FEDERAL UNIVERSITY OF MINAS GERAIS

DOCTORAL THESIS

Imputation of Missing Data Using Gaussian Linear Cluster-Weighted Modeling

Author:
Luis Alejandro MASMELA-CAITA

Institute of Exact Sciences
Department of Statistics

April 26, 2021

FEDERAL UNIVERSITY OF MINAS GERAIS

DOCTORAL THESIS

Imputation of Missing Data Using Gaussian Linear Cluster-Weighted Modeling

Author:
Luis Alejandro
MASMELA-CAITA

Supervisors:
Dra. Thaís
PAIVA GALLETTI
Dr. Marcos
OLIVEIRA PRATES

A thesis submitted in fulfillment of the requirements for the degree of Doctor in Statistics

in the

Institute of Exact Sciences
Department of Statistics

April 26, 2021

Masmela – Caita, Luis Alejandro.

M397i Imputation of missing data using Gaussian linear cluster-weighted modeling [manuscrito] / Luis Alejandro Masmela – Caita. – 2021.
xxiii, 91 f. il.

Orientadora: Thais Paiva Galleti.
Coorientador: Marcos Oliveira Prates
Tese (doutorado) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística.
Referências: f. 89–91.

1. Estatística – Teses. 2. Correlação (Estatistica) – Teses. 3. Crítica de imputação de dados (Estatística) – Teses. 4. Ausência de dados (Estatística) – Teses.5. Processos gaussianos – Teses. I. Galleti, Thais Paiva. II. Prates, Marcos Oliveira III. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística. IV.Título.

CDU 519.2 (043)

Ficha catalográfica elaborada pela bibliotecária Belkiz Inez Rezende Costa
CRB 6ª Região nº 1510

UNIVERSIDADE FEDERAL DE MINAS GERAIS

PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

ATA DA DEFESA DE TESE DE DOUTORADO DO(A) ALUNO(A) LUÍS ALEJANDRO MASMELA CAITA, MATRICULADO(A), SOB O Nº 2017667190, NO PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA, DO INSTITUTO DE CIÊNCIAS EXATAS, DA UNIVERSIDADE FEDERAL DE MINAS GERAIS, REALIZADA NO DIA 26 DE MARÇO DE 2021.

Aos 26 dias do mês de março de 2021, às 15h00, em reunião pública virtual 67 https://meet.google.com/vbq-sbio-uts, (conforme orientações para a atividade de defesa de tese durante a vigência da Portaria PRPG nº 1819), do Instituto de Ciências Exatas da UFMG, reuniram-se os professores abaixo relacionados, formando a Comissão Examinadora homologada pelo Colegiado do Programa de Pós-Graduação em Estatística, para julgar a defesa de tese do(a) aluno(a) Luís Alejandro Masmela Caita, nº 2017667190, intitulada: “Imputation of Missing Data Using Gaussian Linear Cluster-Weighted Modeling", requisito final para obtenção do Grau de doutor em Estatística. Abrindo a sessão, o(a) Senhor(a) Presidente da Comissão, Prof(a). Thais Paiva Galletti, passou a palavra ao(à) aluno(a) para apresentação de seu trabalho. Seguiu-se a arguição pelos examinadores com a respectiva defesa do(a) aluno(a). Após a defesa, os membros da banca examinadora reuniram-se reservadamente sem a presença do(a) aluno(a) e do público, para julgamento e expedição do resultado final. Foi atribuída a seguinte indicação:

(X) Aprovada.
( ) Reprovada com resubmissão do texto em ____ dias.
( ) Reprovada com resubmissão do texto e nova defesa em ____ dias.
( ) Reprovada.

Profa. Thais Paiva Galletti
Orientadora – (DEST/UFMG)

Prof. Marcos Oliveira Prates
Coorientador (DEST/UFMG)

Profa. Lourdes Coral Contreras Montenegro)
(DEST/UFMG)

Profa. Rosangela Helena Loschi
(DEST/UFMG)

Profa. Camila Borelli Zeller)
(DEST/UFJF)

Prof. Daniel Manrique-vallier
(INDIANA UNIVERSITY BLOOMINGTON)

O resultado final foi comunicado publicamente ao(à) aluno(a) pelo(a) Senhor(a) Presidente da Comissão. Nada mais havendo a tratar, o(a) Presidente encerrou a reunião e lavrou a presente Ata, que será assinada por todos os membros participantes da banca examinadora. Belo Horizonte, 26 de março de 2021.

Observações:

1. No caso de aprovação da tese, a banca pode solicitar modificações a serem feitas na versão final do texto. Neste caso, o texto final deve ser aprovado pelo orientador da tese. O pedido de expedição do diploma do candidato fica condicionado à submissão e aprovação, pelo orientador, da versão final do texto.
2. No caso de reprovação da tese com resubmissão do texto, o candidato deve submeter o novo texto dentro do prazo estipulado pela banca, que deve ser de no máximo 6 (seis) meses. O novo texto deve ser avaliado por todos os membros da banca que então decidirão pela aprovação ou reprovação da tese.
3. No caso de reprovação da tese com resubmissão do texto e nova defesa, o candidato deve submeter o novo texto com a antecedência à nova defesa que o orientador julgar adequada. A nova defesa, mediante todos os membros da banca, deve ser realizada dentro do prazo estipulado pela banca, que deve ser de no máximo 6 (seis) meses. O novo texto deve ser avaliado por todos os membros da banca. Baseada no novo texto e na nova defesa, a banca decidirá pela aprovação ou reprovação da tese.

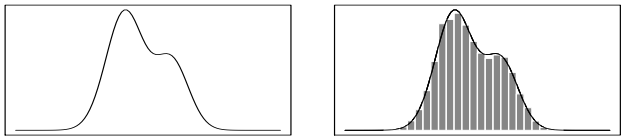
Declaration of Authorship

I, Luis Alejandro MASMELA-CAITA, declare that this thesis titled, "Imputation of Missing Data Using Gaussian Linear Cluster-Weighted Modeling" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____



"I showed my masterpiece to the big people and asked if my drawing scared them. But they replied: Get scared? Why should someone be scared by a ***«mixture of normal distributions»****?"*

The Little Prince.

FEDERAL UNIVERSITY OF MINAS GERAIS

Abstract

Institute of Exact Sciences
Department of Statistics

Doctor in Statistics

Imputation of Missing Data Using Gaussian Linear Cluster-Weighted Modeling

by Luis Alejandro MASMELA-CAITA

Missing data occurs when some values are not stored or observed for variables of interest. However, most of the statistical theory assumes that data is fully observed. An alternative to deal with incomplete databases is to fill in the spaces corresponding to the missing information based on some criteria, this technique is called imputation. We introduce a new imputation methodology for databases with non-response units using additional information from fully observed auxiliary variables. We assume that the non-observed variables are continuous, and that auxiliary variables assist to improve the imputation capacity of the model. In a fully Bayesian framework, our method uses a flexible mixture of multivariate normal distributions to model the response and the auxiliary variables jointly. Under this framework, we use the properties of Gaussian Cluster-Weighted modeling to construct a predictive model to impute the missing values using the information from the covariates. Simulations studies and a real data illustration are presented to show the method imputation capacity under a variety of scenarios and in comparison to other literature methods.

Keywords: Cluster-Weighted Modeling, Gaussian mixture models, imputation method, missing data.

Resumo

Dados ausentes ocorrem quando alguns valores não são armazenados ou observados para variáveis de interesse. No entanto, a maior parte da teoria estatística assume que os dados são totalmente observados. Uma alternativa para lidar com bases de dados incompletas é preencher os espaços correspondentes às informações faltantes com base em alguns critérios, essa técnica é chamada de imputação. Apresentamos uma nova metodologia de imputação para bancos de dados com unidades de não resposta usando informações adicionais de variáveis auxiliares totalmente observadas. Assumimos que as variáveis não observadas são contínuas e que as variáveis auxiliares ajudam a melhorar a capacidade de imputação do modelo. Em uma estrutura totalmente Bayesiana, nosso método usa uma mistura flexível de distribuições normais multivariadas para modelar a resposta e as variáveis auxiliares em conjunto. Sob essa estrutura, usamos as propriedades da modelagem Gaussian Cluster-Weighted para construir um modelo preditivo para imputar os valores ausentes usando as informações das covariáveis. Estudos de simulação e uma ilustração de dados reais são apresentados para mostrar a capacidade de imputação do método sob uma variedade de cenários e em comparação com outros métodos da literatura.

Palavras-chave: Cluster-Weighted Modeling, modelos de mistura Gaussiana, método de imputação, dados faltantes.

Acknowledgements

Inicialmente, agradeço às instituições que me apoiaram financeiramente, pois sem elas minha permanência, bem como minha tranquilidade teriam sido quase impossíveis. Refiro-me à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), por seu apoio através do pagamento de bolsas e auxílios a estudantes de mestrado e doutorado. A la Universidad Distrital Francisco José de Caldas por la comisión de estudios que me fue otorgada. Con el apoyo de estas dos instituciones logré mi sueño de poder complementar mi formación académica en el exterior.

Agradeço à Universidade Federal de Minas Gerais (UFMG), instituição pública que, no marco da política governamental de educação gratuita, permite que uma infinidade de estudantes tenha acesso à educação de alta qualidade.

Agradeço ao Departamento de Estatística da UFMG (DEST-UFMG), formado pelo corpo docente e pelo corpo administrativo. Toda uma equipe que atua conjuntamente com a finalidade de cultivar e desenvolver essa importantíssima e bela área das matemáticas, a Estatística.

Agradeço a meus orientadores, professora Thaís Paiva e professor Marcos Prates. Sua motivação, sua entrega, sua simplicidade, seus conhecimentos, sua compreensão a todo momento, e seu respeito foram atributos que me fizeram refletir sobre o trabalho de ser professor.

Agradeço a todos os professores que de uma forma ou de outra influenciaram meu aprendizado durante minha passagem pelas aulas da UFMG. Muito especialmente, agradeço à professora Rosangela Loschi: seu carisma, sua entrega e interesse se conectaram comigo desde a primeira vez em que assisti sua aula de Inferência Bayesiana.

Agradezco a toda mi familia, por el apoyo que me dieron desde el momento en que se enteraron de este proyecto que enfrentaba. A mi madre Ana Elvira, no pasó una sola noche sin que habláramos y me preguntara cómo fue mi día. A mi hija Laura Alejandra porque mientras estuve lejos, cumplió con su deber de formarse como profesional. A mi hermana Nidia por su apoyo incondicional. Y a mis demás hermanas, Patricia y Mireya, que junto con mis sobrinas Daniela, Camila, Ana y Sofia mantuvieron en unión estrecha mi núcleo familiar mientras estuve lejos.

Agradezco a Danna Lesley, mi compañera de vida, porque gracias a su amor, a su impulso, a su emprendimiento, su diligencia, su sacrificio, a su compañía en cada uno de los momentos de este proceso, logramos juntos cumplir con este gran reto de estudiar en el exterior.

A Cantelli y a Hanny, mis perritas, siempre que me vieron, me recibieron batiendo su cola y echando sus orejitas hacia atrás. Parece algo muy simple, pero te llena de felicidad.

Al Universo, porque conspiró para que todo esto ocurriera.

Contents

List of Figures

List of Tables

List of Abbreviations

CWM	**C**luster **W**eighted **M**odeling
LCWM	**L**inear **C**luster **W**eighted **M**odeling
FMM	**F**inite **M**ixture **M**odels
MRM	**M**ixture (of) **R**egression **M**odels
MCAR	**M**issing **C**ompletly **A**t **R**andom
MAR	**M**issing **A**t **R**andom
MNAR	**M**issing **N**ot **A**t **R**andom
MI	**M**ultiple **I**mputation
ML	**M**aximum **L**ikelihood
EM	**E**xpectation **M**aximization
MCMC	**M**arkov **C**hain **M**onte **C**arlo
MC	**M**onte **C**arlo
WAIC	**W**atanabe **A**kaike **I**nformation **C**riterion
MAP	**M**aximum **A** **P**osteriori
pmm	**p**redict **m**ean **m**atching
KL	**K**ullback **L**eibler
EAM	**E**ncuesta **A**nual **M**anufacturera
CMF	**C**ensus (of) **M**anu**F**acturing
AF	**A**ctivos **F**ijos
GPO	**G**astos (de) **P**ersonal **O**cupado

"Es el tiempo que pasas con tu rosa
lo que hace que tu rosa sea tan importante."
A las rosas que hacen parte de mi vida:
Ana Elvira, quien me dio la vida;
Laura Alejandra, a quien le di la vida;
Danna Lesley, quien me acompaña en la vida;
Patricia, Mireya, Nidia;
Daniela, Camila, Ana y Sofia.
Todas juntas conforman el jardín de mi vida.

Chapter 1

Introduction

"He was only a fox like a hundred thousand other foxes.
But I have made him my friend,
and now he is unique in all the world"

The Little Prince.

1.1 Overview

In studies with statistical data, the vast majority of inference methods start from the assumption that the database that enters the analysis has its information completely observed and the desirable characteristics of the estimators are based on this assumption. However, in many situations, for example, when information is collected through surveys, databases with complete information cannot be guaranteed due to multiple reasons. These reasons must be studied in depth to determine why data is missing or how to avoid incomplete data sets in the information gathering process. To perform statistical analysis with complete databases, the most common solution is to remove individuals with missing information from the database (*Procedures Based on Completely Recorded Units*). Besides the information loss, this approach can lead to estimation biases, particularly when there are differences between the information of those who respond and those who do not. Rubin (1976) discusses the conditions of when the process that caused the missing observations can be ignored. He presents the weakest conditions in the missing data process, so that it is appropriate to ignore this process when making inference about the data distribution.

Several methodologies have been designed to make statistical inference about incomplete data sets. Among them is the *Expectation-Maximization* (EM) algorithm, introduced by Dempster, Laird, and Rubin (1977). It is a likelihood-based procedure that uses only observed data, inferences are made from the observed-data likelihood function. The algorithm allowed to generate robust estimators from the application of the *Maximum Likelihood* (ML) method, where the missing observations are assumed as random variables and the imputed data are generated without the need to adjust models. An alternative proposed in the literature to deal with incomplete databases is to fill in the spaces corresponding to this missing information based on some criteria, this technique is called *imputation*. Imputation is attractive because it makes it easy to implement statistical methods of analysis for complete data sets. One drawback of imputation, followed by the use of full data set analysis methods, is that the resulting inferences can be misleading if the uncertainty due to lack of data has not been addressed (Little and Rubin, 2019). In this regard, Rubin (1987) introduces the concept of *Multiple Imputation* (MI), based on the premise that each missing data must be replaced by various simulated values. The application of this

technique was facilitated with the development of Bayesian simulation methods. From that moment on, the intensive use of these algorithms was encouraged, once routines such as ML and MI could be computationally programmed.

A complete study of the imputation techniques and their classification can be consulted at Little and Rubin (2019). Imputation techniques are also used to generate synthetic data sets in the case of data with disclosure restrictions. Since the synthetic values are not actual observations, they can be published for analysis (Rubin, 1993; Raghunathan, Reiter, and Rubin, 2003). For information obtained through surveys, Kim et al. (2014) propose a fully Bayesian, flexible joint modeling approach for multiple imputation of missing or faulty data subject to linear constraints. The procedure is based on a Dirichlet process mixture of multivariate normal distributions as an imputation engine. The missing data are imputed with values generated from the model adjusted to the observed data. Using a similar statistical model as an imputation engine, Paiva and Reiter (2017) propose a methodology to impute continuous variables with missing data, where the missing data mechanism is non-ignorable. Under a Bayesian approach, the procedure begins by fitting a mixture of multivariate normal distributions based on the observed data. Then, from subsequent samples of the mixture model, an analyst can use the estimated distribution to obtain imputed data in various scenarios.

Our interest is focused on the model used by Kim et al. (2014) and Paiva and Reiter (2017) as an imputation engine, we refer to the Dirichlet process mixture of multivariate normal distributions. According to the results presented by the authors, it is a model that is characterized by great flexibility when it comes to adjusting complex distributional forms. Although the model is implemented for both cases, item and unit of nonresponse[1], our interest is focused on the second. A question that arises and that we want to answer throughout the development of this document is: *how to include auxiliary information in this model in such a way that it is possible to improve the imputation process under some established criteria?*

In the context of information obtained through surveys, when considering the non-response unit pattern, for example, we could include fully observed auxiliary variables for all individuals from other sources of information. Our proposal seeks to include this new information based on the implementation of the Dirichlet process mixture of multivariate normal distributions model, using a completely Bayesian approach. The idea of implementing a regression mixture model arises naturally. In this approach, the covariates are considered deterministic, so that they do not carry information about which group the subject is likely to belong to. When we consider observational data, the covariates may behave differently between groups. In this sense, the idea would be for the model to consider the heterogeneity of the covariate and thus use such information to be able to choose with which component to predict, or in our case to impute from (Hoshikawa, 2013).

One model that brings with it the ability to include these desired features is *Cluster-Weighted Modeling* (CWM), developed by Gershenfeld (1997) in the context of media technology. Ingrassia, Minotti, and Vittadini (2012) proposed to use the CWM

[1] *Unit nonresponse* in a survey occurs when an eligible sample member fails to respond at all whereas *item nonresponse* refers to the absence of answers to specific questions in the survey.

in a statistical environment and showed that it is a general and flexible family of mixture models. In particular, under Gaussian assumptions, they specify characteristics of their probability distribution and related statistical properties and demonstrate links with traditional mixture models in terms of density functions and posterior probabilities. An interesting result is that the Gaussian CWM includes the finite mixture model and the mixture of regression model as special cases. Some recent developments in CWMs can be found in Dang et al. (2017), Punzo and McNicholas (2017), Berta et al. (2016), and Ingrassia et al. (2015).

In this document, *we propose a new imputation methodology for databases with unit nonresponse patterns, on which it is desired to include auxiliary information for all units and which can be considered as helpful to improve the imputation process*. We assume continuous variables with a group structure in the data or where the objective is to explore the data for that structure or where the shape of the data distribution is unknown. We use a fully Bayesian approach to fit the Dirichlet process mixture of multivariate normal distributions to a database that jointly considers the responses and the covariates. With these results, we use properties of the Cluster-Weighted Modeling in the Gaussian case to construct a predictive model that will help us implement the imputation process using, adaptively, the covariates that enter as additional information in the model.

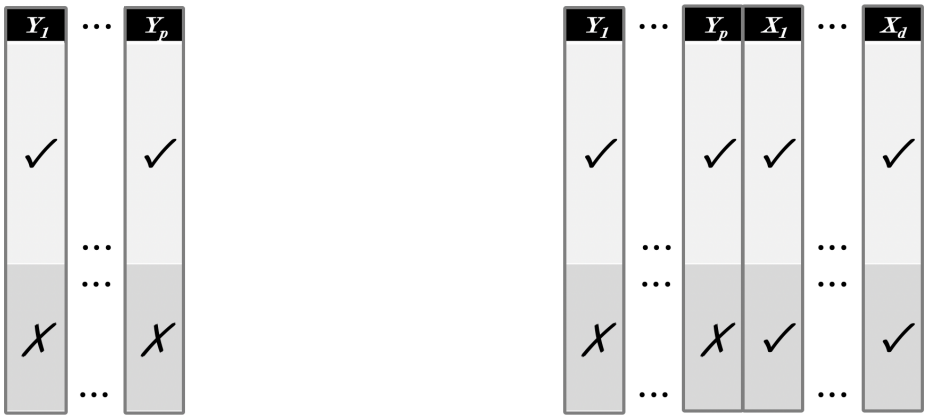
1.2 Motivating Example

As a motivational example, the pattern of missing data on which the implementation of the methodology proposed in Paiva and Reiter (2017) is based is described. Through it, the problem that is intended to be addressed here is illustrated. The authors make use of data from the 2007 United States *Census of Manufacturing* (CMF). The CMF covers all manufacturing establishments in the United States. Data is collected through questionnaires and processed in waves throughout the year, depending on when each establishment mails its form.[2] This can result in *Missing Not at Random* (MNAR) non-response units in some waves. It is assumed that the establishment either sends the complete form, or does not send the form at all. The structure of the dataset can be represented in Figure 1.1a, where the check mark (✓) indicates that the data is observed, while the tag (✗) indicates that data is missing. The first establishments have all the complete information, while the last ones lack the total information. The authors present an approach to inform decisions about nonresponse follow-up sampling. The basic idea is to create complete samples by imputing nonrespondents' data under various assumptions about the nonresponse mechanisms. As part of the methodology, they present a new approach for generating imputations for multivariate continuous data with nonignorable unit nonresponse. They fit mixtures of multivariate normal distributions to the respondents' data, and adjust the probabilities of the mixture components to generate nonrespondents' distributions with desired features.

Assuming that additional information can be obtained from auxiliary data sources, the missing data pattern can take the form shown in Figure 1.1b. In this new pattern, it is observed that new variables $X_1, ..., X_d$ appear with fully observed information for all establishments. Our objective is to use this new information so that the imputations for establishments with missing values in the variables $Y_1, ..., Y_p$ improve

[2] https://www.census.gov/

under some criteria, among others, compared to the imputations obtained, under an assumption of *Missing at Random* (MAR), from the model implemented by Paiva and Reiter (2017).



(a) No auxiliary information.

(b) With auxiliary information.

FIGURE 1.1: Missing data pattern from multivariate databases in the cases of not including and including auxiliary information.

We will approach the study of the problem considered, starting from the structure of the databases in Figure 1.1 for the simplest case, i.e., assuming the value of $p = 1$ and that we will call the univariate model, and then generalize to the multivariate case, that is, when $p > 1$. We will propose a model that allows the auxiliary information to be included in the best way possible. We will analyze the results in the imputations obtained according to the type of variable or variables that enter the model. In this way, we will characterize the type of variable that enters the model and that produces the best results. We will compare the proposed model with other imputation methodologies that are based on a Bayesian approach and that use prediction models to carry out the imputations. Finally, we will implement the proposed model in real data sets to evaluate the performance of our model when comparing it with other imputation procedures that are of interest.

1.3 Outline of the Report

The rest of the document is distributed as follows. Chapter 2 succinctly summarizes the concepts of the required missing data literature. It defines the concepts of pattern and missing data mechanism, it also establishes the imputation model that we will follow from a Bayesian approach. Some ideas of the multiple imputation process, as well as some imputation procedures using prediction are briefly presented. In Chapter 3, we present the finite mixture models, the regression mixture model, and the Cluster-Weighted Modeling specified for the Gaussian case. Some results are presented relating the different models of interest, as well as the Bayesian imputation procedure that will be used. The univariate model is analyzed in Chapter 4. Studies of the results of the model are established through simulations. The performance of the model is observed under the input of various types of variables. The results of the model are also compared with other methodologies of interest. Two sets of real data are used to implement the proposed imputation model and observe its performance, the Faithful database and the Colombian Annual Manufacturing Survey. Similar to the univariate case, an analysis for the multivariate case is

presented in Chapter 5. Simulation studies are carried out considering the type of variables that can enter the model and the results are compared with other imputation methods based on predictions. An example with real data is illustrated using the Iris database. Two missing data patterns are simulated on the database, one under a MAR mechanism and the other under an MNAR mechanism. The results obtained are analyzed and compared with other imputation procedures. In Chapter 6, we present the conclusions obtained and some final observations. Appendices, at the end of the document, present a compendium of graphs as a complement to some of the chapters.

Chapter 2

Missing Data

"Where are the people?" resumed the little prince at last.
"It's a little lonely in the desert..."
"It is lonely when you're among people, too," said the snake.

The Little Prince.

2.1 Overview

Standard statistical methods are generally implemented on data sets organized in matrix form. Matrix entries are almost always real numbers that can represent continuous or categories variables (in nominal or ordinal scales). In statistics, missing values occur when the value of data is not stored for the variable in observation. *Missing data theory* deals with the analysis of data matrices when some of its inputs are not observed. In this case, the validity and efficiency of methods based on complete data cannot be guaranteed when the data is incomplete (Rubin, 1976).

Various ways of dealing with the problem of missing values are studied in the statistical literature. One of them is the approach by likelihood-based methods that use only the observed data, all inferences are based on the observed-data likelihood function (Dempster, Laird, and Rubin, 1977; Rubin, 1987; McCullagh and Nelder, 1989). Sometimes the likelihood function can be complicated to handle and inferences about the parameters of interest can decrease its precision due to missing data. Another alternative proposed to deal with incomplete databases is to fill in the spaces corresponding to this missing information based on some criteria, this technique is called *imputation*. A complete study of imputation techniques and their classification can be found at Little and Rubin (2019) and Zhang (2003). Subsequently, Rubin (1987) introduced the concept of multiple imputation (MI), based on the premise that each missing data must be replaced by several simulated values. The application of this technique was facilitated with the development of Bayesian simulation methods. From that moment on, the intensive use of these algorithms was encouraged, since different MI and ML routines could be programmed computationally.

The following sections summarize some topics from the theory of missing data and that are required for the development of the work that is proposed in this document. To delve into the various topics, bibliographical references are suggested that can be consulted by the reader. In Section 2.2 the concept of missing data patterns will be treated, specifically we will refer to two patterns of interest, unit nonresponses and item nonresponses. Section 2.3 defines the so-called missing data

mechanisms and provides an example to illustrate them. Ways to approach statistical inferences based on incomplete data sets and the imputation model based on the predictive conditional distribution using a Bayesian approach is presented in Section 2.4. The multiple imputation approach to the missing data problem is briefly discussed in Section 2.5. Section 2.6 addresses the MCMC methods to deal with the imputation procedure, specifically we will refer to the Data Augmentation algorithm. Finally, Section 2.7 summarizes the imputation methods that use prediction, some of which will serve as a basis for comparison with our model.

2.2 Missing-Data Patterns

An aspect of special interest regarding the database with missing information has to do with the *missing-data pattern*, with this term we refer to which values are observed in the data matrix and which values are missing. Note that a missing-data pattern simply describes the location of the "holes" in the database and does not explain why the data is missing. Various configurations of missing data patterns can be consulted in the missing data literature. For example, Van Buuren (2018) distinguishes several types of missing data patterns: univariate and multivariate, monotone and non-monotone (or general), connected and unconnected. Enders (2010) additionally presents unit nonresponse, planned missing and latent variable patterns.

In survey data, missing data can be defined as unit nonresponse and item nonresponse. A distinction that turns on whether there is at least one survey item for which a valid response was obtained, or whether the entire unit is missing. Little and Rubin (2019) and Enders (2010) refer to the missing value configuration presented in Figure 1.1b as unit nonresponse pattern. This pattern occurs when $X_1, ..., X_d$ are characteristics that are available for every member of the sampling frame (e.g., census tract data) and $Y_1, ..., Y_p$ are surveys that some respondents refuse to answer. When entire units are missing from a sample, no test or correction for bias is available without obtaining additional data about the targeted respondents who did not respond to the initial survey. Non-response bias refers to the mistake researchers expect to make in estimating a population characteristic based on a sample of survey data in which, due to non-response, certain types of survey respondents are underrepresented (Berg, 2005). Traditionally, survey research has treated unit and item nonresponse as two separate problems with different impacts on data quality, different statistical treatments and adjustments, and different underlying causes (Yan and Curtin, 2010).

Two terms that are important to distinguish are missing data patterns and *missing data mechanisms*. These are terms that researchers sometimes use interchangeably. While missing data pattern is a term we just referred to, missing data mechanisms describe possible relationships between the measured variables and the probability of missing data. We will talk about the latter in the next section.

2.3 Missingness Mechanism

To deal with the problem of missing data effectively, a concept of great interest has to do with the so-called mechanisms that lead to a lack of data and, in particular, the question of whether the fact that the variables have missing observations is related

to the underlying values of the variables in the data set. The crucial role of the mechanism in the analysis of missing data values was largely ignored until the concept was formalized in the theory of Rubin (1976), through the simple idea of treating missing data indicators as random variables and assigning them a distribution.

The notation used in this section to define the so-called missing data mechanisms is set out below. This notation will also be used to describe the MCMC algorithm for the model proposed in later sections.

Let $\boldsymbol{Y}$ denote the $n \times p$ matrix that contains the complete set of data on p variables for all n units. Let $\boldsymbol{R}$ be the response indicator matrix with the same size as matrix $\boldsymbol{Y}$. The elements of $\boldsymbol{Y}$ and $\boldsymbol{R}$ are denoted by y_{ij} and r_{ij}, respectively, where $i = 1, ..., n$ and $j = 1, ..., p$. If y_{ij} is observed, then $r_{ij} = 0$, and if y_{ij} is missing, then $r_{ij} = 1$. An additional notation of quite interest is the partition of the complete data set $\boldsymbol{Y}$ into two sets, the observed data being collectively denoted $\boldsymbol{Y}_{\text{obs}}$ and the missing data collectively denoted $\boldsymbol{Y}_{\text{mis}}$. In this way, the complete data set can be written as $\boldsymbol{Y} = (\boldsymbol{Y}_{\text{obs}}, \boldsymbol{Y}_{\text{mis}})$. It is important to note that here $\boldsymbol{R}$ is assumed to be fully known and that it indicates what we really see, and $\boldsymbol{Y}$ contains values that are all defined.

The missing data mechanism is characterized by the conditional distribution of $\boldsymbol{R}$ given $\boldsymbol{Y}$, say $p(\boldsymbol{R}|\boldsymbol{Y}, \varphi)$. If the missing data does not depend on the values of the data $\boldsymbol{Y}$, missing or observed, that is, if

$$p(\boldsymbol{R}|\boldsymbol{Y}, \varphi) = p(\boldsymbol{R}|\varphi) \text{ for all } \boldsymbol{Y}, \varphi, \tag{2.1}$$

the data is called *Missing Completely At Random* (MCAR). Keep in mind that this assumption does not mean that the mechanism itself is random, but that the lack of information does not depend on the data values. A less restrictive assumption than MCAR is that the missingness depends only on the component $\boldsymbol{Y}_{\text{obs}}$ and not on the missing component. This means that,

$$p(\boldsymbol{R}|\boldsymbol{Y}, \varphi) = p(\boldsymbol{R}|\boldsymbol{Y}_{\text{obs}}, \varphi) \text{ for all } \boldsymbol{Y}_{\text{mis}}, \varphi. \tag{2.2}$$

The mechanism defined in (2.2) is called *Missing At Random* (MAR). Finally, the mechanism is called *Missing Not At Random* (MNAR) if the distribution of $\boldsymbol{R}$ depends on the missing values in the $\boldsymbol{Y}$ data matrix.

Example 2.3.1 (Simulation of missingness mechanism). *Van Buuren (2018) illustrates through a simulation process the three missing data mechanisms, MCAR, MAR and MNAR using R software (R Core Team, 2020). The data $\boldsymbol{Y} = (Y_1, Y_2)$ are drawn from a standard bivariate normal distribution with a correlation between Y_1 and Y_2 equal to 0.5. The response indicator is given by $\boldsymbol{R} = (R_1, R_2)$. Missing data are created in Y_2 using the missing data model,*

$$P(R_2 = 1) = \varphi_0 + \frac{\exp Y_1}{1 + \exp Y_1}\varphi_1 + \frac{\exp Y_2}{1 + \exp Y_2}\varphi_2, \tag{2.3}$$

with different parameters settings for $\varphi = (\varphi_0, \varphi_1, \varphi_2)$. For MCAR we set $\varphi_{MCAR} = (0.5, 0, 0)$, for MAR $\varphi_{MAR} = (0, 1, 0)$ and for MNAR $\varphi_{MNAR} = (0, 0, 1)$. Figure 2.1 displays the distribution of $\boldsymbol{Y}_{obs}$ and $\boldsymbol{Y}_{mis}$ under the three missing data models. As expected, they are similar under MCAR, but become progressively more distinct as we move to the MNAR model.

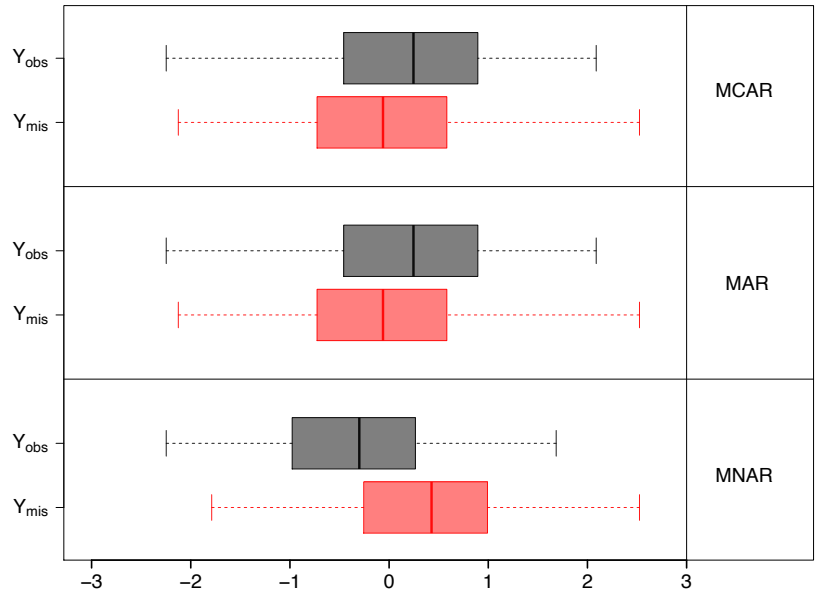


FIGURE 2.1: Distribution of $\boldsymbol{Y}_{\text{obs}}$ and $\boldsymbol{Y}_{\text{mis}}$ under three missing data mechanism.

It can be seen that, to generate the MNAR mechanism, the expression in (2.3) can be written as,

$$P(R_2 = 1) = \text{logit}^{-1}(Y_2)$$

where $\text{logit}(p) = \log[p/(1-p)]$ for $p \in (0,1)$, is the logit function, and logit^{-1} is its inverse function. This transformation will be used later to simulate missing data mechanisms.

2.4 Imputation Model from a Bayesian Framework

To make statistical inference about a population based on a data set, you need the probability model $p(\boldsymbol{Y}|\boldsymbol{\theta})$. In the case of incomplete data, the joint probability model takes the form $p(\boldsymbol{Y}_{\text{obs}}, \boldsymbol{Y}_{\text{mis}}, \boldsymbol{R}|\boldsymbol{\theta}, \boldsymbol{\varphi})$ where $\boldsymbol{Y}_{\text{mis}}$ is unknown, so the likelihood function of this distribution cannot be evaluated. It is then necessary to evaluate the *likelihood function of the observed-data* defined as the function proportional to the marginal distribution of the joint distribution integrated over $\boldsymbol{Y}_{\text{mis}}$, that is,

$$L(\boldsymbol{\theta}, \boldsymbol{\varphi}|\boldsymbol{Y}_{\text{obs}}, \boldsymbol{R}) \propto p(\boldsymbol{Y}_{\text{obs}}, \boldsymbol{R}|\boldsymbol{\theta}, \boldsymbol{\varphi}) \tag{2.4}$$

where,

$$\begin{aligned} p(\boldsymbol{Y}_{\text{obs}}, \boldsymbol{R}|\boldsymbol{\theta}, \boldsymbol{\varphi}) &= \int p(\boldsymbol{Y}_{\text{obs}}, \boldsymbol{Y}_{\text{mis}}, \boldsymbol{R}|\boldsymbol{\theta}, \boldsymbol{\varphi}) d\boldsymbol{Y}_{\text{mis}} \\ &= \int p(\boldsymbol{R}|\boldsymbol{Y}_{\text{obs}}, \boldsymbol{Y}_{\text{mis}}, \boldsymbol{\varphi}) p(\boldsymbol{Y}_{\text{obs}}, \boldsymbol{Y}_{\text{mis}}|\boldsymbol{\theta}) d\boldsymbol{Y}_{\text{mis}}. \end{aligned} \tag{2.5}$$

Under the definitions of the missing data mechanisms MCAR and MAR, (2.5) becomes,

$$p(\boldsymbol{Y}_{\text{obs}}, \boldsymbol{R}|\boldsymbol{\theta}, \boldsymbol{\varphi}) = \begin{cases} p(\boldsymbol{R}|\boldsymbol{\varphi}) p(\boldsymbol{Y}_{\text{obs}}|\boldsymbol{\theta}) & \text{if MCAR,} \\ p(\boldsymbol{R}|\boldsymbol{Y}_{\text{obs}}, \boldsymbol{\varphi}) p(\boldsymbol{Y}_{\text{obs}}|\boldsymbol{\theta}) & \text{if MAR.} \end{cases} \tag{2.6}$$

From a Bayesian approach, the two parameter vectors $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$ are considered *distinct* if the prior joint distribution of $(\boldsymbol{\theta}, \boldsymbol{\varphi})$ can be factored into the prior marginal distributions for $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$, (Rubin, 1976; Rubin, 1987).

So, based on the expression (2.6), if $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$ are distinct, inferences about $\boldsymbol{\theta}$ can be made based only in $p(\boldsymbol{Y}_{\text{obs}}|\boldsymbol{\theta})$, without considering $p(\boldsymbol{R}|\boldsymbol{\varphi})$ or $p(\boldsymbol{R}|\boldsymbol{Y}_{\text{obs}}, \boldsymbol{\varphi})$, that is, ignoring the missing data mechanism. Therefore, for the parameter $\boldsymbol{\theta}$, the observed-data likelihood function ignoring the missingness mechanism can be defined as follows,

$$L(\boldsymbol{\theta}|\boldsymbol{Y}_{\text{obs}}) \propto p(\boldsymbol{Y}_{\text{obs}}|\boldsymbol{\theta}).$$

Likewise, the Bayesian inference in $\boldsymbol{\theta}$ can be based on the *observed-data posterior distribution* defined as follows,

$$p(\boldsymbol{\theta}|\boldsymbol{Y}_{\text{obs}}) \propto L(\boldsymbol{\theta}|\boldsymbol{Y}_{\text{obs}}) \times \pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}), \tag{2.7}$$

where $\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ is the prior distribution of $\boldsymbol{\theta}$.

These inferences use only the observed data $\boldsymbol{Y}_{\text{obs}}$, and are valid as long as the missing data mechanism is ignorable. However, the precision of the inference is lower if a large amount of information is missing (Zhang, 2003).

An alternative proposed in the literature to deal with incomplete databases is to fill in the spaces corresponding to this missing information based on some criteria, this technique is called *imputation*. The idea is to impute the missing data $\boldsymbol{Y}_{\text{mis}}$ and then use standard statistical methods on the complete-data to make inference about $\boldsymbol{\theta}$. For this objective, interest falls on the conditional distribution of $\boldsymbol{Y}_{\text{mis}}$ given $\boldsymbol{Y}_{\text{obs}}$ that can be obtained by integrating under the parameter space $\boldsymbol{\theta}$, that is,

$$p(\boldsymbol{Y}_{\text{mis}}|\boldsymbol{Y}_{\text{obs}}) = \int p(\boldsymbol{Y}_{\text{mis}}|\boldsymbol{Y}_{\text{obs}}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{Y}_{\text{obs}}) d\boldsymbol{\theta}, \tag{2.8}$$

where $p(\boldsymbol{Y}_{\text{mis}}|\boldsymbol{Y}_{\text{obs}}, \boldsymbol{\theta})$ is the *conditional predictive distribution* of $\boldsymbol{Y}_{\text{mis}}$ given $\boldsymbol{Y}_{\text{obs}}$ and $\boldsymbol{\theta}$, and $p(\boldsymbol{\theta}|\boldsymbol{Y}_{\text{obs}})$ is the observed-data posterior distribution of $\boldsymbol{\theta}$. The conditional distribution in (2.8) is called the *predictive posterior distribution* of $\boldsymbol{Y}_{\text{mis}}$ given $\boldsymbol{Y}_{\text{obs}}$ (Rubin, 1987). The two expressions in (2.7) and (2.8) are key to establishing the MCMC sampling algorithm that will be discussed later.

2.5 Multiple Imputation

When doing Multiple Imputation (MI), the goal is to generate multiple possible values for each missing observation in order to obtain two or more complete data sets. Using standard analysis procedures, the researcher can analyze each complete database and then combine these results to obtain the MI estimates.

The set of possible values for missing observations are based on the distribution of the data. The objective is to obtain estimates of the missing values. Estimates of missing values are obtained by simulating random draws from the distribution of the missing variables given the observed variables. Distributions of the data are derived from Bayesian theory, so that the researcher samples values from the posterior

probability distribution of the missing values given the observed variables. The posterior probability distribution of the missing variables given the observed variables is complex, and requires a two-step algorithm, referred to as Data Augmentation (Tanner and Wong, 1987).

MI was proposed in Rubin (1976) and Rubin (1987) as a possible solution to the problem of survey non-response. Following Zhang (2003), MI is a method based on three steps:

1. Generate $M > 1$ complete data sets, filling each missing value using the independent selections of an appropriate imputation model, given the observed values. The imputation model should be constructed to reflect the true distribution of the relationship between the missing values and the observed values.

2. The M imputed complete data sets are analyzed using standard procedures for complete data.

3. The results of the analysis of the M complete databases obtained after imputation are combined in a simple and adequate way.

The analysis of a data set obtained from MI is simple. First, each complete set of data after the imputation process is analyzed using the same method for complete data that would be used in the absence of non-response.

For $m = 1, ..., M$, let $\hat{\theta}_m$ be the estimate for the parameter θ, and W_m the respective associated variance for the estimated parameter θ, calculated from each individual imputed data set m.
The combined estimator for θ is

$$\bar{\theta}_M = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_m.$$

The variability associated with this estimator has two components: the *within imputation* variance,

$$\bar{W}_M = \frac{1}{M} \sum_{m=1}^{M} W_m,$$

and the *between imputation* variance,

$$B_M = \frac{1}{M-1} \sum_{m=1}^{M} (\hat{\theta}_m - \bar{\theta}_M)^2.$$

The total variability associated with $\bar{\theta}_M$ is:

$$T_M = \bar{W}_M + \frac{M+1}{M} B_M,$$

where $(1 + 1/M)$ is a adjustment for M finite. Therefore,

$$\hat{\gamma} = \frac{M+1}{M} \frac{B_M}{T_M}$$

is an estimate of the fraction of information about θ that is missing due to non-response. For large sample sizes and scalar θ, the reference distribution for interval

estimates and significance tests is a $t-$distribution,

$$(\theta - \bar{\theta}_M) T_M^{(-1/2)} \sim t_v,$$

where degrees of freedom,

$$v = (M-1)\left(1 + \frac{1}{M+1}\frac{\bar{W}_M}{B_M}\right)^2,$$

are based on the Satterthwaite approach (Rubin, 1987).

2.6 Markov Chain Monte Carlo

Markov Chain Monte Carlo (MCMC) methods comprise a class of algorithms for sampling from a probability distribution. The MCMC method to impute missing data is applied by Schafer (1997) by implementing the Data Augmentation algorithm developed by Tanner and Wong (1987). The target distribution for our case is the conditional distribution of $\boldsymbol{Y}_{\text{mis}}$ and $\boldsymbol{\theta}$ given $\boldsymbol{Y}_{\text{obs}}$, that is $p(\boldsymbol{Y}_{\text{mis}}, \boldsymbol{\theta}|\boldsymbol{Y}_{\text{obs}})$.

The procedure starts by replacing the missing data $\boldsymbol{Y}_{\text{mis}}$ with some initial values, so $\boldsymbol{\theta}$ can be simulated from the posterior distribution of the complete-data $p(\boldsymbol{\theta}|\boldsymbol{Y}_{\text{obs}}, \boldsymbol{Y}_{\text{mis}})$. Consider the following iterative sampling scheme:

1. Given an updated value $\boldsymbol{\theta}^{(t)}$ of parameter, draw a value of missing data from the conditional predictive distribution of $\boldsymbol{Y}_{\text{mis}}$ given $\boldsymbol{Y}_{\text{obs}}$ and $\boldsymbol{\theta}^{(t)}$, i.e.,

$$\boldsymbol{Y}_{\text{mis}}^{(t+1)} \sim p(\boldsymbol{Y}_{\text{mis}}|\boldsymbol{Y}_{\text{obs}}, \boldsymbol{\theta}^{(t)}). \tag{2.9}$$

2. By conditioning on $\boldsymbol{Y}_{\text{mis}}^{(t+1)}$, the next simulated value of $\boldsymbol{\theta}$ can be draw from its complete-data posterior distribution,

$$\boldsymbol{\theta}^{(t+1)} \sim p(\boldsymbol{\theta}|\boldsymbol{Y}_{\text{obs}}, \boldsymbol{Y}_{\text{mis}}^{(t+1)}). \tag{2.10}$$

Repeating steps (1) and (2) from an initial value $\boldsymbol{\theta}^{(0)}$ generates a Markov chain $\{(\boldsymbol{\theta}^{(t)}, \boldsymbol{Y}_{\text{mis}}^{(t)}) : t = 1, 2, ...\}$. The stationary distribution of the chain is the joint distribution $p(\boldsymbol{\theta}, \boldsymbol{Y}_{\text{mis}}|\boldsymbol{Y}_{\text{obs}})$. The marginal stationary distributions of the subsequences $\{\boldsymbol{\theta}^{(t)} : t = 1, 2, ...\}$ and $\{\boldsymbol{Y}_{\text{mis}}^{(t)} : t = 1, 2, ...\}$ are the posterior distribution of the observed-data $p(\boldsymbol{\theta}|\boldsymbol{Y}_{\text{obs}})$ and the posterior predictive distribution $p(\boldsymbol{Y}_{\text{mis}}|\boldsymbol{Y}_{\text{obs}})$ respectively.

The random selection in (2.9) imputes the missing data $\boldsymbol{Y}_{\text{mis}}$, while the random selection in (2.10) simulates the unknown parameter $\boldsymbol{\theta}$. Therefore, (2.9) and (2.10) are known as the imputation step (**I-step**) and the posterior step (**P-step**), respectively. The steps of the Data Augmentation algorithm are specified in the Algorithm 1.

Algorithm 1: Data Augmentation

Result: Imputed database and parameter estimation

initialization: $\boldsymbol{\theta}^{(0)}$;

for $t = 1, ..., T$ **do**

  I-step: generate $\boldsymbol{y}_{\text{mis}}^{(t)}$ from $p(\boldsymbol{Y}_{\text{mis}}|\boldsymbol{Y}_{\text{obs}}, \boldsymbol{\theta}^{(t-1)})$;

  P-step: generate $\boldsymbol{\theta}^{(t)}$ from $p(\boldsymbol{\theta}|\boldsymbol{Y}_{\text{obs}}, \boldsymbol{y}_{\text{mis}}^{(t)})$;

end

Ideally, the imputations of the missing data should be independent of the observed data. One way to obtain the appropriate multiple imputations is to have a largely separate subsample from a single string. For example, select iterations every l steps after some burn-in time say t_0, with l large enough that the dependency between the imputed values is negligible.

2.7 Synopsis of Imputation Methods Using Prediction

The description below is taken directly from Van Buuren (2018). The author illustrates four ways to create imputations for a single incomplete continuous target variable. Assume $\boldsymbol{Y}$ is the output variable for the univariate case denoted by $\boldsymbol{y}$. Furthermore, using the notation $\boldsymbol{X} = (\boldsymbol{X}_1, ..., \boldsymbol{X}_d)$ for the matrix of the set of input variables, let $\boldsymbol{X}_{\text{obs}}$ be the subset of n_1 rows of $\boldsymbol{X}$ for which $\boldsymbol{y}$ is observed, and $\boldsymbol{X}_{\text{mis}}$ the complementary subset of n_0 rows of $\boldsymbol{X}$ for which $\boldsymbol{y}$ is missing. The vector containing the n_1 observed data in $\boldsymbol{y}$ is denoted by $\boldsymbol{y}_{\text{obs}}$, and the vector of n_0 imputed values in $\boldsymbol{y}$ is indicated by $\dot{\boldsymbol{y}}$.

2.7.1 Predict method

Regression imputation incorporates knowledge of other variables with the idea of producing smarter imputations. The first step is to build a model from the observed data. The predictions for the incomplete cases are then computed under the fitted model and serve as replacements for the missing data. Therefore, the idea is to calculate the regression line and take the imputation of the regression line, i.e., $\dot{\boldsymbol{y}} = \hat{\beta}_0 + \boldsymbol{X}_{\text{obs}}\hat{\beta}_1$ where $\hat{\beta}_0$ and $\hat{\beta}_1$ are least squares estimates calculated from the observed data.

2.7.2 Predict + noise method

We can improve upon the prediction method by adding an appropriate amount of random noise to the predicted value. Let us assume that the observed data are normally distributed around the regression line. The idea now is to draw a random value from a normal distribution with a mean of zero and a estimated standard deviation $\hat{\sigma}$, and add this value to the predicted value. Therefore the imputation is obtained as $\dot{\boldsymbol{y}} = \hat{\beta}_0 + \boldsymbol{X}_{\text{obs}}\hat{\beta}_1 + \dot{\epsilon}$ where $\dot{\epsilon}$ is randomly drawn from the normal distribution as $\dot{\epsilon} \sim \mathcal{N}(0, 1)$. This imputation method is called *stochastic regression imputation*.

2.7.3 Predict + noise + parameter uncertainty

The method in the previous subsection requires that the intercept, the slope and the standard deviation of the residuals are known. However, the values of these parameters are typically unknown, and hence must be estimated from the data. If we had drawn a different sample from the same population, then our estimates for the intercept, slope and standard deviation would be different, perhaps slightly. The amount of extra variability is strongly related to the sample size, with smaller samples yielding more variable estimates. The parameter uncertainty also needs to be included in the imputations. Therefore, the imputed values are obtained for this case as $\dot{\boldsymbol{y}} = \dot{\beta}_0 + \boldsymbol{X}_{\text{obs}}\dot{\beta}_1 + \dot{\epsilon}$ where $\dot{\epsilon}$ is randomly drawn from the normal distribution as $\dot{\epsilon} \sim \mathcal{N}(0,1)$ and $\dot{\beta}_0$, $\dot{\beta}_1$ and $\dot{\sigma}$ are random draws from their posterior distribution, given the data. In this case, the method is called *Bayesian multiple imputation*. If $\dot{\beta}_0$, $\dot{\beta}_1$ and $\dot{\sigma}$ are the least squares estimates calculated from a bootstrap sample taken from the observed data, the method is called *Bootstrap multiple imputation*. Since we will use the Bayesian multiple imputation method later to compare with our proposed model, we go a little deeper into this procedure.

Bayesian multiple imputation

Bayesian sampling draws $\dot{\beta}_0$, $\dot{\beta}_1$ and $\dot{\sigma}$ from their respective posterior distributions. The method draws imputations under the normal linear model using standard non-informative priors for each of the parameters. Specifically, in this Bayesian approach, a prior distribution $p(\beta, \sigma^2) \propto \sigma^{-2}$ is assumed for the conditional model (Rubin, 1987). Algorithm 2 specifies the steps to implement the Bayesian multiple imputation procedure.

Algorithm 2: Bayesian multiple imputation

Input: $\boldsymbol{y}_{\text{obs}}$, $\boldsymbol{X}_{\text{obs}}$, $\boldsymbol{X}_{\text{mis}}$
Output: $\dot{\boldsymbol{y}}$

1 calculate the cross-product matrix $\boldsymbol{S} = \boldsymbol{X}'_{\text{obs}}\boldsymbol{X}_{\text{obs}}$
2 calculate $\boldsymbol{V} = (\boldsymbol{S} + \text{diag}(\boldsymbol{S})\kappa)^{-1}$, with some small ridge parameter κ
3 calculate regression weights $\hat{\beta} = \boldsymbol{V}\boldsymbol{X}'_{\text{obs}}\boldsymbol{y}_{\text{obs}}$
4 draw a random variable $\dot{g} \sim \chi^2_\nu$ with $\nu = n_1 - q$
5 calculate $\dot{\sigma}^2 = (\boldsymbol{y}_{\text{obs}} - \boldsymbol{X}_{\text{obs}}\hat{\boldsymbol{\beta}})'(\boldsymbol{y}_{\text{obs}} - \boldsymbol{X}_{\text{obs}}\hat{\boldsymbol{\beta}})/\dot{g}$
6 draw q independent $\mathcal{N}(0,1)$ variates in vector $\dot{\boldsymbol{z}}_1$
7 calculate $\boldsymbol{V}^{1/2}$ by Cholesky descomposition
8 calculate $\dot{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} + \dot{\sigma}\dot{\boldsymbol{z}}_1\boldsymbol{V}^{1/2}$
9 draw n_0 independent $\mathcal{N}(0,1)$ variates in vector $\dot{\boldsymbol{z}}_2$
10 calculate the n_0 values $\dot{\boldsymbol{y}} = \boldsymbol{X}_{\text{mis}}\dot{\boldsymbol{\beta}} + \dot{\boldsymbol{z}}_2\dot{\sigma}$

2.7.4 Drawing from the observed data

An alternative method of creating imputations consists of finding the predicted values proceeding in the same way as in the previous section, but selecting a small number of candidate donors from the observed data. The selection is made so that the predicted values of the donors are close to the predicted values of the individuals to be imputed. We then randomly select a donor from the candidates and use the observed value that belongs to that donor as a synthetic value. This method is

known as *predictive mean matching* (Little, 1988), and always finds values that have been actually observed in the data.

Predictive mean matching

Predictive mean matching is an easy-to-use and versatile method. It is fairly robust to transformations of the target variable. Imputations are based on values observed elsewhere, so they are realistic. Imputations outside the observed data range will not occur, thus evading problems with meaningless imputations (e.g., negative body height). The model is implicit, which means that there is no need to define an explicit model for the distribution of the missing values. Various metrics are possible to define the distance between the cases. The predictive mean matching metric was proposed by Little (1988). This metric is particularly useful for missing data applications because it is optimized for each target variable separately. The predicted value only needs to be a convenient one-number summary of the important information that relates the covariates to the target. Calculation is straightforward, and it is easy to include nominal and ordinal variables. Algorithm 3 provides the steps used in predictive mean matching using Bayesian parameter draws for $\boldsymbol{\beta}$.

Algorithm 3: Predictive mean matching imputation

Input: $\boldsymbol{y}_{\text{obs}}$, $\boldsymbol{X}_{\text{obs}}$, $\boldsymbol{X}_{\text{mis}}$
Output: $\dot{\boldsymbol{y}}$

1 calculate the cross-product matrix $\boldsymbol{S} = \boldsymbol{X}'_{\text{obs}}\boldsymbol{X}_{\text{obs}}$
2 calculate $\boldsymbol{V} = (\boldsymbol{S} + \text{diag}(\boldsymbol{S})\kappa)^{-1}$, with some small ridge parameter κ
3 calculate regression weights $\hat{\boldsymbol{\beta}} = \boldsymbol{V}\boldsymbol{X}'_{\text{obs}}\boldsymbol{y}_{\text{obs}}$
4 draw a random variable $\dot{g} \sim \chi^2_\nu$ with $\nu = n_1 - q$
5 calculate $\dot{\sigma}^2 = (\boldsymbol{y}_{\text{obs}} - \boldsymbol{X}_{\text{obs}}\hat{\boldsymbol{\beta}})'(\boldsymbol{y}_{\text{obs}} - \boldsymbol{X}_{\text{obs}}\hat{\boldsymbol{\beta}})/\dot{g}$
6 draw q independent $\mathcal{N}(0,1)$ variates in vector $\dot{\boldsymbol{z}}_1$
7 calculate $\boldsymbol{V}^{1/2}$ by Cholesky descomposition
8 calculate $\dot{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} + \dot{\sigma}\dot{\boldsymbol{z}}_1\boldsymbol{V}^{1/2}$
9 calculate $\dot{\eta}(i,j) = |\boldsymbol{X}_{\text{obs},[i]}\hat{\boldsymbol{\beta}} - \boldsymbol{X}_{\text{mis},[j]}\dot{\boldsymbol{\beta}}|$ with $i = 1,...,n_1$ and $j = 1,...,n_0$
10 construct n_0 sets $\boldsymbol{Z}_j$, each containing d candidate donors, from $\boldsymbol{y}_{\text{obs}}$ such that $\sum_d \dot{\eta}(i,j)$ is minimum for all $j = 1,...,n_0$. Break ties randomly
11 draw one donor i_j from $\boldsymbol{Z}_j$ randomly for $j = 1,...,n_0$
12 calculate imputations $\dot{y}_j = y_{i_j}$ for $j = 1,...,n_0$

Additional imputation methods can be consulted in the statistical literature. The methods mentioned in the previous sections are of special interest for comparison and diagnosis issues of our proposed methodology. For the implementation of the same, the `MICE` package (Van Buuren and Groothuis-Oudshoorn, 2011) will be used on R software.

Chapter 3

Cluster Weighted Modeling

"Well, I must endure the presence of a few caterpillars
if I wish to become acquainted with the butterflies."

The Little Prince.

3.1 Overview

Finite mixture models are increasingly exploited due to their convenient way to model unknown distributions and their applications where the data to be analyzed has a group structure or where the objective is to explore the data for said structure. These applications support a variety of techniques in various areas of statistics, including cluster and latent class analysis, discriminant analysis, image analysis, and survival analysis, in addition to its more direct role in inference and data analysis by providing descriptive models for distributions where a single component distribution is apparently inadequate (McLachlan, Lee, and Rathnayake, 2019). Of special interest in this work, the Cluster-Weighted Modeling, proposed by Gershenfeld (1999) under linear and Gaussian assumptions, and its relationship with the finite mixture models is presented. A Bayesian-type approach to the estimation process is also described.

In Section 3.2 finite mixture models are presented, specifically the mixture of distributions and the mixture of regressions in the Gaussian case. Section 3.3 presents the Cluster Weighted-Modeling, in particular the Gaussian linear case. Three results that relate the distribution model, the regression model and the linear Cluster Weighted-Modeling are stated and discussed in the Gaussian context. In Section 3.4 we generalize the linear Gaussian CWM for the multivariate case. Finally, Section 3.5 presents a Bayesian approach through which the estimation and imputation process are implemented from the Linear Gaussian CWM. The results obtained in Sections 3.3 and 3.4 allow to establish the theoretical basis for the algorithm proposed in this last section.

3.2 Finite Mixture Model

3.2.1 Gaussian mixture model

Finite mixture models (FMM) are used to treat heterogeneous data in various experimental situations. Such data arise in practical problems when measurements of the random variable are taken in two or more different conditions. They can be interpreted as if the information came from subpopulations that are called components.

Obtaining these components leads to the estimation of the parameters of the mixture. Several textbooks and documents on studies of the theory of finite mixtures of distributions can be consulted (Frühwirth-Schnatter, 2006; McLachlan and Peel, 2004; Nguyen, 2015).

Consider a population composed of G subgroups, randomly mixed according to the proportions $\alpha_1, ..., \alpha_G$. Suppose the interest is in some random heterogeneous characteristic Y, but homogeneous within subgroups. However, due to heterogeneity, Y has a different probability distribution in each group, but it is generally assumed to come from the same parametric family $p(y|\boldsymbol{\theta})$, with parameter $\boldsymbol{\theta}$ that differs between groups. Groups can be labeled using a discrete indicator variable Z that takes values in the set $\{1, ..., G\}$.

When such a population is randomly sampled, it is possible to record the variable of interest Y together with the indicator for group Z. The sampling probability of the group marked with Z is equal to α_Z, whereas, since Z is known, Y is a random variable that follows the distribution $p(y|\boldsymbol{\theta}_Z)$ with $\boldsymbol{\theta}_Z$ the parameter in the group Z. The joint density $p(y, Z)$ is given by,

$$\begin{aligned} p(y, Z) &= p(y|Z)p(Z) \\ &= p(y|\boldsymbol{\theta}_Z)\alpha_Z. \end{aligned} \tag{3.1}$$

FMM arises if it is not possible to know the group indicator Z; what is observed is only the random variable Y. The marginal density $p(y)$ is given by the mixture of densities,

$$\begin{aligned} p(y) &= \sum_{Z=1}^{G} p(y, Z) \\ &= \sum_{Z=1}^{G} p(y|\boldsymbol{\theta}_Z)\alpha_Z. \end{aligned} \tag{3.2}$$

When the distribution function $p(y|\boldsymbol{\theta}_Z)$ in the expression (3.2) corresponds to a normal distribution in which the parameter $\boldsymbol{\theta}_Z = (\mu_Z, \sigma_Z^2)$, it is called a Gaussian FMM. In the case where $\boldsymbol{Y}$ is a random vector and $\boldsymbol{\theta}_Z = (\boldsymbol{\mu}_Z, \Sigma_Z)$ where $\boldsymbol{\mu}$ is a vector of means and Σ_Z a matrix of variances and covariances, it is known as multivariate Gaussian FMM.

3.2.2 Gaussian mixture of regression model

In data analysis, it is often of more interest to explore the relationship of some random variable Y and a vector of covariates $\boldsymbol{X}$, than to simply explore the distribution of Y alone. Regression analysis is the process of modeling such relationships through the density function for the variable $Y|\boldsymbol{X} = \boldsymbol{x}$, which can be written as $p(y|\boldsymbol{x})$. These types of density functions are called regression models. Like density estimation in general, regression analysis is most commonly performed using parametric models, such that $p(y|\boldsymbol{x}) = p(y|\boldsymbol{x}, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the vector of parameters.

Like the general FMM discussed in Section 3.2.1, assume that there is some discrete indicator variable Z, where the sampling probability of the group marked with Z is equal to α_Z. Since Z is known, suppose that $Y|\boldsymbol{X}, Z$ has by density $p(y|\boldsymbol{x}, \boldsymbol{\theta}_Z)$,

then we have a *mixture of regression model* (MRM) if

$$p(y|\boldsymbol{x}) = \sum_{Z=1}^{G} p(y|\boldsymbol{x}, \boldsymbol{\theta}_Z)\alpha_Z. \tag{3.3}$$

If the density $p(y|\boldsymbol{x}, \boldsymbol{\theta}_Z)$ coincides with the univariate normal density $\phi_1(y; \boldsymbol{b}'_Z\boldsymbol{x} + b_{Z,0}, \sigma^2_Z)$, in which the parameter $\boldsymbol{\theta}_Z = (\boldsymbol{\beta}'_Z, \sigma^2_Z)$ with $\boldsymbol{\beta}_Z = (\boldsymbol{b}'_Z, b_{Z,0})'$, $\boldsymbol{b}_Z \in \mathbb{R}^d$, and $b_{Z,0} \in \mathbb{R}$, the expression in (3.3) is called Gaussian MRM.

The model in (3.3) could be extended to the case where $\boldsymbol{Y}$ is a $p-$dimensional random vector and thus, $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}_Z)$ will coincide with the $p-$variate normal density $\phi_p(\boldsymbol{y}; B'_Z\boldsymbol{x} + \boldsymbol{b}_{Z,0}, \Sigma_Z)$, where the parameter $\boldsymbol{\theta}_Z = (\boldsymbol{B}_Z, \Sigma_Z)$ with $\boldsymbol{B}_Z = (B'_Z, \boldsymbol{b}_{Z,0})'$. Here $B_Z \in \mathbb{R}^{d\times p}$ is a matrix of coefficients and $\boldsymbol{b}_{Z,0} \in \mathbb{R}^p$ for each value of Z. The $i-$th column of the matrix $\boldsymbol{B}_Z$ corresponds to the relationship between the component vector $\boldsymbol{y}_i$ and the vector of covariates $\boldsymbol{x}$. This model is called multivariate Gaussian MRM.

3.3 Cluster-Weighted Modeling

The *Cluster-Weighted Modeling* (CWM) is a procedure that seeks to model the joint probability of data that comes from a heterogeneous population. It is a flexible approach for the statistical modeling of a wide variety of random phenomena, characterized by unobserved heterogeneity. In the context of Sections 3.2.1 and 3.2.2, the CWM seeks to model the densities of the variable Y and the covariates $\boldsymbol{X}$ jointly. CWM was initially introduced by Gershenfeld (1997) for modeling time series data related to musical instrument parameters. Within the framework of the Gaussian MRM, Ingrassia, Minotti, and Vittadini (2012) propose the CWM in a statistical environment, and show that it is a general and flexible family of mixture models.

In a general context, the CWM decomposes the joint probability $p(\boldsymbol{x}, y)$ as follows,

$$p(\boldsymbol{x}, y) = \sum_{Z=1}^{G} p(y|\boldsymbol{x}, Z)p(\boldsymbol{x}|Z)\alpha_Z, \tag{3.4}$$

where $p(y|\boldsymbol{x}, Z)$ is the conditional density of the response variable Y given the predictor vector $\boldsymbol{X}$ in the component indicated by the group Z, $p(\boldsymbol{x}|Z)$ is the probability density of the variable $\boldsymbol{X}$ in the group Z, and α_Z is the sampling probability of the group marked with Z. Therefore, the joint density of $(\boldsymbol{X}, Y)$ can be seen as a mixture of local models $p(y|\boldsymbol{x}, Z)$ weighted (in a broader sense) by both, local densities $p(\boldsymbol{x}|Z)$ and mixing weights α_Z.

For applications whose purposes are classification, the interest is focused on the posterior probability $p(Z|\boldsymbol{x}, y)$ that the observation $(\boldsymbol{x}, y)$ belongs to the component Z given by,

$$p(Z|\boldsymbol{x}, y) = \frac{p(y|\boldsymbol{x}, Z)p(\boldsymbol{x}|Z)\alpha_Z}{\sum_{Z=1}^{G} p(y|\boldsymbol{x}, Z)p(\boldsymbol{x}|Z)\alpha_Z}, \tag{3.5}$$

that is, the classification of each observation depends on both the marginal and conditional densities. Furthermore, with some simple calculations, it can also be obtained that,

$$p(Z|\boldsymbol{x}, y) = \frac{p(y|\boldsymbol{x}, Z)p(Z|\boldsymbol{x})}{\sum_{Z=1}^{G} p(y|\boldsymbol{x}, Z)p(Z|\boldsymbol{x})},$$

with,

$$p(Z|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|Z)\alpha_Z}{\sum_{Z=1}^{G} p(\boldsymbol{x}|Z)\alpha_Z}, \tag{3.6}$$

where $p(Z|\boldsymbol{x})$ is the posterior probability that the observation $\boldsymbol{x}$ belongs to the component Z.

3.3.1 Gaussian Linear CWM

The basic model presented by Ingrassia, Minotti, and Vittadini (2012) is based on considering the marginal and conditional distributions as normal distributions. Thus, $p(\boldsymbol{x}|Z) = \phi_d(\boldsymbol{x}; \boldsymbol{\mu}_Z, \Sigma_Z)$, while $p(y|\boldsymbol{x}, Z) = \phi_1(y; \mu(\boldsymbol{x}, \boldsymbol{\beta}_Z), \sigma_Z^2)$, where the conditional density is based on linear mappings, i.e. $\mu(\boldsymbol{x}, \boldsymbol{\beta}_Z) = \boldsymbol{b}_Z'\boldsymbol{x} + b_{Z,0}$, for some $\boldsymbol{\beta}_Z = (\boldsymbol{b}_Z', b_{Z,0})'$, with $\boldsymbol{b}_Z \in \mathbb{R}^d$ and $b_{Z,0} \in \mathbb{R}$. Under these conditions, the expression in (3.4) takes the form,

$$p(\boldsymbol{x}, y) = \sum_{Z=1}^{G} \phi_1(y; \boldsymbol{b}_Z'\boldsymbol{x} + b_{Z,0}, \sigma_Z^2)\phi_d(\boldsymbol{x}; \boldsymbol{\mu}_Z, \Sigma_Z)\alpha_Z. \tag{3.7}$$

The density in (3.7) is called Gaussian Linear CWM (LCWM). For this case and in the context of the Gaussian LCWM, the posterior probability in (3.5) can be written as

$$p(Z|\boldsymbol{x}, y) = \frac{\phi_1(y; \boldsymbol{b}_Z'\boldsymbol{x} + b_{Z,0}, \sigma_Z^2)\phi_d(\boldsymbol{x}; \boldsymbol{\mu}_Z, \Sigma_Z)\alpha_Z}{\sum_{Z=1}^{G} \phi_1(y; \boldsymbol{b}_Z'\boldsymbol{x} + b_{Z,0}, \sigma_Z^2)\phi_d(\boldsymbol{x}; \boldsymbol{\mu}_Z, \Sigma_Z)\alpha_Z}, \tag{3.8}$$

while, the expression in (3.6) takes the form

$$p(Z|\boldsymbol{x}) = \frac{\phi_d(\boldsymbol{x}; \boldsymbol{\mu}_Z, \Sigma_Z)\alpha_Z}{\sum_{Z=1}^{G} \phi_d(\boldsymbol{x}; \boldsymbol{\mu}_Z, \Sigma_Z)\alpha_Z}. \tag{3.9}$$

3.3.2 Some relationships between FMM, MRM, and LCWM in the Gaussian case

In this section, three results are presented that relate FMM, MRM and LCWM in the Gaussian case. The results that are included for these models have to do with the probability distribution functions and the posterior probabilities.

Let $\boldsymbol{W}$ be a random vector that takes values in $\mathbb{R}^{d+1}$ with joint probability distribution $p(\boldsymbol{w})$. Assume that the density $p(\boldsymbol{w})$ of $\boldsymbol{W}$ corresponds to an FMM, that is,

$$p(\boldsymbol{w}) = \sum_{Z=1}^{G} p(\boldsymbol{w}|Z)\alpha_Z,$$

where $p(\boldsymbol{w}|Z)$ is the probability density of $\boldsymbol{W}|Z$, and $\alpha_Z = p(Z)$ is the mixing weight of the component marked with $Z \in \{1, ..., G\}$. In the case of coinciding with a Gaussian FMM, let $\boldsymbol{\mu}_Z^{(\boldsymbol{w})}$ and $\Sigma_Z^{(\boldsymbol{w})}$ be the vector of means and the covariance matrix of

$\boldsymbol{W}|Z$, respectively.

Suppose that $\boldsymbol{W} = (\boldsymbol{X}', Y)'$, where $\boldsymbol{X}$ is a random vector taking values in $\mathbb{R}^d$ and Y is a random variable. Then,

$$\boldsymbol{\mu}_Z^{(\boldsymbol{w})} = \begin{pmatrix} \boldsymbol{\mu}_Z^{(\boldsymbol{x})} \\ \mu_Z^{(y)} \end{pmatrix} \quad \text{and} \quad \Sigma_Z^{(\boldsymbol{w})} = \begin{pmatrix} \Sigma_Z^{(\boldsymbol{xx})} & \Sigma_Z^{(\boldsymbol{x}y)} \\ \Sigma_Z^{(y\boldsymbol{x})} & \sigma_Z^{2(y)} \end{pmatrix}.$$

A first interesting result indicates that the CWM contains the FMM and, specifically, in the Gaussian context, restricting the CWM to the case LCWM, FMM and LCWM are equivalent. The proofs of the three results presented below can be consulted at Ingrassia, Minotti, and Vittadini (2012) and Nguyen (2015), here only the propositions are stated.

Proposition 3.3.1. *Let $\boldsymbol{W}$ be a random vector that takes values in a subset of $\mathbb{R}^{d+1}$, and suppose that $\boldsymbol{W}|Z \sim \mathcal{N}_{d+1}\left(\boldsymbol{\mu}_Z^{(\boldsymbol{w})}, \Sigma_Z^{(\boldsymbol{w})}\right)$ with $Z \in \{1, ..., G\}$. In particular, the density $p(\boldsymbol{w})$ of $\boldsymbol{W}$ is a Gaussian FMM:*

$$p(\boldsymbol{w}) = \sum_{Z=1}^{G} \phi_{d+1}\left(\boldsymbol{w}; \boldsymbol{\mu}_Z^{(\boldsymbol{w})}, \Sigma_Z^{(\boldsymbol{w})}\right) \alpha_Z.$$

So, $p(\boldsymbol{w})$ can be written similarly to (3.7), that is, a Gaussian LCWM.

From the proof of Proposition 3.3.1 presented in Ingrassia, Minotti, and Vittadini (2012), it is worth highlighting how the density $p(\boldsymbol{w})$ is written to bring it to the structure of a Gaussian LCWM. So,

$$p(\boldsymbol{w}) = \sum_{Z=1}^{G} \phi_1\left(y|\boldsymbol{x}; \mu_Z^{(y|\boldsymbol{x})}, \sigma_Z^{2(y|\boldsymbol{x})}\right) \phi_d\left(\boldsymbol{x}; \boldsymbol{\mu}_Z^{(\boldsymbol{x})}, \Sigma_Z^{(\boldsymbol{xx})}\right) \alpha_Z,$$

where

$$\begin{aligned} \mu_Z^{(y|\boldsymbol{x})} &= \mu_Z^{(y)} + \Sigma_Z^{(y\boldsymbol{x})} \Sigma_Z^{(\boldsymbol{xx})^{-1}} \left(\boldsymbol{x} - \boldsymbol{\mu}_Z^{(\boldsymbol{x})}\right) \\ &= \left[\mu_Z^{(y)} - \Sigma_Z^{(y\boldsymbol{x})} \Sigma_Z^{(\boldsymbol{xx})^{-1}} \boldsymbol{\mu}_Z^{(\boldsymbol{x})}\right] + \left[\Sigma_Z^{(y\boldsymbol{x})} \Sigma_Z^{(\boldsymbol{xx})^{-1}}\right] \boldsymbol{x} \\ &= b_{Z,0} + \boldsymbol{b}_Z \boldsymbol{x}, \end{aligned}$$

and,

$$\begin{aligned} \sigma_Z^{2(y|\boldsymbol{x})} &= \sigma_Z^{2(y)} - \Sigma_Z^{(y\boldsymbol{x})} \Sigma_Z^{(\boldsymbol{xx})^{-1}} \Sigma_Z^{(\boldsymbol{x}y)} \\ &= \sigma_Z^2. \end{aligned}$$

Using arguments similar to those used in the proof of Proposition 3.3.1, it can be concluded that FMM and LCWM have the same posterior probability distribution.

The second result involves the MRM in the Gaussian case given in a general way by (3.3) and that, for the Gaussian case, is specified by

$$p(y|\boldsymbol{x}) = \sum_{Z=1}^{G} \phi_1(y; \boldsymbol{b}_Z' \boldsymbol{x} + b_{Z,0}, \sigma_Z^2) \alpha_Z, \tag{3.10}$$

and for which, the posterior probability $p(Z|\boldsymbol{x}, y)$ is given by the expression

$$p(Z|\boldsymbol{x}, y) = \frac{\phi_1(y; \boldsymbol{b}'_Z\boldsymbol{x} + b_{Z,0}, \sigma_Z^2)\alpha_Z}{\sum_{Z=1}^{G} \phi_1(y; \boldsymbol{b}'_Z\boldsymbol{x} + b_{Z,0}, \sigma_Z^2)\alpha_Z}. \tag{3.11}$$

Proposition 3.3.2. *Consider the Gaussian LCWM given in (3.7), with $\boldsymbol{X}|Z \sim \mathcal{N}_d(\boldsymbol{\mu}_Z, \Sigma_Z)$ and $Z \in \{1, ..., G\}$. If the probability density of $\boldsymbol{X}|Z$ does not depend on the component, that is, $\phi_d(\boldsymbol{x}; \boldsymbol{\mu}_Z, \Sigma_Z) = \phi_d(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma)$ for all $Z \in \{1, ..., G\}$, then it follows that,*

$$p(\boldsymbol{x}, y) = \phi_d(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma)p(y|\boldsymbol{x}), \tag{3.12}$$

where $p(y|\boldsymbol{x})$ is the Gaussian MRM given in (3.10).

Proposition 3.3.2 establishes an expression that relates, in the Gaussian context, the LCWM and the MRM when the covariate $\boldsymbol{x}$ has the same behavior between components.

Finally, a result is presented as a corollary that states, assuming that the covariate $\boldsymbol{x}$ has the same behavior between components, that the posterior probabilities for LCWM and MRM in the Gaussian case coincide.

Corollary 3.3.1. *If the probability density of $\boldsymbol{X}|Z$ does not depend on the component, i.e., $\phi_d(\boldsymbol{x}; \boldsymbol{\mu}_Z, \Sigma_Z) = \phi_d(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma)$ for all $Z \in \{1, ..., G\}$, then the posterior probability in (3.8) coincides with (3.11).*

Additionally, it can be concluded that from the conditions of the Corollary 3.3.1, the posterior probabilities in (3.9) simplify to $p(Z|\boldsymbol{x}) = \alpha_Z$.

3.4 Multivariate Gaussian Linear CWM

In the previous sections, the univariate model was discussed in such a way that it was considered an output random variable $Y \in \mathbb{R}$. This idea can be generalized to an output random vector $\boldsymbol{Y} \in \mathbb{R}^p$. Next, the same results are established that relate the FMM, MRM and LCWM models in the multivariate Gaussian case, a discussion similar to that made in Section 3.3.2 for the univariate case. The propositions and the corollary are presented including their proofs.

Let $\boldsymbol{W}$ be a random vector that takes values in $\mathbb{R}^{d+p}$ with joint probability distribution $p(\boldsymbol{w})$. Assume that the density $p(\boldsymbol{w})$ of $\boldsymbol{W}$ corresponds to a FMM, that is,

$$p(\boldsymbol{w}) = \sum_{Z=1}^{G} p(\boldsymbol{w}|Z)\alpha_Z,$$

where $p(\boldsymbol{w}|Z)$ is the probability density of $\boldsymbol{W}|Z$, and $\alpha_Z = p(Z)$ is the mixing weight of the component marked with $Z \in \{1, ..., G\}$. In the case of coinciding with a Gaussian FMM, let $\boldsymbol{\mu}_Z^{(\boldsymbol{w})}$ and $\Sigma_Z^{(\boldsymbol{w})}$ be the vector of means and the covariance matrix of $\boldsymbol{W}|Z$, respectively.

Suppose that $\boldsymbol{W} = (\boldsymbol{X}', \boldsymbol{Y}')'$, where $\boldsymbol{X}$ is a random vector taking values in $\mathbb{R}^d$, and $\boldsymbol{Y}$ is a random vector in $\mathbb{R}^p$. Then,

$$\boldsymbol{\mu}_Z^{(\boldsymbol{w})} = \begin{pmatrix} \boldsymbol{\mu}_Z^{(\boldsymbol{x})} \\ \boldsymbol{\mu}_Z^{(\boldsymbol{y})} \end{pmatrix} \quad \text{and} \quad \Sigma_Z^{(\boldsymbol{w})} = \begin{pmatrix} \Sigma_Z^{(\boldsymbol{xx})} & \Sigma_Z^{(\boldsymbol{xy})} \\ \Sigma_Z^{(\boldsymbol{yx})} & \Sigma_Z^{(\boldsymbol{yy})} \end{pmatrix}.$$

Proposition 3.4.1 indicates that the CWM contains the FMM, and in the Gaussian context, restricting the CWM to the case LCWM, FMM and LCWM are equivalent.

Proposition 3.4.1. *Let $\boldsymbol{W}$ be a random vector that takes values in a subset of $\mathbb{R}^{d+p}$, and suppose that $\boldsymbol{W}|Z \sim \mathcal{N}_{d+p}\left(\boldsymbol{\mu}_Z^{(\boldsymbol{w})}, \Sigma_Z^{(\boldsymbol{w})}\right)$ with $Z \in \{1, ..., G\}$. In particular, the density $p(\boldsymbol{w})$ of $\boldsymbol{W}$ is a Gaussian FMM:*

$$p(\boldsymbol{w}) = \sum_{Z=1}^{G} \phi_{d+p}\left(\boldsymbol{w}; \boldsymbol{\mu}_Z^{(\boldsymbol{w})}, \Sigma_Z^{(\boldsymbol{w})}\right) \alpha_Z. \tag{3.13}$$

So, $p(\boldsymbol{w})$ can be written similarly to

$$p(\boldsymbol{x}, \boldsymbol{y}) = \sum_{Z=1}^{G} \phi_p(\boldsymbol{y}; B_Z'\boldsymbol{x} + \boldsymbol{b}_{Z,0}, \widetilde{\Sigma}_Z)\phi_d(\boldsymbol{x}; \boldsymbol{\mu}_Z, \Sigma_Z)\alpha_Z, \tag{3.14}$$

that is, a Gaussian LCWM.

Proof. Let us set $\boldsymbol{W} = (\boldsymbol{X}', \boldsymbol{Y}')'$, where $\boldsymbol{X}$ is a $d-$dimensional random vector and $\boldsymbol{Y}$ is a $p-$dimensional random vector. Using properties of the multivariate normal distribution (e.g. Johnson, Wichern, et al., 2002),

$$\begin{aligned} p(\boldsymbol{w}) &= \sum_{Z=1}^{G} \phi_{d+p}\left((\boldsymbol{x}, \boldsymbol{y}); \boldsymbol{\mu}_Z^{(\boldsymbol{w})}, \Sigma_Z^{(\boldsymbol{w})}\right) \alpha_Z \\ &= \sum_{Z=1}^{G} \phi_p\left(\boldsymbol{y}; \boldsymbol{\mu}_Z^{(\boldsymbol{y}|\boldsymbol{x})}, \Sigma_Z^{(\boldsymbol{y}|\boldsymbol{x})}\right) \phi_d\left(\boldsymbol{x}; \boldsymbol{\mu}_Z^{(\boldsymbol{x})}, \Sigma_Z^{(\boldsymbol{x})}\right) \alpha_Z, \end{aligned}$$

where,

$$\begin{aligned} \boldsymbol{\mu}_Z^{(\boldsymbol{y}|\boldsymbol{x})} &= \boldsymbol{\mu}_Z^{(\boldsymbol{y})} + \Sigma_Z^{(\boldsymbol{y}\boldsymbol{x})}\Sigma_Z^{(\boldsymbol{x}\boldsymbol{x})^{-1}}\left(\boldsymbol{x} - \boldsymbol{\mu}_Z^{(\boldsymbol{x})}\right) \\ &= \left[\Sigma_Z^{(\boldsymbol{y}\boldsymbol{x})}\Sigma_Z^{(\boldsymbol{x}\boldsymbol{x})^{-1}}\right]\boldsymbol{x} + \left[\boldsymbol{\mu}_Z^{(\boldsymbol{y})} - \Sigma_Z^{(\boldsymbol{y}\boldsymbol{x})}\Sigma_Z^{(\boldsymbol{x}\boldsymbol{x})^{-1}}\boldsymbol{\mu}_Z^{(\boldsymbol{x})}\right] \\ &= B_Z'\boldsymbol{x} + \boldsymbol{b}_{Z,0} \end{aligned}$$

and

$$\begin{aligned} \Sigma_Z^{(\boldsymbol{y}|\boldsymbol{x})} &= \Sigma_Z^{(\boldsymbol{y}\boldsymbol{y})} - \Sigma_Z^{(\boldsymbol{y}\boldsymbol{x})}\Sigma_Z^{(\boldsymbol{x}\boldsymbol{x})^{-1}}\Sigma_Z^{(\boldsymbol{x}\boldsymbol{y})} \\ &= \widetilde{\Sigma}_Z. \end{aligned}$$

Then, (3.13) can be written as (3.14). □

The expressions used in the proof of Proposition 3.4.1 are used in the programming process of the method that we propose here. Similar to the univariate case, the proposition below establishes an expression that relates the LCWM and the MRM, when the covariate $\boldsymbol{x}$ has the same behavior between components.

Proposition 3.4.2. *Consider the Gaussian LCWM given in (3.14), with $\boldsymbol{X}|Z \sim \mathcal{N}_d(\boldsymbol{\mu}_Z, \Sigma_Z)$ and $Z \in \{1, ..., G\}$. If the probability density of $\boldsymbol{X}|Z$ does not depend on the component, that is, $\phi_d(\boldsymbol{x}; \boldsymbol{\mu}_Z, \Sigma_Z) = \phi_d(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma)$ for all $Z \in \{1, ..., G\}$, then it follows that*

$$p(\boldsymbol{x}, \boldsymbol{y}) = \phi_d(\boldsymbol{\mu}, \Sigma)p(\boldsymbol{y}|\boldsymbol{x}), \tag{3.15}$$

where $p(\boldsymbol{y}|\boldsymbol{x})$ is the Gaussian MRM given by the expression

$$p(\boldsymbol{y}|\boldsymbol{x}) = \sum_{Z=1}^{G} \phi_p(\boldsymbol{y}; B_Z'\boldsymbol{x} + \boldsymbol{b}_{Z,0}, \widetilde{\Sigma}_Z)\alpha_Z. \tag{3.16}$$

Proof. Assume that $\phi_d(\boldsymbol{x}; \boldsymbol{\mu}_Z, \Sigma_Z) = \phi_d(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma)$ for all $Z \in \{1, ..., G\}$, then from the expression in (3.14)

$$\begin{aligned} p(\boldsymbol{x}, \boldsymbol{y}) &= \sum_{Z=1}^{G} \phi_p(\boldsymbol{y}; B_Z'\boldsymbol{x} + \boldsymbol{b}_{Z,0}, \widetilde{\Sigma}_Z)\phi_d(\boldsymbol{x}; \boldsymbol{\mu}_Z, \Sigma_Z)\alpha_Z \\ &= \phi_d(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) \sum_{Z=1}^{G} \phi_p(\boldsymbol{y}; B_Z'\boldsymbol{x} + \boldsymbol{b}_{Z,0}, \widetilde{\Sigma}_Z)\alpha_Z \\ &= \phi_d(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma)p(\boldsymbol{y}|\boldsymbol{x}) \end{aligned}$$

where $p(\boldsymbol{y}|\boldsymbol{x})$ is the Gaussian MRM given in (3.16). □

As the last result in this section, Corollary 3.4.1 states, assuming that the covariate $\boldsymbol{x}$ has the same behavior between components, that the posterior probabilities for LCWM and MRM in the Gaussian case coincide.

Corollary 3.4.1. *If the probability density of $\boldsymbol{X}|Z$ does not depend on the component, that is, $\phi_d(\boldsymbol{x}; \boldsymbol{\mu}_Z, \Sigma_Z) = \phi_d(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma)$ for all $Z \in \{1, ..., G\}$, then the posterior probability given by*

$$p(Z|\boldsymbol{x}, y) = \frac{\phi_p(\boldsymbol{y}; B_Z'\boldsymbol{x} + \boldsymbol{b}_{Z,0}, \widetilde{\Sigma}_Z)\phi_d(\boldsymbol{x}; \boldsymbol{\mu}_Z, \Sigma_Z)\alpha_Z}{\sum_{Z=1}^{G} \phi_p(\boldsymbol{y}; B_Z'\boldsymbol{x} + \boldsymbol{b}_{Z,0}, \widetilde{\Sigma}_Z)\phi_d(\boldsymbol{x}; \boldsymbol{\mu}_Z, \Sigma_Z)\alpha_Z}, \tag{3.17}$$

coincides with

$$p(Z|\boldsymbol{x}, y) = \frac{\phi_p(\boldsymbol{y}; B_Z'\boldsymbol{x} + \boldsymbol{b}_{Z,0}, \widetilde{\Sigma}_Z)\alpha_Z}{\sum_{Z=1}^{G} \phi_p(\boldsymbol{y}; B_Z'\boldsymbol{x} + \boldsymbol{b}_{Z,0}, \widetilde{\Sigma}_Z)\alpha_Z}. \tag{3.18}$$

Proof. Assume that $\phi_d(\boldsymbol{x}; \boldsymbol{\mu}_Z, \Sigma_Z) = \phi_d(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma)$ for all $Z \in \{1, ..., G\}$, from the expression in (3.17) we get,

$$\begin{aligned} p(Z|\boldsymbol{x}, y) &= \frac{\phi_p(\boldsymbol{y}; B_Z'\boldsymbol{x} + \boldsymbol{b}_{Z,0}, \widetilde{\Sigma}_Z)\phi_d(\boldsymbol{x}; \boldsymbol{\mu}_Z, \Sigma_Z)\alpha_Z}{\sum_{Z=1}^{G} \phi_p(\boldsymbol{y}; B_Z'\boldsymbol{x} + \boldsymbol{b}_{Z,0}, \widetilde{\Sigma}_Z)\phi_d(\boldsymbol{x}; \boldsymbol{\mu}_Z, \Sigma_Z)\alpha_Z} \\ &= \frac{\phi_d(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma)\phi_p(\boldsymbol{y}; B_Z'\boldsymbol{x} + \boldsymbol{b}_{Z,0}, \widetilde{\Sigma}_Z)\alpha_Z}{\phi_d(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma)\sum_{Z=1}^{G} \phi_p(\boldsymbol{y}; B_Z'\boldsymbol{x} + \boldsymbol{b}_{Z,0}, \widetilde{\Sigma}_Z)\alpha_Z} \\ &= \frac{\phi_p(\boldsymbol{y}; B_Z'\boldsymbol{x} + \boldsymbol{b}_{Z,0}, \widetilde{\Sigma}_Z)\alpha_Z}{\sum_{Z=1}^{G} \phi_p(\boldsymbol{y}; B_Z'\boldsymbol{x} + \boldsymbol{b}_{Z,0}, \widetilde{\Sigma}_Z)\alpha_Z}, \end{aligned}$$

for $Z \in \{1, ..., G\}$. □

3.5 Bayesian Estimation and Imputation

Since the Propositions 3.3.1 and 3.4.1 establish a mapping between the parameter vectors of the FMM and the LCWM in the Gaussian context, the interest initially lies in the estimation of the FMM parameters. For this purpose, a Bayesian approach is implemented, using for the prior distribution of the mixing weights the stick-breaking representation of a truncated Dirichlet process (Ferguson, 1973; Sethuraman, 1994), since it has been shown that this class of models allows greater flexibility

and better estimation of density (Müller and Mitra, 2013). Below is a brief description of the process.

Each individual in the data set belongs to one of G mixture components, that is, $Z_i \in \{1, ..., G\}$ so that $\boldsymbol{Z} = (Z_1, ..., Z_n)$. The mixing probabilities are given by $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_G)$ with $\alpha_g = P(Z_i = g)$ where $i = 1, ..., n$ and $g = 1, ..., G$.
If $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_G)$ and $\boldsymbol{\Sigma} = (\Sigma_1, ..., \Sigma_G)$ then,

$$\begin{aligned} \boldsymbol{w}_i | Z_i, \boldsymbol{\mu}, \boldsymbol{\Sigma} &\sim \mathcal{N}_p(\boldsymbol{\mu}_{Z_i}, \Sigma_{Z_i}), \\ U_i | \boldsymbol{\alpha} &\sim \text{Multinomial}(\boldsymbol{\alpha}), \end{aligned}$$

where $Z_i = \Gamma' U_i$ with $\Gamma' = (1, ..., G)$.

The prior distribution for $\boldsymbol{\alpha}$ is a stick-breaking representation of a truncated Dirichlet process (Ferguson, 1973; Sethuraman, 1994),

$$\begin{aligned} \alpha_g &= \nu_g \prod_{k<g} (1 - \nu_k) \text{ for } g = 1, ..., G, \\ \nu_g &\sim \text{Beta}(1, \eta) \text{ for } g = 1, ..., G-1; \quad \nu_G = 1, \\ \eta &\sim \text{Gamma}(a_\eta, b_\eta). \end{aligned}$$

Following Kim et al. (2014), we use values of $a_\eta = b_\eta = .25$, which represents a small prior sample size and hence vague specification for Gamma distributions. This ensures that the information from the data dominates the posterior distribution. The specification of prior distributions encourages α_g to decrease stochastically with g. When η is very small, most of the probability in $\boldsymbol{\alpha}$ is allocated to the first few components, thus reducing the risks of over-fitting the data as well as increasing computational efficiency.

For the prior specification of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$,

$$\begin{aligned} \boldsymbol{\mu}_g | \Sigma_g &\sim \mathcal{N}_p(\boldsymbol{\mu}_0, h^{-1}\Sigma_g), \\ \Sigma_g &\sim \text{Inverse Wishart}(f, \Delta). \end{aligned}$$

Here, f is the prior degrees of freedom, and $\Delta = \text{diag}(\delta_1, ..., \delta_p)$ is a diagonal matrix of size $p \times p$ with $\delta_j \sim \text{Gamma}(a_\delta, b_\delta)$ for $j = 1, ..., p$. We use a prior mean of $\boldsymbol{\mu}_0$ equal to the mean of the data set, using $f = p + 1$ degrees of freedom to ensure a proper distribution without overly constraining Σ, and setting $h = 1$ mostly for convenience. We use $a_\delta = b_\delta = .25$, a modest value but not too small, so as to allow substantial prior mass at modest-sized variances (Paiva, 2014; Kim et al., 2014).
Once the estimates for $\boldsymbol{\mu}_g$ and Σ_g have been obtained, using the $\boldsymbol{W} = (\boldsymbol{X}', \boldsymbol{Y}')'$ notation and the results of Proposition 3.4.1 we have

$$\boldsymbol{\mu}_g^{(\boldsymbol{w})} = \begin{pmatrix} \boldsymbol{\mu}_g^{(\boldsymbol{x})} \\ \boldsymbol{\mu}_g^{(\boldsymbol{y})} \end{pmatrix} \quad \text{and} \quad \Sigma_g^{(\boldsymbol{w})} = \begin{pmatrix} \Sigma_g^{(\boldsymbol{xx})} & \Sigma_g^{(\boldsymbol{xy})} \\ \Sigma_g^{(\boldsymbol{yx})} & \Sigma_g^{(\boldsymbol{yy})} \end{pmatrix},$$

the estimates for the parameters of the conditional model are obtained as

$$\begin{aligned} B_g' &= \Sigma_g^{(\boldsymbol{yx})} {\Sigma_g^{(\boldsymbol{xx})}}^{-1}, \\ \boldsymbol{b}_{g,0} &= \boldsymbol{\mu}_g^{(\boldsymbol{y})} - \Sigma_g^{(\boldsymbol{yx})} {\Sigma_g^{(\boldsymbol{xx})}}^{-1} \boldsymbol{\mu}_g^{(\boldsymbol{x})}. \end{aligned}$$

It is important to mention that, throughout the document, values for G are used in two senses. The first when implementing the code for the imputation and estimation procedure (I-step and P-step in Algorithm 1). The second is when the code is used in the process of estimating the distribution mixture model in order to evaluate the imputation procedure, in such a way that, assuming the value of G for a set of data, we fit the distribution to original data (or simulated data) and to imputed data and we compare the two distributions using some kind of measure. In the first of the senses, by specifying initial values, it is possible to use the standard Gibbs sampler algorithm to estimate the posterior distribution. For the number of components G, Kim et al. (2015) recommend starting with a somewhat large value, for example $\widetilde{G} = 30$. At each iteration of the Gibbs sampler, the number of nonempty components is counted. If this count reaches the value assigned to $\widetilde{G}$, it is prudent to increase $\widetilde{G}$ and readjust the model with more components. When the count of nonempty components is less than $\widetilde{G}$, then the value of $\widetilde{G}$ is reasonable.

Because the proposed imputation model is based on Algorithm 1, the previous paragraphs describe the P-step of the procedure in detail. Basically, an independent Bayesian analysis is presented that describes the estimation of the posterior distributions of the model parameters. The computational implementation uses a Gibbs sampler algorithm on which an additional imputation step is included. I-step uses the marginal and conditional distributions of the Gaussian LCWM to classify individuals with missing information and impute incomplete variables using the observed variables. The steps of the implemented procedure are presented in Algorithm 4, and are based on the Gibbs sampler for the Gaussian FMM (Kim et al., 2014).

Algorithm 4: Gaussian Linear CWM imputation

Input: $\boldsymbol{y}_{\text{obs}}, \boldsymbol{X}_{\text{obs}}, \boldsymbol{X}_{\text{mis}}$
Output: $\boldsymbol{y}_{\text{mis}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}$
initialization: $\boldsymbol{y}_{\text{mis}}^{(0)}, \boldsymbol{\alpha}^{(0)}, \boldsymbol{\mu}^{(0)}, \boldsymbol{\Sigma}^{(0)}$
for $j = 1, ..., J$ **do**
  generate $\boldsymbol{u}^{(j)}$ from $p(\boldsymbol{u}|\boldsymbol{y}_{\text{obs}}, \boldsymbol{X}_{\text{obs}}, \boldsymbol{X}_{\text{mis}}, \boldsymbol{y}_{\text{mis}}^{(j-1)}, \boldsymbol{\alpha}^{(j-1)}, \boldsymbol{\mu}^{(j-1)}, \boldsymbol{\Sigma}^{(j-1)})$
  compute $\boldsymbol{z}^{(j)} = f(\boldsymbol{u}^{(j)})$
  generate $\boldsymbol{\nu}^{(j)}$ from $p(\boldsymbol{\nu}|\boldsymbol{z}^{(j)})$
  compute $\boldsymbol{\alpha}^{(j)} = \tilde{f}(\boldsymbol{\nu}^{(j)})$
  for $i = 1, ..., G$ **do**
    generate $\Sigma_i^{(j)}$ from $p(\Sigma_i|\boldsymbol{y}_{\text{obs}}, \boldsymbol{X}_{\text{obs}}, \boldsymbol{X}_{\text{mis}}, \boldsymbol{y}_{\text{mis}}^{(j-1)}, \boldsymbol{z}^{(j)})$
    generate $\boldsymbol{\mu}_i^{(j)}$ from $p(\boldsymbol{\mu}_i|\Sigma_i^{(j)}, \boldsymbol{y}_{\text{obs}}, \boldsymbol{X}_{\text{obs}}, \boldsymbol{X}_{\text{mis}}, \boldsymbol{y}_{\text{mis}}^{(j-1)}, \boldsymbol{z}^{(j)})$
  end
  generate $\boldsymbol{u}_{\text{mis}}^{(j)}$ from $p(\boldsymbol{u}_{\text{mis}}|\boldsymbol{X}_{\text{mis}}, \boldsymbol{\alpha}^{(j)}, \boldsymbol{\mu}^{(j)}, \boldsymbol{\Sigma}^{(j)})$
  compute $\boldsymbol{z}_{\text{mis}}^{(j)} = f(\boldsymbol{u}_{\text{mis}}^{(j)})$
  generate $\boldsymbol{y}_{\text{mis}}^{(j)}$ from $p(\boldsymbol{y}_{\text{mis}}|\boldsymbol{X}_{\text{mis}}, \boldsymbol{z}_{\text{mis}}^{(j)}, \boldsymbol{\mu}^{(j)}, \boldsymbol{\Sigma}^{(j)})$
  sort $\boldsymbol{\alpha}^{(j)}$ in decreasing order
  reorder $\boldsymbol{z}^{(j)}, \boldsymbol{z}_{\text{mis}}^{(j)}, \boldsymbol{\mu}^{(j)}, \boldsymbol{\Sigma}^{(j)}$ based on the order of $\boldsymbol{\alpha}^{(j)}$
end
Result: Imputed database and parameters estimation

The output variable can be partitioned in the form, $\boldsymbol{y} = (\boldsymbol{y}_{\text{obs}}, \boldsymbol{y}_{\text{mis}})$, where $\boldsymbol{y}_{\text{obs}}$

denotes, as before, the observed part, and $\boldsymbol{y}_{\text{mis}}$ denotes the missing part. For classification purposes, the notation $\boldsymbol{z}$ is used for the variable that classifies the observations in the parameter model estimation process, while $\boldsymbol{z}_{\text{mis}}$ is used to classify the observations with missing information for the imputation process. Two steps are incorporated into the algorithm to establish the imputation procedure, the first one updates the variable $\boldsymbol{z}_{\text{mis}}$ and the second one updates the imputations $\boldsymbol{y}_{\text{mis}}$.

Algorithm 4 arises from the imputation procedure implemented in Paiva and Reiter (2017), on which it is intended to enter auxiliary information to improve the imputation process of the variables with missing data. Since the diagnosis of our model starts from the comparison with the imputation engine used by Paiva and Reiter (2017), Algorithm 5 describes the steps used in the imputation procedure used there. It is worth mentioning that from our model it is possible to obtain the procedure used by Paiva and Reiter (2017) as a particular case. Thus, by not considering auxiliary information, Algorithm 4, which specifies a form of regression model, becomes a intercept-only model or naïve model, hence the name we give to Algorithm 5 of *mean imputation*. On the other hand, our model also contains a version of the Bayesian multiple imputation method studied in Section 2.7.3. Algorithm 2 can be approximated considering in our model $G = 1$ as the number of clusters. The difference arises in the non-informative prior distribution used for each case.

Algorithm 5: mean imputation

Input: $\boldsymbol{y}_{\text{obs}}$
Output: $\boldsymbol{y}_{\text{mis}}$, $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\Sigma}}$
initialization: $\boldsymbol{y}_{\text{mis}}^{(0)}$, $\boldsymbol{\alpha}^{(0)}$, $\boldsymbol{\mu}^{(0)}$, $\boldsymbol{\Sigma}^{(0)}$
for $j = 1, ..., J$ **do**
  generate $\boldsymbol{u}^{(j)}$ from $p(\boldsymbol{u}|\boldsymbol{y}_{\text{obs}}, \boldsymbol{y}_{\text{mis}}^{(j-1)}, \boldsymbol{\alpha}^{(j-1)}, \boldsymbol{\mu}^{(j-1)}, \boldsymbol{\Sigma}^{(j-1)})$
  compute $\boldsymbol{z}^{(j)} = f(\boldsymbol{u}^{(j)})$
  generate $\boldsymbol{\nu}^{(j)}$ from $p(\boldsymbol{\nu}|\boldsymbol{z}^{(j)})$
  compute $\boldsymbol{\alpha}^{(j)} = \tilde{f}(\boldsymbol{\nu}^{(j)})$
  for $i = 1, ..., G$ **do**
    generate $\Sigma_i^{(j)}$ from $p(\Sigma_i|\boldsymbol{y}_{\text{obs}}, \boldsymbol{y}_{\text{mis}}^{(j-1)}, \boldsymbol{z}^{(j)})$
    generate $\boldsymbol{\mu}_i^{(j)}$ from $p(\boldsymbol{\mu}_i|\Sigma_i^{(j)}, \boldsymbol{y}_{\text{obs}}, \boldsymbol{y}_{\text{mis}}^{(j-1)}, \boldsymbol{z}^{(j)})$
  end
  generate $\boldsymbol{u}_{\text{mis}}^{(j)}$ from $p(\boldsymbol{u}_{\text{mis}}|\boldsymbol{\alpha}^{(j)})$
  compute $\boldsymbol{z}_{\text{mis}}^{(j)} = f(\boldsymbol{u}_{\text{mis}}^{(j)})$
  generate $\boldsymbol{y}_{\text{mis}}^{(j)}$ from $p(\boldsymbol{y}_{\text{mis}}|\boldsymbol{z}_{\text{mis}}^{(j)}, \boldsymbol{\mu}^{(j)}, \boldsymbol{\Sigma}^{(j)})$
  sort $\boldsymbol{\alpha}^{(j)}$ in decreasing order
  reorder $\boldsymbol{z}^{(j)}$, $\boldsymbol{z}_{\text{mis}}^{(j)}$, $\boldsymbol{\mu}^{(j)}$, $\boldsymbol{\Sigma}^{(j)}$ based on the order of $\boldsymbol{\alpha}^{(j)}$
end
Result: Imputed database and parameters estimation

Chapter 4

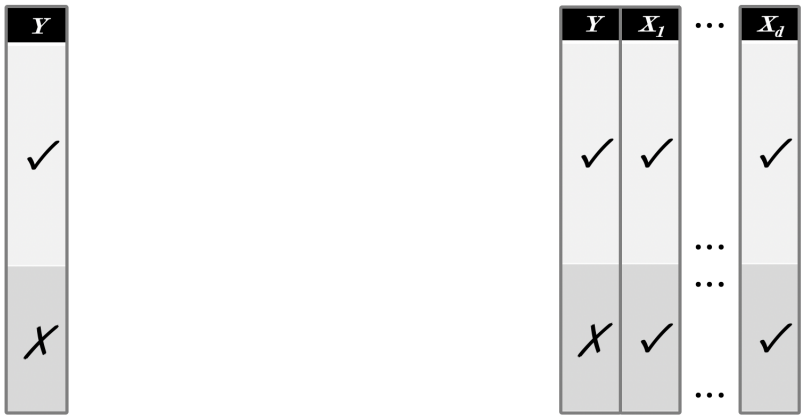
Univariate Gaussian LCWM

"People have forgotten this truth," the fox said.
"But you mustn't forget it.
You become responsible forever for what you've tamed.
You're responsible for your rose."

The Little Prince.

4.1 Overview

In this chapter we present a first implementation of the imputation procedure using the Gaussian LCWM in the simplest case. We will call it the univariate case. Since the objective is explained through the situation presented in the introduction to this document by means of Figure 1.1, we consider as a first approximation that case where $p = 1$ and which is illustrated in Figure 4.1. It is about considering a pattern of unit nonresponse that will be represented by the only output variable Y, which will be imputed making use of auxiliary information represented by the input variables $X_1, ..., X_d$. In other words, we start from the pattern of missing data in Figure 4.1a, where the imputation process does not make use of auxiliary information, and we seek some way to include auxiliary information from fully observed variables to impute units nonresponse, see Figure 4.1b.



(a) No auxiliary information.

(b) With auxiliary information.

FIGURE 4.1: Missing data pattern from univariate databases in the cases of not including and including auxiliary information.

Our proposal starts from the Gaussian FMM used in the process of imputation of the pattern of missing data in Figure 4.1a to consider including auxiliary variables with fully observed information, see Figure 4.1b. For this new pattern of missing

data, considering the new variables as observational and non-deterministic, we can use them adaptively in the imputation procedure. Thus, we propose the use of the Gaussian LCWM as the imputation model, which takes advantage of the flexibility of FMMs to model unknown distribution forms, as well as group structures in the data. We will analyze the performance of the imputation process by entering different types of variables into the model, and we will compare the results with other imputation methods.

Section 4.2 presents simulation studies to evaluate the performance of our model in two ways. The first one seeks to analyze the influence that the variables that enter the model have on the imputation process; the second compares our imputation model with other procedures. Section 4.3 presents two sets of real data on which missing data patterns are simulated to evaluate our imputation model, these are the Faithful database and the data from the Annual Manufacturing Survey of Colombia in 1994.

4.2 Simulation Studies

4.2.1 Model performance when new information is included

To carry out an analysis of the proposed imputation process, a data set was simulated from a mixture of three-dimensional normal distributions with two components. One of the variables was considered as an output variable, while the other two were considered as input variables. The database contains `n = 1000` observations of the form (x_1, x_2, y). The mixing probabilities are $\alpha_1 = 0.6$ and $\alpha_2 = 0.4$, the mean vectors $\boldsymbol{\mu}_1 = (1.0, 9.0, 7.0)$ and $\boldsymbol{\mu}_2 = (1.0, 3.0, 3.0)$, and the covariance matrices are:

$$\Sigma_1 = \begin{pmatrix} 1.00 & 0.50 & 0.50 \\ 0.50 & 1.00 & 0.50 \\ 0.50 & 0.50 & 1.00 \end{pmatrix} \quad \text{and} \quad \Sigma_2 = \begin{pmatrix} 1.00 & 0.50 & -0.50 \\ 0.50 & 1.00 & -0.50 \\ -0.50 & -0.50 & 1.00 \end{pmatrix}.$$

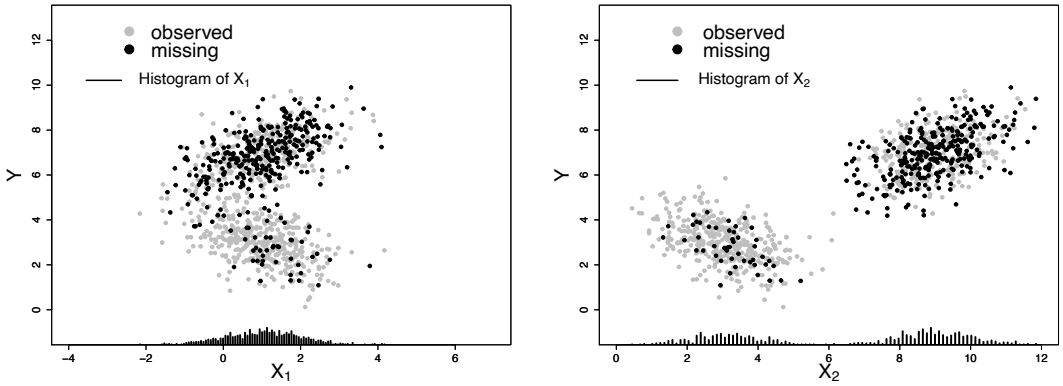
Initially, for cluster 1, 50% of the data was randomly selected and considered missing, while 10% was selected for cluster 2. A summary of how the data was generated is presented in Table 4.1. Scatter plots of the observed and missing data are illustrated in Figure 4.2, with the projection of the plane $X_1 \times Y$ in Figure 4.2a and the projection of the plane $X_2 \times Y$ in Figure 4.2b. Since the probability of missing is the same within each of the components, considering the variable Y as the one with missing information and the variables X_1 and X_2 as fully observed, the missing data mechanism can be assumed as Missing at Random (MAR) (Van Buuren, 2018; Rubin, 1976).

	observed		**missing**		**complete**	
cluster 1	281	*(48.9%)*	294	*(51.1%)*	575	*(100%)*
	(42.7%)		*(86.0%)*		*(57.5%)*	
cluster 2	377	*(88.7%)*	48	*(11.3%)*	425	*(100%)*
	(57.3%)		*(14.0%)*		*(42.5%)*	
total	658	*(68.3%)*	342	*(31.7%)*	1000	*(100%)*
	(100%)		*(100%)*		*(100%)*	

TABLE 4.1: Distribution of observed, missing and complete data by cluster for the simulated database in the univariate case.

Two scenarios are analyzed from the simulated data. In both, the variable Y will be treated as the output variable and the one with missing information. The variables X_1 and X_2 will be considered as fully observed input variables and provide auxiliary information for the imputation processes. In the first scenario (*Scenario 1*), Figure 4.2a, the model imputes the variable Y with the information from the variable X_1. The information provided by X_1 does not allow to conclude with which of the two components to impute. For the second scenario (*Scenario 2*), the variable Y is imputed with information from the variable X_2. Figure 4.2b allows us to conclude that, knowing information on this variable, it is possible to decide correctly which component to impute from.

The histograms at the bottom of the two graphs in Figure 4.2, and which refer to the distributions of the input variables in each case, allow us to observe the difference in their behavior. In the case of the histogram on the right side, where two totally separate groups are observed for the distribution of X_2, it will indicate that the input variable is distributed *separately among components*, this is an ideal behavior for the imputation process.



(a) Projection $X_1 \times Y$.

(b) Projection $X_2 \times Y$.

FIGURE 4.2: Scatter plots for observed and missing data for the simulated database.

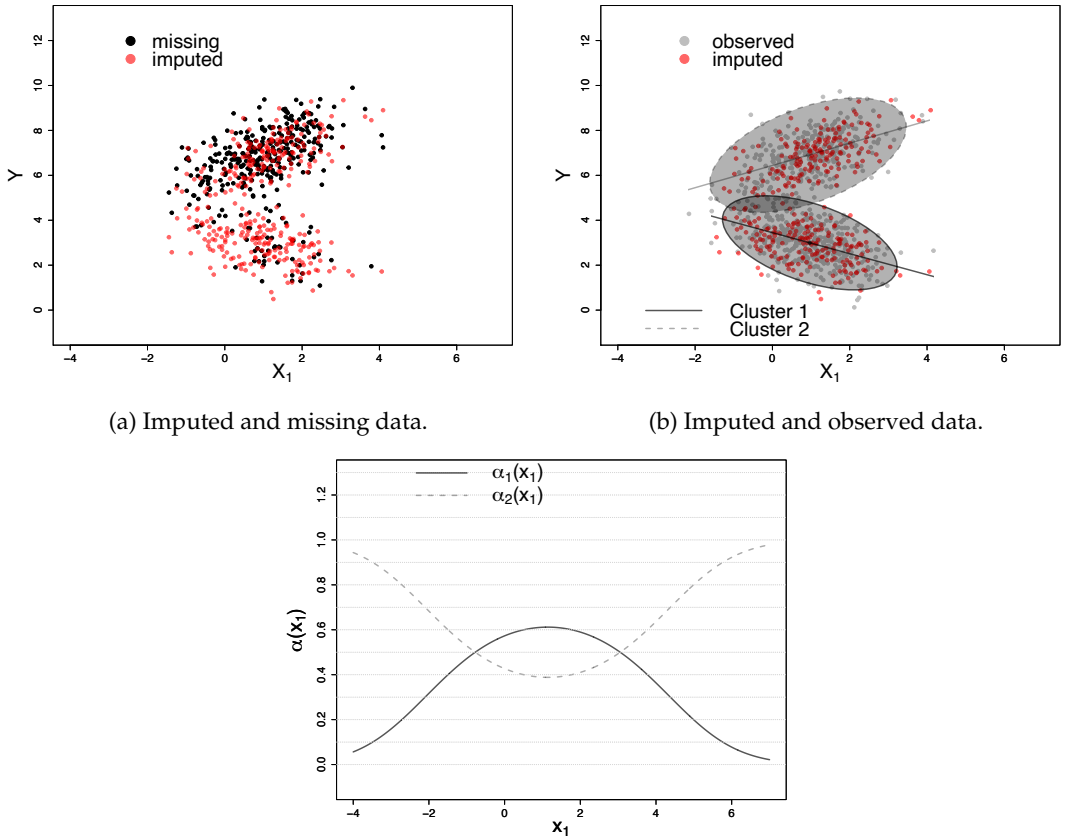
Analysis of the imputation process in the simulated scenarios

In all simulated scenarios, the imputation program was run on R software maintaining similar conditions with `burn-in=10000` and a sample size adjusted for autocorrelation, `effectiveSize=1500`, implemented using `coda` package (Plummer et al., 2006). The number of components was established at the fixed value of `G=10`. In all cases, only the first two components were occupied in the fitted models. The trace plots performed well, guaranteeing the convergence of the chains. To summarize the imputation process, following the idea of Fraley and Raftery (2007), the iteration that maximizes the density *a posteriori* (MAP) is chosen. Thus, all the graphs, the estimates, and the general descriptions provided here are based on the results obtained with this iteration. It should be noted that although we have analyzed our methodology using a single imputed database, the suggestion is to use this imputation procedure that we propose following the MI guidelines. For the graphs corresponding to the posterior probabilities that the observation $\boldsymbol{x}$ belongs to the component Z,

the notation $\alpha_Z(\boldsymbol{x}) = p(Z|\boldsymbol{x})$ is used, following the idea of the notation for mixing probabilities, $\alpha_Z = p(Z)$.

Scenario 1: A variable with low performance in the model

In the first scenario, the imputation model is implemented, seeking to complete the missing information for the output variable Y, using the input variable X_1 as auxiliary information. Since the input variable does not provide information about which component an observation belongs to, special interest is in the behavior of the estimates of the mixing probabilities. For cluster 1, $\widehat{\alpha}_1 = 0.418$, while for cluster 2, $\widehat{\alpha}_2 = 0.582$. These estimates are strongly influenced by the proportion of data observed in each cluster, and determine the proportion of data imputed in each group, (40.9% of the data were imputed in cluster 1, while 59.1% were imputed in cluster 2). Likewise, it can be observed how these proportions of the imputed values per component are considerably different from the proportions of missing data (86.0% for cluster 1 and 14.0% for cluster 2), as seen in Table 4.1.



(a) Imputed and missing data.

(b) Imputed and observed data.

(c) Posterior probability dependent on $\boldsymbol{x}_1$ for each cluster.

FIGURE 4.3: Construction of the imputation model in the case of auxiliary information given by the variable X_1.

Figure 4.3 shows how the imputation model was built and how the missing data was imputed. Figure 4.3a presents a scatter diagram where the imputed data is plotted, as well as the data considered missing. Although the data is imputed around the centers of each component, the proportion in each differs from the proportion

in which they were generated. Figure 4.3b illustrates how the imputation model was built, and presents the observed data that are the basis for the construction of the regression lines. Together with these points, quantile ellipses of 95% allow each of the components to be distinguished; this same graph shows the imputed values around the line. In particular, the regression lines are the result of the so-called conditional distributions, and are used as predictive models in the imputation process. The marginal distributions are used for the classification process and are the basis for the construction of the curves in Figure 4.3c. The graphs of the posterior probabilities dependent on the input variable $\boldsymbol{x}_1$ and defined by the expression (3.9) are shown. For values of $\boldsymbol{x}_1$ close to the estimates of the means in the two components ($\hat{\mu}_1^{(\boldsymbol{x}_1)} = 0.897$, $\hat{\mu}_2^{(\boldsymbol{x}_1)} = 1.109$), the posterior probabilities are strongly influenced by the proportion of observed data.

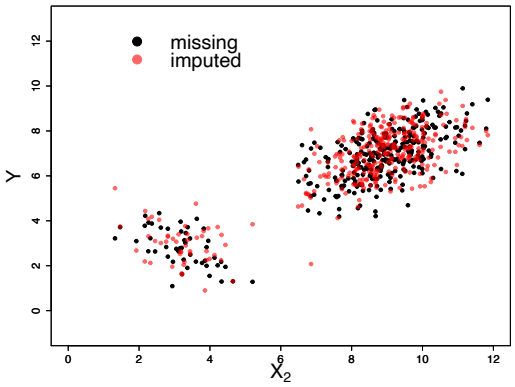
Scenario 2: A variable with high performance in the model

In the second case, where the input variable X_2 is distributed separated by components, after the imputation process, the model returns proportions of imputed data similar to how the missing data was generated. The estimates for the mixing probabilities are $\hat{\alpha}_1 = 0.583$ and $\hat{\alpha}_2 = 0.417$. The fact that the input variable completely separates the components allows the information provided by it to precisely determine the component which to impute from. The proportion of imputed data in each group was 14.0% for component 1 and 86.0% for component 2, the same values presented in Table 4.1 for missing data.

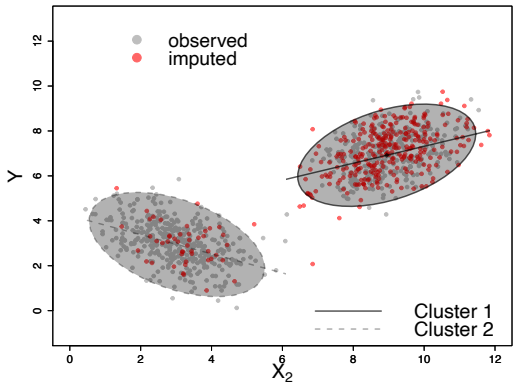
Figure 4.4 shows, for scenario 2, how the data was imputed. For example, Figure 4.4a shows the data that was considered missing and the imputations made by the model. It is evidenced that, in addition to the imputations being made close to the missing data, the proportions in the two components corresponding to missing data and imputed data are the same. Similar to Scenario 1, Figure 4.4b presents how the model was built based on the observed data set and illustrates the distribution of the imputed data. Specifically, the conditional and marginal distributions are responsible for carrying out the imputation and classification processes. The result of the conditional distributions are the regression lines, these are plotted for each component and show, together with the imputed data, the pattern followed to carry out the imputation. We show the 95% confidence ellipses for each component. As product of the marginal distributions together with the mixing probabilities, the posterior probability curves are illustrated in the graph of Figure 4.4c. This graph reflects an ideal behavior regarding the classification process. Punctually, it makes the correct decision regarding which component to impute from, given the value of the input variable. For values of the input variable less than six, the model imputes with probability one in component 1. Likewise, from some value greater than six, the observation is classified in component 2 and imputed with the estimated model for this case. For the values of the input variable around six, there is a transition in the probabilities, in such a way that in the limit, the probabilities to classify in one or another component coincide with the value 0.5.

The scenarios presented attempt to illustrate two extreme cases. The first case, which shows that the input variable does not provide any information to select the component which to impute from, such that the classification remains in the hands of estimating the proportion of data observed in each component, that is, of the mixture probability α_Z as can be deduced from Corollary 3.3.1. The second case is an

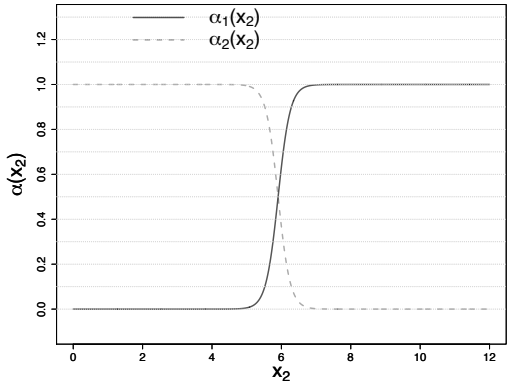
ideal scenario for the implementation of the model, where the information provided by the input variable allows us to determine, in a correct way, with which component to impute from.



(a) Imputed and missing data.



(b) Imputed and observed data.



(c) Posterior probability dependent on x_2 for each cluster.

FIGURE 4.4: Construction of the imputation model in the case of auxiliary information given by the variable X_2.

An additional scenario is presented without discussion in this section. It considers the imputation process under the joint information of the two input variables treated in the previous scenarios. In this situation, the two variables are integrated into an input vector of the form $\boldsymbol{X} = (X_1, X_2)$. This vector allows to consider its distribution as separate between components like the case of Scenario 2, and the results can be consulted in Appendix A. This case is very similar to the case considered in the second scenario.

Results of the imputation processes

Figure 4.5 presents box plots for the output variables in the cases of its complete information ($\boldsymbol{Y}_{\text{com}}$), when only the observed information is considered ($\boldsymbol{Y}_{\text{obs}}$) and after the imputation processes ($\boldsymbol{Y}_{(\cdot)}$). Within each boxplot, the point inside represents the value of the sample mean for each data set.

In the case of the imputation process using the variable X_1 (Scenario 1), since the variable does not provide any information about which component to impute from, the model selects said component based on the estimated mixing probabilities. That

is the reason for the similarity between the distribution of $\boldsymbol{Y}_{\text{obs}}$ and the imputed output variable $\boldsymbol{Y}_{X_1}$. When we use the X_2 variable as an input variable in the model (Scenario 2), since the information it provides allows us to identify the component to which an observation belongs, with the imputed output variable $\boldsymbol{Y}_{X_2}$ we have a closer proximity to the distribution of the original complete variable $\boldsymbol{Y}_{\text{com}}$. Something similar happens with the output variable imputed from the information of the vector (X_1, X_2), denoted as $\boldsymbol{Y}_{(X_1,X_2)}$.

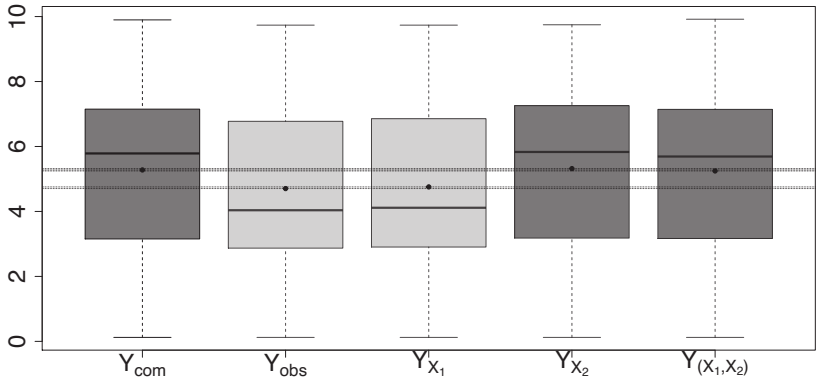


FIGURE 4.5: Box plots for complete, observed, and the imputed variables with information from the input variables X_1, X_2, and (X_1, X_2).

Diagnosis of the imputation process: Kullback-Leibler divergence

Due to the need for a quantitative diagnosis of the imputation process, we used the Kullback-Liebler divergence, a non-symmetric measure of the difference between two probability functions (Kullback and Leibler, 1951). For two density functions $f(\cdot)$ and $g(\cdot)$, in the one-dimensional continuous case, the Kullback-Liebler divergence is defined by the integral

$$\text{KL}(f, g) := \int_{-\infty}^{\infty} f(x) \log \frac{f(x)}{g(x)} dx. \tag{4.1}$$

The divergence $\text{KL}(f, g)$ can be interpreted as the amount of information lost when we want to approximate the f distribution using the g distribution. Unfortunately, for the case where $f(\cdot)$ and $g(\cdot)$ are Gaussian FMM, the expression in (4.1) is intractable. Hershey and Olsen (2007) and Durrieu, Thiran, and Kelly (2012) present several approximations, as well as bounds for the divergence in this case.

To implement the calculation of the KL divergence, the `integrate` function is used in the R software, which allows to approximate integrals of one-dimensional functions over infinite intervals. In this case, it will be used to approximate the integral in (4.1) and will be denoted as KL_{int}. To measure the quality of the imputation process, we calculated a 95% quantile interval based on the KL divergences calculated from `N=10000` replicates of the complete data set, obtained randomly from the original distribution. For this simulation process we use the `mixsmsn` package (Prates, R., and Lachos, 2013). Any value of KL of a variable that is within the interval will allow to conclude that said variable recovers the original distribution. The interval obtained and the KL divergences for the different imputation processes are

presented in Table 4.2. The table also presents the relative distances to the extreme right of the interval, which gives an idea of how far the distribution of interest is from the original distribution. If the value of the KL divergence falls within the interval, it is noted with `WI` (*withing the interval*). In the expression (4.1) the function f will refer to the original distribution, i.e., the one with which the database was generated, while the function g will refer to the distribution estimated from the data set $Y_{(\cdot)}$, we will use the notation $g_{Y_{(\cdot)}}$. To give us an idea of how the imputation model performs when the real distribution is not known, specifically in the case of the examples with real data, Appendix B presents tables of KL divergence values calculated having the estimated distribution based on complete data as reference distribution, that is, it assumes the role of the function f in expression (4.1).

	Approach method	
	KL_{int}	Relative distance
Qu.int. 95.0%	(0, 0.0056)	-
$g_{\boldsymbol{Y}_{com}}$	0.0029	`WI`
$g_{\boldsymbol{Y}_{obs}}$	0.0679	12.25
$g_{\boldsymbol{Y}_{X_1}}$	0.0665	12.00
$g_{\boldsymbol{Y}_{X_2}}$	0.0034	`WI`
$g_{\boldsymbol{Y}_{(X_1,X_2)}}$	0.0036	`WI`

TABLE 4.2: KL divergences and relative distances for the imputed variables with information from the input variables X_1, X_2, and (X_1, X_2).

For $\boldsymbol{Y}_{com}$, it is observed that the KL value is within the quantile interval of 95%. For the distribution of the observed data, $\boldsymbol{Y}_{obs}$, a distance of about twelve units from the original distribution is observed. The variables $\boldsymbol{Y}_{X_2}$ and $\boldsymbol{Y}_{(X_1,X_2)}$ have the best behaviors in terms of proximity to the target distribution, they recover the original distribution. In the case of the imputation process with the vector (X_1, X_2), including in it a variable that is distributed separately among components, as is the case of the input variable X_2, allows obtaining a vector that is distributed separately among components[1].

The same estimates obtained from the imputed variables and used to calculate the KL divergences are used to obtain the graphs of the estimated densities shown in Figure 4.6. A histogram is presented illustrating the distribution of the complete variable, $\boldsymbol{Y}_{com}$. The solid black and gray curves represent the estimated densities for $\boldsymbol{Y}_{com}$ and $\boldsymbol{Y}_{obs}$, respectively. Once again, the good performance of the imputation process can be confirmed with information from the input variable X_2 and the input vector (X_1, X_2). The blue dotted curve and the green dashdot curve corresponding to these two estimated densities, respectively, are closest to the estimated density of the variable $\boldsymbol{Y}_{com}$. This is not the case with the red dashed curve and that corresponds to the variable $\boldsymbol{Y}_{X_1}$.

In conclusion, the presented imputation methodology makes use of the information provided by the input variables. This information can move between two situations: the input variable does not provide any information about which component

[1] A similar performance, when the imputation processes are compared using information from the different auxiliary variables, can be concluded from Table B.1 in Appendix B.

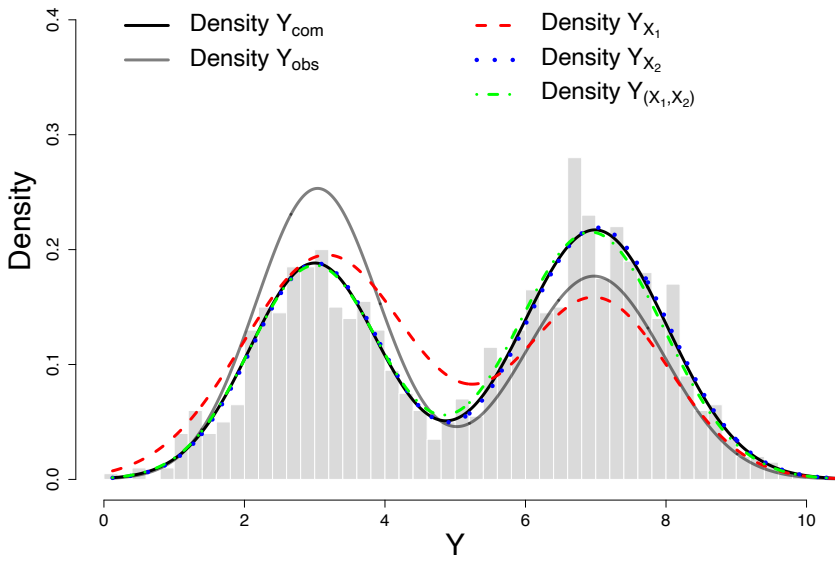


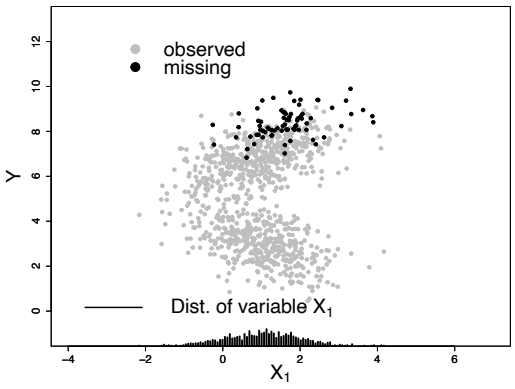
FIGURE 4.6: Histogram of the variable $\boldsymbol{Y}_{\text{com}}$ and estimated densities for the imputed variables with information from the input variables X_1, X_2, and (X_1, X_2).

the observation belongs to; or the input variable is capable of correctly determining the component to which said observation belongs. Information from several input variables can be taken into account together in the classification process. Therefore, to provide an adequate imputation, it is necessary to have inputs that accurately separate the imputation region. If the user does not know which input that is, we suggest the inclusion of an input vector. As shown by our simulations, the performance of the imputation method is not affected by the non-informative inputs, thus continuing to offer an appropriate imputation for the data.

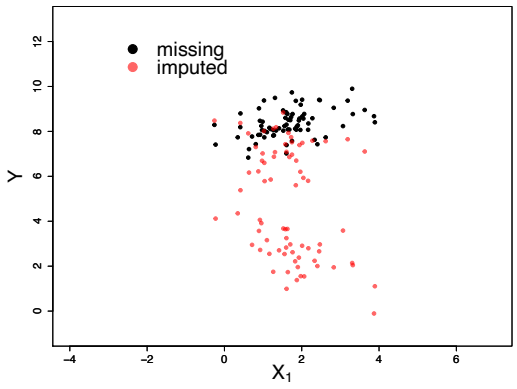
4.2.2 A scenario with missing data from a MNAR mechanism

On the same data set studied in this section, an MNAR mechanism of missing data was simulated to observe how the imputation model behaves. For this case, 20% of data with values of Y greater than 6.5 were randomly selected and considered as missing. This procedure generated 75 missing data points, all belonging to cluster 1. The distribution of the observed and missing data projected on the planes $X_1 \times Y$ and $X_2 \times Y$ can be seen in Figures 4.7a and 4.8a. The missing data is grouped at the top of the point cloud corresponding to cluster 1.

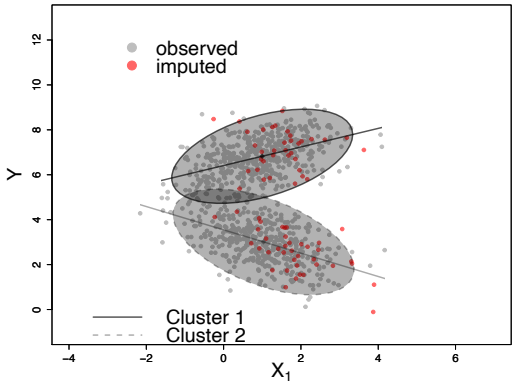
Similar to Scenario 1 presented above, the imputation model was implemented using the input variable X_1 as auxiliary information. Since this variable does not offer any information on which component to impute from, the classification procedure is carried out based on the estimates made with the mixing probabilities, $\hat{\alpha}_1 = 0.526$ and $\hat{\alpha}_2 = 0.474$. Thus, 39 observations are imputed for component 1, while 36 observations are imputed for component 2, as it is seen in Figure 4.7b. Furthermore, using Figure 4.7c as a complement, it can be observed that the imputed data in component 1 tries to cover the width of the cluster conditioned by the values of X_1. Implementing the model with information from the input variable X_2 leads



(a) Distribution of observed and missing data.



(b) Imputed and missing data.



(c) Imputation model with information on variable X_1.

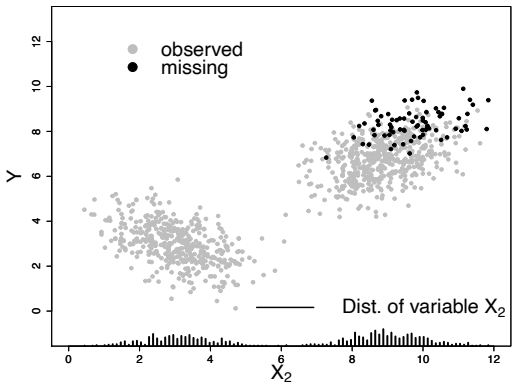
FIGURE 4.7: Simulation of an MNAR mechanism and imputation using the variable X_1 in the univariate case .

to estimates of the mixing probabilities equal to $\hat{\alpha}_1 = 0.565$ and $\hat{\alpha}_2 = 0.435$. However, since the input variable X_2 provides precise information on which component to impute from, all observations are imputed in component 1 as shown in Figure 4.8b. Of utmost importance, it should be noted that, although the model imputes fairly accurately in the correct component, the way it does so is far from how the missing data were generated within the component, see Figure 4.8c. It could be concluded that, although the missing data were generated from a MNAR mechanism, the model imputes within the component assuming MAR.

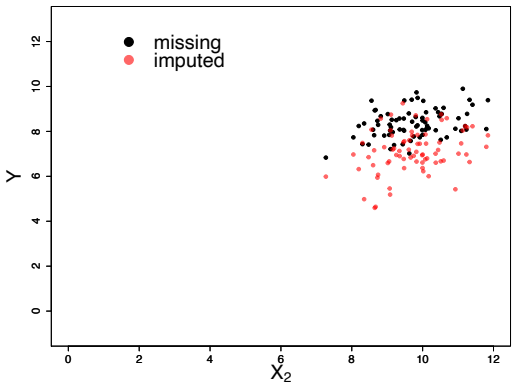
4.2.3 Gaussian LCWM performance relative to other imputation methods

To compare the proposed imputation model with other methods, the package repository was consulted to treat missing data with the R software[2]. `MICE` package (Van Buuren and Groothuis-Oudshoorn, 2011) is one of the commonly used package by R users. Two methods included in the `MICE` package were of special interest in this process, *predictive mean matching* (Little, 1988), implemented using the function `mice.impute.pmm()`, and *Bayesian multiple imputation* (Rubin, 1987), implemented by the function `mice.impute.norm()`. Both methods were described in Chapter 2 in more detail, and the main interest is that they are two predictive methods and use a

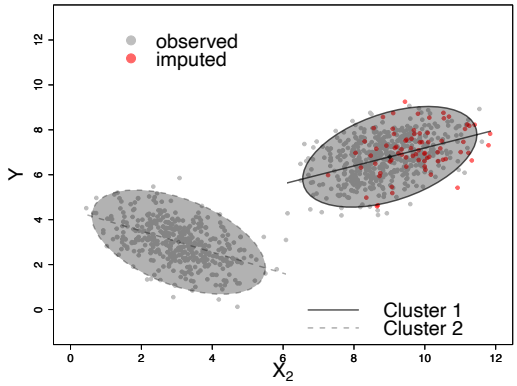
[2] `https://CRAN.R-project.org/view=MissingData`



(a) Distribution of observed and missing data.



(b) Imputed and missing data.



(c) Imputation model with information on variable X_2.

FIGURE 4.8: Simulation of an MNAR mechanism and imputation using the variable X_2 in the univariate case.

Bayesian approach.

We were able to observe that the Bayesian multiple imputation method implemented in `MICE` could be approximated by our model considering the value of $G = 1$ for the number of components. Theoretically, the difference in the two procedures lies in the priori distributions used in each case. To implement the method, we will use our R code and to refer to it we will use the `norm` notation. In the case of predictive mean matching, we will use the `MICE` package and to refer to it we will use the `pmm` notation. For comparison purposes, we also include as a method of interest the procedure implemented in Paiva and Reiter (2017) and that does not use auxiliary information; to refer to this methodology we will use the `mean` notation. Finally, we will refer to our model using the `cwm` notation.

The KL divergence values presented in Table 4.3 were obtained by implementing the imputation methods `mean`, `pmm`, and `norm` on the simulated data set in Section 4.2.1. From the results obtained, a first objective proposed in this work was to include, on the `mean` methodology that does not incorporate auxiliary information, fully observed additional variables that would improve the performance of the imputation process. In the first column of Table 4.3, we can see the results corresponding to our model together with a row on which the KL divergence is shown for the imputation procedure without auxiliary information, $g_{\boldsymbol{Y}_{\texttt{mean}}}$. Even from the

imputation process with the X_1 variable, we can see less loss of information with our methodology. With the variable X_2 and the vector (X_1, X_2), our model was able to recover the true distribution.

Observing the information in the complete Table 4.3, a better performance of the methods `cwm` and `pmm` compared to `norm` can be concluded. The procedure `pmm` has similar results to ours in the case of using variables that are distributed separately among components; in the case of using information from variable X_1, the loss of information is much greater than in the case of our model. Graphs that illustrate the imputation process using the `mean`, `pmm`, and `norm` methodologies can be consulted in Appendix C. Also, in Appendix B, Table B.2 is presented, which includes the values of the KL divergence, assuming that the true distribution is unknown and from which similar conclusions can be obtained.

	`cwm`		`pmm`		`norm`	
	KL_{int}	Relative distance	KL_{int}	Relative distance	KL_{int}	Relative distance
Qu.int. 95.0%	(0,0.0055)	-	(0,0.0055)	-	(0,0.0055)	-
$g_{\mathbf{Y}_{com}}$	0.0029	WI	0.0029	WI	0.0029	WI
$g_{\mathbf{Y}_{obs}}$	0.0679	12.25	0.0679	12.25	0.0679	12.25
$g_{\mathbf{Y}_{mean}}$	0.0854	15.42	-	-	-	-
$g_{\mathbf{Y}_{X_1}}$	0.0665	12.00	0.1130	20.39	0.2093	37.79
$g_{\mathbf{Y}_{X_2}}$	0.0034	WI	0.0022	WI	0.0193	3.48
$g_{\mathbf{Y}_{(X_1,X_2)}}$	0.0036	WI	0.0040	WI	0.0128	2.31

TABLE 4.3: Performance of the `mean`, `cwm`, `pmm` and `norm` methods by calculating the KL divergence for the first simulated data set.

A scenario with better performance of `cwm` versus `pmm`

In this section, a data set is simulated with a pattern of missing data where our methodology performs better than the `pmm` procedure. The database contains `n=1000` observations of the form (x, y). The mixing probabilities are $\alpha_1 = 0.6$ and $\alpha_2 = 0.4$, the mean vectors $\boldsymbol{\mu}_1 = (4.0, 10.0)$ and $\boldsymbol{\mu}_2 = (7.0, 4.0)$, and the covariance matrices are:

$$\Sigma_1 = \begin{pmatrix} 0.50 & 0.35 \\ 0.35 & 0.50 \end{pmatrix} \quad \text{and} \quad \Sigma_2 = \begin{pmatrix} 0.50 & -0.64 \\ -0.64 & 1.00 \end{pmatrix}.$$

The missing data pattern was generated in such a way that all observations with values of the variable X greater than 5 belonging to cluster 1 were considered as missing. A summary of how the data was generated is presented in Table 4.4.

	observed		**missing**		**complete**	
cluster 1	441	*(76.7%)*	134	*(23.3%)*	575	*(100%)*
	(50.9%)		*(100.0%)*		*(57.5%)*	
cluster 2	425	*(100.0%)*	0	*(0.0%)*	425	*(100%)*
	(49.1%)		*(0.0%)*		*(42.5%)*	
total	866	*(86.3%)*	134	*(13.4%)*	1000	*(100%)*
	(100%)		*(100%)*		*(100%)*	

TABLE 4.4: Distribution of observed, missing and complete data by cluster. A scenario with censored data.

The imputation process was implemented using the four methodologies of interest: `mean`, `cwm`, `pmm`, and `norm`. Scatter plots for observed, missing, and imputed data by the four methods are shown in Figure 4.9. A scenario with censored data in cluster 1 is presented. The panels in the figure illustrate the results of the different imputation processes when the four methods are implemented on the database.

The panel in the upper left of Figure 4.9 shows the results of the imputation process with the `mean` method. Since the methodology does not use information from the variable X, the procedure imputes in each cluster proportional to the number of data observed in each case, and imputes around the mean value of each group. For the `cwm` imputation procedure, we can see that the imputed values manage to cover a large part of the region where the missing data was generated, except for some imputations that occur in cluster 2. Regarding the `pmm` method, although some imputed values appear in the region of the missing data, they do not cover the region properly and a considerable number of aligned points are observed, which means that the method imputes with the same observed value many times. This situation occurs when there is little or no information observed in the specific region of imputation (Van Buuren, 2018). Additionally, a considerable amount of points were incorrectly imputed in cluster 2. The panel at the bottom right hand side shows the results of the imputation process with the `norm` method. The procedure erroneously imputes the vast majority of observations, specifically it does so in a region where there is no missingness.

From the analysis presented, it can be concluded that the best performance in terms of the imputation process was the `cwm` methodology. The `mean`, `pmm`, and `norm` methods perform unfavorably in this scenario. A quantitative evaluation of the imputation processes can be carried out from the KL divergence. Table 4.5 presents this measure for the different methods to be compared. The 95% quantile interval for the KL divergence of databases generated with the described specifications is shown in the first line. We assume that a KL divergence value that is within the interval allows us to conclude that the original distribution has been recovered, this will be noted in the relative distance column with `WI` (*within the interval*).

	Approach method	
	KL_{int}	Relative distance
Qu.int. 95.0%	(0, 0.0271)	-
$g_{\boldsymbol{Y}_{com}}$	0.0028	`WI`
$g_{\boldsymbol{Y}_{obs}}$	0.0611	2.25
$g_{\boldsymbol{Y}_{mean}}$	0.3100	11.44
$g_{\boldsymbol{Y}_{cwm}}$	0.0198	`WI`
$g_{\boldsymbol{Y}_{pmm}}$	0.0559	2.06
$g_{\boldsymbol{Y}_{norm}}$	0.2174	8.02

TABLE 4.5: KL divergences and relative distances for the imputation methods in the case of censored missing data.

The values in Table 4.5 allow us to confirm the analysis previously performed. We see that the imputation carried out with `cwm` allows us to recover the original distribution, from here we can conclude its better performance over the other methods. The procedure `mean` shows the worst performance followed by the procedure `norm` among the compared methods. Once again, despite the fact that the method `pmm` has

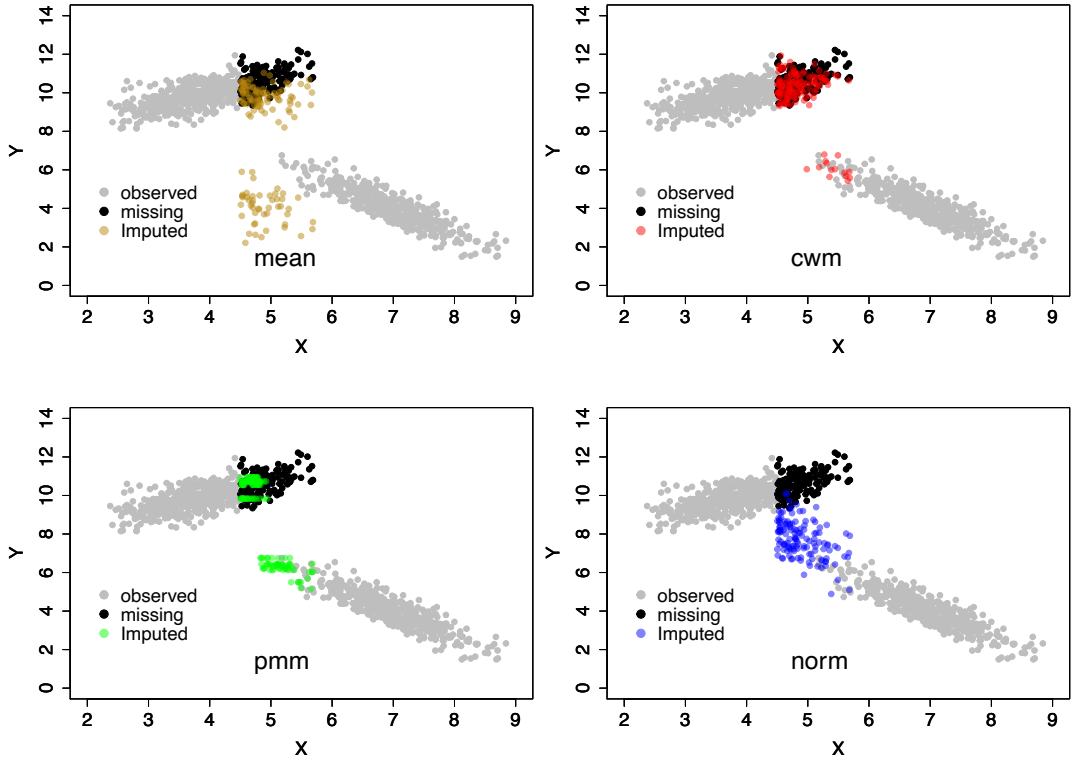


FIGURE 4.9: Scatter plots for the `mean`, `cwm`, `pmm`, and `norm` methods in the case of censored missing data.

a poor performance compared to ours, compared to the procedures `mean` and `norm` it has a lower loss of information when the objective is to recover the original distribution. Similar conclusions are obtained when considering the true distribution as unknown and taking as a reference the distribution of complete data in the calculation of the KL divergence (see Table B.3 in Appendix B).

As a conclusion, an extreme scenario was presented where the pattern of missing data was generated considering censored data and where the distribution of the variables is separated between components. The two clusters were generated from opposite correlations, one group with positive correlation and the other negative. In this situation, the `mean`, `norm`, and `pmm` imputation methods show disadvantages compared to our methodology. Building good imputation models requires analytical skills, and since there is no foolproof method, it is important to do a thorough analysis of the strengths and weaknesses of such models.

4.3 Illustrative Examples with Real Data

4.3.1 Data Set: Faithful

The proposed methodology will be implemented on the database called Faithful, it is a classic data set on geyser eruptions (Härdle et al., 1991). The base contains the waiting times between eruptions and the durations of the Old Faithful geyser eruptions in Yellowstone National Park, Wyoming, United States. In the database, each row represents an observed eruption of the Old Faithful Geyser. The `Faithful` data set

is found in the `datasets` R package, and consists of 272 observations on 2 variables, `eruptions` (represents the duration of the eruption in minutes) and `waiting` (represents the duration in minutes until the next eruption). The use of $G = 2$ components to fit a Gaussian mixture model is reasonable and will be considered here (see, e.g., Benaglia et al., 2009; Prates, R., and Lachos, 2013). Figure 4.10 shows in the left-hand panel a scatter diagram of the data where the `waiting` variable will be assumed as the input variable, while `eruptions` as the output variable. The right-hand graph shows a cluster classification obtained using the Gaussian LCWM implemented by us.

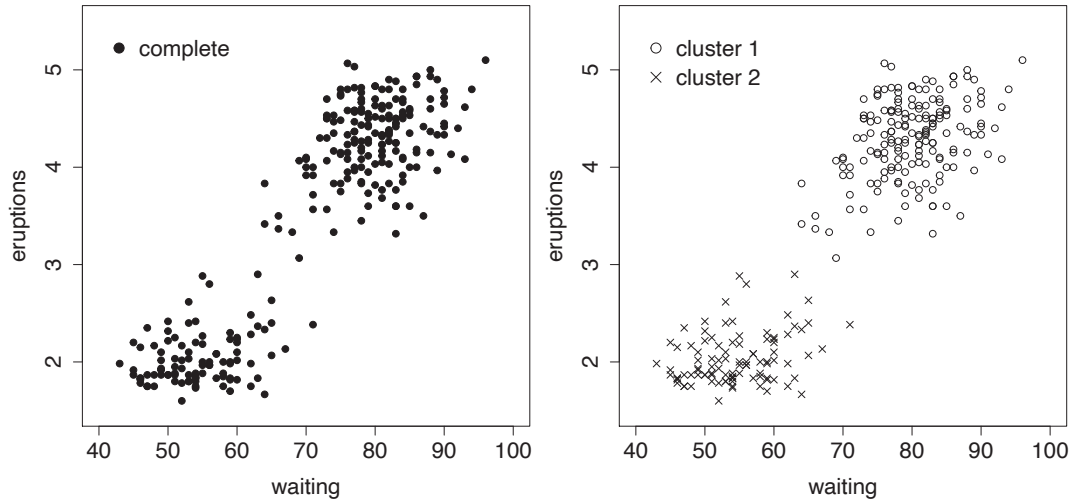


FIGURE 4.10: Scatter plot of the `Faithful` database together with classification in two clusters.

Similar to how missing data patterns were generated in previous cases, the expression in 4.2 was used to simulate the missing data set considering an MNAR mechanism. Values of $\beta_0 = -4.23$ and $\beta_1 = 1.02$ were used allowing larger values of the `eruptions` variable to have a greater probability of being missing. This scheme generated 95 missing data points, out of the 272 that the database has. Figure 4.11 describes the missing data structure through scatter plots, and it also presents the probability graph to generate the missing data. The largest proportion of missing data is concentrated in cluster 1 located in the upper right part of the graph. The observations with the highest values of the variables are located in this region.

We proceed to impute the `Faithful` database, specifically the `eruptions` variable using the information from the fully observed `waiting` variable. Our model is compared to the `mean`, `pmm`, and `norm` procedures. Figure 4.12 presents scatter plots with the observed, missing, and imputed data for the four procedures. We can see that our model better covers the region of missing data. Some observations imputed with `pmm` are far from the missing data, while a considerable amount of data imputed with `norm` is imputed in the middle of the two components, in a region where there is no missing data nor observed. Graphically, we can see that of the four procedures presented, the one corresponding to `cwm` shows better performance, closely followed by `pmm`. Although visually the scatter diagram for the `mean` method presents the worst behavior, we must remember that this procedure does not use the information from the variable `waiting`, however it recognizes the existence of the two components

and imputes, although in wrong proportion, in each one of them.

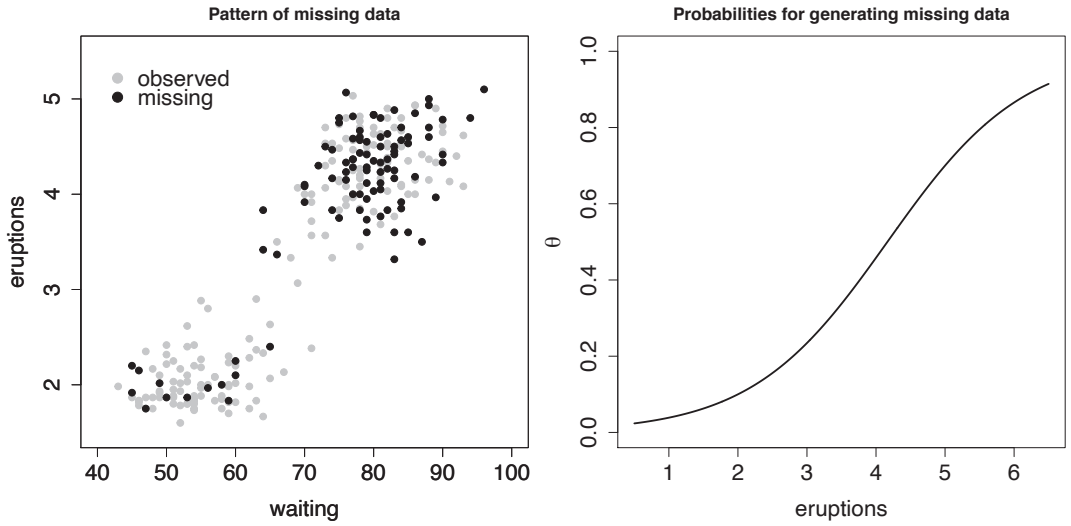


FIGURE 4.11: Construction of the missing data pattern for the `Faithful` data base.

	Approach method	
	KL_{int}	Relative distance
$\texttt{eruptions}_{\texttt{com}}$	-	-
$\texttt{eruptions}_{\texttt{obs}}$	0.0297	1.00
$\texttt{eruptions}_{\texttt{mean}}$	0.0429	1.44
$\texttt{eruptions}_{\texttt{cwm}}$	0.0018	0.06
$\texttt{eruptions}_{\texttt{pmm}}$	0.0070	0.24
$\texttt{eruptions}_{\texttt{norm}}$	0.0518	1.74

TABLE 4.6: KL divergences in relation to the complete data distribution and its relative distances for the `Fainthful` dataset.

The graphical analysis carried out previously can be quantitatively corroborated using the KL divergence. Table 4.6 presents the KL divergence values with respect to the complete-data distribution. This table includes the divergence value for the procedure `mean` that does not use auxiliary information as a reference. The relative distance is taken based on the KL divergence for the variable `eruptions` with only observed data. A value less than one allows us to conclude that the distribution of the imputed variable loses less information than that with only observed data, in such a way that we can conclude the process has a good performance.

The values in Table 4.6 allow us to conclude that the procedures `mean` and `norm` had the worst performances, while the procedures `cwm` and `pmm`, having values lower than one, show the best results. Among these two, our model is the one that presents the lowest KL divergence value, which means that out of the compared methods, the one that loses less information when trying to recover the distribution of the variable with complete data is ours.

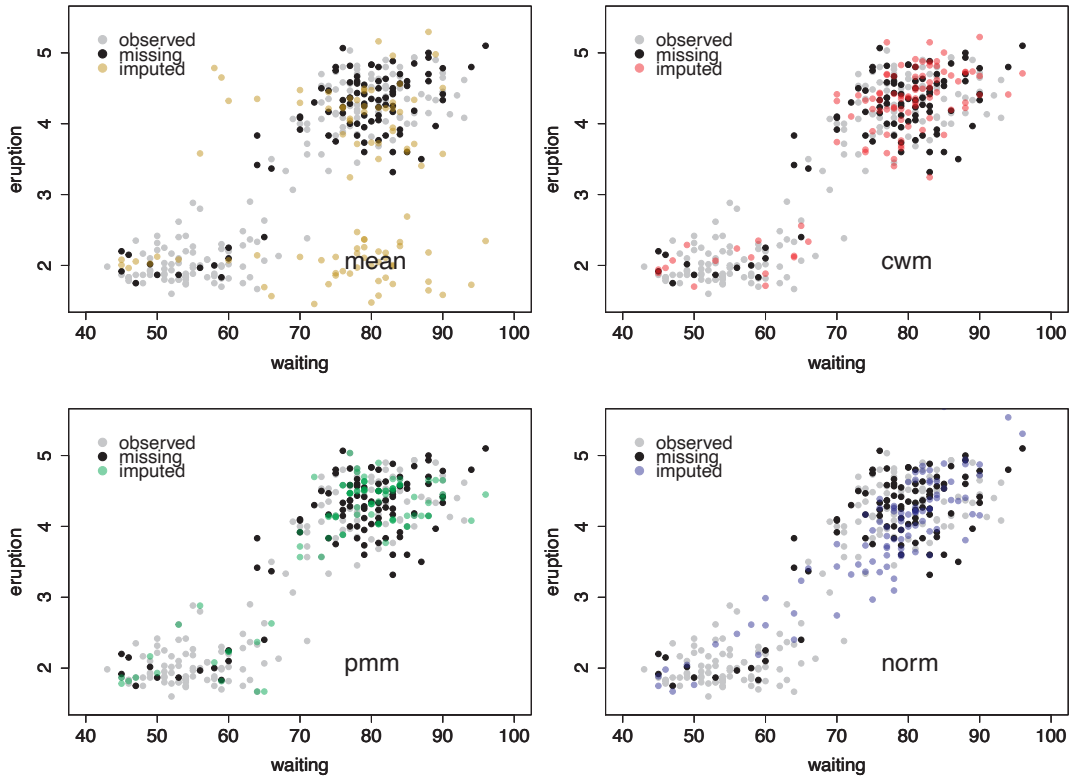


FIGURE 4.12: `Fainthful` dataset imputed using `mean`, `cwm`, `pmm`, and `norm` methods.

4.3.2 Data Set: Annual Manufacturing Survey from Colombia

The imputation process is illustrated with data from the Annual Manufacturing Survey (EAM)[3] from Colombia in 1994. The database contains 580 variables on 7488 companies. To illustrate the application in the univariate case, we selected the variables *Fixed Assets* (AF), which correspond to the assets of the establishment for the development of its industrial activity, and *Employed Personnel Expenses* (GPO), which refers to the sum of the wages and salaries of the personnel hired directly by the establishment. These variables will be considered as output and input variables, respectively. Since the missing data will be simulated, the companies that originally provided incomplete information are removed from the database. This process leaves a base with 7419 companies, that is, less than 1% of companies were eliminated. Due to the right-skewed pattern presented in the distribution of the variables, their values are log-transformed and standardized.

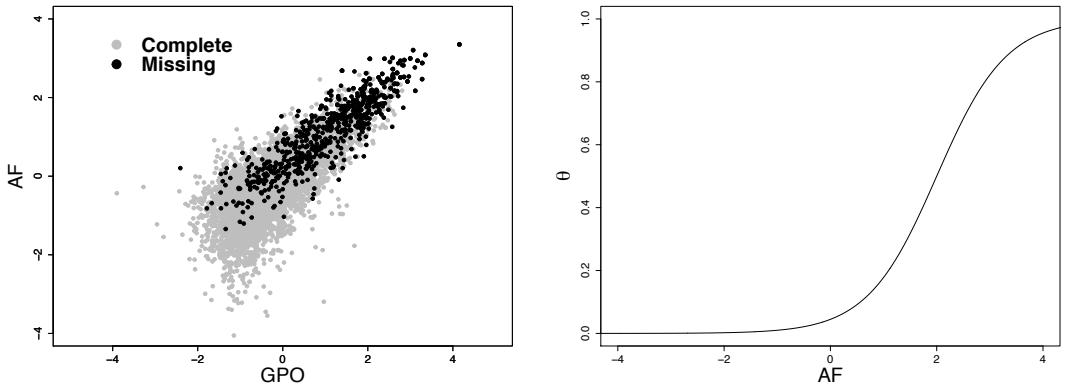
To generate missing data assuming a MNAR mechanism, an indicator variable $R_i \sim \text{Bern}(\theta_i)$ is simulated with missingness probability θ_i for the i-th observation. This probability is obtained from the expression

$$\theta_i = \text{logit}^{-1}(\beta_0 + \beta_1 y_i) \qquad \text{for } i = 1, ..., n. \tag{4.2}$$

The value of θ_i is specified in such a way that individuals with the highest values of the output variable y_i have higher probabilities of not responding. Choosing $\beta_0 = -3.06$ and $\beta_1 = 1.53$ provides a pattern where values of y close to -2 have a

[3] https://www.dane.gov.co/

probability to be missing around 0.002, while values of y close to 2 have a probability to be missing around 0.5. This schema generated 729 missing data points for the output variable AF. Figure 4.13a illustrates how the missing data is distributed within the complete data scatter plot, while Figure 4.13b shows the curve that generates the pattern of missing data given by the expression in (4.2).

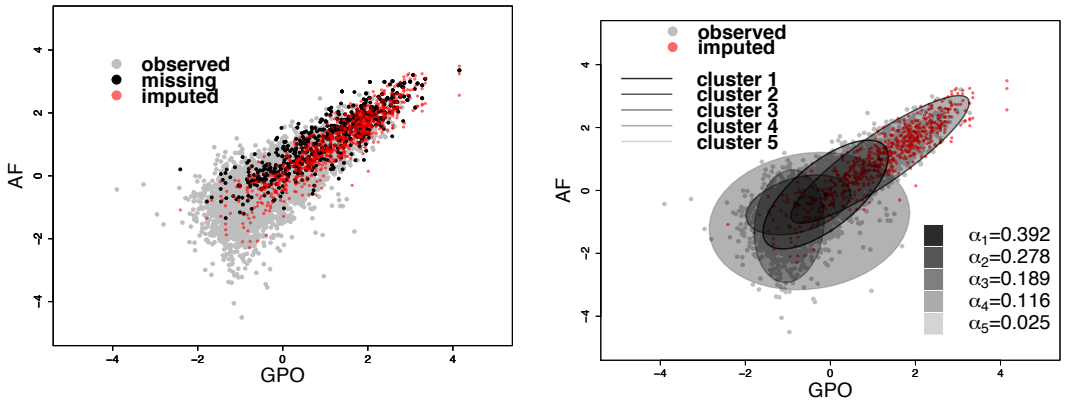


(a) Scatter plot for the GPO and AF variables for the complete and missing data.

(b) Graph of the probability function used to generate missing data.

FIGURE 4.13: Construction of the missing data pattern for the EAM dataset.

To specify the number of components G to use within the imputation process `cmw`, we follow the specifications of Kim et al. (2015). They propose to consider large values for G and, at each iteration of the Gibbs sampler, count the number of components that include at least one observation. If this count reaches the value set for G, it is prudent to increase its value, that is, to readjust the number of components G. When the count of occupied components is less than G, that is, some components in each iteration remain empty, then the choice of G is reasonable. Considering the pair of variables from the EAM data, a value of $G = 20$ was fixed and, after monitoring the number of occupied components, we present an analysis for the first five clusters.



(a) Scatter plot of observed, missing and imputed data points.

(b) Description of the imputation model.

FIGURE 4.14: Imputation model for the Colombian EAM dataset, based on the MAP iteration.

Similar to the simulation studies in Section 4.2, to summarize the classification process and the MCMC results, we selected the iteration that maximizes the density *a posteriori* (MAP), following the idea of Fraley and Raftery (2007). Figure 4.14a shows how the model imputes the missing data set; the imputed data set (red dots) attempts to cover the region where the missing data (black dots) are located. Figure 4.14b shows, using 95% quantile ellipses, how the first five components were constituted for the imputation process of the missing data. Table 4.7 presents the clusters' centers and the mixture weights for each of the first five components, which together cover 99.9% of the total weight. It should be noted that the last component has the greatest variability and coincides with the the smallest mixture probability.

cluster	μ_{GPO}	μ_{AF}	α
1	-0.1130	-0.1307	0.3922
2	1.1661	0.9876	0.2777
3	-0.7558	-0.4606	0.1892
4	-0.9695	-1.1355	0.1158
5	-0.4937	-0.9735	0.0250

TABLE 4.7: Centers and mixing weights by cluster of the imputation model for the EAM dataset.

Appendix D presents graphs that illustrate the construction of the imputation model. It discriminates by component both the 95% quantile ellipse and the regression line, each as a result of the marginal and conditional models, respectively. A graph is also presented with the posterior probabilities dependent on the output variable GPO.

Since our objective is to establish a methodology that allows the inclusion of auxiliary information to the imputation model specified in Paiva and Reiter (2017), we will also use an imputed variable that makes use of the methodology proposed by them. This variable is denoted with $\text{AF}_{\texttt{mean}}$, and will serve as the basis for the diagnosis of the imputation procedure that we present.

The boxplots of Figure 4.15 allow us to compare the distributions of the output variable AF, imputed under different procedures. These correspond to the variable with complete data (AF_{com}), the one with observed data (AF_{obs}), the one imputed with the `mean` procedure ($\text{AF}_{\texttt{mean}}$), and the one imputed with information from the GPO variable using the `cwm` procedure ($\text{AF}_{\texttt{cwm}}$).

This first analysis of Figure 4.15 allows us to observe a greater proximity in the distributions of the variable AF_{obs} and the imputed $\text{AF}_{\texttt{mean}}$, both in light gray. Similarity can also be observed between the distributions of the variables AF_{com} and $\text{AF}_{\texttt{cwm}}$ represented in dark gray. This proximity shows a better performance of the imputation process with auxiliary information compared to that without information, when the objective is to recover the distribution of the complete data, AF_{com}.

Figure 4.16 shows overlapping histograms for observed, missing, and imputed data for the variables $\text{AF}_{\texttt{mean}}$ and $\text{AF}_{\texttt{cwm}}$. The panel on the left side shows, for the variable $\text{AF}_{\texttt{mean}}$, that the distribution of the imputed data is strongly influenced by the distribution of the observed data and does not return the distribution of the missing data. For the case of the variable $\text{AF}_{\texttt{cwm}}$ on the rigth-hand side plot, the auxiliary

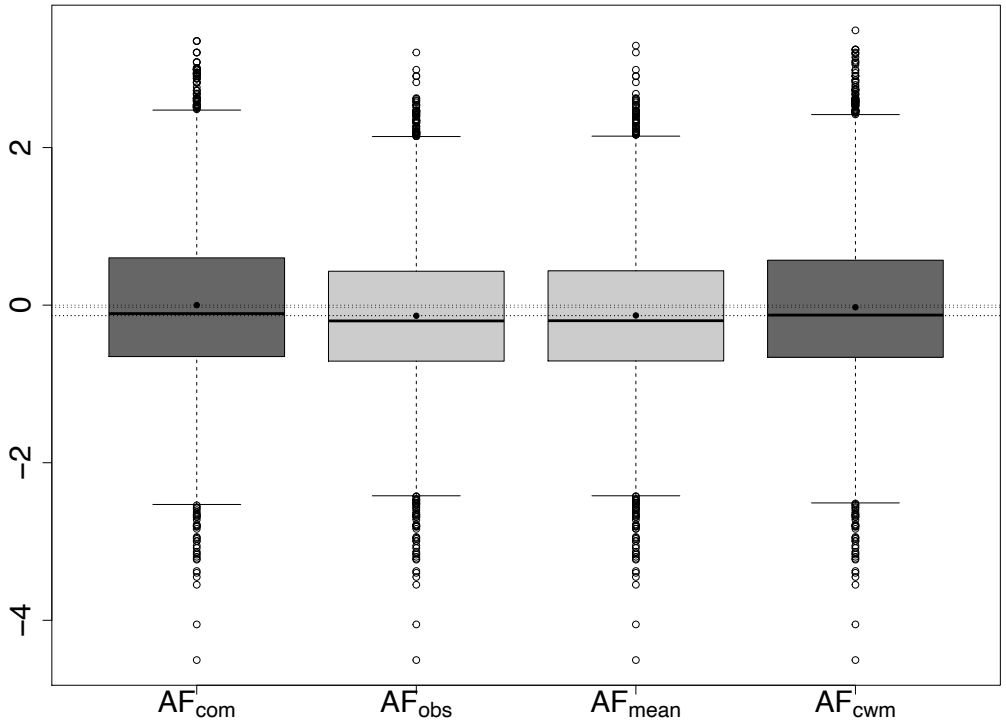


FIGURE 4.15: Boxplots for the cases of the variable AF with complete and observed data, as well as when it was imputed using the `mean` and `cwm` procedures.

information allows the imputed data distribution to recover the missing data distribution. Assuming that a good imputation procedure recovers the distribution of the missing data through the imputed data, the right-hand side histograms show better behavior for the imputed variable $AF_{\texttt{cwm}}$, compared to those corresponding to the variable $AF_{\texttt{mean}}$ on the left-hand side.

Diagnosis of the imputation process of the EAM database

In order to evaluate the imputation procedure, a first decision is to select the model that best fits the original data set. In our case, the choice of the Gaussian LCWM depends on an appropriate choice of the number G of components. Once the number of components has been established, with the imputed database, the Gaussian LCWM is fitted in each case and then evaluated.

Several criteria can be used for model selection among a finite set of models. Watanabe and Opper (2010) define the Watanabe-Akaike information criteria (WAIC). The WAIC is characterized by being a completely Bayesian procedure, in addition, compared to AIC and DIC, WAIC has the desirable property of averaging over the posterior distribution instead of conditioning it to a point estimate (Gelman, Hwang, and Vehtari, 2014). This will be the criterion that we will take into account here.

The model estimation procedure was executed for values of $G = 2, ..., 10$. For each of the estimates obtained, the WAIC was calculated. The results are presented

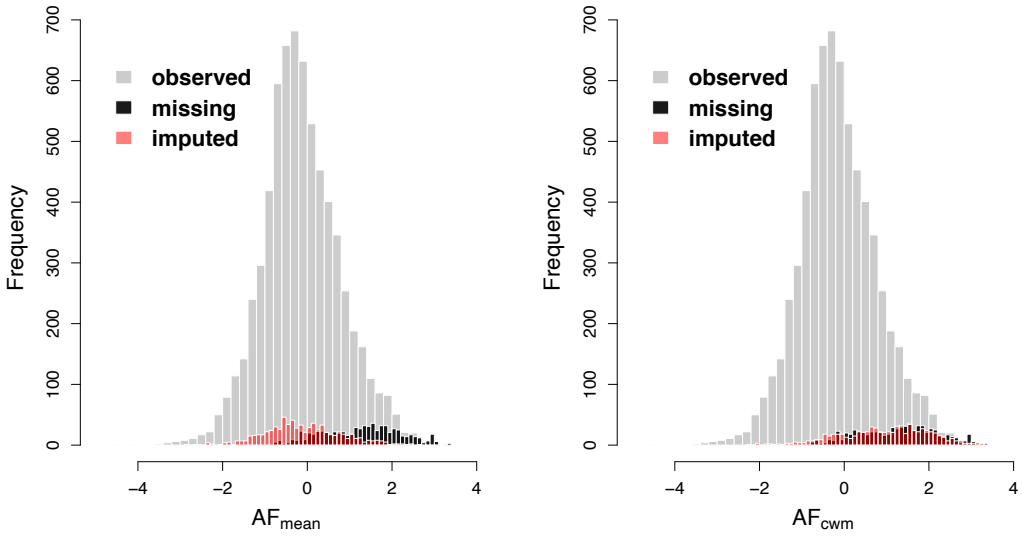


FIGURE 4.16: Histograms of observed, missing, and imputed data for the variables imputed by the two procedures of interest, AF_{mean} and AF_{cwm}.

in Table 4.8, where the last column refers to the number of occupied clusters, examining the occupancy rate between MCMC iterations. With these results and using the parsimony principle, we decided to select a model with $G = 5$ clusters.

G	WAIC	# Occupied clusters	G	WAIC	# Occupied clusters
2	32519.4	2	7	31801.6	7
3	31960.7	3	8	31800.3	6
4	31853.6	4	9	31801.6	6
5	**31799.3**	**5**	10	31800.9	5
6	31802.9	6			

TABLE 4.8: WAIC for different models according to the number of components for the EAM database.

The estimates obtained for each of the imputed databases allow us to calculate the KL divergence to compare the different imputation models. The amount of information lost when the distribution of AF_{com} is approximated by the distributions of AF_{obs}, AF_{mean} and AF_{cwm} is calculated. These values are shown in Table 4.9 calculated using the `integrate` function from the R software. A column is also shown with the relative distance, calculated taking as a reference the KL divergence for the observed data.

The values of Table 4.9 show that the distribution estimated with AF_{cwm} is the one that loses the least information when used to approximate the distribution with the complete data, AF_{com}. The amount of information lost by using AF_{obs} and AF_{mean} was similar. The use of the KL divergence shows a better performance of the imputation model using the information of the variable GPO with the procedure proposed.

As a complement, in Figure 4.17 we can see the comparison of the estimated densities of the imputed variable with auxiliary information (AF_{cwm}) and that imputed without any information (AF_{mean}). The histogram corresponds to the distribution of

	Approach method	
	KL_{int}	Relative distance
AF_{com}	-	-
AF_{obs}	0.02002	1.00
$AF_{\texttt{mean}}$	0.01817	0.91
$AF_{\texttt{cwm}}$	0.00116	0.06

TABLE 4.9: KL divergence and relative distance for the EAM imputed database using `mean` and `cwm` methods.

the variable AF_{com}, while the curves refer to the estimated densities with information from the variables AF_{com}, AF_{obs}, $AF_{\texttt{mean}}$ and $AF_{\texttt{cwm}}$. The graph shows that the density of the variable $AF_{\texttt{mean}}$ (blue dotted line) is closer to that of the variable AF_{obs} (solid gray line), while the density of the variable $AF_{\texttt{cwm}}$ (red dashed line) is closer to that corresponding to the variable AF_{com} (solid black line).

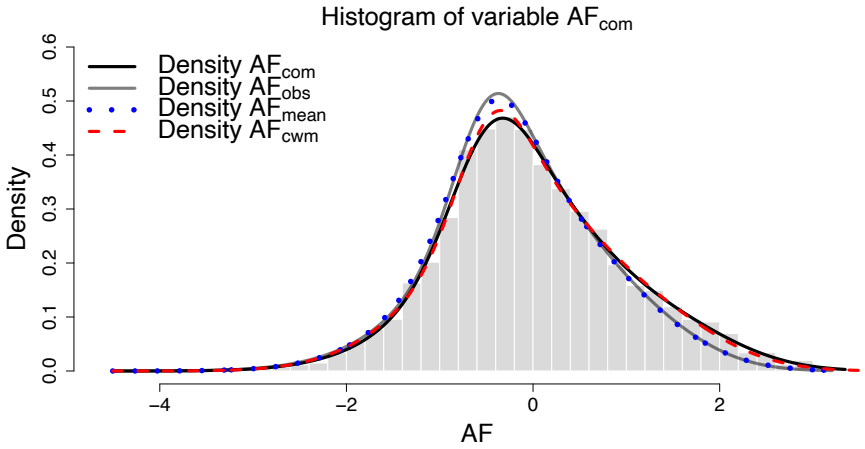


FIGURE 4.17: Estimated densities of observed, missing, and imputed data for the variables imputed by the two procedures of interest, $AF_{\texttt{mean}}$ and $AF_{\texttt{cwm}}$.

In summary, the implementation of our model, which includes auxiliary information, compared to that which does not include it, managed to improve the imputation process. The improvement criterion was based on measuring the amount of information lost when the objective was to approximate the distribution of complete data. We are based on a graphical analysis complemented with a quantitative measure such as the KL divergence.

Chapter 5

Multivariate Gaussian LCWM

"It is madness to hate all roses
because you got scratched with one thorn.
To give up on your dreams
because one didn't come true."

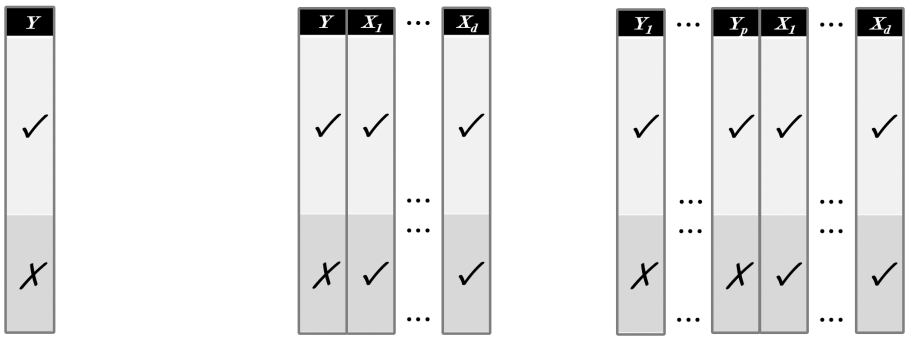
The Little Prince.

5.1 Overview

For this chapter, the primary interest is to generalize our proposed model to the multivariate case. As has been emphasized throughout the document, our interest stems from the model implemented by Paiva and Reiter (2017). The authors present in their work a new approach for generating imputations for multivariate continuous data with nonignorable unit nonresponse. The versatility of their model lies mainly in two characteristics: one, the Gaussian FMM is flexible concerning the modeling of unknown distributional forms; and two, for the case of imputation of databases with a non-ignorable response unit, it facilitates the use of pattern-mixture models by manipulating the mixture probabilities that the FMM initially estimates. In our case, the interest lies in the use of its applications where there is a group structure in the data or where the objective is to explore the data for said structure, or when the data have unknown distributional shapes (McLachlan, Lee, and Rathnayake, 2019). By including additional information through fully observed variables for all individuals, and by assuming such variables as observational, the model is expected to use such information adaptively to decide with which component to impute from, and to do so assuming a MAR missingness mechanism.

The different structures addressed throughout the study can be illustrated in Figure 5.1. The simplest particular case can be represented in Figure 5.1a, and we refer to this as the univariate case that does not include auxiliary information. Starting from this, we include auxiliary information in such a way that our pattern of interest takes the form presented in Figure 5.1b. A detailed study of this case is made in Chapter 4. The most general case, represented by Figure 5.1c, will be the object of study below, and we will refer to it as *multivariate Gaussian LCWM*.

The last section of Chapter 3 presents the theoretical results regarding the multivariate model. In this section, three results are established in two propositions, and a corollary summarizes the relationships for the FMM, MRM and LCWM models in the Gaussian context. In addition, expressions for the mixing probabilities, marginal and conditional distributions corresponding to the Gaussian LCWM are presented and are the basis for the development of this chapter.



(a) No auxiliary information: univariate case (b) With auxiliary information: univariate case (c) With auxiliary information: multivariate case

FIGURE 5.1: Structure of the missing data patterns addressed throughout the study until reaching the Multivariate Gaussian LCWM.

Section 5.2 presents simulation studies for the multivariate model. It explores the behavior of the model from two variables, one that does not give any information about which component to impute from and another where the variable is distributed separately among components. As in the univariate case, a vector is constructed from the two variables that enter the model as auxiliary information and the results obtained are analyzed. These three scenarios are implemented using two additional imputation models used in the context of missing data: *predictive mean matching* (Little, 1988) and *Bayesian multiple imputation* (Rubin, 1987). The performance of the two methodologies is compared with our model and the results obtained when the database is imputed are also included without using additional information. Finally, Section 5.3 implements the proposed model on the `iris` database. Two missing data patterns are simulated on this database, one under the MAR mechanism and the other under an MNAR mechanism, and the results are analyzed.

5.2 Simulation Studies

For the multivariate case, a data set was simulated from a mixture of normal distributions in four dimensions with two components. Two of the variables were considered output variables ($p = 2$), while the other two were considered input variables ($d = 2$). The database contains `n = 1000` observations of the form (x_1, x_2, y_1, y_2). The mixing probabilities are $\alpha_1 = 0.6$ and $\alpha_2 = 0.4$, the mean vectors $\boldsymbol{\mu}_1 = (1.0, 3.0, 4.0, 2.0)$ and $\boldsymbol{\mu}_2 = (1.0, 9.0, 7.0, 6.0)$, and the covariance matrices are:

$$\Sigma_1 = \begin{pmatrix} 1.00 & 0.50 & 0.50 & 0.50 \\ 0.50 & 1.00 & 0.50 & 0.50 \\ 0.50 & 0.50 & 1.00 & 0.50 \\ 0.50 & 0.50 & 0.50 & 1.00 \end{pmatrix} \quad \Sigma_2 = \begin{pmatrix} 1.00 & -0.50 & -0.50 & 0.50 \\ -0.50 & 1.00 & 0.50 & -0.50 \\ -0.50 & 0.50 & 1.00 & -0.50 \\ 0.50 & -0.50 & -0.50 & 1.00 \end{pmatrix}$$

Missing data following a MAR mechanism is generated for the data set, considering the variables Y_1 and Y_2 as the ones with missing information, and the variables X_1 and X_2 as fully observed. For cluster 1, 10% of the data was randomly selected and considered missing, while 50% was selected for cluster 2. A summary of how the data was generated is presented in Table 5.1.

	observed		missing		complete	
cluster 1	516	*(89.7%)*	59	*(10.3%)*	575	*(100%)*
	(69.9%)		*(22.5%)*		*(57.5%)*	
cluster 2	222	*(52.2%)*	203	*(47.8%)*	425	*(100%)*
	(30.1%)		*(77.5%)*		*(42.5%)*	
total	738	*(68.3%)*	262	*(31.7%)*	1000	*(100%)*
	(100%)		*(100%)*		*(100%)*	

TABLE 5.1: Distribution of simulated data and pattern of missing data under a MAR mechanism for the multivariate case.

To complement, the scatter plots matrix of the observed and missing data are illustrated in Figure 5.2. In the graphs, the behavior of the input variables can be observed through projections of the 4-dimensional data set. Two types of behavior are considered similar to the univariate case, an input variable X_1 that does not give information on which component to impute from, and a second input variable X_2 distributed separately among components (see the projections $X_1 \times Y_1$, $X_1 \times Y_2$, $X_2 \times Y_1$, and $X_2 \times Y_2$ in the planes in the lower left corner of Figure 5.2).

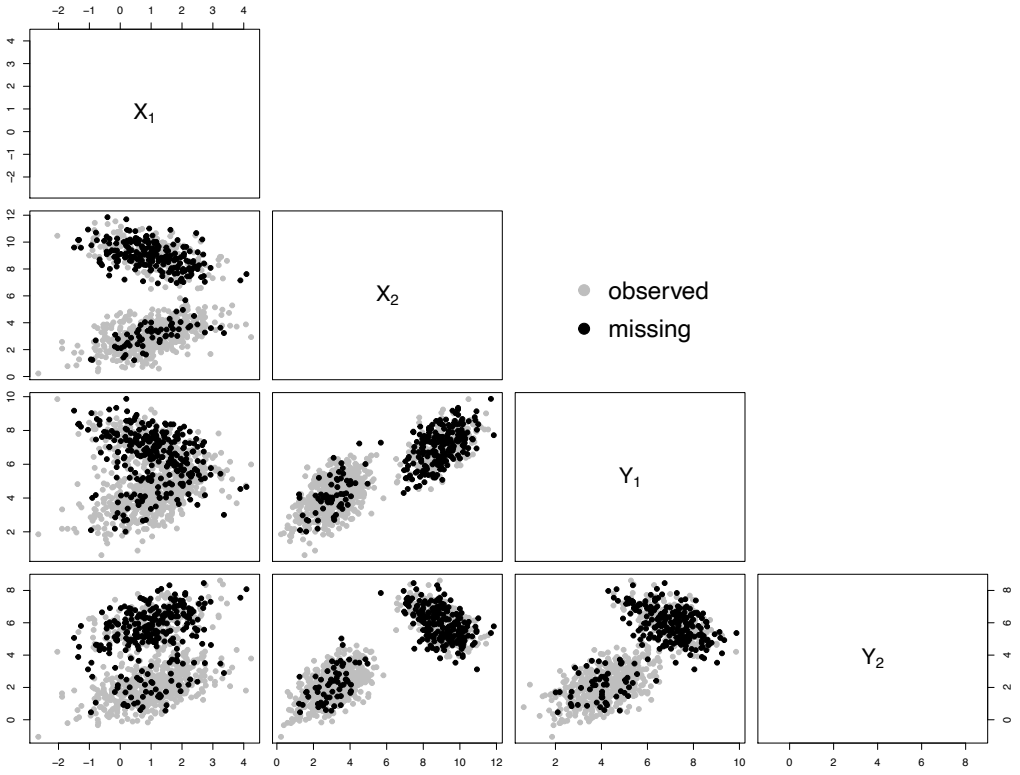


FIGURE 5.2: Pairwise plots of the variables in the simulated database. Observed and missing data generated under a MAR mechanism for the multivariate case.

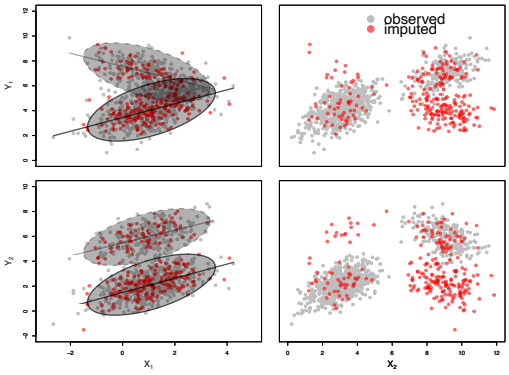
5.2.1 Model performance when new information is included

For this section, the objective is to analyze the type of information that can be entered into the model through the input variables. Based on the simulated data set, we will consider three scenarios. In the first scenario, the variable that enters the model does not provide information on which component to impute from. In the second scenario, the input variable is distributed separately among components, which allows the model to decide in a correct way with which component to impute from. Finally, in the third scenario, an input vector is constructed with the two previous variables, containing a variable that has a separate distribution between components, and the vector inherits this characteristic and its distribution is separated between components.

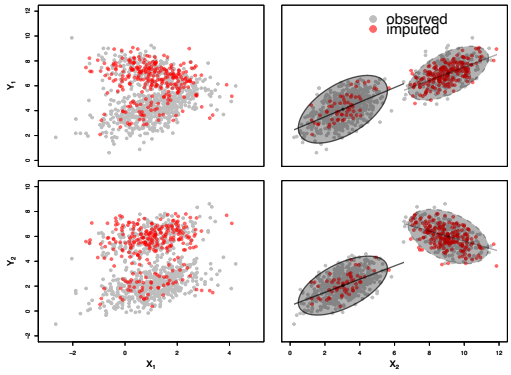
Figure 5.3 shows how the imputation models are built with the information obtained through the three mentioned situations. Figure 5.3a illustrates the construction of the imputation model when the auxiliary information that is entered into the model is done through the variable X_1. This variable does not provide information on which component to impute from. The decision remains in the hands of the estimates of the mixing probabilities, strongly influenced by the number of observations in each cluster. The left-hand panels correspond to the projection plane $X_1 \times Y_1$ and $X_1 \times Y_2$. In them, we observe 95% quantile ellipses for each component and the respective regression lines. We can see that the imputations made by the model are located around the regression lines. The panels on the right hand side illustrate the projections in the planes involving the variable X_2 and allow us to conclude that many imputed observations are made far from the observed data regions.

For the case in which we include information from the X_2 variable, Figure 5.3b illustrates the construction of the model. Similar to the previous case, the 95% quantile ellipses and the regression lines are shown. Observing the four panels together, we can conclude that although it is imputed with only the information of the variable X_2, in all the projection planes, the imputed data is generated from regions with observed information, which seems to indicate the quality of the information that the input variable X_2 delivers to the model. As in the previous case, the imputations are made around the regression lines. Figure 5.3c shows how the Gaussian LCWM proceeds when the auxiliary information considers an input vector made up of the variables X_1 and X_2. It shows a behavior similar to the case where we consider auxiliary information of the input variable X_2. Imputed values are generated from regions with observed information and around the regression lines. It should be noted that, in this case, we have regression planes and quantile ellipsoids for each of the output variables. The figures illustrate their projections on the planes of interest.

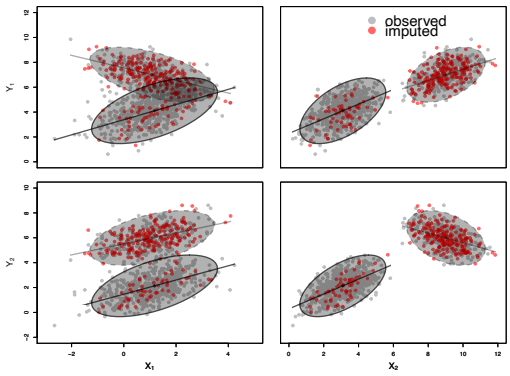
Figure 5.4 presents pairwise plots that illustrate different imputation processes to be compared. The observed, missing and imputed data are plotted on each plane of the scatter plot matrix. We start with the analysis of the imputation process using the input variable X_1. Figure 5.4a shows how the imputations (red dots) behaved with respect to the missing data pattern (black dots). We can see that the procedure incorrectly imputes with respect to the proportion made in each cluster. Cluster 1 has the smallest proportion of missing data, however the method imputes the largest proportion of data here. As we mentioned, this variable does not provide information on which component to make the imputation from. The imputation procedure is established from the estimates of the probabilities of mixtures, which are strongly



(a) With auxiliary information from variable X_1.



(b) With auxiliary information from variable X_2.



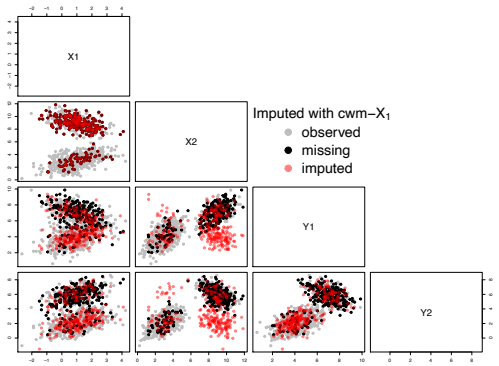
(c) With auxiliary information from vector (X_1, X_2).

FIGURE 5.3: Construction of the imputation process through the Gaussian LCWM using information from different types of variables. Observed and imputed data in the different projection planes. 95% quantile ellipses and regression lines.

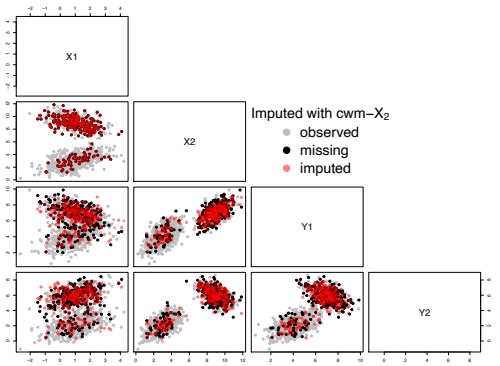
influenced by the proportion of data observed in each component. The projection planes $X_2 \times Y_1$ and $X_2 \times Y_2$ allow us to observe regions with incorrectly imputed observations, regions that do not coincide with either observed data or missing data. These projections allow us to clearly see the error made in the imputation process, which is evident on the $Y_1 \times Y_2$ plane through the difference in the proportions of imputed and missing data in each component.

For the two situations that follow, we can observe a better performance of our model. When we use as auxiliary information a variable or vector that is distributed separately among components, the marginal distribution in the Gaussian LCWM can determine with which component to impute from precisely. Figure 5.4b illustrates the way in which our model imputes using information from the variable X_2. This variable is distributed separately among components, and we observe how the imputations cover the regions corresponding to the missing data in the same proportions in which they appear. Figure 5.4c refers to our model using the input vector (X_1, X_2). A behavior similar to the previous one is shown with respect to the results of the imputation procedure. It is of special interest to refer here to the idea that we mentioned in Chapter 4, regarding an input vector with separate distribution between components. This idea can be reinforced using the multivariate model. We can see in the panel in Figure 5.4c, corresponding to the projection $X_1 \times X_2$, that two

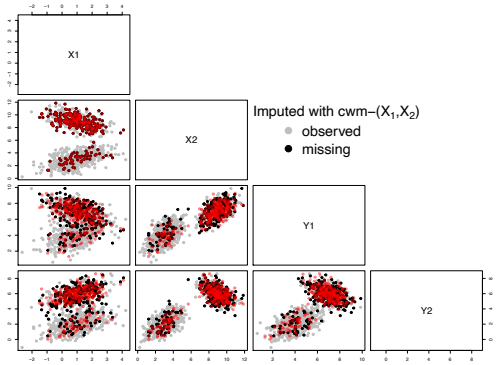
separate groups in the scatter diagram allow us to establish that the distribution of the input vector (X_1, X_2) is separated between components. In the univariate case, we try to illustrate this idea with the bimodal histogram at the bottom of Figure 4.2b.



(a) Imputed with information from variable X_1



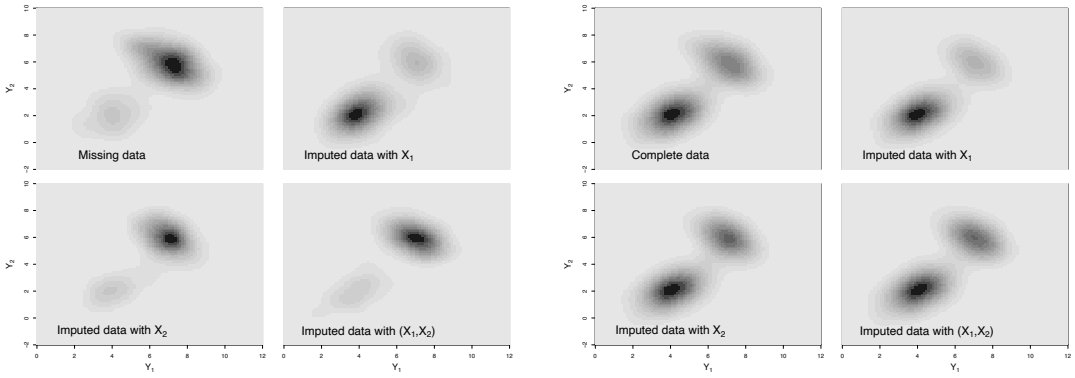
(b) Imputed with information from variable X_2



(c) Imputed with information from vector $(X_1, X2)$.

FIGURE 5.4: Pairwise plots of the simulated multivariate database imputed by Gaussian LCWM with information on different types of variables.

Finally, heat maps for the output variables of the imputed databases using the Gaussian LCWM under the different types of input variables are presented in Figure 5.5. We consider heat maps for imputed values only, in the cases of including as auxiliary information in our model the variables X_1, X_2 and the vector (X_1, X_2). We compare these maps with the one that refers to the data generated as missing to analyze the performance of each type of variable, see Figure 5.5a. Here we can observe the similarity between the maps that correspond to the original missing data and those imputed with information from the variable X_2 and with the vector (X_1, X_2). On the other hand, the map obtained from the database imputed with X_1 differs from the one corresponding to missing data, which allows us to conclude the poor performance of the model with the characteristics of this variable. The heat maps in Figure 5.5b refer to the same cases but for the complete output variables. A greater similarity is observed in the heat maps for complete and imputed data with the variable X_2 and the vector (X_1, X_2). For the case of the data imputed with the variable X_1, a slight difference is observed specifically for the values corresponding to cluster 2.



(a) Missing data (b) Complete data

FIGURE 5.5: Bivariate distribution heatmaps of database imputed using Gaussian LCWM with information on the variables X_1, X_2 and the vector (X_1, X_2).

5.2.2 Gaussian LCWM performance relative to other imputation methods

For this section, we consider comparing our imputation model, the Gaussian LCWM (`cwm`), with various procedures of interest. Among these, is the model that does not include information and we call it the mean method (`mean`), and two methods whose interest is centered on the fact that they are based on a Bayesian approach and on values predicted by a regression model, we refer to these procedures as *predictive mean matching* (`pmm`) and *Bayesian multiple imputation* (`norm`). For these last two procedures that use auxiliary information, we will consider the information coming from the same input variables that we used in the previous section to analyze the performance of our model, we refer to the variables X_1, X_2 and the vector (X_1, X_2).

A graphical analysis of the different imputation processes to be compared can be done from the pairwise plots for the procedures that interest us. Graphs for the procedure `mean`, and for `pmm` and `norm` in the case of information of the variables X_1, X_2 and the vector (X_1, X_2) are shown in Appendix C.2.

As a starting point, the procedure `mean` had similar results to the case of our model with information from the variable X_1. The procedure imputed observations in regions where missing data was not generated, and the proportions of imputed data in each component appear to have been influenced by the data observed in each component. Within each procedure, the performance of the input variables had a behavior similar to the case of the Gaussian LCWM. The variable X_1 had the worst performance, while the variable X_2 and the vector (X_1, X_2) allowed us to observe the best behaviors.

Although for each of the methods a similar behavior occurs when the different input variables are used, something that seems to be advantageous for our methodology is that the values are imputed within the regions established from the observed data, and very close to the missing values. The procedure `norm` performs some imputations outside the regions defined by the clusters, even if the input variable is separated between components. In addition, the imputations that are made within one of the components do not maintain the trend of the data within it. The `pmm` procedure maintains a better performance in the imputations made, closely following

the behavior of the `cwm` methodology.

The analysis carried out can be complemented by using a quantitative measure that allows us to evaluate the different imputation processes. As in the univariate case, the Kullback-Leibler divergence will be used as a metric to compare performance when different types of variables are used as auxiliary information in the proposed model. We will also use it when our objective is to compare with other methods of interest.

5.2.3 Quantitative diagnosis of imputation processes

The use of the Kullback-Leibler divergence aims to compare the estimated distributions for the imputed databases under the different paradigms with the original distribution. Recall that here we seek to measure the amount of information lost when we approximate the true distribution using the distributions estimated from the bases imputed by the different procedures.

The procedure we use to calculate this measure starts from completing the data set from the imputations using the different methods. After this, we estimate the parameters of the model and take the information corresponding to the output variables, means and covariance matrices, together with the estimates of the mixing probabilities. Since there is no closed expression for the calculation of the KL divergence in the case of the Gaussian FMM, we use an approximation by means of Monte Carlo methods (Hershey and Olsen, 2007; Durrieu, Thiran, and Kelly, 2012). To refer to this approximation, we will use the notation KL_{mc}. The results obtained are shown in Table 5.2. At the top of the table, a 95% quantile interval is presented for values of the KL divergence. The calculation is obtained from `N = 10000` sampled databases of size `n = 1000` corresponding to the true distribution of the output variables (Y_1, Y_2). Their parameters are estimated using the `mixsmsn` package and calculating their KL divergence for each base. We assume that any KL divergence value that falls within the interval allows us to conclude that the original distribution was recovered, this will be noted with `WI` in the relative distance column. The values in this column are obtained by calculating the distance between the corresponding KL divergence value and the right limit of the quantile interval.

	`cwm`		`pmm`		`norm`	
	KL_{mc}	Relative distance	KL_{mc}	Relative distance	KL_{mc}	Relative distance
Qu.int. 95.0%	(0,0.0107)	-	(0,0.0107)	-	(0,0.0107)	-
$g_{(Y_1,Y_2)_{com}}$	0.0081	WI	0.0081	WI	0.0081	WI
$g_{(Y_1,Y_2)_{obs}}$	0.0358	3.36	0.0358	3.36	0.0358	3.36
$g_{(Y_1,Y_2)_{mean}}$	0.3140	29.46	-	-	-	-
$g_{(Y_1,Y_2)_{X_1}}$	0.0342	3.22	0.3092	29.00	0.3114	29.21
$g_{(Y_1,Y_2)_{X_2}}$	0.0149	1.40	0.0181	1.75	0.0620	5.81
$g_{(Y_1,Y_2)_{(X_1,X_2)}}$	0.0112	1.05	0.0247	2.31	0.0505	4.74

TABLE 5.2: KL divergence and relative distance for the estimated distributions of the complete, observed and imputed databases in the three cases of interest and the four models to be compared in the multivariate case.

The results of the column corresponding to `cwm` in Table 5.2 allow quantitative confirmation of the analyzes carried out from the graphs of Section 5.2.1. The distributions that lose the least information when trying to approximate the original distribution are those that were imputed using the information from the variable X_2

and the vector (X_1, X_2). Indeed, those are the ones that use information from variables or vectors that are distributed separately among components. Similar to the univariate case, the imputation process with the variable X_1 has a poor performance. However, constructing an input vector with at least one variable that is distributed separately between components, allows to obtain a vector with separated distribution between components and the imputation process with such vector has a good performance. This happened by integrating the variables X_1 and X_2 into a vector and using it as the input vector in our model.

Section 5.2.2 aimed to compare the different imputation methods with the Gaussian LCWM. A first comparison can be carried out using the KL divergence when confronting our model with the procedure that does not include auxiliary information, we refer to that as the `mean` imputation method. We observe that, even using the variable X_1, our model achieved a better performance. The columns labeled `pmm` and `norm` in Table 5.2 present the KL divergence values for similar cases implemented for our model, i.e., imputation of the output variables (Y_1, Y_2) with information from X_1, X_2 and (X_1, X_2). Results similar to those analyzed for the `cwm` method were obtained within each methodology. When comparing the three procedures, we observe a better performance of our model in the three specific situations that are presented. In the case of the use of auxiliary information from the variable X_1, the three procedures show the worst performance, however our imputation model shows by far the smallest loss of information. Similar conclusions can be obtained from Table B.4 of Appendix B where the KL divergence values are calculated with respect to the distribution obtained from complete data.

5.3 Ilustrative Example with Real Data

5.3.1 Data Set: Iris

The Iris flower data set is a multivariate data set collected by Anderson (1935) and first analyzed by Fisher (1936). The data set consists of 50 samples from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor). Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters. Several authors, including McLachlan and Peel (2004) and Frühwirth-Schnatter (2006), have made use of the `iris` database in the framework of the implementation of models related to the Gaussian FMM.

The database variables were divided into two groups, one in which its variables were considered output variables (`Sepal.Length` and `Petal.Width`) and on which the missing data pattern was simulated. The second set of variables was considered as input variables and they are assumed to be fully observed (`Sepal.Width` and `Petal.Length`). Each of the three species was considered as a cluster in the data set, in such a way that, for the diagnosis of the imputation processes, $G = 3$ will be assumed as the true number of components.

The way in which the variables of the `iris` database are distributed is presented in Figure 5.6. The behavior of the input variables approximates the scenarios presented in the case of simulated data. `Petal.Length` variable is characterized by having a distribution that is separated into two groups; the versicolor and virginica species appear next to each other, visually forming a single group. The variable

`Petal.Length` does not give any information on which component to allocate with. Since the missing values will be imputed with the information provided by both variables, the input vector inherits the characteristic of separate distribution between components of the variable `Petal.Length`.

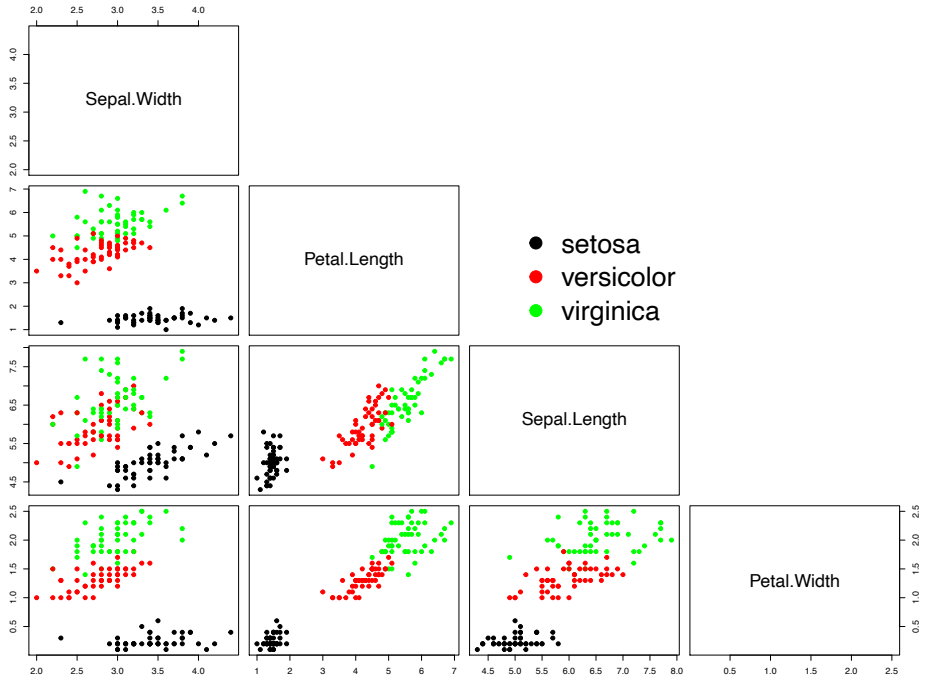


FIGURE 5.6: Pairwise plot of the variables in the `iris` database discriminated by species. At the top left and bottom right, scatter plots of the input variables and output variables are shown respectively. In the lower left corner, the crosses of these variables are displayed.

Two missing data scenarios were simulated, one using a MAR mechanism, while the second used a MNAR mechanism. The databases were imputed using our model together with the three methodologies considered throughout the document.

Simulation of missing data under MAR mechanism

To generate a set of data missing under the MAR mechanism, different probabilities of missingness were established for each species. In this way, 30% of missing data was simulated for the setosa species, 20% for the versicolor species, and 10% for the virginica species. With these proportions thus defined, 20% of missing data is expected in the complete database. The distribution of observed, missing, and complete data obtained by species is shown in Table 5.3.

The database was imputed using the methods presented here. Specifically, the variables `Sepal.Length` and `Petal.Width` are imputed with information from the fully observed variables `Sepal.Width` and `Petal.Length`. For the `mean`, `cwm`, and `norm` imputation procedures, the code designed by us for the Gaussian LCWM was used. In the case of the `mean` method, the option to impute without auxiliary information is specified on the program; for the case of the `norm` procedure, it is implemented by specifying the condition of the number of clusters as $G = 1$, together

	observed		missing		complete	
setosa	36 *(30.5%)*	*(72.0%)*	14 *(43.8%)*	*(28.0%)*	50 *(33.3%)*	*(100%)*
versicolor	37 *(31.4%)*	*(74.0%)*	13 *(40.6%)*	*(26.0%)*	50 *(33.3%)*	*(100%)*
virginica	45 *(38.1%)*	*(90.0%)*	5 *(15.6%)*	*(10.0%)*	50 *(33.3%)*	*(100%)*
total	118 *(100%)*	*(78.7%)*	32 *(100%)*	*(21.3%)*	150 *(100%)*	*(100%)*

TABLE 5.3: Distribution of simulated missing data for the `iris` database under the MAR mechanism.

with the use of auxiliary information. For our `cwm` methodology, we specify a value of $G = 10$ clusters, as well as the use of auxiliary information. In the case of the `pmm` method, the `mice` package is used to implement it.

We present the pairwise plot for the variables of the imputed database using the Gaussian LCWM (see Figure 5.7). Each panel illustrates the observed, missing, and imputed values. The panels show that the regions where missing data appear are well represented by the imputed data. In the panel in the upper left part, the input variables are crossed, since in the imputation procedure this is considered as known information, the points in red appear superimposed on the points in black. In the lower right panel, the output variables are crossed, those with missing information. Here the imputations made by the model can be observed. The four panels in the lower left corner cross the input and output variables and show how the imputation process was conducted.

Pairwise plots similar for each imputation procedure are shown in Appendix E. From these graphs, it can be seen that the `norm` and `pmm` procedures visually present close results. The regions where missing data was generated are covered by imputed data in similar proportions. Although for the component corresponding to the setosa species, the `mean` procedure shows a slightly greater dispersion in the imputed data.

A quantitative analysis is presented using the KL divergence values presented in Table 5.4. In all cases, the adjustment of a Gaussian FMM with $G = 3$ components was considered. For them, the complete database was used, the one with observed data, while in the case of the imputation procedures to be compared, databases imputed by each of the methods of interest were used. We run our code for the estimation procedure with values of `burn-in = 10000` and `effectiveSize = 1000`. With values of the estimates of the mixture probabilities, mean vectors and covariance matrices, and using a procedure based on Monte Carlo methods, the values of the KL divergence were approximate (Hershey and Olsen, 2007; Durrieu, Thiran, and Kelly, 2012).

For this case, we take as a reference the value of the KL divergence for observed data. Thus, the relative distance between the distribution of complete and observed data, based on the KL divergence, will be taken as unity. For imputation procedures with distances less than one, the methodology will be considered good in the sense that less information is lost when using the estimated distribution of the imputed

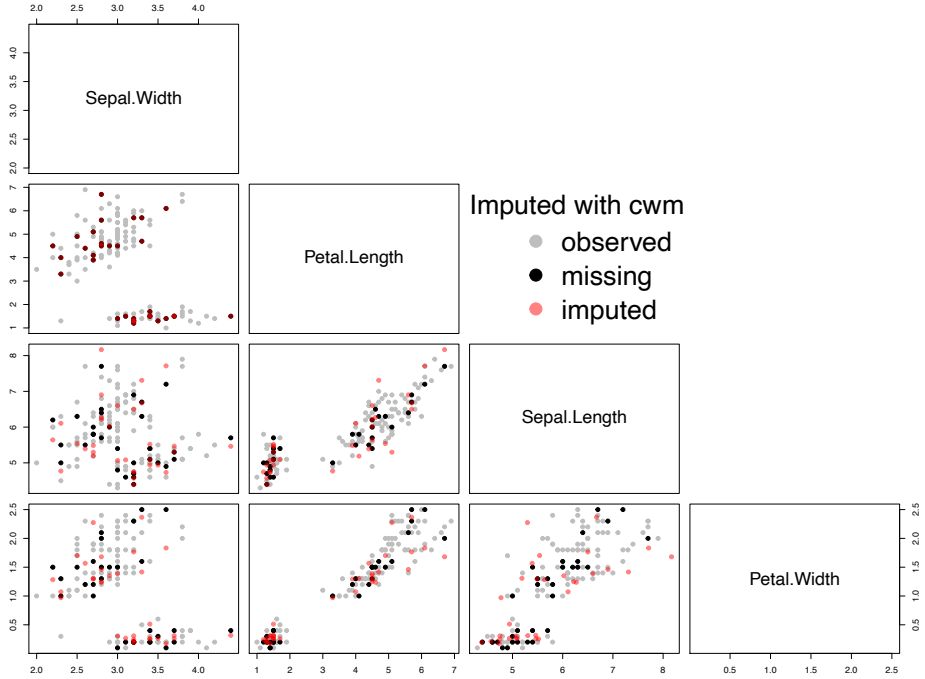


FIGURE 5.7: Pairwise plot of the imputed `iris` database using the Gaussian LCWM. Each panel presents the crossing of two of the variables specifying observed values, missing values and imputed values. Missing data generated using a MAR mechanism.

base to approximate the distribution of complete data, compared to the estimated distribution of observed data. Based on this criterion, the results in Table 5.4 show that the worst performance corresponded to the imputation process without auxiliary information, as expected. Of the methods that used auxiliary information, the one that had the best performance was ours, followed by the `pmm` methodology. The `norm` procedure maintained a loss of information similar to the distribution with only observed data.

Simulation of missing data under MNAR mechanism

A second scenario that was simulated for the missing data set was based on an MNAR mechanism. For this, the missing data set was generated in such a way

	Approach method	
	KL_{mc}	Relative distance
$(PW,SL)_{com}$	-	-
$(PW,SL)_{obs}$	0.0431	1.00
$(PW,SL)_{mean}$	0.1280	2.97
$(PW,SL)_{cwm}$	0.0252	0.58
$(PW,SL)_{pmm}$	0.0396	0.92
$(PW,SL)_{norm}$	0.0429	1.00

TABLE 5.4: KL divergences for the imputed `iris` database. Relative distances taken with reference to the estimated distribution of observed data. Missing data generated from a MAR mechanism.

that they were related to their values. In our case, the missing data was simulated in such a way that the larger values of the `Sepal.Length` and `Petal.Width` output variables had a higher probability of missing. To simulate this missing data mechanism, the expression in (4.2) was used with specific values of β_0 and β_1. The distribution of missing data within the database is summarized in Table 5.5 and establishes the amount of observed, missing, and complete data for each species. Since the highest values of the output variables are found in the virginica species, then the versicolor species and the lowest values in the setosa species, the largest amount of missing data is generated in the virginica category (22), then in the group corresponding to versicolor (7) and finally in the setosa species (2).

	observed		**missing**		**complete**	
setosa	48	*(96.0%)*	2	*(4.0%)*	50	*(100%)*
	(40.3%)		*(6.4%)*		*(33.3%)*	
versicolor	43	*(86.0%)*	7	*(14.0%)*	50	*(100%)*
	(36.1%)		*(22.6%)*		*(33.3%)*	
virginica	28	*(56.0%)*	22	*(44.0%)*	50	*(100%)*
	(23.5%)		*(71.0%)*		*(33.3%)*	
total	119	*(80.0%)*	31	*(20.0%)*	150	*(100%)*
	(100%)		*(100%)*		*(100%)*	

TABLE 5.5: Distribution of simulated missing data for the `iris` database under the MNAR mechanism.

As in the case of missing data under MAR in the previous section, the output variables `Sepal.Length` and `Petal.Width` are imputed with information from the fully observed variables `Sepal.Width` and `Petal.Length`. The conditions to implement the imputation procedures through our programming code on the R software follow the same structure mentioned above, as well as the use of the `MICE` package for the `pmm` procedure.

Figure 5.8 shows pairwise plots for the imputed database in the case of our `cwm` methodology. We can see that our model imputes in the different regions with the same proportions with which the missing data was generated. Where the values of the variables are higher, more imputations appear according to the number of observations that were missing. In the region where the lowest values of the variable are found, a pair of missing values was generated and the model imputed the same. Our imputation process allowed us to cover all the regions where missing data was generated, something that does not seem to have happened with the `norm` and `pmm` methods. In a part of the region occupied by the virginica species, the data disappeared completely and this resulted in the two procedures not imputing data there. Van Buuren (2018) affirms that, although the `pmm` procedure is robust, this last situation that we have just described causes the `pmm` method to impute with the same observations and this becomes a problem when it is imputed with this methodology. This was confirmed by us in the simulation studies in Chapter 4. The pairwise graphs for the other imputation procedures can be consulted in Appendix E.

Once again, the analyzes made from the graphs can be supported by the KL divergence values presented in Table 5.6. In this case, we could observe that the only procedure that presented a good performance compared to the information provided by the set of observed data was ours. The value of the KL divergence for

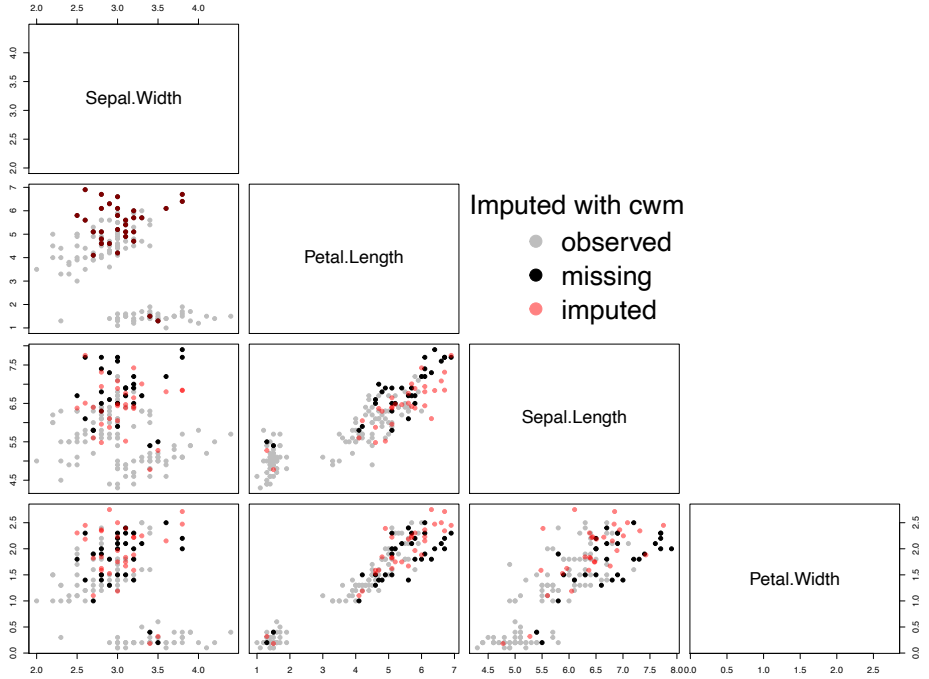


FIGURE 5.8: Pairwise plot of the imputed `iris` database using the Gaussian LCWM. Each panel presents the crossing of two of the variables specifying observed values, missing values and imputed values. Missing data generated using a MNAR mechanism.

the imputation methodology `pmm` matched that for observed data. The `mean` and `norm` procedures performed the worst.

	Approach method	
	KL_{mc}	Relative distance
$(PW,SL)_{com}$	-	-
$(PW,SL)_{obs}$	0.1405	1.00
$(PW,SL)_{mean}$	0.2061	1.46
$(PW,SL)_{cwm}$	0.0762	0.54
$(PW,SL)_{pmm}$	0.1410	1.00
$(PW,SL)_{norm}$	0.2186	1.56

TABLE 5.6: KL divergences for the imputed `iris` database. Relative distances taken with reference to the estimated distribution of observed data. Missing data generated from a MNAR mechanism.

To conclude, two missing data patterns were simulated on the iris database. A first pattern was generated under a MAR mechanism, while the second was made under an MNAR mechanism. In both cases, our imputation model performed better. Initially, when it was compared with the methodology that did not use auxiliary information, and later it performed in a similar way when it was compared with methods of interest that also made use of auxiliary information from fully observed variables. It is important to note that, although the missing data was simulated from MAR and MNAR mechanisms, there is no underlying structure in our model that

considers non-ignorable response mechanisms. Its good performance in the face of missing data under a MNAR mechanism characterizes it as a robust method.

Chapter 6

Conclusions and Final Observations

"One day, I watched the sun setting
forty-four times...
You know... when one is so terribly sad,
one loves sunsets."

The Little Prince.

We present a new methodology called *Gaussian Linear Cluster-Weighted Modeling* for the process of imputation of continuous multivariate data in the case in which the data set has a group-structure or where the objective is to explore the data for such structure or when the data have unknown distributional shapes. Uses the results of the Finite Mixture Models (McLachlan, Lee, and Rathnayake, 2019; Frühwirth-Schnatter, 2006) and the Cluster-Weighted Modeling (Gershenfeld, 1997; Ingrassia, Minotti, and Vittadini, 2012), both restricted to Gaussian distributions. We use a fully Bayesian approach that jointly models and imputes data with missing values using a flexible Dirichlet process mixture of multivariate normal distributions. The imputation model is designed on a non-response unit pattern where variables with missing information are called output variables. For all individuals we assume that it is possible to find auxiliary information from other sources, these fully observed variables are called input variables. Under the assumption of considering the input information as observational, we jointly model input and output variables in such a way that the model uses the input information adaptively to characterize the components. Likewise, the model uses this information to decide with which component to impute.

It is possible to include input variables that move between two extreme scenarios. First, we have variables that do not provide information on which component to impute from, these variables have a similar distribution among the components detected by the model. On the other hand, we have input variables that are distributed *separately among the components*. These are variables with an ideal behavior, which with the information they provide correctly indicate to the model from which component to impute. Furthermore, including in the input vector at least one variable that is distributed separately among components, allows the vector to inherit this property, and its distribution has this same desirable characteristic for the information that is entered into the model. The performance of the type of input variable that enters the imputation model was able to be evaluated through simulated databases where the pattern of missing data was also simulated. A descriptive and graphical analysis was performed, complemented by a quantitative evaluation using the

Kullback-Leibler divergence, a non-symmetric measure of the similarity or difference between two probability distribution functions. For our case, these are the true distribution and the estimated distribution using the imputed database.

Our study was carried out first for the univariate case, and then generalized for the multivariate case. Initially, it was possible, starting from the model that does not consider additional information, to build a model that includes this information and that improves the imputation procedure. This is done under the possibility of selecting input variables with a high correlation with respect to the output variables and that are characterized to be distributed separately among components. Next, we compare the performance of the proposed model with other procedures that were of interest due to their Bayesian approach, and because they use prediction models to obtain the required imputations. These were the *predictive mean matching* and *Bayesian multiple imputation* procedures. The simulated scenarios showed a better performance of our model in the case of databases with group structure, as was the case of the simulated scenarios for univariate and multivariate data. We observed in the *predictive mean matching* methodology a robust procedure. In the different scenarios simulated by us, it always showed results very close to our model. However, when we simulated an extreme scenario, where the missing data pattern considered censored data and with a group structure, we could observe the poor performance of this procedure compared to ours. In this regard, Van Buuren (2018) states that the danger when using the imputation procedure `pmm` is the duplication of the same donor value many times. Also, this problem is more likely to occur if the sample is small or if much more data is missing than the observed data in a particular region of the predicted value.

The Gaussian LCWM imputation procedure in the univariate case was implemented on two real databases in which missing data patterns were simulated. For the Annual Survey of Manufacturing in Colombia dataset, a pattern of missing data was simulated under a MNAR mechanism. Our model allowed entering auxiliary information, achieving a better performance in the imputation process compared to that procedure that did not use additional information. Furthermore, although our model imputes assuming a MAR mechanism, we were able to observe that including appropriate auxiliary information improves the quality of imputations, even if the missing values had been generated from a MNAR mechanism. This allows us to conclude that our model is robust to the type of missing data mechanism. Also, for the univariate case, we implement our model using the Faithful database, which is characterized by having a group-structure. A pattern of missing data was simulated on it under a MNAR mechanism. The variable `eruption` was considered as the output variable, while `waiting` was assumed as the input variable. The missing data was imputed using our model together with three more procedures, the one that does not use auxiliary information, *predictive mean matching*, and *Bayesian multiple imputation*. Once again, the Gaussian LCWM imputation procedure showed the best performance with respect to the presented methods. The group-structure of the data set gives our model advantages over the others, as does the use of the input variable `waiting` which has a separate distribution between the two groups.

The theoretical results for the multivariate model were established, generalizing the results presented by Gershenfeld (1997) and Ingrassia, Minotti, and Vittadini (2012) for the univariate case. With these results and following a procedure similar

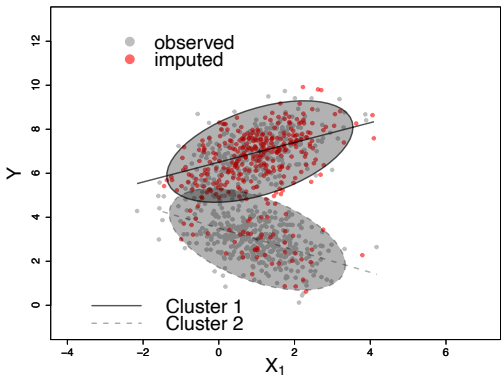
to the univariate case, a set of simulated data was considered on which the performance of different types of input variables for the imputation model was analyzed. Similar characteristics were obtained to the univariate case regarding the performance of different types of input variables. These variables move between two extreme scenarios, variables that do not provide information on which component to impute from and those that are characterized by being distributed separately among components. Due to its performance, this last type of variable is considered desirable to be included as auxiliary information in the proposed imputation model. On the same simulated scenarios, the Gaussian LCWM was compared with the aforementioned imputation procedures, obtaining favorable results for our model. Similarly, our methodology was implemented on the Iris database. On this data set, two input variables and two output variables were considered, in addition, two missing data patterns on MAR and MNAR mechanisms were simulated. The three procedures of interest were implemented for comparison. Once again, our model showed a better performance compared to the other procedures and in both scenarios.

Our imputation procedure was implemented with programming in R software. Several scripts were developed, and at the moment we are debugging them so that they can be shared in a virtual space so that those who are interested in adapting the methodology on databases with missing information have access to them. A direct future work of this study is to extend the methodology to the case of finite mixture models where the considered distributions are other than the normal distribution. For example, Skew Normal or $t-$Mixture Models, among others. It would also be of interest to develop criteria for the selection of meaningful variables that can be used as auxiliaries, as that allows to make the best possible imputation. Another possible extension is to create a `cwm-pmm` hybrid method, that takes advantage of both methodologies and check if it would improve the imputation performance. Finally, in some situations, according to the researcher's knowledge, it might be of interest to allow imputation to be carried in a particular region of the data set. Therefore, an extension to properly having a model that allows for explicit MNAR imputation is of interest.

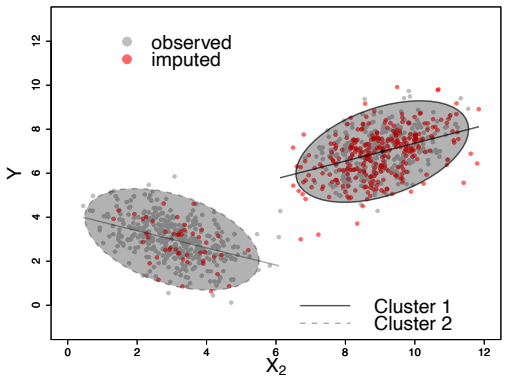
Appendix A

Univariate simulation with information from an input vector

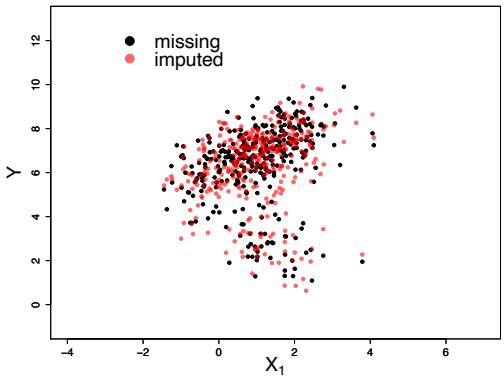
In the simulation studies in Section 4.2, where we evaluate the performance of the variables that enter the imputation procedure under the Gaussian LCWM, we discuss the case of auxiliary information from an input vector. The vector is made up of the two variables that were considered in the analysis, that is, $\boldsymbol{X} = (X_1, X_2)$. Here are the graphs that illustrate this scenario.



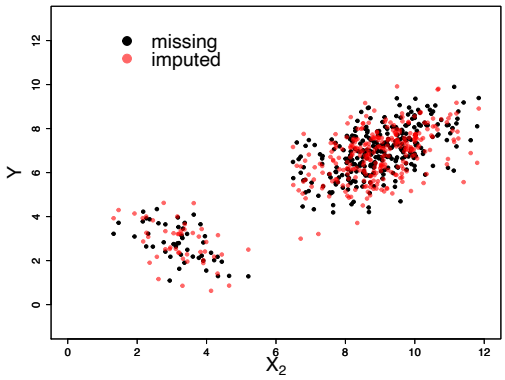
(a) Construction of the model: $X_1 \times Y$ projection plane.



(b) Construction of the model: $X_2 \times Y$ projection plane.

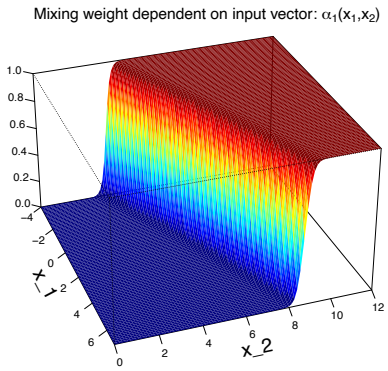


(c) Imputed data: $X_1 \times Y$ projection plane.

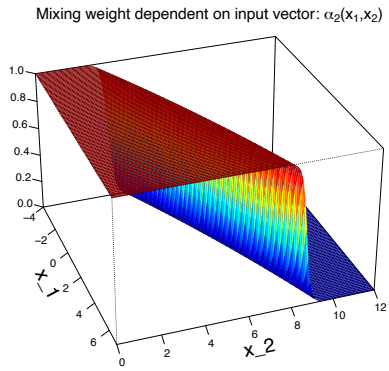


(d) Imputed data: $X_2 \times Y$ projection plane.

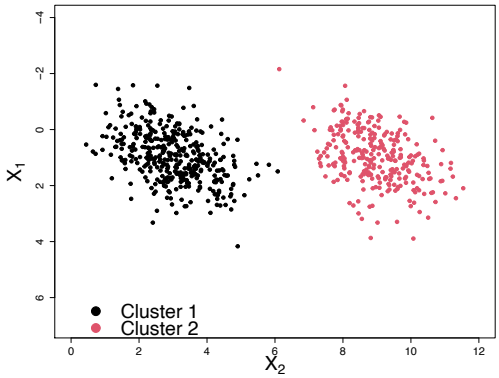
FIGURE A.1: Construction of the univariate imputation model for simulated data considering the vector (X_1, X_2) as auxiliary information.



(a) Surface $\alpha_1(x_1, x_2)$ corresponding to cluster 1.



(b) Surface $\alpha_2(x_1, x_2)$ corresponding to cluster 2.



(c) Scatter plot for the complete data: $X_1 \times X_2$ plane projection

FIGURE A.2: Mixture weights dependent on input vector (X_1, X_2) for the construction of the univariate imputation model.

Appendix B

KL divergence tables for simulated data

To give us an idea of how the performance of the imputation model is when the real distribution is not known, specifically in the case of the examples with real data, below are tables of KL divergence values calculated taking the estimated distribution based on complete data as reference distribution.

	Approach method	
	KL_{int}	Relative distance
$g_{\boldsymbol{Y}_{com}}$	-	-
$g_{\boldsymbol{Y}_{obs}}$	0.0512	1.00
$g_{\boldsymbol{Y}_{X_1}}$	0.0592	1.15
$g_{\boldsymbol{Y}_{X_2}}$	0.0007	0.01
$g_{\boldsymbol{Y}_{(X_1,X_2)}}$	0.0007	0.01

TABLE B.1: KL divergences and relative distances for the imputed variables taking the estimated distribution based on complete data as reference distribution and with information from the input variables X_1, X_2, and (X_1, X_2).

	`cwm`		`pmm`		`norm`	
	KL_{int}	Relative distance	KL_{int}	Relative distance	KL_{int}	Relative distance
$g_{\boldsymbol{Y}_{com}}$	-	-	-	-	-	-
$g_{\boldsymbol{Y}_{obs}}$	0.0512	1.00	0.0512	1.00	0.0512	1.00
$g_{\boldsymbol{Y}_{mean}}$	0.0715	1.40	-	-	-	-
$g_{\boldsymbol{Y}_{X_1}}$	0.0592	1.15	0.1179	2.30	0.1600	3.12
$g_{\boldsymbol{Y}_{X_2}}$	0.0007	0.01	0.0004	0.01	0.0297	0.58
$g_{\boldsymbol{Y}_{(X_1,X_2)}}$	0.0007	0.01	0.0021	0.04	0.0209	0.41

TABLE B.2: Performance of the `mean`, `cwm`, `pmm` and `norm` methods by calculating the KL divergence taking the estimated distribution based on complete data as reference distribution for the univariate simulated data set.

	Approach method	
	KL_{int}	Relative distance
$g_{\boldsymbol{Y}_{com}}$	-	-
$g_{\boldsymbol{Y}_{obs}}$	0.0387	1.00
$g_{\boldsymbol{Y}_{mean}}$	0.2870	7.42
$g_{\boldsymbol{Y}_{cwm}}$	0.0094	0.24
$g_{\boldsymbol{Y}_{pmm}}$	0.0388	1.00
$g_{\boldsymbol{Y}_{norm}}$	0.1905	4.93

TABLE B.3: KL divergences and relative distances taking the estimated distribution based on complete data as reference distribution and for each of the imputation methods in the case of censored missing data.

	cwm		pmm		norm	
	KL_{mc}	Relative distance	KL_{mc}	Relative distance	KL_{mc}	Relative distance
$g_{(Y_1,Y_2)_{com}}$	-	-	-	-	-	-
$g_{(Y_1,Y_2)_{obs}}$	0.0532	1.00	0.0532	1.00	0.0532	1.00
$g_{(Y_1,Y_2)_{mean}}$	0.3509	6.59	-	-	-	-
$g_{(Y_1,Y_2)_{X_1}}$	0.0490	0.92	0.3465	6.51	0.3451	6.49
$g_{(Y_1,Y_2)_{X_2}}$	0.0080	0.14	0.0099	0.18	0.0663	1.24
$g_{(Y_1,Y_2)_{(X_1,X_2)}}$	0.0050	0.09	0.0227	0.43	0.0552	1.04

TABLE B.4: KL divergence and relative distance for the estimated distributions of the observed and imputed databases taking the estimated distribution based on complete data as reference distribution and considering the four models to be compared in the multivariate case.

Appendix C

`mean`, `norm`, and `pmm` imputation procedures.

In the case of the analysis made on the different variables that enter the imputation procedure under the Gaussian LCWM in Sections 4.2 and 5.2, a comparison is made with respect to various imputation methodologies considering the same scenarios. The graphs that illustrate the results of the imputation process are presented below for the univariate and multivariate cases. The average methodology that does not consider the use of variables with auxiliary information is included.

C.1 Imputation with `mean`, `norm`, and `pmm` for the univariate case

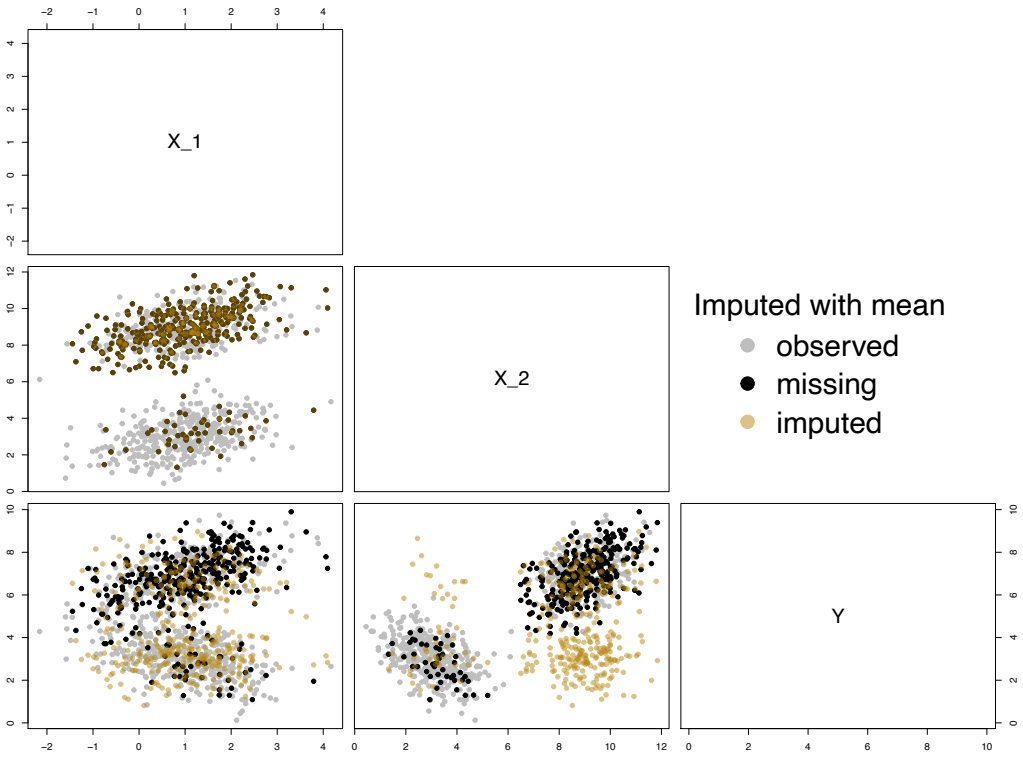


FIGURE C.1: Pair graphs of the univariate database: Simulation 1. Imputed with `mean`.

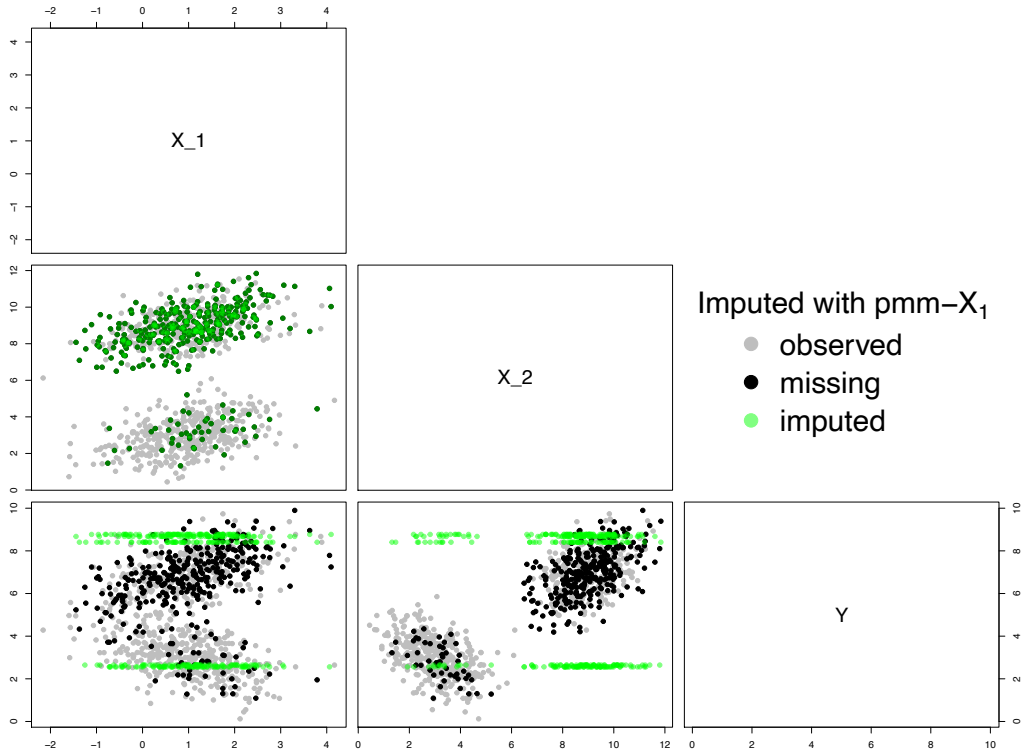


FIGURE C.2: Pair graphs of the univariate database: Simulation 1. Imputed with `pmm` and information from X_1.

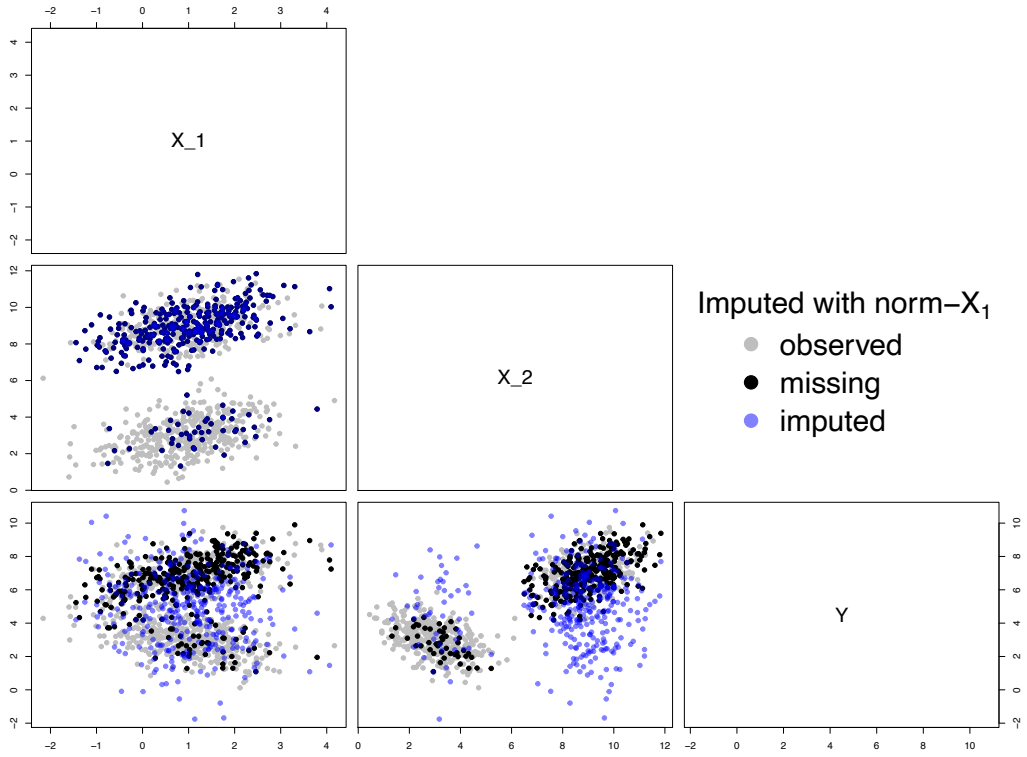


FIGURE C.3: Pair graphs of the univariate database: Simulation 1. Imputed with `norm` and information from X_1.

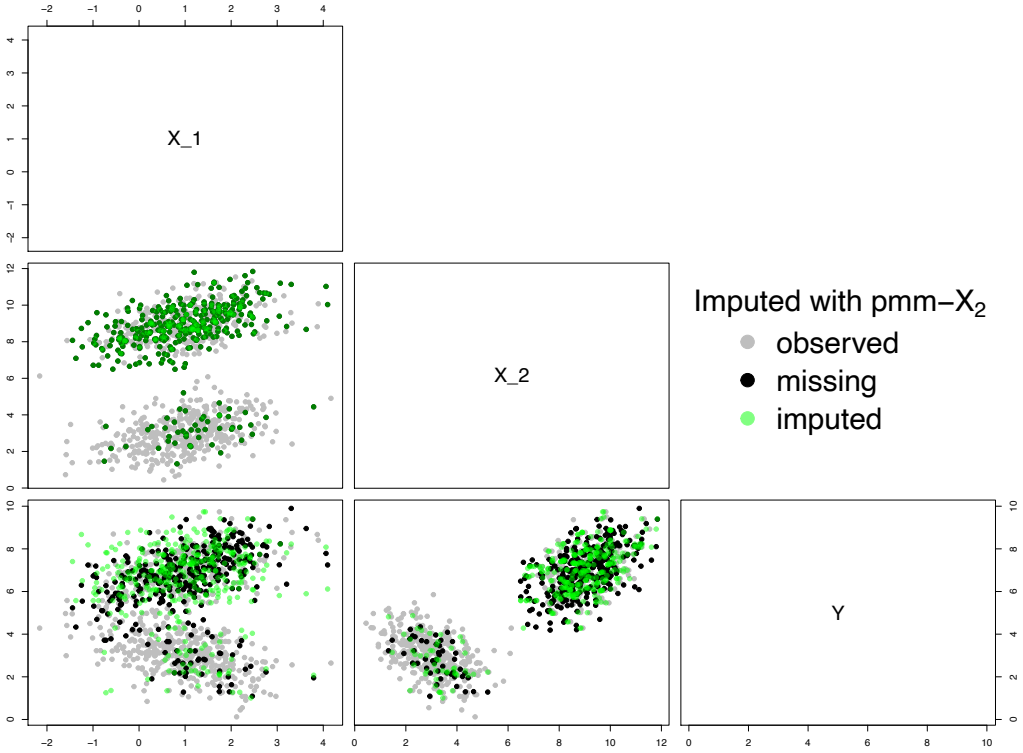


FIGURE C.4: Pair graphs of the univariate database: Simulation 1. Imputed with pmm and information from X_2.

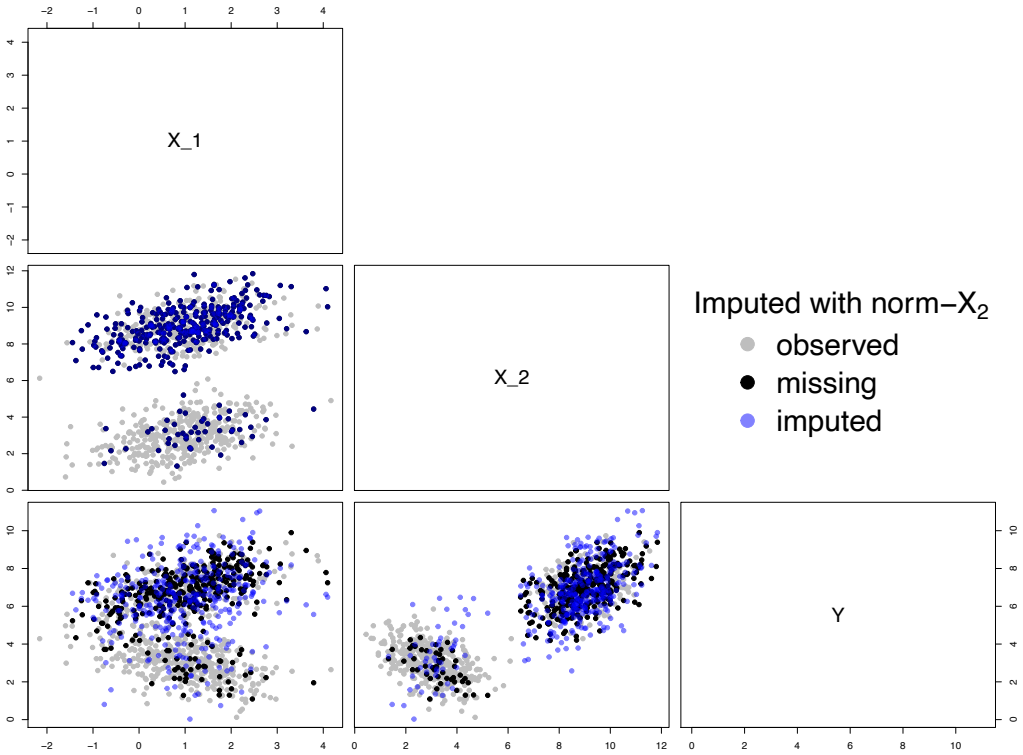


FIGURE C.5: Pair graphs of the univariate database: Simulation 1. Imputed with norm and information from X_2.

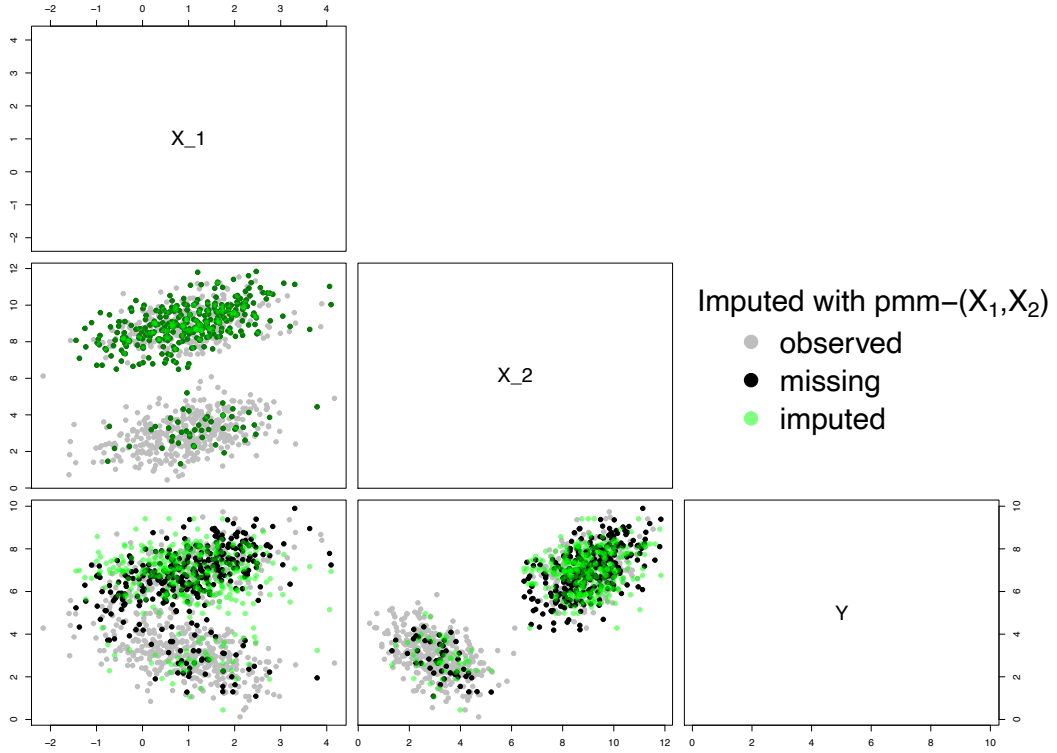


FIGURE C.6: Pair graphs of the univariate database: Simulation 1. Imputed with `pmm` and information from (X_1, X_2).

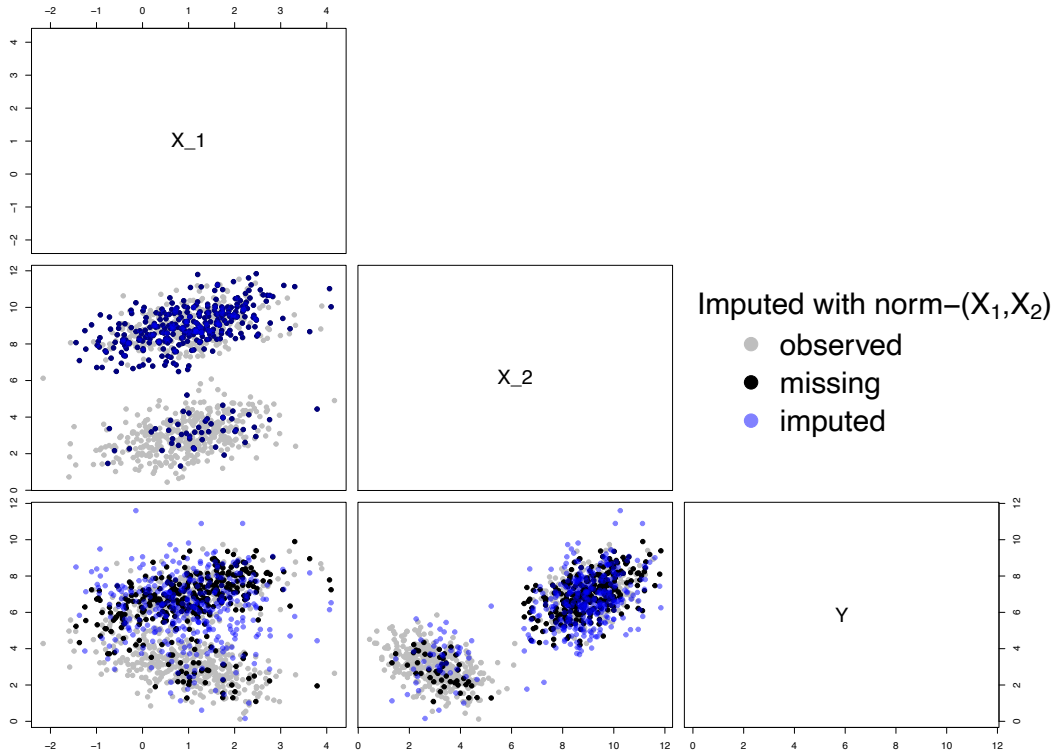


FIGURE C.7: Pair graphs of the univariate database: Simulation 1. Imputed with `norm` and information from (X_1, X_2).

C.2 Imputation with mean, norm and pmm for the multivariate case

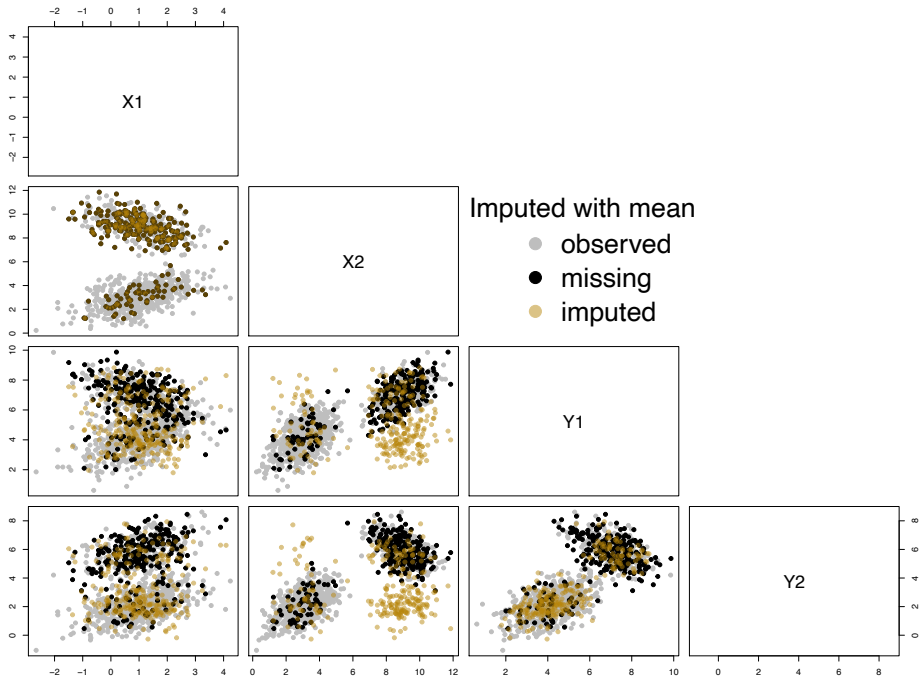


FIGURE C.8: Pair graphs of the multivariate database. Imputed with mean.

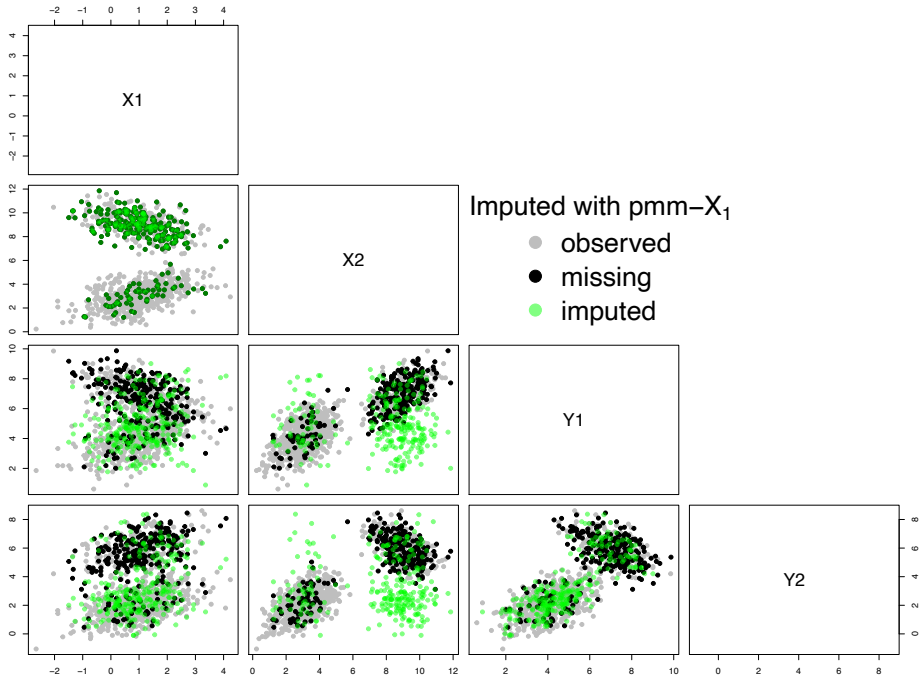


FIGURE C.9: Pair graphs of the multivariate database. Imputed with pmm and information from X_1.

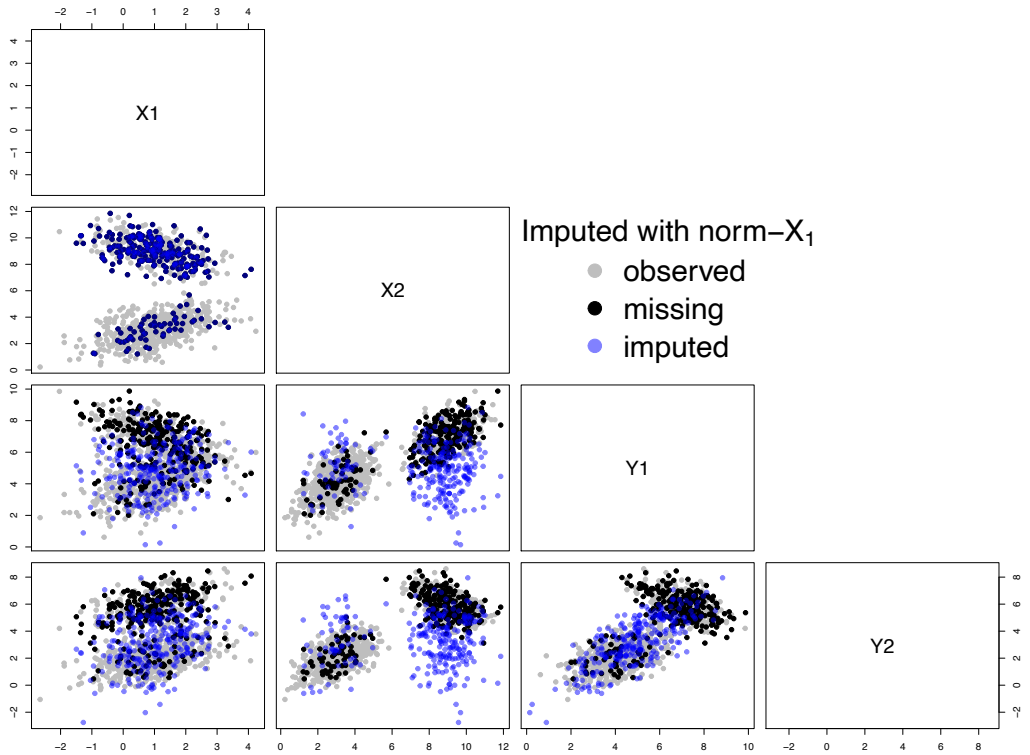


FIGURE C.10: Pair graphs of the multivariate database. Imputed with `norm` and information from X_1.

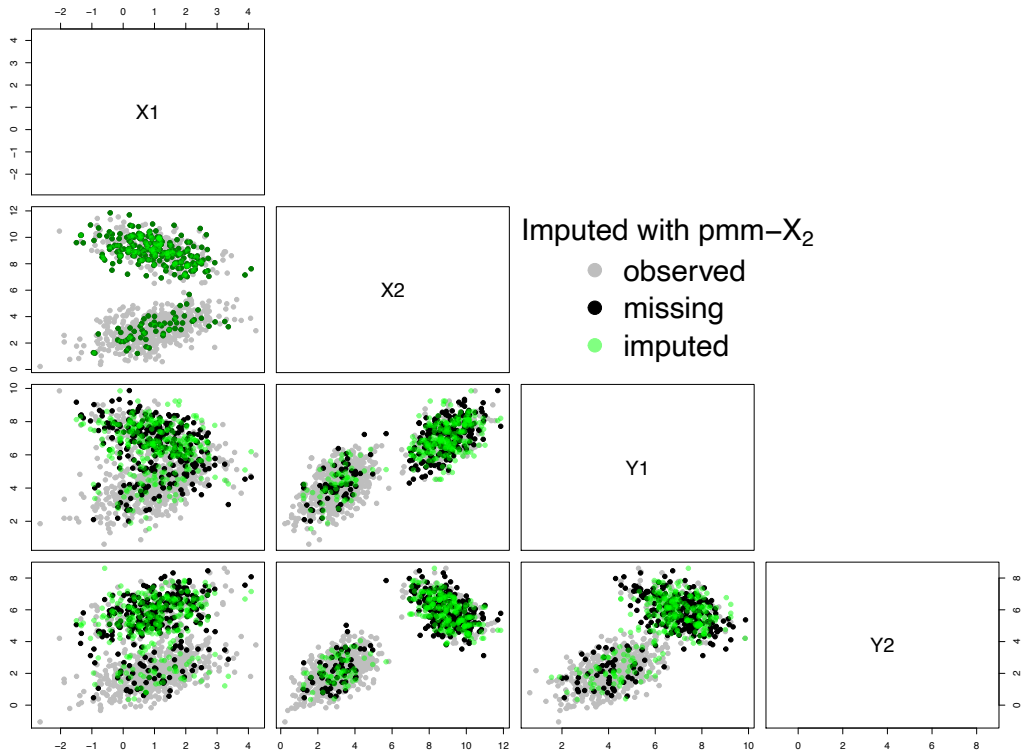


FIGURE C.11: Pair graphs of the multivariate database. Imputed with `pmm` and information from X_2.

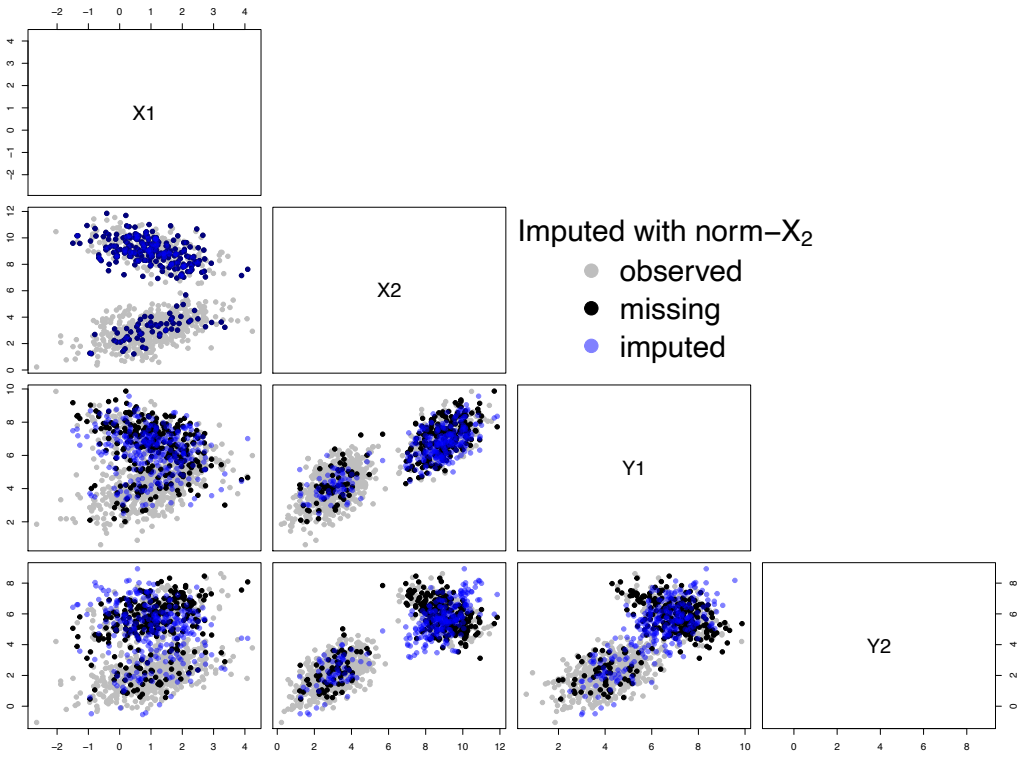


FIGURE C.12: Pair graphs of the multivariate database. Imputed with norm and information from X_2.

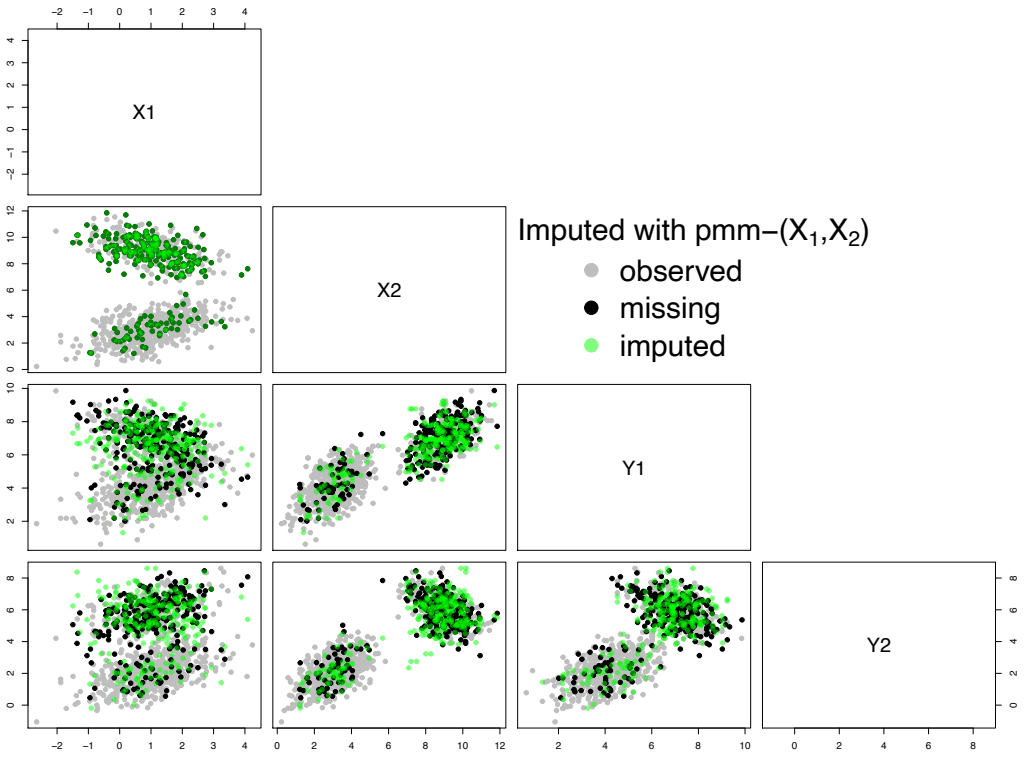


FIGURE C.13: Pair graphs of the multivariate database. Imputed with pmm and information from (X_1, X_2).

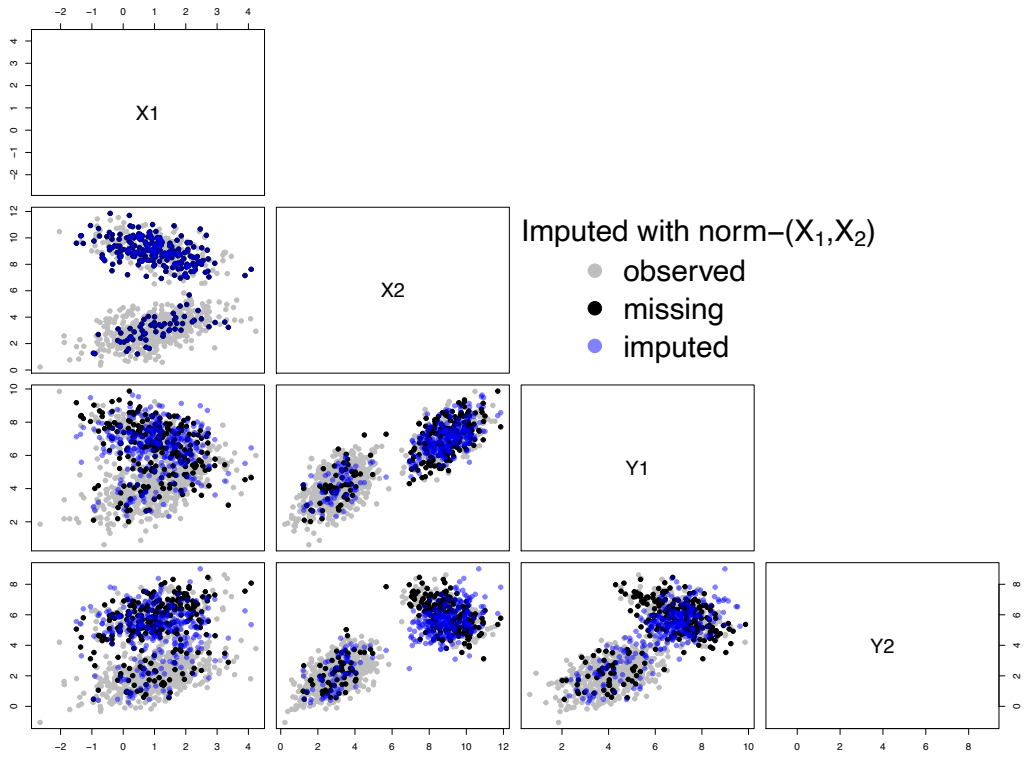
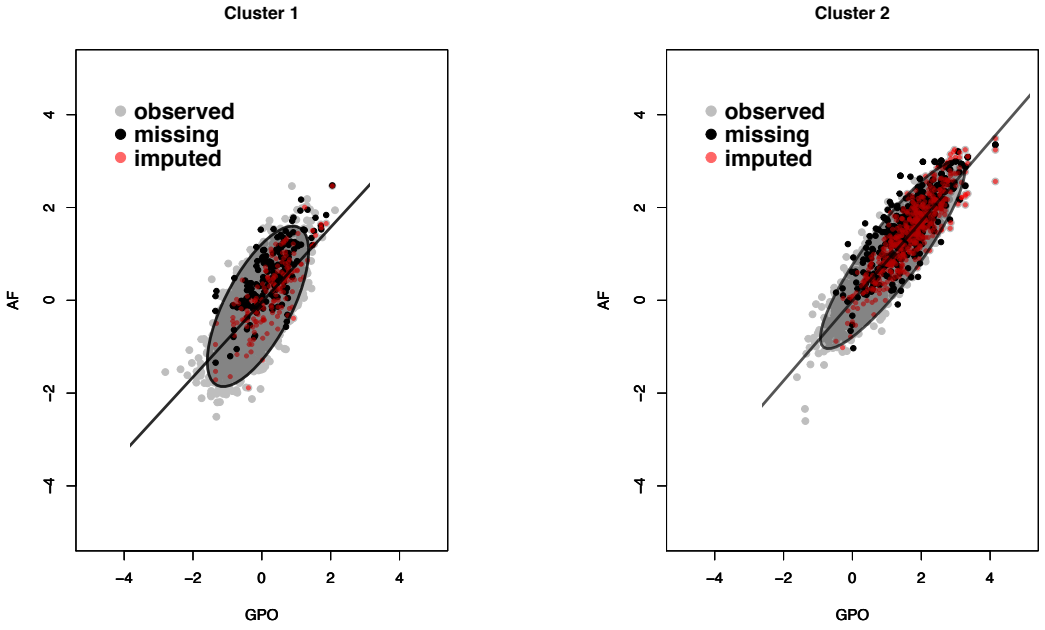


FIGURE C.14: Pair graphs of the multivariate database. Imputed with `norm` and information from (X_1, X_2).

Appendix D

EAM database imputation

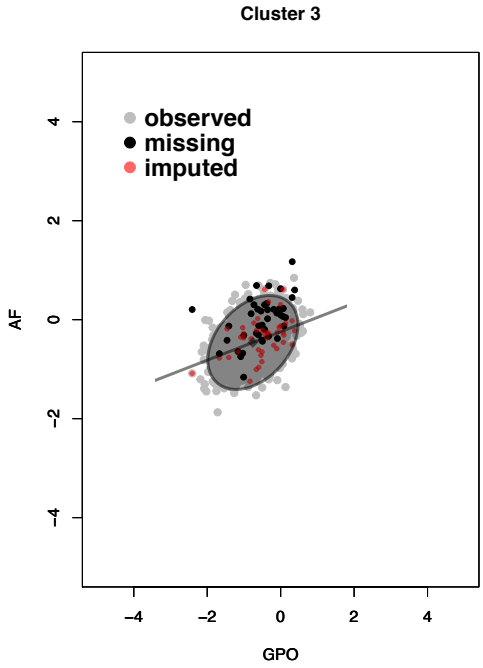
In Section 4.3, we present applications with real data for the Gaussian LCWM in the univariate case. Regarding the application for the data of the Annual Manufacturing Survey in Colombia, the following graphs illustrate the construction of the model discriminated by clusters. In each graph the observed, missing and imputed data, the 95% quantile ellipse and the regression line are presented. The last graph illustrates the curves for the mixture weights dependent on the input variable GPO.



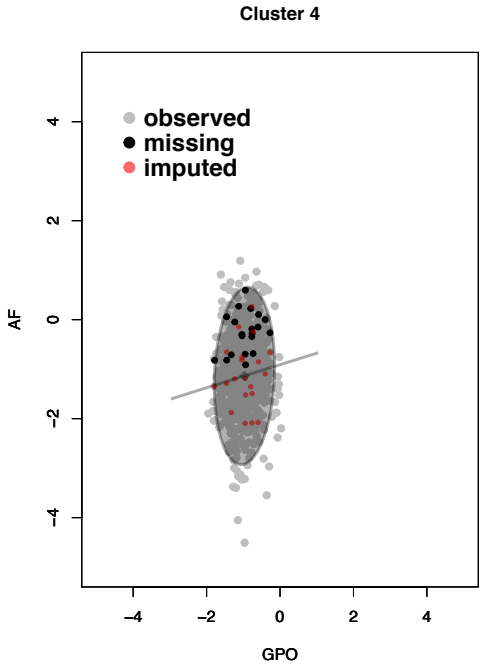
(a) Construction of the model: Cluster 1

(b) Construction of the model: Cluster 2

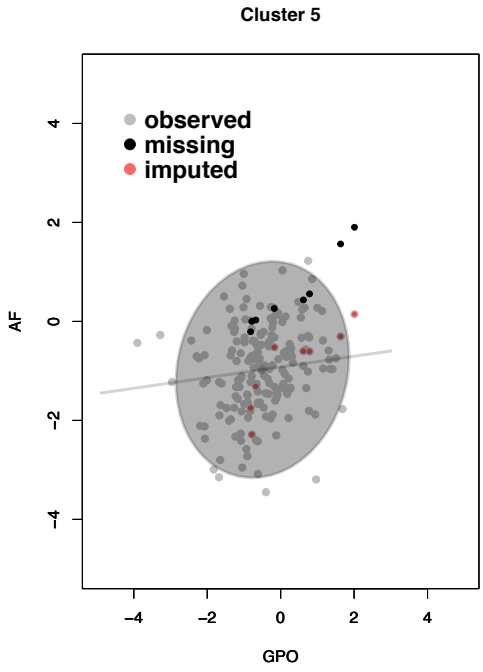
FIGURE D.1: Construction of the imputation model for EAM database: Graphics for clusters 1 and 2.



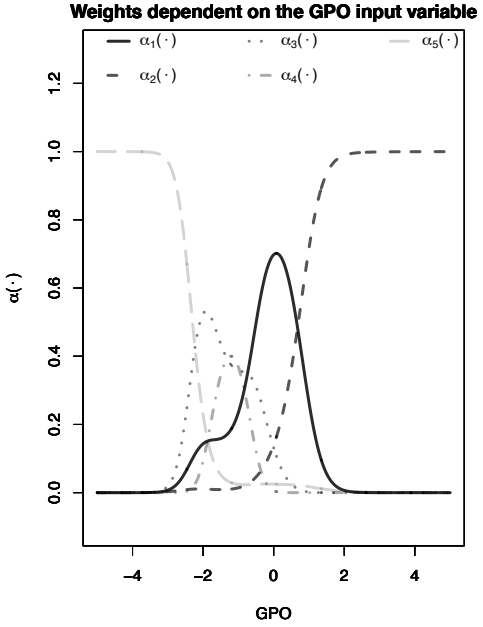
(a) Construction of the model: Cluster 3



(b) Construction of the model: Cluster 4



(c) Construction of the model: Cluster 5



(d) Construction of the model: Weights dependent on the GPO variable

FIGURE D.2: Construction of the imputation model for EAM database: Graphics for clusters 3 to 5 and mixture weights.

Appendix E

Iris database imputation

Section 5.3 presents the application of the Gaussian LCWM in the multivariate case. For that, the Iris database is used and with it, it was simulated two missing data patterns, one under the MAR mechanism and the other under the MNAR mechanism. Pairwise graphs for the `mean`, `norm` and `pmm` imputation procedures are presented below in the two cases of simulated missing data patterns.

E.1 Iris database imputation with missing data MAR mechanism

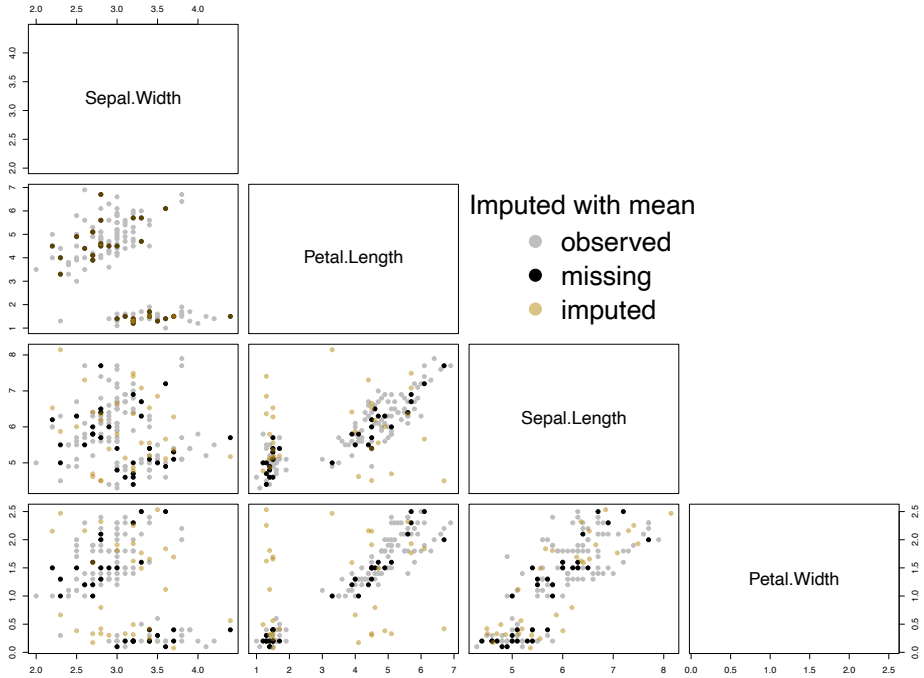


FIGURE E.1: Pairwise plot of the imputed `iris` database using `mean` method. Each panel presents the crossing of two of the variables specifying observed values, missing values and imputed values. Missing data generated using a MAR mechanism.

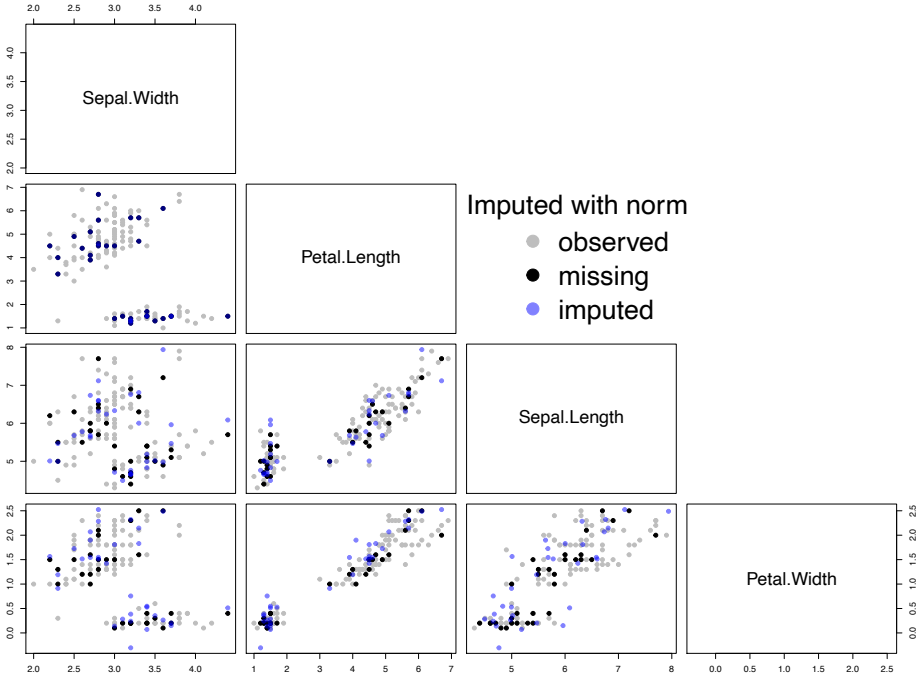


FIGURE E.2: Pairwise plot of the imputed `iris` database using `norm` method. Each panel presents the crossing of two of the variables specifying observed values, missing values and imputed values. Missing data generated using a MAR mechanism.

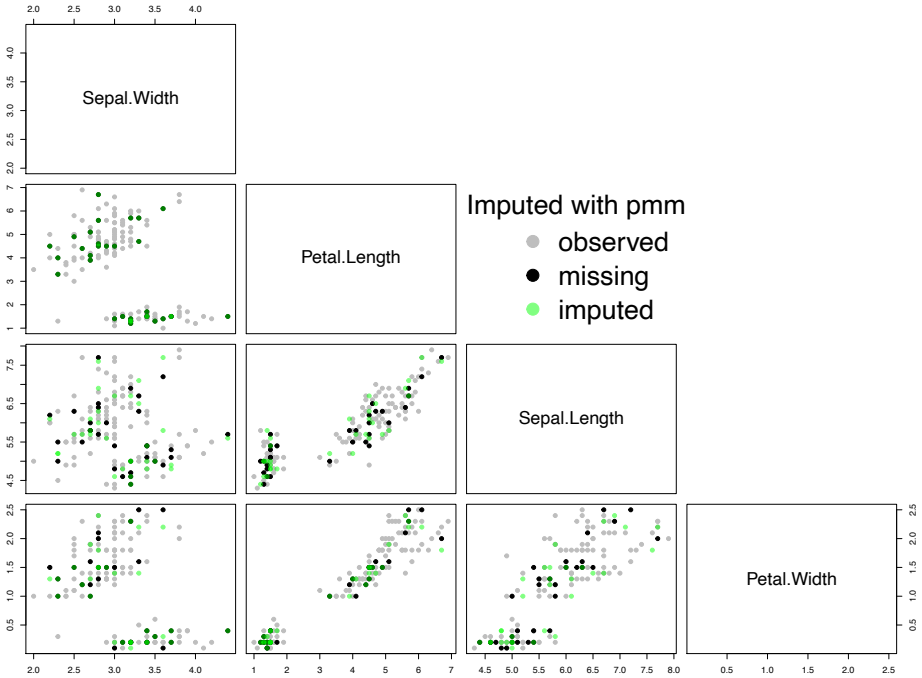


FIGURE E.3: Pairwise plot of the imputed `iris` database using `pmm` method. Each panel presents the crossing of two of the variables specifying observed values, missing values and imputed values. Missing data generated using a MAR mechanism.

E.2 Iris database imputation with missing data MNAR mechanism

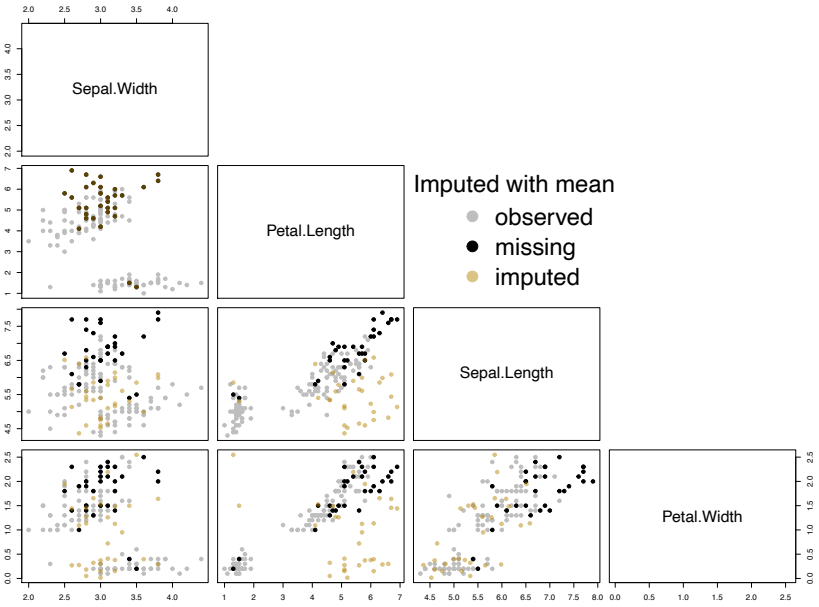


FIGURE E.4: Pairwise plot of the imputed `iris` database using `mean` method. Each panel presents the crossing of two of the variables specifying observed values, missing values and imputed values. Missing data generated using a MNAR mechanism.

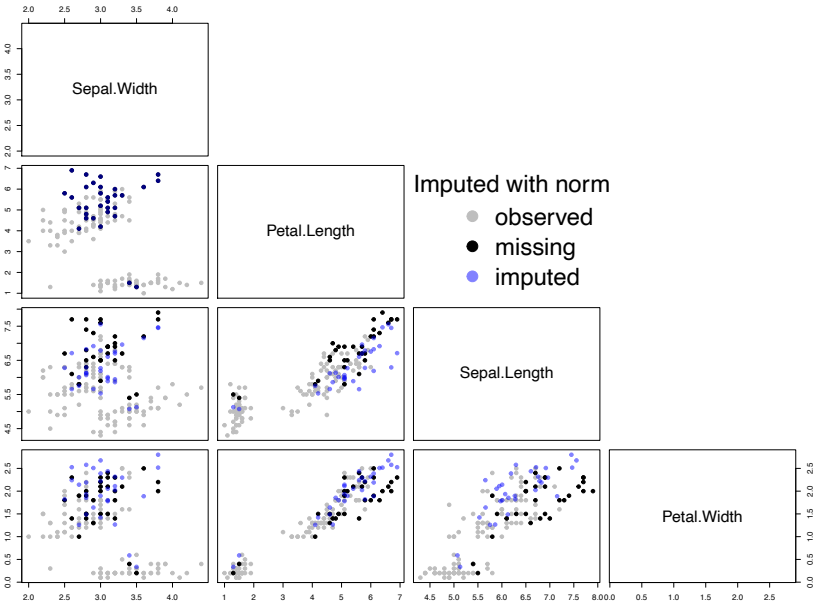


FIGURE E.5: Pairwise plot of the imputed `iris` database using `norm` method. Each panel presents the crossing of two of the variables specifying observed values, missing values and imputed values. Missing data generated using a MNAR mechanism.

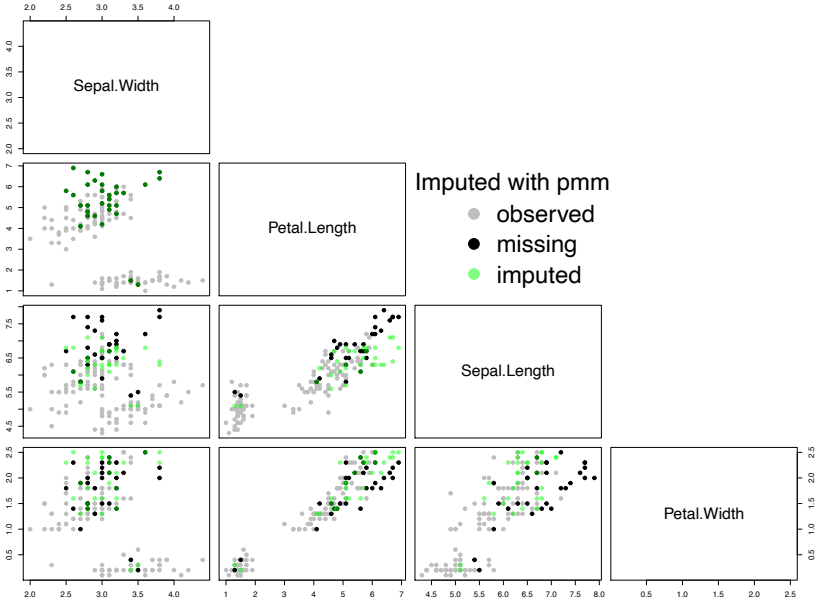


FIGURE E.6: Pairwise plot of the imputed `iris` database using `pmm` method. Each panel presents the crossing of two of the variables specifying observed values, missing values and imputed values. Missing data generated using a MNAR mechanism.

Bibliography

Anderson, Edgar (1935). "The irises of the Gaspe Peninsula". In: *Bull. Am. Iris Soc.* 59, pp. 2–5.

Benaglia, Tatiana, Didier Chauveau, David Hunter, and Derek Young (2009). "mixtools: An R package for analyzing finite mixture models". In: *Journal of Statistical Software* 32.6, pp. 1–29.

Berg, Nathan (2005). "Non-response bias". In: *Encyclopedia of social measurement* 2, pp. 865–873.

Berta, Paolo, Salvatore Ingrassia, Antonio Punzo, and Giorgio Vittadini (2016). "Multilevel cluster-weighted models for the evaluation of hospitals". In: *Metron* 74.3, pp. 275–292.

Dang, Utkarsh J., Antonio Punzo, Paul D. McNicholas, Salvatore Ingrassia, and Ryan P. Browne (2017). "Multivariate response and parsimony for Gaussian cluster-weighted models". In: *Journal of Classification* 34.1, pp. 4–34.

Dempster, Arthur P., Nan M. Laird, and Donald B. Rubin (1977). "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1, pp. 1–22.

Durrieu, J-L, J-Ph Thiran, and Finnian Kelly (2012). "Lower and upper bounds for approximation of the Kullback-Leibler divergence between Gaussian mixture models". In: *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Ieee, pp. 4833–4836.

Enders, Craig K. (2010). *Applied missing data analysis*. Guilford press.

Ferguson, Thomas S. (1973). "A Bayesian analysis of some nonparametric problems". In: *The annals of statistics*, pp. 209–230.

Fisher, Ronald A. (1936). "The use of multiple measurements in taxonomic problems". In: *Annals of eugenics* 7.2, pp. 179–188.

Fraley, Chris and Adrian E. Raftery (2007). "Bayesian regularization for normal mixture estimation and model-based clustering". In: *Journal of classification* 24.2, pp. 155–181.

Frühwirth-Schnatter, Sylvia (2006). *Finite mixture and Markov switching models*. Springer Science & Business Media.

Gelman, Andrew, Jessica Hwang, and Aki Vehtari (2014). "Understanding predictive information criteria for Bayesian models". In: *Statistics and computing* 24.6, pp. 997–1016.

Gershenfeld, Neil A. (1997). "Nonlinear Inference and Cluster-Weighted Modeling". In: *Annals of the New York Academy of Sciences* 808.1, pp. 18–24.

— (1999). *The nature of mathematical modeling*. Cambridge university press.

Härdle, Wolfgang Karl et al. (1991). *Smoothing techniques: with implementation in S*. Springer Science & Business Media.

Hershey, John R. and Peder A. Olsen (2007). "Approximating the Kullback Leibler divergence between Gaussian mixture models". In: *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*. Vol. 4. IEEE, pp. IV–317.

Hoshikawa, Toshiya (2013). "Mixture regression for observational data, with application to functional regression models". In: *arXiv preprint arXiv:1307.0170*.

Ingrassia, Salvatore, Simona C. Minotti, and Giorgio Vittadini (2012). "Local statistical modeling via a cluster-weighted approach with elliptical distributions". In: *Journal of classification* 29.3, pp. 363–401.

Ingrassia, Salvatore, Antonio Punzo, Giorgio Vittadini, and Simona C. Minotti (2015). "Erratum to: The generalized linear mixed cluster-weighted model". In: *Journal of Classification* 32.2, pp. 327–355.

Johnson, Richard A., Dean W. Wichern, et al. (2002). *Applied multivariate statistical analysis*. Vol. 5. 8. Prentice hall Upper Saddle River, NJ.

Kim, Hang J., Lawrence H. Cox, Alan F. Karr, Jerome P. Reiter, and Quanli Wang (2015). "Simultaneous edit-imputation for continuous microdata". In: *Journal of the American Statistical Association* 110.511, pp. 987–999.

Kim, Hang J., Jerome P. Reiter, Quanli Wang, Lawrence H. Cox, and Alan F. Karr (2014). "Multiple imputation of missing or faulty values under linear constraints". In: *Journal of Business & Economic Statistics* 32.3, pp. 375–386.

Kullback, Solomon and Richard A. Leibler (1951). "On information and sufficiency". In: *The annals of mathematical statistics* 22.1, pp. 79–86.

Little, Roderick J. (1988). "Missing-data adjustments in large surveys". In: *Journal of Business & Economic Statistics* 6.3, pp. 287–296.

Little, Roderick J. and Donald B. Rubin (2019). *Statistical analysis with missing data*. Vol. 793. John Wiley & Sons.

McCullagh, P and JA Nelder (1989). *Generalized linear models*.

McLachlan, Geoffrey, Sharon X. Lee, and Suren I. Rathnayake (2019). "Finite mixture models". In: *Annual review of statistics and its application* 6, pp. 355–378.

McLachlan, Geoffrey and David Peel (2004). *Finite mixture models*. John Wiley & Sons.

Müller, Peter and Riten Mitra (2013). "Bayesian nonparametric inference–why and how". In: *Bayesian analysis (Online)* 8.2.

Nguyen, Hien (2015). "Finite mixture models for regression problems". PhD thesis. The University of Queensland.

Paiva, Thais V. (2014). "Multiple Imputation Methods for Nonignorable Nonresponse, Adaptive Survey Design, and Dissemination of Synthetic Geographies". PhD thesis. Duke University.

Paiva, Thais V. and Jerome P. Reiter (2017). "Stop or continue data collection: A nonignorable missing data approach for continuous variables". In: *Journal of Official Statistics* 33.3, pp. 579–599.

Plummer, Martyn, Nicky Best, Kate Cowles, and Karen Vines (2006). "CODA: Convergence Diagnosis and Output Analysis for MCMC". In: *R News* 6.1, pp. 7–11. URL: https://journal.r-project.org/archive/.

Prates, Marcos O., Barbosa Celso R., and Victor H. Lachos (2013). "mixsmsn: Fitting Finite Mixture of Scale Mixture of Skew-Normal Distributions". In: *Journal of Statistical Software* 54.12, pp. 1–20. URL: http://www.jstatsoft.org/v54/i12/.

Punzo, Antonio and Paul D. McNicholas (2017). "Robust clustering in regression analysis via the contaminated Gaussian cluster-weighted model". In: *Journal of Classification* 34.2, pp. 249–293.

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/.

Raghunathan, Trivellore E., Jerome P. Reiter, and Donald B. Rubin (2003). "Multiple imputation for statistical disclosure limitation". In: *Journal of official statistics* 19.1, p. 1.

Rubin, Donald B. (1976). "Inference and missing data". In: *Biometrika* 63.3, pp. 581–592.

— (1987). *Multiple imputation for nonresponse in surveys*. Vol. 81. John Wiley & Sons.

— (1993). "Statistical disclosure limitation". In: *Journal of official Statistics* 9.2, pp. 461–468.

Schafer, Joseph L. (1997). *Analysis of incomplete multivariate data*. CRC press.

Sethuraman, Jayaram (1994). "A constructive definition of Dirichlet priors". In: *Statistica sinica*, pp. 639–650.

Tanner, Martin A. and Wing H. Wong (1987). "The calculation of posterior distributions by data augmentation". In: *Journal of the American statistical Association* 82.398, pp. 528–540.

Van Buuren, Stef (2018). *Flexible imputation of missing data*. CRC press.

Van Buuren, Stef and Karin Groothuis-Oudshoorn (2011). "mice: Multivariate Imputation by Chained Equations in R". In: *Journal of Statistical Software* 45.3, pp. 1–67. URL: https://www.jstatsoft.org/v45/i03/.

Watanabe, Sumio and Manfred Opper (2010). "Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory." In: *Journal of machine learning research* 11.12.

Yan, Ting and Richard Curtin (2010). "The relation between unit nonresponse and item nonresponse: A response continuum perspective". In: *International Journal of Public Opinion Research* 22.4, pp. 535–551.

Zhang, Paul (2003). "Multiple imputation: theory and method". In: *International Statistical Review* 71.3, pp. 581–592.