-\frac{b_{\tau^2_k}}{\tau^2_{Dk}}\right\} \\ &= (\tau^2_{Dk})^{-a_{\tau^2_k}-\frac{\text{rank}(\boldsymbol{K}_{Dk})}{2}-1}\exp\left\{-\frac{1}{\tau^2_{Dk}}\left[b_{\tau^2_k}+\frac{1}{2}\boldsymbol{\beta}'_{Dk}\boldsymbol{K}_{Dk}\boldsymbol{\beta}_{Dk}\right]\right\}, \end{aligned}$$

which is the kernel of an inverse-gamma distribution, then

$$\tau^2_{Dk} \mid \text{else} \sim \text{IG}\left(a_{\tau^2_k} + \frac{\text{rank}(\boldsymbol{K}_{Dk})}{2}, b_{\tau^2_k} + \frac{1}{2}\boldsymbol{\beta}'_{Dk}\boldsymbol{K}_{Dk}\boldsymbol{\beta}_{Dk}\right), \quad k = 1,\ldots,p.$$

Finally, for the latent variables we have

$$\begin{aligned} \mathbb{P}(z_{Dj} = l \mid \text{else}) &= \frac{p(y_{Dj} \mid z_{Dj}=l, \boldsymbol{\theta}_D)\mathbb{P}(z_{Dj}=l)}{\sum_{\ell=1}^{L_D} p(y_{Dj} \mid z_{Dj}=\ell, \boldsymbol{\theta}_D)\mathbb{P}(z_{Dj}=\ell)}, \quad \text{for } l = 1,\ldots,L_D \\ &= \frac{\omega_{Dl}\phi\left(y_{Dj} \mid \boldsymbol{B}_{Dj}\boldsymbol{\beta}_D + \boldsymbol{v}'_{Dj}\boldsymbol{\gamma}_D + \mu_{Dl}, \sigma^2_{Dl}\right)}{\sum_{\ell=1}^{L_D}\omega_{D\ell}\phi\left(y_{Dj} \mid \boldsymbol{B}_{Dj}\boldsymbol{\beta}_D + \boldsymbol{v}'_{Dj}\boldsymbol{\gamma}_D + \mu_{D\ell}, \sigma^2_{D\ell}\right)} \\ &= \pi_{Dj,l}. \end{aligned}$$

Therefore, the full conditional for the latent variables is multinomial with probabilities given by $\boldsymbol{\pi}_{Dj} = (\pi_{Dj,1}, \ldots, \pi_{Dj,L_D})$, that is,

$$z_{Dj} \mid \text{else} \sim \text{Mult}(1, \boldsymbol{\pi}_{Dj}), \quad j = 1, \ldots, n_D.$$

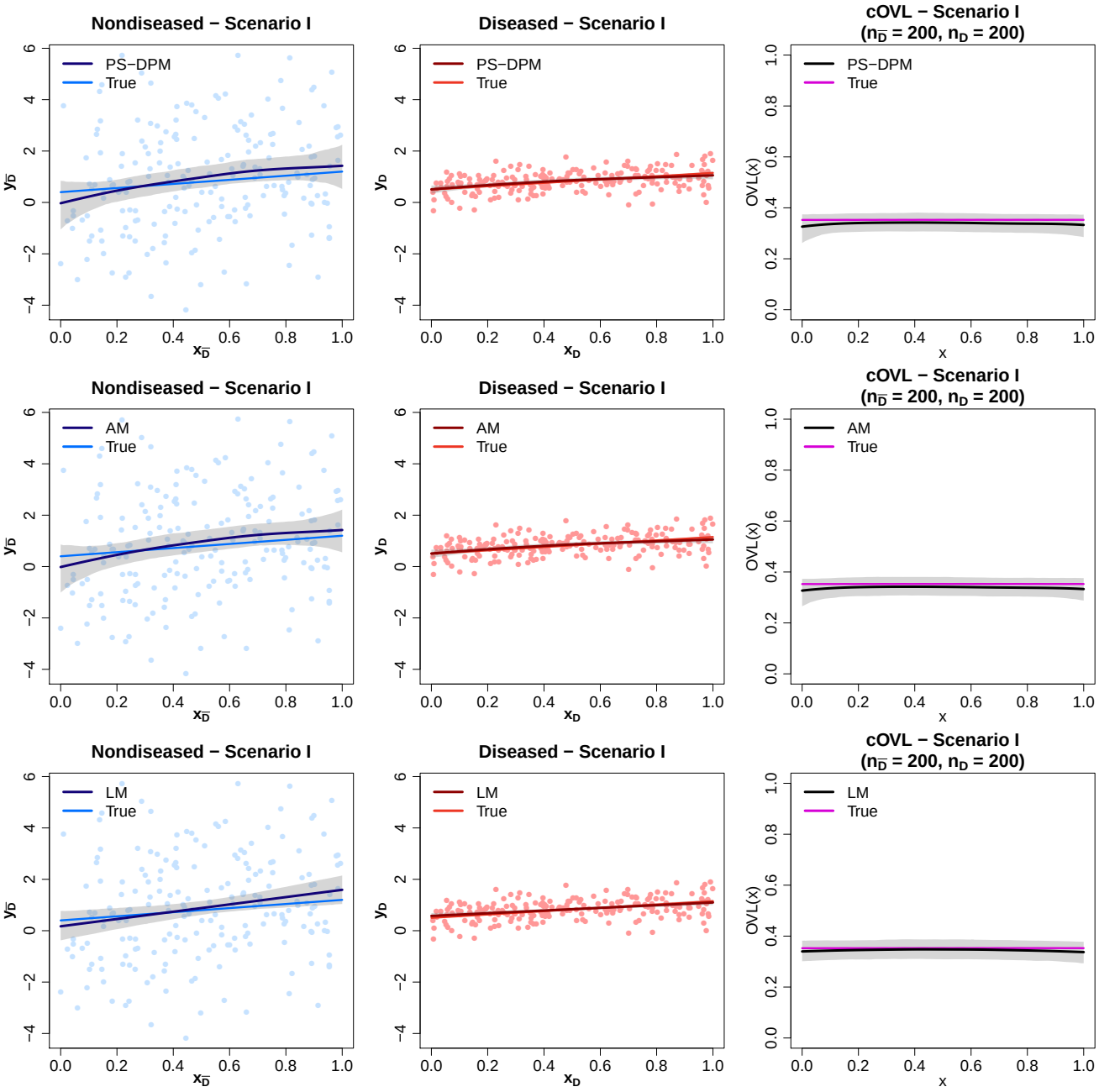

Figure A.1: Results for Sceario I. Mean function: posterior mean and 95% pointwise credible bands (grey area) for nondiseased (first column) and diseased (second column) along with the true ones for each of the three scenarios. Covariate-specific OVL (third column): posterior mean and 95% pointwise credible bands along with the true OVL.

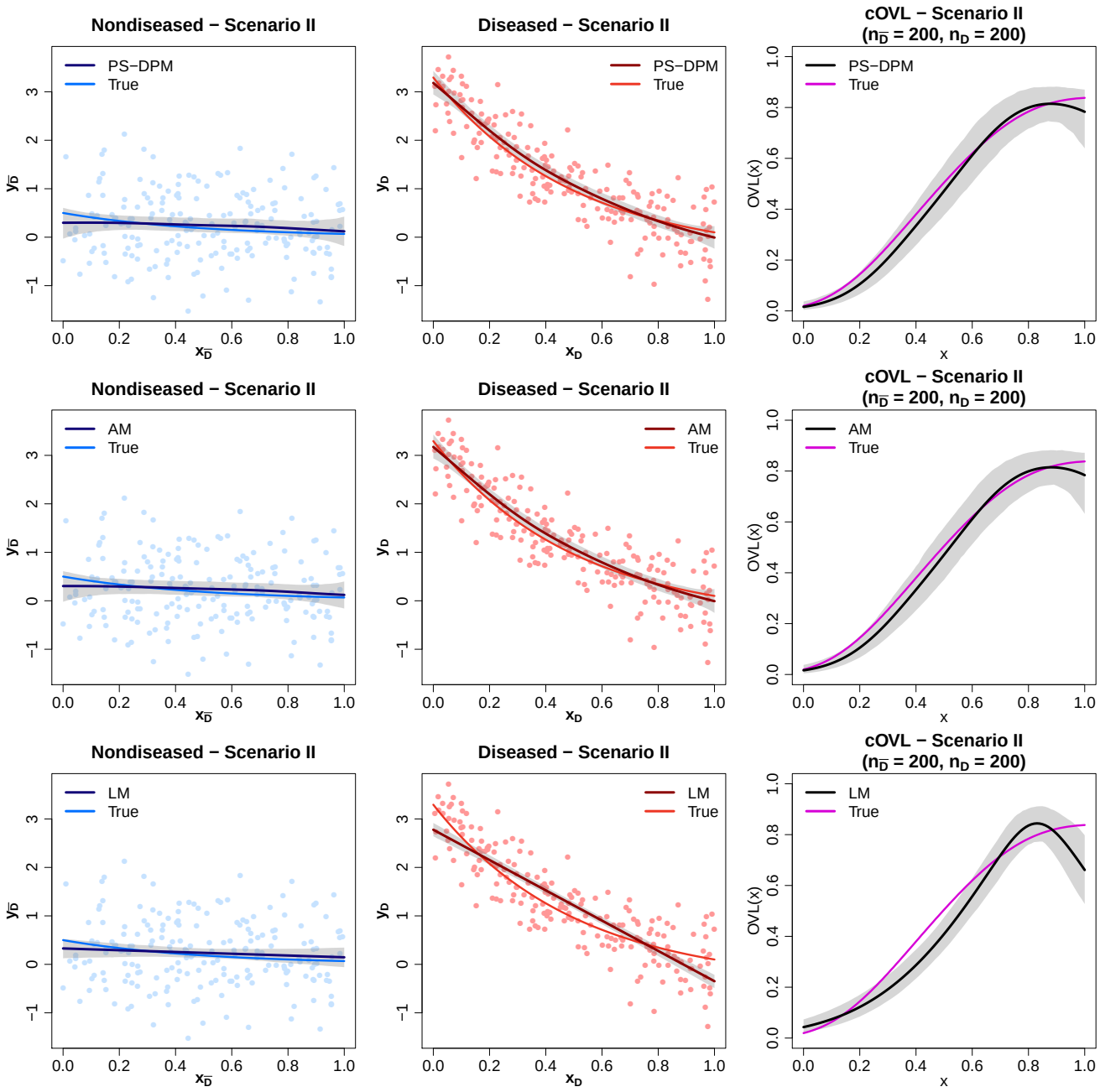

Figure A.2: Results for Scenario II. Mean function: posterior mean and 95% pointwise credible bands (grey area) for nondiseased (first column) and diseased (second column) along with the true ones for each of the three scenarios. Covariate-specific OVL (third column): posterior mean and 95% pointwise credible bands along with the true OVL.

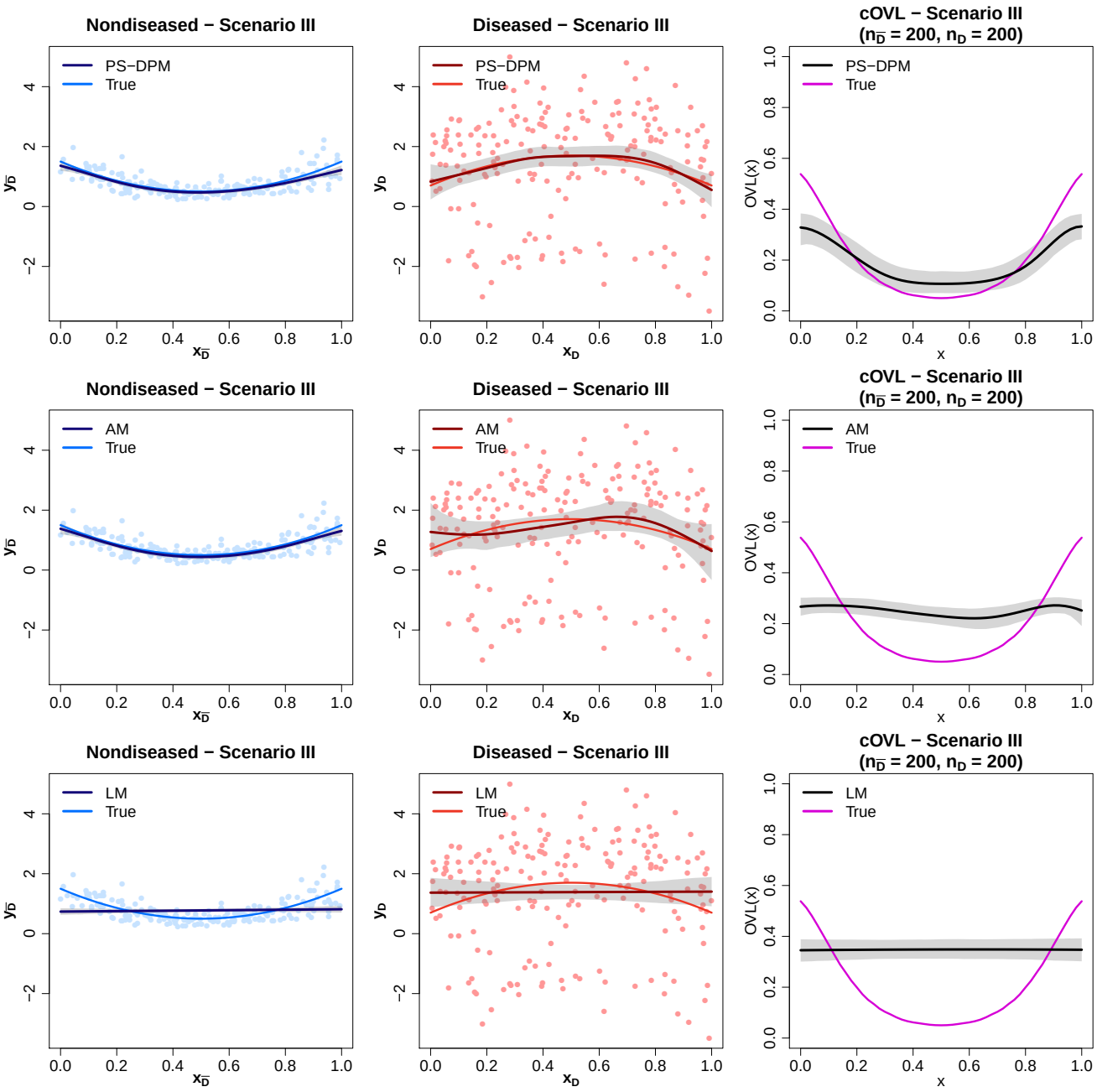

Figure A.3: Results for Scenario III. Mean function: posterior mean and 95% pointwise credible bands (grey area) for nondiseased (first column) and diseased (second column) along with the true ones for each of the three scenarios. Covariate-specific OVL (third column): posterior mean and 95% pointwise credible bands along with the true OVL.

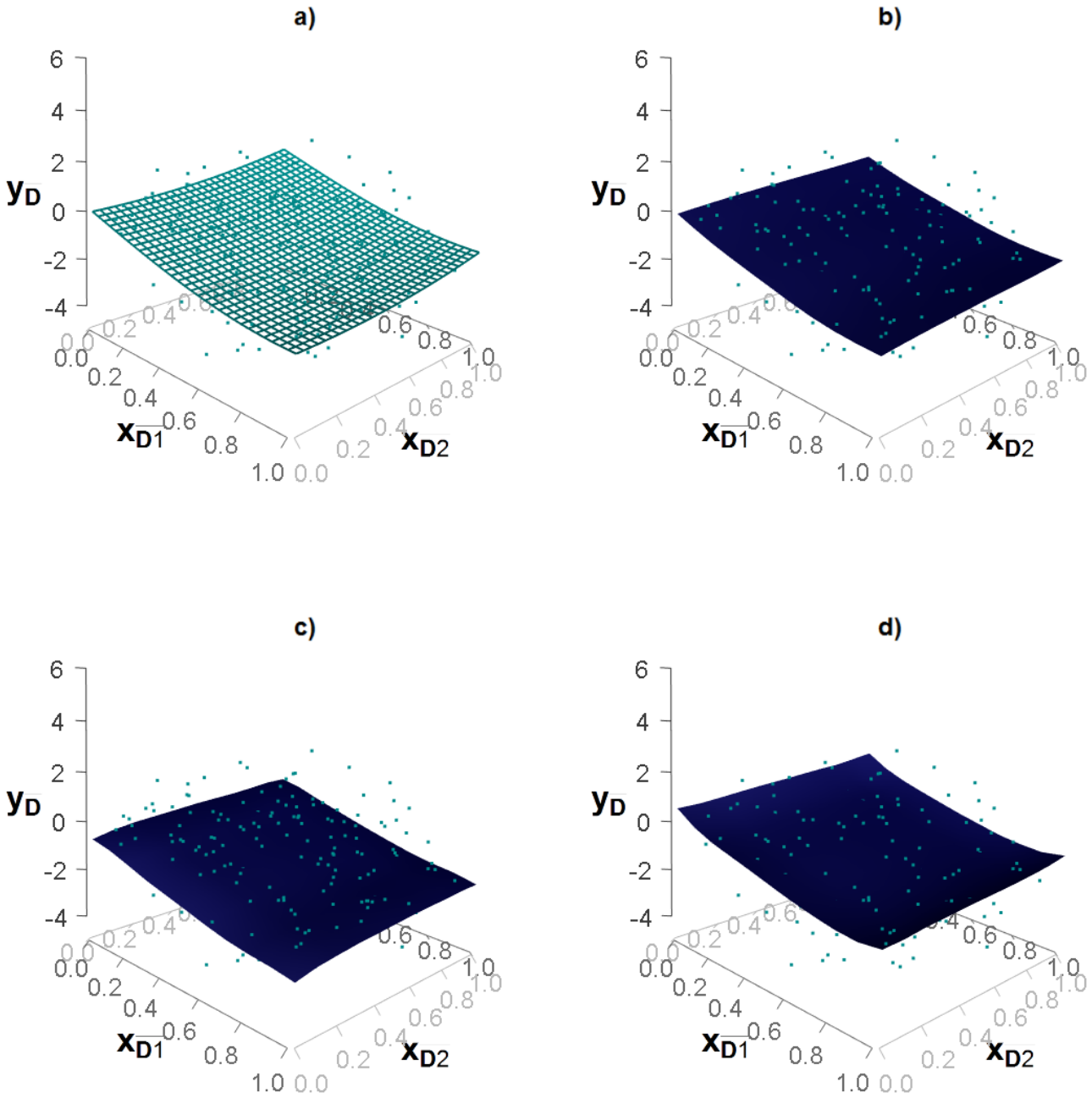

Figure A.4: Results for Scenario IV. a) True mean function for the nondiseased group; b) posterior mean approximation to the surface in a) using PS-DPM; c) and d) 95% pointwise credible surfaces.

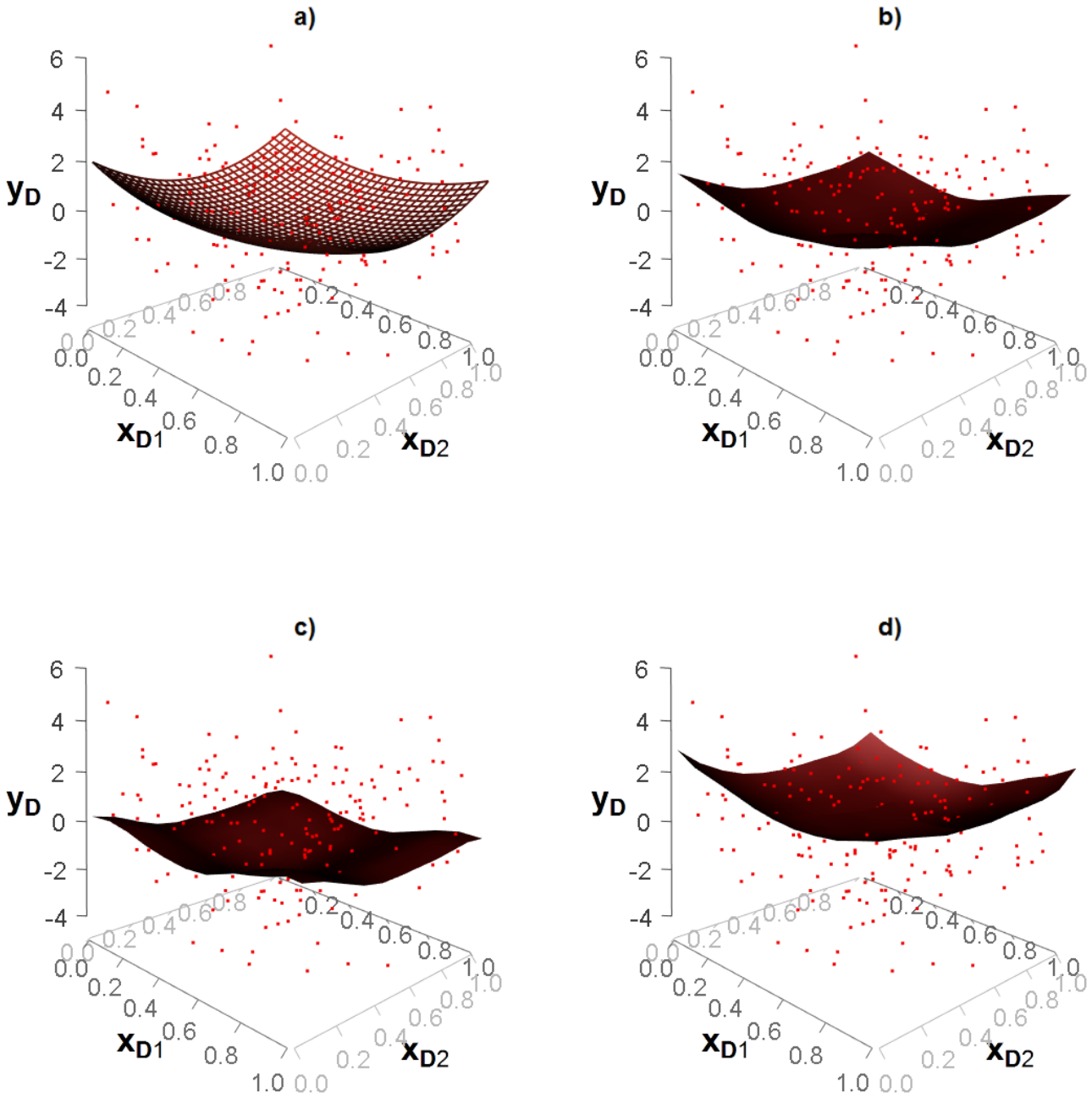

Figure A.5: Results for Scenario IV. a) True mean function for the diseased group; b) posterior mean approximation to the surface in a) using PS-DPM; c) and d) 95% pointwise credible surfaces.

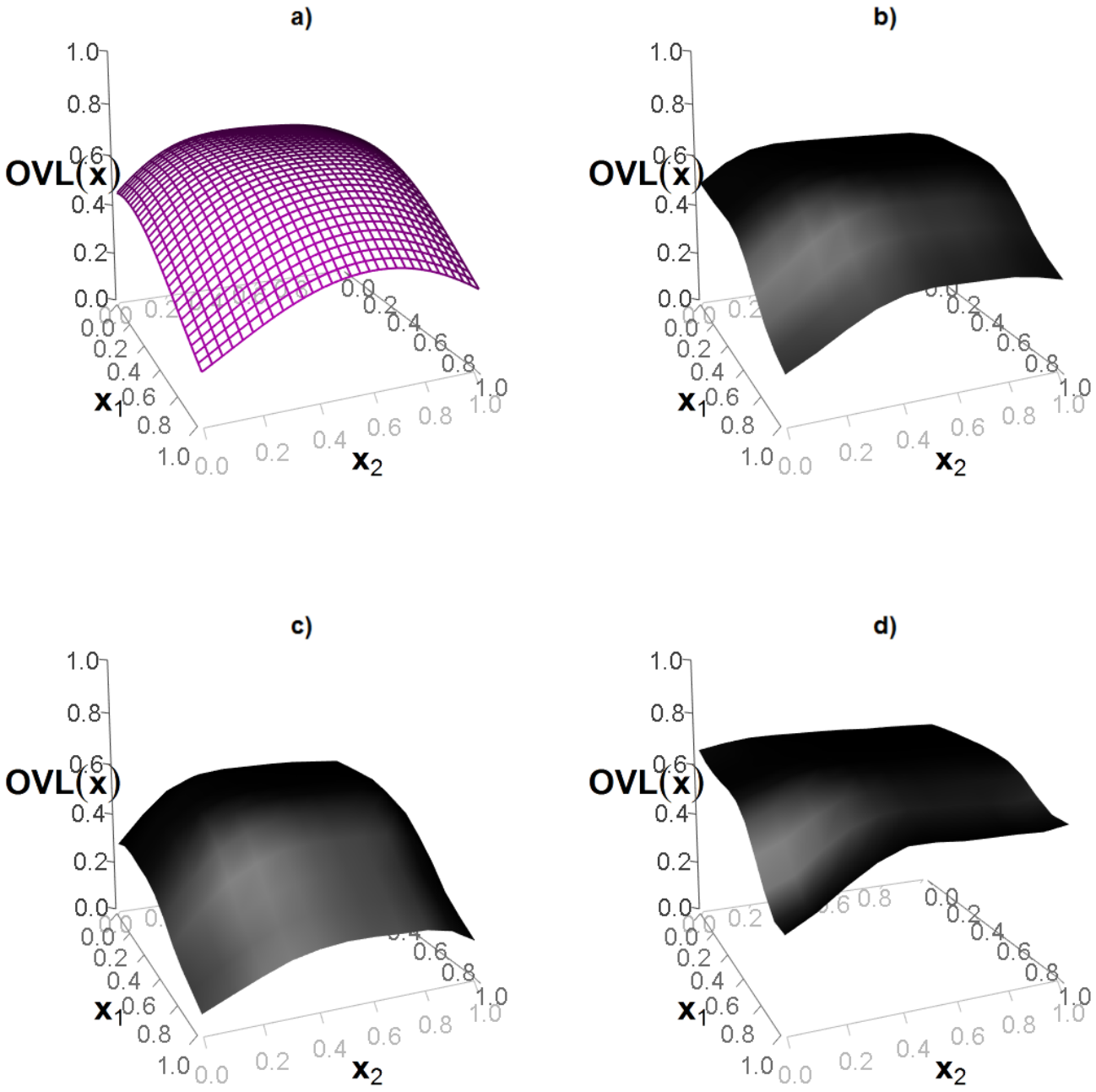

Figure A.6: Results for Scenario IV. a) True cOVL; b) posterior mean approximation to the surface in a) using PS-DPM; c) and d) 95% pointwise credible surfaces.

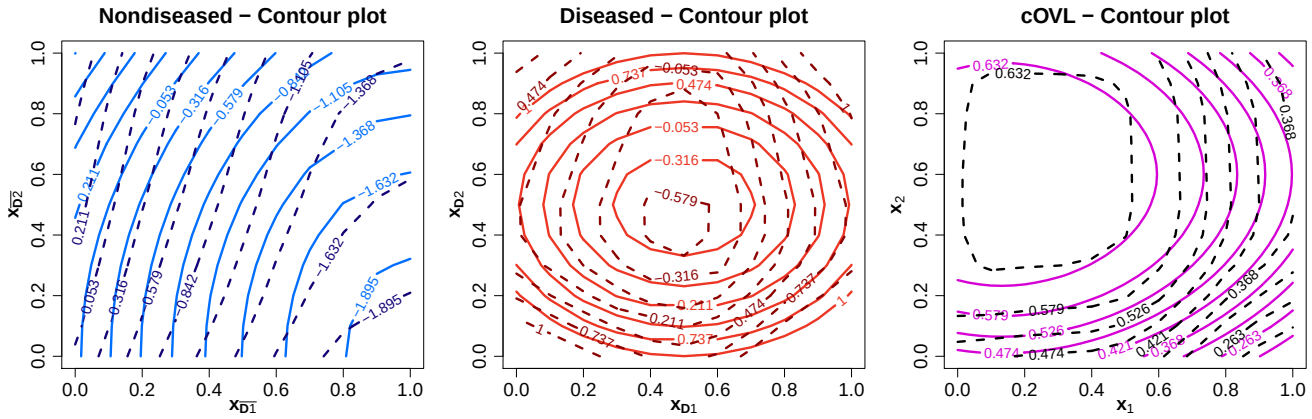

Figure A.7: Results for Scenario IV. Mean function contours: posterior mean for nondiseased (left) and diseased (center) along with the true ones. Covariate-specific OVL contours (right): posterior mean along with the true ones.

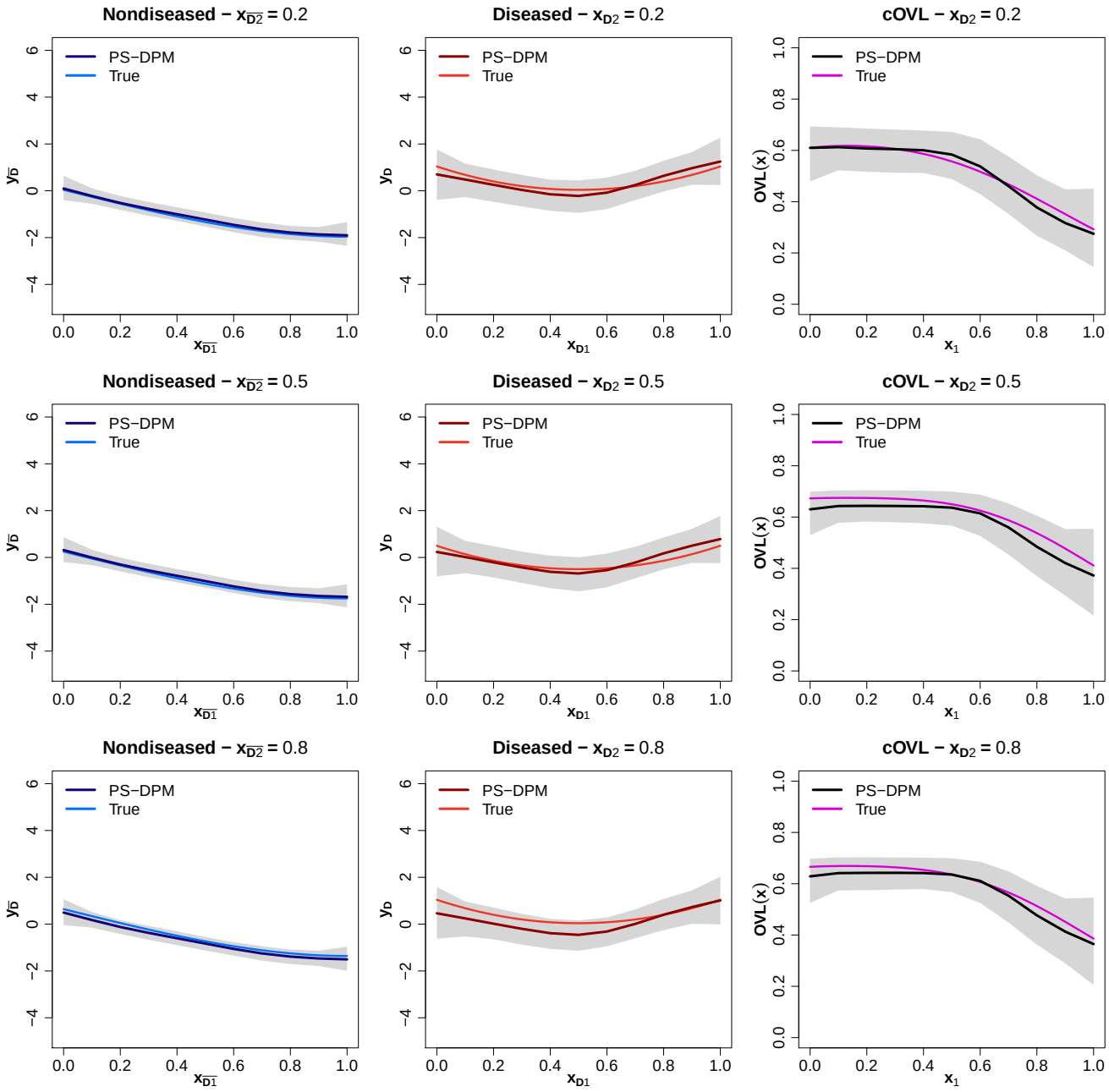

Figure A.8: Results for Scenario IV. Mean function: posterior mean and 95% pointwise credible bands (grey area) for nondiseased (first column) and diseased (second column) along with the true ones given three particular values of the second covariate ($x_2 = 0.2$, $x_2 = 0.5$, $x_2 = 0.8$). Covariate-specific OVL (third column): posterior mean and 95% pointwise credible bands along with the true cOVL for those particular values of x_2.

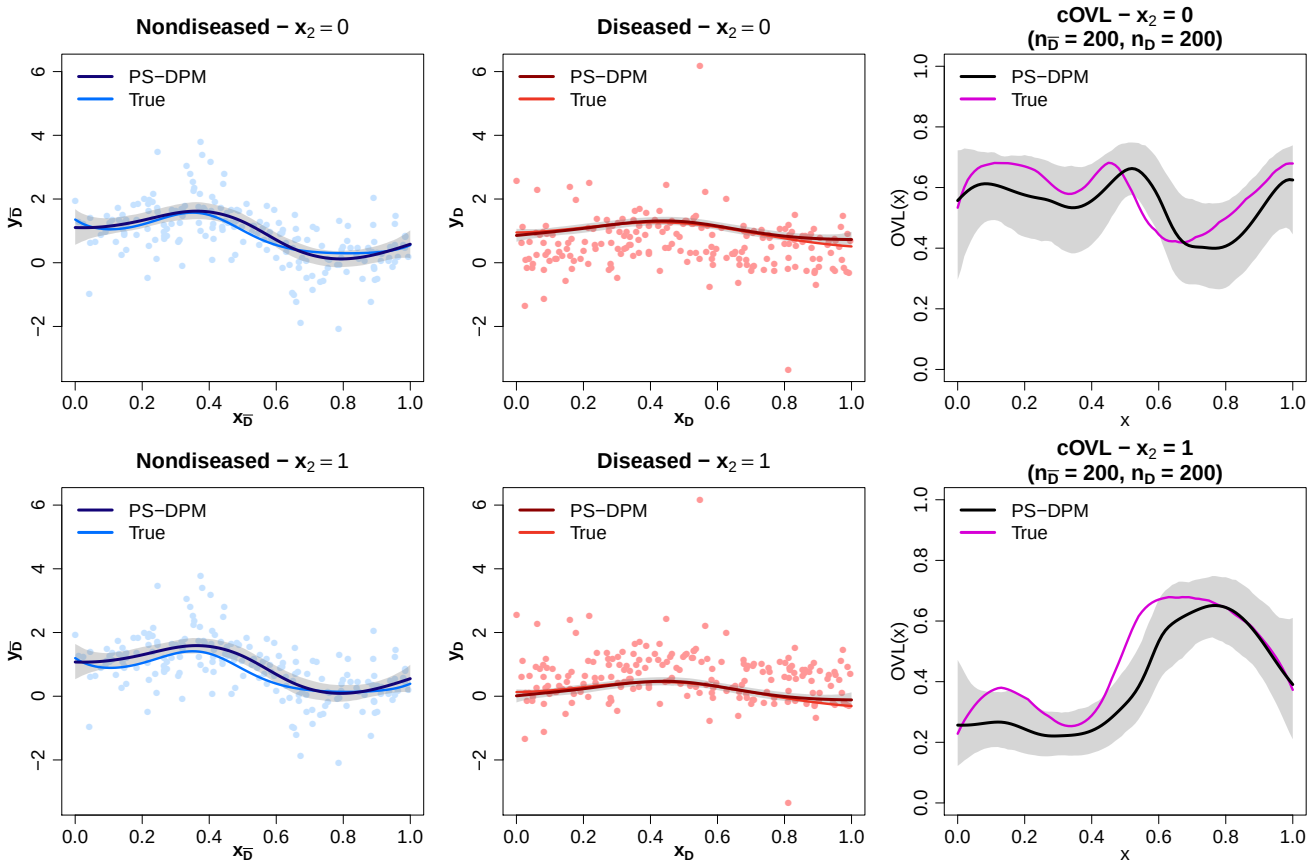

Figure A.9: Results for Scenario V. Mean function: posterior mean and 95% pointwise credible bands (grey area) for nondiseased (first column) and diseased (second column) along with the true ones when $x_2 = 0$ (top row) and $x_2 = 1$ (bottom row). Covariate-specific OVL (third column): posterior mean and 95% pointwise credible bands along with the true cOVL.

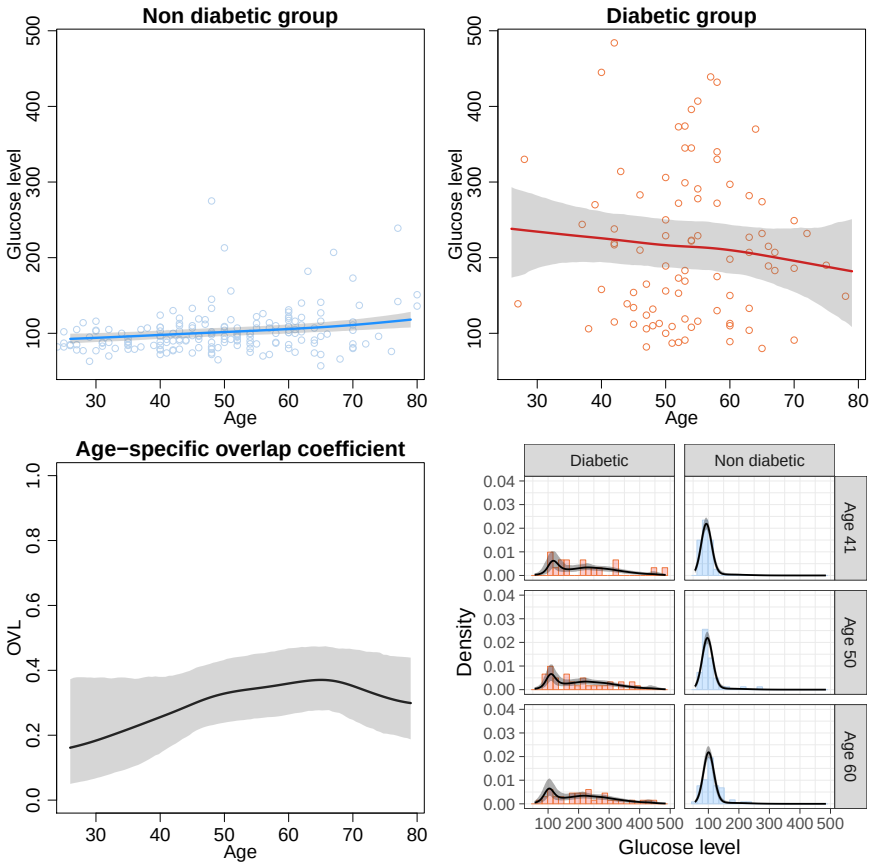

Figure A.10: Top row: Mean function: posterior mean and 95% pointwise credible bands (grey area) for the non diabetic (left) and diabetic (right) groups. Bottom row: Conditional histograms and densities (posterior mean along with 95% credible bands) for three specific ages in the two groups. Covariate-specific OVL: posterior mean and 95% pointwise credible bands.

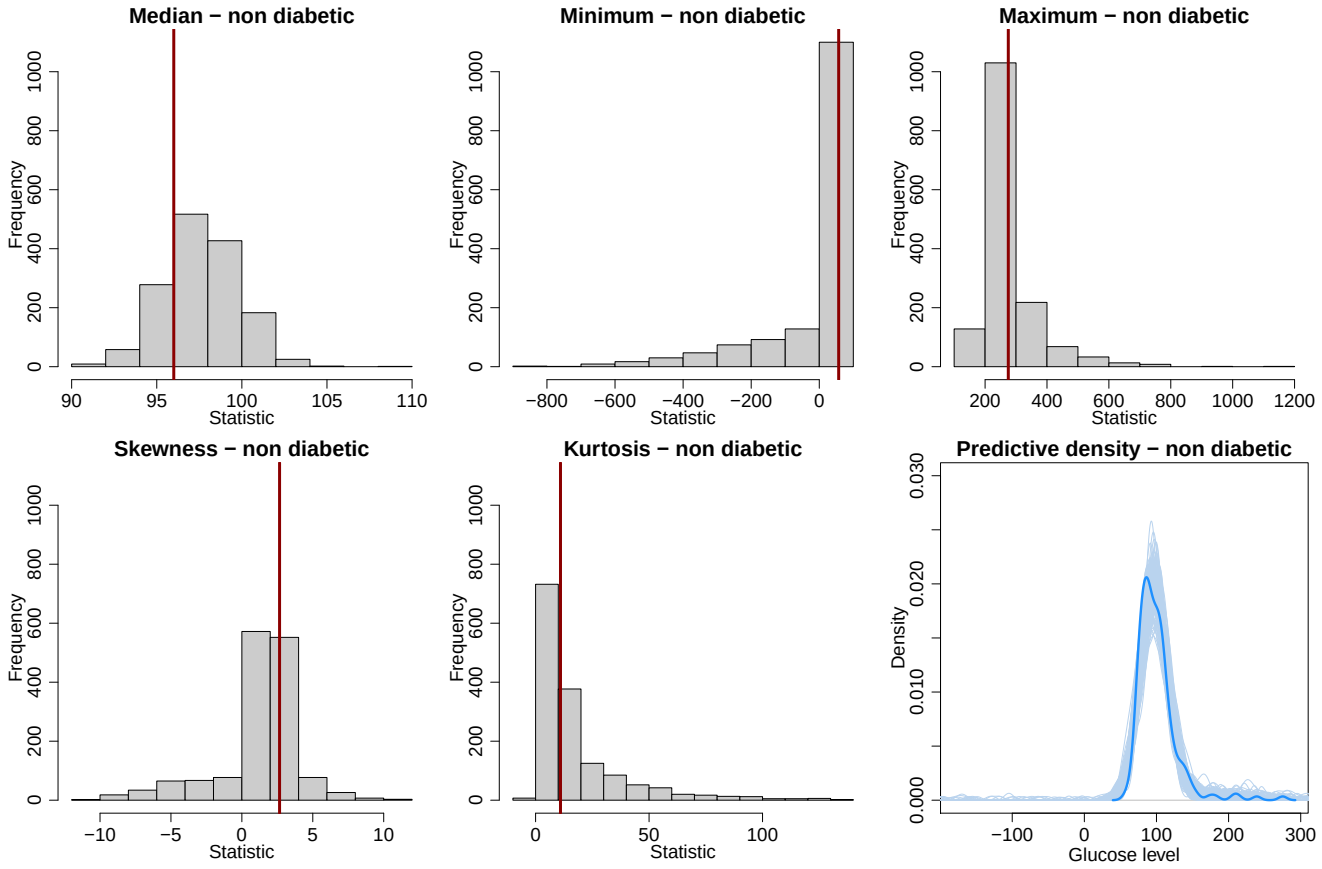

Figure A.11: Predictive checks - non diabetic group. Histograms of different test statistics of the replicated data along with the observed value (vertical line). Predictive densities of the replicated simulations along with the observed data density (thick blue line).

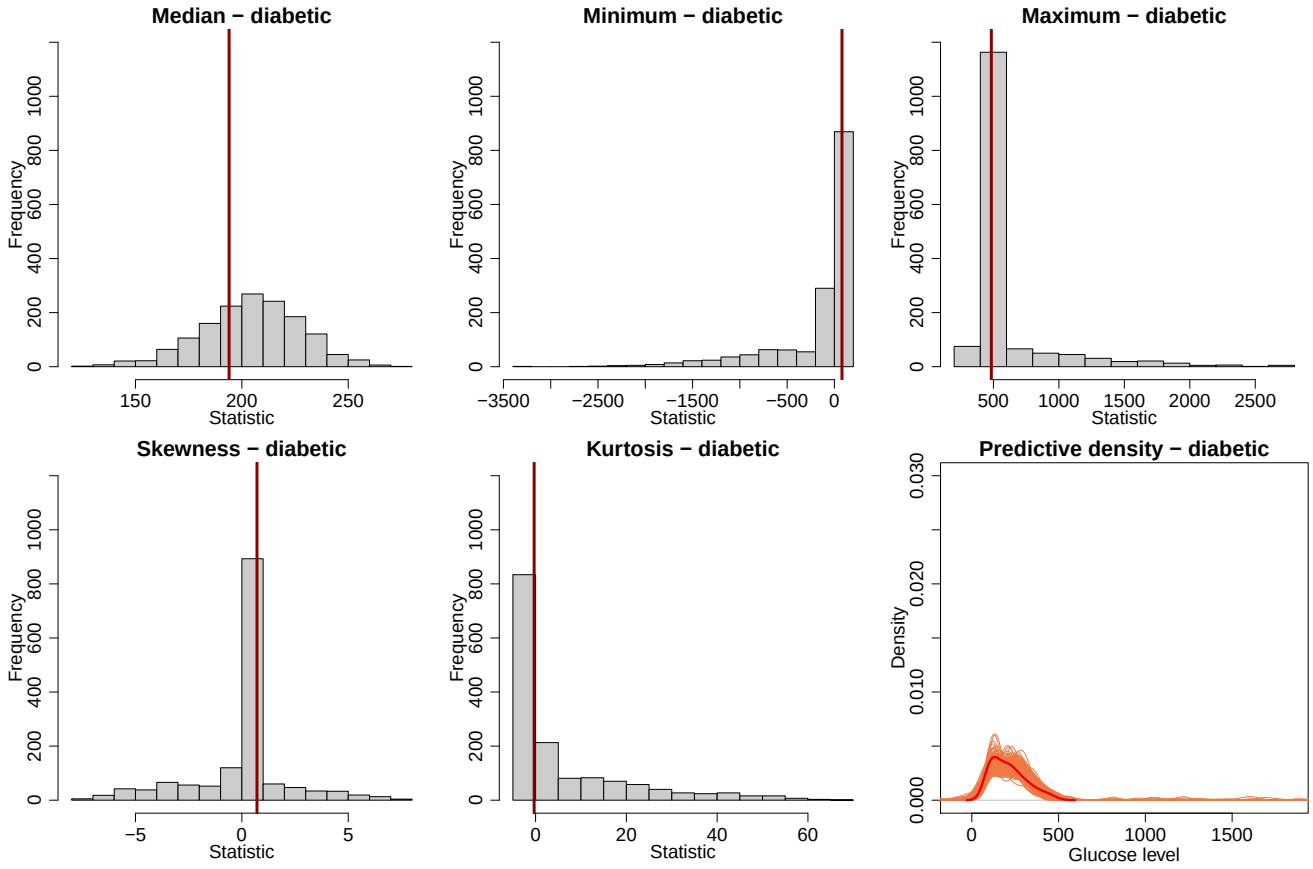

Figure A.12: Predictive checks - diabetic group. Histograms of different test statistics of the replicated data along with the observed value (vertical line). Predictive densities of the replicated simulations along with the observed data density (thick red line).

Bibliography

Berry, D. A. and Stangl, D. (2018), *Bayesian Biostatistics*, CRC Press.

Blackwell, D. and MacQueen, J. B. (1973), 'Ferguson Distributions Via Polya Urn Schemes', *The Annals of Statistics* **1**(2), 353 – 355.

Blennow, K. and Zetterberg, H. (2018), 'Biomarkers for Alzheimer's disease: current status and prospects for the future', *Journal of Internal Medicine* **284**(6), 643–663.

Bolstad, W. M. and Curran, J. M. (2016), *Introduction to Bayesian Statistics*, John Wiley & Sons.

Branscum, A. J., Johnson, W. O., Hanson, T. E. and Gardner, I. A. (2008), 'Bayesian semiparametric ROC curve estimation and disease diagnosis', *Statistics in Medicine* **27**(13), 2474–2496.

Brooks, S. P. and Gelman, A. (1998), 'General methods for monitoring convergence of iterative simulations', *Journal of Computational and Graphical Statistics* **7**(4), 434–455.

Celeux, G., Forbes, F., Robert, C. P. and Titterington, D. M. (2006), 'Deviance information criteria for missing data models', *Bayesian Analysis* **1**(4), 651–673.

Chan, J. C., Henderson, D. J., Parmeter, C. F. and Tobias, J. L. (2017), 'Nonparametric estimation in economics: Bayesian and frequentist approaches', *Wiley Interdisciplinary Reviews: Computational Statistics* **9**(6), e1406.

Chen, K., Ayutyanont, N., Langbaum, J. B., Fleisher, A. S., Reschke, C., Lee, W., Liu, X., Bandy, D., Alexander, G. E., Thompson, P. M. et al. (2011), 'Characterizing Alzheimer's disease using a hypometabolic convergence index', *Neuroimage* **56**(1), 52–60.

Cheng, L., Ramchandran, S., Vatanen, T., Lietzén, N., Lahesmaa, R., Vehtari, A. and Lähdesmäki, H. (2019), 'An additive Gaussian process regression model for interpretable non-parametric analysis of longitudinal data', *Nature Communications* **10**(1), 1–11.

Chib, S. and Greenberg, E. (2010), 'Additive cubic spline regression with Dirichlet process mixture errors', *Journal of Econometrics* **156**(2), 322–336.

Christensen, R., Johnson, W., Branscum, A. and Hanson, T. E. (2011), *Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians*, CRC press.

Clemons, T. E. and Bradley Jr., E. L. (2000), 'A nonparametric measure of the overlapping coefficient', *Computational Statistics & Data Analysis* **34**(1), 51–61.

de Carvalho, M., Barney, B. J. and Page, G. L. (2020), 'Affinity-based measures of biomarker performance evaluation', *Statistical Methods in Medical Research* **29**(3), 837–853.

De Iorio, M., Müller, P., Rosner, G. L. and MacEachern, S. N. (2004), 'An ANOVA model for dependent random measures', *Journal of the American Statistical Association* **99**(465), 205–215.

DeIorio, M. and Robert, C. P. (2002), 'Discussion of Spiegelhalter et al.', *Journal of the Royal Statistical Society, Series B* **64**, 629–630.

Devroye, L. (1986), *Non-Uniform Random Variate Generation*, Springer-Verlag, New York.

Dunn, P. K. and Smyth, G. K. (1996), 'Randomized quantile residuals', *Journal of Computational and Graphical Statistics* **5**(3), 236–244.

Eckhardt, R. (1987), 'Stan Ulam, John von Neumann, and the Monte Carlo method', *Los Alamos Science* **15**(30), 131–136.

Efron, B. (1979), 'Bootstrap methods: another look at the jackknife', *The Annals of Statistics* **7**(1), 1–26.

Eilers, P. H. and Marx, B. D. (1996), 'Flexible smoothing with B-splines and penalties', *Statistical Science* pp. 89–102.

Epifani, I., MacEachern, S. N., Peruggia, M. et al. (2008), 'Case-deletion importance sampling estimators: Central limit theorems and related results', *Electronic Journal of Statistics* **2**, 774–806.

Erkanli, A., Sung, M., Jane Costello, E. and Angold, A. (2006), 'Bayesian semi-parametric ROC analysis', *Statistics in Medicine* **25**(22), 3905–3928.

Escobar, M. D. (1988), Estimating the means of several normal populations by nonparametric estimation of the distribution of the means, PhD thesis, Department of Statistics, Yale University.

Escobar, M. D. (1994), 'Estimating normal means with a Dirichlet process prior', *Journal of the American Statistical Association* **89**(425), 268–277.

Escobar, M. D. and West, M. (1995), 'Bayesian density estimation and inference using mixtures', *Journal of the American Statistical Association* **90**(430), 577–588.

Fahrmeir, L., Kneib, T., Lang, S. and Marx, B. (2013), *Regression: models, methods and applications*, Springer Science & Business Media.

Faraggi, D. (2003), 'Adjusting receiver operating characteristic curves and related indices for covariates', *Journal of the Royal Statistical Society: Series D* **52**(2), 179–192.

Ferguson, T. S. (1973), 'A Bayesian analysis of some nonparametric problems', *The Annals of Statistics* **1**(2), 209–230.

Ferguson, T. S. (1974), 'Prior Distributions on Spaces of Probability Measures', *The Annals of Statistics* **2**(4), 615 – 629.

Ferguson, T. S. (1983), Bayesian density estimation by mixtures of normal distributions, *in* D. Siegmund, J. Rustage and G. G. Rizvi, eds, 'Recent advances in Statistics. Papers in honor of Herman Chernoff on his sixtieth birthday', Bibliohound, pp. 287–302.

Frühwirth-Schnatter, S. and Malsiner-Walli, G. (2019), 'From here to infinity: sparse finite versus Dirichlet process mixtures in model-based clustering', *Advances in Data Analysis and Classification* **13**(1), 33–64.

Gasparini, M. (1995), 'Exact multivariate Bayesian bootstrap distributions of moments', *The Annals of Statistics* **23**(3), 762–768.

Geisser, S. and Eddy, W. F. (1979), 'A predictive approach to model selection', *Journal of the American Statistical Association* **74**(365), 153–160.

Gelfand, A. E. and Dey, D. K. (1994), 'Bayesian model choice: asymptotics and exact calculations', *Journal of the Royal Statistical Society, Series B* **56**(3), 501–514.

Gelfand, A. E. and Smith, A. F. (1990), 'Sampling-based approaches to calculating marginal densities', *Journal of the American Statistical Association* **85**(410), 398–409.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2013), *Bayesian Data Analysis*, CRC press.

Gelman, A., Hwang, J. and Vehtari, A. (2014), 'Understanding predictive information criteria for Bayesian models', *Statistics and Computing* **24**(6), 997–1016.

Gelman, A., Rubin, D. B. et al. (1992), 'Inference from iterative simulation using multiple sequences', *Statistical Science* **7**(4), 457–472.

Geman, S. and Geman, D. (1984), 'Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAMI-6**(6), 721–741.

Geweke, J. (1992), Evaluating the Accuracy of Sampling Based Approaches to Calculating Posterior Moments, *in* J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds, 'Bayesian Statistics', Vol. 4, Oxford University Press, New York.

Gu, J., Ghosal, S. and Roy, A. (2008), 'Bayesian bootstrap estimation of ROC curve', *Statistics in Medicine* **27**(26), 5407–5420.

Hanson, T. E. (2006), 'Modeling censored lifetime data using a mixture of gammas baseline', *Bayesian Analysis* **1**(3), 575–594.

Hastings, W. K. (1970), 'Monte Carlo sampling methods using Markov chains and their applications', *Biometrika* **57**(1), 97–109.

Hjort, N. L. (2003), Topics in nonparametric Bayesian statistics, *in* P. Green, N. Hjort and S. Richardson, eds, 'Highly Structured Stochastic Systems', Oxford University Press. Oxford.

Hjort, N. L., Holmes, C., Müller, P. and Walker, S. G. (2010), *Bayesian Nonparametrics*, Vol. 28, Cambridge University Press.

Hoff, P. D. (2009), *A First Course in Bayesian Statistical Methods*, Springer.

Inácio de Carvalho, V., Jara, A. and de Carvalho, M. (2015), Bayesian nonparametric approaches for roc curve inference, *in* 'Nonparametric Bayesian Inference in Biostatistics', Springer, pp. 327–344.

Inácio de Carvalho, V., Jara, A., Hanson, T. E. and de Carvalho, M. (2013), 'Bayesian nonparametric ROC regression modeling', *Bayesian Analysis* **8**(3), 623–646.

Inácio, V. C. F. (2012), Semiparametric and Nonparametric Modeling of Diagnostic Data, PhD thesis, Universidade de Lisboa, Facultade de Ciências.

Inácio, V., Rodríguez-Álvarez, M. X. and Gayoso-Diz, P. (2020), 'Statistical Evaluation of Medical Tests', *Annual Review of Statistics and its Application* .

Inman, H. F. and Bradley Jr, E. L. (1989), 'The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities', *Communications in Statistics-Theory and Methods* **18**(10), 3851–3874.

Ishwaran, H. and James, L. F. (2001), 'Gibbs sampling methods for stick-breaking priors', *Journal of the American Statistical Association* **96**(453), 161–173.

Ishwaran, H. and Zarepour, M. (2000), 'Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models', *Biometrika* **87**(2), 371–390.

Jack, C. R., Bennett, D. A., Blennow, K., Carrillo, M. C., Feldman, H. H., Frisoni, G. B., Hampel, H., Jagust, W. J., Johnson, K. A., Knopman, D. S., Petersen, R. C., Scheltens, P., Sperling, R. A.

and Dubois, B. (2016), 'A/T/N: an unbiased descriptive classification scheme for Alzheimer disease biomarkers', *Neurology* **87**(5), 539–547.

Johnson, W. O. and de Carvalho, M. (2015), Bayesian Nonparametric Biostatistics, *in* R. Mitra and P. Müller, eds, 'Nonparametric Bayesian Inference in Biostatistics', Frontiers in Probability and the Statistical Sciences. Springer, Cham, pp. 15–54.

Lang, S. and Brezger, A. (2004), 'Bayesian P-splines', *Journal of Computational and Graphical Statistics* **13**(1), 183–212.

Lijoi, A., Prünster, I. et al. (2010), 'Models beyond the Dirichlet process', *Bayesian Nonparametrics* **28**(80), 342.

Liu, J. S. (1996), 'Nonparametric hierarchical Bayes via sequential imputations', *The Annals of Statistics* pp. 911–930.

Lo, A. Y. (1984), 'On a class of Bayesian nonparametric estimates: I. Density estimates', *The Annals of Statistics* **12**(1), 351–357.

Lunn, D. J., Thomas, A., Best, N. and Spiegelhalter, D. (2000), 'WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility', *Statistics and Computing* **10**(4), 325–337.

MacEachern, S. N. (1999), Dependent nonparametric processes, *in* 'ASA Proceedings of the Section on Bayesian Statistical Science, Alexandria, VA.', American Statistical Association.

MacEachern, S. N. (2000), Dependent Dirichlet processes, Technical report, Department of Statistics, Ohio State University.

Martinez-Camblor, P., Corral, N., Rey, C., Pascual, J. and Cernuda-Morollón, E. (2017), 'Receiver operating characteristic curve generalization for non-monotone relationships', *Statistical Methods in Medical Research* **26**(1), 113–123.

Martínez-Camblor, P. and Pardo-Fernández, J. C. (2019), 'Parametric estimates for the receiver operating characteristic curve generalization for non-monotone relationships', *Statistical Methods in Medical Research* **28**(7), 2032–2048.

McElreath, R. (2018), *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*, Chapman and Hall/CRC.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953), 'Equation of state calculations by fast computing machines', *The Journal of Chemical Physics* **21**(6), 1087–1092.

Moyé, L. A. (2016), *Elementary Bayesian Biostatistics*, CRC Press.

Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C., Jagust, W., Trojanowski, J. Q., Toga, A. W. and Beckett, L. (2005), 'The Alzheimer's Disease Neuroimaging Initiative', *Neuroimaging Clinics* **15**(4), 869–877.

Müller, P., Erkanli, A. and West, M. (1996), 'Bayesian curve fitting using multivariate normal mixtures', *Biometrika* **83**(1), 67–79.

Müller, P. and Mitra, R. (2013), 'Bayesian nonparametric inference–Why and How', *Bayesian Analysis* **8**(2), 269–302.

Müller, P. and Quintana, F. A. (2004), 'Nonparametric Bayesian Data Analysis', *Statistical Science* pp. 95–110.

Müller, P., Quintana, F. A., Jara, A. and Hanson, T. (2015), *Bayesian Nonparametric Data Analysis*, Springer.

Núñez-Antonio, G., Mendoza, M., Contreras-Cristán, A., Gutiérrez-Peña, E. and Mendoza, E. (2018), 'Bayesian nonparametric inference for the overlap of daily animal activity patterns', *Environmental and Ecological Statistics* **25**(4), 471–494.

Pastore, M. and Calcagnì, A. (2019), 'Measuring distribution similarities between samples: a distribution-free overlapping index', *Frontiers in Psychology* **10**, 1089.

Pepe, M. S. (1998), 'Three Approaches to Regression Analysis of Receiver Operating Characteristic Curves for Continuous Test Results', *Biometrics* pp. 124–135.

Pepe, M. S. (2003), *The Statistical Evaluation of Medical Tests for Classification and Prediction*, Oxford University Press.

Pepe, M. S., Longton, G., Anderson, G. L. and Schummer, M. (2003), 'Selecting differentially expressed genes from microarray experiments', *Biometrics* **59**(1), 133–142.

Pérez-Fernández, S., Martínez-Camblor, P., Filzmoser, P. and Corral, N. (2021), 'Visualizing the decision rules behind the ROC curves: understanding the classification process', *AStA Advances in Statistical Analysis* **105**(1), 135–161.

Plummer, M. (2003), JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling, *in* K. Hornik, F. Leisch and A. Zeileis, eds, 'Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)', March 20-22, Vienna, Austria.

Plummer, M., Best, N., Cowles, K. and Vines, K. (2006), 'CODA: Convergence Diagnosis and Output Analysis for MCMC', *R News* **6**(1), 7–11.
URL: *https://journal.r-project.org/archive/*

Quackenbush, J. (2002), 'Microarray data normalization and transformation', *Nature Genetics* **32**(4), 496–501.

Quintana, F. A., Mueller, P., Jara, A. and MacEachern, S. N. (2020), 'The dependent Dirichlet process and related models', *arXiv preprint arXiv:2007.06129* .

R Core Team (2020), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

Rahman, M. M., Callaghan, C. K., Kerskens, C. M., Chattarji, S. and O'Mara, S. M. (2016), 'Early hippocampal volume loss as a marker of eventual memory deficits caused by repeated stress', *Scientific Reports* **6**(1), 1–15.

Reiser, B. and Faraggi, D. (1999), 'Confidence intervals for the overlapping coefficient: the normal equal variance case', *Journal of the Royal Statistical Society: Series D* **48**(3), 413–418.

Reitz, C. and Mayeux, R. (2014), 'Alzheimer disease: epidemiology, diagnostic criteria, risk factors and biomarkers', *Biochemical Pharmacology* **88**(4), 640–651.

Richardson, S. and Green, P. J. (1997), 'On Bayesian analysis of mixtures with an unknown number of components (with discussion)', *Journal of the Royal Statistical Society, Series B* **59**(4), 731–792.

Ridout, M. S. and Linkie, M. (2009), 'Estimating overlap of daily activity patterns from camera trap data', *Journal of Agricultural, Biological, and Environmental Statistics* **14**(3), 322–337.

Rodríguez, A. and Martínez, J. C. (2014), 'Bayesian semiparametric estimation of covariate-dependent ROC curves', *Biostatistics* **15**(2), 353–369.

Rodriguez, A. and Müller, P. (2013), 'Nonparametric Bayesian inference', *NSF-CBMS Regional Conference Series in Probability and Statistics* **9**, i–110.

Rossi, P. (2014), *Bayesian Non- and Semi-parametric Methods and Applications*, Princeton University Press.

Rubin, D. B. (1981), 'The Bayesian bootstrap', *The Annals of Statistics* **9**(1), 130–134.

Rue, H., Martino, S. and Chopin, N. (2009), 'Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations', *Journal of the Royal Statistical Society: Series B* **71**(2), 319–392.

Ruf, M., Morgan, O. and Mackenzie, K. (2017), 'Differences between screening and diagnostic tests and case finding', `https://www.healthknowledge.org.uk/public-health-textbook/disease-causation-diagnostic/2c-diagnosis-screening/screening-diagnostic-case-finding`. Accessed: 2021-04-27.

Samawi, H. M., Yin, J., Rochani, H. and Panchal, V. (2017), 'Notes on the overlap measure as an alternative to the Youden index: How are they related?', *Statistics in Medicine* **36**(26), 4230–4240.

Schmid, F. and Schmidt, A. (2006), 'Nonparametric estimation of the coefficient of overlapping—theory and empirical application', *Computational Statistics & Data Analysis* **50**(6), 1583–1596.

Sethuraman, J. (1994), 'A constructive definition of Dirichlet priors', *Statistica Sinica* **4**, 639–650.

Silva-Fortes, C., Turkman, M. A. A. and Sousa, L. (2012), 'Arrow plot: a new graphical tool for selecting up and down regulated genes and genes differentially expressed on samples subgroups', *BMC Bioinformatics* **13**(1), 1–15.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman and Hall/CRC.

Smith, P. and Thompson, T. (1996), 'Correcting for confounding in analyzing receiver operating characteristic curves', *Biometrical Journal* **38**(7), 857–863.

Spiegelhalter, D. J., Abrams, K. R. and Myles, J. P. (2004), *Bayesian Approaches to Clinical Trials and Health-care Evaluation*, Vol. 13, John Wiley & Sons.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and Van Der Linde, A. (2002), 'Bayesian measures of model complexity and fit', *Journal of the Royal Statistical Society, Series B* **64**(4), 583–639.

Stan Development Team (2021), *Stan Modeling Language Users Guide and Reference Manual*. **URL:** *https://mc-stan.org*

Toga, A. W. and Crawford, K. L. (2015), 'The Alzheimer's Disease Neuroimaging Initiative informatics core: a decade in review', *Alzheimer's & Dementia* **11**(7), 832–839.

Turner, D. A. (1978), 'An intuitive approach to receiver operating characteristic curve analysis', *Journal of Nuclear Medicine* **19**(2), 213–220.

Van Havre, Z., White, N., Rousseau, J. and Mengersen, K. (2015), 'Overfitting Bayesian mixture models with an unknown number of components', *PloS ONE* **10**(7), e0131739.

Walker, S. G. (2013), Bayesian nonparametrics, *in* P. Damien, P. Dellaportas, N. G. Polson and D. A. Stephens, eds, 'Bayesian Theory and Applications', Oxford University Press, chapter 13, pp. 266–290.

Walker, S. G., Damien, P., Laud, P. W. and Smith, A. F. (1999), 'Bayesian nonparametric inference for random distributions and related functions', *Journal of the Royal Statistical Society: Series B* **61**(3), 485–527.

Wang, D. and Tian, L. (2017), 'Parametric methods for confidence interval estimation of overlap coefficients', *Computational Statistics & Data Analysis* **106**, 12–26.

Watanabe, S. (2010), 'Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory', *Journal of Machine Learning Research* **11**(Dec), 3571–3594.

Weitzman, M. S. (1970), *Measures of overlap of income distributions of white and Negro families in the United States*, Vol. 22, US Bureau of the Census.

Welge, V., Fiege, O., Lewczuk, P., Mollenhauer, B., Esselmann, H., Klafki, H.-W., Wolf, S., Trenkwalder, C., Otto, M., Kornhuber, J., Wiltfang, J. and Bibl, M. (2009), 'Combined CSF tau, p-tau181 and amyloid-β 38/40/42 for diagnosing Alzheimer's disease', *Journal of Neural Transmission* **116**(2), 203–212.

Wood, S. N. (2017), *Generalized additive models: an introduction with R*, Chapman and Hall/CRC.

World Health Organization (2016), 'Global report on diabetes', `https://apps.who.int/iris/handle/10665/204871`.

World Health Organization (2018), *World health statistics 2018: monitoring health for the SDGs sustainable development goals*, World Health Organization.

World Health Organization (2019), 'Dementia Fact sheet', `https://www.who.int/news-room/fact-sheets/detail/dementia`.

Youden, W. J. (1950), 'Index for rating diagnostic tests', *Cancer* **3**(1), 32–35.

Zhou, X.-H., Obuchowski, N. A. and McClish, D. K. (2011), *Statistical Methods in Diagnostic Medicine*, 2 edn, John Wiley & Sons.

Zou, K. H., Liu, A., Bandos, A. I., Ohno-Machado, L. and Rockette, H. E. (2011), *Statistical Evaluation of Diagnostic Performance: Topics in ROC Analysis*, CRC Press.

Zweig, M. H. and Campbell, G. (1993), 'Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine', *Clinical Chemistry* **39**(4), 561–577.