UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

# Bayesian estimation of dynamic mixture models by wavelets

**Flávia Castro Motta**

Dissertação de Mestrado do Programa Interinstitucional de Pós-Graduação em Estatística (PIPGEs)

ICMC USP SÃO CARLOS

**Flávia Castro Motta**

# Bayesian estimation of dynamic mixture models by wavelets

Dissertation submitted to the Institute of Mathematics and Computer Science – ICMC-USP and to the Department of Statistics – DEs-UFSCar – in accordance with the requirements of the Statistics Interagency Graduate Program, for the degree of Master in Statistics. *FINAL VERSION*

Concentration Area: Statistics

Advisor: Prof. Dr. Michel Helcias Montoril

**USP – São Carlos**
**May 2023**

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

Castro Motta, Flávia
C355b Bayesian estimation of dynamic mixture models by wavelets / Flávia Castro Motta; orientador Michel Helcias Montoril. -- São Carlos, 2023.
112 p.

Dissertação (Mestrado - Programa Interinstitucional de Pós-graduação em Estatística) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2023.

1. Mixture problem. 2. Change-point detection. 3. Wavelets. 4. Spike and slab prior. 5. Wavelet empirical Bayes. I. Helcias Montoril, Michel, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

**Flávia Castro Motta**

# Estimação Bayesiana de modelos de mistura dinâmica por ondaletas

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP e ao Departamento de Estatística – DEs-UFSCar, como parte dos requisitos para obtenção do título de Mestra em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística. *VERSÃO REVISADA*

Área de Concentração: Estatística

Orientador: Prof. Dr. Michel Helcias Montoril

**USP – São Carlos**
**Maio de 2023**

**UNIVERSIDADE FEDERAL DE SÃO CARLOS**

Centro de Ciências Exatas e de Tecnologia
Programa Interinstitucional de Pós-Graduação em Estatística

## Folha de Aprovação

Defesa de Dissertação de Mestrado da candidata Flávia Castro Motta, realizada em 20/04/2023.

### Comissão Julgadora:

Prof. Dr. Michel Helcias Montoril (UFSCar)

Prof. Dr. Chang Chiann (IME-USP)

Prof. Dr. Widemberg da Silva Nobre (UFPR)

O Relatório de Defesa assinado pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa Interinstitucional de Pós-Graduação em Estatística.

*I dedicate this work to my grandparents Aguimar, Magda, Daniel and Sebastião.*

# ACKNOWLEDGEMENTS

At a time when I didn't have much confidence, my mother, Leila, was the one to push me forward and give me the courage that I needed to complete this work. Words are not enough to express how thankful I am for having her as my mother and best friend. Her willingness and wisdom inspire me to live more bravely and generously. Likewise, I am very grateful to my father, Flávio, for his support and love throughout my life. His courage and vision taught me the importance of dreaming big. I also had the absurd luck of having the most sincere and caring human being as a brother. I am so thankful to Conrado for helping me to be kinder and more patient. My godmother, grandparents, uncles, and cousins are also crucial to my life. I am lucky to have the love and support of them all.

I am beyond grateful to my boyfriend, Mateus, whose love and friendship introduced me to a new kind of happiness I never knew was possible. I am so grateful for having such a brilliant person by my side, loving me unconditionally, and making me laugh harder than anyone in the world.

My most sincere gratitude to my advisor, Michel Montoril, who carefully guided and pushed me to learn challenging topics. Thank you for your friendship, patience, and encouragement to improve this work. I could not have asked for a more solicitant advisor. I would also like to extend my gratitude to the professors Andressa Cerqueira, Daiane Zuanetti, Rafael Izbicki, and Rafael Stern, for their remarkable ability to teach and inspire their students.

Throughout my life, I made invaluable friendships that filled me with courage and passion. Among those friends, I particularly appreciate the love and support given by Abritta, Arturo, Ana Victoria, Camila, Carol, Cristiane, Giovanna, Jota, Júlia, and Reis. I also want to acknowledge the friends I have made during my Master's degree. Thanks to the amazing Ana Fernanda, who brought laughter and warmth to our studies together. I am so thankful for her invaluable guidance and support. And to all my other friends from PIPGES, Lubem, Maria Luiza, Rodrigo, and Wellington; I had the best moments studying and talking to you.

My special thanks go to CAPES (Coordenação de Aperfeiçoamento Pessoal de Nível Superior), for providing me the financial support to complete this work. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

Finally, I would like to thank God for giving me His love and protection, and for granting me the intellect and academic tenacity that helped me complete this work.

*“Or, rather, let us be more simple and less vain.*
*Let us limit ourselves to the first sentiments that*
*we find in ourselves, since study always leads us*
*back to them when it has not led us astray.”*
*(Jean-Jacques Rousseau, Emile, or On Education)*

# RESUMO

MOTTA, F. C. **Estimação Bayesiana de modelos de mistura dinâmica por ondaletas**. 2023. 114 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2023.

Modelos de mistura gaussiana são usados com sucesso em várias aplicações de aprendizado estatístico. Os bons resultados fornecidos por esses modelos incentivam diversas generalizações destes. Entre as possíveis adaptações, pode-se supor um comportamento dinâmico para os pesos da mistura para tornar o modelo mais adaptável a diferentes conjuntos de dados. Ao estimar esse comportamento dinâmico, bases de ondaletas surgem como uma alternativa. No entanto, na literatura existente, os métodos baseados em ondaletas apenas estimam os pesos dinâmicos da mistura, não fornecendo estimativas para os parâmetros das componentes do modelo. Neste trabalho, propomos duas abordagens baseadas em ondaletas ortonormais para estimar o comportamento dinâmico do peso da mistura sob algoritmos MCMC eficientes que nos permitem estimar os parâmetros das componentes a partir de suas amostras posteriores. Usamos conjuntos de dados simulados e reais para ilustrar o desempenho de ambas as abordagens. Os resultados indicam que os métodos propostos são alternativas promissoras e computacionalmente eficientes para estimar misturas gaussianas dinâmicas.

**Palavras-chave:** Problema de mistura; Detecção de Ponto de Mudança; Ondaletas; Priori spike e slab; Bayes empírico em ondaletas.

# ABSTRACT

MOTTA, F. C. **Bayesian estimation of dynamic mixture models by wavelets**. 2023. 114 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2023.

Gaussian mixture models are used successfully in various statistical learning applications. The good results provided by these models encourage several generalizations of them. Among possible adaptations, one can assume a dynamic behavior for the mixture weights to make the model more adaptive to different data sets. When estimating this dynamic behavior, wavelet bases have emerged as an alternative. However, in the existing literature, the wavelet-based methods only estimate the dynamic mixing probabilities, failing to provide estimates for the component parameters of the mixture model. In this work, we propose two approaches based on orthonormal wavelets to estimate the dynamic mixture weights under efficient MCMC algorithms that allows us to estimate the component parameters from their posterior samples. We use simulated and real data sets to illustrate both approaches' performances. The results indicate that the proposed methods are promising and computationally efficient alternatives for estimating jointly the dynamic weights and the component parameter of two Gaussian mixtures.

**Keywords:** Mixture problem; Change-point detection; Wavelets; Spike and slab prior; Wavelet empirical Bayes.

# CONTENTS


1 INTRODUCTION . . . 19
1.1 History of Mixture Models . . . 19
1.2 Dynamic mixture weights . . . 20
1.3 Wavelets in statistics . . . 21
1.4 Goals and structure . . . 22

2 WAVELET BACKGROUND . . . 23
2.1 What can wavelets offer? . . . 23
2.2 The wavelet series expansion . . . 24
2.3 Daubechies wavelet bases . . . 28
2.4 The discrete wavelet transform . . . 30

3 WAVELET SHRINKAGE . . . 35
3.1 Denoising problem . . . 35
3.2 Bayesian regularization: spike and slab priors . . . 37
3.3 Wavelet Shrinkage examples . . . 40

4 DYNAMIC GAUSSIAN MIXTURE MODEL . . . 47
4.1 Finite mixture models . . . 47
4.2 The dynamic Gaussian mixture model . . . 49
4.3 Full conditional posterior distributions . . . 51
*4.3.1 The label switching problem* . . . *53*
4.4 Estimation of the dynamic mixture weights . . . 53
*4.4.1 Wavelet regression approach* . . . *53*
*4.4.2 Data augmentation approach* . . . *54*
4.5 Why not sample the "hyperparameters"? . . . 57

5 NUMERICAL STUDIES . . . 65
5.1 Monte Carlo simulations . . . 65
5.2 Application to an array Comparative Genomic Hybridization data set 74

6 CONCLUSIONS . . . 77
6.1 Further research . . . 78

BIBLIOGRAPHY . . . . . 79

APPENDIX A PROOFS . . . . . 85
A.1 Sparsity and vanishing moments . . . . . 85

APPENDIX B POSTERIOR DISTRIBUTION OF THE WAVELET COEFFICIENTS . . . . . 87
B.1 Generic calculations . . . . . 87
*B.1.1 Spike and slab prior - Gaussian component* . . . . . *88*
*B.1.2 Spike and slab prior - Laplace component* . . . . . *90*

APPENDIX C CONDITIONAL POSTERIOR DISTRIBUTIONS . . . 93
C.1 Full conditional posterior of $\mu_k$ . . . . . 93
C.2 Full conditional posterior of $\tau_k^2$ . . . . . 94
C.3 Hyperparameters of *spike and slab prior* . . . . . 95

APPENDIX D MCMC CONVERGENCE DIAGNOSTIC PLOTS . . . 97
D.1 MCMC diagnostics . . . . . 97
D.2 Convergence diagnostic plots for artificial data sets . . . . . 99
*D.2.1 Homogeneous behavior* . . . . . *99*
*D.2.2 Parabolic behavior* . . . . . *101*
*D.2.3 Sinusoidal behavior* . . . . . *103*
*D.2.4 Heavisine behavior* . . . . . *105*
*D.2.5 Bumps behavior* . . . . . *107*
*D.2.6 Blocks behavior* . . . . . *109*
D.3 Convergence diagnostic plots for aCGH data set . . . . . 111

CHAPTER

# 1

# INTRODUCTION

In many situations, we want to model data sets that are not well described by a single unimodal distribution. By allocating the data observations in subpopulations, mixture models are a powerful tool for modeling heterogeneity in a cluster analysis context. For example, in Fu *et al.* (2013), a mixture model coupled with a Dirichlet-process prior was used to cluster noisy measurements of gene expression. For a general methodological review about data clustering using mixture models, see Fraley and Raftery (2002). The flexibility of modeling provided by mixture models enable their use in other major areas of statistics, such as latent class analysis, discriminant analysis, survival analysis, among others. For more details and examples, see McLachlan and Peel (2000).

Due to the importance of these models to this dissertation, in the following section, we give a short historical review of finite mixture models. Then, in section 1.2, we introduce the dynamic extension of the model that plays the central role in this study. In section 1.3, we discuss another essential topic for the development of this dissertation: wavelet bases. At last, in section 1.4, we detail the main purposes and the outline of this research.

## 1.1 History of Mixture Models

One of the first analyses involving mixture models is due to the work of the biostatistician Karl Pearson. In Pearson (1894), the author fitted a mixture of two univariate Gaussian densities with different means and variances to a data set formerly studied by Weldon (1893). The data set consisted of measurements on the ratio of forehead to body length of 1000 crabs sampled from the Bay of Naples. By applying the method of moments to fit the mixture model, Pearson's approach suggested that the asymmetry verified in the data histogram was due to the presence of two different subspecies in the sample (PEARSON, 1894).

During the next 30 years, the method of moments for the mixture problem would continue

to be applied and extended (MCLACHLAN; PEEL, 2000). For example, Charlier and Wicksell (1924) extended the approach to the case of bivariate normal components and Doetsch (1928) considered a model with more than two univariate Gaussian components. However, because of the amount of calculation associated with this approach, various attempts were made over the years to simplify the method (see, e.g., Harding (1948), Preston (1953) and Cassie (1954)).

It was only with the development of high-speed computers that considerable advances were made in mixture models research, in particular with the work by Dempster, Laird and Rubin (1977). In this paper, the authors formalized the Expectation-Maximization (EM) algorithm that enabled an efficient maximum likelihood (ML) estimation of the mixture parameters. Furthermore, this new method provided the theoretical basis for the convergence properties of the ML solution (MCLACHLAN; PEEL, 2000).

The advance of computer technology also played a crucial role in developing the Bayesian approach to the mixture estimation problem. Lavine and West (1992), Diebolt and Robert (1994), Smith and Roberts (1993), Escobar and West (1995) are some of the works that helped to incorporate the concepts of computational Bayesian statistics into the estimation process of mixture models. By using Gibbs sampling and data augmentation methods, these pioneering papers introduced some of the most commonly used Bayesian approaches for obtaining draws from the mixture posterior (FRÜHWIRTH-SCHNATTER, 2006).

With the emergence of different approaches to address the estimation of mixture models, the extent to which they are used has increased considerably. So far, they have been successfully applied to solve problems in astronomy, biology, genetics, medicine, economics, and marketing, among many other areas (FRÜHWIRTH-SCHNATTER; CELEUX; ROBERT, 2018). Furthermore, the possibility of adapting the mixture model to accommodate different data characteristics is another factor that contributes to its applicability in various fields. In the next section, we introduce one possible extension: the dynamic mixture.

## 1.2 Dynamic mixture weights

Among possible adaptations of mixture models, one can assume a dynamic behavior for the mixture weights. This extension is familiar to the generalization of Hidden Markov Models (HMM) to incorporate a "non-homogeneous" structure for the transition probabilities, first described by Hughes and Guttorp (1994). In both scenarios, unobserved probabilities are allowed to vary to make the models more adaptive to different data sets, specially when the observations are indexed by another information such as time or space.

An often-quoted example of an application where a dynamic structure for mixture weights is needed is quality control problems. In these applications, the probability of the supervised system operating in a failure-free regime is most likely not constant across time (NAGY *et al.*, 2011). However, the applicability of these models is not restricted to system supervision

problems. Mixture models with dynamic weights can be applied across a wide range of fields, from traffic flow studies (see Nagy *et al.* (2011)) to applications in genetics (see Montoril, Pinheiro and Vidakovic (2019) and Montoril, Correia and Migon (2021)).

From a frequentist framework, Montoril, Pinheiro and Vidakovic (2019) uses wavelets to estimate the dynamic weights of a mixture of two random variables. Although their wavelet approaches showed good performance in estimating the dynamic mixture weights, their procedure depends on the assumption of known means and variances for the mixture components. In practice, this might be unrealistic. An alternative is to consider a Bayesian framework and use the Gibbs sampling algorithm to obtain joint estimates from both dynamic weights and component parameters.

In this work, we also study the dynamic mixture using wavelet bases. Several properties of wavelets make them profitable tools when it comes to estimating curves. As a result, wavelets have been thrivingly used to address a vast scope of statistical problems (ABRAMOVICH; BAILEY; SAPATINAS, 2000). In the following section, we introduce the main properties of wavelets and provide some references on well-established statistical applications of these mathematical tools.

## 1.3 Wavelets in statistics

The wavelet transform emerged as a synthesis of ideas from multidisciplinary fields, prominently mathematics, physics, and engineering. In general terms, the wavelet transform can provide sparse and informative representations of functions, preserving the local features of these objects. Furthermore, these representations can be very easily obtained through fast algorithms, available in various computer packages (ABRAMOVICH; BAILEY; SAPATINAS, 2000). Due to these properties, wavelet bases have proved to be of significant value to many fields, including statistics.

Among the well-established uses of wavelets within statistical applications, we highlight nonparametric regression (see Donoho and Johnstone (1994) and Cai and Brown (1999)), density estimation (see Donoho (1993); Donoho *et al.* (1996) and Hall and Patil (1995)), survival analysis (see Antoniadis, Gregoire and Nason (1999)), and classification problems (see Chang, Kim and Vidakovic (2003) and Mallat and Hwang (1992)). Wavelet bases also offer significant advantages in time series analysis (see, e.g., Morettin (1996); Priestley (1996); Percival and Walden (1999)). A comprehensive survey of wavelets in statistics can be found in, for instance, Vidakovic (1999) and Ogden (1997).

In this work, wavelets are used to estimate the dynamic weight of a Gaussian mixture model mimicking the nonparametric regression setup. This is made through wavelet shrinkage, which is essentially used to address the denoising problem (see Donoho and Johnstone (1994), Abramovich, Sapatinas and Silverman (1998), Johnstone and Silverman (2005a)). In the next

section, we discuss our central purposes in this research and offer more details about how to use wavelet bases to solve the dynamic mixture estimation problem.

## 1.4 Goals and structure

In this dissertation, the main subject of study is a two-component Gaussian mixture model whose mixture weight is allowed to vary according to some other factor, such as time or space. To address the dynamic structure of the model, we use wavelet bases. This setup is similar to the one approached by Montoril, Pinheiro and Vidakovic (2019), where wavelets are also used within a dynamic mixture model. However, unlike the aforementioned paper, the leading motivation of this research is to provide a Bayesian method capable of jointly estimating the component parameters and the dynamic mixture weights.

Following a Bayesian framework, we implement a straightforward and efficient Gibbs sampling algorithm to carry out the estimation. By giving conjugate prior distributions to the component parameters, we are able to make inference using the distribution of the posterior draws. Regarding the dynamic mixture weights, we implement two wavelet-based approaches within the MCMC algorithm. The first consists of rescaling the original data to obtain a regression setup, where Bayesian wavelet shrinkage techniques can be applied. The second approach adapts the data augmentation method, proposed by Albert and Chib (1993), to efficiently sample from the posterior distribution of the dynamic mixture weights.

The dissertation structure is the following. In Chapter 2, we present some important concepts of wavelet theory, necessary to a better understanding throughout the dissertation. In Chapter 3, we discuss the wavelet shrinkage techniques and their value to statistics denoising applications. We detail the dynamic Gaussian mixture model and the procedures that we implement to estimate the component parameters and the dynamic mixture weights in Chapter 4. In Chapter 5, we present some numerical results based on simulated and real data to illustrate the performance of the proposed methods. Finally, we present some considerations about the results and discuss further directions for this research in Chapter 6. Results concerning Chapter 2, Chapter 3 and Chapter 4 are presented in Appendix A, Appendix B, Appendix C respectively.

CHAPTER

# 2

# WAVELET BACKGROUND

In this second chapter, we introduce basic concepts related to wavelet theory. Our idea is to develop a succinct review of orthonormal wavelet bases, providing a general idea about the wavelet analysis to the reader who may not be conversant with it. We begin by motivating the use of wavelets in section 2.1. In section 2.2, we provide some mathematical exposition related to the construction of orthonormal wavelet bases. In section 2.3, we introduce the Daubechies wavelet families, which correspond to the bases used in this work. At last, in section 2.4, we detail the Discrete Wavelet Transform (DWT) and the pyramidal algorithm used to perform it.

## 2.1 What can wavelets offer?

The word "*wavelets*" means "*small waves*" and, in mathematics, denotes a set of basis functions that represent other functions, signals, and images as a series of successive approximations (HÄRDLE *et al.*, 2012; ABRAMOVICH; BAILEY; SAPATINAS, 2000). The first wavelet basis is due to the work of Haar (1910). However, it was only 70 years later that the mathematical framework needed to develop other wavelet bases was provided (see, e.g., Morlet (1983), Grossmann and Morlet (1984), Meyer (1985), Mallat (1989), Daubechies (1988)). With these contributions, not only new wavelet bases were created, but the whole wavelet theory thrived.

By introducing the concept of multiresolution analysis (MRA) to the wavelet theory, Mallat (1989) was able to design an efficient algorithm to perform the wavelet transform. Another essential contribution was the work by Daubechies (1988), who derived families of orthonormal wavelet bases with compact support. This was an important step for the wavelet theory. Until then, the only orthonormal compactly supported wavelet basis was Haar's, whose discontinuous functions are not appropriate for decomposing smooth signals (VIDAKOVIC, 1999).

Like the Fourier bases, wavelet bases are mathematical tools used in different fields,

from signal processing and numerical analysis to geophysics and astronomy (ABRAMOVICH; BAILEY; SAPATINAS, 2000). However, in contrast to sines and cosines, the oscillations of the wavelet functions are concentrated in a small interval, allowing them to be localized both in frequency and time domain (FARGE, 1992).

The localization property of wavelet functions makes the wavelet transform more advantageous than the Fourier transform to decompose functions that are smooth everywhere, except for a few points. For instance, if the Fourier transform is used to decompose a signal with discontinuities, all Fourier coefficients are affected by these local features. On the other hand, if the wavelet transform is used, the affected coefficients are only those associated with the wavelet functions that overlap those discontinuities (FARGE, 1992).

Besides efficiently capturing local features of signals, wavelet bases are known for providing sparse representations of other functions. This property is among the reasons why wavelets are beneficial tools in a wide range of fields. In statistical applications, for instance, wavelet bases are often used to address the denoising problem. In Chapter 3, we discuss in more detail this topic.

Before discussing the statistical use of wavelets, it is important to emphasize a few mathematical aspects related to this kind of basis. In the following section, we begin with the concept of multiresolution analysis (MRA). Then, we show how the use of wavelets within the MRA allows us to expand functions through different levels of resolution.

## 2.2 The wavelet series expansion

A multiresolution analysis (MRA) is a sequence of nested closed subspaces $\{V_j\}_{j\in\mathbb{Z}}$ that allows one to make approximations of square integrable real functions (VIDAKOVIC, 1999).

**Definition 1.** A function $f : \mathbb{R} \longrightarrow \mathbb{R}$ is said to be a square integrable real function if their $L_2$-norm is finite

$$||f||_2 = \sqrt{\int_{-\infty}^{\infty} |f(x)|^2 \, dx} < \infty.$$

The set of all square integrable real functions corresponds to a space named $L_2(\mathbb{R})$. This space is endowed with the inner product

$$\langle f, g \rangle = \int_{-\infty}^{\infty} f(x) g(x) \, dx.$$

For a sequence of closed subspaces $\{V_j\}_{j\in\mathbb{Z}}$ to constitute a multiresolution analysis, the following conditions must be satisfied:

1. $V_j \in V_{j+1}, \forall j \in \mathbb{Z}$;
2. $f(x) \in V_j \Leftrightarrow f(2^{-j}x) \in V_0, \forall j \in \mathbb{Z}$;

3. $\overline{\bigcup_{j\in\mathbb{Z}} V_j} = L_2(\mathbb{R})$;

4. $\bigcap_{j\in\mathbb{Z}} V_j = \{0\}$;

5. There exists a function $\varphi(\cdot) \in V_0$ such that $\{\varphi(\cdot - k), k \in \mathbb{Z}\}$ is an orthonormal basis for $V_0$.

The last condition requires the existence of $\varphi(\cdot)$. This function is known as *scaling function*. In a broader context, the condition imposed over the integer-translates of $\varphi(\cdot)$, $\{\varphi_{0k}, k \in \mathbb{Z}\} = \{\varphi(\cdot - k), k \in \mathbb{Z}\}$, is that they must constitute a basis for $V_0$, not necessarily orthonormal (Riesz Basis). However, we restrict our study to orthonormal bases for $V_0$ to focus only on the construction of orthonormal wavelet bases, due to their central role in the estimation process of this work. The definition of orthonormal bases follows.

**Definition 2.** A system of functions $\{\varphi(\cdot - k), k \in \mathbb{Z}\}$, where $\varphi_k \in L_2(\mathbb{R})$, is said to be an *orthonormal system* (ONS) if

$$\langle \varphi_j, \varphi_k \rangle = \delta_{jk},$$

where $\delta_{jk}$ is the Kronecker delta, which is zero, when $j \neq k$, and one, when $j = k$. An ONS $\{\varphi(\cdot - k), k \in \mathbb{Z}\}$ is called *orthonormal basis* (ONB) in a subspace $V$ of $L_2(\mathbb{R})$ if any function $f \in V$ can be represented as

$$f(x) = \sum_{k\in\mathbb{Z}} c_k \varphi_k(x),$$

where the coefficients $c_k$ satisfy $\sum_{k\in\mathbb{Z}} |c_k|^2 < \infty$.

It is worth noticing that the containment hierarchy of an MRA, as shown in Figure 1, is constructed such that the subspaces are attached by dyadic scaling of functions (VIDAKOVIC, 1999). Therefore, the system of dilations and translations of $\varphi(x)$, given by

$$\varphi_{jk}(x) = 2^{j/2}\varphi(2^j x - k), \qquad j,k \in \mathbb{Z},$$

allows us to construct orthonormal bases for every subspace $V_j$.

An important condition imposed on the scaling function is $\int_{\mathbb{R}} \varphi(x)\,dx \neq 0$. Furthermore, since $V_0 \subset V_1$, one can represent $\varphi(x)$ as a linear combination of functions from $V_1$,

$$\varphi(x) = \sum_{k\in\mathbb{Z}} h_k \sqrt{2}\varphi(2x - k), \tag{2.1}$$

where the sequence of coefficients $\{h_k\}_{k\in\mathbb{Z}}$ is called *wavelet filter* and (2.1) is known as scaling or dilation equation (VIDAKOVIC, 1999). This equation dictates how the scaling function can be constructed as a linear combination of dyadic rescalings of itself and is fundamental for the development of orthonormal wavelet bases.

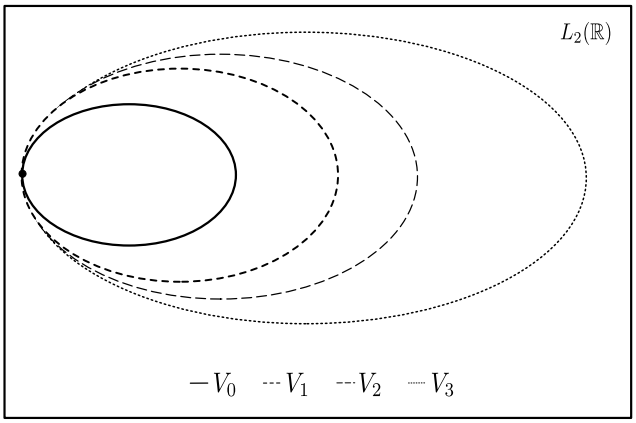



Figure 1 – Illustration of nested subspaces from $V_0$ to $V_3$. The black point represents the null element of space $L_2(\mathbb{R})$.

In an MRA, we can define $W_j$ as the orthogonal complement of $V_j$ in $V_{j+1}$, where

$$V_{j+1} = V_j \oplus W_j, \qquad j \in \mathbb{Z}.$$

Then,

$$V_{j+1} = V_j \oplus W_j = V_{j-1} \oplus W_{j-1} \oplus W_j = \cdots = V_{j_0} \oplus \bigoplus_{l=j_0}^{j} W_l. \tag{2.2}$$

Following the construction of a multiresolution analysis, the union of subspaces $\{V_j\}_{j\in\mathbb{Z}}$ is dense in $L_2(\mathbb{R})$. Therefore, using (2.2), one can obtain

$$L_2(\mathbb{R}) = \overline{V_{j_0} \oplus \bigoplus_{j=j_0}^{\infty} W_j}.$$

This means that every $f \in L_2(\mathbb{R})$ has a unique representation as a convergent series in $L_2(\mathbb{R})$ of the following form

$$f(x) = \sum_{k=-\infty}^{\infty} c_{j_0 k}\varphi_{j_0 k}(x) + \sum_{j=j_0}^{\infty}\sum_{k=-\infty}^{\infty} d_{jk}\psi_{jk}(x), \tag{2.3}$$

where $c_{j_0 k}$'s and $d_{jk}$'s are coefficients of the *multiresolution expansion* of $f$ and $\{\psi_{jk}, k \in \mathbb{Z}\}$ is a general basis for $W_j$. The space $W_j$ is the *resolution level* of an MRA.

In moving from a coarser resolution level $j$ to a finer $j+1$, for $j \in \mathbb{Z}$, we are increasing the resolution at which a function is approximated. However, for (2.3) to be a *wavelet expansion*, $\psi(\cdot)$ must be a wavelet function (HÄRDLE *et al.*, 2012).

**Definition 3.** A function $\psi(x) \in L_2(\mathbb{R})$ is called *wavelet function*, or *mother wavelet*, if it satisfies the *admissibility condition*

$$C_\psi = \int_{-\infty}^{\infty} \frac{|\Psi(\omega)|^2}{|\omega|}\, d\omega < \infty,$$

where $\Psi(\omega)$ is the Fourier transformation of $\psi(x)$. This condition implies that, in the time domain, the average value of $\psi(x)$ must be null,

$$\int_{-\infty}^{\infty} \psi(x)\, dx = 0. \tag{2.4}$$

In words, (2.4) means that $\psi$ has an oscillatory behaviour, which associates it to the name *wavelet*.

According to Vidakovic (1999), a definition of wavelets narrower than that given in Definition 3 is a function whose translations and dyadic dilations,

$$\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k), \qquad j,k \in \mathbb{Z},$$

constitute an orthonormal basis of $L_2(\mathbb{R})$.

Using a wavelet function as $\psi(\cdot)$ in (2.3) transforms the coefficients of the multiresolution expansion into coefficients of a wavelet expansion (HÄRDLE *et al.*, 2012). In this scenario, (2.3) is termed *inhomogeneous wavelet expansion*, $c_{j_0k}$'s are known as *scaling coefficients* and $d_{jk}$'s are called *detail coefficients*. The scaling coefficients capture general aspects of the function, while the detail coefficients give additional and local information about $f(x)$. Notice that the reference space of the expansion (2.3) is $V_{j_0}$, for some $j_0 \in \mathbb{Z}$. Conventionally, one chooses $V_{j_0} = V_0$ (HÄRDLE *et al.*, 2012). We may also consider the *homogeneous wavelet expansion*,

$$f(x) = \sum_{j=-\infty}^{\infty}\sum_{k=-\infty}^{\infty} d_{jk}\psi_{jk}(x),$$

which has no reference space.

Due to the orthonormality of these functions, one can obtain the wavelet coefficients by taking the inner product between $f(x)$ and the functions $\varphi_{j_0k}(x)$ and $\psi_{jk}(x)$, resulting in

$$c_{j_0k} = \langle f, \varphi_{j_0k}\rangle = \int_{-\infty}^{\infty} f(x)\varphi_{j_0k}(x)\,dx, \tag{2.5}$$

$$d_{jk} = \langle f, \psi_{jk}\rangle = \int_{-\infty}^{\infty} f(x)\psi_{jk}(x)\,dx. \tag{2.6}$$

It is important to highlight that, since $\psi(x) \in W_0 \subset V_1$, it can be represented as

$$\psi(x) = \sum_{k\in\mathbb{Z}} g_k\sqrt{2}\varphi(2x-k),$$

where the sequence of coefficients $\{g_k\}_{k\in\mathbb{Z}}$ is a wavelet filter. The coefficients $\{h_k\}$ in (2.1) and $\{g_k\}$ are known in signal processing literature as coefficients to the *quadrature mirror filters*, and they are given by

$$h_k = \sqrt{2}\int_{-\infty}^{\infty} \varphi(x)\varphi(2x-k)\,dx, \tag{2.7}$$

$$g_k = \sqrt{2}\int_{-\infty}^{\infty} \psi(x)\varphi(2x-k)\,dx. \tag{2.8}$$

**Remark 1.** In signal processing literature, $\boldsymbol{h} = \{h_k, k \in \mathbb{Z}\}$ corresponds to coefficients of a low-pass filter and $\boldsymbol{g} = \{g_k, k \in \mathbb{Z}\}$ are coefficients of a high-pass filter. These filters are related by

$$g_k = (-1)^k h_{1-k}. \tag{2.9}$$

The relation (2.9) is known as *quadrature mirror relation*.

## 2.3 Daubechies wavelet bases

One of the simplest orthonormal wavelet basis is the Haar basis, where the scaling and wavelet functions are, respectively,

$$\varphi(x) = \begin{cases} 1, & x \in [0,1], \\ 0 & \text{otherwise.} \end{cases} \qquad \psi(x) = \begin{cases} 1, & x \in [0, \frac{1}{2}), \\ -1, & x \in [\frac{1}{2}, 1), \\ 0 & \text{otherwise.} \end{cases} \tag{2.10}$$

The Haar basis, although orthonormal and compactly supported, is ineffective in approximating smooth functions for several reasons. One of them is the fact that the Haar wavelet function has only one vanishing moment.

**Definition 4.** A function $\psi(x) \in L_2(\mathbb{R})$ has $n$ vanishing moments if

$$\int_{-\infty}^{\infty} x^l \psi(x)\, dx = 0, \qquad l = 0, \ldots, n-1. \tag{2.11}$$

Decomposing a function using a wavelet basis with high-order vanishing moments for its wavelet function $\psi$ may imply a sparse representation. In this case, the detail wavelet coefficients corresponding to regions where the function is very smooth (polynomial regions) are very small or even null (FARGE, 1992), as stated by Proposition 1.

**Proposition 1.** Decomposing any polynomial of degree $m$ (or less) by a wavelet basis whose $\psi$ has $m+1$ vanishing moments results in null detail coefficients.

*Proof.* See Appendix A □

In this perspective, applying orthonormal wavelets, with preassigned number of vanishing moments for $\psi$, brings several advantages in terms of data compression. A key development for the wavelet theory was the work by Daubechies (1988), who constructed compactly supported orthonormal wavelet bases whose wavelet functions have a finite number of vanishing moments. Daubechies (1988) showed that, if the sequence of non-null coefficients $\{h_k\}$ in (2.7) is a Finite Impulse Response (FIR) that satisfies certain conditions, the corresponding wavelet functions generated by it are compactly supported.

**Definition 5.** Let $\boldsymbol{h}$ be a filter and consider the integers $M \leq 0$ and $L > 0$. We say that $\boldsymbol{h}$ is a Finite Impulse Response (FIR) if $h_k = 0$ for $k < M$ or $k > L$, and $h_M, h_L \neq 0$. Then, we can write

$$\boldsymbol{h} = \{h_M, h_{M+1}, \ldots, h_L\}.$$

Daubechies wavelets consist of three families of orthonormal wavelet bases, the *extremal-phase*, also known as *daublets*; the *least-asymmetric*, or *symmlets*, and the *coiflets*. A summary of the properties of each wavelet family is shown in Table 1.

Table 1 – Properties of Daubechies wavelets.

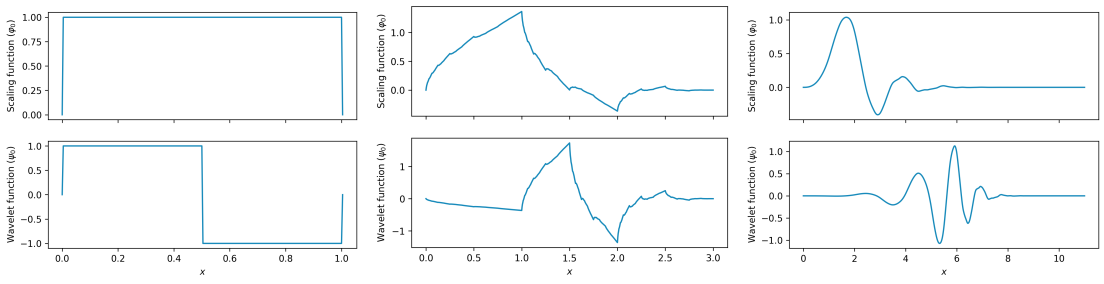
| Property | Daublets | Symmlets | Coiflets |
|---|---|---|---|
| $\psi$ support | $[-N+1, N]$ | $[-N+1, N]$ | $[-N, 2N-1]$ |
| $\varphi$ support | $[0, 2N-1]$ | $[0, 2N-1]$ | $[-N, 2N-1]$ |
| $\int_{-\infty}^{\infty} x^l \psi(x)\, dx = 0$ | $l = 0, \ldots, N-1$ | $l = 0, \ldots, N-1$ | $l = 0, \ldots, 2N-1$ |
| $\int_{-\infty}^{\infty} x^l \varphi(x)\, dx = 0$ | - | - | $l = 1, \ldots, 2N-1$ |

The standard convention is to denote a Daubechies wavelet by its low-pass filter length (VIDAKOVIC, 1999). However, as the number of vanishing moments of the wavelet function, for daublets and symmlets, is half the number of taps of the low-pass filter, these wavelets can also be denoted by the number of vanishing moments they possess. For instance, a daublet with two vanishing moments is frequently called in the literature by D4 or db2. In this work, we use the former standard convention for notation purposes.

## Daublets

Daubechies (1988) constructed 10 classes of daublets, each differing from the others by the number of taps of their low-pass filters and, consequently, by their support width and number of vanishing moments. The coefficients $\{h_k\}$ for daublets with vanishing moments varying from 1 to 10 are tabulated in Daubechies (1992, p. 195).

The daublet whose wavelet function has only one vanishing moment is actually the Haar wavelet. Although discontinuous, the Haar functions are the only Daubechies wavelets that are symmetric and have analytical expression, as shown in (2.10). Haar and other examples of extremal-phase wavelets are presented in Figure 2.

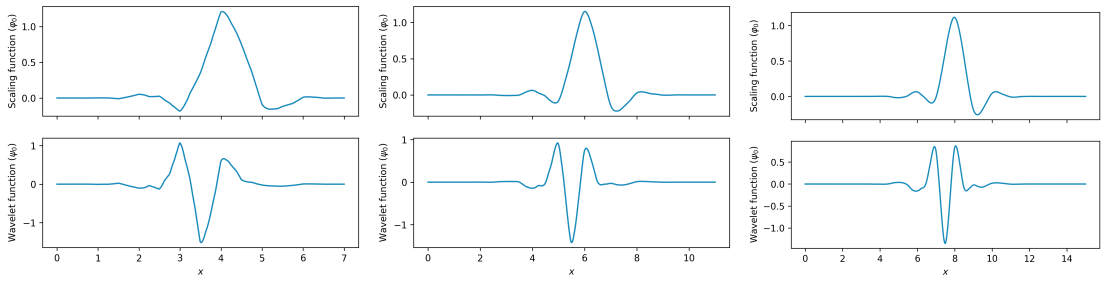



(a) D2. (b) D4. (c) D12.

Figure 2 – Graphs of scaling and wavelet functions of daublets: (a) D2 is a daublet basis (also known as Haar basis) with one vanishing moment; (b) D4 is a daublet basis with two vanishing moments; (c) D12 is a daublet basis with six vanishing moments.

## Symmlets

Daubechies (1988) constructed 7 classes of symmlets. As with daublets, the difference between each symmlet lies in their low-pass filter length and support width. The coefficients




(a) S8. (b) S12. (c) S16.

Figure 3 – Graphs of scaling and wavelet functions of symmlets: (a) S8 is a symmlet basis with four vanishing moments; (b) S12 is a symmlet basis with six vanishing moments; (c) S16 is a symmlet basis with eight vanishing moments.

$\{h_k\}$ for symmlets with vanishing moments varying from 4 to 10 are tabulated in Daubechies (1992, p. 198).

Although symmlets are not symmetric, they were constructed to be as close as possible to symmetry given the support $[0, 2N-1]$. For that reason, Daubechies (1988) called this family of wavelets least-asymmetric. A deeper discussion about the phase properties of these wavelet filters can be found in Percival and Walden (1999, p. 108-116). Figure 3 presents the scaling and wavelet functions for some symmlets.

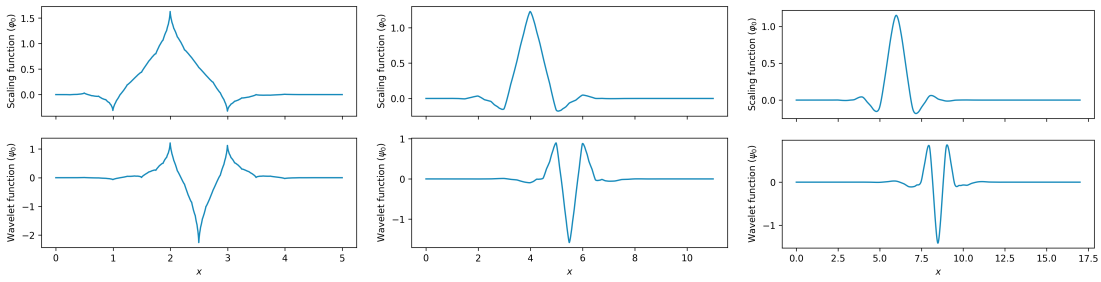
### *Coiflets*

Daubechies named this family of wavelets after R. Coifman, who was the one to suggest that orthonormal wavelet bases with vanishing moments for both wavelet and scaling functions could lead to higher compressibility (DAUBECHIES, 1992; BEYLKIN; COIFMAN; ROKHLIN, 1991). Therefore, in contrast to the daublets and symmlets that have vanishing moments only for the wavelet function, the coiflets were constructed to also have vanishing moments for the scaling function.

Coiflets are even less asymmetric than symmlets. However, because of that, their wavelets have larger support than the symmlets' wavelets (VIDAKOVIC, 1999). The coefficients $\{h_k\}$ for all the 5 coiflets constructed are tabulated in Daubechies (1992, p. 261). Figure 4 presents the scaling and wavelet functions for some wavelets of this family.

## 2.4 The discrete wavelet transform

Up to now, the function space of interest has been $L_2(\mathbb{R})$. However, in many practical situations, the data is sampled over a finite interval, such as the unit interval $[0,1]$ (ABRAMOVICH; BAILEY; SAPATINAS, 2000). In this perspective, transforming a signal within $[0,1]$ to the wavelet domain may require some boundary handling (OGDEN, 1997). To address this condition, a usual approach is to assume that the function that lies within $[0,1]$ is a periodic function with




(a) C6. (b) C12. (c) C18.

Figure 4 – Graphs of scaling and wavelet functions of coiflets: (a) C6 is a coiflet basis with two vanishing moments; (b) C12 is a coiflet basis with four vanishing moments; (c) C18 is a coiflet basis with six vanishing moments.

period one. For more details about different boundary handling approaches see Ogden (1997), Cohen, Daubechies and Vial (1993), and Restrepo and Leaf (1997).

Without loss of generality, consider $f(t) \in L_2([0,1])$ to be the function of interest of some application. In practice, we only have access to $f$ applied to a grid of points in time or space. Thus, let $\boldsymbol{f} = (f(t_1), \ldots, f(t_n))^T$ be a vector of samples of $f(t)$ on a discrete grid of $n$ equally spaced points $t_i$, with $n = 2^J$, for some positive integer $J$. The discrete wavelet transform (DWT) allows us to transform $\boldsymbol{f}$ to the wavelet domain, which means obtaining the scaling and detail coefficients, as in (2.5) and (2.6).

The DWT maps discrete time series from the time domain into the wavelet domain. The outputs of the DWT are a set of wavelet coefficients that preserve the local behavior of the transformed data (VIDAKOVIC, 1999). In matrix notation, the DWT of $\boldsymbol{f}$ is

$$\boldsymbol{\theta} = \boldsymbol{W}\boldsymbol{f},$$

where $\boldsymbol{\theta} = (c_{00}, d_{00}, \boldsymbol{d}_1^T, \ldots, \boldsymbol{d}_{J-1}^T)^T$ is a vector of size $n$, having both scaling and detail coefficients $\boldsymbol{d}_j = (d_{j0}, d_{j1}, \ldots, d_{j2^j-1})^T$, and $\boldsymbol{W}$ is an orthonormal matrix related to the orthonormal wavelet basis chosen to perform the wavelet transformation.

Since $\boldsymbol{W}$ is orthonormal, we can perfectly reconstruct the original vector $\boldsymbol{f}$ from $\boldsymbol{\theta}$, using the inverse discrete wavelet transform (IDWT), that is,

$$\boldsymbol{f} = \boldsymbol{W}^T\boldsymbol{\theta}, \tag{2.12}$$

because $\boldsymbol{W}^T = \boldsymbol{W}^{-1}$, where $\boldsymbol{W}^T$ is the transpose of $\boldsymbol{W}$. The matrix $\boldsymbol{W}$ is constructed by combining the quadrature mirror filters given in (2.7) and in (2.8). In Example 1, we present the transform matrix $\boldsymbol{W}$ related to the Haar basis. For a comprehensive treatment of the construction of the wavelet transform matrix, see, e.g., Fleet (2011).

**Example 1.** Let $\boldsymbol{f} = (1,0,2,-1,0,3,1,2)^T$ be a data set we want to transform by Haar's DWT. Then, the wavelet coefficients can be calculated as

$$\begin{pmatrix} c_{00} \\ d_{00} \\ d_{10} \\ d_{11} \\ d_{20} \\ d_{21} \\ d_{22} \\ d_{23} \end{pmatrix} = \begin{pmatrix} \sqrt{2}/4 & \sqrt{2}/4 & \sqrt{2}/4 & \sqrt{2}/4 & \sqrt{2}/4 & \sqrt{2}/4 & \sqrt{2}/4 & \sqrt{2}/4 \\ \sqrt{2}/4 & \sqrt{2}/4 & \sqrt{2}/4 & \sqrt{2}/4 & -\sqrt{2}/4 & -\sqrt{2}/4 & -\sqrt{2}/4 & -\sqrt{2}/4 \\ 1/2 & 1/2 & -1/2 & -1/2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 & -1/2 & -1/2 \\ 1/\sqrt{2} & -1/\sqrt{2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/\sqrt{2} & -1/\sqrt{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/\sqrt{2} & -1/\sqrt{2} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 2 \\ -1 \\ 0 \\ 3 \\ 1 \\ 2 \end{pmatrix}$$

$$= \begin{pmatrix} 2\sqrt{2} \\ -\sqrt{2} \\ 0 \\ 0 \\ 1/\sqrt{2} \\ 3/\sqrt{2} \\ -3/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix}.$$

The DWT performed by matrix multiplication demands $\mathcal{O}(n^2)$ operations. Therefore, it is not viable when the input vector is very long. Mallat (1989) developed a pyramidal algorithm, also known as cascade algorithm, to perform discrete wavelet transforms more efficiently than by matrix multiplication. Mallat's method takes only $\mathcal{O}(n)$ operations to perform the DWT of a data set of size $n$. The algorithm does this through successive convolution and decimation operations.

**Definition 6.** Let $\boldsymbol{h}$ and $\boldsymbol{a}$ be two bi-infinite sequences. The convolution product $\boldsymbol{y}$ of $\boldsymbol{h}$ and $\boldsymbol{a}$ is a bi-infinite sequence denoted by $\boldsymbol{y} = \boldsymbol{h} \star \boldsymbol{a}$, whose $n$-th component is given by

$$y_n = \sum_{k=-\infty}^{\infty} h_k a_{n-k}.$$

**Corollary 1.** The $n$-th component of a convolution product of a Finite Impulse Response $\boldsymbol{h}$ and an input sequence $\boldsymbol{a}$ (bi-infinite or finite) is a finite sum

$$y_n = \sum_{k=M}^{L} h_k a_{n-k} = h_M a_{n-M} + \cdots + h_{-1} a_{n+1} + h_0 a_n + h_1 a_{n-1} + \cdots + h_L a_{n-L}. \tag{2.13}$$

Mallat's algorithm starts by convolving the original signal with the mirror filters $\boldsymbol{h}$, in (2.7), and $\boldsymbol{g}$, in (2.8). In case of Daubechies wavelets, these filters are FIR. Therefore, the output sequences of the convolution products are given by (2.13). Subsequently, the method discards the sequence values on positions with odd indices. This procedure is called *downsampling* or *decimation* by two.

The outputs are the sequences of detail and scaling coefficients of the finest level $J-1$ of the wavelet decomposition. Then, the algorithm keeps convolving and downsampling the sequence of scaling coefficients until all wavelet coefficients are generated. Figure 5 illustrates the cascade organization of coefficients obtained through this algorithm.

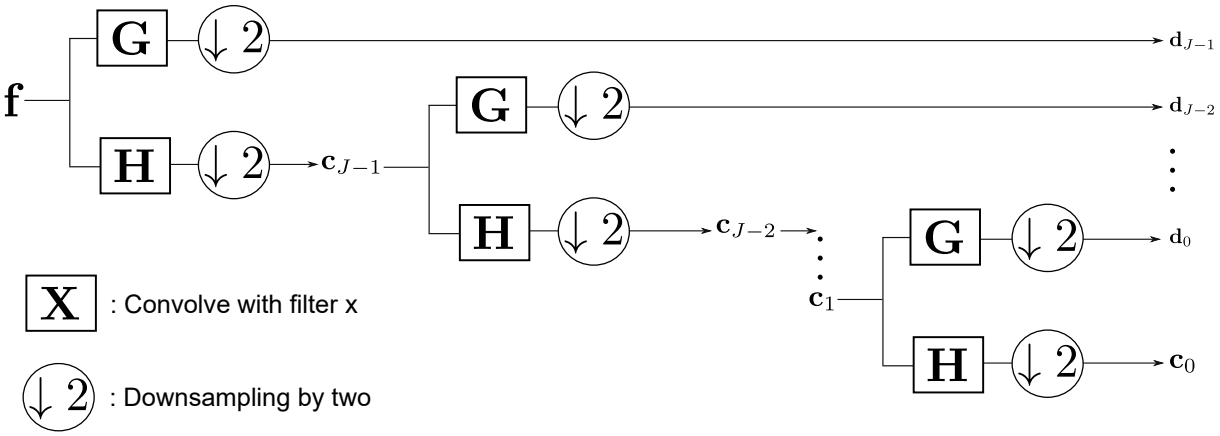



Figure 5 – Illustration of the wavelet decomposition through the cascade algorithm. **H** represents the low-pass filter and **G** represents the high-pass filter.

The cascade algorithm also performs the inverse discrete wavelet transform (IDWT). It first expands the sequences of scaling and detail coefficients of the coarsest level by inserting zeros between each entry. This process is called *upsampling* or *dilation* by two. Subsequently, the algorithm convolves the expanded sequences with the corresponding mirror filters ($\boldsymbol{h}$ for the scaling coefficients and $\boldsymbol{g}$ for the detail coefficients) and sums the outputs.

The result is the sequence of scaling coefficients belonging to the subsequent finer scale. Then, the algorithm keeps upsampling and convolving the sequences of coefficients until the signal is reconstructed from the coefficients of the finest level $J-1$. Figure 6 presents a diagram for the wavelet reconstruction process by the pyramidal algorithm.

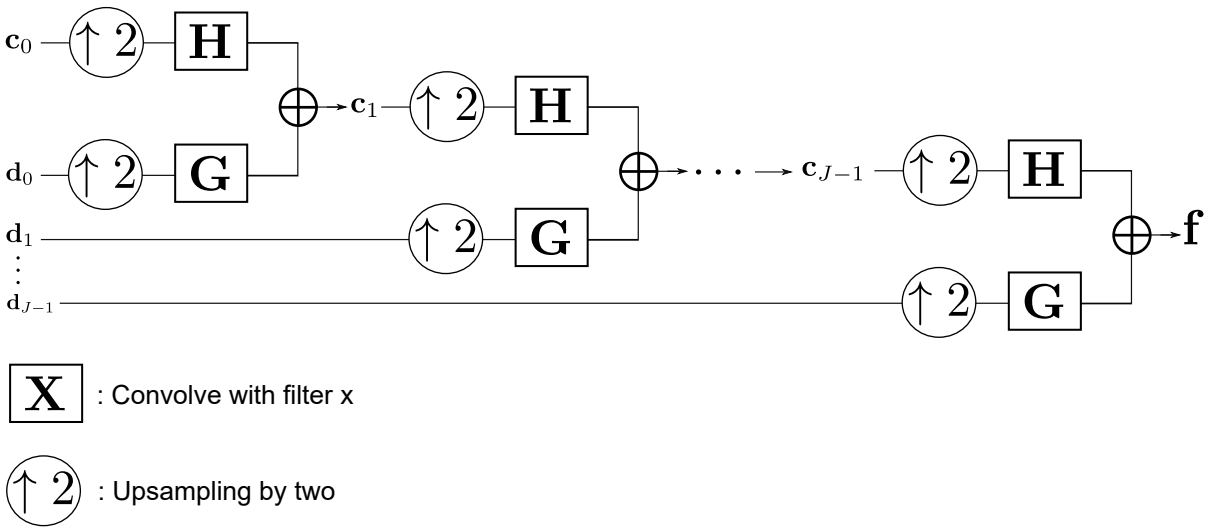



Figure 6 – Illustration of the wavelet reconstruction through the cascade algorithm. **H** represents the low-pass filter and **G** represents the high-pass filter.

CHAPTER

# 3

# WAVELET SHRINKAGE

In this chapter, we present a well-established wavelet procedure used in statistics: wavelet shrinkage. In this approach, wavelets are employed under a problem of nonparametric regression to estimate the regression function. In section 3.1, we introduce this technique, and in section 3.2, we describe the Bayesian framework for wavelet shrinkage. The final section 3.3 shows some examples, on synthetic data sets, of the performance of the presented methods to wavelet shrinkage.

## 3.1 Denoising problem

Consider the nonparametric regression model

$$\boldsymbol{y} = \boldsymbol{f} + \boldsymbol{e}, \tag{3.1}$$

where $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T$ is the vector of observed values, $\boldsymbol{f} = (f(1/n), f(2/n), \ldots, f(n/n))^T$ is the function of interest applied to a grid of equally spaced points, and $\boldsymbol{e} = (e_1, e_2, \ldots, e_n)^T$ is a vector of zero-mean random variables, known in the literature as white noise vector (VIDAKOVIC, 1999). For most applications, unless otherwise specified, $e_t$'s are independent and identically distributed normal random variables with zero mean and constant variance $\sigma^2$.

The goal of nonparametric regression is to recover the unknown function $f$ from the noisy observations $\boldsymbol{y}$. With that in mind, Donoho and Johnstone (1994) proposed a simple method, using wavelets, to estimate $f$. The motivation for this approach, known as wavelet denoising, lies in transforming the noisy observations to the wavelet domain through some orthonormal wavelet basis.

When using an orthonormal basis to perform the DWT of (3.1), the DWT of $\boldsymbol{e}$ is also a vector of independent and identically distributed normal random variables with zero mean and variance $\sigma^2$. Hence, the DWT of $\boldsymbol{y}$ spreads the noise equally over all wavelet coefficients but

concentrates most of the signal related to the function $f$ in a few large coefficients (DONOHO; JOHNSTONE, 1994).

Following this scenario, the authors propose transforming the observations $\boldsymbol{y}$ through the DWT to shrink the noisy wavelet coefficients or even equal them to zero using some threshold. Then, once these coefficients have been shrunk/thresholded, the method applies the IDWT to obtain the estimate of $\boldsymbol{f}$. Let $n$ be a power of two, $n = 2^J$ for some positive integer $J$, then we can represent (3.1) in the wavelet domain as

$$\boldsymbol{d}^* = \boldsymbol{\theta} + \boldsymbol{\varepsilon}, \tag{3.2}$$

where $\boldsymbol{d}^* = \boldsymbol{W}\boldsymbol{y}$, $\boldsymbol{\theta} = \boldsymbol{W}\boldsymbol{f}$, and $\boldsymbol{\varepsilon} = \boldsymbol{W}\boldsymbol{e}$, with $\boldsymbol{W}$ being the matrix associated with the orthonormal wavelet basis chosen to perform the wavelet transformation.

There are many threshold rules to process the coordinates of vector $\boldsymbol{d}^*$. Under all of them, one sets to zero all wavelet coefficients whose absolute values are below a threshold $\lambda$. The difference between distinct rules lies in how they deal with coefficients whose absolute values exceed $\lambda$. Figure 7 depicts the two most common threshold rules: *hard thresholding*,

$$\delta^h(d^*, \lambda) = dI_{\{|d^*|>\lambda\}}, \quad \lambda \geq 0, d^* \in \mathbb{R},$$

and *soft thresholding*,

$$\delta^s(d^*, \lambda) = \text{sgn}(d)\max(0, |d^*| - \lambda) \quad \lambda \geq 0, d^* \in \mathbb{R}.$$

Although these methods determine if the coefficients should be discarded or kept/shrunk, their effectiveness depends on an appropriate choice of threshold. If $\lambda$ is too small, only a few noisy coefficients would be discarded. On the other hand, if $\lambda$ is excessively large, one can compromise features of the unknown function (ABRAMOVICH; BAILEY; SAPATINAS, 2000).

Donoho and Johnstone (1994) proposed the *RiskShrink threshold*, also called *universal threshold*, given by

$$\lambda = \sigma\sqrt{2\log(\iota)},$$

with $\sigma$ being the standard deviation of $\boldsymbol{\varepsilon}$ and $\iota$ the number of coefficients at the finest resolution level, i.e., $\iota = 2^{J-1}$. In real problems, $\sigma$ is replaced by its estimate $\hat{\sigma}$. Donoho and Johnstone (1994) suggested estimating $\sigma$ by computing the median absolute deviation (MAD)[1] of the finest-scale coefficients, because, usually, the empirical wavelet coefficients at that level are essentially noise (NASON, 2008).

Another possible choice for $\lambda$ is the argument that minimizes Stein's unbiased risk estimate (SURE) (STEIN, 1981). Donoho and Johnstone (1995) proposed specifying this threshold level-wise and called this procedure of *SureShrink*. There are several other alternative data-adaptive thresholding rules. We emphasize the Bayesian approaches for thresholding, because it is a topic of interest in this work.

[1] The MAD of a data set is the median of the absolute deviations from the data's median.

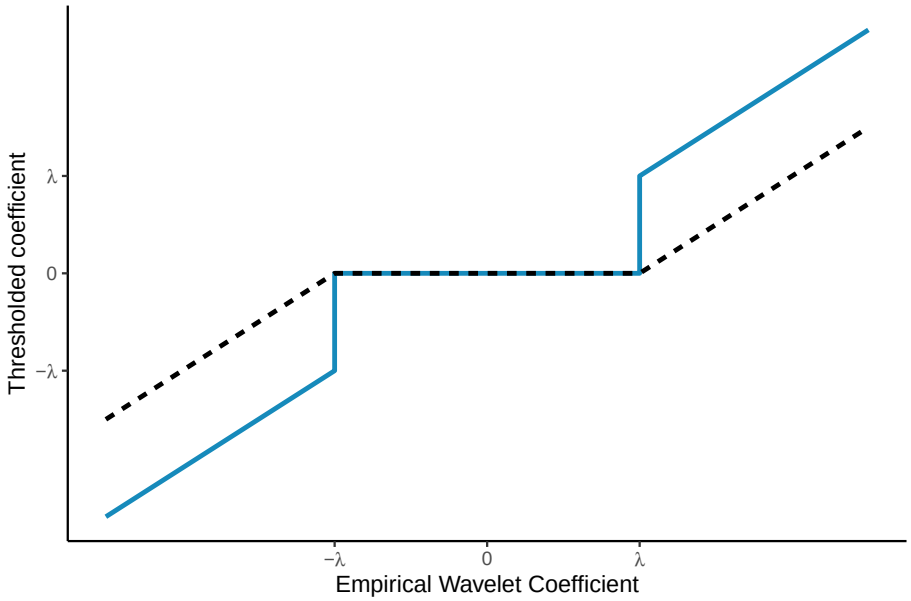



Figure 7 – Hard (full line) and soft thresholding (dashed line) rules for a threshold $\lambda$.

## 3.2 Bayesian regularization: spike and slab priors

A Bayesian wavelet shrinkage consists in placing a prior distribution to each wavelet coefficient of the unknown function. The prior is designed to capture the sparseness associated with most wavelet decompositions. Then, the vector $\boldsymbol{\theta} = (c_{00}, d_{00}, \boldsymbol{d}_1^T, \ldots, \boldsymbol{d}_{J-1}^T)^T$, where $\boldsymbol{d}_j = (d_{j0}, d_{j1}, \ldots, d_{j2^j-1})^T$, is estimated using some Bayes estimator and the resulting posterior distribution of the wavelet coefficients. Applying the IDWT to the estimated $\boldsymbol{\theta}$ gives us an estimation of vector $\boldsymbol{f}$.

One of the earliest contributions to Bayesian wavelet shrinkage is Chipman, Kolaczyk and McCulloch (1997), who propose a mixture between two Gaussian components as prior for each wavelet coefficient related to the function we want to estimate. This prior can be specified as

$$\pi_j \mathrm{N}(0, \upsilon_j^2) + (1 - \pi_j)\mathrm{N}(0, c_j \upsilon_j^2), \tag{3.3}$$

$k = 0, 1, \ldots, 2^j - 1$, $j = 0, 1, \ldots, J-1$, where $\pi_j, c_j$ and $\upsilon_j^2$ are prior parameters to be chosen at each resolution level $j$. This procedure is due to the fact that, typically, at finer resolution levels, the wavelet coefficients contain more noise than signal. The idea is to set considerably small values to $\upsilon_j^2$ and large values to $c_j$, $j = 0, \ldots, J-1$. In this setup, the Gaussian component with the largest variance would describe the behavior of coefficients to be considered as containing some signal, while the other one would capture the null coefficients (CHIPMAN; KOLACZYK; MCCULLOCH, 1997).

In Bayesian variable selection methods, (3.3) is also known as *spike and slab prior*, precisely because the *spike component* shrinks coefficients associated with irrelevant predictors and the *slab component* generates plausible values for the regression coefficients. The spike

and slab prior in (3.3) is originally brought by the Stochastic Search Variable Selection (SSVS) approach, proposed by George and McCulloch (1993).

Another type of spike and slab prior is the one presented by Kuo and Mallick (1998), where the spike component is a point of mass at zero, also called Dirac spike. This spike and slab prior is also incorporated in Bayesian wavelet shrinkage methods with the work of Abramovich, Sapatinas and Silverman (1998), who assume that the detail wavelet coefficients are mutually independent and each one is distributed following

$$\pi_j \mathrm{N}(0, \upsilon_j^2) + (1-\pi_j)\delta_0(\theta_{jk}), \tag{3.4}$$

$k = 0, 1, \ldots, 2^j - 1$, $j = 0, 1, \ldots, J-1$, with $\delta_0$ being a point mass at zero. The hyperparameters $\pi_j$ and $\upsilon_j^2$ are specified appropriately for each resolution level $j$. Observe that, in model (3.4), the $\pi_j$ corresponds to a prior probability that a wavelet coefficient at level $j$ is non-null. Because of that, $\pi_j$ is called *sparsity parameter*.

Abramovich, Sapatinas and Silverman (1998) assume that the hyperparameters $\upsilon_j^2$ and $\pi_j$ are of the form

$$\upsilon_j^2 = 2^{-\alpha j} C_1 \quad \text{and} \quad \pi_j = \min(1, 2^{-\beta j} C_2),$$

$k = 0, 1, \ldots, 2^j - 1$, $j = 0, 1, \ldots, J-1$, where $C_1, C_2, \alpha$ and $\beta$ are non-negative constants. According to Abramovich, Sapatinas and Silverman (1998), $C_1$ and $C_2$ are chosen empirically from the data, while $\alpha$ and $\beta$ are selected in conformity with the user's prior knowledge about the smoothness of the unknown function. In the absence of such knowledge, the authors suggest using the default choice $\alpha = 0.5$ and $\beta = 1$, since it is robust to various degrees of smoothness (ABRAMOVICH; SAPATINAS; SILVERMAN, 1998).

To complete the prior specification, Abramovich, Sapatinas and Silverman (1998) places a diffuse prior on the scaling coefficient at the coarsest level $c_{00}$. Diffuse priors are extensions of the uniform distribution, whose purpose is to minimize the bearing of the prior selection on the inference (MARIN; ROBERT, 2007). As a result, $c_{00}$ is estimated by the sample scaling coefficient obtained from the DWT of the data (ABRAMOVICH; SAPATINAS; SILVERMAN, 1998).

Under the prior (3.4), the posterior distribution for each detail coefficient is also a mixture between a Gaussian distribution and $\delta_0$, given by

$$\begin{aligned} \theta_{jk} | d_{jk}^* &\sim \pi_{\text{post}} \mathrm{N}\left( \frac{\upsilon_j^2}{1+\upsilon_j^2} d_{jk}^*, \frac{\upsilon_j^2}{1+\upsilon_j^2} \right) + (1-\pi_{\text{post}})\delta_0(\theta_{jk}), \\ \pi_{\text{post}} &= \frac{\pi_j g_{\upsilon_j^2}(d_{jk}^*)}{\pi_j g_{\upsilon_j^2}(d_{jk}^*) + (1-\pi_j)\phi(d_{jk}^*)}, \end{aligned} \tag{3.5}$$

$k = 0, 1, \ldots, 2^j - 1$, $j = 0, 1, \ldots, J-1$, where $\phi$ denotes the standard normal density and $g_{\upsilon_j^2}$ denotes the convolution between the slab component in (3.4) (in this case $\mathrm{N}(0, \upsilon_j^2)$) and $\phi$. Using

$\gamma$ to denote the slab density and $\star$ to denote the convolution operator, we can write $g = \gamma \star \phi$. See subsection B.1.1 for the complete derivation of (3.5).

To estimate the vector $\boldsymbol{\theta}$, Abramovich, Sapatinas and Silverman (1998) argue that choosing the traditional $L^2$-loss function as the Bayes rule would only shrink small coefficients towards zero, but not necessarily equal them to zero. Thus, this procedure is known as *shrinkage rule*. If the goal is to set all the estimated noisy coefficients to zero, i.e., if one wants a *thresholding rule*, the authors suggest the posterior median instead of the posterior mean as the Bayes rule. Abramovich, Sapatinas and Silverman (1998) refer to this Bayesian thresholding procedure as *BayesThresh*.

Johnstone and Silverman (2005a, 2005b) propose an approach known as *Empirical Bayes thresholding*. Instead of considering a Gaussian component to describe the behavior of non-null coefficients, the authors consider heavy-tailed distributions. This replacement intends to provide larger estimates for the true signal coefficients than those obtained from Gaussian distributions.

Concerning the slab component, the authors focus on two heavy-tailed distributions: the Laplace density and another density whose tails have the same weight as those of the Cauchy distribution. Because of that, the authors call the latter by *quasi-Cauchy* (JOHNSTONE; SILVERMAN, 2005b). Considering the Laplace density as the slab component, the prior for each detail wavelet coefficient can be written as

$$\pi_j \gamma_a(\theta_{jk}) + (1 - \pi_j)\delta_0(\theta_{jk}), \tag{3.6}$$

$k = 0, 1, \ldots, 2^j - 1,\ j = 0, 1, \ldots, J-1$, where $\gamma_a(x)$ denotes the Laplace density with scale parameter $a > 0$, i.e.,

$$\gamma_a(x) = \frac{a}{2}\exp(-a|x|), \quad x \in \mathbb{R}. \tag{3.7}$$

Johnstone and Silverman (2005a, 2005b) thresholding method is called Empirical Bayes because the hyperparameters $\pi_j$ and $a$ (when the slab component follows a Laplace distribution) are chosen automatically from the data, using a marginal maximum likelihood approach. This means that for each resolution level $j$ of the wavelet transform, the method selects the arguments $\pi_j$ and $a$ that maximize the marginal log-likelihood and plugs them back into the prior. Then, the estimation of $\boldsymbol{\theta}$ is carried out with either posterior medians, posterior means, or other estimators.

Under the prior (3.6), the posterior distribution is given by

$$\begin{aligned} \theta_{jk}|d_{jk} &\sim \pi_{\text{post}} f_1(\theta_{jk}|d_{jk}) + (1 - \pi_{\text{post}})\delta_0(\theta_{jk}), \\ \pi_{\text{post}} &= \frac{\pi_j g_a(d^*_{jk})}{\pi_j g_a(d^*_{jk}) + (1-\pi_j)\phi(d^*_{jk})}, \end{aligned} \tag{3.8}$$

$k = 0, 1, \ldots, 2^j - 1,\ j = 0, 1, \ldots, J-1$, with $f_1(\theta_{jk}|d_{jk})$ being the non-null mixture component and $g_a = \gamma_a \star \phi$. It can be shown that $f_1(\theta_{jk}|d_{jk})$ is a mixture of two truncated normal distributions (see subsection B.1.2 for the complete derivation). Let $X$ have a normal distribution with mean

$\mu$ and variance $\sigma^2$ that lies within the interval $(\alpha,\beta)$, where $-\infty \leq \alpha < \beta \leq \infty$. Thus, we say that $x$, conditional on $\alpha < x < \beta$, has a truncated normal distribution and denote its density by $f_{\text{TN}}(x|\mu,\sigma,\alpha,\beta)$. Then, with a slight abuse of notation, we can write $f_1(\theta_{jk}|d_{jk})$ as

$$\begin{aligned} f_1(\theta_{jk}|d_{jk}) = \eta \times f_{\text{TN}}\left(\theta_{jk}\middle|\frac{d_{jk}}{\sigma_j} - a, 1, 0, +\infty\right) \\ + (1-\eta) \times f_{\text{TN}}\left(\theta_{jk}\middle|\frac{d_{jk}}{\sigma_j} + a, 1, -\infty, 0\right), \end{aligned} \tag{3.9}$$

where

$$\eta = \frac{\exp(-a\frac{d_{jk}}{\sigma_j})\Phi(\frac{d_{jk}}{\sigma_j} - a)}{\exp(a\frac{d_{jk}}{\sigma_j})\tilde{\Phi}(\frac{d_{jk}}{\sigma_j} + a) + \exp(-a\frac{d_{jk}}{\sigma_j})\Phi(\frac{d_{jk}}{\sigma_j} - a)},$$

with $\Phi$ denoting the standard normal cumulative function, and $\tilde{\Phi} = 1 - \Phi$.

## 3.3 Wavelet Shrinkage examples

In the previous sections, we present Frequentist and Bayesian methods that reduce signal noise using wavelet bases. Although this work focuses on the Bayesian approaches of section 3.2, for comparative purposes only, in this section, we use synthetic data sets to illustrate the performance of all the presented methods. We consider five different functions to generate the data sets, with the last three being among the test functions introduced by Donoho and Johnstone (1994):

1. Parabolic:
$$f(t) = 3(t-0.5)^2 + 0.125. \tag{3.10}$$

2. Sinusoidal:
$$f(t) = 0.4\cos(2\pi(t+\pi)) + 0.5. \tag{3.11}$$

3. Heavisine:
$$f(t) = 4\sin(4\pi t) - \text{sgn}(t-0.3) - \text{sgn}(0.72-t). \tag{3.12}$$

4. Blocks:
$$f(t) = \sum_{i=1}^{11} h_i b(t-t_i), \text{ where } b(r) = (1+\text{sgn}(r))/2. \tag{3.13}$$

   - $\{t_i\} = \{0.1, 0.13, 0.15, 0.23, 0.25, 0.40, 0.44, 0.65, 0.76, 0.78, 0.81\}$;
   - $\{h_i\} = \{4, -5, 3, -4, 5, -4.2, 2.1, 4.3, -3.1, 5.1, -4.2\}$.

5. Bumps:
$$f(t) = \sum_{i=1}^{11} h_i b((t-t_i)/s_i), \text{ where } b(r) = (1+|r|)^{-4}. \tag{3.14}$$

- $\{t_i\} = \{0.1, 0.13, 0.15, 0.23, 0.25, 0.40, 0.44, 0.65, 0.76, 0.78, 0.81\}$;
- $\{h_i\} = \{4, 5, 3, 4, 5, 4.2, 2.1, 4.3, 3.1, 5.1, 4.2\}$;
- $\{s_i\} = \{0.005, 0.005, 0.006, 0.01, 0.01, 0.03, 0.01, 0.01, 0.005, 0.008, 0.005\}$.

We apply each function to 1024 equally spaced points on $[0, 1]$. Then, we corrupt these samples with independent normally distributed noise $\mathrm{N}(0, \sigma^2)$. The values for $\sigma$ are taken to satisfy a *signal-to-noise ratio* (SNR)[1] equal to 4. Figure 8 shows the samples with their respective corrupted versions. In real denoising problems, we have access only to the noisy versions. Thus, we aim to reduce the noise and get the most reliable estimate of the signal that is not accessible. In this simulated study, since we have access to the real signals, we can compare them with the estimates provided by each method and obtain an idea of their performances.

Once the noisy data is generated, we obtain its wavelet coefficients by applying the pyramidal algorithm implemented in `wavethresh` (NASON, 2016). Henceforth, we use the coiflet basis with six vanishing moments to perform the transform. It should be stressed that, regarding posterior simulations, using other Daubechies wavelet bases provides similar results to those achieved by this specific coiflet basis. We will not present these supplementary analyses due to space limitations. Additionally, following Donoho and Johnstone (1994), we estimate the standard deviation of the noise by computing the MAD of the finest-scale coefficients.

In order to denoise the coefficients, we apply the four methods briefly introduced in this chapter: universal threshold, SureShrink threshold, BayesThresh approach, and Empirical Bayes thresholding. To apply the BayesThresh procedure, the default choices for $\alpha = 0.5$ and $\beta = 1$ are used. Regarding the Empirical Bayes approach, we use the mixture between a point mass at zero and the Laplace density as prior. In this scenario, the scale parameter $a$ is estimated jointly with the sparsity parameter through the marginal maximum likelihood approach (JOHNSTONE; SILVERMAN, 2005a; JOHNSTONE; SILVERMAN, 2005b). For the Frequentist approaches, we consider the soft thresholding rule, whereas in the Bayesian thresholding methods, we use the posterior median as the point estimate.

Even though this is a simple simulation, it allows us to compare the methods and visualize the denoising contexts where they are eligible. Figures 9 –13 show the estimated signals of all the methods. When the signals are smoother (Parabolic and Sinusoidal), all procedures perform similarly at denoising them. However, the estimates diverge when the functions are rougher (Heavisine, Bumps, and Blocks). For instance, the reconstructions provided by the Universal and SureShrink thresholds for the Heavisine function do not follow signal discontinuities as well as the Bayesian methods do. Likewise, the Universal threshold underperforms when estimating corners in the Blocks function and high peaks in Bumps. Although the estimates for the Blocks

[1] The SNR is the ratio of the sample standard deviation of the signal to the standard deviation of the noise.

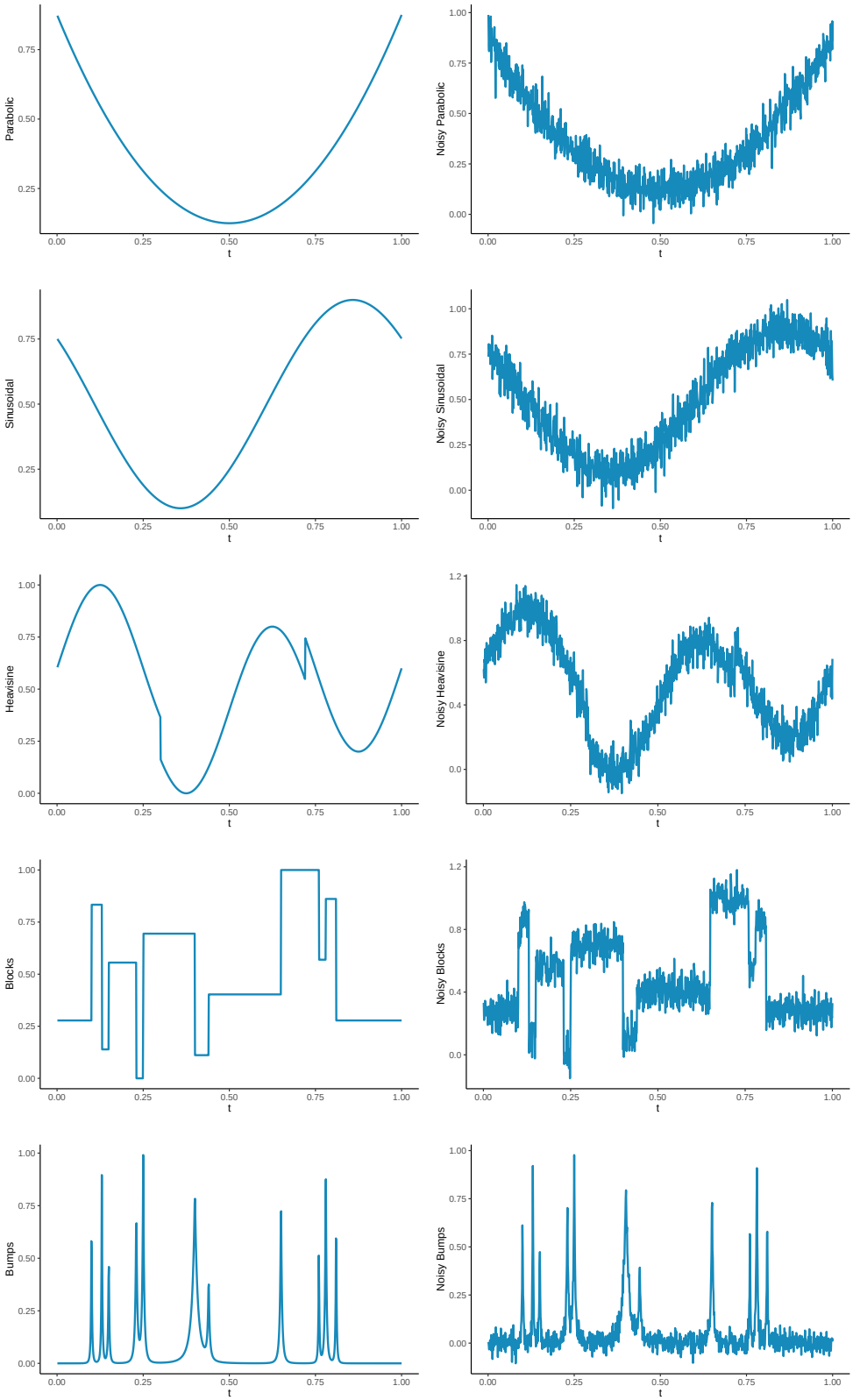
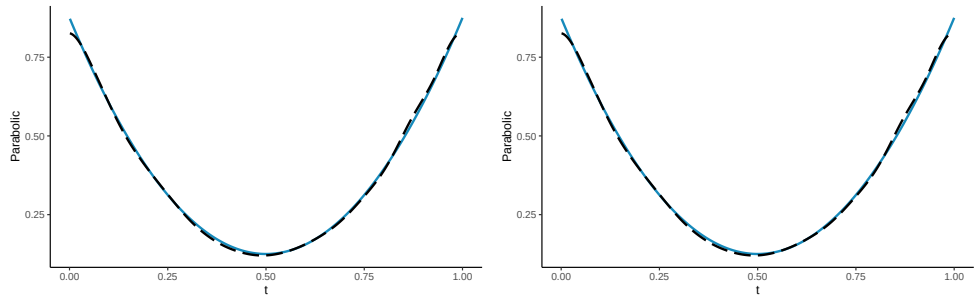


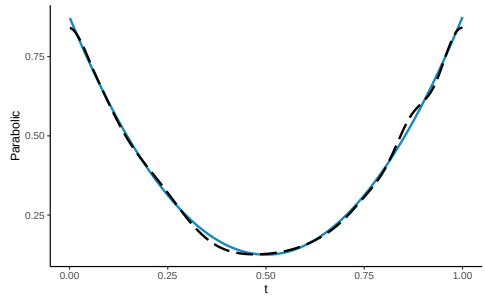

Figure 8 – On the left-hand side are presented,respectively, the functions: Parabolic, Sinusoidal, Heavisine, Blocks and Bumps. On the right-hand, the same functions with added iid Gaussian noise with a variance chosen to achieve a SNR of 4.



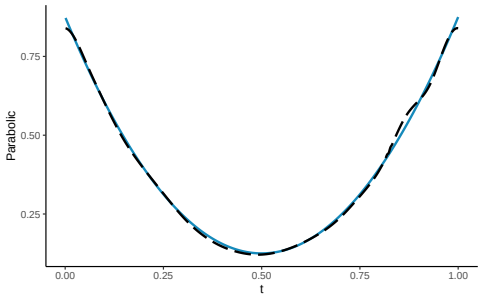

(a) Universal threshold




(b) SureShrink threshold




(c) BayesThresh approach

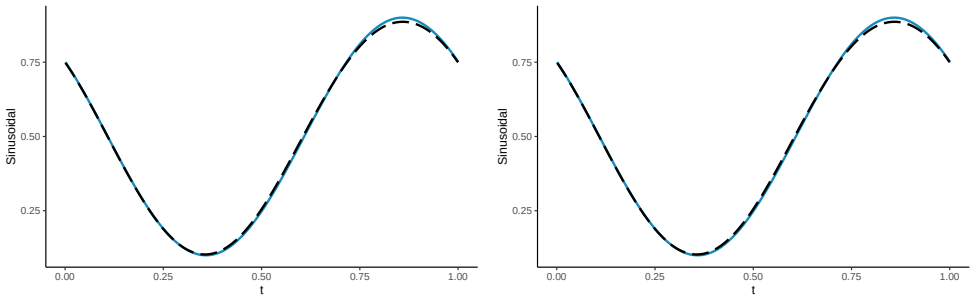



(d) Empirical Bayes thesholding

Figure 9 – Original Parabolic function (full lines) and various reconstructions of the signal based on different thresholding methods (dashed lines).

and Bumps of the other three methods are alike, the estimates of the SureShrink threshold are noisier than the estimates of the Bayesian methods.

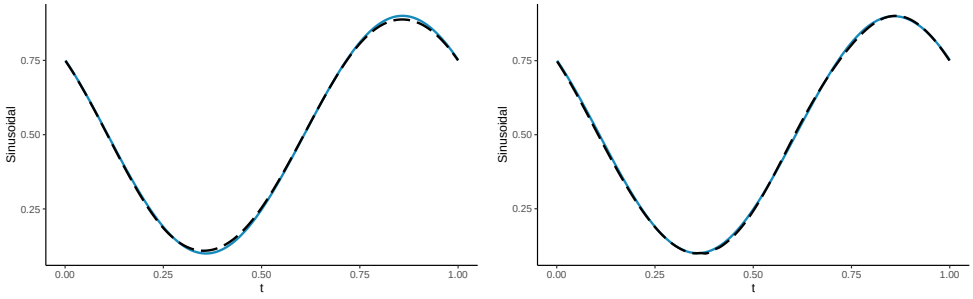
In reason of the good performance of Bayesian approaches in estimating curves through wavelet bases, from now on, they will be employed for these purposes. In the next chapter, we present the dynamic Gaussian mixture model and discuss how to use wavelet bases within the estimation of this model following a Bayesian framework.




(a) Universal threshold




(b) SureShrink threshold

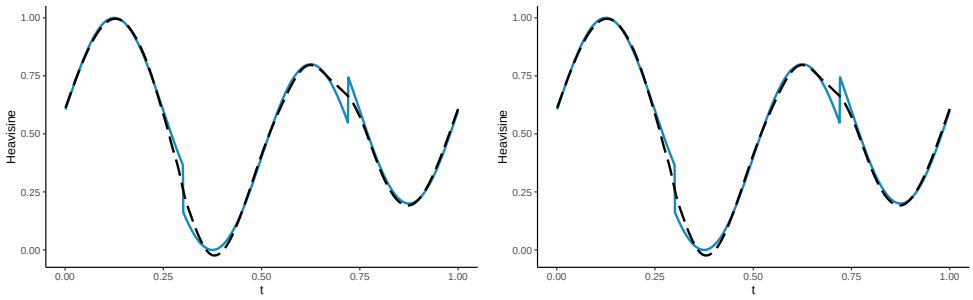



(c) BayesThresh approach




(d) Empirical Bayes thesholding

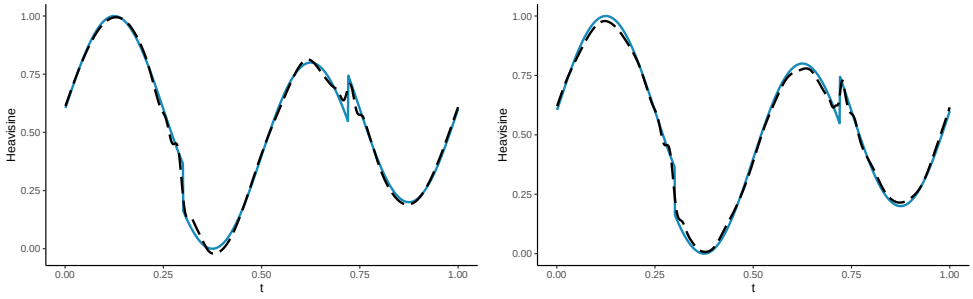
Figure 10 – Original Sinusoidal function (full lines) and various reconstructions of the signal based on different thresholding methods (dashed lines).




(a) Universal threshold




(b) SureShrink threshold

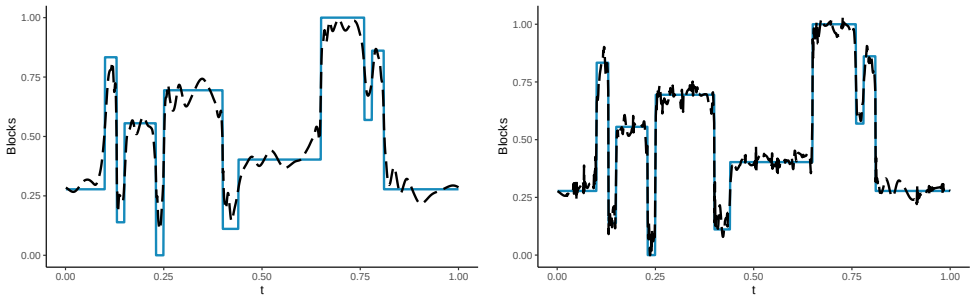



(c) BayesThresh approach




(d) Empirical Bayes thesholding

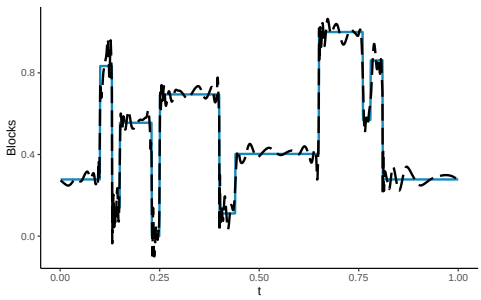
Figure 11 – Original Heavisine function (full lines) and various reconstructions of the signal based on different thresholding methods (dashed lines).



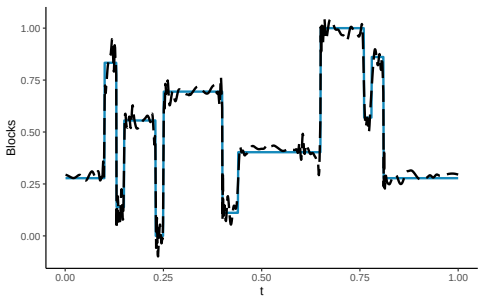

(a) Universal threshold




(b) SureShrink threshold

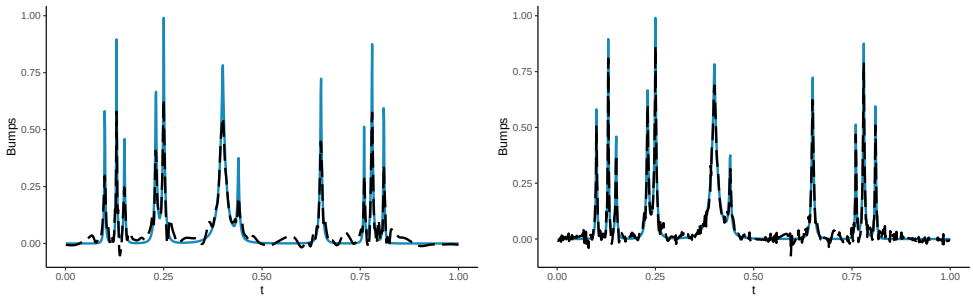



(c) BayesThresh approach




(d) Empirical Bayes thesholding

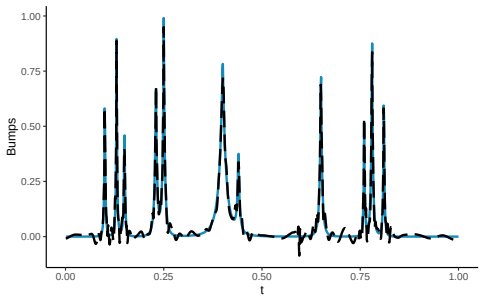
Figure 12 – Original Blocks function (full lines) and various reconstructions of the signal based on different thresholding methods (dashed lines).



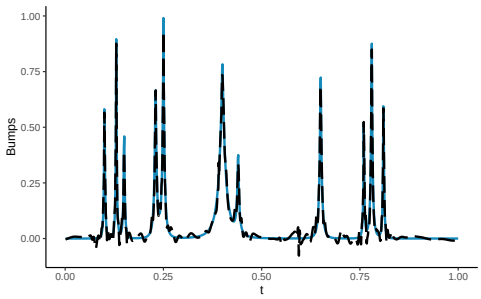

(a) Universal threshold




(b) SureShrink threshold




(c) BayesThresh approach




(d) Empirical Bayes thesholding

Figure 13 – Original Bumps function (full lines) and various reconstructions of the signal based on different thresholding methods (dashed lines).

CHAPTER

# 4

# DYNAMIC GAUSSIAN MIXTURE MODEL

In Chapter 3, we presented some Bayesian wavelet shrinkage methods. In this chapter, we discuss how wavelet bases can be employed to address the estimation problem of a dynamic Gaussian mixture model. We begin introducing the model in section 4.1 and section 4.2. Then, in section 4.3, we comment on the conditional posterior distributions of the mixture parameters used within the MCMC sampling schemes. At last, in section 4.4, we present two estimation approaches for the mixture weights based on Bayesian wavelet shrinkage methods.

## 4.1 Finite mixture models

In several statistical problems, we deal with data that cannot be suitably modeled by a single unimodal distribution. Unlike simple parametric models that may fail to fit data with multimodal behavior, finite mixture models can provide satisfactory approximations of data sets with irregular patterns. They do this by clustering the data observations into subgroups, also known as mixture components.

Usually, mixture models assume that a sample $y_1, \ldots, y_n$ represent i.i.d. realizations from a random variable $Y$ that belongs to a population composed of $K$ subpopulations. Within each subpopulation $g$, the random variable $Y$ is modeled by a distribution with probability density function $f(y|\boldsymbol{\rho}_g)$. The definition of finite mixture models is given below.

**Definition 7.** A random variable $Y$ with support in $\mathcal{Y} \in \mathbb{R}$ follows a finite mixture distribution if its probability density function is given by

$$f(y|\boldsymbol{\rho}, \boldsymbol{\alpha}) = \sum_{g=1}^{K} \alpha_g f(y|\boldsymbol{\rho}_g), \qquad \forall y \in \mathcal{Y}, \tag{4.1}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)^T$ is the vector of mixture weights, such that $\sum_{g=1}^{K} \alpha_g = 1$ and $0 \leq \alpha_g \leq 1$, $g = 1, \ldots, K$, with $\boldsymbol{\rho} = (\boldsymbol{\rho}_1^T, \ldots, \boldsymbol{\rho}_K^T)^T$ being the vector of *component parameters*.

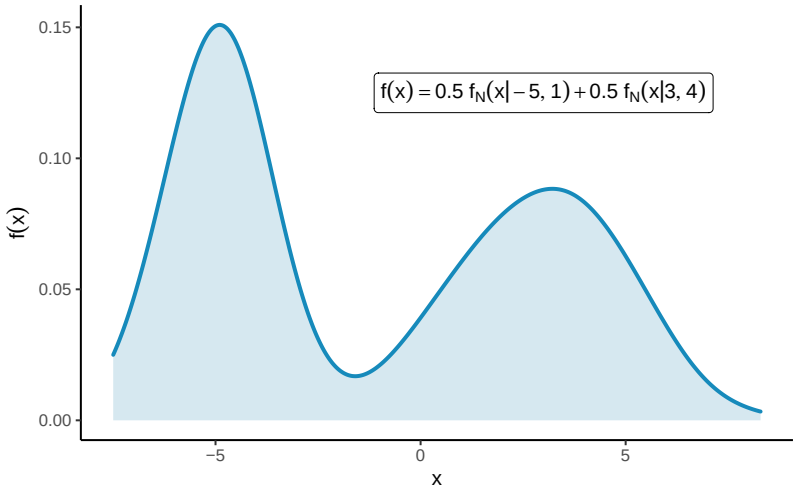



Figure 14 – Density of a mixture of two univariate normal distributions.

A mixture model often found in many context is the mixture model with two components, i.e.,

$$f(y|\boldsymbol{\rho},\boldsymbol{\alpha}) = \alpha f(y|\boldsymbol{\rho}_1) + (1-\alpha) f(y|\boldsymbol{\rho}_2) \qquad \forall y \in \mathbb{R}, \tag{4.2}$$

where $0 \leq \alpha \leq 1$. In clinical and epidemiological settings, for example, a two-component mixture model is frequently used to classify data between the groups "disease absent" and "disease present" (HALL; ZHOU, 2003). For more details see, e.g., Rindskopf and Rindskopf (1986) and Hui and Zhou (1998). Another example of an application based on (4.2) is contamination problems, commonly found in astronomy studies (PATRA; SEN, 2016). In these scenarios, a sample of objects of interest (known as members) from a distant galaxy (e.g., stars) is contaminated to some extent by foreground/background objects (known as contaminants). Therefore, using a two-component mixture model allows separating the objects between groups: members and contaminants (see Walker *et al.* (2009)). In genetics, these models are also used to detect differentially expressed genes within microarray data (see Bordes, Delmas and Vandekerkhove (2006)).

Most of the previous work on this problem frequently assumes that the component densities arise from parametric families of distributions (PATRA; SEN, 2016). In fact, one of the first mixture models fitted to a heterogenous data set was a mixture of two univariate Gaussian densities (see Pearson (1894)). When the mixture components are normal distributions, the mixture model is also called a Gaussian mixture model. Thus, (4.2) becomes

$$f(y|\boldsymbol{\rho},\boldsymbol{\alpha}) = \alpha f_N(y|\mu_1,\sigma_1^2) + (1-\alpha) f_N(y|\mu_2,\sigma_2^2) \qquad \forall y \in \mathbb{R}, \tag{4.3}$$

where $f_N(x|a,b)$ corresponds to the probability density function of a normal random variable with mean $a$ and variance $b$, evaluated at $x$. In Figure 14, we present the density of a two-component mixture of univariate normal distributions.

Observe that in all mixture models discussed, (4.1), (4.2) and (4.3), their mixture components are assigned to mixing probabilities with constant behavior. However, this setting can be restrictive and unsuitable for some applications. For instance, in medical studies concerning the longitudinal effects of treatments, we are interested in classifying the patients between "disease present" and "disease absent" according to their response to medications across time. In this context, the probability of a patient belonging to a certain group is not constant. In fact, it is a function of time, and therefore, it is dynamic (see, e.g., Lu and Song (2012)).

For these scenarios, we can extend the finite mixture model in Definition 7, allowing the mixture weights to vary according to some index, making the model more flexible and adaptive to different data sets. Examples of this extension for a two-component mixture model are shown in Figure 15. In these cases, the mixture weights vary continuously as functions of $t$. Notice how the density behavior in the right column is dictated by the dynamic mixture weight in the left column. From the examples depicted by Figure 15, we can have a general idea about the challenge of estimating mixture weights that have a dynamic behavior. In the next section, we formalize the two-component Gaussian mixture model with dynamic mixing probabilities studied in this work.

## 4.2 The dynamic Gaussian mixture model

Suppose that $y_1, \dots y_n$ represent a random sample from the dynamic Gaussian mixture model

$$\begin{aligned} y_t &= (1-z_t)x_{1t} + z_t x_{2t}, \\ x_{kt}|\mu_k, \tau_k^2 &\sim \mathrm{N}(\mu_k, \tau_k^{-2}), \quad k = 1,2, \\ z_t|\alpha_t &\sim \mathrm{Bern}(\alpha_t), \quad t = 1, \dots, n, \end{aligned} \tag{4.4}$$

where $z_t$'s are non-observable allocation variables that indicate to which mixture component the observations $y_t$'s belong to. The $z_t$ have a Bernoulli distribution with parameter $\alpha_t$, the unknown mixture weight, which has a dynamic behavior. In this sense, if $z_t = 1$, the $t$-th observation belongs to the normal population with mean $\mu_2$ and precision $\tau_2^2$, otherwise $y_t$ is sampled from the normal population with mean $\mu_1$ and precision $\tau_1^2$.

In this context, besides the concern about estimating the component parameters $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$ and $\boldsymbol{\tau^2} = (\tau_1^2, \tau_2^2)^T$, we are also interested in estimating the dynamic behavior of the mixing probability $\alpha_t$. Figure 16 shows the influence of $\alpha_t$ on the shape of the mixture's density curve. Observe the modification of the density curve as $\alpha_t$ varies.

In a frequentist framework, the estimation of the mixture parameters lies in implementing an Expectation-Maximization (EM) algorithm over the mixture likelihood function. Nevertheless, in many cases, the success of this approach depends on both the stopping criterion and the initial values used. Another disadvantage of the EM algorithm is the slow convergence rate that it may present (KARLIS; XEKALAKI, 2003).

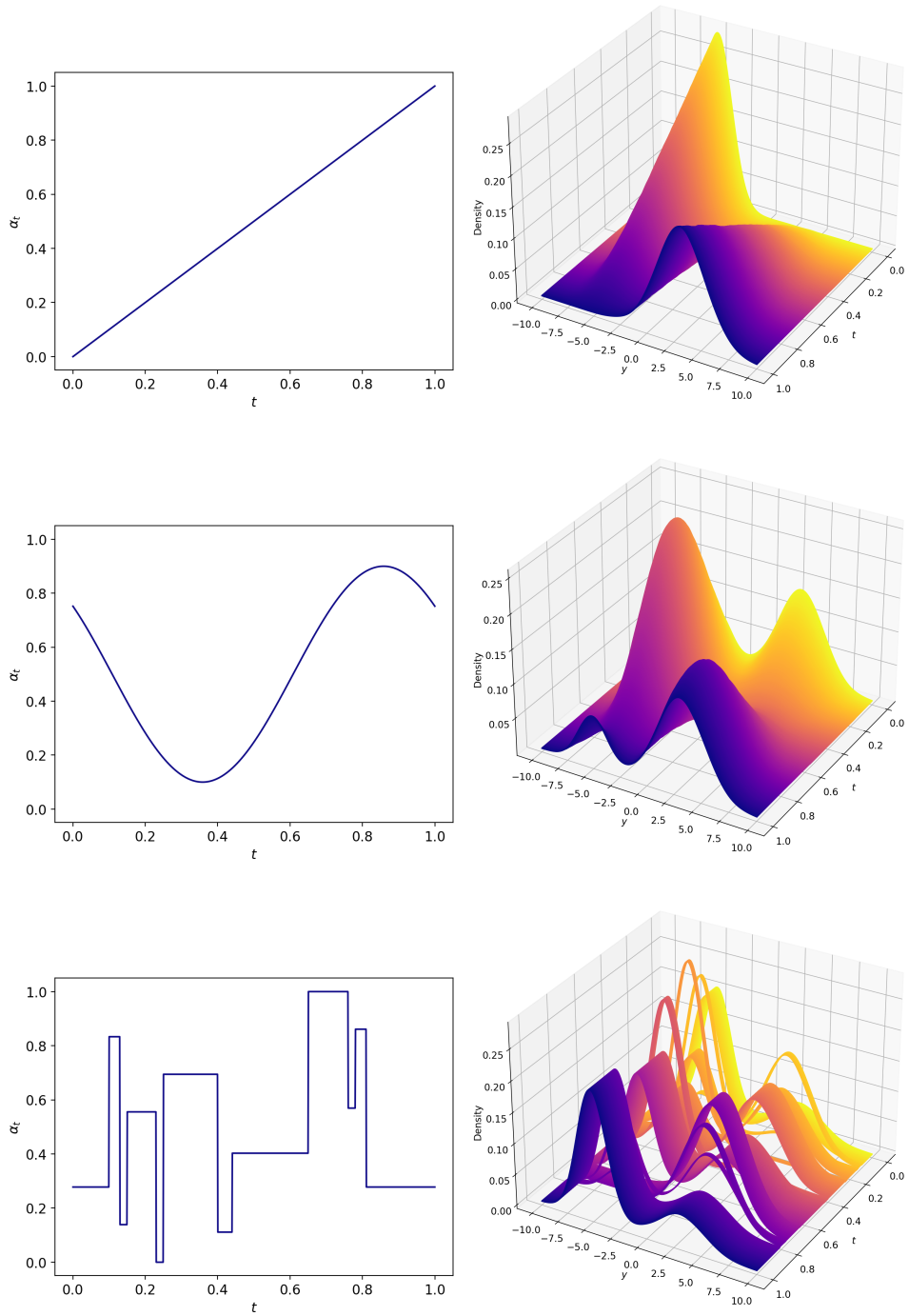



Figure 15 – Functions used to describe the mixture weight behavior are presented on the left-hand side. In the right column, the densities of the mixtures of two univariate normal distributions, whose mixture weight follows the dynamic behavior presented on the left. The mixtures are generated according to $f(y) = (1-\alpha_t) f_N(x|-5,1) + \alpha_t f_N(x|3,4)$.

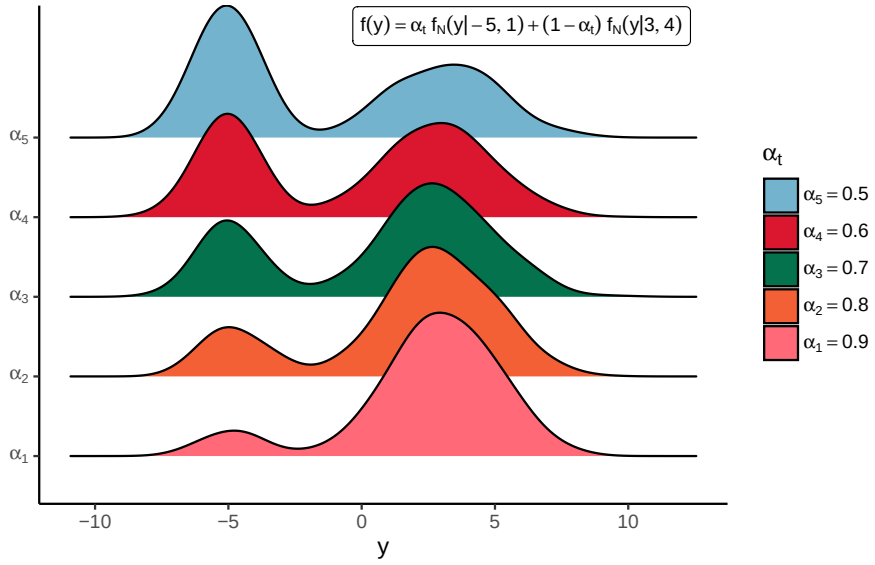



Figure 16 – Illustration of the influence of the dynamic mixture weight on the shape of the density. In this work, we are concerned about estimating the values that $\alpha_t$ assumes for each $t$.

From a Bayesian perspective, a practical estimation of $\boldsymbol{\mu}$ and $\boldsymbol{\tau^2}$ lies in data augmentation and MCMC algorithms (see, e.g., Lavine and West (1992)). Once assigned prior distributions to the unknown parameters, one can use methods such as Monte Carlo Markov Chain to sample from the joint posterior distribution and, therefore, make inferences (e.g., point and credible estimates).

Well-known results from Markov chain theory guarantee that in the long run, under reasonably general conditions, the distribution of the posterior draws of each parameter converges to a stationary distribution, which could be shown to be equal to the posterior distribution of interest (CASELLA; GEORGE, 1992). The Gibbs sampling algorithm is especially useful when one can not analytically determine or directly sample from the complete joint posterior (KRUSCHKE, 2014). Due to these properties, this algorithm is a beneficial approach to obtaining Bayesian inference about unknown parameters.

To obtain the posterior draws from the MCMC sampling scheme, we specify conjugate prior distributions to the component parameters and, through sampling from their posterior distributions, we obtain the posterior draws. In the following section, we derive the conditional posterior distributions of these parameters. Regarding the dynamics of the mixture weight, we study two wavelet-based estimation approaches to this task (see section 4.4).

## 4.3 Full conditional posterior distributions

This section provides the full conditional posterior distributions of $\boldsymbol{\mu}$ and $\boldsymbol{\tau^2}$ and $\boldsymbol{z} = (z_1, \ldots, z_n)^T$. Following the model (4.4), we consider that each observation of the sample $\boldsymbol{y} =$

$(y_1,\ldots,y_n)^T$ is drawn from one of two different normal subpopulations. Relevant group-specific quantities are the number $T_k$ of observations in group $k$ and the group mean $s_k/T_k$, where, for $k=1,2$,

$$T_k = \#\{t : z_t = k-1, t = 1,2,...,n\},$$
$$s_k = \sum_{t:z_t=k-1} y_t.$$

In order to derive the full conditional posterior distribution of $\mu_k$ and $\tau_k^2$, we assume that the data set available $\boldsymbol{y}$ is a time series whose dependence structure is determined by the dynamic mixture weight behavior $\alpha_t$, i.e., $p(\boldsymbol{y}|\boldsymbol{\mu},\boldsymbol{\tau^2},\boldsymbol{z}) = \prod_{t=1}^n p(y_t|z_t,\boldsymbol{\mu},\boldsymbol{\tau^2})$ and $p(\boldsymbol{z}|\alpha_1,\ldots,\alpha_n) = \prod_{t=1}^n p(z_t|\alpha_t)$ (MONTORIL; CORREIA; MIGON, 2021). For the sake of simplicity, let us denote by $[\ldots]$ the set of all remaining variables to be considered for the posterior in use. Then, the complete-data likelihood function is given by

$$p(\boldsymbol{y}|[\ldots]) = \prod_{k=1}^{2}\left(\frac{\tau_k^2}{2\pi}\right)^{T_k/2}\exp\left[-\frac{\tau_k^2}{2}\sum_{t:z_t=k-1}(y_t-\mu_k)^2\right].$$

Each factor of this multiplication carries all information about the parameters of a certain group $k$, and each one, combined with a prior, gives us the posterior density of $\mu_k$ or $\tau_k^2$. Here, we are considering independent priors for the component parameters $\boldsymbol{\mu}$ and $\boldsymbol{\tau^2}$, which means $p(\boldsymbol{\mu},\boldsymbol{\tau_k^2}) = p(\mu_1)p(\tau_1^2)p(\mu_2)p(\tau_2^2)$.

Under the conjugate priors $\mu_k \sim \mathrm{N}(b_{0k},B_{0k})$ and $\tau_k^2 \sim \Gamma(c_{0k},C_{0k})$, one obtains the conditional posterior distributions for each $\mu_k$ and $\tau_k^2$, respectively,

$$\mu_k|[\ldots] \sim \mathrm{N}(b_k,B_k), \tag{4.5}$$
$$\tau_k^2|[\ldots] \sim \Gamma(c_k,C_k), \tag{4.6}$$

where

$$B_k = (B_{0k}^{-1}+\tau_k^2 T_k)^{-1},$$
$$b_k = B_k(\tau_k^2 s_k + B_{0k}^{-1}b_{0k}),$$
$$C_k = C_{0k} + \frac{\sum_{t:z_t=k-1}(y_t-\mu_k)^2}{2},$$
$$c_k = c_{0k} + \frac{T_k}{2}.$$

Given the observations $\boldsymbol{y}$ and the parameters $\boldsymbol{\mu}$, $\boldsymbol{\tau^2}$ and $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_n)^T$, the $z_t$'s are conditionally independent and $p(z_t = 1|\boldsymbol{y},\boldsymbol{\mu},\boldsymbol{\tau^2},\boldsymbol{\alpha}) \propto \alpha_t f_N(y_t|\mu_2,\tau_2^{-2})$. It follows that, for each $t=1,\ldots,n$, the full conditional posterior of $z_t$ is given by

$$z_t|[\ldots] \sim \mathrm{Bern}(\beta_t),$$
$$\beta_t = \frac{\alpha_t f_N(y_t|\mu_2,\tau_2^{-2})}{\alpha_t f_N(y_t|\mu_2,\tau_2^{-2}) + (1-\alpha_t)f_N(y_t|\mu_1,\tau_1^{-2})}. \tag{4.7}$$

### 4.3.1 The label switching problem

The term *label switching* was introduced by Redner and Walker (1984) to denote the invariance of a mixture likelihood function under relabelling the mixture components. This phenomenon is not an issue for maximum likelihood estimation. However, in a Bayesian context, if not properly addressed, the label switching can lead to inadequate estimates of the component parameters (STEPHENS, 2000).

A common strategy to address this problem is to impose an *identifiability constraint* on the parameter space. Thus, whenever an MCMC draw does not fulfill the established constraint, one permutes the labeling of the component parameters to satisfy the restriction. In our approach, we adopt the simple constraint $\mu_1 < \mu_2$ and reorder the pairs $(\mu_k, \tau_k^2)$ accordingly.

## 4.4 Estimation of the dynamic mixture weights

In this section, we present two wavelet-based approaches to estimate the dynamic mixture weights $\alpha_t$. The first consists of transforming the original data into a regression, whose regression function is the mixture weight $\alpha_t$. This rescaling procedure is the same used by Montoril, Pinheiro and Vidakovic (2019). Once rescaled the observations, we use Bayesian wavelet shrinkage techniques to reduce the noise associated with $\alpha_t$.

The second approach adapts the data augmentation method proposed by Albert and Chib (1993) to model the allocation data $z_t|\alpha_t \sim \text{Bern}(\alpha_t)$ in (4.4). As in Albert and Chib (1993), we also combine the probit binary regression model on the $z_t$'s with a normal linear regression on the introduced latent variables. The adaptation consists of using a matrix associated with an orthonormal wavelet basis as the design matrix in the regression.

### 4.4.1 Wavelet regression approach

Let us define

$$m_t = \frac{y_t - \mu_1}{\mu_2 - \mu_1},$$

where $y_t$, $\mu_1$ and $\mu_2$ are specified in (4.4). We can rewrite $m_t$ following a nonparametric regression

$$\begin{aligned} m_t &= \alpha_t + u_t, \quad t = 1, \ldots, n, \\ u_t &= z_t \frac{x_{2t} - x_{1t}}{\mu_2 - \mu_1} - \alpha_t + \frac{x_{1t} - \mu_1}{\mu_2 - \mu_1}, \end{aligned} \tag{4.8}$$

with $u_1, \ldots, u_n$ being independent realizations of a random variable (error). By rewriting $m_t$ as in (4.8), we are transforming the mixture problem into a denoising problem. In this case, the error is the noise that must be reduced to estimate the signal of interest (the dynamic mixture weights $\alpha_t$'s). Given the observed values $y_1, \ldots, y_n$, these errors are uncorrelated, have zero mean, and

present finite variance,

$$\mathbb{E}(u_t) = 0,$$
$$\text{Var}(u_t) = \frac{\tau_1^{-2}}{(\mu_2-\mu_1)^2} + \frac{\tau_2^{-2}-\tau_1^{-2}}{(\mu_2-\mu_1)^2}\alpha_t + \alpha_t(1-\alpha_t),$$

as shown by Montoril, Pinheiro and Vidakovic (2019). Although we do not detail here, the errors do not follow any well-known probability distribution in the literature. In fact, the errors have a complex and heteroscedastic distribution. Because of this, we approximate it by a homoscedastic Gaussian distribution, as usually done in practical situations.

Following this scenario, the estimation of $\alpha_t$, $t = 1,\ldots,n$, can be done through some regularization technique that reduces the noise $u_1,\ldots,u_n$ in (4.8). In this work, we use wavelet shrinkage as the regularization technique to address the estimation. In particular, we apply the BayesThresh method, as described in section 3.2, to the wavelet coefficients obtained by DWT applied to $\boldsymbol{m} = (m_1,\ldots,m_n)^T$. The resulting MCMC procedure is described in Algorithm 1.

**Algorithm 1** – Gibbs sampling algorithm - Wavelet regression - Version 1

1: Choose number of iterations $N$.
2: Specify initial values for $\boldsymbol{\mu}^{(0)}$, $\boldsymbol{\tau^2}^{(0)}$, $\boldsymbol{z}^{(0)} = (z_1^{(0)},\ldots,z_n^{(0)})^T$ and $\boldsymbol{\alpha}^{(0)}$.
3: **for** $i \leftarrow 1$ to $N$ **do**
4: Sample $\mu_1^{(i)} \sim p(\mu_1|[\ldots])$. ▷ See (4.5)
5: Sample $\tau_1^{2(i)} \sim p(\tau_1^2|[\ldots])$. ▷ See (4.6)
6: Sample $\mu_2^{(i)} \sim p(\mu_2|[\ldots])$. ▷ See (4.5)
7: Sample $\tau_2^{2(i)} \sim p(\tau_2^2|[\ldots])$. ▷ See (4.6)
8: **if** $\mu_2 < \mu_1$ **then**
9: Permute the labeling of pairs $(\mu_k^{(i)}, \tau_k^{2(i)})$.
10: **end if**
11: Sample $z_t^{(i)} \sim p(z_t|[\ldots])$, for $t = 1,\ldots,n$. ▷ See (4.7)
12: Calculate $\boldsymbol{m}^{(i)} = (\boldsymbol{y} - \mu_1^{(i)})/(\mu_2^{(i)} - \mu_1^{(i)})$.
13: Compute $\boldsymbol{d}^{(i)} = \boldsymbol{W}\boldsymbol{m}^{(i)}$. ▷ $\boldsymbol{W}$ is the matrix form of the DWT.
14: Apply BayesThresh to $\boldsymbol{d}^{(i)}$.
15: Calculate $\boldsymbol{\alpha}^{(i)} = \boldsymbol{W}^T\boldsymbol{d}^{(i)}$.
16: **end for**

Since the Markov chains may not be in equilibrium at the beginning, we discard the first $B$ iterations as the burn-in period. For posterior inference, we consider draws lagged by each $L$ iterations, in order to reduce correlation. Thus the final size of our marginal posterior samples is $N_{\text{final}} = \frac{N-B}{L}$.

### 4.4.2 Data augmentation approach

Let $z_1,\ldots,z_n$ be independent binary random variables. For each $t = 1,\ldots,n$, there is a vector $\boldsymbol{x}_t = (x_{t1},\ldots,x_{tp})^T$ of known covariates. The probit model consists of defining the

regression $p(z_t = 1) = \Phi(\boldsymbol{x}_t^T \boldsymbol{\theta})$, where $\Phi$ is the standard Gaussian cumulative function and $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)^T$ is a vector of $p$ unknown parameters. Following Albert and Chib (1993), to introduce the data augmentation approach, one creates an auxiliary random variable $l_t$, such that $l_t = \boldsymbol{x}_t^T \boldsymbol{\theta} + e_t$ and $e_t \sim \mathrm{N}(0,1)$. Moreover, $l_t$ is built in a such way that $z_t = 1$, if $l_t > 0$, or $z_t = 0$, otherwise.

The key idea of our approach is to use the transpose of a matrix constructed from an orthonormal wavelet basis $\boldsymbol{W}^T$ as the design matrix $\boldsymbol{X}$ in the probit regression of the data augmentation method proposed by Albert and Chib (1993). Thus, we introduce a random sample $\boldsymbol{l} = (l_1, \ldots, l_n)^T$ of $n$ independent latent variables into the model (4.4), and define the allocation variable $z_t$ to be 1, if $l_t > 0$, and 0, otherwise. Then, for every $t = 1, \ldots, n$, we have

$$\begin{aligned} l_t &= \boldsymbol{h}_t^T \boldsymbol{\theta} + e_t, \\ e_t &\sim \mathrm{N}(0,1), \end{aligned}$$

where $\boldsymbol{h}_t$ corresponds to the $t$-th column of matrix $\boldsymbol{W}$ and $\boldsymbol{\theta} = (c_{00}, d_{00}, \boldsymbol{d}_1^T, \ldots, \boldsymbol{d}_{J-1}^T)^T$ is the vector of wavelet coefficients, such that $p = n = 2^J$. In this scenario, the dynamic mixture weight $\alpha_t$ is the probability of success of $z_t$, which is equal to the binary regression model given by

$$\alpha_t = \Phi(\boldsymbol{h}_t^T \boldsymbol{\theta}).$$

Notice that the latent variables $l_t$'s are unknown. However, given the vector of parameters $\boldsymbol{\theta}$ and the binary data $\boldsymbol{z} = (z_1, \ldots, z_n)^T$, the distribution of $l_t$ follows a truncated normal distribution. Thus, in this work, as in Albert and Chib (1993), we use the Gibbs sampling algorithm to draw $l_1, \ldots, l_n$ from their posterior distribution, that is,

$$\begin{aligned} l_t|[\ldots] &\sim \mathrm{N}(\boldsymbol{h}_t^T \boldsymbol{\theta}, 1) \text{ truncated at left by 0 if } z_t = 1, \\ l_t|[\ldots] &\sim \mathrm{N}(\boldsymbol{h}_t^T \boldsymbol{\theta}, 1) \text{ truncated at right by 0 if } z_t = 0. \end{aligned} \tag{4.9}$$

For the vector of parameters $\boldsymbol{\theta}$, Albert and Chib (1993) derived the posterior distribution of $\boldsymbol{\theta}$ given $\boldsymbol{z}$ and $\boldsymbol{l}$ under a diffuse prior. They also suggested using a proper conjugate Gaussian distribution for $\boldsymbol{\theta}$. In this work, besides considering these priors, we also use mixtures between a point mass at zero and a unimodal probability density function (Gaussian and Laplace). It is important to emphasize that for all priors, we assume that the entries of $\boldsymbol{\theta}$ are mutually independent. In what follows, we detail these possible choices and describe the necessary adaptations for each one.

### *Diffuse and Gaussian priors*

Assigning a diffuse prior and a Gaussian prior with mean $\mu$ and variance $\tau^{-2}$ to $\theta_t$ (the $t$-th element of vector $\boldsymbol{\theta}$), would yield the following Gaussian distributions as the posterior

conditional distributions of $\theta_t$

$$\theta_t|[\dots] \sim \mathrm{N}(\boldsymbol{w}_t^T \boldsymbol{l}, 1), \tag{4.10}$$

$$\theta_t|[\dots] \sim \mathrm{N}\left(\frac{1}{\tau^2+1}(\tau^2\mu + \boldsymbol{w}_t^T \boldsymbol{l}), \frac{1}{\tau^2+1}\right), \tag{4.11}$$

where $\boldsymbol{w}_t$ is a column-vector corresponding to the $t$-th row of matrix $\boldsymbol{W}$. However, when sampling the wavelet coefficients in (4.10) or (4.11), the posterior draws are corrupted with the gaussian noise $e_t$, compromising the final estimates of $\alpha_t$.

An alternative to deal with this issue is regularizing the coefficients through some thresholding method, such as the BayesThresh procedure. In order to illustrate the benefits of this modification, we present in Figure 17 and Figure 18 the estimates of the $\alpha_t$'s curve with and without the "denosing" procedure. For these simulations, we use the same data sets used in section 4.5, where the reader can find more details. Observe that, for every data set, we cannot have proper estimates for the dynamic mixture weights without denoising the coefficients.

### *Spike and Slab priors*

Given that $\boldsymbol{\theta}$ is a vector of wavelet coefficients, another alternative is to use a mixture between a point mass at zero and some unimodal probability density function as the prior for $\theta_t$. For $t = 2^j + k + 1$, $k = 0, \dots, 2^j - 1$ and $j = 0, \dots, J-1$, this kind of prior can be specified as

$$\theta_t \sim (1-\pi_j)\delta_0(\theta_t) + \pi_j \gamma(\theta_t), \tag{4.12}$$

where $\gamma$ can be the Gaussian distribution or the Laplace distribution as presented in (3.4) and in (3.6), respectively. Following Abramovich, Sapatinas and Silverman (1998), the prior specification is completed by placing a diffuse prior on the scaling coefficient at the coarsest level $c_{00}$, in the first entry of vector $\boldsymbol{\theta}$. Thus, for the mixture priors, $\theta_1 = \boldsymbol{w}_1^T \boldsymbol{l}$.

As discussed in section 3.2, using these mixtures as priors for the $\theta_t$'s allows the posterior medians to act like thresholding rules, equating to zero coefficients thought to be noise. In this scenarios, the posterior distribution of $\theta_t$, where $t = 2^j + k + 1$, $k = 0, \dots, 2^j - 1$ and $j = 0, \dots, J-1$, would be of the form

$$\begin{aligned} \theta_t|[\dots] &\sim (1-\pi_{\text{post}})\delta_0(\theta_t) + \pi_{\text{post}} f_1(\theta_t|\boldsymbol{w}_t^T \boldsymbol{l}), \\ \pi_{\text{post}} &= \frac{\pi_j g(\boldsymbol{w}_t^T \boldsymbol{l})}{\pi_j g(\boldsymbol{w}_t^T \boldsymbol{l}) + (1-\pi_j)\phi(\boldsymbol{w}_t^T \boldsymbol{l})}, \end{aligned} \tag{4.13}$$

where $f_1(\theta_t|\boldsymbol{w}_t^T \boldsymbol{l})$ is the posterior non-null mixture component and $g$ is the convolution between $\gamma$ and the standard normal distribution $\phi$, $g = \gamma \star \phi$.

If $\gamma$ in (4.12) is a Gaussian distribution as in (3.4), besides estimating the sparsity parameter $\pi_j$, it is necessary to estimate the variance of the Gaussian component $v_j^2$. The same happens if $\gamma$ is a Laplace as in (3.6), but instead of estimating the variance, it is necessary to

estimate the scale parameter $a$. As suggested by Johnstone and Silverman (2005a, 2005b), these hyperparameters can be estimated jointly by maximizing the marginal log likelihood function, which, for $j = 0, \ldots, J-1$, is given by

$$\sum_{i=1+2^j}^{2^{j+1}} \log\{(1-\pi_j)\phi(\boldsymbol{w}_i^T \boldsymbol{l}) + \pi_j g(\boldsymbol{w}_i^T \boldsymbol{l})\}.$$

Once selected the hyperparameters, we can sample the vector $\boldsymbol{\theta}$ from the posterior distribution in (4.13). Then, we transform these coefficients back to the data domain through the IDWT and calculate $\alpha_t$ by $\Phi(\boldsymbol{h}_t^T \boldsymbol{\theta})$. The resulting MCMC procedure is detailed in Algorithm 2.

**Algorithm 2** – Gibbs sampling algorithm - Data augmentation - Version 1

Choose number of iterations $N$.
Specify initial values for $\boldsymbol{\mu}^{(0)}$, $\boldsymbol{\tau^2}^{(0)}$, $\boldsymbol{z}^{(0)} = (z_1^{(0)}, \ldots, z_n^{(0)})^T$ and $\boldsymbol{\alpha}^{(0)}$.
**for** $i \leftarrow 1$ to $N$ **do**
    Sample $\mu_1^{(i)} \sim p(\mu_1|[\ldots])$. ▷ See (4.5)
    Sample $\tau_1^{2(i)} \sim p(\tau_1^2|[\ldots])$. ▷ See (4.6)
    Sample $\mu_2^{(i)} \sim p(\mu_2|[\ldots])$. ▷ See (4.5)
    Sample $\tau_2^{2(i)} \sim p(\tau_2^2|[\ldots])$. ▷ See (4.6)
    **if** $\mu_2 < \mu_1$ **then**
        Permute the labeling of pairs $(\mu_k^{(i)}, \tau_k^{2(i)})$.
    **end if**
    Sample $z_t^{(i)} \sim p(z_t|[\ldots])$, for $t = 1, \ldots, n$. ▷ See (4.7)
    Sample $l_t^{(i)} \sim p(l_t|[\ldots])$, for $t = 1, \ldots, n$. ▷ See (4.9)
    Select $\upsilon_j^2 / a$ and $\pi_j$ by marginal maximum likelihood.
    Sample $\theta_t^{(i)} \sim p(\theta_t|[\ldots])$, for $t = 1, \ldots, n$. ▷ See (4.10)/(4.11)/(4.13)
    Calculate $\boldsymbol{\alpha}^{(i)} = \Phi(\boldsymbol{W^T \theta})$. ▷ $\boldsymbol{W}$ is the matrix form of the DWT.
**end for**

## 4.5 Why not sample the "hyperparameters"?

A valid question about the approaches in section 4.4 is why not use the structure of the MCMC algorithms to sample the hyperparameters of the spike and slab priors instead of applying the BayesThresh or the marginal maximum likelihood methods? To respond to this inquiry, in this section, we present two alternative algorithms to perform the approaches discussed formerly and compare their results with the ones obtained by Algorithm 1 and Algorithm 2 under simulated studies.

For the wavelet regression approach, we propose assigning (4.12) as prior to the each wavelet coefficient obtained by the DWT applied to $\boldsymbol{m} = (m_1, \ldots, m_n)^T$ and sampling the hyperparameters from their posterior distributions, based on specific priors. For instance, if the spike and slab prior in (3.4) is used, $j = 0, \ldots, J-1$, we could attribute the following priors to $\upsilon_j^{-2}$

and $\pi_j$:

$$\upsilon_j^{-2} \sim \Gamma(\kappa, \xi),$$
$$\pi_j \sim \mathsf{Beta}(\zeta, \rho).$$

Under these circumstances, we sample their values from the posteriors

$$\upsilon_j^{-2}|[\ldots] \sim \Gamma\left(\kappa + \frac{c_j}{2}, \xi + \frac{\sum_{k=0}^{2^j-1}\theta_{jk}^2}{2}\right), \tag{4.14}$$

$$\pi_j|[\ldots] \sim \mathsf{Beta}(\zeta + n_{j1}, \rho + n_{j0}), \tag{4.15}$$

with $c_j$ being the number of coefficients at level $j$, $j = 0, \ldots, J-1$. Let us write $c_j = n_{j0} + n_{j1}$, where $n_{j0}$ and $n_{j1}$ are null and non-null coefficients, respectively. See section C.3 for the complete derivation of (4.14) and (4.15). Once $\upsilon_j^{-2}$ and $\pi_j$ are drawn, we use their values to sample the denoised coefficients from (3.5). In this scenario, modifying Algorithm 1 to accommodate these alterations results in the procedure described in Algorithm 3.

**Algorithm 3** – Gibbs sampling algorithm - Wavelet regression - Version 2

1: Choose number of iterations $N$.
2: Specify initial values for $\boldsymbol{\mu}^{(0)}$, $\boldsymbol{\tau^2}^{(0)}$, $\boldsymbol{z}^{(0)} = (z_1^{(0)}, \ldots, z_n^{(0)})^T$ and $\boldsymbol{\alpha}^{(0)}$.
3: **for** $i \leftarrow 1$ to $N$ **do**
4: Sample $\mu_1^{(i)} \sim p(\mu_1|[\ldots])$. ▷ See (4.5)
5: Sample $\tau_1^{2(i)} \sim p(\tau_1^2|[\ldots])$. ▷ See (4.6)
6: Sample $\mu_2^{(i)} \sim p(\mu_2|[\ldots])$. ▷ See (4.5)
7: Sample $\tau_2^{2(i)} \sim p(\tau_2^2|[\ldots])$. ▷ See (4.6)
8: **if** $\mu_2 < \mu_1$ **then**
9: Permute the labeling of pairs $(\mu_k^{(i)}, \tau_k^{2(i)})$.
10: **end if**
11: Sample $z_t^{(i)} \sim p(z_t|[\ldots])$, for $t = 1, \ldots, n$. ▷ See (4.7)
12: Calculate $\boldsymbol{m}^{(i)} = (\boldsymbol{y} - \mu_1^{(i)})/(\mu_2^{(i)} - \mu_1^{(i)})$.
13: Compute $\boldsymbol{d}^{(i)} = \boldsymbol{W}\boldsymbol{m}^{(i)}$. ▷ $\boldsymbol{W}$ is the matrix form of the DWT.
14: Sample $\upsilon_j^{-2(i)} \sim p(\upsilon_j^{-2}|[\ldots])$, for $j = 0, \ldots, J-1$. ▷ See (4.14)
15: Sample $\pi_j^{(i)} \sim p(\pi_j|[\ldots])$, for $j = 0, \ldots, J-1$. ▷ See (4.15)
16: Sample $\theta_t^{(i)} \sim p(\theta_t|[\ldots])$, for $t = 1, \ldots, n$. ▷ See (3.5)
17: Calculate $\boldsymbol{\alpha}^{(i)} = \boldsymbol{W}^T\boldsymbol{\theta}^{(i)}$.
18: **end for**

Similar modifications are needed in Algorithm 2 to sample $\upsilon_j^{-2}$ and $\pi_j$ from (4.14) and (4.15), instead of selecting their values by the marginal maximum likelihood method in the 13th step. The resulting MCMC procedure is detailed in Algorithm 4.

Given these alternative algorithms, we simulate data sets and apply each algorithm to them to elect the MCMC procedures that provide the best estimates for both component

**Algorithm 4** – Gibbs sampling algorithm - Data augmentation - Version 2

Choose number of iterations $N$.
Specify initial values for $\boldsymbol{\mu}^{(0)}$, $\boldsymbol{\tau^2}^{(0)}$, $\boldsymbol{z}^{(0)} = (z_1^{(0)}, \ldots, z_n^{(0)})^T$ and $\boldsymbol{\alpha}^{(0)}$.
**for** $i \leftarrow 1$ to $N$ **do**
    Sample $\mu_1^{(i)} \sim p(\mu_1|[\ldots])$. ▷ See (4.5)
    Sample $\tau_1^{2(i)} \sim p(\tau_1^2|[\ldots])$. ▷ See (4.6)
    Sample $\mu_2^{(i)} \sim p(\mu_2|[\ldots])$. ▷ See (4.5)
    Sample $\tau_2^{2(i)} \sim p(\tau_2^2|[\ldots])$. ▷ See (4.6)
    **if** $\mu_2 < \mu_1$ **then**
        Permute the labeling of pairs $(\mu_k^{(i)}, \tau_k^{2(i)})$.
    **end if**
    Sample $z_t^{(i)} \sim p(z_t|[\ldots])$, for $t = 1, \ldots, n$. ▷ See (4.7)
    Sample $l_t^{(i)} \sim p(l_t|[\ldots])$, for $t = 1, \ldots, n$. ▷ See (4.9)
    Sample $\upsilon_j^{-2(i)} \sim p(\upsilon_j^{-2}|[\ldots])$, for $j = 0, \ldots, J-1$. ▷ See (4.14)
    Sample $\pi_j{}^{(i)} \sim p(\pi_j|[\ldots])$, for $j = 0, \ldots, J-1$. ▷ See (4.15)
    Sample $\theta_t^{(i)} \sim p(\theta_t|[\ldots])$, for $t = 1, \ldots, n$. ▷ See (3.5)
    Calculate $\boldsymbol{\alpha}^{(i)} = \Phi(\boldsymbol{W^T \theta})$. ▷ $\boldsymbol{W}$ is the matrix form of the DWT.
**end for**

parameters and dynamic mixture weights. We generate the synthetic data sets following the model defined in (4.4), where $\mu_1 = 0$, $\mu_2 = 2$, $\tau_1^2 = 4$ and $\tau_2^2 = 4$. Concerning the dynamic mixture weight $\alpha_t$, we attribute to it a homogeneous (constant) behavior, setting $\alpha_t = 0.75$, and the dynamic behaviors previously introduced in section 3.3: Parabolic, Sinusoidal, Heavisine, Blocks, and Bumps.

For all data sets, we consider the following independent priors for the component parameters: $\mu_1 \sim N(q_1, s^2)$, $\tau_1^2 \sim \Gamma(0.01, 0.01)$, $\mu_2 \sim N(q_3, s^2)$, and $\tau_2^2 \sim \Gamma(0.01, 0.01)$, where $q_1$ and $q_3$ are the first and third quartiles of the observed data and $s^2$ is the sample variance. The idea of using priors derived from the data is to reduce subjectivity, and, by employing the quartiles, to segregate the data into two groups. With respect to the priors of $\upsilon_j^{-2}$ and $\pi_j$ (Algorithm 3 and Algorithm 4), we set $\upsilon_j^{-2} \sim \Gamma(0.01, 0.01)$, $\pi_j \sim \mathsf{Beta}(1,1)$, $j = 0, \ldots, J-1$.

Furthermore, we implement all four algorithms running $N$ = 6,000 iterations with burn-in $B$ = 1,000 and lags of $L$ = 5, which will result in a sample of $N_{\text{final}}$ = 1,000 observations for each parameter. The point estimation is based on the absolute loss, so the estimates are the medians of the MCMC chains. We present the estimates of the four algorithms for the component parameters in Table 2 – Table 5. Notice that, for every behavior of $\alpha_t$, all four algorithms perform well when estimating these parameters.

Regarding to the estimates of the $\alpha_t$'s, we can see in Figure 19 and Figure 20 that Algorithm 1 and Algorithm 2 outperform Algorithm 3 and Algorithm 4, respectively. In every data set, the pointwise estimates based on Algorithm 1 and Algorithm 2 are closer to the curves than the point estimates provided by Algorithm 3 and Algorithm 4. Furthermore, it is worth noting

that the highest posterior density (HPD) intervals generated by Algorithm 1 and Algorithm 2, although sometimes wider, encompass more satisfactory the real functions than Algorithm 3 and Algorithm 4. In reason of these simple simulation results, we prioritize the implementation of Algorithm 1 and Algorithm 2 in the other studies carried out in this dissertation. In the next chapter, we use more detailed numerical studies to evaluate their performances.

Table 2 – Estimates based on Algorithm 1 (95% HPD credible intervals) for the component parameters $\mu_1, \tau_1^2, \mu_2$ and $\tau_2^2$, using the MCMC samples from all data sets.

| $\alpha_t$ | $\mu_1 = 0$ | $\tau_1^2 = 4$ | $\mu_2 = 2$ | $\tau_2^2 = 4$ |
|---|---|---|---|---|
| Homogeneous behavior | 0.03 (-0.04; 0.12) | 4.31 (3.33; 5.39) | 1.99 (1.95; 2.03) | 4.00 (3.58; 4.52) |
| Parabolic behavior | -0.01 (-0.05; 0.04) | 3.93 (3.44; 4.50) | 1.97 (1.90; 2.02) | 4.11 (3.37; 4.90) |
| Sinusoidal behavior | 0.00 (-0.04; 0.06) | 3.78 (3.25; 4.36) | 2.04 (2.00; 2.09) | 4.13 (3.55; 4.84) |
| Heavisine behavior | 0.02 (-0.03; 0.07) | 4.14 (3.52; 4.76) | 1.98 (1.93; 2.03) | 4.21 (3.60; 4.78) |
| Bumps behavior | -0.02 (-0.05; 0.02) | 4.06 (3.66; 4.52) | 1.89 (1.27; 2.17) | 3.20 (0.80; 5.86) |
| Blocks behavior | -0.02 (-0.06; 0.03) | 3.89 (3.37; 4.53) | 2.01 (1.97; 2.06) | 4.10 (3.52; 4.75) |

Table 3 – Estimates based on Algorithm 2 (95% HPD credible intervals) for the component parameters $\mu_1, \tau_1^2, \mu_2$ and $\tau_2^2$, using the MCMC samples from all data sets.

| $\alpha_t$ | $\mu_1 = 0$ | $\tau_1^2 = 4$ | $\mu_2 = 2$ | $\tau_2^2 = 4$ |
|---|---|---|---|---|
| Homogeneous behavior | 0.01 (-0.06; 0.08) | 4.44 (3.45; 5.37) | 1.99 (1.95; 2.03) | 3.97 (3.45; 4.46) |
| Parabolic behavior | -0.01 (-0.05; 0.04) | 3.89 (3.39; 4.42) | 1.97 (1.91; 2.03) | 4.17 (3.35; 4.99) |
| Sinusoidal behavior | 0.00 (-0.04; 0.06) | 3.80 (3.28; 4.39) | 2.04 (1.99; 2.09) | 4.14 (3.51; 4.77) |
| Heavisine behavior | -0.05 (-0.10; 0.00) | 3.68 (3.14; 4.24) | 2.01 (1.96; 2.06) | 4.27 (3.72; 4.94) |
| Bumps behavior | -0.01 (-0.04; 0.02) | 4.02 (3.66; 4.38) | 2.04 (1.85; 2.19) | 4.79 (2.29; 7.66) |
| Blocks behavior | -0.02 (-0.06; 0.03) | 3.93 (3.40; 4.51) | 2.01 (1.96; 2.05) | 4.05 (3.54; 4.70) |

Table 4 – Estimates based on Algorithm 3 (95% HPD credible intervals) for the component parameters $\mu_1, \tau_1^2, \mu_2$ and $\tau_2^2$, using the MCMC samples from all data sets.

| $\alpha_t$ | $\mu_1 = 0$ | $\tau_1^2 = 4$ | $\mu_2 = 2$ | $\tau_2^2 = 4$ |
|---|---|---|---|---|
| Homogeneous behavior | 0.03 (-0.05; 0.12) | 3.59 (2.70; 4.42) | 2.04 (2.01; 2.09) | 4.36 (3.84; 4.96) |
| Parabolic behavior | -0.01 (-0.05; 0.04) | 3.88 (3.34; 4.45) | 1.97 (1.91; 2.03) | 4.17 (3.42; 5.09) |
| Sinusoidal behavior | 0.00 (-0.06; 0.05) | 3.78 (3.17; 4.33) | 2.04 (1.99; 2.09) | 4.13 (3.57; 4.88) |
| Heavisine behavior | 0.02 (-0.03; 0.07) | 4.07 (3.50; 4.68) | 1.99 (1.94; 2.03) | 4.40 (3.78; 5.05) |
| Bumps behavior | -0.01 (-0.04; 0.02) | 3.96 (3.58; 4.33) | 2.05 (1.87; 2.23) | 5.23 (2.40; 8.35) |
| Blocks behavior | -0.02 (-0.07; 0.03) | 3.91 (3.31; 4.45) | 2.01 (1.97; 2.06) | 4.07 (3.49; 4.71) |

Table 5 – Estimates based on Algorithm 4 (95% HPD credible intervals) for the component parameters $\mu_1, \tau_1^2, \mu_2$ and $\tau_2^2$, using the MCMC samples from all data sets.

| $\alpha_t$ | $\mu_1 = 0$ | $\tau_1^2 = 4$ | $\mu_2 = 2$ | $\tau_2^2 = 4$ |
|---|---|---|---|---|
| Homogeneous behavior | 0.01 (-0.06; 0.08) | 4.41 (3.53; 5.49) | 1.99 (1.95; 2.03) | 3.98 (3.51; 4.52) |
| Parabolic behavior | 0.00 (-0.05; 0.04) | 3.89 (3.42; 4.51) | 1.98 (1.91; 2.04) | 4.21 (3.49; 5.06) |
| Sinusoidal behavior | 0.04 (-0.01; 0.08) | 4.18 (3.54; 4.79) | 2.05 (2.01; 2.11) | 4.00 (3.44; 4.70) |
| Heavisine behavior | -0.05 (-0.10; 0.00) | 3.67 (3.12; 4.22) | 2.01 (1.96; 2.06) | 4.27 (3.72; 4.93) |
| Bumps behavior | 0.01 (-0.02; 0.05) | 4.16 (3.76; 4.61) | 1.89 (1.58; 2.17) | 3.00 (1.43; 5.30) |
| Blocks behavior | -0.02 (-0.07; 0.03) | 3.93 (3.30; 4.48) | 2.01 (1.96; 2.06) | 4.04 (3.48; 4.74) |

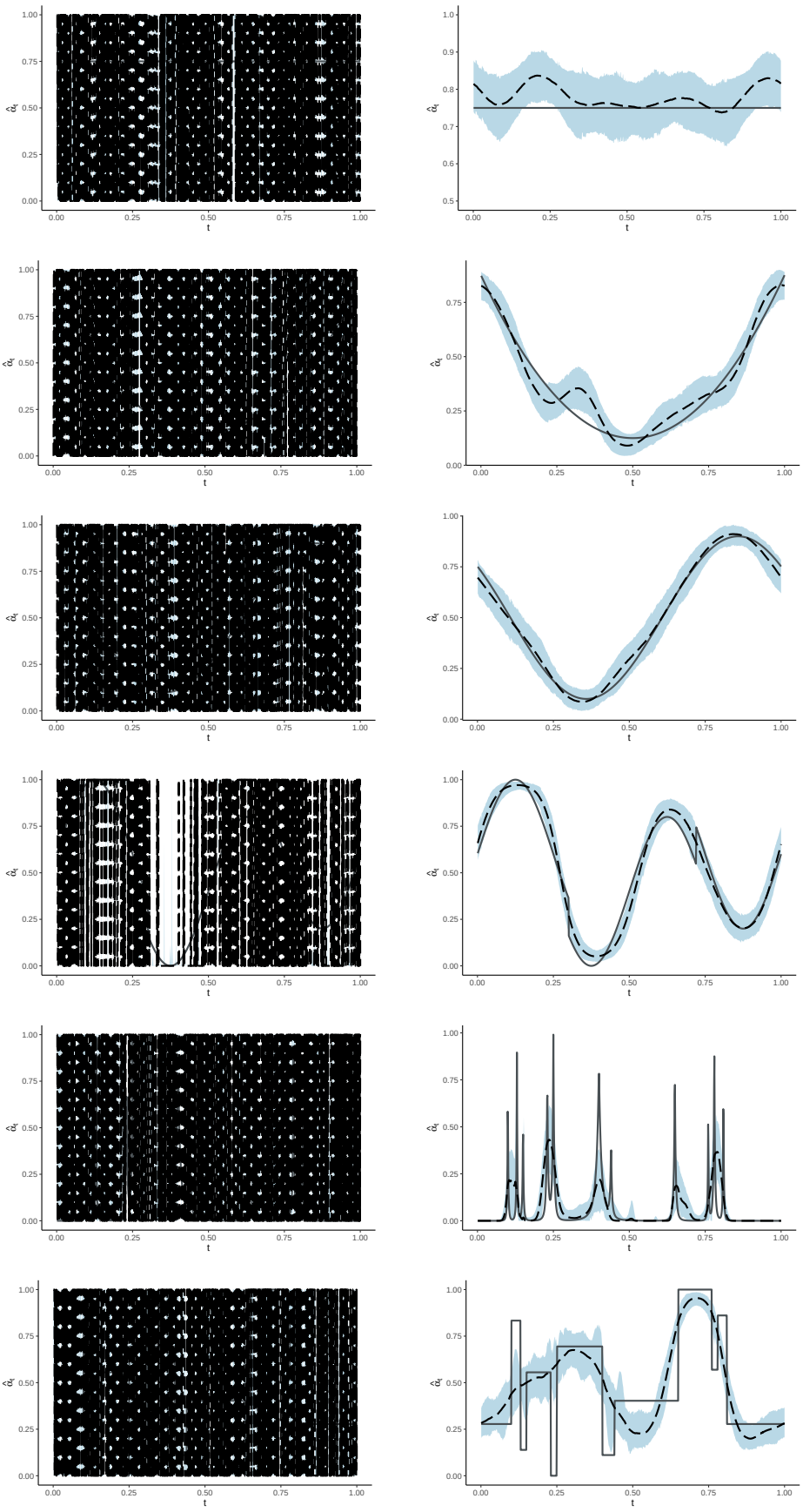


Figure 17 – Pointwise estimates (medians) of the $\alpha_t$'s generated by the data augmentation approach using the diffuse prior to sample the wavelet coefficients. While the first column corresponds to the estimates achieved without the denoising procedure, the second column consists of the estimates provided using the noise reduction by the BayesThresh approach. The full lines correspond to the assigned behavior to $\alpha_t$'s; the dashed lines correspond to the point estimates; and the shaded areas correspond to the 95% HPD intervals.

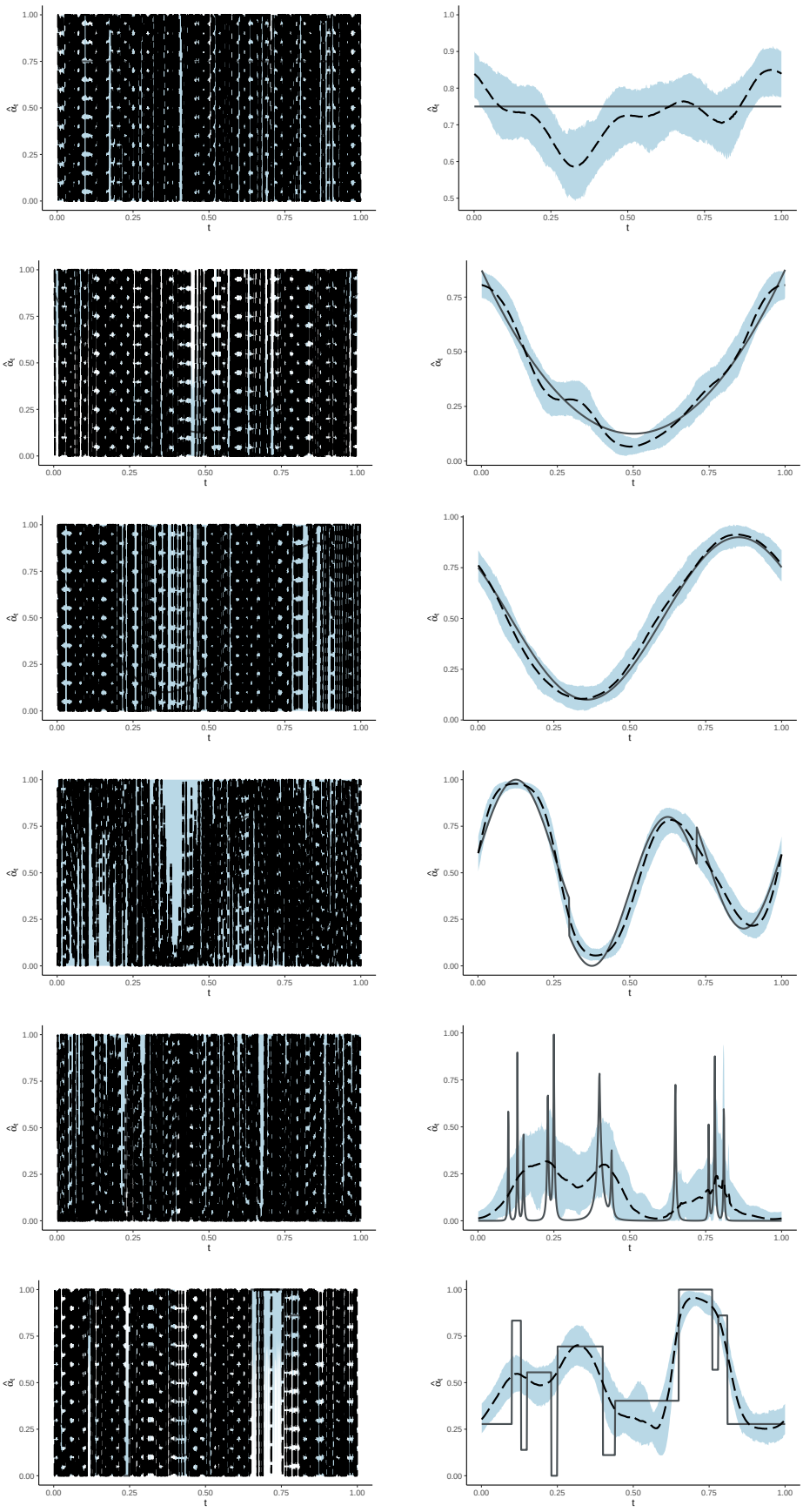


Figure 18 – Pointwise estimates (medians) of the $\alpha_t$'s generated by the data augmentation approach using the Gaussian prior to sample the wavelet coefficients. While the first column corresponds to the estimates achieved without the denoising procedure, the second column consists of the estimates provided using the noise reduction by the BayesThresh approach. The full lines correspond to the assigned behavior to $\alpha_t$'s; the dashed lines correspond to the point estimates; and the shaded areas correspond to the 95% HPD intervals.

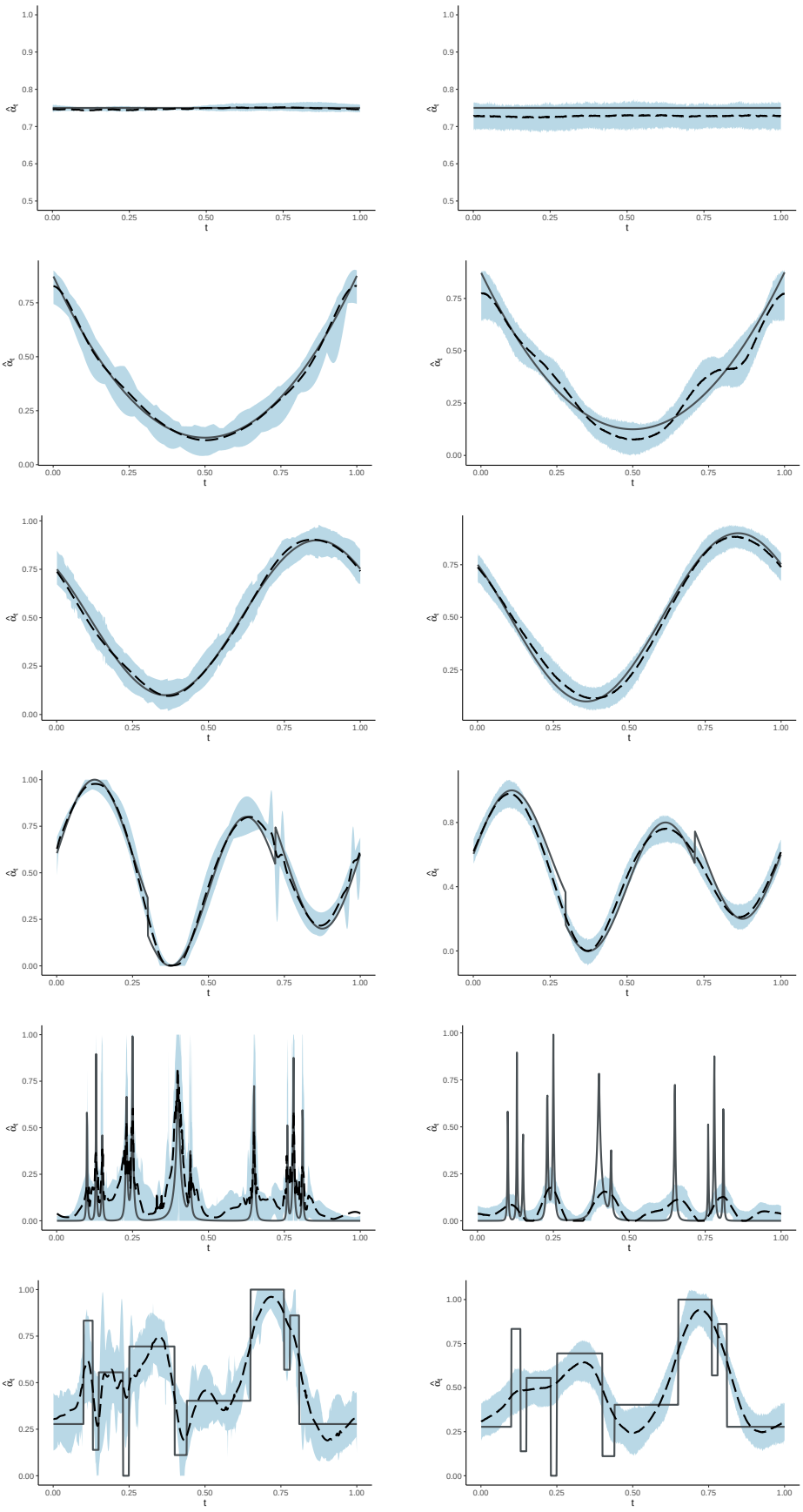


Figure 19 – Pointwise estimates (medians) of the $\alpha_t$'s based on the mixture data sets. The first and second columns represent estimates provided through Algorithm 1 and Algorithm 3, respectively. The full lines correspond to the assigned behavior to $\alpha_t$'s; the dashed lines correspond to the point estimates; and the shaded areas correspond to the 95% HPD intervals.

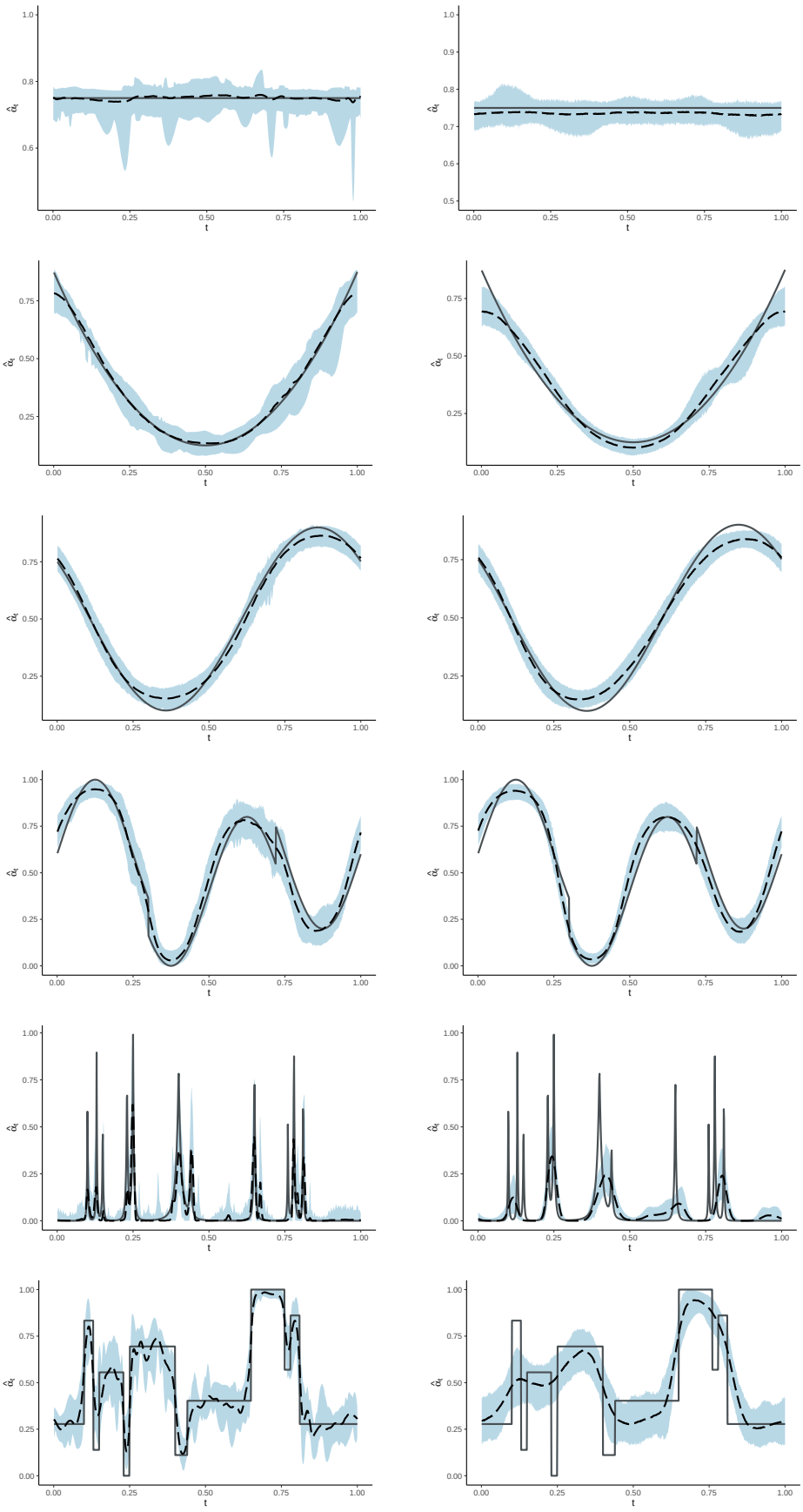


Figure 20 – Pointwise estimates (medians) of the $\alpha_t$'s based on the mixture data sets. The first and second columns represent estimates provided through Algorithm 2 and Algorithm 4, respectively. The full lines correspond to the assigned behavior to $\alpha_t$'s; the dashed lines correspond to the point estimates; and the shaded areas correspond to the 95% HPD intervals.

CHAPTER

# 5

# NUMERICAL STUDIES

In this chapter, we illustrate the performance of the proposed methods through simulated and real data sets. In section 5.1, we conduct Monte Carlo simulations with synthetic data generated by mixtures of two normally distributed groups as defined in (4.4). We use functions with different degrees of smoothness to dictate the dynamic behavior of the mixture weight. In section 5.2, we apply the methods to an array Comparative Genomic Hybridization (aCGH) data set from Glioblastoma cancer studies. In this application, the dynamic mixture weight captures the probability of chromosomal anomalies in genome regions.

## 5.1 Monte Carlo simulations

In our simulated analysis, the synthetic data sets are a mixture of two normally distributed samples with size 1,024, defined as

$$\begin{aligned} y_t &= (1-z_t)x_{1t} + z_t x_{2t}, \\ x_{kt}|\mu_k, \tau_k^2 &\sim \mathrm{N}(\mu_k, \tau_k^{-2}), \quad k = 1,2, \\ z_t|\alpha_t &\sim \mathrm{Bern}(\alpha_t), \quad t = 1,\ldots,1{,}024, \end{aligned}$$

where $\mu_1 = 0$, $\mu_2 = 2$, $\tau_1^2 = 4$ and $\tau_2^2 = 4$. Concerning the dynamic mixture weights, we adopt different curves for $\alpha_t$ to see how the methods perform with smoother and rougher functions. We consider the same behaviors used previously in section 4.5 (where the reader can find more details): constant, parabolic, sinusoidal, heavisine, blocks, and bumps.

For all six behaviors of $\alpha_t$, we conduct Monte Carlo simulations with 1,000 replicates. We consider the Wavelet regression discussed in subsection 4.4.1, hereafter denoted by WR, and the data augmentation method explained in subsection 4.4.2, hereafter called DA. We run Monte Carlo experiments to the DA approach considering all four priors for the distribution of the wavelet coefficients discussed in subsection 4.4.2, namely: diffuse prior, Gaussian prior,

spike and slab prior with Gaussian slab (SSG), and spike and slab prior with Laplace slab (SSL). From now on, we refer to the latter two priors with their respective acronyms: SSG and SSL.

We implement Algorithm 1 or Algorithm 2 for each replication of data running N = 6,000 iterations with burn-in B = 1,000 and lags of L = 5. We consider the following independent priors for the component parameters: $\mu_1 \sim N(q_1, s^2)$, $\tau_1^2 \sim \Gamma(0.01, 0.01)$, $\mu_2 \sim N(q_3, s^2)$, and $\tau_2^2 \sim \Gamma(0.01, 0.01)$, where $q_1$ and $q_3$ are the first and third quartiles, respectively, of the observed data and $s^2$ is the sample variance.

As in section 4.5, the point estimation is based on absolute loss, so the estimates of the component parameters and mixture weights are the medians of the MCMC chains. To evaluate the performance of these point estimates, we compute their averages and the 95% HPD intervals based on the point estimates from the Monte Carlo replicates. The results for the six scenarios of mixture weight are presented below.

1. Constant behavior: in Table 6, we present the estimates of the component parameters considering simulations where the dynamic mixture weights are constant (homogeneous). In general, every approach considered in this work provides good results in estimating these parameters. Furthermore, all 95% HPD intervals encompass the true values of $\mu_1$, $\mu_2$, $\tau_1^2$ and $\tau_2^2$. Regarding the mixture weight estimates, Figure 21 exhibits the results reached by the methods. Except for the DA with the Gaussian prior, which seems to provide evidence of underestimation, all the other approaches provide estimates that match the real curve.

2. Parabolic behavior: we present the estimates of component parameters for the simulations with $\alpha_t$ following the parabolic behavior in Table 7. Note that the methods provide good estimates for the parameters, especially for $\mu_1$ and $\mu_2$. Moreover, all 95% HPD intervals encompass the true values of $\mu_1$, $\mu_2$, $\tau_1^2$ and $\tau_2^2$. Figure 22 shows the estimates for $\alpha_t$'s considering these simulations. We emphasize that all methods' pointwise estimates follow the parabolic curve. Even though none of the approaches mimic the borders of the curve flawlessly, all highest posterior density (HPD) intervals encompass these particular regions of the function.

3. Sinusoidal behavior: for the data sets generated from the sinusoidal weight, we present the estimates of the component parameters in Table 8. It is worth noting that the methods provide good results in estimating these parameters, with the estimates for $\mu_1$ and $\mu_2$ even coinciding with the true parameter values. This also happens for $\tau_1^2$ and $\tau_2^2$ when the prior assigned in the DA approach is the SSG prior. Concerning the estimates for the dynamic mixture weight, in Figure 23, we can see that all the methods succeed in properly mimicking the shapes of the sinusoidal curve.

4. Heavisine behavior: Table 9 exhibits the estimates of $\mu_1$, $\mu_2$, $\tau_1^2$ and $\tau_2^2$ obtained from

the simulations with the heavisine behavior for the dynamic mixture weights. We see that every approach provides estimates for the component parameters close to the true parameter values. Regarding the dynamic behavior of $\alpha_t$ in Figure 24, note that all irregular patterns of the curve are encompassed by the HPD intervals of the WR approach and the DA approach with both spike and slab priors.

5. Bumps behavior: the bumps is the rougher function among all the test functions considered in this work. Therefore, it is responsible for pointing out the most capable methods to provide the best estimates. From Table 10, we see that the DA with the Diffuse prior and the DA with the spike and slab priors (SSG and SSL) are better at estimating the component parameters than the other two approaches (WR and DA with Gaussian prior). Furthermore, in light of Figure 25, the outperformance of the DA with both, SSG and SSL priors, is even more evident. These methods not only estimate the locals where the bumps occur but also their estimates follow the sharp shape of the curves and describe the null values more satisfactorily.

6. Blocks behavior: In Table 11, we present the estimates of the component parameters considering simulations where the blocks function is used to determine the behavior of the dynamic mixture weights. We can see that all approaches provide good estimates for the parameters, with HPD intervals encompassing the true values of $\mu_1$, $\mu_2$,$\tau_1^2$, and $\tau_2^2$. However, when we analyze the estimates provided for the mixture weights, presented in Figure 26, we see a clear outperformance of the DA approach with both, SSG and SSL priors. Their pointwise estimates mimic the shapes of the blocks curve, properly following it in the discontinuity regions. Furthermore, their HPD intervals succeed at encompassing the entire curve.

In general, the Monte Carlo simulations illustrate that all approaches described in Chapter 4 can provide good estimates for the component parameters. However, when it comes to the performance of the approaches at estimating the dynamic weights, it varies according to the smoothness of the mixture weight.

It is evident that the DA approach with the spike and slab priors, i.e., the SSG and the SSL priors, are more successful than the other approaches at estimating the dynamic weights when the functions are "rougher" (bumps and blocks). When the function is smooth, all approaches provide estimates close to the real curve.

Nonetheless, these simulations are conducted under a controlled setting, where the function and the parameters are explicitly established. Therefore, for a deeper evaluation of the approaches, it is also necessary to apply them to real data sets, where the parameters are unknown and the observations are not generated according to the model under analysis. We do this in the next section.

Table 6 – Averages of the point estimates (95% HPD credible intervals) for the component parameters $\mu_1, \tau_1^2, \mu_2$ and $\tau_2^2$, based on 1,000 replications of data sets whose mixture weights follow the constant behavior.

| Method | $\mu_1 = 0$ | $\tau_1^2 = 4$ | $\mu_2 = 2$ | $\tau_2^2 = 4$ |
|---|---|---|---|---|
| WR | 0.00 (-0.08;0.09) | 3.83 (3.21;4.74) | 2.00 (1.96;2.03) | 4.04 (3.49;4.48) |
| DA (Diffuse prior) | 0.01 (-0.04;0.04) | 3.97 (3.10;4.44) | 2.00 (1.98;2.02) | 4.00 (3.65;4.33) |
| DA (Gaussian prior) | 0.01 (-0.08;0.10) | 3.92 (3.04;4.29) | 2.00 (1.97;2.02) | 4.04 (3.72;4.47) |
| DA (SSG prior) | 0.00 (-0.08;0.09) | 3.96 (3.09;4.84) | 2.00 (1.97;2.04) | 4.06 (3.52;4.51) |
| DA (SSL prior) | 0.00 (-0.09;0.09) | 4.02 (3.03;4.99) | 2.00 (1.96;2.05) | 4.01 (3.56;4.55) |

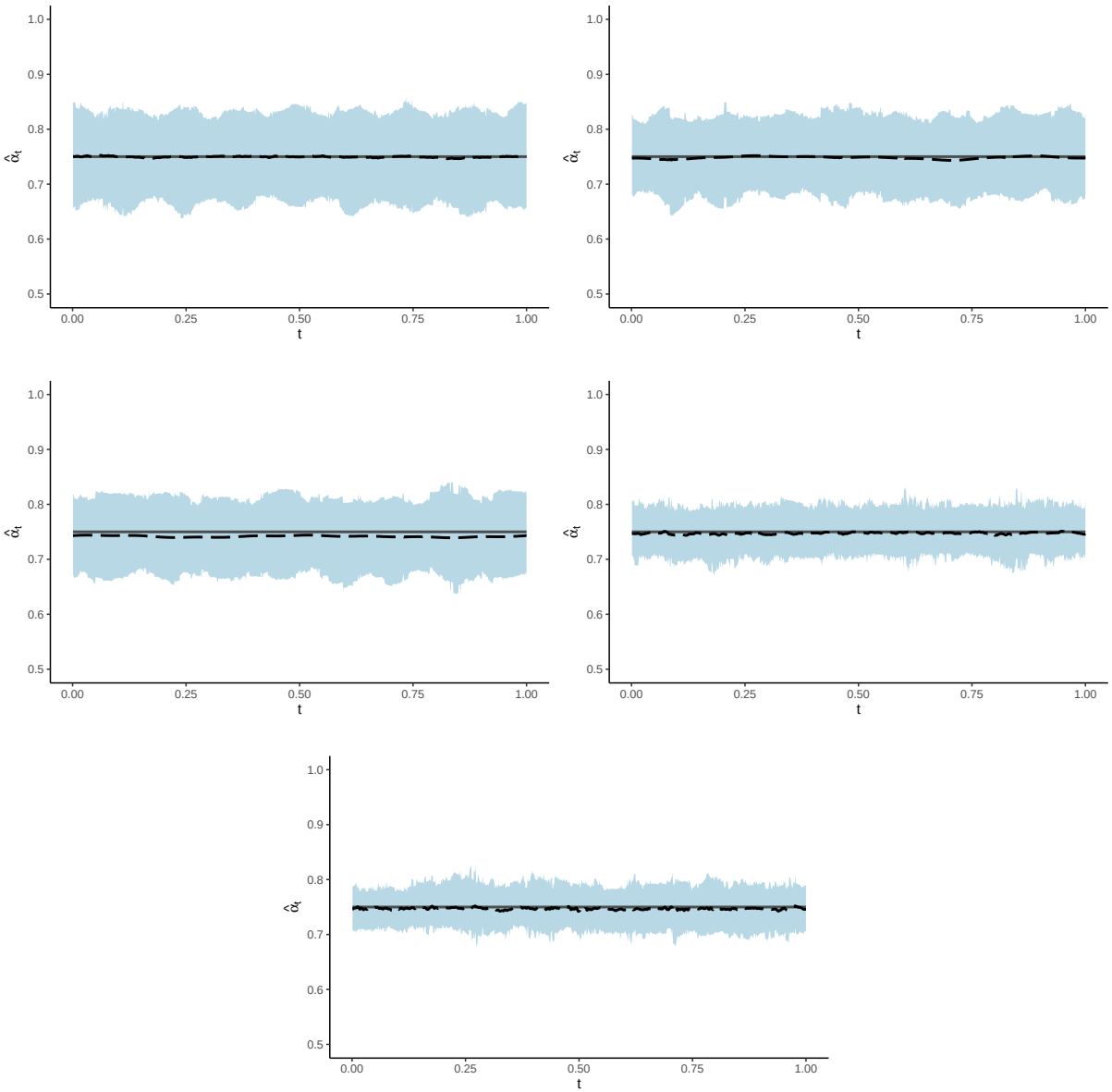



Figure 21 – Estimates of the $\alpha_t$'s provided by the WR approach (top left); the DA approach: with Diffuse prior (top right); Gaussian prior (middle left); SSG prior (middle right); and SSL prior (bottom). The full lines correspond to the constant behavior of $\alpha_t$'s, the dashed lines correspond to the average of the pointwise estimates and the shaded areas correspond to the 95% HPD intervals.

Table 7 – Averages of the point estimates (95% HPD credible intervals) for the component parameters $\mu_1, \tau_1^2, \mu_2$ and $\tau_2^2$, based on 1,000 replications of data sets whose mixture weights follow the parabolic behavior.

| Method | $\mu_1 = 0$ | $\tau_1^2 = 4$ | $\mu_2 = 2$ | $\tau_2^2 = 4$ |
|---|---|---|---|---|
| WR | 0.00 (-0.04;0.04) | 3.99 (3.44;4.54) | 2.00 (1.94;2.06) | 4.02 (3.23;4.69) |
| DA (Diffuse prior) | 0.00 (-0.04;0.04) | 4.02 (3.47;4.56) | 2.00 (1.94;2.05) | 4.00 (3.33;4.76) |
| DA (Gaussian prior) | 0.00 (-0.05;0.04) | 4.03 (3.45;4.52) | 2.00 (1.94;2.06) | 3.97 (3.31;4.70) |
| DA (SSG prior) | 0.00 (-0.03;0.04) | 3.98 (3.30;4.40) | 2.00 (1.94;2.07) | 4.03 (3.40;4.86) |
| DA (SSL prior) | 0.00 (-0.05;0.05) | 4.06 (3.55;4.61) | 2.00 (1.94;2.07) | 3.97 (2.92;4.89) |

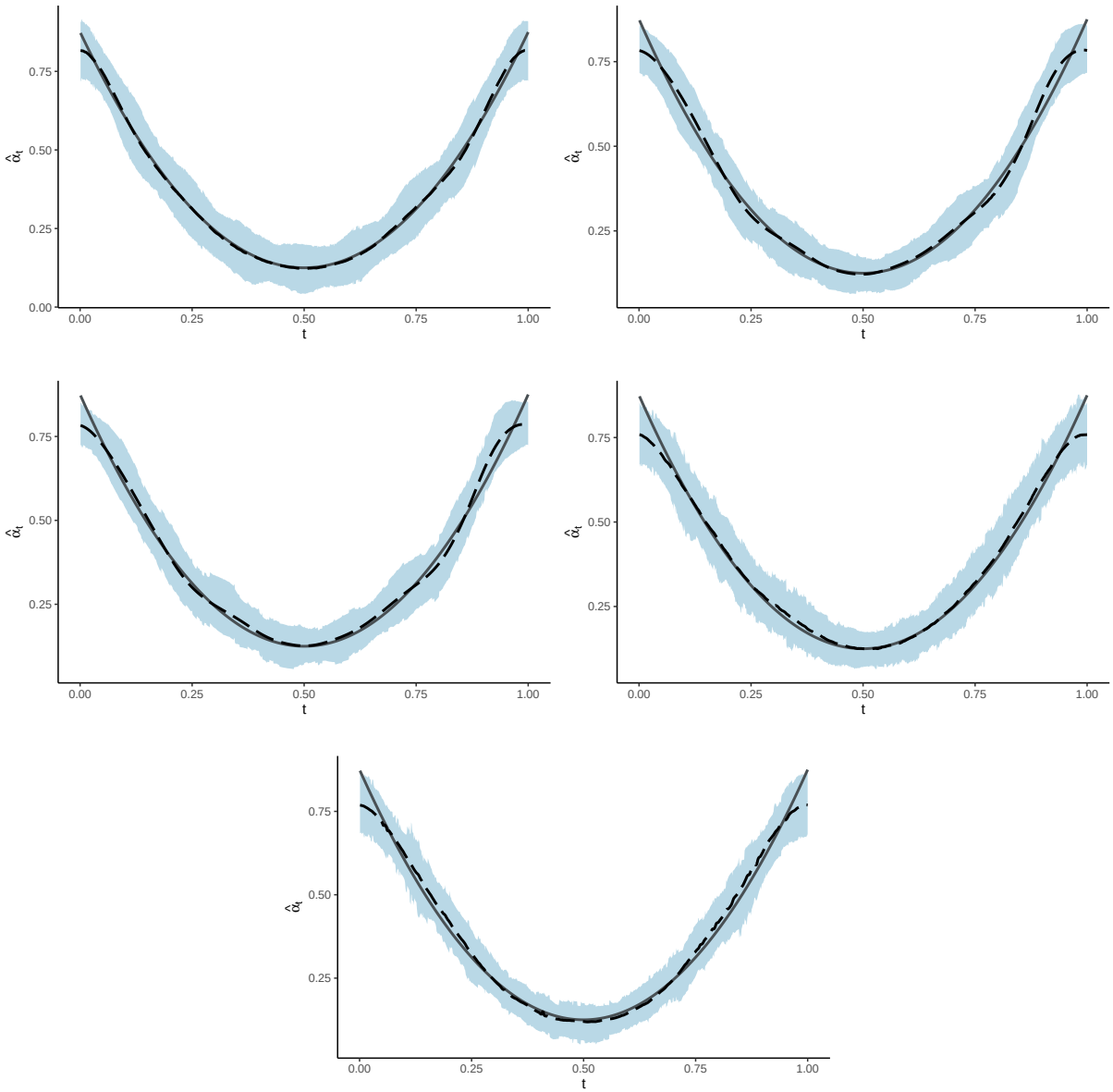



Figure 22 – Estimates of the $\alpha_t$'s provided by the WR approach (top left); the DA approach: with Diffuse prior (top right); Gaussian prior (middle left); SSG prior (middle right); and SSL prior (bottom). The full lines correspond to the parabolic behavior of $\alpha_t$'s, the dashed lines correspond to the average of the pointwise estimates and the shaded areas correspond to the 95% HPD intervals.

Table 8 – Averages of the point estimates (95% HPD credible intervals) for the component parameters $\mu_1, \tau_1^2, \mu_2$ and $\tau_2^2$, based on 1,000 replications of data sets whose mixture weights follow the sinusoidal behavior.

| Method | $\mu_1 = 0$ | $\tau_1^2 = 4$ | $\mu_2 = 2$ | $\tau_2^2 = 4$ |
|---|---|---|---|---|
| WR | 0.00 (-0.05;0.04) | 3.94 (3.48;4.70) | 2.00 (1.94;2.06) | 3.95 (3.48;4.61) |
| DA (Diffuse prior) | 0.00 (-0.04;0.04) | 3.98 (3.50;4.61) | 2.00 (1.94;2.04) | 4.00 (3.50;4.68) |
| DA (Gaussian prior) | 0.00 (-0.05;0.05) | 4.01 (3.37;4.65) | 2.00 (1.96;2.04) | 3.98 (3.40;4.46) |
| DA (SSG prior) | 0.00 (-0.04;0.06) | 4.00 (3.58;4.65) | 2.00 (1.95;2.04) | 4.00 (3.40;4.59) |
| DA (SSL prior) | 0.00 (-0.05;0.05) | 4.05 (3.50;4.62) | 2.00 (1.95;2.04) | 3.99 (3.49;4.50) |

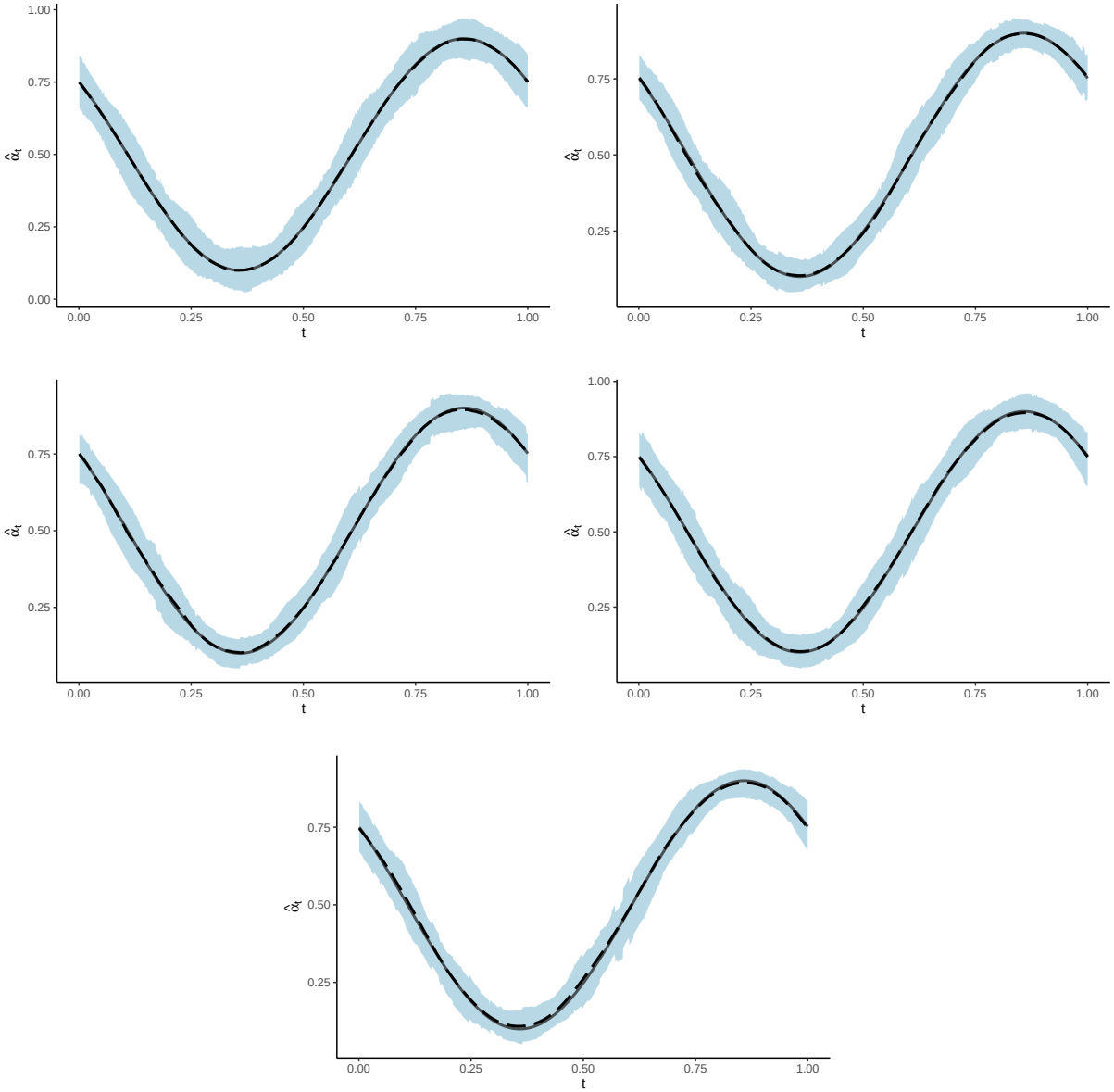



Figure 23 – Estimates of the $\alpha_t$'s provided by the WR approach (top left); the DA approach: with Diffuse prior (top right); Gaussian prior (middle left); SSG prior (middle right); and SSL prior (bottom). The full lines correspond to the sinusoidal behavior of $\alpha_t$'s, the dashed lines correspond to the average of the pointwise estimates and the shaded areas correspond to the 95% HPD intervals.

Table 9 – Averages of the point estimates (95% HPD credible intervals) for the component parameters $\mu_1, \tau_1^2, \mu_2$ and $\tau_2^2$, based on 1,000 replications of data sets whose mixture weights follow the heavisine behavior.

| Method | $\mu_1 = 0$ | $\tau_1^2 = 4$ | $\mu_2 = 2$ | $\tau_2^2 = 4$ |
|---|---|---|---|---|
| WR | 0.00 (-0.05;0.05) | 3.99 (3.36;4.59) | 2.00 (1.95;2.05) | 3.97 (3.40;4.62) |
| DA (Diffuse prior) | 0.00 (-0.04;0.05) | 4.02 (3.40;4.56) | 2.00 (1.95;2.05) | 4.00 (3.39;4.61) |
| DA (Gaussian prior) | 0.00 (-0.04;0.05) | 4.02 (3.34;4.65) | 2.00 (1.95;2.05) | 4.00 (3.35;4.49) |
| DA (SSG prior) | 0.00 (-0.04;0.06) | 4.03 (3.49;4.63) | 2.00 (1.95;2.05) | 4.03 (3.50;4.67) |
| DA (SSL prior) | 0.00 (-0.05;0.04) | 4.02 (3.36;4.59) | 2.00 (1.94;2.04) | 4.00 (3.48;4.59) |

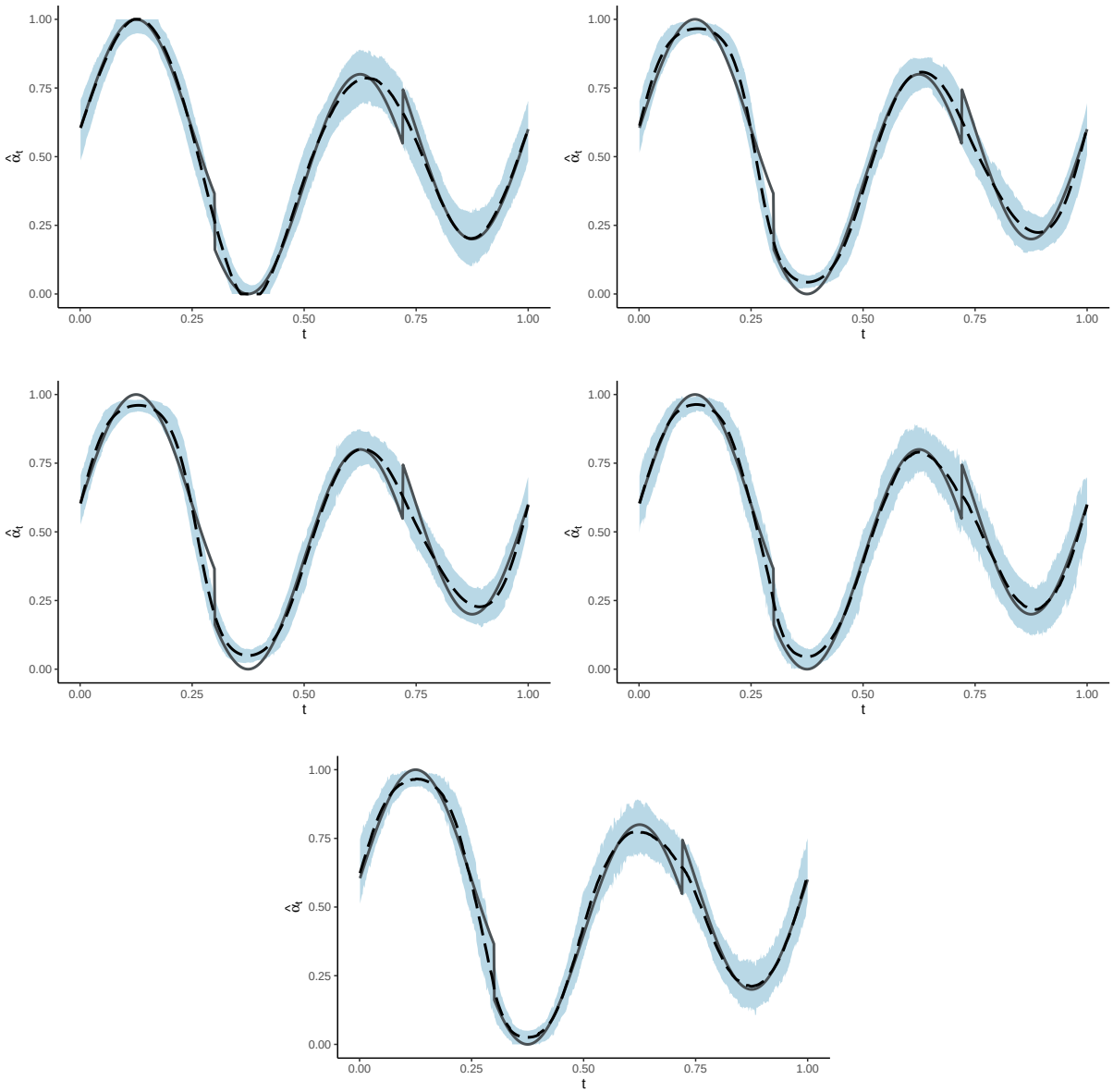



Figure 24 – Estimates of the $\alpha_t$'s provided by the WR approach (top left); the DA approach: with Diffuse prior (top right); Gaussian prior (middle left); SSG prior (middle right); and SSL prior (bottom). The full lines correspond to the heavisine behavior of $\alpha_t$'s, the dashed lines correspond to the average of the pointwise estimates and the shaded areas correspond to the 95% HPD intervals.

Table 10 – Averages of the point estimates (95% HPD credible intervals) for the component parameters $\mu_1, \tau_1^2, \mu_2$ and $\tau_2^2$, based on 1,000 replications of data sets whose mixture weights follow the bumps behavior.

| Method | $\mu_1 = 0$ | $\tau_1^2 = 4$ | $\mu_2 = 2$ | $\tau_2^2 = 4$ |
|---|---|---|---|---|
| WR | -0.02 (-0.06;0.03) | 4.22 (3.70;4.80) | 1.27 (0.71;2.07) | 1.19 (0.95;4.52) |
| DA (Diffuse prior) | 0.00 (-0.04;0.02) | 4.00 (3.60;4.37) | 1.92 (1.69;2.23) | 3.69 (1.54;6.24) |
| DA (Gaussian prior) | -0.02 (-0.06;0.01) | 4.34 (3.75;4.79) | 1.34 (0.76;2.05) | 1.81 (0.78;4.59) |
| DA (SSG prior) | 0.00 (-0.04;0.02) | 4.01 (3.59;4.38) | 1.90 (1.60;2.15) | 3.62 (1.06;6.45) |
| DA (SSL prior) | 0.00 (-0.04;0.03) | 3.96 (3.34;4.53) | 1.89 (1.43;2.22) | 3.66 (0.71;6.56) |

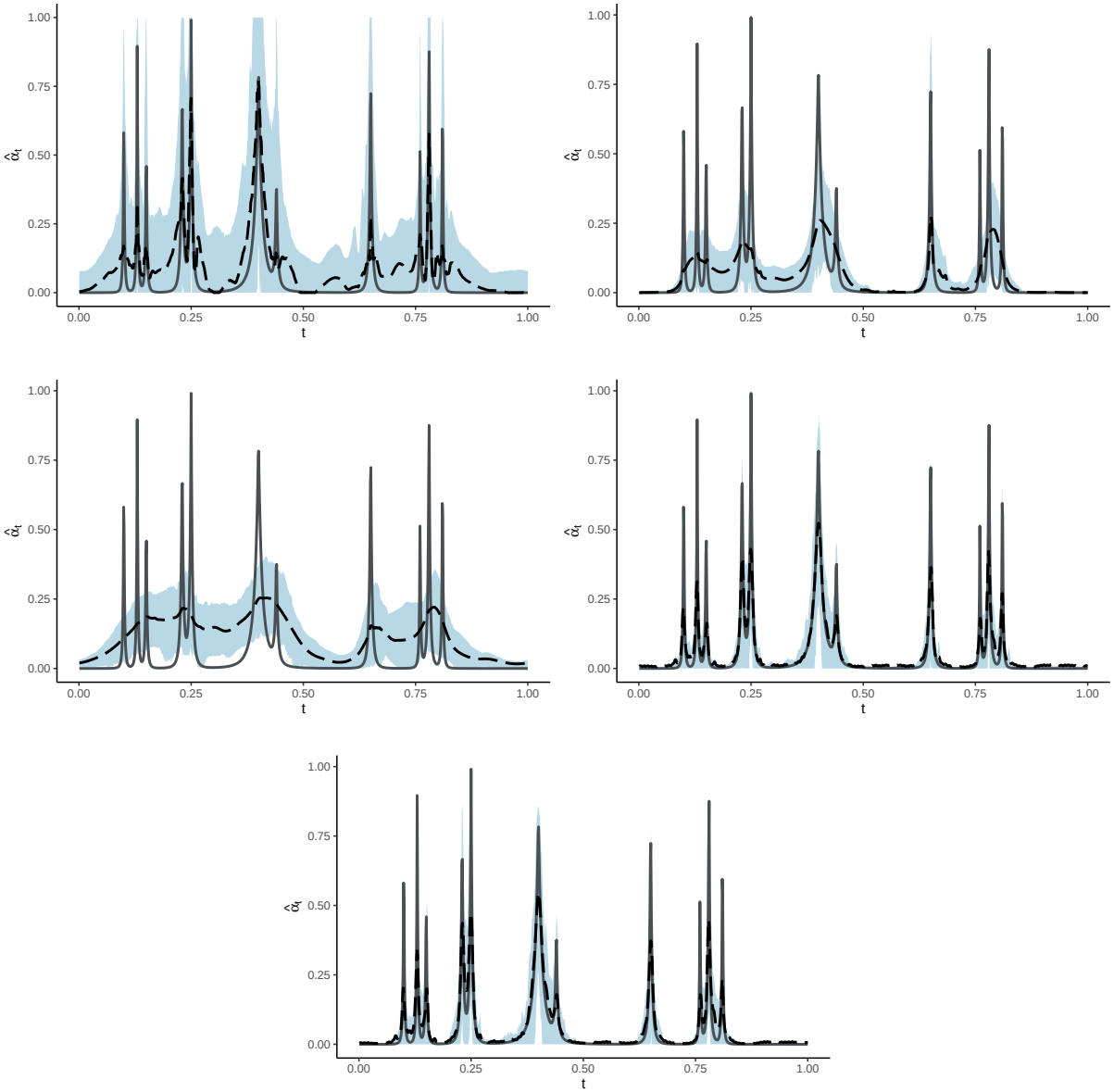



Figure 25 – Estimates of the $\alpha_t$'s provided by the WR approach (top left); the DA approach: with Diffuse prior (top right); Gaussian prior (middle left); SSG prior (middle right); and SSL prior (bottom). The full lines correspond to the bumps behavior of $\alpha_t$'s, the dashed lines correspond to the average of the pointwise estimates and the shaded areas correspond to the 95% HPD intervals.

Table 11 – Averages of the point estimates (95% HPD credible intervals) for the component parameters $\mu_1, \tau_1^2, \mu_2$ and $\tau_2^2$, based on 1,000 replications of data sets whose mixture weights follow the blocks behavior.

| Method | $\mu_1 = 0$ | $\tau_1^2 = 4$ | $\mu_2 = 2$ | $\tau_2^2 = 4$ |
|---|---|---|---|---|
| WR | 0.00 (-0.05;0.05) | 4.00 (3.44;4.71) | 2.00 (1.95;2.05) | 4.03 (3.49;4.69) |
| DA (Diffuse prior) | 0.00 (-0.05;0.05) | 3.94 (3.30;4.39) | 2.00 (1.95;2.05) | 4.06 (3.43;4.56) |
| DA (Gaussian prior) | 0.00 (-0.05;0.05) | 3.97 (3.36;4.48) | 2.00 (1.95;2.04) | 4.00 (3.52;4.64) |
| DA (SSG prior) | 0.00 (-0.04;0.06) | 4.06 (3.41;4.71) | 2.00 (1.95;2.06) | 4.00 (3.50;4.63) |
| DA (SSL prior) | 0.02 (-0.15;0.05) | 3.91 (3.40;5.60) | 1.95 (1.28;2.07) | 3.85 (0.82;4.76) |

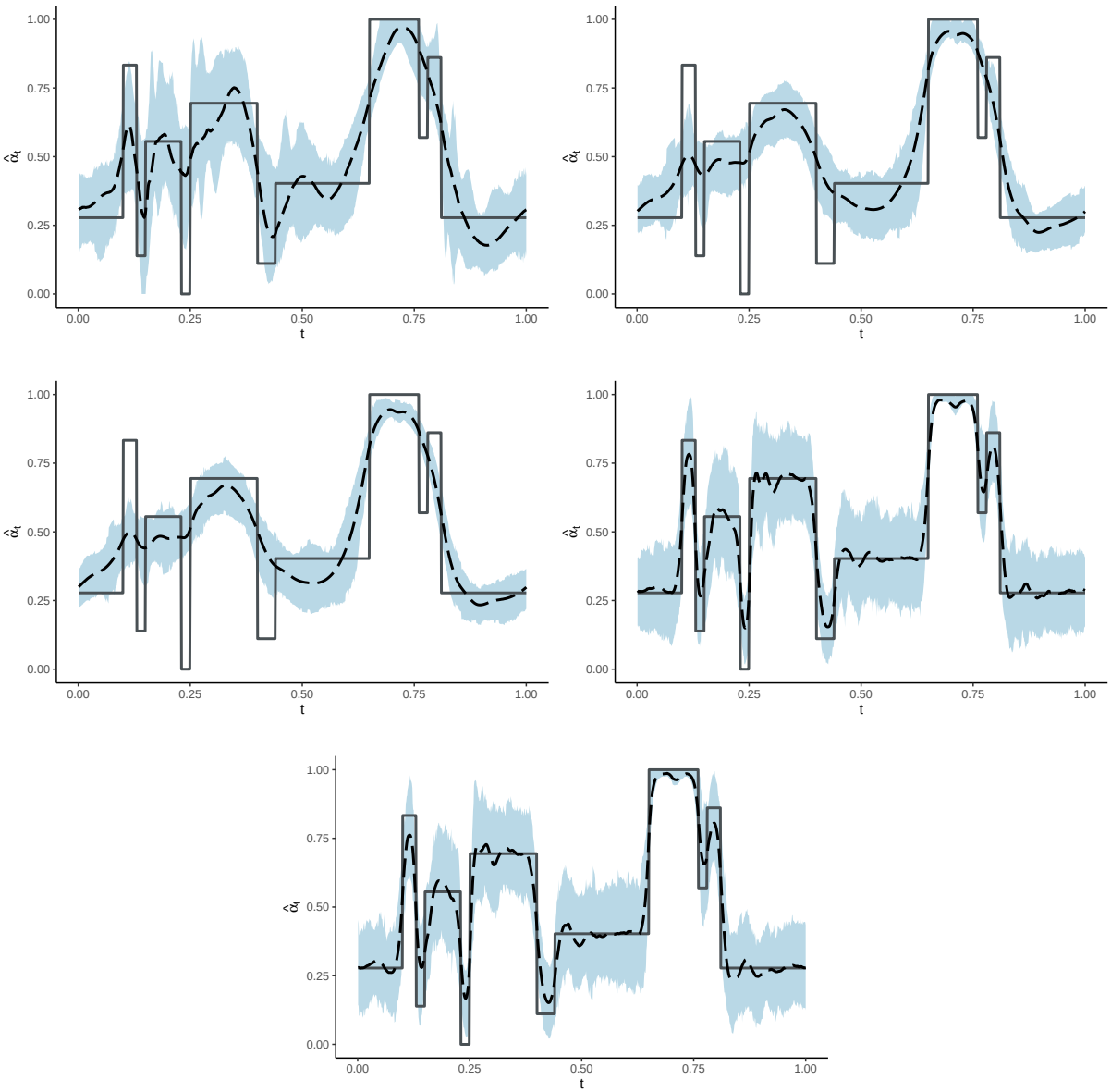



Figure 26 – Estimates of the $\alpha_t$'s provided by the WR approach (top left); the DA approach: with Diffuse prior (top right); Gaussian prior (middle left); SSG prior (middle right); and SSL prior (bottom). The full lines correspond to the blocks behavior of $\alpha_t$'s, the dashed lines correspond to the average of the pointwise estimates and the shaded areas correspond to the 95% HPD intervals.

## 5.2 Application to an array Comparative Genomic Hybridization data set

In general, tumors develop under genomic imbalances, such as deletions, amplifications, and other structural rearrangements of chromosomes and chromosomal segments. Therefore, detecting those imbalances, which are related to alterations in the DNA copy number, is fundamental for cancer diagnosis (PINKEL; ALBERTSON, 2005).

Array comparative genomic hybridization (aCGH) is a technique that, under certain conditions, provides intense fluorescence signals which are used for detecting aberrations in DNA copy number. In aCGH, test and reference genomic DNA samples (the latter being isolated from normal cells) are labeled differently using fluorescent dyes. These samples are then mixed and co-hybridized onto an array CGH platform. Each array spot represents a unique chromosome locus. If the amount of DNA copy number in the test sample is the same as the reference sample, a certain fluorescent signal intensity is measured. Thus, decreases or increases in this intensity ratio can indicate copy number losses or gains in the genome of the test cells (HSU *et al.*, 2005; JANKOVIC *et al.*, 2022). This process is depicted in Figure 27.

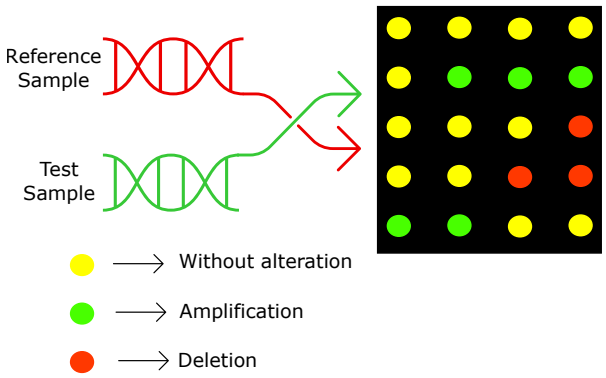



Figure 27 – A diagrammatic depiction of array CGH. Two genomic libraries are differentially labeled and hybridized into a microarray. The fluorescent ratios on each spot are calculated and typically normalized so that the median ratio for the genome is set to some standard value (1.0 on a linear scale or 0.0 on a logarithmic scale). If, to a given spot, both samples hybridize equally, no copy number variation is being detected in the DNA region that matches this spot. However, if the test sample hybridizes more or less intensely than the reference sample, there is some copy number variation in that region (PINKEL; ALBERTSON, 2005; JANKOVIC *et al.*, 2022).

There are several approaches for analyzing array CGH data sets. Lai *et al.* (2005) compared 11 methods found in the literature for data analysis. The authors use simulated and real data from cancer studies to evaluate the performances of these methods in identifying regions of the genome where copy number alterations occurred.

In this work, we use an aCGH data set used in Lai *et al.* (2005) to illustrate how the methods discussed in Chapter 4 can be effective in this kind of application. Figure 28 presents this data set of size $n = 193$. The data consists of an array comparative genomic hybridization

obtained from Glioblastoma Multiforme (GBM) tumor cell samples. GBM is an aggressive malignant brain tumor with a median survival time (MST) of one year (LAI *et al.*, 2005). We aim to detect large signal values that may indicate chromosomal anomalies, as the identification of regions where these proportions are high is important to understanding the pathogenesis.

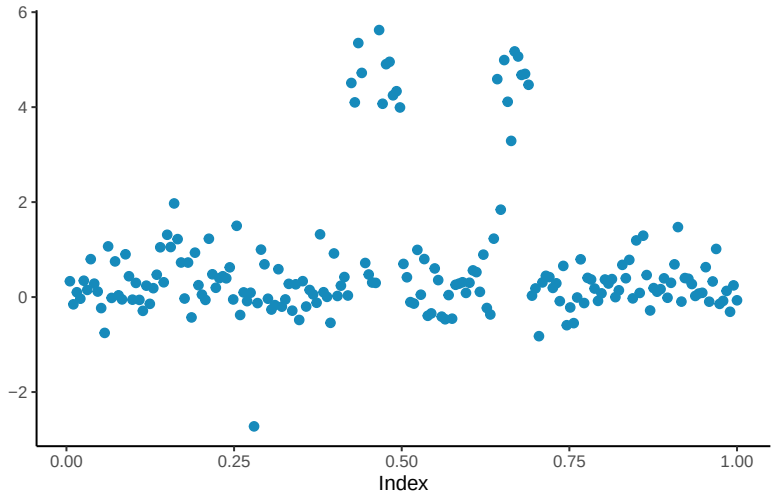



Figure 28 – Observed aCGH values. This data consists of the log-ratios of normalized intensities from disease *versus* control samples, indexed by the physical position of the probes on the genome (LAI *et al.*, 2005).

Algorithm 1 and Algorithm 2 are implemented running N = 6,000 iterations with burn-in B = 1,000 and lags of L = 5. As in section 5.1, we also consider the following independent priors for the component parameters: $\mu_1 \sim N(q_1, s^2)$, $\tau_1^2 \sim \Gamma(0.01, 0.01)$, $\mu_2 \sim N(q_3, s^2)$, and $\tau_2^2 \sim \Gamma(0.01, 0.01)$, where $q_1$ and $q_3$ are the first and third quartiles of the observed data and $s^2$ is the sample variance.

When examining the convergence diagnostic plots for the mixture parameters (see Appendix D) the chains appear to be well mixed and representative of the posterior distribution. We present the estimates of these parameters in Table 12. The estimates of $\mu_1$ and $\mu_2$ are consistent with Killick and Eckley (2014), which estimates the change points in this aCGH data set using a pruned exact linear time (PELT) algorithm.

Table 12 – Medians (95% HPD credible intervals) for the component parameters $\mu_1, \tau_1^2, \mu_2$ and $\tau_2^2$ of the aCGH data set, based on the MCMC samples.

| Method | $\mu_1$ | $\tau_1^2$ | $\mu_2$ | $\tau_2^2$ |
|---|---|---|---|---|
| WR | 0.25 (0.17; 0.32) | 3.51 (2.76; 4.39) | 4.56 (4.28; 4.83) | 3.08 (1.12; 5.59) |
| DA (Diffuse prior) | 0.25 (0.17; 0.33) | 3.49 (2.83; 4.25) | 4.59 (4.33; 4.86) | 3.29 (1.17; 5.57) |
| DA (Gaussian prior) | 0.25 (0.17; 0.33) | 3.49 (2.83; 4.29) | 4.59 (4.34; 4.89) | 3.24 (1.39; 5.48) |
| DA (SSG prior) | 0.25 (0.18; 0.34) | 3.50 (2.80; 4.30) | 4.57 (4.32; 4.81) | 3.28 (1.35; 5.46) |
| DA (SSL prior) | 0.23 (0.16; 0.30) | 5.18 (3.97; 6.51) | 3.61 (2.74; 4.55) | 0.26 (0.13; 0.44) |

Figure 29 exhibits the estimates of the dynamic weights provided by all of the approaches analyzed in this work. The peaks of the curves indicate a higher probability that there are

amplifications in the DNA copy number at that chromosomal position. Note that the WR approach and the DA with SSG and SSL priors detect at least three high-amplitude amplifications, whereas the DA with the diffuse and Gaussian priors combine the first two amplifications detected by the other methods in a single one. According to Lai *et al.* (2005), mapping the indices of these amplifications to their chromosomal positions suggests that there are likely two separate aberrations, not just one.

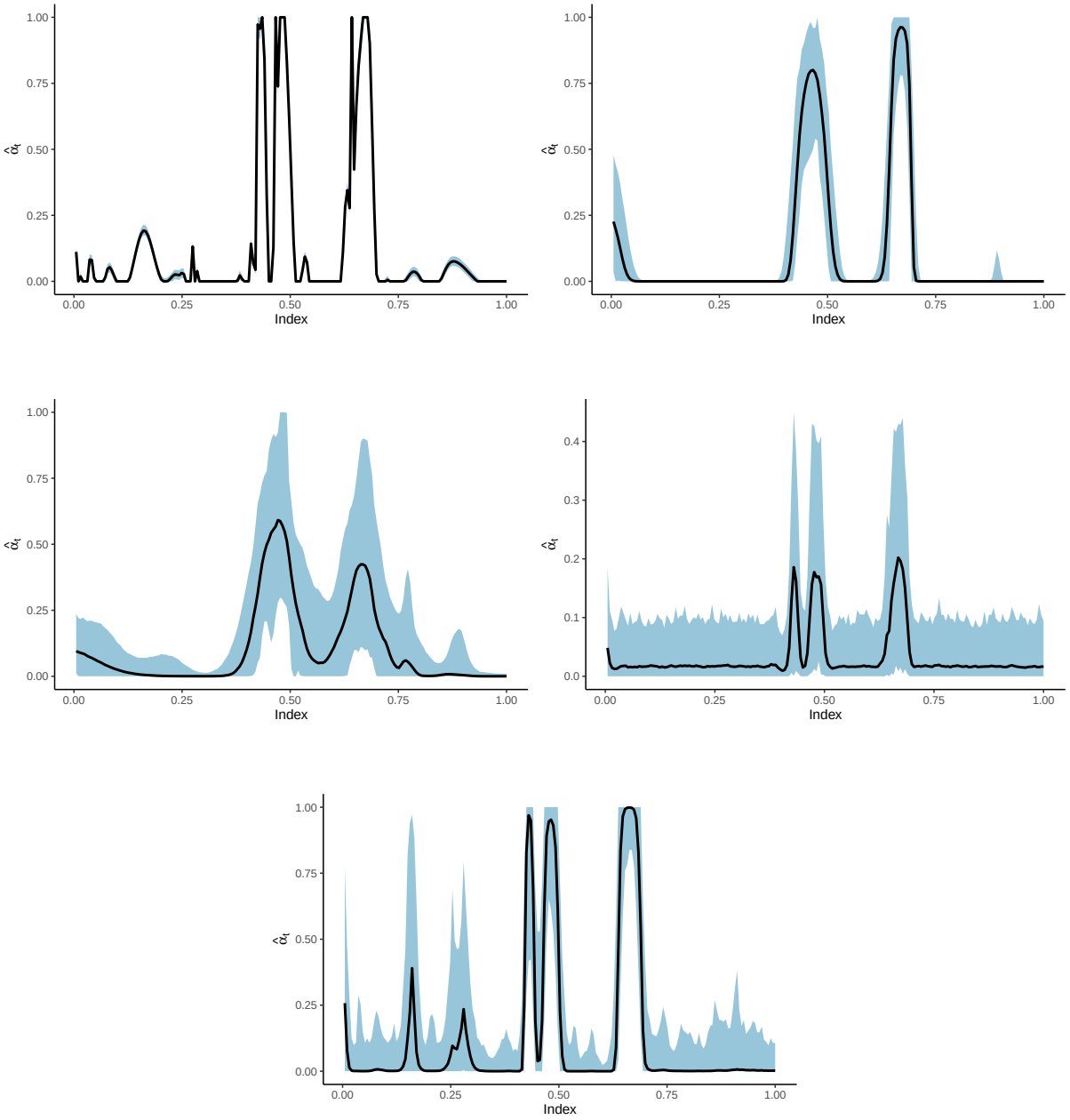



Figure 29 – Estimates of the $\alpha_t$'s of the aCGH data provided by the WR approach (top left); the DA approach: with Diffuse prior (top right); Gaussian prior (middle left); SSG prior (middle right); and SSL prior (bottom). The full lines correspond to the point estimates (medians) and the shaded areas correspond to the 95% HPD intervals.

CHAPTER

# 6

# CONCLUSIONS

The main goal of this work was to study a two-component Gaussian mixture model whose mixture weight is allowed to be dynamic, varying along some index such as time of space. Unlike other frequentist approaches that assume known component parameters (MONTORIL; PINHEIRO; VIDAKOVIC, 2019), here we consider a Bayesian framework and employ efficient Gibbs sampling algorithms to estimate the dynamic weights and the component parameters jointly. To estimate the mixture weights, we use wavelet bases.

In this work, we propose two approaches where Bayesian wavelet regularization techniques are employed to estimate the dynamic mixture weights of the Gaussian mixture model: the Wavelet regression (WR) and the Data augmentation (DA) approaches. The former consists of transforming the original data into a regression, whose regression function is the dynamic mixture weight. Then, once rescaled the observations, we employ regularization techniques to reduce the noise and estimate the dynamic mixture weights. The DA approach consists of adapting the data augmentation method proposed by Albert and Chib (1993), where we use the discrete wavelet transform matrix as the design matrix in the probit binary regression. To evaluate both methods, we use artificial and real data sets.

In the simulated studies, we conduct Monte Carlo simulations using functions with different degrees of smoothness to describe the mixture weight. When the function is very smooth, all approaches provide estimates similar to the real curve. All of them estimate the component parameters satisfactorily. However, when the methods have to estimate rougher functions, the DA approach with spike and slab priors overcome other methods.

In addition, the method is illustrated with a real data set application. This data set consists of an aCGH obtained from Glioblastoma Multiforme (GBM) tumor cell samples. In this scenario, we aim to detect signal values related to chromosomal anomalies in order to identify DNA regions where these aberrations occurred. For this data set, the WR approach and the DA approach with spike and slab priors are the ones to provide the most consistent results. The interested

reader can replicate the real data application by using the computational routines available in <https://github.com/flaviamotta/aCGH_application_wavelets>.

In general, both approaches, WR and DA, deliver satisfactory performances and are implemented with efficient MCMC algorithms. However, regarding the DA approach, it is shown that the spike and slab prior is the best choice for the distribution of wavelet coefficients, given the sparseness associated with them.

## 6.1 Further research

As a future work, one relevant research topic would be developing wavelet-based approaches to estimate the dynamic mixture weights of models with more than two mixture components. This extension would make the dynamic mixture model even more flexible and adaptive to broad clustering and classification problems. In this scenario, besides estimating the component parameters and the dynamic mixture weights, it is necessary to estimate the number of mixture components as well.

Furthermore, a study comparing the performances of both wavelet-based approaches of this work with other nonparametric alternatives to signal denoising, such as splines, Gaussian processes or empirical mode decomposition, would certainly be interesting.

# BIBLIOGRAPHY

ABRAMOVICH, F.; BAILEY, T. C.; SAPATINAS, T. Wavelet analysis and its statistical applications. **Journal of the Royal Statistical Society: Series D (The Statistician)**, Wiley Online Library, v. 49, n. 1, p. 1–29, 2000. Citations on pages 21, 23, 24, 30, and 36.

ABRAMOVICH, F.; SAPATINAS, T.; SILVERMAN, B. W. Wavelet thresholding via a bayesian approach. **Journal of the Royal Statistical Society. Series B (Statistical Methodology)**, [Royal Statistical Society, Wiley], v. 60, n. 4, p. 725–749, 1998. ISSN 13697412, 14679868. Available: <http://www.jstor.org/stable/2985959>. Citations on pages 21, 38, 39, and 56.

ALBERT, J. H.; CHIB, S. Bayesian analysis of binary and polychotomous response data. **Journal of the American Statistical Association**, [American Statistical Association, Taylor & Francis, Ltd.], v. 88, n. 422, p. 669–679, 1993. ISSN 01621459. Available: <http://www.jstor.org/stable/2290350>. Citations on pages 22, 53, 55, and 77.

ANTONIADIS, A.; GREGOIRE, G.; NASON, G. Density and hazard rate estimation for right-censored data by using wavelet methods. **Journal of the Royal Statistical Society. Series B (Statistical Methodology)**, [Royal Statistical Society, Wiley], v. 61, n. 1, p. 63–84, 1999. ISSN 13697412, 14679868. Available: <http://www.jstor.org/stable/2680737>. Citation on page 21.

BEYLKIN, G.; COIFMAN, R. R.; ROKHLIN, V. Fast wavelet transforms and numerical algorithms i. **Communications on Pure and Applied Mathematics**, v. 44, p. 141–183, 1991. Citation on page 30.

BORDES, L.; DELMAS, C.; VANDEKERKHOVE, P. Semiparametric estimation of a two-component mixture model where one component is known. **Scandinavian Journal of Statistics**, v. 33, n. 4, p. 733–752, 2006. Citation on page 48.

CAI, T.; BROWN, L. D. Wavelet estimation for samples with random uniform design. **Statistics & Probability Letters**, v. 42, n. 3, p. 313–321, 1999. ISSN 0167-7152. Available: <https://www.sciencedirect.com/science/article/pii/S0167715298002235>. Citation on page 21.

CASELLA, G.; GEORGE, E. I. Explaining the gibbs sampler. **The American Statistician**, [American Statistical Association, Taylor & Francis, Ltd.], v. 46, n. 3, p. 167–174, 1992. ISSN 00031305. Available: <http://www.jstor.org/stable/2685208>. Citation on page 51.

CASSIE, R. M. Some uses of probability paper in the analysis of size frequency distributions. **Marine and Freshwater Research**, v. 5, n. 3, p. 513–522, 1954. Citation on page 20.

CHANG, W.; KIM, S.-H.; VIDAKOVIC, B. Wavelet-based estimation of a discriminant function. **Applied Stochastic Models in Business and Industry**, v. 19, n. 3, p. 185–198, 2003. Citation on page 21.

CHARLIER, C. V. L.; WICKSELL, S. D. **On the dissection of frequency functions**. [S.l.]: Almqvist & Wiksells boktryckeri, 1924. Citation on page 20.

CHIPMAN, H. A.; KOLACZYK, E. D.; MCCULLOCH, R. E. Adaptive bayesian wavelet shrinkage. **Journal of the American Statistical Association**, [American Statistical Association, Taylor & Francis, Ltd.], v. 92, n. 440, p. 1413–1421, 1997. ISSN 01621459. Available: <http://www.jstor.org/stable/2965411>. Citation on page 37.

COHEN, A.; DAUBECHIES, I.; VIAL, P. Wavelets on the interval and fast wavelet transforms. **Applied and Computational Harmonic Analysis**, v. 1, n. 1, p. 54–81, 1993. ISSN 1063-5203. Citation on page 31.

DAUBECHIES, I. Orthonormal bases of compactly supported wavelets. **Communications on Pure and Applied Mathematics**, v. 41, n. 7, p. 909–996, 1988. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.3160410705>. Citations on pages 23, 28, 29, and 30.

____. **Ten Lectures on Wavelets**. USA: Society for Industrial and Applied Mathematics, 1992. ISBN 0898712742. Citations on pages 29 and 30.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the em algorithm. **Journal of the Royal Statistical Society. Series B (Methodological)**, [Royal Statistical Society, Wiley], v. 39, n. 1, p. 1–38, 1977. ISSN 00359246. Citation on page 20.

DIEBOLT, J.; ROBERT, C. P. Estimation of finite mixture distributions through bayesian sampling. **Journal of the Royal Statistical Society. Series B (Methodological)**, [Royal Statistical Society, Wiley], v. 56, n. 2, p. 363–375, 1994. ISSN 00359246. Citation on page 20.

DOETSCH, G. Die elimination des dopplereffekts bei spektroskopischen feinstrukturen und exakte bestimmung der komponenten. **Zeitschrift für Physik**, Springer, v. 49, n. 9, p. 705–730, 1928. Citation on page 20.

DONOHO, D. L. Nonlinear wavelet methods for recovery of signals, densities, and spectra from indirect and noisy data. In: **In Proceedings of Symposia in Applied Mathematics**. [S.l.]: American Mathematical Society, 1993. p. 173–205. Citation on page 21.

DONOHO, D. L.; JOHNSTONE, I. M. Adapting to unknown smoothness via wavelet shrinkage. **Journal of the American Statistical Association**, Taylor & Francis, v. 90, n. 432, p. 1200–1224, 1995. Available: <https://www.tandfonline.com/doi/abs/10.1080/01621459.1995.10476626>. Citation on page 36.

DONOHO, D. L.; JOHNSTONE, I. M.; KERKYACHARIAN, G.; PICARD, D. Density estimation by wavelet thresholding. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 24, n. 2, p. 508–539, 1996. ISSN 00905364. Citation on page 21.

DONOHO, D. L.; JOHNSTONE, J. M. Ideal spatial adaptation by wavelet shrinkage. **biometrika**, Oxford University Press, v. 81, n. 3, p. 425–455, 1994. Citations on pages 21, 35, 36, 40, and 41.

ESCOBAR, M. D.; WEST, M. Bayesian density estimation and inference using mixtures. **Journal of the American Statistical Association**, [American Statistical Association, Taylor & Francis, Ltd.], v. 90, n. 430, p. 577–588, 1995. ISSN 01621459. Citation on page 20.

FARGE, M. Wavelet transforms and their applications to turbulence. **Annual review of fluid mechanics**, Annual reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA, v. 24, n. 1, p. 395–458, 1992. Citations on pages 24 and 28.

FLEET, P. J. V. **Discrete wavelet transformations: An elementary approach with applications**. [S.l.]: John Wiley & Sons, 2011. Citation on page 31.

FRALEY, C.; RAFTERY, A. E. Model-based clustering, discriminant analysis, and density estimation. **Journal of the American Statistical Association**, [American Statistical Association, Taylor & Francis, Ltd.], v. 97, n. 458, p. 611–631, 2002. ISSN 01621459. Available: <http://www.jstor.org/stable/3085676>. Citation on page 19.

FRÜHWIRTH-SCHNATTER, S. **Finite mixture and Markov switching models**. [S.l.]: Springer, 2006. Citation on page 20.

FRÜHWIRTH-SCHNATTER, S.; CELEUX, G.; ROBERT, C. **Handbook of Mixture Analysis**. [S.l.]: CRC Press, 2018. (Chapman & Hall/CRC). ISBN 9780429508868. Citation on page 20.

FU, A. Q.; RUSSELL, S.; BRAY, S. J.; TAVARÉ, S. Bayesian clustering of replicated time-course gene expression data with weak signals. **The Annals of Applied Statistics**, JSTOR, p. 1334–1361, 2013. Citation on page 19.

GEORGE, E. I.; MCCULLOCH, R. E. Variable selection via gibbs sampling. **Journal of the American Statistical Association**, [American Statistical Association, Taylor & Francis, Ltd.], v. 88, n. 423, p. 881–889, 1993. ISSN 01621459. Available: <http://www.jstor.org/stable/2290777>. Citation on page 38.

GROSSMANN, A.; MORLET, J. Decomposition of hardy functions into square integrable wavelets of constant shape. **SIAM Journal on Mathematical Analysis**, v. 15, n. 4, p. 723–736, 1984. Citation on page 23.

HAAR, A. Zur theorie der orthogonalen funktionensysteme. **Mathematische Annalen**, v. 69, p. 331–371, 1910. Citation on page 23.

HALL, P.; PATIL, P. Formulae for mean integrated squared error of nonlinear wavelet-based density estimators. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 23, n. 3, p. 905–928, 1995. ISSN 00905364. Citation on page 21.

HALL, P.; ZHOU, X.-H. Nonparametric estimation of component distributions in a multivariate mixture. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 31, n. 1, p. 201–224, 2003. Citation on page 48.

HARDING, J. P. The use of probability paper for the graphical analysis of polymodal frequency distributions. **Journal of the Marine Biological Association of the United Kingdom**, Cambridge University Press, v. 28, n. 1, p. 141–153, 1948. Citation on page 20.

HÄRDLE, W.; KERKYACHARIAN, G.; PICARD, D.; TSYBAKOV, A. **Wavelets, approximation, and statistical applications**. [S.l.]: Springer Science & Business Media, 2012. Citations on pages 23, 26, and 27.

HSU, L.; SELF, S. G.; GROVE, D.; RANDOLPH, T.; WANG, K.; DELROW, J. J.; LOO, L.; PORTER, P. Denoising array-based comparative genomic hybridization data using wavelets. **Biostatistics**, Oxford University Press, v. 6, n. 2, p. 211–226, 2005. Citation on page 74.

HUGHES, J. P.; GUTTORP, P. A class of stochastic models for relating synoptic atmospheric patterns to regional hydrologic phenomena. **Water Resources Research**, v. 30, n. 5, p. 1535–1546, 1994. Citation on page 20.

HUI, S. L.; ZHOU, X.-H. Evaluation of diagnostic tests without gold standards. **Statistical Methods in Medical Research**, v. 7, n. 4, p. 354–370, 1998. Citation on page 48.

JANKOVIC, J.; MAZZIOTTA, J. C.; POMEROY, S. L.; NEWMAN, N. J. **Bradley and Daroff's Neurology in Clinical Practice**. [S.l.]: Elsevier, 2022. Citation on page 74.

JOHNSTONE, I.; SILVERMAN, B. W. Ebayesthresh: R programs for empirical bayes thresholding. **Journal of Statistical Software**, v. 12, p. 1–38, 2005. Citations on pages 21, 39, 41, and 57.

JOHNSTONE, I. M.; SILVERMAN, B. W. Empirical bayes selection of wavelet thresholds. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 33, n. 4, p. 1700–1752, 2005. Citations on pages 39, 41, and 57.

KARLIS, D.; XEKALAKI, E. Choosing initial values for the em algorithm for finite mixtures. **Computational Statistics & Data Analysis**, v. 41, n. 3, p. 577–590, 2003. ISSN 0167-9473. Recent Developments in Mixture Model. Available: <https://www.sciencedirect.com/science/article/pii/S0167947302001779>. Citation on page 49.

KILLICK, R.; ECKLEY, I. A. changepoint: An r package for changepoint analysis. **Journal of Statistical Software**, v. 58, n. 3, p. 1–19, 2014. Available: <https://www.jstatsoft.org/index.php/jss/article/view/v058i03>. Citation on page 75.

KRUSCHKE, J. **Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan**. [S.l.]: Academic Press, 2014. Citations on pages 51, 97, and 99.

KUO, L.; MALLICK, B. Variable selection for regression models. **Sankhyā: The Indian Journal of Statistics, Series B (1960-2002)**, Springer, v. 60, n. 1, p. 65–81, 1998. ISSN 05815738. Available: <http://www.jstor.org/stable/25053023>. Citation on page 38.

LAI, W. R.; JOHNSON, M. D.; KUCHERLAPATI, R.; PARK, P. J. Comparative analysis of algorithms for identifying amplifications and deletions in array cgh data. **Bioinformatics**, Oxford University Press, v. 21, n. 19, p. 3763–3770, 2005. Citations on pages 74, 75, and 76.

LAVINE, M.; WEST, M. A bayesian method for classification and discrimination. **The Canadian Journal of Statistics / La Revue Canadienne de Statistique**, [Statistical Society of Canada, Wiley], v. 20, n. 4, p. 451–461, 1992. ISSN 03195724. Available: <http://www.jstor.org/stable/3315614>. Citations on pages 20 and 51.

LU, Z.; SONG, X. Finite mixture varying coefficient models for analyzing longitudinal heterogenous data. **Statistics in Medicine**, v. 31, n. 6, p. 544–560, 2012. Citation on page 49.

MALLAT, S.; HWANG, W. Singularity detection and processing with wavelets. **IEEE Transactions on Information Theory**, v. 38, n. 2, p. 617–643, 1992. Citation on page 21.

MALLAT, S. G. A theory for multiresolution signal decomposition: the wavelet representation. **IEEE transactions on pattern analysis and machine intelligence**, Ieee, v. 11, n. 7, p. 674–693, 1989. Citations on pages 23 and 32.

MARIN, J.-M.; ROBERT, C. P. **Bayesian core: a practical approach to computational Bayesian statistics**. New York: Springer, 2007. Citation on page 38.

MCLACHLAN, G. J.; PEEL, D. **Finite mixture models**. [S.l.]: Wiley New York, 2000. ISBN 0471006262. Citations on pages 19 and 20.

MEYER, Y. Principe d'incertitude, bases hilbertiennes et algèbres d'opérateurs. **Séminaire Bourbaki**, Societe Mathematique de France, v. 662, p. 1985–1986, 1985. Citation on page 23.

MONTORIL, M. H.; CORREIA, L. T.; MIGON, H. S. **Bayesian estimation of dynamic weights in Gaussian mixture models**. arXiv, 2021. Available: <https://arxiv.org/abs/2104.03395>. Citations on pages 21 and 52.

MONTORIL, M. H.; PINHEIRO, A.; VIDAKOVIC, B. Wavelet-based estimators for mixture regression. **Scandinavian Journal of Statistics**, Wiley Online Library, v. 46, n. 1, p. 215–234, 2019. Citations on pages 21, 22, 53, 54, and 77.

MORETTIN, P. A. From fourier to wavelet analysis of time series. In: PRAT, A. (Ed.). **COMPSTAT**. Heidelberg: Physica-Verlag HD, 1996. p. 111–122. ISBN 978-3-642-46992-3. Citation on page 21.

MORLET, J. Sampling theory and wave propagation. In: CHEN, C. H. (Ed.). **Issues in Acoustic Signal — Image Processing and Recognition**. Berlin, Heidelberg: Springer Berlin Heidelberg, 1983. p. 233–261. Citation on page 23.

NAGY, I.; SUZDALEVA, E.; KáRNý, M.; MLYNářOVá, T. Bayesian estimation of dynamic finite mixtures. **International Journal of Adaptive Control and Signal Processing**, v. 25, n. 9, p. 765–787, 2011. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/acs.1239>. Citations on pages 20 and 21.

NASON, G. **Wavelet Methods in Statistics with R**. 1. ed. [S.l.]: Springer Publishing Company, Incorporated, 2008. ISBN 0387759603. Citation on page 36.

____. **wavethresh: Wavelets Statistics and Transforms**. [S.l.], 2016. R package version 4.6.8. Available: <https://CRAN.R-project.org/package=wavethresh>. Citation on page 41.

OGDEN, R. T. **Essential wavelets for statistical applications and data analysis**. [S.l.]: Birkhäuser Boston Inc., 1997. ISBN 0817638644. Citations on pages 21, 30, and 31.

PATRA, R. K.; SEN, B. Estimation of a two-component mixture model with applications to multiple testing. **Journal of the Royal Statistical Society. Series B (Statistical Methodology)**, [Royal Statistical Society, Wiley], v. 78, n. 4, p. 869–893, 2016. Citation on page 48.

PEARSON, K. Iii. contributions to the mathematical theory of evolution. **Philosophical Transactions of the Royal Society of London. (A.)**, v. 185, p. 71–110, 1894. Citations on pages 19 and 48.

PERCIVAL, D. B.; WALDEN, A. T. **Wavelet Methods for Time Series Analysis**. [S.l.]: Cambridge University Press, 1999. (Cambridge Series in Statistical and Probabilistic Mathematics). Citations on pages 21 and 30.

PINKEL, D.; ALBERTSON, D. G. Array comparative genomic hybridization and its applications in cancer. **Nature genetics**, Nature Publishing Group, v. 37, n. 6, p. S11–S17, 2005. Citation on page 74.

PRESTON, E. J. A graphical method for the analysis of statistical distributions into two normal components. **Biometrika**, v. 40, n. 3-4, p. 460–464, 1953. ISSN 0006-3444. Citation on page 20.

PRIESTLEY, M. B. Wavelets and time-dependent spectral analysis. **Journal of Time Series Analysis**, v. 17, n. 1, p. 85–103, 1996. Citation on page 21.

REDNER, R. A.; WALKER, H. F. Mixture densities, maximum likelihood and the em algorithm. **SIAM Review**, Society for Industrial and Applied Mathematics, v. 26, n. 2, p. 195–239, 1984. ISSN 00361445. Available: <http://www.jstor.org/stable/2030064>. Citation on page 53.

RESTREPO, J. M.; LEAF, G. K. Inner product computations using periodized daubechies wavelets. **International Journal for Numerical Methods in Engineering**, v. 40, n. 19, p. 3557–3578, 1997. Citation on page 31.

RINDSKOPF, D.; RINDSKOPF, W. The value of latent class analysis in medical diagnosis. **Statistics in Medicine**, v. 5, n. 1, p. 21–27, 1986. Citation on page 48.

ROY, V. Convergence diagnostics for markov chain monte carlo. **Annual Review of Statistics and Its Application**, v. 7, n. 1, p. 387–412, 2020. Citation on page 97.

SMITH, A. F. M.; ROBERTS, G. O. Bayesian computation via the gibbs sampler and related markov chain monte carlo methods. **Journal of the Royal Statistical Society. Series B (Methodological)**, [Royal Statistical Society, Wiley], v. 55, n. 1, p. 3–23, 1993. ISSN 00359246. Citation on page 20.

STEIN, C. M. Estimation of the mean of a multivariate normal distribution. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 9, n. 6, p. 1135–1151, 1981. ISSN 00905364. Available: <http://www.jstor.org/stable/2240405>. Citation on page 36.

STEPHENS, M. Dealing with label switching in mixture models. **Journal of the Royal Statistical Society. Series B (Statistical Methodology)**, [Royal Statistical Society, Wiley], v. 62, n. 4, p. 795–809, 2000. ISSN 13697412, 14679868. Available: <http://www.jstor.org/stable/2680622>. Citation on page 53.

VIDAKOVIC, B. **Statistical modeling by wavelets**. [S.l.]: John Wiley & Sons, 1999. Citations on pages 21, 23, 24, 25, 27, 29, 30, 31, and 35.

WALKER, M. G.; MATEO, M.; OLSZEWSKI, E. W.; SEN, B.; WOODROOFE, M. Clean kinematic samples in dwarf spheroidals: An algorithm for evaluating membership and estimating distribution parameters when contamination is present. **The Astronomical Journal**, The American Astronomical Society, v. 137, n. 2, p. 3109, jan 2009. Available: <https://dx.doi.org/10.1088/0004-6256/137/2/3109>. Citation on page 48.

WELDON, W. F. R. Ii. on certain correlated variations in carcinus moenas. **Proceedings of the Royal Society of London**, v. 54, p. 318 – 329, 1893. Citation on page 19.

APPENDIX

# A

# PROOFS

## A.1 Sparsity and vanishing moments

In the following, we approach the proof of Proposition 1, where we state that using wavelet bases with high-order vanishing moments to decompose polynomials can provide sparse decompositions. Let $\psi(x)$ be a mother wavelet with $m+1$ vanishing moments and $p(x)$ be a polynomial of degree $m$ or *less*. As shown by (2.6), to decompose $p(x)$ into detail wavelet coefficients using $\psi(x)$, we need to take the inner product between $p(x)$ and $\psi_{jk}(x)$. Thus, we can calculate $d_{jk}$ as

$$\begin{aligned}
d_{jk} &= \langle p, \psi_{jk} \rangle \\
&= \int_{-\infty}^{\infty} p(x)\psi_{jk}(x)\,dx \\
&= \int_{-\infty}^{\infty} (\alpha_m x^m + \alpha_{m-1}x^{m-1} + \cdots + \alpha_0 x^0) 2^{j/2}\psi(2^j x - k)\,dx \\
&= \frac{2^{j/2}}{2^j}\int_{-\infty}^{\infty}\left[\alpha_m\left(\frac{u+k}{2^j}\right)^m + \alpha_{m-1}\left(\frac{u+k}{2^j}\right)^{m-1} + \cdots + \alpha_0\left(\frac{u+k}{2^j}\right)^0\right]\psi(u)\,du,
\end{aligned}$$

using the *binomial theorem*, $(x+y)^n = \sum_{k=0}^{n}\binom{n}{k}x^{n-k}y^k$, we have

$$\begin{aligned}
&= \frac{2^{j/2}}{2^j}\int_{-\infty}^{\infty}\left[\frac{\alpha_m}{2^{jm}}\left(\sum_{i=0}^{m}\binom{m}{i}u^{m-i}k^i\right) + \frac{\alpha_{m-1}}{2^{j(m-1)}}\left(\sum_{i=0}^{m-1}\binom{m-1}{i}u^{m-1-i}k^i\right) + \ldots \right. \\
&\qquad \left. + \alpha_0\right]\psi(u)\,du \\
&= \frac{2^{j/2}}{2^j}\int_{-\infty}^{\infty}\left[\sum_{r=0}^{m}\sum_{i=0}^{r}\frac{\alpha_r}{2^{jr}}\binom{r}{i}u^{r-i}k^i\right]\psi(u)\,du \\
&= \frac{2^{j/2}}{2^j}\int_{-\infty}^{\infty}\left[\sum_{l=0}^{m}\beta_l u^l\right]\psi(u)\,du
\end{aligned}$$

$$= \frac{2^{j/2}}{2^j} \sum_{l=0}^{m} \beta_l \int_{-\infty}^{\infty} u^l \psi(u)\, du$$
$$= 0.$$

Therefore, all detail coefficients $d_{jk}$ of the transformed polynomial, of degree $m$ or less, are null.

APPENDIX

# B

# POSTERIOR DISTRIBUTION OF THE WAVELET COEFFICIENTS

In this chapter, we derive the posterior distributions of the wavelet coefficients under the spike and slab priors discussed in section 3.2: the mixture between a point mass at zero and the Gaussian density and the mixture between a point mass at zero and the Laplace density. In the following section, we begin by introducing generic calculations applicable to both scenarios. Then, we detail the calculus referring to each type of prior.

## B.1 Generic calculations

We normalize the wavelet coefficients by dividing them by $\sigma_j$, the noise standard deviation of each resolution level. Thus, at level $j$, define the sequence $X_{jk} = d_{jk}/\sigma_j$, such that $X_{jk}|\theta_{jk} \sim \mathsf{N}(\theta_{jk}, 1)$. For simplicity, in the next set of expressions, we drop the $j,k$ subscripts, as they add nothing to the current exposition.

Let the prior for the wavelet coefficients distribution be

$$(1-\pi)\delta_0(\theta) + \pi\gamma(\theta), \tag{B.1}$$

where the non-null mixture component $\gamma$ is a symmetric unimodal density. Then, the marginal density can be calculates as

$$\begin{aligned} f(x) &= \int_{-\infty}^{+\infty} f(x|\theta) f_{\text{prior}}(\theta)\, d\theta \\ &= \int_{-\infty}^{+\infty} \phi(x-\theta)[(1-\pi)\delta_0(\theta) + \pi\gamma(\theta)]\, d\theta \\ &= (1-\pi)\int_{-\infty}^{+\infty} \phi(x-\theta)\delta_0(\theta)\, d\theta + \pi\int_{-\infty}^{+\infty} \phi(x-\theta)\gamma(\theta)\, d\theta \\ &= (1-\pi)\phi(x) + \pi g(x), \end{aligned}$$

with $g$ being the convolution between the $\gamma$ density and the standard normal distribution $\phi$ ($g = \gamma \star \phi$).

Therefore, one can easily show that the posterior distribution $f(\theta|x)$ is given by

$$\begin{aligned} f(\theta|x) &= \frac{f(x|\theta) f_{\text{prior}}(\theta)}{f(x)} \\ &= \frac{\phi(x-\theta)[(1-\pi)\delta_0(\theta) + \pi\gamma(\theta)]}{(1-\pi)\phi(x) + \pi g(x)} \\ &= \frac{\phi(x-\theta)(1-\pi)\delta_0(\theta) + \pi\gamma(\theta)\phi(x-\theta)}{(1-\pi)\phi(x) + \pi g(x)} \\ &= \frac{(1-\pi)\phi(x)\delta_0(\theta)}{(1-\pi)\phi(x) + \pi g(x)} + \frac{\pi g(x) f_1(\theta|x)}{(1-\pi)\phi(x) + \pi g(x)} \\ &= (1-\pi_{\text{post}})\delta_0(\theta) + \pi_{\text{post}} f_1(\theta|x), \end{aligned}$$

with $f_1(\theta|x)$ being the non-null mixture component

$$\begin{aligned} f_1(\theta|X=x, \theta \neq 0) &= \frac{f(\theta, X=x) I_{\{\theta \neq 0\}}}{\int_{\mathbb{R}_*} f_{\text{prior}}(\theta) f(x|\theta)\, d\theta} \\ &= \frac{f_{\text{prior}}(\theta) f(x|\theta) I_{\{\theta \neq 0\}}}{\int_{\mathbb{R}_*} \pi\gamma(\theta)\phi(x-\theta)\, d\theta} \\ &= \frac{\pi\gamma(\theta)\phi(x-\theta)}{\pi \int_{\mathbb{R}_*} \gamma(\theta)\phi(x-\theta)\, d\theta} \\ &= \frac{\gamma(\theta)\phi(x-\theta)}{g(x)}, \end{aligned}$$

and $\pi_{\text{post}}$ being the posterior sparsity parameter given by

$$\pi_{\text{post}} = \frac{\pi g(x)}{\pi g(x) + (1-\pi)\phi(x)}.$$

### B.1.1 Spike and slab prior - Gaussian component

Let $\gamma$ in (B.1) be $\text{N}(0, \upsilon_j^2)$, as in (3.4). Then, $g(x)$ can be calculated as

$$\begin{aligned} g(x) &= \int_{-\infty}^{+\infty} \phi(x-\theta)\gamma(\theta)\, d\theta \\ &= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(x-\theta)^2\right] \frac{1}{\sqrt{2\pi}\upsilon} \exp\left[-\frac{1}{2}\frac{\theta^2}{\upsilon^2}\right] d\theta \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\upsilon} \exp\left\{-\frac{1}{2}\left[\frac{\theta^2}{\upsilon^2} + \theta^2 - 2\theta x + x^2\right]\right\} d\theta \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\upsilon} \exp\left\{-\frac{1}{2}\left[\left(\frac{1}{\upsilon^2} + 1\right)\theta^2 - 2\theta x + x^2\right]\right\} d\theta \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\upsilon} \exp\left\{-\frac{1}{2}\left[\left(\frac{1+\upsilon^2}{\upsilon^2}\right)\left(\theta^2 - \frac{2\theta x \upsilon^2}{1+\upsilon^2}\right) + x^2\right]\right\} d\theta \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\upsilon} \exp\left\{ -\frac{1}{2} \left[ \left( \frac{1+\upsilon^2}{\upsilon^2} \right) \left( \theta^2 - 2\frac{\theta x \upsilon^2}{1+\upsilon^2} + \frac{x^2\upsilon^4}{(1+\upsilon^2)^2} \right) - \frac{x^2\upsilon^2}{1+\upsilon^2} + x^2 \right] \right\} d\theta \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\upsilon} \exp\left\{ -\frac{1}{2} \left[ \left( \frac{1+\upsilon^2}{\upsilon^2} \right) \left( \theta - \frac{\theta x \upsilon^2}{1+\upsilon^2} \right)^2 + \frac{x^2}{1+\upsilon^2} \right] \right\} d\theta \\
&= \frac{1}{\sqrt{2\pi}} \exp\left[ -\frac{1}{2} \left( \frac{x^2}{1+\upsilon^2} \right) \right] \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\upsilon} \exp\left\{ -\frac{1}{2} v^2 \right\} \frac{\upsilon}{\sqrt{1+\upsilon^2}} dv \\
&= \frac{1}{\sqrt{2\pi(1+\upsilon^2)}} \exp\left[ -\frac{1}{2} \left( \frac{x^2}{1+\upsilon^2} \right) \right] \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2} v^2 \right\} dv \\
&= \frac{1}{\sqrt{2\pi(1+\upsilon^2)}} \exp\left[ -\frac{1}{2(1+\upsilon^2)} x^2 \right].
\end{aligned}
$$

Thus, the non-null mixture component $f_1(\theta|x)$ is given by

$$
\begin{aligned}
f_1(\theta|x) &= \frac{\phi(x-\theta)\gamma(\theta)}{g(x)} \\
&= \frac{\frac{1}{\sqrt{2\pi}} \exp\left[ -\frac{1}{2}(x-\theta)^2 \right] \frac{1}{\sqrt{2\pi\upsilon^2}} \exp\left[ -\frac{1}{2}\frac{\theta^2}{\upsilon^2} \right]}{\frac{1}{\sqrt{2\pi(\upsilon^2+1)}} \exp\left[ -\frac{1}{2}\left( \frac{x^2}{\upsilon^2+1} \right) \right]} \\
&= \frac{\sqrt{2\pi(\upsilon^2+1)}}{\sqrt{2\pi}\sqrt{2\pi\upsilon^2}} \exp\left\{ -\frac{1}{2} \left[ (x-\theta)^2 + \frac{\theta^2}{\upsilon^2} - \frac{x^2}{\upsilon^2+1} \right] \right\} \\
&= \frac{1}{\sqrt{2\pi\lambda}} \exp\left\{ -\frac{1}{2} \left[ \frac{x^2\upsilon^2 - 2x\theta\upsilon^2 + \theta^2\upsilon^2 + \theta^2 - x^2\lambda}{\upsilon^2} \right] \right\} \\
&= \frac{1}{\sqrt{2\pi\lambda}} \exp\left\{ -\frac{1}{2\lambda} \left[ \frac{x^2\upsilon^2 - 2x\theta\upsilon^2 + \theta^2\upsilon^2 + \theta^2 - x^2\lambda}{1+\upsilon^2} \right] \right\} \\
&= \frac{1}{\sqrt{2\pi\lambda}} \exp\left\{ -\frac{1}{2\lambda} \left[ \frac{x^2\upsilon^2(1+\upsilon^2) - 2x\theta\upsilon^2(1+\upsilon^2) + \theta^2\upsilon^2(1+\upsilon^2)}{(1+\upsilon^2)^2} \right.\right. \\
&\quad \left.\left. + \frac{\theta^2(1+\upsilon^2) - x^2\upsilon^2}{(1+\upsilon^2)^2} \right] \right\} \\
&= \frac{1}{\sqrt{2\pi\lambda}} \exp\left\{ -\frac{1}{2\lambda} \left[ \frac{x^2\upsilon^4 - 2x\theta\upsilon^2 - 2x\theta\upsilon^4 + \theta^2\upsilon^2 + \theta^2\upsilon^4 + \theta^2 + \theta^2\upsilon^2}{(1+\upsilon^2)^2} \right] \right\} \\
&= \frac{1}{\sqrt{2\pi\lambda}} \exp\left\{ -\frac{1}{2\lambda} \left[ \frac{\upsilon^4(x^2 - 2x\theta + \theta^2) - 2\theta\upsilon^2(x-\theta) + \theta^2}{(1+\upsilon^2)^2} \right] \right\} \\
&= \frac{1}{\sqrt{2\pi\lambda}} \exp\left\{ -\frac{1}{2\lambda} \left[ \frac{\upsilon^4(x-\theta)^2 - 2\theta\upsilon^2(x-\theta) + \theta^2}{(1+\upsilon^2)^2} \right] \right\} \\
&= \frac{1}{\sqrt{2\pi\lambda}} \exp\left\{ -\frac{1}{2\lambda} \left[ \frac{(\theta - \upsilon^2(x-\theta))^2}{(1+\upsilon^2)^2} \right] \right\} \\
&= \frac{1}{\sqrt{2\pi\lambda}} \exp\left\{ -\frac{1}{2\lambda} \left[ \frac{\theta - \upsilon^2 x + \upsilon^2\theta}{1+\upsilon^2} \right]^2 \right\} \\
&= \frac{1}{\sqrt{2\pi\lambda}} \exp\left\{ -\frac{1}{2\lambda} \left[ \theta - \frac{\upsilon^2 x}{1+\upsilon^2} \right]^2 \right\} \\
&= \frac{1}{\sqrt{2\pi\lambda}} \exp\left\{ -\frac{1}{2\lambda} [\theta - \lambda x]^2 \right\},
\end{aligned}
$$

where $\lambda = \upsilon^2/(1+\upsilon^2)$. Therefore, the posterior distribution is a mixture of a point mass at zero and $\mathrm{N}\left(\frac{\upsilon^2}{1+\upsilon^2}x, \frac{\upsilon^2}{1+\upsilon^2}\right)$, as stated in (3.5).

### B.1.2 Spike and slab prior - Laplace component

Let $\gamma$ in (B.1) be the Laplace density given in (3.7). Then, the convolution $g$ is

$$
\begin{aligned}
g(x) &= \int_{-\infty}^{+\infty} \phi(x-\theta)\gamma_a(\theta)\,d\theta \\
&= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(x-\theta)^2\right] \frac{a\exp(-a|\theta|)}{2}\,d\theta \\
&= \frac{1}{\sqrt{2\pi}}\frac{a}{2}\int_{-\infty}^{+\infty} \exp\left[-\frac{1}{2}(x-\theta)^2 - a|\theta|\right] d\theta \\
&= \frac{1}{\sqrt{2\pi}}\frac{a}{2}\left\{\int_{-\infty}^{0} \exp\left[-\frac{1}{2}(x-\theta)^2 + a\theta\right] d\theta + \int_{0}^{\infty} \exp\left[-\frac{1}{2}(x-\theta)^2 - a\theta\right] d\theta\right\} \\
&= \frac{1}{\sqrt{2\pi}}\frac{a}{2}\left\{\int_{-\infty}^{0} \exp\left[-\frac{(x-\theta)^2 - 2a\theta}{2}\right] d\theta + \int_{0}^{\infty} \exp\left[-\frac{(x-\theta)^2 + 2a\theta}{2}\right] d\theta\right\} \\
&= \frac{1}{\sqrt{2\pi}}\frac{a}{2}\left\{\int_{-\infty}^{0} \exp\left[-\frac{x^2 - 2x\theta + \theta^2 - 2a\theta}{2}\right] d\theta + \int_{0}^{\infty} \exp\left[-\frac{x^2 - 2x\theta + \theta^2 + 2a\theta}{2}\right] d\theta\right\} \\
&= \frac{1}{\sqrt{2\pi}}\frac{a}{2}\exp\left(\frac{a^2}{2}\right)\left\{e^{ax}\int_{-\infty}^{0} \exp\left[-\frac{[\theta-(x+a)]^2}{2}\right] d\theta + e^{-ax}\int_{0}^{\infty} \exp\left[-\frac{[\theta-(x-a)]^2}{2}\right] d\theta\right\} \\
&= \frac{a}{2}\exp\left(\frac{a^2}{2}\right)\left\{\exp(ax)\int_{-\infty}^{-x-a} \frac{1}{\sqrt{2\pi}}\exp\left[-\frac{u^2}{2}\right] du + \exp(-ax)\int_{-x+a}^{\infty} \frac{1}{\sqrt{2\pi}}\exp\left[-\frac{v^2}{2}\right] dv\right\} \\
&= \frac{a}{2}\exp\left(\frac{a^2}{2}\right)\left\{\exp(ax)\Phi(-x-a) + \exp(-ax)[1-\Phi(-x+a)]\right\} \\
&= \frac{a}{2}\exp\left(\frac{a^2}{2}\right)\left\{\exp(ax)[1-\Phi(x+a)] + \exp(-ax)\Phi(x-a)\right\} \\
&= \frac{a}{2}\exp\left(\frac{a^2}{2}\right)\left\{\exp(ax)\tilde{\Phi}(x+a) + \exp(-ax)\Phi(x-a)\right\},
\end{aligned}
$$

with $\Phi$ denoting the standard normal cumulative function, and $\tilde{\Phi} = 1-\Phi$. Then, we can calculate the non-null mixture component $f_1(\theta|x)$ as

$$
\begin{aligned}
f_1(\theta|x) &= \frac{\phi(x-\theta)\cdot\gamma_a(\theta)}{g(x)} \\
&= \frac{\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}(x-\theta)^2\right]\frac{a\exp(-a|\theta|)}{2}}{\frac{a}{2}\exp\left(\frac{a^2}{2}\right)\left\{\exp(ax)\tilde{\Phi}(x+a) + \exp(-ax)\Phi(x-a)\right\}} \\
&= \frac{\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}(x-\theta)^2 - a|\theta| - \frac{a^2}{2}\right]}{\exp(ax)\tilde{\Phi}(x+a) + \exp(-ax)\Phi(x-a)} \\
&= \frac{\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}(x-\theta)^2 - a\theta - \frac{a^2}{2}\right] I_{\{\theta>0\}}}{\exp(ax)\tilde{\Phi}(x+a) + \exp(-ax)\Phi(x-a)} + \frac{\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}(x-\theta)^2 + a\theta - \frac{a^2}{2}\right] I_{\{\theta<0\}}}{\exp(ax)\tilde{\Phi}(x+a) + \exp(-ax)\Phi(x-a)}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}(x^2-2x\theta+\theta^2+2a\theta+a^2)\right]}{\exp(ax)\tilde{\Phi}(x+a)+\exp(-ax)\Phi(x-a)} I_{\{\theta>0\}} \\
&\quad + \frac{\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}(x^2-2x\theta+\theta^2-2a\theta+a^2)\right]}{\exp(ax)\tilde{\Phi}(x+a)+\exp(-ax)\Phi(x-a)} I_{\{\theta<0\}} \\
&= \frac{\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}(\theta^2-2(x-a)\theta+x^2-2ax+a^2)\right]}{\exp(ax)\tilde{\Phi}(x+a)+\exp(-ax)\Phi(x-a)} \exp(-ax) I_{\{\theta>0\}} \\
&\quad + \frac{\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}(\theta^2-2(x+a)\theta+x^2+2ax+a^2)\right]}{\exp(ax)\tilde{\Phi}(x+a)+\exp(-ax)\Phi(x-a)} \exp(ax) I_{\{\theta<0\}} \\
&= \frac{\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}[\theta-(x-a)]^2\right\}}{\exp(ax)\tilde{\Phi}(x+a)+\exp(-ax)\Phi(x-a)} \exp(-ax) I_{\{\theta>0\}} \\
&\quad + \frac{\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}[\theta-(x+a)]^2\right\}}{\exp(ax)\tilde{\Phi}(x+a)+\exp(-ax)\Phi(x-a)} \exp(ax) I_{\{\theta<0\}} \\
&= \frac{\exp(-ax)\phi[\theta-(x-a)]I_{\{\theta>0\}}}{\exp(ax)\tilde{\Phi}(x+a)+\exp(-ax)\Phi(x-a)} + \frac{\exp(ax)\phi[\theta-(x+a)]I_{\{\theta<0\}}}{\exp(ax)\tilde{\Phi}(x+a)+\exp(-ax)\Phi(x-a)},
\end{aligned}
$$

which is a mixture of two truncated normal distributions, as stated in (3.9),

$$
\begin{aligned}
f_1(\theta|x) &= \frac{\exp(-ax)\phi[\theta-(x-a)]I_{\{\theta>0\}}}{\exp(ax)\tilde{\Phi}(x+a)+\exp(-ax)\Phi(x-a)} \left[\frac{\Phi(x-a)}{1-\Phi(-x+a)}\right] \\
&\quad + \frac{\exp(ax)\phi[\theta-(x+a)]I_{\{\theta<0\}}}{\exp(ax)\tilde{\Phi}(x+a)+\exp(-ax)\Phi(x-a)} \left[\frac{1-\Phi(x+a)}{\Phi(-x-a)}\right] \\
&= \frac{\exp(-ax)\Phi(x-a)}{\exp(ax)\tilde{\Phi}(x+a)+\exp(-ax)\Phi(x-a)} \left\{\frac{\phi[\theta-(x-a)]I_{\{\theta>0\}}}{1-\Phi(-x+a)}\right\} \\
&\quad + \frac{\exp(ax)\tilde{\Phi}(x+a)}{\exp(ax)\tilde{\Phi}(x+a)+\exp(-ax)\Phi(x-a)} \left\{\frac{\phi[\theta-(x+a)]I_{\{\theta<0\}}}{\Phi(-x-a)}\right\} \\
&= \eta f_{TN}(\theta|x-a,1,0,+\infty) + (1-\eta) f_{TN}(\theta|x+a,1,-\infty,0),
\end{aligned}
$$

where

$$
\eta = \frac{\exp(-ax)\Phi(x-a)}{\exp(ax)\tilde{\Phi}(x+a)+\exp(-ax)\Phi(x-a)}.
$$

APPENDIX

# C

# CONDITIONAL POSTERIOR DISTRIBUTIONS

In this chapter, we derive the full conditional posterior distributions for the component parameters, $\mu_k$ in (4.5) and $\tau_k^2$ in (4.6), and for parameters of the *spike and slab prior* $\upsilon_j^{-2}$ and $\pi_j$, discussed in section 4.5.

## C.1 Full conditional posterior of $\mu_k$

Assuming that $\mu_k \sim \mathrm{N}(b_{0k}, B_{0k})$, we have

$$
\begin{aligned}
\mu_k|\tau_k^2, \boldsymbol{y}, \boldsymbol{z} &\propto p(\boldsymbol{y}|\mu_k, \tau_k^2, \boldsymbol{z}) p(\mu_k) \\
&\propto \left(\frac{\tau_k^2}{2\pi}\right)^{\frac{T_k}{2}} \exp\left[-\frac{\tau_k^2}{2} \sum_{t:z_t=k-1} (y_t - \mu_k)^2\right] \frac{1}{\sqrt{2\pi B_{0k}}} \exp\left[-\frac{(\mu_k - b_{0k})^2}{2B_{0k}}\right] \\
&\propto \exp\left\{-\frac{1}{2}\left[\tau_k^2 \sum_{t:z_t=k-1} y_t^2 - 2\tau_k^2 s_k \mu_k + \tau_k^2 T_k \mu_K^2 + \frac{\mu_k^2 - 2b_{0k}\mu_k + b_{0k}^2}{B_{0k}}\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2B_{0k}}\left[\tau_k^2 T_k B_{0k} \mu_K^2 - 2\tau_k^2 s_k B_{0k} \mu_k + \mu_k^2 - 2b_{0k}\mu_k\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2B_{0k}}\left[(\tau_k^2 T_k B_{0k} + 1)\mu_K^2 - 2(\tau_k^2 s_k B_{0k} + b_{0k})\mu_k\right]\right\}.
\end{aligned}
$$

To complete the square, let $p$ and $2pq$ be

$$
\begin{cases}
p^2 = (\tau_k^2 T_k B_{0k} + 1); \\
2pq = 2(\tau_k^2 s_k B_{0k} + b_{0k}).
\end{cases}
$$

$$
\begin{aligned}
&\propto \exp\left\{-\frac{1}{2B_{0k}}\left[p^2\mu_K^2 - 2pq\mu_k\right]\right\}\exp\left(-\frac{q^2}{2B_{0k}}\right)\\
&\propto \exp\left\{-\frac{1}{2B_{0k}}\left[p\mu_K - q\right]^2\right\}\\
&\propto \exp\left\{-\frac{1}{2B_{0k}}\left[\sqrt{\tau_k^2 T_k B_{0k}+1}\,\mu_K - \frac{\tau_k^2 s_k B_{0k}+b_{0k}}{\sqrt{\tau_k^2 T_k B_{0k}+1}}\right]^2\right\}\\
&\propto \exp\left\{-\frac{1}{2B_{0k}}(\tau_k^2 T_k B_{0k}+1)\left[\mu_K - \frac{\tau_k^2 s_k B_{0k}+b_{0k}}{\tau_k^2 T_k B_{0k}+1}\right]^2\right\}\\
&\propto \exp\left\{-\frac{1}{2(B_{0k}^{-1}+\tau_k^2 T_k)^{-1}}\left[\mu_K - (B_{0k}^{-1}+\tau_k^2 T_k)^{-1}(\tau_k^2 s_k + B_{0k}^{-1} b_{0k})\right]^2\right\},
\end{aligned}
$$

which ensures that

$$\mu_k|\tau_k^2, \boldsymbol{y}, \boldsymbol{z} \sim \mathrm{N}(b_k, B_k),$$

where

$$
\begin{aligned}
B_k &= (B_{0k}^{-1} + \tau_k^2 T_k)^{-1},\\
b_k &= B_k(\tau_k^2 s_k + B_{0k}^{-1} b_{0k}).
\end{aligned}
$$

## C.2 Full conditional posterior of $\tau_k^2$

Assuming that $\tau_k^2 \sim \Gamma(c_{0k}, C_{0k})$, we have

$$
\begin{aligned}
p(\tau_k^2|\mu_k, \boldsymbol{y}, \boldsymbol{z}) &\propto p(\boldsymbol{y}|\mu_k, \tau_k^2, \boldsymbol{z})p(\tau_k^2)\\
&\propto \left(\frac{\tau_k^2}{2\pi}\right)^{\frac{T_k}{2}} \exp\left[-\frac{\tau_k^2}{2}\sum_{t:z_t=k-1}(y_t-\mu_k)^2\right](\tau_k^2)^{c_{0k}-1}\exp\left(-C_{0k}\tau_k^2\right)\\
&\propto (\tau_k^2)^{c_{0k}+\frac{T_k}{2}-1}\exp\left\{-\left[C_{0k} + \frac{\sum_{t:z_t=k-1}(y_t-\mu_k)^2}{2}\right]\tau_k^2\right\}.
\end{aligned}
$$

Thus, as stated in (4.6),

$$\tau_k^2|\mu_k, \boldsymbol{y}, \boldsymbol{z} \sim \Gamma(c_k, C_k),$$

where

$$
\begin{aligned}
C_k &= C_{0k} + \frac{\sum_{t:z_t=k-1}(y_t-\mu_k)^2}{2},\\
c_k &= c_{0k} + \frac{T_k}{2}.
\end{aligned}
$$

## C.3 Hyperparameters of *spike and slab prior*

In section 4.5, we discuss sampling $\pi_j$ and $\upsilon_j^{-2}$ from their conditional posterior distributions instead of applying the marginal maximum likelihood. In this section, we derive (4.14) and (4.15). For simplicity, in the next set of expressions, we drop the $j,k$ subscripts, as they add nothing to the current exposition.

Consider the following

$$\begin{aligned}
\upsilon^{-2} &\sim \Gamma(\kappa,\xi) \\
\pi &\sim \mathsf{Beta}(\zeta,\rho) \\
r_t|\pi &\sim \mathsf{Bern}(\pi) \\
\theta_t|\upsilon^2, r_t = 1 &\sim \mathsf{N}(0,\upsilon_j^2) \\
\theta_t|\upsilon^2, r_t = 0 &\sim \delta_0(\theta_t).
\end{aligned}$$

In addition, assume that $\theta_1,\ldots,\theta_n$ are i.i.d. We can calculate the posterior distribution of $\upsilon^{-2}$ as

$$\begin{aligned}
p(\upsilon^{-2}|[\ldots]) &\propto (\upsilon^{-2})^{\kappa-1}\exp(-\xi\upsilon^{-2})\prod_{t=1}^{n}\upsilon^{-1}\exp\left(-\frac{\theta_t^2\upsilon^{-2}}{2}\right) \\
&\propto (\upsilon^{-2})^{\kappa+\frac{n}{2}-1}\exp(-\xi\upsilon^{-2})\exp\left(-\frac{\sum_{t=1}^{n}\theta_t^2\upsilon^{-2}}{2}\right) \\
&\propto (\upsilon^{-2})^{\kappa+\frac{n}{2}-1}\exp\left[-\left(\frac{\sum_{t=1}^{n}\theta_t^2}{2}+\xi\right)\upsilon^{-2}\right].
\end{aligned}$$

Therefore, as stated in (4.14),

$$\upsilon^{-2}|[\ldots] \sim \Gamma\left(\kappa+\frac{n}{2},\frac{\sum_{t=1}^{n}\theta_t^2}{2}+\xi\right).$$

Likewise, the sparsity parameter is given by

$$\begin{aligned}
p(\pi|[\ldots]) &\propto p(\pi)\prod_{t=1}^{n}p(r_t|\pi) \\
&\propto \pi^{\zeta-1}(1-\pi)^{\rho-1}\pi^{n_1}(1-\pi)^{n_0} \\
&\propto \pi^{\zeta+n_1-1}(1-\pi)^{\rho+n_0-1},
\end{aligned}$$

with $n_0 = \#\{t : r_t = 0, t = 1,2,\ldots,n\}$ and $n_1 = n - n_0$. Thus, as stated in (4.15),

$$\pi|[\ldots] \sim \mathsf{Beta}(\zeta+n_1,\rho+n_0).$$

APPENDIX

D

# MCMC CONVERGENCE DIAGNOSTIC PLOTS

This chapter presents the convergence diagnostic plots for the MCMC chains of simulated and real data sets.

## D.1 MCMC diagnostics

When implementing MCMC methods, one must look for convergence of Markov chains to the stationarity and check if the Monte Carlo estimators appear to be appropriately representative of the population quantities (ROY, 2020). In this work, we choose graphical methods for MCMC convergence diagnosis: *the trace plot*, the *autocorrelation plot* and the *density plot*.

The trace plot is a graph of the sampled parameter value in each chain step. This time series plot presents a visual examination of the parameters' trajectory and allows one to determine how well the chain is mixing. Therefore, if the samples are settled into some part of the parameter space, compromising the convergence, we would see flat bits in the trace plot. In general, we expect no particular trends on the chain trajectory since they can indicate that one has not reached stationarity yet (KRUSCHKE, 2014; ROY, 2020). Figure 30 presents two trace diagrams. While the left plot does not exhibit any particular trend that could indicate a lack of convergence, the right plot is an example of a chain whose sampled values are stuck into parts of the parameter space.

The *autocorrelation plot* shows the correlation between the chain values $k$ iterations apart. The vertical axis plots the lag-$k$ autocorrelation function (ACF) and the horizontal axis plots the increasing $k$ values, which represent the number of steps between the MCMC samples, called *lag*. In this graphical method, if the Markov chain was run for a sufficient amount of time, we expect the lag-$k$ autocorrelation values to drop down towards zero quickly as $k$ increases. High lag-$k$ autocorrelation values for higher $k$ may be a sign that each step in the chain is not providing uncorrelated information about the posterior distribution (KRUSCHKE, 2014).

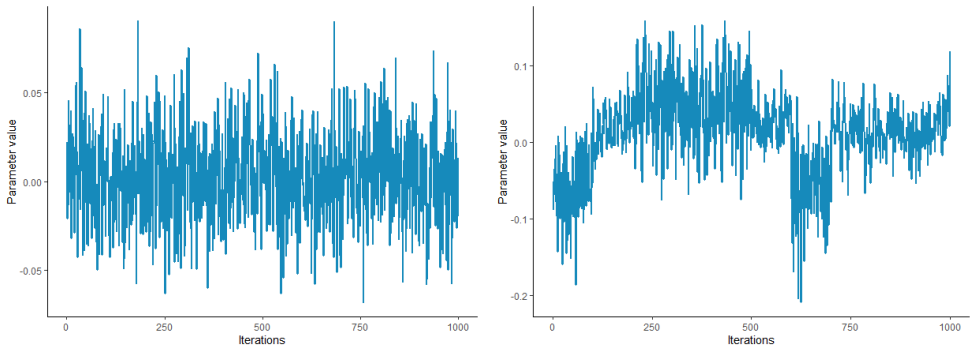



Figure 30 – Illustration of MCMC diagnostics: trace plots. The left plot shows similar pattern throughout the iterations while the right one shows visible trends that indicate slow convergence.

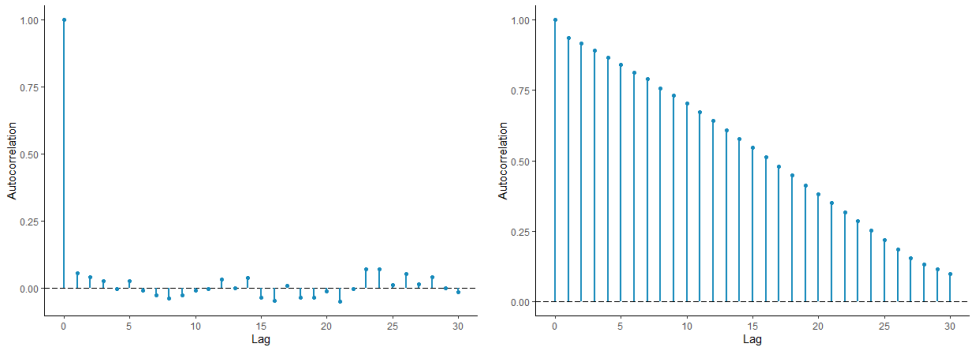



Figure 31 – Illustration of MCMC diagnostics: autocorrelation plots. The left plot suggests that the parameter values at successive steps in the chain provide uncorrelated information about the posterior distribution. In opposition, the right one is an example of a highly autocorrelated chain.

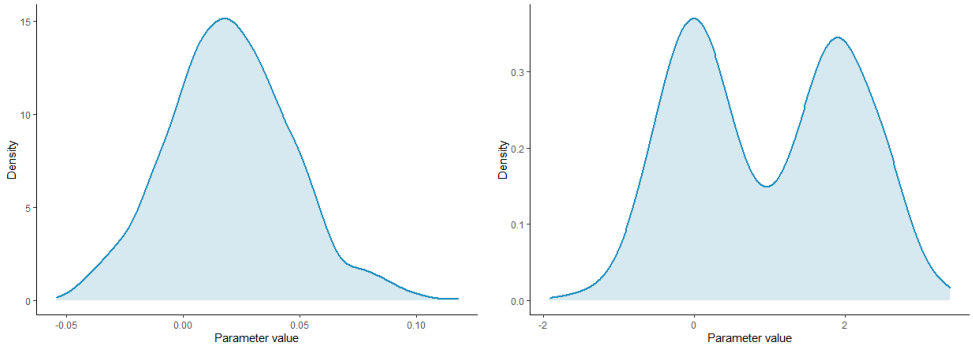



Figure 32 – Illustration of MCMC diagnostics: density plots. The left plot mimics a bell shape as we would expect, while the right plot exhibit multimodality.

Figure 31 presents examples of autocorrelation plots.

Another interesting visual method of diagnosis is the density plot of the MCMC draws, as shown in Figure 32. Since we are dealing with Gaussian mixture models, density plots that exhibit multimodality may indicate that the label switching problem has not been suitably addressed. We expect the density curves to mimic a bell shape. It is important to emphasize that these visual checks of convergence can only probabilistically suggest whether or not the chains are representative of the posterior, they can not guarantee representativeness (KRUSCHKE, 2014).

## D.2 Convergence diagnostic plots for artificial data sets

In this section, we present the trace diagrams, the kernel density estimates, and the autocorrelation plots of the marginal posterior distributions of the component parameters for the synthetic data sets presented in section 4.5. The straight lines, the dashed (black) lines, and dotted (red) lines on trace and density plots represent the true parameter values, the median and the mean of the MCMC draws, respectively. The shaded areas correspond to 95% highest posterior density (HPD) intervals. We present the diagrams separately for each data set considered.

### *D.2.1 Homogeneous behavior*

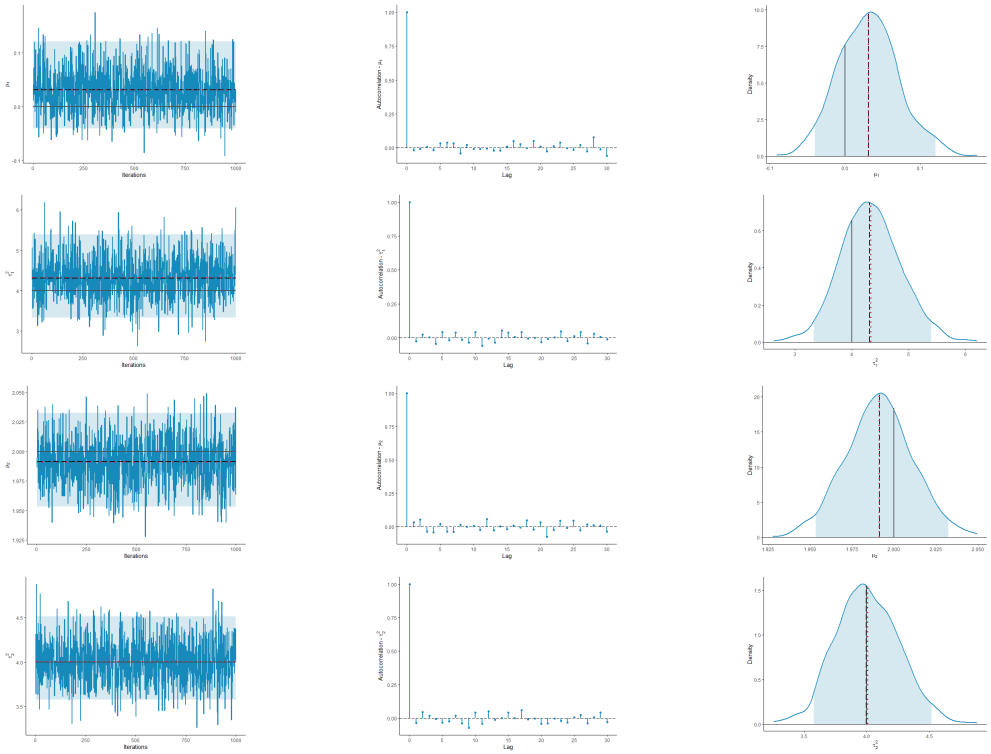


Figure 33 – Algorithm 1 (data set with $\alpha_t$'s following the homogeneous behavior) - convergence diagnostic plots for $\mu_1$ (first row), $\tau_1^2$ (second row), $\mu_2$ (third row), and $\tau_2^2$ (fourth row).

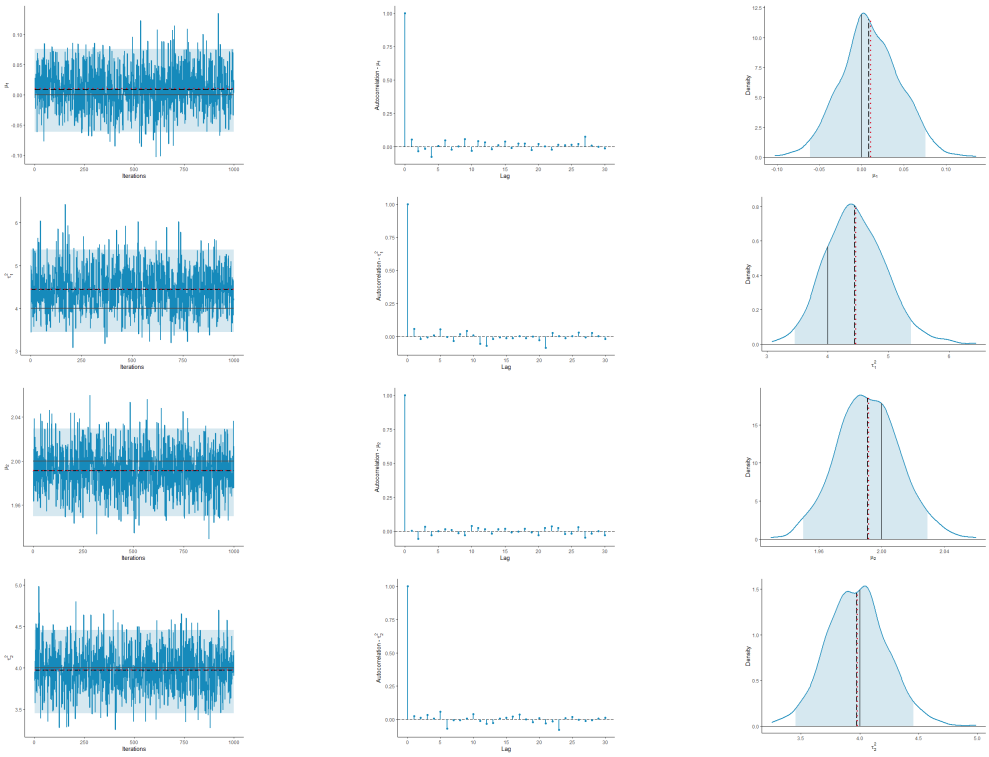


Figure 34 – Algorithm 2 (data set with $\alpha_t$'s following the homogeneous behavior) - convergence diagnostic plots for $\mu_1$ (first row), $\tau_1^2$ (second row), $\mu_2$ (third row), and $\tau_2^2$ (fourth row).

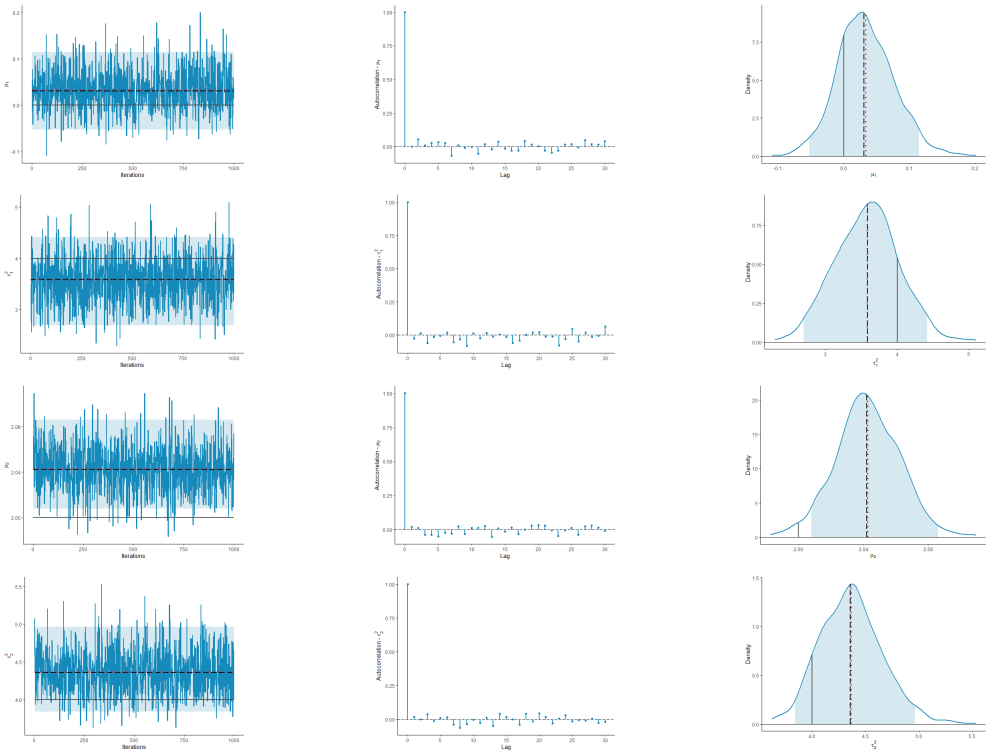


Figure 35 – Algorithm 3 (data set with $\alpha_t$'s following the homogeneous behavior) - convergence diagnostic plots for $\mu_1$ (first row), $\tau_1^2$ (second row), $\mu_2$ (third row), and $\tau_2^2$ (fourth row).

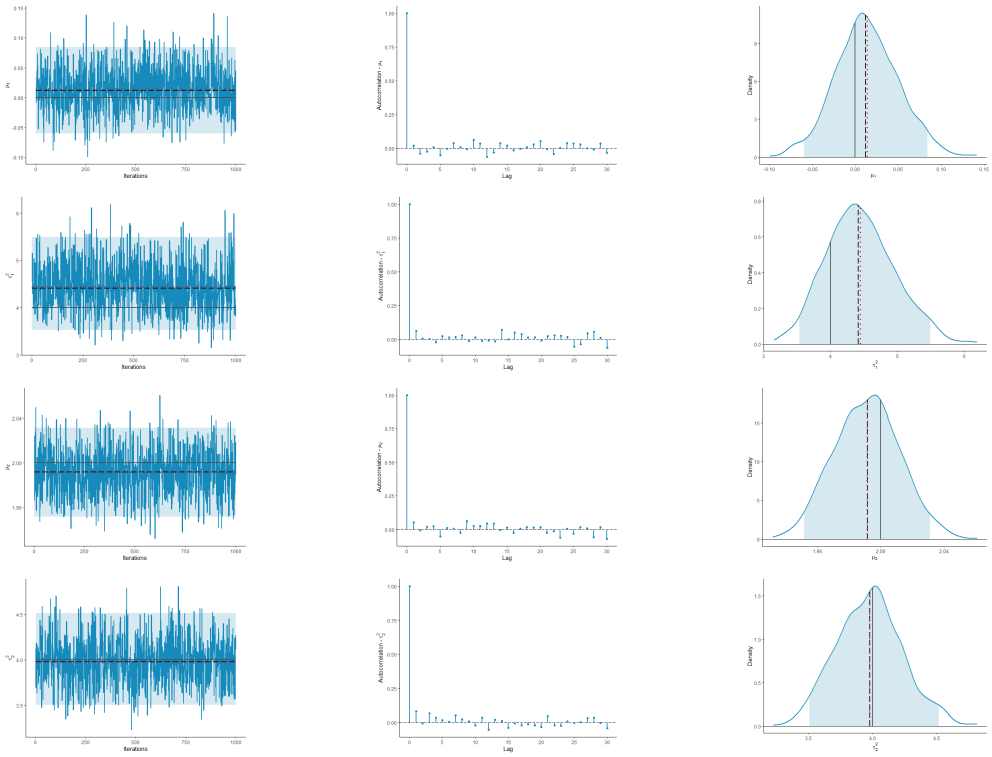


Figure 36 – Algorithm 4 (data set with $\alpha_t$'s following the homogeneous behavior) - convergence diagnostic plots for $\mu_1$ (first row), $\tau_1^2$ (second row), $\mu_2$ (third row), and $\tau_2^2$ (fourth row).

### *D.2.2 Parabolic behavior*

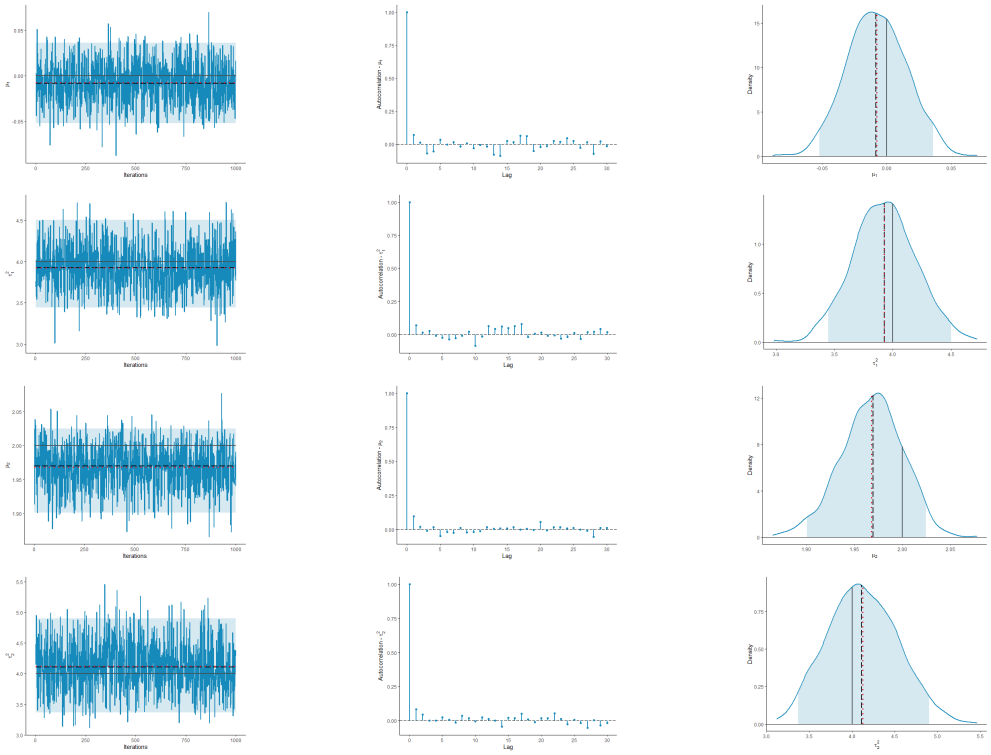


Figure 37 – Algorithm 1 (data set with $\alpha_t$'s following the parabolic behavior) - convergence diagnostic plots for $\mu_1$ (first row), $\tau_1^2$ (second row), $\mu_2$ (third row), and $\tau_2^2$ (fourth row).

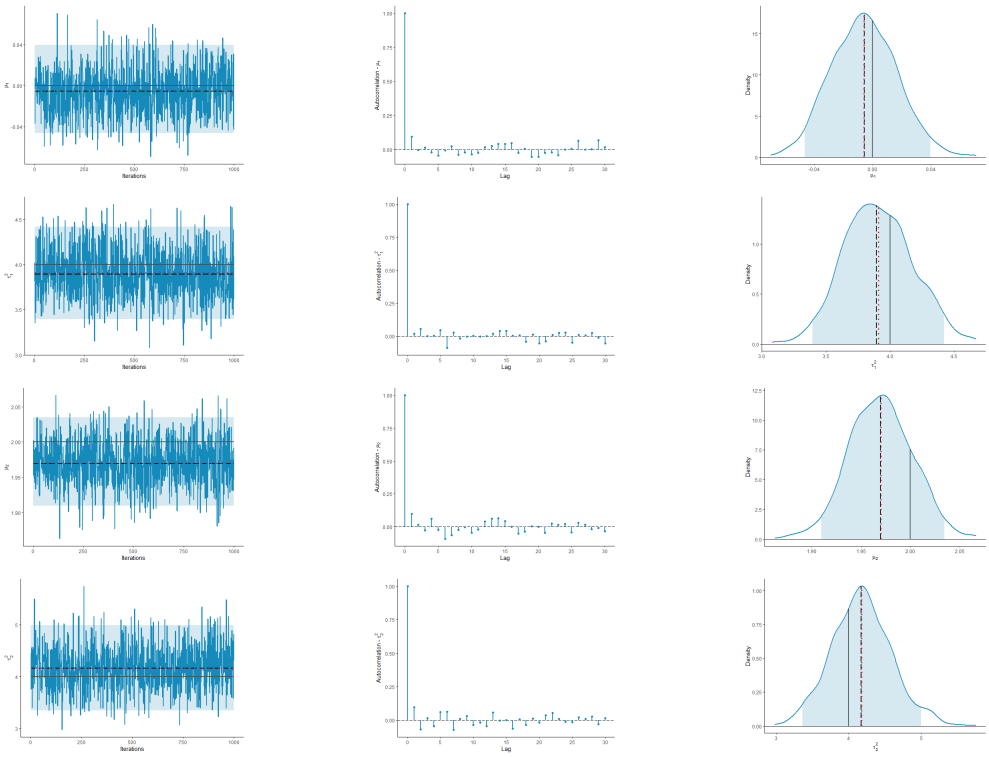


Figure 38 – Algorithm 2 (data set with $\alpha_t$'s following the parabolic behavior) - convergence diagnostic plots for $\mu_1$ (first row), $\tau_1^2$ (second row), $\mu_2$ (third row), and $\tau_2^2$ (fourth row).

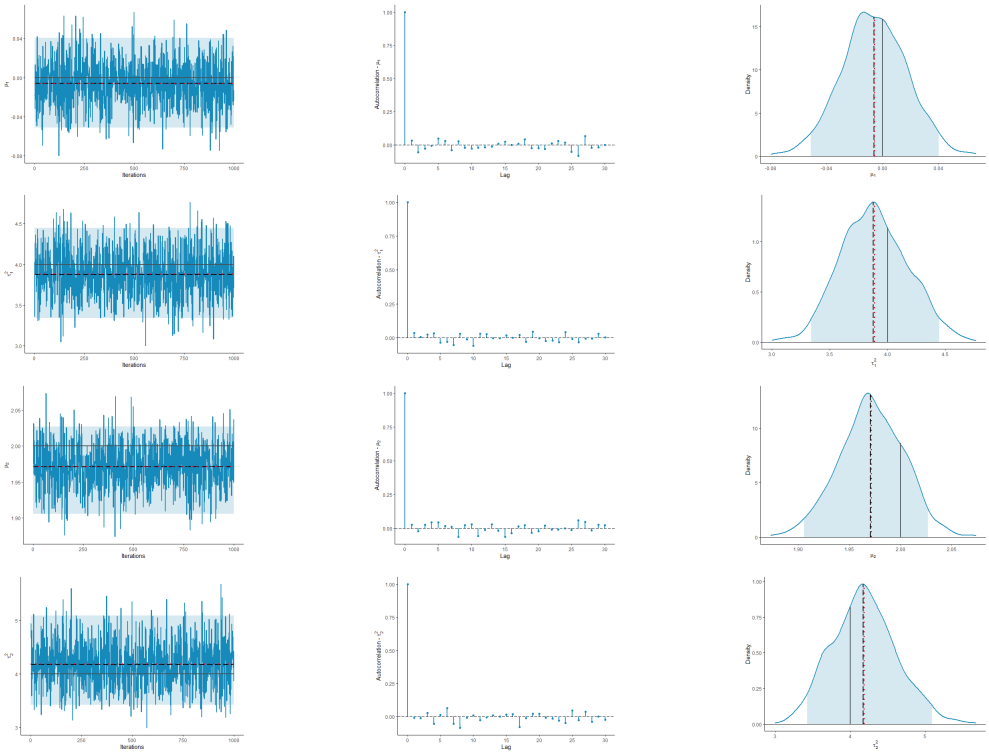


Figure 39 – Algorithm 3 (data set with $\alpha_t$'s following the parabolic behavior) - convergence diagnostic plots for $\mu_1$ (first row), $\tau_1^2$ (second row), $\mu_2$ (third row), and $\tau_2^2$ (fourth row).

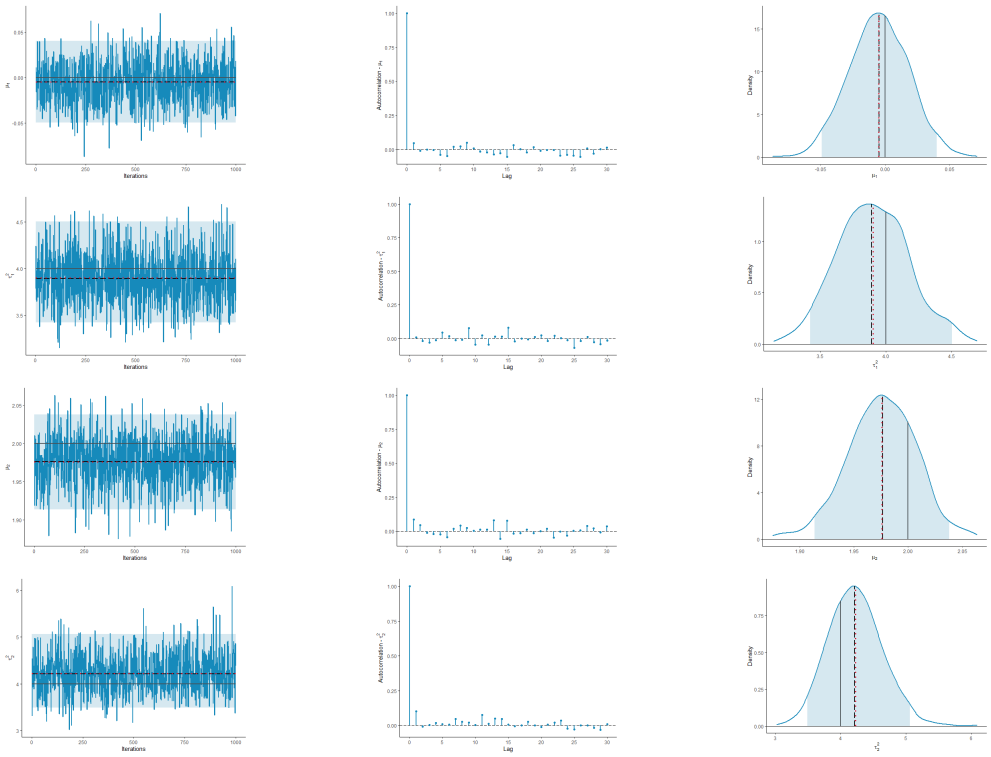


Figure 40 – Algorithm 4 (data set with $\alpha_t$'s following the parabolic behavior) - convergence diagnostic plots for $\mu_1$ (first row), $\tau_1^2$ (second row), $\mu_2$ (third row), and $\tau_2^2$ (fourth row).

### *D.2.3 Sinusoidal behavior*

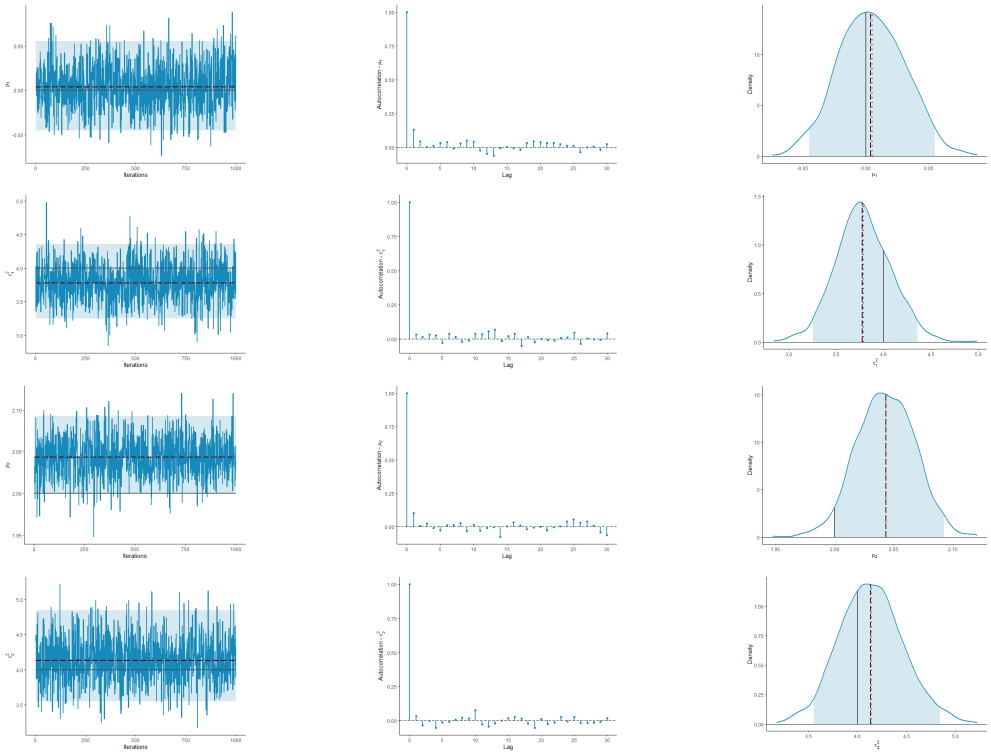


Figure 41 – Algorithm 1 (data set with $\alpha_t$'s following the sinusoidal behavior) - convergence diagnostic plots for $\mu_1$ (first row), $\tau_1^2$ (second row), $\mu_2$ (third row), and $\tau_2^2$ (fourth row).

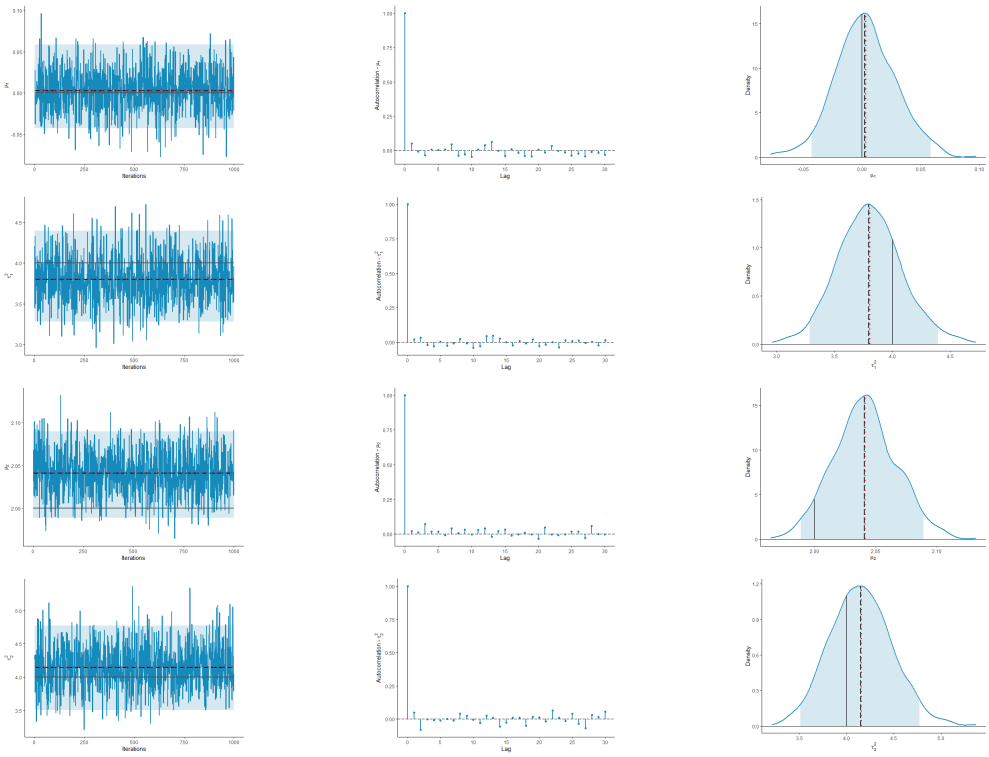


Figure 42 – Algorithm 2 (data set with $\alpha_t$'s following the sinusoidal behavior) - convergence diagnostic plots for $\mu_1$ (first row), $\tau_1^2$ (second row), $\mu_2$ (third row), and $\tau_2^2$ (fourth row).

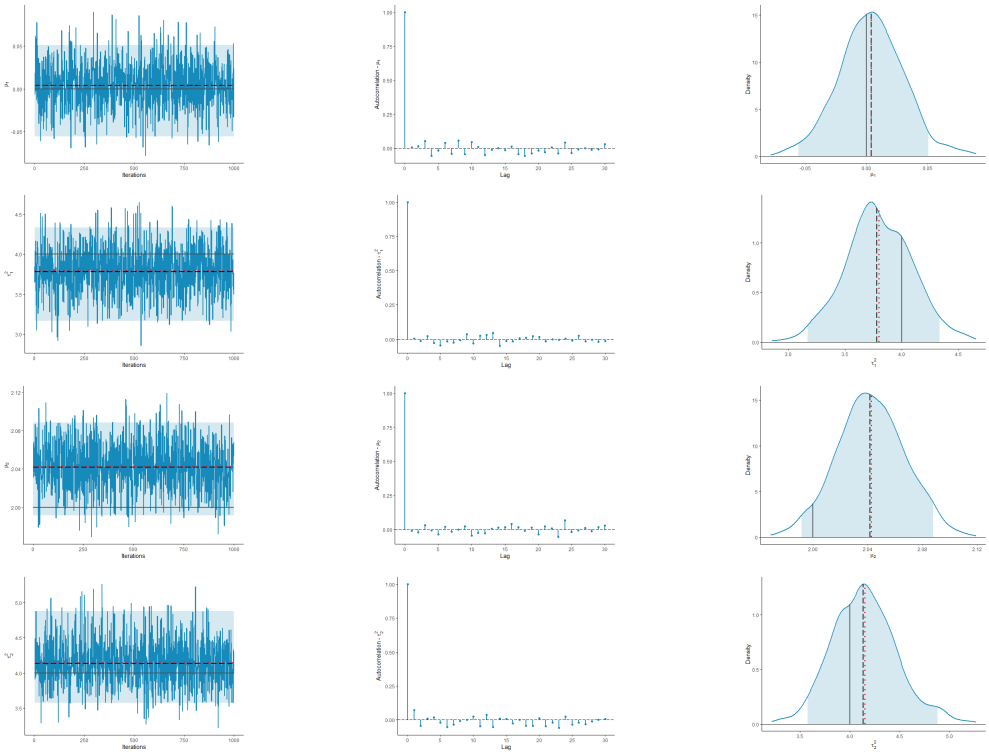


Figure 43 – Algorithm 3 (data set with $\alpha_t$'s following the sinusoidal behavior) - convergence diagnostic plots for $\mu_1$ (first row), $\tau_1^2$ (second row), $\mu_2$ (third row), and $\tau_2^2$ (fourth row).

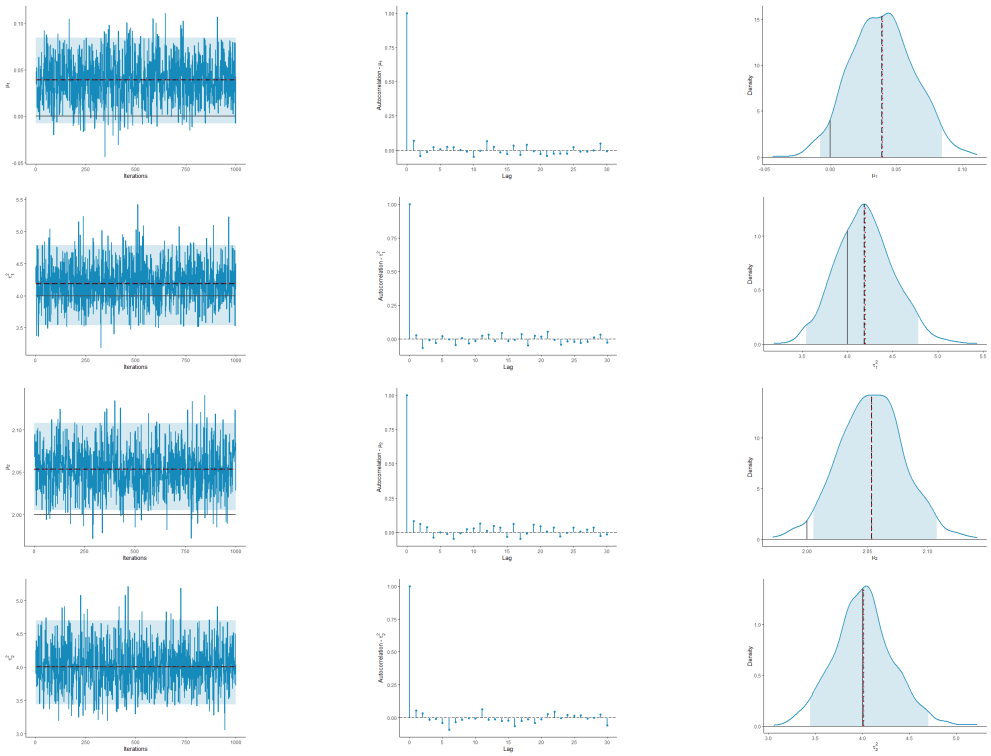


Figure 44 – Algorithm 4 (data set with $\alpha_t$'s following the sinusoidal behavior) - convergence diagnostic plots for $\mu_1$ (first row), $\tau_1^2$ (second row), $\mu_2$ (third row), and $\tau_2^2$ (fourth row).

### *D.2.4 Heavisine behavior*

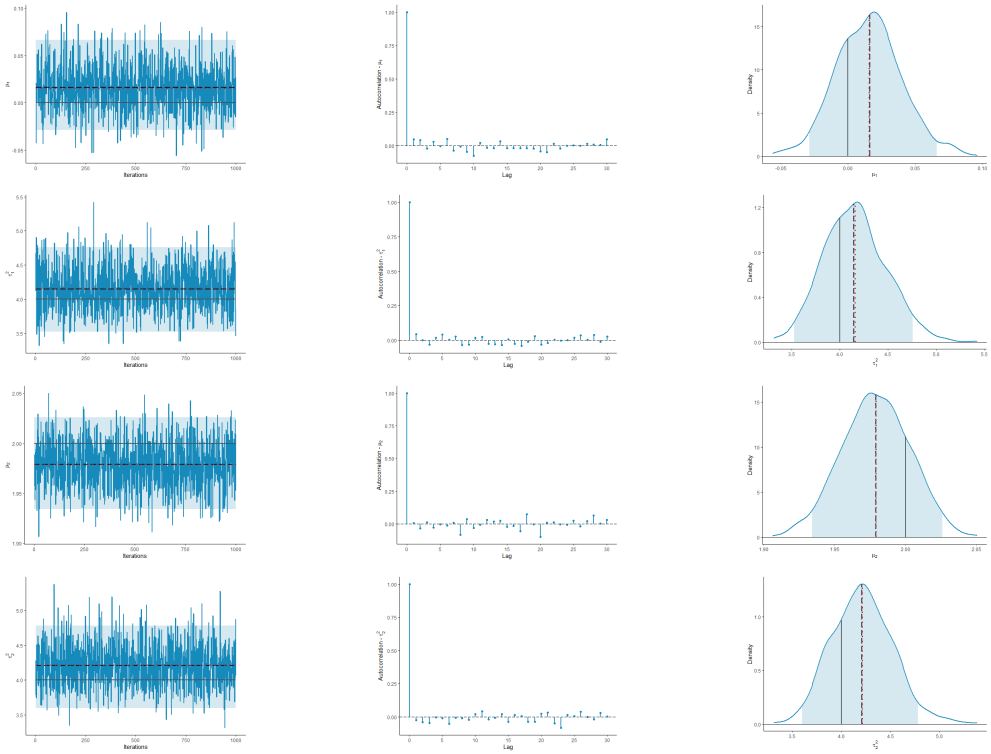


Figure 45 – Algorithm 1 (data set with $\alpha_t$'s following the heavisine behavior) - convergence diagnostic plots for $\mu_1$ (first row), $\tau_1^2$ (second row), $\mu_2$ (third row), and $\tau_2^2$ (fourth row).

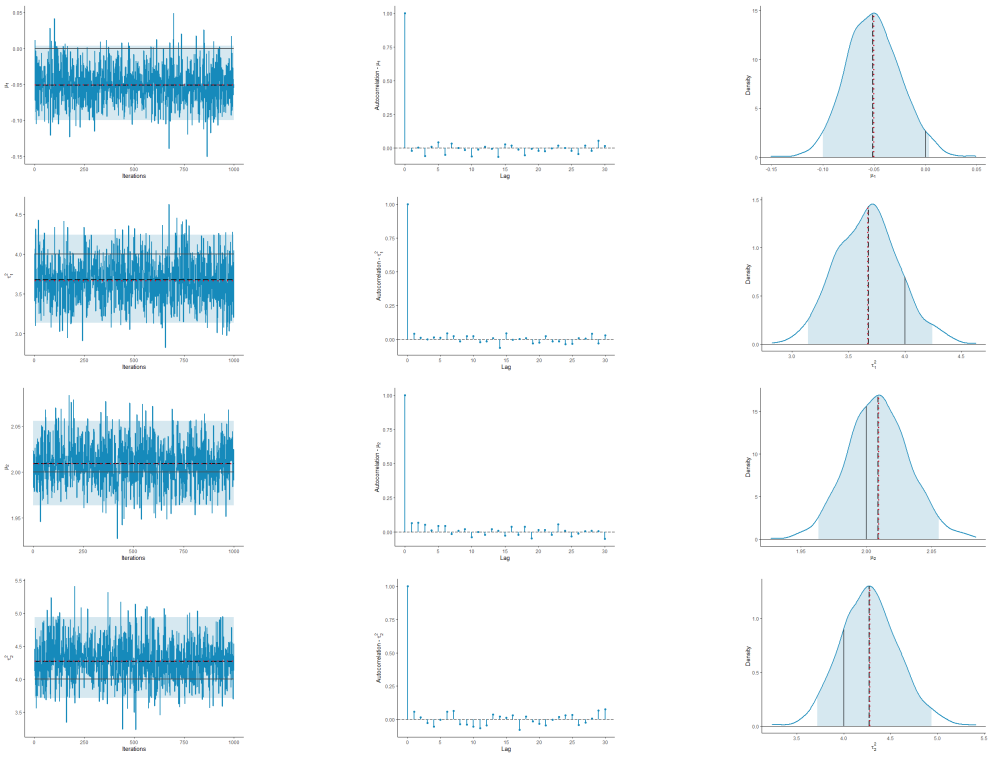


Figure 46 – Algorithm 2 (data set with $\alpha_t$'s following the heavisine behavior) - convergence diagnostic plots for $\mu_1$ (first row), $\tau_1^2$ (second row), $\mu_2$ (third row), and $\tau_2^2$ (fourth row).

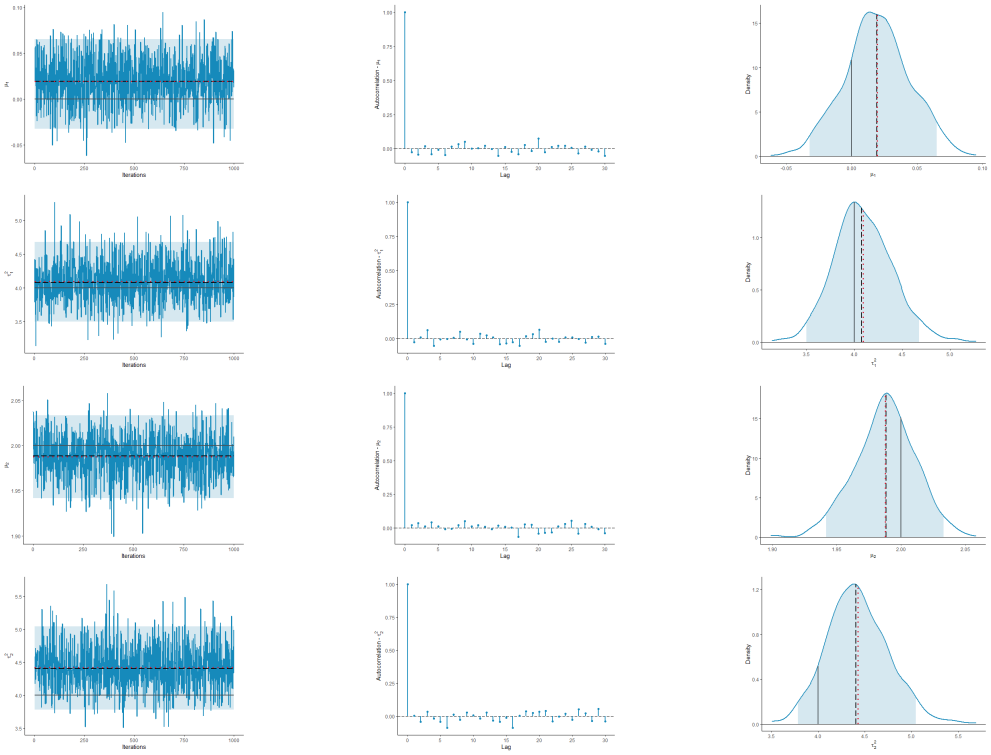


Figure 47 – Algorithm 3 (data set with $\alpha_t$'s following the heavisine behavior) - convergence diagnostic plots for $\mu_1$ (first row), $\tau_1^2$ (second row), $\mu_2$ (third row), and $\tau_2^2$ (fourth row).

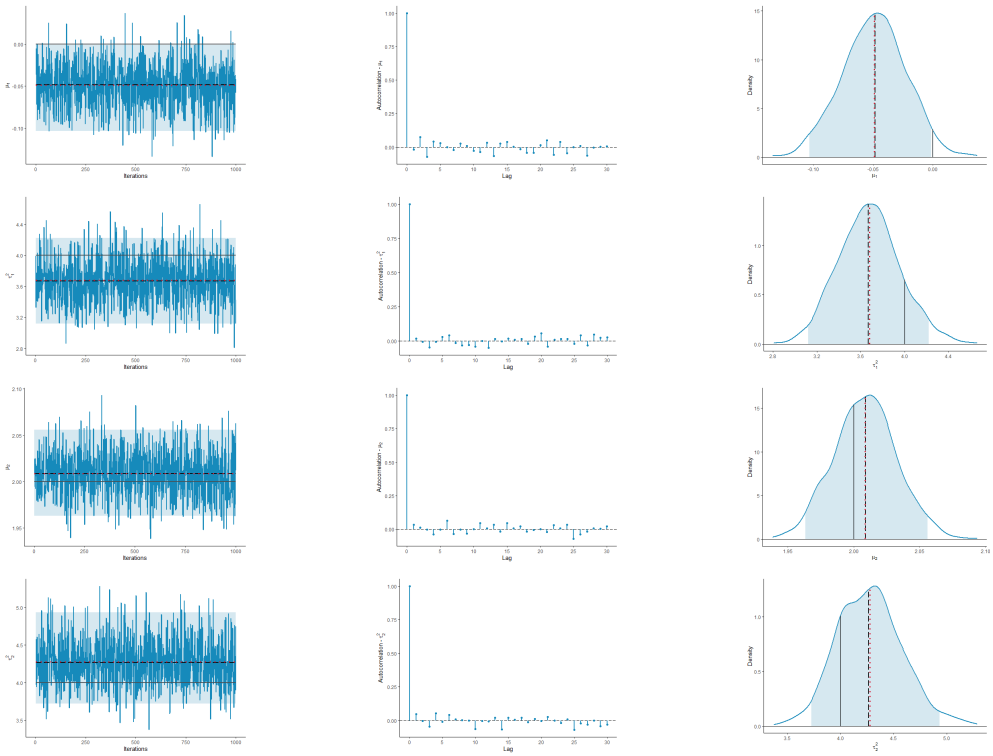


Figure 48 – Algorithm 4 (data set with $\alpha_t$'s following the heavisine behavior) - convergence diagnostic plots for $\mu_1$ (first row), $\tau_1^2$ (second row), $\mu_2$ (third row), and $\tau_2^2$ (fourth row).

### *D.2.5 Bumps behavior*

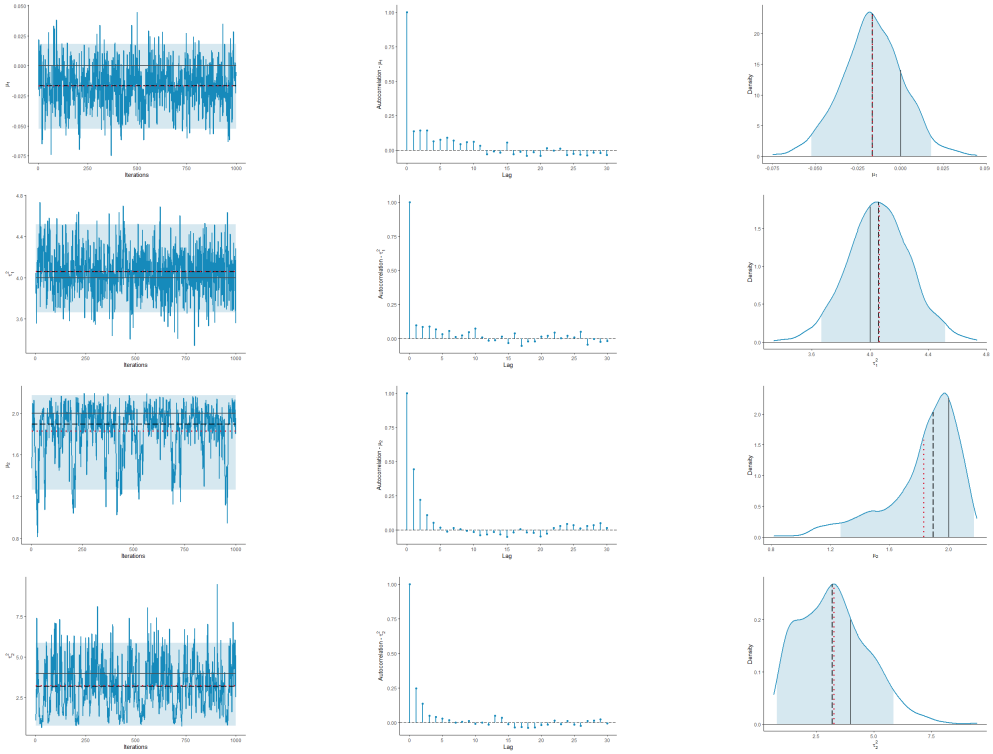


Figure 49 – Algorithm 1 (data set with $\alpha_t$'s following the bumps behavior) - convergence diagnostic plots for $\mu_1$ (first row), $\tau_1^2$ (second row), $\mu_2$ (third row), and $\tau_2^2$ (fourth row).

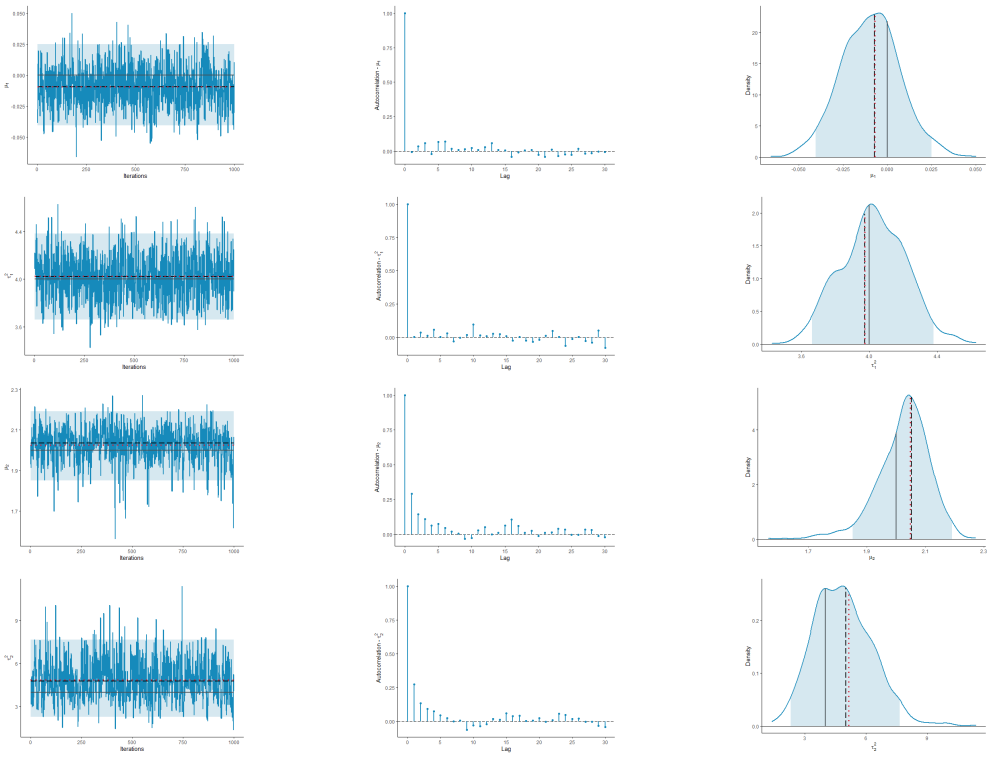


Figure 50 – Algorithm 2 (data set with $\alpha_t$'s following the bumps behavior) - convergence diagnostic plots for $\mu_1$ (first row), $\tau_1^2$ (second row), $\mu_2$ (third row), and $\tau_2^2$ (fourth row).

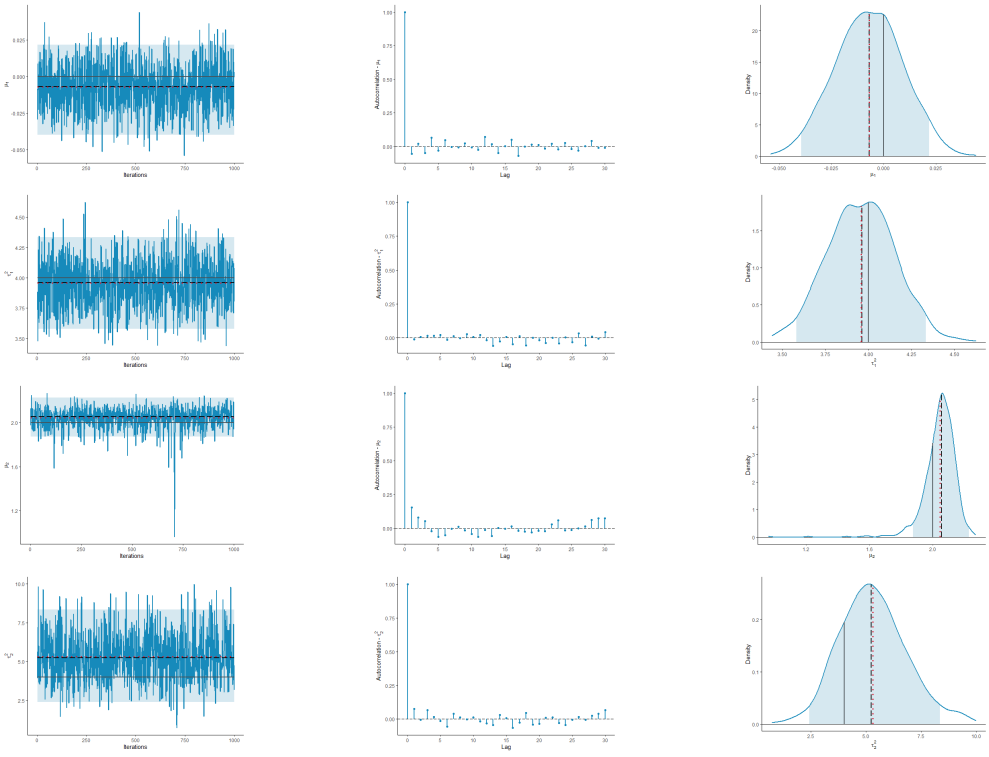


Figure 51 – Algorithm 3 (data set with $\alpha_t$'s following the bumps behavior) - convergence diagnostic plots for $\mu_1$ (first row), $\tau_1^2$ (second row), $\mu_2$ (third row), and $\tau_2^2$ (fourth row).

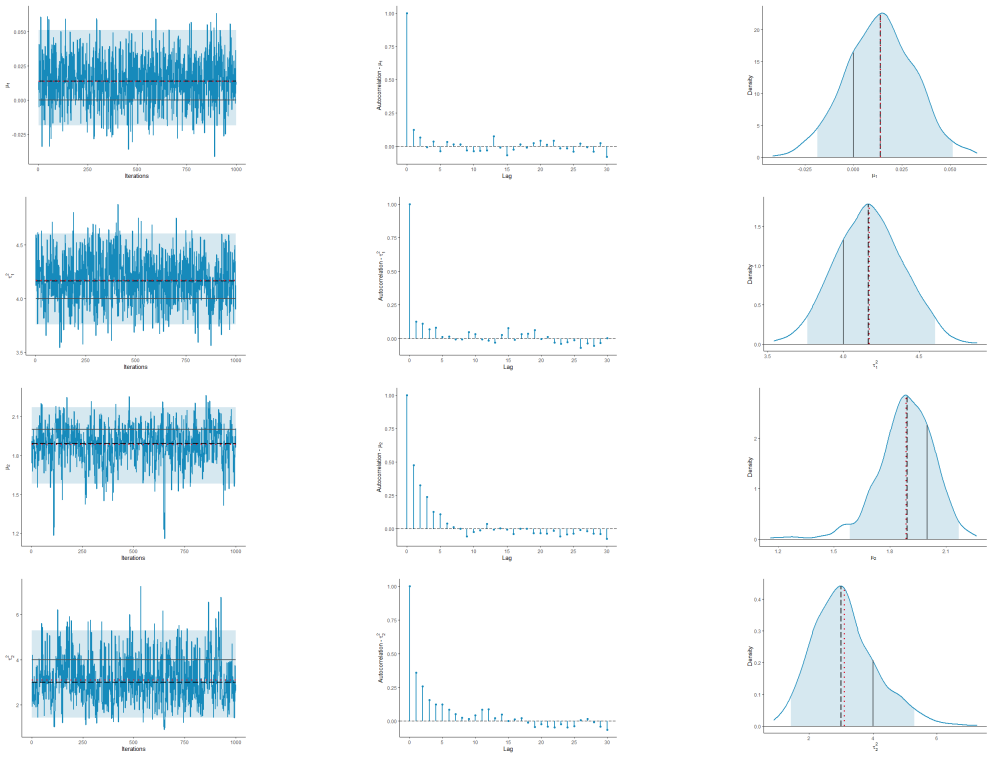


Figure 52 – Algorithm 4 (data set with $\alpha_t$'s following the bumps behavior) - convergence diagnostic plots for $\mu_1$ (first row), $\tau_1^2$ (second row), $\mu_2$ (third row), and $\tau_2^2$ (fourth row).

### *D.2.6 Blocks behavior*

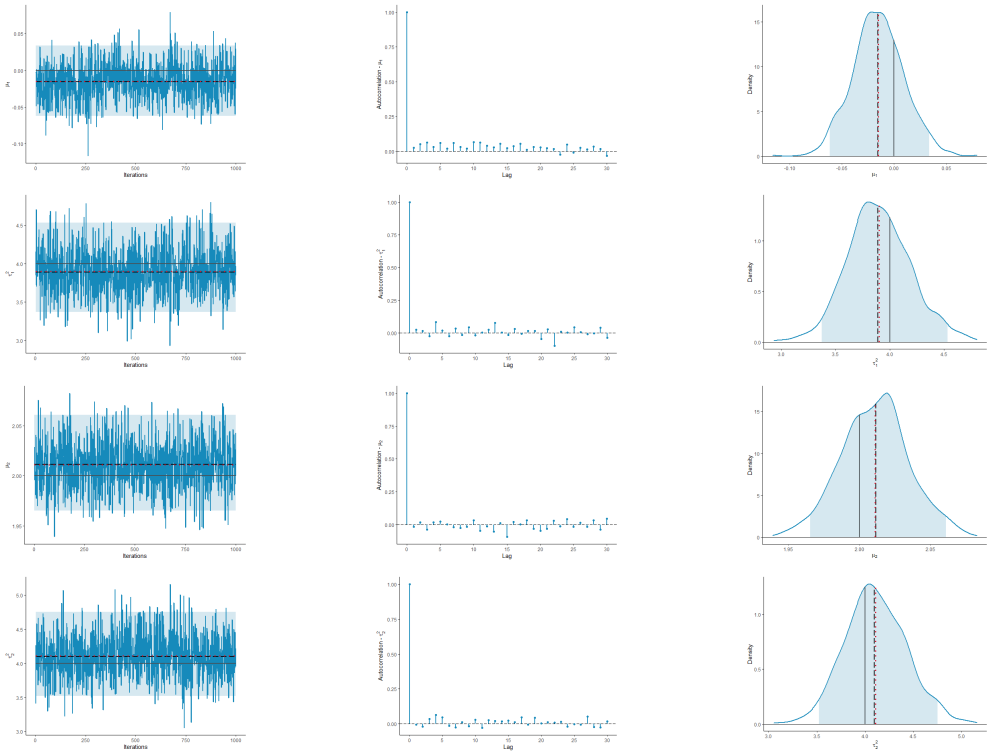


Figure 53 – Algorithm 1 (data set with $\alpha_t$'s following the blocks behavior) - convergence diagnostic plots for $\mu_1$ (first row), $\tau_1^2$ (second row), $\mu_2$ (third row), and $\tau_2^2$ (fourth row).

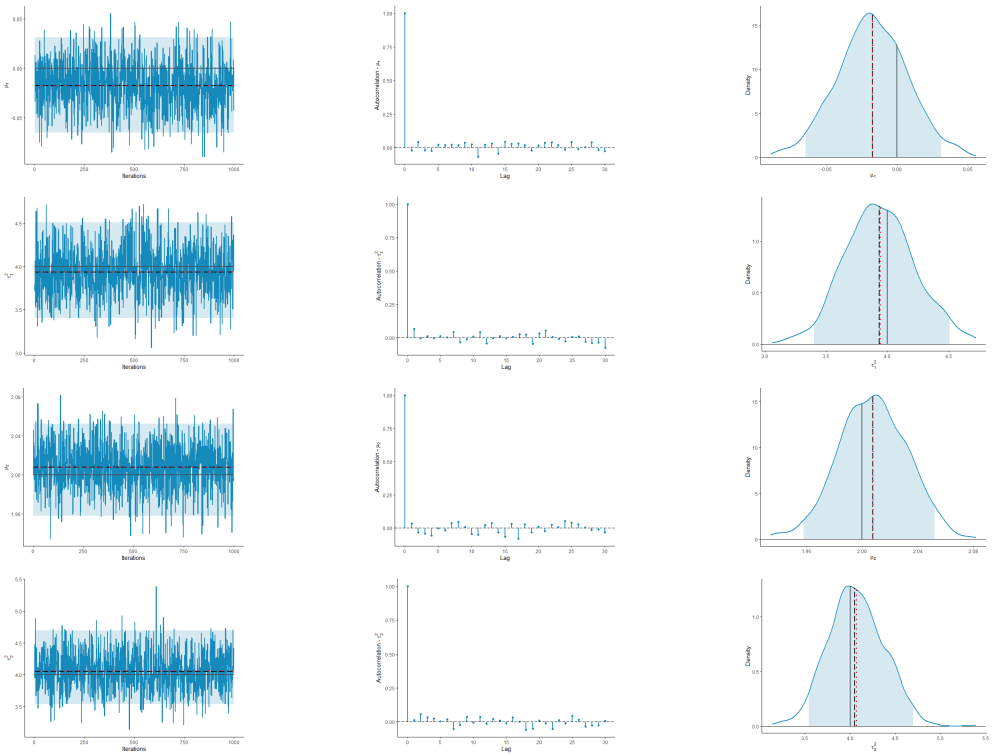


Figure 54 – Algorithm 2 (data set with $\alpha_t$'s following the blocks behavior) - convergence diagnostic plots for $\mu_1$ (first row), $\tau_1^2$ (second row), $\mu_2$ (third row), and $\tau_2^2$ (fourth row).

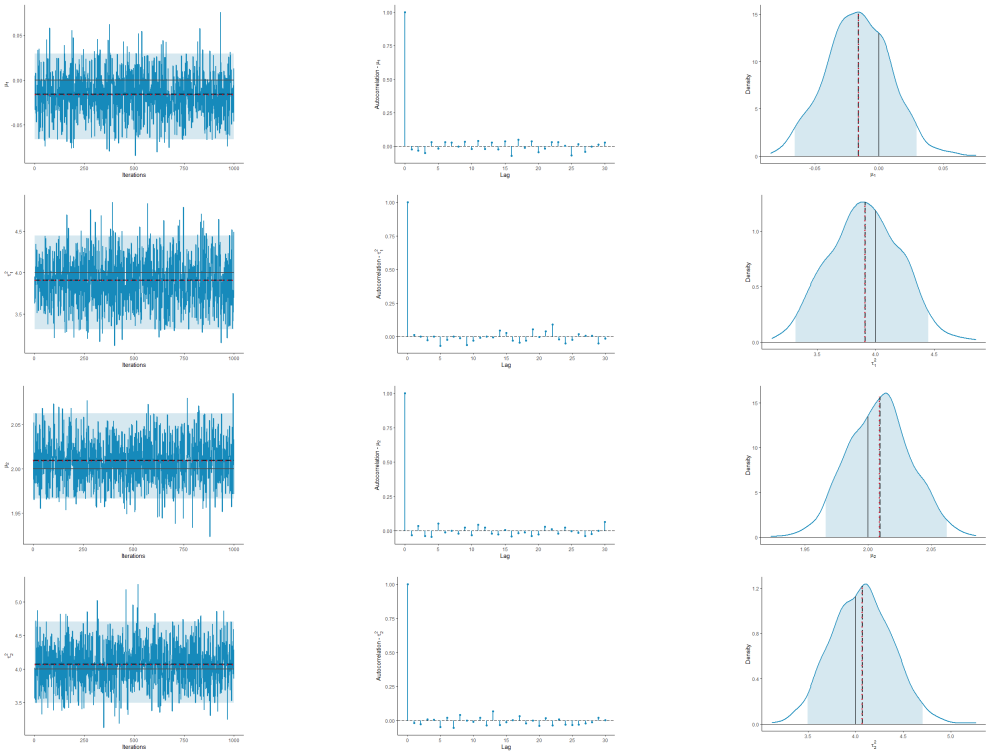


Figure 55 – Algorithm 3 (data set with $\alpha_t$'s following the blocks behavior) - convergence diagnostic plots for $\mu_1$ (first row), $\tau_1^2$ (second row), $\mu_2$ (third row), and $\tau_2^2$ (fourth row).

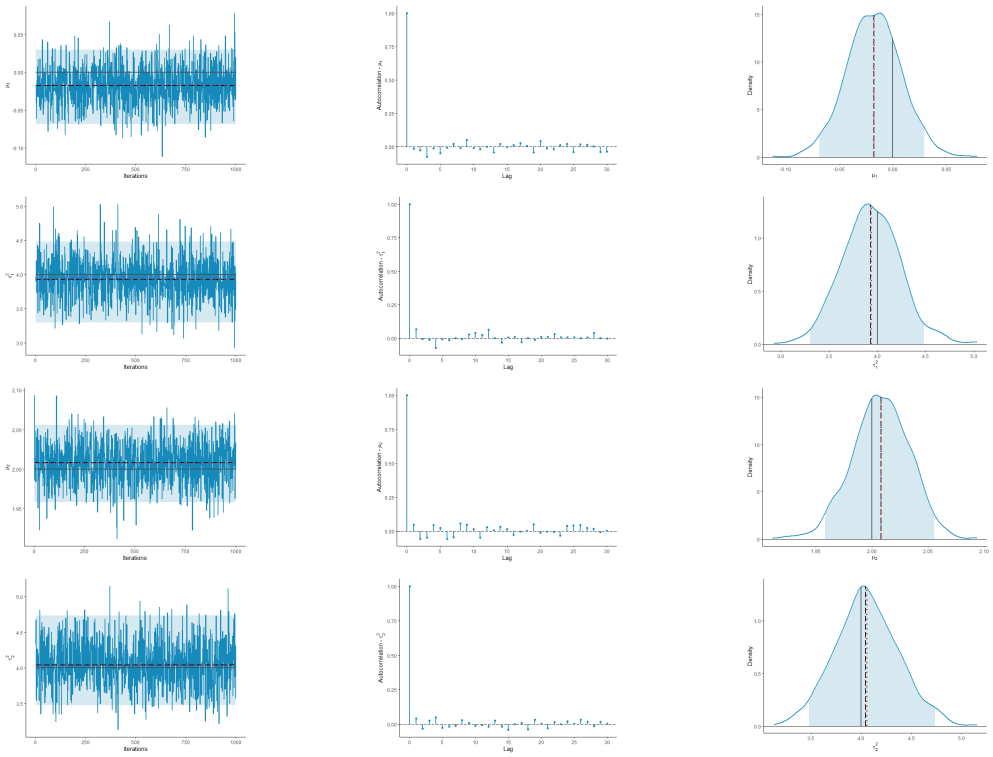


Figure 56 – Algorithm 4 (data set with $\alpha_t$'s following the blocks behavior) - convergence diagnostic plots for $\mu_1$ (first row), $\tau_1^2$ (second row), $\mu_2$ (third row), and $\tau_2^2$ (fourth row).

## D.3 Convergence diagnostic plots for aCGH data set

In this section, we present the trace diagrams, the kernel density estimates, and the autocorrelation plots of the marginal posterior distributions of the component parameters for the aCGH data set. Thus we do not know the true parameter values. The dashed (black) lines and dotted (red) lines on trace and density plots represent the median and the mean of the MCMC draws, respectively. The shaded areas correspond to 95% highest posterior density (HPD) intervals. We present the diagrams separately for each approach considered.

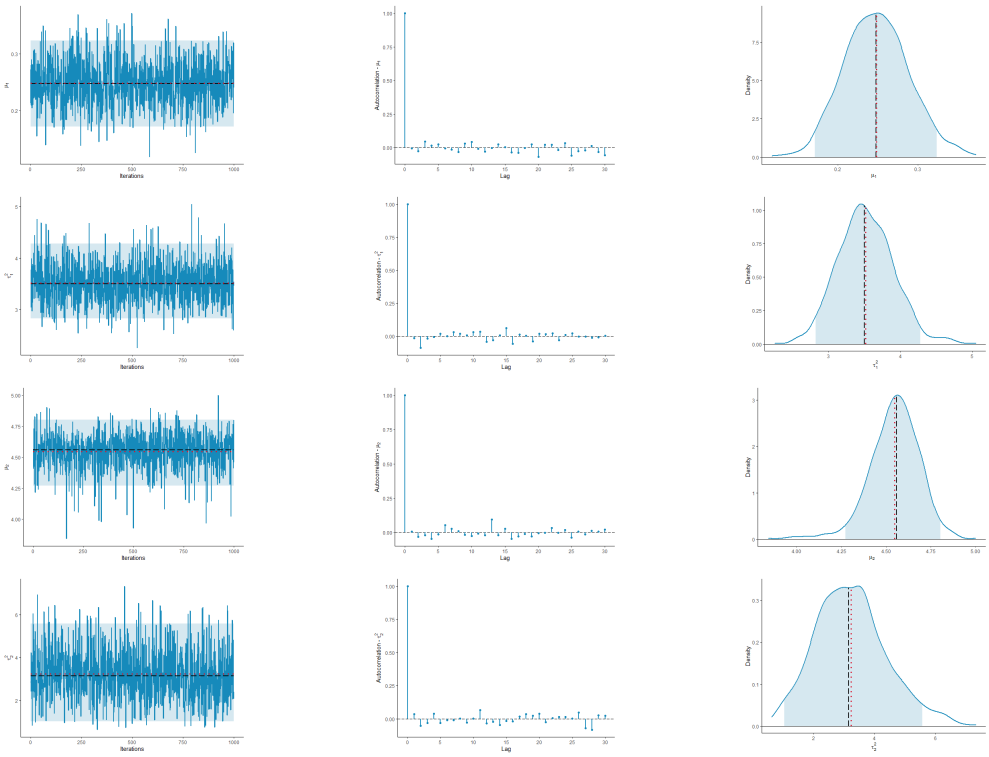


Figure 57 – WR approach - convergence diagnostic plots for $\mu_1$ (first row), $\tau_1^2$ (second row), $\mu_2$ (third row), and $\tau_2^2$ (fourth row).

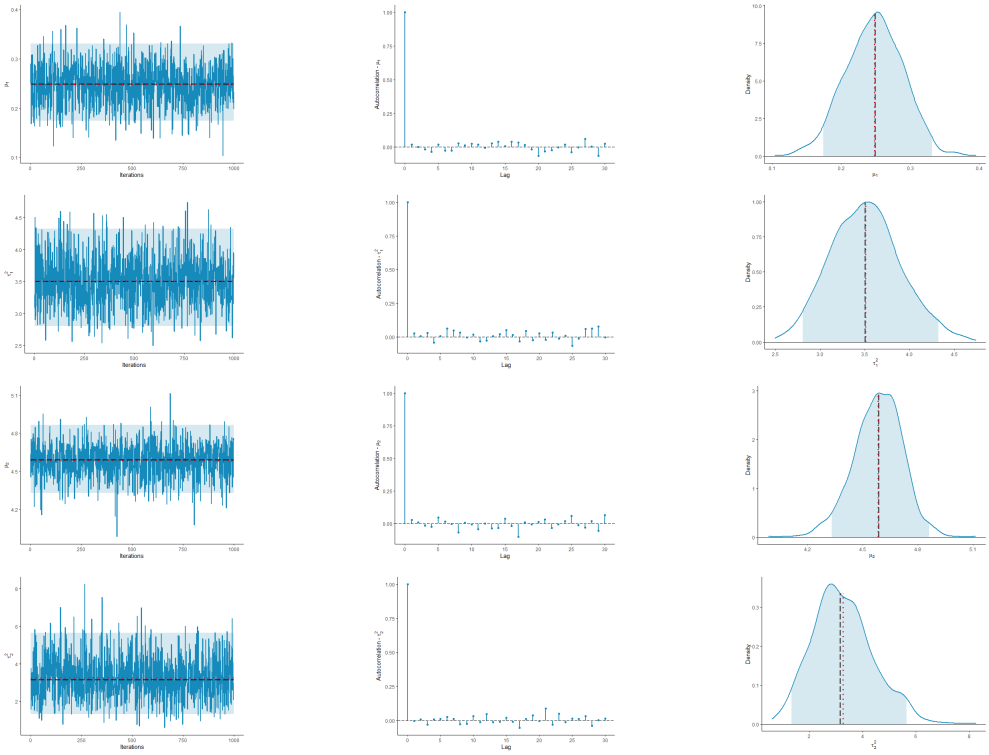


Figure 58 – DA approach: Diffuse prior (aCGH data set) - convergence diagnostic plots for $\mu_1$, $\tau_1^2$, $\mu_2$, and $\tau_2^2$.

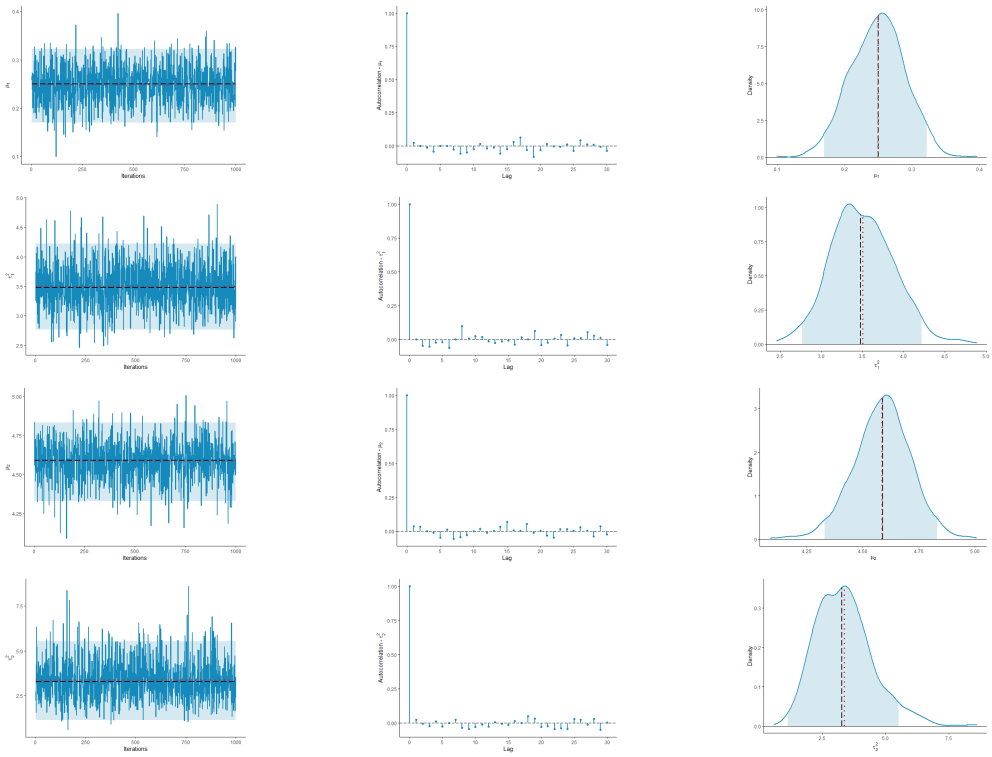


Figure 59 – DA approach: Gaussian prior (aCGH data set) - convergence diagnostic plots for $\mu_1$, $\tau_1^2$, $\mu_2$, and $\tau_2^2$.

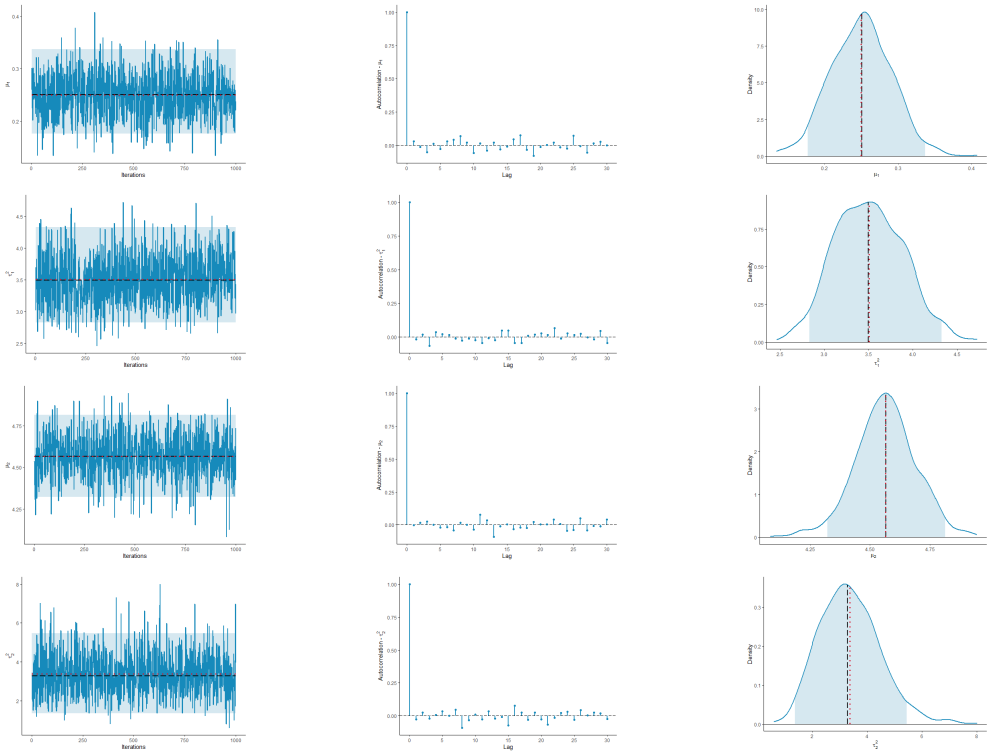


Figure 60 – DA approach: SSG prior (aCGH data set) - convergence diagnostic plots for $\mu_1$, $\tau_1^2$, $\mu_2$, and $\tau_2^2$.

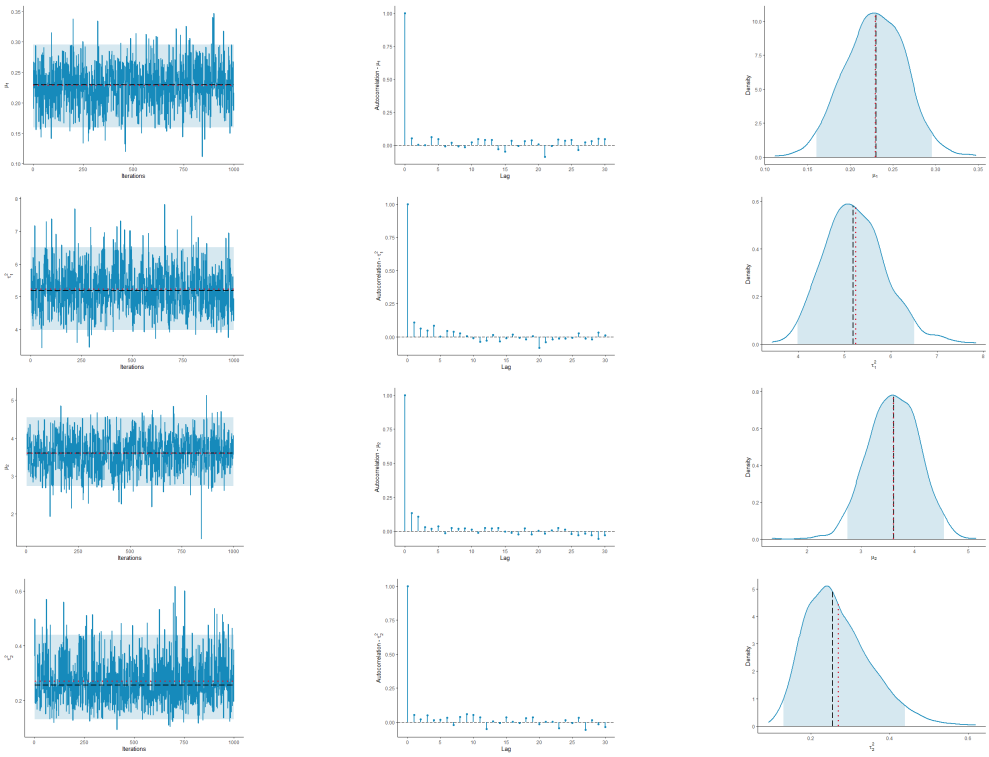


Figure 61 – DA approach: SSL prior (aCGH data set) - convergence diagnostic plots for $\mu_1$, $\tau_1^2$, $\mu_2$, and $\tau_2^2$.


ICMC
USP
SÃO CARLOS