UNIVERSIDADE FEDERAL DO RIO DE JANEIRO

INSTITUTO DE MATEMÁTICA

PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

DANIEL WÜRZLER BARRETO

A MODEL-BASED BAYESIAN APPROACH TO ANOMALY DETECTION VIA MIXTURE MODELS

RIO DE JANEIRO

2023

DANIEL WÜRZLER BARRETO

A MODEL-BASED BAYESIAN APPROACH TO ANOMALY DETECTION VIA MIXTURE MODELS

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Estatística do Instituto de Matemática da Universidade Federal do Rio de Janeiro - UFRJ, como parte dos requisitos necessários para obtenção do título de Mestre em Estatística.

Advisor: Prof.ª Marina Silva Paez, Ph.D.

RIO DE JANEIRO

2023

CIP - Catalogação na Publicação

B273m

Barreto, Daniel Würzler
A model-based Bayesian approach to anomaly detection via mixture models / Daniel Würzler Barreto. -- Rio de Janeiro, 2023.
145 f.

Orientadora: Marina Silva Paez.
Dissertação (mestrado) - Universidade Federal do Rio de Janeiro, Instituto de Matemática, Programa de Pós-Graduação em Estatística, 2023.

1. Bayesian inference. 2. Mixture model. 3. Anomaly detection. I. Paez, Marina Silva, orient. II. Título.

Elaborado pelo Sistema de Geração Automática da UFRJ com os dados fornecidos pelo(a) autor(a), sob a responsabilidade de Miguel Romeu Amorim Neto - CRB-7/6283.

DANIEL WÜRZLER BARRETO

A MODEL-BASED BAYESIAN APPROACH TO ANOMALY DETECTION VIA MIXTURE MODELS

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Estatística do Instituto de Matemática da Universidade Federal do Rio de Janeiro - UFRJ, como parte dos requisitos necessários para obtenção do título de Mestre em Estatística.

Aprovado em _____ de _____ de _____ .

BANCA EXAMINADORA:

_____

Prof.ª Marina Silva Paez, Ph.D.

_____

Prof.ª Mariane Branco Alves, Ph.D.

_____

Prof. Carlos Tadeu Pagani Zanini, Ph.D.

_____

Prof. Vinicius Pinheiro Israel, Ph.D.

ACKNOWLEDGEMENTS

No achievement is ever an individual feat. So, I dedicate this section to those who, in one way or another, provided the guidance and aid needed to conclude this work.

I am grateful for the unconditional love and affection of my family. I cannot put into words the appreciation I feel for my parents, who always believed in, supported, and encouraged me even when I could not see any way forward. To my brother and sister, I thank you for the companionship and understanding when I needed it most.

I am grateful to my friends for providing joy in the good moments and consolation in the bad ones. In particular, it is impossible not to mention Silvaneo Vieira dos Santos Junior, who I could always rely on to share my struggles with and keep me motivated.

I am grateful for my professors, who provided me with the theoretical tools and practical knowledge to challenge my limits. I thank Professor Marina Silva Paez, who had the patience to accompany my growth since my first year of undergraduate studies, for her support and willingness to listen and understand my ideas and suggestions.

I am grateful to the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior and the Laboratório de Matemática Aplicada for their consistent financial support.

Essentially, all models are wrong, but
some are useful.

George E. P. Box

ABSTRACT

In this work, we present a Bayesian model-based approach to anomaly detection using mixture models. Our proposed method, called the filtering model, only requires us to specify a parametric model, that depends on an unknown θ, to describe the behavior of the typical data and uses the chosen model to determine the underlying distribution of the component of the mixture responsible for capturing anomalies. The method is able to simultaneously estimate the classification for each observation and θ, while taking the estimate of θ to be a convex combination of each possible estimate generated by a subsample. For this reason, it can also be used for robust parameter estimation. We consider estimation using Markov chain Monte Carlo techniques, and in particular the Metropolis-Hastings algorithm, and present applications for chemical, health and demographic data.

Keywords: Bayesian inference. Mixture model. Anomaly detection.

CONTENTS

1 INTRODUCTION

The data is the main object of study of a statistician, because even if it is not the only possible source of information, it is generally the primary one used for all inferential procedures. Those, in turn, have an important role on the processes of prediction and explanation of phenomena of interest within multiple fields of study, such as medicine, pharmacology, genetics, finance, economy, psychology, geography, and many others. However, not every data set is exempt of flaws, possibly containing erroneous, inconsistent or atypical entries. The presence of these types of observations throughout inference making may negatively influence results, potentially leading to biased conclusions or decisions. With that in mind, the main objective of this work is to develop a new methodology designed to address the problems of identifying and treating these types of entries.

There are many possible causes that may generate a data set containing problematic observations. The cause may be due to typos during the transcription of the data, noisy measuring instruments, sudden change in behavior, atypical events, corruption of files or even a security breach, since information may be maliciously adulterated for multiple reasons. Naturally, this multiplicity of factors make this a common nuisance during data analysis and consequently attracts the attention of many researches with various backgrounds. Thus, it is not surprising that there are many distinct lines of research within this area, each one based on different techniques.

In this work we propose a methodology for anomaly detection based on mixture models. Our approach uses simple principles to define a mixture component responsible for capturing outliers, given a chosen parametric family of distribution to capture the typical behavior of data. Since this component is model-induced, the resulting model is able to deal with non-identically distributed observations, multivariate data and time series.

For a brief concrete motivational example, Figure 1 shows curves of near infrared absorbance spectra at different wavelengths for 39 gasoline samples. As we can see,

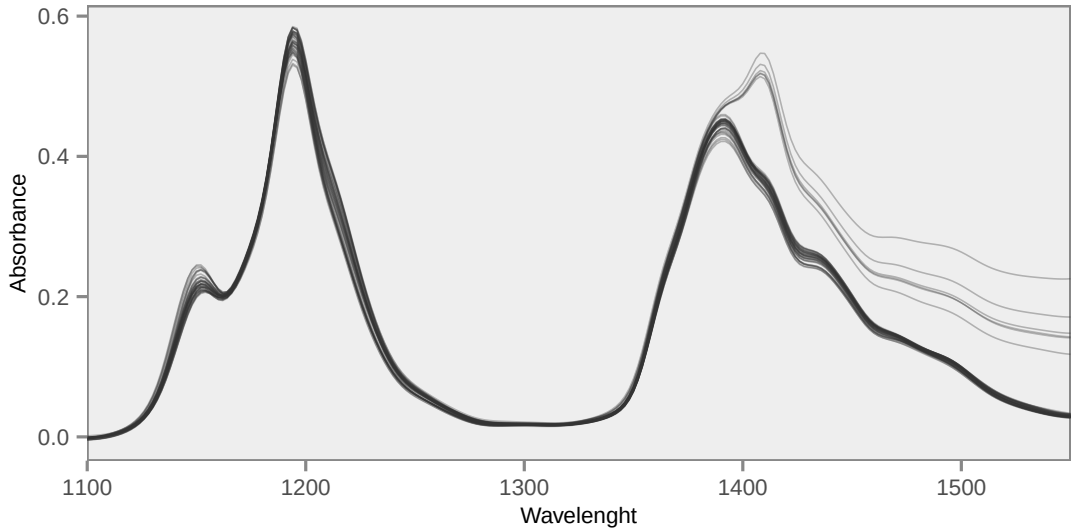

Figure 1: Absorbance curves for gasoline samples from the octane data set at different wavelengths.

some of the observations do match the behavior of the rest, detaching from the other curves for wavelengths greater than $1390nm$. Interestingly, these are observations 25, 26 and 36-39 and known to be outliers because they correspond to gasoline samples that were contaminated with ethanol.

Since our approach is based on mixture models, in principle we would need to choose a family of distributions for the typical and anomalous observations, however, our proposed method only requires us to characterize what we subjectively believe to be a reasonable description of the typical behavior of our data through the use of a parametric model. So, considering the data presented in Figure 1, if we choose a common time series model to capture the typical behavior of the observations, for instance a simple random walk model, then our method is able to incorporate the information from our choice to identify the anomalies without supervision.

In Chapter 2 we present the core theoretical foundations used in this work. In Chapter 3 we propose the *filtering model* and discuss estimation, interpretation, shortcomings, prediction and other topics regarding our method. In Chapter 4 we consider three applications, including the identification of contaminated gasoline samples using the near infrared absorbance spectra, the classification of breast tumors using quantitative features extracted from medical image exams and the iden-

tification of historic events using the estimated mortality rates for the male French population. Lastly, in Chapter 5 we provide some final considerations, including possible research topics for future work.

2 THEORETICAL FOUNDATION

In this chapter we present the theoretical tools used for the construction of the model proposed in Chapter 3 and for the applications of Chapter 4. Firstly, we introduce some concepts of Bayesian parametric inference in section 2.1; in section 2.2 we discuss numerical alternatives for estimation in cases of high analytical complexity; and sections 2.3 and 2.4 are dedicated to present the context of outlier detection methodologies based on mixture models and depth functions, respectively.

2.1 BAYESIAN PARAMETRIC INFERENCE

Given a family of probabilistic models, suppose we wish to find an estimate of an unknown generating process, that is, *the best* model to approximate a phenomenon of interest. Since probabilistic models are characterized by their distribution functions, finding *the best* probabilistic model can be thought as an optimization problem in space S, defined as the set containing all the distribution functions from our chosen family of models. So, in order to avoid the complications that arise from dealing with abstract spaces, it is a common practice to define a bijective mapping function ϕ from a simpler space Θ, e.g. $\mathbb{R}^n$, to S. We denominate Θ the parameter space and each $\theta \in \Theta$ a parametric vector, that represents a distribution function through the relationship $\phi(\theta) = F(\cdot|\theta)$, where $F(\cdot|\theta)$ is a distribution function from S. A family of models that have this type of representation is called parametric, and these families are advantageous because they allow us to change the possibly complicated problem of optimization in S to a simpler problem of optimization in Θ, i.e., we can find our estimate $\hat{F}$ by searching for $\hat{\theta}$, an estimate of *the best* parameter value, such that $\phi(\hat{\theta}) = F(\cdot|\hat{\theta}) = \hat{F}$. So we call parametric inference the process of finding the parameter that maps to *the best* model within a parametric family of distributions. In this section we will consider these inferential processes from the Bayesian perspective, and for a more detailed picture of this subject we redirect reader to Chapter 2 of Migon et al. (2014).

2.1.1 Posterior Distribution

Even though we can, within reason, choose a parametric family of models as candidate approximations of our phenomenon of interest, more information is required in order to find the best candidate. Considering the Bayesian perspective, this additional information can come from two sources: sampled data, assumed to be generated by an element of S that we wish to estimate, and a subjective prior belief about the parameter θ, represented in terms of a distribution on Θ. It is worth mentioning that even the process of choosing the set of distributions S is itself informative, and different families may lead to divergent results. After choosing a distribution function that represents our prior information, we can combine it with the information from our obtained sample to update our knowledge about the parameter, resulting in another distribution called posterior. This new distribution allows us to assess how likely each parameter $\theta \in \Theta$ is given all of the available information, and then we can use additional criteria to find our estimate $\hat{\theta}$.

In order to consider a more mathematical representation of this process, suppose that, given a parametric value θ, a random observation Y of the phenomenon of interest is distributed according to $F_Y(\cdot|\theta)$. Let y be an observed instance of this random variable and let F_θ be the prior distribution on Θ assumed for the parametric vector θ. Also, if for simplicity we assume that $F_Y(\cdot|\theta)$ and F_θ have a probability density function (p.d.f.) or a probability mass function (p.m.f.), respectively given by $f_Y(\cdot|\theta)$ and $\pi(\theta)$, then for all observed samples $y \in S_Y$ and parametric values $\theta \in \Theta$ we can use Bayes' theorem to obtain the posterior distribution given by

$$\pi(\theta|y) = \frac{\pi(\theta) f_Y(y|\theta)}{\int_\Theta \pi(\theta) f_Y(y|\theta)\ d\theta} = \frac{\pi(\theta) f_Y(y|\theta)}{\pi(y)}.$$

Here the integral expression in the denominator is a normalizing constant with respect to θ, denominated (prior) predictive distribution. And, with the exception of some well known examples, this integral typically does not have solutions in terms of elementary functions available, requiring the use of computational methods in order to circumvent algebraic limitations, such as the ones described in section 2.2.

After obtaining the posterior distribution, we can use it to derive a posterior pre-

dictive distribution for an unobserved sample Y^*, commonly assuming conditional independence of Y given θ, by taking

$$\pi(Y^*|y) = \int_\Theta \pi(y^*, \theta|y)\ d\theta = \int_\Theta \pi(y^*|\theta, y)\pi(\theta|y)\ d\theta \overset{ind}{=} \int_\Theta f_Y(y^*|\theta)\pi(\theta|y)\ d\theta.$$

Then the posterior predictive distribution can be used to calculate probabilities with respect to the unobserved sample, e.g. finding a set that contains the new value with high probability. Another possibility is using it to assess whether the estimated model is good enough to explain the already observed sample, since the presence of observations with particularly low predictive value may indicate atypicality or even model inadequacy. This concept will be the backbone of the proposed model presented in Chapter 3.

2.1.2 Elements of Decision Theory

Even though we are able to summarise all of the known information with respect to the problem through the posterior distribution and use it to calculate predictive probabilities, we may still desire to find *the best* model or parameter. So, naturally, we first need to establish what *the best* means in this context, specially considering all of the uncertainties involved. Here we consider the decision theory approach to estimation, so next we introduce some elements of decision theory and establish its connection to Bayesian parametric inference. Since we only briefly present this topic, we redirect the reader to Chapter 4 of Migon et al. (2014) for a better introduction and to Berger (1985) for a more complete understanding.

Let us first introduce the main spaces considered in a decision problem. We denote by Ω the set of all possible results of an experiment, representing all of the possible samples $y \in \Omega$ that could be obtained; Θ is the parameter space, representing all of the possible generating mechanisms of our sample y; and the space of possible actions to be taken we denote by $\mathcal{A}$.

With this concepts at hand, our problem consists of choosing a *decision rule* $\delta : \Omega \to \mathcal{A}$, i.e., a function that establishes what action $a \in \mathcal{A}$ to be taken given that our observed sample is $y \in \Omega$. Next, to differentiate "good" and "bad" decision

rules, we introduce the *loss function* $L : \Theta \times \mathcal{A} \to \mathbb{R}^+$, that can be interpreted as the function that numerically attributes a loss to taking an action $a \in \mathcal{A}$ when the observed sample was generated by $F(\cdot|\theta)$.

Then, in the Bayesian context, we define the risk R of a decision rule δ as the expected posterior loss, that is,

$$R(\delta) = \mathbb{E}_{\theta|y}\left[L(\theta, \delta(Y))|Y = y\right] = \int_{\Theta} L(\theta, \delta(y))\pi(\theta|y)\ d\theta. \tag{2.1}$$

We call Bayes rule the decision rule δ^* that minimizes the risk function, so

$$\delta^* = \arg\min_{\delta} R(\delta), \tag{2.2}$$

and we can choose the Bayes rule as our best decision rule. It is worth noting that, since the risk function depends on the choice of the loss function, distinct agents with the same posterior distribution of θ can find different Bayes rules because of a change in the notion of loss from one perspective to another. Now, returning to the problem of estimation, we can use decision theory to find *the best* parameter value by choosing an appropriate action space $\mathcal{A}$ and a loss function.

If we are interested in choosing a single parameter value $\hat{\theta}$, called point estimate of θ, we take $\mathcal{A} = \Theta$. So our action consists of choosing an estimate $\hat{\theta} \in \Theta$ given an observed sample $y \in \Omega$ and we call the resulting Bayes rule an estimator of θ. Commonly used loss functions for point estimation are: the square loss $L_2(\theta, a) = (\theta - a)^2$, whose estimator is given by the posterior expected value $\mathbb{E}[\theta|Y]$; the absolute loss $L_1(\theta, a) = |\theta - a|$, whose estimator is given by the posterior median $med(\theta|Y)$; and the $0-1$ loss $L_\infty(\theta, a) = \lim_{\varepsilon \to 0} \mathbb{I}_{[\varepsilon, +\infty)}(|\theta - a|)$, whose estimator is given by the posterior mode $mode(\theta|Y)$. Here we $\mathbb{I}_A(x)$ is an indicator function whose output is 1, if $x \in A$, and 0 otherwise.

Acknowledging the uncertainties involved in the process of estimation, one could be interested in the inclusion of some quantification of these uncertainties. A reasonable way of achieving this is to choose, for a given observation $y \in \Omega$, a corresponding region $C(y) \subset \Theta$ as an estimate instead of singular value, meaning that $\mathcal{A}$ is the set of all measurable subsets of Θ. Typically, considering the Bayesian perspective, we choose C_γ such that $\mathbb{P}(\theta \in C_\gamma(Y)|Y) \geq \gamma$, called a $100\gamma\%$ credibility region for θ.

Here, $\gamma \in (0,1)$ represents the credibility level of the region C_γ. It is worth noticing that a $100\gamma\%$ credibility region for θ may not be unique, so a common choice is taking the one with the smallest possible measure, also known as the *highest posterior density* (HPD) $100\gamma\%$ credibility region for θ.

At last, if choose $\mathcal{A}$ to be an indicator function $\mathbb{I}_A(\theta)$, then our decision problem becomes a hypothesis testing. In other words, we choose action 1 if $\theta \in A$ and choose action 0 otherwise. Here, 1 and 0 represent any two complementary hypothesis, so to better illustrate this, we respectively denote them H_1 and H_0. To decide between the two hypothesis from the Bayesian perspective, for $i \in \{0,1\}$ and given an observation $y \in \Omega$, we calculate

$$\mathbb{P}(H_1|y) = \mathbb{P}(\theta \in A|y) \text{ and } \mathbb{P}(H_0|y) = \mathbb{P}(\theta \notin A|y) \tag{2.3}$$

then choose H_1, if $\mathbb{P}(H_1|y) > \mathbb{P}(H_0|y)$, and H_0 otherwise. Even though this procedure is conceptually straightforward, one must proceed with caution when considering Θ continuous and either A or A^c is a set of measure zero. In this case, assuming without loss of generality that the measure of A is zero, choosing a continuous prior distribution for θ implies $\mathbb{P}(H_1) = \mathbb{P}(H_1|y) = 0$, thus, we would always choose H_1. So, to avoid this problem, one can simply attribute positive prior probability to H_0 and H_1.

The idea of hypothesis testing is particularly relevant in the context of anomaly detection, where we want to choose whether to consider an observation from a given sample anomalous or not. So naturally, we consider these concepts when defining our proposed model in Chapter 3.

2.2 STOCHASTIC METHODS FOR APPROXIMATE INFERENCE

Recurrently, during the process of inference making, there are analytical limitations to our capability of finding closed form solution to problem of parametric estimation. In the context of Bayesian inference, this problem usually arises from the difficulty of obtaining the posterior's normalizing constant. The most common way of mitigating this difficulty is to abdicate exact solutions and to develop numeric

methods that allow us to control the approximation error.

In this section, we will exclusively focus on stochastic methods for approximate integration, and more specifically on the class of methods of Markov Chain Monte Carlo, broadly used for inferential procedures considering the Bayesian paradigm. Besides this, we will briefly present two of these algorithms used throughout this work: the Metropolis-Hastings Algorithm and the Gibbs Sampler. For a deeper understanding of the subject, we recommend the read of Gamerman & Lopes (2006) for a general comprehension of this class of methods and Liu (2001) for more advanced topics.

2.2.1 Markov Chain Monte Carlo

The methods of Markov Chain Monte Carlo, commonly referred as MCMC methods, are a hybridization of two separate techniques: sample generation through the use of Markov chains and Monte Carlo integration.

Starting with Monte Carlo methods, suppose $h : \mathbb{R}^d \to \mathbb{R}$ is a function we wish to integrate in a Lebesgue-measurable region S such that $\mu_L(S) < +\infty$, where μ_L represents the Lebesgue measure. In other words, our interest is in the quantity

$$I = \int_S h(x)\ dx.$$

If I cannot be obtained analytically, we can rewrite I considering that

$$I = \int_S h(x)\ dx = \mu_L(S) \int_{\mathbb{R}^d} \frac{h(x)}{\mu_L(S)} \mathbb{I}_S(x)\ dx = \mu_L(S)\mathbb{E}\left[h(X)\right],$$

where $X \sim Uniform(S)$ and $\mu_L(S)$ is the Lebesgue measure of S. Then, we generate an independent sample $x^{(1)}, \ldots, x^{(m)}$ uniformly distributed in S to obtain an approximation $\hat{I}$ considering that

$$I = \mu_L(S)\mathbb{E}_U\left[h(U)\right] \approx \frac{\mu_L(S)}{m} \sum_{j=1}^{m} h\left(x^{(j)}\right) = \hat{I}.$$

In the more general case, denominated importance sampling, we can approximate I generating our sample from different distributions. Let f be the density function of

a random variable Y, of which we are able to sample, with support S and such that $\forall x \in S$ we have $f(x) > 0$ if $h(x) \neq 0$. Then, we can similarly rewrite I considering that

$$I = \int_S h(x)\ dx = \int_S \left[\frac{h(x)}{f(x)}\right] f(x)\ dx = \mathbb{E}\left[\frac{h(Y)}{f(Y)}\right].$$

If the expected value of the above expression is finite, we can again take a simple random sample $\left(y^{(1)}, \ldots, y^{(n)}\right)$ with the distribution of Y and approximate our target value as

$$I = \mathbb{E}\left[\frac{h(Y)}{f(Y)}\right] \approx \frac{1}{m}\sum_{j=1}^{m} \frac{h\left(y^{(j)}\right)}{f\left(y^{(j)}\right)}.$$

Here, the finite expected value ensures that the approximation will eventually improve as m goes to infinity due to the Law of Large Numbers. Futhermore, with some additional suppositions regarding the distribution of Y, by the Central Limit Theorem the approximation is of order $O\left(m^{-\frac{1}{2}}\right)$, guaranteeing that the error can be as small as desired for a sufficiently large sample. More details can be found in Chapter 3 of Gamerman & Lopes (2006).

2.2.2 Sampling Using Markov Chains

The sampling methods using Markov chains consist of building an irreducible Markovian stochastic process so that the stationary distribution of the process is the distribution from which you want to sample. Therefore, if after a sufficiently large number of steps we consider the current state of the chain, it will approximately be distributed according to our target distribution. For a better understanding of the subject we recommend the reading of Chapter 4 of Gamerman & Lopes (2006).

Even though these methods are theoretically well established, in the sense that the forementioned result can be formally proven, it is worth emphasizing that some practical issues still need to be considered. The first practical limitation arises when determining what would be a *sufficiently large* number of chain steps. This problem is known as finding the mixing time of the Markov chain and, since no practical general solution can be provided, it is still an active research field. Instead, we can use a group of techniques that search for signs that the chain is yet to

converge and, in the absence of those, proceed assuming that the chain reached convergence. Usually these techniques consist of a combination of procedures, such as hypothesis testing, graphical analysis and comparison of multiple independent chains with distinct initial states. We then define *burn in* as the states generated before convergence and can take any of the remaining states as a sample point with the desired distribution.

Another practical issue is the burden of generating a possibly long and computationally expensive chain to obtain one sample point. So, in order to avoid this inefficiency, we keep generating states after convergence to increase our sample size. Notice that this introduces a new problem, since by construction the Markovian structure of the stochastic process induces a dependency between consecutive chain states and, consequently, some redundancy in our sample. To mitigate this effect, the most common procedure is to form the sample considering only states with some chosen spacing between them, a technique called *thinning*.

Contextualizing the methods in terms of Bayesian inference, the problem is to calculate quantities of interest, such as moments, probabilities and quantiles, with respect to the posterior distribution when the normalizing constant is unknown or too computationally expensive. We then use Markov chains that only require knowledge of the posterior's kernel to generate a sample, and use Monte Carlo techniques to obtain estimates while controlling the approximation error.

2.2.3 Metropolis-Hastings Algorithm

This algorithm was first proposed by Metropolis et al. (1953) and generalized by Hastings (1970) and is a method of specifying the transition kernels of the Markov chain to guarantee convergence to a target stationary distribution $\pi(\theta)$. It basically consists of proposing the transition to a new state x given the current state y according to a chosen distribution $q(x|y)$ and then introducing an acceptance-rejection step to ensure that the target distribution remains the stationary distribution of the chain. For more details, Chib & Greenberg (1995) provide an intuitive derivation of the algorithm, and the topic is also covered in Chapter 6 of Gamerman & Lopes

(2006) and Chapters 5 and 9 of Liu (2001).

Algorithm 1 Metropolis-Hastings Algorithm

Require: Initial value $\theta^{(0)}$ and proposal distribution $q(x|y)$

for $t \in \{1, \ldots, T\}$ **do**

Sample proposed state θ_{prop} from $q(\theta_{prop}|\theta^{(t-1)})$

Calculate the probability of acceptance $\alpha = \min\left\{1, \frac{\pi(\theta_{prop})q(\theta^{(t-1)}|\theta_{prop})}{\pi(\theta^{(t-1)})q(\theta_{prop}|\theta^{(t-1)})}\right\}$

Sample $U \sim Uniform(0, 1)$

if $U \leq \alpha$ **then**

Take $\theta^{(t)} = \theta_{prop}$

end if

if $U > \alpha$ **then**

Take $\theta^{(t)} = \theta^{(t-1)}$

end if

end for

The Metropolis-Hastings, explicitly presented in Algorithm 1, has the advantage of requiring knowledge regarding the desired stationary distribution only up to a constant, making it of particular interest for Bayesian inference. The reason for this propriety can be traced to the acceptance-rejection probability, that only depends on $\pi(\theta)$ through the ratio

$$\frac{\pi(\theta_{prop})q(\theta^{(t-1)}|\theta_{prop})}{\pi(\theta^{(t-1)})q(\theta_{prop}|\theta^{(t-1)})}, \tag{2.4}$$

so any multiplicative constants in $\pi(\theta)$ cancel out.

Even though the Metropolis-Hastings algorithm can generate samples from a wide variety of choices for $\pi(\theta)$, its main drawback is finding a "good" proposal distribution. If the variance of the proposal distribution q is too high, it will frequently generate samples in regions of low posterior density, leading to a high rejection rate. Conversely, if the variance is too low, then the proposed values tend to be highly correlated, effectively increasing the number of chain steps to obtain an approximately independent sample. Having that in mind, the typical procedure is to fine-tune the proposal distribution until reaching satisfactory results.

Besides the formulation in Algorithm 1, the Metropolis-Hastings algorithm can

be expressed in a more general form to sequentially update individual entries of a parameter vector $\theta = (\theta_1, \ldots, \theta_p)'$ instead of proposing a single θ_{prop} at each step of the chain. This formulation tends to be more common for application in the multivariate case because it usually is a simpler problem to choose a proposal distributions for each entry rather than for the entire vector.

2.2.4 Gibbs Sampler

The Gibbs sampler, originally proposed by Geman & Geman (1984), provides a way of specifying the transition kernels of the Markov chain with a target multivariate stationary distribution $\pi(\theta)$ and is based on the complete conditional distributions of each entry. To clarify, if $\theta = (\theta_1, \ldots, \theta_p)'$, then the conditional distribution of θ_j is given by $\pi(\theta_j|\theta_{-j})$, where $\theta_{-j} = (\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_p)'$. The algorithm then starts with an initial condition $\theta^{(0)} = \left(\theta_1^{(0)} \ldots, \theta_p^{(0)}\right)'$ and then sequentially sampling from the conditional distributions of each entry given the most recently sampled value of the others, as described in Algorithm 2. Interestingly, the Gibbs sampler can be seen as a particular case of the Metropolis-Hastings algorithm, where we use as the proposal distribution for each entry its full conditional, leading to a probability of acceptance of 1.

Algorithm 2 Gibbs Sampler

Require: Initial value $\theta^{(0)} = \left(\theta_1^{(0)} \ldots, \theta_p^{(0)}\right)'$

  for $t \in \{1, \ldots, T\}$ **do**

    for $j \in \{1, \ldots, p\}$ **do**

      Sample $\theta_j^{(t)}$ from $\pi\left(\theta_j \left| \theta_{-j}^{(t)} \right.\right)$, where $\theta_{-j}^{(t)} = \left(\theta_1^{(t)}, \ldots, \theta_{j-1}^{(t)}, \theta_{j+1}^{(t-1)}, \ldots, \theta_p^{(t-1)}\right)'$

    end for

  end for

Some well known variations of the Gibbs sampler can be used to improve the algorithm's convergence properties, such as changing the update order of the entries considering some permutation of the index, blocking highly correlated entries to sample from their joint full conditional distribution and considering the full conditionals of the target distribution marginalized with respect to some of the parameters. For

more details regarding those techniques, we redirect the reader to the Chapter 5 of Gamerman & Lopes (2006) and Chapter 6 of Liu (2001).

It is worth highlighting that, in contrast with the Metropolis-Hastings algorithm, the Gibbs sampler lacks an acceptance-rejection step, so all of the "proposed" states of the chain are accepted with probability 1. Another consequence of this aspect, is that the algorithm does not require the calibration of hyperparameters, unlike some common proposals used when considering the Metropolis-Hastings algorithm. The last noteworthy characteristic of the Gibbs sampler is that sampling from the full conditionals can be performed without knowing the normalizing constant of $\pi(\theta)$, making it particularly attractive considering the Bayesian perspective.

2.3 ANOMALY DETECTION USING DISCRETE MIXTURE MODELS

The earliest instance of the utilization of discrete mixture models in the context of anomaly detection can be traced back to a paper by Newcomb (1886), where the author proposed using a mixture of normal distributions with different variances to model the possible variation in the measurement precision for each data point. Even though he was the first one to propose this treatment to outliers, due to computational limitations, at the time Newcomb was only able to proceed with calculations and estimations based on chosen parameter values instead of using the common inferential procedures known today. It would be only decades later, with Box & Tiao (1968), Guttman (1973), Abraham & Box (1978) and Guttman et al. (1978), that the first methods for estimating the parameters of a mixture model would be proposed, based on the analytical expression of the posterior distribution when assuming conjugate priors. Even so, since the resulting analytical estimators require the computation of 2^n terms, where n represents the sample size, these methods were viable only considering either small samples ($n \leq 20$) or utilizing approximations of the posterior distribution, for instance with the assumption that there were at most a small number of outliers contained in the sample. Considering the classical point of view, no practical methods for obtaining reasonable estimators were available, leading to the preference of other statistical methodologies to identify

anomalies. For a general review of the statistical methods for outlier detection in this period we refer to Barnett & Lewis (1978).

It was only with the advent of efficient computational methods such as the EM algorithm and the Gibbs sampler, by Dempster et al. (1977) and Geman & Geman (1984) respectively, that mixture models became more practical tools for anomaly identification. Since then, multiple algorithms for fitting mixture models have been studied, considering both classical (Aitkin & Wilson (1980); Coretto & Hennig (2011); Yu et al. (2015)) and Bayesian (Verdinelli & Wasserman (1991); Evans et al. (1992)) perspectives and in many different contexts, including clusterization (Banfield & Raftery (1993); Coretto & Hennig (2016); Yin & Wang (2016)), regression (Box & Tiao (1968), Abraham & Box (1978)), functional data (Amovin-Assagba et al. (2022)), sequential data (Brunot (2020)), and convolutional neural networks (Lathuilière et al. (2018)). For a general introduction to mixture models we recommend Frühwirth-Schnatter & Frèuhwirth-Schnatter (2006), which also presents them in the context of outlier detection in Chapter 7. Next, we present a brief overview of the components used for anomaly detection in the literature and we introduce an improper component, proposed independently by Longford & D'Urso (2011) and Coretto & Hennig (2016) in the context of classical estimation, and adapted to the Bayesian perspective by Barreto (2022), that is the basis of the specification of our proposed model in Chapter 3. We could also reference the use of continuous mixtures for anomaly detection throughout the literature, but this lies outside the scope of this work.

2.3.1 Mixture Components for Anomaly Detection

The first discrete mixture model with a component dedicated to capture outliers was the variance inflation model proposed by Tukey (1960) with density

$$f(x|\mu, \sigma^2, k, w) = w\phi(x|\mu, \sigma^2) + (1 - w)\phi(x|\mu, k^2\sigma^2), \tag{2.5}$$

where $\mu \in \mathbb{R}$, $\sigma^2 > 0$, $k > 1$, $w \in (0, 1)$, $\phi(\cdot|\mu, \sigma^2)$ represents the density of an univariate normal random variables with mean μ and variance σ^2, and $(1 - w)$

represents the probability of obtaining an imprecise data measurement. Even though proposed by Tukey (1960), This model was first adjusted by Box & Tiao (1968) to detect anomalous observations in a data set containing differences of heights of fifteen plants. A few years later, Guttman (1973) proposed a similar approach with the location shift model given by

$$f(x|\mu, \theta, \sigma^2, w) = w\phi(x|\mu, \sigma^2) + (1 - w)\phi(x|\mu - \theta, \sigma^2). \tag{2.6}$$

When comparing with the variance inflation model, the parameter k, responsible for increasing the variance of the second component, was removed and in its place $\theta \in \mathbb{R}$ was introduced to allow a translation of the second component. It is easy to see that both the variance inflation and location shift models are particular cases of the normal mixture model with two components, whose full estimation was first discussed by Evans et al. (1992).

Next, the first model to consider a discrete mixture of non-normal distributions for outlier detection was proposed by Banfield & Raftery (1993), considering a problem of clusterization for multivariate data, with a density of the form

$$f(x|\mu, \Sigma, w) = w \sum_{j=1}^{m} \eta_j \phi_d(x|\mu_j, \Sigma_j) + (1 - w)\pi(x), \tag{2.7}$$

where $\mu_j \in \mathbb{R}$, Σ_j is a covariance $d \times d$ matrix, $w \in (0, 1)$, $\eta = (\eta_1, \ldots, \eta_m)$ is such that $\eta_j \geq 0$, $\sum_{j=1}^{m} \eta_j = 1$, $\phi_d(\cdot|\mu, \Sigma)$ represents the density of an d-variate normal random variable with mean vector μ and covariance matrix Σ and $\pi(x)$ represent an uniform distribution on a limited known support of the observations to capture anomalies. Then, following works introduced discrete mixtures considering components with Student-t distribution (Stephens (1997); Peel & McLachlan (2000)), with beta distribution (Bouguessa (2014)) and even mixture of distributions (Amovin-Assagba et al. (2022)), where each component is given by a variance inflation model.

More recent works introduced a new type of component for capturing atypical observations. These were the papers of Longford & D'Urso (2011) and Coretto & Hennig (2016) that independently proposed the use of an improper component in the mixture. This alternative component mainly consists of a heavy tailed function that

achieves relatively low values in the dense regions of the model for typical observations, allowing it to better accommodate outliers while minimizing competition with the main model. The simplest specification presented for this improper component was a constant function over the entire support, representing an uniform improper distribution over the real line. This same improper component was then used by other authors such as Lathuilière et al. (2018) and Inverardi & Taufer (2020). An expression for this model is given by

$$f(x|\theta, \delta, w) = wf(x|\theta) + (1 - w)\delta, \tag{2.8}$$

where, θ represents an unknown parameter, $w \in (0, 1)$, $f(\cdot|\theta)$ is the main model's density function and δ is a positive unknown constant.

Even though the model defined by this mixture with an improper component does not result in a probabilistic model, when considering the maximum likelihood estimator for the resulting *pseudo-likelihood*, usual maximization methods such as the EM algorithm can be adapted to obtain estimates for the parameters. Nevertheless, when looking at this model from the Bayesian perspective, the choice of an improper component leads to an improper predictive distribution unless considering $w = 0$. To overcome this issue, a numerically equivalent approach was proposed by Barreto (2022) considering a reinterpretation of the alternative component and, since it is the basis of our proposed method, we present it in Chapter 3.

2.4 ANOMALY DETECTION USING DEPTH FUNCTION

For any given data set of univariate observations, the maximum and minimum are usually promising candidates when one wishes to search for anomalies. That is because, if we consider the common case of an unimodal distribution, we expect to identify outliers by looking at observations at regions of least density and, in this case, those happen to be close to the minimum or the maximum of the sample. So naturally, multiple outlier detection methods rely on statistics of order and quantiles to identify observations that trespasses a certain established threshold. Despite the effectiveness of this idea, limitations arise when we consider multivariate data. In

two or more dimensions, the lack of a natural ordering of points leads to the absence of an universally agreed upon notion of quantiles, preventing the use of concepts reliant on ordering, such as the sample minima or maxima. Thus, depth function were introduced in order to circumvent this issue, providing ways to introduce an ordering in $\mathbb{R}^d$ based on a probability distribution F.

The first to formally define the notion of a depth function were Zuo & Serfling (2000). Before them, a depth function used to be any function $D(x, F) : \mathbb{R}^d \times \mathbb{F} \to \mathbb{R}$ that induced a F-based center-outward ordering for all $x \in \mathbb{R}^d$, where $F \in \mathbb{F}$ is a probability distribution and $\mathbb{F}$ represents the class of distributions on the Borel sets of $\mathbb{R}^d$. So, considering that this notion relied on somewhat vague descriptive properties, Definition 2.1 aims to formalize it in order to provide a "systematic basis for preferring one function over another", while also generalizing notions such as quantiles and centrality of a distribution. Thus, it was imposed that, if a distribution F has some symmetry around a naturally defined center x_c (the distribution does not have to be symmetric), then x_c is the generalized median of the distribution and it must coincide with the deepest point, i.e., the point x^* that maximizes $D(x, F)$. This definition is presented as follows.

Definition 2.1 (Depth function, Zuo & Serfling (2000))**.** Let the mapping $D : \mathbb{R}^d \times \mathbb{F} \to \mathbb{R}$ be bounded and non-negative. Then D is called a *depth function* if it satisfies the following proprieties:

1. (P1) $D(Ax + b, F_{AX+b}) = D(x, F_X)$ holds for any d-dimensional real random vector X, any $d \times d$ nonsingular matrix A, and any vector $b \in \mathbb{R}^d$,
2. (P2) $D(x_c, F) = \sup_{x \in \mathbb{R}^d} D(x, F)$ holds for any $F \in \mathbb{F}$ having center x_c,
3. (P3) for any $F \in \mathbb{F}$ having deepest point x^*, $D(x, F) \leq D(x^* + \alpha(x - x^*), F)$ holds for all $\alpha \in [0, 1]$, and
4. (P4) $\lim_{||x|| \to +\infty} D(x, F) = 0$, for each $F \in \mathbb{F}$.

Here (P1) makes the depth function invariant to affine transformations, removing the effect of changes in location or scale on the induced ordering of data; (P2)

guarantees that the deepest point lies in the center of the distribution x_c, thus generating a center-outward ordering; (P3) forces the depth function to decrease monotonically when taking a linear trajectory that moves away from its center; and (P4) insures that the depth is arbitrarily small for points sufficiently distant from the center.

It is worth mentioning that, when considering applications, usually the distribution F of the observation is unknown, so estimates of the depth values for each observation can be obtained using an empirical distribution $\hat{F}_n$. Considering the context of outlier detection, after calculating the depth values for each observation, procedures typically consider every observation whose depth is below a certain threshold γ or the k observations with the smallest estimated depth value as outliers.

Since the first uses of depth functions in the literature, this method have been used for anomaly detection in many different contexts, such as multivariate statistical quality control (Liu (1995); Cheng et al. (2000); Hamurkaroğlu et al. (2004)), spatial data (Chen et al. (2008)) and functional data (Febrero et al. (2008); Arribas-Gil & Romo (2014); Sguera et al. (2016); Kuhnt & Rehage (2016); Dai et al. (2020)).

The following subsections of this chapter will focus on providing a few examples of recurrently used depth functions in the literature and later will present the *likelihood depth*, a *pseudo-depth* function that will be relevant for constructing the model proposed in Chapter 3. So, for a more in *depth* review of the field, we redirect the reader to Mosler (2013).

2.4.1 Commonly Used Depth Functions

One of the earliest instances of a depth function was the *halfspace depth*, and it was first introduced by Tukey (1975). To better understand this depth function, let us first take all possible divisions of $\mathbb{R}^d$ in two regions by a $(d-1)$-dimensional hyperplane P. Then, considering a point $x \in \mathbb{R}^d$ and a probability distribution F, we obtain the depth function by taking each division, looking at the region H that contains x (if x lies in P we consider both regions) and determining the smallest

possible value of $\mathbb{P}(H)$. In other words, the *halfspace depth* is the minimal probability $\mathbb{P}(H)$, where H is a closed halfspace containing the observation $x \in \mathbb{R}^d$. The idea here is that, if x is a more central observation, we have a higher value of $\mathbb{P}(H)$ regardless of how we choose a halfspace H containing x, which is not the case for more distal observations. This depth can more formally be expressed as

$$HD(x, F) = \inf\{\mathbb{P}(H) : H \text{ is a closed halfspace containing } x\}, \quad \forall x \in \mathbb{R}^d. \tag{2.9}$$

Another important influential depth was the *simplicial depth*. It was proposed by Liu (1990) and defined to be, fixing a point $x \in \mathbb{R}^d$ and a probability distribution F, the probability of a random simplex in $\mathbb{R}^d$ containing x. In other words,

$$SD(x, F) = \mathbb{P}(x \in S[X_1, \ldots, X_{d+1}]), \quad \forall x \in \mathbb{R}^d, \tag{2.10}$$

where $X_1, \ldots, X_{d+1} \overset{iid}{\sim} F$ and $S[x_1, \ldots, x_{d+1}]$ represents the d-dimensional simplex with vertices $x_1, \ldots, x_{d+1}$, i.e., the smallest d-dimensional convex polytope containing the points $x_1, \ldots, x_{d+1}$. The last depth function we will present on this section is based on the distance defined by Mahalanobis (1936) and was proposed by Liu & Singh (1993), called the *Mahalanobis depth*. It is defined as

$$MD(x, F) = \left[1 + d_M^2(x, \mu|\Sigma)\right]^{-1} = \left[1 + (x - \mu)'\Sigma^{-1}(x - \mu)\right]^{-1}, \quad \forall x \in \mathbb{R}^d, \tag{2.11}$$

where $\mu \in \mathbb{R}^d$ is the expected value of a random variable having distribution F and Σ is its covariance matrix. This depth function belongs to a wider class of distance-based depth functions and other instances of this class can be obtained by swapping the Mahalanobis distance $d_M(\cdot, \mu|\Sigma)$ in the equation above by any other F-based distance function.

2.4.2 The Likelihood Pseudo-Depth

Even though Zuo & Serfling (2000) formally introduced a definition for depth functions, other useful depth-like function can be used to induce an ordering even without satisfying all of the properties mentioned. So we call pseudo-depth any depth-like function that violates one of the proprieties of Definition 2.1, but still

provides an F-based ordering of multivariate observations. One example that is relevant for the construction of the proposed model in Chapter 3 is the *likelihood pseudo-depth*, proposed by Fraiman et al. (1999). Fixing a distribution F with probability density function or probability mass function f, the likelihood depth is given by

$$LD(x, F) = f(x), \quad \forall x \in \mathbb{R}^d, \tag{2.12}$$

and in their work Fraiman et al. (1999) proposed using a kernel estimate of the density function instead of assuming a specific family of distributions for the data.

Interestingly, the likelihood depth may satisfy Definition 2.1 depending on whether or not we impose that F belongs to certain families of distributions. For instance, if we choose F to be the distribution function of a multivariate normal distribution, the resulting ordering is the same as the one generated by the Mahalanobis depth for some choice of μ and Σ, leading to a valid depth function. Conversely, it is not particularly hard to find a distribution with a density that violates the definition. As one example, if we choose f to be the density of a Gamma distribution with shape parameter $\alpha < 1$, then f is not bounded. Another interesting case is considering a mixture of 3 bivariate normal distribution with identity covariance matrices and mean vectors lying on the vertices of an equilateral triangle with center at the origin and side length greater than 5. In this case, the density function f violates $(P2)$, because by rotational symmetry the center of f should be located at the origin, however it differs from the deepest points, that are located near the center of each normal component. Besides this, f also violates $(P3)$, because tracing a trajectory from one of the deepest points to another would generate a linear path that moves away from the deepest point and does not monotonically decrease.

Even though the likelihood pseudo-depth may not satisfy Definition 2.1, it can be argued that, when considering the context of outlier detection, the most sensible ordering of data should consider the density information, instead of a notion of centrality. That is because, assuming we are capable of providing good estimates of the density function, the observations that are less likely to be generated from the distribution F are necessarily those in the regions of lower estimated density.

This makes observations with a particularly low likelihood pseudo-depth reasonable candidates to be considered atypical and this pseudo-depth an attractive ordering function for this purpose.

3 FILTERING MODEL

In this chapter, we present in section 3.1 the construction of the model proposed in this thesis in three stages. Then, section 3.2 introduces the MCMC algorithm utilized for parameter estimation, addressing known issues and the corresponding techniques employed to mitigate them. Lastly, section 3.3 discusses interpretation of the results, model, prior and hyperparameter specification, prediction for future observations and anomaly classification.

3.1 MODEL CONSTRUCTION

In this work, we consider the mixture model approach for the estimation of the probability of an observation being atypical. Even though this is fundamentally arbitrary, we justify our choice by considering a reasoning similar to De Finetti's (1961) when he states that "According to the Bayesian point of view, there exist no observation to be rejected", while discussing how outlier rejection should work within the Bayesian perspective. Loosely following his argument and shifting our focus to the particular context of mixture models, if we consider an observed sample $y_1, \ldots, y_n$, the problem of detecting anomalies is equivalent to estimating an indicator variable z_i such that $z_i = 1$ if the corresponding observation y_i is typical and $z_i = 0$ otherwise. Here, it is worth noting that until now, the notion of *atypicality* is somewhat vague and arbitrary, so for a more precise meaning we consider the discrete mixture

$$\begin{aligned} Y_i|\theta, z_i = 1 &\sim F_{Y_i}(\cdot|\theta), \\ Y_i|\theta^*, z_i = 0 &\sim F_{Y_i^*}(\cdot|\theta^*), \end{aligned} \tag{3.1}$$

for all $i \in \{1, \ldots, n\}$, where $F_{Y_i}(\cdot|\theta)$, $F_{Y_i^*}(\cdot|\theta^*)$ are distribution functions with the same support and θ is our parameter of interest. Now we can specifically call y_i *typical* if it was generated by $F_{Y_i}(\cdot|\theta)$ and *atypical* if it was generated by $F_{Y_i^*}(\cdot|\theta^*)$. We can then consider estimating θ, θ^* and z, the collection of all indicators, by attributing a prior distribution $\pi(\theta, \theta^*, z)$ and obtaining the posterior $\pi(\theta, \theta^*, z|y)$ via Bayes' Theorem. The core idea here is that, if we choose our prior such that $\pi(z_i =$

$c) \neq 1$ for $c \in \{0, 1\}$, i.e., if we do not already know what distribution generated y_i, then $\mathbb{P}(z_i = c|y) \neq 1$ for $c \in \{0, 1\}$. So, since we can never be certain of the value of z_i and knowing that the posterior summarizes all of the available information with respect to θ and z, inference for θ should consider all observations $y_1, \ldots, y_n$, albeit attributing different weights to each observation according to $\mathbb{P}(z_i = 1|y)$.

Even though we here assume a mixture model for the observations, in general this reasoning still holds, as presented by De Finetti (1961), which leads us to the conclusion that no observation should be removed from our sample when estimating θ, and analogously θ^*. However, if our objective demands a classification, we still need to decide whether to consider y_i typical or not. So, since mixture models allows us to estimate θ respecting this principle while also generating classifications in the more generally applicable context of unsupervised learning, we consider this approach for the proposed model of this work.

In the rest of this section, we iteratively construct our proposed model, consisting of a mixture between a parametric model of interest and a model-induced noise component, with a problem-oriented philosophy. Subsection 3.1.1 introduces the *naive* filtering model, that represents the core idea of our proposed methodology, to remove the assumption of identically distributed observations from the model proposed by Longford & D'Urso (2011) and by Coretto & Hennig (2016). Next, subsection 3.1.2 discusses shortcomings of the previous approach, related to an excess removal of observations, and presents the *biased* filtering model as a solution. And finally, subsection 3.1.3 further improves the method presenting the *filtering model* as a way of controlling the bias generated by the previous correction.

3.1.1 Naive Filtering Model

We begin by returning to the motivating problem established in the introduction. Figure 2 presents a visualization of the Octane data set, see Esbensen et al. (2002), colored by their classification as contaminated or not. As we can see, the typical observations (colored in blue) are well behaved, i.e., they are observations for which we could find a reasonable parametric family of distributions that somewhat

accurately represents them, while the atypical ones (colored in red) do not seem to follow an easily describable pattern. Since we are considering a mixture model approach to anomaly detection, we must specify a model for typical and atypical observations, which may be challenging.

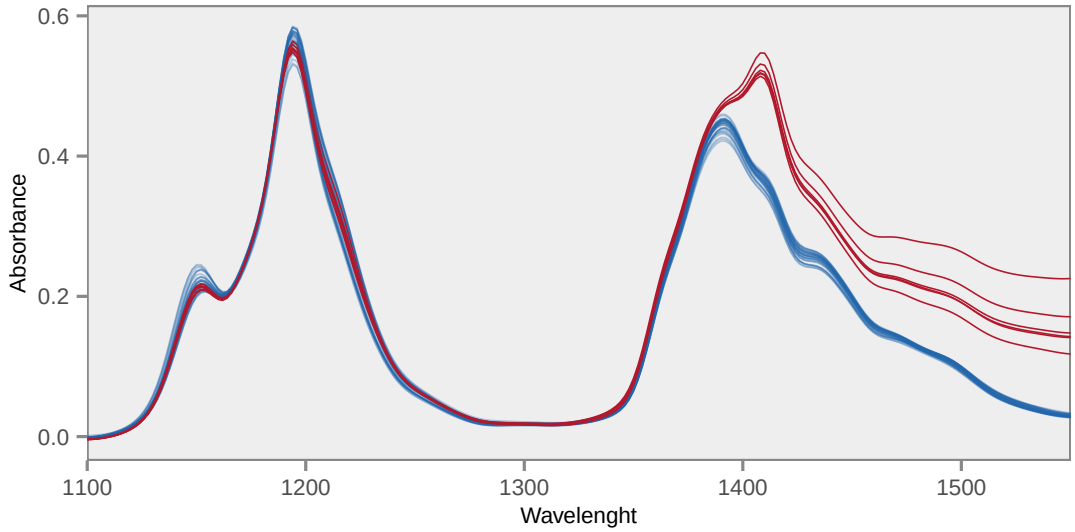

Figure 2: Absorbance curves for gasoline samples from the octane data set at different wavelengths, where in red we highlight the contaminated observations.

Another aspect which is worth considering is that, is this case, from the nature of the data we know that the atypical observations, the ones contaminated with ethanol, present a similar behavior. However, if we are interested in a model capable of identifying any form of contamination in a gasoline sample, then our anomaly detection component must be flexible enough to capture any deviation of the estimated typical behavior of the data, while having a controlled probability of capturing normal observations. So, apparently for this exact reason, the literature gravitated towards heavy-tailed distributions for the anomaly capturing component. As presented in 2.3, the most extreme case of this trend is the mixture model with an improper component, proposed independently by Longford & D'Urso (2011) and Coretto & Hennig (2016).

It is important noticing that, considering the Bayesian perspective, choosing an improper component leads to an improper predictive distribution, resulting in an unnecessary layer of complexity to the estimation process. To specifically address

this issue, Barreto (2022) proposed a reinterpretation of the improper component that results in the following numerically equivalent model

$$\begin{aligned} Y_i|\theta, z_i = 1 &\overset{ind}{\sim} F_{Y_i}(\cdot|\theta), \\ Y_i|\theta, z_i = 0 &\overset{ind}{\sim} Uniform(S), \\ \mu_L(S)^{-1} &= h, \end{aligned} \tag{3.2}$$

where μ_L is the Lebesgue measure and the improper uniform distribution on $\mathbb{R}$ is substituted by a proper uniform on an unknown region S. Here, denoting

$$\mathbb{I}_A(x) = \begin{cases} 1, & \text{if } x \in A \\ 0, & \text{otherwise} \end{cases}, \tag{3.3}$$

we assumed that $\mathbb{I}_S(y_i) = 1$, for all $i \in \{1, \ldots, n\}$ to ensure that the probability of each observation being atypical is positive, and the term h is chosen and estimated considering some heuristics, that are irrelevant for our purposes, to avoid a reduction to the trivial degenerate case, where $\mu_L(S) = 0$ and all observations are atypical with probability 1. Since S is assumed to contain all of the observations regardless of its measure, this component essentially behaves as a constant when only considering the points $y_1, \ldots, y_n$, thus resulting in the same numerical proprieties of the improper component.

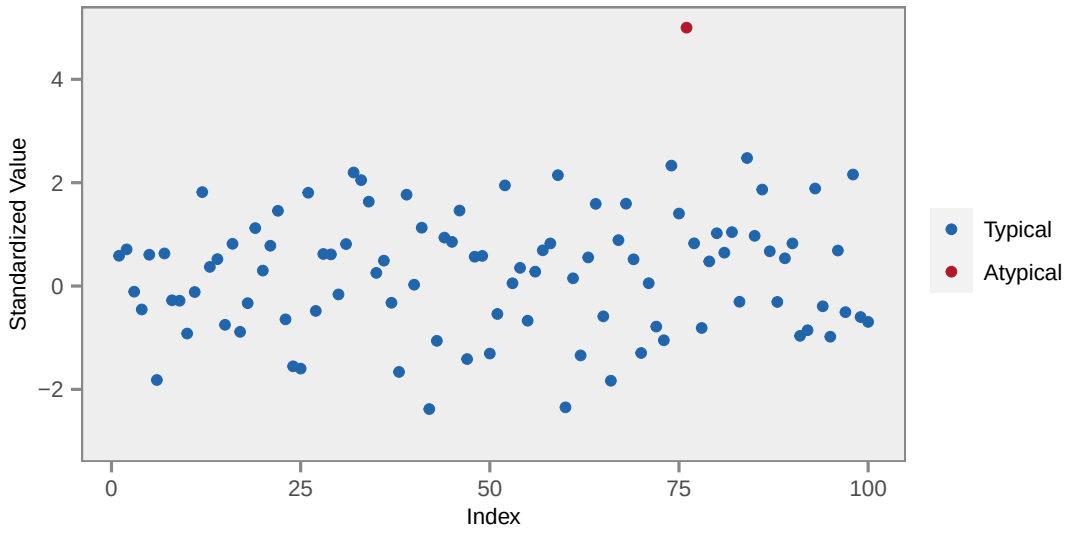

Figure 3: Visualization of the simulated data set 1 with the first 50 observations divided by 100000, where the colors indicate the true classification.

Even though the proposal of Barreto (2022) allows us to consider estimation from the Bayesian point of view, it still incorporates one limitation inherited from the mixture model with an improper component: the component is the same for all observations. In order to emphasize this limitation, we generated a sample such that the first 50 observations come from a $Normal\,(0, 100000^2)$, while the last 50 come from a $Normal(0, 1)$, with the exception of the observation of number 76, that was arbitrarily chosen to be equal to 5 to simulate an anomaly. For future reference, we refer to this sample as the *simulated data set 1*. Figure 3 shows the simulated sample, but we divided the first 50 observations by 100000.

Knowing the true underlying distribution of our data, we assume as the typical component of our model a $Normal\,(0, 100000^2)$ for observations of index 1 to 50 and a $Normal(0, 1)$ for the remaining observations, including observation 76. Now, taking $\pi(z_i) = \frac{1}{2}$ for all observations, we can directly calculate

$$\mathbb{P}(z_i = 1|y) = \frac{f_{Y_i}(y_i|\theta)}{f_{Y_i}(y_i|\theta) + h}, \tag{3.4}$$

so we can consider the simple classification procedure of taking $z_i = 1$, if $\mathbb{P}(z_i = 1|y) \geq \mathbb{P}(z_i = 0|y)$, and $z_i = 0$, otherwise.

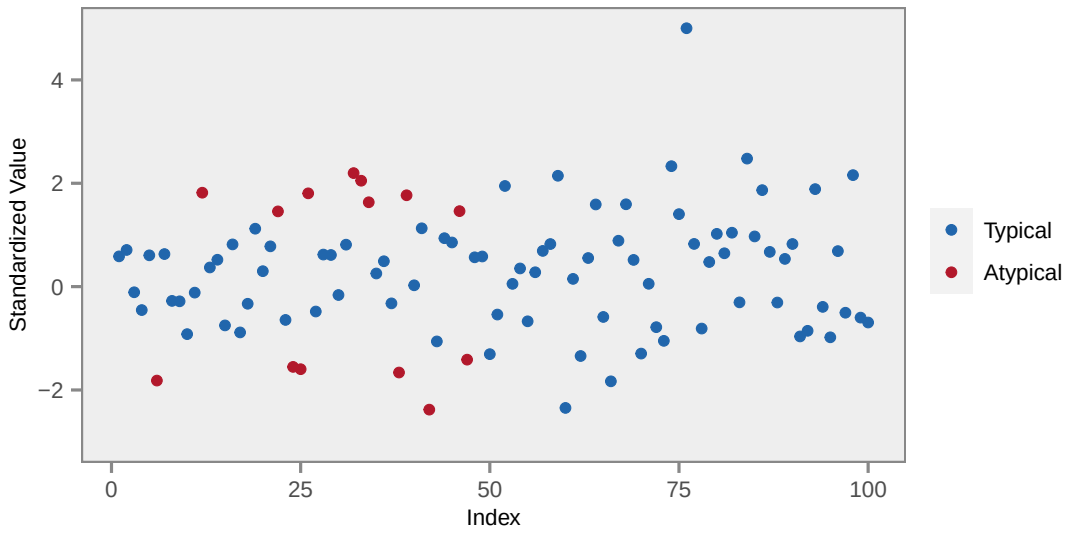

Figure 4: Visualization of the simulated data set 1 with the first 50 observations divided by 100000, where the colors indicate the estimated classification considering $h = 1.4867e - 06$.

Then, with this estimation rule, Figure 4 shows the same visualization of our

sample with the resulting classification when assuming $h = 1.4867e-06$. Examining Figure 4, it becomes clear that if we consider higher values of h we would still remove observations from the first part of the sample, but taking lower values would not help removing the anomaly. From this, we conclude that we cannot define an unique threshold that simultaneously identifies our anomaly while not misclassifying typical observations. This, in spite of being a rather artificial example, showcase the limitations of this assumption, motivating us to propose a modified version of the model to circumvent this issue.

Our solution to this problem involves changing the uniform distribution over S to allow for different levels for each sample point. We achieve this by defining unknown sets $S_1, \ldots, S_n$, one for each observation, with different density levels for each data point. The resulting model then becomes

$$\begin{aligned} Y_i|\theta, z_i = 1 &\overset{ind}{\sim} F_{Y_i}(\cdot|\theta), \\ Y_i|\theta, z_i = 0 &\overset{ind}{\sim} Uniform(S_i), \\ (\theta, z) &\sim \pi(\theta, z), \\ \mu_L(S_i)^{-1} &= h_i(\theta), \end{aligned} \tag{3.5}$$

where $\pi(\theta, z)$ represents the prior distribution of θ and the collection of indicators z. However, it is worth noticing that, if we choose $h_i(\theta) = c_i \in \mathbb{R}$, for all $i \in \{1, \ldots, n\}$, then we cannot estimate the n independent quantities $c_1, \ldots, c_n$, rendering the model useless. Instead, we consider, as the notation already suggests, choosing $h_i(\theta)$ as a function of the parameter θ from the typical component. The idea here is to consider a parametric model that accurately depictures the typical behavior of the data and then chooses $h_i(\theta)$ to specifically capture observations that seem to be too unlikely under the typical model. For this reason, we also call *main component* the component of the mixture responsible for capturing the typical observations and *alternative component* the one used for the anomalies.

In order to guide our choice of $h_i(\theta)$, we will consider the idealized scenario of knowing the value of θ and z_{-i}, that denotes the collection of all indicator variables with the exception of z_i. Even though it is unrealistic, since in section 3.2 we estimate θ and z using MCMC methods, this scenario does occur when sampling z_i

from its full conditional considering a Gibbs step. Having this in mind, we next find the full conditional distribution of z_i, but first, we need to obtain the expression for the posterior distribution. Now, additionally assuming that $F_{Y_i}(\cdot|\theta)$ is an absolute continuous or discrete distribution with density or probability mass function $f_{Y_i}(\cdot|\theta)$ for simplicity, our posterior is given by

$$\begin{aligned}\pi(\theta, z|y) &\overset{Bayes}{=} \frac{\pi(\theta, z)\pi(y|\theta, z)}{\sum_{z\in\{0,1\}^n}\int_\Theta \pi(\theta, z)\pi(y|\theta, z)\ d\theta} \propto \pi(\theta, z)\pi(y|\theta, z) \\ &\overset{ind}{=} \pi(\theta, z)\prod_{i=1}^n\left[f_{Y_i}(y_i|\theta)\right]^{z_i}\prod_{i=1}^n\left[\mu_L(S_i)^{-1}\ \underbrace{\mathbb{I}_{S_i}(y_i)}_{=1}\right]^{1-z_i} \\ &= \pi(\theta, z)\prod_{i=1}^n\left[f_{Y_i}(y_i|\theta)\right]^{z_i}\prod_{i=1}^n\left[h_i(\theta)\right]^{1-z_i}.\end{aligned} \tag{3.6}$$

So, we can obtain the full conditional of z_i as

$$\begin{aligned}&\pi(z_i|\theta, z_{-i}, y) \propto \pi(\theta, z|y) \\ &= \pi(z_i|\theta, z_{-i})\pi(\theta, z_{-i})\left[f_{Y_i}(y_i|\theta)\right]^{z_i}\left[h_i(\theta)\right]^{1-z_i} \\ &\propto w_i^{z_i}(1-w_i)^{1-z_i}\left[f_{Y_i}(y_i|\theta)\right]^{z_i}\left[h_i(\theta)\right]^{1-z_i} \\ &\propto \left(\frac{w_i f_{Y_i}(y_i|\theta)}{w_i f_{Y_i}(y_i|\theta) + (1-w_i)h_i(\theta)}\right)^{z_i}\left(\frac{(1-w_i)h_i(\theta)}{w_i f_{Y_i}(y_i|\theta) + (1-w_i)h_i(\theta)}\right)^{1-z_i}\end{aligned} \tag{3.7}$$

where z_{-i} represents all of the indicators with the exception of z_i and $w_i = \pi(z_i = 1|\theta, z_{-i})$. With this, we can identify that

$$z_i|\theta, z_{-i}, y \sim Bernoulli\left(\frac{w_i f_{Y_i}(y_i|\theta)}{w_i f_{Y_i}(y_i|\theta) + (1-w_i)h_i(\theta)}\right). \tag{3.8}$$

Now notice that, considering a Gibbs step for z_i to sample from the posterior via MCMC, at every step of the chain we are making a random classification of our observation y_i assuming that θ and z_{-i} are known. We then could consider establishing an analogy with hypothesis testing, where our null hypothesis would be $H_0^i : z_i = 1$, or equivalently, $H_0^i : Y_i \sim F_{Y_i}(\cdot|\theta)$, and our alternative hypothesis would be $H_1^i : z_i = 0$, or equivalently, $H_1^i : Y_i$ was not generated by $F_{Y_i}(\cdot|\theta)$, to find a natural way of defining $h_i(\theta)$. However, if we were to reasonably estimate the probability of our alternative hypothesis considering the Bayesian approach, we would have to instantly take $\mathbb{P}(H_1^i|y) = 1$, for all $i \in \{1, \ldots, n\}$, since, with the exception of simulated data, our chosen model for typical $F_{Y_i}(\cdot|\theta)$ is almost surely wrong, forcing us to classify all observations as atypical.

As an alternative, we can notice that, considering the derived full conditional and assuming $w_i = \frac{1}{2}$, for all $i \in \{1, \ldots, n\}$, we would not reject our null hypothesis if $\mathbb{P}(H_0^i|y) \geq \mathbb{P}(H_1^i|y)$, which would be the same as not rejecting H_0^i if $f_{Y_i}(y_i|\theta) \geq h_i(\theta)$. Note however, knowing that $f_{Y_i}(\cdot|\theta)$ can be interpreted as the predictive distribution of Y_i given θ, that we would reject the null hypothesis if the predictive value $f_{Y_i}(y_i|\theta)$ assuming our model is *too small*. So naturally, we are interested in studying the distribution of the random variable given by the transformation $f_{Y_i}(Y_i|\theta)$, i.e., the distribution of the predictive distribution of Y_i, where $Y_i \sim F_{Y_i}(\cdot|\theta)$, to understand which predictive values could be considered *too small*. Next, in order to simplify the notation used, we introduce the definition of an autotransformation to better represent this specific transformation.

Definition 3.1 (Autotransformation)**.** Let X be a d-dimensional random vector with density or probability mass function f_X. Then the *autotransformation* of X is defined as $T_X = f_X(X)$.

For practical purposes, finding the distribution of T_X in terms of elementary functions for a single random variable X is typically an irremediable task, specially considering that it might be the distribution of a non-trivial irreversible transformation of X. Having that in mind, the more general case of seeking the autotransformation for a parametric family of random variables seems to be a fruitless endeavor. Nevertheless, for a sufficiently simple, general and useful class of families of distributions we can use straightforward properties to find the autotransformation of any member in terms of only one of them. This is the class of location and scale families of distributions and the mentioned result can be found in section A.3 of Appendix A. As a side note, from Definition 3.1 we highlight that $T_X > 0$ almost surely, since f_X is always positive in a set that X belongs to with probability 1.

Going back to our objective, we can then define $h_i(\theta)$ as the quantile $1 - \gamma$ of T_{Y_i}, where γ fills a role similar to the credibility level in the context of hypothesis testing. This choice allows us to control, how rare an observation y_i needs to be with respect to our assumed model for us to reject H_0^i. Even though we are technically only sampling from the full conditional at every step of the chain, instead of actually

testing hypothesis, a similar line of thought would allow us to conclude that $h_i(\theta) = F^{-1}_{T_{Y_i}}(1-\gamma|\theta)$ seems to be a reasonable choice in this scenario as well. However, before formally defining the resulting proposed model, we must first consider under what conditions this choice of $h_i(\theta)$ makes sense.

Let us assume, as an example, that $Y_1, \ldots, Y_n|\theta \overset{ind}{\sim} Bernoulli(\theta)$. In this case we simply cannot expect to find particularly small values of $f_{Y_i}(y_i|\theta)$ unless $\theta \approx 1$ or $\theta \approx 0$. And even so, as long as $\theta \in (0,1)$, for sufficiently large values of γ we would still expect to find 0's and 1's regardless of how improbable those outcomes may be, leading us to always sample $z_i = 1$ and rendering the method pointless. More intuitively, if $\theta \in (0,1)$ we cannot differentiate typical and atypical observations because there is no distinctive characteristic to look for, since all the information available to make a decision is that Y_i is either 0 or 1. With this, we can conclude that, for practical purposes, searching for anomalies in a sample of Bernoulli random variables is undesirable. To fix this problem, we can impose as a condition that Y_i must assume values with arbitrarily small positive probabilities, so it is always possible to find unreasonable predictive values $f_{Y_i}(y_i|\theta)$, allowing us to detect anomalies. This line of thought culminates in the condition presented below.

Definition 3.2 (Filtering Condition)**.** Let X be a d-dimensional random vector with density or probability mass function f_X. Then X is said to satisfy the *filtering condition* if for all $\varepsilon > 0$

$$\mathbb{P}(T_X \leq \varepsilon) > 0, \quad \text{or equivalently,} \quad \mathbb{P}(T_X > \varepsilon) < 1. \tag{3.9}$$

Next, we must consider that, since T_{Y_i} can be a discrete or even a mixed random variable, when we define $h_i(\theta) = F^{-1}_{T_{Y_i}}(\cdot|\theta)$ we must determine what this inverse represent in this context. So, before presenting the proposed model, we must firstly present one more definition: the definition of Generalized Inverse Function, which is given below.

Definition 3.3 (Generalized Inverse Function)**.** Let X be a d-dimensional absolutely continuous or discrete random vector with distribution F_X. Then, if X satisfies the

filtering condition, we define the generalized inverse function of T_X, as

$$F_{T_X}^{-1}(y) = \inf\{x \in \mathbb{R} : F_{T_X}(x) \geq y\}, \quad 0 < y \leq 1 \tag{3.10}$$

and $F_{T_X}^{-1}(0) = 0$. In other words, $F_{T_X}^{-1}(y)$ is the smallest value $x \in \mathbb{R} \cup \{-\infty, +\infty\}$ for which $F_{T_X}(x)$ is greater than or equal to y.

To provide some intuition for why the Definition 3.3 would be desirable, notice that our inverse is similar to the left-continuous inverse of F_{T_X}, for which $F_{T_X}^{-1}(0) = -\infty$. However, since an autotransformation T_X is almost surely positive, we want $F_{T_X}^{-1}$ to always assume non-negative values, so considering $F_{T_X}^{-1}(0) = 0$ instead fixes the problem.

Finally, let $Y_1, \ldots, Y_n$ be d-dimensional absolutely continuous or discrete random vectors with respective distribution functions $F_{Y_1}(\cdot|\theta), \ldots, F_{Y_n}(\cdot|\theta)$, where θ is an unknown parameter vector, and such that Y_i satisfy the filtering condition given θ for all $i \in \{1, \ldots, n\}$. Then, the naive filtering model is given by the following hierarchical structure:

$$\begin{aligned} Y_i|\theta, z_i = 1 &\overset{ind}{\sim} F_{Y_i}(\cdot|\theta), \\ Y_i|\theta, z_i = 0 &\overset{ind}{\sim} Uniform(S_i), \\ (\theta, z) &\sim \pi(\theta, z), \\ \mu_L(S_i)^{-1} &= F_{T_{Y_i}}^{-1}(1 - \gamma|\theta) \end{aligned} \tag{3.11}$$

where $\pi(\theta, z)$ is the prior distribution for θ and z, μ_L represents the Lebesgue measure, $F_{T_{Y_i}}^{-1}(\cdot|\theta)$ is the inverse function according to Definition 3.3 and $\gamma \in (0, 1)$ is a hyperparameter to be chosen. Next, considering once more the simulated data set 1, we use the naive filtering model to generate a classification for each data point.

Here, we choose $\gamma = 0.99$, $w_i = \frac{1}{2}$, assume θ known and use the expression for the autotransformation of a normal distribution, presented in section A.4 of Appendix A, for our calculations. Figure 5 presents the posterior expected value of the indicator variables considering the naive filtering model, given by

$$\mathbb{P}(z_i = 1|\,\theta, y) = \mathbb{E}[z_i|\,\theta, y] = \frac{f_{Y_i}(y_i|\theta)}{f_{Y_i}(y_i|\theta) + F_{T_{Y_i}}^{-1}(1 - \gamma|\theta)}. \tag{3.12}$$

As we can see, the naive filtering model correctly identifies the atypical observation and attributes a probability of at least 50% to all of the typical observations. So, if

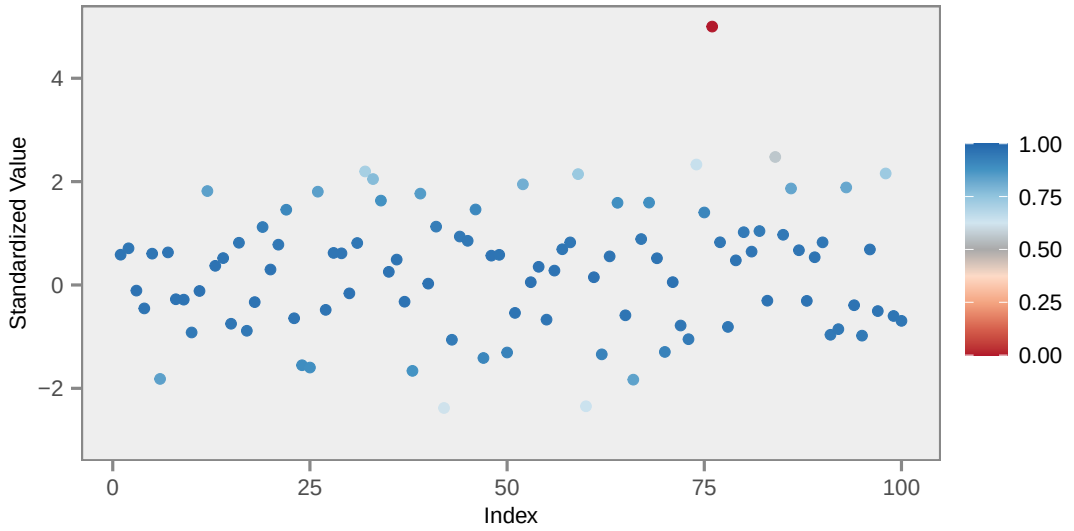

Figure 5: Visualization of the simulated data set 1 with the first 50 observations divided by 100000, where the colors indicate the estimated probability of belonging to the main component according to the naive filtering model.

we were to classify $z_i = 1$, if $\mathbb{P}(z_i = 1|\,\theta, y) \geq \mathbb{P}(z_i = 0|\,\theta, y)$, and $z_i = 0$, otherwise, the resulting accuracy would be of 100%.

Despite the good result obtained in our first toy example, the naive filtering model does present an important shortcoming which we will showcase with a simple example. Let us introduce the *simulated data set 2*, consisting of 3 samples of the same distribution $Normal(0, 1)$. The difference between the samples is their size, so the first sample has a total of 50 observations, the second 500 and the third 5000. We then apply the naive filtering model with $\gamma = 0.80$, $w_i = \frac{1}{2}$ and assuming θ known for each one of the three samples. Table 1 summarizes the results for each one of the samples and Figure 6 presents the posterior expected value of the indicator variables considering the naive filtering model.

As we can see from table 1, as the sample size increases, the number of removed observations $n - n_1$ also increases, attaining a proportion of observations close to the chosen value γ. The overall effect of this removal of observations is essentially a soft truncation of the distribution we wish to estimate, since, even in the case of knowing the exact main component distribution, we still expect to remove $100(1 - \gamma)\%$ of the typical observations that are closest to the tails. The origin of this phenomenon

	n	$n_1 = \sum_{i=1}^{n} z_i$	$\frac{n_1}{n}$
Sample 1	50	44	0.88
Sample 2	500	406	0.812
Sample 3	5000	4021	0.8042

Table 1: Summary of the estimation for the three samples of the simulated data set 2.

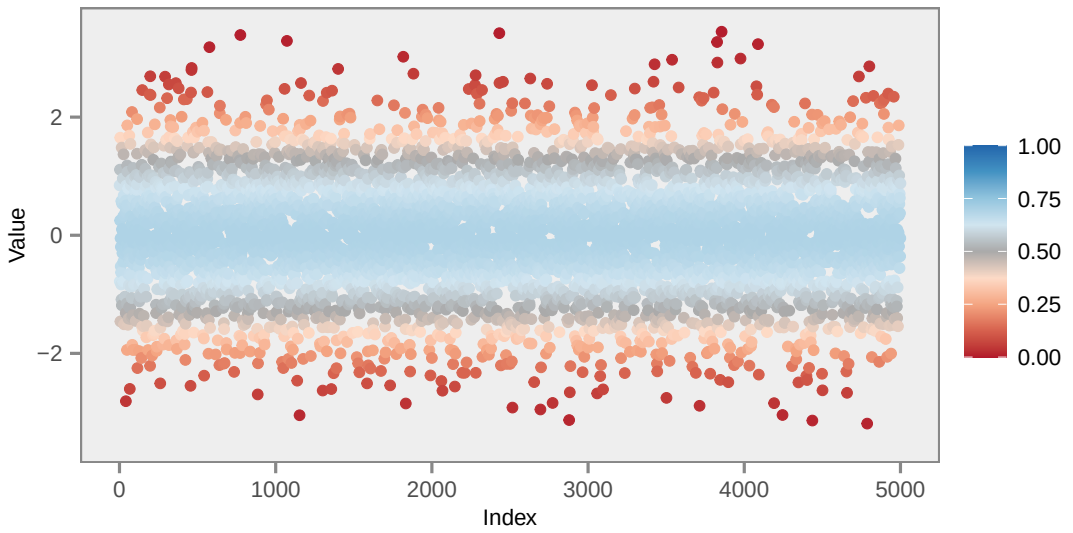

Figure 6: Visualization of the sample 3 of the simulated data set 2, where the colors indicate the estimated probability of belonging to the main component according to the naive filtering model.

can be traced back to the test of hypothesis considered to define the alternative component, since we only *naively* considered controlling the error of rejecting *one* typical observation, when we should account for the total rejection error considering the n decisions made instead.

Another problem worth commenting is masked by our assumption of knowing θ. But if we also considered the estimation of θ, we could expect to obtain estimates of θ with a strong bias because of the truncation, which is undesirable. So, in the next subsection, we discuss how to circumvent this issue with an adaptation of the alternative component.

3.1.2 Biased Filtering Model

Considering the problem of the soft truncation caused by the naive filtering model, we can consider altering our choice of γ to account for the fact that we consider n decisions instead of only one. When considering a hypothesis testing, a common procedure is to use the Bonferroni correction, however, we provide here a different approach to this issue. We established in subsection 3.1.1 that, considering an estimation via MCMC, at every step of the chain we sample z_i given θ and z_{-i} based on the relation between $f_{Y_i}(y_i|\theta)$ and $h_i(\theta)$. What is interesting to point out is that the expression $f_{Y_i}(y_i|\theta)$ can also be interpreted as the value of the likelihood pseudo-depth function for the observation y_i. So, taking inspiration from the re-ordering propriety of depth functions, we can consider defining $h_i(\theta)$ in terms of the *most extreme* observation of our sample of size n. Since, in this case, by *most extreme* we mean having the smallest relative predictive value, one could suppose that we need to compare the predictive values between observations with possibly distinct distributions. However, we can attain a similar effect if we compare how extreme an observation y_i is with respect to the most extreme observation from a virtual sample $Y_i^1, \ldots, Y_i^n \overset{ind}{\sim} F_{Y_i}(\cdot|\theta)$. Then, we can define $h_i(\theta)$ to be the quantile $1-\gamma$ of the random variable given by $\min\{T_{Y_i^1}, \ldots, T_{Y_i^n}\}$, representing the distribution of the predictive value of the most extreme observation in an independent sample of size n. To simplify the notation, we next introduce the definition below.

Definition 3.4 (Autotransformation of Order n)**.** Let X be a d-dimensional absolutely continuous or discrete random vector. Then the *autotransformation of order* n of X is given by $T_X^{(n)} = \min\{T_{X_1}, \ldots, T_{X_n}\}$, where $T_{X_1}, \ldots, T_{X_n}$ are independent and identically distributed autotransformations of X.

Considering Definition 3.4 and the soft truncation problem, we can now choose $h_i(\theta) = F^{-1}_{T^{(n)}_{Y_i}}(1-\gamma|\theta)$, but the interesting part about this choice is that our newly defined function satisfies

$$h_i(\theta) = F^{-1}_{T^{(n)}_{Y_i}}(1-\gamma|\theta) = F^{-1}_{T_{Y_i}}\left(1-\gamma^{\frac{1}{n}}\middle|\,\theta\right), \tag{3.13}$$

so we can interpret the new $h_i(\theta)$ as simply using a correction for the value of γ that accounts for the sample size n. For the result presented above, we use proposition A.2, formally stated and proved in section A.1 of Appendix A.

Notice, however, that the original motivation for adapting the value of γ was to control how many observations *from the main component* are classified as atypical. This means that our choice of $h_i(\theta)$ over-corrects the value of γ and a more precise choice would be to consider taking the minimum of $n_1 + 1$ variables instead of n, where $n_1 = \sum_{i=1}^n z_i$ is the number of typical observations in our sample. Having all of this in mind, we next propose a correction of the naive filtering model.

Let $Y_1, \ldots, Y_n$ be d-dimensional absolutely continuous or discrete random vectors with respective distribution functions $F_{Y_1}(\cdot|\theta), \ldots, F_{Y_n}(\cdot|\theta)$, where θ is an unknown parameter vector, and such that Y_i satisfy the filtering condition given θ for all $i \in \{1, \ldots, n\}$. Then, the biased filtering model is given by the following hierarchical structure:

$$\begin{aligned} Y_i|\theta, z_i = 1 &\overset{ind}{\sim} F_{Y_i}(\cdot|\theta), \\ Y_i|\theta, z_{-i}, z_i = 0 &\overset{ind}{\sim} Uniform(S_i), \\ (\theta, z) &\sim \pi(\theta, z), \\ \mu_L(S_i)^{-1} &= F^{-1}_{T_{Y_i}}\left(1-\gamma^{(n_1+1)^{-1}}\middle|\,\theta\right) \end{aligned} \tag{3.14}$$

where $\pi(\theta, z)$ is the prior distribution for θ and z, z_{-i} represents all of the indicators with the exception of z_i, $n_1 = \sum_{i=1}^n z_i$, μ_L represents the Lebesgue measure, $F^{-1}_{T_{Y_i}}(\cdot|\theta)$ is the inverse function according to Definition 3.3 and $\gamma \in (0, 1)$ is a hyperparameter

to be chosen. The posterior for this model and the full conditional distribution of z_i, as well as other aspects of inference are presented later in this subsection.

Before returning to the problem involving the simulated data set 2, one may be interest in comparing how our chosen corrected quantile performs when compared to the well established Bonferroni correction. In this case, considering a fixed value of n and the corrections as functions of γ, our interest lies in comparing the functions

$$f_{Bon}(\gamma) = \frac{1-\gamma}{n} \quad \text{and} \quad f_{ours}(\gamma) = 1 - \gamma^{\frac{1}{n}}. \tag{3.15}$$

Considering now as a linear approximation of f_{ours} the tangent line of f_{ours} at $\gamma = 1$, we have

$$\begin{aligned} f_{ours}(\gamma) &\approx f_{ours}(1) + \left.\frac{df_{ours}(\gamma)}{d\gamma}\right|_{\gamma=1} (\gamma - 1) = 0 + \left.\frac{d}{d\gamma}\left[1 - \gamma^{\frac{1}{n}}\right]\right|_{\gamma=1} (\gamma - 1) \\ &= \left[-\frac{\gamma^{\frac{1}{n}-1}}{n}\right]\Bigg|_{\gamma=1} (\gamma - 1) = \frac{1-\gamma}{n} = f_{Bon}(\gamma). \end{aligned} \tag{3.16}$$

Which implies that the Bonferroni correction can be seen as, in some sense, the best linear approximation of our correction when $\gamma \approx 1$. With this simple result, we can also expect both approximations to perform similarly when considering relatively high values of γ and we leave the comparison for lower values of γ for future work.

Considering once more the sample 3 of the simulated data set 2, we use the biased filtering model to estimate the expected value of the indicators. Our results are presented in Figure 7, where we used $\gamma = 0.95$, $w_i = \frac{1}{2}$ and assumed θ known. The estimates were made by sampling from the posterior distribution $\pi(z|\theta, y)$ considering a Gibbs algorithm and then estimating the expected value via Monte Carlo integration, as we considered a total of 20000 chain steps and eliminated the first 2000 iterations as burn in. For reference, the approximate computation time required to obtain the sample via MCMC was of 1 minute and 45 seconds. As we can see, the improved version of our proposed model essentially removes the soft truncation, thus providing a satisfactory solution to the problem.

We next present the posterior and the full conditional distribution of z_i for the

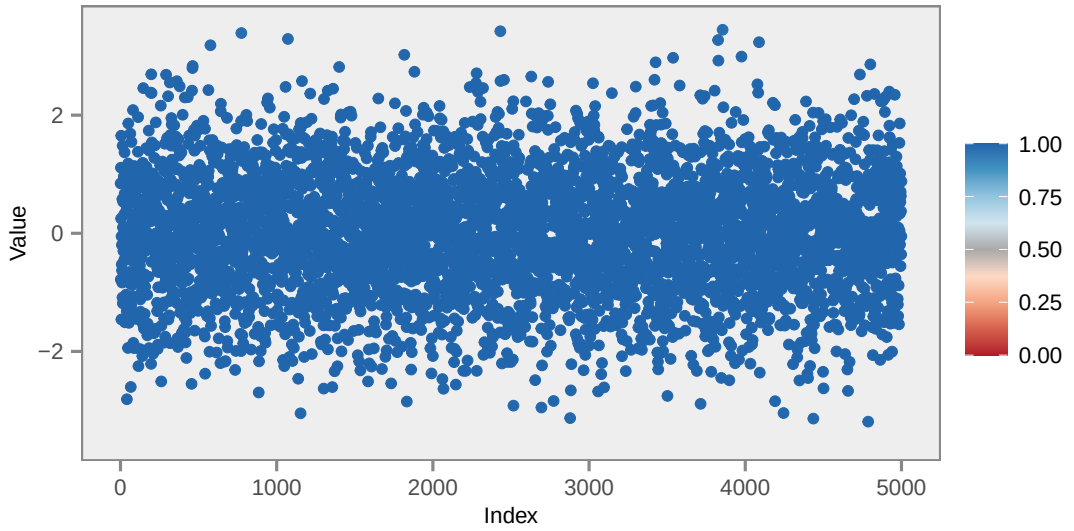

Figure 7: Visualization of the sample 3 of the simulated data set 2, where the colors indicate the estimated probability of belonging to the main component according to the biased filtering model.

model of equation 3.14. The posterior is given by

$$
\begin{aligned}
&\pi(\theta, z|y) \\
&\overset{Bayes}{=} \frac{\pi(\theta, z)\pi(y|\theta, z)}{\sum_z \int_\Theta \pi(\theta, z)\pi(y|\theta, z)\ d\theta} \propto \pi(\theta, z)\pi(y|\theta, z) \\
&\overset{ind}{=} \pi(\theta) \prod_{i=1}^{n} \left[\pi(z_i)\right] \prod_{i=1}^{n} \left[f_{Y_i}(y_i|\theta)\right]^{z_i} \prod_{i=1}^{n} \left[\mu_L(S_i)^{-1} \underbrace{\mathbb{I}_{S_i}(y_i)}_{=1}\right]^{1-z_i} \\
&\overset{ind}{=} \pi(\theta) \prod_{i=1}^{n} \left[w_z^{z_i}(1-w_z)^{1-z_i}\right] \prod_{i=1}^{n} \left[f_{Y_i}(y_i|\theta)\right]^{z_i} \prod_{i=1}^{n} \left[F^{-1}_{T_{Y_i}}\left(1-\gamma^{(n_1+1)^{-1}}\middle|\,\theta\right)\right]^{1-z_i},
\end{aligned}
\tag{3.17}
$$

and for the full conditional distribution of z_k we have

$$
\begin{aligned}
&\pi(z_k|\theta, z_{-k}, y) \propto \pi(\theta, z|y) \\
&\overset{ind}{=} \pi(z_k|\theta, z_{-k})\pi(\theta, z_{-k}) \prod_{i=1}^{n} \left[f_{Y_i}(y_i|\theta)\right]^{z_i} \prod_{i=1}^{n} \left[F^{-1}_{T_{Y_i}}\left(1-\gamma^{(n_1+1)^{-1}}\middle|\,\theta\right)\right]^{1-z_i} \\
&\propto \pi(z_k|\theta, z_{-k}) \left[f_{Y_k}(y_k|\theta)\right]^{z_k} \left[F^{-1}_{T_{Y_k}}\left(1-\gamma^{(n_1^{-k}+z_k+1)^{-1}}\middle|\,\theta\right)\right]^{1-z_k} \\
&\times \prod_{i\neq k} \left[F^{-1}_{T_{Y_i}}\left(1-\gamma^{(n_1^{-k}+z_k+1)^{-1}}\middle|\,\theta\right)\right]^{1-z_i},
\end{aligned}
\tag{3.18}
$$

where $\pi(z_k = 1|\theta, z_{-k}) = w_k$ and $n_j^{-k} = \sum_{i\neq k} \mathbb{I}_{\{z_i=j\}}$, for $j \in \{0, 1\}$. To facilitate

the identification of the resulting Bernoulli distribution, we can calculate

$$\begin{aligned}\pi(z_k = 1|\theta, z_{-k}, y) &\propto \pi(\theta, z|y) \\ &\propto w_k f_{Y_k}(y_k|\theta) \prod_{i \neq k} \left[F^{-1}_{T_{Y_i}} \left(1 - \gamma^{(n_1^{-k}+2)^{-1}} \middle| \theta \right) \right]^{1-z_i}\end{aligned} \tag{3.19}$$

and

$$\begin{aligned}&\pi(z_k = 0|\theta, z_{-k}, y) \propto \pi(\theta, z|y) \\ &\propto (1 - w_k) F^{-1}_{T_{Y_k}} \left(1 - \gamma^{(n_1^{-k}+1)^{-1}} \middle| \theta \right) \prod_{i \neq k} \left[F^{-1}_{T_{Y_i}} \left(1 - \gamma^{(n_1^{-k}+1)^{-1}} \middle| \theta \right) \right]^{1-z_i},\end{aligned} \tag{3.20}$$

which implies

$$z_k|\theta, z_{-k}, y \sim Bernoulli\left(\frac{a_k}{a_k + b_k} \right) \tag{3.21}$$

where

$$a_k = w_k f_{Y_k}(y_k|\theta), \text{ and}$$

$$b_k = (1 - w_k) \underbrace{F^{-1}_{T_{Y_k}} \left(1 - \gamma^{(n_1^{-k}+1)^{-1}} \middle| \theta \right)}_{\text{desired term}} \underbrace{\prod_{i \neq k} \left[\frac{F^{-1}_{T_{Y_i}} \left(1 - \gamma^{(n_1^{-k}+1)^{-1}} \middle| \theta \right)}{F^{-1}_{T_{Y_i}} \left(1 - \gamma^{(n_1^{-k}+2)^{-1}} \middle| \theta \right)} \right]^{1-z_i}}_{\text{bias}}. \tag{3.22}$$

As we can see from the equations above, we introduced a bias on the full conditional when we altered the measure of the unknown regions $S_1, \ldots, S_n$ from the alternative component. If we then consider the particular case of independent and identically distributed response variables, we can better understand the effect of this bias in the parameter estimation. In this case, we can simplify equation 3.22 as follows:

$$\begin{aligned}b_k &= (1 - w_k) F^{-1}_{T_{Y_k}} \left(1 - \gamma^{(n_1^{-k}+1)^{-1}} \middle| \theta \right) \prod_{i \neq k} \left[\frac{F^{-1}_{T_{Y_i}} \left(1 - \gamma^{(n_1^{-k}+1)^{-1}} \middle| \theta \right)}{F^{-1}_{T_{Y_i}} \left(1 - \gamma^{(n_1^{-k}+2)^{-1}} \middle| \theta \right)} \right]^{1-z_i} \\ &\overset{id}{=} (1 - w_k) F^{-1}_{T_{Y_k}} \left(1 - \gamma^{(n_1^{-k}+1)^{-1}} \middle| \theta \right) \prod_{i \neq k} \left[\frac{F^{-1}_{T_{Y_k}} \left(1 - \gamma^{(n_1^{-k}+1)^{-1}} \middle| \theta \right)}{F^{-1}_{T_{Y_k}} \left(1 - \gamma^{(n_1^{-k}+2)^{-1}} \middle| \theta \right)} \right]^{1-z_i} \\ &= (1 - w_k) \underbrace{F^{-1}_{T_{Y_k}} \left(1 - \gamma^{(n_1^{-k}+1)^{-1}} \middle| \theta \right)}_{\text{desired term}} \underbrace{\left[\frac{F^{-1}_{T_{Y_k}} \left(1 - \gamma^{(n_1^{-k}+1)^{-1}} \middle| \theta \right)}{F^{-1}_{T_{Y_k}} \left(1 - \gamma^{(n_1^{-k}+2)^{-1}} \middle| \theta \right)} \right]^{n_0^{-k}}}_{\text{bias}}.\end{aligned} \tag{3.23}$$

Here we can see that the bias in equation 3.23 is the ratio between the chosen quantiles assuming that $z_k = 0$ and $z_k = 1$ to the n_0^{-k}-th power. Furthermore,

notice that

$$
\begin{aligned}
& n_1^{-k} + 1 < n_1^{-k} + 2 \\
& \Rightarrow (n_1^{-k} + 1)^{-1} > (n_1^{-k} + 2)^{-1} \\
& \Rightarrow \gamma^{(n_1^{-k}+1)^{-1}} < \gamma^{(n_1^{-k}+2)^{-1}} \\
& \Rightarrow 1 - \gamma^{(n_1^{-k}+1)^{-1}} > 1 - \gamma^{(n_1^{-k}+2)^{-1}} \\
& \Rightarrow F^{-1}_{T_{Y_k}}\left(1 - \gamma^{(n_1^{-k}+1)^{-1}} \middle| \theta\right) > F^{-1}_{T_{Y_k}}\left(1 - \gamma^{(n_1^{-k}+2)^{-1}} \middle| \theta\right) \\
& \Rightarrow \left[\frac{F^{-1}_{T_{Y_k}}\left(1 - \gamma^{(n_1^{-k}+1)^{-1}} \middle| \theta\right)}{F^{-1}_{T_{Y_k}}\left(1 - \gamma^{(n_1^{-k}+2)^{-1}} \middle| \theta\right)}\right]^{n_0^{-k}+1} > \left[\frac{F^{-1}_{T_{Y_k}}\left(1 - \gamma^{(n_1^{-k}+1)^{-1}} \middle| \theta\right)}{F^{-1}_{T_{Y_k}}\left(1 - \gamma^{(n_1^{-k}+2)^{-1}} \middle| \theta\right)}\right]^{n_0^{-k}} > 1.
\end{aligned}
\tag{3.24}
$$

Thus, the bias favors classifying the observations as anomalies and grows exponentially in significance with the number of atypical observations n_0. We can more directly understand the bias when considering that, according to equation 3.24, $F^{-1}_{T_{Y_i}}\left(1 - \gamma^{(n_1+1)^{-1}} \middle| \theta\right)$ decreases with n_1, so whenever we choose to classify an observation as anomalous we end up increasing the contribution of this term for all of the observations that were already labeled as atypical. In other words, the model has a greater incentive to reject observations than what was chosen by design and this tendentious behavior also leads to a positive feedback loop that can potentially result in the rejection of all observations.

Going back to the more general case of non-identically distributed observations, we can expect a similar behavior for the bias in equation 3.22, since in principle the bias-generating mechanism is the same. Next, we showcase with a numerical example how this bias affects estimation.

Let us consider another simulation study, where we call simulated data set 3 the sample consisting of 20 independent observations with $Normal(0, 1)$ distribution, to represent the main component, and then 60 anomalous observations arbitrarily chosen to be equal to 5. The idea here is to consider a sample with a relatively high number of atypical observations in order to inflate the bias of the model. Figure 8 shows the sampled values for each indicator at each iteration of the Gibbs algorithm. Here we used $\gamma = 0.99$, $w_i = \frac{1}{2}$ and assumed θ known to generate a total of 30 steps of the Markov chain.

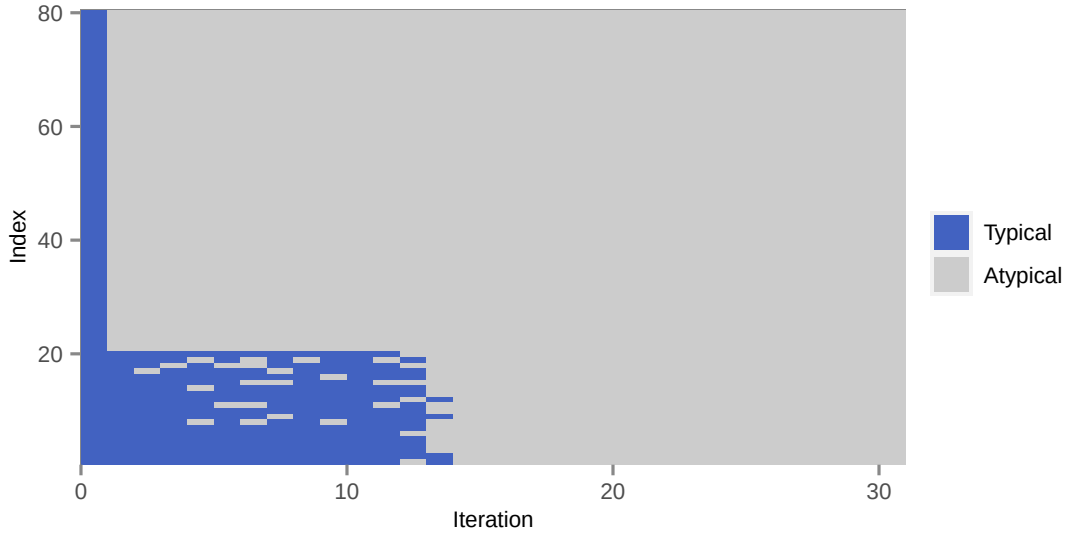

Figure 8: Visualization of the indicators for each observation of the simulated data set 3 throughout the Markov chain considering the biased filtering model. Index 1 to 20 correspond to the typical sample and index 21 to 80 correspond to the anomalous observations.

As Figure 8 shows, even choosing as an initial condition $z_i = 1$, for all $i \in \{1, \ldots, n\}$ and $\gamma = 0.99$, the chain eventually started sampling every indicator as 0 due to the bias previously discussed. So, in this case, estimation simply becomes impracticable.

3.1.3 Filtering Model

In this section, we present the last version of our proposed methodology for this work, the filtering model, which is an improvement over the biased filtering model because it address the problems arising from the dependency on the indicators z for the alternative component. With this modification we mitigate the influence of the bias on parameter estimation and, for some specific cases, are even able to remove it completely.

In order to achieve the desired bias reduction, we first consider a general family of modifications to then choose the most effective correction. Having that in mind, we introduce auxiliary functions $g_1, \ldots, g_n : \mathbb{R} \to \mathbb{R}$ and consider the modification

of the alternative component of the biased filtering model given by

$$\mu_L(S_i)^{-1} = F^{-1}_{T_{Y_i}}\left(1-\gamma^{g_i(n_1+1)^{-1}}\middle|\,\theta\right). \tag{3.25}$$

If we then assume that g_i does not depend on the collection of indicators z, we can analogously recalculate the full conditional for each indicator z_k, obtaining

$$z_k|\theta, z_{-k}, y \overset{ind}{\sim} Bernoulli\left(\frac{a_k}{a_k+b_k}\right) \tag{3.26}$$

where

$$\begin{aligned} &a_k = w_k f_{Y_k}(y_k|\theta), \text{ and} \\ &b_k = (1-w_k)F^{-1}_{T_{Y_k}}\left(1-\gamma^{g_k(n_1^{-k}+1)^{-1}}\middle|\,\theta\right)\prod_{i\neq k}\left[\frac{F^{-1}_{T_{Y_i}}\left(1-\gamma^{g_i(n_1^{-k}+1)^{-1}}\middle|\,\theta\right)}{F^{-1}_{T_{Y_i}}\left(1-\gamma^{g_i(n_1^{-k}+2)^{-1}}\middle|\,\theta\right)}\right]^{1-z_i}, \end{aligned} \tag{3.27}$$

and recalling that $n_j^{-k} = \sum_{i\neq k}\mathbb{I}_{\{z_i=j\}}$, for $j\in\{0,1\}$. However we still require a criteria in order to select *the most effective* auxiliary function. A natural choice would be, if possible, a function that controls the bias considering a simplifying hypothesis, for instance the identically distributed case. And so, focusing on this case we have

$$\begin{aligned} &a_k = w_k f_Y(y_k|\theta), \text{ and} \\ &b_k = (1-w_k)F^{-1}_{T_{Y_k}}\left(1-\gamma^{g_k(n_1^{-k}+1)^{-1}}\middle|\,\theta\right)\left[\frac{F^{-1}_{T_{Y_k}}\left(1-\gamma^{g_k(n_1^{-k}+1)^{-1}}\middle|\,\theta\right)}{F^{-1}_{T_{Y_k}}\left(1-\gamma^{g_k(n_1^{-k}+2)^{-1}}\middle|\,\theta\right)}\right]^{n_0^{-k}}. \end{aligned} \tag{3.28}$$

With the expression above, we can establish that our auxiliary function removes the bias in the independent and identically distributed case if, and only if, it satisfies the equation

$$\begin{aligned} &F^{-1}_{T_{Y_k}}\left(1-\gamma^{g_k(n_1^{-k}+1)^{-1}}\middle|\,\theta\right)\left[\frac{F^{-1}_{T_{Y_k}}\left(1-\gamma^{g_k(n_1^{-k}+1)^{-1}}\middle|\,\theta\right)}{F^{-1}_{T_{Y_k}}\left(1-\gamma^{g_k(n_1^{-k}+2)^{-1}}\middle|\,\theta\right)}\right]^{n_0^{-k}} \\ &= F^{-1}_{T_{Y_k}}\left(1-\gamma^{(n_1^{-k}+1)^{-1}}\middle|\,\theta\right), \end{aligned} \tag{3.29}$$

for all possible values of n_1^{-k}. Next, we introduce the somewhat unprompted definition of a correction function and later we clarify how this function helps us remove the bias of the filtering model.

Definition 3.5 (Correction Function)**.** Let $Y_1, \ldots, Y_n$ be an independent sample from the d-dimensional absolutely continuous or discrete distributions $F_{Y_1}(\cdot|\theta)$, ..., $F_{Y_n}(\cdot|\theta)$ and $\gamma \in (0,1)$ a real scalar. If $Y_1, \ldots, Y_n$ satisfy the filtering condition given θ and denoting h_i the distribution of the autotransformation of the i-th observation, then the *correction function* $g_i = g_i(\cdot|\theta, \gamma)$ for Y_i is the function satisfying the recursive expression given by

$$g_i(x) = \left(\log_\gamma \left[1 - h_i \left(\left[h_i^{-1} \left(1 - \gamma^{x^{-1}} \right) \right]^{\frac{1}{n-x+1}} \left[h_i^{-1} \left(1 - \gamma^{g_i(x+1)^{-1}} \right) \right]^{\frac{n-x}{n-x+1}} \right) \right] \right)^{-1}, \tag{3.30}$$

for $x \in \{1, \ldots, n-1\}$, and satisfying

$$g_i(n) = n, \tag{3.31}$$

for all $i \in \{1, \ldots, n\}$.

From Definition 3.5 it is unclear whether or not such objects exists or even how to obtain numerical values from the expression in equation 3.30, since the correction functions are defined implicitly. So, in section A.2 of Appendix A we will show that correction functions are well defined and provide proof for all of the claims made throughout this section.

It is worth noticing that Definition 3.5 does not assume identically distributed observations, but unfortunately we do not have any specific results to present for this more general case. However, the correction function for Y_i is defined such that it satisfies

$$F_{T_{Y_i}}^{-1}\left(1 - \gamma^{g_i(x)^{-1}} \middle| \theta\right) \left[\frac{F_{T_{Y_i}}^{-1}\left(1 - \gamma^{g_i(x)^{-1}} \middle| \theta\right)}{F_{T_{Y_i}}^{-1}\left(1 - \gamma^{g_i(x+1)^{-1}} | \theta\right)} \right]^{n-x} \leq F_{T_{Y_i}}^{-1}\left(1 - \gamma^{x^{-1}} \middle| \theta\right), \tag{3.32}$$

for $x \in \{1, \ldots, n\}$. Furthermore, if T_{Y_i} is absolutely continuous, then the correction function for Y_i satisfies

$$F_{T_{Y_i}}^{-1}\left(1 - \gamma^{g_i(x)^{-1}} \middle| \theta\right) \left[\frac{F_{T_{Y_i}}^{-1}\left(1 - \gamma^{g_i(x)^{-1}} \middle| \theta\right)}{F_{T_{Y_i}}^{-1}\left(1 - \gamma^{g_i(x+1)^{-1}} | \theta\right)} \right]^{n-x} = F_{T_{Y_i}}^{-1}\left(1 - \gamma^{x^{-1}} \middle| \theta\right), \tag{3.33}$$

for $x \in \{1, \ldots, n\}$ and $i \in \{1, \ldots, n\}$. From equations 3.32 and 3.33 we can finally derive the association between correction functions and the auxiliary functions

present in equation 3.25 using the relation $n_1^{-k} + n_0^{-k} = n - 1$. Notice that, if Y_k is a d-dimensional absolute continuous or discrete random vector and we choose the auxiliary function for Y_k to be the correction function for Y_k, then for the independent and identically distributed case we have

$$\begin{aligned} a_k &= w_k f_Y(y_k|\theta), \text{ and} \\ b_k &= (1-w_k) F^{-1}_{T_{Y_k}}\left(1-\gamma^{g_k(n_1^{-k}+1)^{-1}}\middle|\,\theta\right)\left[\frac{F^{-1}_{T_{Y_k}}\left(1-\gamma^{g_k(n_1^{-k}+1)^{-1}}\middle|\,\theta\right)}{F^{-1}_{T_{Y_k}}\left(1-\gamma^{g_k(n_1^{-k}+2)^{-1}}\middle|\,\theta\right)}\right]^{n_0^{-k}} \\ &\overset{\text{eq.3.32}}{\leq} (1-w_k) F^{-1}_{T_{Y_k}}\left(1-\gamma^{(n_1^{-k}+1)^{-1}}\middle|\,\theta\right), \end{aligned} \tag{3.34}$$

for $k \in \{1, \ldots, n\}$, allowing us to control the probability of classifying an observation as atypical, that is, the probability is at most the desired one. What is even more surprising is that, if T_{Y_k} is also absolutely continuous, then for the independent and identically distributed case we have

$$\begin{aligned} a_k &= w_k f_Y(y_k|\theta), \text{ and} \\ b_k &= (1-w_k) F^{-1}_{T_{Y_k}}\left(1-\gamma^{g_k(n_1^{-k}+1)^{-1}}\middle|\,\theta\right)\left[\frac{F^{-1}_{T_{Y_k}}\left(1-\gamma^{g_k(n_1^{-k}+1)^{-1}}\middle|\,\theta\right)}{F^{-1}_{T_{Y_k}}\left(1-\gamma^{g_k(n_1^{-k}+2)^{-1}}\middle|\,\theta\right)}\right]^{n_0^{-k}} \\ &\overset{\text{eq.3.33}}{=} (1-w_k) F^{-1}_{T_{Y_k}}\left(1-\gamma^{(n_1^{-k}+1)^{-1}}\middle|\,\theta\right), \end{aligned} \tag{3.35}$$

for $k \in \{1, \ldots, n\}$, completely removing the bias. It can also be proven that, for location-scale models with absolutely continuous autotransformations, even with possibly different location and scale parameters for each data point, the model is also unbiased. As an extrapolation, we expect the correction functions to approximately contain the bias' inflation outside of these cases, speculatively leading to an overall improvement of the model. However, further investigations lie outside of the scope of this work, so more research is needed.

Even though the correction functions are useful to control the bias in some specific cases, for parameter estimation we still require a method to calculate its values. So, for the correction function g_i for Y_i consider the following.

1. We only need to evaluate g_i at the integer values $1, \ldots, n$.

2. Equation 3.30 provides a recursive relation that allows us to find the value of $g_i(x)$ given $g_i(x+1)$.

3. For the boundary case, from equation 3.31 we have $g_k(n) = n$.

Then, to find the numerical values for the correction function we start taking $g_i(n) = n$ and then iteratively calculate $g_i(x)$ given $g_i(x+1)$ until obtaining the desired value. As a side note, even though the recursion in equation 3.30 theoretically allows us to calculate these values, this expression is typically numerically unstable, so instead it is recommended using the relation

$$\begin{aligned}&\left[h_i^{-1}\left(1-\gamma^{x^{-1}}\right)\right]^{\frac{1}{n-x+1}}\left[h_i^{-1}\left(1-\gamma^{g_i(x+1)^{-1}}\right)\right]^{\frac{n-x}{n-x+1}}\\&=\exp\left\{\frac{1}{n-x+1}\ln\left[h_i^{-1}\left(1-\gamma^{x^{-1}}\right)\right]+\frac{n-x}{n-x+1}\ln\left[h_i^{-1}\left(1-\gamma^{g_i(x+1)^{-1}}\right)\right]\right\}\end{aligned} \tag{3.36}$$

for a more stable algorithm. Next, we present the last version of the filtering model considered for this work.

Let $Y_1, \ldots, Y_n$ be d-dimensional absolutely continuous or discrete random vectors with respective distribution functions $F_{Y_1}(\cdot|\theta), \ldots, F_{Y_n}(\cdot|\theta)$, where θ is an unknown parameter vector, and such that Y_i satisfy the filtering condition given θ for all $i \in \{1, \ldots, n\}$. Then, the filtering model is given by the following hierarchical structure:

$$\begin{aligned}Y_i|\theta, z_i = 1 &\overset{ind}{\sim} F_{Y_i}(\cdot|\theta),\\ Y_i|\theta, z_{-i}, z_i = 0 &\overset{ind}{\sim} Uniform(S_i),\\ (\theta, z) &\sim \pi(\theta, z),\\ \mu_L(S_i)^{-1} &= F^{-1}_{T_{Y_i}}\left(1-\gamma^{g_i(n_1+1)^{-1}}\middle|\,\theta\right),\end{aligned} \tag{3.37}$$

where $\pi(\theta, z)$ is the prior distributions for θ and z, z_{-i} represents all of the indicators with the exception of z_i, $n_1 = \sum_{i=1}^{n} z_i$, μ_L represents the Lebesgue measure, $F^{-1}_{T_{Y_i}}(\cdot|\theta)$ is the inverse function according to Definition 3.3, $\gamma \in (0, 1)$ is a hyperparameter to be chosen and $g_i = g_i(\cdot|\theta, \gamma)$ is the correction function presented in Definition 3.5.

Now considering once more the simulated data set 3, Figure 9 shows the sampled values for each indicator at each iteration of the Gibbs algorithm. Here we used $\gamma = 0.99$, $w_i = \frac{1}{2}$ and assumed θ known to generate a total of 100 steps of the

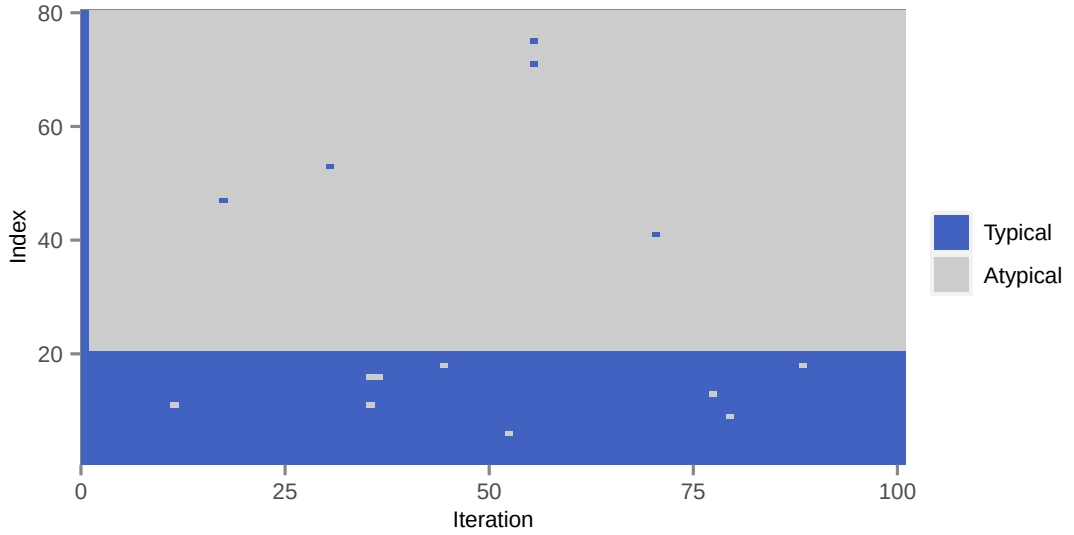

Figure 9: Visualization of the indicators for each observation of the simulated data set 3 throughout the Markov chain considering the (unbiased) filtering model. Index 1 to 20 correspond to the typical sample and index 21 to 80 correspond to the anomalous observations.

Markov chain. As we can see, the correction function adopted to control the bias seems to be working as desired, allowing the model to quickly identify and classify the observations.

3.2 PARAMETER ESTIMATION

Regarding the filtering model of 3.1.3, in this section we consider parameter estimation through MCMC methods, and in particular the Metropolis-Hastings algorithm and the Gibbs sampler described in section 2.2.1. In order to use these methods for model fitting, finding a function proportional to the model's posterior is required. However, the filtering model as presented in equation 3.37 still requires us to choose a model F_{Y_i} for the i-th observation and a prior $\pi(\theta, z)$.

We first consider the most general case possible, where we neither assume a specific distribution for the observations $Y_1, \ldots, Y_n$ nor a particular prior for (θ, z),

so we proceed to express the posterior distribution as follows:

$$\begin{aligned}
&\pi(\theta, z|y) \\
&\overset{Bayes}{=} \frac{\pi(\theta,z)\pi(y|\theta,z)}{\sum_z \int_\Theta \pi(\theta,z)\pi(y|\theta,z)\, d\theta} \propto \pi(\theta,z)\pi(y|\theta,z) \\
&\overset{ind}{=} \pi(\theta,z) \prod_{i=1}^{n} \left[f_{Y_i}(y_i|\theta) \right]^{z_i} \prod_{i=1}^{n} \left[\mu_L(S_i)^{-1} \underbrace{\mathbb{I}_{S_i}(y_i)}_{=1} \right]^{1-z_i} \\
&\overset{ind}{=} \pi(\theta,z) \prod_{i=1}^{n} \left[f_{Y_i}(y_i|\theta) \right]^{z_i} \prod_{i=1}^{n} \left[F^{-1}_{T_{Y_i}} \left(1 - \gamma^{g_i(n_1+1)^{-1}} \middle| \theta \right) \right]^{1-z_i}.
\end{aligned} \tag{3.38}$$

Now, since we wish to use the Metropolis-Hastings algorithm to generate samples from the posterior, we wish to find the full conditionals for every parameter of the model and then proposal distributions for each. So, considering the full conditional distribution of θ, we have

$$\begin{aligned}
&\pi(\theta|z,y) \propto \pi(\theta|z,y)\pi(z|y) = \pi(\theta,z|y) \\
&\propto \pi(\theta|z)\pi(z) \prod_{i=1}^{n} \left[f_{Y_i}(y_i|\theta) \right]^{z_i} \prod_{i=1}^{n} \left[F^{-1}_{T_{Y_i}} \left(1 - \gamma^{g_i(n_1+1)^{-1}} \middle| \theta \right) \right]^{1-z_i} \\
&\propto \pi(\theta|z) \prod_{i=1}^{n} \left[f_{Y_i}(y_i|\theta) \right]^{z_i} \prod_{i=1}^{n} \left[F^{-1}_{T_{Y_i}} \left(1 - \gamma^{g_i(n_1+1)^{-1}} \middle| \theta \right) \right]^{1-z_i},
\end{aligned} \tag{3.39}$$

and, for the full conditional of z_k, from subsection 3.1.3 we have

$$z_k|\theta, z_{-k}, y \overset{ind}{\sim} Bernoulli\left(\frac{a_k}{a_k + b_k} \right) \tag{3.40}$$

where

$$\begin{aligned}
a_k &= w_k f_{Y_k}(y_k|\theta), \\
b_k &= (1-w_k) F^{-1}_{T_{Y_k}} \left(1 - \gamma^{g_k(n_1^{-k}+1)^{-1}} \middle| \theta \right) \prod_{i \neq k} \left[\frac{F^{-1}_{T_{Y_i}} \left(1 - \gamma^{g_i(n_1^{-k}+1)^{-1}} \middle| \theta \right)}{F^{-1}_{T_{Y_i}} \left(1 - \gamma^{g_i(n_1^{-k}+2)^{-1}} \middle| \theta \right)} \right]^{1-z_i},
\end{aligned} \tag{3.41}$$

and recalling that $n_j^{-k} = \sum_{i \neq k} \mathbb{I}_{\{z_i = j\}}$, for $j \in \{0, 1\}$. Here, for z_k we consider a Gibbs step by using as the proposal distribution its own full conditional, surpassing the necessity of including an acceptance-rejection step. Nevertheless, we still need to find a good proposal distribution to sample θ.

Since we made only a few assumptions regarding the distributions of our sample $Y_1, \ldots, Y_n$, it is somewhat difficult to provide general guidance on how to find

reasonable proposals in the general case. So next, in order to help us make some interesting comparisons, we temporarily assume that $\pi(\theta, z) = \pi(\theta)\pi(z)$ and consider the modified full conditional of θ given by

$$\begin{aligned}
&\pi(\theta|z,y) \propto \pi(\theta|z,y)\pi(z|y) = \pi(\theta,z|y) \\
&\propto \pi(\theta|z)\pi(z)\prod_{i=1}^{n}\left[f_{Y_i}(y_i|\theta)\right]^{z_i}\prod_{i=1}^{n}\left[F^{-1}_{T_{Y_i}}\left(1-\gamma^{g_i(n_1+1)^{-1}}\Big|\,\theta\right)\right]^{1-z_i} \\
&\propto \pi(\theta)\prod_{i=1}^{n}\left[f_{Y_i}(y_i|\theta)\right]^{z_i}\prod_{i=1}^{n}\left[F^{-1}_{T_{Y_i}}\left(1-\gamma^{g_i(n_1+1)^{-1}}\Big|\,\theta\right)\right]^{1-z_i} \\
&= \pi(\theta)f(y^1|\theta)\prod_{i=1}^{n}\left[F^{-1}_{T_{Y_i}}\left(1-\gamma^{g_i(n_1+1)^{-1}}\Big|\,\theta\right)\right]^{1-z_i} \\
&\propto \pi(\theta|y^1)\prod_{i=1}^{n}\left[F^{-1}_{T_{Y_i}}\left(1-\gamma^{g_i(n_1+1)^{-1}}\Big|\,\theta\right)\right]^{1-z_i},
\end{aligned} \tag{3.42}$$

where y^1 consists of all observations y_i such that $z_i = 1$. Even though self-evident, it is worth pointing out that, if not for the term from the alternative component, the full conditional for θ would be proportional to the posterior distribution considering only the observations from the main model. With this in mind, if we assume a small number of atypical observations, then a good proposal for $\pi(\theta|y^1)$ should also work well for $\pi(\theta|z,y)$. For a large number of atypical observations, however, we still do not have a general guideline to find good proposals, so this must be considered on a case-by-case basis.

After performing the Metropolis-Hastings algorithm, assuming that from the Markov chain we obtain an independent sample $(\theta^{(1)}, z^{(1)}), \ldots, (\theta^{(m)}, z^{(m)})$ from the model's posterior distribution, we can use this sample to obtain point estimates for our parameters. Here, we consider standard Monte Carlo estimates for the posterior expected values of θ and z_i, respectively given by

$$\hat{\theta} = \frac{1}{m}\sum_{j=1}^{m}\theta^{(j)} \text{ and } \bar{z}_i = \frac{1}{m}\sum_{j=1}^{m}z_i^{(j)}, \tag{3.43}$$

for $i \in \{1, \ldots, n\}$. As presented in section 2.1, we justify this choice considering the Bayes' estimator that minimizes the posterior expect square loss. As an alternative to estimating the expected value of z_k, we can more directly estimate the probability

p_k of allocating the k-th observation to main component of the model by taking

$$\hat{p}_k = \frac{1}{m}\sum_{j=1}^{m} \frac{a_k^{(j)}}{a_k^{(j)} + b_k^{(j)}}, \tag{3.44}$$

where

$$\begin{aligned}
a_k^{(j)} &= w_k^{(j)} f_{Y_k}\left(y_k \left| \theta^{(j)}\right.\right), \\
b_k^{(j)} &= (1 - w_k^{(j)}) F_{T_{Y_k}}^{-1}\left(1 - \gamma^{g_k(n_{1j}^{-k}+1)^{-1}} \middle| \theta^{(j)}\right) \\
&\quad \times \prod_{i \neq k} \left[\frac{F_{T_{Y_i}}^{-1}\left(1 - \gamma^{g_i(n_{1j}^{-k}+1)^{-1}} \middle| \theta^{(j)}\right)}{F_{T_{Y_i}}^{-1}\left(1 - \gamma^{g_i(n_{1j}^{-k}+2)^{-1}} \middle| \theta^{(j)}\right)}\right]^{1-z_i^{(j)}}, \\
n_{1j}^{-k} &= \sum_{i \neq k} z_i^{(j)}, \\
w_k^{(j)} &= \pi\left(z_k = 1 \middle| \theta^{(j)}, z_{-k}^{(j)}\right)
\end{aligned} \tag{3.45}$$

for $k \in \{1, \ldots, n\}$. It is worth noting that even though $\hat{p}_k$ is expensive to compute, we already require the calculation of $a_k^{(j)}$ and $b_k^{(j)}$ for the j-th step of the chain in order to sample from the full conditional of z_k, so the additional cost is practically insignificant in comparison. Besides this, we call attention to the fact that, even though $z_i \in \{0, 1\}$, in this section we only consider estimating the expected value of z_i, so $\bar{z}_i \in [0, 1]$. For a discussion on how to use the sampled values of the chain to classify an observation as atypical or not, we redirect the reader to section 3.3. Next, we present some important characteristics of the filtering model, discuss some of the problems that arise during estimation and then propose a few techniques to circumvent them.

3.2.1 Proprieties of the Filtering Model

The first interesting property of the filtering model is that, under certain conditions, it presents a highly multimodal posterior. To illustrate this, consider the histograms presented in Figure 10. Both of them are a graphical representations of the same sample of size $n = 1000$ generated from a $Normal(0, 1)$, albeit considering different bin sizes. The left plot represents what we typically expect from the histogram of a normally distributed sample, unimodal and approximately symmetric, but the right one shows signs of multimodality.

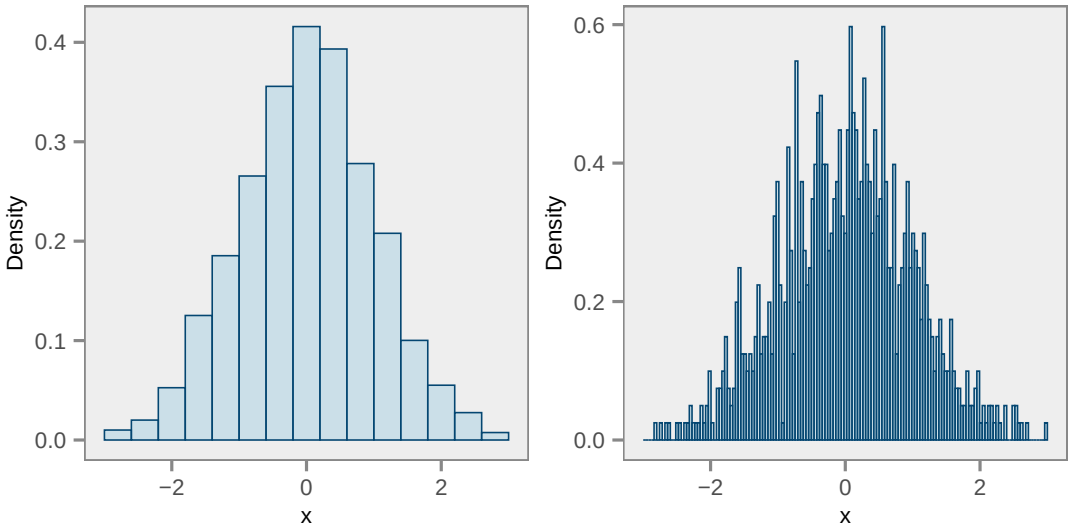

Figure 10: Histograms of the same sample generated from the standard normal distribution, but considering a different number of bins for each one.

This contrast seems to be just an arbitrary illusion, since correctly assuming a normal model for the data forces the likelihood to be unimodal (and for large n the posterior behaves similarly). However, if we consider the filtering model with a normal distribution for the main component, the main component is able to accommodate any specific subset of the observations and then delegate the rest to the alternative component. So, if at some point of the estimation, the chain specifically allocates to the main component a subset of the sample that resembles an isolated mode, it can consistently reject all other observations and get stuck in a local optimum.

To evaluate this effect, we considered multiple simulation studies fitting the filtering model with mixtures of multivariate normal distributions for the main component, but for brevity we omit most of these experiments. From our simulations, we verified that for the univariate normal distribution this effect can have significant impact depending on the initialization, and in section 3.2.2 we showcase one instance of this phenomenon while suggesting corresponding precautions. However, it is interesting to notice that this effect seems to vanish for d-variate normal distributions considering $d \geq 2$. We suspect that, with the increase in dimension, it becomes harder to find isolated subsets capable of trapping the chain.

Next, we call attention to the fact that $\mu(S_i)$ has a dependency on θ. So a natural interest would be determining what is the approximate effect of this dependence on parameter estimation. To advance with this goal, we consider a minimally complex model for which we can compare this influence analytically. Unsurprisingly, one of the simplest nontrivial cases occurs when choosing a normal distribution for the main component of the filtering model. So, assuming the corresponding conjugate prior for the mean μ and the precision τ, we attain the following hierarchical model

$$\begin{aligned} Y_i|\mu,\tau,z_i=1 &\overset{ind}{\sim} Normal\left(\mu,\tau^{-1}\right),\\ Y_i|\mu,\tau,z_{-i},z_i=0 &\overset{ind}{\sim} Uniform(S_i),\\ \mu|\tau &\sim Normal\left(\beta,\lambda^{-1}\tau^{-1}\right),\\ \tau &\sim Gamma(a,b),\\ z_i &\overset{ind}{\sim} Bernoulli(w_z),\\ \mu_L(S_i)^{-1} &= (2\pi)^{-\frac{1}{2}}\tau^{\frac{1}{2}}\exp\left\{-\frac{1}{2}F_G^{-1}\left(\gamma^{g_i(n_1+1)^{-1}}\right)\right\}, \end{aligned} \tag{3.46}$$

where $G\sim Gamma\left(\frac{1}{2},\frac{1}{2}\right)$. For the model above, it can be shown that the full conditional of (μ,τ) is given by

$$\begin{aligned} \mu|\tau,z,y &\sim Normal\left(\bar{\beta},\bar{\lambda}^{-1}\tau^{-1}\right),\\ \tau|z,y &\sim Gamma\left(\bar{a},\bar{b}\right), \end{aligned} \tag{3.47}$$

where $\bar{\lambda}=\lambda+n_1$, $\bar{\beta}=\bar{\lambda}^{-1}\left(\lambda\beta+\sum_{i=1}^n z_i y_i\right)$, $\bar{a}=a+n$, $\bar{b}=b+\sum_{i=1}^n z_i y_i^2+\lambda\beta^2-\bar{\lambda}\bar{\beta}^2$ and $n_k=\sum_{i=1}^n \mathbb{I}_{\{z_i=k\}}$. For more details, see sections A.4 of Appendix A and B.2 of Appendix B. Now, to better understand the effect of this dependency, we can compare the distribution above with what would have been obtained if $\mu(S_i)$ did not depend on θ, i.e., we can compare $\pi(\mu,\tau|z,y)$ with $\pi(\mu,\tau|y^1)$, that is given by

$$\begin{aligned} \mu|\tau,y^1 &\sim Normal\left(\bar{\beta},\bar{\lambda}^{-1}\tau^{-1}\right),\\ \tau|y^1 &\sim Gamma\left(\bar{a}^1,\bar{b}\right), \end{aligned} \tag{3.48}$$

where $\bar{a}^1=a+n_1$ is the only change. Having both expressions at hand, we notice that the location parameter μ is not affected, so we only need to compare the changes for τ. For reference, if $X\sim Gamma(a,b)$, we know that

$$\mathbb{E}[X]=\frac{a}{b} \text{ and } Var(X)=\frac{a}{b^2}, \tag{3.49}$$

so we can compare the expected value and the variance of these models considering that

$$\mathbb{E}[\tau|z,y] = \frac{\bar{a}}{\bar{b}} = \frac{a+n}{\bar{b}} = \frac{a+n_1+n_0}{\bar{b}} = \frac{a+n_1}{\bar{b}} + \frac{n_0}{\bar{b}} = \mathbb{E}[\tau|y^1] + \underbrace{\frac{n_0}{\bar{b}}}_{\geq 0} \tag{3.50}$$

$$\geq \mathbb{E}[\tau|y^1]$$

and

$$Var(\tau|z,y) = \frac{\bar{a}}{\bar{b}} = \frac{a+n}{\bar{b}^2} = \frac{a+n_1+n_0}{\bar{b}^2} = \frac{a+n_1}{\bar{b}^0} + \frac{n_0}{\bar{b}^2} = Var(\tau|y^1) + \underbrace{\frac{n_0}{\bar{b}^2}}_{\geq 0}$$

$$\geq Var(\tau|y^1). \tag{3.51}$$

From the inequalities above, it becomes clear that the filtering model increases the expected value of the precision linearly with n_0 and its standard deviation grows with $\sqrt{n_0}$. One possible interpretation of this phenomenon is that the model uses the observations attributed to the alternative component to implicitly infer which regions of the sample space are unlikely. Consequently, it augments the expected precision of our main component to indicate this gain in information, while also inflating the precision's variance to account for the somewhat vague nature of this gain. It is yet unclear if this interpretation does hold and whether the alternative component exerts other types of influence in the estimation or not. However, since for all applications in Chapter 4 we consider main components with normal response, no further investigation will be conducted regarding this matter in the present work.

3.2.2 Known Issues

The first problem we consider here is a particular consequence of our choice for the alternative component. During the developments of section 3.1, we implicitly assume that $\theta^{(t)}$, the value sampled at the t-th step of the Markov chain, is close to the "correct" value of θ. However, the model also assumes the presence of atypical observations in our sample, that is, samples that were not generated by the distribution of the main component. So, if we assume as an initial condition for z that $z_i^{(0)} = 1$ for all $i \in \{1, \ldots, n\}$, we can expect to sample $\theta^{(1)}$ with a bias. The

question then becomes: can this bias prevent us from properly estimate θ? And unfortunately the answer is yes, but this effect can be mitigated. To see that this bias might be a problem, the following experiment presents one extreme instance of this phenomenon.

Suppose we wish to fit the model of equation 3.46 to the simulated data set 4, from Figure 11. The first 200 observations were generated from a $Normal(0, 100^{-1})$ and the other 100 observations were generated from a $Uniform(-2, 2)$.

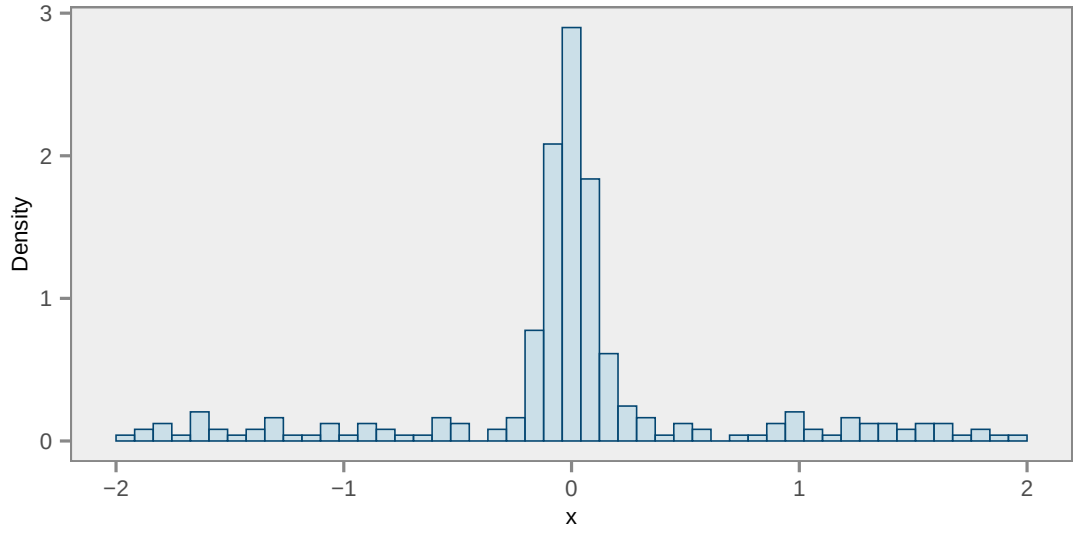

Figure 11: Histograms of the simulated data set 4.

Here, if $X_1 \sim Normal(0, 100^{-1})$, $X_2 \sim Uniform(-2, 2)$ and X_3 is simulated from the mixture of these distributions, then we have

$$\begin{aligned} Var(X_3) &= \frac{200}{200+100} Var(X_1) + \frac{100}{200+100} Var(X_2) \\ &= \frac{2}{3}\left(\frac{1}{100}\right) + \frac{1}{3}\left(\frac{(2-(-2))^2}{12}\right) \approx 0.451. \end{aligned} \tag{3.52}$$

If we assume as an initial condition of the Markov chain that $w_z^{(0)} = \frac{1}{2}$, $\mu^{(0)} = 0$, $\tau^{(0)} = \frac{1}{Var(X_3)}$ and $z_i^{(0)} = 1$, for all $i \in \{1, \ldots, n\}$, the value of γ necessary to have 50% of probability of allocating an observation at $y_i = 2$ to the alternative component is $\gamma = 0.4180$. And, for $\gamma = 0.95$, this probability drops to 6.73%, so the algorithm may have difficulties estimating μ and τ under these conditions.

Having that in mind, we assume as an experiment that the value of γ that represents our subjective prior information is given by $\gamma = 0.95$, then we analyse how

different techniques and initial conditions lead to different results. Here, our goal is to determine how to best estimate μ and τ in the presence of atypical observations that may bias estimation, but without needing to alter our subjective choice for γ. Even though we could use γ as a fine-tunable quantity, we motivate the treatment given here in section 3.3. It is worth pointing out that, for this particular model, we either need to specify the initial values of the chain for μ and τ or for z. However, since we are interested in consistent estimation in the general case, we focus our attention on the choice of $z^{(0)}$. The techniques compared in this work are presented as follows.

1. We consider the naive approach of choosing $z_i^{(0)} = 1$, for all $i \in \{1, \ldots, n\}$ and call it the *default* method.

2. We consider the alleged uninformative initial condition $z_i^{(0)} = 0$, for all $i \in \{1, \ldots, n\}$, and refer to it as the *null* method. This would in theory allow the chain to slowly converge to the "correct" distribution because the proportion of anomalies is smaller than $\frac{1}{2}$, however, there are two main issues with this approach. First, we already established in section 3.2.1 that the posterior distribution can present multimodality for this case, so the chain might get stuck in a local optimum. And, reinforcing this phenomenon, we know that the expected observational precision grows with $n_0 = \sum_{i=1}^{n}(1 - z_i)$, drastically decreasing the probability of a transition between modes. Figure 12 shows the obtained result.

3. We consider a deterministic sequence $\{\gamma_t\}_{t \in \mathbb{N}}$ and use γ_t for the step t of the chain, and we call it the *sequence* method. Here, we choose the sequence such that: for all $t < T^*$ we have $0 < \gamma_t < \gamma^*$ and for all $t \geq T^*$ we have $\gamma_t = \gamma^*$, where $\gamma^* \in (0, 1)$ and T^* is a chosen iteration. The main idea is to temporarily increase the probability of allocating an observation to the alternative component in order the remove atypical observations that may inflate the sampled observational variance of the model. After a certain number of iterations T^*, assuming that all such observations were in fact removed, the sequence turns into a constant value γ^* that correspond to the desired choice of prior.

4. We consider starting the algorithm with only a subset of the sample and progressively including the other observations using the prior-posterior update concept from the Bayes' Theorem, and we call it the *slow* method. So, we select k random observations from our sample and choose a velocity v measured in observations per iteration. Then for the k initial observations we set $z_i^{(0)} = 1$ and consider the rest of them as unobserved (so their indicators is neither 1 nor 0). After initializing, for each iteration t we include new samples to the posterior such that the current number of observation is $n(t) = min\{n, \lfloor k + vt \rfloor\}$ and initialize them with their predicted value using the sampled $\theta^{(t-1)}$. This idea comes from the simulated annealing algorithm, but instead of using a notion of temperature, we use the information from the sample. For large sample sizes we expect the posterior density to concentrate on its modes, reducing the probability of transition from one mode to another, so we start with a small sample size to mitigate this effect.

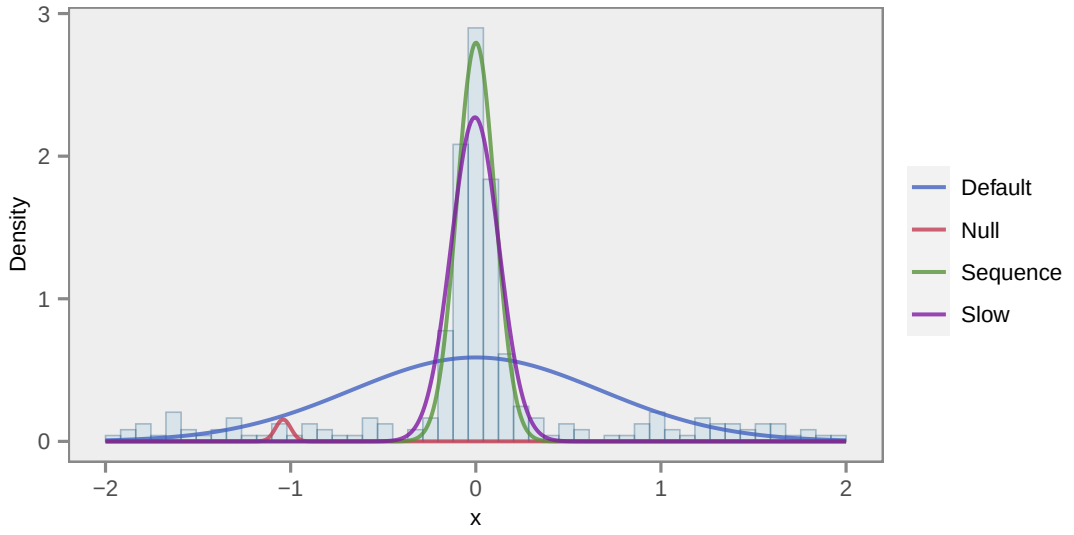

Figure 12: Histogram of the simulated data set 4 with the estimated density for each of the considered methods.

Considering the methods previously described, we simulated 20000 chain iterations for each and discarded the first 2000 as burn in. For the *sequence* method we took $\gamma_t = 0.25$, if $t < 1500$ and $\gamma_t = 0.95$ otherwise. For the *slow* method we took $k = 20$ and $v = 2$. To ensure that all results are comparable, we use the same prior

hyperparameters $w_i = \frac{1}{2}$, $\theta = -1$, $\lambda = 1$, $a = \frac{1}{4}$ and $b = \frac{1}{2}$ for all of them.

Figure 12 shows the density for a normal distribution with the estimated parameter values, multiplied by the posterior proportion of observation allocated to the main component $\frac{n_1}{n}$, for each one of them. As we can seen, the *default* method tries to allocate all of the observations to the main component, consequently obtaining a poor estimate of τ. The *null* method gets stuck in local optima, only allocating few observations to the main component and leading to the highly biased estimates of both μ and τ. It is interesting to notice that, since for the *null* method the initialization is $z_i^{(0)} = 0$, for all $i \in \{1, \ldots, n\}$, the sampled value of $\mu^{(1)}$ is centered on the prior expected value $\theta = -1$, so the chain stays close to -1. Considering now the *sequence* and the *slow* methods, we can see that both of them improve our estimates for τ, although neither of them were able to consistently recuperate the true value of the parameter (for this specific run of the algorithm the *sequence* method performed well). This can be explained if notice that both estimates for the proportion of typical observations $\frac{n_1}{n}$ are significantly greater than the true value of 0.667, i.e. the algorithm used data from the (now truncated) uniform distribution to estimate μ and τ. So, since the distribution of the points allocated to the main component is approximately a mixture of a truncated uniform and a normal distribution, this results in a smaller estimated precision. Unfortunately, in this case, since there is no discerning characteristic between typical observations and the remaining anomalies, the only way correctly estimate μ and τ would be to intentionally classify a random subset of all observations as atypical at each iteration, but in such a way that the remaining sample points have on average the correct normal distribution. However, due to the complexity of such a procedure, we'll address this possible solution in future work.

Other important issue of the filtering model is its computational cost. First, the autotransformation T_{Y_i} is a nontrivial transformation of the original random variable Y_i, so calculating values for $F_{T_{Y_i}}$ and $F^{-1}_{T_{Y_i}}$ can be expensive. Having that in mind, for every step of the chain we need to compute quantiles of T_{Y_i} for each sample point and, contributing to the problem, we also do not have a direct way of calculating the values of the correction functions. The lack of a closed form leads to

Method	μ	τ	$\frac{n_1}{n}$	Time (s)
True Values	0	100	0.667	–
Default	-0.00004 (-0.078, 0.078)	2.19 (1.84, 2.57)	0.997	65.7
Null	-1.04052 (-1.096, -0.916)	579.87 (488.36, 677.80)	0.016	37.34
Slow	-0.00433 (-0.023, 0.014)	60.08 (43.92, 76.76)	0.735	39.48
Sequence	0.00106 (-0.013, 0.016)	94.82 (77.10, 114.61)	0.720	48.96

Table 2: Summary of the parameter estimation for the simulated data set 4; in parenthesis we have a 95% credibility sets.

the recursion formula from Definition 3.5, that greatly adds on the number of times $F_{T_{Y_i}}$ and $F^{-1}_{T_{Y_i}}$ need to be evaluated. In order to mitigate this cost, we propose the *universal correction function* as an alternative to using the correction function from Definition 3.5. However, since proposition A.6 allows us to bypass the necessity of computing the correction function for all of the application of Chapter 4, we present this function in section A.2 of Appendix A.

Another point worth to be considered is that, since we need to estimate n quantities $z_1, \ldots, z_n$, the computational cost grows at least linearly with n. As an alternative, we can choose to update a random subset of size m of the indicators at each iteration, allowing us to reduce the cost of sampling all of them at every step. We can see that this sampler preserves the stationary distribution by showing that increasing the probability of remaining in the same state does not alter the detailed balance equations, but for brevity we omit the proof. It is important noticing that, as expected, this modification of the sampler comes with the self-evident cost of increasing the correlation between sampled states, effectively increasing the number of step of the chain necessary to obtain an approximately independent sample. So, it is still unclear whether the computational cost per independent observation is reduced, but we leave this question to be properly addressed in future work.

3.3 ABOUT THE FILTERING MODEL

In this section we discuss in more detail the impact of the choices for the distribution of the main component, the joint prior for θ and z and the hyperparameter w_z on the filtering model's capacity for anomaly detection. For the hyperparameter γ, we consider both objective and subjective specification, perform a sensitivity analysis to determine a good range of default values, and discuss its interpretation. Next, since by the construction in section 3.1 our model is only partially specified, we present two alternative options for prediction in this context. At last, in this section we also discuss how to classify each data point as typical or atypical from the sampled posterior values.

3.3.1 Main Component and Prior Specification

Considering the model's construction in section 3.1.3 and the notation of section 3.2, we know that an observation such that $\hat{p}_i \approx 0$ indicates a low value of the posterior predictive distribution for that sample point, at least when compared to what was set according to γ and w_z. Having that in mind, the interpretation we derive from the indicator variables is tightly connected to how we interpret what low values for the posterior predictive distribution mean, and just declaring observations with $\hat{p}_i \approx 0$ as atypical is a somewhat incomplete statement. In this case, we can more accurately affirm that these observations have a high probability of not having been generated from the *fitted model*, and, for observations such that $\hat{p}_i \approx 1$, the *fitted model* could have generated them with a high probability.

Knowing the interpretation of $\hat{p}_i$ and that the parametric family of distributions $F_{Y_i}(\cdot|\theta)$ determines what models could possibly be fitted, the importance of the choice of the distribution of the main component becomes clearer. Choosing an inflexible family of distributions leads to an overall poor fit of the data and, consequently, the model compensates this by removing the poorly explained observations. Conversely, if we choose an over-parameterized family, our model will be able to explain any observed sample, so no observation will be removed. From this,

we can see that our precision to detect anomalies is proportional to how accurately the chosen parametric family represents the typical behavior of the data. So, the filtering model heavily relies on $F_{Y_i}(\cdot|\theta)$, which is unsurprising when considering that this is a model-based approach to anomaly detection.

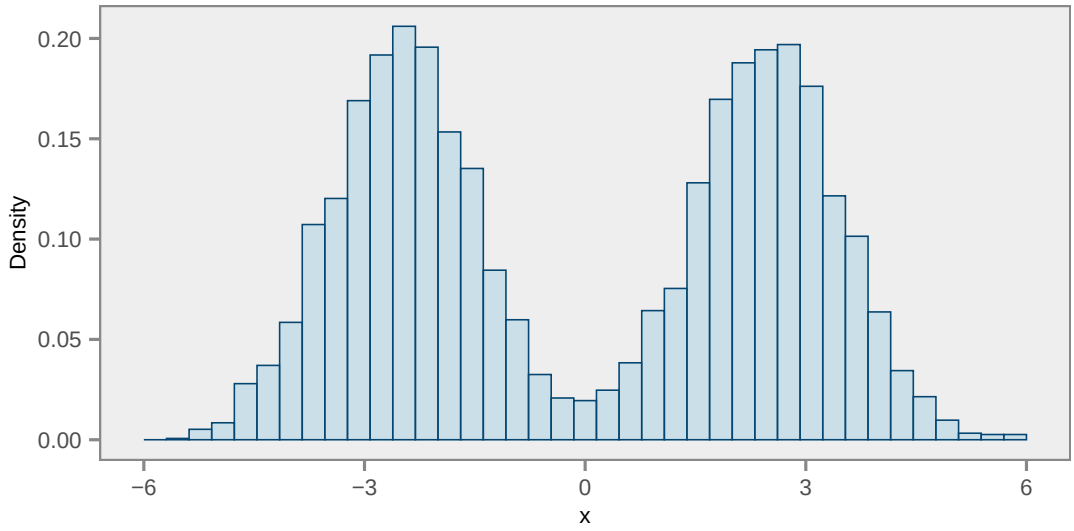

Figure 13: Histogram of a mixture of normal distributions.

Another important matter to consider when choosing $F_{Y_i}(\cdot|\theta)$ is a possible problem of ambiguity. For instance, let us consider fitting the filtering model to the sample from Figure 13, that was simulated from a mixture of two normal distributions. If we then choose as our main component a normal distribution, then by construction it is intrinsically unclear which mode should be treated as the typical behavior of the data, so the filtering model will most likely choose one the modes depending on the initial value of the chain. Here, since distinct subsets of the data could be explained by $F_{Y_i}(\cdot|\theta)$ considering different values of θ, this explains the multimodal behavior of the posterior, which leads to a complex estimation process with multiple reasonable answers. Hence, we can avoid this issue by either imposing restrictions to the family of distributions $F_{Y_i}(\cdot|\theta)$ or informing the model what are reasonable fits to the data through the use of an informative prior distribution $\pi(\theta, z)$.

At last, we call attention the fact that the filtering model assumes that the observations $Y_1, \ldots, Y_n$ all satisfy the filtering condition given θ, so the choice of

$F_{Y_i}(\cdot|\theta)$ must take this into consideration.

For the choice of the prior $\pi(\theta, z)$, a simple assumption we can make is independence between θ and z, since in principle the estimation of θ and z are distinct processes, each having its own goal. However, it is not hard to imagine scenarios where a dependence between them might be useful. As an example, we can imagine that θ somehow quantifies a notion of distance between observations, so a prior dependence between θ and z might represent the knowledge that close observations are likely to have the same classification. Nevertheless, since we wish to discuss next the effect of the prior of θ and of z separately, we assume independence for the rest of this section.

We already established that $\pi(\theta)$ can have an important role in reducing the effect of the multimodality of the posterior depending on the choice $F_{Y_i}(\cdot|\theta)$. But another aspect worth considering when choosing $\pi(\theta)$ is that, if the prior is highly informative, it may lead to a competition between the information from the prior and from the likelihood. This usually is unproblematic during estimation, however, for the filtering model, since by construction the prior is assumed to be a correct representation of our uncertainties but the observations can be ignored, in the case of a conflict the prior will always prevail. Having that in mind, we opt to use uninformative priors whenever possible in the applications of Chapter 4. As a side note, even though we recommend using uninformative priors, it is important to notice that improper priors always result in improper posteriors if $\pi(z)$ is such that

$$\mathbb{P}(z_1 = 0, \ldots, z_n = 0) > 0, \tag{3.53}$$

so caution is advisable.

Considering the choice of $\pi(z)$, we already introduced the concept of distance to add some dependence on θ, but we could also consider using the index distance between two indicators z_i and z_j to convey a similar effect. One natural example is to use the sequential nature of time series, but we can think of more exotic examples of prior dependence structure for this purpose, as we show in the application of section 4.3. It is worth noting, that even if we assume $\pi(z) = \prod_{i=1}^{n} \pi(z_i)$, the posterior may not satisfy $\pi(z|y) = \prod_{i=1}^{n} \pi(z_i|y)$. This implies that part of the dependence

between indicators is already captured by the posterior distribution, so we should only introduce dependence on the prior of z if we wish to reinforce it or if we know of a particular effect that the posterior does not consider.

Another important choice of $\pi(z)$, which is recurrently assumed throughout this work, is taking

$$z_i \overset{ind}{\sim} Bernoulli(w_z), \tag{3.54}$$

for all $i \in \{1, \ldots, n\}$, where w_z is the prior probability of having an anomaly in the data. This is a natural choice whenever we wish to assume that all observation can be anomalies and have no good reason to introduce prior dependence. Here, it is important to highlight that we do not recommend estimating w_z, since the alternative component already implicitly considers the total proportion of anomalies in the sample. So estimating w_z leads to a undesirable "double counting" of this information and consequently disturbs the estimation of z.

The last case we consider here is choosing $\pi(z)$ such that

$$\mathbb{P}(z_i = c_i) = 1, \tag{3.55}$$

for all $i \in S \subset \{1, \ldots, n\}$, where $c_i \in \{0, 1\}$. In this case, we fixate the value of some of the indicator variables, and this can be done for a variety of reasons. We can fixate a subset of the sample in order to remove the ambiguity generated by $F_{Y_i}(\cdot|\theta)$. For instance, considering once more the sample for which a histogram is presented in Figure 13, if we fixate the indicators of some of the observations from the left mode as 1, then we indicate to the filtering model that this mode represents the typical behavior of the data. Another case for which we could fixate the indicator variables is in the context of supervised learning. If we already know beforehand which observations are atypical we can only concern ourselves with estimating θ and then use the fitted model to classify future observations. It is also worth mentioning the particular case of taking

$$\mathbb{P}(z_i = 1) = 1, \tag{3.56}$$

for all $i \in \{1, \ldots, n\}$. Here, we reduce to the case of estimating θ completely disregarding the alternative component of the filtering model, which is equivalent

to just estimating the parametric model of the main component with prior $\pi(\theta)$. As a corollary, we can consider any previously fitted parametric model and convert it into an anomaly detection model by assuming $z_i = 1$ for all $i \in \{1, \ldots, n\}$ and only estimating the indicator z^* of a newly observed y^*.

3.3.2 The Threshold of Maximum Uncertainty

Before discussing the choice of the hyparameter γ, let us consider the following. Going back to the filtering model's construction in section 3.1, we made a comparison between the "classification" made when we sample z_k at every step of the Markov chain and hypothesis testing. Considering this analogy, γ performs a similar role to the level of the credibility region associated with the test (section 6.4 of Migon et al. (2014) explains the relationship between hypothesis testing and credibility regions). However, it is important to notice that when we sample z_k we do not take a deterministic decision based on a critical region of the test because the "decision" made is random. So, if we once more analyze the probability

$$\mathbb{P}\left(z_k \middle| \theta, z_{-k}, y\right) = \frac{a_k}{a_k + b_k}, \tag{3.57}$$

where

$$a_k = w_k f_{Y_k}(y_k|\theta) \text{ and}$$

$$b_k = (1 - w_k) F^{-1}_{T_{Y_k}}\left(1 - \gamma^{g_k(n_1^{-k}+1)^{-1}} \middle| \theta\right) \prod_{i \neq k} \left[\frac{F^{-1}_{T_{Y_i}}\left(1 - \gamma^{g_k(n_1^{-k}+1)^{-1}} \middle| \theta\right)}{F^{-1}_{T_{Y_i}}\left(1 - \gamma^{g_k(n_1^{-k}+2)^{-1}} \middle| \theta\right)}\right]^{1-z_k}, \tag{3.58}$$

we can more accurately understand the role of γ as determining for what values of y_k we have a probability of exactly $\frac{1}{2}$ of accepting an observation. For this reason, since γ determines how much evidence we need to have to maximize the uncertainty in our "decision making", we name it the *threshold of maximum uncertainty*.

With this, since we consider γ a hyperparameter from the prior, the choice of γ should reflect how much evidence is necessary to maximize the subjective uncertainty of the researcher in question. This interpretation allows us to consider anomaly detection in the context of unsupervised learning, nevertheless, we must keep in

mind the issues discussed in subsection 3.2.2. So even though unsupervised, we must still consider robust estimation techniques allied with thorough model validation to ensure reliable inference making.

It is worth pointing out that it may be troublesome justifying a purely subjective choice of γ for some practical applications, specially considering scenarios where the resulting anomaly classification may lead to highly impactful decision. Knowing this, we could consider objectively "estimating" the value of γ in the context of supervised learning, using the common cross-validation or training-validation-test split techniques. However, since in many scenarios we may need to deal with anomaly detection within an unsupervised learning context, it may be impractical to consider either a completely subjective or objective approach to specify γ. So a simple heuristic, e.g. default range of values to choose from, is often the most useful criteria for hyperparameter selection. Having this in mind, we next present a sensitivity analysis on γ to determine its effect across a variety of scenarios.

In our sensitivity analysis, we considered multiple experiments involving simulated data sets to assess how distinct factor affect parameter estimation. For simplicity sake, our data sets consists of n observation generated from a standard normal distribution, of which the last np_{out} are substituted by an anomalous fixed value y_{out}. We then choose the hyperparameter γ and fit our model twice: the first time we consider the model's mean μ and precision τ known and equal to the true values, and then we readjust the model to include the estimation of μ and τ with an uninformative prior. We perform these experiments considering $n \in \{200, 600, 2000\}$, $y_{out} \in \{5, 6, 7\}$, $p_{out} \in \left\{\frac{1}{20}, \frac{1}{10}, \frac{1}{5}\right\}$ and $\gamma \in \{0.01, 0.05, 0.25, 0.75, 0.95\}$, for a total of $3 \times 3 \times 3 \times 5 \times 2 = 270$ simulations. For each experiment, we fitted the filtering model using a particular case of the MCMC algorithm described in section B.2 of Appendix B with 20000 steps of the chain and a burn in of 2000 iterations, and used as the initial state of the chain $z_i^{(0)} = 1$, for all $i \in \{1, \ldots, n\}$. For reference, the total time necessary to run all experiments was approximately 1 hour, 44 minutes and 16 seconds, and we next present some of the main results.

We begin our analysis by considering the the effect of each variable on the total

time of execution of the MCMC algorithm. Unsurprising, the sample size n had the greatest effect on the total time, taking an average of approximately 43.52 seconds for $n = 2000$, 16.70 seconds for $n = 600$, and 9.28 seconds for $n = 200$. Another interesting comparison is of the time it took to fit the model when μ and τ were considered known versus when they were estimated. On average, estimating the mean and precision lead to a 10% increase in the total estimation time, indicating that most of the computational cost came from estimating the indicators variables.

Next, we analyze how well the model was able to detect anomalies for different values of γ when considering μ and τ known, and data with a proportion of anomalies $p_{out} = 0.2$. From the results presented on Table 3, we can see that, as expected, a higher value of γ leads to a more conservative detection of anomalies and a better estimation of the proportion of anomalies. However, it is interesting to notice that, even for lower values of γ, such as 0.01, 0.05 and 0.25, we still obtain reasonable results.

	$\gamma = 0.01$	$\gamma = 0.05$	$\gamma = 0.25$	$\gamma = 0.75$	$\gamma = 0.95$
median($\hat{p}_{out}$)	0.2561	0.2397	0.2211	0.2054	0.2011
min($\hat{p}_{out}$)	0.2214	0.2149	0.2077	0.2019	0.2004
max($\hat{p}_{out}$)	0.3217	0.2873	0.2465	0.2117	0.2024

Table 3: Summary statistics for the estimated proportion of anomalies for different values of γ when considering μ and τ known, and for $p_{out} = 0.2$.

In section 3.2.2, we established that the presence of atypical observations can bias the estimates of the model's parameters, so naturally, the next step is to determine how the values of γ interfere in the estimation of μ and τ. Tables 4 and 5 present the estimated proportion of typical observations, and values of μ and τ for all scenarios considering $\gamma = 0.05$ and $\gamma = 0.95$ respectively. And from these tables, we can observe strong biases regarding the estimates of μ and τ that require some explaining, and, to avoid an entry-by-entry analysis, we group all rows related to a common bias mechanism to address each group individually.

Firstly, consider all rows of Table 4 such that $n = 200$. In this case, we have

$(n, y_{out}, 1-p_{out})$	p	μ	$IC_{0.95}(\mu)$	τ	$IC_{0.95}(\tau)$
$(200, 5, 0.95)$	$5.97e$-3	$4.72e$-3	$(-1.35e$-1$, 1.27e$-1$)$	$1.99e4$	$(1.62e4, 2.39e4)$
$(200, 5, 0.9)$	$5.74e$-3	$3.20e$-3	$(-1.24e$-1$, 1.28e$-1$)$	$1.99e4$	$(1.62e4, 2.39e4)$
$(200, 5, 0.8)$	$4.72e$-3	$2.49e$-4	$(-1.25e$-1$, 1.29e$-1$)$	$1.99e4$	$(1.62e4, 2.40e4)$
$(200, 6, 0.95)$	$5.97e$-3	$4.72e$-3	$(-1.35e$-1$, 1.27e$-1$)$	$1.99e4$	$(1.62e4, 2.39e4)$
$(200, 6, 0.9)$	$5.74e$-3	$3.20e$-3	$(-1.24e$-1$, 1.28e$-1$)$	$1.99e4$	$(1.62e4, 2.39e4)$
$(200, 6, 0.8)$	$4.72e$-3	$2.49e$-4	$(-1.25e$-1$, 1.29e$-1$)$	$1.99e4$	$(1.62e4, 2.40e4)$
$(200, 7, 0.95)$	$5.97e$-3	$4.72e$-3	$(-1.35e$-1$, 1.27e$-1$)$	$1.99e4$	$(1.62e4, 2.39e4)$
$(200, 7, 0.9)$	$5.74e$-3	$3.20e$-3	$(-1.24e$-1$, 1.28e$-1$)$	$1.99e4$	$(1.62e4, 2.39e4)$
$(200, 7, 0.8)$	$4.72e$-3	$2.49e$-4	$(-1.25e$-1$, 1.29e$-1$)$	$1.99e4$	$(1.62e4, 2.40e4)$
$(600, 5, 0.95)$	0.892	0.0253	$(-0.052, 0.104)$	1.409	$(1.222, 1.620)$
$(600, 5, 0.9)$	0.836	0.0242	$(-0.054, 0.104)$	1.558	$(1.341, 1.807)$
$(600, 5, 0.8)$	0.955	0.942	$(0.762, 1.121)$	0.234	$(0.206, 0.265)$
$(600, 6, 0.95)$	0.892	0.0253	$(-0.052, 0.104)$	1.409	$(1.222, 1.620)$
$(600, 6, 0.9)$	0.836	0.0242	$(-0.054, 0.104)$	1.558	$(1.341, 1.807)$
$(600, 6, 0.8)$	0.952	1.104	$(0.891, 1.315)$	0.174	$(0.152, 0.198)$
$(600, 7, 0.95)$	0.892	0.0253	$(-0.052, 0.104)$	1.409	$(1.222, 1.620)$
$(600, 7, 0.9)$	0.836	0.0242	$(-0.054, 0.104)$	1.558	$(1.341, 1.807)$
$(600, 7, 0.8)$	0.810	0.5357	$(-0.066, 1.460)$	1.251	$(0.120, 2.452)$
$(2000, 5, 0.95)$	0.931	0.0228	$(-0.021, 0.066)$	1.145	$(1.069, 1.225)$
$(2000, 5, 0.9)$	0.879	0.0190	$(-0.025, 0.063)$	1.238	$(1.155, 1.327)$
$(2000, 5, 0.8)$	0.986	0.996	$(0.898, 1.093)$	0.217	$(0.204, 0.231)$
$(2000, 6, 0.95)$	0.931	0.0228	$(-0.021, 0.066)$	1.146	$(1.070, 1.225)$
$(2000, 6, 0.9)$	0.879	0.0190	$(-0.025, 0.063)$	1.238	$(1.155, 1.327)$
$(2000, 6, 0.8)$	0.986	1.187	$(1.073, 1.300)$	0.159	$(0.149, 0.169)$
$(2000, 7, 0.95)$	0.931	0.0228	$(-0.021, 0.066)$	1.146	$(1.070, 1.225)$
$(2000, 7, 0.9)$	0.879	0.0190	$(-0.025, 0.063)$	1.238	$(1.155, 1.327)$
$(2000, 7, 0.8)$	0.985	1.379	$(1.247, 1.509)$	0.121	$(0.113, 0.129)$

Table 4: Estimated parameters proportion of typical observations p and parameters μ and τ for $\gamma = 0.05$.

$(n, y_{out}, 1-p_{out})$	p	μ	$IC_{0.95}(\mu)$	τ	$IC_{0.95}(\tau)$
$(200, 5, 0.95)$	0.944	-0.00363	$(-0.134, 0.127)$	1.250	$(1.004, 1.523)$
$(200, 5, 0.9)$	0.995	0.488	$(0.241, 0.731)$	0.338	$(0.272, 0.412)$
$(200, 5, 0.8)$	0.997	0.983	$(0.683, 1.286)$	0.213	$(0.174, 0.258)$
$(200, 6, 0.95)$	0.944	-0.00378	$(-0.135, 0.127)$	1.251	$(1.007, 1.523)$
$(200, 6, 0.9)$	0.895	0.00191	$(-0.137, 0.145)$	1.293	$(1.018, 1.591)$
$(200, 6, 0.8)$	0.997	1.181	$(0.829, 1.537)$	0.155	$(0.127, 0.188)$
$(200, 7, 0.95)$	0.944	-0.00378	$(-0.135, 0.127)$	1.251	$(1.007, 1.523)$
$(200, 7, 0.9)$	0.894	-0.00549	$(-0.137, 0.126)$	1.306	$(1.051, 1.592)$
$(200, 7, 0.8)$	0.997	1.380	$(0.975, 1.788)$	0.118	$(0.096, 0.142)$
$(600, 5, 0.95)$	0.949	0.0327	$(-0.044, 0.111)$	1.144	$(1.014, 1.279)$
$(600, 5, 0.9)$	0.998	0.523	$(0.383, 0.664)$	0.330	$(0.294, 0.369)$
$(600, 5, 0.8)$	0.999	1.017	$(0.843, 1.191)$	0.212	$(0.188, 0.236)$
$(600, 6, 0.95)$	0.949	0.0321	$(-0.045, 0.110)$	1.148	$(1.022, 1.281)$
$(600, 6, 0.9)$	0.998	0.619	$(0.458, 0.783)$	0.251	$(0.223, 0.280)$
$(600, 6, 0.8)$	0.999	1.217	$(1.013, 1.420)$	0.155	$(0.137, 0.172)$
$(600, 7, 0.95)$	0.949	0.0321	$(-0.045, 0.110)$	1.148	$(1.023, 1.281)$
$(600, 7, 0.9)$	0.997	0.715	$(0.532, 0.900)$	0.195	$(0.174, 0.219)$
$(600, 7, 0.8)$	0.999	1.416	$(1.182, 1.649)$	0.117	$(0.104, 0.130)$
$(2000, 5, 0.95)$	0.998	0.269	$(0.205, 0.331)$	0.480	$(0.450, 0.512)$
$(2000, 5, 0.9)$	0.999	0.521	$(0.442, 0.598)$	0.322	$(0.302, 0.342)$
$(2000, 5, 0.8)$	1.000	1.016	$(0.920, 1.113)$	0.211	$(0.198, 0.224)$
$(2000, 6, 0.95)$	0.995	0.299	$(0.227, 0.371)$	0.403	$(0.372, 0.437)$
$(2000, 6, 0.9)$	0.999	0.620	$(0.530, 0.709)$	0.245	$(0.230, 0.260)$
$(2000, 6, 0.8)$	1.000	1.216	$(1.103, 1.329)$	0.154	$(0.145, 0.164)$
$(2000, 7, 0.95)$	0.950	0.0268	$(-0.017, 0.070)$	1.063	$(0.999, 1.130)$
$(2000, 7, 0.9)$	0.999	0.719	$(0.617, 0.819)$	0.191	$(0.179, 0.203)$
$(2000, 7, 0.8)$	1.000	1.416	$(1.286, 1.545)$	0.117	$(0.110, 0.124)$

Table 5: Estimated parameters proportion of typical observations p and parameters μ and τ for $\gamma = 0.95$.

an estimated proportion of typical observations smaller than 0.001, which strongly resembles the estimation obtained for the *null* initialization method presented in section 3.2.2. Having this in mind, once the chain reaches a state with a low enough value of n_1, we can safely assume that the chain rejects almost all of the observations for the same reason. However, this does not explain how the chain reaches such state. To make sense of this, notice that when the value of γ is sufficiently low, even considering the true values of μ and τ, the model starts rejecting some of the typical observations with least density, as shown in Table 3. Because of this, in the case of the normal distribution, this leads to a truncation of the typical points, causing the model to underestimate the variance and, consequently, to reject even more observations until none is left.

To prevent this phenomenon, we could consider an increase in the value of γ, since it would avoid removing typical observations from the main component. However, as shown in Table 5, this introduces another type of problem. Recalling once more section 3.2.2, this resembles the problem with the *default* initialization method, leading us to the suspicion that a simple change of initialization might resolve or mitigate the problem. Nevertheless, from other numerical experiments omitted in this works, even using as an initialization the true values of μ and τ, for most of the cases the chain still converges to the undesired solution. It is worth mentioning that this issue also affected the simulations using $\gamma = 0.05$, allowing us to conclude that no single value of γ is expected to work for all instances.

So, with the considerations above, we propose as an heuristic choosing a low value of γ, e.g. 0.05, as a starting value and, if the chain ever reaches $p \approx 0$, restart the MCMC algorithm with a bigger value of γ. The main justification for this heuristic relies on the fact that, from our experiments, it is way easier to detect a problem when γ is too low compared to the cases for which γ is too high. As an important side note, Table 4 also indicates that the way the hyperparameter affects estimation varies significantly with the sample size n, so the initial chosen value of γ should consider this effect.

3.3.3 Prediction

Even though the filtering model is defined primarily with anomaly detection in sight, it is worth noting that, because we consider an approach via mixture models, it can also be used as a method for robust estimation. That is because we can interpret our estimate $\hat{\theta}$ as a convex combination of the values $\hat{\theta}(z)$ that would have been obtained for every possible value of z, while pondering the influence of each scenario by its posterior probability. This becomes clear when we look at expression of the Bayes estimator $\hat{\theta}_{Bayes}$ that minimizes the expected quadratic loss function *a posteriori*. Since this is equivalent to taking the posterior mean of θ as an estimator, it can be expressed as

$$\hat{\theta}_{Bayes} = \mathbb{E}_\theta\left[\theta|y\right] = \mathbb{E}_z\left[\mathbb{E}_\theta\left[\theta|z,y\right]|y\right] = \sum_z \pi(z|y)\mathbb{E}_\theta\left[\theta|z,y\right] = \sum_z \pi(z|y)\hat{\theta}_{Bayes}(z). \tag{3.59}$$

Notice that, if in our sample an observation y_i strongly disagrees with the model from the main component, by construction the filtering model will attribute a small weight to $\mathbb{P}(z_i = 1|y)$. So the observation y_i will have only a small contribution to $\hat{\theta}$, thus resulting in a robust estimation.

In this context, we might be interested in obtaining a predictive distribution for an unobserved Y^*, instead of using y_* to estimate z^*. So the inquiry then becomes: how to make predictions using a partially specified model?

Considering the method for parameter estimation presented in section 3.2, it is clear that we only have a sample of approximately independent values $(\theta^{(1)}, z^{(1)})$, $\ldots$, $(\theta^{(m)}, z^{(m)})$ to obtain estimates, which cannot help us estimate what the unknown sets $S_1, \ldots, S_n$ are without introducing more assumptions to the model. And noticing that one of the advantages of the filtering model is not having to assume that atypical observations come from an specific region of the support of our observations, losing this propriety by making further assumptions seems to be a step in the wrong direction. Anyway, if one does wish to make them, they must still satisfy $\mathbb{I}_{S_i}(y_i) = 1$, for all $i \in \{1, \ldots, n\}$.

The most natural alternative is to assume that every unobserved Y^* comes from

the main component, i.e., its corresponding indicator z^* is equal to 1. Even though strong, this is a common assumption implicitly made in the context of prediction, so it seems to be a reasonable solution. In this case, since the predictive posterior distribution for the new observation does not depend on the unknown region S^* (assuming conditional independence of the observations given θ), prediction follows as usual. We can see this algebraically considering that

$$\begin{aligned}
&\pi(y^*|y, z^* = 1) \\
&= \sum_{z\in\{0,1\}^n} \int_\Theta \pi(\theta, z, y^*|y, z^* = 1)\ d\theta \\
&= \sum_{z\in\{0,1\}^n} \int_\Theta \frac{\pi(\theta, z, y, y^*|z^* = 1)}{\pi(y|z^* = 1)}\ d\theta \\
&= \sum_{z\in\{0,1\}^n} \int_\Theta \frac{\pi(y, y^*|\theta, z, z^* = 1)\pi(\theta, z|z^* = 1)}{\pi(y|z^* = 1)}\ d\theta \\
&\overset{ind}{=} \sum_{z\in\{0,1\}^n} \int_\Theta \frac{\pi(y^*|\theta, z, z^* = 1)\pi(y|\theta, z, z^* = 1)\pi(\theta, z|z^* = 1)}{\pi(y|z^* = 1)}\ d\theta \\
&\overset{ind}{=} \sum_{z\in\{0,1\}^n} \int_\Theta \pi(y^*|\theta, z^* = 1)\frac{\pi(y|\theta, z)\pi(\theta, z)}{\pi(y)}\ d\theta \\
&= \sum_{z\in\{0,1\}^n} \int_\Theta f_{Y^*}(y^*|\theta)\pi(\theta, z|y)\ d\theta = \mathbb{E}_{(\theta,z)|y}\left[f_{Y^*}(y^*|\theta)\right].
\end{aligned} \tag{3.60}$$

It is important to notice that here we are implicitly assuming that our choice of $\mu_L(S_i)^{-1}$ stays the same for all $i \in \{1, \ldots, n\}$ and, consequently, does not depend on z^*. From the derived expression, we can then use our sample from the Markov chain to obtain Monte Carlo estimates considering that

$$\pi(y^*|y, z^* = 1) = \mathbb{E}_{(\theta,z)|y}\left[f_{Y^*}(y^*|\theta)\right] \approx \frac{1}{m}\sum_{j=1}^{m} f_{Y^*}(y^*|\theta^{(j)}). \tag{3.61}$$

Another approach would be to consider an estimate for the proportions of typical observations, taken as an approximation of the probability p of having a new typical observation, to determine how to inflate the desired credibility level γ to make approximate credibility regions regardless of the presence of anomalies. Having that in mind, let us take

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} \tilde{z}_i = \frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n} z_i^{(j)} \tag{3.62}$$

as an estimator for p and define an inflated credibility level

$$\gamma^* = \frac{\gamma}{\hat{p}}. \tag{3.63}$$

Then, if we choose R such that $\mathbb{P}(Y^* \in R^* | y, z^* = 1) = \gamma^*$ and assume that $\mathbb{P}(z^* = 1|y) \approx p$, we have

$$\begin{aligned}
\mathbb{P}(Y^* \in R|y) &= \mathbb{P}(Y^* \in R, z^* = 1|y) + \mathbb{P}(Y^* \in R, z^* = 0|y) \\
&= \mathbb{P}(Y^* \in R|y, z^* = 1)\mathbb{P}(z^* = 1|y) + \mathbb{P}(Y^* \in R|y, z^* = 0)\mathbb{P}(z^* = 0|y) \\
&\approx \underbrace{\mathbb{P}(Y^* \in R|y, z^* = 1)}_{=\gamma^*}\hat{p} + \underbrace{\mathbb{P}(Y^* \in R|y, z^* = 0)}_{\geq 0}(1 - \hat{p}) \\
&\geq \gamma^*\hat{p} = \frac{\gamma}{\hat{p}}\hat{p} = \gamma.
\end{aligned} \tag{3.64}$$

It is worth noting that, using this method, we can only find a region R with approximate credibility level γ if we have $\gamma \leq \hat{p}$. And even in the cases where this is possible, there is a trade off between the size of the credibility set and its coverage, so we may obtain a credibility set too large to be informative.

3.3.4 Anomaly Classification

Let us consider the problem of using the sampled chain values $(\theta^{(1)}, z^{(1)})$, ..., $(\theta^{(m)}, z^{(m)})$ to determine whether the k-th observation of the sample is typical or not. As an initial thought, notice the following: throughout the generation of the Markov chain, we already made multiple random classifications by sampling the values of the indicators z_k for each observation at every step of the chain. So, one approach to classification is to choose the estimator $\tilde{z}_k$ as the indicator of the event: z_k was classified as 1 more than or as many times as it was classified as 0. This results in the estimator of z_k given by

$$\tilde{z}_k = \begin{cases} 1, & \text{if } \bar{z}_k \geq \frac{1}{2}, \\ 0, & \text{otherwise,} \end{cases} \tag{3.65}$$

for $i \in \{1, \ldots, n\}$, where $\bar{z}_k = \frac{1}{m}\sum_{j=1}^{m} z_k^{(j)}$. However, even though $\tilde{z}_k$ is an intuitive estimator, we next consider a more formal approach.

Our posterior distribution contains all of the information we have regarding the parameters of the model. So, another idea is to use the marginal posterior distribution of z_k to determine whether $\mathbb{P}(z_k = 1|y) \geq \mathbb{P}(z_k = 0|y)$ or not. Nevertheless, we still need to find a way of using the sampled chain values to calculate these probabilities. So, consider the following:

$$\begin{aligned}
&\pi(z_k = 1|y) \\
&= \sum_{z_{-k} \in \{0,1\}^{n-1}} \int_\Theta \pi(\theta, z_k = 1, z_{-k}|y)\ d\theta \\
&= \sum_{z_{-k} \in \{0,1\}^{n-1}} \int_\Theta \pi(z_k = 1|\theta, z_{-k}, y)\pi(\theta, z_{-k}, y)\ d\theta \\
&= \mathbb{E}_{(\theta, z_{-k})|y}\left[\pi(z_k = 1|\theta, z_{-k}, y)\middle| y\right] \\
&= \mathbb{E}_{(\theta, z_{-k})|y}\left[\left(\frac{a_k}{a_k + b_k}\right)^{z_k}\left(1 - \frac{a_k}{a_k + b_k}\right)^{1-z_k}\middle| z_k = 1, y\right], \\
&= \mathbb{E}_{(\theta, z_{-k})|y}\left[\frac{a_k}{a_k + b_k}\middle| y\right],
\end{aligned} \tag{3.66}$$

where

$$\begin{aligned}
&a_k = w_k f_{Y_k}(y_k|\theta) \text{ and} \\
&b_k = (1 - w_k)F^{-1}_{T_{Y_k}}\left(1 - \gamma^{g_k(n_1^{-k}+1)^{-1}}\middle|\theta\right)\prod_{i \neq k}\left[\frac{F^{-1}_{T_{Y_i}}\left(1 - \gamma^{g_k(n_1^{-k}+1)^{-1}}\middle|\theta\right)}{F^{-1}_{T_{Y_i}}\left(1 - \gamma^{g_k(n_1^{-k}+2)^{-1}}\middle|\theta\right)}\right]^{1-z_k}.
\end{aligned} \tag{3.67}$$

Thus, if we consider the approximation

$$\begin{aligned}
\pi(z_k = 1|y) &= \mathbb{E}_{(\theta, z_{-k})|y}\left[\frac{a_k}{a_k + b_k}\middle| y\right] \\
&\approx \frac{1}{m}\sum_{j=1}^{m}\frac{a_k^{(j)}}{a_k^{(j)} + b_k^{(j)}} = \hat{p}_k,
\end{aligned} \tag{3.68}$$

where

$$
\begin{aligned}
a_k^{(j)} &= w_k^{(j)} f_{Y_k}\left(y_k \left| \theta^{(j)} \right.\right), \\
b_k^{(j)} &= (1 - w_k^{(j)}) F_{T_{Y_k}}^{-1}\left(1 - \gamma^{g_k(n_{1j}^{-k}+1)^{-1}} \middle| \theta^{(j)}\right) \\
&\quad \times \prod_{i \neq k} \left[\frac{F_{T_{Y_i}}^{-1}\left(1 - \gamma^{g_i(n_{1j}^{-k}+1)^{-1}} \middle| \theta^{(j)}\right)}{F_{T_{Y_i}}^{-1}\left(1 - \gamma^{g_i(n_{1j}^{-k}+2)^{-1}} \middle| \theta^{(j)}\right)} \right]^{1-z_i^{(j)}}, \\
n_{1j}^{-k} &= \sum_{i \neq k} z_i^{(j)}, \\
w_k^{(j)} &= \pi\left(z_k = 1 \middle| \theta^{(j)}, z_{-k}^{(j)}\right),
\end{aligned}
\tag{3.69}
$$

we can define our estimator of z_k as

$$
\hat{z}_k = \begin{cases} 1, & \text{if } \hat{p}_k \geq \frac{1}{2}, \\ 0, & \text{otherwise,} \end{cases} \tag{3.70}
$$

for $k \in \{1, \ldots, n\}$. Next, we discuss how to estimate z^*, the indicator of a newly observed sample point y^*.

The first thing to notice when trying to classify y^* is that, in our specification of the alternative component, we choose $\mu_L(S_i)^{-1}$ as a function of n_1, where $n_1 = \sum_{i=1}^n z_i$, for all $i \in \{1, \ldots, n\}$. So, since the alternative component for the i-th observation depends on z_{-i}, it is somewhat unclear how to generalise the alternative component for a new observation y^*. The most natural way is to consider the augmented sample $Y_1, \ldots, Y_n, Y_{n+1}$, where $Y_{n+1} = Y^*$, and re-estimate all of the parameters assuming that

$$
\mu_L(S_i)^{-1} = F_{T_{Y_i}}^{-1}\left(1 - \gamma^{g_i(n_1^*+1)^{-1}} \middle| \theta\right), \tag{3.71}
$$

for all $i \in \{1, \ldots, n+1\}$, where $n_1^* = \sum_{i=1}^{n+1} z_i$. Then, we can estimate $z^* = z_{n+1}$, by considering either of the estimators presented previously.

Even though the previous method for estimating z^* is methodologically consistent, it is impractical for most applications because it adds the computational cost required to re-estimate the parameters of the model. It can be specially problematic if our model is already computationally expensive or if the size of the augmented sample becomes too large. So, as an alternative, we can consider using for Y^* an

adapted alternative component given by

$$\mu_L(S^*)^{-1} = F^{-1}_{T_{Y^*}}\left(1-\gamma^{(n_1+1)^{-1}}\middle|\,\theta\right), \tag{3.72}$$

where $n_1 = \sum_{i=1}^{n} z_i$, while keeping the remaining of the alternative components the same. Then we can write our marginal posterior distribution for z^* as

$$\begin{aligned}
\pi(z^*=1|y,y^*) &= \sum_{z\in\{0,1\}^n}\int_\Theta \pi(\theta,z,z^*=1|y,y^*)\ d\theta \\
&= \sum_{z\in\{0,1\}^n}\int_\Theta \frac{\pi(\theta,z,z^*=1,y,y^*)}{\pi(y,y^*)}\ d\theta = \frac{\displaystyle\sum_{z\in\{0,1\}^n}\int_\Theta \pi(\theta,z,z^*=1,y,y^*)\ d\theta}{\displaystyle\sum_{k=0}^{1}\sum_{z\in\{0,1\}^n}\int_\Theta \pi(\theta,z,z^*=k,y,y^*)\ d\theta} \\
&= \frac{\displaystyle\sum_{z\in\{0,1\}^n}\int_\Theta \pi(y^*|\theta,z,z^*=1,y)\pi(y|\theta,z,z^*=1)\pi(z^*=1|\theta,z)\pi(\theta,z)\ d\theta}{\displaystyle\sum_{k=0}^{1}\sum_{z\in\{0,1\}^n}\int_\Theta \pi(y^*|\theta,z,z^*=k,y)\pi(y|\theta,z,z^*=k)\pi(z^*=k|\theta,z)\pi(\theta,z)\ d\theta} \\
&\overset{ind}{=} \frac{\displaystyle\sum_{z\in\{0,1\}^n}\int_\Theta \pi(y^*|\theta,z,z^*=1)\pi(y|\theta,z)\pi(z^*=1|\theta,z)\pi(\theta,z)\ d\theta}{\displaystyle\sum_{k=0}^{1}\sum_{z\in\{0,1\}^n}\int_\Theta \pi(y^*|\theta,z,z^*=k)\pi(y|\theta,z)\pi(z^*=k|\theta,z)\pi(\theta,z)\ d\theta} \\
&= \frac{\displaystyle\sum_{z\in\{0,1\}^n}\int_\Theta \pi(y^*|\theta,z,z^*=1)\pi(z^*=1|\theta,z)\pi(\theta,z|y)\pi(y)\ d\theta}{\displaystyle\sum_{k=0}^{1}\sum_{z\in\{0,1\}^n}\int_\Theta \pi(y^*|\theta,z,z^*=k)\pi(z^*=k|\theta,z)\pi(\theta,z|y)\pi(y)\ d\theta} \\
&= \frac{\displaystyle\sum_{z\in\{0,1\}^n}\int_\Theta \pi(y^*|\theta,z,z^*=1)\pi(z^*=1|\theta,z)\pi(\theta,z|y)\ d\theta}{\displaystyle\sum_{k=0}^{1}\sum_{z\in\{0,1\}^n}\int_\Theta \pi(y^*|\theta,z,z^*=k)\pi(z^*=k|\theta,z)\pi(\theta,z|y)\ d\theta} \\
&= \frac{\mathbb{E}_{(\theta,z)|y}\left[\pi(y^*|\theta,z,z^*=1)\pi(z^*=1|\theta,z)\middle|y^*\right]}{\displaystyle\sum_{k=0}^{1}\mathbb{E}_{(\theta,z)|y}\left[\pi(y^*|\theta,z,z^*=k)\pi(z^*=k|\theta,z)\middle|y^*\right]} = \frac{\mathbb{E}_{(\theta,z)|y}[a_*|y^*]}{\mathbb{E}_{(\theta,z)|y}[a_*|y^*]+\mathbb{E}_{(\theta,z)|y}[b_*|y^*]},
\end{aligned} \tag{3.73}$$

where

$$\begin{aligned}\pi(y^*|\theta,z,z^*) &= \left[f_{Y^*}(y^*|\theta)\right]^{z^*}\left[F^{-1}_{T_{Y^*}}\left(1-\gamma^{(n_1+1)^{-1}}\middle|\,\theta\right)\right],\\ a_* &= w_* f_{Y^*}(y^*|\theta),\\ b_* &= (1-w_*)F^{-1}_{T_{Y^*}}\left(1-\gamma^{(n_1+1)^{-1}}\middle|\,\theta\right),\\ w_* &= \pi(z^*=1|\theta,z).\end{aligned} \tag{3.74}$$

So, if we consider the approximation

$$\begin{aligned}\pi(z^*=1|y,y^*) &= \frac{\mathbb{E}_{(\theta,z)|y}\left[\pi(y^*|\theta,z,z^*=1)\pi(z^*=1|\theta,z)\middle|y^*\right]}{\sum_{k=0}^{1}\mathbb{E}_{(\theta,z)|y}\left[\pi(y^*|\theta,z,z^*=k)\pi(z^*=k|\theta,z)\middle|y^*\right]}.\\ &= \frac{\mathbb{E}_{(\theta,z)|y}[a_*|y^*]}{\mathbb{E}_{(\theta,z)|y}[a_*|y^*]+\mathbb{E}_{(\theta,z)|y}[b_*|y^*]} \approx \frac{\frac{1}{m}\sum_{j=1}^{m}a_*^{(j)}}{\frac{1}{m}\sum_{j=1}^{m}a_*^{(j)}+\frac{1}{m}\sum_{j=1}^{m}b_*^{(j)}} = \hat{p}^*,\end{aligned} \tag{3.75}$$

where

$$\begin{aligned}a_*^{(j)} &= w_*^{(j)} f_{Y^*}\left(y^*\left|\theta^{(j)}\right.\right),\\ b_*^{(j)} &= (1-w_*^{(j)})F^{-1}_{T_{Y^*}}\left(1-\gamma^{(n_1^{(j)}+1)^{-1}}\middle|\,\theta^{(j)}\right),\\ n_1^{(j)} &= \sum_{i=1}^{n} z_i^{(j)}\\ w_*^{(j)} &= \pi\left(z^*=1\left|\theta^{(j)},z^{(j)}\right.\right),\end{aligned} \tag{3.76}$$

then we can define our estimator of z^* as

$$\hat{z}^* = \begin{cases} 1, & \text{if } \hat{p}^* \geq \frac{1}{2},\\ 0, & \text{otherwise.}\end{cases} \tag{3.77}$$

4 APPLICATIONS

Considering the proposed *filtering model*, we chose three applications to assess the model's performance for various scenarios. In section 4.1, we return to our introductory problem of classifying whether or not a given gasoline sample was contaminated with ethanol based on its near infrared spectra. In the second application, in section 4.2, we consider a problem of unsupervised classification, where our interest lies in determining whether a given tumor is benign or malign based on measurements taken from medical image exams. Finally, in section 4.3 we aim to identify historical events based on their effects on the mortality rates for the male population of France from 1816 to 2020. We used the softwares R, by R Core Team (2020), and RStudio, by RStudio Team (2019) for all computational analysis and simulations mentioned throughout this work, and the visualizations shown were made with the aid of the package ggplot2, by Wickham (2016).

4.1 FINDING CONTAMINATION IN GASOLINE SAMPLES

In this section we consider fitting the filtering model to the Octane data set, see Esbensen et al. (2002), consisting of near infrared (NIR) absorbance spectra of $n = 39$ gasoline samples. The absorbance spectra were measured with a regular spacing between for $d = 226$ wavelengths ranging from $1102nm$ to $1552nm$. This data set is provided by CAMO Software and The UnscramblerX, and is freely available for download at `https://www.impopen.com/software/octane-data-set`.

Figure 14 shows a visualization of the data where each one of the $n = 39$ curves represent the absorbance spectra of different gasoline sample. From Figure 14, we can see that some observations detach from the rest for wavelengths above $1390nm$. These, observations 25, 26 and 36-39, are well known outliers consisting of gasoline samples that were contaminated with ethanol.

Considering now the filtering model, since the uncontaminated observations present a relatively homogeneous behavior, we initially assume for the main compo-

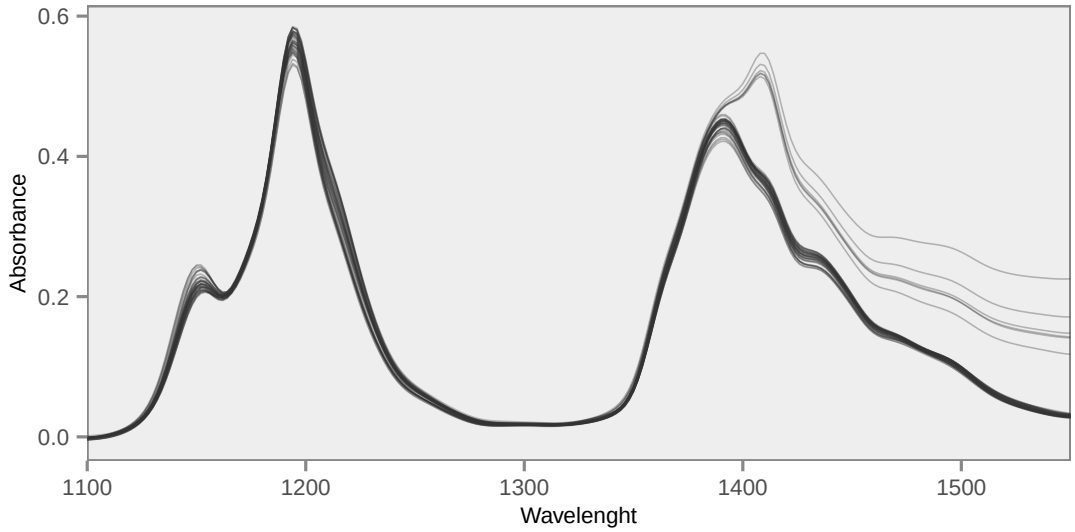

Figure 14: Absorbance curves for gasoline samples from the octane data set at different wavelengths.

nent observations with a multivariate normal distribution with the mean following a normal random walk. We use dynamic linear models (West & Harrison (1997)) to represent the main component, given by

$$\begin{aligned} Y_{ij} &= \mu_j + \nu_{ij}, \qquad \nu_{ij} \overset{ind}{\sim} Normal\left(0, \phi^{-1}\right), \\ \mu_j &= \mu_{j-1} + \omega_j, \quad \omega_{ij} \overset{ind}{\sim} Normal\left(0, \phi_\omega^{-1}\right), \end{aligned} \tag{4.1}$$

for $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, d\}$. Here, Y_{ij} represents the absorbance spectra for the j-th wavelength and i-th observation, μ_j is the common expected absorbance of for the j-th wavelength, ϕ is the precision for the observational error and ϕ_ω is the precision for the random walk. Here, it is important highlighting we assume a constant precision ϕ for all wavelengths for simplicity. For the prior of (θ, z), we assume

$$\pi(\theta, z) = \pi(\theta) \prod_{i=1}^{n} \Big[\pi(z_i)\Big] = \pi(\theta) \prod_{i=1}^{n} \Big[w_z^{z_i}(1 - w_z)^{1-z_i}\Big], \tag{4.2}$$

where $w_z \in w_z$ and the choice of the prior $\pi(\theta)$ is detailed in section B.1 of Appendix B. It is worth noticing that, since our interest is in detecting whether the gasoline sample is contaminated or not, we consider as a sample unit the 39 226-variate observations of absorbance spectra.

To fit the model, we were able to use a Gibbs sampler using the full conditionals described in section B.1 of Appendix B to generate a sample from the posterior

distribution. However, considering the main component presented in equation 4.1, regardless of initialization, choices of γ and $\pi(\theta)$, we consistently estimated the expected value of z_i to be approximately 0 for all observations. Interestingly, this is strong and clear evidence that the chosen main component underfits the typical data, so adjustments are required in order to obtain non-degenerate estimates.

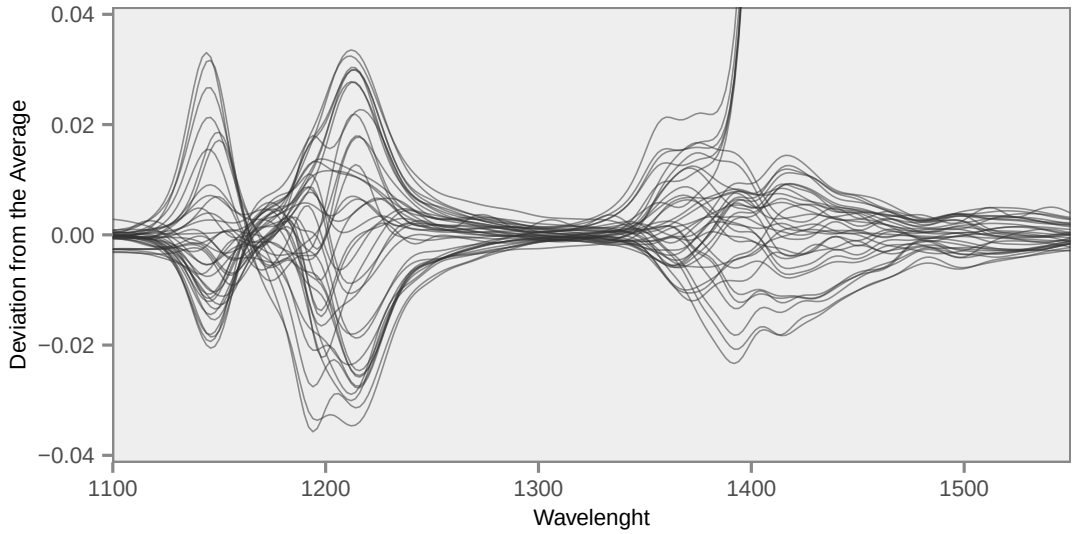

Figure 15: Curves of deviation from the average absorbance for gasoline samples from the octane data set at different wavelengths.

Considering a more detailed exploratory analysis of our data, we were able to identify the unreasonable assumption made. Figure 15 shows the curves of absorbance spectra obtained when subtracting the empirical mean of the uncontaminated samples for each wavelength. As we can see, the assumption of equal observational variances is not only completely violated, but also has a clear wavelength dependence structure. Having this in mind we consider an adapted main component, given by

$$\begin{aligned} Y_{ij} &= \mu_j + \nu_{ij}, \qquad \nu_{ij} \overset{ind}{\sim} Normal\left(0, \phi_j^{-1}\right), \\ \mu_j &= \mu_{j-1} + \omega_j, \quad \omega_{ij} \overset{ind}{\sim} Normal\left(0, \phi_\omega^{-1}\right), \end{aligned} \tag{4.3}$$

where the only difference is the inclusion of a precision ϕ_j for the observational error of the j-th wavelength. It is worth noting that, for simplicity, we do not model the dependence structure between the observational precisions, but a more accurate description of the typical behavior of the data should probably consider it.

We once more fit the model considering a Gibbs sampler using the full conditionals described in section B.1 of Appendix B. We took $\gamma = 0.95$, $w_z = \frac{1}{2}$ and simulated 100000 steps of the chain, taking the first 5000 states as burn in and a thinning of 10. For reference, the approximate computation time required to obtain the sample via MCMC was of 1 hour, 32 minutes and 16 seconds, with most of the cost coming from the estimation of the parameters from the main component. And, as a side note, in this case, the choice of a relatively high value of γ is justified by the small sample size $n = 39$.

Figure 16 presents the estimated classification for each observation and presents a 95% predictive credibility interval for a future uncontaminated observation. From Figure 16 we can also verify that the model was capable of correctly identifying the anomalous observations with high confidence.

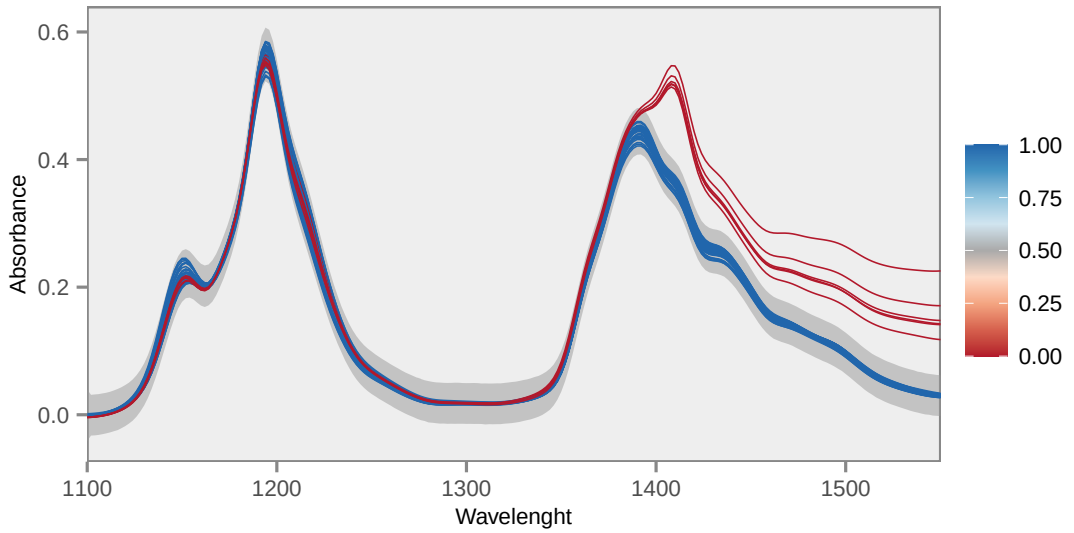

Figure 16: Absorbance curves for gasoline samples from the octane data set at different wavelengths, where the colors indicate the estimated classification from the filtering model.

4.2 BREAST TUMOR CLASSIFICATION

In this section we consider fitting the filtering model to the Breast Cancer Wisconsin (Diagnostic) data set, see Wolberg et al. (1995), consisting of features computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. This

data set contains a total of 30 real-value attributes for each one of the $n = 569$ observations and an indicator of whether the breast mass is benign or malign. And here, our objective is to estimate these indicator variables.

If we denote the dichotomic quantities we wish to estimate by $y_1, \ldots, y_n$, where $y_i = 1$, if the tumor is benign, and $y_i = 2$, otherwise, then it is worth noticing that, since $y_i \in \{1, 2\}$, for all $i \in \{1, \ldots, n\}$, the quantities we wish to estimate do not satisfy the filtering condition regardless of the assumed model. However, notice that this does not necessarily mean that we cannot estimate the y_i's using the filtering model, since the filtering condition is only required for the response variables of the model. Our way of circumventing this issue is to indirectly estimate them using a multivariate mixture model.

Then, if $X_1, \ldots, X_n$ represent the vectors of attributes to for each observation, then we can consider the following structure for the main component:

$$\begin{aligned} X_i|\mu_j, \Omega_j, y_i = j &\overset{ind}{\sim} Normal_d(\mu_j, \Omega_j^{-1}), \\ \mu_j|\Omega_j &\overset{ind}{\sim} Normal_d(\theta_j, \lambda_j^{-1}\Omega_j^{-1}), \\ \Omega_j &\overset{ind}{\sim} Wishart(\nu_j, V_j), \\ y_i|w_y &\overset{ind}{\sim} Categorical(w_y), \\ w_y &\sim Dirichlet(\alpha), \end{aligned} \tag{4.4}$$

for the groups $j \in \{1, 2\}$, where d is the number attributes considered, μ_j represent the average feature vector for an observation with classification j, Ω_j is the precision matrix of each group and w_y represents the vector containing the probability of allocating an observation to each group. However, it is still unclear what do we gain by using the filtering model for the classification.

As an initial justification, we can argue that the filtering model allows for a more robust estimation of the clusters, which is interesting considering a normal-based clustering approach, that is known to be sensitive to the presence of anomalies. However, this could be easily resolved by considering another mixture, for instance, of Student-t distributions. Nevertheless, what we consider here the main advantage of using the filtering model is allowing us to treat this task as an open set classification problem, i.e., our fitted model is able to classify each observation as benign,

malign or neither one of them.

Typically, a classification model is used to allocate an unlabeled observation to one of the groups it was trained to identify. However, this has the arguably undesirable consequence of attributing a label to an observation even if it is completely isolated from all other sample points. So, in the case of a strong anomaly, unless the observation has an approximately equal distance to the two closest clusters, the classification tends to have a high confidence. Thus, only accounting for the uncertainty of the estimation is insufficient to identify these cases. With this in mind, considering problems that may have a significant impact in someone's life, for instance determining whether or not an individual should be diagnosed with cancer, controlling this extrapolation error is seems to be a desirable propriety.

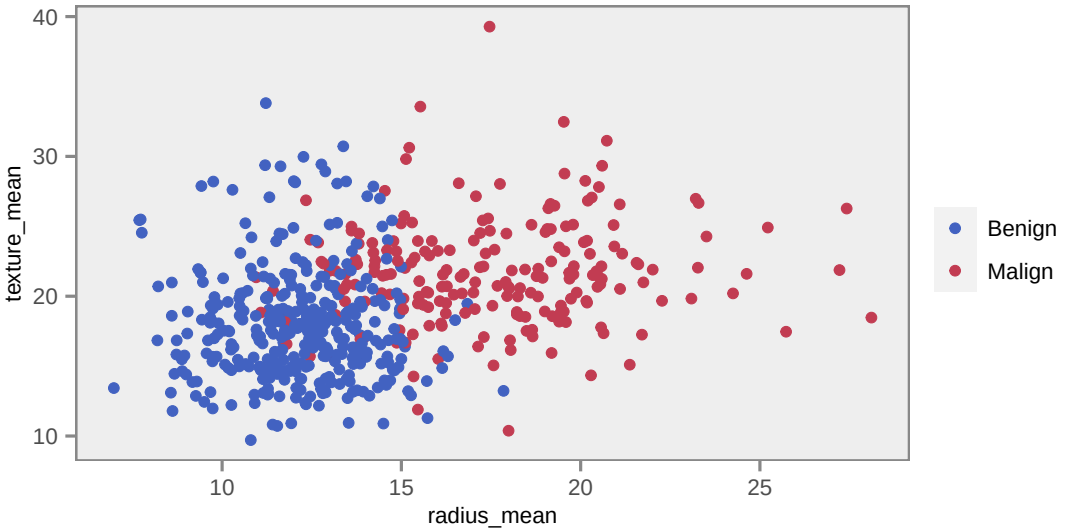

Figure 17: Scatter plot for the attributes radius_mean and texture_mean, where the color indicate whether the observation is benign (blue) or malign (red).

Firstly, knowing that our methodology is model based, we need to asses whether or not the assumed model reasonably describes the typical behavior of the data. Knowing that for a multivariate normal random vector each subset of the entries are also normally distributed, we can use some graphical analysis to select attributes that seem to approximately follow a multivariate normal distribution. So, since we already know the classification, we analyzed the scatter plot for every pair of attributes and subjectively discarded the variables for which one of the groups seemed

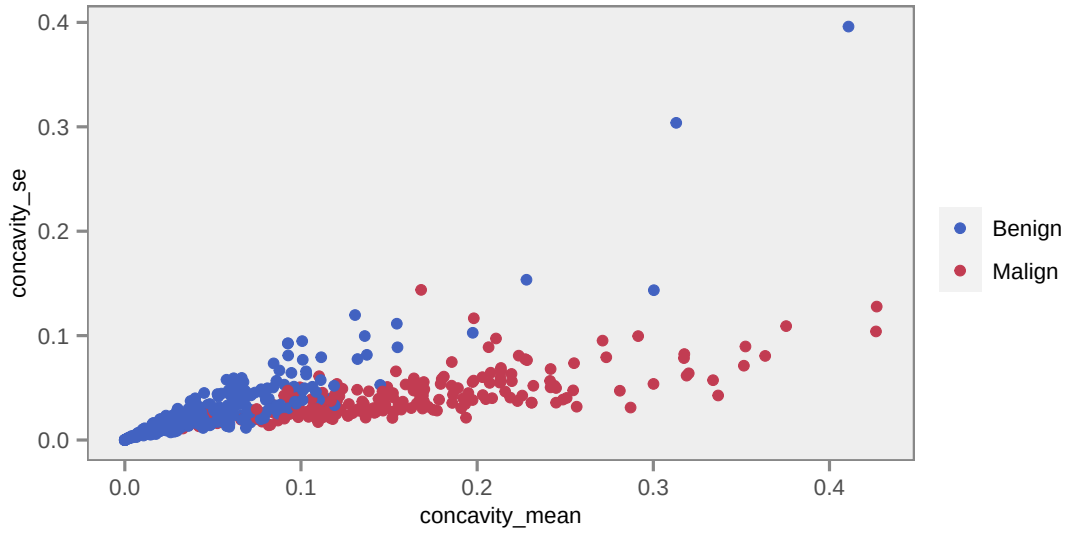

Figure 18: Scatter plot for the attributes concavity_mean and concavity_se, where the color indicate whether the observation is benign (blue) or malign (red).

to violate normality. As an example of our analysis, Figure 17 shows the scatter plot for two of the accepted variables: radius_mean and texture_mean; and Figure 18 shows the scatter plot for two rejected variables: concavity_mean and concavity_se.

To fit the filtering model with main component given by equation 4.4, we were able to use a Gibbs sampler using the full conditionals described in section B.2 of Appendix B to generate a sample from the posterior distribution. We took $\gamma = 0.75$, $w_z = \frac{1}{2}$ and simulated 100000 steps of the chain, taking the first 5000 states as burn in and a thinning of 10. For reference, the approximate computation time required to obtain the sample via MCMC was of 10 minutes and 43 seconds, with approximately 57.85% of the cost coming from the estimation of the parameters from the main component. And, as a side note, in this case, we consider a relatively high value of γ because, from the exploratory analysis and pre-treatment of the data, we could rule out the possibility of a strong bias from a consistent group of anomalous observations negatively impacting our estimates. Figure 19 presents the estimated classification for each observation and, comparing with Figure 17, we can see that the model provided a reasonable classification.

For comparison, we also fitted a conventional normal mixture model to the same data, i.e. we considered the indicators z_i known and equal to 1 for all i, and present

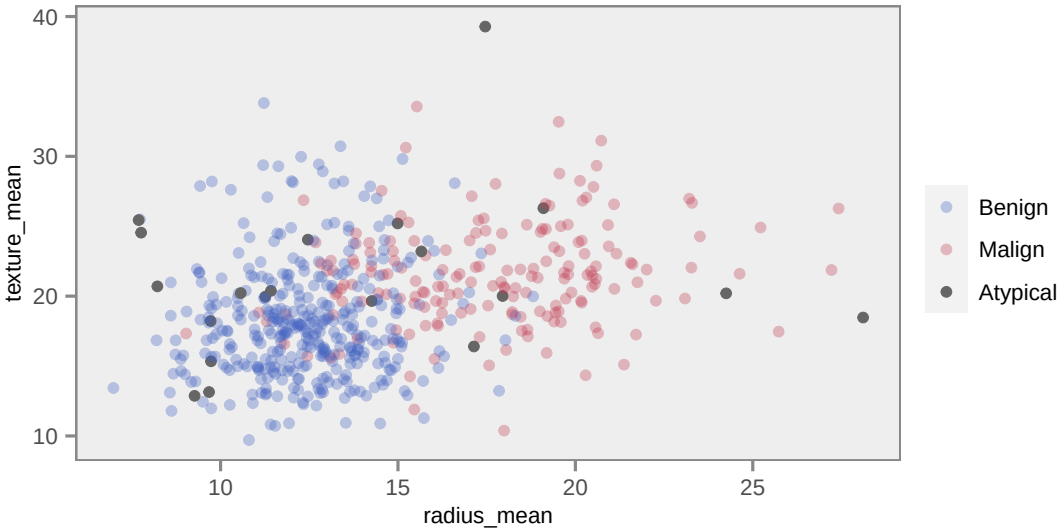

Figure 19: Scatter plot for the attributes radius_mean and texture_mean, where the color indicate the estimated classification from the filtering model.

our results in table 6. To construct this table we considered three classifiers: the filtering model with open set classification, the filtering model with closed set classification and the conventional normal mixture model. Here, both the open and closed set classifiers use the fitted filtering model. The different between them is that the open one use the value of z_i and y_i to label the i-th observation as either benign, malign or atypical, while the closed one uses only the information from y_i to determine whether a given sample point is benign or malign.

Model	accuracy (%)	atypical (%)
Filtering Model (Open)	90.86	3.51
Filtering Model (Closed)	92.97	0.00
Normal Mixture	92.31	0.00

Table 6: Summary of the classification from the estimated models for the Breast Cancer Wisconsin (Diagnostic) data set.

Considering the accuracy of each of the classifiers we can state that they are approximately equivalent, however, it is worth reinforcing that the open set classifier provides us with more information for decision making, allowing us to better quantify our uncertainties associated with the classification task.

4.3 IDENTIFICATION OF HISTORIC EVENTS

In this section we consider fitting the filtering model to a data set containing the mortality rates of the French male population from 1816 to 2020. The mortality rates we used in this work were obtained at the Human Mortality Database (2000) and are freely available at `https://www.mortality.org/`. Figure 20 shows a heatmap of the mortality rates for ages ranging from 0 to 100 and for years from 1816 to 2020. As we can see, the mortality rates are well behaved, represented by a mostly smooth heatmap, with the exception of some abrupt changes, so our objective is to use these sudden changes of behavior in order to identify historic events.

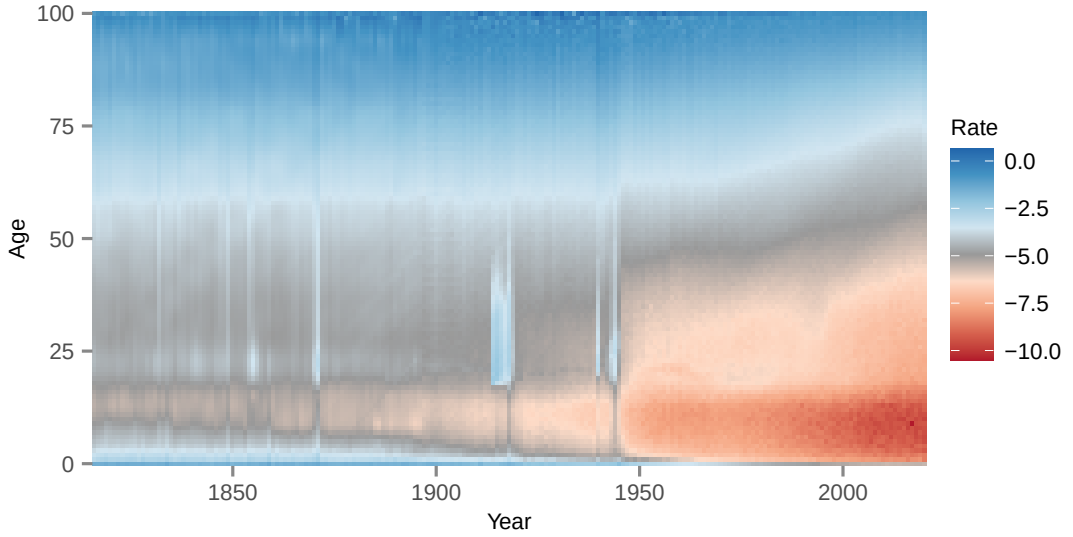

Figure 20: Heatmap of mortality rates of the French male population for ages ranging from 0 to 100 and for years from 1816 to 2020.

For our main component, we here consider the *Dynamic Improvement Model* (translated from the original name "Modelo de Improvement Dinâmico") proposed by Sartório (2018). This model is an extension of the model proposed by Lee & Carter (1992), that considers estimating the best decomposition of the form

$$Y_{it} = \alpha_i + \beta_i \kappa_t + \varepsilon_{it}, \tag{4.5}$$

where Y_{it} is the natural logarithm of the mortality rate of for i-th age group and at the t-th year, the vector α captures the average mortality rate for each age, the vector κ is the only term dependent on the time, so it is responsible for capturing

the temporal evolution of mortality rates, the vector β modulates the intensity of the temporal effect from κ for each age and ε_{it} represents an homoscedastic normal error term. In their original article, Lee & Carter (1992) propose the use of a two step estimation: first a singular value decomposition technique to estimate α and β, and later using an ARIMA model to estimate κ and make predictions while using the estimated values of α and β from the first step. Due to an identifiability issue, the following restriction were added to ensure an unique solution:

$$\sum_i \beta_i = 1 \quad \text{and} \sum_t \kappa_t = 0. \tag{4.6}$$

A few years later, Pedroza (2006) proposed a Bayesian approach to estimation by rewriting the model in the form of a dynamic linear model. This generalization allowed for one-step estimation and also provided a consistent framework to flexibilize the original model's assumptions, such as homoscedasticity. Then, Sartório (2018) proposed an extension based on this dynamic linear model representation, which is the model we use as the main component for this application. The resulting model is given by

$$\begin{aligned}
Y_t &= \alpha + \beta_t \kappa_t + \nu_t, & \nu_t &\overset{ind}{\sim} Normal_{(n+1)}(0, diag(\phi)^{-1}),\\
\kappa_t &= \kappa_{t-1} + \delta_{t-1} + \omega_t^{(\kappa)}, & \omega_t^{(\kappa)} &\overset{ind}{\sim} Normal(0, \phi_\kappa^{-1}),\\
\delta_t &= \delta_{t-1} + \omega_t^{(\delta)}, & \omega_t^{(\delta)} &\overset{ind}{\sim} Normal(0, \phi_\delta^{-1}),\\
\beta_t^* &= \beta_{t-1}^* + \omega_t^{(\beta^*)}, & \omega_t^{(\beta^*)} &\overset{ind}{\sim} Normal_n(0, diag(\phi_{\beta^*})^{-1}).
\end{aligned} \tag{4.7}$$

where Y_t is a vector representing the natural logarithm of the mortality rates at time t for all age groups, β becomes a matrix of values for each age group i and time t, κ follows a random walk with drift δ, to capture a long term consistent mortality reduction trend, $\phi = (\phi_0, \ldots, \phi_n)'$ and $\phi_{\beta^*} = (\phi_{\beta_1^*}, \ldots, \phi_{\beta_n^*})'$ are precision vectors, we denote by $diag(x)$ the diagonal matrix with elements given by vector x and β^* is a transformation of β we explain ahead. It is worth mentioning that the expression presented here in equation 4.7 is a slight simplification of the original model, so we refer to the thesis by Sartório (2018) for the complete specification.

As well as in Lee & Carter (1992), Sartório (2018) also imposed restrictions due

to identifiability issues, given by

$$\sum_i \beta_{it}^2 = 1 \quad \text{and} \quad \sum_t \kappa_t = 0, \tag{4.8}$$

for all $t \in \{1, \ldots, T\}$, however, we adapt these restrictions for computational convenience. The first restriction we consider is

$$\sum_{i=0}^{n} \beta_{it} = n + 1, \tag{4.9}$$

which means that each β_{it} can be written as an affine combination of the rest, for a fixed value of $t \in \{1, \ldots, T\}$. Having this in mind, we can impose this restriction by estimating $\beta_t^* = (\beta_{1t}, \ldots, \beta_{nt})'$ instead of $\beta_t = (\beta_{0t}, \ldots, \beta_{nt})'$ and writing β_{0t} as function of β_t^*. In matrix form, we equivalently have

$$\beta_t = \begin{bmatrix} n+1 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} -1 & -1 & \cdots & -1 & -1 \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix} \beta_t^*, \tag{4.10}$$

for all $t \in \{1, \ldots, n\}$, which is a convenient representation. Next, we adapt the restriction for κ by taking

$$\sum_{t=1}^{T} z_{it} \beta_{it} \kappa_t = 0, \tag{4.11}$$

for all $i \in \{0, \ldots, n\}$, where the prior for the indicators z_{it}'s will be introduced later in this section, and we apply it to the full conditional of α, presented in section B.3. It is interesting to notice that

$$\begin{aligned} \mathbb{E}\left[\frac{1}{n_{i\cdot 1}} \sum_{t=1}^{T} z_{it} Y_{it} \middle| \theta, z \right] &= \frac{1}{n_{i\cdot 1}} \sum_{t=1}^{T} z_{it} \mathbb{E}\left[Y_{it} \middle| \theta, z \right] = \frac{1}{n_{i\cdot 1}} \sum_{t=1}^{T} z_{it} \left[\alpha_i + \beta_{it} \kappa_t \right] \\ &= \frac{1}{n_{i\cdot 1}} \sum_{t=1}^{T} z_{it} \alpha_i + \frac{1}{n_{i\cdot 1}} \sum_{t=1}^{T} \beta_{it} \kappa_t = \alpha_i, \end{aligned} \tag{4.12}$$

where θ is the collection of all of the model's parameters and $n_{i\cdot 1} = \sum_{t=1}^{T} z_{it}$, so with this restriction we can interpret α_i as the expected average log-mortality rate for the age group i considering only the typical observations.

We then consider the filtering model with main component given by the model presented in 4.7. The full model specification along with the prior structure is detailed in section B.3 of Appendix B. For the choice of prior for the matrix of indicators z, we consider two distinct structures. The first, which we call the *independent indicator structure*, treats the classification of any two indicators as independent tasks, and is given by

$$z_{it} \overset{ind}{\sim} Bernoulli\Big(w_z\Big), \tag{4.13}$$

for all age groups $i \in \{0, \dots n\}$ and year indexes $t \in \{1, \dots, T\}$, where w_z represents the prior probability of classifying an observation as atypical. As an alternative, we next consider the *correlated indicator structure*, given by

$$\begin{aligned} z_{it}|z_{(i-1)t}, \rho &\sim Bernoulli\Big(z_{(i-1)t}\rho + (1 - z_{(i-1)t})(1 - \rho)\Big), \\ z_{(-1)t} &\overset{ind}{\sim} Bernoulli(\rho_0), \\ \rho &\sim Beta(a, b), \end{aligned} \tag{4.14}$$

for all age groups $i \in \{0, \dots n\}$ and year indexes $t \in \{1, \dots, T\}$, where $z_{(-1)t}$ are virtual indicators used as auxiliary variables for algebraic convenience, $\rho, \rho_0 \in (0, 1)$ and $a, b > 0$. Here, this structure represents the assumption that, for a given year t, the classification for observations at age groups i and $i - 1$ should be correlated with some intensity ρ.

To fit the filtering model with main component given by equation 4.7, we considered a Gibbs sampler, using the full conditionals described in section B.3 of Appendix B to generate a sample from the posterior distribution. Here we took $\gamma = 0.05$, $w_z = \frac{1}{2}$, $\rho_0 \in \frac{1}{2}$, $a = 1000$, $b = 1000$ and simulated 10000 steps of the chain, taking the first 1000 states as burn in and a thinning of 10. For reference, the approximate computation time required to obtain the sample via MCMC was of 11 hour, 48 minutes and 50 seconds for the model with independent indicator structure and of 12 hours, 34 minutes and 32 seconds for the model with correlated indicator structure, with most of the cost coming from the estimation of the parameters from the main component. And, as a side note, in this case, a relatively low value of γ was chosen in an attempt to compensate the possible bias generated by the presence of highly anomalous observations known to be present in the data, e.g. mortality

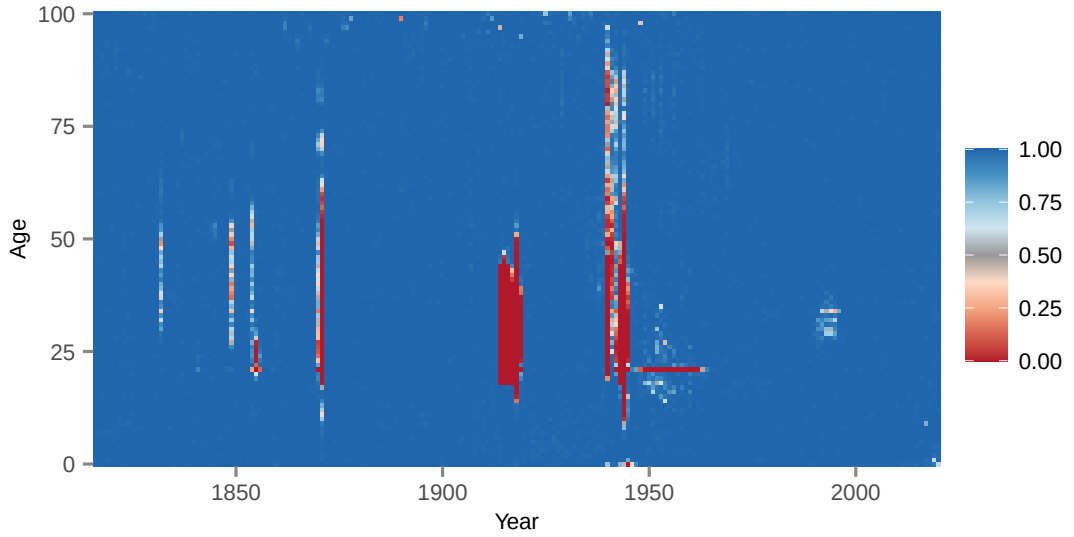

Figure 21: Heatmap of the estimated classification made by the filtering model with independent indicator structure for each age group and year considered.

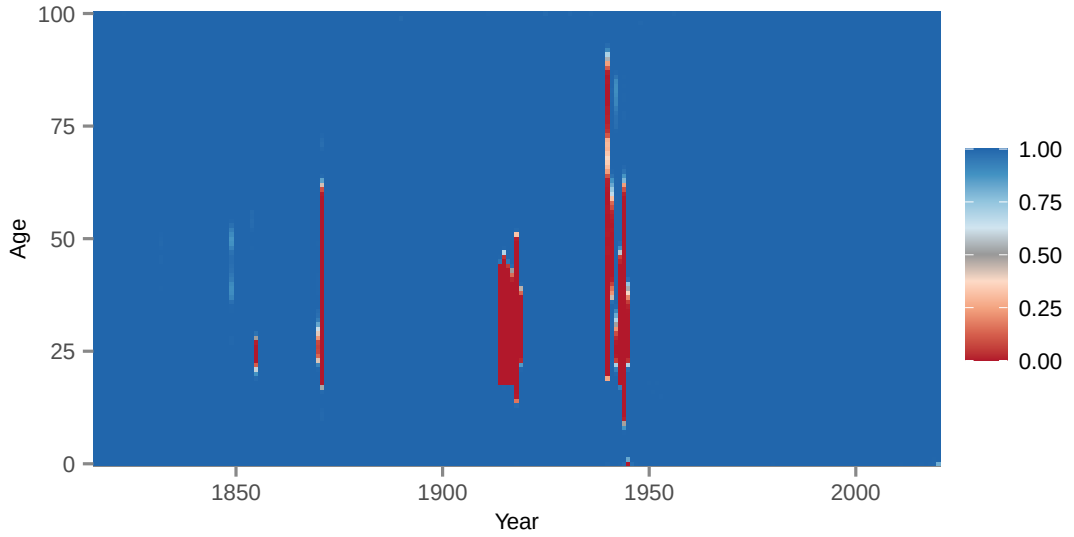

Figure 22: Heatmap of the estimated classification made by the filtering model with correlated indicator structure for each age group and year considered.

during World Wars I and II.

Figures 21 and 22 present the heatmap with estimated classification considering the model with independent and correlated indicator structures, respectively. As we can see, both fitted models were able to indentify some well known historic events, such as World War I, between years 1914 and 1918, and World War II, from 1939 to 1945, satisfactorily, even though the classifications are qualitatively different when comparing the prior structures considered. For the independent indicator structure, we can see a more noisy and error-prone classification, while the one provided by the correlated indicator structure is smoother. However, it is interesting to notice that the added structure in the correlated estimation necessarily is an improvement, since it ends up being more hesitant to classify an observation as an anomaly.

5 FINAL CONSIDERATIONS

In this work we propose a Bayesian model-based anomaly detection using a mixture of a chosen parametric model and an uniform distribution, whose measure is given by a quantile of the autotransformation of the response variable. Our methodology is flexible, dealing with the case of non-identically distributed observations, models with significant hierarchical complexity and even dependence structure on the indicator variables, at the cost of assuming independent response variables satisfying the filtering condition. Estimation is considered using MCMC methods, with the Metropolis-Hastings algorithm or, under some rare and specific circumstances, the Gibbs sampler, and accounts for all of the uncertainties involved in the process. The method also has a broad scope of applications, dealing with problems of clusterization, unsupervised and/or open set classification and regression. However, it is important pointing out that the filtering model still has important issues and poorly understood properties that require further studies.

We already established that the analytical and computational cost may prevent the practical use of our method in more general scenarios and, as evidence, we recall that all of the applications presented in this work relied on assuming normally distributed response variables. To address this, in future work we hope to use a combination of Extreme Value Theory and approximate inference techniques, such as variational methods, to propose reasonable and practical approximations of the filtering model.

Another important issue is the multimodality and possible lack of identifiability of the proposed model, as both of them contribute to the complexity of the estimation and sensibility to the initial condition of the MCMC algorithms. As a path to mitigate this problem, we wish to consider reinforcement learning techniques to find a better balance between allocating observations to the alternative component and seeking a better fit to our data. Another related matter is the understanding of how the alternative component affects the posterior distribution of the parameter from the main component, which may have a relevant role in the estimation complexity problem.

Lastly, we highlight that, since we consider a model-based approach, our results heavily rely on an appropriate choice of the distribution family of the main component and, depending on the case, may be sensitive to the specification of the prior distribution. In particular, the choice of the threshold of maximum uncertainty showed to be sensitive and requires further analysis. Having this in mind, even though the method is able to perform complete unsupervised anomaly identification, we strongly recommend thorough model validation and, if possible, using supervision to ensure proper inference and decision making.

References

Abraham, B., & Box, G. E. P. (1978). Linear models and spurious observations. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *27*(2), 131–138.
URL http://www.jstor.org/stable/2346940

Aitkin, M., & Wilson, G. T. (1980). Mixture models, outliers, and the em algorithm. *Technometrics*, *22*(3), 325–331.
URL http://www.jstor.org/stable/1268316

Amovin-Assagba, M., Gannaz, I., & Jacques, J. (2022). Outlier detection in multivariate functional data through a contaminated mixture model.

Arribas-Gil, A., & Romo, J. (2014). Shape outlier detection and visualization for functional data: the outliergram. *Biostatistics*, *15*(4), 603–619.

Banfield, J. D., & Raftery, A. E. (1993). Model-based gaussian and non-gaussian clustering. *Biometrics*, *49*(3), 803–821.
URL http://www.jstor.org/stable/2532201

Barnett, V., & Lewis, T. (1978). *Outliers in statistical data (1st ed.)*. John Wiley & Sons.

Barreto, D. W. (2022). *Modelo para Filtragem de Observações Não-Conformes*. Bachelor's thesis, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis*. Springer-Verlag New York, Inc.

Bouguessa, M. (2014). A mixture model-based combination approach for outlier detection. *International Journal on Artificial Intelligence Tools*, *23*(04), 1460021.
URL https://doi.org/10.1142/S0218213014600215

Box, G. E. P., & Tiao, G. C. (1968). A bayesian approach to some outlier problems. *Biometrika*, *55*(1), 119–129.
URL http://www.jstor.org/stable/2334456

Brunot, M. (2020). A gaussian uniform mixture model for robust kalman filtering. *IEEE Transactions on Aerospace and Electronic Systems*, *56*(4), 2656–2665.

Chen, Y., Dang, X., Peng, H., & Bart, H. L. (2008). Outlier detection with the kernelized spatial depth function. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(2), 288–305.

Cheng, A. Y., Liu, R. Y., & Luxhøj, J. T. (2000). Monitoring multivariate aviation safety data by data depth: control charts and threshold systems. *IIE Transactions*, *32*(9), 861–872.

Chib, S., & Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The american statistician*, *49*(4), 327–335.

Coretto, P., & Hennig, C. (2011). Maximum likelihood estimation of heterogeneous mixtures of gaussian and uniform distributions. *Journal of Statistical Planning and Inference*, *Vol. 141*, p. 462–473.

Coretto, P., & Hennig, C. (2016). Robust improper maximum likelihood: Tuning, computation, and a comparison with other methods for robust gaussian clustering. *Journal of the American Statistical Association*, *111*(516), 1648–1659.
URL `https://doi.org/10.1080/01621459.2015.1100996`

Dai, W., Mrkvička, T., Sun, Y., & Genton, M. G. (2020). Functional outlier detection and taxonomy by sequential transformations. *Computational Statistics & Data Analysis*, *149*, 106960.

De Finetti, B. (1961). The Bayesian approach to the rejection of outliers. Proc. 4th Berkeley Symp. Math. Stat. Probab. 1, 199-210 (1961).

de Haan, L., & Ferreira, A. (2006). *Extreme value theory: an introduction*, vol. 3. Springer.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, *39*(1), 1–22.

Esbensen, K. H., Guyot, D., Westad, F., & Houmoller, L. P. (2002). *Multivariate data analysis: in practice: an introduction to multivariate data analysis and experimental design*. Multivariate Data Analysis.

Evans, M., Guttman, I., & Olkin, I. (1992). Numerical aspects in estimating the parameters of a mixture of normal distributions. *Journal of Computational and Graphical Statistics - J COMPUT GRAPH STAT*, *1*, 351–365.

Febrero, M., Galeano, P., & González-Manteiga, W. (2008). Outlier detection in functional data by depth measures, with application to identify abnormal nox levels. *Environmetrics: The official journal of the International Environmetrics Society*, *19*(4), 331–345.

Fisher, R. A., & Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Mathematical Proceedings of the Cambridge Philosophical Society*, *24*(2), 180–190.

Fraiman, R., Meloche, J., García-Escudero, L. A., Gordaliza, A., He, X., Maronna, R., Yohai, V. J., Sheather, S. J., McKean, J. W., Small, C. G., et al. (1999). Multivariate l-estimation. *Test*, *8*, 255–317.

Frühwirth-Schnatter, S., & Frèuhwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*, vol. 425. Springer.

Gamerman, D., & Lopes, H. F. (2006). *Markov Chain Monte Carlo - Stochastic Simulation for Bayesian Inference (2nd ed.)*. 6000 Broken Sound Parkway NW, Suite 300: Chapman & Hall/CRC.

Geman, S., & Geman, D. (1984). Stochastic relaxation, gibbs distribution, and the bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, *6*, 721–741.

Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une serie aleatoire. *Annals of mathematics*, (pp. 423–453).

Guttman, I. (1973). Care and handling of univariate or multivariate outliners in detecting spuriosity—a bayesian approach. *Technometrics*, *15*(4), 723–738.
URL https://doi.org/10.1080/00401706.1973.10489107

Guttman, I., Dutter, R., & Freeman, P. R. (1978). Care and handling of univariate outliers in the general linear model to detect spuriosity: A bayesian approach. *Technometrics*, *20*(2), 187–193.
URL http://www.jstor.org/stable/1268712

Hamurkaroğlu, C., Mehmet, M., & Saykan, Y. (2004). Nonparametric control charts based on mahalanobis depth. *Hacettepe Journal of Mathematics and Statistics*, *33*, 57–67.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, *57*(1), 97–109.

Huijben, I. A., Kool, W., Paulus, M. B., & Van Sloun, R. J. (2022). A review of the gumbel-max trick and its extensions for discrete stochasticity in machine learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(2), 1353–1371.

Human Mortality Database (2000). Max planck institute for demographic research (germany), university of california, berkeley (usa), and french institute for demographic studies (france). Available at www.mortality.org. Data downloaded on 07/05/2023.

Inverardi, P. L. N., & Taufer, E. (2020). Outlier detection through mixtures with an improper component. *Electronic Journal of Applied Statistical Analysis*, *13*(1).
URL http://siba-ese.unisalento.it/index.php/ejasa/article/view/21006

Kuhnt, S., & Rehage, A. (2016). An angle-based multivariate functional pseudo-depth for shape outlier detection. *Journal of Multivariate Analysis*, *146*, 325–340.

Lathuilière, S., Mesejo, P., Alameda-Pineda, X., & Horaud, R. (2018). Deepgum: Learning deep robust regression with a gaussian-uniform mixture model.

Lee, R. D., & Carter, L. R. (1992). Modeling and forecasting us mortality. *Journal of the American statistical association*, *87*(419), 659–671.

Liu, J. S. (2001). *Monte Carlo Strategies In Scientific Computing*. 175 Fifth Avenue, New York, NY 10010, USA: Springer-Verlag New York, Inc.

Liu, R. Y. (1990). On a notion of data depth based on random simplices. *The Annals of Statistics*, (pp. 405–414).

Liu, R. Y. (1995). Control charts for multivariate processes. *Journal of the American Statistical Association*, *90*(432), 1380–1387.

Liu, R. Y., & Singh, K. (1993). A quality index based on data depth and multivariate rank tests. *Journal of the American Statistical Association*, *88*(421), 252–260.

Longford, N. T., & D'Urso, P. (2011). Mixture models with an improper component. *Journal of Applied Statistics*, *38*(11), 2511–2521.
URL `https://doi.org/10.1080/02664763.2011.559208`

Magalhães, M. N. (2006). *Probabilidade e variáveis aleatórias*. Edusp.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India*, *Vol. 2*(1), 49–55.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, *21*(6), 1087–1092.

Migon, H. S., Gamerman, D., & Louzada, F. (2014). *Statistical inference: an integrated approach*. CRC press.

Mosler, K. (2013). Depth statistics. *Robustness and complex data structures: Festschrift in Honour of Ursula Gather*, (pp. 17–34).

Newcomb, S. (1886). A generalized theory of the combination of observations so as to obtain the best result. *American Journal of Mathematics*, *8*(4), 343–366.
URL `http://www.jstor.org/stable/2369392`

Pedroza, C. (2006). A bayesian forecasting model: predicting us male mortality. *Biostatistics*, *7*(4), 530–550.

Peel, D., & McLachlan, G. J. (2000). Robust mixture modelling using the t distribution. *Statistics and Computing*, *Vol. 10*, p. 339–348.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
URL `https://www.R-project.org/`

RStudio Team (2019). *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA.
URL `http://www.rstudio.com/`

Sartório, V. S. (2018). *Modelagem da Evolução de Mortalidade Considerando Dinâmicas Temporais Etárias Específicas*. Bachelor's thesis, Escola Nacional de Ciências Estatísticas, Rio de Janeiro, Brasil.

Sguera, C., Galeano, P., & Lillo, R. E. (2016). Functional outlier detection by a local depth with application to no x levels. *Stochastic environmental research and risk assessment*, *30*(4), 1115–1130.

Stephens, M. (1997). *Bayesian methods for mixtures of normal distributions*. Ph.D. thesis, Magdalen College, Oxford.

Tukey, J. W. (1960). A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, (pp. 448–485).

Tukey, J. W. (1975). Mathematics and the picturing of data. *Proceedings of the International Congress of Mathematicians [editor, Ralph D. James]*, *Vol. 2*, p. 523–531.

Verdinelli, I., & Wasserman, L. (1991). Bayesian analysis of outlier problems using the gibbs sampler. *Statistics and Computing*, *1*, 105–117.

West, M., & Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models, (2nd ed.)*. 175 Fifth Avenue, New York, NY 10010, USA: Springer-Verlag New York, Inc.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.
URL `https://ggplot2.tidyverse.org`

Wolberg, W., Mangasarian, O., Street, N., & Street, W. (1995). Breast cancer wisconsin (diagnostic). UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5DW2B.

Yin, J., & Wang, J. (2016). A model-based approach for text clustering with outlier detection. In *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*, (pp. 625–636). IEEE.

Yu, C., Chen, K., & Yao, W. (2015). Outlier detection and robust mixture modeling using nonconvex penalized likelihood. *Journal of Statistical Planning and Inference*, *164*, 27–38.

Zuo, Y., & Serfling, R. (2000). General notions of statistical depth function. *Annals of statistics*, (pp. 461–482).

A AUTOTRANSFORMATIONS AND CORRECTION FUNCTIONS

In this Appendix we will provide some additional information to better introduce some of the definitions presented in Chapter 3. In section A.1 we will provide the proof of some the results stated in section 3.1 related to autotrasformations. Section A.2 shows some results related to correction functions and presents, in some sense, a convenient universal correction function based on the Weibull distribution. Then, section A.3 particularizes some results regarding the general class of location and scale models and section A.4 further considers these results for the multivariate normal distribution, that has significant importance for all of the applications in Chapter 4.

A.1 GENERAL PROPRIETIES OF AUTOTRANSFORMATIONS

We begin stating and proving proposition A.1, that shows us how to write the distribution of an autotransformation of order n in terms of the distribution of the autotransformation of the same random variable.

Proposition A.1. *Let X be a d-dimensional absolutely continuous or discrete random vector. Then the distribution function of $T_X^{(n)}$ is given by the expression*

$$F_{T_X^{(n)}}(x) = 1 - [1 - F_{T_X}(x)]^n . \tag{A.1}$$

Proof. The following proof is equivalent to the one used to find the distribution of the minimum of independent random variables. So, we have

$$\begin{aligned} F_{T_X^n}(x) &= \mathbb{P}\left(T_X^n \leq x\right) = \mathbb{P}\left(\min\{T_{X_1}, \ldots, T_{X_n}\} \leq x\right) \\ &= 1 - \mathbb{P}\left(\min\{T_{X_1}, \ldots, T_{X_n}\} > x\right) \\ &= 1 - \mathbb{P}\left(T_{X_1} > x, \ldots, T_{X_n} > x\right) \\ &\overset{ind}{=} 1 - \prod_{i=1}^{n} \mathbb{P}\left(T_{X_i} > x\right) = 1 - \prod_{i=1}^{n} \left[1 - \mathbb{P}\left(T_{X_i} \leq x\right)\right] \\ &\overset{id}{=} 1 - \left[1 - F_{T_X}(x)\right]^n . \end{aligned} \tag{A.2}$$

□

Next, analogously to what we obtained in proposition A.1, we can find the quantile function of $T_{Y^{(n)}}$ in terms of the quantile function of the autotransformation T_Y, as stated in propositions A.2.

Proposition A.2. *Let X be a d-dimensional absolutely continuous or discrete random vector. Then, the quantile function of $T_X^{(n)}$ is given by the expression*

$$F^{-1}_{T_X^{(n)}}(x) = F^{-1}_{T_X}\left(1-(1-x)^{\frac{1}{n}}\right). \tag{A.3}$$

Proof. From propositions A.1 and we have

$$\begin{aligned} F_{T_X^{(n)}}(x) \geq y &\iff 1-[1-F_{T_X}(x)]^n \geq y \\ &\iff 1-F_{T_X}(x) \leq (1-y)^{\frac{1}{n}} \\ &\iff F_{T_X}(x) \geq 1-(1-y)^{\frac{1}{n}}, \end{aligned} \tag{A.4}$$

so we can infer that, for all $y \in (0,1]$,

$$\begin{aligned} F^{-1}_{T_X^{(n)}}(y) &= \inf\left\{x \in \mathbb{R} : F_{T_X^{(n)}}(x) \geq y\right\} \\ &= \inf\left\{x \in \mathbb{R} : F_{T_X}(x) \geq 1-(1-y)^{\frac{1}{n}}\right\} \\ &= F^{-1}_{T_X}\left(1-(1-y)^{\frac{1}{n}}\right) \end{aligned} \tag{A.5}$$

and for $y = 0$ the equality

$$F^{-1}_{T_X^{(n)}}(y) = F^{-1}_{T_X^{(n)}}(0) \overset{def}{=} 0 \overset{def}{=} F^{-1}_{T_X}(0) = F^{-1}_{T_X}\left(1-(1-0)^{\frac{1}{n}}\right) = F^{-1}_{T_X}\left(1-(1-y)^{\frac{1}{n}}\right) \tag{A.6}$$

is satisfied by definition, so we have

$$F^{-1}_{T_X^{(n)}}(x) = F^{-1}_{T_X}\left(1-(1-x)^{\frac{1}{n}}\right). \tag{A.7}$$

□

Both propositions A.1 and A.2 are useful because they lead to an important reduction on the number of distribution functions considered in the filtering model's definition, which substantially diminishes the analytical and/or computational burden of the method's implementation.

A.2 CORRECTION FUNCTIONS

In this section we will clarify some of the results presented in subsection 3.1.3 related to correction functions. But first, we will introduce and prove a necessary proposition related to the generalized inverse function of Definition 3.3 that follows.

Proposition A.3. *Let X be a d-dimensional absolutely continuous or discrete random vector with distribution F_X and satisfying the filtering condition. Then, the generalized inverse of $F_{T_X^{(n)}}$ must satisfy*

$$F^{-1}_{T_X^{(n)}}(y) \leq t \iff y \leq F_{T_X^{(n)}}(t) \tag{A.8}$$

for all $0 \leq y \leq 1$, $t \geq 0$ and $n \in \mathbb{N}^*$.

Proof. The generalized inverse function of Definition 3.3 is a modification of the generalized inverse presented in Definition 3.11 of Magalhães (2006) and both definitions coincide for $0 < y \leq 1$. So the proof of this proposition for $0 < y \leq 1$ is the same as the one from proposition 3.4 of Magalhães (2006), and for this reason the proof for this case will be omitted. Then, considering $y = 0$, we need to show that

$$F^{-1}_{T_X^{(n)}}(0) \leq t \iff 0 \leq F_{T_X^{(n)}}(t), \tag{A.9}$$

for all $t \geq 0$ and $n \in \mathbb{N}^*$. Considering the left-hand side of the equivalence we notice that, since $t \geq 0$ and by definition $F^{-1}_{T_X^{(n)}}(0) = 0$, the inequality always holds. Now, looking at the right-hand side of the equivalence we know that the inequality always holds for $t \geq 0$, because $T_X^{(n)} > 0$ almost surely. So, since both sides are always true, the equivalence holds for $y = 0$. □

Next, we show that the correction functions according to equations 3.30 and 3.31 are well defined. To do this, we need to show that the operations from the recursive relation in equation 3.30, given the initial value $g_i(n) = n$, do not violate the domain constraints of the generalized inverse of $F_{T_{Y_i}}(\cdot|\theta)$ and provide clarification on some edge cases, where the definition is unclear. For simplicity we will recurrently adopt the notation: $h_i = F_{T_{Y_i}}(\cdot|\theta)$.

First, we establish as a convention that

- $1 - \gamma^{0^{-1}} = 1 - \gamma^{+\infty} = 1$,
- $\left[\log_\gamma(0)\right]^{-1} = [+\infty]^{-1} = 0$, and
- $\left[\log_\gamma(1)\right]^{-1} = 0^{-1} = +\infty$.

Now, we need to show that, for all $x \in \{1, \ldots, n\}$,

$$0 < 1 - \gamma^{g_i(x)^{-1}} \leq 1. \tag{A.10}$$

Notice that, considering our convention we have

$$\begin{aligned} 0 < 1 - \gamma^{g_i(x)^{-1}} \leq 1 &\iff -1 < -\gamma^{g_i(x)^{-1}} \leq 0 \\ &\iff 1 > \gamma^{g_i(x)^{-1}} \geq 0 \\ &\iff 0 < g_i(x)^{-1} \leq +\infty \\ &\iff 0 \leq g_i(x) < +\infty \end{aligned} \tag{A.11}$$

So instead, we prove the equivalent condition $0 \leq g(x) < +\infty$, for all $x \in \{1, \ldots, n\}$, by induction in x.

Before following with the proof, it is important highlighting that from the filtering condition we have that $F_{T_{Y_i}}(\varepsilon|\theta) > 0$, for all $\varepsilon > 0$, and we know that $T_{Y_i} > 0$ almost surely, which imply that

$$F_{T_{Y_i}}(x) = 0 \iff x \leq 0. \tag{A.12}$$

From the equivalence above and using that distribution functions are right-continuous, we can also conclude that, for $y \in (0, 1]$,

$$F^{-1}_{T_{Y_i}}(y) = \inf\left\{x \in \mathbb{R} : F_{T_{Y_i}}(x) \geq y > 0\right\} > 0. \tag{A.13}$$

So, considering that $F^{-1}_{T_{Y_i}}(0) = 0$, we similarly infer that

$$F_{T_{Y_i}}(y) = 0 \iff y = 0. \tag{A.14}$$

Now, back to the proof by induction we have:

Proof. Fixating $n \in \mathbb{N}^*$, for the initial case we know that $0 \leq g_i(n) = n < +\infty$. Next, we need to show that, for all $x \in \{1, \ldots, n-1\}$,

$$0 \leq g(x+1) < +\infty \implies 0 \leq g(x) < +\infty. \tag{A.15}$$

Consider the following:

$$
\begin{aligned}
&0 \leq g_i(x+1) < +\infty \\
&\Rightarrow 0 < 1 - \gamma^{g_i(x+1)^{-1}} \leq 1 \\
&\Rightarrow 0 < h_i^{-1}\left(1 - \gamma^{g_i(x+1)^{-1}}\right) \leq h^{-1}(1) \\
&\Rightarrow 0 < \left[h_i^{-1}\left(1 - \gamma^{g_i(x+1)^{-1}}\right)\right]^{\overbrace{\frac{n-x}{n-x+1}}^{>0}} \leq \left[h_i^{-1}(1)\right]^{\frac{n-x}{n-x+1}} \\
&\Rightarrow 0 < \underbrace{\left[h_i^{-1}\left(1 - \gamma^{x^{-1}}\right)\right]^{\frac{1}{n-x+1}}}_{>0} \left[h_i^{-1}\left(1 - \gamma^{g_i(x+1)^{-1}}\right)\right]^{\frac{n-x}{n-x+1}} \\
&\leq \left[h_i^{-1}\left(1 - \gamma^{x^{-1}}\right)\right]^{\frac{1}{n-x+1}} \left[h_i^{-1}(1)\right]^{\frac{n-x}{n-x+1}} \leq \left[h_i^{-1}(1)\right]^{\frac{1}{n-x+1}} \left[h_i^{-1}(1)\right]^{\frac{n-x}{n-x+1}} = h_i^{-1}(1) \\
&\Rightarrow 0 < h_i\left(\left[h_i^{-1}\left(1 - \gamma^{x^{-1}}\right)\right]^{\frac{1}{n-x+1}} \left[h_i^{-1}\left(1 - \gamma^{g_i(x+1)^{-1}}\right)\right]^{\frac{n-x}{n-x+1}}\right) \leq h_i\left(h_i^{-1}(1)\right) \leq 1 \\
&\Rightarrow 0 \leq 1 - h_i\left(\left[h_i^{-1}\left(1 - \gamma^{x^{-1}}\right)\right]^{\frac{1}{n-x+1}} \left[h_i^{-1}\left(1 - \gamma^{g_i(x+1)^{-1}}\right)\right]^{\frac{n-x}{n-x+1}}\right) < 1 \\
&\Rightarrow 0 < \log_\gamma\left[1 - h_i\left(\left[h_i^{-1}\left(1 - \gamma^{x^{-1}}\right)\right]^{\frac{1}{n-x+1}} \left[h_i^{-1}\left(1 - \gamma^{g_i(x+1)^{-1}}\right)\right]^{\frac{n-x}{n-x+1}}\right)\right] \leq +\infty \\
&\Rightarrow 0 < [g_i(x)]^{-1} \leq +\infty \\
&\Rightarrow 0 \leq g_i(x) < +\infty.
\end{aligned}
\tag{A.16}
$$

□

The next proposition is the main result regarding correction functions of this work. It allows us to control or even eliminate the bias from the filtering model considering the identically distributed case and it is restated below.

Proposition A.4. *Let $Y_1, \ldots, Y_n$ be an independent sample from the d-dimensional absolutely continuous or discrete distributions $F_{Y_1}(\cdot|\theta)$, ..., $F_{Y_n}(\cdot|\theta)$ and $\gamma \in (0,1)$ a real scalar. If $Y_1, \ldots, Y_n$ satisfy the filtering condition given θ, then the correction function $g_i = g_i(\cdot|\theta, \gamma)$ for the i-th observation satisfies*

$$
F_{T_{Y_i}}^{-1}\left(1 - \gamma^{g_i(x)^{-1}}\middle|\,\theta\right)\left[\frac{F_{T_{Y_i}}^{-1}\left(1 - \gamma^{g_i(x)^{-1}}\middle|\,\theta\right)}{F_{T_{Y_i}}^{-1}\left(1 - \gamma^{g_i(x+1)^{-1}}|\,\theta\right)}\right]^{n-x} \leq F_{T_{Y_i}}^{-1}\left(1 - \gamma^{x^{-1}}\middle|\,\theta\right), \tag{A.17}
$$

for $x \in \{1, \ldots, n\}$. Furthermore, if T_{Y_i} is absolutely continuous, then correction

function $g_i = g_i(\cdot|\theta,\gamma)$ *for the* i*-th observation satisfies*

$$F_{T_{Y_i}}^{-1}\left(1-\gamma^{g_i(x)^{-1}}\middle|\,\theta\right)\left[\frac{F_{T_{Y_i}}^{-1}\left(1-\gamma^{g_i(x)^{-1}}\middle|\,\theta\right)}{F_{T_{Y_i}}^{-1}\left(1-\gamma^{g_i(x+1)^{-1}}|\,\theta\right)}\right]^{n-x} = F_{T_{Y_i}}^{-1}\left(1-\gamma^{x^{-1}}\middle|\,\theta\right), \quad \text{(A.18)}$$

for $x \in \{1,\ldots,n\}$.

Proof. Regarding the first statement, adopting the notation $h_i = F_{T_{Y_i}}(\cdot|\theta)$, consider that

$$\begin{aligned}
&g_i(x) = \left[\log_\gamma\left[1-h_i\left(\left[h_i^{-1}\left(1-\gamma^{x^{-1}}\right)\right]^{\frac{1}{n-x+1}}\left[h_i^{-1}\left(1-\gamma^{g_i(x+1)^{-1}}\right)\right]^{\frac{n-x}{n-x+1}}\right)\right]\right]^{-1}\\
&\Rightarrow g_i(x) \geq \left[\log_\gamma\left[1-h_i\left(\left[h_i^{-1}\left(1-\gamma^{x^{-1}}\right)\right]^{\frac{1}{n-x+1}}\left[h_i^{-1}\left(1-\gamma^{g_i(x+1)^{-1}}\right)\right]^{\frac{n-x}{n-x+1}}\right)\right]\right]^{-1}\\
&\Leftrightarrow g_i(x)^{-1} \leq \log_\gamma\left[1-h_i\left(\left[h_i^{-1}\left(1-\gamma^{x^{-1}}\right)\right]^{\frac{1}{n-x+1}}\left[h_i^{-1}\left(1-\gamma^{g_i(x+1)^{-1}}\right)\right]^{\frac{n-x}{n-x+1}}\right)\right]\\
&\Leftrightarrow \gamma^{g_i(x)^{-1}} \geq 1-h_i\left(\left[h_i^{-1}\left(1-\gamma^{x^{-1}}\right)\right]^{\frac{1}{n-x+1}}\left[h_i^{-1}\left(1-\gamma^{g_i(x+1)^{-1}}\right)\right]^{\frac{n-x}{n-x+1}}\right)\\
&\Leftrightarrow 1-\gamma^{g_i(x)^{-1}} \leq h_i\left(\left[h_i^{-1}\left(1-\gamma^{x^{-1}}\right)\right]^{\frac{1}{n-x+1}}\left[h_i^{-1}\left(1-\gamma^{g_i(x+1)^{-1}}\right)\right]^{\frac{n-x}{n-x+1}}\right)\\
&\Leftrightarrow h_i^{-1}\left(1-\gamma^{g_i(x)^{-1}}\right) \leq \left[h_i^{-1}\left(1-\gamma^{x^{-1}}\right)\right]^{\frac{1}{n-x+1}}\left[h_i^{-1}\left(1-\gamma^{g_i(x+1)^{-1}}\right)\right]^{\frac{n-x}{n-x+1}}\\
&\Leftrightarrow \frac{\left[h_i^{-1}\left(1-\gamma^{g_i(x)^{-1}}\right)\right]^{n-x+1}}{\left[h_i^{-1}\left(1-\gamma^{g_i(x+1)^{-1}}\right)\right]^{n-x}} \leq h_i^{-1}\left(1-\gamma^{x^{-1}}\right)\\
&\Leftrightarrow h_i^{-1}\left(1-\gamma^{g_i(x)^{-1}}\right)\left[\frac{h_i^{-1}\left(1-\gamma^{g_i(x)^{-1}}\right)}{h_i^{-1}\left(1-\gamma^{g_i(x+1)^{-1}}\right)}\right]^{n-x} \leq h_i^{-1}\left(1-\gamma^{x^{-1}}\right),
\end{aligned} \quad \text{(A.19)}$$

for all $x \in \{1,\ldots,n\}$ and $i \in \{1,\ldots,n\}$. It is worth pointing out that this relation does hold for $x = n$ even though $g_i(n+1)$ is undefined. For clarity, notice that

considering $x = n$ we have

$$\begin{aligned}
& h_i^{-1}\left(1-\gamma^{g_i(n)^{-1}}\right)\left[\frac{h_i^{-1}\left(1-\gamma^{g_i(n)^{-1}}\right)}{h_i^{-1}\left(1-\gamma^{g_i(n+1)^{-1}}\right)}\right]^{n-n} \leq h_i^{-1}\left(1-\gamma^{n^{-1}}\right) \\
& \Leftrightarrow h_i^{-1}\left(1-\gamma^{g_i(n)^{-1}}\right) \leq h_i^{-1}\left(1-\gamma^{n^{-1}}\right) \\
& \Leftrightarrow 1-\gamma^{g_i(n)^{-1}} \leq h_i\left(h_i^{-1}\left(1-\gamma^{n^{-1}}\right)\right) \\
& \Leftrightarrow \gamma^{g_i(n)^{-1}} \geq 1-h_i\left(h_i^{-1}\left(1-\gamma^{n^{-1}}\right)\right) \\
& \Leftrightarrow g_i(n)^{-1} \leq \log_\gamma\left[1-h_i\left(h_i^{-1}\left(1-\gamma^{n^{-1}}\right)\right)\right] \\
& \Leftrightarrow g_i(n) \geq \left(\log_\gamma\left[1-h_i\left(h_i^{-1}\left(1-\gamma^{n^{-1}}\right)\right)\right]\right)^{-1}
\end{aligned} \tag{A.20}$$

and

$$\begin{aligned}
& h_i^{-1}\left(1-\gamma^{n^{-1}}\right) \leq h_i^{-1}\left(1-\gamma^{n^{-1}}\right) \\
& \Leftrightarrow 1-\gamma^{n^{-1}} \leq h_i\left(h_i^{-1}\left(1-\gamma^{n^{-1}}\right)\right) \\
& \Leftrightarrow \gamma^{n^{-1}} \geq 1-h_i\left(h_i^{-1}\left(1-\gamma^{n^{-1}}\right)\right) \\
& \Leftrightarrow n^{-1} \leq \log_\gamma\left[1-h_i\left(h_i^{-1}\left(1-\gamma^{n^{-1}}\right)\right)\right] \\
& \Leftrightarrow n \geq \left(\log_\gamma\left[1-h_i\left(h_i^{-1}\left(1-\gamma^{n^{-1}}\right)\right)\right]\right)^{-1},
\end{aligned} \tag{A.21}$$

so, since $g_i(n) = n$, we have

$$g_i(n) = n \Rightarrow h_i^{-1}\left(1-\gamma^{g_i(n)^{-1}}\right)\left[\frac{h_i^{-1}\left(1-\gamma^{g_i(n)^{-1}}\right)}{h_i^{-1}\left(1-\gamma^{g_i(n+1)^{-1}}\right)}\right]^{n-n} \leq h_i^{-1}\left(1-\gamma^{n^{-1}}\right). \tag{A.22}$$

Now, considering the case where T_{Y_i} is absolutely continuous we can obtain analogous results substituting the inequalities from equivalences A.19 and A.22 for equalities. For brevity, we skip to the conclusion that

$$\begin{aligned}
& g_i(x) = \left[\log_\gamma\left[1-h_i\left(\left[h_i^{-1}\left(1-\gamma^{x^{-1}}\right)\right]^{\frac{1}{n-x+1}}\left[h_i^{-1}\left(1-\gamma^{g_i(x+1)^{-1}}\right)\right]^{\frac{n-x}{n-x+1}}\right)\right]\right]^{-1} \\
& \Leftrightarrow h_i^{-1}\left(1-\gamma^{g_i(x)^{-1}}\right)\left[\frac{h_i^{-1}\left(1-\gamma^{g_i(x)^{-1}}\right)}{h_i^{-1}\left(1-\gamma^{g_i(x+1)^{-1}}\right)}\right]^{n-x} = h_i^{-1}\left(1-\gamma^{x^{-1}}\right),
\end{aligned} \tag{A.23}$$

and

$$g_i(n) = n \Leftrightarrow h_i^{-1}\left(1-\gamma^{g_i(n)^{-1}}\right)\left[\frac{h_i^{-1}\left(1-\gamma^{g_i(n)^{-1}}\right)}{h_i^{-1}\left(1-\gamma^{g_i(n+1)^{-1}}\right)}\right]^{n-n} = h_i^{-1}\left(1-\gamma^{n^{-1}}\right). \tag{A.24}$$

The equivalences A.23 and A.24 also allows us to state that, if T_{Y_i} is an absolutely continuous random variable, the correction function g_i is the unique function satisfying the desired propriety. □

Next we will introduce the *universal correction function* g^*, prove some of its interesting proprieties, and provide some intuition regarding its supposed "universality".

In section 3.1.3 the correction function is usually denoted by g_i to avoid an overloaded notion, however, a more precise notation for the function would be $g_i(\cdot|\theta, \gamma)$, since it has a dependence on the parametric vector θ and on the hyperparameter γ but does not depend on the collection of indicators z. With this explicit dependence, it becomes clear, considering the Metropolis-Hastings algorithm considered in section 3.2, that the values of the correction function need to be recalculated at every iteration for the acceptance-rejection step, due to the change of the current value $\theta^{(t-1)}$ and/or the proposed value θ_{prop}. Furthermore, we know that no general analytical solution for the correction, quantile and distribution functions for auto-transformations can be provided, drastically increasing the expected computational cost of model. Having that in mind, if instead we were able to provide a reasonable approximation g^* of the correction function that had no dependence on θ, γ and z, only depending on the total sample size n, we could avoid the additional cost of recalculating the value of $g_i(\cdot|\theta, \gamma)$ for each iteration of the Metropolis-Hastings algorithm and each sample. Better yet, the values $g^*(1), \ldots, g^*(n)$ could be calculated preceding the algorithm's initialization, then stored and accessed on demand, completely surpassing the need of computing the correction function throughout the iterations.

Even though we do not have a proven best approximation or even the error control for our proposed universal correction function g^*, we do have theoretical reasons to believe that $g^*(x)$ is a good approximation of $g_i(x|\theta, \gamma)$ for large values of n and x. Our proposal relies on Extreme Value Theory, see de Haan & Ferreira (2006) for an introduction on the topic, and uses a correction function based on the Weibull distribution as an approximation. Consider now the following central

theorem from Extreme Value Theory:

Theorem A.1 (Fisher & Tippett (1928) and Gnedenko (1943))**.** *Let $X_1, X_2 \dots$ be a sequence of independent and identically distributed random variables. Suppose there exists a sequence of constants $a_n > 0$, and $b_n \in \mathbb{R}$ (for $n \in \{1, \dots, n\}$) such that*

$$\frac{\max\{X_1, \dots, X_n\} - b_n}{a_n} \tag{A.25}$$

has a non degenerate limit distribution, i.e.

$$\lim_{n \to +\infty} F^n(a_n x + b_n) = G(x), \tag{A.26}$$

for every point x of G, and G a non degenerate distribution function. Then, the class of possible limiting distributions G is given by

$$G(x|a, b, \xi) = \exp\left\{ -\left(1 + \xi\left(\frac{x-b}{a}\right)\right)^{-\frac{1}{\xi}} \right\}, \quad 1 + \gamma x > 0, \tag{A.27}$$

where $a > 0$, $b, \xi \in \mathbb{R}$ and the expression inside the exponent should be interpreted as $\exp\{-\frac{x-b}{a}\}$.

With the theorem above we know that, if we consider the normalized maximum of any random variable it can only converge to a fairly restricted family of distributions, which is the generalized extreme value distribution. As a side note, the limit of the appropriately normalized minimum of independent random variables can be obtained from the results for maximum using that

$$\min\{X_1, \dots, X_n\} = -\max\{-X_1, \dots, -X_n\}. \tag{A.28}$$

Now, since for any d-dimensional absolutely continuous or discrete random variable X we have $T_X^{(n)} = \min\{T_{X_1}, \dots, T_{X_n}\}$, where $T_{X_1}, \dots, T_{X_n}$ are independent autotransformations of X, we can speculate that for large n the autotransformation of order n will behave similarly to one of the possible limits of G for the minimum. It can also be proven, see de Haan & Ferreira (2006) for more details, that if the limit exists in this context, the Gumbel and the Weibull are the only possible limiting distributions, because $T_X^{(n)} > 0$ almost surely, justifying the use of the Weibull

distribution for our next results. So, if we take $Y \sim Weibull(\alpha, \beta)$, considering the parametrization for the density given by

$$f_Y(y|\alpha, \beta) = \beta\alpha y^{\alpha-1} \exp\left\{-\beta x^{\alpha}\right\}, \quad y > 0, \tag{A.29}$$

where $\alpha > 0$ and $\beta > 0$, we know that for its quantile function we have

$$F_Y^{-1}(x) = \left[-\frac{1}{\beta}\ln(1-x)\right]^{\frac{1}{\alpha}}. \tag{A.30}$$

Next, we define the universal correction function based on the Weibull distribution as follows.

Definition A.1 (Universal Correction Function)**.** Let $Y_1, \ldots, Y_n$ be independent d-dimensional absolutely continuous or discrete distributions random variables satisfy the filtering condition given a parameteric vector θ. Then the *universal correction function* g^* for any of the Y_i's is given by

$$g^*(x) = x^{\frac{1}{n-x+1}} \left[g^*(x+1)\right]^{\frac{n-x}{n-x+1}}, \tag{A.31}$$

for $x \in \{1, \ldots, n\}$.

It is worth highlighting from A.1 that the universal correction function does not depend on the unknown parametric vector θ or on a hypeparameter $\gamma \in (0, 1)$, unlike most of the correction functions, thus accomplishing our simplification goal. It also does not depend in any form on the distribution of the sample $Y_1, \ldots, Y_n$, only on its size, hence the *universal* in its name. Next, we provide a result analogous to proposition A.4, that allows us to better understand the purpose of Definition A.1 and establish its connection to the Weibull distribution.

Proposition A.5. *Let $Y_1, \ldots, Y_n$ be independent random variables, g^* the universal correction function and $\gamma \in (0, 1)$ a real scalar. If $Y_1, \ldots, Y_n$ are such that $T_{Y_i} \sim Weibull(\alpha_i, \beta_i)$ for $i \in \{1, \ldots, n\}$, then the correction function for Y_i satisfies:*

$$g_i(x|\alpha, \beta, \gamma) = g_i^*(x), \tag{A.32}$$

for all $x \in \{1, \ldots, n\}$.

Proof. First, notice that for $x = n$ we have

$$g^*(n) = n^{\frac{1}{n-n+1}} \left[g^*(n+1)\right]^{\frac{n-n}{n-n+1}} = n = g_i(n). \tag{A.33}$$

Now, from proposition A.4 we know that $g_i = g_i(\cdot|\alpha_i, \beta_i)$ is the only function that satisfies

$$F^{-1}_{T_{Y_i}}\left(1-\gamma^{g_i(x)^{-1}}\middle|\,\alpha_i,\beta_i\right)\left[\frac{F^{-1}_{T_{Y_i}}\left(1-\gamma^{g_i(x)^{-1}}\middle|\,\alpha_i,\beta_i\right)}{F^{-1}_{T_{Y_i}}\left(1-\gamma^{g_i(x+1)^{-1}}|\,\alpha_i,\beta_i\right)}\right]^{n-x} = F^{-1}_{T_{Y_i}}\left(1-\gamma^{x^{-1}}\middle|\,\alpha_i,\beta_i\right) \tag{A.34}$$

for all $x \in \{1,\ldots,n\}$, so we have

$$\begin{aligned}
&F^{-1}_{T_{Y_i}}\left(1-\gamma^{g_i(x)^{-1}}\middle|\,\alpha_i,\beta_i\right)\left[\frac{F^{-1}_{T_{Y_i}}\left(1-\gamma^{g_i(x)^{-1}}\middle|\,\alpha_i,\beta_i\right)}{F^{-1}_{T_{Y_i}}\left(1-\gamma^{g_i(x+1)^{-1}}|\,\alpha_i,\beta_i\right)}\right]^{n-x} = F^{-1}_{T_{Y_i}}\left(1-\gamma^{x^{-1}}\middle|\,\alpha_i,\beta_i\right)\\
&\Leftrightarrow \left[-\frac{1}{\beta_i}\ln\left(1-\left(1-\gamma^{g_i(x)^{-1}}\right)\right)\right]^{\frac{1}{\alpha_i}}\left[\frac{\left[-\frac{1}{\beta_i}\ln\left(1-\left(1-\gamma^{g_i(x)^{-1}}\right)\right)\right]^{\frac{1}{\alpha_i}}}{\left[-\frac{1}{\beta_i}\ln\left(1-\left(1-\gamma^{g_i(x+1)^{-1}}\right)\right)\right]^{\frac{1}{\alpha_i}}}\right]^{n-x}\\
&= \left[-\frac{1}{\beta_i}\ln\left(1-\left(1-\gamma^{x^{-1}}\right)\right)\right]^{\frac{1}{\alpha_i}}\\
&\Leftrightarrow \left(-\frac{1}{\beta_i}\ln\left(\gamma^{g_i(x)^{-1}}\right)\left[\frac{-\frac{1}{\beta_i}\ln\left(\gamma^{g_i(x)^{-1}}\right)}{-\frac{1}{\beta_i}\ln\left(\gamma^{g_i(x+1)^{-1}}\right)}\right]^{n-x}\right)^{\frac{1}{\alpha_i}} = \left[-\frac{1}{\beta_i}\ln\left(\gamma^{x^{-1}}\right)\right]^{\frac{1}{\alpha_i}}\\
&\Leftrightarrow \ln\left(\gamma^{g_i(x)^{-1}}\right)\left[\frac{\ln\left(\gamma^{g_i(x)^{-1}}\right)}{\ln\left(\gamma^{g_i(x+1)^{-1}}\right)}\right]^{n-x} = \ln\left(\gamma^{x^{-1}}\right)\\
&\Leftrightarrow g_i(x)^{-1}\ln(\gamma)\left[\frac{g_i(x)^{-1}\ln(\gamma)}{g_i(x+1)^{-1}\ln(\gamma)}\right]^{n-x} = x^{-1}\ln(\gamma)\\
&\Leftrightarrow g_i(x)^{-1}\frac{g_i(x)^{-(n-x)}}{g_i(x+1)^{-(n-x)}} = x^{-1}\\
&\Leftrightarrow g_i(x) = x^{\frac{1}{n-x+1}}\left[g_i(x+1)\right]^{\frac{n-x}{n-x+1}}
\end{aligned} \tag{A.35}$$

Since g_i and g^* coincide for the initial value $x = n$ and their recursive formulas are identical, then they must be the same for all $x \in \{1,\ldots,n\}$. □

It is worth noting that proposition A.5 shows that the universal correction function does not depend on the choices of α_i, β_i and γ. This allows us to conclude that,

if there is a Weibull distribution with parameters $\alpha_i(\theta,\gamma)$ and $\beta_i(\theta,\gamma)$ that approximates well the distribution $F_{T_{Y_i}}(\cdot|\theta)$, then g^* should be a reasonable approximation of $g_i(\cdot|\theta,\gamma)$.

As a side note, the numerical values for the universal correction function can also be computed using its initial value and then the recursive formula, but we recommend using the equivalent, albeit more numerically stable, expression given by

$$g^*(x) = \exp\left\{\frac{1}{n-x+1}\ln(x) + \frac{n-x}{n-x+1}\ln\left[g^*(x+1)\right]\right\}, \tag{A.36}$$

for $x \in \{1,\ldots,n\}$.

A.3 LOCATION-SCALE MODELS

In this section we will present some useful proprieties related to the autotransformations of location-scale models. So let Z be a d-dimensional absolutely continuous random vector, $\mu \in \mathbb{R}^d$ be a location parameter and A be a $d \times d$ lower triangular matrix with positive diagonal entries such that $\Sigma = AA'$ is a positive-definite matrix. Then, if $X = AZ + \mu$, the density of X is given by

$$f_X(x|\mu,\Sigma) = \left[\det(\Sigma)\right]^{-\frac{1}{2}} f_Z\left(A^{-1}(x-\mu)\right). \tag{A.37}$$

With the expression above, we can easily rewrite the autotransformation of X in terms of the autotransformation of Z considering that

$$\begin{aligned} T_X = f_X(X|\mu,\Sigma) &= \left[\det(\Sigma)\right]^{-\frac{1}{2}} f_Z\left(A^{-1}(X-\mu)\right) = \left[\det(\Sigma)\right]^{-\frac{1}{2}} f_Z\left(Z\right) \\ &= \left[\det(\Sigma)\right]^{-\frac{1}{2}} T_Z. \end{aligned} \tag{A.38}$$

Here, it is worth noting that T_X does not depend on the location parameter μ and its dependence on Σ is only multiplicative, leading to some straightforward way of finding the distribution and quantile functions of T_X. And this can be done considering that

$$\begin{aligned} F_{T_X}(x|\Sigma) &= \mathbb{P}\left(T_X \leq x|\,\Sigma\right) = \mathbb{P}\left(\left[\det(\Sigma)\right]^{-\frac{1}{2}} T_Z \leq x \middle|\, \Sigma\right) \\ &= \mathbb{P}\left(T_Z \leq \left[\det(\Sigma)\right]^{\frac{1}{2}} x \middle|\, \Sigma\right) = F_{T_Z}\left(\left[\det(\Sigma)\right]^{\frac{1}{2}} x\right) \end{aligned} \tag{A.39}$$

and, consequently, we have

$$F_{T_X}^{-1}(x|\Sigma) = [\det(\Sigma)]^{-\frac{1}{2}} F_{T_Z}^{-1}(x)\,. \tag{A.40}$$

Let us now consider the case of discrete d-dimensional random vector X whose domain is the set $S_X \subset \mathbb{R}^d$. In this case, the autotransformation T_X is also a discrete random variable whose domain is the set

$$S_{T_X} = \{y \in \mathbb{R} : S_X^*(y) \neq \emptyset\}\,, \tag{A.41}$$

where

$$S^*(y) = \left\{x \in \mathbb{R}^d : \mathbb{P}(X = x) = y\right\}. \tag{A.42}$$

With this, we can calculate probabilities involving T_X noticing that, for all $y \in \mathbb{R}$,

$$\mathbb{P}(T_X = y) = \sum_{x \in S_X^*(y)} \mathbb{P}(X = x). \tag{A.43}$$

Interestingly, if we then consider $Y = t(X)$, where t is a reversible function, then

$$\begin{aligned}\mathbb{P}(T_Y = y) &= \sum_{x \in S_Y^*(y)} \mathbb{P}(Y = x) = \sum_{x \in S_Y^*(y)} \mathbb{P}(t(X) = x) \\ &= \sum_{x \in S_Y^*(y)} \mathbb{P}(X = t^{-1}(x)) = \sum_{x \in S_X^*(y)} \mathbb{P}(X = x) = \mathbb{P}(T_X = y)\end{aligned} \tag{A.44}$$

for all $y \in \mathbb{R}$, which implies that T_Y has the same distribution of T_X. And, in particular, if we consider an affine reversible transformation the resulting distribution will not change as well, so the autotransformation of a discrete random variable with location and scale parameters does not depend on either of them. We can also interpret this propriety as a "label invariance" of discrete distributions, i.e., if we rename the elements of S_X and change the notion of order the distribution of the autotransformation does not change, since we do not alter the way we distribute probabilities in the transformed set when compared to the original one. Next we present the last result related to location-scale models of this work.

Proposition A.6. *Let $Y_1, \ldots, Y_n$ be an independent sample from the d-dimensional absolutely continuous or discrete distributions $F_{Y_1}(\cdot|\theta), \ldots, F_{Y_n}(\cdot|\theta)$ and $\gamma \in (0, 1)$ a real scalar. Besides, let $\mu_1, \ldots, \mu_n \in \mathbb{R}^d$ be location vectors, $A_1, \ldots, A_n$ be $d \times d$*

lower triangular matrices with positive diagonal entries such that $\Sigma_k = A_k A_k'$ *is a positive-definite matrix, for all* $k \in \{1, \ldots, n\}$*, and* $\theta = (A_1, \ldots, A_n, \mu_1, \ldots, \mu_n)$*. If a* d*-dimensional random vector* Z *whose distribution does not depend on* θ *and such that* $Y_k = A_k Z + \mu_k$ *exists for all* $k \in \{1, \ldots, n\}$*, then the correction function* g_k *for* Y_k *satisfies*

$$\begin{aligned} & F^{-1}_{T_{Y_k}}\left(1-\gamma^{g_k(n_1^{-k}+1)^{-1}}\middle|\,\theta\right)\prod_{i\neq k}\left[\frac{F^{-1}_{T_{Y_i}}\left(1-\gamma^{g_i(n_1^{-k}+1)^{-1}}\middle|\,\theta\right)}{F^{-1}_{T_{Y_i}}\left(1-\gamma^{g_i(n_1^{-k}+2)^{-1}}\middle|\,\theta\right)}\right]^{1-z_i} \\ & \leq F^{-1}_{T_{Y_k}}\left(1-\gamma^{(n_1^{-k}+1)^{-1}}\middle|\,\theta\right), \end{aligned} \tag{A.45}$$

for all $n_1^{-k} \in \{0, \ldots, n-1\}$ *and* $k \in \{1, \ldots, n\}$*, where* $z_1, \ldots, z_n \in \{0, 1\}$ *and are such that* $n_1^{-k} = \sum_{i \neq k} z_i$*. Furthermore, if* T_Z *is absolutely continuous we have*

$$\begin{aligned} & F^{-1}_{T_{Y_k}}\left(1-\gamma^{g_k(n_1^{-k}+1)^{-1}}\middle|\,\theta\right)\prod_{i\neq k}\left[\frac{F^{-1}_{T_{Y_i}}\left(1-\gamma^{g_i(n_1^{-k}+1)^{-1}}\middle|\,\theta\right)}{F^{-1}_{T_{Y_i}}\left(1-\gamma^{g_i(n_1^{-k}+2)^{-1}}\middle|\,\theta\right)}\right]^{1-z_i} \\ & = F^{-1}_{T_{Y_k}}\left(1-\gamma^{(n_1^{-k}+1)^{-1}}\middle|\,\theta\right), \end{aligned} \tag{A.46}$$

for all $n_1^{-k} \in \{0, \ldots, n-1\}$ *and* $k \in \{1, \ldots, n\}$*.*

Proof. Let us first show that the correction functions does not depend on θ. For the discrete case we already know that $T_{Y_k} = T_Z$, so, since the correction function g_k depends on Y_k only through $F_{T_{Y_k}}(\cdot|\theta) = F_{T_Z}$, we can trivially see that it does not depend on θ. Now, assuming Y_k absolutely continuous and adopting the notation $h_k = F_{T_{Y_k}}(\cdot|\theta)$ for simplicity, from Definition 3.5 and equations A.39 and A.40 we have

$$\begin{aligned} & g_k(x) = \left(\log_\gamma\left[1-h_k\left(\left[h_k^{-1}\left(1-\gamma^{x^{-1}}\right)\right]^{\frac{1}{n-x+1}}\left[h_k^{-1}\left(1-\gamma^{g_k(x+1)^{-1}}\right)\right]^{\frac{n-x}{n-x+1}}\right)\right]\right)^{-1} \\ & \Leftrightarrow 1-\gamma^{g_k(x)^{-1}} = h_k\left(\left[h_k^{-1}\left(1-\gamma^{x^{-1}}\right)\right]^{\frac{1}{n-x+1}}\left[h_k^{-1}\left(1-\gamma^{g_k(x+1)^{-1}}\right)\right]^{\frac{n-x}{n-x+1}}\right) \\ & = h_k\left(\left[[\det(\Sigma_k)]^{-\frac{1}{2}}F_{T_Z}^{-1}\left(1-\gamma^{x^{-1}}\right)\right]^{\frac{1}{n-x+1}}\left[[\det(\Sigma_k)]^{-\frac{1}{2}}F_{T_Z}^{-1}\left(1-\gamma^{g_k(x+1)^{-1}}\right)\right]^{\frac{n-x}{n-x+1}}\right) \\ & = F_{T_Z}\left[[\det(\Sigma_k)]^{\frac{1}{2}}[\det(\Sigma_k)]^{-\frac{1}{2}}\left[F_{T_Z}^{-1}\left(1-\gamma^{x^{-1}}\right)\right]^{\frac{1}{n-x+1}}\left[F_{T_Z}^{-1}\left(1-\gamma^{g_k(x+1)^{-1}}\right)\right]^{\frac{n-x}{n-x+1}}\right] \\ & = F_{T_Z}\left(\left[F_{T_Z}^{-1}\left(1-\gamma^{x^{-1}}\right)\right]^{\frac{1}{n-x+1}}\left[F_{T_Z}^{-1}\left(1-\gamma^{g_k(x+1)^{-1}}\right)\right]^{\frac{n-x}{n-x+1}}\right), \end{aligned} \tag{A.47}$$

which implies that

$$g_k(x) = \left[\log_\gamma \left[1 - F_{T_Z}\left(\left[F_{T_Z}^{-1}\left(1-\gamma^{x^{-1}}\right)\right]^{\frac{1}{n-x+1}}\left[F_{T_Z}^{-1}\left(1-\gamma^{g_k(x+1)^{-1}}\right)\right]^{\frac{n-x}{n-x+1}}\right)\right]\right]^{-1} \tag{A.48}$$

for all $x \in \{1,\ldots,n-1\}$, and, since $g_k(n) = n$, that the correction functions are identical for all $k \in \{1,\ldots,n\}$. We can also notice, from the expressions obtained in absolutely continuous case and from the argument used in the discrete case, that g_k is the same as the correction function for Z (assuming a sample of n independent random variables with the same distribution), so we denote it g instead.

Now, from proposition A.4 we have

$$F_{T_Z}^{-1}\left(1-\gamma^{g(x)^{-1}}\right)\left[\frac{F_{T_Z}^{-1}\left(1-\gamma^{g(x)^{-1}}\right)}{F_{T_Z}^{-1}\left(1-\gamma^{g_i(x+1)^{-1}}\right)}\right]^{n-x} \leq F_{T_Z}^{-1}\left(1-\gamma^{x^{-1}}\right), \tag{A.49}$$

so, assuming $Y_1,\ldots,Y_n$ discrete, we obtain that

$$\begin{aligned}
&F_{T_{Y_k}}^{-1}\left(1-\gamma^{g_k(n_1^{-k}+1)^{-1}}\middle|\,\theta\right)\prod_{i\neq k}\left[\frac{F_{T_{Y_i}}^{-1}\left(1-\gamma^{g_i(n_1^{-k}+1)^{-1}}\middle|\,\theta\right)}{F_{T_{Y_i}}^{-1}\left(1-\gamma^{g_i(n_1^{-k}+2)^{-1}}\middle|\,\theta\right)}\right]^{1-z_i}\\
&= F_{T_Z}^{-1}\left(1-\gamma^{g(n_1^{-k}+1)^{-1}}\right)\prod_{i\neq k}\left[\frac{F_{T_Z}^{-1}\left(1-\gamma^{g(n_1^{-k}+1)^{-1}}\right)}{F_{T_Z}^{-1}\left(1-\gamma^{g(n_1^{-k}+2)^{-1}}\right)}\right]^{1-z_i}\\
&= F_{T_Z}^{-1}\left(1-\gamma^{g(n_1^{-k}+1)^{-1}}\right)\left[\frac{F_{T_Z}^{-1}\left(1-\gamma^{g(n_1^{-k}+1)^{-1}}\right)}{F_{T_Z}^{-1}\left(1-\gamma^{g(n_1^{-k}+2)^{-1}}\right)}\right]^{n-(n_1^{-k}+1)}\\
&\leq F_{T_Z}^{-1}\left(1-\gamma^{(n_1^{-k}+1)^{-1}}\right) = F_{T_{Y_k}}^{-1}\left(1-\gamma^{(n_1^{-k}+1)^{-1}}\middle|\,\theta\right),
\end{aligned} \tag{A.50}$$

for all $n_1^{-k} \in \{0,\ldots,n-1\}$ and $k \in \{1,\ldots,n\}$. Then, assuming $Y_1,\ldots,Y_n$ continu-

ous, we have

$$
\begin{aligned}
&F_{T_{Y_k}}^{-1}\left(1-\gamma^{g_k(n_1^{-k}+1)^{-1}}\middle|\,\theta\right)\prod_{i\neq k}\left[\frac{F_{T_{Y_i}}^{-1}\left(1-\gamma^{g_i(n_1^{-k}+1)^{-1}}\middle|\,\theta\right)}{F_{T_{Y_i}}^{-1}\left(1-\gamma^{g_i(n_1^{-k}+2)^{-1}}\middle|\,\theta\right)}\right]^{1-z_i}\\
&=\left[\det(\Sigma_k)\right]^{-\frac{1}{2}}F_{T_Z}^{-1}\left(1-\gamma^{g(n_1^{-k}+1)^{-1}}\right)\prod_{i\neq k}\left[\frac{\left[\det(\Sigma_i)\right]^{-\frac{1}{2}}F_{T_Z}^{-1}\left(1-\gamma^{g(n_1^{-k}+1)^{-1}}\right)}{\left[\det(\Sigma_i)\right]^{-\frac{1}{2}}F_{T_Z}^{-1}\left(1-\gamma^{g(n_1^{-k}+2)^{-1}}\right)}\right]^{1-z_i}\\
&=\left[\det(\Sigma_k)\right]^{-\frac{1}{2}}F_{T_Z}^{-1}\left(1-\gamma^{g(n_1^{-k}+1)^{-1}}\right)\left[\frac{F_{T_Z}^{-1}\left(1-\gamma^{g(n_1^{-k}+1)^{-1}}\right)}{F_{T_Z}^{-1}\left(1-\gamma^{g(n_1^{-k}+2)^{-1}}\right)}\right]^{n-(n_1^{-k}+1)}\\
&\leq\left[\det(\Sigma_k)\right]^{-\frac{1}{2}}F_{T_Z}^{-1}\left(1-\gamma^{(n_1^{-k}+1)^{-1}}\right)=F_{T_{Y_k}}^{-1}\left(1-\gamma^{(n_1^{-k}+1)^{-1}}\middle|\,\theta\right),
\end{aligned}
\tag{A.51}
$$

for all $n_1^{-k}\in\{0,\ldots,n-1\}$ and $k\in\{1,\ldots,n\}$. At last, assuming T_Z absolutely continuous, from proposition A.4 we also know that

$$
F_{T_Z}^{-1}\left(1-\gamma^{g(x)^{-1}}\right)\left[\frac{F_{T_Z}^{-1}\left(1-\gamma^{g(x)^{-1}}\right)}{F_{T_Z}^{-1}\left(1-\gamma^{g_i(x+1)^{-1}}\right)}\right]^{n-x}=F_{T_Z}^{-1}\left(1-\gamma^{x^{-1}}\right).
\tag{A.52}
$$

So we can obtain the desired result repeating the procedures above, but considering the equality instead. For brevity, with jump to conclude that

$$
\begin{aligned}
&F_{T_{Y_k}}^{-1}\left(1-\gamma^{g_k(n_1^{-k}+1)^{-1}}\middle|\,\theta\right)\prod_{i\neq k}\left[\frac{F_{T_{Y_i}}^{-1}\left(1-\gamma^{g_i(n_1^{-k}+1)^{-1}}\middle|\,\theta\right)}{F_{T_{Y_i}}^{-1}\left(1-\gamma^{g_i(n_1^{-k}+2)^{-1}}\middle|\,\theta\right)}\right]^{1-z_i}\\
&=F_{T_{Y_k}}^{-1}\left(1-\gamma^{(n_1^{-k}+1)^{-1}}\middle|\,\theta\right),
\end{aligned}
\tag{A.53}
$$

for all $n_1^{-k}\in\{0,\ldots,n-1\}$ and $k\in\{1,\ldots,n\}$, finishing our demonstration. □

At last, considering the full conditional distribution in equation 3.26, it is not hard to see that proposition A.6 implies that the filtering model is always unbiased for location-scale models, regardless of whether the observations are identically distributed or not. Furthermore, proving of proposition A.6 we also show that the correction function for each observation depends only on the distribution of Z and on the hyperparameter γ, losing its dependence on the models parameters.

A.4 MULTIVARIATE NORMAL DISTRIBUTION

Considering the case of the multivariate normal distribution, we can find its distribution in terms of a known distribution function. Let $X \sim Normal_d(\mu, \Sigma)$, then it is known that $Y = (X-\mu)'\Sigma^{-1}(X-\mu) \sim Gamma\left(\frac{d}{2}, \frac{1}{2}\right)$, see Chapter 4 of Gamerman & Lopes (2006) for instance. Now let $T_X = f_X(X|\mu, \Sigma)$ be the autotransformation of X. We can then obtain f_{T_X} considering that, for $0 < x < (2\pi)^{-\frac{d}{2}} \left[\det(\Sigma)\right]^{-\frac{1}{2}}$,

$$\begin{aligned}
F_{T_X}(x) &= \mathbb{P}\left(T_X \leq x\right) = \mathbb{P}\left(f_X(X|\mu,\Sigma) \leq x\right) \\
&= \mathbb{P}\left((2\pi)^{-\frac{d}{2}} \left[\det(\Sigma)\right]^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(X-\mu)\Sigma^{-1}(X-\mu)\right\} \leq x\right) \\
&= \mathbb{P}\left((2\pi)^{-\frac{d}{2}} \left[\det(\Sigma)\right]^{-\frac{1}{2}} \exp\left\{-\frac{Y}{2}\right\} \leq x\right) \\
&= \mathbb{P}\left(\exp\left\{-\frac{Y}{2}\right\} \leq (2\pi)^{\frac{d}{2}} \left[\det(\Sigma)\right]^{\frac{1}{2}} x\right) \\
&= \mathbb{P}\left(Y \geq -2\ln\left[(2\pi)^{\frac{d}{2}} \left[\det(\Sigma)\right]^{\frac{1}{2}} x\right]\right) \\
&= 1 - F_Y\left(-2\ln\left[(2\pi)^{\frac{d}{2}} \left[\det(\Sigma)\right]^{\frac{1}{2}} x\right]\right).
\end{aligned} \tag{A.54}$$

Besides this, after noticing that $0 \leq f_X(x) \leq (2\pi)^{-\frac{d}{2}} \left[\det(\Sigma)\right]^{-\frac{1}{2}}$, for all $x \in \mathbb{R}^d$, it can be trivially seen that $F_{T_X}(x) = 0$, for all $x \leq 0$, and that $F_{T_X}(x) = 1$, for $x \geq (2\pi)^{-\frac{d}{2}} \left[\det(\Sigma)\right]^{-\frac{1}{2}}$.

We can find the quantile function as well, considering that

$$\begin{aligned}
F_{T_X}(x) = y &\iff 1 - F_Y\left(-2\ln\left[(2\pi)^{\frac{d}{2}} \left[\det(\Sigma)\right]^{\frac{1}{2}} x\right]\right) = y \\
&\iff -2\ln\left[(2\pi)^{\frac{d}{2}} \left[\det(\Sigma)\right]^{\frac{1}{2}} x\right] = F_Y^{-1}(1-y) \\
&\iff (2\pi)^{\frac{d}{2}} \left[\det(\Sigma)\right]^{\frac{1}{2}} x = \exp\left\{-\frac{1}{2}F_Y^{-1}(1-y)\right\} \\
&\iff x = (2\pi)^{-\frac{d}{2}} \left[\det(\Sigma)\right]^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}F_Y^{-1}(1-y)\right\},
\end{aligned} \tag{A.55}$$

and thus, we conclude that

$$F_{T_X}^{-1}(x) = (2\pi)^{-\frac{d}{2}} \left[\det(\Sigma)\right]^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}F_Y^{-1}(1-x)\right\}. \tag{A.56}$$

Besides, we can use proposition A.2 to obtain

$$F_{T_X^{(n)}}^{-1}(x) = (2\pi)^{-\frac{d}{2}} \left[\det(\Sigma)\right]^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}F_Y^{-1}\left((1-x)^{\frac{1}{n}}\right)\right\}. \tag{A.57}$$

B MODELS FOR APPLICATIONS

In this Appendix we detail the models used for the applications in Chapter 4, including the priors used and the necessary full conditionals for the derivation of the Gibbs sampler.

B.1 RANDOM WALK MODEL

Let us consider the filtering model with the normal random walk model as the main component. We can represent the resulting model, including all of the chosen prior distributions for each parameter, in the hierarchical form given by

$$\begin{aligned}
Y_i|\mu,\phi,z_i=1 &\overset{ind}{\sim} Normal_d\left(\mu, diag(\phi)^{-1}\right),\\
Y_i|\phi,z_{-i},z_i=0 &\overset{ind}{\sim} Uniform(S_i),\\
\mu_j|\mu_{j-1},\phi_\omega &\sim Normal\left(\mu_{j-1},\phi_\omega^{-1}\right),\\
\mu_0 &\sim Normal(m_0,C_0),\\
\phi_j &\overset{ind}{\sim} Gamma\left(\frac{a_j}{2},\frac{b_j}{2}\right),\\
\phi_\omega &\sim Gamma\left(\frac{a_\omega}{2},\frac{b_\omega}{2}\right),\\
z_i &\overset{ind}{\sim} Bernoulli(w_z),\\
\mu_L(S_i)^{-1} &= (2\pi)^{-\frac{d}{2}}\left[\prod_{j=1}^{d}\phi_j\right]^{\frac{1}{2}}\exp\left\{-\frac{1}{2}F_G^{-1}\left(\gamma^{g_i(n_1+1)^{-1}}\right)\right\}.
\end{aligned} \tag{B.1}$$

for $i\in\{1,\ldots,n\}$ and $j\in\{1,\ldots,d\}$, where $n_1=\sum_{i=1}^{n}z_i$, F_G represents the distribution function of a $Gamma\left(\frac{d}{2},\frac{1}{2}\right)$, g_i is the correction function considering the assumed distribution for each sample element and μ_L is the Lebesgue measure. We also denote μ the collection of scalars $(\mu_0,\ldots,\mu_d)$, ϕ the d-dimensional vector $(\phi_1,\ldots,\phi_d)'$ and z the collection of indicators $z_1,\ldots,z_n$. Here, the real scalar m_0, the positive scalars C_0, $a_1,\ldots,a_d$, $b_1,\ldots,b_d$, a_ω, b_ω and the scalars $w_z,\gamma\in(0,1)$ are hyperparameters that require specification. It is also worth restating that the function $diag:\mathbb{R}\to\mathbb{M}_{n\times n}(\mathbb{R})$, where $\mathbb{M}_{n\times n}(\mathbb{R})$ is the set of all real valued $n\times n$

matrices, is defined as

$$diag(x) = diag\left((x_1, \ldots, x_n)'\right) = \begin{bmatrix} x_1 & 0 & \cdots & 0 & 0 \\ 0 & x_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & x_{n-1} & 0 \\ 0 & 0 & \cdots & 0 & x_n \end{bmatrix}, \tag{B.2}$$

for every vector $x = (x_1, \ldots, x_n)' \in \mathbb{R}^n$. Furthermore, we assume the dependence structure necessary to obtain the following prior factorization:

$$\begin{aligned} \pi(\mu, \phi, \phi_\omega, z) &\overset{ind}{=} \pi(\mu, \phi_\omega)\pi(\phi)\pi(z) = \pi(\mu|\phi_\omega)\pi(\phi_\omega)\pi(\phi)\pi(z) \\ &\overset{ind}{=} \prod_{j=1}^{d}\left[\pi(\mu_j|\mu_{j-1}, \phi_\omega)\right]\pi(\mu_0)\pi(\phi_\omega)\prod_{j=1}^{d}\left[\pi(\phi_j)\right]\prod_{i=1}^{n}\left[\pi(z_i)\right]. \end{aligned} \tag{B.3}$$

Next, assuming an observed sample y, a collection of d-dimensional vector $y_1, \ldots, y_n$, and representing our parameters $\Theta = (\mu, \phi, \phi_\omega, z)$, we can use Bayes' Theorem to express the posterior as

$$\begin{aligned} \pi(\Theta|y) &\overset{Bayes}{=} \frac{\pi(\Theta)\pi(y|\Theta)}{\int \pi(\Theta)\pi(y|\Theta)\, d\Theta} \propto \pi(\Theta)\pi(y|\Theta) \\ &= \pi(\mu, \phi, \phi_\omega, z)\pi(y|\mu, \phi, \phi_\omega, z) \\ &= \prod_{j=1}^{d}\left[\pi(\mu_j|\mu_{j-1}, \phi_\omega)\right]\pi(\mu_0)\pi(\phi_\omega)\prod_{j=1}^{d}\left[\pi(\phi_j)\right]\prod_{i=1}^{n}\left[\pi(z_i)\right] \\ &\times \prod_{i=1}^{n}\left[\pi(y_i|\mu, \phi, z_i = 1)\right]^{z_i}\prod_{i=1}^{n}\left[\pi(y_i|\phi, z_i = 0, z_{-i})\right]^{1-z_i} \end{aligned} \tag{B.4}$$

Then, with the expression for the posterior acquired, we begin obtaining the full conditional for μ by noticing that, given ϕ, ϕ_ω and z, we can rewrite the model in the form of a dynamic linear model with known evolution matrices, covariance matrices and drift vectors as follows:

$$\begin{aligned} Y_t^1 &= F_t^1\theta_t + \nu_t^1, \quad & \nu_t^1 &\overset{ind}{\sim} Normal_{n_1}\left(v_t^1, V_t^1\right), \\ \theta_t &= G_t\theta_{t-1} + \omega_t, \quad & \omega_t &\overset{ind}{\sim} Normal_n(w_t, W_t), \\ & & \theta_0 &\sim Normal_n(m_0, C_0), \end{aligned} \tag{B.5}$$

where $n_1 = \sum_{i=1}^{n} z_{it}$, $\theta_t = \mu_t$, the superscript ^{1} indicates that only the rows i such that $z_i = 1$ are considered from the corresponding column vectors/matrices and we

have

$$F_t = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}_{n\times 1}, \quad V_t = diag(\phi)^{-1}, \quad v_t = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}_{n\times 1}, \tag{B.6}$$
$$G_t = \begin{bmatrix} 1 \end{bmatrix}_{1\times 1}, \quad W_t = \begin{bmatrix} \phi_\omega^{-1} \end{bmatrix}_{1\times 1}, \quad w_t = \begin{bmatrix} 0 \end{bmatrix}_{1\times 1}.$$

Then, we are able to apply the FFBS algorithm described in Chapter 15 of West & Harrison (1997) to directly sample from the full conditional of μ.

As a side note, since the filtering component does not depend on location parameters, the full conditional of μ is the usual full conditional obtained for the original random walk model, with the modification of only taking the observations from the main component in consideration, i.e., it considers all of the y_i's such that $z_i = 0$ as if they were not observed.

Then, considering the full conditional of ϕ, notice that

$$\begin{aligned}
&\pi(\phi|\mu,\phi_\omega,z,y) \propto \pi(\mu,\phi,\phi_\omega,z|y) \\
&\propto \prod_{j=1}^{d}\left[\pi(\phi_j)\right]\prod_{i=1}^{n}\left[\pi(y_i|\mu,\phi,z_i=1)\right]^{z_i}\prod_{i=1}^{n}\left[\pi(y_i|\phi,z_i=0,z_{-i})\right]^{1-z_i} \\
&\propto \prod_{j=1}^{d}\left[\phi_j^{\frac{a_j}{2}-1}\exp\left\{-\frac{b_j}{2}\phi_j\right\}\right]\prod_{i=1}^{n}\left[(2\pi)^{-\frac{d}{2}}\left[\prod_{j=1}^{d}\phi_j\right]^{\frac{1}{2}}\exp\left\{-\frac{1}{2}\sum_{j=1}^{d}\phi_j(y_{ij}-\mu_j)^2\right\}\right]^{z_i} \\
&\times \prod_{i=1}^{n}\left[(2\pi)^{-\frac{d}{2}}\left[\prod_{j=1}^{d}\phi_j\right]^{\frac{1}{2}}\exp\left\{-\frac{1}{2}F_G^{-1}\left(\gamma^{g_i(n_1+1)^{-1}}\right)\right\}\right]^{1-z_i} \\
&\propto \prod_{j=1}^{d}\left[\phi_j^{\frac{a_j}{2}-1}\phi_j^{\frac{n_1}{2}}\phi_j^{\frac{n-n_1}{2}}\exp\left\{-\frac{b_j}{2}\phi_j\right\}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\mathbb{I}_{\{z_i=1\}}\phi_j(y_{ij}-\mu_j)^2\right\}\right] \\
&\propto \prod_{j=1}^{d}\left[\phi_j^{\frac{a_j+n}{2}-1}\exp\left\{-\frac{1}{2}\left[b_j+\sum_{i=1}^{n}\mathbb{I}_{\{z_i=1\}}(y_{ij}-\mu_j)^2\right]\phi_j\right\}\right],
\end{aligned} \tag{B.7}$$

and thus, identifying the distribution's kernel we have

$$\phi_j|\mu,\phi_\omega,z,y \overset{ind}{\sim} Gamma\left(\frac{\bar{a}_j}{2},\frac{\bar{b}_j}{2}\right), \tag{B.8}$$

where $\bar{a}_j = a_j + n$ and $\bar{b}_j = b_j + \sum_{i=1}^{n}\mathbb{I}_{\{z_i=1\}}(y_{ij}-\mu_j)^2$.

Now, taking the full conditional of ϕ_ω into consideration, notice that

$$\begin{aligned}
&\pi(\phi_\omega|\mu,\phi,z,y) \propto \pi(\mu,\phi,\phi_\omega,z|y) \\
&\propto \prod_{j=1}^{d}\left[\pi(\mu_j|\mu_{j-1},\phi_\omega)\right]\pi(\phi_\omega) \\
&\propto \prod_{j=1}^{d}\left[\phi_\omega^{\frac{1}{2}}\exp\left\{-\frac{\phi_\omega}{2}(\mu_j-\mu_{j-1})^2\right\}\right]\phi_\omega^{\frac{a_\omega}{2}-1}\exp\left\{-\frac{b_\omega}{2}\phi_\omega\right\} \\
&= \phi_\omega^{\frac{a_\omega}{2}-1}\phi_\omega^{\frac{d}{2}}\exp\left\{-\frac{b_\omega}{2}\phi_\omega\right\}\exp\left\{-\frac{\phi_\omega}{2}\sum_{j=1}^{d}(\mu_j-\mu_{j-1})^2\right\} \\
&= \phi_\omega^{\frac{a_\omega+d}{2}-1}\exp\left\{-\frac{1}{2}\left[b_\omega+\sum_{j=1}^{d}(\mu_j-\mu_{j-1})^2\right]\phi_\omega\right\}
\end{aligned} \tag{B.9}$$

and thus, identifying the distribution's kernel we have

$$\phi_\omega|\mu,\phi,z,y \sim Gamma\left(\frac{\bar{a}_\omega}{2},\frac{\bar{b}_\omega}{2}\right), \tag{B.10}$$

where $\bar{a}_\omega = a_\omega + d$ and $\bar{b}_\omega = b_\omega + \sum_{j=1}^{d}(\mu_j-\mu_{j-1})^2$.

At last, for the full conditional of z notice that

$$\begin{aligned}
&\pi(z_k|\mu,\phi,\phi_\omega,z_{-k},y) \propto \pi(\mu,\phi,\phi_\omega,z|y) \\
&\propto \prod_{i=1}^{n}\left[\pi(z_i)\right]\prod_{i=1}^{n}\left[\pi(y_i|\mu,\phi,z_i=1)\right]^{z_i}\prod_{i=1}^{n}\left[\pi(y_i|\phi,z_i=0,z_{-i})\right]^{1-z_i} \\
&= \prod_{i=1}^{n}\left[w_z^{z_i}(1-w_z)^{1-z_i}\right]\prod_{i=1}^{n}\left[(2\pi)^{-\frac{d}{2}}\left[\prod_{j=1}^{d}\phi_j\right]^{\frac{1}{2}}\exp\left\{-\frac{1}{2}\sum_{j=1}^{d}\phi_j(y_{ij}-\mu_j)^2\right\}\right]^{z_i} \\
&\times \prod_{i=1}^{n}\left[(2\pi)^{-\frac{d}{2}}\left[\prod_{j=1}^{d}\phi_j\right]^{\frac{1}{2}}\exp\left\{-\frac{1}{2}F_G^{-1}\left(\gamma^{g_i(n_1+1)^{-1}}\right)\right\}\right]^{1-z_i} \\
&\propto \prod_{i=1}^{n}\left[w_z^{z_i}(1-w_z)^{1-z_i}\right]\prod_{i=1}^{n}\left[\exp\left\{-\frac{1}{2}\sum_{j=1}^{d}\phi_j(y_{ij}-\mu_j)^2\right\}\right]^{z_i} \\
&\times \exp\left\{-\frac{n_0}{2}F_G^{-1}\left(\gamma^{g(n_1+1)^{-1}}\right)\right\} \\
&\propto \left[w_z^{z_k}(1-w_z)^{1-z_k}\right]\exp\left\{-\frac{z_k}{2}\sum_{j=1}^{d}\phi_j(y_{ij}-\mu_j)^2\right\} \\
&\times \exp\left\{-\frac{n_0^{-k}+1-z_k}{2}F_G^{-1}\left(\gamma^{g(n_1^{-k}+z_k+1)^{-1}}\right)\right\}
\end{aligned} \tag{B.11}$$

where $n_j^{-k} = \sum_{i \neq k} \mathbb{I}_{\{z_i = j\}}$ and $g = g_i$ (recall that all of the observations are identically distributed). Since we will use the Gumbel-max trick, see Huijben et al. (2022) for a description of the method, to sample from the resulting categorical distribution, we will also calculate the logarithm of the non-normalized probabilities for each indicator function. So, for $k \in \{1, \ldots, n\}$, we have

$$\ln\left[\pi(z_k = 1|\mu, \phi, \phi_\omega, z_{-k}, y)\right] = \ln(w_z) - \frac{1}{2}\sum_{j=1}^{d} \phi_j (y_{ij} - \mu_j)^2 + C \tag{B.12}$$

and

$$\begin{aligned} &\ln\left[\pi(z_k = 0|\mu, \phi, \phi_\omega, z_{-k}, y)\right] = \ln(1 - w_z) \\ &- \frac{1}{2}\left[(n_0^{-k} + 1) F_G^{-1}\left(\gamma^{g_i (n_1^{-k}+1)^{-1}}\right) - n_0^{-k} F_G^{-1}\left(\gamma^{g_i (n_1^{-k}+2)^{-1}}\right)\right] + C \\ &= \ln(1 - w_z) - \frac{1}{2} F_G^{-1}\left(\gamma^{(n_1^{-k}+1)^{-1}}\right) + C. \end{aligned} \tag{B.13}$$

It is worth pointing out that simplified expression for the equation B.13 is due to proposition A.4, as detailed in section A.2.

B.2 MULTIVARIATE NORMAL MIXTURE MODEL

Let us consider the filtering model with a multivariate normal mixture model as the main component. We can represent this model, including all of the chosen prior distributions for each parameter, in the hierarchical form given by

$$\begin{aligned} X_i|\mu_j, \Omega_j, y_i = j, z_i = 1 &\overset{ind}{\sim} Normal_d(\mu_j, \Omega_j^{-1}) \\ X_i|\mu, \Omega, z_{-i}, y_i = j, z_i = 0 &\overset{ind}{\sim} Uniforme(S_i) \\ \mu_j|\Omega_j &\overset{ind}{\sim} Normal_d(\theta_j, \lambda_j^{-1}\Omega_j^{-1}) \\ \Omega_j &\overset{ind}{\sim} Wishart(\nu_j, V_j) \\ y_i|w_y &\overset{ind}{\sim} Categorica(w_y) \\ w_y &\sim Dirichlet(\alpha) \\ z_i &\overset{ind}{\sim} Categorica(w_z) \\ \mu_L(S_i)^{-1} = (2\pi)^{-\frac{d}{2}} \left[\det(\Omega_j)\right]^{\frac{1}{2}} &\exp\left\{-\frac{1}{2} F_G^{-1}\left(\gamma^{g_i (n_{\cdot 1}+1)^{-1}}\right)\right\}, \end{aligned} \tag{B.14}$$

for $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, m\}$, where $n_{\cdot k} = \sum_{i=1}^{n} \mathbb{I}_{\{z_i = k\}}$, F_G represents the distribution function of a $Gamma\left(\frac{d}{2}, \frac{1}{2}\right)$, g_i is the correction function considering the assumed distribution for the i-th observation and μ_L is the Lebesgue measure. We also denote μ the collection of d-dimensional vectors $\mu_1, \ldots, \mu_m$, Ω the collection of $d \times d$ matrices $\Omega_1, \ldots, \Omega_m$, y the collection of class indicators $y_1, \ldots, y_n$ and z the collection of anomaly indicators $z_1, \ldots, z_n$. Here, the d-dimensional vectors $\theta_1, \ldots, \theta_m$, the positive scalars $\lambda_1, \ldots, \lambda_m$, the scalar $\nu_1, \ldots, \nu_m > d - 1$, the $d \times d$ matrices $V_1, \ldots, V_m$, the m-dimensional vector $\alpha = (\alpha_1, \ldots, \alpha_n)'$, the prior anomaly probability w_z and the scalar $\gamma \in (0, 1)$ are hyperparameters that require specification. Furthermore, we assume the dependence structure necessary to obtain the following prior factorization:

$$\begin{aligned}
\pi(\mu, \Omega, w_y, y, z) &\overset{ind}{=} \pi(\mu, \Omega)\pi(y, w_y)\pi(z) \\
&\overset{ind}{=} \prod_{j=1}^{n} \Big[\pi(\mu_j, \Omega_j)\Big] \pi(y|w_y)\pi(w_y)\pi(z) \\
&\overset{ind}{=} \prod_{j=1}^{m} \Big[\pi(\mu_j|\Omega_j)\pi(\Omega_j)\Big] \prod_{i=1}^{n} \Big[\pi(y_i|w_y)\pi(z_i)\Big] \ \pi(w_y).
\end{aligned} \tag{B.15}$$

Next, assuming an observed sample x from this model, that is a collection of d-dimensional vectors $x_1, \ldots, x_n$, and representing our parameters as $\Theta = (\mu, \Omega, w_y, y, z)$, we can use Bayes's Theorem to express the posterior as

$$\begin{aligned}
&\pi(\Theta|x) \overset{Bayes}{=} \frac{\pi(\Theta)\pi(x|\Theta)}{\int \pi(\Theta)\pi(x|\Theta)\ d\Theta} \propto \pi(\Theta)\pi(y|\Theta) \\
&= \pi(\mu, \Omega, w_y, y, z)\pi(x|\mu, \Omega, w_y, y, z) \\
&\overset{ind}{=} \prod_{i=1}^{n} \Big[\pi(x_i|\mu, \Omega, w_y, y_i, z)\pi(y_i|w_y)\pi(z_i)\Big] \prod_{j=1}^{m} \Big[\pi(\mu_j|\Omega_j)\pi(\Omega_j)\Big]\pi(w_y) \\
&\overset{ind}{=} \prod_{j=1}^{m} \prod_{k=0}^{1} \prod_{i=1}^{n} \Big[\pi(x_i|\mu, \Omega, w_y, y_i = j, z_i = k, z_{-i})\pi(y_i = j|w_y)\pi(z_i = k)\Big]^{\mathbb{I}_{\{y_i = j\}}\mathbb{I}_{\{z_i = k\}}} \\
&\times \prod_{j=1}^{m} \Big[\pi(\mu_j|\Omega_j)\pi(\Omega_j)\Big]\pi(w_y).
\end{aligned} \tag{B.16}$$

To start the process of identifying the full conditionals, we will first consider a blocking technique, briefly presented in section 2.2.4, to improve the mixing proprieties by sampling from the joint conditional $\pi(\mu, \Omega|y, z, w_y, x)$. This can be accom-

plished using the product rule decomposition

$$\pi(\mu, \Omega | w_y, y, z, x) = \pi(\mu | \Omega, w_y, y, z, x) \pi(\Omega | w_y, y, z, x), \tag{B.17}$$

i.e., by sequentially sampling

$$\Omega^* \sim \pi(\Omega | w_y, y, z, x) \tag{B.18}$$

and then

$$\mu^* \sim \pi(\mu | \Omega^*, w_y, y, z, x). \tag{B.19}$$

to obtain a sample (μ^*, Ω^*) from the desired distribution. So we begin obtaining the full conditional for μ by combining the quadratic form from the likelihood with the one from the prior, as shown below:

$$\begin{aligned}
&\sum_{i \in C_{j1}} (x_i - \mu_j)' \Omega_j (x_i - \mu_j) + \lambda_j (\mu_j - \theta_j)' \Omega_j (\mu_j - \theta_j) \\
&= \sum_{i \in C_{j1}} x_i' \Omega_j x_i - \sum_{i \in C_{j1}} x_i' \Omega_j \mu_j - \mu_j' \Omega_j \sum_{i \in C_{j1}} x_i + n_{j1} \mu_j' \Omega_j \mu_j \\
&+ \lambda_j \mu_j' \Omega_j \mu_j - \lambda_j \theta_j' \Omega_j \mu_j - \lambda_j \mu_j' \Omega_j \theta_j + \lambda_j \theta_j' \Omega_j \theta_j \\
&= (\lambda_j + n_{j1}) \mu_j' \Omega_j \mu_j - \mu_j' \Omega_j \left(\sum_{i \in C_{j1}} x_i + \lambda_j \theta_j \right) - \left(\sum_{i \in C_{j1}} x_i + \lambda_j \theta_j \right)' \Omega_j \mu_j \\
&+ \sum_{i \in C_{j1}} x_i' \Omega_j x_i + \lambda_j \theta_j' \Omega_j \theta_j \\
&= \left(\mu_j - \bar{\theta}_j \right)' \bar{\Omega}_j \left(\mu_j - \bar{\theta}_j \right) + \sum_{i \in C_{j1}} x_i' \Omega_j x_i + \lambda_j \theta_j' \Omega_j \theta_j - \bar{\theta}_j' \bar{\Omega}_j \bar{\theta}_j,
\end{aligned} \tag{B.20}$$

where $\bar{\Omega}_j = (\lambda_j + n_{j1}) \Omega_j$, $\bar{\theta}_j = \frac{1}{\lambda_j + n_{j1}} \left[\lambda_j \theta_j + \sum_{i \in C_{j1}} x_i \right]$, $C_{jk} = \{ i \in \{1, \ldots, n\} : y_i = j, z_i = k \}$ and $n_{jk} = \sum_{i=1}^{n} \mathbb{I}_{\{y_i = j\}} \mathbb{I}_{\{z_i = k\}}$. With the new quadratic form obtained,

notice that

$$\begin{aligned}
&\pi(\mu|\Omega, w_y, y, z, x) \\
&\propto \prod_{j=1}^{m} \prod_{i \in C_{j1}} \left[\pi(x_i|\mu_j, \Omega_j, y_i = j, z_i = 1)\right] \prod_{j=1}^{m} \left[\pi(\mu_j|\Omega_j)\right] \\
&\propto \prod_{j=1}^{m} \exp\left\{-\frac{1}{2} \sum_{i \in C_{j1}} (x_i - \mu_j)'\Omega_j(x_i - \mu_j)\right\} \prod_{j=1}^{m} \exp\left\{-\frac{1}{2}\lambda_j(\mu_j - \theta_j)'\Omega_j(\mu_j - \theta_j)\right\} \\
&= \prod_{j=1}^{m} \exp\left\{-\frac{1}{2}\left[\sum_{i \in C_{j1}} (x_i - \mu_j)'\Omega_j(x_i - \mu_j) + \lambda_j(\mu_j - \theta_j)'\Omega_j(\mu_j - \theta_j)\right]\right\} \\
&\propto \prod_{j=1}^{m} \exp\left\{-\frac{1}{2}\left(\mu_j - \bar{\theta}_j\right)' \bar{\Omega}_j \left(\mu_j - \bar{\theta}_j\right)\right\},
\end{aligned} \tag{B.21}$$

and thus, identifying the distribution's kernel we have

$$\mu_j|\Omega, w_y, y, z, x \overset{ind}{\sim} Normal_d\left(\bar{\theta}_j, \bar{\Omega}_j^{-1}\right). \tag{B.22}$$

As a side note, since the filtering component does not depend on location parameters, the full conditional of μ is the usual full conditional obtained for the multivariate normal mixture model, with the modification of only taking the observations from the main component in consideration, i.e., it considers all of the x_i's such that $z_i = 0$ as if they were not observed. Next, notice that

$$\begin{aligned}
&\pi(\Omega|w_y, y, z, x) = \int \pi(\Omega, \mu|w_y, y, z, x)\ d\mu \\
&= \int \prod_{j=1}^{m} \prod_{k=0}^{1} \prod_{i=1}^{n} \left[\pi(x_i|\mu, \Omega, w_y, y_i = j, z_i = k)\right]^{\mathbb{I}_{\{y_i=j\}}\mathbb{I}_{\{z_i=k\}}} \prod_{j=1}^{m} \left[\pi(\mu_j|\Omega_j)\pi(\Omega_j)\right]\ d\mu \\
&\propto \int \prod_{j=1}^{m} \prod_{i \in C_{j1}} \left[(2\pi)^{-\frac{d}{2}} \left[\det(\Omega_j)\right]^{\frac{1}{2}} \exp\left\{-\frac{1}{2}(x_i - \mu_j)'\Omega_j(x_i - \mu_j)\right\}\right] \\
&\times \prod_{j=1}^{m} \prod_{i \in C_{j0}} \left[(2\pi)^{-\frac{d}{2}} \left[\det(\Omega_j)\right]^{\frac{1}{2}} \exp\left\{-\frac{1}{2}F_G^{-1}\left(\gamma^{g_i(n_{\cdot 1}+1)^{-1}}\right)\right\}\right] \\
&\times \prod_{j=1}^{m} \left[(2\pi)^{-\frac{d}{2}} \left[\det(\Omega_j)\right]^{\frac{1}{2}} \exp\left\{-\frac{\lambda_j}{2}(\mu_j - \theta_j)'\Omega_j(\mu_i - \theta_j)\right\}\right] \\
&\times \prod_{j=1}^{m} \left[\left[\det(\Omega_j)\right]^{\frac{\nu_j - d - 1}{2}} \exp\left\{-\frac{1}{2}tr\left(\Omega_j V_j\right)\right\}\right]\ d\mu \\
&\propto \int \prod_{j=1}^{m} \prod_{i \in C_{j1}} \left[\exp\left\{-\frac{1}{2}(x_i - \mu_j)'\Omega_j(x_i - \mu_j)\right\}\right]
\end{aligned}$$

$$
\begin{aligned}
&\times \prod_{j=1}^{m} \left[[\det(\Omega_j)]^{\frac{1}{2}} \exp\left\{ -\frac{\lambda_j}{2}(\mu_j - \theta_j)'\Omega_j(\mu_i - \theta_j) \right\} \right] d\mu \\
&\times \prod_{j=1}^{m} \left[[\det(\Omega_j)]^{\frac{\nu_j + n_{j\cdot} - d - 1}{2}} \exp\left\{ -\frac{1}{2} tr\left(\Omega_j V_j\right) \right\} \right] \\
&\propto \prod_{j=1}^{m} \underbrace{\int \left[(2\pi)^{-\frac{d}{2}} \left[\det(\bar{\Omega}_j)\right]^{\frac{1}{2}} \exp\left\{ -\frac{1}{2}(\mu_j - \bar{\theta}_j)'\bar{\Omega}_j(\mu_i - \bar{\theta}_j) \right\} \right] d\mu_j}_{=1, \text{ density of a } Normal_d\left(\bar{\mu}_j, \bar{\Omega}_j^{-1}\right)} \\
&\times \prod_{j=1}^{m} \left[\exp\left\{ -\frac{1}{2} \left[\sum_{i \in C_{j1}} x_i'\Omega_j x_i + \lambda_j \theta_j'\Omega_j\theta_j - \bar{\theta}_j'\bar{\Omega}_j\bar{\theta}_j \right] \right\} \right] \\
&\times \prod_{j=1}^{m} \left[[\det(\Omega_j)]^{\frac{\nu_j + n_{j\cdot} - d - 1}{2}} \exp\left\{ -\frac{1}{2} tr\left(\Omega_j V_j\right) \right\} \right] \\
&= \prod_{j=1}^{m} \left[[\det(\Omega_j)]^{\frac{\nu_j + n_{j\cdot} - d - 1}{2}} \exp\left\{ -\frac{1}{2} tr\left(\Omega_j \left[V_j + \sum_{i \in C_{j1}} x_i x_i' + \lambda_j \theta_j \theta_j' - \bar{\lambda}_j \bar{\theta}_j \bar{\theta}_j' \right] \right) \right\} \right] \\
&= \prod_{j=1}^{m} \left[[\det(\Omega_j)]^{\frac{\bar{\nu}_j - d - 1}{2}} \exp\left\{ -\frac{1}{2} tr\left(\Omega_j \bar{V}_j\right) \right\} \right], \qquad \text{(B.23)}
\end{aligned}
$$

and thus, identifying the distribution's kernel we have

$$
\Omega_j | \mu, y, z, w_y, x \overset{ind}{\sim} Wishart\left(\bar{\nu}_j, \bar{V}_j\right), \qquad \text{(B.24)}
$$

where $\bar{\nu}_j = \nu_j + n_{j\cdot}$, $\bar{V}_j = V_j + \sum_{i \in C_{j1}} x_i x_i' + \lambda_j \theta_j \theta_j' - \bar{\lambda}_j \bar{\theta}_j \bar{\theta}_j'$, $\bar{\lambda}_j = \lambda_j + n_{j1}$, $\bar{\theta}_j = \frac{1}{\lambda_j + n_{j1}} \left[\lambda_j \theta_j + \sum_{i \in C_{j1}} x_i \right]$ and $n_{j\cdot} = \sum_{i=1}^{n} \mathbb{I}_{\{y_i = j\}}$.

Now, for the full conditional of w_y, notice that

$$
\begin{aligned}
&\pi(w_y | \mu, \Omega, y, z, x) \\
&\propto \prod_{j=1}^{m} \prod_{i=1}^{n} \left[\pi(y_i = j | w_y) \right]^{\mathbb{I}_{\{y_i = j\}}} \pi(w_y) \propto \prod_{j=1}^{m} w_{yj}^{n_{j\cdot}} w_{yj}^{\alpha_j - 1} = \prod_{j=1}^{m} w_{yj}^{\alpha_j + n_{j\cdot} - 1}, \qquad \text{(B.25)}
\end{aligned}
$$

and thus, identifying the distribution's kernel we have

$$
w_y | \mu, \Omega, y, z, x \overset{ind}{\sim} Dirichlet\left(\bar{\alpha}\right), \qquad \text{(B.26)}
$$

where $\bar{\alpha} = (\bar{\alpha}_1, \ldots, \bar{\alpha}_m)' = (\alpha_1 + n_{1\cdot}, \ldots, \alpha_m + n_{m\cdot})'$.

And at last, for the full conditional of y and z we will once more consider blocking to improve the chain mixing properties by sampling from the joint full

conditional $\pi(y_l, z_l|\mu, \Omega, w_y, y_{-l}, z_{-l}, x)$, where $y_{-l} = (y_1, \ldots, y_{l-1}, y_{l+1}, \ldots, y_n)'$ and $z_{-l} = (z_1, \ldots, z_{l-1}, z_{l+1}, \ldots, z_n)'$.

$$
\begin{aligned}
&\pi(y_l, z_l|\mu, \Omega, w_y, y_{-l}, z_{-l}, x) \\
&\propto \prod_{j=1}^{m}\prod_{k=0}^{1}\prod_{i=1}^{n}\left[\pi(x_i|\mu, \Omega, y_i = j, z_i = k)\pi(y_i = j|w_y)\pi(z_i = k)\right]^{\mathbb{I}_{\{y_i=j\}}\mathbb{I}_{\{z_i=k\}}} \\
&\propto \prod_{j=1}^{m}\prod_{i=1}^{n}\left[w_{yj}w_z\phi_d(x_i|\mu_j, \Omega_j^{-1})\right]^{\mathbb{I}_{\{y_i=j\}}\mathbb{I}_{\{z_i=1\}}}\left[w_{yj}(1-w_z)\mu_L(S_i)^{-1}\right]^{\mathbb{I}_{\{y_i=j\}}\mathbb{I}_{\{z_i=0\}}} \\
&= \prod_{j=1}^{m}\prod_{i=1}^{n}\left[w_{yj}w_z(2\pi)^{-\frac{d}{2}}\left[\det(\Omega_j)\right]^{\frac{1}{2}}\exp\left\{-\frac{1}{2}(x_i-\mu_j)'\Omega_j(x_i-\mu_j)\right\}\right]^{\mathbb{I}_{\{y_i=j\}}\mathbb{I}_{\{z_i=1\}}} \\
&\times \prod_{j=1}^{m}\prod_{i=1}^{n}\left[w_{yj}(1-w_z)(2\pi)^{-\frac{d}{2}}\left[\det(\Omega_j)\right]^{\frac{1}{2}}\exp\left\{-\frac{1}{2}F_G^{-1}\left(\gamma^{g_i(n_{\cdot 1}+1)^{-1}}\right)\right\}\right]^{\mathbb{I}_{\{y_i=j\}}\mathbb{I}_{\{z_i=0\}}} \\
&\propto \prod_{j=1}^{m}\left[w_{yj}w_z\left[\det(\Omega_j)\right]^{\frac{1}{2}}\exp\left\{-\frac{1}{2}(x_l-\mu_j)'\Omega_j(x_l-\mu_j)\right\}\right]^{\mathbb{I}_{\{y_l=j\}}\mathbb{I}_{\{z_l=1\}}} \\
&\times \prod_{j=1}^{m}\prod_{i=1}^{n}\left[w_{yj}(1-w_z)\left[\det(\Omega_j)\right]^{\frac{1}{2}}\right]^{\mathbb{I}_{\{y_i=j\}}\mathbb{I}_{\{z_i=0\}}} \\
&\times \left[\exp\left\{-\frac{1}{2}F_G^{-1}\left(\gamma^{g_i\left(n_{\cdot 1}^{-l}+\mathbb{I}_{\{z_l=1\}}+1\right)^{-1}}\right)\right\}\right]^{n_{\cdot 0}^{-l}+\mathbb{I}_{\{z_l=0\}}} \\
&\propto \prod_{j=1}^{m}\left[w_{yj}\left[\det(\Omega_j)\right]^{\frac{1}{2}}\right]^{\mathbb{I}_{\{y_l=j\}}} w_z^{\mathbb{I}_{\{z_l=1\}}}(1-w_z)^{\mathbb{I}_{\{z_l=0\}}} \\
&\times \exp\left\{-\frac{\mathbb{I}_{\{z_l=1\}}}{2}(x_l-\mu_{y_l})'\Omega_{y_l}(x_l-\mu_{y_l})\right\} \\
&\times \exp\left\{-\frac{n_{\cdot 0}^{-l}+\mathbb{I}_{\{z_l=0\}}}{2}F_G^{-1}\left(\gamma^{g\left(n_{\cdot 1}^{-l}+\mathbb{I}_{\{z_l=1\}}+1\right)^{-1}}\right)\right\},
\end{aligned}
\tag{B.27}
$$

where $n_{\cdot k}^{-l} = \sum_{i\neq l}\mathbb{I}_{\{z_i=k\}}$ and $g = g_i$ (recall that equality is attained because for the normal distribution the correction function does not depend on the location or scale parameters, so g is the same regardless of the class attribution y_i). Since we will use the Gumbel-max trick, see Huijben et al. (2022) for a description of the method, to sample from the resulting categorical distribution, we will also calculate the logarithm of the non-normalized probabilities for each possible category. So, for

$j \in \{1, \ldots, m\}$, we have

$$\begin{aligned} &\ln\left[\pi(y_l = j, z_l = 1|\mu, \Omega, w_y, y_{-l}, z_{-l}, x)\right] \\ &= \ln(w_{yj}) + \frac{1}{2}\ln\left[\det(\Omega_j)\right] + \ln(w_z) - \frac{1}{2}(x_l - \mu_j)'\Omega_j^{-1}(x_l - \mu_j) + C, \end{aligned} \tag{B.28}$$

and

$$\begin{aligned} &\ln\left[\pi(y_l = j, z_l = 0|\mu, \Omega, w_y, y_{-l}, z_{-l}, x)\right] \\ &= \ln(w_{yj}) + \frac{1}{2}\ln\left[\det(\Omega_j)\right] + \ln(1 - w_z) \\ &- \frac{1}{2}\left[(n_{\cdot 0}^{-l} + 1)F_G^{-1}\left(\gamma^{g_i\left(n_{\cdot 1}^{-l}+1\right)^{-1}}\right) - n_{\cdot 0}^{-l}F_G^{-1}\left(\gamma^{g_i\left(n_{\cdot 1}^{-l}+2\right)^{-1}}\right)\right] + C \\ &= \ln(w_{yj}) + \frac{1}{2}\ln\left[\det(\Omega_j)\right] + \ln(1 - w_z) - \frac{1}{2}F_G^{-1}\left(\gamma^{\left(n_{\cdot 1}^{-l}+1\right)^{-1}}\right) + C. \end{aligned} \tag{B.29}$$

It is worth pointing out that simplified expression for the equation B.29 is due to proposition A.4, as detailed in section A.2.

B.3 DYNAMIC IMPROVEMENT MODEL

Let us consider the filtering model with the dynamic improvement model of Sartório (2018) as the main component. We can represent this model, including all of the chosen prior distributions for each parameter, in the hierarchical form given

by

$$
\begin{aligned}
Y_{it}|\alpha, \beta_t^*, \kappa_t, \phi_i, z_{it} = 1 &\overset{ind}{\sim} Normal(\alpha_i + \beta_{it}\kappa_t, \phi_i^{-1}),\\
Y_{it}|\phi_i, z_{-(it)}, z_{it} = 0, &\overset{ind}{\sim} Uniform(S_{it}),\\
\kappa_t|\kappa_{t-1}, \delta_{t-1}, \phi_\kappa &\overset{ind}{\sim} Normal(\kappa_{t-1} + \delta_{t-1}, \phi_\kappa^{-1}),\\
\delta_t|\delta_{t-1}, \phi_\delta &\overset{ind}{\sim} Normal(\delta_{t-1}, \phi_\delta^{-1}),\\
\beta_t^*|\beta_{t-1}^*, \phi_{\beta^*}, &\overset{ind}{\sim} Normal(\beta_{t-1}^*, diag(\phi_{\beta^*})^{-1}),\\
\alpha_i &\overset{ind}{\sim} Normal(\mu_{\alpha_i}, \phi_{\alpha_i}^{-1}),\\
\begin{bmatrix}\kappa_0\\ \delta_0\end{bmatrix} &\sim Normal_2(m_0, C_0),\\
\beta_0^* &\sim Normal_n(m_0^*, C_0^*),\\
\phi &\overset{ind}{\sim} Gamma\left(\frac{a_i}{2}, \frac{b_i}{2}\right),\\
\phi_\kappa &\overset{ind}{\sim} Gamma\left(\frac{a_\kappa}{2}, \frac{b_\kappa}{2}\right),\\
\phi_\delta &\overset{ind}{\sim} Gamma\left(\frac{a_\delta}{2}, \frac{b_\delta}{2}\right),\\
\phi_{\beta_i^*} &\overset{ind}{\sim} Gamma\left(\frac{a_{\beta_i}}{2}, \frac{b_{\beta_i}}{2}\right),\\
z_{it}|z_{(i-1)t}, \rho &\sim Bernoulli\Big(z_{(i-1)t}\rho + (1 - z_{(i-1)t})(1 - \rho)\Big),\\
z_{(-1)t} &\overset{ind}{\sim} Bernoulli(\rho_0),\\
\rho &\sim Beta(a, b),\\
\mu_L(S_{it})^{-1} &= (2\pi)^{-\frac{1}{2}}\phi_i^{\frac{1}{2}} \exp\left\{-\frac{1}{2}F_G^{-1}\left(\gamma^{g_{it}(n_{\cdot\cdot 1}+1)^{-1}}\right)\right\}.
\end{aligned}
\tag{B.30}
$$

for $i \in \{0, \ldots, n\}$ representing the age and $t \in \{1, \ldots, T\}$ representing a time index, where $n_{\cdot\cdot k} = \sum_{i=0}^{n}\sum_{t=1}^{T}\mathbb{I}_{\{z_{it}=k\}}$, F_G represents the distribution function of a $Gamma\left(\frac{1}{2}, \frac{1}{2}\right)$, g_{it} is the correction function considering the assumed distribution for the mortality rate at age i and year given by index t and μ_L is the Lebesgue measure. We also denote α the $(n+1)$-dimensional vector $(\alpha_0, \ldots, \alpha_n)'$, β^* the collection of n-dimensional vectors $\beta_0^*, \ldots, \beta_T^*$, β the collection of $(n+1)$-dimensional vectors $\beta_0, \ldots, \beta_T$, κ the collection of real valued scalars $\kappa_0, \ldots, \kappa_T$, δ the collection of real valued scalars $\delta_0, \ldots, \delta_T$, ϕ the $(n+1)$-dimensional vector with positive entries $(\phi_0, \ldots, \phi_n)'$, ϕ_{β^*} the n-dimensional vector with positive entries $(\phi_{\beta_1^*}, \ldots, \phi_{\beta_n^*})'$, z the

$(n+2)\times T$ indicator matrix such that $[z]_{it} = z_{it}$ (here the rows' index starts at -1 and the columns' at 1) and $z_{-(it)}$ the collection of all entries of z with the exception of z_{it}. Here, the $(n+1)$-dimensional vector $\mu_\alpha = (\mu_{\alpha_0}, \ldots, \mu_{\alpha_n})'$, the $(n+1)$-vector with positive entries $\phi_\alpha = (\phi_{\alpha_0}, \ldots, \phi_{\alpha_i})'$, the positive scalars $a_0, \ldots, a_n$, $b_0, \ldots, b_n$, $a_{\beta_1}, \ldots, a_{\beta_n}$, $b_{\beta_1}, \ldots, b_{\beta_n}$, a_κ, b_κ, a_δ, b_δ, the 2-dimensional vector m_0, the 2×2 matrix C_0, the n-dimensional vector m_0^*, the $n\times n$ matrix C_0^*, the positive scalars a, b and the scalars $\rho_0, \gamma \in (0,1)$ are hyperparameters that require specification. It is also worth restating that

$$\beta_t = \begin{bmatrix} \beta_{0t} \\ \beta_{1t} \\ \beta_{2t} \\ \vdots \\ \beta_{(n-1)t} \\ \beta_{nt} \end{bmatrix} = \begin{bmatrix} n+1 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} -1 & -1 & \cdots & -1 & -1 \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix} \beta_t^*. \tag{B.31}$$

and that the function $diag : \mathbb{R}^n \to \mathbb{M}_{n\times n}(\mathbb{R})$, where $\mathbb{M}_{n\times n}(\mathbb{R})$ is the set of all real valued $n\times n$ matrices, is defined as

$$diag(x) = \begin{bmatrix} x_1 & 0 & \cdots & 0 & 0 \\ 0 & x_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & x_{n-1} & 0 \\ 0 & 0 & \cdots & 0 & x_n \end{bmatrix}, \tag{B.32}$$

for every vector $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$. Furthermore, we assume the dependence structure necessary to obtain the following prior factorization:

$$
\begin{aligned}
&\pi(\alpha, \beta^*, \kappa, \delta, \phi, \phi_{\beta^*}, \phi_\kappa, \phi_\delta, z, \rho) \\
&\overset{ind}{=} \pi(\alpha)\pi(\beta^*, \phi_{\beta^*})\pi(\kappa, \delta, \phi_\kappa, \phi_\delta)\pi(\phi)\phi(z, \rho) \\
&\overset{ind}{=} \prod_{i=0}^{n}\Big[\pi(\alpha_i)\Big]\pi(\beta^*|\phi_{\beta^*})\pi(\phi_{\beta^*})\pi(\kappa, \delta|\phi_\kappa, \phi_\delta)\pi(\phi_\kappa, \phi_\delta)\prod_{i=0}^{n}\Big[\pi(\phi_i)\Big]\phi(z|\rho)\pi(\rho) \\
&\overset{ind}{=} \prod_{i=0}^{n}\Big[\pi(\alpha_i)\Big]\prod_{t=1}^{T}\Big[\pi(\beta_t^*|\beta_{t-1}^*, \phi_{\beta^*})\Big]\pi(\beta_0^*)\prod_{i=1}^{n}\Big[\pi(\phi_{\beta_i^*})\Big] \\
&\times \prod_{t=1}^{T}\Big[\pi(\kappa_t, \delta_t|\kappa_{t-1}, \delta_{t-1}, \phi_\kappa, \phi_\delta)\Big]\pi(\kappa_0, \delta_0)\pi(\phi_\kappa, \phi_\delta)\prod_{i=0}^{n}\Big[\pi(\phi_i)\Big] \\
&\times \prod_{t=1}^{T}\Big[\prod_{i=0}^{n}\Big[\phi(z_{it}|z_{(i-1)t}, \rho)\Big]\pi(z_{(-1)t})\Big]\,\pi(\rho) \\
&\overset{ind}{=} \prod_{i=0}^{n}\Big[\pi(\alpha_i)\Big]\prod_{t=1}^{T}\Big[\pi(\beta_t^*|\beta_{t-1}^*, \phi_{\beta^*})\Big]\pi(\beta_0^*)\prod_{i=1}^{n}\Big[\pi(\phi_{\beta_i^*})\Big]\prod_{t=1}^{T}\Big[\pi(\kappa_t|\kappa_{t-1}, \delta_{t-1}, \phi_\kappa)\Big] \\
&\times \prod_{t=1}^{T}\Big[\pi(\delta_t|\delta_{t-1}, \phi_\delta)\Big]\pi(\kappa_0, \delta_0)\pi(\phi_\kappa)\pi(\phi_\delta)\prod_{i=0}^{n}\Big[\pi(\phi_i)\Big] \\
&\times \prod_{t=1}^{T}\Big[\prod_{i=0}^{n}\Big[\phi(z_{it}|z_{(i-1)t}, \rho)\Big]\pi(z_{(-1)t})\Big]\,\pi(\rho).
\end{aligned}
\tag{B.33}
$$

Next, assuming an observed $(n+1) \times T$ sample matrix y of mortality rates, and representing our parameters $\Theta = (\alpha, \beta^*, \kappa, \delta, \phi, \phi_{\beta^*}, \phi_\kappa, \phi_\delta, z, \rho)$, we can use Bayes'

Theorem to express the posterior as

$$\begin{aligned}
\pi(\Theta|y) &\overset{Bayes}{=} \frac{\pi(\Theta)\pi(y|\Theta)}{\int \pi(\Theta)\pi(y|\Theta)\ d\Theta} \propto \pi(\Theta)\pi(y|\Theta) \\
&= \pi(\alpha, \beta^*, \kappa, \delta, \phi, \phi_{\beta^*}, \phi_\kappa, \phi_\delta, z, \rho)\pi(y|\alpha, \beta^*, \kappa, \delta, \phi, \phi_{\beta^*}, \phi_\kappa, \phi_\delta, z, \rho) \\
&= \prod_{i=0}^{n}\left[\pi(\alpha_i)\right]\prod_{t=1}^{T}\left[\pi(\beta_t^*|\beta_{t-1}^*, \phi_{\beta^*})\right]\pi(\beta_0^*)\prod_{i=1}^{n}\left[\pi(\phi_{\beta_i^*})\right] \\
&\times \prod_{t=1}^{T}\left[\pi(\kappa_t|\kappa_{t-1}, \delta_{t-1}, \phi_\kappa)\right]\prod_{t=1}^{T}\left[\pi(\delta_t|\delta_{t-1}, \phi_\delta)\right]\pi(\kappa_0, \delta_0)\pi(\phi_\kappa)\pi(\phi_\delta) \\
&\times \prod_{i=0}^{n}\left[\pi(\phi_i)\right]\prod_{t=1}^{T}\left[\prod_{i=0}^{n}\left[\phi(z_{it}|z_{(i-1)t}, \rho)\right]\pi(z_{(-1)t})\right]\pi(\rho) \\
&\times \prod_{i=0}^{n}\prod_{t=1}^{T}\left[\pi(y_i|\alpha_i, \beta_t^*, \kappa_t, \phi_i, z_{it}=1)\right]^{z_{it}}\prod_{i=0}^{n}\prod_{t=1}^{T}\left[\pi(y_i|\phi_i, z_{-(it)}, z_{it}=0)\right]^{1-z_{it}}.
\end{aligned} \tag{B.34}$$

Then, with the expression for the posterior acquired, we begin obtaining the full conditional for α by combining the quadratic form from the likelihood with the one from the prior, as shown below:

$$\begin{aligned}
&\phi_{\alpha_i}(\alpha_i - \mu_{\alpha_i})^2 + \phi_i\sum_{t=1}^{T} z_{it}(y_{it} - \alpha_i - \beta_{it}\kappa_t)^2 \\
&= \phi_{\alpha_i}(\alpha_i^2 - 2\alpha\mu_{\alpha_i}\mu_{\alpha_i}^2) + \phi_i\sum_{t=1}^{T} z_{it}(\alpha_i - (y_{it} - \beta_{it}\kappa_t))^2 \\
&= \phi_{\alpha_i}\alpha_i^2 - 2\phi_{\alpha_i}\alpha_i\mu_{\alpha_i} + \phi_{\alpha_i}\mu_{\alpha_i}^2 + \phi_i\sum_{t=1}^{T} z_{it}(\alpha_i^2 - 2\alpha_i(y_{it} - \beta_{it}\kappa_t) + (y_{it} - \beta_{it}\kappa)^2) \\
&= \phi_{\alpha_i}\alpha_i^2 - 2\phi_{\alpha_i}\alpha_i\mu_{\alpha_i} + \phi_{\alpha_i}\mu_{\alpha_i}^2 + \phi_i n_{i\cdot 1}\alpha_i^2 - 2\phi_i\alpha_i\sum_{t=1}^{T} z_{it}y_{it} + 2\phi_i\alpha_i\underbrace{\sum_{t=1}^{T} z_{it}\beta_{it}\kappa_t}_{=0} \\
&+ \phi_i\sum_{t=1}^{T} z_{it}(y_{it} - \beta_{it}\kappa_t)^2 \\
&= (\phi_{\alpha_i} + \phi_i n_{i\cdot 1})\alpha_i^2 - 2\alpha_i(\phi_i n_{i\cdot 1}\bar{y}_{i\cdot 1} + \phi_{\alpha_i}\mu_{\alpha_i}) + \phi_i\sum_{t=1}^{T} z_{it}(y_{it} - \beta_{it}\kappa_t)^2 + \phi_{\alpha_i}\mu_{\alpha_i}^2 \\
&= (\phi_{\alpha_i} + \phi_i n_{i\cdot 1})\left(\alpha_i - \frac{\phi_{\alpha_i}\mu_{\alpha_i} + \phi_i n_{i\cdot 1}\bar{y}_{i\cdot 1}}{\phi_{\alpha_i} + \phi_i n_{i\cdot 1}}\right)^2 + \phi_i\sum_{t=1}^{T} z_{it}(y_{it} - \beta_{it}\kappa_t)^2 + \phi_{\alpha_i}\mu_{\alpha_i}^2 \\
&- (\phi_{\alpha_i} + \phi_i n_{i\cdot 1})\left(\frac{\phi_{\alpha_i}\mu_{\alpha_i} + \phi_i n_{i\cdot 1}\bar{y}_{i\cdot 1}}{\phi_{\alpha_i} + \phi_i n_{i\cdot 1}}\right)^2
\end{aligned}$$

$$= \bar{\phi}_{\alpha_i}(\alpha_i - \bar{\mu}_{\alpha_i})^2 + \phi_i \sum_{t=1}^{T} z_{it}(y_{it} - \beta_{it}\kappa_t)^2 + \phi_{\alpha_i}\mu_{\alpha_i}^2 - \bar{\phi}_{\alpha_i}\bar{\mu}_{\alpha_i}^2 \tag{B.35}$$

where $n_{i\cdot 1} = \sum_{t=1}^{T} z_{it}$, $\bar{y}_{i\cdot 1} = \frac{1}{n_{i\cdot 1}} \sum_{t=1}^{T} z_{it} y_{it}$, $\bar{\phi}_{\alpha_i} = \phi_{\alpha_i} + \phi_i n_{i\cdot 1}$ and $\bar{\mu}_{\alpha_i} = \frac{\phi_{\alpha_i}\mu_{\alpha_i} + \phi_i n_{i\cdot 1}\bar{y}_{i\cdot 1}}{\phi_{\alpha_i} + \phi_i n_{i\cdot 1}}$. With the new quadratic form obtained, notice that

$$\begin{aligned}
&\pi(\alpha|\beta^*, \kappa, \delta, \phi, \phi_{\beta^*}, \phi_\kappa, \phi_\delta, z, \rho, y) \propto \pi(\alpha, \beta^*, \kappa, \delta, \phi, \phi_{\beta^*}, \phi_\kappa, \phi_\delta, z, \rho|y) \\
&\propto \prod_{i=0}^{n} \left[\pi(\alpha_i)\right] \prod_{i=0}^{n}\prod_{t=1}^{T} \left[\pi(y_i|\alpha_i, \beta_t^*, \kappa_t, \phi_i, z_{it} = 1)\right]^{z_{it}} \\
&\propto \prod_{i=0}^{n} \left[(2\pi)^{-\frac{1}{2}} \phi_{\alpha_i}^{\frac{1}{2}} \exp\left\{-\frac{\phi_{\alpha_i}}{2}(\alpha_i - \mu_{\alpha_i})^2\right\}\right] \\
&\times \prod_{i=0}^{n}\prod_{t=1}^{T} \left[(2\pi)^{-\frac{1}{2}} \phi_i^{\frac{1}{2}} \exp\left\{-\frac{\phi_i}{2}(y_{it} - \alpha_i - \beta_{it}\kappa_t)^2\right\}\right]^{z_{it}} \\
&\propto \prod_{i=0}^{n} \left[\exp\left\{-\frac{1}{2}\left[\phi_{\alpha_i}(\alpha_i - \mu_{\alpha_i})^2 + \phi_i \sum_{t=1}^{T} z_{it}(y_{it} - \alpha_i - \beta_{it}\kappa_t)^2\right]\right\}\right] \\
&\propto \prod_{i=0}^{n} \left[\exp\left\{-\frac{\bar{\phi}_{\alpha_i}}{2}(\alpha_i - \bar{\mu}_{\alpha_i})^2\right\}\right],
\end{aligned} \tag{B.36}$$

and thus, identifying the distribution's kernel we have

$$\alpha_i|\beta^*, \kappa, \delta, \phi, \phi_{\beta^*}, \phi_\kappa, \phi_\delta, z, \rho, y \overset{ind}{\sim} Normal\left(\bar{\mu}_{\alpha_i}, \bar{\phi}_{\alpha_i}^{-1}\right). \tag{B.37}$$

Now, for the full conditional of β^* we can notice that, given α, κ, ϕ, ϕ_{β^*} and z, we can rewrite the model in the form of a dynamic linear model with known evolution matrices, covariance matrices and drift vectors as follows:

$$\begin{aligned}
Y_t^1 &= (F_t^*)^1\theta_t^* + (\nu_t^*)^1, \quad (\nu_t^*)^1 \overset{ind}{\sim} Normal_{n_{\cdot t1}}\left((v_t^*)^1, (V_t^*)^1\right), \\
\theta_t^* &= G_t^*\theta_{t-1}^* + \omega_t^*, \qquad\qquad \omega_t^* \overset{ind}{\sim} Normal_n(w_t^*, W_t^*), \\
&\qquad\qquad\qquad\qquad\quad \theta_0^* \sim Normal_n(m_0^*, C_0^*),
\end{aligned} \tag{B.38}$$

where $n_{\cdot t1} = \sum_{i=0}^{n} z_{it}$, $\theta_t^* = \beta_t^*$, the superscript ^{1} indicates that only the rows i such that $z_{it} = 1$ are considered from the corresponding column vectors/matrices and we

have

$$F_t^* = \kappa_t \begin{bmatrix} -1 & -1 & \cdots & -1 & -1 \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix}_{(n+1)\times n}, \quad G_t^* = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix}_{n\times n}, \tag{B.39}$$

$$v_t^* = \alpha + \kappa_t \begin{bmatrix} (n+1) \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{(n+1)\times 1}, \qquad w_t^* = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{n\times 1},$$

$V_t^* = diag(\phi)^{-1}$ and $W_t^* = diag(\phi_{\beta^*})^{-1}$. Then, we are able to apply the FFBS algorithm described in Chapter 15 of West & Harrison (1997) to directly sample from the full conditional of β^*. It is worth mentioning that in the case where $n_{\cdot t}^1 = 0$ for some t, we can still obtain a sample using the Kalman filter adapted for missing observations in the *forward filtering* part of the FFBS algorithm. For a precise description we direct the reader to Chapter 4 of West & Harrison (1997).

Next, for the joint full conditional of κ and δ we can similarly notice that, given α, β^*, ϕ, ϕ_κ, ϕ_δ and z, we can rewrite the model in the form of a dynamic linear model with known evolution matrices, covariance matrices and drift vectors as follows:

$$\begin{aligned} Y_t^1 &= F_t^1\theta_t + \nu_t^1, \quad & \nu_t^1 &\overset{ind}{\sim} Normal_{n_{\cdot t1}}\left(v_t^1, V_t^1\right), \\ \theta_t &= G_t\theta_{t-1} + \omega_t, \quad & \omega_t &\overset{ind}{\sim} Normal_n(w_t, W_t), \\ & & \theta_0 &\sim Normal_n(m_0, C_0), \end{aligned} \tag{B.40}$$

where $n_{\cdot t1} = \sum_{i=0}^n z_{it}$, $\theta_t = (\kappa_t, \delta_t)'$, the superscript ^{1} indicates that only the rows i such that $z_{it} = 1$ are considered from the corresponding column vectors/matrices and we have

$$F_t = \kappa_t \begin{bmatrix} \beta_{0t} & 0 \\ \vdots & \vdots \\ \beta_{nt} & 0 \end{bmatrix}_{(n+1)\times 2}, \quad G_t = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}_{2\times 2}, \tag{B.41}$$

$v_t = \alpha$, $w_t = (0,0)'$, $V_t = diag(\phi)^{-1}$ and $W_t = diag((\phi_\kappa, \phi_\delta)')^{-1}$. Then, we again are able to apply the FFBS algorithm described in Chapter 15 of West & Harrison (1997) to directly sample from the joint full conditional of κ and δ. It is worth mentioning that in the case where $n_{\cdot t1} = 0$ for some t, we can still obtain a sample using the Kalman filter adapted for missing observations in the *forward filtering* part of the FFBS algorithm. For a more precise description we direct the reader to Chapter 4 of West & Harrison (1997).

As a side note, since the filtering component does not depend on location parameters, the full conditionals of α, β^* and κ are the usual full conditionals obtained for the original dynamic improvement model, with the modification of only taking the observations from the main component in consideration, i.e., it considers all of the y_{it}'s such that $z_{it} = 0$ as if they were not observed.

Then, considering the full conditional of ϕ, notice that

$$
\begin{aligned}
&\pi(\phi|\alpha,\beta^*,\kappa,\delta,\phi_{\beta^*},\phi_\kappa,\phi_\delta,z,\rho,y) \propto \pi(\alpha,\beta^*,\kappa,\delta,\phi,\phi_\kappa,\phi_\delta,\phi_{\beta^*},z,\rho|y) \\
&\propto \prod_{i=0}^{n}\left[\pi(\phi_i)\right]\prod_{i=0}^{n}\prod_{t=1}^{T}\left[\pi(y_i|\alpha_i,\beta_t^*,\kappa_t,\phi_i,z_{it}=1)\right]^{z_{it}} \\
&\times \prod_{i=0}^{n}\prod_{t=1}^{T}\left[\pi(y_i|\phi_i,z_{-(it)},z_{it}=0)\right]^{1-z_{it}} \\
&\propto \prod_{i=0}^{n}\left[\phi_i^{\frac{a_i}{2}-1}\exp\left\{-\frac{b_i}{2}\phi_i\right\}\right]\prod_{i=0}^{n}\prod_{t=1}^{T}\left[(2\pi)^{-\frac{1}{2}}\phi_i^{\frac{1}{2}}\exp\left\{-\frac{\phi_i}{2}(y_{it}-\alpha_i-\beta_{it}\kappa_t)^2\right\}\right]^{z_{it}} \\
&\times \prod_{i=0}^{n}\prod_{t=1}^{T}\left[(2\pi)^{-\frac{1}{2}}\phi_i^{\frac{1}{2}}\exp\left\{-\frac{1}{2}F_G^{-1}\left(\gamma^{g_{it}(n_{\cdot\cdot 1}+1)^{-1}}\right)\right\}\right]^{1-z_{it}} \\
&\propto \prod_{i=0}^{n}\left[\phi_i^{\frac{a_i}{2}-1}\exp\left\{-\frac{b_i}{2}\phi_i\right\}\phi_i^{\frac{n_{i\cdot 1}}{2}}\phi_i^{\frac{T-n_{i\cdot 1}}{2}}\exp\left\{-\frac{\phi_i}{2}\sum_{t=1}^{T}z_{it}(y_{it}-\alpha_i-\beta_{it}\kappa_t)^2\right\}\right] \\
&\propto \prod_{i=0}^{n}\left[\phi_i^{\frac{a_i+T}{2}-1}\exp\left\{-\frac{1}{2}\left[b_i+\sum_{t=1}^{T}z_{it}(y_{it}-\alpha_i-\beta_{it}\kappa_t)^2\right]\phi_i\right\}\right],
\end{aligned}
\tag{B.42}
$$

and thus, identifying the distribution's kernel we have

$$
\phi_i|\alpha,\beta^*,\kappa,\delta,\phi_{\beta^*},\phi_\kappa,\phi_\delta,z,\rho,y \overset{ind}{\sim} Gamma\left(\frac{\bar{a}_i}{2},\frac{\bar{b}_i}{2}\right), \tag{B.43}
$$

where $\bar{a}_i = a_i + T$ and $\bar{b}_i = b_i + \sum_{t=1}^{T} z_{it}(y_{it} - \alpha_i - \beta_{it}\kappa_t)^2$.

Now, taking the full conditional of ϕ_{β^*} into consideration, notice that

$$\begin{aligned}
&\pi(\phi_{\beta^*}|\alpha,\beta^*,\kappa,\delta,\phi\phi_\kappa,\phi_\delta,z,\rho,y) \propto \pi(\alpha,\beta^*,\kappa,\delta,\phi,\phi_{\beta^*},\phi_\kappa,\phi_\delta,z,\rho|y)\\
&\propto \prod_{t=1}^{T}\left[\pi(\beta_t^*|\beta_{t-1}^*,\phi_{\beta^*})\right]\prod_{i=1}^{n}\left[\pi(\phi_{\beta_i^*})\right]\\
&\propto \prod_{i=1}^{n}\prod_{t=1}^{T}\left[(2\pi)^{-\frac{1}{2}}\phi_{\beta_i^*}^{\frac{1}{2}}\exp\left\{-\frac{\phi_{\beta_i^*}}{2}(\beta_{it}^*-\beta_{i(t-1)}^*)^2\right\}\right]\prod_{i=1}^{n}\left[\phi_{\beta_i^*}^{\frac{a_{\beta_i}}{2}-1}\exp\left\{-\frac{b_{\beta_i}}{2}\phi_{\beta_i^*}\right\}\right]\\
&\propto \prod_{i=1}^{n}\left[\phi_{\beta_i^*}^{\frac{a_{\beta_i}}{2}-1}\phi_{\beta_i^*}^{\frac{T}{2}}\exp\left\{-\frac{b_{\beta_i}}{2}\phi_{\beta_i^*}\right\}\exp\left\{-\frac{\phi_{\beta_i^*}}{2}\sum_{t=1}^{T}(\beta_{it}^*-\beta_{i(t-1)}^*)^2\right\}\right]\\
&\propto \prod_{i=1}^{n}\left[\phi_{\beta_i^*}^{\frac{a_{\beta_i}+T}{2}-1}\exp\left\{-\frac{1}{2}\left[b_{\beta_i}+\sum_{t=1}^{T}(\beta_{it}^*-\beta_{i(t-1)}^*)^2\right]\phi_{\beta_i^*}\right\}\right]
\end{aligned} \tag{B.44}$$

and thus, identifying the distribution's kernel we have

$$\phi_{\beta_i^*}|\alpha,\beta^*,\kappa,\delta,\phi,\phi_\kappa,\phi_\delta,z,\rho,y \overset{ind}{\sim} Gamma\left(\frac{\bar{a}_{\beta_i}}{2},\frac{\bar{b}_{\beta_i}}{2}\right), \tag{B.45}$$

where $\bar{a}_{\beta_i} = a_{\beta_i} + T$ and $\bar{b}_{\beta_i} = b_{\beta_i} + \sum_{t=1}^{T}(\beta_{it}^* - \beta_{i(t-1)}^*)^2$. Alternatively, we can use a discount factors to obtain estimates of the precision of β^* instead of directly estimating the ϕ_{β^*}, as presented in Chapters 2 and 6 of West & Harrison (1997).

Next, for the full conditional of ϕ_κ notice that

$$\begin{aligned}
&\pi(\phi_\kappa|\alpha,\beta^*,\kappa,\delta,\phi,\phi_{\beta^*},\phi_\delta,z,\rho,y) \propto \pi(\alpha,\beta^*,\kappa,\delta,\phi,\phi_{\beta^*},\phi_\kappa,\phi_\delta,z,\rho|y)\\
&\propto \prod_{t=1}^{T}\left[\pi(\kappa_t|\kappa_{t-1},\delta_{t-1},\phi_\kappa)\right]\pi(\phi_\kappa,\phi_\delta)\\
&\propto \prod_{t=1}^{T}\left[(2\pi)^{-\frac{1}{2}}\phi_\kappa^{\frac{1}{2}}\exp\left\{-\frac{\phi_\kappa}{2}(\kappa_t-\kappa_{t-1}-\delta_{t-1})^2\right\}\right]\phi_\kappa^{\frac{a_\kappa}{2}-1}\exp\left\{-\frac{b_\kappa}{2}\phi_\kappa\right\}\\
&\propto \phi_\kappa^{\frac{a_\kappa}{2}-1}\phi_\kappa^{\frac{T}{2}}\exp\left\{-\frac{b_\kappa}{2}\phi_\kappa\right\}\exp\left\{-\frac{\phi_\kappa}{2}\sum_{t=1}^{T}(\kappa_t-\kappa_{t-1}-\delta_{t-1})^2\right\}\\
&\propto \phi_\kappa^{\frac{a_\kappa+T}{2}-1}\exp\left\{-\frac{1}{2}\left[b_\kappa+\sum_{t=1}^{T}(\kappa_t-\kappa_{t-1}-\delta_{t-1})^2\right]\phi_\kappa\right\}
\end{aligned} \tag{B.46}$$

and thus, identifying the distribution's kernel we have

$$\phi_\kappa|\alpha,\beta^*,\kappa,\delta,\phi,\phi_{\beta^*},\phi_\delta,z,\rho,y \overset{ind}{\sim} Gamma\left(\frac{\bar{a}_\kappa}{2},\frac{\bar{b}_\kappa}{2}\right), \tag{B.47}$$

where $\bar{a}_\kappa = a_\kappa + T$ and $\bar{b}_\kappa = b_\kappa + \sum_{t=1}^{T}(\kappa_t - \kappa_{t-1} - \delta_{t-1})^2$. Alternatively, we can use a discount factors to obtain estimates of the precision of κ instead of directly estimating the ϕ_κ, as presented in Chapters 2 and 6 of West & Harrison (1997).

Now, for the full conditional of ϕ_δ notice that

$$\begin{aligned}
&\pi(\phi_\delta|\alpha,\beta^*,\kappa,\delta,\phi,\phi_{\beta^*},\phi_\kappa,z,\rho,y) \propto \pi(\alpha,\beta^*,\kappa,\delta,\phi,\phi_{\beta^*},\phi_\kappa,\phi_\delta,z,\rho|y)\\
&\propto \prod_{t=1}^{T}\left[\pi(\delta_t|\delta_{t-1},\phi_\delta)\right]\pi(\phi_\delta)\\
&\propto \prod_{t=1}^{T}\left[(2\pi)^{-\frac{1}{2}}\phi_\delta^{\frac{1}{2}}\exp\left\{-\frac{\phi_\delta}{2}(\delta_t-\delta_{t-1})^2\right\}\right]\phi_\delta^{\frac{a_\delta}{2}-1}\exp\left\{-\frac{b_\delta}{2}\phi_\delta\right\}\\
&\propto \phi_\delta^{\frac{a_\delta}{2}-1}\phi_\delta^{\frac{T}{2}}\exp\left\{-\frac{b_\delta}{2}\phi_\delta\right\}\exp\left\{-\frac{\phi_\delta}{2}\sum_{t=1}^{T}(\delta_t-\delta_{t-1})^2\right\}\\
&\propto \phi_\delta^{\frac{a_\delta+T}{2}-1}\exp\left\{-\frac{1}{2}\left[b_\delta+\sum_{t=1}^{T}(\delta_t-\delta_{t-1})^2\right]\phi_\delta\right\}
\end{aligned} \tag{B.48}$$

and thus, identifying the distribution's kernel we have

$$\phi_\delta|\alpha,\beta^*,\kappa,\delta,\phi,\phi_{\beta^*},\phi_\kappa,z,\rho,y \overset{ind}{\sim} Gamma\left(\frac{\bar{a}_\delta}{2},\frac{\bar{b}_\delta}{2}\right), \tag{B.49}$$

where $\bar{a}_\delta = a_\delta + T$ and $\bar{b}_\delta = b_\delta + \sum_{t=1}^{T}(\delta_t-\delta_{t-1})^2$. Alternatively, we can use a discount factors to obtain estimates of the precision of δ instead of directly estimating the ϕ_δ, as presented in Chapters 2 and 6 of West & Harrison (1997).

For the full conditional of ρ, notice that

$$\begin{aligned}
&\pi(\rho|\alpha,\beta^*,\kappa,\delta,\phi,\phi_{\beta^*},\phi_\kappa,\phi_\delta,z,y) \propto \pi(\alpha,\beta^*,\kappa,\delta,\phi,\phi_{\beta^*},\phi_\kappa,\phi_\delta,z,\rho|y)\\
&\propto \prod_{t=1}^{T}\left[\prod_{i=0}^{n}\phi(z_{it}|z_{(i-1)t},\rho)\right]\pi(\rho)\\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\rho^{a-1}(1-\rho)^{b-1}\prod_{t=1}^{T}\prod_{i=0}^{n}\left[z_{(i-1)t}\rho+(1-z_{(i-1)t})(1-\rho)\right]^{z_{it}}\\
&\times\prod_{t=1}^{T}\prod_{i=0}^{n}\left[1-z_{(i-1)t}\rho-(1-z_{(i-1)t})(1-\rho)\right]^{1-z_{it}}\\
&\propto \rho^{a-1}(1-\rho)^{b-1}\prod_{t=1}^{T}\prod_{i=0}^{n}\left[\rho^{\mathbb{I}_{\{z_{it}=z_{(i-1)t}\}}}(1-\rho)^{\mathbb{I}_{\{z_{it}\neq z_{(i-1)t}\}}}\right]\\
&\propto \rho^{\bar{a}-1}(1-\rho)^{\bar{b}-1}
\end{aligned} \tag{B.50}$$

and thus, identifying the distribution's kernel we have

$$\rho|\alpha,\beta^*,\kappa,\delta,\phi,\phi_{\beta^*},\phi_\kappa,\phi_\delta,z,y \sim Beta\left(\bar{a},\bar{b}\right), \tag{B.51}$$

where

$$\bar{a} = a + \sum_{t=1}^{T}\sum_{i=0}^{n}\mathbb{I}_{\{z_{it}=z_{(i-1)t}\}}, \quad \text{and} \quad \bar{b} = b + \sum_{t=1}^{T}\sum_{i=0}^{n}\mathbb{I}_{\{z_{it}\neq z_{(i-1)t}\}}. \tag{B.52}$$

At last, for the full conditional of z notice that

$$
\begin{aligned}
&\pi(z|\alpha,\beta^*,\kappa,\delta,\phi,\phi_{\beta^*},\phi_\kappa,\phi_\delta,\rho,y) \propto \pi(\alpha,\beta^*,\kappa,\delta,\phi,\phi_{\beta^*},\phi_\kappa,\phi_\delta,z,\rho|y)\\
&\propto \prod_{t=1}^{T}\left[\prod_{i=0}^{n}\left[\phi(z_{it}|z_{(i-1)t},\rho)\right]\pi(z_{(-1)t})\right]\prod_{i=0}^{n}\prod_{t=1}^{T}\left[\pi(y_i|\alpha_i,\beta_t^*,\kappa_t,\phi_i,z_{it}=1)\right]^{z_{it}}\\
&\times\prod_{i=0}^{n}\prod_{t=1}^{T}\left[\pi(y_i|\phi_i,z_{-(it)},z_{it}=0)\right]^{1-z_{it}}\\
&\propto \prod_{t=1}^{T}\prod_{i=0}^{n}\left[z_{(i-1)t}\rho+(1-z_{(i-1)t})(1-\rho)\right]^{z_{it}}\\
&\times\prod_{t=1}^{T}\prod_{i=0}^{n}\left[1-z_{(i-1)t}\rho-(1-z_{(i-1)t})(1-\rho)\right]^{1-z_{it}}\prod_{t=1}^{T}\left[\rho_0^{z_{(-1)t}}(1-\rho_0)^{1-z_{(-1)t}}\right]\\
&\times\prod_{i=0}^{n}\prod_{t=1}^{T}\left[(2\pi)^{-\frac{1}{2}}\phi_i^{\frac{1}{2}}\exp\left\{-\frac{\phi_i}{2}(y_{it}-\alpha_i-\beta_{it}\kappa_t)^2\right\}\right]^{z_{it}}\\
&\times\prod_{i=0}^{n}\prod_{t=1}^{T}\left[(2\pi)^{-\frac{1}{2}}\phi_i^{\frac{1}{2}}\exp\left\{-\frac{1}{2}F_G^{-1}\left(\gamma^{g_{it}(n_{\cdot\cdot 1}+1)^{-1}}\right)\right\}\right]^{1-z_{it}}.\\
&\propto \prod_{t=1}^{T}\prod_{i=0}^{n}\left[\rho^{\mathbb{I}_{\left\{z_{it}=z_{(i-1)t}\right\}}}(1-\rho)^{\mathbb{I}_{\left\{z_{it}\neq z_{(i-1)t}\right\}}}\right]\prod_{t=1}^{T}\left[\rho_0^{z_{(-1)t}}(1-\rho_0)^{1-z_{(-1)t}}\right]\\
&\times\prod_{i=0}^{n}\prod_{t=1}^{T}\exp\left\{-z_{it}\frac{\phi_i}{2}(y_{it}-\alpha_i-\beta_{it}\kappa_t)^2\right\}\exp\left\{-\frac{n_{\cdot\cdot 0}}{2}F_G^{-1}\left(\gamma^{g(n_{\cdot\cdot 1}+1)^{-1}}\right)\right\},
\end{aligned}
\tag{B.53}
$$

where $n_{\cdot\cdot k}=\sum_{i=0}^{n}\sum_{t=1}^{T}\mathbb{I}_{\{z_{it}=k\}}$ and $g=g_{it}$ (recall that equality is attained because for the normal distribution the correction function does not depend on the location or scale parameters, so g is the same regardless of the indexes i and t). We will now split the calculations in three distinct cases. Let us first consider $j\in\{0,\ldots,n-1\}$

and $k \in \{1, \ldots, T\}$, then

$$\begin{aligned}
&\pi(z_{jk}|\alpha, \beta^*, \kappa, \delta, \phi, \phi_{\beta^*}, \phi_\kappa, \phi_\delta, z_{-(jk)}, \rho, y) \propto \pi(\alpha, \beta^*, \kappa, \delta, \phi, \phi_{\beta^*}, \phi_\kappa, \phi_\delta, z, \rho|y) \\
&\propto \prod_{t=1}^{T}\prod_{i=0}^{n}\left[\rho^{\mathbb{I}_{\left\{z_{it}=z_{(i-1)t}\right\}}}(1-\rho)^{\mathbb{I}_{\left\{z_{it}\neq z_{(i-1)t}\right\}}}\right]\prod_{t=1}^{T}\left[\rho_0^{z_{(-1)t}}(1-\rho_0)^{1-z_{(-1)t}}\right] \\
&\times \prod_{i=0}^{n}\prod_{t=1}^{T}\exp\left\{-z_{it}\frac{\phi_i}{2}(y_{it}-\alpha_i-\beta_{it}\kappa_t)^2\right\}\exp\left\{-\frac{n_{..0}}{2}F_G^{-1}\left(\gamma^{g_{it}(n_{..1}+1)^{-1}}\right)\right\} \\
&\propto \rho^{\mathbb{I}_{\left\{z_{(j+1)k}=z_{jk}\right\}}+\mathbb{I}_{\left\{z_{jk}=z_{(j-1)k}\right\}}}(1-\rho)^{\mathbb{I}_{\left\{z_{(j+1)k}\neq z_{jk}\right\}}+\mathbb{I}_{\left\{z_{jk}\neq z_{(j-1)k}\right\}}} \\
&\times \exp\left\{-z_{jk}\frac{\phi_j}{2}(y_{jk}-\alpha_j-\beta_{jk}\kappa_k)^2\right\} \\
&\times \exp\left\{-\frac{n_{..0}^{-(jk)}+(1-z_{jk})}{2}F_G^{-1}\left(\gamma^{g(n_{..1}^{-(jk)}+z_{jk}+1)^{-1}}\right)\right\}.
\end{aligned} \tag{B.54}$$

where $n_{..l}^{-(jk)} = \sum_{i=0}^{n}\sum_{t=1}^{T}\mathbb{I}_{\{z_{it}=l\}} - \mathbb{I}_{\{z_{jk}=l\}}$. Next, considering $j = n$ and $k \in \{1, \ldots, T\}$, we have

$$\begin{aligned}
&\pi(z_{nk}|\alpha, \beta^*, \kappa, \delta, \phi, \phi_{\beta^*}, \phi_\kappa, \phi_\delta, z_{-(nk)}, \rho, y) \propto \pi(\alpha, \beta^*, \kappa, \delta, \phi, \phi_{\beta^*}, \phi_\kappa, \phi_\delta, z, \rho|y) \\
&\propto \prod_{t=1}^{T}\prod_{i=0}^{n}\left[\rho^{\mathbb{I}_{\left\{z_{it}=z_{(i-1)t}\right\}}}(1-\rho)^{\mathbb{I}_{\left\{z_{it}\neq z_{(i-1)t}\right\}}}\right]\prod_{t=1}^{T}\left[\rho_0^{z_{(-1)t}}(1-\rho_0)^{1-z_{(-1)t}}\right] \\
&\times \prod_{i=0}^{n}\prod_{t=1}^{T}\exp\left\{-z_{it}\frac{\phi_i}{2}(y_{it}-\alpha_i-\beta_{it}\kappa_t)^2\right\}\exp\left\{-\frac{n_{..0}}{2}F_G^{-1}\left(\gamma^{g_{it}(n_{..1}+1)^{-1}}\right)\right\} \\
&\propto \rho^{\mathbb{I}_{\left\{z_{nk}=z_{(n-1)k}\right\}}}(1-\rho)^{\mathbb{I}_{\left\{z_{nk}\neq z_{(n-1)k}\right\}}} \\
&\times \exp\left\{-z_{nk}\frac{\phi_n}{2}(y_{nk}-\alpha_n-\beta_{nk}\kappa_k)^2\right\} \\
&\times \exp\left\{-\frac{n_{..0}^{-(nk)}+(1-z_{nk})}{2}F_G^{-1}\left(\gamma^{g(n_{..1}^{-(nk)}+z_{nk}+1)^{-1}}\right)\right\}.
\end{aligned} \tag{B.55}$$

And now, considering $j = -1$ and $k \in \{1, \ldots, T\}$ we have

$$\begin{aligned}
&\pi(z_{(-1)k}|\alpha, \beta^*, \kappa, \delta, \phi, \phi_{\beta^*}, \phi_\kappa, \phi_\delta, z_{-((-1)k)}, \rho, y) \propto \pi(\alpha, \beta^*, \kappa, \delta, \phi, \phi_{\beta^*}, \phi_\kappa, \phi_\delta, z, \rho|y) \\
&\propto \prod_{t=1}^{T}\prod_{i=0}^{n}\left[\rho^{\mathbb{I}_{\left\{z_{it}=z_{(i-1)t}\right\}}}(1-\rho)^{\mathbb{I}_{\left\{z_{it}\neq z_{(i-1)t}\right\}}}\right]\prod_{t=1}^{T}\left[\rho_0^{z_{(-1)t}}(1-\rho_0)^{1-z_{(-1)t}}\right] \\
&\times \prod_{i=0}^{n}\prod_{t=1}^{T}\exp\left\{-z_{it}\frac{\phi_i}{2}(y_{it}-\alpha_i-\beta_{it}\kappa_t)^2\right\}\exp\left\{-\frac{n_{\cdot\cdot 0}}{2}F_G^{-1}\left(\gamma^{g_{it}(n_{\cdot\cdot 1}+1)^{-1}}\right)\right\} \\
&\propto \rho^{\mathbb{I}_{\left\{z_{0k}=z_{(-1)k}\right\}}}\rho_0^{\mathbb{I}_{\left\{z_{(-1)k}=1\right\}}}(1-\rho)^{\mathbb{I}_{\left\{z_{0k}\neq z_{(-1)k}\right\}}}(1-\rho_0)^{\mathbb{I}_{\left\{z_{(-1)k}=0\right\}}} \\
&\times \exp\left\{-z_{jk}\frac{\phi_j}{2}(y_{jk}-\alpha_j-\beta_{jk}\kappa_k)^2\right\}.
\end{aligned} \tag{B.56}$$

Since we will use the Gumbel-max trick, see Huijben et al. (2022) for a description of the method, to sample from the resulting categorical distribution, we will also calculate the logarithm of the non-normalized probabilities for each possible category. So, for $j \in \{0, \ldots, n-1\}$ and $k \in \{1, \ldots, T\}$, we have

$$\begin{aligned}
&\ln\left[\pi(z_{jk}=1|\alpha, \beta^*, \kappa, \delta, \phi, \phi_{\beta^*}, \phi_\kappa, \phi_\delta, z_{-(jk)}, \rho, y)\right] \\
&= \left(\mathbb{I}_{\left\{z_{(j+1)k}=1\right\}}+\mathbb{I}_{\left\{z_{(j-1)k}=1\right\}}\right)\ln(\rho)+\left(\mathbb{I}_{\left\{z_{(j+1)k}=0\right\}}+\mathbb{I}_{\left\{z_{(j-1)k}=0\right\}}\right)\ln(1-\rho) \\
&- \frac{\phi_j}{2}(y_{jk}-\alpha_j-\beta_{jk}\kappa_k)^2+C
\end{aligned} \tag{B.57}$$

and

$$\begin{aligned}
&\ln\left[\pi(z_{jk}=0|\alpha, \beta^*, \kappa, \delta, \phi, \phi_{\beta^*}, \phi_\kappa, \phi_\delta, z_{-(jk)}, \rho, y)\right] \\
&= \left(\mathbb{I}_{\left\{z_{(j+1)k}=0\right\}}+\mathbb{I}_{\left\{z_{(j-1)k}=0\right\}}\right)\ln(\rho)+\left(\mathbb{I}_{\left\{z_{(j+1)k}=1\right\}}+\mathbb{I}_{\left\{z_{(j-1)k}=1\right\}}\right)\ln(1-\rho) \\
&- \frac{1}{2}\left[\left(n_{\cdot\cdot 0}^{-(jk)}+(1-z_{jk})\right)F_G^{-1}\left(\gamma^{g(n_{\cdot\cdot 1}^{-(jk)}+1)^{-1}}\right)-n_{\cdot\cdot 0}^{-(jk)}F_G^{-1}\left(\gamma^{g(n_{\cdot\cdot 1}^{-(jk)}+2)^{-1}}\right)\right]+C \\
&= \left(\mathbb{I}_{\left\{z_{(j+1)k}=0\right\}}+\mathbb{I}_{\left\{z_{(j-1)k}=0\right\}}\right)\ln(\rho)+\left(\mathbb{I}_{\left\{z_{(j+1)k}=1\right\}}+\mathbb{I}_{\left\{z_{(j-1)k}=1\right\}}\right)\ln(1-\rho) \\
&- \frac{1}{2}F_G^{-1}\left(\gamma^{g(n_{\cdot\cdot 1}^{-(jk)}+1)^{-1}}\right)+C.
\end{aligned} \tag{B.58}$$

Next, for $j = n$ and $k \in \{1, \ldots, T\}$ we have

$$\begin{aligned}
&\ln\left[\pi(z_{nk}=1|\alpha, \beta^*, \kappa, \delta, \phi, \phi_{\beta^*}, \phi_\kappa, \phi_\delta, z_{-(nk)}, \rho, y)\right] \\
&= \mathbb{I}_{\left\{z_{(n-1)k}=1\right\}}\ln(\rho)+\mathbb{I}_{\left\{z_{(n-1)k}=0\right\}}\ln(1-\rho)-\frac{\phi_j}{2}(y_{nk}-\alpha_n-\beta_{nk}\kappa_k)^2+C
\end{aligned} \tag{B.59}$$

and

$$\begin{aligned}
&\ln\left[\pi(z_{nk}=0|\alpha,\beta^*,\kappa,\delta,\phi,\phi_{\beta^*},\phi_\kappa,\phi_\delta,z_{-(nk)},\rho,y)\right]\\
&=\mathbb{I}_{\left\{z_{(n-1)k}=0\right\}}\ln(\rho)+\mathbb{I}_{\left\{z_{(n-1)k}=1\right\}}\ln(1-\rho)\\
&-\frac{1}{2}\left[\left(n_{\cdot\cdot 0}^{-(nk)}+(1-z_{nk})\right)F_G^{-1}\left(\gamma^{g(n_{\cdot\cdot 1}^{-(nk)}+1)^{-1}}\right)-n_{\cdot\cdot 0}^{-(nk)}F_G^{-1}\left(\gamma^{g(n_{\cdot\cdot 1}^{-(nk)}+2)^{-1}}\right)\right]+C\\
&=\mathbb{I}_{\left\{z_{(n-1)k}=0\right\}}\ln(\rho)+\mathbb{I}_{\left\{z_{(n-1)k}=1\right\}}\ln(1-\rho)-\frac{1}{2}F_G^{-1}\left(\gamma^{g(n_{\cdot\cdot 1}^{-(nk)}+1)^{-1}}\right)+C.
\end{aligned}\tag{B.60}$$

And finally, for $j=-1$ and $k\in\{1,\ldots,T\}$ we have

$$\begin{aligned}
&\ln\left[\pi(z_{(-1)k}=1|\alpha,\beta^*,\kappa,\delta,\phi,\phi_{\beta^*},\phi_\kappa,\phi_\delta,z_{-((-1)k)},\rho,y)\right]\\
&=\mathbb{I}_{\{z_{0k}=1\}}\ln(\rho)+\mathbb{I}_{\{z_{0k}=0\}}\ln(1-\rho)+\ln(\rho_0)+C
\end{aligned}\tag{B.61}$$

and

$$\begin{aligned}
&\ln\left[\pi(z_{(-1)k}=1|\alpha,\beta^*,\kappa,\delta,\phi,\phi_{\beta^*},\phi_\kappa,\phi_\delta,z_{-((-1)k)},\rho,y)\right]\\
&=\mathbb{I}_{\{z_{0k}=0\}}\ln(\rho)+\mathbb{I}_{\{z_{0k}=1\}}\ln(1-\rho)+\ln(1-\rho_0)+C.
\end{aligned}\tag{B.62}$$

It is worth pointing out that simplified expressions for the equations B.58 and B.60 is due to proposition A.4, as detailed in section A.2.