UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

# Variable selection algorithms for non-homogeneous hidden Markov models

**Gustavo A. Sabillón Lee**

Tese de Doutorado do Programa Interinstitucional de Pós-Graduação em Estatística (PIPGEs)

ICMC USP
SÃO CARLOS

**Gustavo A. Sabillón Lee**

# Variable selection algorithms for non-homogeneous hidden Markov models

Doctoral dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP and to the Department of Statistics – DEs-UFSCar, in partial fulfillment of the requirements for the degree of the Doctorate Interagency Program Graduate in Statistics. *FINAL VERSION*

Concentration Area: Statistics

Advisor: Profa. Dra. Daiane Aparecida Zuanetti

**USP – São Carlos**
**December 2024**

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

Sabillón Lee, Gustavo Alexis
S116v Variable selection algorithms for non-homogeneous hidden Markov models / Gustavo Alexis Sabillón Lee; orientadora Daiane Zuanetti. -- São Carlos, 2024.
140 p.

Tese (Doutorado - Programa Interinstitucional de Pós-graduação em Estatística) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2024.

1. Hidden Markov models. 2. Variable Selection. 3. LASSO. 4. Stochastic Expectation-Maximization. 5. Time Series Forecasting. I. Zuanetti, Daiane, orient. II. Título.

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

**Gustavo A. Sabillón Lee**

# Algoritmos para seleção de variáveis em modelos Markovianos ocultos não-homogêneos

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

Área de Concentração: Estatística

Orientadora: Profa. Dra. Daiane Aparecida Zuanetti

**USP – São Carlos**
**Dezembro de 2024**

**UNIVERSIDADE FEDERAL DE SÃO CARLOS**

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

## Folha de Aprovação

Defesa de Tese de Doutorado do candidato Gustavo Alexis Sabillón Lee, realizada em 29/10/2024.

### Comissão Julgadora:

Profa. Dra. Daiane Aparecida Zuanetti (UFSCar)

Prof. Dr. Rafael Izbicki (UFSCar)

Prof. Dr. Erlandson Ferreira Saraiva (UFMS)

Prof. Dr. Vinicius Diniz Mayrink (UFMG)

Prof. Dr. Carlos Tadeu Pagani Zanini (UFRJ)

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

*Para ustedes, abuela Mercedes, abuela Gloria y abuelo Julio.*

# ACKNOWLEDGEMENTS

Primeiro agradeço a Deus, nosso Padre e criador. Ele me acompanohou nesta joranda, me deu a sabedoria e a perseverança para conseguir cumprir com meus objetivos.

Aos meus pais, Norman e Gioconda, pelo seus exemplos de amor, bondade e superação. Em cada um dos momentos, eles estiveram comigo, me acompanhando e apoiando constantemente. Sem o apoio deles, nunca teria chegado onde estou.

A meu irmão Wilberto, por ser o melhor exemplo de honestidade e humildade que tenho na minha vida. A minha querida irmã Gloria Mercedes, por todo o carinho e a dedicação à família. Aos dois agradeço também por ser os melhores exemplos de pai e mãe que eu posso ter. Sem o exemplo deles, não teria sido possível realizar este sonho.

A minha orientadora, a professora Daiane Aparecida Zuanetti, por ter depositado a sua confiança em mim. Pela sua enorme paciência e por sempre ter a melhor disposição para ajudar nos momentos difíceis. Certamente, sem a orientação dela, eu não teria consegui meus objetivos. Muito obrigado por ser uma referência profissional e de excelência para mim.

Aos professores da USP e da UFSCar pelos valiosos ensinamentos, por fornecer uma excelente base para meu encaminhamento na área da Estatística.

Ao meus amigos Jadson Marcelino e Alex de la Cruz, por suas amizades, ensinamentos e disponibilidade para ajudar em todo momento. A todos os meus colegas do PIPGEs, pelos bons tempos compartilhados e pelas muitas horas de estudo.

A todas as pessoas que de alguma forma ajudaram, quero deixar um agradecimento especial, porque sem eles isto não teria sido possível.

Finalmente, quero agradecer às duas pessoas que significam todo para mim, minha esposa Jeny, e meu filho Samuel. Vocês são a inspiração que faz que eu continue me esforçando para ser melhor pai, melhor profissional e melhor pessoa. São a minha força no dia a dia. Sem vocês, nunca teria sonhado em chegar tão longe. Amo vocês com todo meu coração.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

# RESUMO

SABILLÓN LEE, G.A. **Algoritmos para seleção de variáveis em modelos Markovianos ocultos não-homogêneos**. 2024. 140 p. Tese (Doutorado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2024.

Modelos Markovianos ocultos não-homogêneos são um paradigma estatístico no qual uma sequência de estados não observáveis gera uma sequência de valores observáveis. Transições entre os estados não observáveis são controladas por coeficientes de transição e covariáveis. Contudo, estudos referentes a seleção de variáveis para este modelo têm sido pouco explorados. Devido a isto, o objetivo central desta tese é propor métodos de seleção de variáveis que melhorem o desempenho preditivo do modelo. Propomos duas versões do LASSO para o modelo Markoviano oculto não-homogêneo, o LASSO Global e LASSO Individual. Os métodos propostos são testados em um estudo de simulação para analisar seu desempenho sob condições controladas. As métricas de avaliação utilizadas são o erro quadrático médio preditivo, a precisão na predição da sequência não-observável e a eficiência do encolhimento dos coeficientes. Com relação ao erro quadrático médio preditivo, as propostas consistentemente mostram um desempenho preditivo melhor do que o ARIMA e a Regressão Linear Penalizada. Elas apresentam um desempenho muito bom na previsão da sequência de estados não observáveis que gera os valores observáveis. Em termos de eficiência do encolhimento dos coeficientes, as propostas mostram um desempenho excelente em todos os cenários de simulação, ao selecionar variáveis por meio do encolhimento dos coeficientes. Esse ganho no desempenho preditivo, bem como a capacidade de realizar a seleção de variáveis, torna os métodos propostos uma opção interessante para aplicação com o modelo. Finalmente, os métodos são aplicados para caracterizar e prever o regime de chuvas na cidade de São Carlos, Brasil, exibindo um bom desempenho na previsão das quantidades de chuva na região, bem como na seleção de covariáveis relevantes para o modelo.

**Palavras-chave:** Seleção de variáveis, LASSO, EM Estocástico, modelo Markoviano oculto, previsão de séries temporais.

# ABSTRACT

SABILLÓN LEE, G.A. **Variable selection algorithms for non-homogeneous hidden Markov models**. 2024. 140 p. Tese (Doutorado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2024.

Non-homogeneous hidden Markov models are a statistical paradigm in which a sequence of non-observable states generates a sequence of observable. Transitions between the non-observable states are controlled by transition coefficients and covariates. Because variable selection has been hardly explored for this model, the central purpose of this thesis is to propose variable selection methods which improve predictive performance of the model. We propose two versions of the LASSO for the non-homogeneous hidden Markov model, the Global LASSO and Individual LASSO. The proposed methods are tested in a simulation study, to analyze their performance under controlled conditions. Evaluation metrics used are the mean squared prediction error, non-observable sequence prediction accuracy and coefficient shrinkage efficiency. Regarding the mean squared prediction error, the proposals consistently show better predictive performance than ARIMA and Penalized Linear Regression. They show very good performance when predicting the non-observable state sequence which generates the observable values. In terms of coefficient shrinkage efficiency, the proposals show excellent performance in all simulation scenarios when selecting variables via coefficient shrinkage. This gain in predictive performance as well as the ability to perform variable selection makes the proposed methods an interesting option to apply with the model. Finally, the methods are applied to characterize and predict the rainfall regime in the city of São Carlos, Brazil, displaying good performance when predicting rainfall quantities in the region as well as selecting relevant covariates for the model.

**Keywords:** Variable selection, LASSO, stochastic-EM, hidden Markov models, time-series forecasting.

# CONTENTS

CHAPTER

# 1

# INTRODUCTION

Among the many challenges that researchers encounter in statistical data analysis, is the presence of sub-populations within the general population being studied. The problem of heterogeneity in data is amplified by the fact that in real-life applications, the individual observations in these sub-populations are rarely identified as belonging to one of the specific sub-populations. Several paradigms have been developed over the years to accurately model phenomena such as these. One of the most common and most explored of these paradigms in longitudinal data is called the hidden Markov model. Different authors have used several names for the hidden Markov model. Baum and Petrie (1966) first introduced the term *probabilistic functions of Markov chains*. Later on, Engel (1994) used the name *Markov switching model*, most likely relating this term to the dynamic nature of the latent components of these models. Afterwards, Jacquet, Seroussi and Szpankowski (2008) used the term *hidden Markov process* to refer to these models. In this thesis, we will use the term hidden Markov model.

## 1.1 Hidden Markov models

Hidden Markov models (HMM) are a type of model in which the system that is being described is considered a Markov process with non-observable states. The concept of Markov models has existed for quite some time and early authors such as Rabiner and Juang (1986) define them as a doubly stochastic process with an underlying stochastic process that is not observable, but can only be observed through another stochastic process that produces a sequence of observable values. Dymarski (2011) extends this definition by explaining that the non-observable stochastic process is a Markov chain, that is characterized by discrete states and transition probabilities. The second stochastic process emits observable values based on a state-dependent probability distribution. Poritz (1988) provides an analogous definition, stating that if this model is seen as a generative model where observations are emitted by the non-observable states, then the Markov chain synthesizes a sequence of states, called a path. The state-dependent probability

distributions then transform this path into a time-series. It is always relevant to mention that the term "hidden", when referring to a hidden Markov model, describes the states of the underlying Markov chain and not its parameters.

These models can be classified according to several criteria. The most common classification found in the literature separates these models according to the dynamics of the transition matrix of the hidden stochastic process. Using this criterion, the resulting classifications are named homogeneous and non-homogeneous. Huang, Huang and He (2019) mentions that the homogeneous HMM (HHMM) assumes constant transition probabilities, while non-homogeneous HMM (NHMM) assumes a varying transition probabilities which depend on a set of covariates. While both assumptions may seem plausible in different applications, a homogeneous HMM may not adequately model phenomena in which external factors influence the dynamics of the underlying non-observable stochastic process.

Applications of these models are widespread in many areas of research. Rabiner (1989a) presents an early example of applications of these models in speech recognition, in which HMMs are used to build an isolated word recognizer. This word recognizer identifies individual words in a vocabulary and models each word using a distinct HMM. These models are applied by Krogh *et al.* (1994) to deal with database searching and multiple sequence alignment of protein families and protein domains. According to Gollery (2005), sequence alignment is a way of arranging the sequences of DNA, RNA, or proteins to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences. Krogh *et al.* (1994) construct a HHMM and uses it to obtain multiple alignment of all the training sequences. It is also used to search the SWISSPROT database for other sequences that are members of the given protein family. The resulting HHMM performs well when it is tasked with the production of multiple alingments of different protein structures. Eddy (2004) describes HMMs as a formal foundation for making probabilistic models of linear sequence labeling problems. The author portrays this quality of HMMs with a toy example regarding biological sequence analysis. In this example, an HMM is used to identify sites in genes where specific sequences of DNA called exons and introns can be found.

Spezia (2006) uses NHMMs to analyse a time series of the maxima per day of the hourly mean concentrations of ozone gas in San Giorgio, Bergamo (Italy). The main focus of the investigation was the classification of the hidden states of the environment, according to their conditions which can favor ozone pollution. The results show that an NHMM performs well when attempting to separate the hidden states in terms of factors which may influence ozone pollution. Banachewicz, Lucas and Vaart (2008) apply the NHMM to model portfolio defaults using empirical U.S. default data. Among the conclusions, Banachewicz, Lucas and Vaart (2008) determines that GDP growth, the term structure of interest rates and stock market returns impact the state transition probabilities of the non-observable chain. The impact, however, is not uniform across different industries which were studied, therefore indicating a weak correspondence

between industry credit cycle dynamics and general business cycles.

## 1.2 Variable selection algorithms

In recent times, data sets with large numbers of covariates have become very common. This situation, which is greatly linked to a notable increase in processing power, as well an increased proficiency to capture data of different types and with high dimensionality, has generated an infamous question among researchers: How to select a subset of variables to fit the model with the best predictive performance? In the context of NHMM, this problem is of particular interest to researchers because covariates directly influence the dynamics of the non-observable Markov chain. This, in turn, will directly affect the observable values which are generated at every instant of time.

According to Adams and Beling (2019), variable selection is the process of reducing the number of collected features (covariates) to a relevant subset of features and is often used to combat the problem of high dimensionality. Variable selection can increase the performance of models by eliminating noise in the data, increase the training and prediction speed of the model, improve model interpretation as well as decrease the risk of overfitting. Common and traditional variable selection methods include forward selection and backward selection. The names of these methods are directly related to the direction in which the significant variable search is performed. Forward selection begins its search with no variables in the candidate model, and adds a new variable from the set of available variables after determining if the addition of the variable improves some previously selected statistical criterion to measure model quality. Backward selection involves the opposite procedure, in which you begin with a complete model, and it removes the variable which is least important to the model based on the previously selected criteria. These methods are adequate for several types of regression models, and are easy to implement. However, they have many setbacks. Blanchet, Legendre and Borcard (2008) explains that even though forward selection may be applied on a dataset with more covariates than observations, it is prone to overestimating the amount of explained variance, which is measured by the coefficient of multiple determination $R^2$. Chowdhury and Turin (2020) mentions that a significant drawback of backward variable selection is that once a variable is eliminated from the model it is not re-entered again. This may be a problem because a dropped variable may become significant later in the final model as it is inputted into the model with other variables.

Recently proposed methods that rely on penalization have gained popularity due to the fact that they can effectively deal with data whose number of variables is greater than the number of observations. Some of these methods include Ridge regression which was first introduced by Hoerl and Kennard (1970), the Smoothly Clipped Absolute Deviation penalization method, proposed by Fan and Li (2001), and Elastic Net Regularization, described by Zou and Hastie (2005). Essentially, these methods rely on specific penalty terms to perform shrinkage

on regression coefficients, such that those coefficients which are not relevant will effectively be shrunk approximately to zero.

### 1.2.1 Least Absolute Shrinkage and Selection Operator (LASSO)

Among these penalty driven methods, we also find the Least Absolute Shrinkage and Selection Operator (LASSO). It was first proposed by Tibshirani (1996) and as explained, LASSO minimizes the residual sum of squares subjected to the sum of the absolute value of the coefficients of the model, under the assumption that it is less than a constant. Tibshirani applied his proposed method on prostate cancer data from a study conducted by Stamey *et al.* (1989) that examines the correlation between PSA (prostate specific antigen) and several clinical measures in men who were about to undergo a prostatectomy. In the application, several factors related to PSA levels in patients who were about to receive a prostatectomy were used to determine a patient's pre-operation PSA level. LASSO successfully performs shrinkage on regression coefficients related to these factors, which otherwise had greater values when compared to other variable selection methods, such as subset selection methods.

After its inception, LASSO has been widely applied by researchers in several fields of knowledge. A study conducted by Steyerberg, Eijkemans and Habbema (2001) predict 30-day mortality after acute myocardial infarction. Huang, Ma and Zhang (2008) study the asymptotic properties of the adaptive LASSO estimators in sparse, high-dimensional, linear regression models when the number of covariates may increase with the sample size. In this paper, the authors consider variable selection using the adaptive LASSO, where the $L_1$ norms in the penalty are re-weighted by data-dependent weights. They apply the adaptive LASSO to a data set regarding tissue harvested from the eyes of 120 twelve week old lab rats. Their interest was finding the genes whose expression are correlated with that of a specific gene, namely gene TRIM32, which was found to cause Bardet-Biedl syndrome (CHIANG *et al.*, 2006), which is a genetically heterogeneous disease of multiple organ systems including the retina. Their results indicate that adaptive LASSO and traditional LASSO are similar, with the adaptive LASSO showing a slight advantage in terms of mean squared errors. Wu *et al.* (2009) evaluated the performance of LASSO penalized logistic regression in case–control disease gene mapping with a large number of SNPs (single nucleotide polymorphisms) predictors. The strength of the LASSO penalty was tuned to select a predetermined number of the most relevant SNPs and other predictors. Among the results of the study, the researchers found that analysis of simulated data demonstrated that LASSO penalized regression is easily capable of identifying pertinent predictors in grossly under-determined problems.

The previously cited studies show that applications of the general and specific versions of the LASSO are widespread. However, in the context of NHMMs, we find few instances of applications in which LASSO or other penalization methods are used for variable selection. Städler and Mukherjee (2013) propose a penalized log-likelihood procedure involving the $L_1$

norms of state specific inverse covariance matrices, carrying out optimization using a traditional Expectation-Maximization algorithm. In their study, the authors propose a fixed universal penalization parameter, which depends only on the sample size and number of covariates. The main purpose of this approach was to avoid problems with preprocessing of the covariates, assuming that the scale for each hidden state may be different, as well as decreasing computational costs for parameter estimation. This approach differs from the common method in which the regularization parameter is usually chosen empirically to minimize the prediction error. Choi *et al.* (2013) use a sparsely correlated HMM, which is a compromise between an independent HMM and a fullHMM, to perform genome-wide location analysis on multiple regulatory proteins or epigenetic marks. It is of great interest to understand the correlation among the factors that carry out some biological processes, a sparsely correlated HMM approach captures a small subset of non-ignorable correlations among data series to avoid modeling all pairwise correlations. This sparsity property is achieved by adopting a regularization regression strategy. Their results show that the scHMM algorithm is able to take advantage of the interactions between correlated series to improve the inference of hidden state, as well as reduce computational effort related to estimating parameters. Other authors such as Sabillón and Zuanetti (2023) use model selection criteria such as BIC and AICc for carrying out a simple variable selection in NHMMs when analyzing rainfall patterns in Honduras.

### 1.2.2 Bayesian methods

With the technological advances which have emerged in recent times and the increase in computational power available to researchers, we perceive an increase in the use of Bayesian statistical methods over the last few years. These methods tend to be computationally costly, but they present many attractive advantages when compared to frequentist methods. Under the Bayesian perspective, methods which are usually used to select variables are reversible jump algorithm or the use of shrinkage priors.

In a Bayesian context, shrinkage occurs in terms of the selection of the prior distribution of the parameters being targeted for shrinkage. Among the authors that have worked with shrinkage priors, Carvalho, Polson and Scott (2009) presents a general, fully Bayesian framework for sparse supervised-learning problems based on the horseshoe prior. The horseshoe prior is a member of the family of multivariate scale mixtures of normals, and is related to Laplacian priors (e.g. the LASSO). The proposed framework presents robustness at handling unknown sparsity and large outlying signals. In their work, the authors provide a theoretical and conceptual comparison of nine different shrinkage priors and parameterize the priors, if possible, in terms of scale mixture of normal distributions to facilitate comparisons. The authors' goal in this work is to characterize the horseshoe estimator as a default procedure that is well-behaved, that is computationally tractable, and that seems to outperform its competitors in a wide variety of sparse situations. Erp, Oberski and Mulder (2019) point out that Bayesian penalization is becoming

increasingly popular, in which the prior distribution performs a function similar to that of the penalty term in classical penalization. Specifically, the so-called shrinkage priors in Bayesian penalization aim to shrink small effects to zero while maintaining true large effects. Compared to classical penalization techniques, Bayesian penalization techniques perform similarly or sometimes even better, and they offer additional advantages such as readily available uncertainty estimates, automatic estimation of the penalty parameter, and more flexibility in terms of penalties that can be considered. Finally, a recently published article by Zhou and Song (2023) introduces the adaptive LASSO to accommodate the local sparsity of functional coefficients. This is done using Laplace priors in a Bayesian approach to jointly conduct estimation, variable selection, and the detection of zero-effect regions. This proposed approach incorporates the dependent Dirichlet process with stick-breaking prior for accommodating the unspecified distribution of the random effect and a blocked Gibbs sampler for efficient posterior sampling. The authors evaluate the performance of the proposed method through simulation studies, and the utility of the methodology is demonstrated by an application to the analysis of air pollution and meteorological data.

Spezia (2020) proposes a novel evolutionary Monte Carlo (EMC) algorithm for the selection of exogenous variables affecting the different rows of the transition matrices in a NHMM. The EMC is an MCMC method which processes a population of chains in parallel, with a different temperature attached to each chain. They use their model in an application to ozone dynamics, where results about covariate selection, choice of the number of hidden states, parameter estimation, hidden chain reconstruction, and classification are explored. Another algorithm which is becoming popular in recent years is the reversible jump algorithm. Some early examples of its usage can be found in Green (1995). In the article, the authors propose a new framework for the construction of reversible Markov chain samplers that jump between parameters spaces of differing dimensionality, with the intent of solving model determination problems. They illustrate their proposed framework using applications dealing with multiple change-point analysis in one and two dimensions, and to a Bayesian comparison of binomial experiments. Meligkotsidou and Dellaportas (2011) introduce an MCMC reversible jump algorithm for predictive inference of NHMMs, allowing for model uncertainty regarding the set of covariates that affect the transition matrix. They apply the model and the proposed algorithm to a data set of interest rates returns in the United States. Their results indicate that for the analyzed data set, the proposed methodology has better predictive ability than a standard HHMMs, and their general model formulation and algorithm improve the predictive ability of standard NHMMs. An extension of the work developed in Meligkotsidou and Dellaportas (2011) can be found in Koki, Meligkotsidou and Vrontos (2020). In the article, the authors model time series via predictive regressions with state dependent coefficients and time varying transition probabilities that depend on the predictors via a logistic function. In a hidden Markov setting, inference for logistic regression coefficients becomes complicated and in some cases impossible due to convergence issues. The authors paper address this problem using a new latent variable scheme

that utilizes the Pólya-Gamma class of distributions for the transition coefficients. Predictor selection and inference on the model parameters are based on a MCMC scheme using reversible jump. Single-step and multiple-steps-ahead predictions are obtained by the most probable model, median probability model or a Bayesian model averaging approach.

## 1.3 Proposals and structure

As we can clearly perceive, NHMM are being widely applied in several fields of knowledge. They are adequate to model and explain many different time-series or sequential phenomena. It is also apparent that variable selection methods for these models have not been explored in depth. These two key factors are the main motivation for this thesis.

In this thesis, we will propose and explore two variable selection methods for NHMMs. These methods are the LASSO for NHMMs, presented in two separate proposals named the Global LASSO and Individual LASSO. We have selected this method for its ability to perform variable selection by shrinking coefficients which are not relevant and for being computationally efficient. The document is organized as follows: Chapter 2 covers some theoretical concepts of NHMMs. In this chapter, the elements of an NHMM will be described, as well as the likelihood function and some particularities related to the model's structure. Chapter 3 presents the proposed variable selection method, LASSO for NHMMs. The LASSO is applied with the Stochastic EM algorithm, presented in Sabillón and Zuanetti (2023). An extensive simulation study is conducted and presented in Chapter 4 to test the performance of each method under controlled conditions. Chapter 5 applies the proposed methods to a rainfall dataset and, finally, in Chapter 6 we show a brief discussion about the results found and future studies

CHAPTER

2

# NON-HOMOGENEOUS HIDDEN MARKOV MODELS

The central focus of this work is exploring and proposing variable selection algorithms for non-homogeneous hidden Markov models. For this reason, this chapter will be of utmost importance to enhance our understanding of the non-homogeneous hidden Markov model, its inner workings and nuances.

As mentioned in the previous chapter, hidden Markov models may be characterized by a finite set of states, each being associated to a probability distribution. Transitions between these states are controlled by a set of probabilities contained in a probability matrix, commonly called transition matrix. Observations will be generated from the probability distribution associated to each hidden state. Hidden Markov models can be classified as being homogeneous or non-homogeneous, regarding the dynamics of the transition probabilities between the states of the non-observable Markov chain. These transition probabilities may be constant as is the case of the homogeneous hidden Markov model (HHMM), or they may be dependent on a set of covariates as is the case of the non-homogeneous hidden Markov model (NHMM).

The vast majority of phenomena which can be modeled using hidden Markov models are characterized by having transition probabilities which are dynamic over time. Several authors clearly establish that the constant nature of the transition matrix of a HHMM makes them inadequate to model many time-series phenomena. Holsclaw *et al.* (2017) mentions that the homogeneity of the HHMM is a factor which makes this model less flexible in practical applications, and one approach to improve this limitation is to make the transition probabilities dependent on a series of covariates. Lagona, Maruotti and Picone (2011) states that NHMM are a generalization of the class of HHMM by allowing the transition probabilities to vary over time as a function of a set of covariates. These facts make NHMM very versatile and widely applicable in different fields of knowledge.

## 2.1 Definition of a NHMM

As explained by Taylor (2020), the hidden Markov model has two defining properties. Firstly, it assumes that the observation at time $t$ is generated by a process whose hidden state $S_t$ is not revealed to an external observer. Secondly, it assumes that the state of this hidden process satisfies the Markov property, that is, given the value of $S_{t-1}$, the current state $S_t$ is independent of all the states before $t-1$. These properties are crucial because they are the distinguishing characteristics of a hidden Markov model, and they extend to our model of interest, the NHMM. The elements which comprise an NHMM are:

1. A discrete space of non-observable states, $\varphi = \{1, 2, \ldots, K\}$;

2. A set of observable values which may be discrete or continuous, depending on the variable which will be observed, $\omega = \{0, 1, 2, \ldots, M\}$, $\omega = \{0, 1, 2, \ldots\}$ or $\omega = \mathbb{R}$;

3. A discrete random variable $S_t$ for $t = 1, 2, \ldots, T$, which assumes the values of the discrete space of non-observable values $\varphi$, at different moments over time;

4. A random variable $Y_t$ for $t = 1, 2, \ldots, T$, which assumes the values of the set of observable values $\omega$, at different moments over time;

5. An initial probability distribution for the non-observable states $\mathbf{p} = \{p_i\}$ for $i = 1, 2, \ldots, K$, such that $p_i = P(S_1 = i)$ and $\sum_{i=1}^{K} p_i = 1$;

6. A row vector $\mathbf{X}_t = (X_{t1}, \ldots, X_{tD})$, for $t = 1, 2, \ldots, T$, which represents the values of the $D$ observed covariates which influence the transition probabilities between hidden states at times $t-1$ and $t$;

7. A coefficient matrix $\boldsymbol{\beta}$ containing elements $\boldsymbol{\beta}_{ij} = \begin{pmatrix} \beta_{ij1} \\ \cdot \\ \cdot \\ \cdot \\ \beta_{ijD} \end{pmatrix}$ which are vectors containing $D$ coefficients, each associated to an observed covariate, where $\beta_{ij1}$ may be considered the intercept, accompanied by covariate $X_{t1} = 1$ for all $t$;

8. A probability distribution for the transitions between non-observable states $\mathbf{A}_t = \left\{a_{ijt}\right\}$, such that $a_{ijt} = h(\mathbf{X}_t \boldsymbol{\beta}_{ij})$ where $h(\cdot)$ is a link function such that $0 \leq h(\cdot) \leq 1$ and $\sum_{j=1}^{K} h(\mathbf{X}_t \boldsymbol{\beta}_{ij}) = 1$;

9. A probability distribution for the observations associated to each non-observable state indexed by parameters $\boldsymbol{\theta}_j$, and whose density or probability function is given by $f(y_t \mid S_t = j) = f(y_t \mid \boldsymbol{\theta}_j)$.

### 2.1.1 Marginal distribution of the observable values

As mentioned in the previous section, an NHMM is characterized by the probability of the initial state of the Markov chain, $\mathbf{p} = \{p_i\}$ such that $p_i = P(S_1 = i)$ and by the transition probabilities given by $P(S_t = j \mid S_{t-1} = i) = a_{ijt}$. Using conditional probability, we now have that the joint distribution of non-observable variables is given by

$$\begin{aligned} P(\mathbf{S} \mid \mathbf{p}, \boldsymbol{\beta}, \mathbf{X}) =& P(S_1, \ldots, S_T \mid \mathbf{p}, \boldsymbol{\beta}, \mathbf{X}) \\ =& P(S_1 \mid \mathbf{p}) P(S_2 \mid S_1, \boldsymbol{\beta}, \mathbf{X}) \ldots P(S_T \mid S_{T-1}, \boldsymbol{\beta}, \mathbf{X}) \\ =& p_{s_1} a_{s_1 s_2 2} \ldots a_{s_T s_{T-1} T} \\ =& p_{s_1} \prod_{t=2}^{T} a_{s_{t-1} s_t t}. \end{aligned} \tag{2.1}$$

By assuming that the observations are conditionally independent when the $\mathbf{S}$ sequence is given, the conditional joint distribution for the observable variables is given by

$$\begin{aligned} P(\mathbf{Y} \mid \mathbf{S}, \boldsymbol{\theta}) =& \prod_{t=1}^{T} P(Y_t \mid S_t, \boldsymbol{\theta}) \\ =& \prod_{t=1}^{T} f(y_t | \boldsymbol{\theta}_{s_t}), \end{aligned} \tag{2.2}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$. We must mention that when $Y_t$ is discrete, then $f(y_t | \boldsymbol{\theta}_{s_t}) = P(Y_t = y_t \mid \boldsymbol{\theta}_{s_t})$. From Equations (2.1) and (2.2), the joint probability of $\mathbf{Y}$ and $\mathbf{S}$ can be written as

$$P(\mathbf{Y}, \mathbf{S} \mid \mathbf{p}, \boldsymbol{\beta}, \mathbf{X}, \boldsymbol{\theta}) = P(\mathbf{Y} \mid \mathbf{S}, \boldsymbol{\theta}) P(\mathbf{S} \mid \mathbf{p}, \boldsymbol{\beta}, \mathbf{X}). \tag{2.3}$$

Finally, the sum over all possible hidden state sequences $\mathbf{s}$ will be calculated on Equation (2.3) resulting in the following equation

$$\begin{aligned} P(\mathbf{Y} \mid \mathbf{p}, \boldsymbol{\beta}, \mathbf{X}, \boldsymbol{\theta}) =& \sum_{s_1, s_2, \ldots, s_T} P(\mathbf{Y} \mid \mathbf{S}, \boldsymbol{\theta}) P(\mathbf{S} \mid \mathbf{p}, \boldsymbol{\beta}, \mathbf{X}) \\ =& \sum_{s_1, s_2, \ldots, s_T} p_{s_1} f(y_1 \mid \theta_{s_1}) \prod_{t=2}^{T} a_{s_{t-1} s_t t} f(y_t | \boldsymbol{\theta}_{s_t}), \end{aligned} \tag{2.4}$$

which denotes the marginal distribution of $\mathbf{Y}$.

### 2.1.2 Augmented likelihood function

An important function which will be used for parameter estimation in later sections is the likelihood function for the parameters of interest. The likelihood function of the parameters

of the model is given by

$$\begin{aligned}\mathscr{L}(\boldsymbol{\theta},\mathbf{p},\boldsymbol{\beta} \mid \mathbf{y},\mathbf{s},\mathbf{X}) =& P(\mathbf{S}=\mathbf{s} \mid \mathbf{p},\boldsymbol{\beta},\mathbf{X})P(\mathbf{Y}=\mathbf{y} \mid \mathbf{S}=\mathbf{s},\boldsymbol{\theta}) \\ =& P(S_1=s_1 \mid \mathbf{p}) \left(\prod_{t=2}^{T} P(S_t=s_t \mid S_{t-1}=s_{t-1},\boldsymbol{\beta},\mathbf{X})\right)\left(\prod_{t=1}^{T} P(Y_t=y_t \mid S_t=s_t,\boldsymbol{\theta})\right).\end{aligned} \tag{2.5}$$

The specific formulation of the likelihood function shown will be determined by the specific characteristics of the following elements:

- The distribution of the observable variables $Y_t \mid S_t$ and
- The link function, $h(\mathbf{X}_t\boldsymbol{\beta}_{ij})$, used to calculated $P(S_t=j \mid S_{t-1}=i,\boldsymbol{\beta},\mathbf{X})$.

Specific formulations will be discussed in detail in upcoming sections and chapters.

### 2.1.3 The non-homogeneous transition matrix

The characteristic which distinguishes the NHMM from the HHMM is the dynamic nature of the transition matrix. Due to this fact, this section will be dedicated to offering a clear perspective on the structure of the matrix and its role in the transitions between hidden states. For such purposes, we introduce a toy example, as presented in Sabillón (2020).

Consider a NHMM with two non-observable states. For such model, the transition coefficient matrix will be given by the following expression:

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_{11} & \boldsymbol{\beta}_{12} \\ \boldsymbol{\beta}_{21} & \boldsymbol{\beta}_{22} \end{bmatrix}, \tag{2.6}$$

where each element of the transition coefficient matrix is a column vector containing $D$ positions. If we consider $D=2$ covariates, in which one of the covariates is associated to the intercept, then every element has two positions. For example, $\boldsymbol{\beta}_{21}$ is given by the following expression:

$$\boldsymbol{\beta}_{21} = \begin{pmatrix} \beta_{211} \\ \beta_{212} \end{pmatrix}, \tag{2.7}$$

where $\beta_{211}$ is the intercept and $\beta_{212}$ will be multiplied by the covariate $X$ at time $t$ when we calculate the probability of transitioning from state 2 to state 1 at time $t$. The entries (covariates) at time $t$ are given by the following expression:

$$\mathbf{X}_t = \begin{pmatrix} 1 & x_t \end{pmatrix}. \tag{2.8}$$

The transition probability between state $i$ and $j$ is then calculated by the link function $h(\mathbf{X}_t\boldsymbol{\beta}_{ij})$ described in Element 8 found of Section 2.1.

Let us consider that the link function for the model is the *SoftMax* link function, defined by Gao and Pavel (2018). This choice of link function will be discussed with greater detail in Section 4.2.1. The *SoftMax* function is defined by

$$a_{ijt} = \frac{exp(\mathbf{X}_t \boldsymbol{\beta}_{ij})}{\sum_{l=1}^{K} exp(\mathbf{X}_t \boldsymbol{\beta}_{il})}, \qquad 1 \leq i, j \leq K \tag{2.9}$$

where $a_{ijt}$ refers to the transition probability from state $i$ to state $j$ at time $t$.

In order to clearly grasp the function of the transition matrix, let us suppose that the hidden Markov chain is at non-observable state 2 at time $t$. Because of the fact that this example refers to a model with only two hidden states, then there are only two possible transitions:

1. Exit state 2 at time $t$ and enter state 1 at time $t+1$;
2. Continue on state 2 at $t+1$.

To calculate the probabilities of being in state 2 at time $t$ and either going to state 1 or continuing on state 2 at time $t+1$, we will utilize the previously mentioned elements to calculate the quantities $a_{21t}$ and $a_{22t}$.

Therefore, we have that $a_{21t}$ is given by:

$$a_{21t} = \frac{exp(\mathbf{X}_t \boldsymbol{\beta}_{21})}{\sum_{j=1}^{K} exp(\mathbf{X}_t \boldsymbol{\beta}_{2j})} = \frac{exp(\beta_{211} + \beta_{212} \cdot x_t)}{exp(\beta_{211} + \beta_{212} \cdot x_t) + exp(\beta_{221} + \beta_{222} \cdot x_t)} \tag{2.10}$$

and we also have that $a_{22t}$ is given by:

$$a_{22t} = \frac{exp(\mathbf{X}_t \boldsymbol{\beta}_{22})}{\sum_{j=1}^{K} exp(\mathbf{X}_t \boldsymbol{\beta}_{2j})} = \frac{exp(\beta_{221} + \beta_{222} \cdot x_t)}{exp(\beta_{211} + \beta_{212} \cdot x_t) + exp(\beta_{221} + \beta_{222} \cdot x_t)}. \tag{2.11}$$

In the case of the previously discussed example, we can state that $S_t \mid S_{t-1} = 2$ follows a *Discrete* distribution, with two classes whose probabilities are given by $a_{21t}$ e $a_{22t}$, respectively, or:

$$S_t \mid S_{t-1} = 2 \sim Discrete(a_{21t}, a_{22t}), \qquad a_{21t} + a_{22t} = 1. \tag{2.12}$$

This process is repeated at every moment in time throughout the entire hidden Markov chain, in order to calculate the probability of occurrence of each hidden state at every moment in time.

## 2.2 The 3 challenges of a NHMM

When dealing with HMM, and specifically with NHMM, we will be confronted with 3 problems which are intrinsic of the model. These problems have been approached by many

authors over the years. Rabiner (1989b) discusses these problems in detail, and states that in order for the model to be useful in real-world applications, these problems must be solved first. To simplify the notation for the description of these problems, we have that $\eta = (\mathbf{p}, \mathbf{A}, \boldsymbol{\theta})$. The 3 problems are:

1. Given model $\eta = (\mathbf{p}, \mathbf{A}, \boldsymbol{\theta})$ and the sequence of observable random variables $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_T)$, how do we efficiently calculate the probability that sequence was generated by model $\eta$? In other words, how to efficiently calculate $P(\mathbf{Y} \mid \eta)$ which is part of the likelihood function?

2. Given model $\eta = (\mathbf{p}, \mathbf{A}, \boldsymbol{\theta})$ and the sequence of observed values $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_T)$, among the different sequences of non-observable states which could have generated the sequence of observed values, which is the most likely?

3. How do we estimate the parameters of model $\eta = (\mathbf{p}, \mathbf{A}, \boldsymbol{\theta})$?

Zuanetti and Milan (2017) clearly describes these 3 *canonical* problems as follows: the first problem pertains to calculating the probability that a specific sequence of observable values occurs, given that the structure and the parameters of the distributions involved are known. In other words, we want to find the marginal distribution of $\mathbf{Y}$. Rabiner (1989a) calls this problem, the "evaluation problem", and explains it as the process of scoring how well a given model matches a given observation sequence.

The second problem is defined as identifying the most probable sequence of non-observable states given the sequence of observations and the model. It deals with the prediction of $\mathbf{S}$, which is extremely useful in practical applications. Rabiner (1989a) outlines this problem as attempting to uncover the hidden part of the model or, in other words, finding the “correct” non-observable state sequence. However, it should be understood that there is no such thing as a “correct” state sequence to be found. What is usually done is that an optimality criterion is established to reach a solution to this problem as best as possible.

The third problem is self-explanatory, as described by Zuanetti and Milan (2017), and it involves estimating the model parameters. Rabiner (1989a) calls this problem the "training problem" and describes it as attempting to optimize the model parameters to best describe how an observation sequence is generated. The solution of this third problem is extremely important for applications of NHMM, since it allows to adapt model parameters to observed data, creating optimal models for real phenomena.

The most traditional approach for dealing with these problems and estimating the parameters of a HMM is still using the Baum-Welch algorithm for HMMs (RABINER, 1989a; ZUCCHINI; MACDONALD; LANGROCK, 2017; MARUOTTI; ROCCI, 2012). It is a maximum likelihood estimation method which combines the Expectation-Maximization (EM) algorithm

(DEMPSTER; LAIRD; RUBIN, 1977) with forward-backward and Viterbi algorithms in order to gain computational efficiency.

To work around several problems related to the EM algorithm, such as slow convergence, convergence towards local solutions and starting point sensitivity, the stochastic EM algorithm and Bayesian methods have been proposed and used to estimate a HMM. More detail about these methods are given in Sabillón and Zuanetti (2023). Here, given that we will combine estimation and variable selection in the same method, we discuss the proposed algorithms in next chapters.

CHAPTER

# 3

# STOCHASTIC EXPECTATION-MAXIMIZATION WITH PENALIZATION VIA LASSO METHOD

In this chapter we present our proposal for variable selection in NHMMs, the LASSO. We formulate two different versions of the LASSO penalization term, a global penalization and a state-specific penalization. These penalizations will be applied along with the Stochastic EM algorithm to perform parameter estimation.

## 3.1 Introduction

The application of NHMMs has become widespread in recent years. This fact has confronted the NHMM with a habitual problem for all statistical models: variable selection. High dimensionality and sparsity pose two great challenges for model fitting in any family of models. When there are large amounts of information available, predictive models will usually perform better, however, there are a wide range of issues associated with high-dimensional data. Among these issues, we have increased model training time, algorithm complexity as well as computational issues such as data storage space. Particularly, in NHMMs, the impact of high dimensionality and sparsity on transition probabilities may represent the difference between an adequate fit or overfitting.

There are several methods to perform variable selection which have been applied in the general context of statistical modelling. However, there are few works in literature which apply these methods in NHMMs. Some of these methods applied to models which are similar to NHMMs include filter feature selection. Filter feature selection has been applied for HHMMs as shown in Zhu, Hong and Wong (2008), where a discriminant feature selection approach for HHMM is applied to study micro-milling tool conditions. In this previously mentioned study,

a modification of Fisher's linear discriminant analysis (FDA) is used. Another type of feature selection is called wrapper variable selection, and an example is presented in Günter and Bunke (2003). Their proposal is intended to work with HMMs in the context of speech recognition. The authors' feature selection algorithm uses a new objective function that quickly computes an approximation of the recognition rate on a validation set to assign a measure of quality to each feature set. In general, feature selection methods have been seldom explored, in the context of NHMMs. This becomes more evident when the variable selection method of interest is penalized estimation. One of the few examples of applications of penalized methods on NHMMs can be found in the work of Städler and Mukherjee (2013). In their work, the authors propose penalized estimation for a HMM with a multivariate Normal observable random vector. Specifically, they apply $L_1$-penalized estimation on the state-specific inverse covariance matrices. Their proposal is based on a universal penalization parameter, which is fixed and is constructed as a function of the number of observations and the number of covariates. The main purpose of their approach is to obtain sparse inverse covariance matrices which can be interpreted as state-specific conditional independence graphs or networks. They apply their approach in Genome Biology, a field in which sparsity is particularly useful due to the fact that it is very common to have a large amount of covariates available. Finally, it is important to mention the work of Zhou and Song (2023). In this recently published article, the authors develop a Bayesian approach to jointly conduct estimation and variable selection in a functional HMM. They make use of the adaptive LASSO through Laplace priors to induce shrinkage of the transition coefficients. They demonstrate the utility of the proposed methodology in an application dealing with the analysis of air pollution and meteorological data. Under a Bayesian perspective, the implementation of Laplace priors in an MCMC framework is analogous to applying LASSO in an optimization procedure. Nonetheless, given that their proposed method relies on MCMC sampling, the computational cost of such method is high, leading to greater processing times.

To the extent of our knowledge, there are no instances in literature of works dealing with penalized estimation of the transition coefficients of a NHMM under a frequentist perspective. Due to this fact, the main motivation of this chapter is to propose a penalized estimation method for the regression coefficients of the transition matrix. This is of utmost importance in the context of NHMMs because, as described in Section 2.1, the transition coefficients are the driving component behind the transition dynamics of the NHMM. The transition coefficients, along with the covariates, determine what will be the non-observable state at time $t$, and this in turn, will determine the state-specific distribution which will generate an observable value. In a predictive setting, variable selection via penalized estimation will greatly improve predictive capabilities of an NHMM. This chapter is organized as follows: Section 3.2 presents a univariate NHMM with Normal response variable as well as some of its particular characteristics and Section 3.3 introduces our proposals for penalized estimation of the transition parameters of an NHMM using a Stochastic EM algorithm.

## 3.2 Model specification

In this section we will describe the univariate Normal NHMM, based on the structure and characteristics presented in Section 2.1. With the intent of setting up a simulation scheme, as well as improving model comprehensibility, we will now establish that the elements of the NHMM fulfill the following assumptions:

- Assume that $\theta_i = (\mu_i, \sigma_i)$, for $i = 1, \ldots, K$;
- Assume that $Y_t | S_t = i \sim Normal(\mu_i, \sigma_i^2)$ for $i = 1, \ldots, K$ and $t = 1, \ldots, T$ where $\mu_i$ and $\sigma_i$ are the mean and standard deviation of the *Normal* distribution corresponding to the observable values, and;
- Assume that $n_1, \ldots, n_K$ represent the sample sizes for the groups of observations belonging to each of the non-observable states. That said, we have that $n_1 + \ldots + n_K = T$.

Using these specific facts, from Equation (2.5) we now have that the augmented likelihood function for the parameters is given by the following expression:

$$
\begin{aligned}
\mathscr{L}(\boldsymbol{\theta}, \mathbf{p}, \boldsymbol{\beta} \mid \mathbf{y}, \mathbf{s}, \mathbf{X}) =& P(\mathbf{Y} = \mathbf{y} \mid \mathbf{S} = \mathbf{s}, \boldsymbol{\theta}) P(\mathbf{S} = \mathbf{s} \mid \mathbf{p}, \boldsymbol{\beta}, \mathbf{X}) \\
=& P(S_1 = s_1 \mid \mathbf{p}) \left( \prod_{t=2}^{T} P(S_t = s_t \mid S_{t-1} = s_{t-1}, \boldsymbol{\beta}, \mathbf{X}) \right) \left( \prod_{t=1}^{T} P(Y_t = y_t \mid S_t = s_t, \boldsymbol{\theta}) \right) \\
=& \left( \prod_{j=1}^{K} p_j^{\mathbb{I}_{s_1}(j)} \right) \prod_{t=1}^{T} \left[ \prod_{j=1}^{K} \left( \frac{1}{\sqrt{2\pi\sigma_j^2}} exp\left\{ -\frac{(y_t - \mu_j)^2}{2\sigma_j^2} \right\} \right)^{\mathbb{I}_{s_t}(j)} \right. \\
& \left. \times \left( \prod_{i=1}^{K} \prod_{j=1}^{K} \left( \frac{exp(\mathbf{X}_t \boldsymbol{\beta}_{ij})}{\sum_{l=1}^{K} exp(\mathbf{X}_t \boldsymbol{\beta}_{il})} \right)^{\mathbb{I}_{s_{t-1}}(i) \mathbb{I}_{s_t}(j)} \right) \right] \\
=& \left( \frac{1}{\sqrt{2\pi}} \right)^T \prod_{j=1}^{K} \left[ \left( \prod_{t:s_t=j} \frac{1}{\sqrt{\sigma_j^2}} \right) exp\left\{ -\sum_{t:s_t=j} \frac{(y_t - \mu_j)^2}{2\sigma_j^2} \right\} \right] \left( \prod_{j=1}^{K} p_j^{\mathbb{I}_{s_1}(j)} \right) \\
& \times \left( \prod_{t=2}^{T} \prod_{i=1}^{K} \prod_{j=1}^{K} \left( \frac{exp(\mathbf{X}_t \boldsymbol{\beta}_{ij})}{\sum_{l=1}^{K} exp(\mathbf{X}_t \boldsymbol{\beta}_{il})} \right)^{\mathbb{I}_{s_{t-1}}(i) \mathbb{I}_{s_t}(j)} \right) \\
=& \left( \frac{1}{\sqrt{2\pi}} \right)^T \prod_{j=1}^{K} \left[ p_j^{\mathbb{I}_{s_1}(j)} \left( \left( \frac{1}{\sqrt{\sigma_j^2}} \right)^{n_j} \right) exp\left\{ -\sum_{t:s_t=j} \frac{(y_t - \mu_j)^2}{2\sigma_j^2} \right\} \right. \\
& \left. \times \prod_{i=1}^{K} \prod_{(t:t \geq 2, (s_{t-1}, s_t) = (i,j))} \frac{exp(\mathbf{X}_t \boldsymbol{\beta}_{ij})}{\sum_{l=1}^{K} exp(\mathbf{X}_t \boldsymbol{\beta}_{il})} \right].
\end{aligned}
\tag{3.1}
$$

A commonly used practice is to apply the logarithm on the likelihood function of any model of interest. This will facilitate the calculation of the maximum likelihood estimators for the

parameters which have a closed form. Calculating the logarithm of the likelihood poses no problem in the context of optimization because the logarithm is a monotonically increasing function. This fact is important because it ensures that the maximum value of the log of the probability occurs at the same point as the original probability function. Therefore, our maximum likelihood estimation will be performed using the simpler log-likelihood instead of the original likelihood. By applying the logarithm on Equation (3.1), we have that the log-likelihood of the parameters of the model is given by:

$$log\mathscr{L}(\boldsymbol{\theta},\mathbf{p},\boldsymbol{\beta} \mid \mathbf{y},\mathbf{s},\mathbf{X}) = -\frac{T}{2}log(2\pi) + \sum_{j=1}^{K}\left[\mathbb{I}_{s_1}(j)log(p_j) - \frac{n_j}{2}log(\sigma_j^2) - \sum_{t:s_t=j}\frac{(y_t-\mu_j)^2}{2\sigma_j^2} + \sum_{i=1}^{K}\sum_{(t:t\geq 2,(s_{t-1},s_t)=(i,j))}\mathbf{X}_t\boldsymbol{\beta}_{ij} - log\sum_{l=1}^{K}exp(\mathbf{X}_t\boldsymbol{\beta}_{il})\right]. \tag{3.2}$$

Here, we highlight that the distribution of observable variables can be exchanged for other univariate or multivariate distribution and the method can be straightforward adapted.

## 3.3 Parameter estimation using stochastic Expectation-Maximization algorithm

An important point for any statistical model is the estimation of its parameters of interest, and in this section we will explore the estimation of those parameters using a variation of the traditional EM algorithm, specifically developed for NHMMs.

The traditional EM algorithm has been widely used in the context of models with latent variables, because of the fact that can handle missing data. Several authors apply the EM algorithm to estimate sets of parameters of hidden Markov models. Cappé (2011) propose an online parameter estimation algorithm that combines two key ideas: an Expectation-Maximization (EM) methodology, consisting in reparameterizing the problem using complete-data sufficient statistics and exploiting a purely recursive form of smoothing in HMMs based on an auxiliary recursion. The proposed online EM algorithm resembles a classical stochastic approximation (or Robbins–Monro) algorithm, and the authors' objective was to resist conventional analysis of convergence. They provide limited results which identify the potential limiting points of the recursion as well as the large-sample behavior of the quantities involved in their EM algorithm. Gao and Song (2011) develop an extension of the EM algorithm in the framework of composite likelihood estimation given missing data or latent variables. The authors establish key theoretical properties of the composite likelihood EM (CLEM) algorithm: the ascent property, algorithmic convergence, and convergence rate. Their proposed EM algorithm is applied to estimate the transition probabilities in a multivariate hidden Markov model. Bietti, Bach and Cont (2015)

introduce a new incremental EM algorithm for HMMs and show that it compares favorably to existing online EM algorithms. The motivation behind their work is to provide online algorithms that perform segmentation and clustering in one pass. Rather than separately detecting changes and finding similarities, the algorithm performs online unsupervised joint segmentation and clustering. This enables real-time applications as well as scalability of such systems to very large databases and signals. Finally, they test their proposal and present results for real-time segmentation of musical notes and acoustic scenes.

The proposal we introduce in this thesis consists of performing penalized estimation of model parameters using a Stochastic EM algorithm. This Stochastic EM is introduced by Sabillón and Zuanetti (2023). The aforementioned algorithm uses randomly drawn values of the non-observable states of the Markov chain generated at every iteration of the algorithm, instead of using expectation to remove these missing values as is done in traditional EM algorithms. This fact makes the Stochastic EM less prone to finding local maxima as solutions to the optimization problem being solved and also simplifies the calculations and implementation of the algorithm. We will now describe our proposal and all the steps it includes.

A prerequisite before applying any EM-type algorithm is to determine maximum likelihood estimators for the parameters of interest, in case that those estimators (maximum points) have a known form. As shown in Equation (3.2), there is a constraint for the set of initial probabilities of the hidden Markov chain. This constraint can be described as $\sum_{j=1}^{K} p_j = 1$. Due to this restriction on the initial probabilities $p_j$, we will make use of Lagrange multipliers to perform maximization. We now have that the log-likelihood of the parameters along with the inclusion of the Lagrangian multiplier term is given by:

$$\begin{aligned}\ell(\boldsymbol{\theta},\mathbf{p},\boldsymbol{\beta} \mid \mathbf{y},\mathbf{s},\mathbf{X}) = &-\frac{T}{2}log(2\pi)+\sum_{j=1}^{K}\left[\mathbb{I}_{s_1}(j)log(p_j)-\frac{n_j}{2}log(\sigma_j^2)-\sum_{t:s_t=j}\frac{(y_t-\mu_j)^2}{2\sigma_j^2}\right.\\ &\left.+\sum_{i=1}^{K}\sum_{(t:t\geq 2,(s_{t-1},s_t)=(i,j))}\mathbf{X}_t\boldsymbol{\beta}_{ij}-log\sum_{l=1}^{K}exp(\mathbf{X}_t\boldsymbol{\beta}_{il})\right]\\ &+\alpha_0\left(1-\sum_{j=1}^{K}p_j\right),\end{aligned} \tag{3.3}$$

where $\alpha_0$ is the Lagrangian multiplier used for the maximization of the equation. To obtain the maximum likelihood estimators for the parameters of interest, we now derive Equation (3.3) in terms of each of the parameters and equal the result of that derivative to zero. First, to obtain the estimator for $\mu_j$ we have

$$\begin{aligned}\frac{\partial \ell\ (\boldsymbol{\theta},\mathbf{p},\boldsymbol{\beta} \mid \mathbf{y},\mathbf{s},\mathbf{X})}{\partial \mu_j} &= 0\\ \hat{\mu}_j &= \frac{\sum_{t:s_t=j} y_t}{n_j}.\end{aligned} \tag{3.4}$$

Afterwards, to obtain the maximum likelihood estimator for $\sigma_j^2$ we have

$$\begin{aligned} \frac{\partial \ell\ (\boldsymbol{\theta}, \mathbf{p}, \boldsymbol{\beta} \mid \mathbf{y}, \mathbf{s}, \mathbf{X})}{\partial \sigma_j^2} &= 0 \\ \hat{\sigma}_j^2 &= \frac{\sum_{t:s_t=j}(y_t - \hat{\mu}_j)^2}{n_j}. \end{aligned} \tag{3.5}$$

To find the maximum likelihood estimator for the initial probabilities of the non-observable Markov chain, we will now have to find the value of the $\alpha_0$ constant. Once more, by obtaining the derivative of Equation (3.3) and then equaling to zero, we have that

$$\begin{aligned} \frac{\partial \ell\ (\boldsymbol{\theta}, \mathbf{p}, \boldsymbol{\beta} \mid \mathbf{y}, \mathbf{s}, \mathbf{X})}{\partial p_j} &= 0 \\ \frac{\mathbb{I}_{s_1}(j)}{p_j} - \alpha_0 &= 0 \\ \hat{p}_j &= \frac{\mathbb{I}_{s_1}(j)}{\alpha_0}. \end{aligned} \tag{3.6}$$

To be able calculate the value of the $\alpha_0$ constant, we now rely on the fact that:

$$p_1 + p_2 + \ldots + p_K = 1 \Rightarrow \sum_{j=1}^{K} \hat{p}_j = 1. \tag{3.7}$$

Then, we can now deduce the value of $\alpha_0$ as,

$$\sum_{j=1}^{K} \hat{p}_j = 1 \Rightarrow \sum_{j=1}^{K} \frac{\mathbb{I}_{s_1}(j)}{\alpha_0} = 1 \Rightarrow \sum_{j=1}^{K} \mathbb{I}_{s_1}(j) = \alpha_0. \tag{3.8}$$

Therefore, the maximum likelihood estimator for the initial probabilities, $p_j$ is given by:

$$\hat{p}_j = \mathbb{I}_{s_1}(j), \tag{3.9}$$

since $\sum_{j=1}^{K} \mathbb{I}_{s_1}(j) = 1$.

After deriving the estimators for $p_j$, $\mu_j$ and $\sigma_j$, we will calculate the estimators for $\boldsymbol{\beta}_{ij}$. Because of the nature of the model's likelihood function and, specifically, the nature of the link function, the estimators for $\boldsymbol{\beta}_{ij}$ will not have a closed form. Given this fact, we will use numerical optimization methods to calculate the value of the estimates of $\boldsymbol{\beta}_{ij}$.

The main focus of this chapter is to present a penalized estimation method for the transition coefficients of the NHMM. For this purpose, we now introduce a penalization term on the log-likelihood function based on the $L_1$-norm as presented by Tibshirani (1996) and Friedman, Hastie and Tibshirani (2010). Let's assume that $log\mathscr{L}(\boldsymbol{\beta} \mid \mathbf{y})$ is the log-likelihood function for some set of parameters of interest. These authors describe the LASSO estimate for the parameters as

$$\hat{\boldsymbol{\beta}} = \arg\max \{log\mathscr{L}(\boldsymbol{\beta} \mid \mathbf{y})\} \text{ subject to } \sum_j |\beta_j| \leq c, \tag{3.10}$$

where $c \geq 0$ is a tuning parameter, and controls the amount of shrinkage that is applied to the estimates of the parameters. Again using Lagrange multipliers, this means that the actual objective function of the optimization process to find parameters' estimates will have the following form:

$$log\mathscr{L}(\boldsymbol{\beta} \mid \mathbf{y}) - \lambda \sum_{j} |\beta_j|, \text{ where } \lambda \geq 0. \tag{3.11}$$

Using the general penalization term described in Equation (3.11), we have constructed 2 different versions of the objective function of NHMM model which will be subjected to penalization. In the first version, the $\lambda$ parameter which controls penalization is fixed for the regressions related to all the transitions. In other words, each regression will receive the same penalization term, and shrinkage of the transition coefficients will be performed globally. We call this first proposal the Global LASSO. Since the target of this optimization procedure is to obtain the $\boldsymbol{\beta}_{ij}$ transition coefficients then the first version of the objective function with a global $\lambda$ parameter will have the following form:

$$\left[\sum_{i=1}^{K}\sum_{j=1}^{K}\sum_{(t:t\geq 2,(s_{t-1},s_t)=(i,j))} \mathbf{X}_t\boldsymbol{\beta}_{ij} - log\sum_{l=1}^{K} exp(\mathbf{X}_t\boldsymbol{\beta}_{il})\right] - \lambda(\sum_{i=1}^{K}\sum_{j=1}^{K}\sum_{m=2}^{D}|\beta_{ijm}|). \tag{3.12}$$

For the second version of the objective function, we will now consider a vector of tuning parameters, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)$. Each element in this vector represents the tuning parameter for the regression function related to the $i-th$ hidden state. Instead of having a global tuning parameter, which will equally penalize the coefficients of the regressions for each transition, we will introduce a specific tuning parameter for the regression related to each hidden state. This specific tuning parameter means that the regression function for each hidden state will be individually penalized considering its own particular tuning parameter. This will ensure greater tuning precision, given the fact that each transition has different transition coefficients. We will call this second proposal, Individual LASSO. The second version of the objective function with transition specific $\lambda$ tuning parameter will be given by:

$$\left[\sum_{i=1}^{K}\sum_{j=1}^{K}\sum_{(t:t\geq 2,(s_{t-1},s_t)=(i,j))} \mathbf{X}_t\boldsymbol{\beta}_{ij} - log\sum_{l=1}^{K} exp(\mathbf{X}_t\boldsymbol{\beta}_{il})\right] - \sum_{i=1}^{K}(\lambda_i\sum_{j=1}^{K}\sum_{m=2}^{D}|\beta_{ijm}|), \tag{3.13}$$

where $\lambda_1 \geq 1 \ldots \lambda_K \geq 0$.

Essentially, the constraint introduced as the sum of the absolute value of the transition coefficients in Equation (3.12) will induce shrinkage of the those coefficients whose real value is in the vicinity of zero.

The $\lambda$ hyper-parameter controls the intensity of the penalization and may be selected during the estimation procedure using cross-validation while targeting the optimization of a specific metric. In most cases, the metric which is used to select $\lambda$ is the classification error. However, in the context of NHMMs this metric cannot be used because of the fact that the

sequence of hidden states is not observable and we assume a continuous observable variable. Therefore, the metric which will be used to select the optimum value of $\lambda$ will be the mean squared prediction error. An in-depth explanation of the metrics used to evaluate the proposed methods will be presented in Section 4.1.

In many cases the LASSO will perform automatic variable selection by setting some of the least relevant coefficients exactly to zero. This characteristic of LASSO makes it very appealing in the context of Statistical Learning, as it naturally produces sparse models. A detailed description of the inner workings of the LASSO regularization can be found in several works throughout the literature. These works include, and are not limited to Tibshirani (1996), Friedman, Hastie and Tibshirani (2010) and Städler and Mukherjee (2013).

Given that the regression coefficients related to the $i$-th hidden state do not impact the coefficients related to the other hidden states directly and to gain computational efficiency and more simplicity of estimation, we maximize the objective functions separately for each state $i$, that is, for $i = 1, 2, \ldots, K$, we maximize

$$\left[ \sum_{j=1}^{K} \sum_{(t:t\geq 2,(s_{t-1},s_t)=(i,j))} \mathbf{X}_t \boldsymbol{\beta}_{ij} - log \sum_{l=1}^{K} exp(\mathbf{X}_t \boldsymbol{\beta}_{il}) \right] - \lambda \left( \sum_{j=1}^{K} \sum_{m=2}^{D} |\beta_{ijm}| \right) \tag{3.14}$$

in the Global LASSO, and

$$\left[ \sum_{j=1}^{K} \sum_{(t:t\geq 2,(s_{t-1},s_t)=(i,j))} \mathbf{X}_t \boldsymbol{\beta}_{ij} - log \sum_{l=1}^{K} exp(\mathbf{X}_t \boldsymbol{\beta}_{il}) \right] - \lambda_i \sum_{j=1}^{K} \sum_{m=2}^{D} |\beta_{ijm}| \tag{3.15}$$

in the Individual LASSO. We also tried to estimate the regression coefficients jointly (considering all the hidden states in the objective functions but the results show less favorable performance when compared to separately estimating the coefficients).

After describing the procedures to obtain the maximum likelihood estimators for each of the sets of parameters of interest, we now move on to the description of the Stochastic EM algorithm. It is important to understand that if the $\mathbf{S}$ sequence was observable such as $\mathbf{Y}$, then the maximum likelihood estimates for $\mathbf{p}$, $\boldsymbol{\theta}$ and $\boldsymbol{\beta}_{ij}$ would be obtained using the observed data and the equations described in the previous paragraphs. However, this is not the case, so we apply the iterative stochastic EM algorithm which predicts the $\mathbf{s}$ sequence, and then calculates the value of the parameters which maximize the likelihood function. According to results shown in Sabillón and Zuanetti (2023), this stochastic EM algorithm (SEM) presents faster convergence in several scenarios as well as facilitating the calculations necessary for its implementation. The SEM algorithm combined with LASSO to select and estimate a NHMM is described as:

1. Assign arbitrary initial values to $\mathbf{S}$ and $\boldsymbol{\beta}$, and fix a value for $\lambda$ or $\boldsymbol{\lambda}$;
2. Using the observed $\mathbf{y}$ sequence and the values assigned to the $\mathbf{s}$ sequence, estimate:
    - $p_j$, $\mu_j$, $\sigma_j$ using the maximum likelihood estimators for $j = 1, \ldots, K$, and

- $\boldsymbol{\beta}_{ij}$ applying numerical methods, such as the BFGS algorithm, on Expression (3.12) or (3.13);

3. Based on the updated values of the parameters, update the values of the $\mathbf{s}$ sequence using Bayes theorem:

$$\begin{aligned} P(S_t = j \mid S_{t-1} = i, Y_t = y_t, \mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \mathbf{p}, \boldsymbol{\beta}) &= \frac{P(S_t = j, Y_t = y_t \mid S_{t-1} = i, \mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \mathbf{p}, \boldsymbol{\beta})}{P(Y_t = y_t \mid S_{t-1} = i, \mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \mathbf{p}, \boldsymbol{\beta})} \\ &= \frac{P(S_t = j, Y_t = y_t \mid S_{t-1} = i, \mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \mathbf{p}, \boldsymbol{\beta})}{\sum_{l=1}^{K} P(S_t = l, Y_t = y_t \mid S_{t-1} = i, \mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \mathbf{p}, \boldsymbol{\beta})} \\ &= \frac{P(S_t = j \mid S_{t-1} = i, \mathbf{p}, \boldsymbol{\beta}, \mathbf{X}) P(Y_t = y_t \mid S_t = j, \boldsymbol{\mu}, \boldsymbol{\sigma})}{\sum_{l=1}^{K} P(S_t = l \mid S_{t-1} = i, \mathbf{p}, \boldsymbol{\beta}, \mathbf{X}) P(Y_t = y_t \mid S_t = l, \boldsymbol{\mu}, \boldsymbol{\sigma})} \\ &= \frac{a_{ijt} \frac{1}{\sqrt{2\pi\sigma_j^2}} exp\left\{ -\frac{(y_t - \mu_j)^2}{2\sigma_j^2} \right\}}{\sum_{l=1}^{K} a_{ilt} \frac{1}{\sqrt{2\pi\sigma_l^2}} exp\left\{ -\frac{(y_t - \mu_l)^2}{2\sigma_l^2} \right\}}, \end{aligned} \tag{3.16}$$

that is, from

$$S_t \mid S_{t-1} = i, Y_t = y_t, \mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \mathbf{p}, \boldsymbol{\beta} \sim Discrete(p_{i1t}, p_{i2t}, \ldots, p_{iKt}); \tag{3.17}$$

where $p_{ijt} = P(S_t = j \mid S_{t-1} = i, Y_t = y_t, \mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \mathbf{p}, \boldsymbol{\beta})$. For $t = 1$, we only need to change $a_{ijt}$ by $p_j$ in the conditional probability of $S_t$;

4. Repeat steps 2 and 3 until convergence is attained. The most commonly used convergence criterion is the difference between the log-likelihood between iterations $q$ and $q+1$. When this difference becomes less than a previously established tolerance value, then the algorithm ends its iterations.

In the context of NHMMs, a notable disadvantage of EM-type algorithms is that it is necessary for all possible transitions between hidden states to be observed at least once and, in the situation assumed here, to have at least two observations allocated in each hidden state to be able to calculate the variance related to that hidden state. If a specific transition was not be observed, then the algorithm encounters a logical error, and it would not continue iterating. To avoid this problem, the $\mathbf{s}$ sequence generated at every iteration is tested to see if all possible transitions have occurred. If all possible transitions have not occurred, then 2 points are randomly chosen along the length of the hidden state sequence, and the transitions which were not observed are assigned to these randomly chosen points. The stochastic EM algorithm described in this section was implemented using R, developed by R Core Team (2013).

Step number 2 of the previous algorithm mentions the use of BFGS as a numerical optimization method. The BFGS algorithm is a second order optimization algorithm. Its name is an acronym which includes the names of its co-discoverers: Broyden, Fletcher, Goldfarb, and Shanno. It is mostly intended for convex optimization problems with a single optima and

belongs to a class of Quasi-Newtonian methods. The Hessian matrix does not have to be precisely calculated for the application of this type of methods, and rather an approximation the inverse of it is used. This means that the Hessian matrix does not need to be available in each iteration of the algorithm. More detailed information on the BFGS algorithm can be found in Yuan (1991) and Dai (2002). The importance of selected numerical optimization method is analyzed in detail in Section 4.3, given that it is a crucial part of the proper functioning of the proposed algorithms.

CHAPTER

# 4

# SIMULATION STUDIES

In order to test the performance of the proposed methods, we have designed an in-depth simulation study. We will subject our proposals to situations which stress the performance of the algorithms, to be able to understand the behavior of these methods under controlled conditions. This chapter is organized as follows: Section 4.1 presents a description of the metrics which will be used to evaluate the proposed methods. Section 4.2 presents some general conditions that apply across all scenarios presented in the study. Additionally, this section explains the reasoning behind our choice of link function. Sections 4.3 and 4.4 describe simulation scenarios which evaluate the shrinkage and predictive performance of the proposals, respectively. Finally, Section 4.5 offers concluding remarks regarding the simulation study.

## 4.1 Evaluation metrics

To objectively evaluate the performance of the proposed methods, we have selected some common metrics to have a frame of reference for comparison. For the scenarios presented in Section 4.3, the primary purpose is to illustrate the impact of the selected numerical optimization method on the shrinkage efficiency of the proposed algorithms. Section 4.3 of this document originates from initial unsuccessful experiences related to shrinkage, and subsequent discoveries related to the numerical optimization method being utilized which lead to the attainment of successful shrinkage with good results. Other objectives of the section also include showing that transition coefficient shrinkage is achieved, and is working properly. With this purpose in mind, the scenarios in the aforementioned section will only test and analyze metrics related to shrinkage of the transition coefficients of the model, focusing on some specific factors such as initial value of the transition coefficients, and value of $D$, the amount of coefficients in every regression equation. Synthetic data sets will not be separated into training, validation and test data sets because we are only interested in monitoring how shrinkage occurs, rather than actual selection of the best model via cross-validation.

In the case of scenarios presented in Section 4.4, the main interests of the study are to determine the predictive power of the proposed algorithms. We also analyze the estimation of the parameters of distribution of the observable random variables, understand the algorithms' accuracy in classifying observations into their corresponding non-observable states, and establish how shrinkage of the transition coefficients occurs under controlled conditions. For such purposes, all data sets throughout the different scenarios in Section 4.4 will be separated into training, validation and test data sets. Sometimes these data sets are also called training set, out-of-sample set and out-of-time set, respectively. The chosen percentages are 80% of the data set as training data set, 15% of the data set as validation data set and 5% of the data set as testing data set. As we have time series data sets, the testing data corresponds to the 5% final observations. Tuning of the LASSO penalization parameter will be performed on the validation data set. Predictive performance will be evaluated using the test data set.

The first set of metrics to be evaluated are metrics regarding the parameters of the observable random variables. For these parameters, we calculate the mean squared error, bias, standard deviation, and confidence interval. To define these metrics, let $\theta$ be any parameter of interest, and $\hat{\theta}$ the estimator for such parameter. We have that the mean square error of $\hat{\theta}$ is defined in Equation (4.1) as

$$\text{MSE}(\hat{\theta}) = \left[\text{Bias}(\hat{\theta})\right]^2 + \text{Var}(\hat{\theta}). \tag{4.1}$$

The bias of $\hat{\theta}$ is defined in Equation (4.2) as

$$\text{Bias}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta \tag{4.2}$$

and the standard deviation of $\hat{\theta}$ is estimated in Equation (4.3) as

$$\text{SD}(\hat{\theta}) = \sqrt{\frac{\sum_{r=1}^{R}(\hat{\theta}_r - \bar{\theta})^2}{R}}, \tag{4.3}$$

where $\bar{\theta}$ is the observed average of $\hat{\theta}$s and $R$ is the number of replications.

The next set of metrics, and possibly the set of greatest interest, are metrics which evaluate the quality of the predictions on the test data set. For such purpose we will calculate the mean squared prediction error on the test data set. The best model fit using the Global or Individual LASSO will be that which minimizes this mean squared prediction error. Let $Y_t$ be the real value of the response variable and $\hat{Y}_t$ the value predicted by the model at position $t$ of the chain, and $n_{test}$ the length of the test data set. The mean squared prediction error is given by:

$$\text{MSPE} = \frac{\sum_{t=1}^{n_{test}}(Y_t - \hat{Y}_t)^2}{n_{test}}. \tag{4.4}$$

As shown in Sabillón and Zuanetti (2023), in a NHMM, the predicted value may be calculated as $\hat{y}_t = \hat{E}(Y_t|S_{t-1} = \hat{s}_{t-1}) = \sum_{j=1}^{K} \hat{a}_{\hat{s}_{t-1}jt}\hat{\mu}_j$ and $\hat{s}_t = \arg_j \max \hat{a}_{\hat{s}_{t-1}jt}$. This predicted $\hat{s}_t$ is used to calculate $\hat{a}_{\hat{s}_t jt}$ for the observation $t+1$ and so on.

In order to have a comparative frame for the performance of the proposed algorithms, all simulations in Section 4.4 include fitting and evaluation of two other models for predictive comparison. This first of these models is the Auto-regressive Integrated Moving Average (ARIMA) model. This model is fit using the `auto.arima` function from the R Core Team (2013) programming language. The function conducts a search over all possible models within the set of provided order constraints and returns the best ARIMA model according to either AIC, AICc or BIC value. It allows external regressors and chooses the order of first-differencing and seasonal-differencing based on specific tests. More details on the inner workings and structure for this model can be found in Hyndman and Khandakar (2008). Specific details on how the selection of seasonal differences is carried out for this model can be found in Wang, Smith-Miles and Hyndman (2006). From this point on, we will refer to this model as ARIMA and we consider it is one of the most flexible and complete parametric models for time series prediction. The second model selected for predictive comparison is the Penalized Linear Regression with LASSO. To fit this model, we use the `cv.glmnet` function from the `glmnet` package. This package was created by Friedman, Tibshirani and Hastie (2010). We set the parameters of this function to ensure that LASSO is the applied penalization method. Details on the inner workings of LASSO with linear regression can be found in Friedman, Tibshirani and Hastie (2010). From this point on, we refer to this model as Penalized Linear Regression. Although Penalized Linear Regression is not an appropriate method to apply with time series data since it does not include correlation information among the observations, we have selected it due to its widely known predictive capability.

For both of these comparative methods, we use the training and validation database for estimation of the models and use the test database to calculate the prediction measures. In this sense, these methods have an advantage over NHMM because they consider a larger database and more recent information in the estimation.

A third set of metrics which will be calculated are metrics regarding the accuracy in predicting the non-observable sequence of states. However, it is important to understand that in a real data setting, evaluating this aspect of the model is not viable, because the labels for the non-observable sequence are unknown (hence the name hidden Markov model). Nonetheless, because we are in a simulated setting, we will take advantage of the fact that the non-observable sequence is known to the researcher, and we will calculate the accuracy of the proposed methods when trying to predict the non-observable sequence on the test data set. To define predictive accuracy, let $S_t$ be the real value of the non-observable state and $\hat{S}_t$ the predicted value, at position $t$ of the non-observable chain. The algorithm to calculate the predictive accuracy for the non-observable test sequence is as follows:

1. Initialize $hitcounter = 0$, a variable to accumulate the quantity of hits when predicting the non-observable test sequence;

2. For every position in the test sequence, generate a predicted non-observable state $\hat{s}_t$ using the estimated parameters, as $\hat{s}_t = \arg_j \max \hat{a}_{\hat{s}_{t-1}jt}$;

3. Verify if $s_t$ is equal to $\hat{s}_t$. If $s_t$ is equal to $\hat{s}_t$ then add 1 to $hitcounter$, if not, take no action;

4. Repeat steps 2 and 3, for all positions in the non-observable test sequence;

5. The predictive accuracy will be calculated by dividing $hitcounter/n_{test}$, and will be reported as a decimal ranging from 0 (No hits) to 1 (100% accuracy).

As a fourth set of metrics we will be observing the LASSO's ability to set the transitions coefficients exactly to zero. For such purposes we will adopt a binary approach to measuring the performance of the LASSO. As a cut-off value we have set $cut-off = 0.05$. Any coefficient whose estimated absolute value is greater than 0.05 will be considered as not being shrunk to 0 by the LASSO, and any coefficient whose estimated absolute value is less than or equal to 0.05 will be considered as shrunk to 0 by the LASSO. After this, we define the following classification indicators:

- True Positive (TP): A coefficient whose real value is zero, and whose estimated value was shrunken to zero;
- True Negative (TN): A coefficient whose real value is different than zero, and whose estimated value was not shrunken to zero;
- False Positive (FP): A coefficient whose real value is different than zero, and whose estimated value was shrunken to zero;
- False Negative (FN): A coefficient whose real value is zero, and whose estimated value was not shrunken to zero.

For each of the indicators mentioned previously, we will test the estimated coefficients, and count the occurrence of each indicator. Using the classification indicators we have defined and calculated, we now construct classical classification metrics to evaluate the performance of the LASSO when shrinking coefficients to zero. These metrics are specificity, sensitivity and accuracy. There formulas are given by:

$$\text{Specificity} = \frac{TN}{TN+FP}, \tag{4.5}$$

$$\text{Sensitivity} = \frac{TP}{TP+FN}, \tag{4.6}$$

$$\text{Accuracy} = \frac{TN+TP}{TP+FP+TN+FN}. \tag{4.7}$$

Finally, we will collect some metrics regarding execution time for each replication, and total execution time for each simulation scenario.

## 4.2 General conditions for simulations

As previously mentioned, this simulation study consists of 2 distinct types of simulations. The first type of simulation is solely designed to test effects of the selected numerical optimization on shrinkage performance of the proposed algorithms under different conditions. The second type of simulation will verify all aspects of the proposed algorithms, by varying different factors and parameters which may influence the performance of our proposals. As we present and explore each scenario within these two simulation designs, the specific conditions and description of each scenario will be listed and described along with all the observed results of in scenario. Besides that, we now list some general conditions which are applied throughout the simulation scenarios:

- The number of non-observable states was fixed at $K = 2, 3$;
- Six chain lengths are chosen for simulations, $T = 400, 600, 800$ when $K = 2$ and $T = 800, 1000, 1200$ when $K = 3$;
- With no loss of generality, and with the purpose of simplifying programming and comparison among the different algorithms, all covariates are generated as $X_{td} \sim N(0,1)$, for $d = 2, \ldots, D$; $t = 1, \ldots, T$. $X_t 1$ is the covariate which corresponds to the intercept in all regressions;
- For the simulations presented in Section 4.3, $R = 30$ replications of the data set are randomly generated for each value in the $\lambda$ vector;
- For the simulations presented in Section 4.4, $R = 50$ replications of the data set are randomly generated for $K = 2$. $R = 30$ replications of the data set are generated when $K = 3$;
- Two values of $D$, the number of covariates for each non-observable state, are used depending on the value of $K$. For $K = 2$ the number of covariates is set to $D = 10, 20$. For $K = 3$, the number of covariates is set to $D = 8, 15$;
- The choice of values for the transition coefficients $\boldsymbol{\beta}$ and the parameters of the distribution of the observable random variables, $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ will be discussed in detail in the following sections for each simulation scenario.

Notice that for $K$ non-observable states, the model presents $K^2$ possible transitions. This also indicates that the transition coefficient matrix will have $K^2$ vector elements. Each one of these vectors will contain $D$ covariates. If we contextualize for a model which has $K = 3$ and

$D = 8$, when $T = 800$ it means that the 800 observations have to be shared among 9 different sets of coefficients, belonging to 9 distinct transitions. If such is the case, then each transition will have approximately 88 observations allocated to it, under the assumption that observations are uniformly distributed among all possible transitions, and will be used to estimate 8 regression coefficients.

However, uniformly distributed observations among all transitions is a very strong assumption, which overlooks the complex dynamics that different sets of values for the transition coefficients create. After much testing before the simulations, we can at the very least assume that each transition will have a reasonable number of observations allocated to it in order to perform estimation and prediction. If a transition with no allocated observations occurs, then the mechanism described at the end of Section 3.3 to force the occurrence of all possible transitions is triggered to allow for the proposed algorithms to continue their execution. The triggering of the described mechanism is a step we would prefer to avoid.

### 4.2.1 Choice of link function

As mentioned in Section 2.1.3, the driving factor behind the dynamics of the non-observable Markov chain is the non-homogeneous transition matrix. The core of this process lies in the link function which serves the purpose of generating transition probabilities between non-observable states at different moments of time, as a function of the transition coefficients $\boldsymbol{\beta}$ and the covariate vector at time $t$, $\mathbf{X}_t$. According to Gao and Pavel (2018), in the realm of multi-class logistic regression, the Softmax function maps a vector of covariates to a probability distribution pertaining to the different classes. Wolfe *et al.* (2017) reinforces this idea by stating that in a Softmax regression, a classification is made between multiple classes, using the probabilities generated by the Softmax function given by:

$$a_{ijt} = h(\mathbf{X}_t \boldsymbol{\beta}_{ij}) = \frac{exp(\mathbf{X}_t \boldsymbol{\beta}_{ij})}{\sum_{l=1}^{K} exp(\mathbf{X}_t \boldsymbol{\beta}_{il})}. \tag{4.8}$$

However, in the context of NHMMs and considering the estimation methods that will be applied, the Softmax function as a link function introduces a problem of non-identifiability of the transition parameters.

Souza (2010) defines a family of densities as being identifiable if the values of different parameters determine different members of a family of densities. In the case of NHMMs, different and infinite sets of $\boldsymbol{\beta}$ could calculate the same value for the transition probabilities through the Softmax function. Due to this fact, the choice of link function for all simulations and applications of the NHMM should be the multinomial-logistic function, also known as mlogit. Essentially, the mlogit link function is a special case of the Softmax function in which transition coefficients

related to the first non-observable state are set to 0. The mlogit link function is given by:

$$a_{ijt} = h(\mathbf{X}_t\boldsymbol{\beta}_{ij}) = \begin{cases} \frac{1}{1+\sum_{l=2}^{K} exp(\mathbf{X}_t\boldsymbol{\beta}_{il})}, \text{if } j = 1 \\ \frac{exp(\mathbf{X}_t\boldsymbol{\beta}_{ij})}{1+\sum_{l=2}^{K} exp(\mathbf{X}_t\boldsymbol{\beta}_{il})}, \text{ if } j > 1, \end{cases} \tag{4.9}$$

for $i = 1, \ldots, K$.

Although the model was presented using the Softmax link function, we adopt the mlogit link function in practice, which is a special case of the Softmax function.

### 4.2.2 Data simulation procedure

An important part of any simulation study, is the careful planning and design of the mechanism which simulates data. This will allow the researcher the capability to simulate data under controlled conditions to test specific characteristics and tolerances of any estimation procedure which is being investigated or proposed. Artificial data simulation must mimic the behavior of the natural mechanism which has the structure of the studied model, and at the same time allow for the researcher the flexibility of controlling characteristics of the synthetic data.

We have designed a procedure which simulates artificial data from a NHMM. The procedure is described in Algorithm 1.

**Algorithm 1** – Data simulation algorithm for the NHMM

1: Set desired values for $\boldsymbol{\beta}$, $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, $\mathbf{p}$, $K$, $D$, and $T$.
2: Simulate $T$ values for $D$ covariates where $X_{ti} \sim N(0,1)$ for $i = 1, \ldots, D$ and $t = 1, \ldots, T$.
3: Initialize vectors $\mathbf{S}$ and $\mathbf{Y}$, representing the non-observable chain and the corresponding observable values.
4: Simulate a value between 1 and $K$ from a Discrete distribution with probabilities given by $\mathbf{p}$, for the first position in the non-observable chain, $s_1$.
5: Simulate a value for $Y_1$ where $Y_1 \sim N(\mu_{s_1}, \sigma_{s_1})$.
6: **for** $t \leftarrow 2$ to $T$ **do**
7: Calculate the probabilities of transitioning from non-observable state $S_{t-1}$ to any state, 1 to $K$, at position $S_t$ using Equation (4.9).
8: Simulate a value between 1 and $K$ for $s_t$ using the probabilities calculated in Step 6.
9: Simulate a value for $Y_t$ where $Y_t \sim N(\mu_{s_t}, \sigma_{s_t})$.
10: **end for**

The data simulation procedure described in Algorithm 1 is used in all simulations scenarios presented in this thesis.

## 4.3 Simulation scenarios: shrinkage performance

The first part of the simulation study focuses exclusively on understanding how shrinkage of the transition coefficients occurs. The main motivation for having a separate section dedicated

to exploring shrinkage performance is due to experiences in earlier unsuccessful simulation trials and the impact of the numerical algorithm used for optimization. In the aforementioned first round of simulations, we explored several aspects of the proposals' performance but in spite of this, we did not manage to obtain favorable results related to shrinkage metrics. With the intent of providing an additional frame of comparison for the shrinkage performance of the proposals, the results and discussion related to that first round of simulations can be found in Appendix A.

As we state in Section 3.3, estimation of regression coefficients is carried out using numerical optimization methods. In the first round of simulation trials, we only tested the Nelder-Mead algorithm for doing this. After much trial-and-error and testing different parameter configurations, we came upon the idea of of testing different numerical optimization methods to perform estimation of the $\boldsymbol{\beta}_{ij}$ transition coefficients. This led to the discovery of great differences in performance when choosing other numerical methods such as the BFGS optimization method.

The scenarios presented in this section focus on capturing shrinkage metrics. For such purposes, we generate 30 data sets for each value of $\lambda$. This is done to be able to calculate mean values of the shrinkage metrics, as well as to be able to understand the variability in shrinkage performance which will occur through out the set of values of $\lambda$ being tested.

The following sections which describe simulation Scenarios A, B and C illustrate the impact of the selected optimization method on shrinkage of the transition coefficients. We also show the effect of different initial values of the transition coefficients on the final estimates, and also display how shrinkage behaves for different values of $D$, the amount of covariates present in the regression equations which calculate the probabilities of transitioning between non-observable states.

### 4.3.1 Scenario A

Scenario A will compare the shrinkage performance of the BFGS and Nelder-Mead methods when used as optimization procedures with the Global LASSO. This scenario involves the following parameter configuration:

- The selected value for the number of non-observable states is $K = 2$;
- The selected amount of covariates is $D = 20$;
- The selected chain length is $T = 600$;
- 40 values for $\lambda$ are used ranging from 0 to 20, in incremental steps with a value of 0.5;
- The parameters for the probability distribution of the observable random variables are set to $\mu_1 = 60$, $\sigma_1 = 2$, and $\mu_2 = 70$, $\sigma_2 = 3$;
- Initial values for the transition coefficients are randomly simulated from $N(0,1)$ distribution.

The matrix of transition coefficients selected for Scenario A is shown in Equation (4.10) as

$$\boldsymbol{\beta} = \begin{bmatrix} \begin{pmatrix} \beta_{111} = 0.0 \\ \beta_{112} = 0.0 \\ \beta_{113} = 0.0 \end{pmatrix} & \begin{pmatrix} \beta_{121} = -1.5 \\ \beta_{122} = -1.5 \\ \beta_{123} = -2.6 \end{pmatrix} \\ \\ \begin{pmatrix} \beta_{211} = 0.0 \\ \beta_{212} = 0.0 \\ \beta_{213} = 0.0 \end{pmatrix} & \begin{pmatrix} \beta_{221} = -2.0 \\ \beta_{222} = 2.6 \\ \beta_{223} = 1.4 \end{pmatrix} \end{bmatrix}. \tag{4.10}$$

Therefore, only the first three transition coefficients (including the intercept) are relevant for both hidden states and the remaining 17 transition coefficients are set to zero. Due to space limitations and to improve readability, we only show 3 coefficients in each $\boldsymbol{\beta}_{ij}$, however all $\boldsymbol{\beta}_{ij}$ contain 20 coefficients.

It is important to remember that, given the structure of the NHMM and using de mlogit link function, for any value of $K > 0$ and $D > 0$, we will have $D \times K \times (K-1)$ transition parameters being estimated. This structure of the transition matrix implies that there are 6 coefficients whose real value is non-zero and 34 coefficients whose real value is 0.

Figures 1 (left column plots) and 2 (right column plots) show shrinkage results for the Global LASSO when using BFGS and Nelder-Mead numerical methods, respectively. The blue line represents the mean value for the metric being plotted. The gray area that is enclosed by dashed red lines represents the calculated 95% CI for the values obtained along the 30 replications for each value of $\lambda$.

The plot for sensitivity in Figure 1 shows that the BFGS method yields values of sensitivity which considerably increase as $\lambda$ increases, attaining a value approximately greater than 95%. This increase occurs with a relatively small variability, implying that using the BFGS method imbues consistency into the Global LASSO. On the other hand, from Figure 2, we perceive that values for sensitivity obtained when using Nelder-Mead as the optimization method are erratic. Besides this, the mean value for this metric does not reach a value greater than 25%.

Regarding specificity, we perceive that it is essentially constant for the BFGS method. It will begin to display variability for the greatest tested values of $\lambda$. This is expected, because as values of $\lambda$ increase, penalization increases and therefore some transition coefficients which should not be shrunken past the threshold value will, in fact, be shrunken. However, performance for the Global LASSO using the BFGS method is excellent in this metric. When analyzing specificity while using the Nelder-Mead method in the Global LASSO, once again we come across considerable variability, even for smaller values of $\lambda$, showing the poor performance of the Global LASSO when the Nelder-Mead method is used as numerical optimization method.

In terms of accuracy, we can once again see the same tendency shown in previous metrics, where the BFGS method shows far superior performance than the Nelder-Mead method. This is

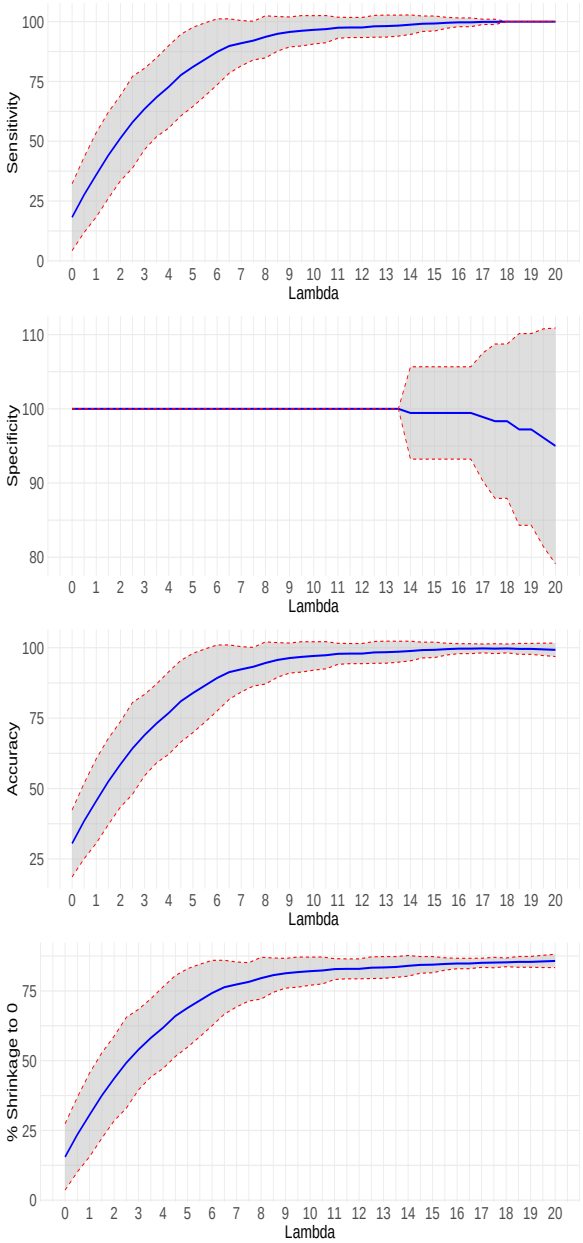



Figure 1 – BFGS optimization Metrics. For specificity, sometimes the upper credibility limit is greater than 100 because we use the Gaussian distribution to build it and it does not respect the maximum value of the metric.

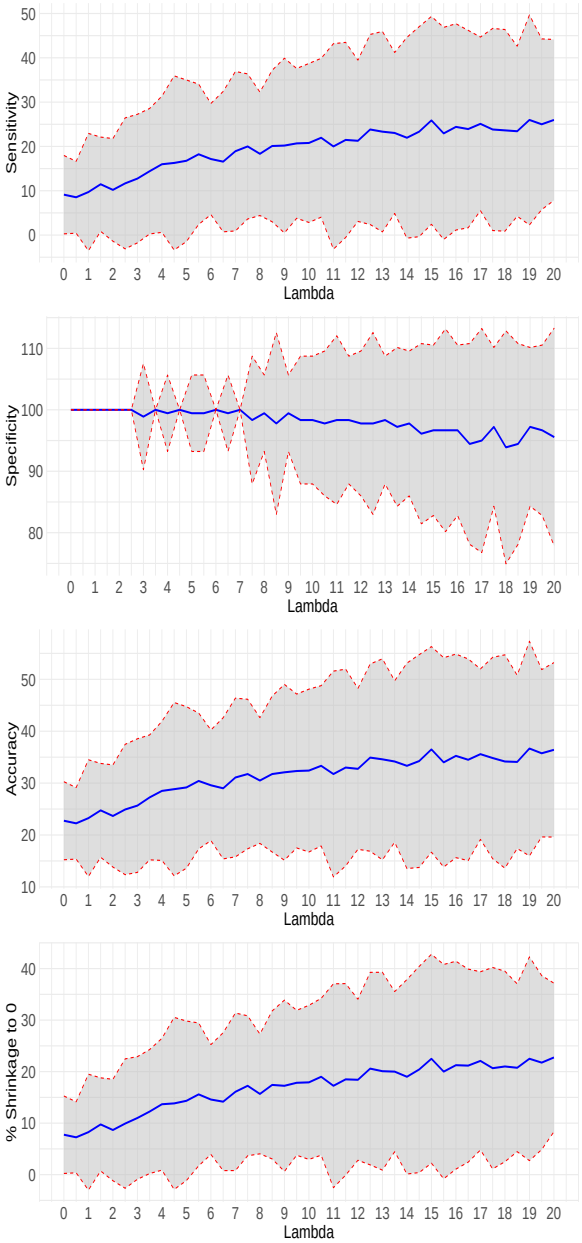
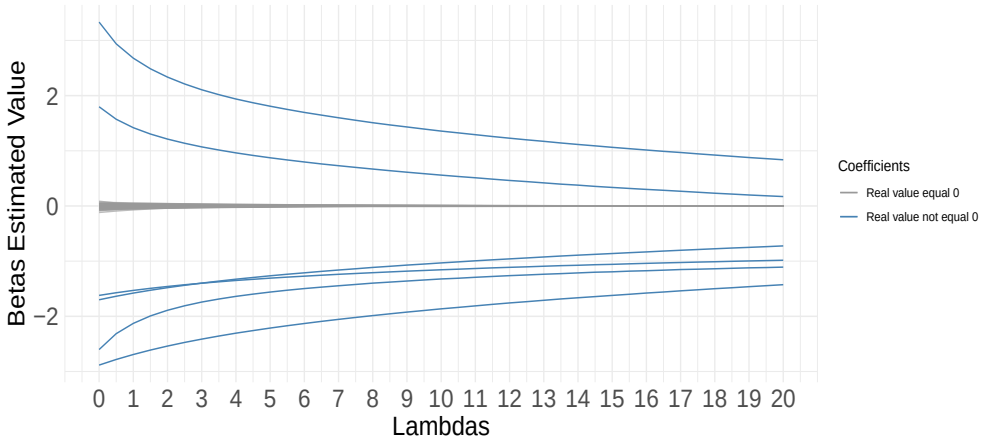



Figure 2 – Nelder-Mead optimization metrics. For specificity, sometimes the upper credibility limit is greater than 100 because we use the Gaussian distribution to build it and it does not respect the maximum value of the metric.

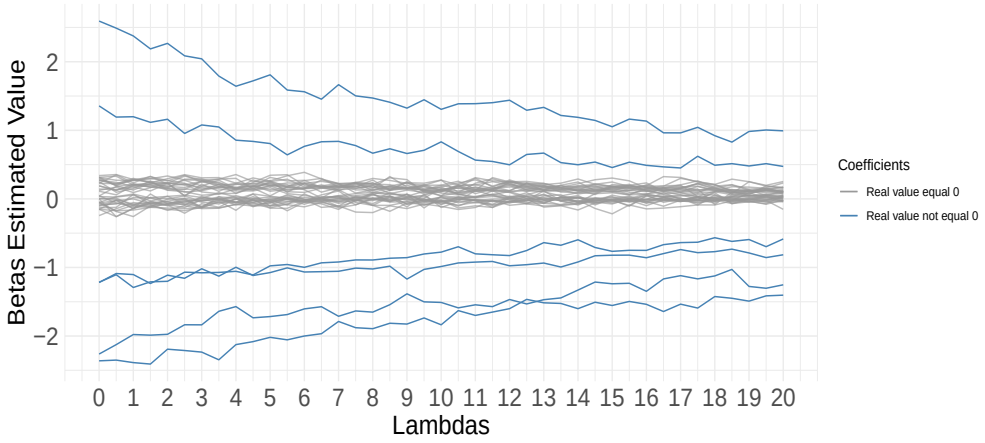
expected, because of the fact that accuracy is a function of specificity and sensitivity, and the BFGS performed better in both of the previous metrics.

To support the idea that the Global LASSO is actually effective in shrinking down coefficients, we calculate the general proportion of coefficients that were effectively shrunken to 0 (this metric is not presented in Section 4.1) without distinguishing if said coefficients have a real value of zero or non-zero. If our proposal is working properly then this metric should steadily

increase as the value of $\lambda$ increases. As can be perceived from the last (bottom) plot in Figure 1, Global LASSO with BFGS as the numerical optimization method attains a mean shrinkage rate of 87%. This indicates that, in fact, the Global LASSO achieves shrinkage of transition coefficients when using BFGS as the numerical optimization method. We can also perceive the considerable decrease in shrinkage when comparing the BFGS method to the Nelder-Mead method.




(a) BFGS optimization.




(b) Nelder-Mead optimization.

Figure 3 – Average estimated values for transition coefficients.

Finally, to further illustrate the effects of shrinkage, we plot the average estimated parameter values for the 30 replications for each $\lambda$, to show the progression of these estimated values as penalization increases. Figure 3 shows the plots for the average estimated values for both optimization methods. Although at first glance it seems that in the case of coefficients whose real value is 0, the Global LASSO with Nelder-Mead method produces estimates which are apparently being shrunk to 0. This is due to the estimation accuracy of the stochastic EM that manages to produce estimates which are close to 0. However, we emphasize that coefficient estimates must be shrunk beyond the established zero-threshold to be considered efficiently

shrunk to 0. This will only happen if penalization is being performed correctly by the Global LASSO.

When analyzing the Nelder-Mead plot in detail, we perceive that a considerable percentage of these coefficients whose real value is 0 are consistently estimated as being close to zero (hence the almost constant gray plotted lines), but not past the zero-threshold. The opposite is true when analyzing the plot for the Global LASSO using BFGS method. It seems that for very small values of $\lambda$, we can already perceive that most (if not all) of the coefficients whose real value is zero, have been shrunken past the zero threshold. These observations provide graphical confirmation of the satisfactory shrinkage performance when using the BFGS algorithm.

The observed results in this scenario lead us to believe that the most adequate numerical optimization method to be used with Global LASSO, among the two methods we are testing, is the BFGS method. We continue to verify this hypothesis in Sections 4.3.2 and 4.3.3.

### 4.3.2 Scenario B

The main objective of Scenario B is to analyze and understand the performance of the Nelder-Mead and BFGS methods as optimization methods for the Global LASSO, specifically in the presence of a greater amount of transition coefficients. For such purpose, the parameter configuration for this scenario has been established as follows:

- The selected value for the number of non-observable states is $K = 3$;
- The selected amount of covariates is $D = 15$;
- The selected chain length is $T = 1000$;
- 40 values for $\lambda$ are used ranging from 0 to 20, in incremental steps with a value of 0.5;
- The parameters for the probability distribution of the observable random variables are set to $\mu_1 = 60$, $\sigma_1 = 2$, $\mu_2 = 70$, $\sigma_2 = 1.5$ and $\mu_2 = 80$, $\sigma_2 = 3$;
- Initial values for the transition coefficients are randomly generated from $N(0,1)$ distribution.

An important observation that should always be considered when analyzing estimation of NHMMs is that the model uses the mlogit link function, a particular case of the Softmax function, therefore the coefficients related to the transitions from the first non-observable state are all set

to zero, in order to achieve parameter identifiability. Equation (4.11) given by

$$\boldsymbol{\beta} = \begin{bmatrix} \begin{pmatrix} \beta_{111} = 0.0 \\ \beta_{112} = 0.0 \\ \beta_{113} = 0.0 \end{pmatrix} & \begin{pmatrix} \beta_{121} = -1.5 \\ \beta_{122} = -1.5 \\ \beta_{123} = -2.6 \end{pmatrix} & \begin{pmatrix} \beta_{131} = -1.3 \\ \beta_{132} = -3.2 \\ \beta_{133} = -2.4 \end{pmatrix} \\ \begin{pmatrix} \beta_{211} = 0.0 \\ \beta_{212} = 0.0 \\ \beta_{213} = 0.0 \end{pmatrix} & \begin{pmatrix} \beta_{221} = -2.0 \\ \beta_{222} = 2.6 \\ \beta_{223} = 1.4 \end{pmatrix} & \begin{pmatrix} \beta_{231} = -1.3 \\ \beta_{232} = 1.7 \\ \beta_{233} = 1.3 \end{pmatrix} \\ \begin{pmatrix} \beta_{311} = 0.0 \\ \beta_{312} = 0.0 \\ \beta_{313} = 0.0 \end{pmatrix} & \begin{pmatrix} \beta_{321} = 2.4 \\ \beta_{322} = 2.1 \\ \beta_{323} = -1.5 \end{pmatrix} & \begin{pmatrix} \beta_{331} = -1.3 \\ \beta_{332} = -2.7 \\ \beta_{333} = -2.5 \end{pmatrix} \end{bmatrix} \tag{4.11}$$

shows transition coefficients used to generate data in Scenario B, considering $K = 3$. As mentioned in the previous scenario we only show 3 entries due to readability and space limitations.

Figures 4 (left column plots) and 5 (right column plots) show shrinkage results for the Global LASSO when using BFGS and Nelder-Mead numerical methods in Scenario B, respectively. The differences in shrinkage performance between the two tested numerical optimization methods becomes very evident in this scenario, most likely due to the increase in the number of transition coefficients. For this scenario, we have 18 coefficients whose real value is non-zero, and 72 coefficients whose real value is zero.

This first plot in the left column of Figure 4 shows that the BFGS method produces values of sensitivity which notably increase as the values of $\lambda$ increases, attaining a value of approximately greater than 98%. Another interesting fact is that it seems that the values of sensitivity continue to increase even for greater values of $\lambda$. The values also show remarkably low variability, showing consistent performance of the Global LASSO when using BFGS as optimization method. In contrast, from the first plot in Figure 5 we perceive that, in terms of sensitivity, very poor performance is obtained when using Nelder-Mead as the optimization method. The average is never greater than 12% and the observed variability implies that, in some cases, the attained sensitivity can be equal to 0.

In terms of specificity (second plot in Figure 4), we observe that for the BFGS method, specificity decreases steadily as values of $\lambda$ decrease. As mentioned in the previous scenario, this behavior is natural and expected for any penalization procedure given that more and more transition coefficients which should not be shrunken past the threshold value will be shrunken as $\lambda$ increases. Once again, Global LASSO with using the BFGS method shows very good performance. When analyzing specificity with the Nelder-Mead method in the Global LASSO (second plot of the right column), we perceive that there is little change in the specificity as the values of $\lambda$ increase. This is indicating that the Global LASSO loses shrinkage effectiveness when using the Nelder-Mead optimization method.

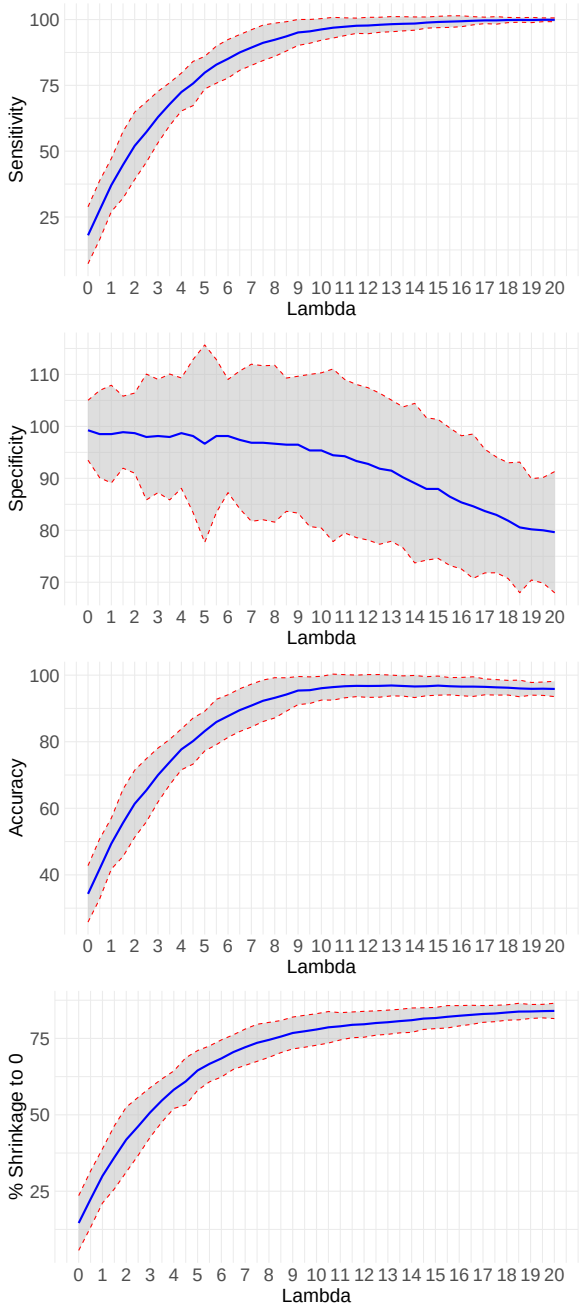



Figure 4 – BFGS optimization Metrics. For specificity, sometimes the upper credibility limit is greater than 100 because we use the Gaussian distribution to build it and it does not respect the maximum value of the metric.

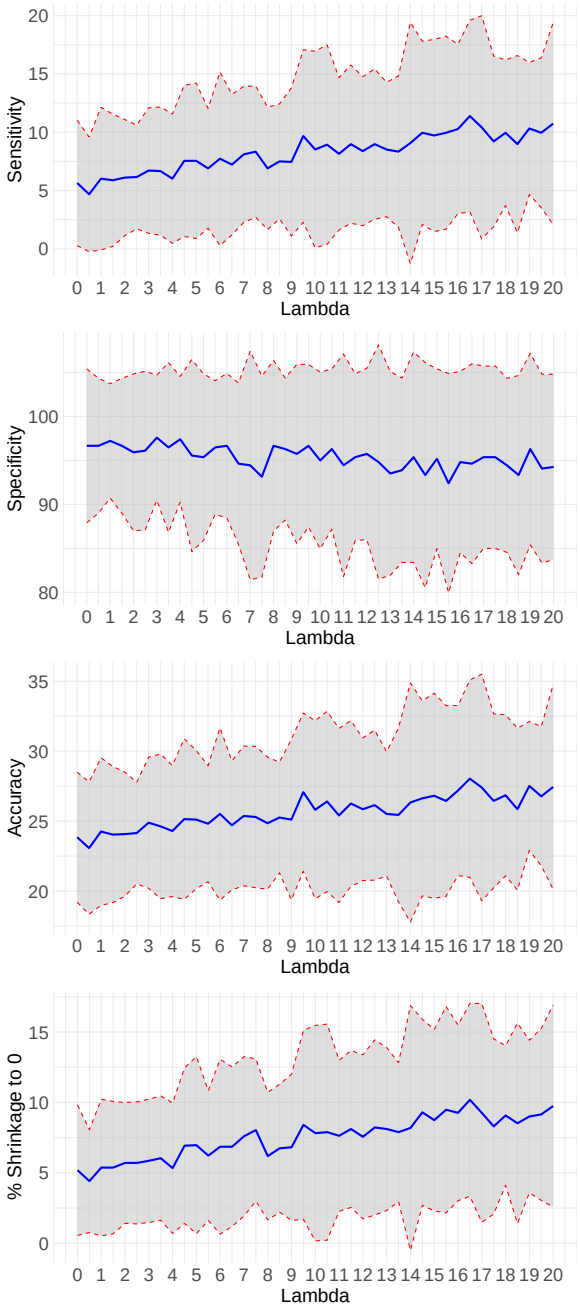
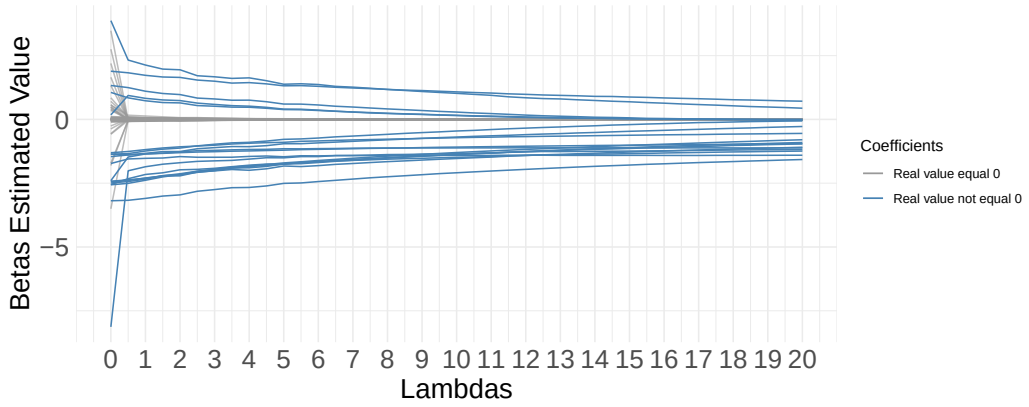


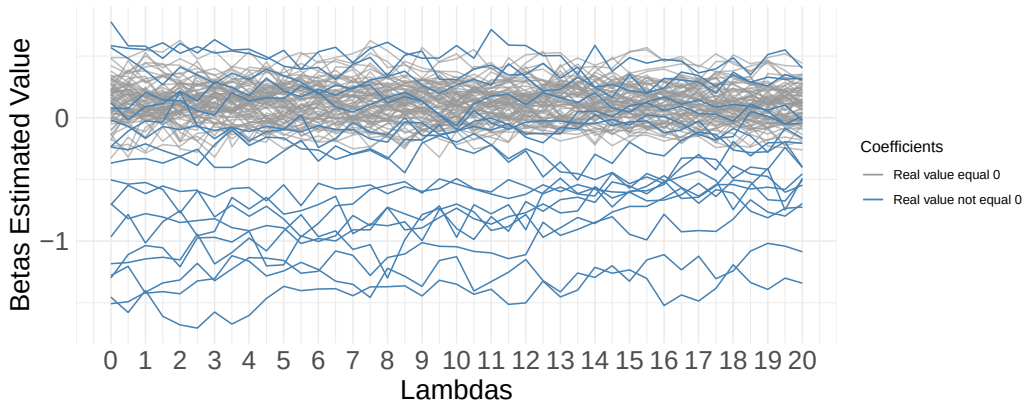

Figure 5 – Nelder-Mead optimization metrics. For specificity, sometimes the upper credibility limit is greater than 100 because we use the Gaussian distribution to build it and it does not respect the maximum value of the metric.

Regarding accuracy, we can once again see the same tendency shown in previous metrics, where the BFGS method shows far superior performance than the Nelder-Mead method. Regarding the total percentage of coefficients shrunk to zero, we perceive from the last (bottom) plot in Figure 4 that the Global LASSO with BFGS has a mean shrinkage rate which steadily increases until reaching an average value greater than 85%. From the plot, we can infer that

if greater values of $\lambda$ would have been included in the tested range, the mean values for this metric would have continued to increase. This is the expected behavior for this metric if the proposed algorithm is in functioning properly. This behavior shows that the Global LASSO achieves shrinkage of transition coefficients in this scenario when using BFGS as the numerical optimization method. In the case of the Nelder-Mead method, from the plot displaying total percentage of coefficient shrunk to zero, we can observe that shrinkage is not occurring, given that there is little or no change in the total percentage of coefficients being shrunken to zero as the values of $\lambda$ increase.




(a) BFGS optimization.




(b) Nelder-Mead optimization.

Figure 6 – Average estimated values for transition coefficients.

As was done in Scenario A, we illustrate the evolution of shrinkage of the estimated values of the coefficients as values of $\lambda$ increase in Figure 6. Once again, we notice a pattern in which the transition coefficients estimated using BFGS as the numerical optimization method are correctly estimated. Coefficients whose real value is zero are shrunken down properly towards zero, even for smaller values of $\lambda$, while coefficients whose real value is non-zero are shrunken down only when the greatest tested values of $\lambda$ are attained.

Meanwhile, performance regarding shrinkage when applying the Nelder-Mead optimiza-

tion with Global LASSO is notably poor. Estimated values of the transition coefficients seem to wander erratically throughout different regions of the plot. Estimates of the coefficients whose real value is zero seem to cluster around a value of 0, however it appears that they are not converging to zero, as should be the case for coefficients which are being effectively shrunken down to 0. Estimates for non-zero coefficients show no evidence of being slightly shrunken down, as should happen for non-zero coefficients.

The results we observe in this scenario reinforce the notion that the BFGS method is the most appropriate numerical optimization method to be used with Global LASSO among the two methods we are testing. Further evidence will be provided in Scenario 4.3.3.

### 4.3.3 Scenario C

In Scenario C, we continue to analyze the impact of the selected optimization method on transition coefficient shrinkage and estimation, as well as including variations in the initial values for the transition to determine if the Global LASSO is sensitive to different initial value of the transition parameters when different numerical optimization methods are used. Parameter configuration for this scenario is established as follows:

- The selected value for the number of non-observable states is $K = 2$;
- The selected amount of covariates is $D = 20$;
- The selected chain length is $T = 800$;
- 40 values for $\lambda$ are used ranging from 0 to 20, in incremental steps with a value of 0.5;
- The parameters for the probability distribution of the observable random variables are set to $\mu_1 = 60$, $\sigma_1 = 2$, and $\mu_2 = 70$, $\sigma_2 = 3$;
- Initial values for the transition coefficients are randomly generated from $N(10,3)$ distribution.

Transition coefficients used to simulate data in Scenario C are

$$\boldsymbol{\beta} = \begin{bmatrix} \begin{pmatrix} \beta_{111} = 0.0 \\ \beta_{112} = 0.0 \\ \beta_{113} = 0.0 \end{pmatrix} & \begin{pmatrix} \beta_{121} = -1.5 \\ \beta_{122} = -1.5 \\ \beta_{123} = -2.6 \end{pmatrix} \\ \\ \begin{pmatrix} \beta_{211} = 0.0 \\ \beta_{212} = 0.0 \\ \beta_{213} = 0.0 \end{pmatrix} & \begin{pmatrix} \beta_{221} = -2.0 \\ \beta_{222} = 2.6 \\ \beta_{223} = 1.4 \end{pmatrix} \end{bmatrix}. \tag{4.12}$$

Again, we only show 3 entries due to readability and space limitations.

Figure 7 shows results for shrinkage metrics when using the Global LASSO with BFGS as the selected numerical optimization method. It is displaying very similar tendencies to what the two previous scenarios show. Essentially, using the BFGS method yields much better performance of the LASSO in terms of all the analyzed metrics. Values of sensitivity considerably increase as $\lambda$ increases, attaining a maximum mean value apparently greater than 97% with relatively small variability. This indicates that the BFGS method improves the robustness and consistency into the Global LASSO. The opposite can be said, implied from Figure 8, where we observe that values for sensitivity when using Nelder-Mead as the optimization method, since they are unstable, with a mean value barely greater than 0% and displaying variability which is enough to assume that for several values of $\lambda$ the observed sensitivity was actually 0%.

Regarding specificity, the plots reveal that both optimization methods obtain a mean value for specificity equal to 100%. For the Nelder-Mead method, the values display variability along different values of $\lambda$.

Given that accuracy is a function of specificity and sensitivity, we observe that Nelder-Mead method obtains poor performance in terms of accuracy. This is due to the fact that it had poor performance in terms of sensitivity. In the case of the BFGS method, performance is superb across all analyzed metrics.

Finally, to further illustrate the effects of shrinkage, we plot the average estimated parameter values for the 30 replications for each $\lambda$, to show the progression of these estimated values as penalization increases in Figure 9. The figure very clearly displays the impact of the initial values on the estimated values of the transition coefficients. When observing the plot for BFGS optimization, we notice that different starting values have little, if any, impact on the quality of the estimated values for the transition coefficients. The coefficients whose real value is zero are correctly shrunken down to zero while the non-zero coefficients remain above the zero-threshold. On the other hand, when we analyze the estimated values for the transition coefficients with Nelder-Mead optimization, it seems that the values are predominantly concentrated around 10 for the coefficients whose real value is 0, and with greater dispersion for coefficients whose real value is non-zero. This shows us that, not only are coefficients whose real value is zero not being shrunken towards zero, but the coefficients whose real value is non-zero are not being properly estimated when the choice of numerical optimization is Nelder-Mead and the initial values are poor or very different from the actual value of the coefficients.

## 4.4 Simulation scenarios: inferential and predictive performance

The second part of the simulation studies focuses on evaluating the inferential and predictive capabilities of the proposed methods using the metrics presented in Section 4.1. This section presents 3 different simulation scenarios. The scenarios are designed to test specific

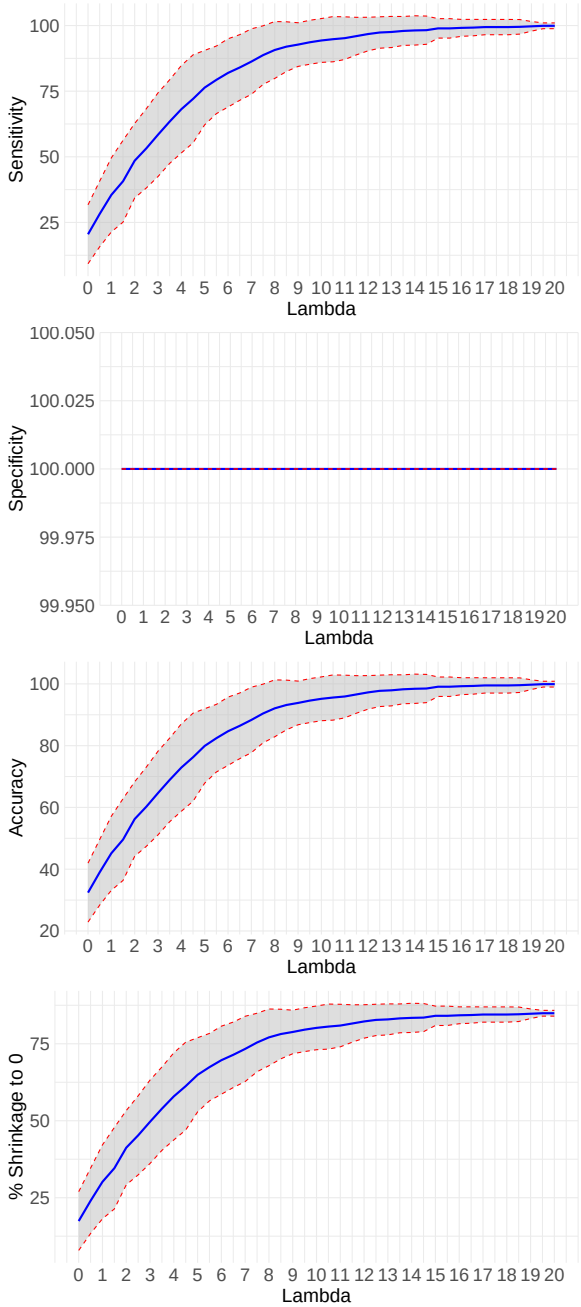



Figure 7 – BFGS optimization Metrics. For specificity, sometimes the upper credibility limit is greater than 100 because we use the Gaussian distribution to build it and it does not respect the maximum value of the metric.

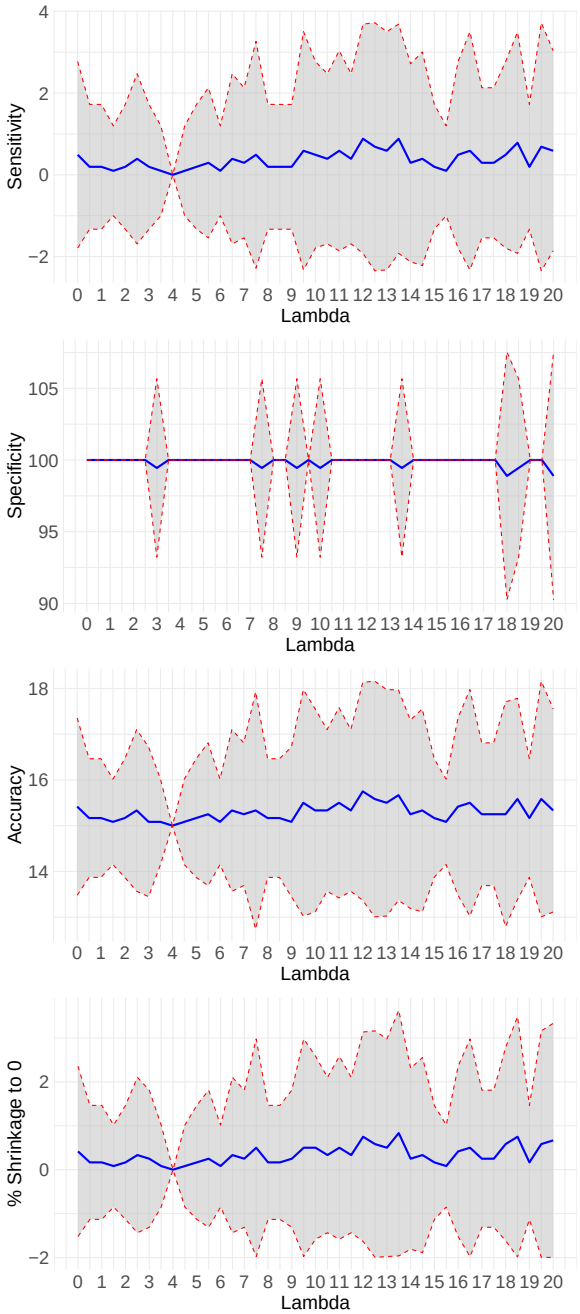
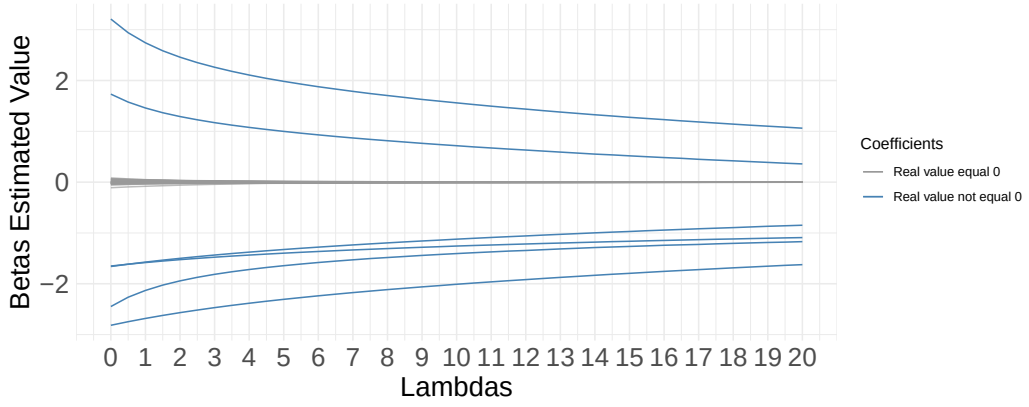
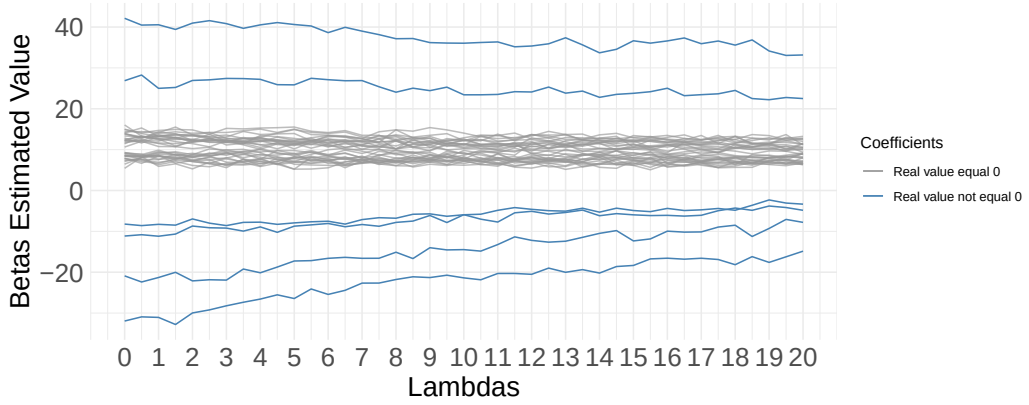



Figure 8 – Nelder-Mead optimization metrics. For specificity, sometimes the upper credibility limit is greater than 100 because we use the Gaussian distribution to build it and it does not respect the maximum value of the metric.

factors which the researcher believes may impact the performance of the proposed methods, as well as to analyze the general behavior of the methods under some previously established conditions. Specific details related to all the characteristics of each scenario are presented in their corresponding sections.




(a) BFGS optimization.




(b) Nelder-Mead optimization.

Figure 9 – Average estimated values for transition coefficients.

### 4.4.1 Scenario 1

The first scenario is designed as a general test of the proposed methods, to assess if the proposals are working correctly rather than to test some of their specific characteristics. For this scenario, the real values of the parameters for the probability distribution of the observable random variables have been set as follows:

- For $K = 2$ the parameters are set to $\mu_1 = 60$, $\sigma_1 = 2$, and $\mu_2 = 70$, $\sigma_2 = 3$, and
- For $K = 3$ the parameters are set to $\mu_1 = 60$, $\sigma_1 = 2$, $\mu_2 = 70$, $\sigma_2 = 1.5$, and $\mu_3 = 80$, $\sigma_3 = 3$.

In this scenario, we use two values for the amount of covariates in the model. When $K = 2$ the value of $D$ is set to 10, and when $K = 3$ the value of $D = 8$. It is important to remember that, given the structure of the NHMM, for any value of $K > 0$ and $D > 0$, we will have $D \times K \times (K-1)$ transition parameters being estimated. The real values for the transition coefficients used in this simulation scenario are shown in the following matrix equations. Due to

space limitations and to improve readability, we only show 3 coefficients in each $\boldsymbol{\beta}_{ij}$, however all $\boldsymbol{\beta}_{ij}$ contain 10 coefficients when $K = 2$ and 8 coefficients when $K = 3$. The rest of the coefficients which are not shown in each entry are set to 0. Equation (4.13) given by

$$\boldsymbol{\beta} = \begin{bmatrix} \begin{pmatrix} \beta_{111} = 0.0 \\ \beta_{112} = 0.0 \\ \beta_{113} = 0.0 \end{pmatrix} & \begin{pmatrix} \beta_{121} = -1.5 \\ \beta_{122} = -1.5 \\ \beta_{123} = -2.6 \end{pmatrix} \\ \\ \begin{pmatrix} \beta_{211} = 0.0 \\ \beta_{212} = 0.0 \\ \beta_{213} = 0.0 \end{pmatrix} & \begin{pmatrix} \beta_{221} = -2.0 \\ \beta_{222} = 2.6 \\ \beta_{223} = 1.4 \end{pmatrix} \end{bmatrix} \tag{4.13}$$

shows the transition coefficients used when $K = 2$.

It is important to remember that since we are using the mlogit link function, then the coefficients related to the transitions from the first non-observable state are all set to zero, in order to achieve parameter identifiability. Equation (4.14) given by

$$\boldsymbol{\beta} = \begin{bmatrix} \begin{pmatrix} \beta_{111} = 0.0 \\ \beta_{112} = 0.0 \\ \beta_{113} = 0.0 \end{pmatrix} & \begin{pmatrix} \beta_{121} = -1.5 \\ \beta_{122} = -1.5 \\ \beta_{123} = -2.6 \end{pmatrix} & \begin{pmatrix} \beta_{131} = -1.3 \\ \beta_{132} = -3.2 \\ \beta_{133} = -2.4 \end{pmatrix} \\ \\ \begin{pmatrix} \beta_{211} = 0.0 \\ \beta_{212} = 0.0 \\ \beta_{213} = 0.0 \end{pmatrix} & \begin{pmatrix} \beta_{221} = -2.0 \\ \beta_{222} = 2.6 \\ \beta_{223} = 1.4 \end{pmatrix} & \begin{pmatrix} \beta_{231} = -1.3 \\ \beta_{232} = 1.7 \\ \beta_{233} = 1.3 \end{pmatrix} \\ \\ \begin{pmatrix} \beta_{311} = 0.0 \\ \beta_{312} = 0.0 \\ \beta_{313} = 0.0 \end{pmatrix} & \begin{pmatrix} \beta_{321} = 2.4 \\ \beta_{322} = 2.1 \\ \beta_{323} = -1.5 \end{pmatrix} & \begin{pmatrix} \beta_{331} = -1.3 \\ \beta_{332} = -2.7 \\ \beta_{333} = -2.5 \end{pmatrix} \end{bmatrix} \tag{4.14}$$

shows transition coefficients used to simulate data when $K = 3$. As mentioned previously, we only show 3 entries due to readability and space limitations.

Regarding the tuning parameter grid, the chosen set of values of the $\lambda$ penalization parameter for the Global LASSO ranged from 0 to 20. To ensure that the vector of tuning parameters was sufficiently fine, 40 values were sequentially generated within this range in incremental steps with a value of 0.5. This guarantees a fine penalization grid to perform tuning. The values for $\lambda_i$ chosen for the Individual LASSO ranged from 2.5 to 12.5. We sequentially consider 14 values (using steps with a value of 0.75) for each $\lambda_i$ in the selected range from 2.5 to 12.5. For this scenario, using the 14 values for each $\lambda_i$ means that when $K = 3$, $14^3$ different models will be tested (all possible combinations for $(\lambda_1, \lambda_2, \lambda_3)$). We verified that this range contained the combination of $\lambda_i$ which is optimal for tuning by observing that the last combination in the range was not commonly selected as the chosen for tuning. Initial values for the transition coefficients are randomly generated from a $N(0,1)$ distribution.

One of the main factors that determines the difficulty of estimating and predicting a model is the amount of overlapping that occurs among the distributions of the random variables for the observable values. That is, if there is a large percentage of observations which have high probabilities of belonging to more than one of the non-observable states then the algorithm will have greater chances of incorrectly identifying the non-observable state in which each observation should be classified and therefore, can produce biased parameter estimates which will, in turn, decrease the quality of predictions.

To illustrate this idea and observe that, in general terms, this scenario may be considered relatively less difficult for the proposals to perform estimation and prediction, we select a specific replication for a specific chain length, and we plot the distribution of the simulated observable values along with the corresponding non-observable state to which they belong to. For $K = 2$, we have randomly selected replication 20 for $T = 600$. For $K = 3$, we have randomly selected replication 17 for $T = 1200$.

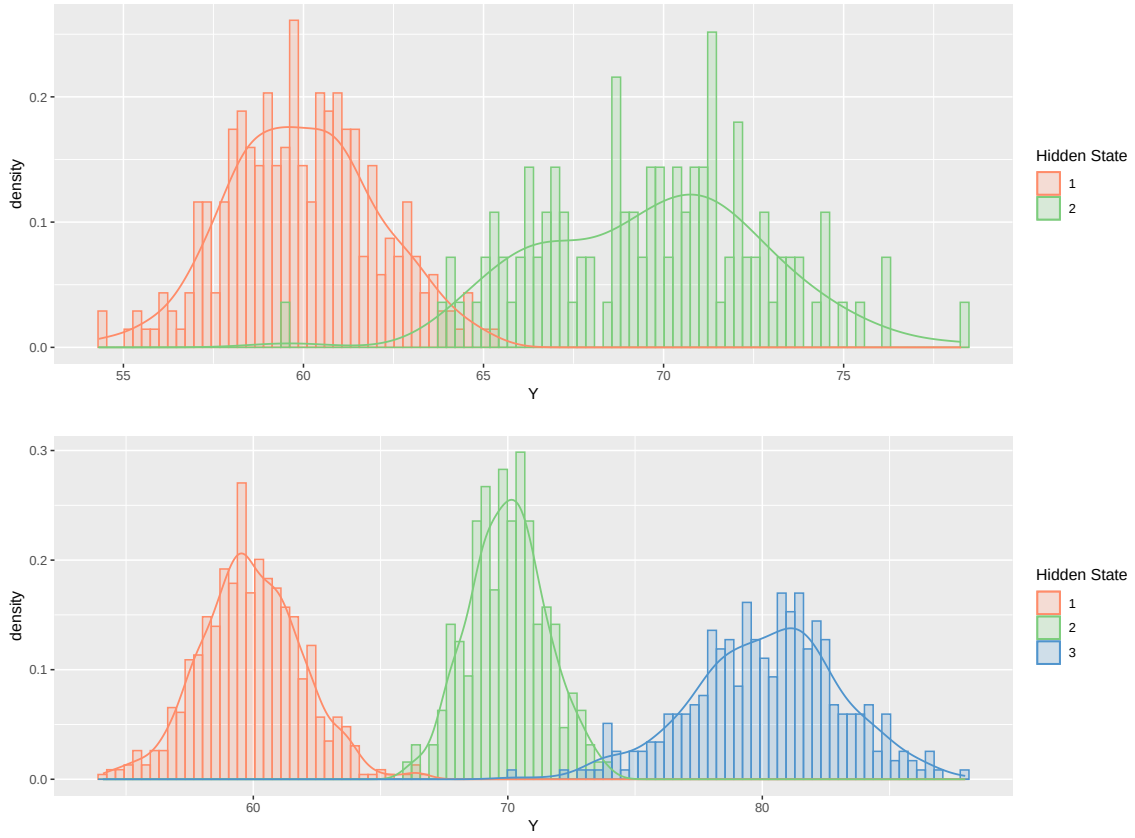



Figure 10 – Distribution of the observable values for $K = 2$ and $K = 3$.

As shown in Figure 10, there is very slight apparent overlapping of the distributions of the observable values of each component. According to the empirical rule, in a Normal distribution, we can expect 68.26% of observations to fall within $\pm 1$ standard deviation from the mean of the distribution, 95.44% will fall within $\pm 2$ standard deviations from the mean and 99.7% will fall

within $\pm 3$ standard deviations from the mean. We can clearly perceive this idea in the graphs shown in Figure10.

For example, when $K = 2$ we have that 95.44% of that data of the distribution of the first non-observable state will be in the interval $[56, 64]$; for the second non-observable state, 95.44% of the observations will fall between $[64, 76]$. These two intervals are not overlapping. Theoretically, there remains approximately 2.28% of observations which have a very high probability of overlapping at the right tail of the distribution for the observable values of non-observable state 1 and the left tail of the distribution for the observable values of non-observable state 2. Since 2.28% of the observations of each distribution is a relatively small quantity of overlapping observations, we label this scenario as not very challenging for estimation and prediction.

A similar situation occurs when $K = 3$. We notice that the intervals that corresponds to $\pm 2$ standard deviations from the mean are $[56, 64]$ for the distribution of the observable values corresponding to non-observable state 1, $[67, 73]$ for non-observable state 2 and $[74, 86]$ for non-observable state 3. As was the case for $K = 2$, we notice that there is possible to have some slight overlapping of these distributions. This is particularly true when we analyze the left tail of the distribution for the observable values of non-observable state 3 and the right tail of the distribution for the observable values of non-observable state 2. However, given the fact that the percentage of overlapping observations is small, the assumption that this scenario is relatively not challenging is still valid. It is important to understand that these are theoretical percentages, and there could be potentially be slightly more (or less) overlapping. Figure 10 gives a clear perception of the actual overlapping for a specific replication when $K = 2$ and $K = 3$.

Table 1 shows the results for the estimation of the parameters of the probability distribution for the observable random variables for $K = 2$. Estimation results are excellent for this scenario for all chain lengths. MSE for the parameters seems to gradually decrease as chain length increases, as does the standard deviation and the magnitude of the bias. Parameter confidence intervals, in general, have relatively small lengths. For most chain lengths, it seems that the magnitude of the bias decreases with chain length. All the aforementioned factors clearly indicate that estimation of this set of parameters is carried out successfully.

Table 2 present the results for estimating the parameters of the probability distribution for observable random variables with $K = 3$. As was the case with $K = 2$, results are consistent and very satisfactory across all chain lengths in this scenario. The mean squared error for the parameters also decreases gradually with increasing chain lengths, as does both the standard deviation and the magnitude of the bias. This shows that for this scenario, retrieval of parameter values has been performed successfully. One particular observation is that parameter confidence interval seem to be slightly greater for the parameters related to non-observable state 2. This is most likely due to the overlapping that occurs among the distribution of this observable values for non-observable state 2 and the other two non-observable states. However, parameter estimates

Table 1 – Estimation results for the parameters of the observable distributions for $K = 2$.

| $T$ | Algo. | Param. | Real | Mean | Median | SD | Bias | MSE | 95%CI |
|---|---|---|---|---|---|---|---|---|---|
| 400 | Global | $\mu_1$ | 60.00 | 60.14 | 59.96 | 0.57 | -0.14 | 0.34 | 59.49, 61.74 |
| | | $\mu_2$ | 70.00 | 69.82 | 69.80 | 0.65 | 0.18 | 0.46 | 68.60, 70.99 |
| | | $\sigma_1$ | 2.00 | 2.20 | 1.91 | 0.75 | -0.20 | 0.61 | 1.66, 4.48 |
| | | $\sigma_2$ | 3.00 | 3.23 | 3.16 | 0.52 | -0.23 | 0.32 | 2.42, 4.20 |
| | Individual | $\mu_1$ | 60.00 | 59.93 | 59.93 | 0.20 | 0.07 | 0.04 | 59.56, 60.27 |
| | | $\mu_2$ | 70.00 | 69.83 | 69.90 | 0.54 | 0.17 | 0.33 | 68.73, 70.67 |
| | | $\sigma_1$ | 2.00 | 1.90 | 1.91 | 0.17 | 0.10 | 0.04 | 1.60, 2.21 |
| | | $\sigma_2$ | 3.00 | 3.15 | 3.04 | 0.40 | -0.15 | 0.19 | 2.54, 4.00 |
| 600 | Global | $\mu_1$ | 60.00 | 59.98 | 59.94 | 0.26 | 0.02 | 0.07 | 59.61, 60.40 |
| | | $\mu_2$ | 70.00 | 69.76 | 69.77 | 0.55 | 0.24 | 0.36 | 68.76, 70.72 |
| | | $\sigma_1$ | 2.00 | 2.01 | 1.98 | 0.34 | -0.01 | 0.12 | 1.63, 2.41 |
| | | $\sigma_2$ | 3.00 | 3.15 | 3.05 | 0.36 | -0.15 | 0.15 | 2.62, 3.85 |
| | Individual | $\mu_1$ | 60.00 | 59.99 | 59.92 | 0.42 | 0.01 | 0.18 | 59.60, 61.30 |
| | | $\mu_2$ | 70.00 | 69.54 | 69.76 | 1.06 | 0.46 | 1.34 | 66.07, 70.47 |
| | | $\sigma_1$ | 2.00 | 2.08 | 1.91 | 0.90 | -0.08 | 0.82 | 1.68, 5.18 |
| | | $\sigma_2$ | 3.00 | 3.13 | 3.10 | 0.32 | -0.13 | 0.12 | 2.62, 3.74 |
| 800 | Global | $\mu_1$ | 60.00 | 60.01 | 60.03 | 0.19 | -0.01 | 0.03 | 59.64, 60.27 |
| | | $\mu_2$ | 70.00 | 69.94 | 70.05 | 0.47 | 0.06 | 0.22 | 68.96, 70.54 |
| | | $\sigma_1$ | 2.00 | 2.04 | 2.02 | 0.20 | -0.04 | 0.04 | 1.69, 2.38 |
| | | $\sigma_2$ | 3.00 | 3.04 | 2.99 | 0.26 | -0.04 | 0.07 | 2.63, 3.66 |
| | Individual | $\mu_1$ | 60.00 | 59.90 | 59.91 | 0.16 | 0.10 | 0.04 | 59.62, 60.15 |
| | | $\mu_2$ | 70.00 | 69.74 | 69.81 | 0.43 | 0.26 | 0.26 | 68.78, 70.48 |
| | | $\sigma_1$ | 2.00 | 1.91 | 1.91 | 0.15 | 0.09 | 0.03 | 1.67, 2.23 |
| | | $\sigma_2$ | 3.00 | 3.13 | 3.14 | 0.33 | -0.13 | 0.13 | 2.54, 3.72 |

are still considered very good.

When comparing retrieval of parameter values by the two proposed algorithms, we can perceive there are marginal differences in their performance. For all chain lengths, in both $K = 2$ and $K = 3$, we observe that parameter confidence intervals have similar lengths and it is apparent that differences in performance are marginal. With these facts, we can state that both proposals have obtained satisfactory parameter estimates.

Table 3 shows the results for the MSPE on the test data set for the two proposals as well as the two methods selected for comparison when $K = 2$. As we can observe, the proposals seem to show better performance when predicting values of the test data set. The Global and Individual LASSO algorithms have consistently obtained lower values for the median and mean of the MSPEs collected over all chain lengths. This indicates that, in general, the two proposed algorithms show better predictive performance when compared to penalized Linear Regression and ARIMA. Similar results can be observed in Table 4 when $K = 3$, where the two proposals also consistently obtained lower values of the MSPE for all chain lengths.

To have a more detailed understanding of the specific predictive performance of the algorithms in Scenario 1, we plot the MSPE along the 50 replications for $K = 2$ and along the

Table 2 – Estimation results for the parameters of the observable distributions for $K = 3$.

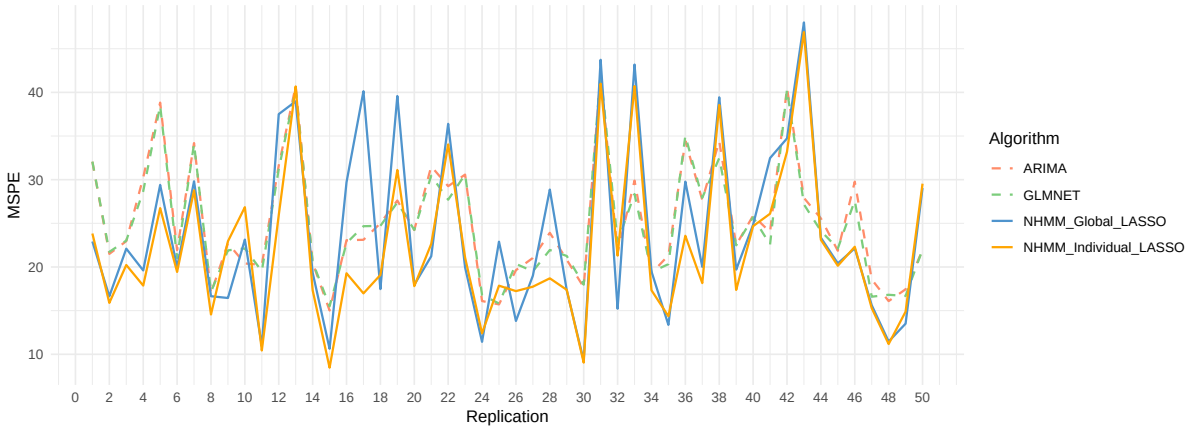
| $T$ | Algo. | Param. | Real | Mean | Median | SD | Bias | MSE | 95%CI |
|---|---|---|---|---|---|---|---|---|---|
| 800 | Global | $\mu_1$ | 60.00 | 59.83 | 59.91 | 0.26 | 0.17 | 0.10 | 59.13, 60.11 |
| | | $\mu_2$ | 70.00 | 69.30 | 69.85 | 1.56 | 0.70 | 2.90 | 65.75, 70.73 |
| | | $\mu_3$ | 80.00 | 80.10 | 80.30 | 0.69 | -0.10 | 0.49 | 78.13, 80.62 |
| | | $\sigma_1$ | 2.00 | 1.86 | 1.90 | 0.19 | 0.14 | 0.06 | 1.45, 2.15 |
| | | $\sigma_2$ | 1.50 | 1.76 | 1.52 | 1.22 | -0.26 | 1.56 | 1.27, 5.21 |
| | | $\sigma_3$ | 3.00 | 2.97 | 2.82 | 0.60 | 0.03 | 0.37 | 2.52, 5.07 |
| | Individual | $\mu_1$ | 60.00 | 59.89 | 59.92 | 0.38 | 0.11 | 0.16 | 59.16, 60.53 |
| | | $\mu_2$ | 70.00 | 68.78 | 69.81 | 2.61 | 1.22 | 8.29 | 61.21, 70.46 |
| | | $\mu_3$ | 80.00 | 79.72 | 80.15 | 1.17 | 0.28 | 1.46 | 76.29, 80.72 |
| | | $\sigma_1$ | 2.00 | 1.83 | 1.88 | 0.23 | 0.17 | 0.08 | 1.33, 2.14 |
| | | $\sigma_2$ | 1.50 | 1.58 | 1.48 | 1.39 | -0.08 | 1.92 | 1.10, 5.58 |
| | | $\sigma_3$ | 3.00 | 3.27 | 2.90 | 0.92 | -0.27 | 0.91 | 2.39, 5.56 |
| 1000 | Global | $\mu_1$ | 60.00 | 59.84 | 59.88 | 0.23 | 0.16 | 0.08 | 59.29, 60.12 |
| | | $\mu_2$ | 70.00 | 69.68 | 69.88 | 1.02 | 0.32 | 1.15 | 67.40, 71.07 |
| | | $\mu_3$ | 80.00 | 80.22 | 80.18 | 0.49 | -0.22 | 0.29 | 79.38, 81.08 |
| | | $\sigma_1$ | 2.00 | 1.85 | 1.89 | 0.17 | 0.15 | 0.05 | 1.45, 2.06 |
| | | $\sigma_2$ | 1.50 | 1.82 | 1.67 | 1.09 | -0.32 | 1.29 | 1.39, 4.70 |
| | | $\sigma_3$ | 3.00 | 2.78 | 2.77 | 0.42 | 0.22 | 0.22 | 2.30, 3.51 |
| | Individual | $\mu_1$ | 60.00 | 59.80 | 59.82 | 0.25 | 0.20 | 0.10 | 59.34, 60.16 |
| | | $\mu_2$ | 70.00 | 69.29 | 69.64 | 1.25 | 0.71 | 2.06 | 66.91, 71.03 |
| | | $\mu_3$ | 80.00 | 80.09 | 80.16 | 0.68 | -0.09 | 0.46 | 78.53, 81.19 |
| | | $\sigma_1$ | 2.00 | 1.81 | 1.86 | 0.18 | 0.19 | 0.07 | 1.46, 2.04 |
| | | $\sigma_2$ | 1.50 | 1.77 | 1.68 | 1.30 | -0.27 | 1.76 | 1.28, 5.05 |
| | | $\sigma_3$ | 3.00 | 2.92 | 2.80 | 0.60 | 0.08 | 0.37 | 2.26, 4.39 |
| 1200 | Global | $\mu_1$ | 60.00 | 59.82 | 59.86 | 0.21 | 0.18 | 0.08 | 59.35, 60.09 |
| | | $\mu_2$ | 70.00 | 69.43 | 69.83 | 0.83 | 0.57 | 1.02 | 67.28, 70.26 |
| | | $\mu_3$ | 80.00 | 80.20 | 80.24 | 0.32 | -0.20 | 0.14 | 79.32, 80.64 |
| | | $\sigma_1$ | 2.00 | 1.86 | 1.88 | 0.13 | 0.14 | 0.04 | 1.62, 2.01 |
| | | $\sigma_2$ | 1.50 | 1.56 | 1.54 | 1.01 | -0.06 | 1.02 | 1.36, 4.31 |
| | | $\sigma_3$ | 3.00 | 2.84 | 2.81 | 0.24 | 0.16 | 0.08 | 2.54, 3.46 |
| | Individual | $\mu_1$ | 60.00 | 59.85 | 59.88 | 0.15 | 0.15 | 0.05 | 59.44, 60.05 |
| | | $\mu_2$ | 70.00 | 69.89 | 69.86 | 0.84 | 0.11 | 0.71 | 68.31, 71.46 |
| | | $\mu_3$ | 80.00 | 80.30 | 80.31 | 0.50 | -0.30 | 0.34 | 79.31, 81.17 |
| | | $\sigma_1$ | 2.00 | 1.88 | 1.89 | 0.11 | 0.12 | 0.03 | 1.67, 2.03 |
| | | $\sigma_2$ | 1.50 | 1.78 | 1.73 | 1.07 | -0.28 | 1.22 | 1.26, 4.47 |
| | | $\sigma_3$ | 3.00 | 2.77 | 2.70 | 0.38 | 0.23 | 0.20 | 2.23, 3.55 |

30 replications for $K = 3$. The plotted chain lengths were selected randomly; the selected chain length for $K = 2$ is $T = 600$ and for $K = 3$ is $T = 1200$.

Figure 11 shows the MSPE for the proposals as well as the two algorithms which we are comparing. Although the graph shows that there are few replications in which proposed methods have a slightly higher MSPE, it is evident that for the majority of replications, the LASSOs for NHMMs have obtained lower values for the MSPE. To be precise, when $K = 2$, the Individual LASSO has obtained lower MSPEs than ARIMA in 80% of replications and lower

Table 3 – Mean squared predictive error results for the $K = 2$.

| $T$ | **Algo.** | **Median** | **Mean** | **SD** |
|---|---|---|---|---|
| 400 | Global | 24.98 | 26.65 | 8.61 |
| | Individual | 24.39 | 26.05 | 9.11 |
| | ARIMA | 27.51 | 28.70 | 8.00 |
| | Pen. Linear Reg. | 27.52 | 28.11 | 7.08 |
| 600 | Global | 20.81 | 23.97 | 9.93 |
| | Individual | 19.78 | 22.24 | 8.69 |
| | ARIMA | 23.09 | 25.24 | 6.88 |
| | Pen. Linear Reg. | 22.68 | 24.96 | 6.73 |
| 800 | Global | 20.62 | 22.74 | 7.26 |
| | Individual | 20.73 | 22.66 | 7.58 |
| | ARIMA | 25.12 | 26.21 | 6.00 |
| | Pen. Linear Reg. | 25.17 | 25.67 | 5.63 |




(a) $K = 2$


(b) $K = 3$

Figure 11 – Values of the MSPE along all replications for scenario 1.

than Penalized Linear Regression in 78% of replications. The Global LASSO has obtained lower MSPEs than both ARIMA and Penalized Linear Regression in 70% of replications. When $K = 3$, the Individual LASSO has obtained lower MSPEs than ARIMA in 70% of replications and lower than Penalized Linear Regression in 76% of replications. The Global LASSO has obtained lower MSPEs than ARIMA in 73% of replications and lower than Penalized Linear Regression in 73% of replications. With this details, we can conclude that in general, the proposals show better performance than the comparison algorithms.

Table 4 – Mean squared predictive error results for the $K = 3$.

| $T$ | **Algo.** | **Median** | **Mean** | **SD** |
|---|---|---|---|---|
| 800 | Global | 58.73 | 56.85 | 11.76 |
| | Individual | 53.22 | 55.58 | 11.96 |
| | ARIMA | 60.93 | 64.02 | 17.18 |
| | Pen. Linear Reg. | 60.14 | 60.63 | 8.48 |
| 1000 | Global | 58.14 | 59.69 | 13.31 |
| | Individual | 57.47 | 59.61 | 15.56 |
| | ARIMA | 64.62 | 67.59 | 13.53 |
| | Pen. Linear Reg. | 62.78 | 65.11 | 9.90 |
| 1200 | Global | 58.63 | 57.63 | 12.18 |
| | Individual | 57.63 | 57.61 | 11.67 |
| | ARIMA | 61.55 | 63.07 | 8.24 |
| | Pen. Linear Reg. | 60.98 | 62.66 | 7.51 |

To further comprehend the predictive performance of the two proposed algorithms and to analyze a broad range of situations including one in which performance is not as satisfactory as might be expected, we have selected a specific replication for further analysis where the proposals did not have superior performance. Given the graphical results shown in Figure 11, we select replication 26 when $K = 3$ and $T = 1200$. As shown in the figure, it seems that in replication 26, the MSPE for the proposed algorithms was relatively higher than the two comparison algorithms, as well as being the replication in which the two proposals scored the highest MSPE. We analyze this replication in detail in Figure 12.

As can be perceived in Figure 12, the simulated data for this specific scenario is extremely erratic, presenting great difficulties for all algorithms which are being compared. There are some intervals in which the proposals seem to capture the general behavior of the real data better than the comparison algorithms. Examples of these intervals are indices 0 - 5, 14-18, 24 - 28. However, in general, it seems that this particular replications has presented generalized problems for all algorithms to perform prediction. We believe that the erratic nature of this particular replication has randomly produced an greater MSPE for the proposals than for the comparison algorithms.

Table 5 and Table 6 show results for the amount of hits when predicting **S** test sequence. Given that the ARIMA and Penalized Linear Regression algorithms do not perform classification, we will only be comparing results for the Global and Individual LASSO. The tables show that

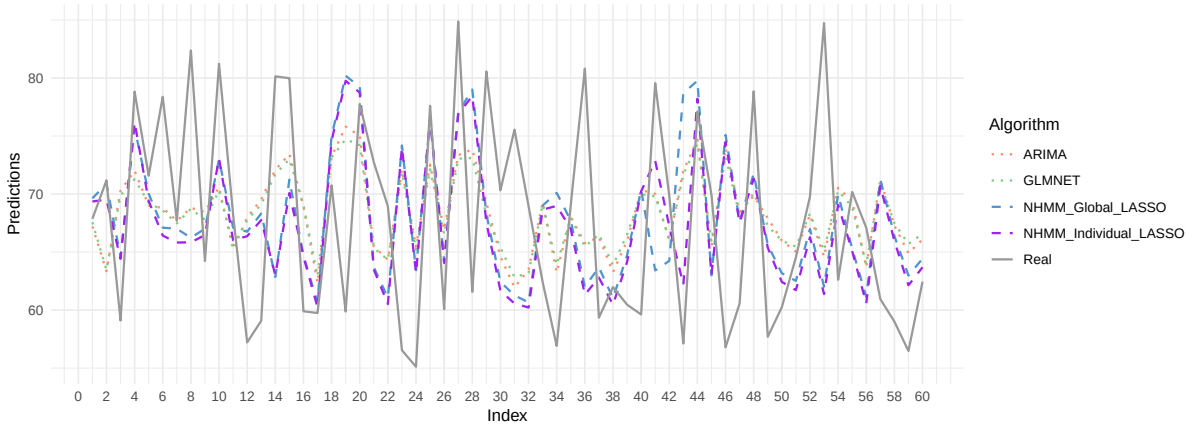



Figure 12 – Predicted values for the proposed and comparison algorithms for replication 26, when $K = 3$ and $T = 1200$.

Table 5 – Results for the hit frequency when predicting **S** test sequence for $K = 2$.

| $T$ | **Algo.** | **Median** | **Mean** | **SD** |
|---|---|---|---|---|
| 400 | Global | 0.79 | 0.80 | 0.09 |
| | Individual | 0.80 | 0.80 | 0.07 |
| 600 | Global | 0.76 | 0.77 | 0.15 |
| | Individual | 0.78 | 0.80 | 0.13 |
| 800 | Global | 0.77 | 0.78 | 0.08 |
| | Individual | 0.77 | 0.78 | 0.07 |

results for hit frequency are very good for all chain lengths. We must stress the fact that for every position in the non-observable sequence, the algorithm must choose the correct non-observable state among one of the $K$ possible values. This intuitively indicates that the values for the hit frequency collected in this scenario show performance which is more than satisfactory.

Table 6 – Results for the hit frequency when predicting **S** test sequence for $K = 3$.

| $T$ | **Algo.** | **Median** | **Mean** | **SD** |
|---|---|---|---|---|
| 800 | Global | 0.71 | 0.71 | 0.08 |
| | Individual | 0.71 | 0.70 | 0.08 |
| 1000 | Global | 0.69 | 0.71 | 0.09 |
| | Individual | 0.69 | 0.70 | 0.09 |
| 1200 | Global | 0.69 | 0.70 | 0.09 |
| | Individual | 0.69 | 0.70 | 0.08 |

When comparing the two proposed algorithms, we perceive that the differences in accuracy when predicting the non-observable **S** test sequence are almost non-existent. However, because of the fact that in a real data setting we do not have the true values of the labels for the the non-observable states, this metric is only useful when we are in a simulated data setting.

To have a more detailed understanding of how prediction of the **S** test sequence occurs,

we have also constructed a confusion matrix which shows the predicted values and true values for the **S** Test sequence. We randomly select a single replication for specific chain length for one of the proposed algorithms. We select replication 5 for $T = 1200$ for the Global LASSO. The results for the confusion matrix when $K = 3$ are shown in Table 7.

Table 7 – Confusion matrix for real vs. predicted values of **S** test sequence in replication 5 when $T = 1200$ for the Global LASSO.

| | | ***Predicted Value*** | | |
|---|---|---|---|---|
| | | **1** | **2** | **3** |
| ***Real Value*** | **1** | 31 | 0 | 3 |
| | **2** | 5 | 3 | 2 |
| | **3** | 4 | 1 | 11 |

As we perceive from Table 7, the algorithm is most accurate when identifying state 1, having an accuracy of 77.5% (31 hits out of 40 attempts). It is least accurate when identifying non-observable state 3, obtaining an accuracy of 68.75% (11 hits out of 16 attempts). In spite of that lesser accuracy for non-observable state 3, we observe a predictive accuracy of 75% for the specific replication displayed. Considering 3 non-observable states, this is very good performance.

The next set of results are related to the proposals' capacity to shrink transition coefficients to 0. Here, it is important to mention the impact of selecting the correct optimization algorithm. In previous simulation experiments, the researcher obtained poor results in metrics related shrinkage. It was later ascertained through experimentation that these poor results were due to selecting the incorrect optimization procedure for our objective functions. Tables 8 and 9 show results for shrinkage of the transition coefficients. As we can perceive, the two algorithms have obtained excellent results regarding their capacity to shrink coefficients whose actual value is 0. For all chain lengths and for both proposals, we observe that mean values for specificity are always greater than 0.95, meaning that the proposals rarely shrink a coefficient to 0, when its actual value is different than 0. The mean values for sensitivity are all greater than 0.80 for $K = 2$ and greater than 0.70 for $K = 3$. This shows that the proposals are successful in identifying and shrinking to 0 the majority of the coefficients whose actual value is 0.

One interesting fact we notice when we observe Tables 8 and 9, is that standard deviations for sensitivity (and therefore, for accuracy as well) are considerably greater for the Global LASSO. This is most likely due to the fact that in the Individual LASSO, there is an individual penalization parameter for the regressions corresponding to each non-observable state and their tuning is performed individually, therefore leading to more accurate shrinkage of the transition coefficients. On the other hand, the Global LASSO has a single penalization parameter for all regressions. This might lead to under-penalization (or over-penalization) of transition coefficients for one of the non-observable states in one or more specific replications, and therefore leading to greater

Table 8 – Results for shrinkage of the transition coefficients for $K = 2$.

| $T$ | Algo. | Specificity | | | Sensitivity | | | Accuracy | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | *Median* | *Mean* | *SD* | *Median* | *Mean* | *SD* | *Median* | *Mean* | *SD* |
| 400 | Global | 1.00 | 0.95 | 0.09 | 0.93 | 0.80 | 0.26 | 0.93 | 0.84 | 0.18 |
| | Individual | 1.00 | 0.96 | 0.07 | 0.89 | 0.85 | 0.15 | 0.90 | 0.88 | 0.11 |
| 600 | Global | 1.00 | 0.97 | 0.06 | 0.96 | 0.81 | 0.29 | 0.95 | 0.86 | 0.20 |
| | Individual | 1.00 | 0.98 | 0.08 | 0.86 | 0.83 | 0.15 | 0.90 | 0.88 | 0.11 |
| 800 | Global | 1.00 | 0.99 | 0.04 | 1.00 | 0.93 | 0.13 | 1.00 | 0.95 | 0.09 |
| | Individual | 1.00 | 1.00 | 0.02 | 0.93 | 0.88 | 0.12 | 0.95 | 0.92 | 0.09 |

values of standard deviation for sensitivity. Even though the Global LASSO presents these greater standard deviations in the specificity, we must conclude that both proposals demonstrate excellent performance in terms of shrinkage and variable selection.

When comparing results shown in Tables 8 and 9 to results observed in the first round of simulations in Appendix A, we perceive great improvement regarding shrinkage performance. Particularly, if we compare Tables 54 and 55 to Table 8 which correspond to similar parameter configurations for simulated data, we observe notable differences in shrinkage performance. For specificity, we see differences of 75% in some cases, displaying great improvement in this metric. The same is observed when we compare Table 9 to Tables 63 and 64, which come from simulated data with similar parameter configurations. In such comparison we see an increase in performance of around 80%, in terms of specificity.

Table 9 – Results for shrinkage of the transition coefficients for $K = 3$.

| $T$ | Algo. | Sensitivity | | | Specificity | | | Accuracy | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | *Median* | *Mean* | *SD* | *Median* | *Mean* | *SD* | *Median* | *Mean* | *SD* |
| 800 | Global | 1.00 | 0.94 | 0.09 | 0.73 | 0.71 | 0.29 | 0.81 | 0.80 | 0.16 |
| | Individual | 0.89 | 0.92 | 0.07 | 0.82 | 0.81 | 0.12 | 0.85 | 0.85 | 0.07 |
| 1000 | Global | 0.94 | 0.94 | 0.08 | 0.83 | 0.75 | 0.26 | 0.89 | 0.82 | 0.16 |
| | Individual | 0.89 | 0.91 | 0.08 | 0.80 | 0.79 | 0.10 | 0.84 | 0.83 | 0.06 |
| 1200 | Global | 0.92 | 0.92 | 0.08 | 0.92 | 0.80 | 0.25 | 0.92 | 0.84 | 0.14 |
| | Individual | 0.94 | 0.93 | 0.07 | 0.82 | 0.81 | 0.09 | 0.84 | 0.85 | 0.06 |

Tables 10 and 11 show observed processing times for Scenario 1. As expected, the Global LASSO has a shorter processing time. This is due to the fact the amount of models that will be tested to find the best is the number of tuning parameters considered, while in Individual LASSO it is an exponential of it. For example, when $K = 3$ and considering 14 different values for the tuning parameter, then $14^3$ models will be tested in the Individual LASSO, amounting to 2744 models, while in the Global LASSO only 14 models will be tested. With this, we begin to perceive the computational costs of using a greater quantity of non-observable states with the Individual LASSO. This problem could be somewhat mitigated by designing a procedure to optimal select a grid of values for the tuning parameter. This would allow for less values to be tested in the tuning process and would reduce computational costs for the Individual LASSO

proposal.

Table 10 – Processing times (in minutes) for $K = 2$, where total time considers running the $R = 50$ replications.

| $T$ | Algo. | Total Time | Avg. Time/Rep | Median Time/Rep | SD Time/Rep |
|---|---|---|---|---|---|
| 400 | Global | 35.17 | 0.70 | 0.70 | 0.03 |
| | Individual | 591.17 | 11.82 | 11.56 | 1.34 |
| 600 | Global | 46.37 | 0.93 | 0.93 | 0.02 |
| | Individual | 920.39 | 18.41 | 18.40 | 0.20 |
| 800 | Global | 56.31 | 1.13 | 1.13 | 0.03 |
| | Individual | 1,128.76 | 22.58 | 22.58 | 0.36 |

Table 11 – Processing times (in minutes) for $K = 3$, where total time considers running the $R = 30$ replications.

| $T$ | Algo. | Total Time | Avg. Time/Rep | Median Time/Rep | SD Time/Rep |
|---|---|---|---|---|---|
| 800 | Global | 76.15 | 2.54 | 2.54 | 0.15 |
| | Individual | 5,101.27 | 170.04 | 170.83 | 7.36 |
| 1000 | Global | 89.62 | 2.99 | 3.00 | 0.06 |
| | Individual | 6,163.31 | 205.44 | 204.01 | 6.72 |
| 1200 | Global | 99.60 | 3.32 | 3.33 | 0.15 |
| | Individual | 6,899.92 | 230.00 | 230.64 | 9.60 |

### 4.4.2 Scenario 2

The next scenario is designed to test the capacity of the proposed algorithms to identify transition coefficients whose actual value is relatively closer to 0 than the rest of the coefficients in the model. For this scenario, the real values of the parameters for the probability distribution of the observable random variables have been set as follows:

- For $K = 2$ the parameters are set to $\mu_1 = 60$, $\sigma_1 = 2$, and $\mu_2 = 70$, $\sigma_2 = 3$; and
- For $K = 3$ the parameters are set to $\mu_1 = 60$, $\sigma_1 = 2$, $\mu_2 = 70$, $\sigma_2 = 1.5$, and $\mu_3 = 80$, $\sigma_3 = 3$.

In this scenario, we use two values for the amount of covariates in the model. When $K = 2$ the value of $D$ is set to 10, and when $K = 3$ the value of $D = 8$. The previously described parameter configuration is analogous to Scenario 1. The real values for the transition coefficients used in this simulation scenario are shown below. As is done in previous scenarios, we only display the values for non-zero coefficients in each $\boldsymbol{\beta}_{ij}$, however all $\boldsymbol{\beta}_{ij}$ vectors contain 10 coefficients when $K = 2$ and 8 coefficients when $K = 3$. The rest of the coefficients in each entry are set to 0.

Equation (4.15) shows the transition coefficients used when $K = 2$, given by

$$\boldsymbol{\beta} = \begin{bmatrix} \begin{pmatrix} \beta_{111} = 0.0 \\ \beta_{112} = 0.0 \\ \beta_{113} = 0.0 \\ \beta_{114} = 0.0 \end{pmatrix} & \begin{pmatrix} \beta_{121} = -1.5 \\ \beta_{125} = -1.5 \\ \beta_{126} = -2.6 \\ \beta_{129} = 0.15 \end{pmatrix} \\ \\ \begin{pmatrix} \beta_{211} = 0.0 \\ \beta_{212} = 0.0 \\ \beta_{213} = 0.0 \\ \beta_{214} = 0.0 \end{pmatrix} & \begin{pmatrix} \beta_{221} = -2.0 \\ \beta_{224} = 2.6 \\ \beta_{226} = 1.4 \\ \beta_{228} = 0.05 \end{pmatrix} \end{bmatrix}. \tag{4.15}$$

As can be observed, there is one entry in each $\boldsymbol{\beta}_{ij}$ whose value is relatively close to 0. This should present a challenge to the proposals in term performing accurate variable selection. Equation (4.16) given by

$$\boldsymbol{\beta} = \begin{bmatrix} \begin{pmatrix} \beta_{111} = 0.0 \\ \beta_{112} = 0.0 \\ \beta_{113} = 0.0 \\ \beta_{114} = 0.0 \end{pmatrix} & \begin{pmatrix} \beta_{121} = -1.5 \\ \beta_{124} = -1.5 \\ \beta_{125} = 0.2 \\ \beta_{127} = -2.6 \end{pmatrix} & \begin{pmatrix} \beta_{131} = -1.3 \\ \beta_{132} = 0.1 \\ \beta_{136} = -3.2 \\ \beta_{138} = -2.4 \end{pmatrix} \\ \\ \begin{pmatrix} \beta_{211} = 0.0 \\ \beta_{212} = 0.0 \\ \beta_{213} = 0.0 \\ \beta_{214} = 0.0 \end{pmatrix} & \begin{pmatrix} \beta_{221} = -2.0 \\ \beta_{223} = 2.6 \\ \beta_{224} = 0.15 \\ \beta_{227} = 1.4 \end{pmatrix} & \begin{pmatrix} \beta_{231} = -1.3 \\ \beta_{232} = 1.7 \\ \beta_{233} = 1.3 \\ \beta_{237} = 0.25 \end{pmatrix} \\ \\ \begin{pmatrix} \beta_{311} = 0.0 \\ \beta_{312} = 0.0 \\ \beta_{313} = 0.0 \\ \beta_{314} = 0.0 \end{pmatrix} & \begin{pmatrix} \beta_{321} = 2.4 \\ \beta_{324} = 2.1 \\ \beta_{327} = -1.5 \\ \beta_{328} = 0.3 \end{pmatrix} & \begin{pmatrix} \beta_{331} = -1.3 \\ \beta_{332} = -2.7 \\ \beta_{335} = 0.3 \\ \beta_{338} = -2.5 \end{pmatrix} \end{bmatrix} \tag{4.16}$$

shows transition coefficients used to simulate data when $K = 3$. As mentioned previously, we only display 4 entries due to readability and space constraints.

Regarding the tuning parameter grid, the chosen set of values of the $\lambda$ penalization parameter for the Global LASSO ranged from 0 to 20. To ensure that the vector of tuning parameters was sufficiently fine, 40 values were sequentially considered within this range in incremental steps with a value of 0.5. This guarantees a fine penalization grid to perform tuning. The values for $\lambda_i$ chosen for the Individual LASSO ranged from 2.5 to 12.5. We sequentially consider 14 values (using steps with a value of 0.75) for each $\lambda_i$ in the selected range from 2.5 to 12.5.

The parameter configuration used in this scenario is analogous to that which is used in Scenario 1. This classifies this scenario, in general, as not difficult for the proposals to perform

estimation and prediction. However, regarding the metrics related to shrinkage accuracy, the parameter configuration of this scenario presents an additional difficulty factor, which is the presence of coefficients whose actual value is relatively close to 0. This aspect will be analyzed in detail in this scenario.

As was done in previous scenarios, we select a specific replication for a specific chain length, and we plot the distribution of the simulated observable values along with the corresponding non-observable state to which they belong to, in order to have a visual perception and understanding of overlapping in this particular scenario. For $K = 2$, we have randomly selected replication 22 for $T = 800$. For $K = 3$, we have randomly selected replication 7 for $T = 1000$.


Figure 13 – Distribution of the observable values for $K = 2$ and $K = 3$.

Graphical results in Figure 13 indicate that overlapping of the distributions of the observable values of each non-observable state occurs in the same manner as in Scenario 1. Therefore, the reader may refer to Section 4.4.1 to understand theoretical overlapping percentages for this scenario. In general, and without considering the specific challenge of having to select the coefficients whose actual value is close to 0, then this scenario presents a situation which should not be difficult for the proposed algorithms.

Table 12 displays the results for the estimation of the parameters of the probability distribution for the observable random variables for $K = 2$. The proposed algorithms are notably successful when estimating this set of parameters for all chain lengths. MSE for the parameters

Table 12 – Estimation results for the parameters of the observable distributions for $K = 2$.

| $T$ | Algo. | Param. | Real | Mean | Median | SD | Bias | MSE | 95%CI |
|---|---|---|---|---|---|---|---|---|---|
| 400 | Global | $\mu_1$ | 60.00 | 60.09 | 59.95 | 0.50 | -0.09 | 0.26 | 59.45, 61.61 |
| | | $\mu_2$ | 70.00 | 69.81 | 69.95 | 0.63 | 0.18 | 0.43 | 68.68, 70.94 |
| | | $\sigma_1$ | 2.00 | 2.14 | 1.92 | 0.66 | -0.14 | 0.46 | 1.59, 4.44 |
| | | $\sigma_2$ | 3.00 | 3.24 | 3.18 | 0.49 | -0.24 | 0.30 | 2.44, 4.18 |
| | Individual | $\mu_1$ | 60.00 | 59.92 | 59.93 | 0.19 | 0.08 | 0.04 | 59.48, 60.30 |
| | | $\mu_2$ | 70.00 | 69.82 | 69.93 | 0.58 | 0.18 | 0.36 | 68.52, 70.77 |
| | | $\sigma_1$ | 2.00 | 1.90 | 1.91 | 0.16 | 0.10 | 0.04 | 1.60, 2.21 |
| | | $\sigma_2$ | 3.00 | 3.15 | 3.09 | 0.38 | -0.15 | 0.17 | 2.50, 3.88 |
| 600 | Global | $\mu_1$ | 60.00 | 59.95 | 59.94 | 0.22 | 0.05 | 0.05 | 59.59, 60.38 |
| | | $\mu_2$ | 70.00 | 69.74 | 69.80 | 0.57 | 0.26 | 0.39 | 68.72, 70.64 |
| | | $\sigma_1$ | 2.00 | 1.97 | 1.90 | 0.24 | 0.03 | 0.06 | 1.66, 2.60 |
| | | $\sigma_2$ | 3.00 | 3.17 | 3.19 | 0.36 | -0.17 | 0.15 | 2.60, 3.85 |
| | Individual | $\mu_1$ | 60.00 | 59.95 | 59.95 | 0.31 | 0.05 | 0.10 | 59.63, 60.31 |
| | | $\mu_2$ | 70.00 | 69.68 | 69.82 | 0.87 | 0.32 | 0.86 | 68.72, 70.51 |
| | | $\sigma_1$ | 2.00 | 2.00 | 1.90 | 0.66 | 0.00 | 0.44 | 1.63, 2.23 |
| | | $\sigma_2$ | 3.00 | 3.09 | 3.05 | 0.31 | -0.09 | 0.10 | 2.58, 3.76 |
| 800 | Global | $\mu_1$ | 60.00 | 60.02 | 60.03 | 0.20 | -0.02 | 0.04 | 59.62, 60.32 |
| | | $\mu_2$ | 70.00 | 69.93 | 70.04 | 0.49 | 0.07 | 0.25 | 68.90, 70.55 |
| | | $\sigma_1$ | 2.00 | 2.05 | 2.04 | 0.21 | -0.05 | 0.05 | 1.67, 2.50 |
| | | $\sigma_2$ | 3.00 | 3.05 | 2.99 | 0.30 | -0.05 | 0.09 | 2.59, 3.76 |
| | Individual | $\mu_1$ | 60.00 | 59.91 | 59.94 | 0.18 | 0.09 | 0.04 | 59.61, 60.15 |
| | | $\mu_2$ | 70.00 | 69.73 | 69.78 | 0.43 | 0.27 | 0.26 | 68.98, 70.48 |
| | | $\sigma_1$ | 2.00 | 1.92 | 1.89 | 0.16 | 0.08 | 0.03 | 1.70, 2.28 |
| | | $\sigma_2$ | 3.00 | 3.14 | 3.13 | 0.32 | -0.14 | 0.12 | 2.59, 3.72 |

seems to gradually decrease as chain length increases. The standard deviation and the magnitude of the bias also seems to decrease as chain length increases, with the exception of Individual LASSO for $T = 600$. In spite of this greater variability, parameter estimates are still excellent for that specific parameter set. Parameter confidence intervals, in general, have relatively small lengths. For most chain lengths, it seems that the magnitude of the bias decreases with chain length. All the previously described characteristics of this set of estimates clearly indicate that estimation of this set of parameters is carried out successfully and with excellent results.

Table 13 shows results for estimation of the parameters corresponding to the probability distribution for observable random variables when $K = 3$. As was similarly observed when $K = 2$, results are consistent and very satisfactory across all chain lengths in this scenario. It seems that the mean squared error for the parameters also decreases gradually with increasing chain lengths. In general, retrieval of parameter values has been performed successfully and obtaining very good results. As was observed in Scenario 1, parameter confidence intervals seem to be slightly greater for the parameters related to non-observable state 2. We observe that values for the standard deviation of these parameters related to non-observable state 2 is relatively larger when compared to parameters for the other two non-observable states. This must be due to the

Table 13 – Estimation results for the parameters of the observable distributions for $K = 3$.

| $T$ | Algo. | Param. | Real | Mean | Median | SD | Bias | MSE | 95%CI |
|---|---|---|---|---|---|---|---|---|---|
| 800 | Global | $\mu_1$ | 60.00 | 59.85 | 59.87 | 0.16 | 0.15 | 0.05 | 59.52, 60.11 |
| | | $\mu_2$ | 70.00 | 69.51 | 69.81 | 0.95 | 0.49 | 1.14 | 67.69, 70.77 |
| | | $\mu_3$ | 80.00 | 80.26 | 80.37 | 0.45 | -0.26 | 0.27 | 79.28, 80.90 |
| | | $\sigma_1$ | 2.00 | 1.86 | 1.86 | 0.15 | 0.14 | 0.04 | 1.60, 2.10 |
| | | $\sigma_2$ | 1.50 | 1.72 | 1.76 | 1.14 | -0.22 | 1.35 | 1.33, 3.96 |
| | | $\sigma_3$ | 3.00 | 2.87 | 2.78 | 0.39 | 0.13 | 0.17 | 2.40, 3.83 |
| | Individual | $\mu_1$ | 60.00 | 59.83 | 59.86 | 0.20 | 0.17 | 0.07 | 59.45, 60.12 |
| | | $\mu_2$ | 70.00 | 69.54 | 69.82 | 1.26 | 0.46 | 1.80 | 67.51, 71.66 |
| | | $\mu_3$ | 80.00 | 80.05 | 80.27 | 0.74 | -0.05 | 0.55 | 78.22, 80.91 |
| | | $\sigma_1$ | 2.00 | 1.86 | 1.91 | 0.18 | 0.14 | 0.05 | 1.57, 2.13 |
| | | $\sigma_2$ | 1.50 | 1.68 | 1.71 | 1.08 | -0.18 | 1.20 | 1.21, 3.63 |
| | | $\sigma_3$ | 3.00 | 3.02 | 2.79 | 0.70 | -0.02 | 0.49 | 2.21, 4.73 |
| 1000 | Global | $\mu_1$ | 60.00 | 59.83 | 59.88 | 0.21 | 0.17 | 0.07 | 59.39, 60.17 |
| | | $\mu_2$ | 70.00 | 69.60 | 69.73 | 0.93 | 0.40 | 1.02 | 67.55, 71.15 |
| | | $\mu_3$ | 80.00 | 80.27 | 80.19 | 0.38 | -0.27 | 0.22 | 79.75, 81.19 |
| | | $\sigma_1$ | 2.00 | 1.83 | 1.84 | 0.15 | 0.17 | 0.05 | 1.53, 2.04 |
| | | $\sigma_2$ | 1.50 | 1.75 | 1.69 | 1.04 | -0.25 | 1.14 | 1.45, 3.84 |
| | | $\sigma_3$ | 3.00 | 2.72 | 2.75 | 0.22 | 0.28 | 0.13 | 2.34, 3.17 |
| | Individual | $\mu_1$ | 60.00 | 59.81 | 59.89 | 0.23 | 0.19 | 0.09 | 59.41, 60.15 |
| | | $\mu_2$ | 70.00 | 69.69 | 69.69 | 1.36 | 0.31 | 1.96 | 67.53, 72.29 |
| | | $\mu_3$ | 80.00 | 80.02 | 80.09 | 0.74 | -0.02 | 0.54 | 78.25, 80.92 |
| | | $\sigma_1$ | 2.00 | 1.84 | 1.86 | 0.18 | 0.16 | 0.06 | 1.51, 2.19 |
| | | $\sigma_2$ | 1.50 | 1.68 | 1.72 | 0.95 | -0.18 | 0.93 | 1.30, 3.69 |
| | | $\sigma_3$ | 3.00 | 2.90 | 2.79 | 0.52 | 0.10 | 0.29 | 2.31, 4.40 |
| 1200 | Global | $\mu_1$ | 60.00 | 59.82 | 59.86 | 0.21 | 0.18 | 0.08 | 59.35, 60.09 |
| | | $\mu_2$ | 70.00 | 69.43 | 69.83 | 0.83 | 0.57 | 1.02 | 67.28, 70.26 |
| | | $\mu_3$ | 80.00 | 80.20 | 80.24 | 0.32 | -0.20 | 0.14 | 79.32, 80.64 |
| | | $\sigma_1$ | 2.00 | 1.86 | 1.88 | 0.13 | 0.14 | 0.04 | 1.62, 2.01 |
| | | $\sigma_2$ | 1.50 | 1.59 | 1.67 | 1.01 | -0.09 | 1.03 | 1.36, 3.51 |
| | | $\sigma_3$ | 3.00 | 2.84 | 2.81 | 0.24 | 0.16 | 0.08 | 2.54, 3.46 |
| | Individual | $\mu_1$ | 60.00 | 59.86 | 59.86 | 0.15 | 0.14 | 0.04 | 59.57, 60.12 |
| | | $\mu_2$ | 70.00 | 69.59 | 69.77 | 0.88 | 0.41 | 0.94 | 67.23, 71.03 |
| | | $\mu_3$ | 80.00 | 80.21 | 80.17 | 0.52 | -0.21 | 0.31 | 79.28, 81.12 |
| | | $\sigma_1$ | 2.00 | 1.88 | 1.87 | 0.15 | 0.12 | 0.04 | 1.52, 2.07 |
| | | $\sigma_2$ | 1.50 | 1.61 | 1.69 | 1.06 | -0.11 | 1.13 | 1.18, 3.65 |
| | | $\sigma_3$ | 3.00 | 2.86 | 2.83 | 0.42 | 0.14 | 0.19 | 2.41, 3.71 |

overlapping that occurs among the distribution of this observable values for non-observable state 2 and the other two non-observable states. In spite of these details, all parameter estimates can be considered satisfactory.

When comparing parameter estimates produced by the two proposals, for all chain lengths in both $K = 2$ and $K = 3$, we can observe that the difference in performance between the two algorithms is marginal. It seems that in the case of the parameters related to non-observable state 2, the Individual LASSO is obtaining marginally better results in all chain lengths, except

$T = 1200$. However, these differences are small, and we can conclude that both proposals have obtained satisfactory parameter estimates.

Table 14 – Mean squared predictive error results for the $K = 2$.

| $T$ | **Algo.** | **Median** | **Mean** | **SD** |
|---|---|---|---|---|
| 400 | Global | 26.48 | 27.46 | 8.86 |
| | Individual | 25.20 | 25.86 | 8.63 |
| | ARIMA | 27.78 | 28.68 | 8.16 |
| | Pen. Linear Reg. | 27.65 | 28.11 | 7.24 |
| 600 | Global | 22.15 | 23.83 | 9.40 |
| | Individual | 20.30 | 22.45 | 8.77 |
| | ARIMA | 23.54 | 25.49 | 6.94 |
| | Pen. Linear Reg. | 23.00 | 25.05 | 6.74 |
| 800 | Global | 20.99 | 22.62 | 6.81 |
| | Individual | 21.81 | 22.93 | 7.35 |
| | ARIMA | 25.14 | 26.12 | 5.76 |
| | Pen. Linear Reg. | 25.13 | 25.70 | 5.50 |

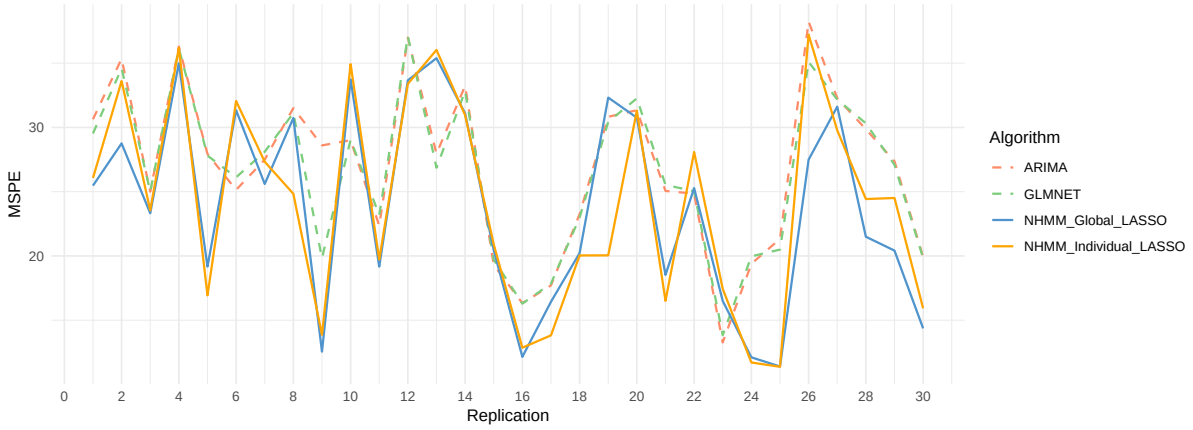
Table 14 shows the results for the MSPE on the test data set for the two proposals as well as the two methods selected for comparison when $K = 2$. As can be perceived, the proposed algorithms display better performance when predicting values of the test data set, consistently obtaining lower values for the median and mean of the MSPEs collected over all chain lengths. In general, this is an indicator that the two proposals perform prediction better when compared to penalized Linear Regression and ARIMA. Similar results are observed in Table 15 when $K = 3$, where the two proposals also consistently obtain lower values of the MSPE for all chain lengths.

To promote thorough understanding of the predictive performance of the algorithms, we plot the MSPE for the 50 replications when $K = 2$ and for the 30 replications when $K = 3$ in order to have a visual perspective of the behavior of this indicator. The plotted chain lengths were selected randomly; the selected chain length for $K = 2$ is $T = 800$ and for $K = 3$ is $T = 1000$.

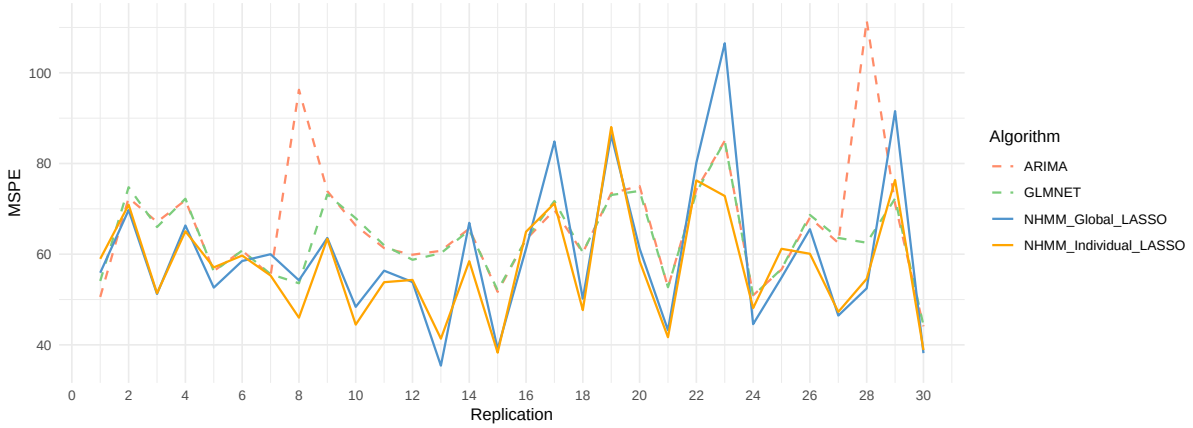
Table 15 – Mean squared predictive error results for the $K = 3$.

| $T$ | **Algo.** | **Median** | **Mean** | **SD** |
|---|---|---|---|---|
| 800 | Global | 55.30 | 55.22 | 10.68 |
| | Individual | 56.25 | 55.57 | 11.63 |
| | ARIMA | 60.75 | 61.09 | 8.90 |
| | Pen. Linear Reg. | 61.50 | 61.25 | 8.61 |
| 1000 | Global | 56.13 | 59.98 | 16.48 |
| | Individual | 57.76 | 57.53 | 12.20 |
| | ARIMA | 64.44 | 66.29 | 13.82 |
| | Pen. Linear Reg. | 62.90 | 63.42 | 9.07 |
| 1200 | Global | 61.55 | 60.35 | 12.98 |
| | Individual | 59.27 | 58.80 | 12.69 |
| | ARIMA | 63.71 | 64.86 | 9.06 |
| | Pen. Linear Reg. | 62.68 | 63.63 | 7.75 |

Figure 14 shows the MSPE for the proposals as well as the two algorithms which we are comparing. The graph shows that there are a small number of replications in which our proposals have a higher MSPE, it is evident that for the majority of replications, the LASSOs for NHMMs display better performance in terms of MSPE. To be precise, when $K = 2$, the Individual LASSO has obtained lower MSPEs than ARIMA in 70% of replications and lower than Penalized Linear Regression in 76% of replications. The Global LASSO has obtained lower MSPEs than both ARIMA in 73% of replications and lower than Penalized Linear Regression in 70% of replications. When $K = 3$, the Individual LASSO has obtained lower MSPEs than ARIMA in 76% of replications and lower than Penalized Linear Regression in 80% of replications. The Global LASSO has obtained lower MSPEs than ARIMA in 73% of replications and lower than Penalized Linear Regression in 70% of replications. This is evidence that the two proposals, in general, have better predictive performance than the two algorithms chosen for comparison.




(a) $K = 2$




(b) $K = 3$

Figure 14 – Values of the MSPE along all replications for Scenario 2.

Table 16 and Table 18 show results for the amount of hits when predicting **S** test sequence.

As a reminder to the reader, given the fact that the ARIMA and Penalized Linear Regression algorithms do not perform classification, we will only be comparing results for the Global and Individual LASSO. The tables show that results for hit frequency are very satisfactory for all chain lengths. In general, we see that differences in performance between the two algorithms are slight. An interesting point is that the Individual LASSO seems to marginally perform better for smaller chain lengths. However, the magnitude of the standard deviations apparently do not allow us to conclude that there is statistical difference between the performance of the two proposals. In spite of this, we can mention that results have been very good for this particular set of metrics.

Table 16 – Results for the hit frequency when predicting **S** test sequence for $K = 2$.

| $T$ | **Algo.** | **Median** | **Mean** | **SD** |
|---|---|---|---|---|
| 400 | Global | 0.72 | 0.72 | 0.12 |
| | Individual | 0.75 | 0.72 | 0.13 |
| 600 | Global | 0.77 | 0.76 | 0.15 |
| | Individual | 0.80 | 0.78 | 0.12 |
| 800 | Global | 0.79 | 0.77 | 0.08 |
| | Individual | 0.78 | 0.77 | 0.08 |

As was done in the previous scenario, we illustrate how prediction of the **S** test sequence is carried out using a confusion matrix. For Scenario 2, we select replication 14 for $T = 800$, for the Individual LASSO. The results are shown in Table 17.

Table 17 – Confusion matrix for real vs. predicted values of S test sequence in replication 14 when $T = 800$ for the Individual LASSO.

| | | ***Predicted Value*** | | |
|---|---|---|---|---|
| | | **1** | **2** | **3** |
| ***Real Value*** | **1** | 17 | 2 | 2 |
| | **2** | 3 | 4 | 1 |
| | **3** | 2 | 1 | 8 |

In general, the results for the confusion matrix show good accuracy in predicting the **S** test sequence. We can perceive that non-observable state 2 presents the lowest accuracy, correctly identifying non-observable state 2 in 57% of attempts (4 hits out of 7 attempts). Non-observable state 1 presents an accuracy of 77.2% (17 hits in 22 attempts). The general accuracy for this specific replication is 72.5%, a reasonably good performance for this metric, considering $K = 3$.

The next set of metrics which we analyze are metrics related to shrinkage of the transition coefficients. For Scenario 2 the we purposefully included transition coefficients in each $\boldsymbol{\beta}_{ij}$ whose real absolute value is no greater than 0.3. This scenario was designed specifically to test if the proposed algorithms are able to identify covariates whose coefficient may be close to zero, but not have an actual value of zero.

Table 18 – Results for the hit frequency when predicting **S** test sequence for $K = 3$.

| $T$ | **Algo.** | **Median** | **Mean** | **SD** |
|---|---|---|---|---|
| 800 | Global | 0.70 | 0.70 | 0.07 |
| | Individual | 0.70 | 0.71 | 0.09 |
| 1000 | Global | 0.68 | 0.68 | 0.08 |
| | Individual | 0.70 | 0.69 | 0.07 |
| 1200 | Global | 0.69 | 0.70 | 0.08 |
| | Individual | 0.68 | 0.68 | 0.09 |

From Table 19 we perceive that values of specificity are somewhat lower than in Scenario 1. This is indicating that some of the coefficients whose real value is non-zero are being shrunken past the established threshold. This is most likely due to the inclusion of coefficients whose value is very close to the threshold. In spite of this fact, the performance is still very good across all metrics related to shrinkage. Specificity is rather consistent for both algorithms across all sample sizes. The median and mean values are approximately 0.86 for both proposed algorithms in all chain lengths, which indicates that despite the inclusion of coefficients whose actual value is relatively close to 0, the proposals still perform well when identifying coefficients whose actual value is non-zero. The standard deviations are relatively low, indicating low variability in performance across all replications.

Regarding sensitivity, we observe high median values across all sample sizes, with the Global LASSO obtaining an excellent median performance in greater chain lengths. We can also perceive that the Global LASSO has considerably greater standard deviations. This may be related to the fact that there is only one $\lambda$ parameter which penalizes the coefficients for the regression corresponding to all non-observable states, possibly making the Global LASSO less precise when performing the aforementioned penalization. However, when we analyze the mean values for sensitivity we can perceive that there is a slight decrease in values. This fact along with the high values of standard deviation indicates that at least one replication obtained relatively lower performance in this metric.

The results for accuracy show us that the Global algorithm has higher variability in performance, especially with smaller chain lengths. However, as the chain length increases, the performance of the proposals improves for the largest chain size. This suggests that the Global algorithm benefits significantly more from larger chain lengths, reducing coefficient misidentification. The Individual LASSO also displays good performance across all sample sizes. However, it is important to mention the considerably lower variability shown by the Individual LASSO. This consistency in the accuracy suggests that the Individual LASSO is more robust and performs well even with smaller chain sizes. In general, both algorithms display good performance in this set of metrics when $K = 2$.

Table 20 exhibits the results for shrinkage of the transition coefficients when $K = 3$. The results in the table are leading to similar conclusions that were reached when analyzing the results

Table 19 – Results for shrinkage of the transition coefficients for $K = 2$.

| $T$ | Algo. | Specificity | | | Sensitivity | | | Accuracy | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | *Median* | *Mean* | *SD* | *Median* | *Mean* | *SD* | *Median* | *Mean* | *SD* |
| 400 | Global | 0.86 | 0.85 | 0.11 | 0.92 | 0.79 | 0.28 | 0.90 | 0.81 | 0.16 |
| | Individual | 0.86 | 0.85 | 0.09 | 0.92 | 0.87 | 0.14 | 0.90 | 0.86 | 0.08 |
| 600 | Global | 0.86 | 0.86 | 0.07 | 1.00 | 0.82 | 0.27 | 0.90 | 0.83 | 0.16 |
| | Individual | 0.86 | 0.86 | 0.11 | 0.85 | 0.85 | 0.16 | 0.85 | 0.85 | 0.09 |
| 800 | Global | 0.87 | 0.86 | 0.06 | 1.00 | 0.93 | 0.14 | 0.95 | 0.91 | 0.08 |
| | Individual | 0.86 | 0.88 | 0.06 | 0.92 | 0.87 | 0.12 | 0.90 | 0.88 | 0.08 |

in Table 19. Just as was observed in the Table 19, the values for all metrics are lower for this scenario than in previous scenarios, once again supporting the statement that the proposals can have some difficulties in identifying coefficients whose actual value is very close to 0. Another interesting observation is that the standard deviations of the 3 metrics are considerably greater for the Global LASSO. This is a valid statement for all chain lengths. Even though the means for all the metrics are relatively similar, the previous fact may indicate that the Global LASSO produced relatively lower results in at least one replication.

When analyzing specificity, it seems that the performance of the Individual LASSO improves slightly as chain lengths increase, while maintaining smaller values of standard deviation. In terms of sensitivity, the Global LASSO consistently displays better median values, however, when analyzing the mean values for sensitivity it is clear that Individual LASSO shows better performance in all chain lengths except $T = 1200$. As a matter of fact, even though the Global LASSO shows greater variability in all chain lengths, its mean value is greater than that obtained by the Individual LASSO when $T = 1200$. This may be related to the magnitude of the penalization and the fact there is only one penalization parameter which shrinks the coefficients of all regressions in the Global LASSO.

Table 20 – Results for shrinkage of the transition coefficients for $K = 3$.

| $T$ | Algo. | Specificity | | | Sensitivity | | | Accuracy | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | *Median* | *Mean* | *SD* | *Median* | *Mean* | *SD* | *Median* | *Mean* | *SD* |
| 800 | Global | 0.75 | 0.75 | 0.14 | 0.92 | 0.81 | 0.24 | 0.78 | 0.78 | 0.08 |
| | Individual | 0.71 | 0.74 | 0.09 | 0.85 | 0.83 | 0.10 | 0.79 | 0.78 | 0.06 |
| 1000 | Global | 0.79 | 0.79 | 0.11 | 0.83 | 0.73 | 0.26 | 0.79 | 0.76 | 0.10 |
| | Individual | 0.79 | 0.79 | 0.08 | 0.79 | 0.76 | 0.12 | 0.77 | 0.77 | 0.06 |
| 1200 | Global | 0.71 | 0.74 | 0.12 | 0.94 | 0.82 | 0.24 | 0.81 | 0.78 | 0.07 |
| | Individual | 0.79 | 0.77 | 0.08 | 0.79 | 0.79 | 0.12 | 0.79 | 0.78 | 0.05 |

This means that, in general, it is likely that the magnitude of penalization parameter chosen over the replications was large enough to shrink many of the transition coefficients to 0, leading to a greater sensitivity. However, this has the opposite effect on specificity. At first glance, it might seem that the magnitude of the penalization selected by the Global LASSO could have slightly over-penalized transition coefficients whose actual value is non-zero, therefore

lowering the specificity or, at the very least, resulting in greater variability (at least one replication may have performance which is considerably lower than the Individual LASSO). This would indicate that for a situation in which there are parameters whose real value is closer to 0, the Individual LASSO will carry out a finer penalization, maintaining a balance between specificity and sensitivity.

Table 21 shows the percentage of replications in which each non-zero coefficient whose actual value is in the vicinity of 0 is identified as actually being non-zero. In general, it appears that the Global LASSO is more precise when performing penalization and selecting non-zero coefficients. It has displayed better performance when identifying and correctly shrinking coefficients whose value is close to 0, in four out of six of the coefficients. This would indicate that individual penalization of the regressions for the non-observable states is actually allowing over-penalization of coefficients whose actual value is non-zero but close to the zero-threshold. Curiously, the Individual LASSO performed better when identifying coefficients $\beta_{328}$ and $\beta_{335}$. Both of these coefficients are pertaining to non-observable state 3. This leads us to think that individual penalization was indeed effective in performing finer tuning, at least to some extent, given that for that particular non-observable state, the Global LASSO over-penalized the coefficients.

In general, both of the proposals continue to display satisfactory performance when shrinking transition coefficients of the regressions, but the Individual LASSO shows a slight advantage maintaining a balance of the metrics, most likely due to the fact that penalization is carried out individually for the regressions of each non-observable state. Regarding non-zero coefficients whose actual value is closer to the zero-threshold, the Global LASSO correctly identified non-zero coefficients in 63 out of 180 possible attempts (30 replications for 6 coefficients) which represents 35.00% of the attempts while the Individual LASSO correctly identified non-zero coefficients whose value is close to 0 in 59 out of 180 attempts, representing 32.78% correct identification rate. While these metric values may not seem impressive, we believe that given the difficulty of identifying coefficients which are very close to the zero-threshold, the proposals still retain reasonable performance.

Table 21 – Percentage of correct shrinkage of coefficients whose real value is close to 0 when $K = 3$.

| **Coefficient** | **Real Value** | **Global LASSO** | **Individual LASSO** |
|---|---|---|---|
| $\beta_{125}$ | 0.20 | 43.33% | 36.67% |
| $\beta_{132}$ | 0.10 | 23.33% | 13.33% |
| $\beta_{224}$ | 0.15 | 36.67% | 30.00% |
| $\beta_{237}$ | 0.25 | 30.00% | 23.33% |
| $\beta_{328}$ | 0.30 | 33.33% | 40.00% |
| $\beta_{335}$ | 0.30 | 43.33% | 53.33% |

We also observe notable differences in shrinkage performance when comparing results shown in Tables 19 and 20 with results obtained in the first round of simulations found in

Appendix A. In the case of $K = 2$, we compare the results of Table 19 with results in any tables corresponding to $K = 2$ such as Tables 45, 46, 54 and 55, which are tables that show results using data that was simulated with similar parameter configurations. In the comparison, we perceive considerable differences in shrinkage performance, particularly in terms of sensitivity. For all cases, the difference in performance is at least 25% for the Global LASSO and close to least 30% for the Individual LASSO. This shows the great improvement in performance through out the development of this research project. When comparing results for $K = 3$, this improvement is even more evident. Table 20 shows differences of up to 60% when comparing with results shown in Tables 63 and 64 found in Appendix A. These two tables display results with slightly similar parameter configurations in the simulated data. This differences are observable in the sensitivity metric.

Table 22 – Processing times (in minutes) for $K = 2$, where total time considers running the $R = 50$ replications.

| $T$ | Algo. | Total Time | Avg. Time/Rep | Median Time/Rep | SD Time/Rep |
|---|---|---|---|---|---|
| 400 | Global | 32.09 | 0.64 | 0.63 | 0.03 |
| | Individual | 701.92 | 14.04 | 14.37 | 1.13 |
| 600 | Global | 43.67 | 0.87 | 0.87 | 0.04 |
| | Individual | 967.63 | 19.35 | 19.31 | 0.55 |
| 800 | Global | 57.04 | 1.14 | 1.14 | 0.06 |
| | Individual | 1,220.30 | 24.41 | 24.35 | 0.74 |

Tables 22 and 23 show observed processing times for Scenario 3. As we have observed in all previous scenarios and as is expected, the Global LASSO has a smaller processing times for all chain lengths. Once again, as we have emphasized in previous scenarios, we can acknowledge the increasing computational costs of the Individual LASSO as the amount of non-observable states, $K$, increases.

Table 23 – Processing times (in minutes) for $K = 3$, where total time considers running the $R = 30$ replications.

| $T$ | Algo. | Total Time | Avg. Time/Rep | Median Time/Rep | SD Time/Rep |
|---|---|---|---|---|---|
| 800 | Global | 86.33 | 2.88 | 2.91 | 0.17 |
| | Individual | 5,832.20 | 194.41 | 196.21 | 9.33 |
| 1000 | Global | 102.68 | 3.42 | 3.43 | 0.12 |
| | Individual | 7,000.72 | 233.36 | 233.04 | 8.43 |
| 1200 | Global | 117.69 | 3.92 | 3.92 | 0.15 |
| | Individual | 7,977.81 | 265.93 | 264.87 | 8.76 |

### 4.4.3 Scenario 3

Scenario 3 is designed to test the performance of the proposals under less favorable conditions. Specifically, for this scenario, we have purposefully simulates data with a greater percentage of overlapping observations. This should present a considerable challenge for both

estimation and prediction, as well as the other metrics which are being analyzed. An additional difficulty introduced in this scenario is the presence of more coefficients in every regression as will be shown in the description for the parameter configuration. For this scenario, the real values of the parameters for the probability distribution of the observable random variables have been set as follows:

- For $K = 2$ the parameters are set to $\mu_1 = 60$, $\sigma_1 = 3.5$, and $\mu_2 = 70$, $\sigma_2 = 2.8$; and

- For $K = 3$ the parameters are set to $\mu_1 = 60$, $\sigma_1 = 3.5$, $\mu_2 = 70$, $\sigma_2 = 2.8$, and $\mu_3 = 80$, $\sigma_3 = 4$.

Scenario 3 includes the use of a greater number of coefficients to simulate the data sets. We use two values for the amount of covariates in the model. When $K = 2$ the value of $D$ is set to 20, and when $K = 3$ the value of $D = 15$. This parameter configuration is designed to test if the proposed algorithms can perform shrinkage of a greater number of transition coefficients whose real value is zero. The real values for the transition coefficients used in this simulation scenario are shown in the following matrix equations. As has been customary so far, only non-zero coefficients are displayed in each $\boldsymbol{\beta}_{ij}$, however all $\boldsymbol{\beta}_{ij}$ vectors contain 20 coefficients when $K = 2$ and 15 coefficients when $K = 3$. The rest of the coefficients in each entry are set to 0. Equation (4.17) shows the transition coefficients used when $K = 2$, given by

$$\boldsymbol{\beta} = \begin{bmatrix} \begin{pmatrix} \beta_{111} = 0.0 \\ \beta_{112} = 0.0 \\ \beta_{113} = 0.0 \end{pmatrix} & \begin{pmatrix} \beta_{121} = -1.5 \\ \beta_{122} = -1.5 \\ \beta_{123} = -2.6 \end{pmatrix} \\ \\ \begin{pmatrix} \beta_{211} = 0.0 \\ \beta_{212} = 0.0 \\ \beta_{213} = 0.0 \end{pmatrix} & \begin{pmatrix} \beta_{221} = -2.0 \\ \beta_{222} = 2.6 \\ \beta_{223} = 1.4 \end{pmatrix} \end{bmatrix}. \tag{4.17}$$

Observe that the non-zero transition coefficients are not the same for different non-obsersable states and are not in sequence anymore.

Equation (4.18) shows transition coefficients used to simulate data when $K = 3$, given

by

$$\boldsymbol{\beta} = \begin{bmatrix} \begin{pmatrix} \beta_{111} = 0.0 \\ \beta_{112} = 0.0 \\ \beta_{113} = 0.0 \end{pmatrix} & \begin{pmatrix} \beta_{121} = -1.5 \\ \beta_{122} = -1.5 \\ \beta_{123} = -2.6 \end{pmatrix} & \begin{pmatrix} \beta_{131} = -1.3 \\ \beta_{132} = -3.2 \\ \beta_{133} = -2.4 \end{pmatrix} \\ \begin{pmatrix} \beta_{211} = 0.0 \\ \beta_{212} = 0.0 \\ \beta_{213} = 0.0 \end{pmatrix} & \begin{pmatrix} \beta_{221} = -2.0 \\ \beta_{222} = 2.6 \\ \beta_{223} = 1.4 \end{pmatrix} & \begin{pmatrix} \beta_{231} = -1.3 \\ \beta_{232} = 1.7 \\ \beta_{233} = 1.3 \end{pmatrix} \\ \begin{pmatrix} \beta_{311} = 0.0 \\ \beta_{312} = 0.0 \\ \beta_{313} = 0.0 \end{pmatrix} & \begin{pmatrix} \beta_{321} = 2.4 \\ \beta_{322} = 2.1 \\ \beta_{323} = -1.5 \end{pmatrix} & \begin{pmatrix} \beta_{331} = -1.3 \\ \beta_{332} = -2.7 \\ \beta_{333} = -2.5 \end{pmatrix} \end{bmatrix}. \tag{4.18}$$

As mentioned previously, we only display 3 entries in each coefficient vector to improve readability and space.

We must once again stress that the presence of more coefficients than previous scenarios, and specially the greater percentage of overlapping observations classify this scenario as challenging for the proposals to perform estimation and prediction. Regarding the tuning parameter grid, the chosen set of values of the $\lambda$ penalization parameter for the Global LASSO and Individual Lasso is equal to those considered in the previous scenarios.

To further illustrate the nature of the simulated data for this scenario, and to display the additional difficulty implied in the overlapping distributions of the observable values, we select a specific replication for a specific chain length, and we plot the distribution of the simulated observable values along with the corresponding non-observable state to which they belong to. As was the case in previous scenarios, this will aid in having a visual perception and understanding of overlapping in this particular scenario. For $K = 2$, we have randomly selected replication 24 for $T = 600$. For $K = 3$, we have randomly selected replication 13 for $T = 1000$.

From Figure 15 we perceive that there is considerable overlapping of the distributions of the observable values. This is a factor of great importance when judging the difficulty which the scenario presents for the proposed algorithms to perform estimation. Therefore, we have deemed this scenario as presenting a challenging situation for the proposed algorithms.

Regarding estimation of the parameters of the distribution of the observable random variables, we begin to perceive some of the difficulties encountered by the proposals. Table 24 shows the results for estimation of the previously mentioned set of parameters when $K = 2$. We observe some confidence intervals which are considerably larger than observed in previous scenarios. It is also evident that the algorithms encountered difficulties when estimating the variance parameters. This is consistent among all chain lengths. It seems that overlapping of the distributions may have caused some of the observable values to be incorrectly classified into a non-observable state to which they presumably do not belong, therefore leading to estimates

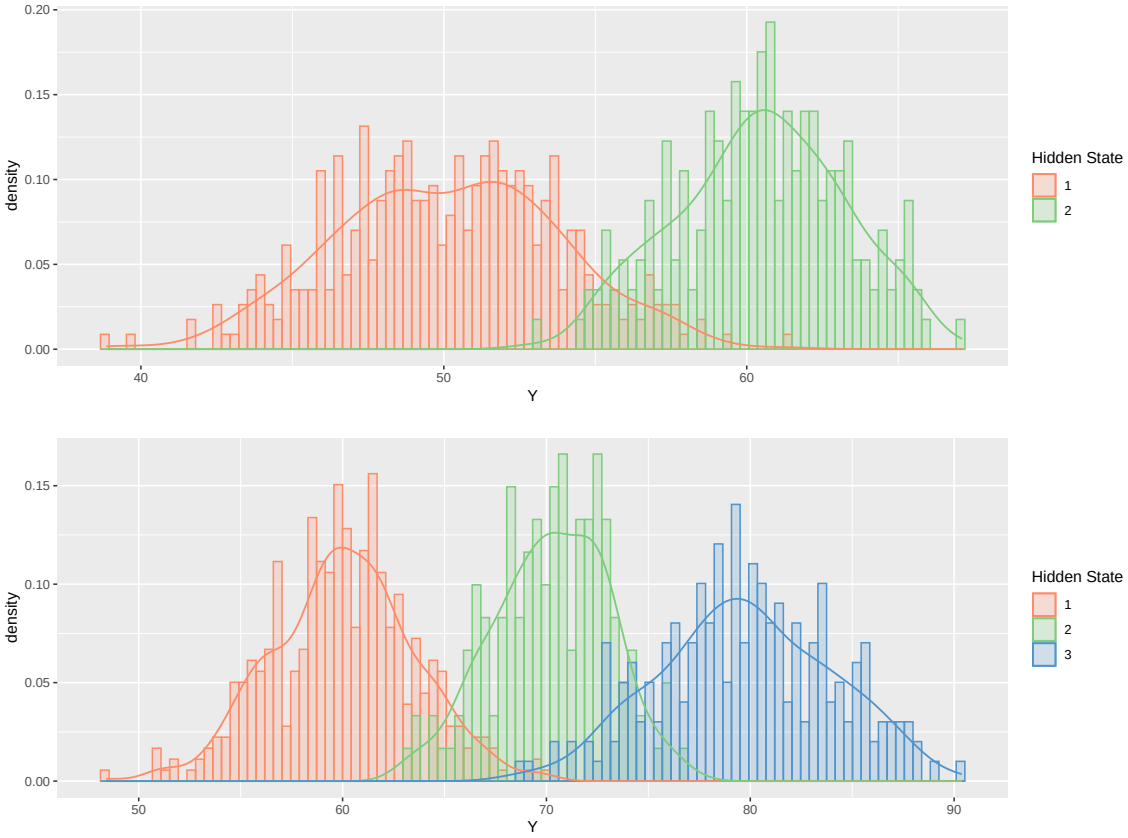



Figure 15 – Distribution of the observable values for $K = 2$ and $K = 3$.

of this set of parameters which seem to be relatively more distant from the actual value than in previous scenarios. For the Individual LASSO, some 95% confidence interval do not contain the real value of the parameter. Such is the case of $\sigma_1$ when $T = 600$ and $T = 800$, as well as $\mu_2$ when $T = 800$. For this scenario, it is clearly evident that the degree of overlapping is impacting the performance of the proposals.

It is important to mention that in spite of this considerable difficulty, the estimates remain at the very least, reasonable. We observe mean and median values which are still relatively close to the real value of the parameters. Standard deviations are still relatively small, with the exception of parameters related to the second non-observable state. For $K = 2$, we state that estimates remain relatively fair.

Table 25 shows estimation results for the parameters of the random variables for the observable values. We see similar results as observed for $K = 2$. Several of the parameter confidence intervals do not contain the real value for the parameter being estimated. This happens for the Global LASSO when $T = 800$ for the $\mu_1$ and $\sigma_1$ parameters. For the Individual LASSO, this happens for $\mu_1$ when $T = 1000$. In the case of parameter $\mu_2$, we also observe some relatively large confidence intervals for both proposed algorithms in all chain lengths. The standard deviation and bias of $\mu_2$ is also relatively large when compared to that of other parameters. As a consequence, we observe greater values of the mean squared error for this

Table 24 – Estimation results for the parameters of the observable distributions for $K = 2$.

| $T$ | Algo. | Param. | Real | Mean | Median | SD | Bias | MSE | 95%CI |
|---|---|---|---|---|---|---|---|---|---|
| 400 | Global | $\mu_1$ | 60.00 | 59.59 | 59.38 | 0.71 | 0.41 | 0.67 | 58.80, 61.43 |
| | | $\mu_2$ | 70.00 | 69.28 | 69.38 | 0.84 | 0.72 | 1.21 | 67.86, 70.43 |
| | | $\sigma_1$ | 3.50 | 3.21 | 3.10 | 0.66 | 0.29 | 0.52 | 2.43, 5.04 |
| | | $\sigma_2$ | 2.80 | 3.14 | 3.00 | 0.66 | -0.34 | 0.55 | 2.50, 4.81 |
| | Individual | $\mu_1$ | 60.00 | 59.37 | 59.38 | 0.43 | 0.63 | 0.58 | 58.57, 60.17 |
| | | $\mu_2$ | 70.00 | 69.20 | 69.29 | 0.71 | 0.80 | 1.14 | 67.52, 70.20 |
| | | $\sigma_1$ | 3.50 | 2.95 | 2.96 | 0.31 | 0.55 | 0.40 | 2.42, 3.53 |
| | | $\sigma_2$ | 2.80 | 3.10 | 3.06 | 0.59 | -0.30 | 0.44 | 2.51, 4.60 |
| 600 | Global | $\mu_1$ | 60.00 | 59.40 | 59.34 | 0.50 | 0.60 | 0.61 | 58.55, 60.33 |
| | | $\mu_2$ | 70.00 | 69.32 | 69.36 | 0.57 | 0.68 | 0.79 | 67.97, 70.24 |
| | | $\sigma_1$ | 3.50 | 3.07 | 3.07 | 0.37 | 0.43 | 0.32 | 2.49, 3.66 |
| | | $\sigma_2$ | 2.80 | 3.07 | 3.04 | 0.23 | -0.27 | 0.12 | 2.66, 3.53 |
| | Individual | $\mu_1$ | 60.00 | 59.47 | 59.45 | 0.36 | 0.53 | 0.41 | 58.92, 60.20 |
| | | $\mu_2$ | 70.00 | 69.37 | 69.41 | 0.77 | 0.63 | 1.00 | 67.81, 70.25 |
| | | $\sigma_1$ | 3.50 | 3.02 | 3.04 | 0.26 | 0.48 | 0.30 | 2.52, 3.38 |
| | | $\sigma_2$ | 2.80 | 2.98 | 2.84 | 0.60 | -0.18 | 0.39 | 2.58, 3.48 |
| 800 | Global | $\mu_1$ | 60.00 | 59.46 | 59.41 | 0.37 | 0.54 | 0.42 | 58.89, 60.06 |
| | | $\mu_2$ | 70.00 | 69.36 | 69.37 | 0.60 | 0.64 | 0.77 | 68.38, 70.25 |
| | | $\sigma_1$ | 3.50 | 3.06 | 3.03 | 0.26 | 0.44 | 0.26 | 2.61, 3.62 |
| | | $\sigma_2$ | 2.80 | 3.04 | 2.98 | 0.43 | -0.24 | 0.24 | 2.58, 3.63 |
| | Individual | $\mu_1$ | 60.00 | 59.50 | 59.47 | 0.43 | 0.50 | 0.43 | 58.99, 60.61 |
| | | $\mu_2$ | 70.00 | 68.89 | 69.38 | 1.46 | 1.11 | 3.38 | 64.62, 69.86 |
| | | $\sigma_1$ | 3.50 | 2.84 | 2.99 | 0.46 | 0.66 | 0.65 | 1.61, 3.31 |
| | | $\sigma_2$ | 2.80 | 3.43 | 2.94 | 1.23 | -0.63 | 1.91 | 2.62, 6.62 |

parameter through out all simulations in Scenario 3. Another curious phenomenon which we observe is an apparent inversion of estimated values between the $\sigma_1$ and $\sigma_2$ parameters. This is particularly noticeable for both the Global and Individual LASSO when $T = 800$, and $T = 1000$. It seems to be slightly mitigated when $T = 1200$ for both algorithms. This might suggest that when the distributions of the random variables of the observable values have a higher degree of overlap, greater chain lengths will be necessary in order to perform estimation with good results.

The anomalies observed in estimation for both $K = 2$ and $K = 3$ are most likely due to overlapping of the distributions, as was previously mentioned. As we have stated in previous scenarios, overlapping establishes a difficulty for the algorithms to successfully identify the non-observable state to which each of the observations belong, specially in regions near the tails of the distributions, where it is evident that observations could belong to either of the non-observable states.

Table 26 displays results for the mean square predictive error when $K = 2$. As the table shows, in spite of the difficulties encountered in the estimation of the parameters shown in Tables 24 and 25, the proposed algorithms continue to consistently show better predictive performance than the comparison algorithms. For all chain lengths, the Global and Individual LASSO show

Table 25 – Estimation results for the parameters of the observable distributions for $K = 3$.

| $T$ | Algo. | Param. | Real | Mean | Median | SD | Bias | MSE | 95%CI |
|---|---|---|---|---|---|---|---|---|---|
| 800 | Global | $\mu_1$ | 60.00 | 59.13 | 59.23 | 0.46 | 0.87 | 0.98 | 58.10, 59.65 |
| | | $\mu_2$ | 70.00 | 68.40 | 68.89 | 1.44 | 1.60 | 4.65 | 64.45, 70.21 |
| | | $\mu_3$ | 80.00 | 80.37 | 80.45 | 0.71 | -0.37 | 0.64 | 78.73, 81.38 |
| | | $\sigma_1$ | 3.50 | 2.83 | 2.84 | 0.20 | 0.67 | 0.49 | 2.42, 3.12 |
| | | $\sigma_2$ | 2.80 | 3.26 | 3.24 | 0.50 | -0.46 | 0.46 | 2.36, 4.17 |
| | | $\sigma_3$ | 4.00 | 3.67 | 3.57 | 0.39 | 0.33 | 0.26 | 3.22, 4.64 |
| | Individual | $\mu_1$ | 60.00 | 59.26 | 59.41 | 0.64 | 0.74 | 0.96 | 57.74, 60.10 |
| | | $\mu_2$ | 70.00 | 68.51 | 68.97 | 2.04 | 1.49 | 6.38 | 63.22, 71.19 |
| | | $\mu_3$ | 80.00 | 80.43 | 80.51 | 0.99 | -0.43 | 1.17 | 78.42, 81.94 |
| | | $\sigma_1$ | 3.50 | 2.82 | 2.85 | 0.35 | 0.68 | 0.58 | 2.14, 3.59 |
| | | $\sigma_2$ | 2.80 | 3.45 | 3.35 | 0.86 | -0.65 | 1.17 | 1.96, 5.16 |
| | | $\sigma_3$ | 4.00 | 3.66 | 3.55 | 0.52 | 0.34 | 0.39 | 2.94, 4.77 |
| 1000 | Global | $\mu_1$ | 60.00 | 59.01 | 59.05 | 0.54 | 0.99 | 1.28 | 57.89, 60.86 |
| | | $\mu_2$ | 70.00 | 68.64 | 68.46 | 1.58 | 1.36 | 4.34 | 65.13, 70.69 |
| | | $\mu_3$ | 80.00 | 80.42 | 80.46 | 0.73 | -0.42 | 0.71 | 79.12, 81.64 |
| | | $\sigma_1$ | 3.50 | 2.84 | 2.78 | 0.29 | 0.66 | 0.52 | 2.44, 3.59 |
| | | $\sigma_2$ | 2.80 | 3.38 | 3.34 | 0.49 | -0.58 | 0.58 | 2.59, 4.15 |
| | | $\sigma_3$ | 4.00 | 3.59 | 3.59 | 0.44 | 0.41 | 0.36 | 2.78, 4.58 |
| | Individual | $\mu_1$ | 60.00 | 59.01 | 59.05 | 0.54 | 0.99 | 1.28 | 57.93, 59.90 |
| | | $\mu_2$ | 70.00 | 68.64 | 68.46 | 1.58 | 1.36 | 4.34 | 65.09, 70.82 |
| | | $\mu_3$ | 80.00 | 80.42 | 80.46 | 0.73 | -0.42 | 0.71 | 79.03, 81.67 |
| | | $\sigma_1$ | 3.50 | 2.84 | 2.78 | 0.29 | 0.66 | 0.52 | 2.41, 3.62 |
| | | $\sigma_2$ | 2.80 | 3.38 | 3.34 | 0.49 | -0.58 | 0.58 | 2.61, 4.11 |
| | | $\sigma_3$ | 4.00 | 3.59 | 3.59 | 0.44 | 0.41 | 0.36 | 2.79, 4.61 |
| 1200 | Global | $\mu_1$ | 60.00 | 59.33 | 59.26 | 0.42 | 0.67 | 0.63 | 58.49, 60.21 |
| | | $\mu_2$ | 70.00 | 68.75 | 68.92 | 1.94 | 1.25 | 5.35 | 65.33, 71.07 |
| | | $\mu_3$ | 80.00 | 80.44 | 80.39 | 0.78 | -0.44 | 0.80 | 78.71, 81.56 |
| | | $\sigma_1$ | 3.50 | 3.00 | 2.93 | 0.40 | 0.50 | 0.42 | 2.58, 3.77 |
| | | $\sigma_2$ | 2.80 | 3.17 | 3.16 | 0.45 | -0.37 | 0.34 | 2.41, 3.93 |
| | | $\sigma_3$ | 4.00 | 3.55 | 3.49 | 0.46 | 0.45 | 0.42 | 2.90, 4.64 |
| | Individual | $\mu_1$ | 60.00 | 59.49 | 59.53 | 0.42 | 0.51 | 0.44 | 58.57, 60.06 |
| | | $\mu_2$ | 70.00 | 69.08 | 68.92 | 1.94 | 0.92 | 4.61 | 65.19, 71.12 |
| | | $\mu_3$ | 80.00 | 80.36 | 80.39 | 0.78 | -0.36 | 0.77 | 78.74, 81.62 |
| | | $\sigma_1$ | 3.50 | 3.08 | 2.93 | 0.40 | 0.42 | 0.34 | 2.60, 3.86 |
| | | $\sigma_2$ | 2.80 | 3.11 | 3.16 | 0.45 | -0.31 | 0.30 | 2.39, 3.91 |
| | | $\sigma_3$ | 4.00 | 3.62 | 3.49 | 0.48 | 0.38 | 0.37 | 2.83, 4.61 |

lower mean and median values for the MSPE than the ARIMA and penalized linear regression algorithms. We notice that ARIMA and penalized linear regression consistently show lower values for the standard deviation, meaning that the results obtained are more concentrated around the mean values shown in the table than those for the proposal algorithms.

We observe similar results in Table 27, in which the proposals also obtain better mean and median performance than the comparison algorithms. For $K = 3$, the standard deviations show similar values for all tested algorithms, with the exception of ARIMA when $T = 800$. In the

case of $K = 3$, difference in performance between the proposed algorithms and the comparison algorithms become more apparent as chain size increases.

To promote a clear understanding of the predictive displayed performance and to have a detailed comparison between the proposed and comparison algorithms, we will once again plot the MSPE along all replications for a specific chain length for both $K = 2$ and $K = 3$.

Table 26 – Mean squared predictive error results for the $K = 2$.

| $T$ | **Algo.** | **Median** | **Mean** | **SD** |
|---|---|---|---|---|
| 400 | Global | 26.97 | 30.27 | 10.68 |
| | Individual | 26.44 | 27.87 | 11.84 |
| | ARIMA | 30.14 | 29.93 | 9.22 |
| | Pen. Linear Reg. | 29.13 | 30.52 | 8.85 |
| 600 | Global | 28.85 | 30.16 | 10.74 |
| | Individual | 27.70 | 29.44 | 9.30 |
| | ARIMA | 33.19 | 33.34 | 7.35 |
| | Pen. Linear Reg. | 31.81 | 32.02 | 6.78 |
| 800 | Global | 29.17 | 30.86 | 9.20 |
| | Individual | 29.92 | 29.81 | 8.19 |
| | ARIMA | 31.64 | 31.24 | 5.88 |
| | Pen. Linear Reg. | 31.10 | 31.32 | 6.12 |

Figure 16 shows the results for the MSPE for the proposed algorithms as well as the comparison algorithms. The chosen chain lengths for this figure are $T = 800$ for $K = 2$ and $T = 1000$ for $K = 3$. From the figure, we observe that for the majority of replications, the proposed algorithms display lower values of MSPE than the comparison algorithms. To be exact, when $K = 2$, the Global LASSO has lower values of MSPE than ARIMA in 60% of all replications, lower than penalized linear regression in 63% of all replications. The Individual LASSO has a lower MSPE than ARIMA in 70% of all replications and lower than penalized linear regression in 63% of all replications. When comparing the two algorithms, we observe that Individual LASSO has a lower MSPE than the Global LASSO in 60% of all replications.

When we calculate these percentages for $K = 3$, the differences in performance become more apparent. The Global LASSO has lower values of MSPE for both ARIMA and penalized linear regression in 83% of all replications. The Individual LASSO has lower values of MSPE than ARIMA in 80% of all replications, while having lower values than penalized linear regression in 70% of all replications.

While we cannot state that there are significant differences between the performance of the algorithms, the previously mentioned percentages show that the proposed algorithms, in fact, perform better than the comparison algorithms in this scenario.

To further illustrate performance of the algorithms, we select a single replication and plot the predicted values for the test data set. To contrast Scenario 1 in which the replication was selected visually based on displaying bad performance, and Scenario 2 in which the replication

Table 27 – Mean squared predictive error results for the $K = 3$.

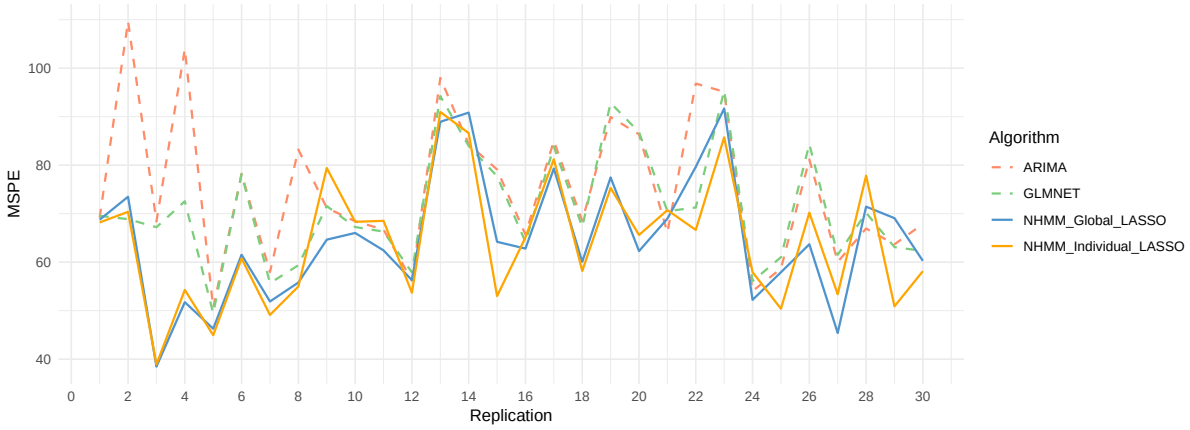
| $T$ | Algo. | Median | Mean | SD |
|---|---|---|---|---|
| 800 | Global | 70.77 | 69.36 | 12.61 |
| | Individual | 71.44 | 71.03 | 15.03 |
| | ARIMA | 72.42 | 77.35 | 26.13 |
| | Pen. Linear Reg. | 72.31 | 73.52 | 12.66 |
| 1000 | Global | 63.24 | 64.78 | 13.03 |
| | Individual | 64.34 | 64.31 | 13.07 |
| | ARIMA | 68.83 | 75.04 | 15.48 |
| | Pen. Linear Reg. | 69.22 | 71.01 | 11.96 |
| 1200 | Global | 68.98 | 68.90 | 11.04 |
| | Individual | 68.58 | 69.12 | 11.45 |
| | ARIMA | 77.66 | 76.31 | 12.94 |
| | Pen. Linear Reg. | 74.67 | 73.65 | 12.06 |

was chosen randomly, we now select a replication in which the proposals seem to obtain better performance than the comparison algorithms. We select replication 3 for $T = 1000$ when $K = 3$. Figure 17 shows the predicted values for the 4 algorithms as well as the real values of the test set. As can be perceived, the proposed algorithms capture the behavior of the values of the test data set better than the comparison algorithms, showing different intervals in which that behavior is modeled very accurately. There are some peaks which all algorithms have difficulty in capturing, but it seems that the proposal algorithms still manage to retain the general behavior of the real test values.

Regarding hit frequency, Tables 28 and 29 show the results obtain by the proposed algorithms when predicting the **S** test sequence which simulates the observable values. We perceive that there is a slight drop in performance in this metric when compared to the previous scenarios. This is evident in the median values of hit frequency when $K = 2$ and the mean values when $K = 3$. This phenomenon is most likely due to the fact that Scenario 3 purposefully introduces greater overlapping among the distributions of the random variables of the observable values in order to stress the proposed algorithms. When comparing to values of hit frequency for Scenario 2, we observe drops in performance of approximately 2% for several chain lengths. In the case of $K = 2$, this becomes more evident for longer chain lengths. In some cases, there is a drop of performance of 7%, (Individual LASSO in Scenario 2 scored 80%). For $T = 400$ and $K = 2$, the Individual LASSO displays a relatively large value of the standard deviation. This indicates that in at least one of the replications, it must have obtained relatively poor performance.

For $K = 3$, we observe that the proposed algorithms present similar results to those shown for $K = 2$. There is a perceivable decrease in performance when comparing results to those of previous scenarios. Once again, we attribute this slight decrease to the degree of overlapping among the distributions of the random variables of the observable values. The proposals also show relatively small values for standard deviation, which indicates consistency in the obtained results.


(a) $K = 2$




(b) $K = 3$

Figure 16 – Values of the MSPE along all replications for Scenario 3.

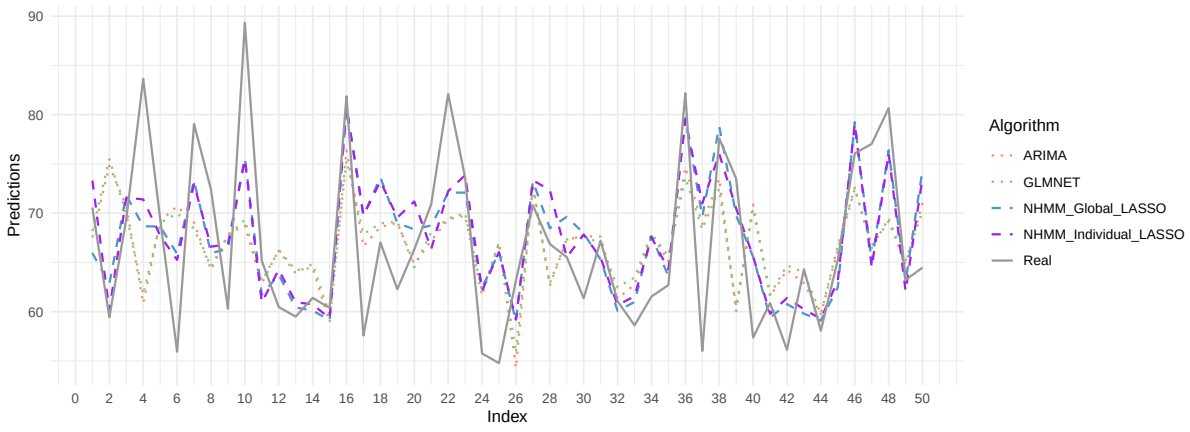



Figure 17 – Predicted values for the proposed and comparison algorithms for replication 3, when $K = 3$ and $T = 1000$.

Table 28 – Results for the hit frequency when predicting **S** test sequence for $K = 2$.

| $T$ | **Algo.** | **Median** | **Mean** | **SD** |
|---|---|---|---|---|
| 400 | Global | 0.73 | 0.73 | 0.14 |
| | Individual | 0.72 | 0.76 | 0.19 |
| 600 | Global | 0.74 | 0.75 | 0.12 |
| | Individual | 0.73 | 0.76 | 0.12 |
| 800 | Global | 0.74 | 0.76 | 0.11 |
| | Individual | 0.75 | 0.77 | 0.10 |

Table 29 – Results for the hit frequency when predicting **S** test sequence for $K = 3$.

| $T$ | **Algo.** | **Median** | **Mean** | **SD** |
|---|---|---|---|---|
| 800 | Global | 0.66 | 0.68 | 0.06 |
| | Individual | 0.67 | 0.67 | 0.07 |
| 1000 | Global | 0.67 | 0.67 | 0.07 |
| | Individual | 0.68 | 0.68 | 0.09 |
| 1200 | Global | 0.68 | 0.68 | 0.05 |
| | Individual | 0.68 | 0.67 | 0.06 |

However, in spite of these slight issues, the algorithms continue to perform classification of the observations reasonably well even under stressful conditions such as those presented in Scenario 3.

As was done in Scenarios 1 and 2, we illustrate how prediction of the **S** test sequence occurs using a confusion matrix. In this scenario, we randomly select replication 11 for $T = 1000$, for the Individual LASSO. The results are shown in Table 30.

The results shown in Table 30 illustrate the difficulty encountered by the algorithms due to the greater overlapping of the distributions of the random variables for the observable values. This is particularly evident regarding identification of non-observable state 2, in which we perceive an accuracy of 20%, (1 hit in 5 attempts), displaying the worst performance in this replication. This might be caused by the fact that non-observable state 2 overlaps with both non-observable state 1 and 3 at the tails of the distributions, therefore making proper prediction of non-observable state 2 more difficult. Non-observable states 1 and 3 present much greater accuracy, at 74.3% and 66%, respectively.

Table 30 – Confusion matrix for real vs. predicted values of **S** test sequence in replication 11 when $T = 1000$ for the Individual LASSO.

| | | ***Predicted Value*** | | |
|---|---|---|---|---|
| | | **1** | **2** | **3** |
| ***Real Value*** | **1** | 29 | 2 | 1 |
| | **2** | 5 | 1 | 1 |
| | **3** | 5 | 2 | 4 |

Table 31 displays results for transition coefficient shrinkage when $K = 2$ for Scenario

3. Once again, we perceive that shrinkage is performed successfully and the proposal obtain excellent results. For specificity, both the Global and Individual LASSO obtain mean values greater than 94% for all chain lengths. Median values show a 100% success rate for all chain lengths. This indicates that both proposals are being highly successful when attempting to identify coefficients whose actual value is non-zero. Regarding sensitivity, the proposals show very good performance, obtaining mean scores higher than 84% for all chain lengths. The performance shown in sensitivity and specificity is directly transmitted to the accuracy metric, where we observe excellent performance for both proposals in all chain lengths.

As was observed in previous scenarios, the standard deviation of sensitivity is considerably greater for the Global LASSO. As was previously discussed, this may be due to the fact that Global LASSO uses only one penalization parameter to penalize the coefficients of the regressions of all non-observable states. This may lead to have over-penalization (or under-penalization) for one of regression equation corresponding to one of the non-observable states in at least one of the replications, therefore producing greater variability in the metric.

Table 31 – Results for shrinkage of the transition coefficients for $K = 2$.

| $T$ | **Algo.** | **Specificity** | | | **Sensitivity** | | | **Accuracy** | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | *Median* | *Mean* | *SD* | *Median* | *Mean* | *SD* | *Median* | *Mean* | *SD* |
| 400 | Global | 1.00 | 0.95 | 0.08 | 0.97 | 0.88 | 0.18 | 0.98 | 0.89 | 0.15 |
| | Individual | 1.00 | 0.94 | 0.12 | 0.84 | 0.84 | 0.12 | 0.86 | 0.86 | 0.11 |
| 600 | Global | 1.00 | 0.97 | 0.07 | 1.00 | 0.85 | 0.25 | 0.98 | 0.87 | 0.21 |
| | Individual | 1.00 | 0.99 | 0.05 | 0.88 | 0.87 | 0.11 | 0.90 | 0.89 | 0.10 |
| 800 | Global | 1.00 | 0.99 | 0.03 | 0.97 | 0.87 | 0.20 | 0.98 | 0.88 | 0.17 |
| | Individual | 1.00 | 0.98 | 0.09 | 0.88 | 0.86 | 0.13 | 0.90 | 0.88 | 0.11 |

Table 32 shows results for shrinkage performance when $K = 3$ for Scenario 3. The proposals show very good performance for $K = 3$ in this scenario. We observe values for specificity and sensitivity that are very satisfactory. Identification of non-zero and zero transition coefficients is carried out successfully, attaining mean scores for specificity which are greater than 90% and scores for sensitivity which are greater than 80% for all chain lengths. As a result, the accuracy displayed by both proposals is excellent, consistently obtaining mean values greater than 82% for both proposals in all chain lengths. Finally, we must highlight the occurrence of greater variability in the sensitivity metric for the Global LASSO. The possible reasons for which this occurs have been highlighted in previous scenarios, as well as in the discussion for shrinkage results regarding $K = 2$ in this scenario.

As commented in previous scenarios, we perceive great differences in performance when comparing results found in Tables 31 and 32, to tables found in Appendix A. For all chain lengths and for both proposals, we perceive considerable improvement in terms of identifying coefficients whose actual value is 0 and shrinking them past the established zero threshold.

Our last set of metrics, shown in Tables 33 and 34, show processing times for the

Table 32 – Results for shrinkage of the transition coefficients for $K = 3$.

| $T$ | Algo. | Specificity | | | Sensitivity | | | Accuracy | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | *Median* | *Mean* | *SD* | *Median* | *Mean* | *SD* | *Median* | *Mean* | *SD* |
| 800 | Global | 0.94 | 0.92 | 0.08 | 0.90 | 0.80 | 0.22 | 0.90 | 0.82 | 0.17 |
| | Individual | 0.89 | 0.90 | 0.08 | 0.81 | 0.81 | 0.11 | 0.84 | 0.83 | 0.09 |
| 1000 | Global | 0.94 | 0.92 | 0.09 | 0.93 | 0.86 | 0.14 | 0.92 | 0.87 | 0.11 |
| | Individual | 0.94 | 0.92 | 0.08 | 0.82 | 0.82 | 0.09 | 0.85 | 0.84 | 0.07 |
| 1200 | Global | 0.94 | 0.94 | 0.07 | 0.94 | 0.85 | 0.19 | 0.94 | 0.87 | 0.14 |
| | Individual | 0.96 | 0.95 | 0.08 | 0.94 | 0.88 | 0.10 | 0.96 | 0.90 | 0.08 |

proposed algorithms. Results show that Global LASSO has considerably shorter processing times for both $K = 2$ and $K = 3$. As previously stated, this is completely expected given the nature of the proposed algorithms.

Table 33 – Processing times (in minutes) for $K = 2$, where total time considers running the $R = 50$ replications.

| $T$ | Algo. | Total Time | Avg. Time/Rep. | Median Time/Rep. | SD Time/Rep. |
|---|---|---|---|---|---|
| 400 | Global | 27.14 | 0.54 | 0.53 | 0.01 |
| | Individual | 762.10 | 15.24 | 15.35 | 0.42 |
| 600 | Global | 36.78 | 0.74 | 0.76 | 0.01 |
| | Individual | 1,010.95 | 20.22 | 20.20 | 0.22 |
| 800 | Global | 47.49 | 0.95 | 0.98 | 0.01 |
| | Individual | 1,241.16 | 24.82 | 24.83 | 0.45 |

Table 34 – Processing times (in minutes) for $K = 3$, where total time considers running the $R = 30$ replications.

| $T$ | Algo. | Total Time | Avg. Time/Rep | Median Time/Rep | SD Time/Rep |
|---|---|---|---|---|---|
| 800 | Global | 83.08 | 2.77 | 2.74 | 0.04 |
| | Individual | 5,605.84 | 186.86 | 186.29 | 1.91 |
| 1000 | Global | 96.66 | 3.22 | 3.22 | 0.05 |
| | Individual | 6,522.65 | 217.42 | 216.92 | 2.39 |
| 1200 | Global | 111.78 | 3.73 | 3.69 | 0.04 |
| | Individual | 7,632.28 | 254.41 | 252.36 | 3.59 |

## 4.5 Closing remarks for simulations

The first part of the simulation study focuses on demonstrating that the proposed algorithms perform shrinkage of the transition coefficients satisfactorily, as well as determining which optimization method is most adequate to use. The effect of factors such as amount of covariates, $D$, in each regression equation, and the initial values for the transition coefficients is also tested and analyzed along with the optimization methods. From the results of Scenarios A, B and C, we can conclude that the most appropriate optimization method to be used is the BFGS

method. In all scenarios, the use of this method along with the proposed algorithm has yielded outstanding results for all metrics related to transition coefficient shrinkage. Besides that fact, we discover that in general, the Nelder-Mead method yields poor results for shrinkage under any of the studied scenarios. It is very sensitive to different sets of initial values, and it seems that in all scenarios, shrinkage does not happen correctly when Nelder-Mead is the selected optimization method.

The Individual LASSO was not tested in Section 4.3, merely because of visualization considerations. The fact that there are 3 penalization parameters, ($\lambda_i$ for $i = 1, \ldots, K$) to analyze in the Individual LASSO, makes obtaining a graphical notion of how values of the selected metrics behave with different values of $\lambda$ not viable, or at the very least, extremely challenging. However, through our understanding of the inner workings of the two proposed algorithms and because of the independence between the regressions for the transitions between non-observable states when given the non-observable sequence $\mathbf{S}$, we assume that performance of the Individual LASSO should be similar to what was observed in Scenarios A, B and C. In fact, if we consider that in the Global LASSO, the transition coefficients are penalized equally by the same value of $\lambda$, that is, $\lambda_1 = \ldots = \lambda_K$, then it is a special combination that is also tested in the Individual LASSO, since it tests all possible combinations for $(\lambda_1, \ldots, \lambda_K)$.

The second part of the simulation study focuses on inferential and predictive performance, as well as other general aspects of the proposed algorithms. We carry out several test to understand the capabilities of the proposed algorithms in terms of estimation, prediction, classification and shrinkage. Regarding estimation, Scenarios 1 and 2 show very good results for all parameters being estimated. Some slight issues related to the parameter confidence intervals related to non-observable state 2 were noticed, but in spite of this, mean and median estimated values were satisfactory. Scenario 3 demonstrates the impact of overlapping distributions of the observable values on parameter estimation. Some real values of confidence intervals do not included the real parameter, and the size of several of the confidence interval is considerably larger than those observed in other scenarios. However, in spite of these observed difficulties, parameter estimates for Scenario 3 may be considered reasonable.

In terms of predictive capabilities, the proposed algorithms show better performance than the comparison models in all scenarios. Both mean and median values of MSPE are lower for the proposed algorithms when compared with ARIMA and Penalized Linear Regression. In general, the proposed algorithms capture the behavior of the test data sets well, and show good performance when predicting values of the test data set.

When performing classification, we observe very good results in all scenarios. We must mention that Scenario 3 presents a slight drop in classification performance, most likely due to the greater degree of overlapping between the distributions of the observable values. We must highlight the fact that in a real data setting, classification cannot be properly evaluated. However, since the real values of the non-observable sequence are available in a simulated setting, we have

clearly established that the proposed algorithms have very good performance when classifying observations into their corresponding non-observable states.

Regarding shrinkage, we perceive that the proposed algorithms have outstanding performance. In all scenarios, shrinkage of the transition coefficients occurs successfully along all analyzed metrics. Scenario 2 shows that when coefficient real values are in the proximity of zero, then the algorithms present a decrease in their performance. A possible reason for this is the zero threshold used in the simulations. A lower value for the threshold might have allowed for better identification of coefficients whose real value is not zero. In spite of the phenomenon observed in Scenario 2, shrinkage still occurs with excellent results throughout all scenarios. A noteworthy fact regarding shrinkage is the remarkable progress displayed between the first round of simulations shown in Appendix A, and the simulations presented in Chapter 4. We now understand that this progress is due to the selection of an adequate numerical optimization method such as BFGS. Such discovery inspired the design of simulations presented in Section 4.3.

Finally, we verify the computational costs of using Individual and Global LASSO. As expected, the Individual LASSO requires significantly more time than the Global LASSO due to the having an individual penalization parameter for the regression equations of for the transitions to each non-observable state. However, it is in general slightly better at variable selection.

CHAPTER

# 5

# RAINFALL PREDICTION USING NHMM AND VARIABLE SELECTION

The versatility of NHMMs is well known in several fields of knowledge, particularly in the presence of time-series data. These models have been widely used to explain, classify and predict different phenomena in which there is temporal data, or at least, some type of temporal dependence among sequential observations. The model's capacity to capture the intricate dynamics of first-order dependence among observations makes it ideal to model meteorological phenomena such as precipitation occurrence.

Several authors have used NHMMs to model precipitation. Holsclaw *et al.* (2017) develop an efficient Bayesian MCMC sampling scheme and apply it on a NHMM to model rainfall quantities in India over a 30-year period, starting in 1981 and ending in 2010. Applying NHMMs produced much insight into the duration and occurrence of rainfall periods across the country, each with specific characteristics. Sabillón and Zuanetti (2023) propose a Bayesian MCMC methodology, as well as a Stochastic EM algorithm to explain the monthly amount of days with rainfall in the capital city of Honduras, Tegucigalpa. Sabillón and Zuanetti (2023) clearly identify 3 non-observable states, each presenting distinct characteristics related to average temperature, humidity and rainfall. The identification of these 3 non-observable states is coherent with the weather patterns observed in the studied region.

To the extent of the researcher's knowledge, there are no known instances in literature of variable selection being used to choose the best set of covariates to calculate transition probabilities between non-observable states, specifically in the context of improving the NHMM's capability to predict the occurrence of rainfall. In this chapter, we present an application of the NHMM and the proposed algorithms to perform variable selection (shrinkage), parameter estimation, classification of observations and prediction for rainfall data collected in the city of São Carlos, in the state of São Paulo, in southeastern Brazil.

## 5.1 Description of the phenomenon and data

The city of São Carlos is located in the state of São Paulo, in the southeastern region of Brazil. It is located at approximately $22°0'55''$ S, $47°53'28''$ W, latitude and longitude coordinates. It has a mean elevation of 854 m above sea level.

The data being used in this application was collected at a weather station operated by the INMET (Instituto Nacional de Meteorologia), the Brazilian national meteorological organization. It is responsible for monitoring and forecasting weather and climate conditions across the country. The catalog number of the weather station is A711. This weather station was installed and initiated data collection on September 3, 2006. The exact location of the weather station is $21°58'49.2''$ S, $47°53'2.1''$ W. Data from this weather station and other stations throughout the different locations in Brazil is readily available for download at https://portal.inmet.gov.br/paginas/catalogoaut.

The starting date for the data set being used is July 14, 2007. We decide to use this starting date because previous dates show a large amount of missing values. The last date with recorded data in the data set corresponds to August 14, 2024. The sensors in weather station A711 collect data on an hourly basis, therefore, the time frame of the collected data amounts to 149788 hourly data points. The variables in the data set include, mean, maximum and minimum recorded values for temperature, dew point, atmospheric pressure, and humidity. Wind velocity, wind gust velocity, wind direction are also recorded. Our target variable, amount of precipitation is also measured and recorded on an hourly basis.

Table 35 – Summary statistics for the variables in the São Carlos precipitation data set.

| Variable | Unit | Min | Median | Mean | Max | SD |
|---|---|---|---|---|---|---|
| Rainfall | mm | 0.00 | 11.60 | 26.48 | 283.80 | 36.35 |
| MeanTemp | $°C$ | 12.85 | 21.48 | 21.06 | 28.20 | 2.57 |
| MaxTemp | $°C$ | 13.52 | 22.16 | 21.74 | 29.06 | 2.59 |
| MinTemp | $°C$ | 12.25 | 20.91 | 20.41 | 27.37 | 2.56 |
| MeanHumidity | % | 33.42 | 70.02 | 68.51 | 91.00 | 10.42 |
| MaxHumidity | % | 36.32 | 72.93 | 71.28 | 92.04 | 10.18 |
| MinHumidity | % | 30.53 | 67.15 | 65.65 | 89.88 | 10.59 |
| MeanDewPoint | $°C$ | 2.32 | 14.78 | 14.26 | 20.11 | 3.63 |
| MaxDewPoint | $°C$ | 2.65 | 15.29 | 14.82 | 20.43 | 3.62 |
| MinDewPoint | $°C$ | 2.02 | 14.22 | 13.74 | 19.88 | 3.65 |
| MeanPressure | $hPa$ | 913.04 | 919.52 | 919.90 | 928.92 | 2.80 |
| MaxPressure | $hPa$ | 913.29 | 919.79 | 920.16 | 929.14 | 2.79 |
| MinPressure | $hPa$ | 912.80 | 919.27 | 919.65 | 928.70 | 2.82 |
| WindSpeed | $m/s$ | 25.00 | 1.73 | 1.69 | 3.45 | 0.63 |
| WindDirection | $°$ | 71.52 | 166.15 | 177.36 | 306.91 | 51.79 |
| WindGust | $m/s$ | 0.96 | 4.84 | 4.79 | 8.04 | 1.09 |

Since our simulation study dealt with much shorter chain lengths, we decide to summarize the data into weekly time periods. For all covariates, we calculate the mean over the weekly periods, and for the target variable we calculate the sum of the observed values over each weekly

period. After this aggregation, and further removal of missing values, the resulting data set will contain 881 data points. This resulting chain length is similar to those which are tested throughout our simulation study in Chapter 4. However, the methods may be applied to longer sequences. Summary statistics for the data set are presented in Table 35. It is also important to understand the distribution of the covariates available for model fitting. For such purposes, we plot the distribution of the covariates and display the resulting plots in Figure 18. Yearly time series plots of each of the covariates are also made available in Appendix B to be able to have a clear perception of behavior and tendencies of each covariate along the complete studied time period. As was done in the simulation studies in Chapter 4, all covariates are standardized to have a mean of 0 and a standard deviation of 1.

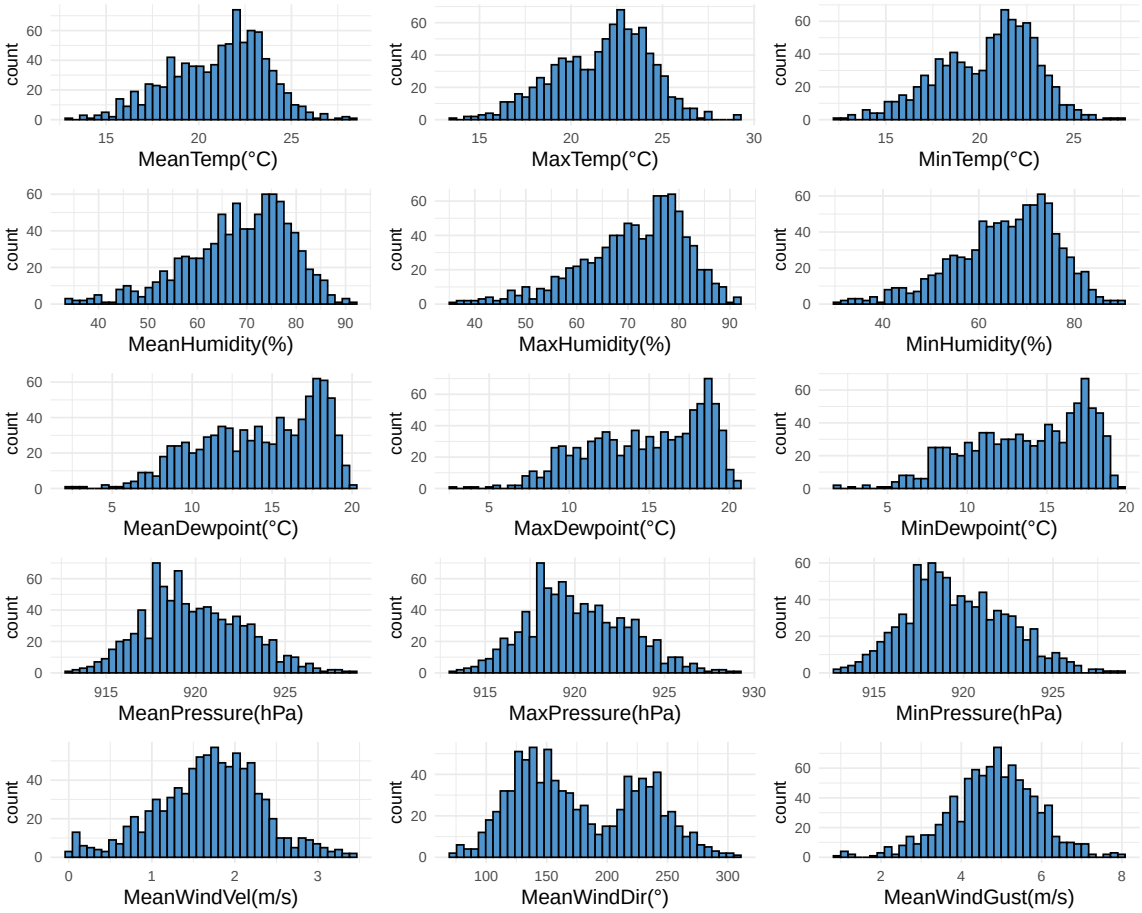



Figure 18 – Histograms for the variables in the database.

After the previously described treatment of the data set, we proceed to model fitting. We fit models using both proposed algorithms, as well as the ARIMA and Penalized Linear Regression algorithms. Our evaluation metric is the mean square predictive error. As described in Chapter 4, we calculate this metric on the test data set. Separation of the data base into training, validation and test data set is carried out as was done in the simulation study, allocating 80% of observations to the training data set, 15% to the validation set and 5% to the test data set. We fit models using both of the proposed algorithms for values of $K = 2, 3$. All models are fitted with 2

versions of the data set, the first is described in the previous paragraph in which observations and covariates correspond to the same time period (week), and a second version of the data set in which observations are lagged by one unit of time. In this second set of models, we will use values of the covariates at time $t$ to predict values of precipitation at time $t+1$. This is a highly advantageous, given that it will allow to use data from the current week to predict volumes of precipitation for the upcoming week. Initial values for the transition parameters are generated from a $N(0,1)$ distribution.

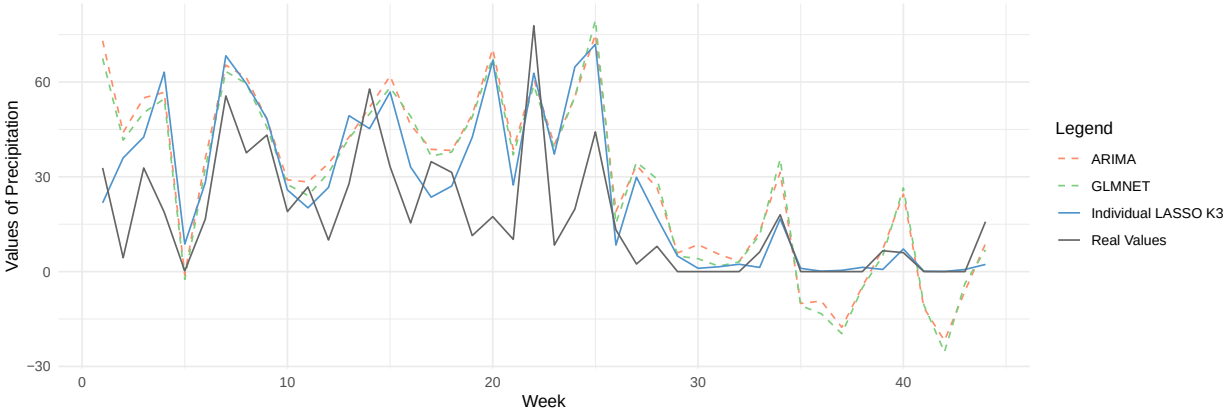
Results for all model fits are shown in Table 36. Results for models fitted with the lagged version of the data set are labeled with the word "lag" in parenthesis.

Table 36 – MSPE for all fitted algorithms.

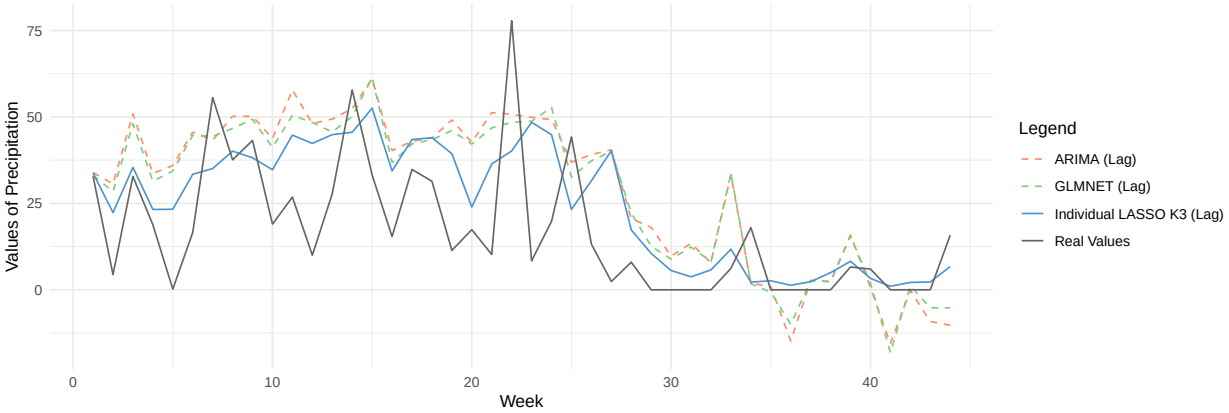
| **Algorithm** | **MSPE** |
|---|---|
| ARIMA | 473.9 |
| Pen. Linear Reg. | 460.5 |
| Global LASSO K2 | 431.1 |
| Individual LASSO K2 | 426.2 |
| Global LASSO K3 | 383.7 |
| Individual LASSO K3 | 335.2 |
| ARIMA (Lag) | 483.2 |
| Pen. Linear Reg. (Lag) | 433.8 |
| Global LASSO K2 (Lag) | 305.7 |
| Individual LASSO K2 (Lag) | 302.3 |
| Global LASSO K3 (Lag) | 293.8 |
| Individual LASSO K3 (Lag) | 288.9 |

As we can perceive from Table 36, the proposed algorithms have better predictive performance than ARIMA and Penalized Linear Regression for both $K = 2$ and $K = 3$. We also observe that all models except the ARIMA model perform better on the data set with the lagged response variable. For both the lagged data set and the original data set, models which have the better predictive performance are those fitted with $K = 3$. In other words, using 3 non-observable states to predict precipitation seems to be most adequate for the particular phenomenon. The model that shows the best predictive performance for either data set is the Individual LASSO fitted with $K = 3$ non-observable states, but the Global LASSO presents a similar performance. We select the Individual LASSO model with $K = 3$ for both the original and lagged data set for further analysis. We will compare their performance to the ARIMA and Penalized Linear Regression models.

From Figure 19a, we observe that the 3 models fit with the response variable with no lag seem to perform better when capturing some of the nuances and peaks of the real time series. However, errors in capturing behavior of the real time series also become much more evident. Such is the case for predictions in weeks 20 and 25 , in which the 3 models greatly overestimate the amount of precipitation. In spite of that evident overestimate, the Individual LASSO performs better than ARIMA and Penalized Linear Regression for that point in the time




(a) Models fit with no lag data




(b) Models fit with lagged data

Figure 19 – Predictions and real values of the test data set for original and lagged Individual LASSO with $K = 3$, and for comparison models.

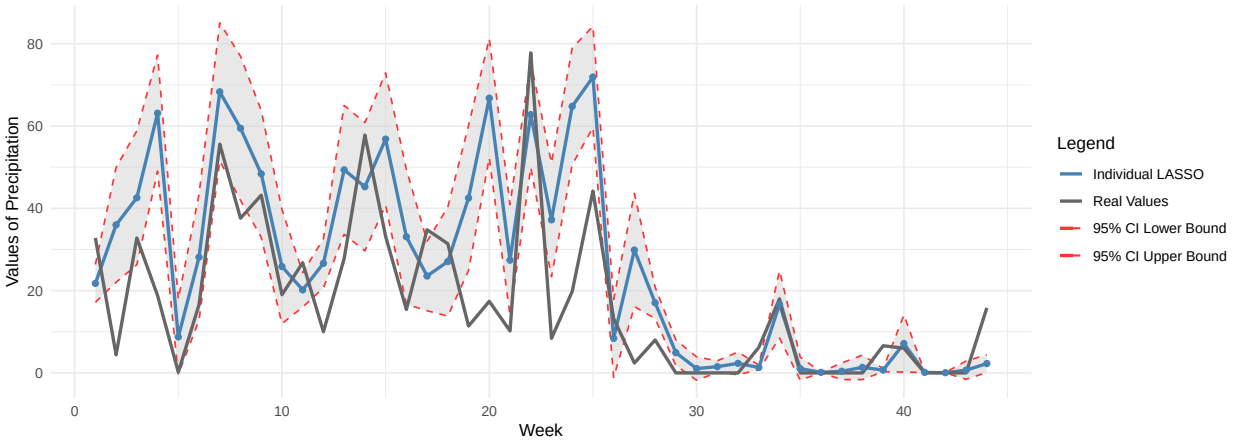
series. As a matter of fact, if we perform a point-by-point comparison between the errors for the ARIMA and Penalized Linear Regression models with the errors for the Individual LASSO with no lag, we observe that the errors for the Individual LASSO with no lag are less than the errors for the Penalized Linear Regression model in 32 of the 44 data points, which represents 72.7% of observations. The Individual LASSO with no lag has errors with lesser magnitude than the ARIMA model in 33 of 44 data points, representing 75% of observations in the test data set.

In contrast, when analyzing Figure 19b, we perceive that ARIMA and Penalized Linear Regression with lagged data have notably worse performance than the Individual LASSO with lag. It also seems that the Individual LASSO with lag tends to be somewhat smoother than models without lag. This means that the Individual LASSO with lag may not fully capture the behavior of the real time series, (particularly at peaks and extreme data points, such as 5, 12 and 22) but it performs better at modeling the real time series by predicting values which minimize errors as much as possible, while still retaining some of the general behavior of the real time

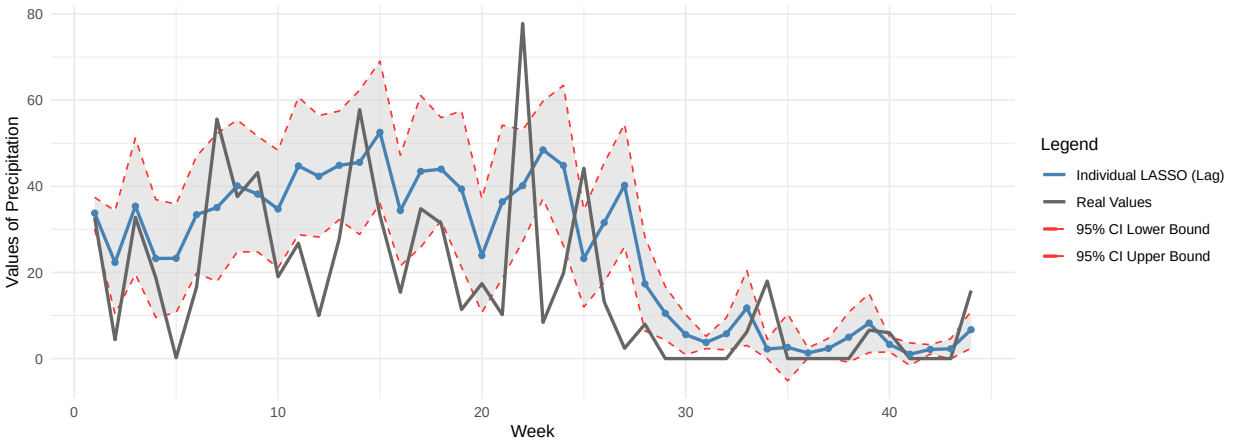
series.

By performing a point-by-point comparison between the errors for the lagged ARIMA and lagged Penalized Linear Regression models with the errors for the Individual LASSO with lag, we observe that the errors for the Individual LASSO with lag are less than the errors for the Penalized Linear Regression model in 33 of the 44 data points, which represents 75% of observations. The Individual LASSO with lag has errors that are smaller than the lagged ARIMA model in 35 of 44 data points, representing 79.4% of observations in the test data set.

To further understand the behavior of both of the Individual LASSOs that obtained the best predictive performances, we generate a 95% confidence intervals for each data point and analyze the results graphically. These credibility intervals are generated by randomly sampling 30 values for each point in the time-series, based on the transition probabilities calculated for that point in time. Afterwards, we use a Gaussian distribution with the 30 generated values to calculate the upper and lower bounds of the interval.




(a) Models fit with no lag data




(b) Models fit with lagged data

Figure 20 – Real values, predictions and 95% CI for the Individual LASSO.

From Figure 20, we perceive that, in spite of modeling the real time series relatively well, some of the data points of the real time series are not contained within the 95% confidence interval of the predictions of both of the Individual LASSOs. This is particularly evident when observing extreme points in the time series, which the 95% confidence interval is rarely containing. For the Individual LASSO with no lag, the percentage of observations contained within the confidence interval is 70.4%, while for the Individual LASSO with lag we have 61.3% of observations contained in the interval. However, in spite of this fact, we still judge the performance of the proposed methods to be reasonably good.

Table 37 – Estimates for the parameters of the distributions of the observable values.

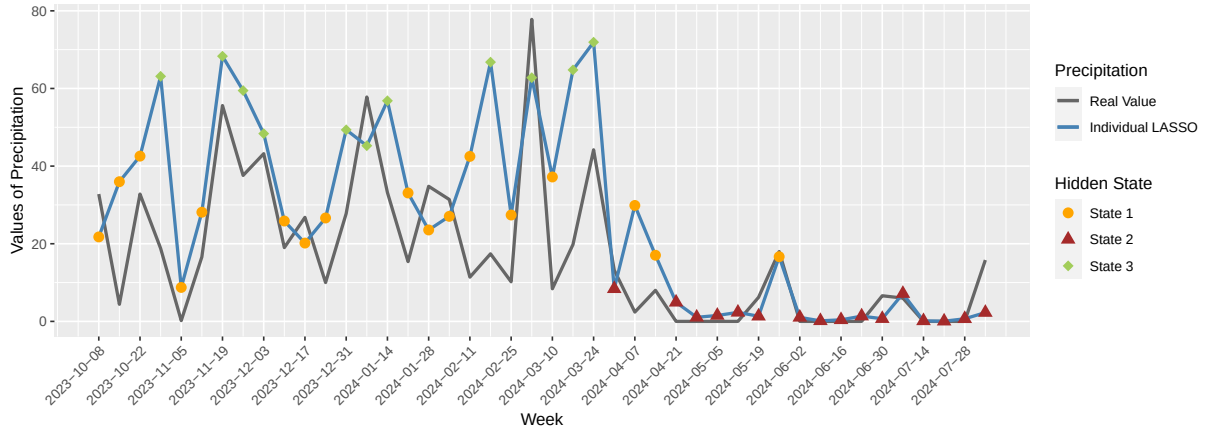
| **Parameter** | **Individual LASSO** | **Individual LASSO (Lag)** |
|---|---|---|
| $\mu_1$ | 14.71 | 16.21 |
| $\mu_2$ | 0.07 | 0.16 |
| $\mu_3$ | 76.18 | 78.63 |
| $\sigma_1$ | 10.41 | 10.71 |
| $\sigma_2$ | 0.18 | 0.35 |
| $\sigma_3$ | 36.49 | 36.09 |

Another important task that the proposed methods perform is estimation of the parameters of the distributions of the observable values. As was done in Chapter 4, we have assumed Normal distributions for the observable values for all non-observable states. Since the models with the lowest MSPE have $K = 3$ non-observable states, we will have 3 sets of these parameters, one for each non-observable state. The results for estimation are shown in Table 37.

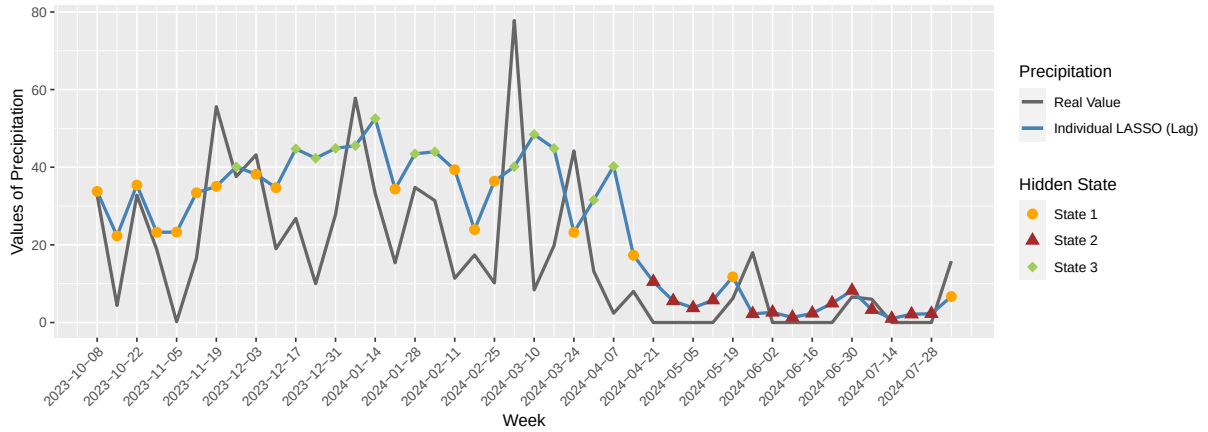
From Table 37, we perceive the characteristics of the non-observable states inferred by the models. Parameters for the distribution of the observable values for state 1 describe time periods with intermediate amounts of precipitation, and intermediate variability in those amounts. State 2 describes time periods with little or no precipitation, and very low variability in the amounts of precipitation observed in the time periods that correspond to this state. Finally, state 3 describes time periods with considerably larger amounts of precipitation, and displaying much greater variability. State 3 probably accounts for the peaks observed in the data.

An interesting capability of the NHMM is the fact that the non-observable states predicted by the model may be used as classification label. Figure 21 shows the values which both the Lagged and no Lag Individual LASSO predict for the test data set, and their corresponding predicted non-observable states. For the test data, we observe that both models perform very well when allocating observations into non-observable states. This particularly evident for observations classified into non-observable state 2 in which, we perceive that errors between the predicted and observed values have the smallest magnitudes for both models. States 1 and 3 have errors with greater magnitudes, mostly because of the common occurrence of peaks in the time series for these two non-observable states.

For the test data set, we observe that predictions which are allocated to non-observable




(a) Models fit with no lag data




(b) Models fit with lagged data

Figure 21 – Real values, predictions for the Individual LASSO and corresponding non-observable states.

Table 38 – MSPE for predictions for every non-observable state.

| | **Individual LASSO** | | **Individual LASSO (Lag)** | |
|---|---|---|---|---|
| | *MSPE* | *# of Predictions* | *MSPE* | *# of Predictions* |
| State 1 | 305 | 17 | 257 | 17 |
| State 2 | 18.9 | 16 | 35.5 | 14 |
| State 3 | 842 | 11 | 603 | 13 |

state 2 occur between April and July, while most predictions which are allocated to states 1 and 3 occur from October to March. This is very coherent with the observed weather patterns of the city of São Carlos, given that the dry season extends from late April to approximately the end of September or early October, and non-observable state 2 describes periods with very low or no precipitation and low variability in the values of precipitation. On the other hand, states 1 and 3 describe periods with considerably greater amounts of precipitation that state 2 and more variability in the observed values, which corresponds to the months in which states 1 and 3 are predominantly allocated in the test data set.

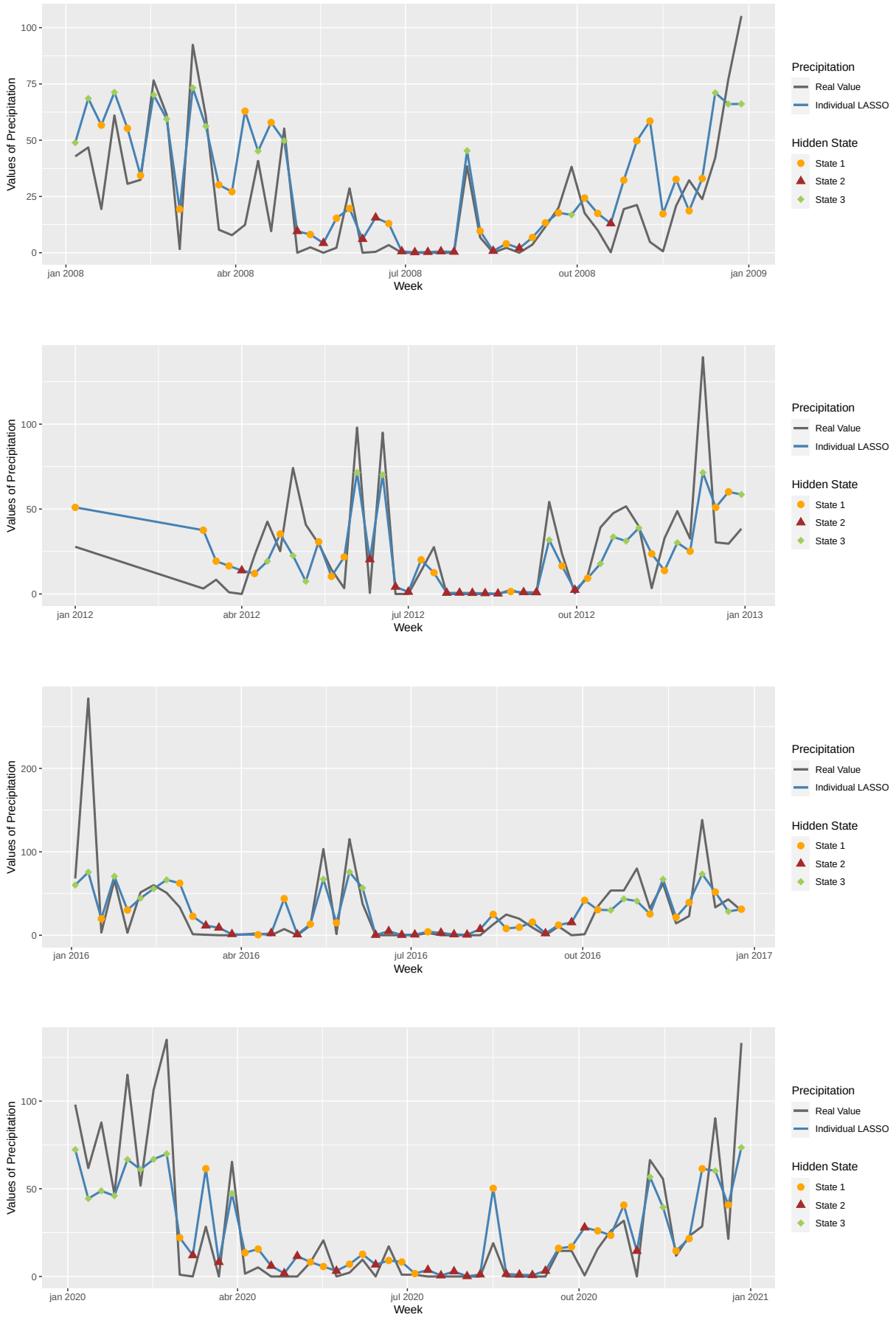



Figure 22 – Real values of precipitation, fitted values for Individual LASSO (no Lag) and their corresponding non-observable states for 2008, 2012, 2016 and 2020.

To further support the good performance of the NHMM in identifying periods with distinct values of precipitation and classifying predictions and/or observations into these inferred periods, we select 4 years of data from the data set, and plot the real values of precipitation, fitted values for the Lag and no Lag Individual LASSO along with their corresponding predicted non-observable states.

From Figures 22 and 23 we perceive classification patterns which are very coherent with observed weather patterns in the city of São Carlos. In general, we observe weeks with lower amounts of precipitation between the months of April and October. By late September or early-October, we observe an increase in the amounts of precipitation with increased variability in the weekly values. This increase in precipitation is maintained until late April or early May, when the dry season arrives. By analyzing the non-observable states allocated to the observations in the months cited in the previous description, we can perceive that classification of observations is performed reasonably well.

Finally, we discuss the task of interest for the proposed methods, that is variable selection via coefficient shrinkage. For such purpose, it is important to understand that given that model fitting was carried out with 15 covariates and considering 3 hidden states, the resulting transition coefficient matrix will be a $3 \times 3$ matrix, in which each element is a vector containing 16 transition coefficients. To facilitate understanding of the interpretations, we will present the coefficients in a table format. Coefficients corresponding to the transitions to non-observable state 1 will not be included in the table, given that they are all set to 0 because we use the mlogit function as link function.

Table 39 – Estimated transition coefficients for the Individual LASSO with no Lag. Coefficients within the zero threshold are marked with (*).

| **Variable** | **Coefficient** | $\boldsymbol{\beta_{12}}$ | $\boldsymbol{\beta_{13}}$ | $\boldsymbol{\beta_{22}}$ | $\boldsymbol{\beta_{23}}$ | $\boldsymbol{\beta_{32}}$ | $\boldsymbol{\beta_{33}}$ |
|---|---|---|---|---|---|---|---|
| Intercept | $\beta_{..1}$ | -1.554 | -2.617 | -1.790 | -1.661 | -2.050 | -1.778 |
| Avg. Temperature | $\beta_{..2}$ | -0.035 | 0.003* | 5.173 | 0.001* | -0.001* | 0.001* |
| Avg. Max. Temperature | $\beta_{..3}$ | -1.438 | 5.550 | 0.002* | -0.001* | -0.002* | 2.502 |
| Avg. Min. Temperature | $\beta_{..4}$ | -2.998 | 2.306 | -6.350 | 7.571 | -5.019 | 6.009 |
| Avg. Humidity | $\beta_{..5}$ | -5.951 | 0.036* | -2.683 | 8.023 | -0.003* | 10.920 |
| Avg. Max. Humidity | $\beta_{..6}$ | -1.901 | 4.366 | -4.814 | 0.001* | -7.293 | 0.000* |
| Avg. Min. Humidity | $\beta_{..7}$ | 1.369 | 6.559 | 4.218 | 2.705 | -0.001* | 0.001* |
| Avg. Dew Point | $\beta_{..8}$ | 0.031* | 0.003* | 0.005* | 0.001* | 2.988 | -0.004* |
| Avg. Max. Dew Point | $\beta_{..9}$ | 3.510 | 0.003* | 0.003* | -1.905 | 0.009 | 0.000* |
| Avg. Min. Dew Point | $\beta_{..10}$ | 2.458 | -10.472 | 2.736 | -9.427 | 4.224 | -11.686 |
| Avg. Pressure | $\beta_{..11}$ | 0.002* | 0.002* | 5.281 | -0.013 | 0.005* | -0.001* |
| Avg. Max. Pressure | $\beta_{..12}$ | -9.088 | 11.105 | -0.001* | -0.003* | -0.001* | 14.724 |
| Avg. Min. Pressure | $\beta_{..13}$ | 9.716 | -11.613 | 7.156 | -1.612 | 1.110 | -14.914 |
| Avg. Wind Speed | $\beta_{..14}$ | 0.641 | -1.280 | 1.876 | -1.242 | 1.877 | -0.271 |
| Avg. Wind Direction | $\beta_{..15}$ | -0.101 | -0.268 | -0.202 | 0.427 | 0.007* | -0.363 |
| Avg. Wind Gusts | $\beta_{..16}$ | -1.014 | 1.949 | -2.394 | 1.495 | -2.716 | 0.075 |

Table 39 shows the estimates for the transition coefficients for the Individual LASSO with

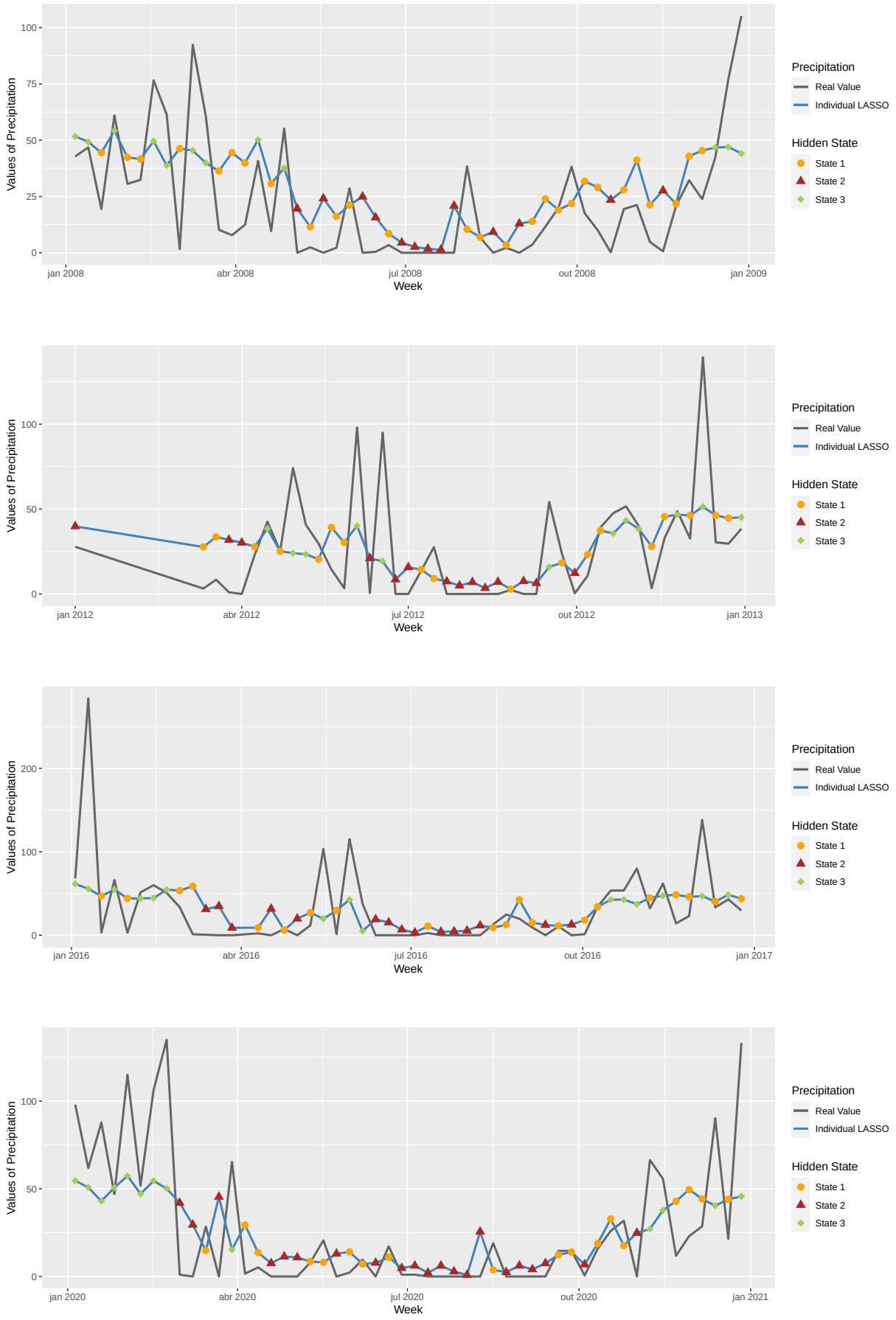



Figure 23 – Real values of precipitation, fitted values for Individual LASSO (Lagged) and their corresponding non-observable states for 2008, 2012, 2016 and 2020.

$K = 3$ and data with no lag. In the following interpretations, to facilitate referring to any specific transitions, we will simply refer to the name of the vector which contains the coefficients of given transition. For example, transition from state 1 to state 3 will simply be cited as transition $\boldsymbol{\beta_{13}}$. The table shows that, if we apply the same criteria used in Chapter 4 to determine if coefficients have been shrunken to 0, then several coefficients have not been selected in the different transition coefficient vectors.

Among some of the most noteworthy observations, we perceive that average temperature was only selected for the transition $\boldsymbol{\beta_{22}}$. Average maximum temperature is not important for transitions $\boldsymbol{\beta_{22}}$ and $\boldsymbol{\beta_{23}}$ which imply leaving state 2. Average minimum temperature is important to calculate the transition probabilities between all non-observable states. Interestingly, we observe that for transition $\boldsymbol{\beta_{12}}$, both average maximum and minimum temperature have negative coefficients, indicating that greater values of theses covariates will decrease the probability of transitioning from state 1, which describes periods of little to no precipitation. This is coherent with observed weather patterns of São Carlos, given that temperatures between the months of April and August tend to decrease, precisely when the lowest values of precipitation are observed in the region. Another notable fact is that for transitions $\boldsymbol{\beta_{13}}$, $\boldsymbol{\beta_{23}}$ and $\boldsymbol{\beta_{33}}$, average minimum temperature was important to calculate transition probabilities, and was always positive. This indicates that the greater the minimum average temperature the greater the probability of transitioning to state 3, which is the state that describes periods of rain with considerably greater amounts of precipitation. In general, The magnitude, sign and impact of coefficients which correspond to covariates related to temperature seem to be coherent with the interaction between temperature and precipitation, as has been commented by some authors. Fraedrich and Williams (2003) states that climate models show that an increase in surface temperatures results in more water vapor in the atmosphere, which can lead to increased precipitation rates. Other authors such as Sr. and Pearce (2012), mention that rising temperatures are expected to increase evaporation rates and atmospheric moisture, thereby enhancing the potential for more intense and frequent precipitation extremes. These ideas provide additional support for the parameter estimates of the temperature covariates, given that Figure 24, 25, and 26 in the Appendix B show that the lowest temperatures occur between weeks 20 and 35 in most years for which we have data. The time period between weeks 20 and 35 is the time period when we most commonly observe weeks with lowest amounts of precipitation.

The relationship between humidity and precipitation is also very clearly established. Several authors such as Zhang and Lee (2023) state that atmospheric humidity plays a crucial role in shaping precipitation patterns. Regions with higher humidity levels experience more frequent and intense precipitation, as increased moisture content in the atmosphere facilitates the development of precipitation. In the case of the estimated coefficients, we have that average humidity covariate is not important for transition probability calculation in the case of two transitions, $\boldsymbol{\beta_{13}}$ and $\boldsymbol{\beta_{32}}$. Average maximum humidity seems to have little impact in the calculating transition probabilities to transition into non-observable state 3, and average minimum

humidity is not relevant when calculating transition probabilities when leaving non-observable state 3. Among some interesting points regarding humidity covariates is their coherence with precipitation phenomena. In case of transition $\boldsymbol{\beta_{32}}$, which refers to transitioning from a state with considerably greater amounts of precipitation to state 2, which is the state with the lowest amount of precipitation, we observe that the coefficient for average maximum humidity is negative with relatively large magnitude. This indicates that the greater the average maximum humidity, then the smaller the probabilities of transitioning to state 2. This is mostly coherent with the weather patterns of São Carlos, given that the periods with the lowest amount of precipitation tend to have the lowest recorded humidity,as seen in Figures 27, 28, and 29 in Appendix B. For both transitions $\boldsymbol{\beta_{13}}$ and $\boldsymbol{\beta_{23}}$, we observe that humidity covariates are either not impacting transition probabilities, or they are increasing the probabilities of the transition occurring as the value of the covariate increases. In particular, we observe that for transition $\boldsymbol{\beta_{23}}$, the greater the value of average humidity and average maximum humidity, then the greater the probability of transitioning from state 2, that has very low amounts of precipitation to state 3 which has the greatest amount of observed precipitation of all 3 non-observable states. One more time, most if not all of estimated coefficients show, coherence with the phenomenon being studied.

Dew point is another key indicator that influences the occurrence of precipitation. Houghton, Mehta and Wilson (2018) explain that high dew point temperatures indicate the presence of ample moisture in the air, which is crucial for the formation of clouds and precipitation. When the dew point approaches the air temperature, the air becomes saturated, and the likelihood of condensation and subsequent precipitation increases. Smith, Anderson and Jones (2020) agrees with this notion and comments that precipitation is highly dependent on the dew point because it reflects the amount of moisture available in the atmosphere. When air with a high dew point is lifted and cooled, it reaches saturation more quickly, leading to cloud formation and potential precipitation. In the case of the estimated coefficients, we observe that several have been deemed as not being important for the calculation of transitions of the model. It seems that the average minimum dew point coefficients are presenting opposite interpretation to what is expected. For example, in the case of transition $\boldsymbol{\beta_{23}}$, we observe that an increase in the average minimum dew point would decrease the probability of transitioning from state 2 to state 3. This is counter-intuitive, given that increases in dew points usually means an expected increase in precipitation. The same occurs when analyzing transition $\boldsymbol{\beta_{32}}$, which indicates that an increase in the average minimum dew point would increase the probabilities of transition to state 2, which has the least amount of observed precipitation. This inverse tendency to expected behavior is observed among in coefficients in all transitions for the minimum average temperature variable. Coincidentally, as Figures 30, 32 and 31 in Appendix B show, the lowest observed average, minimum and maximum dew points are observed between weeks 20 and 40 of all the years in the data. These weeks correspond to periods with the lowest precipitation.

Regarding atmospheric pressure covariates, we perceive what seems to be some multicolinearity issues between the average minimum and maximum pressure covariates. This

is somewhat evident in transitions $\boldsymbol{\beta_{12}}$, $\boldsymbol{\beta_{13}}$ and, $\boldsymbol{\beta_{33}}$. Further analyses could be required to determine if these covariates should be removed from the model due to these issues. In the case of transitions $\boldsymbol{\beta_{22}}$, $\boldsymbol{\beta_{23}}$ and $\boldsymbol{\beta_{32}}$, we perceive much adherence to what is commonly known about the relationship between pressure and precipitation. The coefficients related to average pressure and average minimum pressure for transition $\boldsymbol{\beta_{22}}$ indicate that as pressure increases, there are higher probabilities of remaining in state 2, a behavior which is consistent with the phenomenon, given that state 2 is the state with least amounts of precipitation and higher pressure induce less precipitation. In the case of transition $\boldsymbol{\beta_{32}}$, an increase in the average minimum pressure will indicate an increase in the probability of transitioning to state 3 from state 2, in other words, transitioning from a period of greater amounts of precipitation to periods with the least amount of precipitation. In the case of transition $\boldsymbol{\beta_{23}}$, as the average minimum pressure increases, probabilities of transitioning to state 3 which present greater amounts of precipitation decrease. The magnitude and sign of the coefficients related to pressure covariates are reflective of the precipitation patterns in the city of São Carlos. These behaviors are also supported by the behavior of the covariates throughout periods of the year. As can be seen in Figures 33, 34 and 35 in Appendix B, the greatest observed values of average, maximum and minimum pressure occur between weeks 20 and 40, throughout the years. This is consistent with the phenomenon, given that the aforementioned weeks correspond to the periods of precipitation which usually present the lowest amounts of precipitation.

In the case of covariates related to wind, we observe that the occurrence of greater values of wind speed seems to positively influence transitioning into state 2, in which we have the lowest amounts of precipitation. The relationship is inverse when we observe the values of wind gusts in which greater values of wind gusts seem to decrease the probabilities of transitioning to state 2. Even though wind direction was selected as being relevant to all transitions, in general, it has small impact on calculating transition probabilities for all transitions. The influence of wind direction on precipitation is specific for any studied region, and there are no instances of references or research which establish a relationship between wind direction and precipitation in the region of São Carlos.

CHAPTER

6

# DISCUSSION AND FUTURE STUDIES

In this thesis, we initially review the NHMM. Elements, structure, and important functions are thoroughly explained to enhance the reader's comprehension of the model. This review of the NHMM is vital, given that the proposals deal with variable selection methods for the model. After this, we introduce two proposals, the Global LASSO and Individual LASSO for the NHMM. The theoretical structure of the two proposals and their usage coupled together with the Stochastic EM algorithm is explained in detail.

We test the performance of these proposals in two separate simulation studies. The first is designed to analyze and understand the importance of the selected numerical optimization method in the shrinkage performance of the proposed algorithms. We conclude that among the tested methods, the BFGS algorithm yields superior results and, therefore, the most appropriate method to be applied with the proposed algorithms. The second simulation study is designed to test all aspects of the proposed algorithms. Under controlled conditions, the two proposals show excellent performance when comparing their MSPE to the two comparison models, ARIMA and Penalized Linear Regression. In terms of prediction of the $\mathbf{S}$ sequence, both proposals achieve satisfactory results across all scenarios. Regarding shrinkage of the transition coefficients, the proposed algorithms show very good performance in performing shrinkage, and consequently variable selection, even in the presence of values which are close to the established zero threshold. In general, the advantages shown by the two proposals are noteworthy, given that for all simulation scenarios, the proposals have shown very good performance regarding predictive metrics. We would like to highlight that the proposed methods, in addition to being compatible with other methods in terms of prediction in time series, also select variables and classify observations into homogeneous periods, which helps us understand the evolution of the time series and the factors that impact its behavior.

We consider a Normal distribution for the observable variables, but the proposals may be easily adapted for other continuous and discrete distributions as well as for multivariate distributions.

Finally, we apply the proposed methods on a rainfall data set that contains precipitation records from the city of São Carlos, Brazil. The methods show good performance when predicting rainfall amounts. Predictions are classified into different states that correspond to periods of precipitation with different amounts of observed precipitation and variability in rainfall quantities. The proposed methods select the most important variables for rainfall prediction by shrinking the coefficients of those variables which are not important for specific transitions, improving predictive performance related to two comparison models, as well as gaining great insight on which are the most important covariates that explain transitions between the different rainfall regimes which are observed in the city of São Carlos. In spite of the real data application presented in this thesis is related to Meteorology, the methods may also be applied to financial, economic and social time series.

Individual and Global LASSO, the two methods proposed in this thesis, show great promise as tools for improving the predictive performance of the NHMM, as well as for variable selection by shrinkage. The performance shown in the simulation studies and the real data application exemplify their excellent qualities. We have laid the groundwork for other researchers to continue applying and improving the proposed methods in future projects.

## 6.1 Future proposals

The proposed methods have shown great potential in increasing the predictive capabilities of the NHMM as well as performing automatic variable selection for the model. There are no known instances of variable selection methods for the covariates of the non-homogeneous transition matrix that make use of frequentist approaches such as the Stochastic EM algorithm that is scalable. This thesis has established a starting point for several ideas which are yet to be explored. Among these we have:

- **Optimized $\lambda$ vector**: Using a vector of $\lambda$s which is selected by applying optimization will greatly improve the computational efficiency of the Individual LASSO, and might also additionally improve the shrinkage efficiency. An idea might be to find the values of each $\lambda_i$ for which all coefficients are set to 0, and to use this value as an upper limit for the $\lambda$ vector. We will explore this idea in future steps.

- **Regressions for the observable values**: Instead of having the emissions of the observable values be generated from a distribution having a general mean, we may include a regression in the mean parameter to emit these observable values. This will introduce an additional improvement in predictive accuracy of the model, and even allow for the possibility of penalization on such regressions.

- **In-depth study of numerical methods**: While we have made some important discoveries regarding the impact of the numerical method chosen for optimization with the proposed

algorithms, we believe that a much more detailed study regarding different numerical optimization methods is necessary to fully understand the impact of the selected optimization method on the performance of the LASSOs.

- **Bayesian methods**: The possibility of exploring the data-driven reversible jump (DDRJ), in the context of variable selection for the non-homogeneous transition matrix. The DDRJ has been applied in variable selection for hidden Markov models, but we find few instances in which it is specifically applied to perform variable selection for the covariates of the non-homogeneous transition matrix.

# BIBLIOGRAPHY

ADAMS, S.; BELING, P. A. A survey of feature selection methods for gaussian mixture models and hidden Markov models. **Artificial Intelligence Review**, Springer, v. 52, p. 1739–1779, 2019. Citation on page 17.

BANACHEWICZ, K.; LUCAS, A.; VAART, A. V. D. Modelling portfolio defaults using hidden Markov models with covariates. **Econometrics Journal**, v. 11, n. 1, p. 155–171, 2008. Citation on page 16.

BAUM, L. E.; PETRIE, T. Statistical inference for probabilistic functions of finite state Markov chains. **The Annals of Mathematical Statistics**, Institute of Mathematical Statistics, v. 37, n. 6, p. 1554–1563, 1966. ISSN 00034851. Available: <http://www.jstor.org/stable/2238772>. Citation on page 15.

BIETTI, A.; BACH, F.; CONT, A. An online EM algorithm in hidden (semi-)Markov models for audio segmentation and clustering. In: IEEE. **2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)**. [S.l.], 2015. p. 1881–1885. Citation on page 34.

BLANCHET, F. G.; LEGENDRE, P.; BORCARD, D. Forward selection of explanatory variables. **Ecology**, JSTOR, v. 89, p. 2623–2632, 2008. Citation on page 17.

CAPPé, O. Online EM algorithm for hidden Markov models. **Journal of Computational and Graphical Statistics**, Taylor & Francis, v. 20, n. 3, p. 728–749, 2011. Available: <https://doi.org/10.1198/jcgs.2011.09109>. Citation on page 34.

CARVALHO, C. M.; POLSON, N. G.; SCOTT, J. G. Handling sparsity via the horseshoe. In: PMLR. **Artificial intelligence and statistics**. [S.l.], 2009. p. 73–80. Citation on page 19.

CHIANG, A. P.; BECK, J. S.; YEN, H.-J.; TAYEH, M. K.; SCHEETZ, T. E.; SWIDERSKI, R. E.; NISHIMURA, D. Y.; BRAUN, T. A.; KIM, K.-Y. A.; HUANG, J. Homozygosity mapping with SNP arrays identifies trim32, an E3 ubiquitin ligase, as a bardet–biedl syndrome gene (BBS11). **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 103, n. 16, p. 6287–6292, 2006. Citation on page 18.

CHOI, H.; FERMIN, D.; NESVIZHSKII, A. I.; GHOSH, D.; QIN, Z. S. Sparsely correlated hidden Markov models with application to genome-wide location studies. **Bioinformatics**, Oxford University Press, v. 29, n. 5, p. 533–541, 2013. Citation on page 19.

CHOWDHURY, M. Z. I.; TURIN, T. C. Variable selection strategies and its importance in clinical prediction modelling. **Family Medicine and Community Health**, BMJ Specialist Journals, v. 8, n. 1, p. e000262, 2020. Citation on page 17.

DAI, Y.-H. Convergence properties of the bfgs algoritm. **SIAM Journal on Optimization**, SIAM, v. 13, n. 3, p. 693–701, 2002. Citation on page 40.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the em algorithm. **Journal of the royal statistical society: series B (methodological)**, Wiley Online Library, v. 39, n. 1, p. 1–22, 1977. Citation on page 29.

DYMARSKI, P. **Hidden Markov models: Theory and applications**. [S.l.]: BoD–Books on Demand, 2011. Citation on page 15.

EDDY, S. R. What is a hidden Markov model? **Nature biotechnology**, Nature Publishing Group, v. 22, n. 10, p. 1315–1316, 2004. Citation on page 16.

ENGEL, C. Can the Markov switching model forecast exchange rates? **Journal of International Economics**, v. 36, n. 1, p. 151–165, 1994. ISSN 0022-1996. Available: <https://www.sciencedirect.com/science/article/pii/0022199694900620>. Citation on page 15.

ERP, S. V.; OBERSKI, D. L.; MULDER, J. Shrinkage priors for bayesian penalized regression. **Journal of Mathematical Psychology**, Elsevier, v. 89, p. 31–50, 2019. Citation on page 19.

FAN, J.; LI, R. Variable selection via nonconcave penalized likelihood and its oracle properties. **Journal of the American statistical Association**, Taylor & Francis, v. 96, n. 456, p. 1348–1360, 2001. Citation on page 17.

FRAEDRICH, K. D.; WILLIAMS, M. J. E. The influence of temperature on precipitation: A climate model study. **Journal of Climate**, American Meteorological Society, v. 16, n. 15, p. 2552–2560, 2003. Available: <https://journals.ametsoc.org/view/journals/clim/16/15/1520-0442_2003_016_2552_tiotpa_2.0.co_2.xml>. Citation on page 106.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. Regularization paths for generalized linear models via coordinate descent. **Journal of statistical software**, NIH Public Access, v. 33, n. 1, p. 1–22, 2010. Citations on pages 36 and 38.

FRIEDMAN, J.; TIBSHIRANI, R.; HASTIE, T. Regularization paths for generalized linear models via coordinate descent. **Journal of Statistical Software**, v. 33, n. 1, p. 1–22, 2010. Citation on page 43.

GAO, B.; PAVEL, L. **On the Properties of the Softmax Function with Application in Game Theory and Reinforcement Learning**. 2018. Citations on pages 27 and 46.

GAO, X.; SONG, P. X.-K. Composite likelihood EM algorithm with applications to multivariate hidden Markov model. **Statistica Sinica**, JSTOR, v. 21, n. 1, p. 165–185, 2011. Citation on page 34.

GOLLERY, M. Bioinformatics: sequence and genome analysis. **Clinical Chemistry**, American Association for Clinical Chemistry, Inc., v. 51, n. 11, p. 2219–2220, 2005. Citation on page 16.

GREEN, P. J. Reversible jump Markov chain monte carlo computation and bayesian model determination. **Biometrika**, Oxford University Press, v. 82, n. 4, p. 711–732, 1995. Citation on page 20.

GÜNTER, S.; BUNKE, H. Fast feature selection in an HMM-based multiple classifier system for handwriting recognition. In: SPRINGER. **Pattern Recognition: 25th DAGM Symposium, Magdeburg, Germany, September 10-12, 2003. Proceedings 25**. [S.l.], 2003. p. 289–296. Citation on page 32.

HOERL, A. E.; KENNARD, R. W. Ridge regression: applications to nonorthogonal problems. **Technometrics**, Taylor & Francis, v. 12, n. 1, p. 69–82, 1970. Citation on page 17.

HOLSCLAW, T.; GREENE, A.; ROBERTSON, A.; SMYTH, P. Bayesian non-homogeneous Markov models via Pólya-Gamma data augmentation with applications to rainfall modeling. **The Annals of Applied Statistics**, v. 11, n. 1, p. 393–426, 01 2017. Citations on pages 23 and 95.

HOUGHTON, J. T.; MEHTA, L. G.; WILSON, R. J. Understanding and using the dew point. **Journal of Climate and Weather**, v. 45, p. 123–134, 2018. Citation on page 107.

HUANG, J.; MA, S.; ZHANG, C.-H. Adaptive lasso for sparse high-dimensional regression models. **Statistica Sinica**, Institute of Statistical Science, Academia Sinica, v. 18, n. 4, p. 1603–1618, 2008. ISSN 10170405, 19968507. Available: <http://www.jstor.org/stable/24308572>. Citation on page 18.

HUANG, M.; HUANG, Y.; HE, K. Estimation and testing nonhomogeneity of hidden Markov model with application in financial time series. **Statistics and Its Interface**, International Press of Boston, v. 12, n. 2, p. 215–225, 2019. Citation on page 16.

HYNDMAN, R. J.; KHANDAKAR, Y. Automatic time series forecasting: The forecast package for r. **Journal of Statistical Software**, v. 27, n. 3, p. 1–22, 2008. Available: <https://www.jstatsoft.org/index.php/jss/article/view/v027i03>. Citation on page 43.

JACQUET, P.; SEROUSSI, G.; SZPANKOWSKI, W. On the entropy of a hidden Markov process. **Theoretical Computer Science**, v. 395, n. 2, p. 203–219, 2008. ISSN 0304-3975. SAIL – String Algorithms, Information and Learning: Dedicated to Professor Alberto Apostolico on the occasion of his 60th birthday. Available: <https://www.sciencedirect.com/science/article/pii/S0304397508000364>. Citation on page 15.

KOKI, C.; MELIGKOTSIDOU, L.; VRONTOS, I. Forecasting under model uncertainty: Non-homogeneous hidden Markov models with Pòlya-Gamma data augmentation. **Journal of Forecasting**, Wiley Online Library, v. 39, n. 4, p. 580–598, 2020. Citation on page 20.

KROGH, A.; BROWN, M.; MIAN, I. S.; SJÖLANDER, K.; HAUSSLER, D. Hidden Markov models in computational biology: Applications to protein modeling. **Journal of molecular biology**, Elsevier, v. 235, n. 5, p. 1501–1531, 1994. Citation on page 16.

LAGONA, F.; MARUOTTI, A.; PICONE, M. A non-homogeneous hidden Markov model for the analysis of multi-pollutant exceedances data. **Hidden Markov Models, Theory and Applications**, InTech Rijeka, Croatia, p. 207–222, 2011. Citation on page 23.

MARUOTTI, A.; ROCCI, R. A mixed non-homogeneous hidden Markov model for categorical data, with application to alcohol consumption. **Statistics in Medicine**, v. 31, n. 9, p. 871–886, 2012. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4478>. Citation on page 28.

MELIGKOTSIDOU, L.; DELLAPORTAS, P. Forecasting with non-homogeneous hidden Markov models. **Statistics and Computing**, Springer, v. 21, n. 3, p. 439–449, 2011. Citation on page 20.

PORITZ, A. B. Hidden Markov models: a guided tour. In: **ICASSP**. [S.l.: s.n.], 1988. v. 88, p. 7–13. Citation on page 15.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2013. Available: <http://www.R-project.org/>. Citations on pages 39 and 43.

RABINER, L.; JUANG, B. An introduction to hidden Markov models. **IEEE ASSP Magazine**, v. 3, n. 1, p. 4–16, 1986. Citation on page 15.

RABINER, L. R. A tutorial on hidden Markov models and selected applications in speech recognition. **Proceedings of the IEEE**, Ieee, v. 77, n. 2, p. 257–286, 1989. Citations on pages 16 and 28.

____. A tutorial on hidden Markov models and selected applications in speech recognition. In: **Proceedings of the IEEE**. [S.l.: s.n.], 1989. p. 257–286. Citation on page 28.

SABILLÓN, G. A. **Algoritmos de estimação para modelos Markovianos não-homogêneos**. Master's Thesis (Master's Thesis) — Universidade de São Paulo, 2020. Citation on page 26.

SABILLÓN, G. A.; ZUANETTI, D. A. Analyzing the rainfall pattern in honduras through non-homogeneous hidden markov models. **Journal of Data Science**, School of Statistics, Renmin University of China, v. 21, n. 4, p. 799–817, 2023. Citations on pages 19, 21, 29, 35, 38, 42, and 95.

SMITH, M. K.; ANDERSON, L. J.; JONES, P. G. The role of dew point in weather prediction. **Meteorological Research Letters**, v. 32, p. 78–90, 2020. Citation on page 107.

SOUZA, F. L. de. **Mistura de Distribuições Extremais**. Master's Thesis (Master's Thesis) — Universidade de Brasília, 2010. Citation on page 46.

SPEZIA, L. Bayesian analysis of non-homogeneous hidden Markov models. **Journal of Statistical Computation and Simulation**, Taylor & Francis, v. 76, n. 8, p. 713–725, 2006. Citation on page 16.

____. Bayesian variable selection in non-homogeneous hidden Markov models through an evolutionary Monte Carlo method. **Computational Statistics & Data Analysis**, Elsevier, v. 143, p. 106840, 2020. Citation on page 20.

SR., R. A. P.; PEARCE, R. Climate change and precipitation extremes: An overview. **Weather and Climate Extremes**, Elsevier, v. 1, p. 1–12, 2012. Available: <https://www.sciencedirect.com/science/article/pii/S2212094712000022>. Citation on page 106.

STÄDLER, N.; MUKHERJEE, S. Penalized estimation in high-dimensional hidden Markov models with state-specific graphical models. **The Annals of Applied Statistics**, JSTOR, v. 7, n. 4, p. 2157–2179, 2013. Citations on pages 18, 32, and 38.

STAMEY, T. A.; KABALIN, J. N.; MCNEAL, J. E.; JOHNSTONE, I. M.; FREIHA, F.; REDWINE, E. A.; YANG, N. Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. ii. radical prostatectomy treated patients. **The Journal of Urology**, Elsevier, v. 141, n. 5, p. 1076–1083, 1989. Citation on page 18.

STEYERBERG, E.; EIJKEMANS, M.; HABBEMA, J. Application of shrinkage techniques in logistic regression analysis: a case study. **Statistica Neerlandica**, Wiley Online Library, v. 55, n. 1, p. 76–88, 2001. Citation on page 18.

TAYLOR, S. **Markov Models: An Introduction to Markov Models**. [S.l.]: Steven Taylor, 2020. Citation on page 24.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 58, n. 1, p. 267–288, 1996. Citations on pages 18, 36, and 38.

WANG, X.; SMITH-MILES, K.; HYNDMAN, R. Characteristic-based clustering for time series data. **Data Min. Knowl. Discov.**, v. 13, p. 335–364, 09 2006. Citation on page 43.

WOLFE, J.; JIN, X.; BAHR, T.; HOLZER, N. Application of Softmax regression and its validation for spectral-based land cover mapping. **ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences**, XLII-1/W1, p. 455–459, 2017. Citation on page 46.

WU, T. T.; CHEN, Y. F.; HASTIE, T.; SOBEL, E.; LANGE, K. Genome-wide association analysis by lasso penalized logistic regression. **Bioinformatics**, v. 25, n. 6, p. 714–721, 2009. ISSN 1367-4803. Available: <https://doi.org/10.1093/bioinformatics/btp041>. Citation on page 18.

YUAN, Y.-x. A modified bfgs algorithm for unconstrained optimization. **IMA Journal of Numerical Analysis**, Oxford University Press, v. 11, n. 3, p. 325–332, 1991. Citation on page 40.

ZHANG, L.; LEE, S. J. The impact of atmospheric humidity on precipitation patterns across different climate zones. **Atmospheric Science Letters**, v. 24, n. 3, p. e1168, 2023. Available: <https://rmets.onlinelibrary.wiley.com/doi/full/10.1002/asl.1168>. Citation on page 106.

ZHOU, X.; SONG, X. Functional concurrent hidden Markov model. **Statistics and Computing**, Springer, v. 33, n. 3, p. 57, 2023. Citations on pages 20 and 32.

ZHU, K. P.; HONG, G. S.; WONG, Y. S. A comparative study of feature selection for hidden Markov model-based micro-milling tool wear monitoring. **Machining Science and Technology**, Taylor & Francis, v. 12, n. 3, p. 348–369, 2008. Citation on page 31.

ZOU, H.; HASTIE, T. Regularization and variable selection via the elastic net. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 67, n. 2, p. 301–320, 2005. Citation on page 17.

ZUANETTI, D. A.; MILAN, L. A. Second-order autoregressive hidden markov model. **Brazilian Journal of Probability and Statistics**, JSTOR, v. 31, n. 3, p. 653–665, 2017. Citation on page 28.

ZUCCHINI, W.; MACDONALD, I.; LANGROCK, R. **Hidden Markov Models for Time Series: An Introduction Using R, Second Edition**. CRC Press, 2017. (Chapman & Hall/CRC Monographs on Statistics and Applied Probability). ISBN 9781482253849. Available: <https://books.google.com.br/books?id=KlWzDAAAQBAJ>. Citation on page 28.

APPENDIX

A

# QUALIFICATION EXAM SIMULATION RESULTS, ANALYSIS AND DISCUSSIONS

This appendix retains the memory of results obtained in our first round of simulations performed for Qualification Exam. As the reader may notice, results obtained related to shrinkage of the transition coefficients in this initial round of simulations were not satisfactory. Our intent with the inclusion of this appendix is for the reader to have a clear perception of the evolution of the proposed algorithms along the stages of this research project. All discussions, analyses and texts retain most, if not all of the original content submitted in the Qualification exam phase of this research project.

## A.1 Qualification Exam Simulation Study

We now carefully describe each simulation scenario, along with results and particular considerations regarding each scenario. To promote the reader's understanding of the development of the simulation study, we will now emphasize several key points of the study.

This work includes two proposals for penalized estimation of the transition coefficients of the NHMM. As mentioned before, and so far we have found no examples of works dealing with penalized estimation of the transition coefficients of a NHMM. Due to this fact, and to be able to establish a frame of reference, we now define a list of terms used in the simulation study.

- **Individual LASSO**: All results labeled as "Individual LASSO" refer to the output of the algorithm in which the regression coefficients of each hidden state are individually penalized.

- **Global LASSO**: All results labeled as "Global LASSO" refer to the output of the algorithm in which the regressions for all hidden states are jointly penalized.

- **No-LASSO**: Results showing the label "No-LASSO model" refer to results in which the value of the penalization parameter $\lambda$ is zero, which is the equivalent of applying no penalization on the coefficients;

- **All Lambdas**: To establish one more point of comparison throughout the simulation study, we have collected the metrics mentioned in Section 4.1 for all the fitted models with the different values of $\lambda$ used in each replication. For all these models with different values of $\lambda$, we collect and record all the described metrics, and calculate the mean, median and standard deviation(SD) of these metrics in every replication. We call these combined results "All Lambdas".

As established earlier, all metrics are collected over the course of 30 replications in order to have a considerable sample size.

For all simulation scenarios, the initial values of the transition parameters for the estimation algorithm are randomly generated from a $N(0,3)$ distribution. The non-observable sequence is initialized by randomly generating a value from 1 to $K$ for each position along the sequence.

### A.1.1 Scenario # 1: K=2 and D=6

For this first scenario, we use 6 covariates, of which three will be set to zero. The real values of the parameters for the probability distribution of the observable random variables have been set as follows: $\mu_1 = 30$, $\sigma_1 = 1$ and $\mu_2 = 80$, $\sigma_2 = 3$. The real values for the transition coefficients used in this simulation scenario are shown in the following matrix:

$$\boldsymbol{\beta} = \begin{bmatrix} \begin{pmatrix} 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \end{pmatrix} & \begin{pmatrix} 1.5 \\ 6.2 \\ -2.6 \\ 0.0 \\ 0.0 \\ 0.0 \end{pmatrix} \\ \begin{pmatrix} 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \end{pmatrix} & \begin{pmatrix} 1.3 \\ -2.7 \\ 2.5 \\ 0.0 \\ 0.0 \\ 0.0 \end{pmatrix} \end{bmatrix}. \tag{A.1}$$

It is important to remember that since we are using the mlogit link function, then the coefficients related to the transitions from the first non-observable state are all set to zero, in order to achieve parameter identifiability. The range for $\lambda$ was chosen for Global LASSO ranged from 0 to 23.

Table 40 – Estimation results for the parameters of the observable distributions.

| $T$ | **Parameter** | **Real** | **Median** | **Mean** | **SD** | **Bias** | **MSE** | **95% CI** |
|---|---|---|---|---|---|---|---|---|
| 250 | $\mu_1$ | 30 | 29.96 | 29.97 | 0.12 | 0.03 | 0.02 | (29.77, 30.20) |
| | $\mu_2$ | 80 | 79.92 | 79.94 | 0.22 | 0.06 | 0.05 | (79.55, 80.32) |
| | $\sigma_1$ | 1 | 1.01 | 1.01 | 0.07 | -0.01 | 0.00 | (0.89, 1.14) |
| | $\sigma_2$ | 2 | 2.98 | 2.97 | 0.17 | 0.03 | 0.03 | (2.64, 3.29) |
| 500 | $\mu_1$ | 30 | 29.98 | 29.98 | 0.07 | 0.02 | 0.00 | (29.86, 30.09) |
| | $\mu_2$ | 80 | 80.01 | 80.04 | 0.17 | -0.04 | 0.03 | (79.75, 80.30) |
| | $\sigma_1$ | 1 | 1.01 | 1.00 | 0.06 | 0.00 | 0.00 | (0.89, 1.12) |
| | $\sigma_2$ | 2 | 3.01 | 3.01 | 0.10 | -0.01 | 0.01 | (2.82, 3.17) |
| 1000 | $\mu_1$ | 30 | 30,00 | 29,99 | 0,05 | 0,01 | 0,00 | (29.91, 30.08) |
| | $\mu_2$ | 80 | 80,02 | 80,02 | 0,14 | -0,02 | 0,02 | (79.78, 80.26) |
| | $\sigma_1$ | 1 | 0,98 | 0,99 | 0,04 | 0,01 | 0,00 | (0.94, 1.07) |
| | $\sigma_2$ | 2 | 3,01 | 3,00 | 0,11 | 0,00 | 0,01 | (2.78, 3.16) |

To ensure that the vector of tuning parameters was sufficiently fine, 500 values were sequentially generated within this range. This guaranteed a fine penalization grid to carry out tuning of the penalization parameter to be applied on the transition coefficients.

The range for $\lambda$ chosen for the Individual LASSO also ranged from 0 to 23. However, since penalization is carried out individually for each hidden state, we use a smaller sequence of values for each $\lambda_i$, because of computational costs. For the range from 0 to 23, we sequentially generate 50 values of $\lambda_i$. If we have $K$ hidden states, then 50 values of $\lambda_i$ for each hidden state means that $50^K$ models will be fitted and tested. For $K = 2$, this is still viable, but as the number of hidden states increases, tuning with a greater amount of $\lambda$ becomes computationally non-viable.

Table 40 shows the results for the estimation of the parameters of the probability distribution for the observable random variables. We notice that estimation results are good for this particular scenario for all chain lengths. MSE for the parameters seems to decrease as chain length increases, as well as the length of the 95% confidence interval. This shows that estimation of this set of parameters is carried out successfully and with no problems.

Table 41 shows the results of the MSPE on the test data set for the Global LASSO. The first interesting point is that models using the Global LASSO with the best selected $\lambda$ have median MSPE which is consistently smaller than the MSPE of No-LASSO models. On average, all fitted models(all lambdas) have MSPE which is noticeably larger than the average for models using the best value of $\lambda$. This is evidence that the proposed penalization method using the best $\lambda$ greatly improves the predictive performance of the NHMM.

A similar situation occurs when we observe the results in Table 42, in which the Individual LASSO consistently obtains lower MSPE values for all chain lengths, in comparison to all lambda models and to the no-LASSO model. When we compare the Individual LASSO result in Table 41 with the Global LASSO results in Table 42, we notice that the Individual LASSO has achieved

Table 41 – MSPE results for the Global LASSO, No-LASSO and for the mean of all lambdas.

| | **Global LASSO** | | | **No-LASSO** | | | **All Lambdas** | | |
|---|---|---|---|---|---|---|---|---|---|
| *T* | *Median* | *Mean* | *SD* | *Median* | *Mean* | *SD* | *Median* | *Mean* | *SD* |
| 250 | 284.15 | 288.64 | 100.94 | 682.82 | 765.53 | 297.48 | 659.78 | 698.78 | 138.35 |
| 500 | 410.27 | 384.49 | 100.99 | 692.21 | 687.27 | 258.15 | 642.14 | 640.61 | 113.21 |
| 1000 | 469.53 | 470.85 | 84.65 | 725.54 | 712.66 | 137.13 | 686.86 | 673.68 | 103.39 |

Table 42 – MSPE results for the Individual LASSO, No-LASSO and for the mean of all lambdas.

| | **Individual LASSO** | | | **No-LASSO** | | | **All Lambdas** | | |
|---|---|---|---|---|---|---|---|---|---|
| *T* | *Median* | *Mean* | *SD* | *Median* | *Mean* | *SD* | *Median* | *Mean* | *SD* |
| 250 | 287.37 | 279.30 | 104.44 | 682.82 | 765.53 | 297.48 | 672.79 | 700.49 | 141.21 |
| 500 | 413.55 | 383.90 | 93.83 | 692.21 | 687.27 | 258.15 | 645.22 | 643.45 | 114.63 |
| 1000 | 470.72 | 474.98 | 85.14 | 725.54 | 712.66 | 137.13 | 694.17 | 676.93 | 102.60 |

a slightly lower mean MSPE in some chain lengths. However, in general the results are quite similar for both of the proposals, and in all chain lengths the median MSPE is lower for the Global LASSO.

Tables 43 and 44 show results for the amount of hits when predicting **S** test sequence. Results for proposals with their best selected $\lambda$ show excellent performance in predicting **S** test sequence. It is important to remember that it is non-viable to calculate this metric in a real data setting, but in this simulated setting, we have calculated this metric to have one more characteristic upon which to judge the proposals' performances. A curious phenomenon which occurs in both proposals is that as chain length increases, we observe a slight drop in predictive accuracy of the **S** test sequence. Our hypothesis regarding this phenomenon was that as the sequence length increases (since we always consider the final 20% of the observations as test data set), there is an expected drop in accuracy. In the case of $T = 1000$, the test data set size will be 200, representing a prediction window 4 times greater than that of a $T = 250$, where the prediction window will have a size of 50 observations.

To verify this, we randomly selected a replication from the simulations where $T = 1000$. The selected replication was replication 6. Instead of measuring the accuracy in predicting the full **S** test sequence, we measured the accuracy in predicting the first 50 hidden states of the **S** test sequence. When carrying out this small test, our hypothesis was confirmed, given that the predictive accuracy for the Global LASSO was 0.87 on the first 50 hidden states of the **S** test sequence, and 0.89 for the Individual LASSO. In other words, the proposed methods retain high predictive accuracy of the **S** test sequence when the prediction window is shorter, as is natural in most time-series models.

In either one of the two cases, Individual and Global LASSO with the best $\lambda$ clearly outperform a model with no penalization, effectively showing the improvement in predictive accuracy which the two proposals bring in the context of NHMMs, specifically in a hypothetical scenario when attempting to predict the **S** test sequence. For this set of metrics, the Global

Table 43 – Hit frequency when predicting **S** sequence - results for the Global LASSO, No-LASSO and for all lambdas.

| | **Global LASSO** | | | **No-LASSO** | | | **All Lambdas** | | |
|---|---|---|---|---|---|---|---|---|---|
| *T* | *Median* | *Mean* | *SD* | *Median* | *Mean* | *SD* | *Median* | *Mean* | *SD* |
| 250 | 0.88 | 0.87 | 0.06 | 0.68 | 0.65 | 0.13 | 0.65 | 0.65 | 0.08 |
| 500 | 0.83 | 0.82 | 0.06 | 0.67 | 0.67 | 0.13 | 0.66 | 0.67 | 0.06 |
| 1000 | 0.77 | 0.77 | 0.05 | 0.66 | 0.67 | 0.06 | 0.66 | 0.67 | 0.05 |

Table 44 – Hit frequency when predicting **S** sequence - results for the Individual LASSO, No-LASSO and for all lambdas.

| | **Individual LASSO** | | | **No-LASSO** | | | **All Lambdas** | | |
|---|---|---|---|---|---|---|---|---|---|
| *T* | *Median* | *Mean* | *SD* | *Median* | *Mean* | *SD* | *Median* | *Mean* | *SD* |
| 250 | 0.86 | 0.87 | 0.07 | 0.68 | 0.65 | 0.13 | 0.66 | 0.64 | 0.08 |
| 500 | 0.82 | 0.82 | 0.06 | 0.67 | 0.67 | 0.13 | 0.67 | 0.67 | 0.06 |
| 1000 | 0.77 | 0.76 | 0.05 | 0.66 | 0.67 | 0.06 | 0.66 | 0.66 | 0.06 |

Table 45 – Shrinkage of the transition coefficients - results for the Global LASSO, No-LASSO and for all lambdas.

| | **Sensitivity** | | | **Specificity** | | | **Accuracy** | | |
|---|---|---|---|---|---|---|---|---|---|
| *T* | *Median* | *Mean* | *SD* | *Median* | *Mean* | *SD* | *Median* | *Mean* | *SD* |
| 250 | 0.50 | 0.47 | 0.28 | 1.00 | 0.96 | 0.08 | 0.67 | 0.71 | 0.13 |
| 500 | 0.50 | 0.47 | 0.25 | 1.00 | 0.99 | 0.04 | 0.75 | 0.73 | 0.12 |
| 1000 | 0.50 | 0.52 | 0.20 | 1.00 | 1.00 | 0.00 | 0.75 | 0.76 | 0.10 |

LASSO has a slightly better performance than the Individual LASSO.

The set of metrics which is shown in Tables 45 and 46 refers to the proposals' capacity to effectively shrink the transition coefficients to a value of zero. As previously explained, a "binary approach" was adopted when evaluating the proposals' capacity to shrink coefficients. As we can perceive, the proposals achieve an intermediate rate of shrinkage of coefficients, attaining a maximum mean specificity of 50% in any chain length. In other words, on average, the Global LASSO effectively sets 50% of the transition coefficients to zero, whose real value is zero. It is important to notice that in a situation where $D = 6$ and $K = 2$, we are trying to perform shrinkage on 12 transition coefficients. One reason for which the rate of shrinkage is not higher could be that transition coefficients related to some transition (for example, transition from hidden state 1 to hidden state 2), may have not been subjected to significant shrinkage due to the fact that such transition was not observed a reasonable amount of times, as to perform such shrinkage.

Even though shrinkage of some coefficients whose actual value is zero did not occur, the coefficients of the model with the best selected $\lambda$ show outstanding performance in predicting both the **S** test sequence, as well as consistently obtaining low MSPE values. This statement is holds true for both of the proposal.

Finally, Tables 47 and 48 show processing times for the proposals of this chapter. As

Table 46 – Shrinkage of the transition coefficients - results for the Individual LASSO, No-LASSO and for all lambdas.

| | **Sensitivity** | | | **Specificity** | | | **Accuracy** | | |
|---|---|---|---|---|---|---|---|---|---|
| *T* | *Median* | *Mean* | *SD* | *Median* | *Mean* | *SD* | *Median* | *Mean* | *SD* |
| 250 | 0.33 | 0.37 | 0.25 | 1.00 | 0.98 | 0.06 | 0.67 | 0.67 | 0.12 |
| 500 | 0.50 | 0.43 | 0.24 | 1.00 | 0.97 | 0.06 | 0.71 | 0.70 | 0.12 |
| 1000 | 0.50 | 0.42 | 0.29 | 1.00 | 1.00 | 0.00 | 0.75 | 0.71 | 0.14 |

Table 47 – Processing times for the Global LASSO, where total time considers running the $R = 30$ replications.

| *T* | **Avg Time/Replication** | **SD time/Replication** | **Total Time** |
|---|---|---|---|
| 250 | 0.20 mins | 0.002 mins | 6.01 mins |
| 500 | 0.33 mins | 0.003 mins | 9.96 mins |
| 1000 | 0.59 mins | 0.004 mins | 17.61 mins |

Table 48 – Processing times for the Individual LASSO, where total time considers running the $R = 30$ replications.

| *T* | **Avg Time/Replication** | **SD time/Replication** | **Total Time** |
|---|---|---|---|
| 250 | 1.02 mins | 0.007 mins | 30.89 mins |
| 500 | 1.67 mins | 0.072 mins | 50.11 mins |
| 1000 | 2.84 mins | 0.086 mins | 85.44 mins |

expected, the Global LASSO has a smaller processing time. This is due to the fact the amount of models that will be tested to find the best $\lambda$ is exactly the same as the amount of values in the $\lambda$ vector. This is not true for the Individual LASSO, because the amount of models to be tested will depend on the amount of values in the $\lambda$ grid, as well as the amount of hidden states in the sequence.

To exemplify these situations, if a grid of 50 values of $\lambda$ is used to with the Global LASSO, then 50 models will be tested. If the same grid is used for the Individual LASSO, the $50^K$ models will be tested. If we are in a setting which assumes 3 hidden states, then we will be testing 125,000 models, which will evidently require more processing time.

### A.1.2 Scenario 2: K=2 and D=10

In the second scenario, we have 10 covariates for each non-observable state, of which 7 will be set to zero. The real values of the parameters for the probability distribution of the observable random variables are established as follows: $\mu_1 = 30$, $\sigma_1 = 1$ and $\mu_2 = 80$, $\sigma_2 = 3$. The real values for the transition coefficients used in this simulation scenario are shown in the

following matrix:

$$\boldsymbol{\beta} = \begin{bmatrix} \begin{pmatrix} 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \end{pmatrix} & \begin{pmatrix} 1.5 \\ 6.2 \\ -2.6 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \end{pmatrix} \\ \begin{pmatrix} 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \end{pmatrix} & \begin{pmatrix} 1.3 \\ -2.7 \\ 2.5 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \end{pmatrix} \end{bmatrix}. \tag{A.2}$$

Once again, we must point out that since we are using the mlogit link function, then the coefficients related to the transitions from the first non-observable state are all set to zero, in order to achieve parameter identifiability. The range for $\lambda$ was chosen for Global LASSO ranged from 0 to 23. To ensure that the vector of tuning parameters was sufficiently fine, 500 values were sequentially generated within this range. This guarantees a fine penalization grid to perform penalization parameter tuning.

The values for $\lambda$ chosen for the Individual LASSO also ranged from 0 to 23. Because of the reasons explained in Section A.1.1 (Scenario 1), we sequentially generate 50 values of $\lambda_i$ for the selected range from 0 to 23. We verified that this range contained a value of $\lambda$ which is optimal for tuning.

Table 49 shows the results for the estimation of the parameters of the probability distribution for the observable random variables. Estimation results are very good for this scenario for all chain lengths. MSE for the parameters seems to decrease as chain length increases, as well as the length of the 95% confidence interval. This shows that estimation of this set of parameters is carried out satisfactorily. For the simulation where $T = 500$, we observed a single replication, replication 23, in which there was high variability when estimating the parameters. However, this was an isolated case, and for the remaining 29 replications, the algorithm executed successfully

Table 49 – Estimation results for the parameters of the observable distributions.

| $T$ | Parameter | Real | Median | Mean | SD | Bias | MSE | 95% CI |
|---|---|---|---|---|---|---|---|---|
| 250 | $\mu_1$ | 30 | 30.03 | 30.02 | 0.11 | -0.02 | 0.01 | (29.81, 30.19) |
| | $\mu_2$ | 80 | 80.12 | 80.02 | 0.30 | -0.02 | 0.09 | (79.44, 80.39) |
| | $\sigma_1$ | 1 | 1.01 | 1.01 | 0.07 | -0.01 | 0.01 | (0.87, 1.11) |
| | $\sigma_2$ | 3 | 2.95 | 2.98 | 0.21 | 0.02 | 0.04 | (2.67, 3.34) |
| 500 | $\mu_1$ | 30 | 30.00 | 31.10 | 5.62 | -1.10 | 32.82 | (29.82, 39.83) |
| | $\mu_2$ | 80 | 80.00 | 79.40 | 3.53 | 0.60 | 12.82 | (74.42, 80.55) |
| | $\sigma_1$ | 1 | 1.01 | 2.09 | 4.54 | -1.09 | 21.80 | (0.93, 14.11) |
| | $\sigma_2$ | 3 | 3.02 | 3.69 | 3.86 | -0.69 | 15.41 | (2.71, 8.96) |
| 1000 | $\mu_1$ | 30 | 30.00 | 30.01 | 0.05 | -0.01 | 0.00 | (29.91, 30.08) |
| | $\mu_2$ | 80 | 79.98 | 79.97 | 0.13 | 0.03 | 0.02 | (79.71, 80.23) |
| | $\sigma_1$ | 1 | 1.00 | 1.00 | 0.04 | 0.00 | 0.00 | (0.93, 1.08) |
| | $\sigma_2$ | 3 | 3.00 | 3.01 | 0.10 | -0.01 | 0.01 | (2.80, 3.18) |

Table 50 – MSPE results for the Global LASSO, No-LASSO and for the mean of all lambdas.

| | **Global LASSO** | | | **No-LASSO** | | | **All Lambdas** | | |
|---|---|---|---|---|---|---|---|---|---|
| $T$ | *Median* | *Mean* | *SD* | *Median* | *Mean* | *SD* | *Median* | *Mean* | *SD* |
| 250 | 251.44 | 272.86 | 176.08 | 666.11 | 763.60 | 324.32 | 786.81 | 803.95 | 136.91 |
| 500 | 448.02 | 468.72 | 109.67 | 786.35 | 803.39 | 160.36 | 822.38 | 835.79 | 116.82 |
| 1000 | 495.40 | 501.83 | 97.05 | 782.53 | 772.05 | 136.55 | 794.30 | 792.79 | 120.57 |

Table 51 – MSPE results for the Individual LASSO, No-LASSO and the mean of all lambdas.

| | **Individual LASSO** | | | **No-LASSO** | | | **All Lambdas** | | |
|---|---|---|---|---|---|---|---|---|---|
| $T$ | *Median* | *Mean* | *SD* | *Median* | *Mean* | *SD* | *Median* | *Mean* | *SD* |
| 250 | 267.96 | 281.31 | 167.47 | 666.11 | 763.60 | 324.32 | 824.09 | 812.29 | 144.01 |
| 500 | 469.13 | 457.48 | 103.50 | 786.35 | 803.39 | 160.36 | 812.97 | 830.92 | 113.92 |
| 1000 | 495.80 | 481.54 | 98.82 | 782.53 | 772.05 | 136.55 | 800.69 | 791.43 | 118.80 |

and with no problems. This situation impacted the results of SD, Bias, MSE and CI but not the final point estimates of the parameters, specially when considering the median.

Table 50 shows the results of the MSPE on the test data set for the Global LASSO. We observe that models using the best value of $\lambda$ have MSPE which is perceivably smaller than the MSPE of No-LASSO models, and also have MSPE which is much lower than the average for all models fitted in any replication. This is clearly shows that the proposed penalization method shows better predictive performance of the NHMM.

Table 51 shows similar results to those observed in Table 50, in which the Individual LASSO, in general, obtains lower MSPE values in comparison to all lambdas models and to the no-LASSO model. When the results for the two proposals are compared, we notice that the Individual LASSO has achieved a slightly lower mean MSPE for some chain lengths. However, results in this set of metrics are very similar, and some of the median MSPE values are lower for the Individual LASSO .

Table 52 – Hit frequency when predicting **S** sequence - results for the Global LASSO, No-LASSO and for all lambdas.

| | **Global LASSO** | | | **No-LASSO Model** | | | **All Lambdas Models** | | |
|---|---|---|---|---|---|---|---|---|---|
| *T* | *Median* | *Mean* | *SD* | *Median* | *Mean* | *SD* | *Median* | *Mean* | *SD* |
| 250 | 0.90 | 0.89 | 0.08 | 0.72 | 0.67 | 0.14 | 0.65 | 0.64 | 0.05 |
| 500 | 0.81 | 0.78 | 0.08 | 0.65 | 0.65 | 0.07 | 0.64 | 0.63 | 0.05 |
| 1000 | 0.78 | 0.78 | 0.05 | 0.66 | 0.66 | 0.06 | 0.64 | 0.65 | 0.05 |

Table 53 – Hit frequency when predicting **S** sequence - results for the Individual LASSO, No-LASSO and for all lambdas.

| | **Individual LASSO** | | | **No-LASSO** | | | **All Lambdas** | | |
|---|---|---|---|---|---|---|---|---|---|
| *T* | *Median* | *Mean* | *SD* | *Median* | *Mean* | *SD* | *Median* | *Mean* | *SD* |
| 250 | 0.88 | 0.87 | 0.08 | 0.72 | 0.67 | 0.14 | 0.64 | 0.64 | 0.06 |
| 500 | 0.80 | 0.78 | 0.08 | 0.65 | 0.65 | 0.07 | 0.64 | 0.63 | 0.05 |
| 1000 | 0.79 | 0.79 | 0.04 | 0.66 | 0.66 | 0.06 | 0.64 | 0.65 | 0.05 |

Tables 52 and 53 show results for the amount of hits when predicting **S** test sequence. Results for both the Global and Individual LASSO with their best selected $\lambda$ show good performance in predicting **S** test sequence. A slight drop in predictive accuracy of the **S** test sequence as chain length increases was also observed in this scenario, just as in Section A.1.1 (Scenario 1).

Once again, we randomly selected a replication from the simulations where $T = 1000$. The selected replication was replication 17. Instead of measuring the accuracy in predicting the full **S** test sequence, we measured the accuracy in predicting the first 50 hidden states of the **S** test sequence. When carrying out this small test, the predictive accuracy for the Global LASSO was 0.88 on the first 50 hidden states of the **S** test sequence, and 0.91 for the Individual LASSO. The proposed methods retain high predictive accuracy of the **S** test sequence when the prediction window is shorter, as is common in time-series models and as was observed in the previous scenario (Scenario 1).

The Individual and Global LASSO with the best $\lambda$ outperform a model with no penalization, showing a considerable improvement in predictive accuracy in a hypothetical scenario regarding prediction of the **S** test sequence. For this set of metrics, the Global LASSO has a slightly better performance than the Individual LASSO.

The set of metrics which is shown in Tables 54 and 55 refers to the proposals' capacity to effectively shrink the transition coefficients to a value of zero. As we can observe, the proposals did not achieve a high rate of shrinkage of coefficients. However, it is important to understand that in a situation where $D = 10$ and $K = 2$, we are trying to perform shrinkage on 20 transition coefficients. It may be possible that transition coefficients related to a particular transition (for example, transition from hidden state 1 to hidden state 2), may have not been subjected to significant shrinkage due to the fact that such transition was not observed a reasonable amount

Table 54 – Shrinkage of the transition coefficients - results for the Global LASSO, No-LASSO and for all lambdas.

| | **Sensitivity** | | | **Specificity** | | | **Accuracy** | | |
|---|---|---|---|---|---|---|---|---|---|
| *T* | *Median* | *Mean* | *SD* | *Median* | *Mean.* | *SD* | *Median* | *Mean* | *SD* |
| 250 | 0.11 | 0.15 | 0.14 | 1.00 | 0.95 | 0.10 | 0.37 | 0.39 | 0.08 |
| 500 | 0.11 | 0.11 | 0.10 | 1.00 | 0.98 | 0.06 | 0.35 | 0.37 | 0.07 |
| 1000 | 0.14 | 0.13 | 0.09 | 1.00 | 1.00 | 0.00 | 0.40 | 0.39 | 0.06 |

Table 55 – Shrinkage of the transition coefficients - results for the Individual LASSO, No-LASSO and for all lambdas.

| | **Sensitivity** | | | **Specificity** | | | **Accuracy** | | |
|---|---|---|---|---|---|---|---|---|---|
| *T* | *Median* | *Mean* | *SD* | *Median* | *Mean.* | *SD* | *Median* | *Mean* | *SD* |
| 250 | 0.14 | 0.15 | 0.11 | 1.00 | 1.00 | 0.00 | 0.40 | 0.41 | 0.08 |
| 500 | 0.11 | 0.13 | 0.11 | 1.00 | 0.97 | 0.07 | 0.35 | 0.38 | 0.08 |
| 1000 | 0.11 | 0.14 | 0.11 | 1.00 | 1.00 | 0.00 | 0.38 | 0.40 | 0.07 |

Table 56 – Processing times for the Global LASSO, where total time considers running the $R = 30$ replications.

| *T* | **Avg Time/Replication** | **SD time/Replication** | **Total Time** |
|---|---|---|---|
| 250 | 0.21 mins | 0.01 mins | 6.40 mins |
| 500 | 0.35 mins | 0.01 mins | 10.70 mins |
| 1000 | 0.62 mins | 0.01 mins | 18.87 mins |

of times, as to perform such shrinkage.

Another possible explanation for this phenomenon could be related to the metric that was used to select the best value for the tuning parameter. The selected metric was the MSPE. The model with the lowest MSPE on the test data set was selected as the best model, however, shrinkage of the transition coefficients is related to predicting the most likely **S** sequence. Perhaps this difference in criteria could have led to the selection of a model with the lowest MSPE, whose transition coefficients may not have necessarily been shrunken all the way to zero. However, more testing will be necessary to verify either of these hypotheses.

Despite the previously mentioned facts, we can clearly observe that even though full shrinkage of all coefficients whose actual value was zero did not occur, the coefficients of the model with the best selected $\lambda$ show excellent performance in predicting both the **S** test sequence, as well as consistently obtaining low MSPE values. This statement is valid for both of the proposals.

Finally, Tables 56 and 57 show processing times for the proposals of this chapter. As expected, the Global LASSO has a smaller processing time. As was mentioned in the previously presented scenario, this is due to the fact the amount of models that will be tested to find the best $\lambda$ is $50^K$ in the Individual LASSO, while in the Global LASSO it will be 500.

If a smaller tuning grid is selected, the Individual LASSO might present advantages

Table 57 – Processing times for the Individual LASSO, where total time considers running the $R = 30$ replications.

| $T$ | **Avg Time/Replication** | **SD time/Replication** | **Total Time** |
|---|---|---|---|
| 250 | 1.04 mins | 0.01 mins | 31.30 mins |
| 500 | 1.61 mins | 0.07 mins | 48.43 mins |
| 1000 | 2.78 mins | 0.09 mins | 83.5 mins |

in processing, given that penalization is realized individually for each non-observable state. However this depends on optimizing the selection of the values in the tuning grid. As mentioned before one of the future proposals is establishing a method to select an optimum range of values for the tuning parameters.

### A.1.3 Scenario 3: K=3 and D=6

For this third scenario, we consider a situation with 3 non-observable states. The real values of the parameters for the probability distribution of the observable random variables are set as follows: $\mu_1 = 20$, $\sigma_1 = 1$; $\mu_2 = 150$, $\sigma_2 = 2$ and $\mu_3 = 250$, $\sigma_3 = 3$. The real values for the transition coefficients used to simulate data in this scenario are shown in the following matrix:

$$\boldsymbol{\beta} = \begin{bmatrix} \begin{pmatrix} 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \end{pmatrix} & \begin{pmatrix} -1.5 \\ 1.5 \\ 2.6 \\ 0.0 \\ 0.0 \\ 0.0 \end{pmatrix} & \begin{pmatrix} 1.3 \\ 3.2 \\ 2.4 \\ 0.0 \\ 0.0 \\ 0.0 \end{pmatrix} \\ \begin{pmatrix} 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \end{pmatrix} & \begin{pmatrix} -2.0 \\ 4.6 \\ 1.4 \\ 0.0 \\ 0.0 \\ 0.0 \end{pmatrix} & \begin{pmatrix} -3.3 \\ 1.7 \\ 1.3 \\ 0.0 \\ 0.0 \\ 0.0 \end{pmatrix} \\ \begin{pmatrix} 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \end{pmatrix} & \begin{pmatrix} 2.4 \\ 1.1 \\ -1.5 \\ 0.0 \\ 0.0 \\ 0.0 \end{pmatrix} & \begin{pmatrix} -2.3 \\ -2.7 \\ -3.5 \\ 0.0 \\ 0.0 \\ 0.0 \end{pmatrix} \end{bmatrix}. \tag{A.3}$$

As is the case in the previous scenarios, we are using the mlogit link function, therefore, the coefficients related to the transitions from the first non-observable state are all set to zero, in

Table 58 – Estimation results for the parameters of the observable distributions.

| $T$ | Parameter | Real | Median | Mean | SD | Bias | MSE | 95% CI |
|---|---|---|---|---|---|---|---|---|
| 250 | $\mu_1$ | 20.00 | 20.02 | 20.01 | 0.10 | -0.01 | 0.01 | (19.86, 20.15) |
| | $\mu_2$ | 150.00 | 150.05 | 150.06 | 0.16 | -0.06 | 0.03 | (149.80, 150.33) |
| | $\mu_3$ | 250.00 | 250.07 | 250.06 | 0.42 | -0.06 | 0.18 | (249.28, 250.84) |
| | $\sigma_1$ | 1.00 | 0.96 | 0.99 | 0.10 | 0.01 | 0.01 | (0.84, 1.20) |
| | $\sigma_2$ | 2.00 | 2.05 | 2.04 | 0.15 | -0.04 | 0.02 | (1.77, 2.25) |
| | $\sigma_3$ | 3.00 | 2.98 | 2.94 | 0.23 | 0.06 | 0.06 | (2.54, 3.29) |
| 500 | $\mu_1$ | 20.00 | 20.00 | 20.00 | 0.09 | 0.00 | 0.01 | (19.85, 20.15) |
| | $\mu_2$ | 150.00 | 150.02 | 150.04 | 0.17 | -0.04 | 0.03 | (149.78, 150.39) |
| | $\mu_3$ | 250.00 | 250.06 | 250.02 | 0.32 | -0.02 | 0.10 | (249.42, 250.47) |
| | $\sigma_1$ | 1.00 | 0.99 | 0.98 | 0.06 | 0.02 | 0.00 | (0.89, 1.09) |
| | $\sigma_2$ | 2.00 | 2.04 | 2.03 | 0.12 | -0.03 | 0.02 | (1.84, 2.24) |
| | $\sigma_3$ | 3.00 | 2.94 | 2.95 | 0.17 | 0.05 | 0.03 | (2.69, 3.30) |
| 1000 | $\mu_1$ | 20.00 | 19.99 | 19.98 | 0.07 | 0.02 | 0.00 | (19.82, 20.07) |
| | $\mu_2$ | 150.00 | 150.00 | 150.00 | 0.09 | 0.00 | 0.01 | (149.88, 150.23) |
| | $\mu_3$ | 250.00 | 250.11 | 250.06 | 0.24 | -0.06 | 0.06 | (249.65, 250.45) |
| | $\sigma_1$ | 1.00 | 1.00 | 1.00 | 0.03 | 0.00 | 0.00 | (0.93, 1.05) |
| | $\sigma_2$ | 2.00 | 1.99 | 1.99 | 0.08 | 0.01 | 0.01 | (1.83, 2.11) |
| | $\sigma_3$ | 3.00 | 3.02 | 3.00 | 0.14 | 0.00 | 0.02 | (2.72, 3.20) |

order to achieve parameter identifiability.

The range of values for $\lambda$ that was chosen for Global LASSO ranged from 0 to 23. To ensure that the vector of tuning parameters was sufficiently fine, 500 values were sequentially generated within this range. This guarantees a fine penalization grid to perform penalization parameter tuning.

The values for $\lambda_i$ chosen for the Individual LASSO also ranged from 0 to 23. We sequentially generate 20 values of for each $\lambda_i$ for the selected range from 0 to 23. For this scenario, using 20 different values for each $\lambda_i$ means that $20^3$ different models will be tested. As was described in Section A.1.1, we verified that this range contained a value of $\lambda$ which is optimal for tuning.

Table 58 shows the results for the estimation of the parameters of the probability distribution for the observable random variables. Estimation results are excellent for this scenario for all chain lengths. MSE for the parameters seems to gradually decrease as chain length increases, as does the standard deviation and the magnitude of the bias. This clearly shows that estimation of this set of parameters is carried out successfully.

Table 59 shows the results of the MSPE on the test data set for the Global LASSO. The same pattern that was observed in previous scenarios occurs in this scenario, where models using the best value of $\lambda$ have MSPE which is noticeably smaller than the MSPE of models with no penalization, and also have MSPE which is considerably smaller than the average for all models fitted in any replication. This is valid for all chain lenghts, and demonstrates that the proposed

Table 59 – MSPE results for the Global LASSO, No-LASSO and for the mean of all lambdas.

| | **Global LASSO** | | | **No-LASSO** | | | **All Lambdas** | | |
|---|---|---|---|---|---|---|---|---|---|
| *T* | *Median* | *Mean* | *SD* | *Median* | *Mean* | *SD* | *Median* | *Mean* | *SD* |
| 250 | 3924.98 | 4059.01 | 1363.08 | 7977.80 | 8149.09 | 2395.37 | 9157.59 | 9081.98 | 1590.26 |
| 500 | 4995.57 | 5050.11 | 1162.50 | 8165.94 | 8782.73 | 2400.15 | 8600.79 | 8852.72 | 1551.54 |
| 1000 | 6316.17 | 6178.25 | 893.80 | 8352.59 | 8324.82 | 959.26 | 8733.86 | 8702.09 | 1038.70 |

Table 60 – MSPE results for the Individual LASSO, No-LASSO and for the mean of all lambdas.

| | **Individual LASSO** | | | **No-LASSO** | | | **All Lambdas** | | |
|---|---|---|---|---|---|---|---|---|---|
| *T* | *Median* | *Mean* | *SD* | *Median* | *Mean* | *SD* | *Median* | *Mean* | *SD* |
| 250 | 3719.54 | 3793.49 | 1281.21 | 7977.80 | 8149.09 | 2395.37 | 9220.43 | 9169.43 | 1668.98 |
| 500 | 4444.61 | 4870.91 | 1318.90 | 8165.94 | 8782.73 | 2400.15 | 8629.00 | 8876.67 | 1599.37 |
| 1000 | 5770.96 | 5897.23 | 864.32 | 8352.59 | 8324.82 | 959.26 | 8729.77 | 8695.93 | 1081.17 |

Table 61 – Hit frequency when predicting **S** sequence - results for the Global LASSO, No-LASSO and for all lambdas.

| | **Global LASSO** | | | **No-LASSO** | | | **All Lambdas** | | |
|---|---|---|---|---|---|---|---|---|---|
| *T* | *Median* | *Mean* | *SD* | *Median* | *Mean* | *SD* | *Median* | *Mean* | *SD* |
| 250 | 0.78 | 0.78 | 0.07 | 0.61 | 0.63 | 0.12 | 0.54 | 0.55 | 0.06 |
| 500 | 0.76 | 0.75 | 0.06 | 0.64 | 0.62 | 0.08 | 0.58 | 0.58 | 0.06 |
| 1000 | 0.68 | 0.69 | 0.04 | 0.60 | 0.60 | 0.07 | 0.59 | 0.59 | 0.04 |

penalization method shows better predictive performance.

Table 60 shows similar results to those observed in Table 59. In these results, the Individual LASSO obtains lower MSPE values in comparison to all lambda models and to the no-LASSO model for all considered chain lengths. This is a clear indicator of the gain in predictive performance that this proposal brings to the NHMM. When comparing the results of the two proposals, we observe that ice that the Individual LASSO has achieved a slightly lower mean and median MSPE for all chain lengths. This can be attributed to the fact that tuning is done individually for each non-observable state, and therefore, will yield a more finely tuned model than the Global LASSO.

Tables 61 and 62 show results for the amount of hits when predicting **S** test sequence. Results for both the Global and Individual LASSO with their best selected $\lambda$ show indicate a considerable gain in performance when predicting **S** test sequence, compared to a model with no penalization. The same phenomenon which manifests a slight drop in predictive accuracy of the **S** test sequence as chain length increases was also observed in this scenario. However, as was done in previous scenarios, we randomly select a replication from the simulations where $T = 1000$. The selected replication was replication 21. We measure the accuracy in predicting the first 50 hidden states of the **S** test sequence. When carrying out this minor verification, the predictive accuracy for the Global LASSO was 0.78 on the first 50 hidden states of the **S** test sequence, and 0.82 for the Individual LASSO. This shows that the proposed methods achieve higher predictive accuracy of the **S** test sequence when the prediction window is shorter.

Table 62 – Hit frequency when predicting **S** sequence - results for the Individual LASSO, No-LASSO and for all lambdas.

| | **Individual LASSO** | | | **No-LASSO** | | | **All Lambdas** | | |
|---|---|---|---|---|---|---|---|---|---|
| *T* | *Median* | *Mean* | *SD* | *Median* | *Mean* | *SD* | *Median* | *Mean* | *SD* |
| 250 | 0.80 | 0.78 | 0.06 | 0.61 | 0.63 | 0.12 | 0.54 | 0.55 | 0.06 |
| 500 | 0.76 | 0.74 | 0.06 | 0.64 | 0.62 | 0.08 | 0.59 | 0.58 | 0.06 |
| 1000 | 0.72 | 0.71 | 0.04 | 0.60 | 0.60 | 0.07 | 0.59 | 0.60 | 0.04 |

Table 63 – Shrinkage of the transition coefficients - results for the Global LASSO, No-LASSO and for all lambdas.

| | **Sensitivity** | | | **Specificity** | | | **Accuracy** | | |
|---|---|---|---|---|---|---|---|---|---|
| *T* | *Median* | *Mean* | *SD* | *Median* | *Mean* | *SD* | *Median* | *Mean* | *SD* |
| 250 | 0.06 | 0.09 | 0.10 | 1.00 | 0.98 | 0.03 | 0.53 | 0.53 | 0.04 |
| 500 | 0.08 | 0.10 | 0.09 | 1.00 | 0.96 | 0.05 | 0.53 | 0.53 | 0.05 |
| 1000 | 0.06 | 0.08 | 0.07 | 1.00 | 0.98 | 0.03 | 0.53 | 0.53 | 0.04 |

In this third scenario, we observe results which are similar to previous scenarios, having the Individual and Global LASSO with the best $\lambda$ consistently obtain better results than a model with no penalization, by showing a considerable improvement in predictive accuracy in a hypothetical scenario regarding prediction of the **S** test sequence. As observed in the previous set of metrics, the Individual LASSO has a slightly better performance than the Global LASSO, however, in general results are very similar.

Tables 63 and 64 show metrics related to the proposals' capacity to shrink the transition coefficients to a value of zero. In this scenario, the proposals were not very successful in performing shrinkage of coefficients. It is important to point out that in a situation where $D = 6$ and $K = 3$, we are trying to perform shrinkage on 36 transition coefficients. As stated in Scenario A.1.2, it is possible that transition coefficients related to a particular transition may have not been subjected to significant shrinkage due to the fact that such transition was not observed a reasonable amount of times, as to perform such shrinkage. Essentially, this means that there isn't enough data for that specific transition to perform shrinkage. The issue of the choice of metric to select the best value for the tuning parameter also applies here, given that the MSPE is related to the observable values, and not to the non-observable chain. Perhaps this difference in criteria could have led to the selection of a model with the lowest MSPE, whose transition coefficients may not have necessarily been shrunken all the way to zero. We have yet to test either of these hypothesis in a controlled setting.

However, in spite of the fact that the proposals did not achieve full shrinkage of most coefficients whose actual value was zero, the coefficients of the model with the best selected $\lambda$ continue to show satisfactory performance in predicting both the **S** test sequence, as well as consistently obtaining low MSPE values. When comparing either proposal to a model with no penalization, this gain in performance is very evident, and this statement is valid for both of the proposals.

Table 64 – Shrinkage of the transition coefficients - results for the Individual LASSO, No-LASSO and for all lambdas.

| | **Sensitivity** | | | **Specificity** | | | **Accuracy** | | |
|---|---|---|---|---|---|---|---|---|---|
| *T* | *Median* | *Mean* | *SD* | *Median* | *Mean* | *SD* | *Median* | *Mean* | *SD* |
| 250 | 0.17 | 0.13 | 0.09 | 1.00 | 0.97 | 0.04 | 0.56 | 0.55 | 0.04 |
| 500 | 0.06 | 0.08 | 0.08 | 1.00 | 0.97 | 0.03 | 0.53 | 0.53 | 0.04 |
| 1000 | 0.06 | 0.07 | 0.06 | 1.00 | 0.99 | 0.03 | 0.53 | 0.53 | 0.04 |

Table 65 – Processing times for the Global LASSO, where total time considers running the $R = 30$ replications.

| *T* | **Avg Time/Replication** | **SD time/Replication** | **Total Time** |
|---|---|---|---|
| 250 | 0.48 mins | 0.04 mins | 14.36 mins |
| 500 | 0.70 mins | 0.07 mins | 20.93 mins |
| 1000 | 1.21 mins | 0.29 mins | 36.39 mins |

Table 66 – Processing times for the Individual LASSO, where total time considers running the $R = 30$ replications.

| *T* | **Avg Time/Replication** | **SD time/Replication** | **Total Time** |
|---|---|---|---|
| 250 | 6.76 mins | 0.99 mins | 202.87 mins |
| 500 | 9.73 mins | 0.23 mins | 291.90 mins |
| 1000 | 17.43 mins | 6.25 mins | 522.87 mins |

Finally, Tables 65 and 66 show processing times for the proposals of this chapter. The Global LASSO has a shorter processing time, as expected. This is due to the fact the amount of models that will be tested to find the best $\lambda$ is $20^3$ in the Individual LASSO, amounting to 8000 models. In the Global LASSO the amount of models that will be tested is fixed and only depends on the size of the tuning grid, which is a grid of 500 values for all scenarios.

In this particular scenario we begin to perceive the computational costs of using a greater quantity of non-observable states with the Individual LASSO. This problem will be somewhat mitigated with the implementation of a method to optimally select a grid of values for the tuning parameter. This would allow for less values to be tested in the tuning process and would reduce computational costs for the Individual LASSO proposal.

## A.2 Closing Remarks

From this simulation study, we perceive that the proposals introduced in this thesis clearly improve the predictive performance of the NHMM, when considering different sets of evaluation metrics. For the different chain lengths studied in the simulation scenario, both proposals consistently show better performance than a model with no penalization, when analyzing their MSPE. In a hypothetical scenario in which we attempt to predict the **S** sequence, results for both proposals show better performance than a model with no penalization in all scenarios. All scenarios presented a slight drop in predictive accuracy when the chain length increased, but

as we established this is due to greater prediction window for greater chain sizes. Regarding shrinkage of the coefficients, one scenario showed intermediate success in performing shrinkage of the transition coefficients, while two other scenarios showed little success in performing shrinkage. We have several hypotheses to test regarding this less than satisfactory performance in this set of metrics.

When comparing the MSPE, hit frequency and shrinkage results for the two proposals among, we perceive that in general, results are very similar. However, when comparing processing times, the Global LASSO shows clear advantages in processing times. As mentioned before, this is expected, given the structure of the objective function, and the fact that there is only lambda for the regressions of all non-observable states. The two proposals show great promise as methods to improve predictive potential of the NHMM, as well as performing variable selection.

APPENDIX

B

# PLOTS FOR PRECIPITATION DATA SET

Appendix B contains plots for all covariates contained in the data set used in Chapter 5. These plots will aid in understanding the relationship between each covariate and the response variable, Precipitation, as well as help in the interpretation of estimated coefficients for fitted models.

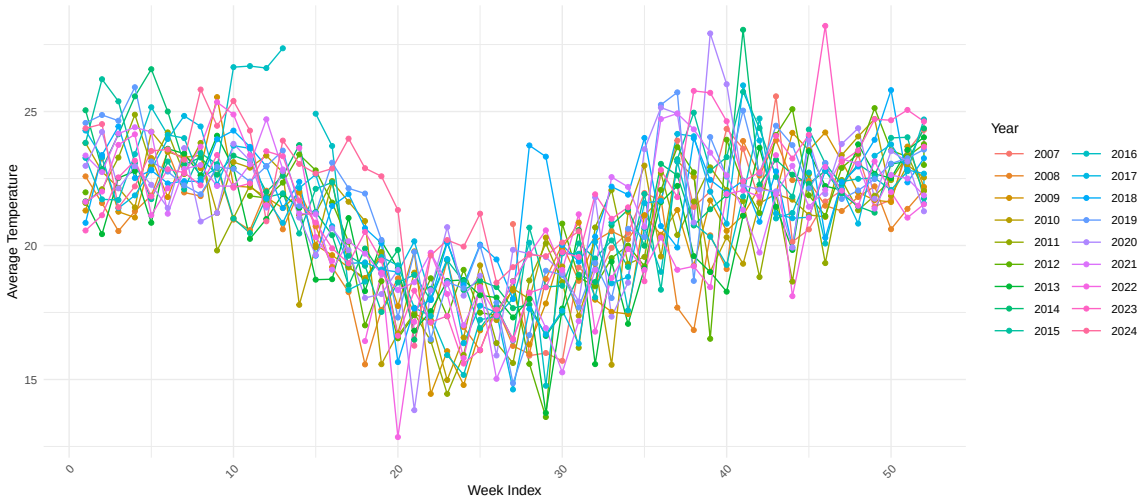



Figure 24 – Average temperatures observed from 2007 through 2024.

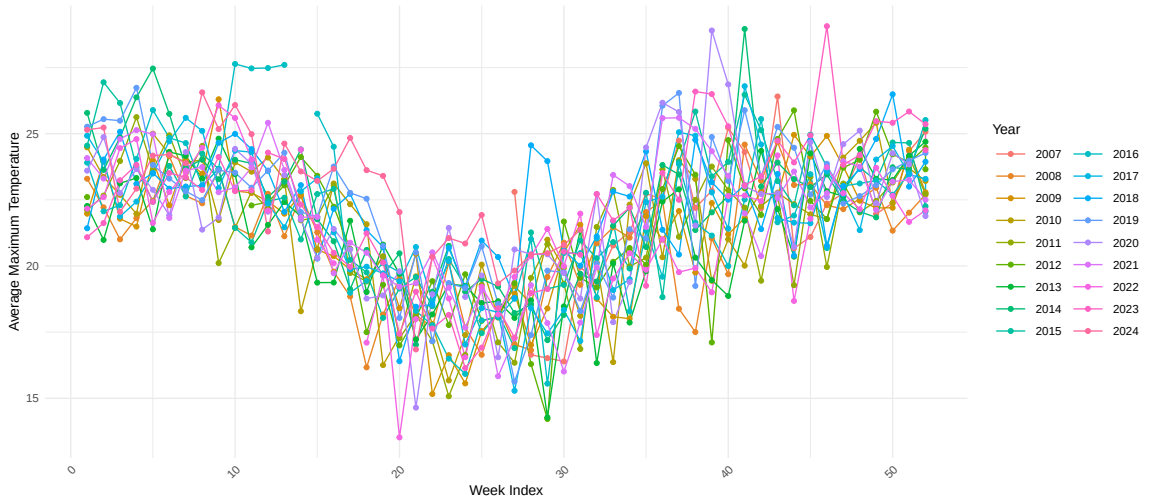



Figure 25 – Average maximum temperatures observed from 2007 through 2024.

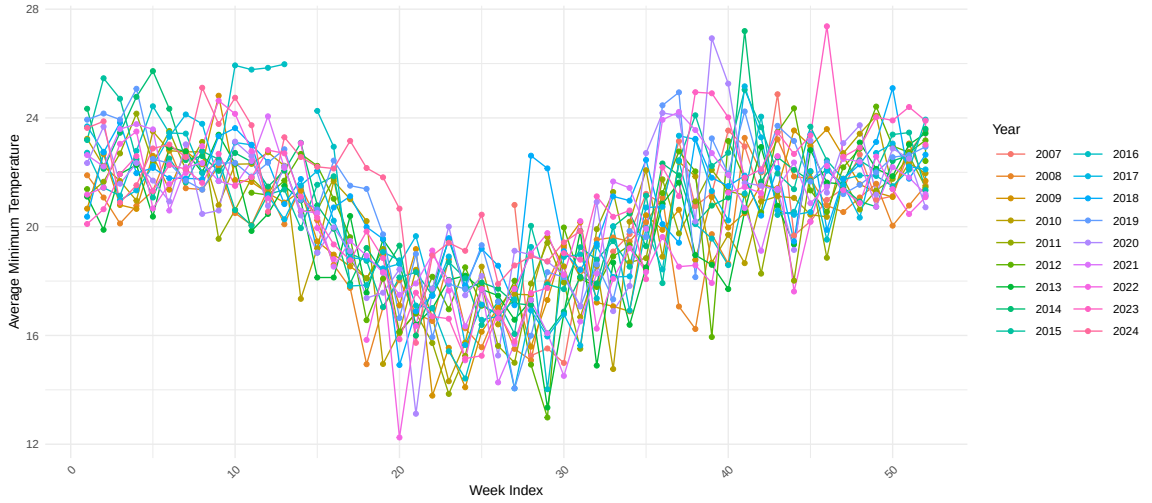



Figure 26 – Average minimum temperatures observed from 2007 through 2024.

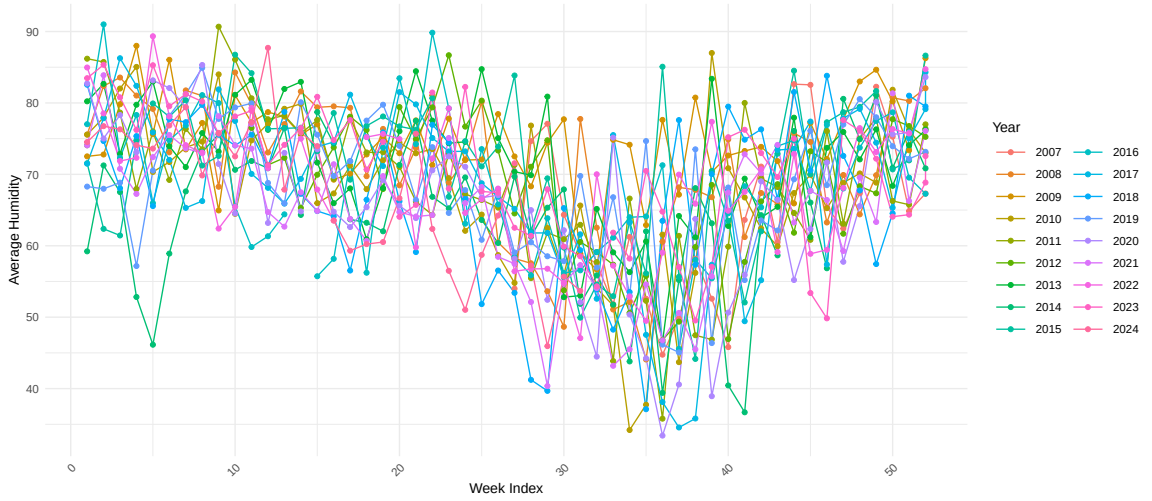



Figure 27 – Average humidity observed from 2007 through 2024.

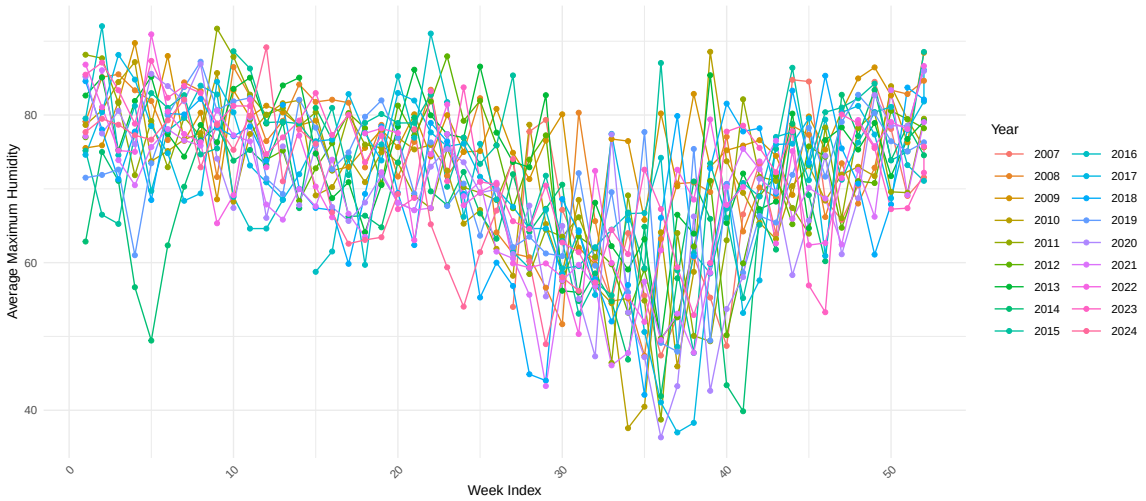



Figure 28 – Average maximum humidity observed from 2007 through 2024.

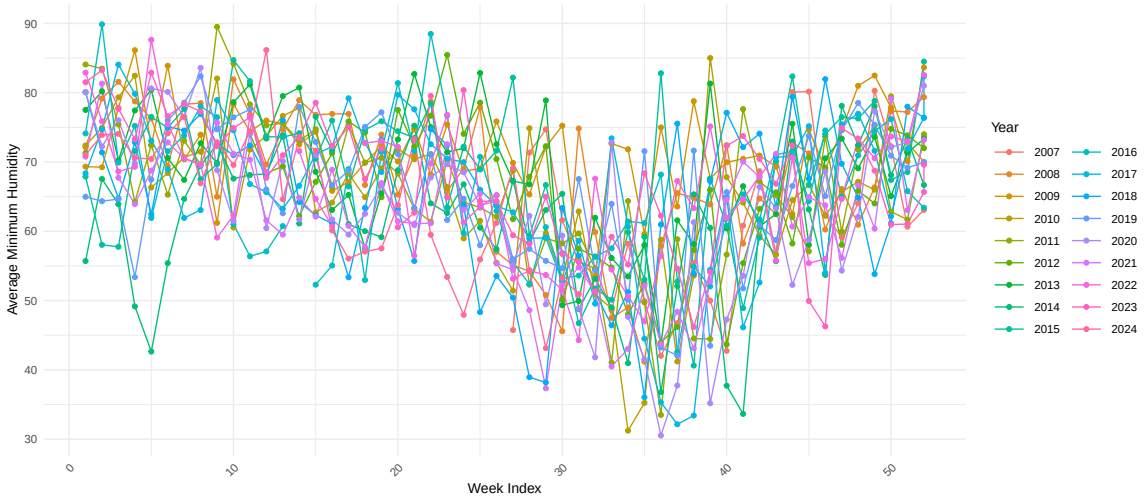



Figure 29 – Average minimum humidity observed from 2007 through 2024.

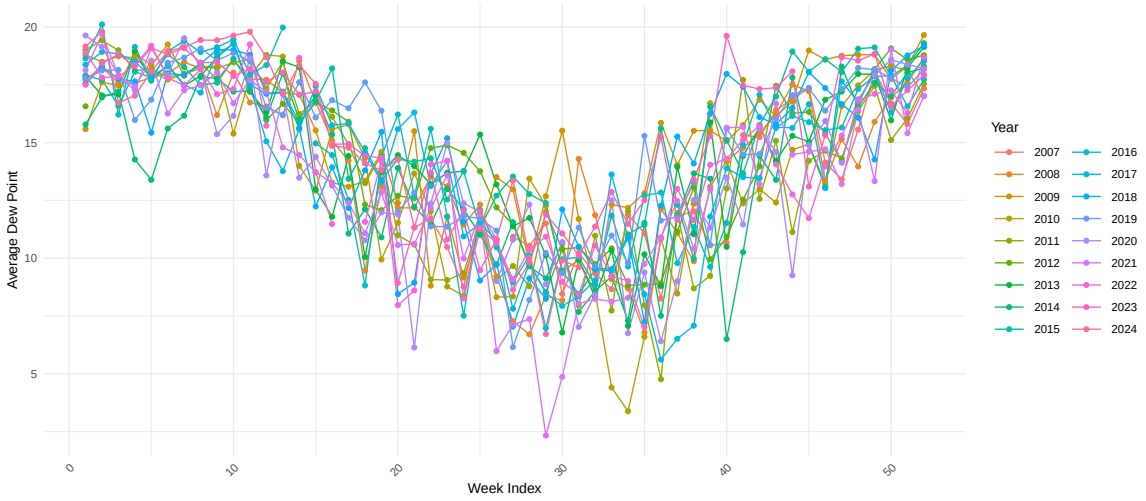



Figure 30 – Average dew point observed from 2007 through 2024.

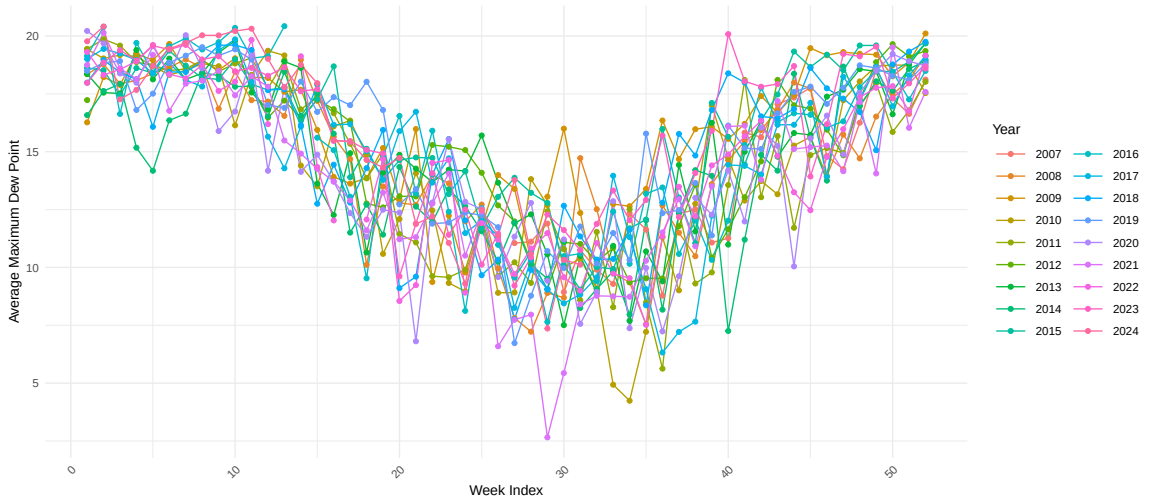



Figure 31 – Average maximum dew point observed from 2007 through 2024.

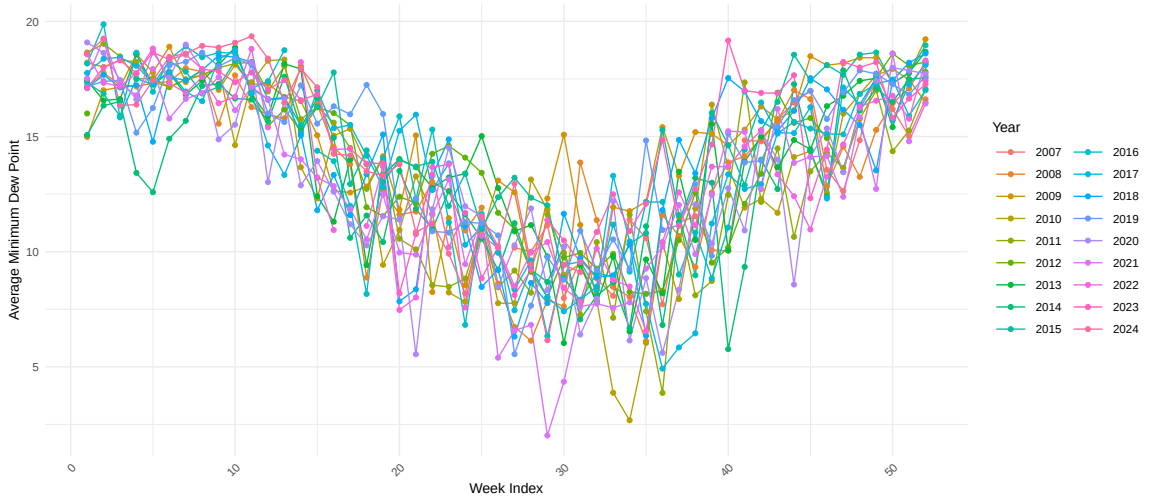



Figure 32 – Average minimum dew point observed from 2007 through 2024.

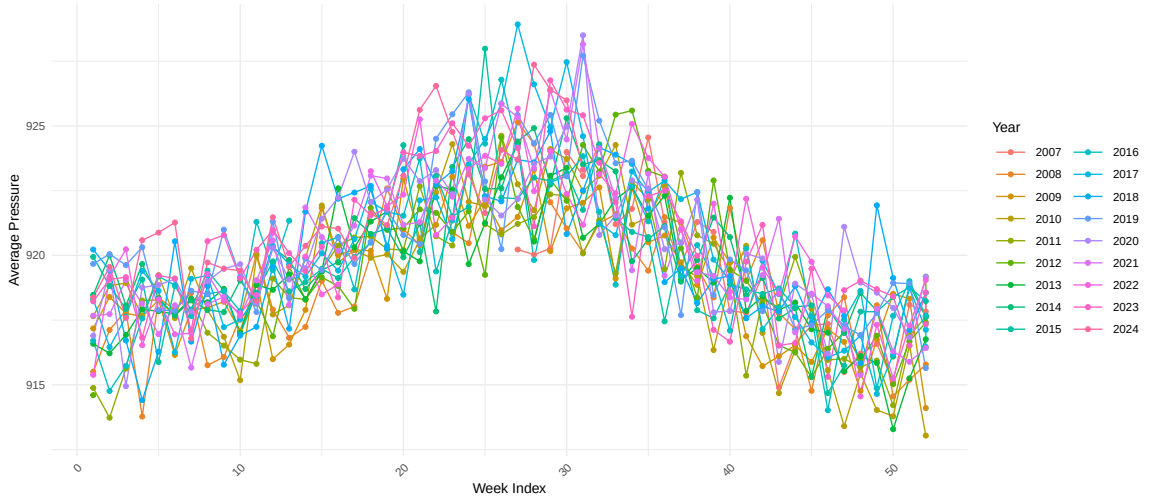



Figure 33 – Average pressure observed from 2007 through 2024.

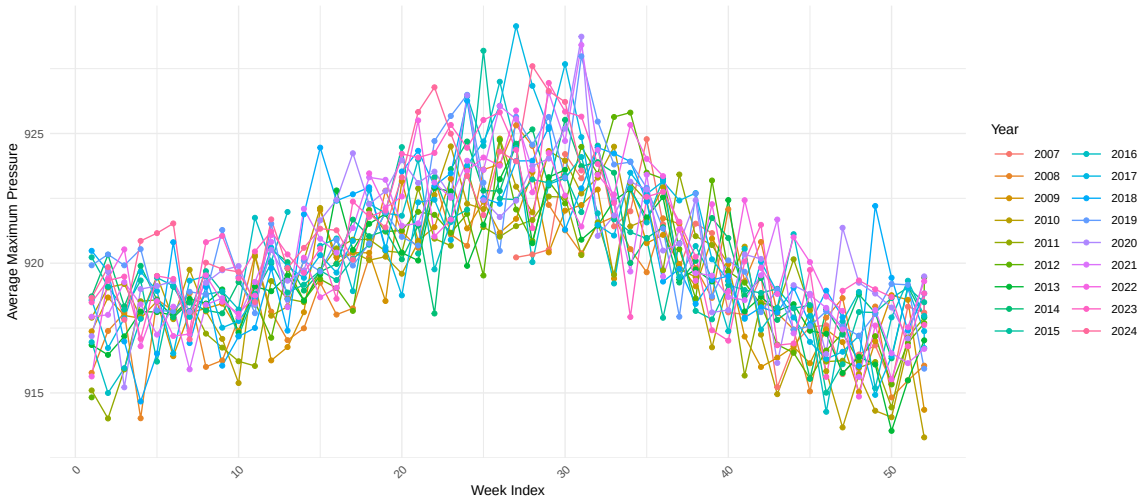



Figure 34 – Average maximum pressure observed from 2007 through 2024.

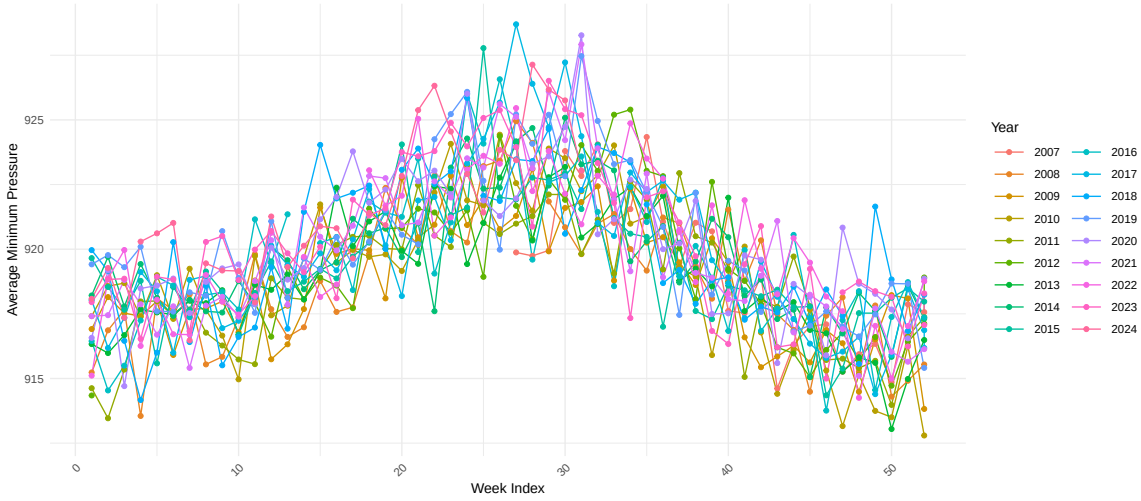



Figure 35 – Average minimum pressure observed from 2007 through 2024.

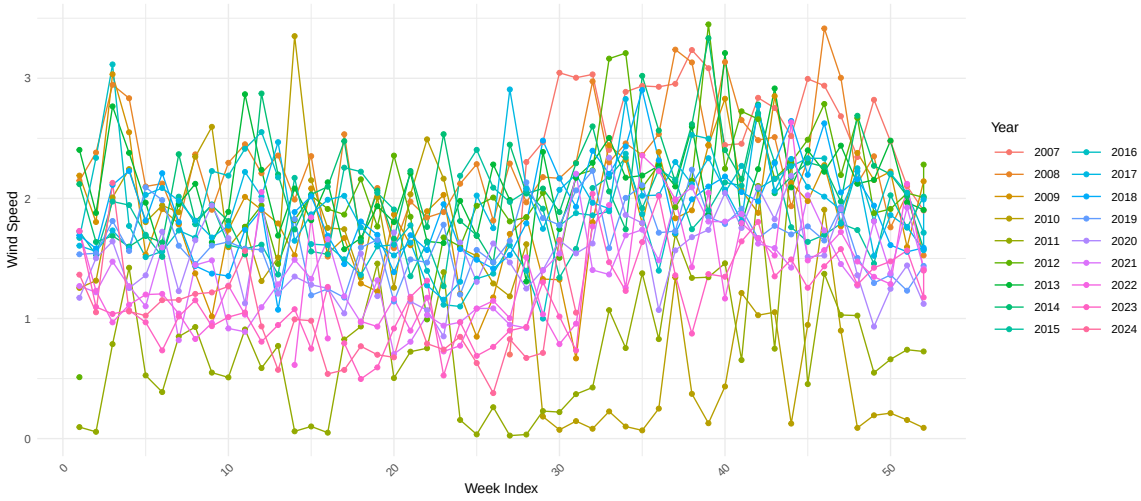



Figure 36 – Average wind speed observed from 2007 through 2024.

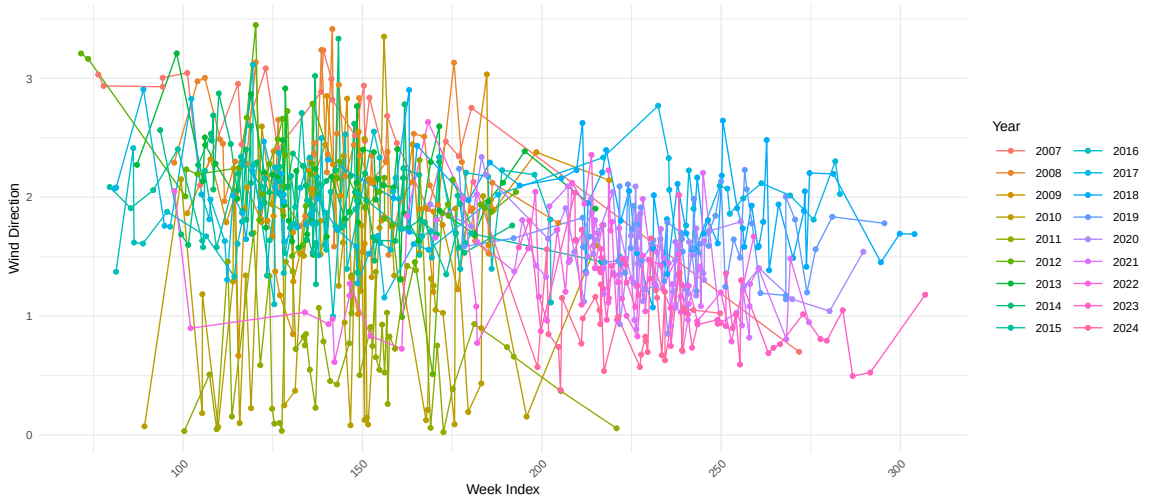



Figure 37 – Average wind direction observed from 2007 through 2024.

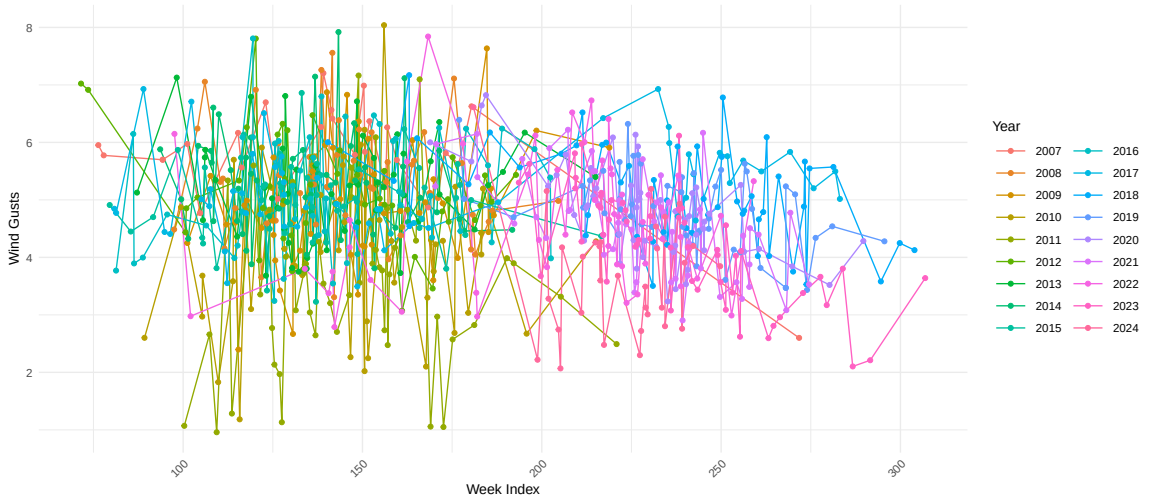



Figure 38 – Average wind gusts observed from 2007 through 2024.


ICMC
USP
SÃO CARLOS