Bayesian Inference for Dynamic Spatio-temporal Models

The University Of Sheffield.

Sofia Maria Karadimitriou

School of Mathematics and Statistics

Submitted for the degree of *Doctor of Philosophy in Probability & Statistics*

Supervisors:
Dr. Kostas Triantafyllopoulos
Dr. Timothy Heaton

September, 2018

To my family; Dimitris, Kassiani, Nikos and Sofia ...

Acknowledgments

I would like to thank and acknowledge my supervisory team, composed of Kostas Triantafyllopoulos and Timothy Heaton but also my advisor Prof. David Applebaum. Thank you for the guidance and assistance during my PhD, especially during the hard times of stress. I would also like to thank Allan Chalk, who kindly provided me with the traffic flow data used for the second part of the work presented in this thesis.

I will always humbly thank my family, who have been incredibly important to me, for their unconditional moral support, the faith and trust they put in me, and their wise advice through hard times that I encountered during my PhD. Thank you to my friends in Greece, Christy and Oldie who proved to be exceptionally selfless and supportive even through a big distance between us.

Thank you to all my friends in Sheffield, Voula, Christos, Rahul and Veronica that have been my support especially during the final year.

The biggest thank you to Dimitris, who never ceased to believe in me, who have been through so much together and still we support each other unconditionally.

Abstract

Spatio-temporal processes are phenomena evolving in space, either by being a point, a field or a map and also they vary in time. A stochastic process may be proposed as a vehicle to infer and hence offer predictions of the future. In this era high dimensional datasets can be available where measurements are observed daily or even hourly at more than one locations along with many predictors. Therefore, what we would like to infer is high dimensional and the analysis is difficult to come through due to high complexity of calculations or efficiency from a computational aspect.

The first Reduced-dimension Dynamic Spatio Temporal Models (DSTMs) were developed to jointly describe the spatial and temporal evolution of a function observed subject to noise. A basic state space model is adopted for the discrete temporal variation, while a continuous autoregressive structure describes the continuous spatial evolution. Application of DTSMs rely upon the pre-selection of a suitable reduced set of basis functions and this can present a challenge in practice.

In this thesis we propose a Hierarchical Bayesian framework for high dimensional spatio-temporal data based upon DTSMs which attempts to resolve this issue allowing the basis to adapt to the observed data. Specifically, we present a wavelet decomposition for the spatial evolution but where one would typically expect parsimony. This believed parsimony can be achieved by placing a Spike and Slab prior distribution on the wavelet coefficients. The aim of using the Spike and Slab prior, is to filter wavelet coefficients with low contribution, and thus achieve the dimension reduction with significant computational savings.

We then propose an Hierarchical Bayesian State-space model, for the estimation of which we offer an appropriate Forward Filtering Backward Sampling algorithm under an MCMC procedure. Then, we extend this model for estimating Poisson counts and Multinomial cell probabilities through proposing a Conditional Particle Filtering framework.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

There is a plethora of phenomena that can be expressed and evolve through time and space. Historical processes such as the growth or decline of populations; migrations; epidemics; or even physical, biological, environmental processes are showing changes of spatial patterns through time. An important and revolutionary consideration in statistical science for these kind of processes was the marriage of time series and spatial processes. This marriage is of paramount importance as it is not deemed sufficient to consider these processes as solely one spatial process at a given time point or as a time series process at a given location; the behaviour and correlation of those patterns are important to be captured in a spatio-temporal manner.

Several models have been developed in order to predict the spatial features of those processes, to forecast the temporal patterns, but also to account for dependence in the parameter inference. Cressie's comment on Handcock and Wallis (1994) and the development in Goodall and Mardia (1994) provided us with the introduction of spatio-temporal models. These models were then extended in terms of a combination of spatial and dynamical prediction, known as the kriged Kalman Filter modelling (Mardia et al., 1998). Maximum likelihood estimation algorithms were introduced in Brown et al. (2001) to estimate the rainfall levels in Lancashire, England. Furthermore, a Bayesian paradigm of these models was applied to rainfall precipitation data in Venezuela (Sansó and Guenni, 1999) while the Hierarchical Bayesian spatio-temporal framework where this thesis is based on was proposed in Wikle et al. (1998). Finally, extensions for non-stationary spatio-temporal datasets were developed by Stroud et al. (2001).

However, nowadays, we have access to large datasets for which we want to extract information of their spatio-temporal patterns. Additionally, we want to conduct inference on the parameters without having to deal with the curse of dimensionality as there is limited computational power. Given that, spatio-temporal models of such nature have been developed by considering the use of smart and efficient approximations of orthogonal functions. Wikle and Cressie (1999) introduce the Reduced-dimension Dynamic Spatio-temporal Models (DSTMs). Their estimation procedure is based on the kriged Kalman Filter of Mardia et al. (1998) with the novel step of conducting dimension reduction through orthonormal basis decompositions. Since then, extensions have been proposed from an inferential perspective, such as in Xu and Wikle (2007) from an E-M algorithm estimation. Bayesian approaches to this model were introduced in Wikle et al. (2001), Wikle (2003) and Xu et al. (2005). Finally, a further extension have been proposed via a kernel representation in Wikle (2002) and more specifically for Poisson count data for the cloud coverage intensity.

Having said that, the scope of this thesis is to extend the model of Wikle and Cressie (1999) under a Bayesian Hierarchical setting (Wikle et al., 1998) where we can still introduce sparsity but also preserve the important features and dependencies of those patterns. Another significant feature of our proposed methodology is the use of wavelet basis functions for the orthogonal decomposition in order to approximate the processes of interest. Wavelets are strong detectors of discontinuities but their key characteristic is that only a few bases can approximate very well an underlying function. Daubechies (1992) provides a thorough review on wavelets which are proved to be efficient in numerous applications of different fields where functions subject to noise need to be estimated (such us in Antonini et al. (1992), Leung et al. (1998), Demiralp et al. (1999) or Xue et al. (2003)).

Consequently, the expected sparsity of wavelets can be then taken advantage by a prior belief which introduces sparsity. Such priors have been developed by Park and Casella (2008) and Ishwaran and Rao (2005) in order to shrink the regression coefficients and extract the important predictors. Thus, through this thesis we will take advantage of the shrinkage properties of the Spike and Slab prior (Ishwaran and Rao, 2005) and output important spatio-temporal patterns with the help of wavelets. This already brings our modelling framework to a Bayesian setting. It is known that Markov Chain Monte Carlo (MCMC) approaches are used by far in the literature in order to extract posterior estimates of the parameters of interest. However, due to the tempo-

ral aspect of our modelling framework, the sampling of specific dynamic components through MCMC leads to non-convergence. Thus, for these parameters, the Forward Filtering Backward Sampling (FFBS) algorithm will be used for the posterior sampling.

Additionally, we extend our modelling framework for Poisson and Multinomial spatio-temporal processes. However, as the observations are not Gaussian, the same inferential framework on the dynamic components cannot be used. For that reason, we introduce a Particle Filtering framework with the inclusion of Gibbs sampling and Metropolis-Hastings steps while we also use static parameter estimation (Storvik, 2002). These algorithms have been developed by Lindsten et al. (2014) and Andrieu et al. (2010) and have not been used yet in the context of spatio-temporal models.

Therefore, our proposed methology is then split in two parts, the first one is estimating Gaussian spatio-temporal process and conducting inference via MCMC procedures; the second part is predicting non-Gaussian spatio-temporal process through Particle Filtering (PF) methods. One challenge of our approach is that we wish to introduce adaptivity in the proposed model through the dataset but also keep its computational efficiency and approximation to the spatio-temporal patterns at a high level. Indeed, it is commonly admitted that there is no panacea for these types of datasets, yet, especially when there are complex underlying characteristics on top of the high dimensions. That is the reason why we firstly introduce sparsity with our model but also through Bayesian techniques we let the data adaptively inform us of the spatio-temporal dependencies that lie in the background.

Furthermore, in order to test the predictive ability of the proposed models, we provide in each case simulation studies and applications on pollution and traffic flow datasets. Our findings include advantages and drawbacks of the proposed framework. One drawback is the computational intensity of our modelling approach. However, there are only a few big data applications that can run efficiently in commercial laptops or desktops. Thus, in the following thesis we provide a literature of the different tools that we combine together into a flexible modelling approach. We then proceed on our proposed methodologies with simulation and application results.

Chapter 2 offers a literature review of the different topics combined in the thesis. In a first section, we give an introduction to spatio-temporal processes. Specifically, we firstly explain what a time and a spatial process is and then we proceed on the spatio-

temporal one. We then provide a presentation of the Reduced-dimension Dynamic Spatio-temporal Model in which our thesis is based on. Finally, in the last section we will provide a thorough detail of the wavelet basis decomposition and wavelet shrinkage as wavelets will be an important feature in the proposed methodology.

Chapter 3 describes our whole Bayesian Modeling approach (framework, assumptions, parameters) of Reduced-dimension DSTMs under a wavelet basis decomposition for the approximation of the processes of interest in a Gaussian setting. Furthermore, we provide simulation studies to evaluate our model's performance and then we proceed to an application to pollution data in Athens.

Chapter 4 extends the Reduced-dimension DSTMs for Poisson distributed spatio-temporal temporal processes. We then propose two modeling procedures based on the application of interest. Additionally, due to observations being non-Gaussian, we provide new algorithmic procedures for the inferential part, that being the Particle Filter (PF). Furthermore, we provide simulation studies to evaluate our model's performance and then we proceed to an application to traffic flow data in one of the biggest motorways in the UK, the M6.

In Chapter 5 we provide an extension to the Reduced-dimension DSTMs for Multinomial distributed spatio-temporal temporal processes under Particle Filtering techniques for the parameter estimation. We then show a simulation study for the model's efficiency and we provide an application by revisiting the traffic flow data of Chapter 4.

In Chapter 6 we summarise the findings, advantages and drawbacks of the proposed methodology but also comment on further considerations and future work.

Chapter 2

Literature Review

2.1 Spatio-temporal Modelling

In this section we introduce what is a spatial process and a time series process with their combined version that results in a spatio-temporal process. Then we will move on to the dynamic modelling formulation of a spatio-temporal process, the known Dynamic Spatio-temporal Models (DSTMs) in section 2.2 and more specifically on the Dimension Reduced one that introduces an effective suggestion for dimension reduction. Our main focus of research evolves around the Dimension Reduced DSTMs where is the most volume of our literature review that is explained.

2.1.1 Spatial processes

Spatial processes represent how a phenomenon can be evolved in a spatial region. Suppose we consider the observed measurements $Y_{\boldsymbol{s}} = \{Y_{s_1}, \ldots, Y_{s_n}\}$ at locations $s_1, \ldots, s_n \in D_s$, where D_s specifies a continuous spatial region. Then, there are three fundamental types of spatial data:

- Point-referenced or geostatical processes, where the data $Y_{\boldsymbol{s}}$ are randomly measured at selected locations which can vary through an area such as air pollution in the rural and suburban areas. In the rest of the thesis, our main focus will be on processes of that nature.
- Lattice Processes, which are summaries of variables in partitioned grids of an area with boundaries that are defined either from the user or because the nature of the

data have already pre-defined regions, such as the number of accidents per post code in Leeds.

- Point processes, in which $Y_{\boldsymbol{s}}$ are happening randomly and in random locations of a specified area, such as looking at breast cancer cases in the county of Yorkshire.

In spatial processes, the way measurements are sampled spatially plays a crucial role as there are different approaches and correlation structures that we can consider for the spatial dependences and the expansion of these processes in space. These examples and numerous modelling approaches are discussed in textbooks such as in Cressie (1992), Banerjee et al. (2014) and Cressie and Wikle (2015).

2.1.2 Time series processes

A time series process sampled at discrete time is defined as a series of observations $Y_t = \{Y_1, Y_2, \ldots, Y_T\}$ observed at the discrete time points $t = 1, 2, \ldots, T \in D_T$, where t denotes the time index.

These processes for example can explain phenomena such as the daily prices of financial assets, daily temperatures in a particular city, annual precipitation levels of a lake, number of earthquakes in a city and so forth. What makes time series different for spatial processes is the way the time points are collected, as it introduces particular correlation or dependence structure between measurements or observations. Therefore, in order to understand and model these processes, it is necessary to use statistical models that consider dependence structure in time. The study of such models is known as time series analysis and has been discussed in many textbooks, see e.g. Brockwell and Davis (1991), Shumway and Stoffer (2000) and Lindsey (2004).

2.1.3 Spatio-temporal Processes

Spatio-temporal processes are a combination of a spatial process and a time series process, i.e., they evolve through both time and space. Let $\{Y(s,t) : \boldsymbol{s} \in D_s, t \in D_t\}$ denote the spatio-temporal process that expresses a random phenomenon evolving in the spatio-temporal set $D_s \times D_t$. This for instance can be the daily wind speed at a specific coordinate system $\boldsymbol{s} = (\text{latitude}, \text{longitude}, \text{altitude})^\top$ and D_t denoting the days, or it can express the monthly air pollution of $n = 100$ weather stations in the UK with $\boldsymbol{s} = (s_1, \ldots, s_{100})^\top$ and D_t denoting the months.

The process can be formulated accordingly in terms of either a spatial varying time series model or a temporally varying spatial model by writing $Y(\boldsymbol{s}, t) = Y_{\boldsymbol{s}}(t)$ and $Y(\boldsymbol{s}, t) = Y_t(\boldsymbol{s})$ respectively. The study of such models has been discussed in many textbooks, see e.g. Banerjee et al. (2014) and Cressie and Wikle (2015) while for the rest of the thesis we will focus on the Dynamic Spatio-temporal Models and more specifically the Dimension Reduced ones.

2.2 The Dynamic Spatio-temporal Model

In this section we start by reviewing the work that was conducted by Wikle and Cressie (1999). Specifically, we will discuss the general framework of modelling point referenced spatio-temporal processes. Then, we will present how they adapted this approach to allow application to high dimensional datasets through truncated basis decomposition in section 1.1.2 before describing in section 1.1.3 the implications of this truncation on their model and inference. Finally, in section 1.1.4 we will talk about the limitations of their problem and how we attempt to address these in latter chapters.

Consider the measurements of a spatio-temporal process $Y(s, t)$, with $s \in D_s \subset \mathbb{R}$ denoting the locations and $t = 1, 2, 3, \ldots$ denoting the discrete time points. Wikle and Cressie (1999) assumed that these $Y(s, t)$ correspond to noisy observation of an unobserved, underlying process $X(s, t)$ which is denoised and smoother than the observed process $Y(s, t)$ and can be written in terms of a linear model

$$Y(s, t) = X(s, t) + \epsilon(s, t) \tag{2.1}$$

where $\epsilon(s, t)$ is the measurement error that we will consider as a spatio-temporal white noise process with zero mean and $\text{Var}(\epsilon(s, t)) = \sigma_\epsilon^2 \boldsymbol{I}_n$, $\forall s, t$.

Wikle and Cressie (1999) then consider the smooth process $X(s, t)$ to be further composed of X_K, which is modelled to evolve over time through a stochastic integro-difference equation and a spatially correlated error process, i.e.,

$$X(s, t) = X_K(s, t) + \nu(s, t) \tag{2.2}$$

where $\nu(s,t)$ represents the small scale spatial random variation which is considered as a spatially variant zero mean process which is static in time.

The component $X_K(s,t)$ is assumed to unfold in terms of a state integro–difference equation with an autoregressive structure

$$X_K(s,t) = \int_D w_s(u) X_K(u, t-1) du + \eta(s,t) \tag{2.3}$$

where $\eta(s,t)$ signifies the spatially coloured error process and $w_s(u)$ is a weighting function or redistribution kernel that represents the contribution of location u at $t-1$ to the value at location s at time t with $|\int_D w_s(u)du| < 1$ which indicates spatial stationarity. For more information on spatial, temporal and spatio-temporal stationarity the reader should refer to Cressie and Wikle (2015).

Several comments are in order. There are assumptions to be met in order to consider this model plausible and meet spatial, temporal and spatio-temporal criteria. Firstly, the innovation process ϵ is uncorrelated with the underlying process X and both the spatial and temporal errors ν and η. Furthermore, we require several independence criteria between the separate components. Specifically, for all $s, r \in D \subset \mathbb{R}$ and discrete time points $t \neq \tau$:

$$\begin{aligned} E(\epsilon(s,t)\epsilon(r,\tau)) &= 0 \\ E(\nu(s,t)\nu(r,\tau)) &= 0 \\ E(\eta(s,t)\eta(r,\tau)) &= 0 \end{aligned}$$

Additionally, for all r, s, t and τ, between the spatial and temporal variation components ν and η we should have independence, i.e., $E(\nu(s,t)\eta(r,\tau)) = 0$. Finally, for all r, s and t there should be independence between the stochastic integro- difference equation (2.3) and the spatial variation component ν and between the stochastic integro-difference equation in previous time and another location with the temporal component η as well, i.e.,

$$\begin{aligned} E(\nu(s,t)X_K(r,t)) &= 0 \\ E(\eta(s,t)X_K(r,t-1)) &= 0 \end{aligned}$$

2.2.1 Dimension Reduction through Orthonormal Basis

Enviromental datasets are often high dimensional due to measurements that can be available anytime from weather stations around the globe, for instance air pollution measurements which are observed daily or even hourly at more than one hundred weather stations or locations along with many predictors, especially in large countries. Therefore, it is high dimensional and the analysis is difficult due to high complexity of calculations or efficiency from a computational aspect.

To tackle the difficulties caused by dimensionality Wikle and Cressie (1999) proposed a dimension reduction technique for spatio-temporal processes which allowed them to make inference about their model. Specifically, they approximate the functions of interest $X_K(s, t-1)$ and $w_s(u)$ in (2.3) by a truncated set of orthonormal basis functions, in which case a basis decomposition of a process $X_K(s,t)$ is derived as

$$X_K(s,t) = \sum_{j=1}^{K} \alpha_t(j)\phi_j(\boldsymbol{s}) \tag{2.4}$$

for locations $s \in D \subset \mathbb{R}$ and discrete time points $t = 1, 2, \ldots$ where $\boldsymbol{\alpha}_t = (\alpha_t(1), \ldots, \alpha_t(K))^\top$, are assumed as zero-mean time series which represent the basis coefficients that change through each time point and the sequence $\phi_j(\boldsymbol{s})$, with $i = 1, 2, \ldots, K$ are the basis functions chosen for each location which are complete and orthonormal.

Thus, by defining the vector $\boldsymbol{\phi}(\boldsymbol{s}) = (\phi_1(\boldsymbol{s}), \ldots, \phi_K(\boldsymbol{s}))^\top$ we can also define the $n \times K$ basis matrix $\boldsymbol{\Phi}$ where K indicates the amount of truncation that is induced,i.e.,

$$\boldsymbol{\Phi} = \begin{bmatrix} \phi(\boldsymbol{s}_1)^\top \\ \vdots \\ \phi(\boldsymbol{s}_n)^\top \end{bmatrix}$$

which finally gives us the decomposition

$$\boldsymbol{X}_K(t) = \boldsymbol{\Phi}\boldsymbol{\alpha}_t \tag{2.5}$$

with $\boldsymbol{X}_K(t) = (X_k(s_1,t), \ldots, X_k(s_n,t))^\top$. Due to the completeness of the orthonormal basis the weighting function can also be decomposed in terms of the chosen basis

functions, each one for a different location.

$$w_s(u) = \sum_{i=1}^{K} b_i(\boldsymbol{s})\phi_i(\boldsymbol{u}), \tag{2.6}$$

for $u, s \in D$ and where $\boldsymbol{s} = \{s_1, \ldots, s_n\}$ and $\boldsymbol{b}(\boldsymbol{s}) = (b_1(\boldsymbol{s}), \ldots, b_K(\boldsymbol{s}))^\top$ are unknown but unstochastic parameters which define the $K \times n$ matrix

$$\boldsymbol{B} = \begin{bmatrix} b(\boldsymbol{s}_1)^\top \\ \vdots \\ b(\boldsymbol{s}_n)^\top \end{bmatrix}$$

which again gives us the decomposition

$$\boldsymbol{w} = \boldsymbol{B}^\top \boldsymbol{\Phi} \tag{2.7}$$

with $\boldsymbol{w} = (w_{s_1}(\boldsymbol{u}), \ldots, w_{s_n}(\boldsymbol{u}))^\top$.

It is believed is that, for many real datasets, the important dynamics of the process can be accurately modelled with just a few basis functions. However, this requires us to choose the most appropriate basis functions. This is what we are trying to tackle in the next chapters.

2.2.2 State-space representation

The key factor into resorting into a basis decomposition is to reduce an initially complex model with extra diffusive dynamics into a much simpler one. In that simpler model a linear combination of those complicated processes can be incorporated. The introduction of this linearity allows us to use known statistical inference approaches.

Hence, by making use of the orthonormality properties of the orthonormal basis $\boldsymbol{\Phi}$ and substituting into the measurement equation (2.1) and state process (2.3) we can derive the observation equation of $Y(s,t)$ in terms of the decomposed terms of the

desired smooth process $X(s,t)$

$$\begin{aligned} Y(s,t) &= \boldsymbol{\Phi}\boldsymbol{\alpha}_t + \nu(s,t) + \epsilon(s,t) \\ \nu(s,t) &\sim \mathrm{N}(0, \sigma_\nu^2 \boldsymbol{S}) \\ \epsilon &\sim \mathrm{N}(0, \sigma_\epsilon^2 I) \end{aligned} \tag{2.8}$$

with $\boldsymbol{S}$ specifying an appropriate spatial correlation function and by the model assumptions we can derive that the zero-mean time series components in (2.3) are independent with the spatial components ν and ϵ , i.e., $E(\alpha_{it} v(s))$ and $E(\alpha_{it}\epsilon(s,t)) = 0$ respectively, for all s, t and $i = 1, \ldots, K$, while for the state equation (2.2) we can derive

$$\boldsymbol{\Phi}\boldsymbol{\alpha}_t = \boldsymbol{B}\boldsymbol{\alpha}_{t-1} + \boldsymbol{\eta}_t \tag{2.9}$$

By writing everything in a matrix form for each s_i with $i = 1, \ldots, n$ and solving for α_t we can write the state equation (2.9) as,

$$\boldsymbol{\alpha}_t = \boldsymbol{H}\boldsymbol{\alpha}_{t-1} + \boldsymbol{J}\boldsymbol{\eta}_t \tag{2.10}$$

where we assume that $n \geq K$, with $\boldsymbol{\eta}_t = (\eta(s_1,t), \ldots, \eta(s_n,t))^\top$ and we define $\boldsymbol{J} = (\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\top$ and $(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}$ non singular and $\boldsymbol{H}_{K\times K} = \boldsymbol{J}\boldsymbol{B}$. If the truncation parameter K changes with time, then $\boldsymbol{H}$ becomes a time varying evolution matrix and can be estimated appropriately under the state space framework. More details on state space models can be found on Shumway and Stoffer (2000).

Considering the orthogonality property of the basis, equation (2.10) can be simplified with $\boldsymbol{H} = \boldsymbol{\Phi}^\top\boldsymbol{B}$ and $\boldsymbol{J}\boldsymbol{\eta}_t = \boldsymbol{\Phi}^\top\boldsymbol{\eta}_t$ and thus we write the final state equation as

$$\boldsymbol{\alpha}_t = \boldsymbol{\Phi}^\top\boldsymbol{B}\boldsymbol{\alpha}_{t-1} + \boldsymbol{\Phi}^\top\boldsymbol{\eta}_t, \;\; \boldsymbol{\eta}_t \sim \mathrm{N}(0, \boldsymbol{\Sigma}_\eta) \tag{2.11}$$

Throughout this thesis we are using the same basis functions for $w_s(u)$ and $\boldsymbol{X}_K(t)$, however, one can choose different ones which will result in a slightly different form state equation in (2.11). More details and approaches can be found in Cressie and Wikle (2015).

2.2.3 Prediction and Inference

Wikle and Cressie (1999) fitted their model partly by using the Kalman Filter (Kalman, 1960) framework for the state vector $\boldsymbol{\alpha}_t$ and simple spatial kriging for the derivation of the underlying process $X_k(s,t)$.

Specifically, given the measurement equation (2.8) and the state process equation (2.11), the optimal predictor for $\boldsymbol{\alpha}_t$ up to t is expressed recursively according to a Kalman Filter:

$$\hat{\boldsymbol{\alpha}}_{t|t} \equiv \mathrm{E}[\boldsymbol{\alpha}_t|\boldsymbol{Y}_t,\ldots,\boldsymbol{Y}_1] = \hat{\boldsymbol{\alpha}}_{t|t-1} + \boldsymbol{G}_t\boldsymbol{Y}_t - \boldsymbol{\Phi}\hat{\boldsymbol{\alpha}}_{t|t-1} \tag{2.12}$$

with mean square error for $t \geq 1$

$$\boldsymbol{P}_{t|t} \equiv \mathrm{E}[(\boldsymbol{\alpha}_{t|t} - \hat{\boldsymbol{\alpha}}_{t|t})(\boldsymbol{\alpha}_{t|t} - \hat{\boldsymbol{\alpha}}_{t|t})^\top] = \boldsymbol{P}_{t|t-1} - \boldsymbol{G}_t\boldsymbol{\Phi}\boldsymbol{P}_{t|t-1} \tag{2.13}$$

with $\boldsymbol{G}_t$ specifying the Kalman gain at time t given by

$$\boldsymbol{G}_t = \boldsymbol{P}_{t|t-1}\boldsymbol{\Phi}^\top[\mathrm{Var}(\boldsymbol{\epsilon}_t) + \mathrm{Var}(\boldsymbol{\nu}_t) + \boldsymbol{\Phi}\boldsymbol{P}_{t|t-1}\boldsymbol{\Phi}^\top]^{-1} \tag{2.14}$$

with $\hat{\boldsymbol{\alpha}}_{t|t-1}$ indicating the mean $\mathrm{E}(\boldsymbol{\alpha}_t|\boldsymbol{Y}_{1:t-1})$, i.e., the forecast of $\boldsymbol{\alpha}_t$ at time t given up to time $t-1$. Similarly $\boldsymbol{P}_{t|t-1}$ indicates the the forecast of the covariance matrix $\boldsymbol{P}_t$ at time t given up to time $t-1$. This means that when $\boldsymbol{Y}_t$ is observed, the dataset is updated to $\boldsymbol{Y}_{1:t}$ and then $\hat{\boldsymbol{\alpha}}_{t|t}$ is the mean $\mathrm{E}(\boldsymbol{\alpha}_t|\boldsymbol{Y}_{1:t})$.

After deriving the optimal estimation $\hat{\boldsymbol{\alpha}}_t$, then, the optimal predictor for the unobserved process can be obtained through linearity in the Gaussian case:

$$\hat{X}(s,t|t) = \phi(\boldsymbol{s})^\top\hat{\boldsymbol{\alpha}}_{t|t} + C_\nu(\boldsymbol{s})^\top(C_0^Y)^{-1}\boldsymbol{Y}_t, \tag{2.15}$$

with $C_v(\boldsymbol{s}) = E[\nu(s,t),\boldsymbol{\nu}_t] \equiv (C_v(\boldsymbol{s},s_1),\ldots,C_v(\boldsymbol{s},s_n))^\top$, $C_0^Y = Cov(\boldsymbol{Y}_t,\boldsymbol{Y}_t)$ and the conditional prediction error variance for the optimal predictor is

$$\begin{aligned}\mathrm{E}[X(s,t) - \hat{X}(s,t|t)|\boldsymbol{Y}_t,\ldots,\boldsymbol{Y}_1] = {}& \phi(\boldsymbol{s})^\top\boldsymbol{P}_{t|t}\phi(s) - C_v(s)^\top(C_0^Y)^{-1}C_v(s) \\ & - 2\phi(\boldsymbol{s})^\top\mathrm{Cov}(\hat{\boldsymbol{\alpha}}_{t|t},\boldsymbol{Y}_t)(C_0^Y)^{-1}C_v(s) + C_v(s,s)\end{aligned} \tag{2.16}$$

A few comments are in order. The second term of (2.15) is a type of simple kriging applied to the spatial error term $\nu(s,t)$. The higher truncation level we have, i.e., the lower K, the more $X(s,t)$ looks like the simple-kriging predictor in the presence of

measurement error. The first term in (2.16) represents the prediction error variance from X_K while the second and third represent the simple-kriging prediction variance of the ν process while the last term is a correction term that was derived from the covariance between the Kalman filter prediction of X_K through both the use of the Kalman Filter prediction of α and the simple kriging predictor. Analogously to the optimal predictor, as the truncation integer K decreases, the prediction error variance looks more and more like the simple kriging variance.

For the parameter estimation under the DSTM, Wikle and Cressie (1999) mention that in order to derive the optimal predictor, the covariance parameters and basis functions should be known. However, in practice, the covariance parameters and the coefficient matrix $\boldsymbol{B}$ will be subject to estimation. That means that the exact conditional expectations cannot be derived. This approach is similar to simply conducting a Kalman Filter in time or simple kriging in space where the covariance parameters need to be estimated.

Additionally, it is assumed that the underlying process $\boldsymbol{X}_K$ includes any non stationary structure through the evolution matrix $\boldsymbol{\Phi}^\top \boldsymbol{B}$ and the temporal error process η_t. The non-dynamic spatial components are assumed to be composed of ν and ϵ by considering a nugget effect, i.e., estimating the variance σ_ϵ while considering ν to be an isotropic process. This nugget effect σ_ϵ is estimated empirically through an appropriate choice of a variogram. Moreover, the covariance of the process ν is again estimated empirically through $\sigma_\nu = C_0^X - C_0^{X_K}$ where $C_0^X = \text{Cov}(X(t), X(t))$ and $C_0^{X_K} = \text{Cov}(X_K(t), X_K(t)) = \boldsymbol{\Phi} J C_0^X J^\top \boldsymbol{\Phi}^\top$ for chosen basis functions $\boldsymbol{\Phi}$ and thus by examining different choices of covariance functions for the measurement error variance, the truncation parameter K is chosen arbitrarily. Finally, when these estimates are derived, the Kalman Filter is then used in order to estimate the model parameters.

Finally, Wikle and Cressie (1999) suggest a plethora of complete and orthonormal basis functions such as wavelets and orthogonal polynomials. In their study, where they modeled winder surface data, they used an empirical orthogonal function basis set, which are widely used for spatial prediction (Cohen and Jones, 1969) and are helpful when there is an anisotropic and heterogeneous covariance structure (Creutin and Obled, 1982).

2.2.4 Criticism and Motivation for our work

In this section we discuss the issues and weaknesses of Wikle and Cressie (1999) original approach that we attempt to address in this thesis. We provide a summary of the work of others in addressing the same issues and outline our proposed solutions. These are described in more detail in our later chapters.

Truncation The choice of the type and number of basis functions play a crucial role in the quality of inference. The number of truncated locations and time points is solely according to an empirical decomposition and thus the predetermination of the truncation parameter K is considered an arbitrary choice. In the case where the data are on an irregular grid then in order to implement their approach, Wikle and Cressie (1999) mention that an interpolation is needed onto the observations to transform them into a regular gridded setting. This is because in real life datasets the space is continuous and the data are not evenly distributed spatially. Wikle and Cressie (1999) suggest the use of empirical orthogonal functions. Specifically, since the state and observation processes vary continuously in space, a 'pre-gridding' procedure (Karl et al., 1982) should be applied. They use a simple space-time prediction framework in order to obtain smooth predictions of the underlying process on the prediction grid of interest given the irregularly spaced data. They then perform a principal component decomposition on the field of interest in order to derive the empirical orthogonal basis functions on a regular grid. Algorithmically, they firstly predict the underlying process X at each spatial location for each time point by using biharmonic splines (Sandwell, 1987). They then, separately, for each spatial location on the prediction grid, smooth over time using a local Gaussian kernel with a time-smoothing parameter chosen by generalised crossvalidation (Hastie and Tibshirani, 1990, pg. 49-52). This procedure gives the gridded underlying process X^o, from which the covariance estimate C_0^o is estimated through the method of moments. Finally, another suggested method is the use of Karhunen-Loève expansion in the continues space in order to derive the basis functions under a space-time Kalman Filtering framework (Wikle, 1996).

Another argument is that the data which are gathered often have very different dimensions in space and in time. Specifically, they may have many locations but only a few time points or vice versa. If there exist only a few time points then firstly, K is required to be small and makes the choice of the basis a challenge in practice. Secondly, the inference is uncertain, since we have way more parameters to infer on than data

points. Therefore, the arbitrary choice of the number of basis and thus the truncation level for such datasets has many drawbacks in both inference and in model uncertainty in terms of fitting which thus makes the choice of the type of basis a challenge in practice.

Ideally one would desire an adaptive procedure where the truncation parameter is chosen during the estimation of the parameters. Furthermore, since we would wish to retain a dimension reduction and thus a parsimonious modeling framework, a basis decomposition which reduces the inference complexity and induces this sparsity should be considered. Therefore, we propose a Wavelet Basis decomposition in Chapter 3 and we continue using it in the rest of the thesis. The reason we choose Wavelets especially for the spatial evolution is because we would only expect parsimony since only a few basis can already provide us with a very good approximation of the functions or processes of interest.

Inference Wikle and Cressie (1999) in order to estimate the model parameters under the DSTM are using the method of moments which is an extremely inefficient in the case of many parameters to be estimated. Novel inferential stages are given in Xu and Wikle (2007) from an E-M algorithm estimation perspective while the Bayesian approach to the model was worked out long before that, starting with Wikle et al. (2001), Wikle (2003) and Xu et al. (2005). The hierarchical Bayesian approach is by far the most flexible way to model Dynamic Spatio Temporal Models for real-world problems. Our innovative approach is that under a Bayesian Hierarchical representation we are taking advantage of the wavelets sparsity via using a Spike and Slab prior while the choice of the covariance structure of the spatial wavelet coefficients can be chosen by the researcher anytime. This framework is introduced in Chapter 3, while it will be similar to the rest of the thesis.

Non-Gaussian extensions In real life applications, most of the problems are non-Gaussian. For instance, some pollutants are Gamma distributed, or if we would like to work on the number of accidents varying spatially and through the years we would use a Poisson or a Multinomial distribution. For instance, if we consider that we observe accident counts in specific areas then the signal is Poisson distributed, i.e., $\boldsymbol{Y}_t \sim \text{Po}(\boldsymbol{\lambda}_t)$ which consequently signifies that the underlying process subject to estimate is the mean vector $\boldsymbol{\lambda}_t$. Only a few extensions have been developed for multivariate non-Gaussian processes of that nature while there is a plethora of datasets where there is a clear

non-normality. Some attempts have been done via a kernel representation in Wikle (2002) for Poisson count data for the cloud coverage intensity. Wikle and Cressie (1999) though do not provide a non-Gaussian version of their methodology.

In order to expand our proposed DSTM for non-Gaussian data, in Chapter 4 we will investigate the Poisson case along with a novel step of spatial autoregressive mean effect and finally we will investigate the Multinomial case in Chapter 5.

2.3 Wavelets

The main goal of using wavelets throughout our work is to approximate and denoise functions subject to noise. This approximation can be achieved via orthonormal basis decompositions which under wavelets' framework we can introduce sparsity that can then be taken advantage of. More specifically, consider the vector of observations $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ that is derived from the following model of non-parametric regression and we wish to estimate the unknown function f through the noisy process $\boldsymbol{y}$, for $x \in [0, 1]$:

$$y_i = f(x_i) + e_i \tag{2.17}$$

with $i = 1, \ldots, n$ and where $x_i = i/n$ and the i.i.d. $e_i \sim \mathrm{N}(0, \sigma^2)$.

In order to achieve this kind of estimation in general but more specifically for our work, we are going to use wavelet basis for the denoising and approximation of spatio-temporal functions. Therefore, in this section we will start by reviewing what an Orthonormal Basis in section 2.3.1 and consequently we will define a Wavelets basis in section 2.3.2. In section 2.3.3 we will describe the Discrete Wavelet Transform (DWT). Furthermore, in section 2.3.4 we will describe known Wavelets that can be used for decomposition. Finally, in section 2.3.5 and 2.3.6 we will talk about the applications of wavelets in denoising and the Bayesian approaches that have been developed on wavelets.

2.3.1 Definition of Orthonormal Basis

Orthonomal Bases for Vectors A basis of a vector space $\boldsymbol{S}$ is defined as a subset of $u_1, \ldots, u_n$ vectors that are linearly independent and span $\boldsymbol{S}$, i.e., Span$\{x : x =$

$\alpha_1 u_1 + \alpha_2 u_2, \ldots, \alpha_n u_n, \alpha_1, \alpha_2, \ldots \alpha_n \in \mathbb{R}\} = S$ which $\alpha_1, \ldots, \alpha_n$ are called the coefficients of the base field.

Then, by considering $< . >$ being the inner product, a basis is called orthogonal if $< u_i, u_j >= 0$ for $i \neq j$ and $< u_i, u_i >= \|u\|$, for all i. When $< u_i, u_j >= 1$, for $i = j$ then the basis vector $\{u_j\}$ is called orthonormal. Thus, if $\boldsymbol{\Phi} = \{\phi_j\}_{j \in I}$ is an orthonormal basis, then every element of $\boldsymbol{x}$ can be written as $\boldsymbol{x} = \sum_{\boldsymbol{\phi} \in \boldsymbol{\Phi}} \boldsymbol{< x, \phi >} \boldsymbol{\phi}$.

Orthonomal Bases for Functions For a real valued function f in $\boldsymbol{L^2(\mathbb{R})}$, i.e. $\int_{-\infty}^{\infty} |f(x)|^2 dx < \infty$, $\boldsymbol{\Phi} = \{\phi_i(x)\}_{i \in \mathbb{N}}$ is an orthonormal basis in $\boldsymbol{L^2(\mathbb{R})}$ for an interval $\boldsymbol{I}$ if, by defining the inner product between f and ϕ being $< f, \phi >= \int f(x)\phi(x)dx$ we have

- $< \phi_i(x), \phi_j(x) >= \int_I \phi_i(x)\phi_j(x)dx = 0$ for $i \neq j$
- $< \phi_i(x), \phi_i(x) >= \int_I \phi_i(x)\phi_i(x)dx = \|\phi_i(x)\|_2 = 1$

and then, $f(x)$ can be decomposed as $f(x) = \sum_{i=1}^{n} \alpha_i \phi_i(x)$ where the coefficients can be derived through orthogonality as $\alpha_i = \int_I f(x)\phi_i(x)dx =< f, \phi_i >$.

The most common orthonormal basis is the Fourier series where by taking $\boldsymbol{I} = [-\pi, \pi]$ it can be shown that

$$\left\{ \frac{1}{\sqrt{2\pi}}, \frac{\sin(x)}{\sqrt{\pi}}, \frac{\sin(x)}{\sqrt{2\pi}}, \ldots, \frac{\sin(nx)}{\sqrt{\pi}}, \frac{\cos(x)}{\sqrt{\pi}}, \ldots \frac{\cos(nx)}{\sqrt{\pi}} \right\}, n \in \mathbb{N}$$

forms an orthonormal basis on $\boldsymbol{I}$.

2.3.2 Definition of Wavelets

Definition 2.3.1. Wavelets are a family of functions $\{\psi_{jk}, \phi_{jk}\}$ on the Hilbert space $\boldsymbol{L}^2(\mathbb{R})$ of one or several variables $\boldsymbol{L}^2(\mathbb{R}^d)$ that can represent (or interpolate) other functions and are satisfying three properties (Jorgensen, 2006):

- They form a basis for $L^2(\mathbb{R})$ or $L^2(\mathbb{R}^d)$ and own orthogonality properties
- The individual ψ_{jk} and ϕ_{jk} arise from two generating functions called the mother and father wavelet respectively which relate the two operations of scaling and translating described bellow.

Daubechies (1992) defined the ψ wavelet function in the Hilbert space $L^2(\mathbb{R})$ of squared integrable functions on real line such that

$$\psi_{jk} = 2^{j/2}\psi(2^j x - k) \quad j, k \in \mathbb{Z} \tag{2.18}$$

while $\phi_{jk}(x)$ is derived from the *father* or scaling wavelet function $\phi(x)$ and it can be calculated from the dyadic dilations of ϕ defined as

$$\phi_{jk}(x) = 2^{j/2}\phi(2^j x - k) \tag{2.19}$$

and the whole family is an orthonormal basis in $\boldsymbol{L^2(\mathbb{R})}$.

- They are indexed by integer translations for and any powers of scaling.

For a function $f \in \boldsymbol{L}^2(\mathbb{R})$ we can decompose a function $\psi(x)$ on $\boldsymbol{L}^2(\mathbb{R})$ such that, for a chosen J with $\{\phi_{Jk}(x), \{\psi_{jk}(x)\}_{j<J}\}_{k\in\mathbb{Z}}$ that gives us

$$f(x) = \sum_{k\in\mathbb{Z}} \alpha_k \phi_{Jk}(x) + \sum_{j<J, k\in\mathbb{Z}} \beta_{jk}\psi_{jk}(z),$$

with

$$\alpha_{jk} = \int_{\mathbb{R}} f(x)\phi_{jk}(x)dx \text{ and } \beta_{jk} = \int_{\mathbb{R}} f(x)\psi_{jk}(x)dx$$

with J defined as the maximum resolution.

Parsimony in Wavelet Decompositions For many smooth functions, or smooth ones with some jump discontinuities or inhomogeneities, the decomposition is sparse. Sparsity is the most desirable attribute of the wavelet coefficients which is interpreted as lots of the wavelet coefficients being zero or close to zero. Moreover, by considering the energy, $\sum_i f(x_i)^2 = \sum_{i,j} \alpha_{j,k}^2$, and by taking sparsity into consideration, this means that the energy of f is now focused into fewer coefficients without losing any information. This means that only a few basis functions can provide us with a good approximation of a function which for high dimensional problems is a key factor.

Finally, wavelets are designed to have compact support which means increased sparsity and it allows localisation in both time and frequency.

Vanishing Moments The main idea is that if and only if the wavelet scaling function can generate polynomials up to degree $L-1$, then we say that a wavelet has L vanishing moments. More vanishing moments means that the scaling function can represent more complex functions.

So, we would like to construct a wavelet function $\psi(x)$ which corresponds to differencing up to degree x^{L-1}, so that $\int x^l \psi(x)dx = 0$, for $l = 0, \ldots, L$ which L stands for the moments. Since a function can be written in McLaurin series, we would have that $\int f(x)\psi(x)dx = 0 + 0 + 0 + \cdots + \int \frac{x^L}{L!} + \ldots$.

That means that the more vanishing moments a wavelet function can have, the more complex functions can be represented with sparse wavelet coefficients since they can ignore certain trends and can be sensitive to higher degree oscillations.

2.3.3 Discrete Wavelet Transform

By assuming that f was observed or sampled discretely, as $f(x_i)$, with $i = 1, \ldots, N$, $N = 2^j$, $t_i = i/N$) and $\psi_{jk} = (\psi_{jk}(1/N), \psi_{jk}(2/N), \ldots)^T$ then the discrete wavelet transform is defined as

$$\boldsymbol{f} = \sum_{k=-\infty}^{\infty} \alpha_{jk}\psi_{jk}$$

with $\psi_{jk} = 2^{j/2}\psi(2^j x - k)$ and so the coefficients can be estimated via

$$\boldsymbol{\alpha} = \boldsymbol{W}\boldsymbol{f}$$

where W is an orthonormal matrix, i.e., $WW^T = I$ and the i-th row of W is $\psi_{jk}(x_1), \psi_{jk}(x_2), \ldots$. Since $\boldsymbol{W}$ is an orthogonal matrix, then $\|\boldsymbol{\alpha}\|^2 = \boldsymbol{\alpha}^T\boldsymbol{\alpha} = (\boldsymbol{W}\boldsymbol{f})^T(\boldsymbol{W}\boldsymbol{f}) = \boldsymbol{f}^T(\boldsymbol{W}^T\boldsymbol{W})\boldsymbol{f} = \|\boldsymbol{f}\|$.

2.3.4 Examples of Wavelets

Haar Wavelets The Haar Wavelet is the simplest wavelet which is also discrete. The scaling function is defined as $\phi(x) = \mathbf{1}_{[0\leq x\leq 1]}$ while the mother wavelet can be defined

as:

$$\psi(x) = \begin{cases} 1, & 0 \leq x < \frac{1}{2} \\ -1, & \frac{1}{2} \leq x < 1 \\ 0, & \text{otherwise} \end{cases}$$

Any continuous real function with compact support which can be approximated by the linear combinations of the dyadic dilations of $\phi(2^n x)$ and their shifted functions. Furthermore, any continuous real function defined on [0, 1] can be approximated by the linear combinations of the constant function 1, $\psi(2^n x)$ and their shifted functions.

Haar Transform The Haar transform can be derived through as we call the Haar matrix but via its normalised form. Thus, the Haar transform of a value x_n is derived as $y_n = H_n x_n$. Due to orthogonality properties $H^{-1} = H^\top$ and thus the inverse Haar transform is given as $x_n = H_n^\top y_n$.

For example, the normalised Haar matrix for a vector x of $n = 4$ is given as:

$$H_4 = \frac{1}{2} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ \sqrt{2} & -\sqrt{2} & 0 & 0 \\ 0 & 0 & \sqrt{2} & -\sqrt{2} \end{bmatrix}$$

and consider the vector $x = (0.82, 0.48, -1.26, -1.18)^\top$. The discrete Haar transform is derived as

$$y_4 = H_4 x_4 = \frac{1}{2} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ \sqrt{2} & -\sqrt{2} & 0 & 0 \\ 0 & 0 & \sqrt{2} & -\sqrt{2} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 2 \\ 0 \end{bmatrix}$$

Daubechies Wavelet Daubechies (1992) defined the Daubechies wavelets that generate orthonormal basis in $L^2(\mathbb{R})$ where the scaling and translation functions are related

to each other with the following way

$$\psi(x) = \sqrt{2}\sum_{k=0}^{L-1} g_k \psi(2x-k)$$
$$\phi(x) = \sqrt{2}\sum_{k=0}^{L-1} h_k \phi(2x-k)$$

where L stands for the vanishing moments. Generally the wavelets of a class L vanishing moments have to be composed such that they satisfy the properties defined in section 2.2.3. For more details on the mathematical formulation of these wavelets, the reader can direct to Daubechies (1992).

For instance, the decomposition under the normalised Daubechies matrix of vanishing moments $L = 10$ for a vector x of $n = 4$ is given as:

$$y_4 = D_4 x_4 = \begin{bmatrix} 0.500 & 0.500 & 0.500 & 0.500 \\ 0.837 & -0.483 & -0.129 & -0.224 \\ -0.129 & -0.224 & 0.837 & -0.483 \\ 0.183 & 0.683 & -0.183 & -0.683 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \\ 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1.414 \\ 1.414 \\ 0 \end{bmatrix}$$

In practice one does not need to decompose by matrix multiplications as it creates computational inefficiency for high dimensions. Therefore, the Fast Wavelet Transform (FWT) algorithm was developed for filtering and decimating. The combination of computational efficiency and approximating discontinuous functions makes FTW in some applications better than the Fast Fourier Transform (FFT) (Daubechies and Sweldens, 1998) and this is what we are going to focus on for the rest of the thesis.

2.3.5 Applications of Wavelets to denoising

Due to wavelets' sparsity properties and ability to create fast decompositions, they have been used in numerous applications. More specifically, in image compression (Antonini et al., 1992), in spectroscopy for characterisation of substances (Leung et al., 1998), for modeling event potentials for cognitive information processing (Demiralp et al., 1999), in electroencephalogram analysis (Xue et al., 2003) or for the computation of connection weights in networks (Jemai et al., 2011).

We are mainly particularly interested in their application based on denoising and esti-

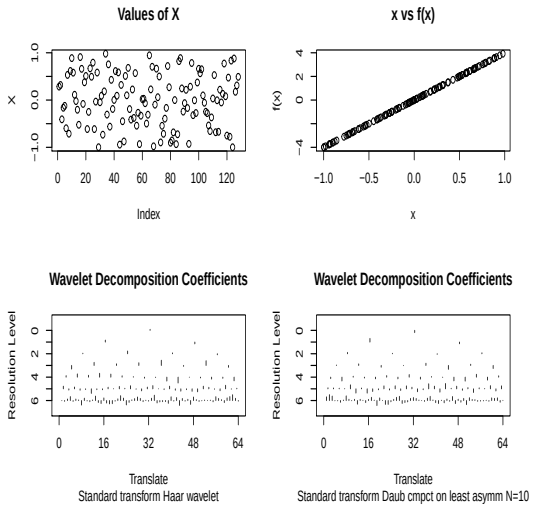

Figure 2.1: Coefficients of the wavelet transforms under a linear function $f(x) = 4x$ for vector $f(x)$. On the bottom left the Haar wavelet was used. On the bottom right Daubechies with $m = 10$ vanishing moments was used.

mation of a function. Returning to the problem specified in (2.17), the process can be easily multiplied with a wavelet matrix W due to linearity as follows:

$$\begin{aligned} \boldsymbol{z} = \boldsymbol{W}\boldsymbol{y} &= \boldsymbol{W}(\boldsymbol{f} + \boldsymbol{e}) \\ \boldsymbol{z} &= \boldsymbol{W}\boldsymbol{f} + \boldsymbol{W}\boldsymbol{e} \\ \boldsymbol{z} &= \boldsymbol{w}^* + \boldsymbol{\eta} \end{aligned} \tag{2.20}$$

with i.i.d. $\eta \sim \mathrm{N}(0, \sigma^2)$.

Our main focus is that through the specification of the wavelet matrix W to estimate the new process z which is an approximate version of the underlying process y. As wavelets provide us with parsimonious representations but nonetheless, good approximation, the key factor into estimating the new process z is via thresholding procedures on the wavelet coefficients w^*. In order to conduct thresholding, some techniques have been developed in the literature, either Frequentist or Bayesian, where the threshold levels for the choice of the important coefficients is specified or estimated. For the non-important ones that would indicate that they are bellow the threshold level and thus should be set as zero. Additionally, the noise of the transformed model will also be white noise and hence is spread evenly over all wavelet coefficients. The estimation and thresholding procedures under wavelet decomposition are called shrinkage methods

and are explained in the following section.

2.3.6 Shrinkage Methods

Considering the parsimonious representation that wavelets introduce, frequentist and Bayesian methods have been developed for the estimation of w^* in (2.20). These methods are called shrinkage as in order to estimate w^* in (2.20) eventually a thresholding (shrinking) estimation can be build. That means that given a wavelet coefficient z and a threshold we can remove the ones that are smaller than a specified threshold $\lambda > 0$ and keep the ones that are larger as they will be interpreted as the important ones to be kept.

Following (2.20), for the successful application of wavelet shrinkage some criteria are needed to be satisfied. Consider the observed subject-to-noise function f in (2.17). Then, through the discrete wavelet transform that is adopted and due to linearity, a modification or shrinkage is accomplished to the noisy function's wavelet coefficients, and afterwards through the inverse wavelet transform the function is estimated.

Under a frequentist framework, the discrete estimator $\hat{f} = (\hat{f}_1, \ldots, \hat{f}_n)$ will be judged by the mean squared error $\text{MSE}(\hat{f}, f) = n^{-1}E\|\hat{f} - f\|_{\ell_2}^2 = n^{-1}\sum_i = n^{-1}E[\hat{f}(x_i) - f(x_i)]$. Furthermore, the estimator $\hat{f}$ requires to be with high probability as smooth as f and should achieve almost minimax risk over one of a wide range classes, including the class in which linear estimators do not achieve the minimax case (Vidakovic, 2009, pg. 168).

Several thresholding choosers have been proposed in the literature, such as the hard and soft thresholding functions (Donoho and Johnstone, 1994). Moreover, the SURE thresholding method which however, this does not work well when the true signal coefficients are highly sparse and thus a hybrid of universal and SURE thresholding maneuver can be performed again only on certain levels above a given primary resolution. Additionally, a cross-validation techinque which was introduced by Nason (1996) and a two fold cross validation as well (Nason, 2002) and finally the False discovery rate (FDR) method by Abramovich and Benjamini (1996) which uses hypothesis testing for the sparsity of the wavelet coefficients. Abramovich et al. (2006) demonstrated that there is a connection between FDR and Minimax estimators. That is that FDR is at the same time asymptotically minimax for a wide range of loss functions and parameter spaces. The algorithm and a wider explanation on FDR is on (Nason, 2010, pg. 100)

and (Abramovich and Benjamini, 1996, pg. 5).

Adaptive Bayesian Shrinkage The main idea is to establish a Bayesian shrinkage rule by imposing a prior onto the wavelet coefficients. This will result into posterior summaries where the appropriate shrinkage will be decided adaptively.

Consider again the model in (2.17). Under a Bayesian approach what we like is to capture the expected sparsity of the wavelet coefficients through a suitable model, typically a prior on the coefficients $\boldsymbol{w^*}$. There are a number of different approaches to this problem but we will focus on two which are described in Nason (2010), as they are similar to our Bayesian approach that will be presented and discussed in the following chapters.

A model of hierarchical structure and was proposed by Chipman et al. (1997). Their approach is based on the Stochastic Search Variable Selection model introduced by George and McCulloch (1994) with the assumption that σ^2 is known. Specifically, they consider the model:

$$z|w^*, \sigma^2 \sim \mathrm{N}(w^*, \sigma^2)$$

with a prior distribution of mixed Gaussians defined as:

$$w^*|\gamma_j \sim \gamma_j \mathrm{N}(0, (c_j\tau_j)^2) + (1-\gamma_j)\mathrm{N}(0, \tau_j^2) \text{ with}$$
$$\gamma_j \sim \mathrm{Ber}(p_j)$$

where the hyperparameters p_j, c_j and τ_j depend on the level j to which corresponding w^* belongs.

The prior parameteter τ_j is set to be small and Chipman et al. (1997) proposed that values that are contained in the interval $(-3\tau_j, 3\tau_j)$ should be thought as zero. Additionally, c_j^2 is set to be much larger than one and thus it can be noted that a wavelet coefficient a-priori has the possibility to be very large with probability p_j or small with probability $(1-p_j)$.

The posterior distribution $w^*|z$ can be derived through the Bayes Theorem, consid-

ering that we have a mixture of priors as follows:

$$p(w^*|z,\gamma) = p(w^*|z,\gamma_j = 1)P(\gamma_j = 1|z) + p(w^*|z,\gamma_j = 0)P(\gamma_j = 0|z)$$

while obviously the marginals can be derived as

$$P(\gamma_j = 1|z) = \frac{p_j p(z|\gamma_j = 1)}{p_j p(z|\gamma_j = 1) + (1-p_j)p(z|\gamma_j = 0)} \quad \text{and}$$

$$P(\gamma_j = 0) = 1 - P(\gamma_j = 1|d) = (\frac{(1-p_j)p(z|\gamma_j = 0)}{p_j p(z|\gamma_j = 1) + (1-p_j)p(z|\gamma_j = 0)}$$

and the conditionals as:

$$w^*|z,\gamma = 1 \sim \mathrm{N}(\frac{(c\tau)^2}{\sigma^2 + (c\tau)^2}, \frac{\sigma^2(c\tau)^2}{\sigma^2 + (c\tau)^2}) \quad \text{and}$$

$$w^*|z,\gamma = 0 \sim \mathrm{N}(\frac{\tau^2}{\sigma^2 + \tau^2}, \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2})$$

and thus the Bayes rule under squared error loss for w^* at level j, which is also the shrinkage rule is calculated as:

$$\hat{w^*}_{\pi(z)} = \mathbb{E}(w^*|z) = P(\gamma_j = 1) \cdot \frac{(c_j\tau_j)^2}{\sigma^2 + (c_j\tau_j)^2} + P(\gamma_j = 0) \cdot \frac{\tau_j^2}{\sigma^2 + \tau_j^2}$$

This quantity can be considered as a smooth interpolation between two lines through the origin with slopes $\frac{\sigma^2(c\tau)^2}{\sigma^2 + (c\tau)^2}$ and $\frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}$ (Vidakovic, 2009, pg. 251).

Another approach takes into consideration the high level of sparsity of the wavelet coefficients, i.e. many of them are exactly zero and thus excluded from the wavelet regression and only a few non-zero. Hence, the previous method is quite inappropriate to capture this level of sparsity. However, if we would like a mixture of something that is exactly zero and something else, then this problem would be solved. With this justification Clyde et al. (1998) suggested a mixture of an indicator function based on a weight parameter γ_j on the variance of a Gaussian distribution which brings us into

the form of a Spike and Slab prior. By setting the priors:

$$w^*|\gamma_j, \sigma^2 \sim \mathrm{N}(0, (1-\gamma_j) + \gamma_j c_j \sigma^2)$$
$$\lambda_\nu/\sigma^2 \sim \chi^2_\nu$$

where λ and ν are fixed hyperparameters and the i.i.d. $\gamma_j \sim Ber(p_j)$ which also are indicating the selected basis element, i.e., column of W. Specifically, if γ_j is close to 1, then w^* for each element j will have a higher variance deviating from zero, while if it is closer to zero then it will vary from zero with a variance of one.

The posterior mean of $w^*|\boldsymbol{\gamma}$ can be calculated as

$$\mathbb{E}(\boldsymbol{w^*}|\boldsymbol{\gamma}) = \Gamma(I_n + C^{-1})^{-1}\boldsymbol{z}$$

where Γ and C are diagonal matrices with γ_{ij} and c_{ij} respectively.

The posterior mean is attained by averaging over all models. Model averaging leads to the multiple shrinkage estimator of $\boldsymbol{w^*}$

$$\mathbb{E}(\boldsymbol{w^*}|\boldsymbol{z}) = \sum_{\boldsymbol{\gamma}} p(\boldsymbol{\gamma}|\boldsymbol{z})\Gamma(I_n + C^{-1})^{-1}z$$

where $\pi(\boldsymbol{\gamma}|\boldsymbol{z})$ is the posterior probability of a particular subset $\boldsymbol{\gamma}$. Because of the complexity of 2^n calculations for the posterior probabilities, Clyde et al. (1996) approximated it by either conditioning on σ^2 or by assuming independence on the elements in $\boldsymbol{\gamma}$.

The approximate model probabilities, for the conditional case, are functions of the data through the regression sum of squares and are given by:

$$\pi(\boldsymbol{\gamma}, \boldsymbol{z}) \approx \pi_{approx}(\boldsymbol{\gamma}|y) = \prod_{j,k} \rho_{jk}^{\gamma_{j,k}}(1-\rho_{jk})^{1-\gamma_{jk}}$$

$$\rho_{jk}(\boldsymbol{z}, \sigma) = \frac{\alpha_{jk}(\boldsymbol{z}, \sigma)}{1 + \alpha_{jk}(\boldsymbol{z}, \sigma)}$$

$$\alpha_{jk}(\boldsymbol{z}, \sigma) = \frac{p_{jk}}{1-p_{jk}}(1+c_{jk})^{-1/2} \cdot \exp\left(\frac{S^2_{jk}}{2\sigma^2}\right)$$

$$S_{jk}^2 = z_{jk}^2/(1 + c_{jk}^{-1})$$

The p_{jk} can be used to obtain a direct approximation to the multiple shrinkage Bayes rule. The posterior mean for w_{jk}^* is $\approx \rho_{jk}(1 + c_{jk}^{-1})^{-1}d_{jk}$ and it can be seen as a level dependent wavelet shrinkage rule, generating a variety of non linear rules.

Donoho and Johnstone (1994) and Donoho et al. (1996) proposed for the selection of the universal threshold rejection regions of suitable hypotheses tests. Testing a precise hypothesis in the Bayesian analysis, requires a prior that has a point mass component which then under the null hypothesis the Bayes factor is calculated. Another approach is that from Abramovich et al. (1998), where they use weighted absolute error loss and use empirical Bayes for the thresholding hyperparameters.

In Chapter 3 we discuss the advantages of using a non-conventional spike and slab prior setting in terms of adaptivity but also flexibility in the wavelet coefficients' inference. Furthermore, we propose a matrix variate normal extension for high dimensional systems.

Application of wavelets on DSTMs Throughout this thesis we will get engaged with high dimensional Spatio-temporal processes and more specifically, the Dimension Reduced DSTMs under a wavelet basis decomposition. Considering that the existence of discontinuities in a spatial weight function, wavelets would capture those discontinuities and reduce the parameter space. Furthermore, a Bayesian approach for shrinkage similar to the approach of Clyde et al. (1998) will be used in order to estimate the spatial wavelet coefficients. Finally, for the temporal ones, Kalman Filtering methods (Kalman, 1960) will be used under a Gaussian DSTMs in Chapter 3 while for the non-Gaussian ones, a combination of particle methods will be used.

2.3.7 Comparison of Wavelets to other bases

Wavelets combine an elegant mathematical approach and computational efficiency. Specifically, in comparison to the Fourier bases, both wavelets and splines are either equally efficient or even computationally cheaper. Furthermore, wavelets provide localisation in contrast to Fourier series. They are strong detectors of discontinuities while they preserve a cheap computational decomposition under shrinkage properties as the large coefficients are shrunk while the low ones are treated as noise and thus

discarded. Specifically, a DWT is conducted under a level of $O(n)$ iterations while FFT is conducted in terms of $O(n\log(n))$ (Ramsay and Silverman, 2007).

Additionally, wavelets combine the frequency-specific approximating power of the Fourier series in time but also the spatial localised properties of splines. Thus, in the context of a spatio-temporal system, wavelets can provide us with the advantages of both the Fourier series and splines. Moreover, their theoretical properties adapt well to different degrees of smoothness and regularity. For instance, splines are mostly used when there is a lot of smoothness where more derivatives are used and the knots can be even chosen for each sample point which makes the user to have a system of a high number of basis functions and thus parameters. Furthermore, in B-splines, the knots are mostly regarded as fixed in order for the loss function to produce low bias and for computational convenience. On the other hand, the use of free-knot splines, which is a much more realistic scenario for complex and discontinuous data, consists of choosing an appropriate loss function which is problematic and the computational challenges are severe. Added to that, a low dimensional B-spline system is preferable to a higher one due to the trade-off between bias and variance in terms of the loss function. That means that a tolerance value should be set on the bias in order to achieve a stable estimation of the smoothness interpolation of the data. On the contrary, with wavelets we need less basis functions to estimate even a smooth function.

Chapter 3

Gaussian Reduced-dimension Dynamic Spatio-Temporal Models

3.1 Introduction

In this chapter we are going to focus on the reduced-dimension DSTM introduced by Wikle and Cressie (1999). Specifically, we are choosing a decomposition of wavelet basis functions. The advantage of using wavelets is that they introduce parsimony and only a few basis can be used in order to achieve a good approximation of a function and also localise possible discontinuities. This means that if the function of interest shows these discontinuities, then this will influence only the wavelets ψ_{jk} near it and consequently only those coefficients will be affected. Therefore, their compact support makes it easier for approximation of non-smooth functions. Thus, through a wavelet decomposition for the approximation of a process under a Reduced-dimension DSTM, one would typically expect sparsity. This believed parsimony can be then imposed through an appropriate prior belief in a Bayesian framework. This chosen prior belief is a complex form of a Spike and Slab prior (Ishwaran and Rao, 2005) where it introduces adaptivity to our data. The motivation of our work lies in the combination of Wavelets and DSTM as they are not widely used together.

Recall the model framework explained in Chapter 2,

$$\begin{aligned} Y(s,t) &= X(s,t) + \epsilon_t \\ X(s,t) &= X_K(s,t) + \nu_t \\ X_K(s,t) &= \int_D w_s(u) X_K(u,t-1)du + \eta(s,t) \end{aligned} \tag{3.1}$$

with $X_K(s,t)$ being the underlying process to be approximated through the basis decomposition $X_K(s,t) = \boldsymbol{\Phi}^T \boldsymbol{\alpha}_t$. Additionally, let $w_s(u)$ being the spatial contribution of location u at time $t-1$ to the location s at time t which again is approximated via $\boldsymbol{w} = \boldsymbol{B}^\top \boldsymbol{\Phi}$. Finally, we model the measurement component $\epsilon_t \sim \mathrm{N}(0, \sigma_\epsilon)$, the spatial error component $\nu_t \sim \mathrm{N}(0, \sigma_\nu \boldsymbol{S})$ and the temporal error component $\eta_t \sim \mathrm{N}(0, \boldsymbol{\Sigma}_\eta)$.

The proposed methodology that will be discussed in details in the following sections considers a Bayesian framework for the model parameters in (3.1). Specifically, we use a Spike and Slab prior for the weight function onto the elements of $\boldsymbol{B}$. Through that, we can focus on weights where we would expect to have a high contribution of one location to another, while the negligible ones would be considered close to being zero. Therefore, an adaptive framework for dimension reduction can be achieved. Additionally, we combine the Forward Filtering Backward Sampling (FFBS) algorithm for the stochastic parameters—or states— $\boldsymbol{\alpha}_t$ into our Bayesian framework, which under these models have not been used considering a Spike and Slab hierarchy and wavelet basis decomposition.

Furthermore, we offer a flexible modelling procedure for the covariance estimation under a preferable Bayesian setting. Specifically, an appropriate spatial correlation matrix can be used and through our Bayesian framework, the smoothing and/or scaling parameters can be estimated. For the temporal covariance of the evolution process, $\boldsymbol{\Sigma}_\eta$, numerous structures can be specified and therefore different Bayesian techniques can be used according to how we expect the underlying process to vary in time. For the measurement process' error variances σ_ϵ and σ_ν, due to the signal-to-noise ratio the inference on those variances have been proved to be challenging and we offer a few suggestions for a stable system.

In this chapter, we test our methodology's performance on simulated data where our results show that our methodology is advantageous. Specifically, our model can approximate the underlying process $\boldsymbol{X}_K$ well and reconstruct it in scenarios where we

had both a discontinuous process and a smoother one. Additionally, the majority of the Spike and Slab elements $\boldsymbol{B}$ are estimated fairly well and we can successfully reconstruct the weight function which is a challenge in real life applications. However, we noticed that due to the high signal-to-noise ratio, Spike and Slab has a difficulty to estimate perfectly all of the elements $\boldsymbol{B}$. Finally, the covariance estimation is successfully conducted, however, there is a trade off in the computational complexity as we need more computational power for the system to converge. Finally, the higher amount of locations n we have, again the computational complexity increases exponentially. Last but not least, at the end of this chapter we offer a real life application to pollution data under our proposed methodology. Our findings include that the proposed methodology is adapted for trended observations but we also derive causal relationships between locations from the estimation of the weight function.

3.2 Details of the Problem

In most real applications, the spatial part of the process $Y(s,t)$ is usually high dimensional which consequently brings a high dimension in the parameter space. This results to the problem of both the choice of basis and the truncation level as well. As discussed in Chapter 2, Wikle and Cressie (1999) predefine the truncation level K by the empirical covariance matrix. In any case, this method works under fixing the truncation level K and then conducting the appropriate inference. If we consider a non parsimonious basis function, for instance, such as Fourier basis, but introduce sparsity via a small number of K, the model will miss in sense of adaptivity and then important information of the data might be lost. That will consequently affect the inference of the parameters and thus the reconstruction of the process X_K and the weight function $w_s(u)$.

Considering the choice of basis under a Reduced Dimension DSTM, researchers used orthogonal polynomials, empirical orthogonal functions (EOF), wavelets, process normal modes, or splines (e.g., Wikle (1996); Mardia et al. (1998); Wikle and Cressie (1999); Berliner et al. (2000); Stroud et al. (2001); Wikle et al. (2001); Hsu et al. (2004); Xu et al. (2005); Johannesson et al. (2007); Cressie et al. (2009)). Finally in Wikle et al. (2001), a combination of equatorial normal mode (ENM) orthogonal basis functions and Wavelet for small scales under autoregressive priors was used.

Furthermore, apart from the truncation level, Wikle and Cressie (1999) are using a combination of the method of moments, a simple spatial kriging and Kalman Filter (Kalman, 1960) inferential procedure. The drawback of fixing K is that the lower the truncation level is, the higher the prediction error variance will be. Additionally, for the derivation of the optimal spatial predictor, essentially, the covariance parameters are considered known which instead due to the Kalman Filter recursions they should always be subject to estimation. Researchers in the past have used several inferential strategies under DSTMs, such as generalised expectation maximisation (GEM) algorithms instead of the method of moments or the simple EM algorithm but in order to increase the efficiency, Xu and Wikle (2007) had to ensure that the likelihood moves monotonically. Furthermore, a Group Lasso technique has been used by Bigot et al. (2011) but only on the covariance parameters while Berliner et al. (2000) used a Bayesian setting with the use of empirical orthogonal functions (EOFs) as basis functions. More traditional techniques, such as REML was used for the covariance parameter estimation as Furrer et al. (2006). Finally, a hierarchical Bayesian framework was used in Wikle et al. (2001) for modelling tropical surface winds, however, the authors have considered extra parameters that dealt with the wind direction.

In this thesis we will engage with three main issues. Firstly, we will tackle the curse of dimensionality in the parameter space by using an adaptive framework of Bayesian inference while preserving a parsimonious representation in basis functions. This makes our metholodogy efficient in the sense that we will not decompose the covariance parameters into basis representation which could add up more dimensions to the parameter space as previous authors did. Secondly, the truncation level instead of being manually chosen for the spatial coefficient matrix $\boldsymbol{B}$, will become adaptive while preserving most of the spatial information. Finally, the filtering and smoothing techniques deployed for the inference of $\boldsymbol{\alpha}_t$ provide a fast and accurate estimation for approximating the underlying process $X_K(s,t)$.

3.3 Proposed Methodology

In order to address the problems in Wikle and Cressie (1999), three concepts will be discussed. The first deals with the choice of the basis of the orthonormal decomposition of the underlying process and weight function; in this thesis a wavelet basis is proposed for this purpose. The second which was mentioned in the previous section, deals with

Bayesian inference which considers a multivariate form of a Spike and Slab prior for the spatial coefficient matrix $\boldsymbol{B}$. The combination of these two concepts brings us into an adaptive and efficient scheme in our methodology for the truncation level. Furthermore, the Forward Filtering Backward Sampling (FFBS) algorithm will be introduced along with the Kalman Filtering recursions for the Bayesian estimation of the temporal parameters $\boldsymbol{\alpha}_t$. Finally, for the covariance estimation numerous Bayesian approaches will be introduced according to the desired structure one wishes to considered based on the type of the application.

3.3.1 Bayesian Framework

The main goal is to be able to estimate the redistribution's kernel coefficient matrix $\boldsymbol{B}$, where the spatial contribution dynamics can be explained by using a sparse representation. Thus, by considering a prior kernel which induces parsimony, we will produce an adaptive framework where dimension reduction can be achieved. An intuitive explanation of Spike and Slab hierarchy is given in section 3.3.2 while the final model hierarchy is introduced in section 3.3.3.

Additionally, we are interested in the estimation of the dynamic coefficients $\boldsymbol{\alpha}_t$, where each future value depends on the previous past value. One may consider this problem through state space models and work accordingly via the Kalman Filter recursions (Kalman, 1960). More details will be provided bellow in the separate inference sections for each parameter.

For the variance and covariance parameters some comments are in order. Firstly, σ_v and σ_ϵ cannot be inferred together. This is due to their ratio σ_ϵ/σ_v which is the signal to noise ratio, i.e., how much the initial disturbance disperses when the variance of the kernel in the weighting function is larger. Therefore, either the ratio should be considered known and conduct inference into one of the two, or consider one of the two variances to be known.

Moreover, for the covariance structure of the temporal components $\boldsymbol{\alpha}_t$, $\boldsymbol{\Sigma}_\eta$, can be inferred through standard Bayesian procedures if it is time invariant, either with an Inverse Wishart prior, or if it considered to be diagonal, i.e., $\Sigma = \sigma_\eta \boldsymbol{I}_{n\times n}$, with a common inverse gamma prior on σ_η or it can be considered diagonal with a different variance for each location s, then n separate inverse gamma priors should be considered.

3.3.2 Spike and Slab prior on the Coefficient matrix B

Spike and Slab priors (Ishwaran and Rao, 2005) is a hierarchy of mixtures of a point mass distribution close to zero and a right tailed one. The main idea is that the coefficients in $\boldsymbol{B}$ that are expected to be zero will be filtered out through the hierarchy while the rest will be estimated appropriately. The logic behind this prior is similar to the Bayesian Lasso (Park and Casella, 2008), however, it is more powerful in case that the number of parameters is greater than the data points where in DSTMs that is mostly the case. This will yield posterior estimates that tend to involve only a small proportion of the parameters, hopefully avoiding over-fitting of the data. Prior to introducing the spike and slab hierarchy, lets us introduce the notions of vec and Kronecker product.

For a $K \times n$ matrix $\boldsymbol{B}$ the vec operator rearranges the elements of $\boldsymbol{B}$ into a vector, by stacking the columns of $\boldsymbol{B}$ one after the other. For instance, if $\boldsymbol{B} = [\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_n]$ where each $\boldsymbol{b}_i = [b_{1i}, b_{2i}, \ldots, b_{Ki}]^\top$ represents the i-th column of $\boldsymbol{B}$, for $i = 1, \ldots, n$ then

$$\text{vec}(\boldsymbol{B}) = \begin{bmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \\ \vdots \\ \boldsymbol{b}_n \end{bmatrix}$$

is a $Kn \times 1$ vector.

The Kronecker product of an $m \times n$ matrix $\boldsymbol{\Gamma}$ and a $p \times q$ matrix $\boldsymbol{\Psi}$ is the $mp \times nq$ matrix defined by

$$\boldsymbol{\Gamma} \otimes \boldsymbol{\Psi} = (\gamma_{ij}\boldsymbol{\Psi})$$

For example, if $m = n = p = q = 2$ then we have

$$\boldsymbol{\Gamma} \otimes \boldsymbol{\Psi} = \begin{bmatrix} \gamma_{11}\boldsymbol{\Psi} & \gamma_{12}\boldsymbol{\Psi} \\ \gamma_{21}\boldsymbol{\Psi} & \gamma_{22}\boldsymbol{\Psi} \end{bmatrix} = \begin{bmatrix} \gamma_{11}\psi_{11} & \gamma_{11}\psi_{12} & \gamma_{12}\psi_{11} & \gamma_{12}\psi_{12} \\ \gamma_{11}\psi_{21} & \gamma_{11}\psi_{22} & \gamma_{12}\psi_{21} & \gamma_{12}\psi_{22} \\ \gamma_{21}\psi_{11} & \gamma_{21}\psi_{12} & \gamma_{22}\psi_{11} & \gamma_{22}\psi_{12} \\ \gamma_{21}\psi_{21} & \gamma_{21}\psi_{22} & \gamma_{22}\psi_{21} & \gamma_{22}\psi_{22} \end{bmatrix}$$

Therefore, under a Spike and Slab hierarchy we have the following formulation according to our case by considering that our data satisfy (2.8) and (2.11) and by defining $\text{vec}(\boldsymbol{B}) = (\boldsymbol{\beta}_1(s), \ldots, \boldsymbol{\beta}_K(s))^\top$ and a $K \times K$ diagonal matrix $\boldsymbol{\Gamma} = diag(\gamma_1, \ldots, \gamma_K)$ where each $\gamma_k = \rho_k \tau_k^2$ represents the variance of each level $K = 1, \ldots$ of the spatial

wavelet coefficients the hierarchy is written as

$$\begin{aligned}
\boldsymbol{Y}_t|\boldsymbol{\alpha}_t, \sigma_v^2, \sigma_\epsilon^2 &\sim \mathrm{N}(\boldsymbol{\Phi}\boldsymbol{\alpha}_t, \sigma_v^2\boldsymbol{S} + \sigma_\epsilon^2 I) \\
\boldsymbol{\alpha}_t|\boldsymbol{B}, \boldsymbol{\Sigma}_\eta &\sim \mathrm{N}(\boldsymbol{\Phi}^\top \boldsymbol{B}\boldsymbol{\alpha}_{t-1}, \boldsymbol{\Phi}^\top \boldsymbol{\Sigma}_\eta \boldsymbol{\Phi}) \\
\mathrm{vec}(\boldsymbol{B})|\boldsymbol{\Gamma} &\sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Gamma} \otimes \boldsymbol{I}) \\
\rho_k|v_0, q &\sim (1-q)\delta_{v_0}(.) + q\delta_1(.) \\
\tau_k^{-2}|\omega_1, \omega_2 &\sim \mathrm{G}(\omega_1, \omega_2)
\end{aligned} \tag{3.2}$$

where $q \sim \mathrm{U}(0,1)$ or can be Beta distributed and σ_ϵ^2,σ_v^2 and $\boldsymbol{\Sigma}_\eta$ can be considered either known or inferred. Additionally, v_0 specifies a number very close to zero and the function $\delta(.)$ is the dirac delta or point mass function.

A few comments are in order. Firstly, the formulation of $\mathrm{vec}(\boldsymbol{B})|\boldsymbol{\Gamma} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Gamma} \otimes \boldsymbol{I})$ is the vectorised multivariate normal version of the matrix normal distribution of the random variable matrix $\boldsymbol{B}$, i.e., $\boldsymbol{B} \sim \mathrm{N}(\boldsymbol{B}_0, \boldsymbol{I}, \boldsymbol{\Gamma})$ where in this case $\boldsymbol{B}_0$ is a $n \times K$ location matrix of zeros and the identity $n \times n$ matrix $\boldsymbol{I}$ and the $K \times K$ matrix $\boldsymbol{\Gamma}$ being the scale matrices. In this case, as $\boldsymbol{\Gamma}$ signifies the variance for each level of the wavelet coefficients, the main idea is to filter out the zero wavelet coefficients in each level by deriving a very small posterior mean, through the hypervariances γ_k. This means that each level of coefficient has the same variance— or else it has the same magnitude of shrinkage or non shrinkage— for all the locations. Small hypervariances will give a high concentration over zero, while high hypervariances will escalate the non zero ones. Furthermore, as v_0 is considered to be a value very close to zero and is chosen along with the gamma density so, the hypervariances have a spike at v_0 and a right continuous tail which actually gives us a bimodal distribution for γ_k. The complexity parameter q controls the size of the model since it adjusts how likely the latent variable ρ_k is equal to one or v_0. The usage of a continuous prior over q offers an adaptive kind of estimation for the actual size of $\mathrm{vec}(\boldsymbol{B})$.

In order to derive the conditional posteriors, we will consider the prior belief that is in the vectorised form of the coefficient matrix $\boldsymbol{B}$ which is a multivariate Gaussian with a zero vector of means and the Kronecker product of the two covariance matrices and $\boldsymbol{\Gamma}_{K\times K}$ and $\boldsymbol{I}_{K\times K}$. However, in order to sample $\boldsymbol{B}$ the rest of the hierarchy parameters need to be updated as well. This will be achieved via Gibb's sampling steps for each parameter. In the following sections we will describe how each update is performed.

Finally, the use of this flavour of spike and slab prior instead of a conventional one by placing straightforwardly a mixture of a point mass at zero and a Gaussian provides us with more flexibility and adaptivity. Specifically, by considering the mixture in the variance components of the coefficient matrix we can choose the desirable variability among levels and locations. For instance, in our setting we suggest that the levels will vary differently as we would expect specific levels to capture more detail than others. One could generalise for both varying levels and locations. Consequently, this variability can be adaptively estimated and interpreted through this modelling process.

3.3.3 Summary of the Modeling Framework

In this part, we provide a summary of the proposed approach. Specifically, in table 3.1 we give a summary of the framework of the proposed methodology, the model, the parameters to be estimated and their relative priors, followed by the parameters' update via Gibbs sampling, along with the deterministic steps that are used for the estimation of the underlying process and the weighting function. In order to fit the model in table 3.1 via Gibbs sampling, it is required to update the parameters in an iterative procedure based upon their conditional distributions. In each section we will explain each of the following updates in more detail.

- Update $\boldsymbol{B}|\boldsymbol{\alpha}_t, \boldsymbol{\alpha}_{t-1}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}_\eta$ through the Spike and Slab hierarchy via Gibbs sampling
- Update the hyperparameters of Spike and Slab via Gibbs sampling, i.e.,

  1 Update for each level $\boldsymbol{\beta}_k|\alpha, \boldsymbol{\Gamma}, \rho, \tau_k^2, q$
  2 Update $\rho_k|\boldsymbol{\beta}_k, \tau_k^2, q, v_0$
  3 Update $\tau_k^2|\boldsymbol{\beta}_k, \rho_k, q, \omega_1, \omega_2$
  4 Update $q|\rho_k, v_0$

- Update $\boldsymbol{\alpha}_t|\boldsymbol{B}, \boldsymbol{\Sigma}_\eta, \boldsymbol{Y}_t$ via Kalman Filter and Forward Filtering Backward Sampling algorithm
- Update $\boldsymbol{\Sigma}_\eta|\boldsymbol{\alpha}_t, \boldsymbol{\alpha}_{t-1}, \boldsymbol{B}$ of the temporal components $\boldsymbol{\alpha}_t$ based on three structures

  1 If $\boldsymbol{\Sigma}_\eta = \sigma_\eta^2 \boldsymbol{I}$ under an inverse gamma prior then update via Gibbs sampling $\sigma_\eta^2|\boldsymbol{\alpha}_t, \boldsymbol{\alpha}_{t-1}, \boldsymbol{B}$
  2 $\boldsymbol{\Sigma}_\eta = \boldsymbol{\sigma}_\eta \boldsymbol{I}$ with $\boldsymbol{\sigma}_\eta^2 = (\sigma_{1\eta},^2 \ldots, \sigma_{n\eta}^2)^\top$ with n independent inverse gamma priors, i.e., update $\boldsymbol{\sigma}_\eta^2|\boldsymbol{\alpha}_t, \boldsymbol{\alpha}_{t-1}, \boldsymbol{B}$ via Gibbs sampling

3 Full structure on $\boldsymbol{\Sigma}_\eta$ with an Inverse Wishart prior, i.e., updating via Gibbs sampling $\boldsymbol{\Sigma}_\eta|\boldsymbol{\alpha}_t, \boldsymbol{\alpha}_{t-1}, \boldsymbol{B}$

- Update $\sigma^2_\nu, \sigma^2_\epsilon|\boldsymbol{\alpha}_t$ which signify the spatial covariance and measurement variance respectively through:

  1 Update $\sigma^2_\nu|\sigma^2_\epsilon, \boldsymbol{\alpha}_t$ by fixing σ^2_ϵ to be known under an inverse gamma prior

  2 Update $\sigma^2_\nu|\sigma^2_\epsilon, \boldsymbol{\alpha}_t$ by fixing $\sigma^2_\epsilon/\sigma^2_\nu$ to be known under an inverse gamma prior

- Update spatial correlation function's $\boldsymbol{S}$ parameters under gamma prior for selected correlation functions.

Data:	
Noisy spatio-temporal process:	$\mathbf{Y}$, $T \times n$ matrix
Smooth spatio-temporal process:	$\mathbf{X}$, $T \times n$ matrix
Integro-difference component:	$\boldsymbol{X}_K$, $T \times n$ matrix
Redistribution kernel:	$\mathbf{w}$, $n \times n$ matrix
Approximations:	
$\boldsymbol{X}_K = \boldsymbol{\alpha}\Phi$,	$\boldsymbol{\alpha}_{T\times n}$, coefficients of Wavelet matrix $\Phi_{n\times K}$
$\boldsymbol{w}_s = \boldsymbol{B}\Phi$,	$\boldsymbol{B}_{K\times n}$, coefficients of Wavelet matrix $\Phi_{n\times K}$
Model:	
$\boldsymbol{Y}_t = \boldsymbol{\Phi}\boldsymbol{\alpha}_t + \boldsymbol{v}_s + \boldsymbol{\epsilon}_t$	$v_s \sim \mathrm{N}(\mathbf{0}, \sigma^2_v \boldsymbol{S}_\lambda)$, $\boldsymbol{\epsilon}_t \sim \mathrm{N}(\mathbf{0}, \sigma^2_\epsilon \boldsymbol{I})$
$\boldsymbol{\alpha}_t = \boldsymbol{\Phi}^\top \boldsymbol{B}\boldsymbol{\alpha}_{t-1} + \boldsymbol{\eta}_t$	$\boldsymbol{\eta}_t \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_\eta)$
Parameters and Prior distributions:	
$\boldsymbol{\alpha}_{0\|0} \sim \mathrm{N}(\boldsymbol{m}_0, \boldsymbol{P}_{0\|0})$	$\boldsymbol{m}_0$, $\boldsymbol{P}_{0\|0}$ prior mean and covariance respectively.
$\mathrm{vec}(\boldsymbol{B})\|\boldsymbol{\Gamma} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Gamma} \otimes \boldsymbol{I})$	$\boldsymbol{\Gamma} = \mathrm{diag}\{\gamma_1, \ldots, \gamma_k\}$, $\gamma_k = \rho_k \tau^2_k$
$\rho_k\|v_0, q \sim (1-q)\delta_{v_0}(.) + q\delta_1(.)$	$q \sim \mathrm{U}(0,1)$
$\tau^{-2}_k\|\omega_1, \omega_2 \sim \mathrm{G}(\omega_1, \omega_2)$	$\boldsymbol{\beta}_k \sim \mathrm{N}(\mathbf{0}, \gamma_k \boldsymbol{I})$
$\sigma^2_v \sim \mathrm{IG}(\delta_0, \xi_0)$, $\lambda \sim \mathrm{G}(u_1, u_2)$	$\boldsymbol{\Sigma}_\eta \sim \mathrm{IW}(\nu, Q)$ or $\sigma^2_\eta \sim \mathrm{IG}(\psi_1, \psi_2)$

Table 3.1: Framework of the model

3.4 Updating the parameters

In the following parts we provide thorough proofs of updating the parameters of the model formulation in Table 3.1. In section 3.4.1 the Spike and Slab hierarchy parameters' full conditional posterior results are derived. Then, in section 3.4.2 the explanation

of the Kalman Filter recursions and Forward Filtering Backward Sampling algorithm for the sampling of $\boldsymbol{\alpha}_t$ is provided. Furthermore, in section 3.4.3 the updating of temporal covariance structure $\boldsymbol{\Sigma}_\eta$ is provided for the three different scenarios. Finally, in sections 3.4.5 and 3.4.6 the updating of the measurement variances σ_ϵ^2 and σ_ν^2 along with the spatial correlation parameters is provided.

3.4.1 Updating $\boldsymbol{B}|\boldsymbol{\alpha}_t, \boldsymbol{\alpha}_{t-1}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}_\eta$

In order to reduce the complexity of calculations we will rewrite the likelihood of $\boldsymbol{\alpha}_t$ in a vectorised form. Thus, by setting $\boldsymbol{J} = \boldsymbol{\Phi}^\top \boldsymbol{\Sigma}_\eta^{-1} \boldsymbol{\Phi}$ we can derive

$$\begin{aligned} p(\mathrm{vec}(\boldsymbol{\alpha})|\boldsymbol{Y}_{1:T}, \boldsymbol{B}, \boldsymbol{\Sigma}_\eta) \propto \prod_{t=2}^{T} \exp[\mathrm{vec}(\boldsymbol{\alpha}_t)^\top - \mathrm{vec}(\boldsymbol{B})^\top (\boldsymbol{\alpha}_{t-1}^\top \otimes \boldsymbol{\Phi}^\top)^\top \boldsymbol{J}^{-1} \\ \times (\mathrm{vec}(\boldsymbol{\alpha}_t) - \boldsymbol{\alpha}_{t-1}^\top \otimes \boldsymbol{\Phi}^\top)] \\ = \exp(\sum_{t=2}^{T} (\mathrm{vec}(\boldsymbol{\alpha}_t)^\top - \mathrm{vec}(\boldsymbol{B})^\top (\boldsymbol{\alpha}_{t-1}^\top \otimes \boldsymbol{\Phi}^\top)^\top \boldsymbol{J}^{-1} (\mathrm{vec}(\boldsymbol{\alpha}_t) - \boldsymbol{\alpha}_{t-1}^\top \otimes \boldsymbol{\Phi}^\top))) \end{aligned} \tag{3.3}$$

which will help us in the update of the coefficient matrix $\boldsymbol{B}$.

Updating $\boldsymbol{B}|\boldsymbol{\alpha}_t, \boldsymbol{\Sigma}_\eta, \boldsymbol{\Gamma}$ Under the multivariate normal prior set on $\mathrm{vec}(\boldsymbol{B})$ with a mean zero vector, i.e., $\mathrm{vec}(\boldsymbol{B}) \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Gamma} \otimes \boldsymbol{I})$, with $\boldsymbol{\Gamma} = \mathrm{diag}(\gamma_1, \ldots, \gamma_K)$ and $\gamma_k = \rho_k \tau_k^2$, the posterior distribution is derived as the product of the likelihood function (3.3) and the prior $p(\mathrm{vec}(\boldsymbol{B})|\boldsymbol{\Gamma})$, i.e.,

$$
\begin{aligned}
p(\boldsymbol{B}|\boldsymbol{\Sigma}_\eta, \boldsymbol{\alpha}_{1:T}, Y_{1:T}) \propto & p(\text{vec}(\boldsymbol{\alpha})|\boldsymbol{Y}_{1:T}, \boldsymbol{B}, \boldsymbol{\Sigma}_\eta) p(\text{vec}(\boldsymbol{B})|\boldsymbol{\Gamma}) \\
& \propto \exp[\sum_{t=2}^{T} (\text{vec}(\boldsymbol{\alpha}_t)^\top - \text{vec}(\boldsymbol{B})^\top (\boldsymbol{\alpha}_{t-1}^\top \otimes \boldsymbol{\Phi}^\top)^\top \boldsymbol{J}^{-1} \\
& \times (\text{vec}(\boldsymbol{\alpha}_t) - \boldsymbol{\alpha}_{t-1}^\top \otimes \boldsymbol{\Phi}^\top)) + \text{vec}(\boldsymbol{B})^\top (\boldsymbol{\Gamma}^{-1} \otimes \boldsymbol{I}) vec(\boldsymbol{B})] \\
& = \exp[\sum_{t=2}^{T} (\text{vec}(\boldsymbol{B})^\top (\boldsymbol{\alpha}_t^\top \otimes \boldsymbol{\Phi}^\top)^\top \boldsymbol{J}^{-1} (\boldsymbol{\alpha}_{t-1}^\top \otimes \boldsymbol{\Phi}^\top) \text{vec}(\boldsymbol{B}) \\
& - 2\text{vec}(\boldsymbol{\alpha}_t)^\top \boldsymbol{J}^{-1} (\boldsymbol{\alpha}_{t-1}^\top \otimes \boldsymbol{\Phi}^\top) \text{vec}(\boldsymbol{B})) + \text{vec}(\boldsymbol{B})^\top (\boldsymbol{\Gamma}^{-1} \otimes \boldsymbol{I}) \text{vec}(\boldsymbol{B})] \\
& \propto \exp \Bigg(\text{vec}(\boldsymbol{B})^\top (\sum_{t-1}^{T} ((\boldsymbol{\alpha}_t^\top \otimes \boldsymbol{\Phi}^\top)^\top \boldsymbol{J}^{-1} (\boldsymbol{\alpha}_{t-1}^\top \otimes \boldsymbol{\Phi}^\top)) \\
& + \boldsymbol{\Gamma}^{-1} \otimes \boldsymbol{I}) \text{vec}(\boldsymbol{B}) - 2C_1 \Bigg) \qquad (3.4)
\end{aligned}
$$

where $C_1 = -2\sum_{t=2}^{T} (vec(\boldsymbol{\alpha}_t)^\top \boldsymbol{J}^{-1})$ with $\boldsymbol{J} = \boldsymbol{\Phi}^\top \boldsymbol{\Sigma}_\eta^{-1} \boldsymbol{\Phi}$. Thus, (3.4) is the exponential part of a multivariate Normal distribution with a mean vector and covariance matrix dependent on $\boldsymbol{\alpha}_t$, i.e., $\boldsymbol{B}|\boldsymbol{\Sigma}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}_\eta, \boldsymbol{\alpha}_{1:T}, \boldsymbol{Y}_{1:T} \sim \text{N}(\tilde{\boldsymbol{\mu}}, \boldsymbol{D})$ where $\tilde{\boldsymbol{\mu}} = \boldsymbol{\mu}_1 * \boldsymbol{D}$ with $\boldsymbol{\mu}_1 = \sum_{t=2}^{T} ((\boldsymbol{\alpha}_{t-1}^\top \otimes \boldsymbol{\Phi}^\top) \boldsymbol{\Phi}^\top \boldsymbol{\Sigma}_\eta^{-1} \boldsymbol{\Phi} \boldsymbol{\alpha}_t^\top)$ and $\boldsymbol{D} = (\boldsymbol{\mu}_1 + (\boldsymbol{\Gamma} \otimes \boldsymbol{I})^{-1})^{-1}$.

The mean vector $\tilde{\boldsymbol{\mu}}$ indicates that the contribution of one location to another will be affected and expanded by both the measurements at time t and $t-1$, while they will as affect the magnitude of scaling in that contribution which is how we interpret the integro-difference equation (3.1).

3.4.2 Updating Spike and Slab hyperparameters

Updating $\boldsymbol{\beta}_k|\alpha, \Gamma, \rho, \tau_k^2, q$ for each level k For the rest of the spike and slab hierarchy, we will need to calculate the separate conditional posteriors of each individual coefficient $n \times 1$ vector $\boldsymbol{\beta}_k$ in order to improve the mathematical calculations. Thus, each vector $\boldsymbol{\beta}_k$ is marginally normally distributed with zero mean vector and covariance matrix $\gamma_k \boldsymbol{I}_{K \times K}$. Thus, for each level of $\boldsymbol{\beta}$ we are expecting to have the same scale for all locations. Therefore, the likelihood of each level can be written as $\boldsymbol{\beta}_k \sim \text{N}(\boldsymbol{0}, \gamma_k \boldsymbol{I})$,

i.e,

$$\begin{aligned} p(\boldsymbol{\beta}_k|\rho_k,\tau_k^2) &= \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\gamma_k^2}}\exp\left(-\frac{\boldsymbol{\beta}_k^\top\boldsymbol{\beta}_k}{2\gamma_k^2}\right) \\ &\propto (\gamma_k)^{-n/2}\exp\left(-\frac{\boldsymbol{\beta}_k^\top\boldsymbol{\beta_k}}{2\gamma_k^2}\right) = (\rho_k\tau_k^2)^{-n/2}\exp\left(-\frac{\boldsymbol{\beta}_k^\top\boldsymbol{\beta}_k}{2\rho_k\tau_k^2}\right) \end{aligned} \tag{3.5}$$

Updating $\rho_k|\boldsymbol{\beta}_k,\tau_k^2,q,v_0$ The individual full conditional posteriors for the latent parameter ρ_k can be derived as a mixture distribution for $k=1,\ldots,K$. The likelihood of the latent parameter ρ_k is given as:

$$p(\rho_k|v_0,q) = (1-q)\delta_{v0}(.) + q\delta_1(.) \tag{3.6}$$

where $q\sim U(0,1)$. Under the point mass $\delta_{v_0}(.)$ and under $\delta_1(.)$ the respective distribution of $\boldsymbol{\beta}_k$ can be then rewritten as:

$$\begin{aligned} p(\boldsymbol{\beta}_k|v_0,\tau_k^2) &\propto (v_0)^{-n/2}\exp\left(-\frac{\boldsymbol{\beta}_k^\top\boldsymbol{\beta}_k}{v_0\tau_k^2}\right) \text{ for } \rho_k = v_0 \\ p(\boldsymbol{\beta}_k|\tau_k) &\propto \exp\left(-\frac{\boldsymbol{\beta}_k^\top\boldsymbol{\beta}_k}{\tau_k^2}\right) \text{ for } \rho_k = 1 \end{aligned} \tag{3.7}$$

Thus, by using (3.6) and (3.7), the full conditional of the latent parameter $p(\rho_k|\boldsymbol{\beta}_k,\tau_k,q)$ can be written as:

$$\begin{aligned} \rho_k|\text{vec}(\boldsymbol{B}),\tau,q \sim \frac{q_{1,k}}{q_{1,k}+q_{2,k}}\delta_{v_0}(.) + \frac{q_{2,k}}{q_{1,k}+q_{2,k}}\delta_1(.) \quad \text{with} \\ q_{1,k} = (1-q)v_0^{-n/2}\exp\left(-\frac{\boldsymbol{\beta}_k^\top\boldsymbol{\beta}_k}{2v_0\tau_k^2}\right) \\ q_{2,k} = q\exp\left(-\frac{\boldsymbol{\beta}_k^\top\boldsymbol{\beta}_k}{\tau_k^2}\right) \end{aligned} \tag{3.8}$$

Updating $\tau_k^2|\boldsymbol{\beta}_k,\rho_k,q,\omega_1,\omega_2$ The latent parameter's weights q will be conditionally dependent on the magnitude of the scaling parameter τ_k^2. As each individual scaling parameter has an inverse gamma prior distribution, i.e., $\tau_k^2\sim \text{IG}(\omega_1,\omega_2)$, then τ_k^2 will

be updated via the conditional distribution for each $k = 1, \ldots, K$

$$\begin{aligned} p(\tau_k^{-2}|\boldsymbol{\beta}_k, \rho_k, q, \omega_1, \omega_2) =& p(\tau_k^2|\omega_1, \omega_2)p(\boldsymbol{\beta}_k|\rho_k, \tau_k) \\ \propto& (\tau_k^2)^{\omega_1 - 1} e^{-\frac{\omega_2}{\tau_k}} (\tau_k^2)^{-n/2} \exp\left(-\frac{\boldsymbol{\beta}_k^\top \boldsymbol{\beta}_k}{2\rho_k \tau_k^2} \right) \\ =& (\tau_k^2)^{-(n/2+\omega+1)} \exp\left(-\frac{1}{\tau_k^2}(\frac{\boldsymbol{\beta}_k^\top \boldsymbol{\beta}_k}{\rho_k}) + \omega_2 \right) \end{aligned} \tag{3.9}$$

which is an inverse-Gamma distribution,i.e., $\tau_k^{-2} \sim \text{IG}(\omega_1 + n/2, \omega_2 + \frac{\boldsymbol{\beta}_k^\top \boldsymbol{\beta}_k}{2\rho_k})$ which is as well conditionally updated according to the latent parameter ρ_k.

Updating $q|\rho_k, v_0$ Finally, the complexity parameter q will be updated according to the latent parameter ρ_k

$$q|\rho_k \sim \text{Beta}(1 + \#\{k : \rho_k = 1\}, 1 + \#\{k : \rho_k = v_0\}) \tag{3.10}$$

where $\#\{k : \rho_k = 1\}$ indicates the number of times the latent parameter sampled 1 for the $k - th$ level and $\#\{k : \rho_k = v_0\}$ specifies the number of times the latent parameter sampled v_0. This explains the reduce of the dimension or else the choice of the non-zero coefficients of $\boldsymbol{B}$ which are updated adaptively.

3.4.3 Updating $\boldsymbol{\alpha}_t|\boldsymbol{B}, \boldsymbol{\Sigma}_\eta, \boldsymbol{Y}_t$

As mentioned in Chapter 2, $\boldsymbol{\alpha}_t$ are unknown stochastic parameters under a state space model but they also meet the Markov property. That means that given the present value $\boldsymbol{\alpha}_t$ the past $\boldsymbol{\alpha}_{t-i}$ and the future $\boldsymbol{\alpha}_{t+j}$ are conditionally independent for any i and j. Our aim for conducting inference for these temporal parameters is to estimate the state vector $\boldsymbol{\alpha}_t$ firstly forward in time which is known as filtering and then estimate backward in time which is known as smoothing given the available dataset $\boldsymbol{Y}_t$. In the next paragraphs we briefly discuss the filtering and the sampling steps which are respectively known as Kalman filter (Kalman, 1960) and Forward Filtering Backward Sampling (Carter and Kohn (1994) and Frühwirth-Schnatter (2001)) under the model (2.8) and (2.11).

The Kalman filter relies upon the specification of the distribution of an initial state vector $\boldsymbol{\alpha}_0$; which is referred to as prior distribution because it is set prior to ob-

serving any data. Thus, by considering a prior at the initial time point for $\boldsymbol{\alpha}_0$, i.e., $\boldsymbol{\alpha}_0 \sim \mathrm{N}(\hat{\boldsymbol{\alpha}}_{0|0}, \boldsymbol{P}_{0|0})$ with $\hat{\boldsymbol{\alpha}}_{0|0}$ being a $K \times 1$ prior vector and $K \times K$ prior covariance matrix $\boldsymbol{P}_{0|0}$, the Kalman Filter recursions (Kalman, 1960) based on the observations observed up time t, i.e., $\boldsymbol{Y}_{1:t}$, for $t = 1, \ldots, T$ are given as:

- For the forecast distribution of $\boldsymbol{\alpha}_t$ at $t-1$:

$$\begin{aligned} \boldsymbol{\alpha}_t | \boldsymbol{Y}_{1:t-1}, \boldsymbol{\Sigma}, \boldsymbol{\Sigma}_\eta, \boldsymbol{B} \sim \mathrm{N}(\hat{\boldsymbol{\alpha}}_{t|t-1}, \boldsymbol{P}_{t|t-1}) \text{ , where} \\ \hat{\boldsymbol{\alpha}}_{t|t-1} = \boldsymbol{\Phi}^\top \boldsymbol{B} \hat{\boldsymbol{\alpha}}_{t-1|t-1} \text{ and} \\ \boldsymbol{P}_{t|t-1} = \boldsymbol{\Phi}^\top \boldsymbol{B} \boldsymbol{P}_{t-1|t-1} (\boldsymbol{\Phi}^\top \boldsymbol{B}^\top) + \boldsymbol{\Phi} \boldsymbol{\Sigma}_\eta \boldsymbol{\Phi}^\top \end{aligned} \tag{3.11}$$

- The posterior of $\boldsymbol{\alpha}_t$ at time t is

$$\begin{aligned} \boldsymbol{\alpha}_t | \boldsymbol{Y}_{1:t}, \boldsymbol{\Sigma}, \boldsymbol{\Sigma}_\eta, \boldsymbol{B}, \sim \mathrm{N}(\hat{\boldsymbol{\alpha}}_{t|t}, \boldsymbol{P}_{t|t}) \quad \text{where} \\ \hat{\boldsymbol{\alpha}}_{t|t} = \hat{\boldsymbol{\alpha}}_{t|t-1} + \boldsymbol{K}_t \boldsymbol{e}_t \quad \text{and} \\ \boldsymbol{P}_{t|t} = \boldsymbol{P}_{t|t-1} - \boldsymbol{K}_t \boldsymbol{F}_{t|t-1} \boldsymbol{K}_t^\top \quad \text{with} \\ \boldsymbol{Y}_{t|t-1} = \boldsymbol{\Phi}^\top \hat{\boldsymbol{\alpha}}_{t|t-1}, \boldsymbol{e}_t = \boldsymbol{Y}_t - \hat{\boldsymbol{Y}}_{t|t-1}, \\ \boldsymbol{F}_t = \boldsymbol{\Phi} \boldsymbol{P}_{t|t-1} \boldsymbol{\Phi}^\top + \boldsymbol{\Sigma} \text{ and } \boldsymbol{K}_t = \boldsymbol{P}_{t|t-1} \boldsymbol{\Phi} \boldsymbol{F}_{t|t-1}^{-1} \end{aligned} \tag{3.12}$$

with $\boldsymbol{\Sigma} = \sigma_\epsilon \boldsymbol{I}_{n \times n} + \sigma_v^2 \boldsymbol{S}$.

A few comments are in order. $\hat{\boldsymbol{\alpha}}_{t|t-1}$ represents the mean $\mathrm{E}(\boldsymbol{\alpha}_t | \boldsymbol{Y}_{1:t-1})$, i.e. the forecast of $\boldsymbol{\alpha}_t$ at time t, given the observations up to $t-1$. Then, when $\boldsymbol{Y}_t$ is observed, the data set is updated to $\boldsymbol{Y}_{1:t}$ and we derive the filtered estimate $\hat{\boldsymbol{\alpha}}_{t|t}$ which is the mean $\mathrm{E}(\boldsymbol{\alpha}_t | \boldsymbol{Y}_{1:t})$. Analogously, $\hat{\boldsymbol{Y}}_{t|t-1}$ is the one-step ahead forecast of Y_t given the information up to time $t-1$ while e_t is the one-step prediction error while $\boldsymbol{K}_t$ is known as Kalman gain. Additionally, as we have a recursive update over time, $p(\boldsymbol{\alpha}_t | \boldsymbol{Y}_{1:t-1})$ is the prior distribution at time t of $\boldsymbol{\alpha}_t$, given the past $\boldsymbol{Y}_{1:t-1}$ and prior of observing $\boldsymbol{Y}_t$ at time t. The posterior distribution of $\boldsymbol{\alpha}_t$ at time t is considered after both $\boldsymbol{Y}_t$ and $\boldsymbol{Y}_{t-1}$ are observed.

Smoothing is an important development of the theory and application of state space models and has been discussed, in Carlin et al. (1992), Catlin (2012), De Jong (1989), Angus (1992) and in Durbin and Koopman (2012). As we work under a Bayesian setting, the posterior distribution $p(\boldsymbol{\alpha}_t | \boldsymbol{\alpha}_{-t}, \boldsymbol{Y}_{1:T})$, can be obtained as a multivariate Gaussian, with $\boldsymbol{\alpha}_{-t}$ signifying the rest of time points in states $\boldsymbol{\alpha}$ except the one at

time t. Since we can sample from $p(\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{-t})$, this provides a single step of the Gibbs sampler where it is noted that at time $t = T$ we sample from the posterior $\boldsymbol{\alpha}_T|\boldsymbol{Y}_{1:T}$, which by the Kalman Filter recursions is again Gaussian. This approach was proposed by Carlin et al. (1992) together with extensions to non-linear and non-Gaussian state space models. Unfortunately, this approach can be very inefficient, because the prior correlation imposed in the system of state vectors $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1^\top, \ldots, \boldsymbol{\alpha}_T^\top]$ is largely transferred to the posterior state vectors $\boldsymbol{\alpha}|\boldsymbol{Y}_{1:T}$. The aforesaid chain correlation along with the high dimensional state space imposed by the time series introduces convergence problems in the Gibbs sampler and it slows it down considerably.

As a result alternative Gibbs sampling schemes are proposed in the literature, Carter and Kohn (1994) and Frühwirth-Schnatter (1994) where independently proposed a block application of Gibbs sampling, which is considerably more stable and orders of magnitude faster than the above scheme, as reported in Shephard (1994). According to this instead of sampling from $\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{-t}, \boldsymbol{Y}_{1:T}$ we can successively sample from just $\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t+1}, Y_{1:T}$.

In Gibbs sampling for the states our target full conditional distribution for the state vector $\boldsymbol{\alpha}_t$ is $p(\boldsymbol{\alpha}_{1:T}|\boldsymbol{Y}_{1:T}, \boldsymbol{\Sigma}, \boldsymbol{\Sigma}_\eta, \boldsymbol{B})$. Carter and Kohn (1994) developed the method mentioned above which is known as Forward Filter Backward Sampling algorithm. In this context we describe the general approach where we have a multivariate model with unknown covariance matrices subject to be estimated. By rewriting

$$\begin{aligned} p(\boldsymbol{\alpha}_{1:T}|\boldsymbol{Y}_{1:T}, \sigma_v, \sigma_\epsilon, \boldsymbol{\Sigma}_\eta, \boldsymbol{B}) &= \prod_{t=1}^{T-1} p(\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t+1}, \boldsymbol{Y}_{1:T}, \sigma_v, \sigma_\epsilon, \boldsymbol{\Sigma}_\eta, \boldsymbol{B}) \\ &\quad \times p(\boldsymbol{\alpha}_T|\boldsymbol{Y}_{1:T}, \sigma_v, \sigma_\epsilon, \boldsymbol{\Sigma}_\eta, \boldsymbol{B}) \\ &= \prod_{t=1}^{T-1} p(\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t+1}, \boldsymbol{Y}_{1:T}, \sigma_v, \sigma_\epsilon, \boldsymbol{\Sigma}_\eta, \boldsymbol{B}) \\ &\quad \times p(\boldsymbol{\alpha}_T|\boldsymbol{Y}_{1:T}, \sigma_v, \sigma_\epsilon, \boldsymbol{\Sigma}_\eta, \boldsymbol{B}) \qquad (3.13) \\ &= \prod_{t=1}^{T-1} p(\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t+1}, \boldsymbol{Y}_{1:t}, \sigma_v, \sigma_\epsilon, \boldsymbol{\Sigma}_\eta, \boldsymbol{B}) p(\boldsymbol{\alpha}_T|\boldsymbol{Y}_{1:T}, \sigma_v, \sigma_\epsilon, \boldsymbol{\Sigma}_\eta, \boldsymbol{B}) \qquad (3.14) \end{aligned}$$

where $\boldsymbol{\alpha}_{1:T} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_T)$. We note that (3.13) is obtained due to conditional independence of $\boldsymbol{\alpha}_t, \boldsymbol{\alpha}_{t+1}|\boldsymbol{\alpha}_{t+2}, \ldots, \boldsymbol{\alpha}_T$ (i.e., given the present, the future and the past are

conditionally independent). Analogously, jointly, $\boldsymbol{Y}_t, \boldsymbol{\alpha}_t | \boldsymbol{Y}_{t+1}, \ldots, \boldsymbol{Y}_T$ are conditionally independent and we end up on (3.14).

At the end of the filtering step we obtain $\boldsymbol{\alpha}_T$ from $N(\boldsymbol{\alpha}_{T|T}, \boldsymbol{P}_{T|T})$ and then we use the following smoothing recursions in order to draw $\boldsymbol{\alpha}_t$ for $t = T, \ldots, 1$:

$$\begin{aligned} \boldsymbol{\alpha}_t | \boldsymbol{\alpha}_{t+1}, Y_{1:T}, \boldsymbol{\Sigma}, \boldsymbol{\Sigma}_\eta, \boldsymbol{B} \sim N(\hat{\boldsymbol{\alpha}}_{t|t+1}, \boldsymbol{P}_{t|t+1}) \quad \text{where} \\ \hat{\boldsymbol{\alpha}}_{t|t+1} = \hat{\boldsymbol{\alpha}}_{t|t} + \boldsymbol{L}_t(\boldsymbol{\alpha}_{t+1} - \hat{\boldsymbol{\alpha}}_{t+1|t}) \text{ and} \\ \boldsymbol{P}_{t|t+1} = \boldsymbol{P}_{t|t} - \boldsymbol{L}_t \boldsymbol{P}_{t+1|t} \boldsymbol{L}_t^\top \quad \text{with} \\ \boldsymbol{L}_t = \boldsymbol{P}_{t|t}(\boldsymbol{\Phi}^\top \boldsymbol{B})^\top \boldsymbol{P}_{t+1|t}^{-1} \end{aligned} \tag{3.15}$$

It is noticeable that the FFBS algorithm does not require a prior state vector $\boldsymbol{\alpha}_0$. In other words, the Kalman Filter provides a learned procedure in order for the smoothing recursions to take place. Secondly, the covariance matrix $\boldsymbol{P}_{t|t} = \boldsymbol{P}_{t|t+1}$ does not depend on the future value at $t+1$ and can be provided by the Kalman Filter in the first step. This can result in significant computational savings, as only the computation of the mean vector $\boldsymbol{\alpha}_{t|t+1}$ is needed to simulate the vector $\boldsymbol{\alpha}_{t+1}$ at each Gibbs iteration.

3.4.4 Updating $\boldsymbol{\Sigma}_\eta | \boldsymbol{\alpha}_t, \boldsymbol{\alpha}_{t-1}, \boldsymbol{B}$

The FFBS algorithm stated above that was proposed in Carter and Kohn (1994) considered the covariance inference by placing improper priors on the observation and transition variances, resulting in proper inverse gamma priors for these variances in Fearnhead (2002) and Carter and Kohn (1994).

On the other hand, Carter and Kohn (1994) introduced the d-inverse gamma state space model, whereby the variance of the transition innovation vector η_t is diagonal, each element of its main diagonal independently following a priori an inverse gamma distribution.

We introduce the simplest diagonal case for the covariance matrix $\boldsymbol{\Sigma}_\eta$, the d-inverse gamma approach and finally the inferential stage by considering a correlation structure under the DSTM model.

$\boldsymbol{\Sigma}_\eta = \sigma_\eta^2 \boldsymbol{I}$ with common elements. If we consider a diagonal structure with common variances on the transition covariance matrix, i.e., $\boldsymbol{\Sigma}_\eta = \sigma_\eta^2 \boldsymbol{I}$, then the locations

$\boldsymbol{s} = \{s_1, \ldots, s_n\}$ are considered to be temporally independent to each other but all locations to themselves will vary under the same magnitude σ_η^2 autoregressively. Under this approach, it s plausible that a gamma prior will be placed onto the precision element.

Therefore, we set $\boldsymbol{\Sigma}_\eta = \sigma_\eta^2 \boldsymbol{I}_{n\times n}$, where $\sigma_\eta^2 \sim \text{IG}(\psi_1, \psi_2)$. The posterior distribution $p(\sigma_\eta^2|\boldsymbol{\alpha}_t, \boldsymbol{\alpha}_{t-1}, \psi_1, \psi_2)$ can be derived as a product of the prior $p(\sigma_\eta^2|\psi_1, \psi_2)$ and the likelihood $p(\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1}, B, \sigma_\eta^2)$, i.e.,

$$\begin{aligned} p(\sigma_\eta^2|\boldsymbol{\alpha}_t, \boldsymbol{\alpha}_{t-1}, \psi_1, \psi_2) \propto & p(\sigma_\eta^2|\psi_1, \psi_2) p(\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1}, \boldsymbol{B}, \sigma_\eta^2) \\ \propto & \frac{1}{\sqrt{2\pi}} (\sigma_\eta^2)^{-(n+T)/2} (\sigma_\eta^2)^{-\psi_1 - 1} \exp(-\frac{\psi_2}{\sigma_\eta^2}) \\ & \times \exp(-\frac{1}{2\sigma_\eta^2} \sum_{i=1}^{n} \sum_{t=2}^{T+1} (\boldsymbol{\alpha}_t - \boldsymbol{\Phi}^\top \boldsymbol{B} \boldsymbol{\alpha}_{t-1})^\top (\boldsymbol{\alpha}_t - \boldsymbol{\Phi}^\top \boldsymbol{B} \boldsymbol{\alpha}_{t-1})) \\ = & (\sigma_\eta^2)^{-\psi_1 - \frac{(n+T)}{2} - 1} \exp(-\frac{C/2 + \psi_2}{\sigma_\eta^2}) \end{aligned} \tag{3.16}$$

where $C = \sum_{i=1}^{n} \sum_{t=2}^{T+1} (\boldsymbol{\alpha}_t - \boldsymbol{\Phi}^\top \boldsymbol{B} \boldsymbol{\alpha}_{t-1})^\top (\boldsymbol{\alpha}_t - \boldsymbol{\Phi}^\top \boldsymbol{B} \boldsymbol{\alpha}_{t-1})$. The equation (3.16) provides us with the form of an inverse gamma distribution with shape and scale parameters $\psi_1 + (n+T)/2$ and $C/2 + \psi_2$ respectively. Therefore, from (3.16) we conclude that the full conditional posterior for the common transition variance is distributed as $\sigma_\eta^2|\boldsymbol{\alpha}_t, \boldsymbol{\alpha}_{t-1}, \psi_1, \psi_2 \sim \text{IG}(\psi_1 + (n+T)/2, C/2 + \psi_2)$.

Diagonal $\boldsymbol{\Sigma}_\eta$ with different elements— inverse d-gamma approach Analogously to the simple diagonal case, the transition covariance matrix is considered to have a temporal independent structure for the locations $\boldsymbol{s} = \{s_1, \ldots, s_n\}$ to each other but with the difference that the locations themselves will vary with a different magnitude autoregressively, i.e, $\boldsymbol{\Sigma}_\eta = \text{diag}\{\sigma_{\eta_1}^2, \ldots, \sigma_{\eta_n}^2\}$. Consequently, n independent inverse-gamma priors should be considered for each variance element $\sigma_{\eta i}$, with $i = 1, \ldots, n$.

Hence, if we set $\boldsymbol{\Sigma}_\eta = \text{diag}\{\sigma_{\eta_1}^2, \ldots, \sigma_{\eta_n}^2\}$ where each $\sigma_{\eta_i}^2 \sim \text{IG}(\psi_1, \psi_2)$ with $i = 1, \ldots, n$ and each posterior distribution $p(\sigma_{\eta_i}^2|\alpha_{t,i}, \alpha_{t-1,i}, \psi_1, \psi_2)$ can be derived as a product of the prior $p(\sigma_{\eta_i}^2|\psi_1, \psi_2)$ and the univariate likelihood of α_{ti}, $i = 1, \ldots, n$ for each

location, i.e., $p(\alpha_{t,i}|\alpha_{t-1,i}, \boldsymbol{B}, \sigma^2_{\eta_i})$:

$$\begin{aligned} p(\sigma^2_{\eta_i}|\alpha_{t,i}, \alpha_{t-1,i}, \psi_1, \psi_2) \propto & p(\sigma^2_{\eta_i}|\psi_1, \psi_2) p(\alpha_{t,i}|\alpha_{t-1,i}, \boldsymbol{B}, \sigma^2_{\eta}) \\ \propto & \frac{1}{\sqrt{2\pi}} (\sigma^2_{\eta_i})^{-T/2} (\sigma^2_{\eta_i})^{-\psi_1 - 1} \exp(-\frac{\psi_2}{\sigma^2_{\eta_i}}) \\ & \times \exp(-\frac{1}{2\sigma^2_{\eta_i}} \sum_{t=2}^{T+1} (\alpha_{t,i} - \phi_i^\top \beta_j \alpha_{t-1,i})^2) \\ = & (\sigma^2_{\eta_i})^{-\psi_1 - \frac{T}{2} - 1} \exp(-\frac{C_i/2 + \psi_2}{2\sigma^2_{\eta_i}}) \end{aligned} \tag{3.17}$$

where $C_i = \sum_{t=2}^{T+1} (\alpha_{t,i} - \phi_i^\top \beta_j \alpha_{t-1,i})^2$ with ϕ_i indicating the i-th row of the Wavelet matrix $\boldsymbol{\Phi}$ and β_j indicating the $j = i$-th column of the coefficient matrix $\boldsymbol{B}$. The full conditional posterior (3.17) is the form of an inverse gamma distribution with shape and scale parameters $\psi_1 + T/2$ and $C_i/2 + \psi_2$ respectively and thus we conclude that the updating for each variance element $\sigma^2_{\eta i}$ will be performed through $\sigma^2_{\eta_i}|\alpha_{t,i}, \alpha_{t-1,i}, \psi_1, \psi_2 \sim \text{IG}(\psi_1 + T/2, C_i/2 + \psi_2)$.

Covariance structure on Σ_η The more general approach is placing an inverse Wishart prior into the transition covariance matrix Σ_η and thus allow us to estimate the correlation between the elements η_t and thus the temporal correlation between the locations' evolution.

By using the likelihood formulation through the FFBS scheme in (3.14) and by considering that *a priori* Σ_η follows an inverse Wishart distribution, i.e.,

$$p(\boldsymbol{\Sigma}_\eta) = c|\boldsymbol{\Sigma}_\eta|^{-(\nu+n+1)/2} \exp(-\frac{1}{2} \text{trace}(\boldsymbol{Q}\boldsymbol{\Sigma}_\eta^{-1})) \tag{3.18}$$

where $\boldsymbol{\Sigma}_\eta$ is the prior scale matrix, ν the prior degrees of freedom and c is the proportionality constant, then the conditional distribution $p(\boldsymbol{\Sigma}_\eta|\boldsymbol{\alpha}_{1:T}, \boldsymbol{Y}_{1:T}, \boldsymbol{B})$ is

$$
\begin{aligned}
p(\boldsymbol{\Sigma}_\eta|\boldsymbol{\alpha}_{1:T}, Y_{1:T}, B) \propto & p(\boldsymbol{\alpha}_{1:T}|\boldsymbol{\Sigma}_\eta, \boldsymbol{B})p(\boldsymbol{\Sigma}_\eta) = \prod_{t=1}^{T} p(\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1}, \boldsymbol{\Sigma}_\eta)p(\boldsymbol{\Sigma}_\eta) \\
& \propto \prod_{t=1}^{T} |\boldsymbol{\Sigma}_\eta|^{-1/2} \exp[-\frac{1}{2}\operatorname{trace}\{(\boldsymbol{\alpha}_t - \boldsymbol{\Phi}^\top \boldsymbol{B}\boldsymbol{\alpha}_{t-1})^\top \\
& \times(\boldsymbol{\alpha}_t - \boldsymbol{\Phi}^\top \boldsymbol{B}\boldsymbol{\alpha}_{t-1})\boldsymbol{\Sigma}_\eta^{-1}\}]|\boldsymbol{\Sigma}_\eta|^{-(\nu+d+1)/2} \exp(-\frac{1}{2}\operatorname{trace}(\boldsymbol{Q}\boldsymbol{\Sigma}_\eta^{-1})) \\
& \propto |\boldsymbol{\Sigma}_\eta|^{-(\nu+T+n+1)/2} \exp[-\frac{1}{2}\operatorname{trace}\{(\boldsymbol{C}+\boldsymbol{Q})\boldsymbol{\Sigma}_\eta^{-1}\}] \qquad (3.19)
\end{aligned}
$$

with $\boldsymbol{C} = \sum_{t=1}^{T}(\boldsymbol{\alpha}_t - \boldsymbol{\Phi}^\top \boldsymbol{B}\boldsymbol{\alpha}_{t-1})^\top(\boldsymbol{\alpha}_t - \boldsymbol{\Phi}^\top \boldsymbol{B}\boldsymbol{\alpha}_{t-1})$ which is proportional to an inverse Wishart distribution with parameters $\nu+T$ and $\boldsymbol{C}+\boldsymbol{Q}$ respectively, i.e., $\boldsymbol{\Sigma}_\eta|\boldsymbol{\alpha}_{1:T}, \boldsymbol{Y}_{1:T}, \boldsymbol{B} \sim \text{IW}(\nu+T, \boldsymbol{C}+\boldsymbol{Q})$. From the above full conditional distributions and given the posteriors for $\boldsymbol{\alpha}$ and $\boldsymbol{B}$ and according to the desired structure of the covariance matrix $\boldsymbol{\Sigma}_\eta$ we can easily sample the variance estimates. The choice of the structure of the covariance matrix is dependent solely on the application.

3.4.5 Updating $\sigma_\epsilon^2, \sigma_\nu^2|\boldsymbol{\alpha}_t, \boldsymbol{Y}_t$

Estimating both σ_ν^2 and σ_ϵ^2 is challenging. Specifically, a mutual sampling for σ_ϵ^2 and σ_v^2 provides difficulties for multi-layer models such as DSTMs. Therefore, the most flexible choice is to either consider σ_ϵ^2 or their ratio —which is known as the signal-to-noise ratio— to be known and sample σ_v^2 accordingly. The approaches aforementioned are being explained along with our investigations into their relative success in the next paragraphs.

Jointly sampling of $\sigma_\epsilon^2, \sigma_\nu^2|\boldsymbol{\alpha}_t, \boldsymbol{Y}_t$ Many approaches were conducted to estimate jointly σ_ϵ^2 and σ_v^2 via the Metropolis-Hastings algorithm (Chib and Greenberg, 1995) by working on the joint full conditional $\sigma_\epsilon^2, \sigma_v^2|\boldsymbol{Y}_t, \boldsymbol{\alpha}_t, \boldsymbol{B}, \boldsymbol{\Sigma}_\eta$. Specifically, two separate inverse gamma priors on both variance parameters were considered, i.e., $\sigma_\nu^2 \sim \text{IG}(\delta_0, \xi_0)$ and $\sigma_\epsilon^2 \sim \text{IG}(\delta_0, \xi_0)$. The joint full conditional can be written as

$$p(\sigma_\epsilon^2, \sigma_v^2|\boldsymbol{Y}_t, \boldsymbol{\alpha}_t, \boldsymbol{B}, \boldsymbol{\Sigma}_\eta) \propto |\sigma_v^2\boldsymbol{S} + \sigma_\epsilon^2\boldsymbol{I}|^{-n/2}(\sigma_v^2)^{-\delta_0-1}\exp(-\xi_0/\sigma_v^2)(\sigma_\epsilon^2)^{-\delta_0-1}$$
$$\times \exp\Bigg(-\frac{1}{2}\,\text{trace}\,\big[\sum_{t=1}^{T}(\boldsymbol{Y}_t - \boldsymbol{\Phi}^\top\boldsymbol{\alpha}_t)(\boldsymbol{Y}_t - \boldsymbol{\Phi}^\top\boldsymbol{\alpha}_t)^\top$$
$$\times (\sigma_v^2\boldsymbol{S} + \sigma_\epsilon^2\boldsymbol{I})^{-1}\big]\Bigg)\exp(-\xi_0/\sigma_\epsilon^2) \tag{3.20}$$

where a known form of distribution cannot be derived. Unfortunately, a Metropolis-Hastings iterative step cannot be considered for the joint vector in this case neither a Metropolis-adjusted Langevin algorithm (Roberts and Rosenthal, 1998). A Metropolis-Hastings step as stated above results into non-convergence when the covariance parameters are jointly estimated. This is due to the presence of an unidentifiable parameter vector by the likelihood. For that reason, in the next paragraphs we investigate two approaches that provided us convergence and fair estimations.

Fixing σ_ϵ^2 to update $\sigma_\nu^2|\boldsymbol{\alpha}_t, \boldsymbol{Y}_t$ Based on the application we might have information on the measurement error variance σ_ϵ either due to previous modelling procedures or for instance we have prior knowledge of the instrument that we got the measurement from. Thus, by fixing σ_ϵ to be known, then (3.20) takes the form of the product of the observed process's likelihood and the prior of σ_v

$$p(\sigma_v^2|\boldsymbol{Y}_t, \boldsymbol{\alpha}_t, \boldsymbol{B}, \boldsymbol{\Sigma}_\eta) \propto |\sigma_v^2\boldsymbol{S} + \sigma_\epsilon^2 I|^{-n/2} \times (\sigma_v^2)^{-\delta_0-1}\exp(-\xi_0/\sigma_v^2)$$
$$\times \exp\Bigg(-\frac{1}{2}tr\big[\sum_{t=1}^{T}(\boldsymbol{Y}_t - \boldsymbol{\Phi}^\top\boldsymbol{\alpha}_t)(\boldsymbol{Y}_t - \boldsymbol{\Phi}^\top\boldsymbol{\alpha}_t)^\top(\sigma_v^2\boldsymbol{S} + \sigma_\epsilon^2\boldsymbol{I})^{-1}\big]\Bigg) \tag{3.21}$$

and can be sampled under the univariate Metropolis-Hastings contexts that are described above.

Fixing $\sigma_\epsilon^2/\sigma_\nu^2$ to update $\sigma_\nu^2|\boldsymbol{\alpha}_t, \boldsymbol{Y}_t$ Finally, the most convenient approach is if the signal to noise ratio is considered known instead, i.e., $\sigma_\epsilon^2/\sigma_v^2 = c$. This knowledge can be as well obtained based on the application and the instrument that we extracted the observations $\boldsymbol{Y}_t$ from. Thus, inference on σ_v is conducted, then we have a more

simplified version than (3.21) and thus the posterior distribution is of known form, i.e., (3.21) takes the form:

$$\begin{aligned}
p(\sigma_v^2|\boldsymbol{Y}_t,\boldsymbol{\alpha}_t,\boldsymbol{B},\boldsymbol{\Sigma}_\eta,\sigma_\epsilon^2) \propto & |\sigma_v^2\boldsymbol{S}+\sigma_\epsilon^2\boldsymbol{I}|^{-(n+T)/2}(\sigma_v^2)^{-\delta_0-1}\exp(-\xi_0/\sigma_v^2) \\
& \times\exp\bigg(-\frac{1}{2}[\sum_{t=1}^{T}(\boldsymbol{Y}_t-\boldsymbol{\Phi}^\top\boldsymbol{\alpha}_t)^\top(\sigma_v^2\boldsymbol{S}+\sigma_\epsilon^2\boldsymbol{I})^{-1}(\boldsymbol{Y}_t-\boldsymbol{\Phi}^\top\boldsymbol{\alpha}_t)]\bigg) \\
= & |\sigma_v^2\boldsymbol{S}+c\sigma_v^2\boldsymbol{I}|^{-(n+T)/2}\times(\sigma_v^2)^{-\delta_0-1}\exp(-\xi_0/\sigma_v^2) \\
& \times\exp\bigg(-\frac{1}{2}[\sum_{t=1}^{T}(\boldsymbol{Y}_t-\boldsymbol{\Phi}^\top\boldsymbol{\alpha}_t)^\top(\sigma_v^2 S+c\sigma_v^2 I)^{-1}(\boldsymbol{Y}_t-\boldsymbol{\Phi}^\top\boldsymbol{\alpha}_t)]\bigg) \\
= & (\sigma_v^2)^{-(n+T)/2-\delta_0-1}|\boldsymbol{S}+c\boldsymbol{I}|^{-(n+T)/2}\exp\bigg(-\frac{1}{\sigma_v^2}(\frac{\boldsymbol{C}}{2}+\xi_0)\bigg) \qquad (3.22)
\end{aligned}$$

with $\boldsymbol{C}=\big[\sum_{t=1}^{T}(\boldsymbol{Y}_t-\boldsymbol{\Phi}^\top\boldsymbol{\alpha}_t)^\top(\boldsymbol{S}+c\boldsymbol{I})^{-1}(\boldsymbol{Y}_t-\boldsymbol{\Phi}^\top\boldsymbol{\alpha}_t)\big]$, which is the form of an inverse gamma distribution with shape $n/2+\delta_0$ and scale $\boldsymbol{C}/2+\xi_0$, i.e., $\sigma_v|\boldsymbol{Y}_t,\boldsymbol{\alpha}_t,\boldsymbol{B},\boldsymbol{\Sigma}_\eta,\sigma_\epsilon^2\sim$ IG$((n+T)/2+\delta_0,\boldsymbol{C}/2+\xi_0)$ and can be sampled with Gibbs steps in the hierarchy (3.2).

3.4.6 Updating the spatial correlation function's S parameters

As discussed in Chapter 2, the spatially varying error ν_s is affected by an appropriately selected correlation function and the scaling variance σ_ν. The choice of that correlation function depends on the application, however, all of them have scaling (and smoothing) parameters to be inferred unless we choose to fix them. There are numerous isotropic parametric spatial correlation functions in the bibliography (Yaglom (1987); Abrahamsen (1997); MacKay (1998)) which are based on the spatial distance h and a spatial scale parameter λ such as the Exponential or the Gaussian correlation functions. In such cases, *a priori* information can be placed into the scaling parameter and work under a Metropolis-Hastings framework.

We provide an example under an exponential and a Matèrn correlation function (Matérn, 1960) but in most correlation functions the proofs are analogous. The spatial correlation exponential function is given as

$$S(h;\theta)=\begin{cases}(1-\exp(-|h/\lambda|)) & h>0,\lambda>0\\ 0 & \text{otherwise}\end{cases}$$

with h indicating the distance between the locations and λ being the scaling parameter. Generally, λ controls how fast the correlation decays with distance, which determines the coarse-scale behaviour of the sample path that is generated by the stochastic process of interest with the given correlation function (i.e., $\boldsymbol{X}_t$).

In order to conduct inference on the scaling parameter λ, due to the nature of λ, a gamma prior $\mathrm{G}(u_1, u_2)$ is appropriate and thus the posterior $p(\lambda|\boldsymbol{Y}_t, \boldsymbol{\alpha}_t, \sigma_\epsilon, \sigma_v)$ is the product of the process's likelihood and the prior:

$$\begin{aligned} p(\lambda|\boldsymbol{Y}_t, \alpha_t, \sigma_\epsilon, \sigma_v) \propto & |\sigma_v^2 \boldsymbol{S} + \sigma_\epsilon^2 \boldsymbol{I}|^{-n/2} \times \lambda^{-u_1-1} \exp(-u_2\lambda) \\ & \times \exp\Bigg(-\frac{1}{2}\big[\sum_{t=1}^{T} (\boldsymbol{Y}_t - \boldsymbol{\Phi}^\top \boldsymbol{\alpha}_t^\top)(\sigma_v^2 \boldsymbol{S}_\lambda + \sigma_\epsilon^2 \boldsymbol{I})^{-1} (\boldsymbol{Y}_t - \boldsymbol{\Phi}^\top \boldsymbol{\alpha}_t)\big]\Bigg) \end{aligned} \tag{3.23}$$

which is not of known form and can be sampled through Metropolis-Hastings steps.

Another example of correlation function is provided by the Matèrn family (Matérn, 1960) which is defined as a function of the spatial distance h and two other parameters:

$$S(h;\theta) = \begin{cases} \frac{1}{\Gamma(\lambda)} (\frac{\zeta h}{2})^\lambda 2K_\lambda(\zeta h) & h > 0, \zeta, \lambda > 0 \\ 0 & \text{otherwise} \end{cases}$$

where K_λ is a modified Bessel function (see Matérn (1960)), ζ is the parameter that defines the extend of the spatial dependence and λ is a smoothing parameter, which applies smoothness over the dependence range and controls how many times differentiable is the correlation function at $h = 0$. Diggle et al. (2003) suggests that appropriate value for λ can be chosen to reflect scientific knowledge about the smoothness of the process under study. Therefore, it can be either consider fixed, or a suitable prior knowledge can be considered and then resort to a bivariate Metropolis-Hastings estimation.

3.4.7 Summary and pseudo-code of the algorithm

To sum up, the inferential stage is consisted of an adaptive MCMC procedure for the covariance inference updated from (3.16) to (3.19) for Σ_η and (3.20) to (3.22) for σ_v^2 and σ_ϵ^2. Furthermore, the Spike and Slab hierarchy parameters are updated in Gibbs

Sampling steps through the posterior densities (3.8) to (3.10) for the parameters ρ_k, τ_k and u with the final update on the matrix $\boldsymbol{B}$ through the posterior (3.4). Then, the weight function $w_s(u)$ is calculated deterministically from the Inverse Discrete Wavelet Transform (IDWT). Finally, The underlying process' coefficients $\boldsymbol{\alpha}_t$ are inferred firstly through the Kalman Filter recursions (3.11) and (3.12) and then the smoothed estimates are derived through Forward Filter Backward Sampling algorithm steps through (3.15). Finally, the underlying process $X_k(s,t)$ is again deterministically calculated under the IDWT framework.

Initial step:
Draw $\sigma_\epsilon^0, \sigma_v^0, w^0, \tau^0, \rho^0, \boldsymbol{\Sigma}_\eta^0, \boldsymbol{\alpha}^0$ and $\boldsymbol{B}^0$ from the priors
For $i \geq 1$ assign $\sigma_\epsilon = \sigma_\epsilon^{(i-1)}$,$\sigma_v = \sigma_v^{(i-1)}$, $\boldsymbol{\Sigma}_\eta = \boldsymbol{\Sigma}_\eta^{(i-1)}$
Update B: Spike and Slab hierarchy
Sample:
$\rho_k^{(i+1)}$ from the conditional mixture distribution (3.8)
$\tau_k^{-2^{(i+1)}}$ from the conditional Gamma distribution (3.9)
$\boldsymbol{B}^{(i+1)}$ from the Normal distribution (3.4)
the complexity parameter $q^{(i+1)}$ from Beta distribution (3.10)
Deterministically calculate $w_s = \boldsymbol{B\Phi}$
Update $\boldsymbol{\alpha}$: FFBS algorithm
For each $t = 1, \ldots, T$
Run Kalman Filter to obtain $\hat{\boldsymbol{\alpha}}_{t+1\|t}$,$\boldsymbol{\alpha}_{t\|t}$,$\boldsymbol{P}_{t\|t}$, $\boldsymbol{P}_{t+1\|t}$ from (3.12)
For each $t = T, T-1, \ldots, 1$
Use the smooth recursions to derive $\boldsymbol{\alpha}_t^{(i)}$ from (3.15)
Deterministically calculate $X_K = \boldsymbol{\alpha\Phi}$
Update Covariance and correlation parameters
If σ_ϵ known then update σ_v through Metropolis-Hastings from (3.21)
If σ_ϵ/σ_v known then update σ_v through the inverse Gamma distribution (3.22)
Sample $\boldsymbol{\Sigma}_\eta$ with equations (3.16) -(3.19) based on the structure
Update the correlation parameters λ through Metropolis-Hastings

Table 3.2: Pseudo-code of the MCMC approach

3.5 Posterior Predictive Distribution

In spatio-temporal framework the aim is to achieve spatial interpolation of the process of unmonitored locations and temporal forecasting. In this section we consider temporal ℓ-step ahead forecasting at the monitoring locations, then by defining $\boldsymbol{D}_t = \{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_t\}$, it is given by the following posterior predictive distribution:

$$p(\boldsymbol{Y}_{t+\ell}|\boldsymbol{D}_t) = \int \cdots \int \prod_{h=1}^{\ell} p(\boldsymbol{Y}_{t+h}|\boldsymbol{\alpha}_{t+h}, \sigma_\epsilon^2, \sigma_\nu^2, \boldsymbol{\theta}, \boldsymbol{\Sigma_\eta}) p(\boldsymbol{\alpha}_{t+h}|\boldsymbol{\alpha}_{t+h-1}, \boldsymbol{B}, \boldsymbol{\Sigma}_\eta) \times p(\boldsymbol{\alpha}|\boldsymbol{D}_t) p(\boldsymbol{B}) p(\boldsymbol{\Sigma}_\eta) p(\sigma_\nu^2) p(\boldsymbol{\theta}) d\boldsymbol{\alpha} d\tilde{\boldsymbol{\alpha}} d\boldsymbol{B} d\boldsymbol{\Sigma}_\eta d\sigma_\nu^2 d\boldsymbol{\theta} \quad (3.24)$$

where $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_T)$ and $\tilde{\boldsymbol{\alpha}} = (\boldsymbol{\alpha}_{t+1}, \ldots, \boldsymbol{\alpha}_{t+\ell})$. The integral at (3.24) is approximated by

$$p(\boldsymbol{Y}_{t+\ell}|\boldsymbol{D}_t) \approx \frac{1}{M} \sum_{m=1}^{M} p(\boldsymbol{Y}_{t+\ell}|\boldsymbol{\alpha}_{t+\ell}^{(m)}, \sigma_\nu^{(m)}, \boldsymbol{B}^{(m)}, \boldsymbol{\Sigma}_\eta^{(m)}, \boldsymbol{\theta}^{(m)}) \quad (3.25)$$

where m denotes the m-th MCMC and FFBS iteration of the samples from the posteriors of $(\sigma_\nu^2, \boldsymbol{B}, \boldsymbol{\Sigma}_\eta, \boldsymbol{\theta})$ and of $(\boldsymbol{\alpha}_{t+1}, \ldots, \boldsymbol{\alpha}_{t+\ell})$ respectively. Samples can be obtained by propagating $\boldsymbol{\alpha}_{t+\ell}$ following the transition equation and through the samples from the posterior distribution of the parameter vector.

Assume we want to predict the process of a vector of dimension ℓ, with ungauged locations at an observed time point $t \in T$, i.e., $\tilde{\boldsymbol{Y}}_t = (\tilde{Y}_t(s_1), \ldots, \tilde{Y}_t(s_\ell))^\top$. Considering the spatio-temporal vector $\boldsymbol{Y} = (Y(s_1, 1), \ldots, Y(s_n, 1), \ldots, Y(s_1, T), \ldots, Y(s_n, T))^\top$ which represents the vector of time series observed at n monitoring locations, the predictive posterior distribution is given by

$$p(\tilde{\boldsymbol{Y}}_t|\boldsymbol{Y}) = \int_{\boldsymbol{\theta}} p(\tilde{\boldsymbol{Y}}_t|\boldsymbol{\theta}, \boldsymbol{Y}) p(\boldsymbol{\theta}|\boldsymbol{Y}) d\boldsymbol{\theta} \quad (3.26)$$

where $\boldsymbol{\theta}$ is the parameter vector that encompasses all the unknowns in the model.

Thus, by considering that the parameters σ_ϵ^2, σ_ν^2 and the correlation function's parameters are included in the variance matrix $\boldsymbol{\Sigma} = \sigma_\epsilon^2 \boldsymbol{I}_{n\times n} + \sigma_\nu^2 \boldsymbol{S}_{\boldsymbol{\theta}}$, then, (3.26) can be explicitly written as

$$p(\tilde{\boldsymbol{Y}}_t|\boldsymbol{Y}) = \int \cdots \int p(\tilde{\boldsymbol{Y}}_t|\boldsymbol{Y}, \sigma_\epsilon^2, \sigma_\nu^2, \boldsymbol{\alpha}, \boldsymbol{\theta}) p(\sigma_\epsilon^2, \sigma_\nu^2, \boldsymbol{\alpha}, \boldsymbol{\theta}|\tilde{\boldsymbol{Y}}_t) d\sigma_\epsilon^2 d\sigma_\nu^2 d\boldsymbol{\theta} d\boldsymbol{\alpha} \quad (3.27)$$

Spatial interpolation can be obtained by considering the jointly $(\boldsymbol{Y}_t, \tilde{\boldsymbol{Y}}_t)$ conditional on the unknown parameters and the prior beliefs $\boldsymbol{\theta}_0$, i.e.,

$$\begin{pmatrix} \boldsymbol{Y}_t \\ \tilde{\boldsymbol{Y}}_t \end{pmatrix} | \boldsymbol{\alpha}, \boldsymbol{\Sigma} \sim \mathrm{N}\left(\begin{pmatrix} \boldsymbol{\Phi}\boldsymbol{\alpha}_t \\ \tilde{\boldsymbol{\Phi}}\boldsymbol{\alpha}_t \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{Y}} & \boldsymbol{\Sigma}_{\boldsymbol{Y}\tilde{\boldsymbol{Y}}} \\ \boldsymbol{\Sigma}_{\boldsymbol{Y}\tilde{\boldsymbol{Y}}} & \boldsymbol{\Sigma}_{\tilde{\boldsymbol{Y}}} \end{pmatrix} \right) \tag{3.28}$$

which then gives us the marginal conditional posterior distribution for spatial interpolation

$$\tilde{\boldsymbol{Y}}_t | \boldsymbol{Y}_t, \sigma_\epsilon^2, \sigma_\nu^2, \boldsymbol{\alpha}_t, \boldsymbol{\theta} \sim \mathrm{N}\left(\tilde{\boldsymbol{\Phi}}\boldsymbol{\alpha}_t + \boldsymbol{\Sigma}_{\boldsymbol{Y}\tilde{\boldsymbol{Y}}}(\boldsymbol{\Sigma}_{\boldsymbol{Y}})^{-1}(\boldsymbol{Y}_t - \boldsymbol{\Phi}\boldsymbol{\alpha}_t), \boldsymbol{\Sigma}_{\tilde{\boldsymbol{Y}}} - \boldsymbol{\Sigma}_{\boldsymbol{Y}\tilde{\boldsymbol{Y}}}(\boldsymbol{\Sigma}_{\boldsymbol{Y}})^{-1}\boldsymbol{\Sigma}_{\boldsymbol{Y}\tilde{\boldsymbol{Y}}} \right). \tag{3.29}$$

However, due to the nature of DWT, interpolation under a new location vector is non feasible. This is because DWT does not provide us with a smooth interpolation and immediately a change of basis should be considered which then changes the estimation itself. Our methodology is able to perform temporal forecasting, however, if someone would like to conduct spatial interpolation, then a choice of a different basis, such as B-Splines should be considered and then sample from the conditional predictive distribution (3.29) based on this basis. Otherwise, if someone would prefer to work under wavelet basis decomposition, a lifting scheme could be considered by treating these locations as missing. This issue and a methodology is discussed in Heaton and Silverman (2008). Creating a scheme for spatial interpolation under this framework would be a valuable area for further research.

3.6 Other Considerations

In this section we address alternative approaches for the covariance inference of σ_ϵ^2 and σ_ν^2 which due to the limitation of time we were not able to investigate further. Furthermore, we discuss the choice of wavelet functions one would consider under the proposed methodology.

3.6.1 Inference on σ_ϵ^2 and σ_ν^2

In a more complex context the spatial covariance can be modeled in terms of the next few eigenfunctions of the empirical orthogonal function (EOF) decomposition to account for the spatial structure that is lost while conducting dimension reduction (Berliner et al., 2000). Furthermore, one could typically expand v_t in terms of an

additional basis set, e.g. $v_t = \Psi c_t$, and model the c_t coefficients hierarchically to absorb the residual spatial dependence as in Wikle et al. (2001), however, depending on the basis functions, one might have to put constraints on c_t. Finally, another approach could be the use of a slice sampler to improve the convergence problems that are arising if we jointly sample σ_ϵ and σ_v through Metropolis-Hastings.

3.6.2 Wavelet choice of basis

There are several types of wavelets that can be chosen. Firstly, if we were to employ a Discrete Wavelet Transform (DWT) for a point referenced spatio-temporal process then equally spaced time points and locations should be considered. Otherwise, one would have to explore the DWT under the lifting scheme (Sweldens (1996a) and Sweldens (1998)) where the theory is analogous to the initial one but biorthogonal Wavelets are used instead. Secondly, the choice of which type of basis should be used has received a lot of criticism, however, the same applies for other known basis such as Fourier or Bessel. Thus, in choosing the appropriate wavelet function there are numerous factors that should be taken into consideration (Farge, 1992).

Mostly for time series analysis, an aperiodic fluctuation in the process produces a different wavelet spectrum. Therefore, the wavelet basis should express the type of oscillations that exist in the time series. For instance, if the series produces sharp jumps or discontinuities, one would choose a step function such as the Harr wavelet, while for a less noisy or discontinuous one, a much smoother basis is more appropriate, i.e., the Daubechies family.

In the past, under a Reduced Dimension DSTM context, wavelets were only used in Wikle et al. (2001) where a Hierarchical Model approach with Gaussian Vector Autoregressive priors for the wavelets coefficients under a Gibbs sampler inferential framework was adopted. However, the challenge with using wavelet bases for Dynamic Spatio-temporal Models is adequately capturing the necessary interaction across scales while still preserving the dimension reduction. However, if a wavelet decomposition is used for the approximation of a process under a Reduced-dimension DSTM, one would typically expect parsimony. This believed parsimony can be then imposed through an appropriate prior belief in a Bayesian framework. Therefore, by considering a prior kernel which induces sparsity for the wavelet parameters, an adaptive framework for dimension reduction can be achieved. This novel Bayesian framework is presented in

the next section.

Choice of prior parameters in Spike and Slab In Nason (2010) it is noted that for the Adaptive Bayesian Thresholding with mixture priors, in order for the Bayesian inference to work well, the hyperparameters should be carefully chosen. Empirical Bayes methods can be used, however, in our concept there exists a high dimensional parameter space so it is inefficient. Through our simulations and applications we have observed that the tuning of v_0, ω_1 and ω_2 plays a big role into introducing sparsity to our model. Higher values of v_0 provides with non-parsimonious representations while the choice of ω_1 and ω_2 that results to high precision similarly provides us with high values of coefficients. Thus, in all simulations and applications we considered the hyperparameters to extract low values of coefficients.

3.7 Simulation study

Before applying our methodology to real data, in order to investigate its efficiency and especially whether the estimation approach is able to capture discontinuities well, simulation studies should be conducted. The importance of capturing the discontinuities and spatial propagation is of paramount importance since in real stochastic systems, the redistribution kernel $w_s(u)$ is very difficult to be guessed or approximated. Thus, we wish to test our method in three different settings; we achieve this through the design of these respective simulation studies:

- No discontinuity in weight function $w_s(u)$— we still get good approximations
- Discontinuity in weight function $w_s(u)$ and covariance inference— we wish to show that our method can adapt to discontinuities and we can estimate fairly well the covariance parameters
- No discontinuity in weight function $w_s(u)$ but more locations— we want to see how our methodology can estimate as the parameter space increases

In section 3.7.1 we introduce a simulation scheme of a Reduced-Dimension DSTM under wavelet basis decomposition. Furthermore, instead of simulating the matrix $\boldsymbol{B}$ through the Spike and Slab prior, a kernel is chosen for $w_s(u)$ and through that $\boldsymbol{B}$ is calculated through DWT. Additionally, as mentioned above, in sections 3.7.2 to 3.7.4 we conduct inference on the processes' parameters simulated under the simulation scheme in section

3.7.1. Finally, it has to be noted that due to limited computational resources we were not able to provide a more high dimensional example.

3.7.1 Creating a Gaussian DSTM under Wavelet decomposition

In this section we introduce a simulation scheme based on model (2.8) and (2.11) via Wavelet decomposition.

Essentially, since wavelets are strong detectors of discontinuities, a good approximation is expected for simulated data which for instance, one location point has a very high intensity while the rest of the location points have very low or no spatial intensity. This example provides us already with the idea of a discontinuous kernel $w_s(u)$. In this study, we will consider a grid of points with a barrier (or discontinuity) in between the grid which separates it with different intensities. Let us define the discontinuous kernel $w_s(u)$ for locations s and u in the 1-D interval $[a, c]$ as

$$w_s(u) = \begin{cases} f(\|s-u\|^2) & \text{if } s,u \in [a,b] \\ g(\|s-u\|^2) & \text{if } s,u \in (b,c] \\ 0 & \text{otherwise} \end{cases} \tag{3.30}$$

with f and g being two different chosen kernel functions. In the case where we do not want a discontinuous kernel (3.30) is simplified as

$$w_s(u) = \begin{cases} f(\|s-u\|^2) & \text{if } s,u \in [a,c] \\ 0 & \text{otherwise} \end{cases} \tag{3.31}$$

with f being the chosen kernel function for all locations.

Such examples of simulating such processes that have a spatially discontinuous kernel are described and compared with non-discontinuous simulated ones while pseudo code is given below. For the simulation scheme we define the Euclidean distance between two locations being $d = \|s-u\|^2$ and the parameter vectors of the kernel densities being $\boldsymbol{\theta}$.

1 Start by considering a number of equally spaced locations n $[a, c] \in D \subset \mathbb{R}$ and T time points, a Wavelet matrix $\boldsymbol{\Phi}_{n\times n}$ a covariance matrix $\boldsymbol{\Sigma}_\eta$, a spatial correlation

matrix $\boldsymbol{S}$ and variances $\sigma_\epsilon^2, \sigma_v^2$

- Define a barrier between the locations
- For the two parts according to the barrier, consider initial constants at $t = 1$ being c_1, c_2

2 Building the weight matrix

- For each of the locations calculate d between the locations s and u.

  If a barrier is desired then use (3.30):

  If both locations are in $[a, b]$, set $w_s(u) = f(d, \boldsymbol{\theta}_1)$

  If both locations are in $(b, c]$ then set $w_s(u) = g(d, \boldsymbol{\theta}_2)$

  else $w_s(u) = 0$.

  If a barrier is not desired then use (3.31):

  $w_s(u) = f(d, \boldsymbol{\theta})$
- Normalise the weights:

$$w_s^*(u) = \frac{w_s(u)}{\sum_{j=1}^n w_s(u)}$$

3 For $t = 1$

- Calculate the coefficient matrix $\boldsymbol{B} = \boldsymbol{w}^* \boldsymbol{\Phi}^{-1}$
- Calculate $\boldsymbol{\alpha}_1 = \boldsymbol{X}_{K1} \boldsymbol{\Phi}^{-1}$

4 For $t \geq 2$

- Calculate $\boldsymbol{\alpha}_t = \boldsymbol{\Phi}^\top \boldsymbol{B} \boldsymbol{\alpha}_{t-1} + \boldsymbol{\Phi}^\top \boldsymbol{\eta}_t$, $\boldsymbol{\eta}_t \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_\eta)$
- Calculate $\boldsymbol{X}_{Kt} = \boldsymbol{\alpha}_t \boldsymbol{\Phi}$
- Calculate $\boldsymbol{X}_t = \boldsymbol{X}_{Kt} + \boldsymbol{v}_t$, $\boldsymbol{v}_t \sim N(\boldsymbol{0}, \sigma_v^2 \boldsymbol{S})$
- Calculate $\boldsymbol{Z}_t = \boldsymbol{Y}_t + \boldsymbol{\epsilon}_t$, $\boldsymbol{\epsilon}_t \sim N(\boldsymbol{0}, \sigma_\epsilon^2 \boldsymbol{I})$

A few comments are in order. The choice of the type of the $\boldsymbol{w}$ matrix, which is the redistribution kernel of a location to another at time point $t-1$ to the time point t, will affect the spreading of the intensity. For instance, if all the locations have Gaussian kernels with location parameter θ_1 and scale parameter θ_2, then the process will propagate and diffuse in the same way in both barrier sides. On the other hand, if for instance there are two different Gaussian kernels in which the scale parameter θ_2 is

greater for one barrier, then the dilation of that kernel is bigger and thus the process becomes more diffusive on that side. In the case where the kernels are not of the same distribution, then according to the location and scale parameters, both sides will have a different spatial propagation and diffusion over time.

As an example of how the process's propagation can differ according to different expansion and noise, under the same seed we simulated processes with $n = 32$ locations in the 1-D space [0, 5] with a total of $T = 256$ time points. Additionally, we considered the same number of locations and starting values for $t = 0$ being 1 and 12 for the locations that lie inside [0, 2.5] and (2.5, 5] respectively and the same diagonal covariance matrix $\Sigma_\eta = 10 * I$. The measurement error variance was chosen to be $\sigma_\epsilon^2 = 8$, the spatial variation error variance $\sigma_\nu^2 = 1$ and an exponential spatial correlation function with scale parameter $\lambda = 3$ was considered. All the examples have been produced under a Daubechies level 10 wavelet decomposition. In the upper plots in Figure 3.3 the weighting function that we created depends on one barrier in between the spatial 1-D space [0,5], i.e.,

$$w_s(u) = \begin{cases} N(\|s-u\|^2 \,|0, 0.5) & \text{if } s, u \in [0, 2.5] \\ \exp(\|s-u\|^2 \,|1) & \text{if } s, u \in (2.5, 5] \\ 0 & \text{otherwise} \end{cases} \tag{3.32}$$

This means that in each time point the intensity of the process on location points which reside in $[0, 2.5]$ will have a zero contribution for the locations that lie in $(2.5, 5]$. Additionally, $[0, 2.5]$ diffuses with a Gaussian contribution with a scale of 0.5 while the points that lie in (2.5, 5] diffuse exponentially with rate parameter 1. This example of discontinuity can be perfectly incorporated by the wavelets approximation.

Furthermore, in the bottom plots in Figure 3.3 , the weight function is considered again to be discontinuous but now under the same kernel but only with different parameters, i.e.,

$$w_s(u) = \begin{cases} N(\|s-u\|^2 \,|0, 0.5) & \text{if } s, u \in [0, 2.5] \\ N(\|s-u\|^2 \,|0.1, 1) & \text{if } s, u \in (2.5, 5] \\ 0 & \text{otherwise} \end{cases} \tag{3.33}$$

It is notable that the diffusion in both sides of the barrier expands differently which provides us with a discontinuous example to be approximated by wavelet basis decom-

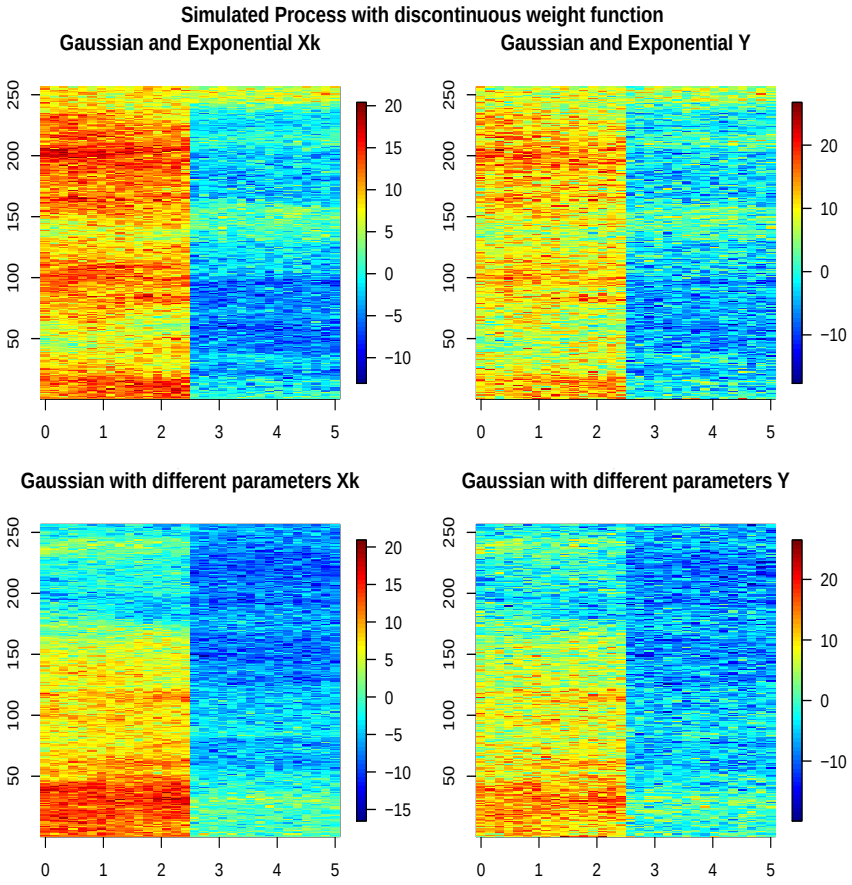

Figure 3.1: Image plots of the noisy process Y (on the right) and the underlying process X_K (on the left) for $T = 256$, $n = 32$ where $\boldsymbol{s} \in [0, 5]$ under a signal to noise ratio $\sigma_\epsilon/\sigma_v = 8$. On the upper panel a discontinuous $w_s(u)$ is considered with f being a Gaussian kernel with mean and variance 0 and 0.5 respectively and g being an exponential kernel with rate parameter 1. On the bottom panel a discontinuous $w_s(u)$ is considered with f being a Gaussian kernel with mean and variance 0 and 0.5 respectively and g being a Gaussian kernel with mean and variance 0.1 and 1 respectively.

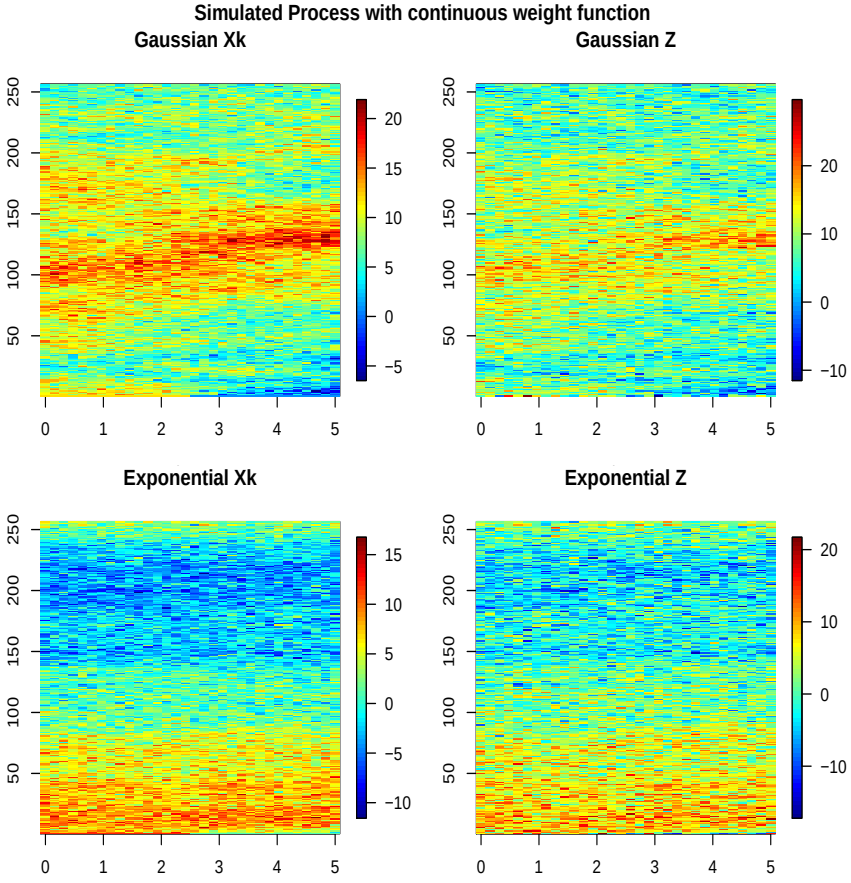

Figure 3.2: Image plots of the noisy process Y (on the right) and the underlying process X_K (on the left) for $T = 256$, $n = 32$ where $\boldsymbol{s} \in [0, 5]$ under a signal to noise ratio $\sigma_\epsilon/\sigma_v = 8$. On the upper panel a Gaussian kernel with mean and variance 0 and 0.5 respectively is considered. On the bottom panel an exponential kernel with rate parameter 0.5 is considered.

position. Finally, in Figure 3.4 we provide an example of continuous kernel functions for the spatial diffusion. In these cases, we want to approximate the weight function but also the underlying process X_K.

3.7.2 No discontinuity in weight function $w_s(u)$

In this analysis, we ran our model under known covariance structure under wavelet basis decomposition with a wavelet Daubechies bases of level 10. The simulation was conducted under a signal to noise ratio $\sigma_\epsilon/\sigma_v = 3$, with $\sigma_\epsilon^2 = 3$ and $\sigma_v^2 = 1$ and a diagonal temporal variation covariance matrix $\boldsymbol{\Sigma}_\eta = 5 * \boldsymbol{I}_{n\times n}$. Furthermore, a Gaussian redistribution kernel with mean and variance 0 and 0.5 respectively was used, i.e.,

$$w_s(u) = \begin{cases} \mathrm{N}((\|s-u\|^2)|0, 0.5) & \text{if } s, u \in [0,5] \\ 0 & \text{otherwise} \end{cases} \tag{3.34}$$

while an exponential covariance function with a scaling parameter $\theta = 3$ was considered, i.e.,

$$S(h;3) = \begin{cases} (1-\exp(-|h/3|)) & h > 0 \\ 0 & \text{otherwise} \end{cases} \tag{3.35}$$

For the inferential part, we considered the Spike and Slab hyperparameters $v_0 = 0.05$ and $\omega_1 = 2$, $\omega_2 = 20$ for the point mass and variance components respectively which gives as an informative priors for $\boldsymbol{B}$ close to zero as we expect parsimony. Finally, the Kalman Filter recursions for $\boldsymbol{\alpha}_t$ for $t = 0$ a non informative prior was considered with mean and covariance matrix $m_{0|0} = \mathbf{0}$ and $\boldsymbol{P}_{0|0} = 10^3 \boldsymbol{I}_{n\times n}$ respectively.

Comparing the posterior mode of the underlying process and the simulated (real) one in Figure 3.5 it can observed that even if it is a point estimate, our methodological model performs very well. The high intensities in the latter time points are captured nicely, while the low ones seem underestimated, however, trendwise we have a successful prediction of the underlying process $\boldsymbol{X}_k$ (Figure 3.6). Furthermore, the spatial wavelet coefficients $\boldsymbol{B}$ under our Spike and Slab hierarchy were fairly estimated whereas a tendency for overestimation has been observed (Figure 3.7). Notably, even if the elements of the matrix $\boldsymbol{B}$ are not greatly estimated, the reconstruction of the weight function (Figure 3.8) for most of the elements is precise, with an average associate error $L-2$ norm equal to 0.4.

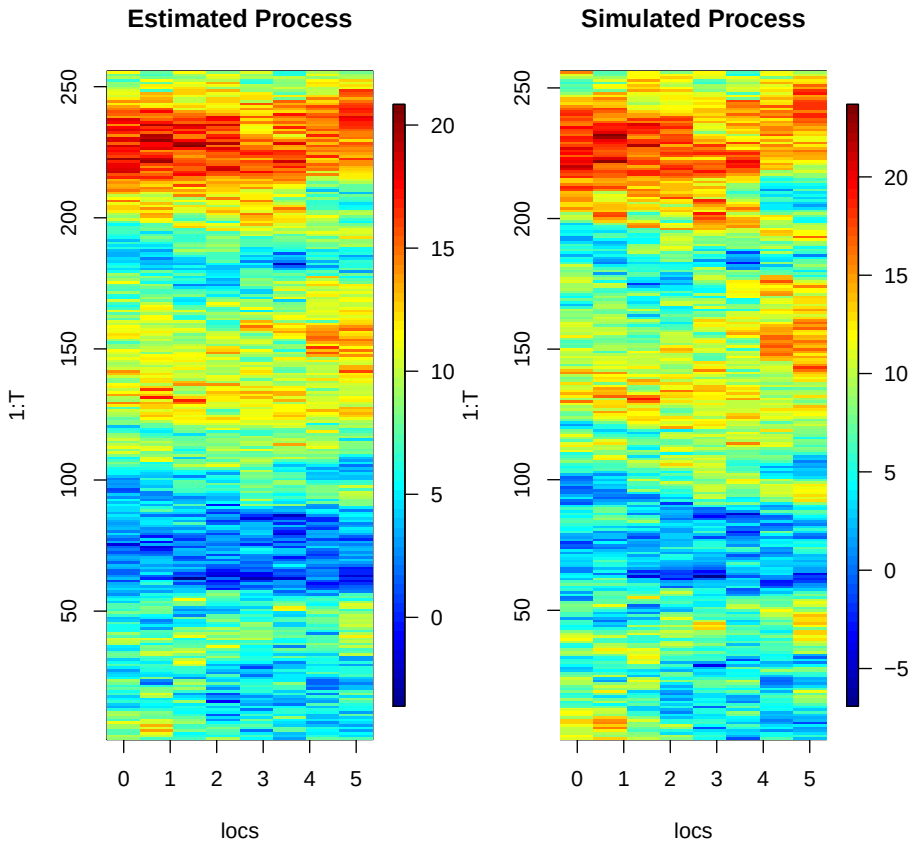

Figure 3.3: Image plots of the estimated process (on the left) and the simulated underlying process X_K (on the right) for $T = 256$, $n = 8$, $s \in [0, 5]$ under a Daubechies wavelet of smoothness level 10 and a Gaussian weight function $w_s(u)$. The processes are approximated through IDWT under the smoothed estimates of $\boldsymbol{\alpha}_t$ for $N = 10^5$ Gibbs iterations with a burn in period of $i = 50000$.

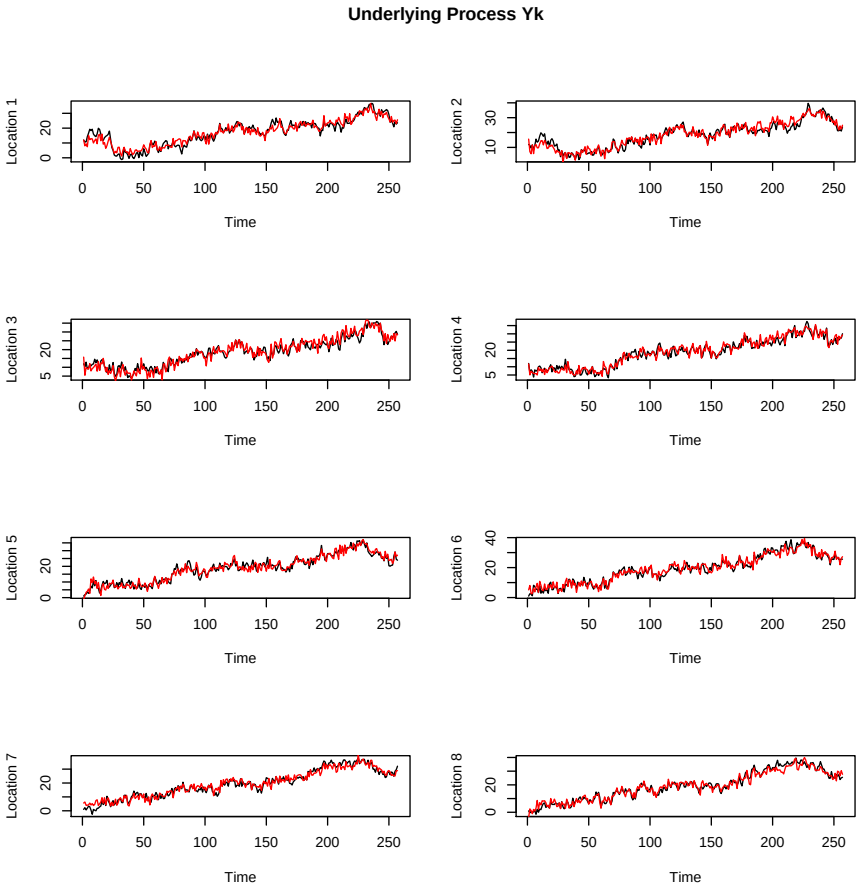

Figure 3.4: Time series plots of the underlying process for the locations (black) and the estimated one (red) for $T = 256$, $n = 8$, $s \in [0, 5]$ and $N = 10^5$ Gibbs iterations with a burn in period of $i = 50000$.

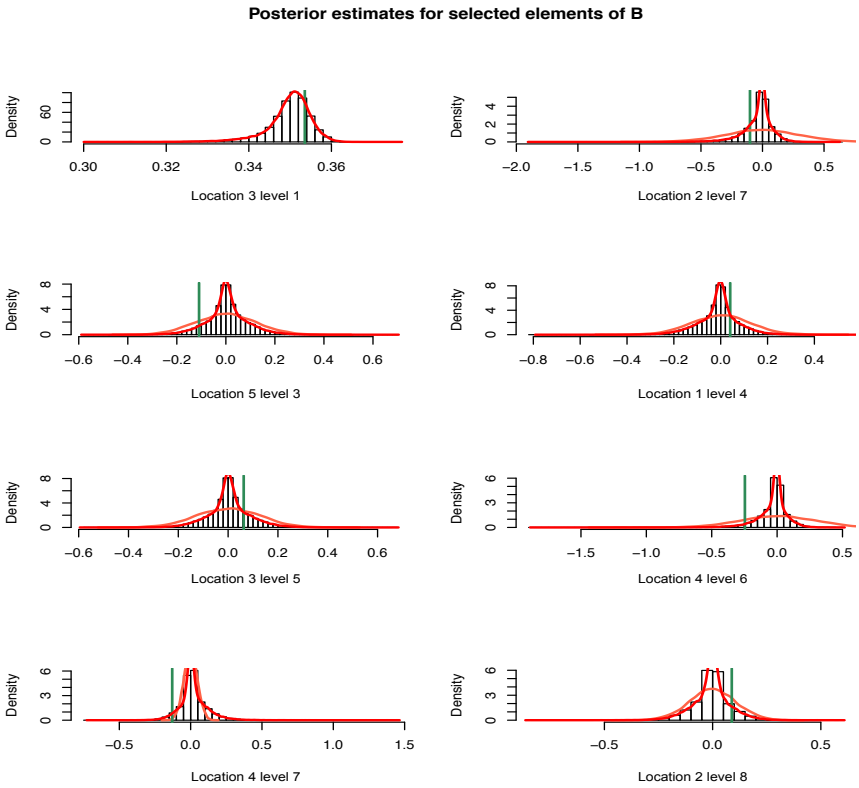

Figure 3.5: Histograms of the posteriors for selected elements of B. The red line indicates the kernel density estimation of the posterior estimates, the orange line indicates the prior belief while the vertical green line specifies the actual value of $\beta_{k,s}$. The hyperparameter values were set to $v_0 = 0.05$, $\omega_1 = 2$ and $\omega_2 = 20$. The inference was conducted under $N = 10^5$ Gibbs iterations with a burn in period of $i = 50000$.

Furthermore, we have observed that there is a tendency of overestimating the negative elements of $\boldsymbol{B}$, while for the almost zero or zero ones, the posterior estimates are precise. Thus, there is an indication that the spike and slab hierarchy is a very sensible tool for a weight function that tends to show large contributions between many spatial

locations. However, it seems to be the appropriate approach when there exists a sparse weight function.

The argument stated above can be justified from the posterior estimates for the hypervariances γ_k (Figure 3.9). In all levels the bimodal nature of the distribution creates a tendency of deriving elements of β_k very close to zero, in which case under a high dimensional dataset we would expect a sparse representation of both the matrix B and the weight function $\boldsymbol{w}$. Moreover, for the non-zero elements the hypervariance is very small to sample higher values of $|\beta_k|$. A panacea for this would be to be able to 'tune' the hyperparameter v_0, however, if treat that parameter as very small or significantly larger, then an underestimation or overestimation phenomenon will be again observed in these respective cases.

In conclusion, all the prameters under the MCMC scheme have converged after $N = 100000$ iterations and a burn-in period of $i = 50000$ decided under convergence diagnostics, such as autocorrelation plots and trace plots. The efficiency of the algorithm for this amount of locations and time points did seem fast, however, the more the locations are increased, the more the computational complexity and thus the iterations for the parameters to converge.

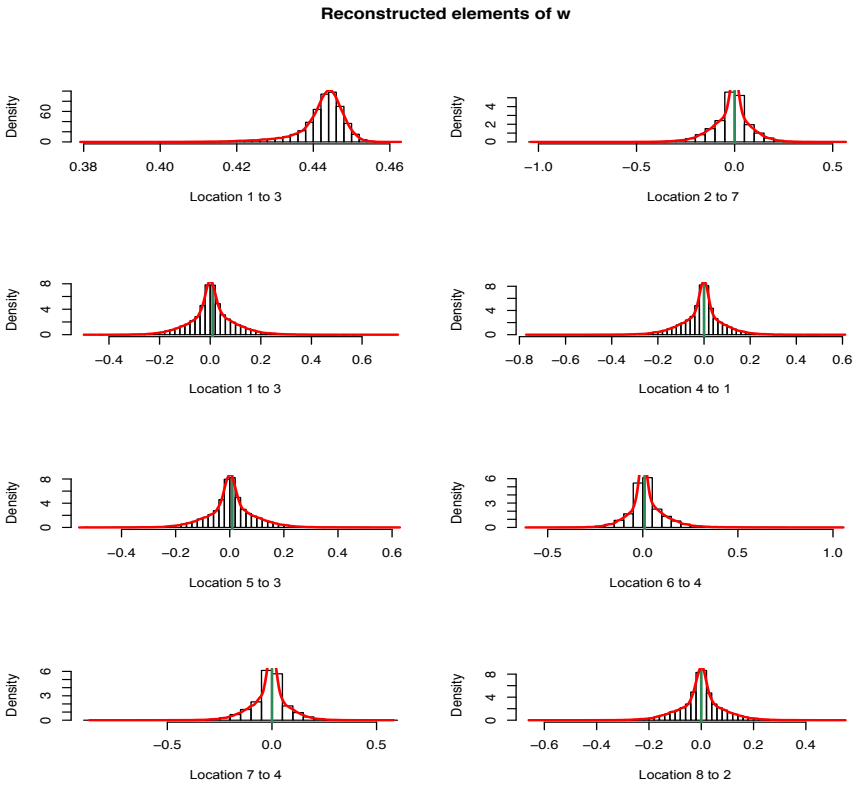

Figure 3.6: Histograms of the reconstructed elements of the redistributional kernel $w_s(u)$ produced from IDWT by the posterior estimates of $\boldsymbol{B}$. The red line indicates the empirical distribution of each element. The green vertical line indicates the actual value of the weight function. The associate distributions are produced under $N = 10^5$ Gibbs iterations with a burn in period of $i = 50000$.

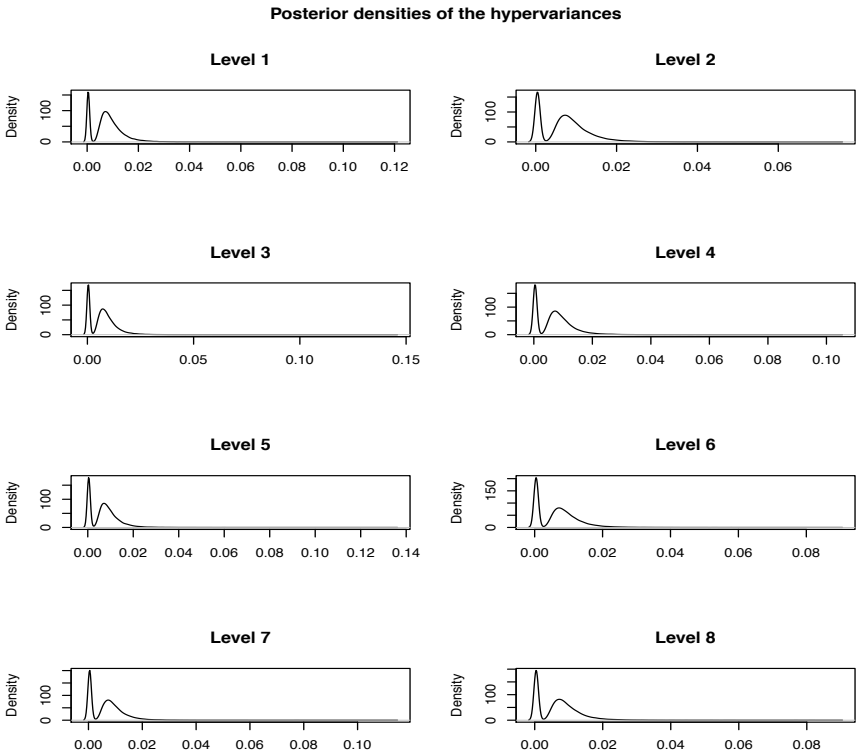

Figure 3.7: Posterior density estimates for the hypervariances γ_k for each level of the wavelet coefficient $\boldsymbol{\beta}_k$. The bimodal distribution is due to the point mass prior and the hyperparameter values that were set, i.e., $v_0 = 0.05$, $\omega_1 = 2$ and $\omega_2 = 20$. The inference was conducted under $N = 10^5$ Gibbs iterations with a burn in period of $i = 50000$.

3.7.3 Discontinuity in weight function $w_s(u)$ and covariance inference

This simulation was conducted under the same covariance parameters and spatial correlation function in (3.35). However, we have produced a process based on a discontinuous kernel. Specifically, a Gaussian redistribution kernel with mean and variance $\theta_{11} = 0$

and $\theta_{12} = 0.5$ respectively was used for the locations that lie in the area $[0, 2.5)$ while a Laplace redistribution kernel was used with $\theta_{12} = 0$ and $\theta_{12} = 1$ in the area $[2.5, 5]$, i.e.,

$$w_s(u) = \begin{cases} \mathrm{N}(\|s-u\|^2 \,|0, 0.5) & \text{if } s, u \in [a, b] \\ \mathrm{Laplace}(\|s-u\|^2 \,|0, 1) & \text{if } s, u \in (b, c] \\ 0 & \text{otherwise} \end{cases} \tag{3.36}$$

For the inferential part, the same hyperparameters were considered, while for the covariance inference we consider σ_ν to be the parameter to estimated under an inverse gamma prior with $\delta_0 = 2500$ and $\xi_0 = 0.01$ being the shape and scale parameters respectively which provides us with informative prior of small variance. Analogously, for the covariance parameter σ_η an inverse gamma prior with $\psi_1 = 13000$ and $\psi_2 = 100$ being the shape and scale parameters respectively was considered which again provides us with an informative prior of a small variance.

Essentially, we expected that since wavelets are strong detectors of discontinuites we would be able to track the very high and very low intensity areas. As seen in Figure 3.10, we managed to reconstruct the underlying process $\boldsymbol{X}_k$ and successfully captured the discontinuity level of these intensities pretty well. Between the 20th and 100th time point however the very low intensity locations are overestimated while gradually the estimation gets better. This has to do with the FFBS as well since the more we approach the final point, the better the process will be approximated.

Similarly to the previous scenario, the posterior modes for the different elements of $\boldsymbol{B}$ tend to be very good when the coefficient is zero while for the non-zero ones for a small amount of cases we derive over or under estimations. This has to do with Spike and Slab being sensitive when there is a high signal-to-noise ratio. The posterior mode of the covariance parameters are 4.7 and 5.29 for σ_ϵ^2 and σ_η^2 respectively. There is an indication that the variances are hyperparameter sensitive since we have always a large summation in the scale parameter in the conditional posterior which could give us very high values for the variances and thus the system would stop running due to singularity. It can be observed through the posterior distributions (Figure 3.13) that we placed highly informative priors, which for instance for σ_η was not a good choice and we still have posterior modes which are not far from the truth. Notably though, the estimation of σ_ν is difficult to come through since the measurement error variance σ_ϵ is indirectly included in the estimation. We ran the algorithm for a number of iterations

$N = 200000$ and under a burn-in period of $i = 100000$. It seems that we needed more iterations for the variances to converge, however due to computational power, the user has to resort to more efficient computer resources.

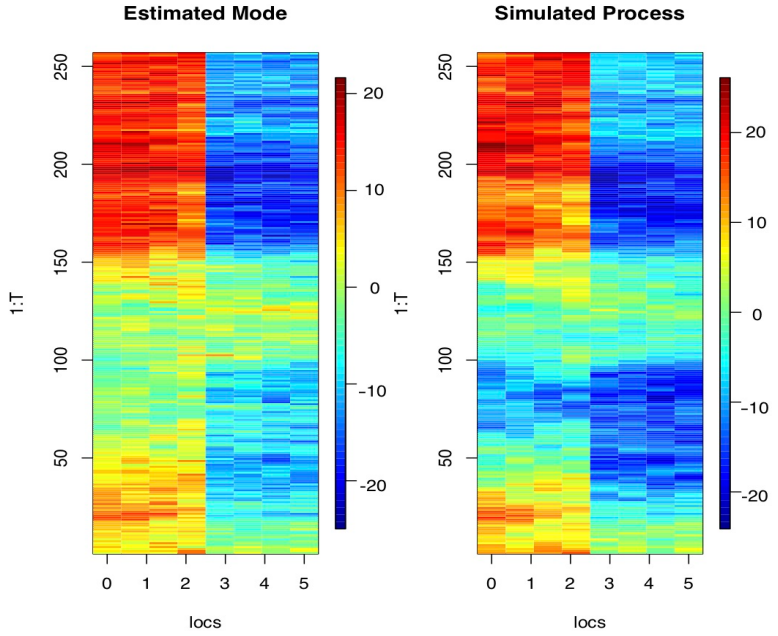

Figure 3.8: Image plots of the estimated process (on the left) and the simulated underlying process X_K (on the right) for $T = 256$, $n = 8$, $s \in [0, 5]$ under a Daubechies wavelet of smoothness level 10 and a discontinuous weight function $w_s(u)$. The processes are approximated through IDWT under the smoothed estimates of $\boldsymbol{\alpha}_t$ for $N = 2 * 10^5$ Gibbs iterations with a burn in period of $i = 10^6$.

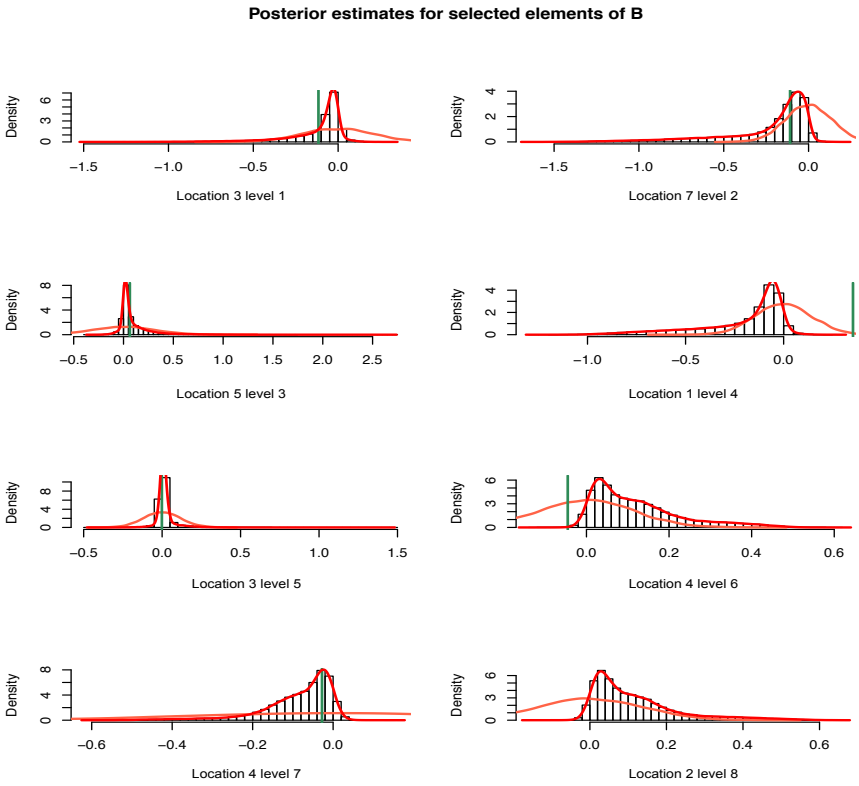

Figure 3.9: Histograms of the posteriors for selected elements of B. The red line indicates the empirical distributions of the estimates, the orange line indicates the prior belief while the vertical green line specifies the actual value of $\beta_{k,s}$. The hyperparameter values were set to $v_0 = 0.05$, $\omega_1 = 2$ and $\omega_2 = 20$. The inference was conducted under $N = 2 * 10^5$ Gibbs iterations with a burn in period of $i = 10^5$.

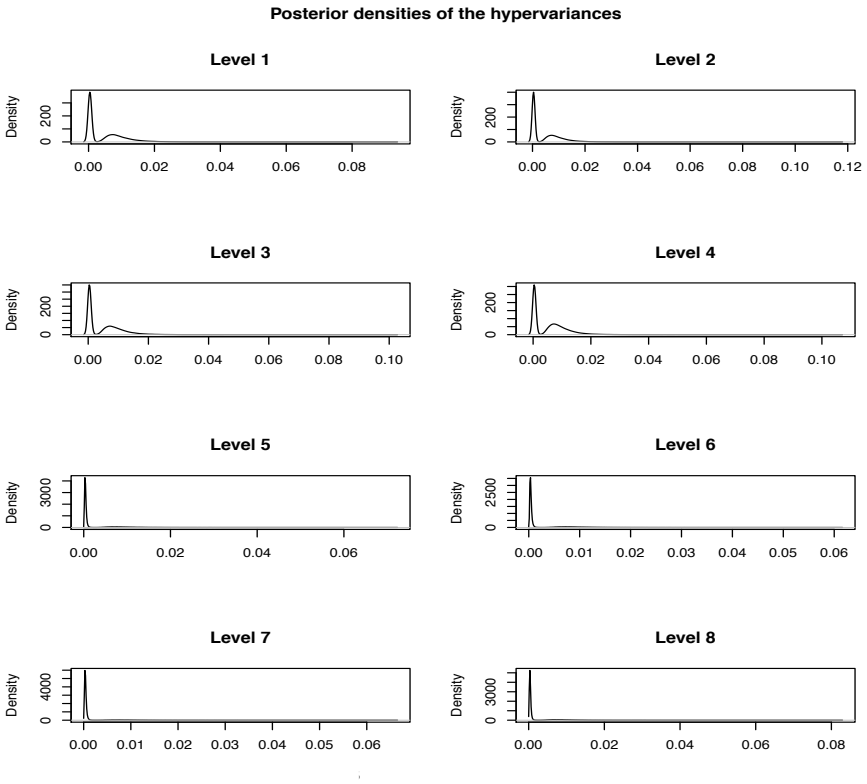

Figure 3.10: Posterior density estimates for the hypervariances γ_k for each level of the wavelet coefficient $\boldsymbol{\beta}_k$. The bimodal distribution is due to the point mass prior and the hyperparameter values that were set, i.e., $v_0 = 0.05$, $\omega_1 = 2$ and $\omega_2 = 20$. The inference was conducted under $N = 2 * 10^5$ Gibbs iterations with a burn in period of $i = 10^5$.

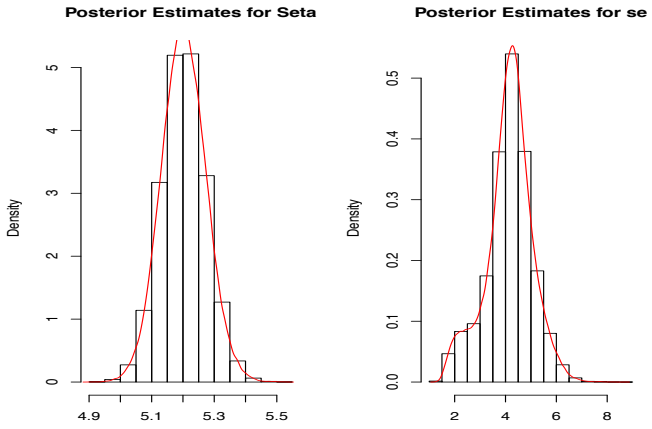

Figure 3.11: Posterior density estimates for the covariance parameters σ_ν and σ_η. Red lines indicate the empirical distribution of the estimates. The prior that was set for the spatial covariance σ_ν is an inverse gamma with $\delta_0 = 2500$ and $\xi_0 = 0.01$ being the shape and scale parameters respectively while for the covariance parameter σ_η an inverse gamma prior with $\psi_1 = 13000$ and $\psi_2 = 100$ being the shape and scale parameters respectively was considered. The inference was conducted under $N = 2 * 10^5$ Gibbs iterations with a burn in period of $i = 10^5$.

3.7.4 No discontinuity in weight function $w_s(u)$ but more locations

After assessing our model's capability with known and unknown covariances for various weight functions under a small amount of locations, we conducted a simulation study for an amount of $n = 32$ locations as well. Notably, due to high computational complexity we noticed that compared to eight locations, it was far more expensive to run and we needed more iterations for the parameters to converge.

Noticeably, the trend of the underlying process was captured pretty well with only a few locations and time points showing over and under estimations (Figure 3.14). The high intensity pattern between the 150th time point and the 200th is estimated pretty well for the majority of the locations, while for the latter time points, most of the locations are estimated pretty well. A slight capture of the pattern for the first time

points can be noticed while for the very low ones we have a slight overestimation.

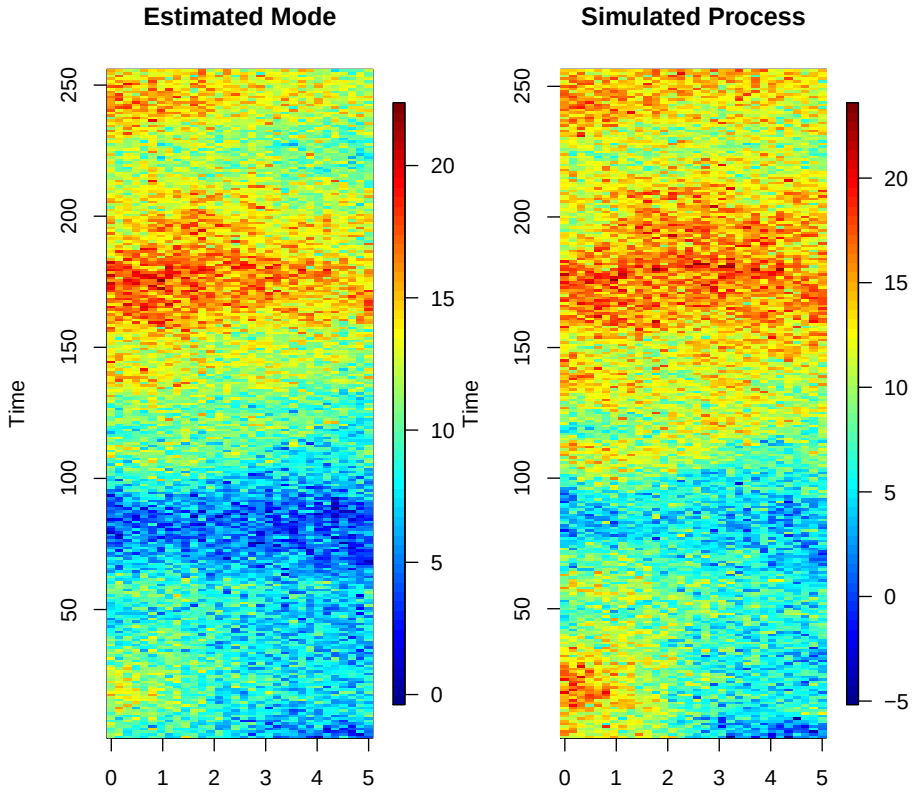

Figure 3.12: Image plots of the estimated process (on the left) and the simulated underlying process X_K (on the right) for $T = 256$, $n = 32$, $s \in [0, 5]$ under a Daubechies wavelet of smoothness level 10 and a Gaussian kernel weight function $w_s(u)$. The processes are approximated through IDWT under the smoothed estimates of $\boldsymbol{\alpha}_t$ for $N = 4 * 10^5$ Gibbs iterations with a burn in period of $i = 3 * 10^5$.

Thus, we can still get a very satisfying estimation of the underlying process by considering that we did not have the computational power for it to converge. Furthermore, since the parameters for the spatial wavelet coefficients $\boldsymbol{B}$ are 1024 parameters in total, we have observed again that the close to zero and zero ones are perfectly estimated while for the positive ones we derive slightly overestimated posterior modes, however, for the negative ones we get an underestimation. This was observed for the eight locations as well.

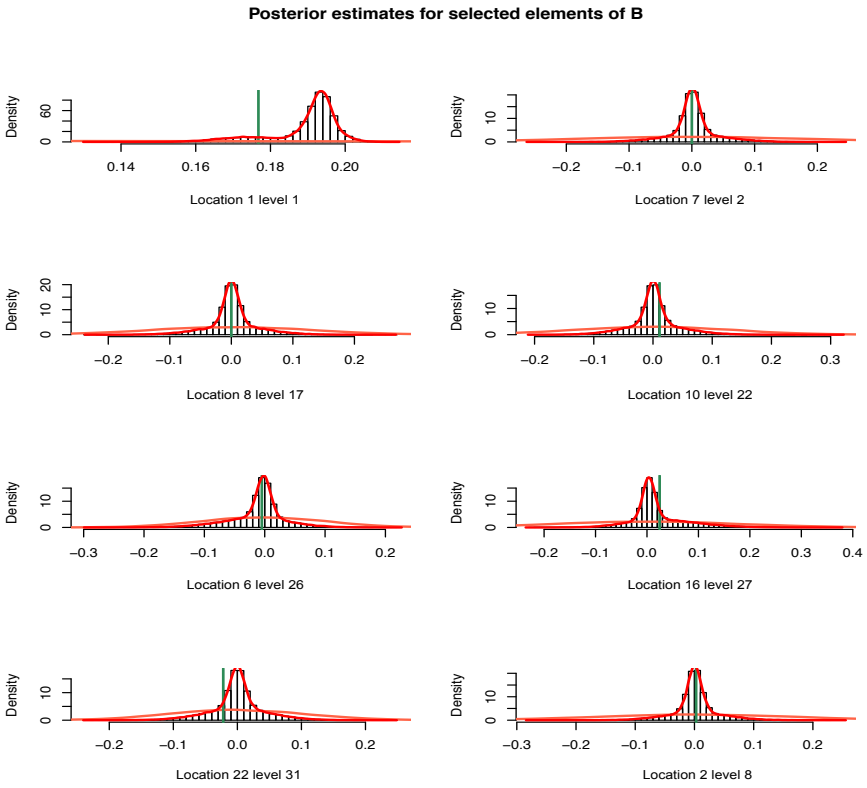

Figure 3.13: Histograms of the posteriors for selected elements of $\boldsymbol{B}$. The red line indicates the empirical distributions of the estimates, the orange line indicates the prior belief while the vertical green line specifies the actual value of $\beta_{k,s}$. The hyperparameter values were set to $v_0 = 0.05$, $\omega_1 = 2$ and $\omega_2 = 20$. The inference was conducted under $N = 4 * 10^5$ Gibbs iterations with a burn in period of $i = 3 * 10^5$.

3.8 Application to Pollution Data

In this section we will consider air pollution data, consisting of a response spatio-temporal point referenced process variable $Y(s,t)$ with values of the pollutant Nitro-

gen Oxide (NO) in milligrams per square meter. This data set is collected on a daily frequency over a period of 256 days, from 20th April 2008 until 31th December 2008, picked in order to avoid seasonality effects. The data, are obtained by eight monitoring stations in the city of Athens, which are shown in Figure 3.16. Previous analysis on this dataset has been conducted in order to track when higher levels can cause risk in public health (Bersimis et al., 2018). Furthermore, analysis of air pollution datasets under spatio-temporal models have been thoroughly conducted in the literature such in Wikle et al. (1998), Riccio et al. (2006) or Schmidt and Gelfand (2003) and further analysis has been conducted for modelling Ozone levels (Huerta et al. (2004) Sahu et al. (2007), Sahu et al. (2009), Sahu and Bakar (2012)).

As these data are measurements by older instruments, a large amount of signal to noise ratio is expected. Therefore, what we would like to infer is NO across time and the eight spatial locations around the city of Athens. Furthermore, they were given with an amount of missing values, which were imputed via a Moving Average (MA) prior to the analysis. Therefore, there are several interesting questions we may ask but the most interesting one is whether we can find any causal relationships between the locations themselves through the estimation of the weight function.

3.8.1 Lifting scheme for multidimensional spatially irregural data

The weather station's locations are defined in a two-dimensional grid and are irregularly spaced. Thus, a standard use of orthogonal wavelet transform as in 1-D is not plausible for this application. In cases like this, second generation wavelet techniques called 'lifting' for two dimensions have been developed which can handle multidimensional irregularly spaced data that commonly arise in statistics (Jansen et al., 2009). These techniques specifically for these kind of data have been created in Jansen et al. (2009) and are developed and built on Jansen et al. (2001). For a quick introduction to the lifting scheme the reader can see Sweldens (1996b).

In the 2-D dimension many concepts of a neighbourhood structure can be considered. Jansen et al. (2009) suggested the usage of Voronoi polygons to define the neighbourhood structure, which are employed by a lifting scheme. The basic idea is to construct, at each stage, a triangulation of the data locations. The neighbours of any location are then the locations that are entered by edges within the triangulation. As soon as a detail coefficient corresponding to a particular location has been found, the triangulation

is appropriately modified to exclude that location.

However, it is important for the variance structure of the new lifted wavelet coefficients to be carefully analysed. Jansen et al. (2009) use a novel Bayesian wavelet shrinkage technique by considering the artificial levels. Thus, under this scope, by extracting the new lifted coefficients with the use of aforementioned lifting scheme under Voronoi polygons for the neighbourhood structure and by extracting as well their scaling variances, new artificial groups or levels can be decided. One way to decide the artificial levels of grouping is an adaptation of the dyadic structure of the standard discrete wavelet transform. Specifically, the coefficients can be split into levels in some arbitrary way, and one possibility is simply to impose an artificial dyadic split, with the highest level containing the half of the coefficients with finest scale, and subsequently lower levels successively a quarter, an eighth, and so on, of the total number of coefficients in the order that is defined by the lifting scheme.

In this application, we employed lifting scheme under Voronoi polygons for the neighbourhood structure and extracted the scales of the coefficients. Then, a spike and slab prior for each level was imposed. As eight locations is not a big number to be considered for grouping artificial levels, in this case we model them without any grouping. However, a generalisation for a dataset with more locations in terms of the posterior distributions that consider the grouping for artificial levels of coefficient matrix $\boldsymbol{B}$ can be analogously derived.

3.8.2 Analysis and Results

In this analysis we are considering our Bayesian framework of the Dimension Reduced DSTM under the lifting scheme with a Voronoi neighbourhood structure for the decomposition of NO with a known signal-to-noise ratio ($\sigma_\epsilon/\sigma_\nu = 1.5$) with the hierarchy in Table 3.1. under the pseudo-code in Table 3.2 under the Spike and Slab prior with $v_0 = 0.02$. Moreover, we noticed that some locations are more volatile than others, therefore a diagonal structure with difference elements on $\boldsymbol{\Sigma}_\eta$, i.e, $\boldsymbol{\Sigma}_\eta = \boldsymbol{\sigma}_\eta^2 \boldsymbol{I}$ with $\boldsymbol{\sigma}_\eta^2 = (\sigma_{\eta 1}, \ldots, \sigma_{\eta 8})$ each one with an inverse gamma prior with shape and scale parameters 10^3 and 0.01 respectively. Finally, an exponential spatial correlation function was considered with scaling parameter $\theta = 3$.

Several findings are in order. Wikle and Cressie (1999) consider a detrended pro-

cess $Y(s,t)$ under the Dimension Reduced DSTM. However, in order to check the performance of the proposed methodology, no detrending methods were applied. Interestingly, even if the noisy process $Y(s,t)$ follows specific trends through time, the estimated underlying process $X_K(s,t)$ of nitrogen oxide (NO) captures these trends pretty well, as seen in Figure 3.17 and 3.18. This provides evidence of the adaptivity of the proposed methodology.

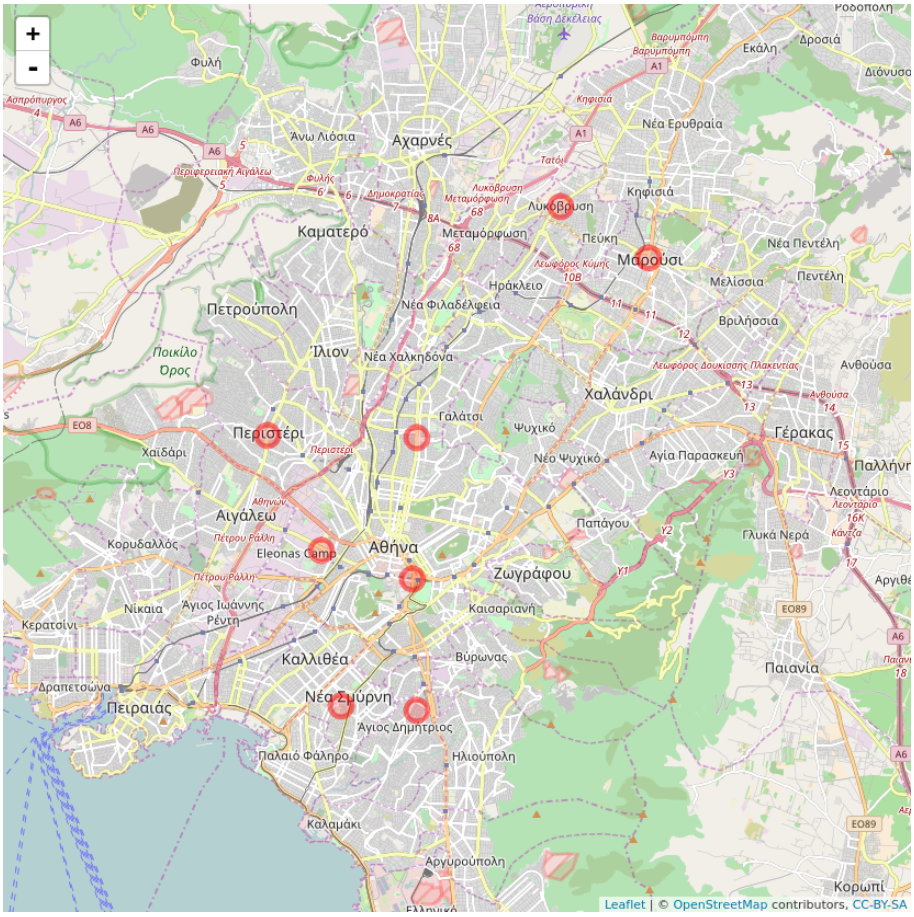

Figure 3.14: Map of Athens. The eight weather stations are represented with red circles. The weather stations are located at central and suburban locations in Athens.

Additionally, the model itself produces trended underlying process values where the moving average imputed values are fairly estimated under a trend. We can observe a pattern for most of the locations. For the final time points, which represent the winter months, there is increased volatility in terms of NO emissions. This can be justified as during summer months most of the citizens in Athens take their vacation time off

and thus there is way less traffic on the streets. Furthermore, during winter months, as traffic is high we would expect higher emissions of NO. Finally, only for the last location, Peristeri, we can observe exactly the opposite effect. Peristeri is a more rural area which is near the west coast of Athens. Therefore, during summer months most people travel around that area as it is a national network in order to reach the coast in Athens peninsula and but as well most of the rest mainland of Greece (southern towards Peloponnese and northern and east towards Thiva and Chalkis respectively).

Secondly, considering the reconstructed estimates of $w_s(u)$ (Figure 3.18) the most important finding is the causal relationships of the weight function $w_s(u)$ between locations. Specifically, for the central area of Athens we observe that the weight function leans towards the negative contribution to the suburban Likovrisi area. This is reasonable as Likovrisi is a residential only area with a low traffic activity while the central area of Athens hs a high activity of drivers that come through all suburban areas to work and/or study.

Furthermore, it is known that Athens is an area where most of the air pollution, both by cars and industry is gathered around the city centre while the suburban areas are greener, sparse inhabitant and have only a few residential cars. Intuitively we would expect negative relationships in the pollutants between central weather stations and suburban ones. Moreover, as the measurements are daily, the interpretation of the reconstructed weight function $w_s(u)$ describes the contribution in the amount of NO levels at one location at time $t-1$ to another location at time t which means we would expect a negative causal relationship at the NO levels between a central weather station and a suburban one. Additionally, Likovrisi is a neighbouring area of Marousi which is a combination of industrial and residential neighbourhoods. The weight function of Marousi to Likovrisi suggests that if there is a high level of the pollutant NO on the first day, we would expect a high level of NO in Likovrisi as well. It is reasonable as we would expect the pollution and the traffic activity in neighbouring areas to be similar. Finally, Likovrisi and Pireus are very distant neighbours with a complete different structure. Pireus is both an industrial, residential and trading area. In Pireus lies the Athens port for trading and commuting to the islands but there are plenty of offices and factories as well. It is reasonable to have a non-causal relationship for these two areas due to their large distance and the different social and geological structure they own. On the other side, we would expect Pireus and Geoponiki to interact with each other as they are both big areas, reasonably close to each other with two central highways

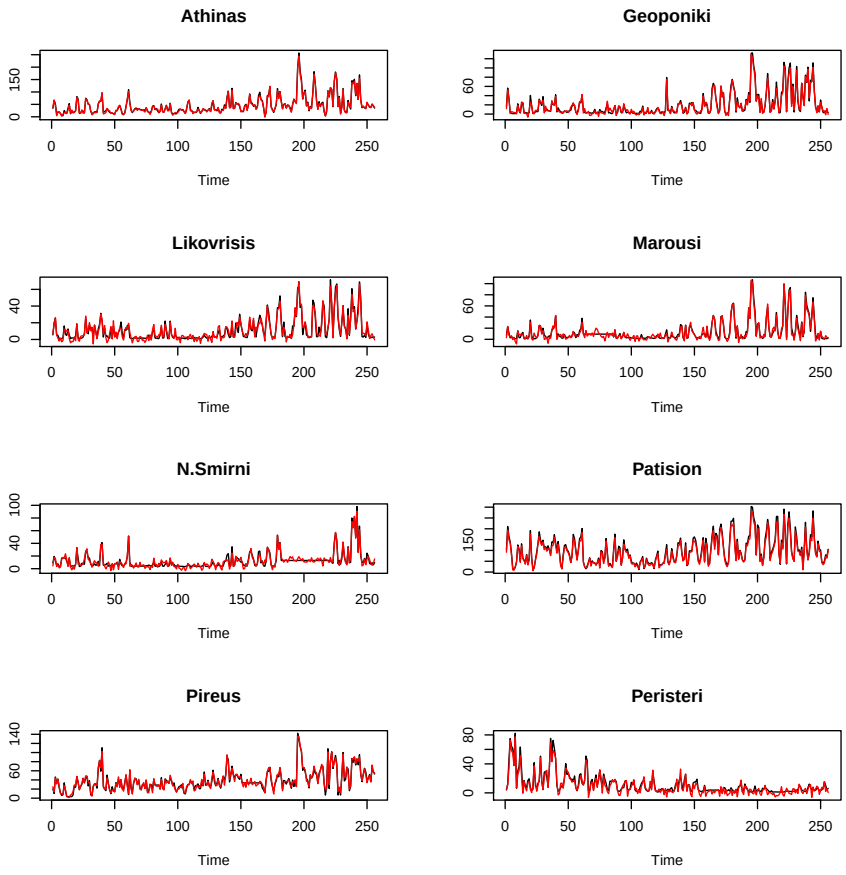

Figure 3.15: Noisy Nitrogen Oxide (NO) measurements (black) vs approximated underlying process (red) for eight weather stations consisted of central and suburban locations in Athens. The processes are approximated through IDWT under a Haar wavelet through the smoothed estimates of $\boldsymbol{\alpha}_t$ for $N = 2 * 10^6$ Gibbs iterations with a burn in period of $i = 10^6$.

connecting them. Noticeably, our model estimates that there is no causality between them. Furthermore, Geoponiki is very close to the centre of Athens, being connected with highways and dense populated areas in between. It is suggested by our model that there exists a positive causal spatial contribution from Geoponiki to the central area of Athens.

In addition, Peristeri shows temporally a different pattern of levels of emission compared to the rest of the locations. That sensibly contributes into understanding the negative causal relationship created in terms to Athens. The weight function indicates that if one day we have high emissions in Peristeri, we would expect in Athens to have way lower the day after. This is intuitively a prediction of their contrast in emission levels between the summer and winter months. Moreover, Peristeri receives no causal contribution from Pattision, which is a highly dense populated area with universities and offices as well. Pattision, throughout all time points shows the highest emission levels, only with a small decrease during September. Even if Peristeri has a contrasting pattern with the rest of the locations, it seems that it does not affect the emissions across that area. One reason for that is that their distance is far with a bad road connection between them. Finally, despite N. Smirni and Marousi appearing to be similar residential areas (they are far from each other but they share a similar distance to the centre), there seems to be a negative causality between them. If we observe the emissions of NO throughout all the time points for both locations, they have a similar pattern, except of the autumn months where the data set includes numerous missing values, which where imputed via moving average for N. Smirni with very low levels, and very high ones for Marousi. We suspect that this affected the estimation of the weighting function and gave us a contrast between these two locations.

The posterior variance densities for all locations are shown in Figure 3.19. Specifically, while we gave highly informative priors, adaptively the data gave us different posterior estimates. Moreover, the posterior modes are very close to the empirical variances for all locations which is reasonable. Finally, the spatial variance σ_v^2 gave us a posterior mode of 6.3. The level of spatial variance is low indicating us that all locations, i.e., both rural and urban areas share similar levels of NO. This is very plausible considering the aforementioned comments that in suburban areas factories and companies exist while in the centre there are high levels of traffic. Furthermore, as we considered a signal to noise ratio of 1.5, that gives us an estimated measurement error variance of $\sigma_\epsilon^2 = 1.5 * 6.3 = 9.45$ providing us with an estimation higher than the

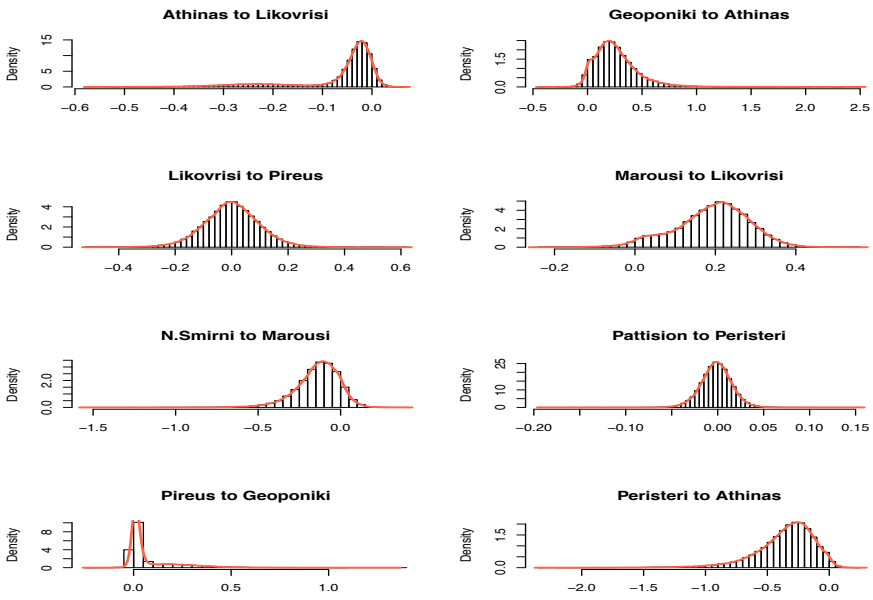

Figure 3.16: Histograms of selected reconstructed elements of the redistributional kernel $w_s(u)$ produced from IDWT under Haar wavelet basis and the posterior estimates of $\boldsymbol{B}$. The red line indicates the empirical distribution of each element. The selected pairs are representing causality of one location to another. A weight function with a posterior mode being around zero indicates zero causality, otherwise, either a positive or negative causal relationship is assumed. The associate distributions are produced under $N = 2 * 10^6$ Gibbs iterations with a burn in period of $i = 10^6$.

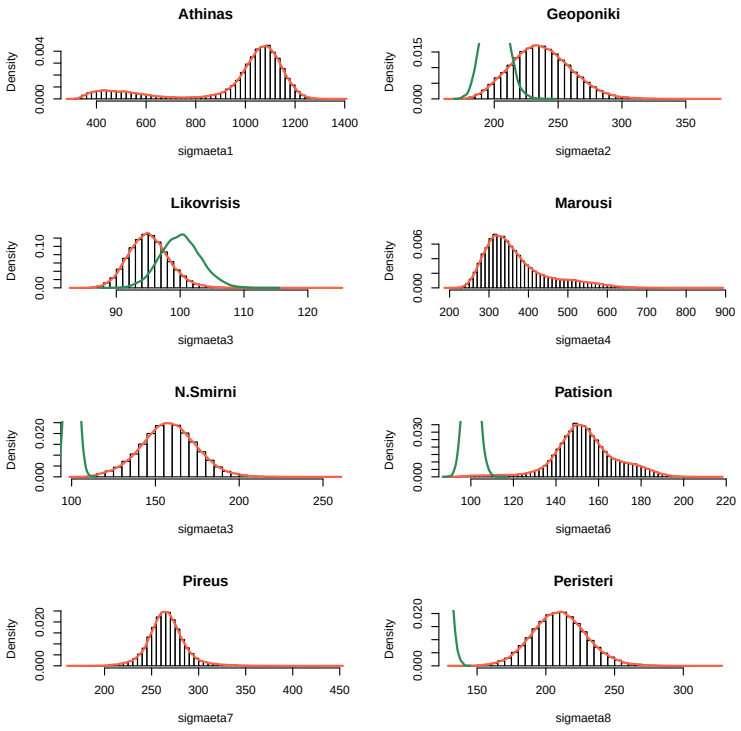

Figure 3.17: Posterior density estimates for the temporal variance elements of states $\boldsymbol{\alpha}_t$, σ_η. Red lines indicate the empirical distribution of the estimates. Green lines indicate the prior distribution for the covariance elements, an inverse gamma with 10^3 and 0.01 being the shape and scale parameters which indicates a high informative prior.The inference was conducted under $N = 2 * 10^6$ Gibbs iterations with a burn in period of $i = 10^6$.

spatial variance σ_ν^2. This is plausible as the variability of NO is high and the model considers the extra variability through the error variance parameter σ_ϵ^2.

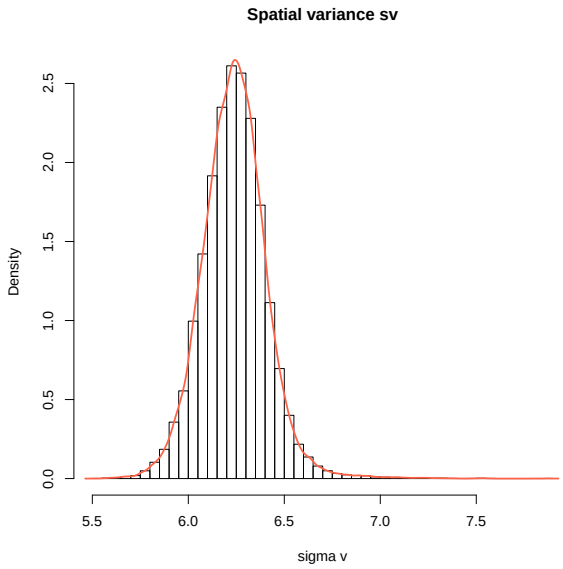

Figure 3.18: Posterior density estimates for the spatial variance of the observed measurements $\boldsymbol{Y}_t$, σ_ν. The orange line indicates the empirical distribution of the estimates. An inverse gamma with 10^3 and 0.01 being the shape and scale parameters which indicates a highly informative prior.The inference was conducted under $N = 2 * 10^6$ Gibbs iterations with a burn in period of $i = 10^6$.

Missing Value Treatment In this application a simple moving average imputation was used for the calculation of the missing values. The disadvantages of moving average analysis center around its simplicity and subjective flexibility. A simple moving average places the exact same weight on for instance, ten time points into the past, i.e., $t - 10$ in the observations that took place at $t - 1$. Thus, while it is simple to impute a missing value, it cannot possibly capture volatile trends as a whole. Furthermore, even with the use of shorter moving averages that can be used to model volatility, it's very challenging to decide on the correct window to use.

As we are working under a Bayesian setting, we can treat the missing values as parameters subject to estimation or else, derive posterior inference for their prediction. Let $\boldsymbol{Y}_t = (\boldsymbol{Y}_t^{mis}, \boldsymbol{Y}_t^{obs})$ and consider the missing data indicator $\boldsymbol{m}$ which indicates as 1 being a missing value and as 0 an observed value. Then, the model can be factorised

as:

$$p(\boldsymbol{Y}_t, \boldsymbol{m}) = p(\boldsymbol{m}|\boldsymbol{Y}_t^{mis}, \boldsymbol{Y}_t^{obs}, \boldsymbol{\alpha}_t, \sigma_\epsilon^2, \sigma_\nu^2) p(\boldsymbol{Y}_t^{mis}, \boldsymbol{Y}_t^{obs}|\boldsymbol{\alpha}_t, \sigma_\epsilon^2, \sigma_\nu^2)$$

when the missing data mechanism $\boldsymbol{m}$ is ignorable, then we can ignore $p(\boldsymbol{m}|\boldsymbol{Y}_t^{mis}, \boldsymbol{Y}_t^{obs}$ and just fit the analysis model from $p(\boldsymbol{Y}_t^{mis}, \boldsymbol{Y}_t^{obs}|\boldsymbol{\alpha}_t, \sigma_\epsilon^2, \sigma_\nu^2)$ which is the usual likelihood that we would specify under a fully observed response $\boldsymbol{Y}_t$. Thus, estimating the missing responses $\boldsymbol{Y}_t^{mis}$ is equivalent to posterior prediction from the model fitted to the observed data. Obviously, one would need to conduct forecast on the equivalent state vector $\boldsymbol{\alpha}_t$ through the FFBS recursions while also using the sampling parameters under the $i-th$ Gibbs iteration.

3.9 Conclusion

We have introduced an adaptive Bayesian procedure for Gaussian Dimension Reduced DSTMs. This truncation makes use of an efficient sparse wavelet decomposition where its spatial coefficients are inferred through a Spike and Slab prior. Furthermore, an efficient filtering and smoothing procedure under a Bayesian framework is used for the estimation of the temporal wavelet coefficients. Lastly, a flexible Bayesian estimation is provided with preferable covariance structures. Last but not least, a simulation scheme was introduced for Gaussian Dimension Reduced DSTM processes under wavelet basis decomposition.

Firstly, case studies for both a small and relatively higher number of locations have proved the effectiveness of our methodology on approximating an underlying spatio-temporal process with complex dynamics. Moreover, the proposed methodology successfully approximates processes with spatial discontinuities (section 3.7.3). Additionally, the reconstruction of the weight function, which is a challenge in practice to be estimated, was predicted fairly well. It has to be noted that known kernels where used to be simulated instead of the Spike and Slab hierarchy itself which makes our model a successful detector of spatial kernels. As for the inferential procedure of the covariance structure, it was found to be successfully adaptive and flexible as we predicted all our covariance parameters pretty well. Furthermore, we tested our methodology on real data without using any detrending procedures in contrast to Wikle and Cressie (1999). Still, our model managed to capture both the trend and seasonal effects of the processes and we managed to approximate the processes very well. Furthermore, we achieved to produce spatial causality between locations. These spatial causalities seemed reason-

able based on the geology, social and economical aspects but mostly the distance of those areas. Finally, the most stable process to be sampled is $\boldsymbol{\alpha}$, since Kalman Filter and smoothing recursions are very adaptive since they take into consideration the past and future values and therefore the prior information given at zero time is washed out by the model itself.

However, no methodology is perfect for such complex models and we encountered hindrances during our estimation. Firstly, the hyperparameters of the Spike and Slab hierarchy under this framework should be considered with care. For instance, if we are expecting to have a large amount of sparsity, then v_0 parameter should be set very small close to 0. However, since it affects the variance elements of β_k it may provide a structure to $\boldsymbol{B}$ which produces elements generally very close to zero and thus it results into a very complicated thresholding approach. On the other hand, if v_0 is large and closer to 0.5, it tends to produce elements that are non-zero but there are only a few zero ones. However, the value of v_0 should be considered together with the hyperparameter values for the precision τ_k^2. For instance, if very high or very low mean values are set, then it will result either into very large elements on the matrix B or very low ones combined with the respective value that will be put to v_0. Moreover, the more parameters we have under the Spike and Slab, combined with the signal-to-noise ratio, the more we will tend to fail to estimate the large values of $\boldsymbol{B}$. Finally, the application on real data showed as that the model is sensitive in predicting the causality when a large amount of missing data exists.

In terms of the hyperparameters of the variance elements in $\boldsymbol{\Sigma}_\eta$ and σ_v, non-informative priors do tend to affect the sampling in our simulations and the estimation in terms of calculations. Additionally, it was observed that the more we increase the number of locations, the more computational power we need for the model to run. The rate of iterations for the parameters to converge by n is exponentially increasing, while the more complicated covariance structure we have, the rate of iterations is affected in an additive way.

Finally, one could resort to another type of efficient inferential approach through *Stan* which is a probabilistic programming language (*http://mc-stan.org*). This software makes the required computation automatically using the most elegant and up to date techniques including automatic differentiation through Hamiltonian Monte Carlo. Feeding the data into Stan makes the inference faster as it is compiled to C++, while

there is a package developed in *R* which extends to Bayesian Inference as well. Although *Stan* already provides efficient inference for a wide range of models, it has its limitations. Hierarchical models often suffer from inefficiencies due to distributions with difficult posterior geometries, and in many cases reparameterization can help. Other sources of problems can be highly non-linear dependency structures (leading to banana-shaped, curved posteriors), multi-modal posterior distributions, and long-tailed distributions. Thus, an implementation of our methodology through *Stan* would definitely increase the computational efficiency, though its implementation should be examined with caution.

Chapter 4

Poisson Reduced-dimension Dynamic Spatio-Temporal Models

4.1 Introduction

In the previous chapter we proposed an adaptive Bayesian modeling procedure with the help of wavelet basis decomposition for the model of Wikle and Cressie (1999). That model as well as our methodology considers that the observed measurements are normally distributed which means that both the observation and state equations are linear. However, many spatio-temporal data are generated from non-Gaussian distributions. For instance, if we would like to model traffic accidents in specific areas, then we have Poisson counts, or if we would like to model the cancer rates per location, then we have Multinomial proportions. Therefore, the implementation of Gaussian DSTMs in those datasets is false as the observation equation is non-linear anymore. Additionally, combining the non-linearity of observations and that spatially the problem is high dimensional, then the inference and a realistic implementation makes it a challenge in practice.

Based on the aforementioned arguments, in this chapter we are going to focus on the reduced-dimension DSTM under Poisson counts proposed by Wikle (2002). Specifically, we will address the limitations of that modeling procedure which is based solely on a specific application in cloud intensity data by considering an overall spatial mean

effect for all locations of interest. Thus, we will expand this modeling procedure to a general case in terms of the spatial mean effect but also we will introduce an autoregressive framework. Additionally, we will consider in this chapter the wavelet basis decompositions due to their efficiency and good localisation properties. However, the inferential procedure unlike in the previous chapter will not be based solely on Gibbs sampling for the temporal components $\boldsymbol{\alpha}_t$ of the model. Thus, we propose an efficient inferential procedure through particle filtering (PF) methods where we will provide thorough explanation in the latter sections.

Simulation implementations were conducted in both proposed modelling frameworks. Our findings are promising for processes of counts. Specifically, the mean of Poisson processes was captured fairly well even in abrupt peaks, which is in general a challenge in practice. Furthermore, the weight function was successfully reconstructed, however, updating the elements of $\boldsymbol{B}$ is much slower than the Gaussian case where we solely conducted MCMC inference. Additionally, the covariance estimation, in which in this case less parameters are considered, was conducted fairly well.

Finally, at the end of this chapter we offer a real life application to traffic flow data under our proposed methodology. Our findings include that the proposed methodology is approximating fairly well mean intensities under a low number of time points and imputed missing values. Last but not least, we derive causal relationships in the traffic flow between counties in the M6 motorway from the reconstruction of the weight function.

4.2 Discussion of Poisson DSTM

In this section we discuss the spatio-temporal approach of Wikle (2002) which is based on the non-Gaussian spatial modeling approach of Diggle et al. (1998) and the hierarchical representation of Wikle et al. (1998). It is assumed that conditional on a Poisson intensity process at all spatial and temporal locations of interest, the data are distributed as independent Poisson random variables.

4.2.1 Model Formulation

Consider the count spatio-temporal process $Y(s,t)$ under a spatio-temporal Poisson intensity $\lambda(s,t)$, with $s \in D \subset \mathbb{R}$ denoting the locations and $t = 1, 2, \ldots, T$ denoting the discrete time points, i.e.,

$$Y(s,t)|\lambda(s,t) \sim \text{Poi}(\lambda(s,t)) \tag{4.1}$$

where $\lambda(s,t)$ is the Poisson intensity process at spatial location s and time point t. In order to bring the process into a linear framework, $\lambda(s,t)$ can be modeled via a generalised dynamic linear model, i.e.,

$$\log(\lambda(s,t))|\mu, \boldsymbol{\alpha}_t, \sigma_\epsilon \sim \text{N}(\mu + \nu\boldsymbol{\phi}_s\boldsymbol{\alpha}_t, \sigma_\epsilon \boldsymbol{I}) \tag{4.2}$$

where μ is the overall mean effect, same for all locations and σ_ϵ represents the extra Poisson variability and let $\boldsymbol{\alpha}_t$ evolve as in (2.11) and ν is a scaling parameter while ϕ_s represents the $s-th$ row of the wavelet matrix $\boldsymbol{\Phi}$.

A few comments are in order. Firstly, Wikle (2002) considers an overall mean effect for all locations which means that the spatial contribution on the intensities $\lambda(s,t)$ for each location should be similar. This approach was applied to cloud intensity data where someone would expect a similar spatial aspect as there is no geological attributes to consider. For instance, if the random variable of interest was the number of events of a river flooding within a month, one would expect a different mean contributing to the intensity as the inner and outer ground attributes may vary from one location to another.

Secondly, the spatial kernel $w_s(u)$ in Wikle (2002)) is considered to be Gaussian while a spectral decomposition is conducted on the translation and dilation parameters where this brings a hierarchical framework for the estimation of those parameters. This means that the weight function is not approximated anymore but is already predefined and the focus of introducing parsimony is on each spatially varying kernel parameters. Specifically, he considers a Fourier basis decomposition on the kernel parameters where they will vary for each location and then conducts MCMC inference on them.

Finally, Wikle (2002) used Gibbs sampling as an inferential procedure for $\boldsymbol{\alpha}_t$. In Chapter 3 we explained the hindrances of conducting Gibbs sampling for the inference of parameters $\boldsymbol{\alpha}_t$. Furthermore, if we like to use the Forward Filtering Backward Sam-

pling algorithm, under this model, it makes it unreliable. For this reason we suggest that particle filtering approaches for the particles $\boldsymbol{\alpha}_t$ with Gibbs sampling steps for the rest of parameters instead.

4.2.2 Details of the problem and suggestions

Under the non-Gaussian framework several approaches have been implemented in terms of a transformation of the process into the hierarchy structure in order to bring it into a generalised dynamic framework. These models were developed by Diggle et al. (1998) while they have been broadly used in the case of spatio-temporal processes, such as in Brix and Diggle (2001), Wikle (2003), Wikle and Hooten (2006), Hooten et al. (2007) and Hooten and Wikle (2008) where each author uses an exponential family approach of generalised linear models. Moreover, there is a plethora of approaches for datasets of non-linear nature. Specifically, Wendt et al. (2004) proposed a Bayesian waypoint analysis on nonlinear equations of motion, Wikle and Holan (2011) consider a hierarchical framework under Integro-difference equations (IDEs) with stochastic variable selection. For more details on nonlinear approaches there is a full literature in Cressie and Wikle (2015).

Given that the methodology in Wikle (2002) is developed and applied under a specific application, it is sensible for us to propose a generalisation of it and consider some approaches for μ. Specifically, if we want to induce this notion for other applications, it would be sensible for the mean effect to consider $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$, i.e, to differ for each location and incorporate that into the intensity process. Furthermore, an alternative and even more flexible approach is to consider an autoregressive structure for the mean vector $\boldsymbol{\mu}$. This relies on the fact that in many applications we would expect the mean effect to be conditionally dependent on the previous time point, for instance, the number of trees at time t in one area would be affected from the number of trees in the same and the rest of locations at $t-1$.

Additionally, under our suggestion on using wavelet decomposition on the weight function, we are able to successfully approximate discontinuities of very low and high contributions that affect in an autoregressive manner the intensity process without predefining the weighting function kernel as in Wikle (2002). This means that the data adaptively will guide us to understand the nature of that kernel while also introducing parsimony.

Finally, it is notable that due to the nature of the observations we can either observe intensities or counts. In the case where we are receiving count measurements σ_ϵ^2 will not be needed in the model anymore, however, in the case of different application where our observed measurements are intensities, one can consider it in the model. In the rest of the thesis we will focus on count observations.

On the computational aspect, in nonlinear approaches that were developed for spatio-temporal processes, authors had to tackle with the unrealistic formulation under a generalised framework. Even so, generalised dynamic state-space models have always been a flexible and efficient way to model non-Gaussian processes by considering temporal dependencies. Thus, considering these formulations under Bayesian settings has been a debatable issue. Extended Kalman Filter (EKF) methods have been a solution to this problem, where the first two moments provide a good approximation to the distribution of the measurement and state processes. Alternatively, Gibbs sampling steps within Metropolis-Hastings in order to sample the full conditionals, only by considering linearised model distributions (Cressie and Wikle, 2015). However, MCMC methods for high dimensional non-linear state space models are difficult to implement and need a lot of information in terms of tuning. A new evolving efficient procedure that is fairly used for spatial processes' modelling is the Integrated nested Laplace Approximation. This method approximates posterior marginals efficiency in models with latent Gaussian processes, however, if there is a high dimensional parameter space of non-Gaussian hyperparameters and spatial structures that cannot be coerced, it is questionable if INLA could perform well under non-Gaussian DSTMs (Cressie and Wikle, 2015).

In Chapter 3 we conducted inference via the FFBS algorithm that was implemented for the inference of the temporal wavelet coefficients α_t. However, in this modeling framework it cannot be used as the Gaussianity in the observation equation is violated. This already brings an inferential problem in terms of smoothing. Secondly, as it will be shown, Poisson counts are very difficult to be modeled temporally in the case of abrupt peaks in the series, even if wavelets help us to track the discontinuities. Finally, one question is how one can take into consideration the spatial correlation that exists between the locations as the structure of the error terms is not similar to the Gaussian DSTMs. Having said that, we will consider the implementation of Particle Filtering (PF) techniques for the temporal components. Therefore, an appropriate particle filtering algorithm will be used in order to derive the filtered posterior estimates $\boldsymbol{\alpha}_t$. In

order to achieve that we will consider many scenarios, but we will keep the Spike and Slab prior belief of $\boldsymbol{B}$ the same as in the Gaussian case in Chapter 3.

4.3 Proposed Approach

In order to address the limitations in Wikle (2002), two concepts will be discussed. The first one deals with the choice in the formulation of the overall mean effect μ for all spatial locations s. The second one deals with the Bayesian inference which consists of a multivariate form of a Spike and Slab prior for the spatial coefficient matrix $\boldsymbol{B}$ and particle filter methods which the combination of the two brings us into an efficient inferential scheme. Finally, the resulting Conditional Particle Filtering (CPF) and Particle Metropolis-Hastings (PMH) algorithms both under Static parameter estimation (Storvik, 2002) will be presented for the inference of the temporal parameters $\boldsymbol{\alpha}_t$ (and $\boldsymbol{\mu}_t$).

4.3.1 Spatially varying mean effect

As stated in the previous sections, Wikle (2002) considers the spatial overall mean effect μ to be static and same for all locations. However, by considering most real life applications this is an unrealistic assumption as we would not expect spatial locations to behave in a similar way unless we have concrete evidence and knowledge from experts on the application.

Suppose we observe the Poisson distributed spatio-temporal measurements $Y(s,t)$ with mean intensity $\lambda(s,t)$. Let us consider the spatially mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top$ that contributes in the spatial aspect of the intensity process $\lambda(t,s)$ and the state equation in (2.11), i.e.,

$$\begin{aligned} \log(\lambda(s,t)) &= \mu_s + \boldsymbol{\Phi}^\top \boldsymbol{\alpha}_t \\ \boldsymbol{a}_t &= \boldsymbol{\Phi}^\top \boldsymbol{B} \boldsymbol{\alpha}_{t-1} + \boldsymbol{\Phi}^\top \boldsymbol{\eta}_t \end{aligned} \tag{4.3}$$

where μ_s is now the mean effect for the spatial location s and $\boldsymbol{\eta}_t \sim \mathrm{N}(0, \boldsymbol{\Sigma}_\eta)$. The equations in (4.3) suggest that the intensity $\lambda(s,t)$ will be affected from $\lambda(u, t-1)$ in an autoregressive manner through the state equation of $\boldsymbol{\alpha}_t$. Moreover, an extra spatial contribution is incorporated into the autoregressive structure under a temporal error $\boldsymbol{\eta}_t$. The framework on this modelling approach is summarised in Table 4.1.

Data:	
Observed spatio-temporal process:	$\mathbf{Y}$, $T \times n$ matrix
Intensity spatio-temporal process:	$\boldsymbol{\lambda}$, $T \times n$ matrix
Spatially varying mean effect:	$\boldsymbol{\mu}$, $1 \times n$ vector
Redistribution kernel:	$\mathbf{w}$, $n \times n$ matrix
Approximations:	
$\boldsymbol{\lambda} = \boldsymbol{\alpha}\Phi$,	$\boldsymbol{\alpha}_{T\times n}$, coefficients of matrix $\Phi_{n\times K}$
$\boldsymbol{w}_s = \boldsymbol{B}\Phi$,	$\boldsymbol{B}_{K\times n}$, coefficients of matrix $\Phi_{n\times K}$
Model:	
$\log(\boldsymbol{\lambda}_t) = \boldsymbol{\mu} + \boldsymbol{\Phi}\boldsymbol{\alpha}_t$	$\boldsymbol{\lambda}_t = \exp(\boldsymbol{\mu} + \boldsymbol{\Phi}\boldsymbol{\alpha}_t)$
$\boldsymbol{\alpha}_t = \boldsymbol{\Phi}^\top \boldsymbol{B}\boldsymbol{\alpha}_{t-1} + \boldsymbol{\eta}_t$	$\boldsymbol{\eta}_t \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_\eta)$
Parameters and Prior distributions:	
$\boldsymbol{\alpha}_{0\|0} \sim \mathrm{N}(\boldsymbol{m}_0, \boldsymbol{P}_{0\|0})$	$\boldsymbol{m}_0$, $\boldsymbol{P}_{0\|0}$ prior mean & covariance
$\mathrm{vec}(\boldsymbol{B})\|\boldsymbol{\Gamma} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Gamma} \otimes \boldsymbol{I})$	$\boldsymbol{\Gamma} = \mathrm{diag}\{\gamma_1, \ldots, \gamma_k\}$, $\gamma_k = \rho_k \tau_k^2$
$\rho_k \| v_0, q \sim (1-q)\delta_{v_0}(.) + q\delta_1(.)$	$q \sim \mathrm{U}(0,1)$
$\tau_k^{-2} \| \omega_1, \omega_2 \sim G(\omega_1, \omega_2)$	$\boldsymbol{\beta}_k \sim \mathrm{N}(\mathbf{0}, \gamma_k \boldsymbol{I})$
$\boldsymbol{\mu}_{0\|0} \sim \mathrm{N}(\boldsymbol{\mu}_0, \boldsymbol{M}_{0\|0})$	$\boldsymbol{\Sigma}_\eta \sim \mathrm{IW}(\nu, Q)$ or $\sigma_\eta \sim \mathrm{IG}(\psi_1, \psi_2)$

Table 4.1: Framework of the model under spatially varying $\boldsymbol{\mu}$

A few comments are in order. The spatial effect $\boldsymbol{\mu}$ in this case is considered static. This means that each location s spatially will behave in the same magnitude in its own intensity $\lambda(s,t)$ while the rest of the locations will affect it in an autoregressive manner via $\boldsymbol{\alpha}_t$. In the case where the overall mean effect is the same for all locations, we would expect to be estimated appropriately. In the Gaussian case of Chapter 3 we considered a spatial variation σ_ν under a spatial correlation function. In this case we can complement the loss of not using a spatial covariance function through introducing a separate mean effect which can affect the locations separately while the approximation of the weight function under the matrix $\boldsymbol{B}$ will provide us with the spatial diffusion dynamics. As noticed, we have removed the scaling parameter v. As we are now considering more separate spatial parameters we consider the spatial scale aspect to be affected by the combination of each μ_s and the matrix $\boldsymbol{B}$.

Additionally, this approach considers a strict spatial progression as we expect the locations across time to always meet the same mean effect. If a process of interest has sudden jumps across time, we would expect that this model would be unsuitable for that process as we would like a more flexible one that could track the spatial discontinuities across time combines with the wavelet basis decomposition.

4.3.2 Autoregressive structure on μ

Suppose again we observe the Poisson distributed spatio-temporal measurements $Y(s,t)$ with mean intensity $\lambda(s,t)$ but with dynamic structure on overall mean vector $\boldsymbol{\mu}$ due to the dynamic coefficients α_t which may contribute into the posterior mean of $\boldsymbol{\mu}$. Therefore, presume to give $\boldsymbol{\mu}$ an autoregressive structure, i.e.,

$$\mu(s,t) = \psi_s \mu(s,t-1) + \zeta_t \tag{4.4}$$

with the temporal error $\zeta_t \sim \mathrm{N}(0, \sigma_\zeta)$ and ψ_s specifying the autocorrelation of the mean effect μ_s between time $t-1$ and t. Thus, by defining $\boldsymbol{\mu}_t = (\mu_{1t}, \ldots, \mu_{nt})^\top$ and the diagonal $n \times n$ matrix $\boldsymbol{\Psi} = \mathrm{diag}(\psi_1, \ldots, \psi_n)$, the hierarchy (4.3) changes to:

$$\begin{aligned} \log(\boldsymbol{\lambda}_t)) &= \boldsymbol{\mu}_t + \boldsymbol{\Phi}^\top \boldsymbol{\alpha}_t \\ \boldsymbol{\alpha}_t &= \boldsymbol{\Phi}^\top \boldsymbol{B} \boldsymbol{\alpha}_{t-1} + \boldsymbol{\Phi}^\top \boldsymbol{\eta}_t \text{ with } \boldsymbol{\eta}_t \sim \mathrm{N}(0, \boldsymbol{\Sigma}_\eta) \\ \boldsymbol{\mu}_t &= \boldsymbol{\Psi} \boldsymbol{\mu}_{t-1} + \boldsymbol{\zeta}_t \text{ with } \boldsymbol{\zeta}_t \sim \mathrm{N}(0, \sigma_\zeta \boldsymbol{I}) \end{aligned} \tag{4.5}$$

where a simple diagonal structure for the variance of the error component ζ_t is considered. In the case that $\psi_s = 1$ then (4.4) takes the form of a random walk model. The autocorrelation matrix $\boldsymbol{\Psi}$ is considered diagonal as all of the spatially and temporally spatial information are included in matrix $\boldsymbol{B}$ and the individual temporal vector $\boldsymbol{\mu}_t$. Furthermore, for the complexity of our model and parameter space, we consider the variance σ_ζ to own a simple structure as we would not expect $\boldsymbol{\mu}_t$ in each spatial location to vary greatly since we are incorporating that variation in $\boldsymbol{\Sigma}_\eta$ as well.

For the models in (4.3) and (4.5) we have combined the Spike and Slab prior belief as in Chapter 3, and the model is summarised in Table 4.2. However, we are introducing a new inferential framework. For each case we have alternative approaches according to the choice of the overall mean vector $\boldsymbol{\mu}$, however, both lie in the particle filtering (PF) algorithms. These particle filtering methods, along with the extensions that we are using in our inferential procedure are provided in the following section.

Data:	
Observed spatio-temporal process:	$\mathbf{Y}$, $T \times n$ matrix
Intensity spatio-temporal process:	$\boldsymbol{\lambda}$, $T \times n$ matrix
Autoregressive spatially varying mean effect:	$\boldsymbol{\mu}$, $T \times n$ vector
Autocorrelation parameters:	$\boldsymbol{\Psi}$, $n \times n$ matrix
Redistribution kernel:	$\mathbf{w}$, $n \times n$ matrix
Approximations:	
$\boldsymbol{\lambda} = \boldsymbol{\alpha}\Phi$,	$\boldsymbol{\alpha}_{T\times n}$, coefficients of matrix $\Phi_{n\times K}$
$\boldsymbol{w}_s = \boldsymbol{B}\Phi$,	$\boldsymbol{B}_{K\times n}$, coefficients of matrix $\Phi_{n\times K}$
Model:	
$\log(\boldsymbol{\lambda}_t) = \boldsymbol{\mu} + \boldsymbol{\Phi}\boldsymbol{\alpha}_t$	$\boldsymbol{\lambda}_t = \exp(\boldsymbol{\mu} + \boldsymbol{\Phi}\boldsymbol{\alpha}_t)$
$\boldsymbol{\alpha}_t = \boldsymbol{\Phi}^\top \boldsymbol{B}\boldsymbol{\alpha}_{t-1} + \boldsymbol{\eta}_t$	$\boldsymbol{\eta}_t \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_\eta)$
$\boldsymbol{\mu}_t = \boldsymbol{\Psi} + \boldsymbol{\zeta}_t$	$\boldsymbol{\zeta}_t \sim \mathrm{N}(\mathbf{0}, \sigma_\zeta \boldsymbol{I})$
Parameters and Prior distributions:	
$\boldsymbol{\alpha}_{0\|0} \sim \mathrm{N}(\boldsymbol{m}_0, \boldsymbol{P}_{0\|0})$	$\boldsymbol{m}_0$, $\boldsymbol{P}_{0\|0}$ prior mean & covariance
$\mathrm{vec}(\boldsymbol{B})\|\boldsymbol{\Gamma} \sim N(\mathbf{0}, \boldsymbol{\Gamma} \otimes \boldsymbol{I})$	$\boldsymbol{\Gamma} = \mathrm{diag}\{\gamma_1, \ldots, \gamma_k\}$, $\gamma_k = \rho_k \tau_k^2$
$\rho_k\|v_0, q \sim (1-q)\delta_{v_0}(.) + q\delta_1(.)$	$q \sim \mathrm{U}(0,1)$
$\tau_k^{-2}\|\omega_1, \omega_2 \sim \mathrm{G}(\omega_1, \omega_2)$	$\boldsymbol{\beta}_k \sim \mathrm{N}(\mathbf{0}, \gamma_k \boldsymbol{I})$
$\boldsymbol{\mu}_{0\|0} \sim \mathrm{N}(\boldsymbol{\mu}_{0\|0}, \boldsymbol{M}_{0\|0})$	$\boldsymbol{\mu}_{0\|0}$, $\boldsymbol{M}_{0\|0}$ prior mean & covariance
$\mathrm{diag}(\boldsymbol{\Psi}) \sim \mathrm{N}(\boldsymbol{\psi}_0, \boldsymbol{v}_0 \boldsymbol{I})$	$\boldsymbol{\Sigma}_\eta \sim \mathrm{IW}(\nu, Q)$ or $\sigma_\eta \sim \mathrm{IG}(\psi_1, \psi_2)$

Table 4.2: Framework of the model under autoregressive $\boldsymbol{\mu}$

4.3.3 Inference through Particle Filtering

Particle Filtering (PF) methods have been widely used in state-space models in order to update recursively the conditional posterior distributions when Kalman Filter techniques are not efficient or when the model is of nonlinear nature (for details, see Cappé et al. (2007)). Particle filtering is an extension of importance sampling in the evaluation of the expected value of $f(x)$ with respect to the distribution $\pi(x)$, i.e., $\mathrm{E}_\pi[f(x)] = \int f(x)\pi(x)dx$. If an importance density is considered then one can approximate that expected value $\mathrm{E}_\pi[f(x)] = \int f(x)\frac{\pi(x)}{g(x)}g(x)dx = \mathrm{E}_g[f(x)w^*(x)]$ instead, with $w^*(x) = \frac{\pi(x)}{g(x)}$ being called the importance function.

Through sequential Monte Carlo an approximate estimate is provided such as

$$\begin{aligned} \mathrm{E}_\pi[f(x)] &\approx \frac{1}{N} f(x^{(i)}) w^*(x) = \frac{1/N \sum_{i=1}^N f(x^{(i)}) \tilde{w}^{(i)}}{C} \\ &\approx \frac{\sum_{i=1}^N f(x^{(i)}) \tilde{w}^{(i)}}{\sum_{i=1}^N \tilde{w}^{(i)}} = \sum_{i=1}^N f(x^{(i)}) w^{(i)} \end{aligned}$$

with $\tilde{w}^{(i)}(x) = C w^*(x^{(i)})$, with C being the normalising constant and is evaluated during the unnormalised particle weights $\tilde{w}^{(i)}(x)$. The normalised weights $w^{(1)}, \ldots, w^{(N)}$, i.e., $w^{(i)} = \tilde{w}^{(i)}(x) / \sum_{j=1}^N \tilde{w}^{(j)}(x)$ sum to one and give us along with the sample particles $x^{(1)}, \ldots, x^{(N)}$ a discrete approximation of the target function π, i.e., $\hat{\pi} = \sum_{i=1}^N w^{*(i)} \delta_{x^{(i)}}$, with w^* and δ_x being the normalising weights and the point mass function for x respectively. For more details, particle filtering in terms of state-space models are explained in Petris et al. (2009).

Thus, given the observed data $\boldsymbol{Y}_t$ we can conduct inference of the states $\boldsymbol{\alpha}_t$ by assuming the variance matrices known based on the approximation of the posterior through particle filtering with N random sampled particles—or trajectories— and via the log transform in calculating $\boldsymbol{\lambda}_t$ as well. The algorithm is a combination of importance sampling and resampling techniques.

By having a prior and an importance density $p(\boldsymbol{\alpha}_0)$, $q(\boldsymbol{\alpha}_0|\boldsymbol{Y}_0)$ respectively for $t = 0$ and $p(\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1})$, $q(\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1}, \boldsymbol{Y}_t)$ respectively for $t \geq 1$ then we can define the importance weights as:

$$\tilde{w}_0^{(i)} = \frac{p(\boldsymbol{\alpha}_0^{(i)}) p(\boldsymbol{Y}_0|\boldsymbol{\alpha}_0^{(i)})}{q(\boldsymbol{\alpha}_0^{(i)}|\boldsymbol{Y}_0)}, \tag{4.6}$$

$$\tilde{w}_t^{(i)} = \frac{p(\boldsymbol{\alpha}_t^{(i)}|\boldsymbol{\alpha}_{t-1}^{(i)}) p(\boldsymbol{Y}_t|\boldsymbol{\alpha}_t^{(i)})}{q(\boldsymbol{\alpha}_t^{(i)}|\boldsymbol{\alpha}_{t-1}^{(i)} \boldsymbol{Y}_t)} \tilde{w}_{t-1}^{(i)} \tag{4.7}$$

where $p(\boldsymbol{\alpha}_0)$ specifies the prior for time $t = 0$ and $p(\boldsymbol{\alpha}_t^{(i)}|\boldsymbol{\alpha}_{t-1}^{(i)})$ for $t \geq 1$ and $p(\boldsymbol{Y}|\boldsymbol{\alpha}^{(i)})$ being the density of the observations.

The choice of the importance function can be either the prior or of a different form. The most flexible approach, which we will consider in this chapter, is the known Bootstrap particle filtering that considers the importance function being the same as the prior and then the algorithm is simplified. More specifically, if we set $p(\boldsymbol{\alpha}_0^{(i)}) = q(\boldsymbol{\alpha}_0|\boldsymbol{Y}_0^{(i)})$

and $p(\boldsymbol{\alpha}_t^{(i)}|\boldsymbol{\alpha}_{t-1}^{(i)}) = q(\boldsymbol{\alpha}_t^{(i)}|\boldsymbol{\alpha}_{t-1}^{(i)})$ then the weights are simplified as

$$\tilde{w}_0^{(i)} = p(\boldsymbol{Y}_0|\boldsymbol{\alpha}_0^{(i)}), \tag{4.8}$$

$$\tilde{w}_n^{(i)} = p(\boldsymbol{Y}_t|\boldsymbol{\alpha}_t^{(i)})\tilde{w}_{t-1}^{(i)} \tag{4.9}$$

The final step in the updating process is consisted of scaling the unnormalised weights

$$w_t^{*(i)} = \frac{w_t^{(i)}}{\sum_{i=1}^{N} w_t^{(i)}}$$

However, especially at the early stage of particle filtering, some particles may have large weights while others may have really small ones and this phenomenon leads to a dislocation in the Monte Carlo approximation, the so called path degeneracy. In order to avoid this and consequently poor estimates, a function-free diagnostic tool should be considered. One such diagnostic which is widely used is the idea of effective sample size (N_{eff}). The effective sample size is defined as $N_{eff} = (\sum_{i=1}^{N}(w_t^{*(i)}))^{-1}$ and ranges between N if all particles are equal and one if one particle has a weight of one. Therefore, if N_{eff} falls under a threshold N_0 then we should conduct a resampling step with new particles and reset the respective weights to $1/N$. The resampling does not change the expected value of the targets but it increases its Monte Carlo variance and can be conducted in many different ways, however, in this thesis we will use a multinomial resampling which is the simplest one. Specifically, it is consisted of sampling N particles from $\hat{p}$ and by using the sampled points, with equal weights as the new approximations of the target function.

In the next sections we will explain each of the following updates in more detail and we will provide separate algorithms for each case.

- Update $\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1}, \boldsymbol{\lambda}_t, \boldsymbol{Y}_t$ and $\boldsymbol{\lambda}_t|\boldsymbol{\alpha}_t, \boldsymbol{\alpha}_{t-1}, \boldsymbol{Y}_t$ via Particle Filtering
- Update $\boldsymbol{B}|\boldsymbol{\alpha}_t, \boldsymbol{\alpha}_{t-1}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}_\eta$ through the Spike and Slab hierarchy via Conditional Particle Filtering with Ancestor Resampling (CPF-AS) (Lindsten et al., 2014)
- Update $\boldsymbol{\Sigma}_\eta|\boldsymbol{\alpha}_t, \boldsymbol{\alpha}_{t-1}, \boldsymbol{B}$ of the temporal components $\boldsymbol{\alpha}_t$ and $\sigma_\epsilon|\boldsymbol{\lambda}_t, \boldsymbol{\alpha}_t|\mu$ via Conditional Particle Filtering with Ancestor Resampling (CPF-AS)
- Update spatially varying $\boldsymbol{\mu}|\boldsymbol{\alpha}_t, \boldsymbol{\lambda}_t, \boldsymbol{Y}_t$ through:

  Sampling through Particle Metropolis-Hastings (PMH) (Andrieu et al., 2010) or

  Sampling through static parameter estimation (Storvik, 2002)

- Update the autoregressive mean effect $\boldsymbol{\mu}_t|\boldsymbol{\alpha}_t, \boldsymbol{\Psi}, \boldsymbol{\lambda}_t, \boldsymbol{Y}_t$ through particle filtering
- Update $\sigma_\zeta|\boldsymbol{\mu}_t, \boldsymbol{\Psi}$, the error variance of $\boldsymbol{\mu}_t$, via Conditional Particle Filtering with Ancestor Resampling
- Update the autocorrelation components $\boldsymbol{\Psi}|\boldsymbol{\mu}, \boldsymbol{\alpha}_t, \boldsymbol{\Psi}, \boldsymbol{\lambda}_t, \boldsymbol{Y}_t$ through Conditional Particle Filtering with Ancestor Resampling

    Sampling through Conditional Particle Filtering with Ancestor Resampling or

    Sampling through static parameter estimation

The next section will provide us with thorough details on the algorithmic and mathematical procedure for the estimation of the parameters of interest. Based on the preferred modelling framework different algorithms are considered. However, due to the high computational intensity we pick the most efficient based on the proposed models.

4.4 Updating the parameters

A summary of the proposed algorithms is shown in Table 4.3 based on the model formulation that is chosen. Furthermore, we indicate which algorithms are efficient. These algorithms will be explained in the next sections thoroughly. It can be noted that if we consider a static spatial mean vector $\boldsymbol{\mu}$ then the most efficient algorithm to be considered if all parameters are unknown is Algorithm 4.9 while the slowest, even with known variances is Algorithm 4.7. If we consider the autoregressive modelling framework (4.4), then the most efficient algorithm is Algorithm 4.13 and the slowest is Algorithm 4.11.

Table	*Algorithm*	*Model of* $\boldsymbol{\mu}$	*Parameters*	*Efficiency*
4.4	Bootstrap Particle Filtering	Static	$\boldsymbol{\alpha}_t$, $\boldsymbol{\lambda}_t$	Fast
4.5	Conditional Particle Filtering with Ancestor Resampling	Static	$\boldsymbol{\alpha}_t$, $\boldsymbol{\lambda}_t$, $\boldsymbol{B}$	Moderate
4.6	Conditional Particle Filtering with Ancestor Resampling	Static	$\boldsymbol{\alpha}_t$, $\boldsymbol{\lambda}_t$, $\boldsymbol{B}$, $\boldsymbol{\Sigma}$, σ_ϵ	Moderate
4.7	Particle Metropolis-Hastings	Static	$\boldsymbol{\alpha}_t$, $\boldsymbol{\lambda}_t$, $\boldsymbol{\mu}$	Slow
4.8	Bootstrap Particle Filtering Static parameter estimation	Static	$\boldsymbol{\alpha}_t$, $\boldsymbol{\lambda}_t$, $\boldsymbol{\mu}$	Fast
4.9	Conditional Particle Filtering with Ancestor Resampling Static parameter estimation	Static	$\boldsymbol{\alpha}_t$, $\boldsymbol{\lambda}_t$ $\boldsymbol{B}$, $\boldsymbol{\mu}$ σ_ϵ, $\boldsymbol{\Sigma}_\eta$	Moderately Slow
4.10	Bootstrap Particle Filtering	Autoregressive	$\boldsymbol{\alpha}_t$, $\boldsymbol{\lambda}_t$, $\boldsymbol{\mu}_t$	Fast
4.11	Conditional Particle Filtering with Ancestor Resampling	Autoregressive	$\boldsymbol{\alpha}_t$, $\boldsymbol{\lambda}_t$, $\boldsymbol{B}$, $\boldsymbol{\mu}$ $\boldsymbol{\Psi}$, $\boldsymbol{\Sigma}_\eta$, σ_ζ, σ_ϵ	Moderately Slow
4.12	Bootstrap Particle Filtering Static parameter estimation	Autoregressive	$\boldsymbol{\alpha}_t$, $\boldsymbol{\lambda}_t$, $\boldsymbol{\mu}_t$, $\boldsymbol{\Psi}$	Fast
4.13	Conditional Particle Filtering Ancestor Resampling Static parameter estimation	Autoregressive	$\boldsymbol{\alpha}_t$, $\boldsymbol{\lambda}_t$, $\boldsymbol{\mu}_t$ $\boldsymbol{B}$, $\boldsymbol{\Psi}$ $\boldsymbol{\Sigma}_\eta$, σ_ϵ, σ_ζ	Moderate

Table 4.3: Summary of proposed algorithms

4.4.1 Updating $\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1}, \boldsymbol{\lambda}_t, \boldsymbol{Y}_t$ and $\boldsymbol{\lambda}_t|\boldsymbol{\alpha}_t, \boldsymbol{\alpha}_{t-1}, \boldsymbol{Y}_t$

Under our model assumptions and by assuming that for now $\boldsymbol{\mu}$ is known, the density distribution of the measurements $\boldsymbol{Y}_t$ for each time point t can be written as

$$\begin{aligned} p(\boldsymbol{Y}_t|\boldsymbol{\lambda}_t, \boldsymbol{\alpha}_t, \boldsymbol{\mu}) &= \prod_{s=1}^{n} p(Y(t,s)|\lambda(t,s)) = \prod_{s=1}^{n} \frac{e^{-\lambda(t,s)}\lambda(t,s)^{Y(t,s)}}{Y(t,s)!} \\ &= e^{-\sum_{s=1}^{n}\lambda(t,s)} \prod_{s=1}^{n} \frac{\lambda(t,s)^{Y(t,s)}}{Y(t,s)!} \end{aligned} \tag{4.10}$$

with $\lambda(t,s) = \exp(\mu_s + \boldsymbol{\Phi}\boldsymbol{\alpha}_t)$, while the particles follow a Gaussian distribution density, i.e., $\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1}, \boldsymbol{B}, \boldsymbol{\Sigma}_\eta \sim \mathrm{N}(\boldsymbol{\Phi}^\top \boldsymbol{B}\boldsymbol{\alpha}_{t-1}, \boldsymbol{\Phi}^\top \boldsymbol{\Sigma}_\eta \boldsymbol{\Phi})$ with $\boldsymbol{B}$ and $\boldsymbol{\Sigma}_\eta$ being the unknown parameters to be estimated. Since $\lambda(t,s)$ and $\boldsymbol{\alpha}_t$ are related, during the sampling of

the N particles of $\boldsymbol{\alpha}_t$, the intensities $\lambda(t,s)$ will be calculated so that we can derive the approximate filtered estimates, i.e., $\hat{\lambda}(t,s)|\boldsymbol{\alpha}_t = \sum_{i=1}^{N} w^{(i)}\lambda^{(i)}(t,s)$. The bootstrap particle filter algorithm for $\boldsymbol{\alpha}_t$ and $\lambda(t,s)$ is summarised bellow by considering that the parameters $\boldsymbol{B}$ and $\boldsymbol{\Sigma}_\eta$ are known:

Initial step:
Simulate N particles $\boldsymbol{\alpha}_0^{(1)}, \ldots, \boldsymbol{\alpha}_0^{(N)}$ from $p(\boldsymbol{\alpha}_0)$
Calculate $\boldsymbol{\lambda}_0^{(1)}, \ldots, \boldsymbol{\lambda}_0^{(N)}$
Set $w_0^{(i)} 1/N$, $i = 1, \ldots, N$
Particle Sampling:
For $t = 1, \ldots, T$:
Sample $\boldsymbol{\alpha}_t^{(1)}, \ldots, \boldsymbol{\alpha}_t^{(N)}$ from the importance function $g(\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1}^{(i)}, Y_t)$
Calculate $\boldsymbol{\lambda}^{(i)}(t,s) = \exp(\mu_s + \boldsymbol{\Phi}\boldsymbol{\alpha}_t^{(i)})$
Calculate the weights $\tilde{w}_t^{(i)}$ from (4.9)
Normalise the weights $w_t^{(i)} = \frac{\tilde{w}_t^{(i)}}{\sum_{i=1}^{N} \tilde{w}_t^{(i)}}$
Resampling step: Multinomial Resampling
Calculate the effective sample size $\mathrm{N}_{eff} = (\sum_{i=1}^{N}(w_t^{(i)})^2)^{-1}$
Draw N indices $i_1, \ldots, i_N$ from the discrete distribution $P(\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_t^{(i)}) = w_t^{(i)}$
Relabel the sample $\boldsymbol{\alpha}_t^{(i)} = \boldsymbol{\alpha}_t^{(i_j)}$, for $i = 1, 2, \ldots, N$
Update to equal weights by $w_t^{(i)} = 1/N$
Posterior Estimation Approximate the posteriors
$\hat{p}(\boldsymbol{\alpha}_t|\boldsymbol{Y}_{1:t}) = \sum_{i=1}^{N} w_t^{(i)}\delta(\boldsymbol{\alpha}_t - \hat{\boldsymbol{\alpha}}_t)$, where $\hat{\boldsymbol{\alpha}}_t = \sum_{i=1}^{N} w_t^{(i)}\boldsymbol{\alpha}_t^{(i)}$
$\hat{p}(\boldsymbol{\lambda}_t|\boldsymbol{Y}_{1:t}, \boldsymbol{\alpha}_t) = \sum_{i=1}^{N} w_t^{(i)}\delta(\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_t)$, where $\hat{\boldsymbol{\lambda}}_t = \sum_{i=1}^{N} w_t^{(i)}\boldsymbol{\lambda}_t^{(i)}$

Table 4.4: Bootstrap Particle Filtering Pseudo Code for $\boldsymbol{\alpha}_t$ and $\boldsymbol{\lambda}_t$ for the Poisson DSTM (4.3) for known $\boldsymbol{\mu}$, $\boldsymbol{B}$ and $\boldsymbol{\Sigma}_\eta$.

4.4.2 Updating $\boldsymbol{B}|\boldsymbol{\alpha}_t, \boldsymbol{\alpha}_{t-1}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}_\eta$

The prior and posterior hierarchy under the Spike and Slab for $\boldsymbol{B}$ remain the same as in Chapter 3 since $\boldsymbol{B}$ is a part of the linear states $\boldsymbol{\alpha}_t$ which remains the same. The posterior updating will involve separate Gibbs steps for the estimation of $\boldsymbol{B}$ combined with particle filtering. Specifically, in each Gibbs iteration a new particle filtering algorithm will be run in order to sample randomly a particle vector α_t and use it in the posterior of $\boldsymbol{B}$. This procedure is called conditional particle filtering with ancestor sampling (CPF-AS). The method was firstly introduced from the PMCMC methods by Andrieu et al. (2010), and was evolved by Lindsten et al. (2014).

Specifically, the CPF-AS is similar to the regular particle filter as shown in Table 4.4. The extra step is that in each Gibbs sampling iteration the N-th particle trajectory is a particle vector $\boldsymbol{\alpha}_{1:t}[m]$ which is specified a priori where m signifies the $m-th$ iteration of the Gibbs sampling. This means that CPF-AS generates N weighted particle trajectories $\{\boldsymbol{\alpha}^i_{1:T}, w^i_T\}^N_{i=1}$ but under the formulation of the conditional particle filter in Andrieu et al. (2010), one of these trajectories is marked as $\boldsymbol{\alpha}_{1:T}[m]$. By using the ancestor sampling, then the CPF-AS is obtained and the resulting trajectories $\{\boldsymbol{\alpha}^i_{1:T}, w^i_T\}^N_{i=1}$ are still influenced by that selected $\boldsymbol{\alpha}_{1:T}[m]$ (Svensson et al., 2015).

Therefore, an extra iterative step can be included in the algorithm of Table 4.4 where a particle vector $\boldsymbol{\alpha}_{0:T}$ is drawn and through that the associate vector $\boldsymbol{\lambda}_{0:T}$ is calculated. Then, these vectors are used in the conditional posterior distribution of $\boldsymbol{B}$ in order to sample the $m-th$ estimate. The CPF-AS algorithm (Table 4.5) for the model in (4.3) is provided bellow.

Initial step:
Initialise at $m = 1$:
Set $\text{vec}(\boldsymbol{B})[1]$ arbitrarily or through the prior $p(\text{vec}(\boldsymbol{B}))$
Particle Sampling:
For $m = 1, \ldots, M$:
Implement Algorithm 4.4 with $\boldsymbol{B} = \boldsymbol{B}[m-1]$
Draw $\omega \sim C(\{w_T\}_{i=1}^N)$ and output the trajectory $\boldsymbol{\alpha}_{0:T}^{\omega}[m]$
Calculate $\boldsymbol{\lambda}_{0:T}^{\omega}[m] \vert \boldsymbol{\alpha}_{0:T}^{\omega}[m]$
Gibbs sampling step:
Draw $\boldsymbol{B}[m]$ from $\text{vec}(\boldsymbol{B}) \sim p(\text{vec}(\boldsymbol{B}) \vert \boldsymbol{\alpha}_{0:T}^{\omega}[m], \boldsymbol{\lambda}_{0:T}^{\omega}[m])$

Table 4.5: Conditional Particle Filtering Pseudo Code for $\boldsymbol{\alpha}_t$ and $\boldsymbol{\lambda}_t$ and $\boldsymbol{B}$ for the Poisson DSTM (4.3) for known $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}_\eta$.

4.4.3 Updating $\boldsymbol{\Sigma}_\eta | \boldsymbol{\alpha}_t, \boldsymbol{\alpha}_{t-1}, \boldsymbol{B}$ and $\sigma_\epsilon | \boldsymbol{\lambda}_t, \boldsymbol{\alpha}_t, \mu$

According to the preferred structure of $\boldsymbol{\Sigma}_\eta$, the posterior distribution will be of the same as in Chapter 3. Therefore, similarly, the posterior estimates for the covariance structure of $\boldsymbol{\Sigma}_\eta$ can be derived under the conditional particle filtering algorithm introduced in Table 4.5. Additionally, in the case where we want to incorporate the extra Poisson variability σ_ϵ for modeling observed intensities, under an inverse gamma prior with shape and scale parameters ψ_1 and ψ_2 respectively we can derive:

$$\begin{aligned} p(\sigma_\epsilon | \boldsymbol{\lambda}_t, \boldsymbol{\alpha}_t, \mu) \propto & p(\log(\boldsymbol{\lambda}_t) | \boldsymbol{\alpha}_t, \sigma_\epsilon) \times p(\sigma_\epsilon) \\ = & (2\pi\sigma_\epsilon^2)^{-(n+T)/2} (\sigma_\epsilon^2)^{-\psi_1 - 1} \exp\left(- \frac{\psi_2}{\sigma_\epsilon^2} \right) \\ & \times \exp\left(- \frac{1}{2\sigma_\epsilon} \sum_{t=1}^{T} (\log(\boldsymbol{\lambda} t) - (\boldsymbol{\mu} + \boldsymbol{\Phi}\boldsymbol{\alpha_t}))^\top (\log(\boldsymbol{\lambda} t) - (\boldsymbol{\mu} + \boldsymbol{\Phi}\boldsymbol{\alpha_t})) \right) \\ \propto & (\sigma_\epsilon^2)^{-\psi_1 - \frac{n+T}{2} - 1} \exp\left(- \frac{\boldsymbol{C}/2 + \psi_2}{\sigma_\epsilon^2} \right) \end{aligned} \tag{4.11}$$

with $\boldsymbol{C} = \sum_{t=1}^{T} (\log(\boldsymbol{\lambda}_t) - (\boldsymbol{\mu} + \boldsymbol{\Phi}\boldsymbol{\alpha_t}))^\top (\log(\boldsymbol{\lambda}_t) - (\boldsymbol{\mu} + \boldsymbol{\Phi}\boldsymbol{\alpha_t}))$. Thus, two extra steps are incorporated in the conditional particle filter algorithm of Table 4.5 for the covariance inference and the new algorithm is provided in Table 4.6.

Initial step:
Initialise at $m = 1$:
Set $\text{vec}(\boldsymbol{B})[1]$ arbitrarily or through the prior $p(\text{vec}(\boldsymbol{B}))$
Set $\boldsymbol{\Sigma}_\eta[1]$ arbitrarily or through the prior $p(\boldsymbol{\Sigma}_\eta))$
Particle Sampling:
For $m = 1, \ldots, M$:
Implement Algorithm 4.4 with $\boldsymbol{B} = \boldsymbol{B}[m-1]$
Draw $\omega \sim C(\{w_T\}_{i=1}^N)$ and output the trajectory $\boldsymbol{\alpha}_{0:T}^{\omega}[m]$
Calculate $\boldsymbol{\lambda}_{0:T}^{\omega}[m] \vert \boldsymbol{\alpha}_{0:T}^{\omega}[m]$
Gibbs sampling step:
Draw $\boldsymbol{B}[m]$ from $\text{vec}(\boldsymbol{B}) \sim p(\text{vec}(\boldsymbol{B}) \vert \boldsymbol{\alpha}_{0:T}^{\omega}[m], \boldsymbol{\lambda}_{0:T}^{\omega}[m])$
Draw $\boldsymbol{\Sigma}_\eta \sim p(\boldsymbol{\Sigma}_\eta \vert \boldsymbol{\alpha}_{0:T}^{\omega}[m], \boldsymbol{\lambda}_{1:T}[m])$
If σ_ϵ is assumed, then draw $\sigma_\epsilon \sim p(\sigma_\epsilon \vert \boldsymbol{\lambda}_t, \boldsymbol{\alpha}_t \vert \boldsymbol{\mu})$

Table 4.6: Conditional Particle Filtering Pseudo Code for α_t and $\boldsymbol{\lambda}_t$, $\boldsymbol{B}$, $\boldsymbol{\Sigma}_\eta$ for the Poisson DSTM (4.3) for known $\boldsymbol{\mu}$.

4.4.4 Update spatially varying $\boldsymbol{\mu} \vert \boldsymbol{\alpha}_t, \boldsymbol{\lambda}_t, \boldsymbol{Y}_t$

Wikle (2002) considers a uniform prior $\text{U}[-10, 10]$ for the application on cloud intensity data and infers via Gibbs sampling. However, for a general application someone needs explicit information on the hyperparameter values of μ, otherwise, the estimation should be adaptive. In this thesis three approaches of inference are considered for the estimation of the spatial mean effect $\boldsymbol{\mu}$.

Particle Metropolis-Hastings for Static $\boldsymbol{\mu}$ If we consider the case where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top$ is static, as we are interested into modeling Poisson counts under a Gaussian or Uniform prior, the posterior distribution of $\boldsymbol{\mu}$ is not of a known form. This covers the special case where the overall mean effect is the same in every location as in (4.2). Thus, the use of Metropolis-Hastings steps will have to be included in the particle filtering algorithm. This procedure is called Particle Metropolis-Hastings (PMH) and was developed by Andrieu et al. (2010). Specifically, Particle Metropolis-Hastings is an iterative procedure where in each iteration a particle filter is employed in order to derive an unbiased estimation of the likelihood followed by a Metropolis-Hastings procedure as described in Chapter 3 in order to approximate the posterior distribution of the parameter, which in this case is $p(\boldsymbol{\mu} \vert \boldsymbol{\lambda}_{1:T}, \boldsymbol{\alpha}_{1:T})$. Under a Random

Walk Metropolis-Hastings and by considering a symmetric proposal in each iterative step the likelihood can be estimated as

$$\log \hat{p}_\theta^N(y_{1:t}) = \log \hat{p}_\theta^N(Y_{1:t-1}) + \{w_{max} + \sum_{i=1}^{N} w_t^i - \log N\} \tag{4.12}$$

and therefore can be used in order to calculate the acceptance ratio. Thus, under an appropriate proposal distribution we summarise the algorithm for the Particle Metropolis-Hastings (PMH) in Table 4.7.

Initial step:
Initialise at $m = 1$:
Set $\boldsymbol{\mu}$ arbitrarily or via $p(\boldsymbol{\mu})$
Run Algorithm 1 and estimate the likelihood (4.12)
Particle Sampling:
For $m = 1, \ldots, M$:
Implement Algorithm 4.4 with proposed value $\boldsymbol{\mu}_{prop}\|\boldsymbol{\mu}[m-1]$
Extract the likelihood (4.12) and the estimates $\boldsymbol{\alpha}_{1:T}$, $\lambda_{1:T}$
Metropolis-Hastings Acceptance step:
Calculate the log-likelihood difference between $\boldsymbol{\mu}_{prop}$ and $\boldsymbol{\mu}[m-1]$
Sample $u \sim \mathrm{U}(0,1)$:
if $u <$ acceptance probability then update $\boldsymbol{\mu}[m] = \boldsymbol{\mu}_{prop}$
else $\boldsymbol{\mu}[m] = \boldsymbol{\mu}[m-1]$

Table 4.7: Particle Metropolis-Hastings under static parameter estimation Pseudo Code for $\boldsymbol{\alpha}_t$, $\boldsymbol{\lambda}_t$, and $\boldsymbol{\mu}$ for the Poisson DSTM (4.3).

Initial step:
Simulate N particles $\boldsymbol{\alpha}_0^{(1)}, \ldots, \boldsymbol{\alpha}_0^{(N)}$ from $p(\boldsymbol{\alpha}_0)$
Sample $\boldsymbol{\mu}^{(0)}$ from $p(\boldsymbol{\mu})$
Calculate $\boldsymbol{\lambda}_0^{(1)}, \ldots, \boldsymbol{\lambda}_0^{(N)}$
Set $w_0^{(i)} 1/N$, $i = 1, \ldots, N$
Particle Sampling:
For $t = 1, \ldots, T$:
Sample $\boldsymbol{\mu} \sim f_t(\boldsymbol{\mu} \vert \boldsymbol{\alpha}_{0:t-1}^{(i)}, \boldsymbol{Y}_{1:t})$
Sample $\boldsymbol{\alpha}_t^{(1)}, \ldots, \boldsymbol{\alpha}_t^{(N)}$ from importance function $g(\boldsymbol{\alpha}_t \vert \boldsymbol{\alpha}_{t-1}^{(i)}, \boldsymbol{Y}_t, \boldsymbol{\mu})$
Calculate $\boldsymbol{\lambda}_t^{(i)} = \exp(\boldsymbol{\mu} + \boldsymbol{\Phi}\boldsymbol{\alpha}_t^{(i)})$
Calculate the weights $\tilde{w}_t^{(i)}$ from (4.9)
Normalise the weights $w_t^{(i)} = \frac{\tilde{w}_t^{(i)}}{\sum_{i=1}^N \tilde{w}_t^{(i)}}$
Resampling step: Multinomial Resampling
Calculate the effective sample size $\mathrm{N}_{eff} = (\sum_{i=1}^N (w_t^{(i)})^2)^{-1}$
Draw N indices $i_1, \ldots, i_N$ from discrete distribution $P(\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_t^{(i)}) = w_t^{(i)}$
Relabel the sample $\boldsymbol{\alpha}_t^{(i)} = \boldsymbol{\alpha}_t^{(i_j)}$, for $i = 1, 2, \ldots, N$
Relabel the sample $T_t^{(i)} = T(T_{t-1}^{i_j}, \boldsymbol{\alpha}_t^{(i)})$, for $i = 1, 2, \ldots, N$
Update to equal weights by $w_t^{(i)} = 1/N$
Posterior Estimation Approximate the posteriors
$\hat{p}(\boldsymbol{\alpha}_t \vert \boldsymbol{Y}_{1:t}) = \sum_{i=1}^N w_t^{(i)} \delta(\boldsymbol{\alpha}_t - \hat{\boldsymbol{\alpha}}_t)$, where $\hat{\boldsymbol{\alpha}}_t = \sum_{i=1}^N w_t^{(i)} \boldsymbol{\alpha}_t^{(i)}$
$\hat{p}(\boldsymbol{\lambda}_t \vert Y_{1:t}, \boldsymbol{\alpha}_t) = \sum_{i=1}^N w_t^{(i)} \delta(\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_t)$, where $\hat{\boldsymbol{\lambda}}_t = \sum_{i=1}^N w_t^{(i)} \boldsymbol{\lambda}_t^{(i)}$

Table 4.8: Bootstrap Particle Filtering under static parameter estimation Pseudo Code for $\boldsymbol{\alpha}_t$ and $\boldsymbol{\lambda}_t$ and $\boldsymbol{\mu}$ for the Poisson DSTM (4.3) for known $\boldsymbol{B}$ and Σ_η.

Static Parameter Alternatively, a static parameter estimation under particle filtering (Storvik, 2002)) can be considered where sufficient statistics based on the observations $\boldsymbol{Y}$ and the states $\boldsymbol{\alpha}$, such as, $\boldsymbol{T}_t(\boldsymbol{Y}_{1:t}, \boldsymbol{\alpha}_{1:t})$ are used to recursively update the posterior distribution of $\boldsymbol{\mu}$. Specifically, at each previous iterative point $t-1$, a new trajectory vector $\boldsymbol{\alpha}_{t-1}$ is available from the posterior distribution $p(\boldsymbol{\alpha}_{1:t-1} \vert \boldsymbol{Y}_{1:t-1})$, however, we then use the additional step of simulating $\boldsymbol{\mu}$ based on the sufficient statistics $\boldsymbol{T}_{t-1}$, i.e., $\boldsymbol{\mu} \sim p(\boldsymbol{\mu} \vert \boldsymbol{T}_{t-1})$. As the sampling of $\boldsymbol{\mu}$ at each time point is not dependent on the values simulated in previous time points, this approach can improve the efficiency and flexibility for the current framework and produces better proposal distributions, state and parameter estimates. Thus, by considering a prior for $\boldsymbol{\mu}$, the particle filtering

algorithm under static parameter estimation is summarised on Table 4.8. Thus, if we treat the rest of parameters unknown and combine the algorithm in Table 4.4 and Table 4.9 we bring the inferential procedure into a Conditional Particle Filtering under static parameter estimation framework.

Initial step:
Initialise at $m = 1$:
Set $\text{vec}(\boldsymbol{B})[1]$ arbitrarily or through the prior $p(\text{vec}(\boldsymbol{B}))$
Set $\boldsymbol{\Sigma}_\eta[1]$ arbitrarily or through the prior $p(\boldsymbol{\Sigma}_\eta)$
Particle Sampling:
For $m = 1, \ldots, M$:
Implement Algorithm 4.8 with $\boldsymbol{B} = B[m-1]$
Draw $\omega \sim \text{C}(\{w_T\}_{i=1}^N)$ and output the trajectory $\alpha_{0:T}^{\omega}[m]$
Calculate $\lambda_{0:T}^{\omega}[m] \vert \alpha_{0:T}^{\omega}[m]$
Gibbs sampling step:
Draw $\boldsymbol{B}[m]$ from $\text{vec}(\boldsymbol{B}) \sim p(\text{vec}(\boldsymbol{B}) \vert \alpha_{0:T}^{\omega}[m], \lambda_{0:T}^{\omega}[m])$
Draw $\boldsymbol{\Sigma}_\eta \sim p(\boldsymbol{\Sigma}_\eta \vert \alpha_{0:T}^{\omega}[m], \boldsymbol{\lambda}_{1:T}[m])$
If σ_ϵ is assumed, then draw $\sigma_\epsilon \sim p(\sigma_\epsilon \vert \boldsymbol{\lambda}_t, \boldsymbol{\alpha}_t \vert \boldsymbol{\mu})$

Table 4.9: Conditional Particle Filtering under static parameter estimation Pseudo Code for α_t and $\boldsymbol{\lambda}_t$, $\boldsymbol{B}$ and $\boldsymbol{\mu}$ for the Poisson DSTM (4.3).

4.4.5 Updating the autoregressive mean effect $\boldsymbol{\mu}_t$

Finally, if we treat the mean effect as a temporal varying vector $\boldsymbol{\mu}_t = (\mu_{1t}, \ldots, \mu_{nt})^\top$ as in (4.4), then we will consider an analogous estimation to $\boldsymbol{\alpha}_t$ and $\boldsymbol{\lambda}_t$. Specifically, each time varying vector $\boldsymbol{\mu}_t$ will be included into the particle filtering algorithm and sampled with the same weights w_t as $\boldsymbol{\alpha}_t$. In this thesis we will consider again the importance function $q(\boldsymbol{\mu}_t^{(i)} \vert \boldsymbol{\mu}_{t-1}^{(i)}) = p(\boldsymbol{\mu}_t^{(i)} \vert \boldsymbol{\mu}_{t-1}^{(i)})$ which gives us the bootstrap particle filter algorithm for model (4.5) in Table 4.10.

Initial step:
Simulate N particles $\boldsymbol{\alpha}_0^{(1)}, \ldots, \boldsymbol{\alpha}_0^{(N)}$ from $p(\boldsymbol{\alpha}_0)$
Simulate N particles $\boldsymbol{\mu}_0^{(1)}, \ldots, \boldsymbol{\mu}_0^{(N)}$ from $p(\boldsymbol{\mu}_0)$
Calculate $\boldsymbol{\lambda}_0^{(1)}, \ldots, \boldsymbol{\lambda}_0^{(N)}$
Set $w_0^{(i)} 1/N$, $i = 1, \ldots, N$
Particle Sampling:
For $t = 1, \ldots, T$:
Sample $\boldsymbol{\alpha}_t^{(1)}, \ldots, \boldsymbol{\alpha}_t^{(N)}$ from the importance function $g_1(\boldsymbol{\alpha}_t \vert \boldsymbol{\alpha}_{t-1}^{(i)}, Y_t)$
Sample $\boldsymbol{\mu}_t^{(1)}, \ldots, \boldsymbol{\mu}_t^{(N)}$ from the importance function $g_2(\boldsymbol{\mu}_t \vert \boldsymbol{\mu}_{t-1}^{(i)}, Y_t)$
Calculate $\boldsymbol{\lambda}_t^{(i)} = \exp(\boldsymbol{\mu}_t + \boldsymbol{\Phi}\boldsymbol{\alpha}_t^{(i)})$
Calculate the weights $\tilde{w}_t^{(i)}$ from (4.9)
Normalise the weights $w_t^{(i)} = \frac{\tilde{w}_t^{(i)}}{\sum_{i=1}^N \tilde{w}_t^{(i)}}$
Resampling step: Multinomial Resampling
Calculate the effective sample size $N_{eff} = (\sum_{i=1}^N (w_t^{(i)})^2)^{-1}$
Draw N indices $i_1, \ldots, i_N$ from the discrete distribution
$P((\boldsymbol{\alpha}_t, \boldsymbol{\mu}_t)^\top = (\boldsymbol{\alpha}_t^{(i)}, \boldsymbol{\mu}_t^{(i)})^\top) = w_t^{(i)}$
Relabel the sample $\boldsymbol{\alpha}_t^{(i)} = \boldsymbol{\alpha}_t^{(i_j)}$, for $i = 1, 2, \ldots, N$
Relabel the sample $\boldsymbol{\mu}_t^{(i)} = \boldsymbol{\mu}_t^{(i_j)}$, for $i = 1, 2, \ldots, N$
Update to equal weights by $w_t^{(i)} = 1/N$
Posterior Estimation Approximate the posteriors
$\hat{p}(\boldsymbol{\alpha}_t \vert \boldsymbol{Y}_{1:t}) = \sum_{i=1}^N w_t^{(i)} \delta(\boldsymbol{\alpha}_t - \hat{\boldsymbol{\alpha}}_t)$, where $\hat{\boldsymbol{\alpha}}_t = \sum_{i=1}^N w_t^{(i)} \boldsymbol{\alpha}_t^{(i)}$
$\hat{p}(\boldsymbol{\mu}_t \vert \boldsymbol{Y}_{1:t}) = \sum_{i=1}^N w_t^{(i)} \delta(\boldsymbol{\mu}_t - \hat{\boldsymbol{\mu}}_t)$, where $\hat{\boldsymbol{\mu}}_t = \sum_{i=1}^N w_t^{(i)} \boldsymbol{\mu}_t^{(i)}$
$\hat{p}(\boldsymbol{\lambda}_t \vert \boldsymbol{Y}_{1:t}, \boldsymbol{\alpha}_t) = \sum_{i=1}^N w_t^{(i)} \delta(\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_t)$, where $\hat{\boldsymbol{\lambda}}_t = \sum_{i=1}^N w_t^{(i)} \boldsymbol{\lambda}_t^{(i)}$

Table 4.10: Bootstrap Particle Filtering Pseudo Code for $\boldsymbol{\alpha}_t$, $\boldsymbol{\lambda}_t$ and $\boldsymbol{\mu}_t$ for the Poisson DSTM (4.5) for known $\boldsymbol{B}$, $\boldsymbol{\Sigma}_\eta$, $\boldsymbol{\Psi}$ and σ_ζ.

4.4.6 Updating $\sigma_\zeta | \boldsymbol{\mu}_t, \boldsymbol{\Psi}$

Considering the inference of the variance component σ_ζ, by considering an inverse gamma prior, i.e., $\sigma_\zeta \sim \text{IG}(\omega_1, \omega_2)$ the full conditional posterior can be calculated as a

product of the likelihoood in (4.15) and the prior $p(\sigma_\zeta|\omega_1,\omega_2)$ such as

$$\begin{aligned} p(\sigma_\zeta|\boldsymbol{\mu}_{1:T},\psi_s) &= L(\boldsymbol{\mu}_{1:T}|\Psi,\sigma_\zeta)\dot{p}(\sigma_\zeta|\omega_1,\omega_2) \\ &= (2\pi\sigma_\zeta^2)^{-\frac{T+n-1}{2}}\exp\Big(-\frac{1}{2\sigma_\zeta}\sum_{t=2}^{T}\sum_{s=1}^{n}(\mu_{s,t}-\psi_s\mu_{s,t})^2\Big) \\ &\times (\sigma_\zeta^2)^{-\omega_1-1}\exp\big(-\omega_2/\sigma_\zeta^2\big) \\ &\propto (\sigma_\zeta^2)^{-(\frac{T+n-1}{2}+\omega_1+1)}\exp\Big(-\frac{1}{\sigma_\zeta^2}(\frac{C}{2}+\omega_2)\Big) \end{aligned} \tag{4.13}$$

with $C=\sum_{t=2}^{T}\sum_{s=1}^{n}(\mu_{s,t}-\psi_s\mu_{s,t})^2$ and is the form of an inverse gamma distribution with shape and scale parameters $\omega_1+T/2$ and $C/2+\omega_2$ respectively. Therefore, from (4.13) we conclude that the conditional posterior of the variance component on the autoregressive structure (4.4) is distributed as $\sigma_\zeta^2|\boldsymbol{\mu}_{1:T},\Psi\sim\text{IG}(\omega_1+T/2,C/2+\omega_2)$.

4.4.7 Updating $\boldsymbol{\Psi}|\boldsymbol{\mu},\boldsymbol{\alpha}_t,\sigma_\zeta,\boldsymbol{\lambda}_t,\boldsymbol{Y}_t$

Sampling through CPF-AS The autocorrelation matrix $\boldsymbol{\Psi}$ will be estimated by Gibbs sampling steps in the CPF-AS framework. Specifically, since each ψ_s indicates the autocorrelation between the mean effect at location s between time point t and $t-1$, that means that is bound to own values that lie in the interval $[-1,1]$. Thus, a reasonable prior to be considered is either a uniform prior, i.e., $\psi_s\sim\text{U}[-1,1]$ or a truncated normal distribution with $a=-1$ and $b=1$ being the lower and upper truncation points respectively.

We assume a truncated normal prior $p(\psi_s)$ for each location, i.e., $\psi_s\sim\text{N}_{[-1,1]}(\psi_0,c_0)$. The likelihood for each location $L(\mu_{s,1:T}|\psi_s,\sigma_\zeta)$ is written as

$$\begin{aligned} L(\mu_{s,1:T}|\psi_s,\sigma_\zeta) &= \prod_{t=2}^{T}\frac{1}{\sqrt{2\pi\sigma_\zeta^2}}\exp\Big(-\frac{1}{2\sigma_\zeta}(\mu_{s,t}-\psi_s\mu_{s,t})^2\Big) \\ &= (2\pi\sigma_\zeta^2)^{\frac{T-1}{2}}\exp\Big(-\frac{1}{2\sigma_\zeta}\sum_{t=2}^{T}(\mu_{s,t}-\psi_s\mu_{s,t})^2\Big) \end{aligned} \tag{4.14}$$

while the likelihood for all n locations can be written as a product of n independent normals for each location s, i.e.,

$$\begin{aligned} L(\boldsymbol{\mu}_{1:T}|\boldsymbol{\Psi},\sigma_\zeta) &= \prod_{s=1}^{n}\prod_{t=2}^{T}\frac{1}{\sqrt{2\pi\sigma_\zeta^2}}\exp\Big(-\frac{1}{2\sigma_\zeta}(\mu_{s,t}-\psi_s\mu_{s,t})^2\Big) \\ &= (2\pi\sigma_\zeta^2)^{\frac{T+n-1}{2}}\exp\Big(-\frac{1}{2\sigma_\zeta}\sum_{t=2}^{T}\sum_{s=1}^{n}(\mu_{s,t}-\psi_s\mu_{s,t})^2\Big) \end{aligned} \tag{4.15}$$

The conditional posterior of each ψ_s is the product of the likelihood in (4.14) and the truncated normal prior $p(\psi_s|\psi_0,c_0)$, i.e,

$$\begin{aligned} p(\psi_s|\mu_{s,1:T},\sigma_\zeta) &= L(\mu_{s,1:T}|\psi_s,\sigma_\zeta)\dot{p}(\psi_s|\psi_0,c_0) \\ &= (2\pi\sigma_\zeta^2)^{\frac{T-1}{2}}\exp\Big(-\frac{1}{2\sigma_\zeta}\sum_{t=2}^{T}(\mu_{s,t}-\psi_s\mu_{s,t})^2\Big) \\ &\times \mathbf{1}_{\psi_s\in[-1,1]}(2\pi c_0^2)^{-1/2}\exp\Big(-\frac{1}{2c_0^2}(\psi_s-\psi_0)^2\Big) \\ &\propto \mathbf{1}_{\psi_s\in[-1,1]}\exp\Big(-\frac{1}{2}\Big[\frac{\sum_{t=2}^{T}(\mu_{s,t}-\psi_s\mu_{s,t})^2}{\sigma_\zeta^2}+\frac{(\psi_s-\psi_0)^2}{c_0^2}\Big]\Big) \\ &= \exp\Big(-\frac{1}{2}\Big[\frac{\sum_{t=2}^{T}\mu_{s,t}^2}{\sigma_\zeta}-\frac{2\psi_s\sum_{t=2}^{T}\mu_{s,t}\mu_{s,t-1}}{\sigma_\zeta}+\frac{\psi_s^2\sum_{t=2}^{T}\mu_{t-1}^2}{\sigma_\zeta^2}\Big]\Big) \\ &\times \mathbf{1}_{\psi_s\in[-1,1]}\exp\Big(-\frac{1}{2}\Big[\frac{\psi_s^2}{c_0^2}-\frac{2\psi_s\psi_0}{c_0^2}+\frac{\psi_0^2}{c_0^2}\Big]\Big) \end{aligned}$$

which then is expanded as:

$$\begin{aligned} p(\psi_s|\mu_{s,1:T},\sigma_\zeta) &\propto \mathbf{1}_{\psi_s\in[-1,1]}\exp\Big(-\frac{1}{2}\Big[\psi_s^2\big(\frac{\sum_{t=2}^{T}\mu_{t-1}^2}{\sigma_\zeta}+\frac{1}{c_0^2}\big) \\ &- 2\psi_s\big(\frac{\sum_{t=2}^{T}\mu_{s,t}\mu_{s,t-1}}{\sigma_\zeta^2}+\frac{\psi_0}{c_0^2}\big)\Big]\Big) \\ &= \mathbf{1}_{\psi_s\in[-1,1]}\exp\Big(-\frac{1}{2}\Big[\psi_s^2\big(\frac{\sum_{t=2}^{T}\mu_{t-1}^2}{\sigma_\zeta}+\frac{1}{c_0^2}\big) \\ &- 2\psi_s\big(\frac{\sum_{t=2}^{T}\mu_{s,t}\mu_{s,t-1}}{\sigma_\zeta^2}+\frac{\psi_0}{c_0^2}\big)\Big]\Big) \end{aligned} \tag{4.16}$$

and gives us a truncated normal conditional posterior for ψ_s at $[-1,1]$ with $m_\psi = (\sum_{t=2}^{T} \mu_{s,t}\mu_{s,t-1}/\sigma_\zeta^2 + \psi_0/c_0^2)c_\psi^2$ and $c_\psi^2 = (\sum_{t=2}^{T} \mu_{s,t-1}^2/\sigma_\zeta^2 + 1/c_0^2)^{-1}$ being the posterior mean and variance respectively. In the case where $\psi_s \sim \mathrm{U}[-1,1]$, the conditional posterior (4.16) is simplified as a truncated normal at $[-1,1]$ with $m_\psi = (\sum_{t=2}^{T} \mu_{s,t}\mu_{s,t-1}/\sigma_\zeta^2)c_\psi^2$ and $c_\psi^2 = (\sum_{t=2}^{T} \mu_{s,t-1}^2/\sigma_\zeta^2)^{-1}$ being the posterior mean and variance respectively. Considering the derivation of (4.16) and (4.13), the CPF-AS algorithm for all parameters for the model in (4.5) is summarised in Table 4.11.

Static parameter estimation As the number of parameters in $\boldsymbol{\Psi}$ increases with the number of locations, considering CPF-AS inference might be inefficient. Therefore in high dimensional cases, it can be treated as static parameter for the autoregressive equation (4.4) we can replace it to $\boldsymbol{\mu}$ in the algorithm of Table 4.5. This means that a static parameter vector $\boldsymbol{\psi} = \mathrm{diag}(\boldsymbol{\Psi})$ can be considered with prior distributions being a truncated normal distribution at $[-1,1]$ for each diagonal element ψ_s. This would give us a sampling in each iterative step for all parameters based on the sufficient statistics of $\boldsymbol{\mu}_{0:t-1}$ and $\boldsymbol{Y}_{1:t}$, i.e., $\psi_s \sim f_1(\psi_s|\boldsymbol{\mu}_{0:t-1}^{(i)}, \boldsymbol{Y}_{1:t})$. Thus, in Table 4.12 the Bootstrap Particle Filter under static parameter estimation for $\boldsymbol{\psi}$ is provided. By combining the updating of the rest of the parameters, consequently we resort to the CPF-AS under static parameter estimation algorithm in Table 4.13 for the model (4.5).

Initial step:
Initialise at $m = 1$:
Set $\text{vec}(\boldsymbol{B})[1]$ arbitrarily or through the prior $p(\text{vec}(\boldsymbol{B}))$
Set $\boldsymbol{\Sigma}_\eta[1]$ arbitrarily or through the prior $p(\boldsymbol{\Sigma}_\eta)$
Set $\boldsymbol{\psi}[1]$ arbitrarily or through the prior $p(\boldsymbol{\psi})$
Set $\sigma_\zeta[1]$ arbitrarily or through the prior $p(\sigma)\zeta)$
Particle Sampling:
For $m = 1, \ldots, M$:
Implement Algorithm 4.10 with:
$\boldsymbol{B} = \boldsymbol{B}[m-1]$, $\boldsymbol{\Sigma}_\eta = \boldsymbol{\Sigma}_\eta[m-1]$
$\boldsymbol{\psi} = \boldsymbol{\psi}[m-1]$, $\sigma_\zeta[m] = \sigma_\zeta[m-1]$
(If σ_ϵ is assumed $\sigma_\epsilon[m] = \sigma_\epsilon[m-1]$
Draw $\omega \sim \text{C}(\{w_T\}_{i=1}^N)$ and output the trajectories:
$\boldsymbol{\alpha}_{0:T}^\omega[m]$, $\boldsymbol{\mu}_{0:T}^\omega[m]$
Calculate $\boldsymbol{\lambda}_{0:T}^\omega[m] \vert \boldsymbol{\alpha}_{0:T}^\omega[m], \boldsymbol{\mu}_{0:T}^\omega[m]$
Gibbs sampling step:
Draw $\boldsymbol{B}[m]$ from $vec(\boldsymbol{B}) \sim p(\text{vec}(\boldsymbol{B}) \vert \boldsymbol{\alpha}_{0:T}^\omega[m], \boldsymbol{\lambda}_{0:T}^\omega[m], \boldsymbol{\mu}_{0:T}^\omega[m])$
Draw $\boldsymbol{\Sigma}_\eta[m] \sim p(\boldsymbol{\Sigma}_\eta \vert \boldsymbol{\alpha}_{0:T}^\omega[m], \boldsymbol{\lambda}_{1:T}[m], \boldsymbol{\mu}_{0:T}^\omega[m])$
Draw $\boldsymbol{\psi}_\eta[m] \sim p(\boldsymbol{\Psi} \vert \boldsymbol{\alpha}_{0:T}^\omega[m], \boldsymbol{\lambda}_{1:T}[m], \boldsymbol{\mu}_{0:T}^\omega[m])$
Draw $\sigma_\zeta[m] \sim p(\sigma_\zeta \vert, \boldsymbol{\Psi}, \boldsymbol{\alpha}_{0:T}^\omega[m], \boldsymbol{\lambda}_{1:T}[m], \boldsymbol{\mu}_{0:T}^\omega[m])$
(If σ_ϵ is assumed, then draw $\sigma_\epsilon[m] \sim p(\sigma_\epsilon \vert \boldsymbol{\alpha}_{0:T}^\omega[m], \boldsymbol{\lambda}_{1:T}[m], \boldsymbol{\mu}_{0:T}^\omega[m]))$

Table 4.11: Conditional Particle Filtering Pseudo Code for all parameters in the Poisson DSTM (4.5).

Initial step:
Simulate N particles $\boldsymbol{\alpha}_0^{(1)}, \ldots, \boldsymbol{\alpha}_0^{(N)}$ from $p(\alpha_0)$
Simulate N particles $\boldsymbol{\mu}_0^{(1)}, \ldots, \boldsymbol{\mu}_0^{(N)}$ from $p(\boldsymbol{\mu}_0)$
Sample $\boldsymbol{\psi}^{(0)}$ from $p(\boldsymbol{\psi})$
Calculate $\boldsymbol{\lambda}_0^{(1)}, \ldots, \boldsymbol{\lambda}_0^{(N)}$
Set $w_0^{(i)} 1/N$, $i = 1, \ldots, N$
Particle Sampling:
For $t = 1, \ldots, T$:
Sample $\boldsymbol{\psi} \sim f_t(\boldsymbol{\psi} \vert \boldsymbol{\alpha}_{0:t-1}^{(i)}, \boldsymbol{\mu}_{0:t-1}^{(i)}, \boldsymbol{Y}_{1:t})$
Sample $\boldsymbol{\mu}_t^{(1)}, \ldots, \boldsymbol{\mu}_t^{(N)}$ from the importance function $g_1(\boldsymbol{\mu}_t \vert \boldsymbol{\mu}_{t-1}^{(i)}, \boldsymbol{Y}_t, \boldsymbol{\psi}, \sigma_\zeta)$
Sample $\boldsymbol{\alpha}_t^{(1)}, \ldots, \boldsymbol{\alpha}_t^{(N)}$ from the importance function $g_2(\boldsymbol{\alpha}_t \vert \boldsymbol{\alpha}_{t-1}^{(i)}, \boldsymbol{Y}_t, \boldsymbol{\mu}_t^{(i)})$
Calculate $\boldsymbol{\lambda}_t^{(i)} = \exp(\boldsymbol{\mu}_t^{(i)} + \boldsymbol{\Phi}\boldsymbol{\alpha}_t^{(i)})$
Calculate the weights $\tilde{w}_t^{(i)}$ from (4.9)
Normalise the weights $w_t^{(i)} = \frac{\tilde{w}_t^{(i)}}{\sum_{i=1}^N \tilde{w}_t^{(i)}}$
Resampling step: Multinomial Resampling
Calculate the effective sample size $\mathrm{N}_{eff} = (\sum_{i=1}^N (w_t^{(i)})^2)^{-1}$
Draw N indices $i_1, \ldots, i_N$ from the discrete distribution
$P((\boldsymbol{\alpha}_t, \boldsymbol{\mu}_t)^\top = (\boldsymbol{\alpha}_t^{(i)}, \boldsymbol{\mu}_t^{(i)})^\top) = w_t^{(i)}$
Relabel the sample $\boldsymbol{\alpha}_t^{(i)} = \boldsymbol{\alpha}_t^{(i_j)}$, for $i = 1, 2, \ldots, N$
Relabel the sample $\boldsymbol{\mu}_t^{(i)} = \boldsymbol{\mu}_t^{(i_j)}$, for $i = 1, 2, \ldots, N$
Relabel the sample $T_t^{(i)} = T(T_{t-1}^{i_j}, \boldsymbol{\alpha}_t^{(i)}, \boldsymbol{\mu}_t^{(i)})$, for $i = 1, 2, \ldots, N$
Update to equal weights by $w_t^{(i)} = 1/N$
Posterior Estimation Approximate the posteriors
$\hat{p}(\boldsymbol{\alpha}_t \vert Y_{1:t}) = \sum_{i=1}^N w_t^{(i)} \delta(\boldsymbol{\alpha}_t - \hat{\boldsymbol{\alpha}}_t)$, where $\hat{\boldsymbol{\alpha}}_t = \sum_{i=1}^N w_t^{(i)} \boldsymbol{\alpha}_t^{(i)}$
$\hat{p}(\boldsymbol{\mu}_t \vert \boldsymbol{Y}_{1:t}) = \sum_{i=1}^N w_t^{(i)} \delta(\boldsymbol{\mu}_t - \hat{\boldsymbol{\mu}}_t)$, where $\hat{\boldsymbol{\mu}}_t = \sum_{i=1}^N w_t^{(i)} \boldsymbol{\mu}_t^{(i)}$
$\hat{p}(\boldsymbol{\lambda}_t \vert \boldsymbol{Y}_{1:t}, \boldsymbol{\alpha}_t) = \sum_{i=1}^N w_t^{(i)} \delta(\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_t)$, where $\hat{\lambda}_t = \sum_{i=1}^N w_t^{(i)} \boldsymbol{\lambda}_t^{(i)}$

Table 4.12: Bootstrap Particle Filtering under static parameter estimation Pseudo Code for $\boldsymbol{\alpha}_t$, $\boldsymbol{\lambda}_t$ and $\boldsymbol{\mu}_t$ for the Poisson DSTM (4.5) for known $\boldsymbol{B}$, $\boldsymbol{\Sigma}_\eta$ and σ_ζ.

Initial step:
Initialise at $m = 1$:
Set $\text{vec}(\boldsymbol{B})[1]$, $\boldsymbol{\Sigma}$ arbitrarily or through the prior $p(\text{vec}(\boldsymbol{B}))$
Set $\boldsymbol{\Sigma}_\eta[1]$ arbitrarily or through the prior $p(\boldsymbol{\Sigma}_\eta)$
Set $\sigma_\zeta[1]$ arbitrarily or through the prior $p(\sigma_\zeta)$
Particle Sampling:
For $m = 1, \ldots, M$:
Implement Algorithm 4.12 with:
$\boldsymbol{B} = \boldsymbol{B}[m-1]$, $\boldsymbol{\Sigma}_\eta = \boldsymbol{\Sigma}_\eta[m-1]$, $\sigma_\zeta = \sigma_\zeta[m-1]$
(If σ_ϵ is assumed $\sigma_\epsilon[m] = \sigma_\epsilon[m-1]$
Draw $\omega \sim \text{C}(\{w_T\}_{i=1}^N)$ and output the trajectories:
$\boldsymbol{\alpha}_{0:T}^\omega[m]$, $\boldsymbol{\mu}_{0:T}^\omega[m]$
Calculate $\boldsymbol{\lambda}_{0:T}^\omega[m] \vert \boldsymbol{\alpha}_{0:T}^\omega[m], \boldsymbol{\mu}_{0:T}^\omega[m]$
Gibbs sampling step:
Draw $\boldsymbol{B}[m]$ from $\text{vec}(\boldsymbol{B}) \sim p(\text{vec}(\boldsymbol{B}) \vert \boldsymbol{\alpha}_{0:T}^\omega[m], \boldsymbol{\lambda}_{0:T}^\omega[m], \boldsymbol{\mu}_{0:T}^\omega[m])$
Draw $\boldsymbol{\Sigma}_\eta[m] \sim p(\boldsymbol{\Sigma}_\eta \vert \boldsymbol{\alpha}_{0:T}^\omega[m], \boldsymbol{\lambda}_{1:T}[m], \boldsymbol{\mu}_{0:T}^\omega[m])$
Draw $\sigma_\zeta[m] \sim p(\sigma_\zeta \vert, \boldsymbol{\Psi}, \boldsymbol{\alpha}_{0:T}^\omega[m], \boldsymbol{\lambda}_{1:T}[m], \boldsymbol{\mu}_{0:T}^\omega[m])$
(If σ_ϵ is assumed, then draw $\sigma_\epsilon[m] \sim p(\sigma_\epsilon \vert \boldsymbol{\alpha}_{0:T}^\omega[m], \boldsymbol{\lambda}_{1:T}[m], \boldsymbol{\mu}_{0:T}^\omega[m]))$

Table 4.13: Conditional Particle Filtering Pseudo Code for all parameters in the Poisson DSTM (4.5) under static parameter estimation for $\boldsymbol{\psi}$.

4.4.8 Theoretical Convergence of CPF-AS and PMH

Andrieu et al. (2010) discuss the theoretical convergence of Particle Markov Chain Monte Carlo (PMCMC) methods and more specifically on PMH. Specifically, general PMCMC methods target the conditional posteriors for any $N \geq 1$ fixed number of particles and they leave the target density invariant. They can as well be considered as standard MCMC updates and will lead to theoretical convergence under mild assumptions. Furthermore, the PMH algorithm that is introduced in Andrieu et al. (2010) leaves the conditional posteriors and the target distributions invariant while the acceptance probability for $N \to \infty$ converges and justifies the Metropolis-Hastings terminology. Additionally, they argue that as in CPF the proposal density is bypassed, the use of conditional Sequential Monte Carlo update is used as a special type of PMCMC and it converges under mild assumptions for $M \to \infty$. Therefore, under this logic, the CPF-AS introduced by Lindsten et al. (2014) as a special case of PMCMC and more

specifically of CPF, show that it retains the theoretical convergence and invariance properties.

4.4.9 Summary of proposed pseudocodes

Modelling static $\boldsymbol{\mu}$ If a static spatially mean effect $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$, i.e., the model in (4.3) is considered, we propose the Conditional Particle Filtering with Ancestor Resampling scheme under static parameter estimation to be considered (Table 4.9). Due to the high parameter space, resulting to an extra Metropolis-Hastings step will increase the computational inefficiency plus the more locations are considered, the higher the parameter space only for PMH will be. Therefore, a static parameter estimation provides us with a much more flexible approach as we will not have to deal with the curse of disconvergence.

Modelling autoregressive $\boldsymbol{\mu}_t$ Analogously, by considering the two approaches for $\boldsymbol{\Psi}$ and the updating procedure of σ^2_ζ, the algorithm in Table 4.13 is the most efficient for (4.5) , i.e., a CPF-AS under static parameter estimation for the diagonal elements of $\boldsymbol{\Psi}$.

4.5 Posterior Predictive Distribution

Similarly to the Gaussian DSTM in Chapter 3, consider the temporal ℓ-steps ahead forecasting at the monitoring locations. Then, by defining $\boldsymbol{D}_t = \{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_t\}$ and have obtained the samples from the Particle Filtering algorithms stated in the previous section, for any positive integer ℓ, the ℓ-step ahead forecast distribution for the model in (4.3) is

$$p(\boldsymbol{Y}_{t+\ell}|\boldsymbol{D}_t) = \int p(\boldsymbol{Y}_{t+\ell}|\boldsymbol{\alpha}_{t+\ell}, \sigma^2_\epsilon, \boldsymbol{\mu}) p(\boldsymbol{\alpha}_{t+\ell}|\boldsymbol{D}_t) d\boldsymbol{\alpha}_{t+\ell} \tag{4.17}$$

and is approximated by

$$\hat{p}(\boldsymbol{Y}_{t+\ell}|\boldsymbol{D}_t) = \sum_{i=1}^{N} p(\boldsymbol{Y}_{t+\ell}|\boldsymbol{\alpha}^{(i)}_{t+\ell}) w^{(i)}_t, \tag{4.18}$$

where by writing recurrently the evolution of $\boldsymbol{\alpha}_{t+\ell}$ as $\boldsymbol{\alpha}_{t+\ell} = (\boldsymbol{\Phi}^\top \boldsymbol{B})^\ell \boldsymbol{\alpha}_t + \sum_{h=1}^{\ell} \eta_{t+h}$, we use $\boldsymbol{\alpha}^{(i)}_{t+\ell} = (\boldsymbol{\Phi}^\top \boldsymbol{B})^\ell \boldsymbol{\alpha}^{(i)}_t$, for $i = 1, \ldots, N$.

Analogously, for the model in (4.5) , the ℓ-step ahead predictive distribution is

$$p(\boldsymbol{Y}_{t+\ell}|\boldsymbol{D}_t) = \int\int p(\boldsymbol{Y}_{t+\ell}|\boldsymbol{\alpha}_{t+\ell}, \sigma^2_\epsilon, \boldsymbol{\mu}_t)p(\boldsymbol{\alpha}_{t+\ell}|\boldsymbol{D}_t)p(\boldsymbol{\mu}_{t+\ell}|\boldsymbol{D}_t)d\boldsymbol{\alpha}_{t+\ell}d\boldsymbol{\mu}_{t+\ell} \tag{4.19}$$

and is approximated by

$$\hat{p}(\boldsymbol{Y}_{t+\ell}|\boldsymbol{D}_t) = \sum_{i=1}^{N} p(\boldsymbol{Y}_{t+\ell}|\boldsymbol{\alpha}^{(i)}_{t+\ell}, \boldsymbol{\mu}^{(i)}_{t+\ell})w^{(i)}_t. \tag{4.20}$$

with the same propagation for $\boldsymbol{\alpha}_{t+\ell}$ but the extra recurrent evolution for $\boldsymbol{\mu}_{t+\ell}$. That is, $\boldsymbol{\mu}_{t+\ell} = \boldsymbol{\Psi}^\ell\boldsymbol{\mu}_t + \sum_{h=1}^{\ell}\zeta_{t+h}$and we use $\boldsymbol{\mu}^{(i)}_{t+\ell} = \boldsymbol{\Psi}^\ell\boldsymbol{\mu}^{(i)}_t$, for $i = 1, \ldots, N$.

Therefore, in each MCMC m-th iteration we acquire the samples of the particle estimates for $\boldsymbol{\alpha}_{t+\ell}$ and for the model in (4.5) the estimates for $\boldsymbol{\mu}_{t+\ell}$ through the sampling of non-dynamic unknown parameters conducted in the previous iteration. During the Particle Filtering steps, we calculate $\boldsymbol{\lambda}^{(i)}_{t+\ell}$ and thus sample $\boldsymbol{Y}_{t+\ell}$ where its distribution is approximated by the summation stated above.

Analogously to the Gaussian case, we would like to conduct spatial interpolation under the new ungauged spatial vector of length ℓ, at an observed time point $t \in T$, i.e., $\tilde{\boldsymbol{Y}}_t = (\tilde{Y}_t(s_1), \ldots, \tilde{Y}_t(s_\ell))^\top$. In this case we shall use the posterior predictive distribution of the link of the intensity process, i.e., $\log(\tilde{\boldsymbol{\lambda}}_t)$ which is written as

$$p(\log(\tilde{\boldsymbol{\lambda}}_t|\boldsymbol{Y})) = \int_{\boldsymbol{\theta}} p(\log(\tilde{\boldsymbol{\lambda}}_t)|\boldsymbol{Y}, \boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{Y})d\boldsymbol{\theta} \tag{4.21}$$

where again $\boldsymbol{\theta}$ is the parameter vector associated to the link equation in (4.3) or (4.5). Thus, the distribution in (4.21) can be written as

$$p(\log(\tilde{\boldsymbol{\lambda}}_t|\boldsymbol{Y})) = \int\cdots\int p(\log(\tilde{\boldsymbol{\lambda}}_t)|\boldsymbol{Y}, \boldsymbol{\alpha}, \boldsymbol{\mu}, \sigma^2_\epsilon)p(\boldsymbol{\alpha}, \boldsymbol{\mu}, \sigma^2_\epsilon|\boldsymbol{Y})d\boldsymbol{\alpha}d\boldsymbol{\mu}d\sigma^2_\epsilon. \tag{4.22}$$

Spatial interpolation of the intensity process $\log(\tilde{\boldsymbol{\lambda}}_t)$ can be derived by considering jointly $\log(\tilde{\boldsymbol{\lambda}}_t), \log(\boldsymbol{\lambda}_t)$ conditional on the unknown parameters and the model structure of (4.3), i.e.,

$$\begin{pmatrix} \log(\boldsymbol{\lambda}_t) \\ \log(\tilde{\boldsymbol{\lambda}}_t) \end{pmatrix} |\boldsymbol{\alpha}, \boldsymbol{\mu}, \sigma^2_\epsilon \sim \mathrm{N}\left(\begin{pmatrix} \boldsymbol{\mu} + \boldsymbol{\Phi}\boldsymbol{\alpha}_t \\ \tilde{\boldsymbol{\mu}} + \tilde{\boldsymbol{\Phi}}\boldsymbol{\alpha}_t \end{pmatrix}, \begin{pmatrix} \sigma^2_\epsilon\boldsymbol{I}_{n\times n} & \boldsymbol{\Sigma}_{\boldsymbol{\lambda}\tilde{\boldsymbol{\lambda}}} \\ \boldsymbol{\Sigma}_{\boldsymbol{\lambda}\tilde{\boldsymbol{\lambda}}} & \sigma^2_\epsilon\boldsymbol{I}_{\ell\times\ell} \end{pmatrix}\right) \tag{4.23}$$

where $\tilde{\boldsymbol{\Phi}}$ is the basis matrix for the new locations and $\tilde{\boldsymbol{\mu}}$ the predicted mean spatial effect of the new ungauged locations and $\boldsymbol{\Sigma}_{\boldsymbol{\lambda}\tilde{\boldsymbol{\lambda}}}$ the covariance between $\log(\tilde{\boldsymbol{\lambda}}_t)$ and $\log(\boldsymbol{\lambda}_t)$. We then have the marginal conditional posterior for $\log(\tilde{\boldsymbol{\lambda}}_t)$, i.e.,

$$\log(\tilde{\boldsymbol{\lambda}}_t)|\log(\boldsymbol{\lambda}_t), \boldsymbol{\alpha}_t, \boldsymbol{\mu}, \sigma_\epsilon^2 \sim \mathrm{N}\left(\tilde{\boldsymbol{\mu}} + \tilde{\boldsymbol{\Phi}}\boldsymbol{\alpha}_t + \boldsymbol{\Sigma}_{\boldsymbol{\lambda}\tilde{\boldsymbol{\lambda}}}(\sigma_\epsilon^2 \boldsymbol{I}_{n\times n})^{-1}(\log(\boldsymbol{\lambda}_t) - \boldsymbol{\mu} - \boldsymbol{\Phi}\boldsymbol{\alpha}_t, \boldsymbol{V}\right) \tag{4.24}$$

where $\boldsymbol{V} = \sigma_\epsilon^2 \boldsymbol{I}_{\ell\times\ell} - \boldsymbol{\Sigma}_{\boldsymbol{\lambda}\tilde{\boldsymbol{\lambda}}}(\sigma_\epsilon^2 \boldsymbol{I}_{n\times n}^2)^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\lambda}\tilde{\boldsymbol{\lambda}}}$. Equivalently, if we consider the modelling approach in (4.5), then we have an autoregressive mean effect and thus

$$\begin{pmatrix} \log(\boldsymbol{\lambda}_t) \\ \log(\tilde{\boldsymbol{\lambda}}_t) \end{pmatrix} |\boldsymbol{\alpha}, \boldsymbol{\mu}_t, \sigma_\epsilon^2 \sim \mathrm{N}\left(\begin{pmatrix} \boldsymbol{\mu}_t + \boldsymbol{\Phi}\boldsymbol{\alpha}_t \\ \tilde{\boldsymbol{\mu}}_t + \tilde{\boldsymbol{\Phi}}\boldsymbol{\alpha}_t \end{pmatrix}, \begin{pmatrix} \sigma_\epsilon^2 \boldsymbol{I}_{n\times n} & \boldsymbol{\Sigma}_{\boldsymbol{\lambda}\tilde{\boldsymbol{\lambda}}} \\ \boldsymbol{\Sigma}_{\boldsymbol{\lambda}\tilde{\boldsymbol{\lambda}}} & \sigma_\epsilon^2 \boldsymbol{I}_{\ell\times\ell} \end{pmatrix}\right) \tag{4.25}$$

with the marginal conditional posterior for $\log(\boldsymbol{\lambda}_t)$ being

$$\log(\tilde{\boldsymbol{\lambda}}_t)|\log(\boldsymbol{\lambda}_t), \boldsymbol{\alpha}_t, \boldsymbol{\mu}_t, \sigma_\epsilon^2 \sim \mathrm{N}\left(\tilde{\boldsymbol{\mu}}_t + \tilde{\boldsymbol{\Phi}}\boldsymbol{\alpha}_t + \boldsymbol{\Sigma}_{\boldsymbol{\lambda}\tilde{\boldsymbol{\lambda}}}(\sigma_\epsilon^2 \boldsymbol{I}_{n\times n})^{-1}(\log(\boldsymbol{\lambda}_t) - \boldsymbol{\mu}_t - \boldsymbol{\Phi}\boldsymbol{\alpha}_t, \boldsymbol{V}\right) \tag{4.26}$$

where $\boldsymbol{V} = \sigma_\epsilon^2 \boldsymbol{I}_{\ell\times\ell} - \boldsymbol{\Sigma}_{\boldsymbol{\lambda}\tilde{\boldsymbol{\lambda}}}(\sigma_\epsilon^2 \boldsymbol{I}_{n\times n}^2)^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\lambda}\tilde{\boldsymbol{\lambda}}}$.

As it was aforementioned in Chapter 3, a spatial interpolation under new locations cannot be directly implemented. The user should resort to different schemes or choose smooth basis functions.

4.6 Simulation Study

Similarly to the Gaussian Reduced-dimension DSTM in Chapter 3, we developed an appropriate simulation by considering a Dimension-reduced Poisson DSTM under wavelet decomposition based on the complete model in (4.5) which is described in 4.5.1. Then we are examining numerous cases where in each case the most important matrix of interest, $\boldsymbol{B}$, is always considered unknown. Due to our limited computational power, a few comments will be in order as in some cases we needed more iterations and thus computational power to reach convergence. Thus, we wish to test our method in three different settings; we achieve this through the design of these respective simulation studies:

- Discontinuity in weight function $w_s(u)$ and static known $\boldsymbol{\mu}$— we wish to show that our method can adapt to discontinuities and we can estimate fairly well the parameters

- No discontinuity in weight function $w_s(u)$ and autoregressive $\boldsymbol{\mu}$ with autocorrelation inference— we wish to show that our method performs well under the proposed autoregressive structure while the static parameters $\boldsymbol{\Psi}$ can be identified
- No discontinuity in weight function $w_s(u)$, autoregressive $\boldsymbol{\mu}$, autocorrelation and covariance inference— we want to see how our methodology can estimate the covariance parameters under the autoregressive structure while the static parameters $\boldsymbol{\Psi}$ can be identified

In section 4.6.1 we introduce the simulation scheme of a Poisson Reduced-Dimension DSTM under wavelet basis decomposition. Furthermore, as in Chapter 3, instead of simulating the matrix $\boldsymbol{B}$ through the Spike and Slab prior, a kernel is chosen for $w_s(u)$ and through that $\boldsymbol{B}$ is calculated through DWT. Additionally, as mentioned above, in sections 4.6.2 to 4.6.4 we conduct inference on the processes' parameters simulated under the simulation scheme in section 4.6.1.

4.6.1 Simulation of a Poisson DSTM under Wavelet decomposition

1 Start by considering a number of equally spaced n locations in an interval $[c_1, c_2] \in D \subset \mathbb{R}$ and T time points, a Wavelet matrix $\Phi_{n\times n}$ and a covariance matrix Σ_η.

- If we want model (4.3), then consider a vector $\boldsymbol{\mu}$.
- If we want model (4.5), then consider an autocorrelation matrix $\boldsymbol{\Psi}$ and a variance σ^2_ζ.

2 Building the weight matrix

- For each of the locations calculate d, where d is the Euclidean distance between the locations $\boldsymbol{s}$
- Choose weight function $\boldsymbol{w}$ (discontinuous or continuous) to calculate the spatial contribution
- Spatial stationary weights:

$$w^*_{i,j} = 0.9 * \frac{w_{i,j}}{\sum_{j=1}^{n} w_{i,j}}$$

3 For $t = 1$

- Calculate the coefficient matrix $\boldsymbol{B} = \boldsymbol{w}^*\boldsymbol{\Phi}^{-1}$
- Initialise $\boldsymbol{\lambda}_1$ (and $\boldsymbol{\mu}_1$ if in model (4.5))
- Calculate $\boldsymbol{\alpha}_1$

  $\boldsymbol{\alpha}_1 = \boldsymbol{\Phi}^\top(\log(\boldsymbol{\lambda}_1) - \boldsymbol{\mu})$ if in model (4.3)

  $\boldsymbol{\alpha}_1 = \boldsymbol{\Phi}^\top(\log(\boldsymbol{\lambda}_1) - \boldsymbol{\mu}_1)$ if in model (4.5)

4 For $t \geq 2$

- $\boldsymbol{\alpha}_t = \boldsymbol{\Phi}^\top \boldsymbol{B}\boldsymbol{\alpha}_{t-1} + \boldsymbol{\Phi}^\top \eta_t$, $\eta_t \sim \mathrm{N}(0, \boldsymbol{\Sigma}_\eta)$
- If in (4.5) then calculate $\boldsymbol{\mu}_t = \boldsymbol{A_s}\boldsymbol{\mu}_{t-1} + \zeta_t$, $\zeta_t \sim \mathrm{N}(0, \sigma_\zeta)$
- Perform IDWT on $\boldsymbol{\lambda}_t$:

  If in model (4.3) $\boldsymbol{\lambda}_t = e^{\boldsymbol{\mu}+\boldsymbol{\Phi}\boldsymbol{\alpha}_t}$

  In if model (4.5) $\boldsymbol{\lambda}_t = e^{\boldsymbol{\mu}_t+\boldsymbol{\Phi}\boldsymbol{\alpha}_t}$
- Simulate i.i.d $\boldsymbol{Y}_t \sim \mathrm{Poi}(\boldsymbol{\lambda}_t)$

A few comments are in order. The higher autocorrelation we have in (4.4), the more difficult the estimation will be. Moreover, if we have high values for $\boldsymbol{\mu}_t$, as $\boldsymbol{\alpha}_t$ have a zero mean (no trend), we will encounter a numerical problem in the calculations of the exponential component for the calculation of $\boldsymbol{\lambda}_t$. Furthermore, the spatial stationarity especially in the Poisson DSTM seemed necessary as the simulation would blow up. Nonetheless, the behaviour of the redistribution kernel is similar to the Gaussian DSTM as under the Poisson case, the choice of it does not affect the observation equation.

4.6.2 Discontinuity in weight function $w_s(u)$ and static known $\boldsymbol{\mu}$

In this simulation we considered the model in (4.3) for $n = 8$ locations and $T = 256$ in the 1-D space $[0, 5]$. Two different kernels were considered for the weight function $w_s(u)$. Specifically, for the locations lying in $[0, 2.5]$ we considered a Gaussian kernel with mean and variance being 1 and 4 respectively, while those lying in $[2.5, 5]$ a Laplace kernel was considered with mean and rate parameters being 0 and 1 respectively, i.e,

$$w_s(u) = \begin{cases} \mathrm{N}(\|s-u\|^2 \,|1, 4) & \text{if } s, u \in [0, 0.25] \\ \mathrm{Laplace}(\|s-u\|^2 \,|0, 1) & \text{if } s, u \in (2.5, 5] \\ 0 & \text{otherwise} \end{cases} \tag{4.27}$$

The spatial mean effect is considered to be static and known with varying values of $\boldsymbol{\mu} = (0.1, 0.2, -0.15, 0.3, -0.4, 0.25, -0.6, -0.05)^\top$. The temporal covariance matrix was set to be of a simple diagonal structure, i.e., $\Sigma_\eta = 0.5 \cdot \boldsymbol{I}$. Finally, the wavelet basis that was used for the decomposition of the weight function $w_s(u)$ was a Haar basis. We conducted the Conditional Particle Filtering with Ancestor Resampling Algorithm 4.2 of Table 4.5 with N = 500 particles and M = 10^4 Gibbs iterations with a burn-in period of i = 5000 which was decided through traceplots and autocorrelation plots as diagnostic criteria.

For the inferential part, we considered again the Spike and Slab hyperparameters $v_0 = 0.05$ and $\omega_1 = 2$, $\omega_2 = 20$ for the point mass and variance components respectively. Comparing the posterior mode of the state processes $\boldsymbol{\alpha}_t$ and the simulated (real) ones in Figure 4.1 it can observed that they are being estimated fairly well. This can be consecutively seen in the estimation of the intensity processes $\boldsymbol{\lambda}_t$ as most of the peaks, or else, the very high intensities are captured nicely, even not exactly at a high level, with only a few exceptions, such as for location 7 (Figure 4.2). Poisson distributed time series are very difficult to be estimated perfectly as there can be jumps

and discontinuities which are very difficult to be detected. The combination of the efficiency of particle filtering algorithms, the wavelet basis and the Spike and Slab prior gave us a successful estimation of those discontinuities.

Furthermore, the spatial wavelet coefficients under our Spike and Slab hierarchy were estimated very well (Figure 4.3), with both the close to zero and non zero real values being very close to the posterior modes. This provided us with a good approximation of the weight function as well, which in the next simulation scenarios we will show that for more unknown parameters our model performs satisfyingly as well.

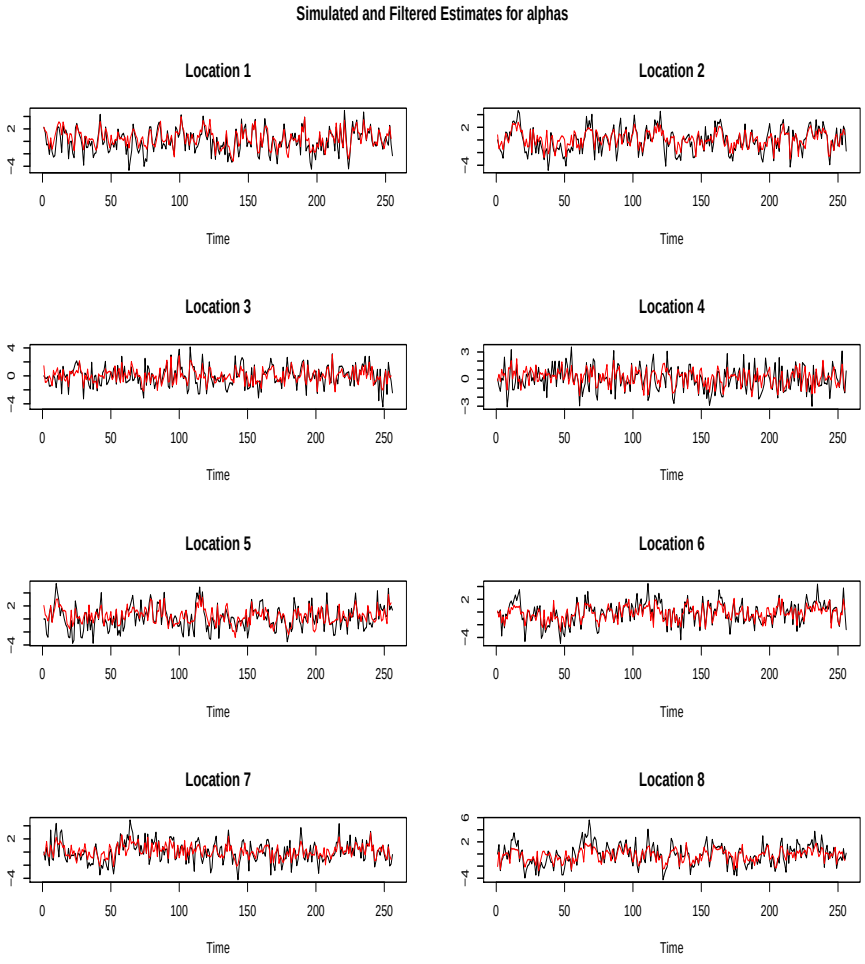

Figure 4.1: Time series plots of simulated states $\boldsymbol{\alpha}_t$ for the locations (black) and the estimated filtered ones (red) for N = 500, M = 10^4, $T = 256$, $n = 8$ under a burn-in period of i = 5000.

Finally, during our inference we were tracking the effective sample size (N_{eff}) under the particle filtering framework. We have observed that we had many particles with very low weights which resulted in resampling. However, the trend shows that throughout the inference it can still give us high values. Figure 4.4 shows us the values of the effective sample size for the last Gibbs iteration of the CPF-AS algorithm.

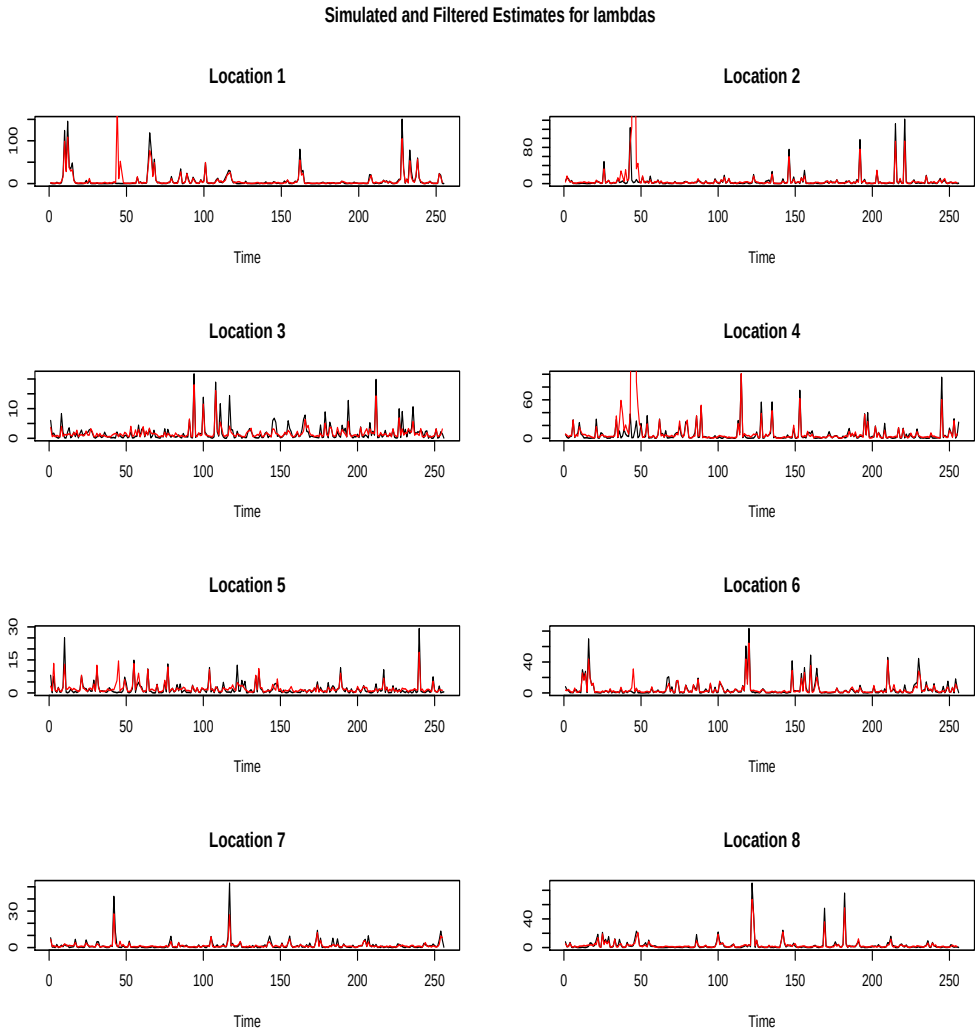

Figure 4.2: Time series plots of the simulated mean intensity process $\boldsymbol{\lambda}_t$ for the locations (black) and the estimated filtered one (red) for N = 500, M = 10^4, $T = 256$, $n = 8$ under a burn-in period of i = 5000.

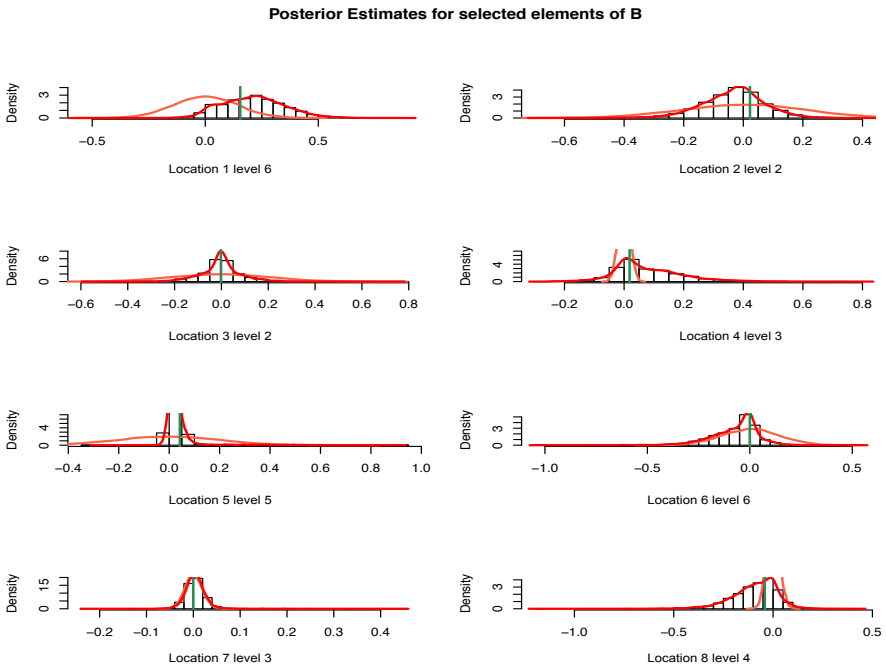

Figure 4.3: Posterior distribution for selected elements of $\boldsymbol{B}$, density posterior estimation is marked with red, the prior distribution with orange and the real value with a green vertical line, for N = 500, M = 10^4, $T = 256$, $n = 8$ under a burn-in period of i = 5000. The hyperparameter values were set to $v_0 = 0.05$, $\omega_1 = 2$ and $\omega_2 = 20$.

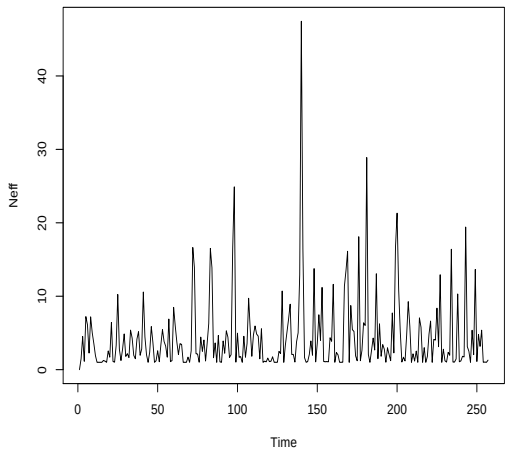

Figure 4.4: Effective sample size of the final Gibbs iteration M under particle filtering for N = 500, M = 10^4, $T = 256$, $n = 8$ under a burn-in period of i = 5000.

4.6.3 No discontinuity in weight function $w_s(u)$ and autoregressive $\boldsymbol{\mu}$ with autocorrelation inference

In this simulation scenario we considered the model in (4.5) with the autoregressive equation (4.4) for $n = 8$ locations and $T = 256$ in the 1-D space $[0, 1]$. Moreover, a Gaussian kernel with mean and variance being 0.5 and 0.2 respectively was considered,i.e.,

$$w_s(u) = \begin{cases} \mathrm{N}((\|s-u\|^2)|0.5, 0.5) & \text{if } s, u \in [0,5] \\ 0 & \text{otherwise} \end{cases} \tag{4.28}$$

which means we would expect all locations to contribute almost equally to the spatial dilation of the rest. The spatial mean effect is considered to be an autoregressive spatially varying process and thus we consider the autocorrelation components $\boldsymbol{\Psi} = (0.1, 0.2, -0.6, 0.35, 0.6, -0.4, 0.8, -0.2)^\top \cdot \boldsymbol{I}$ as static parameters subject to estimation while σ_ζ is considered to be known and equal to 1. Additionally, the temporal covariance matrix is considered known and with a diagonal structure, i.e., $\Sigma_\eta = 2 \cdot \boldsymbol{I}$. Finally, the wavelet basis that was used for the decomposition of the weight function $w_s(u)$ was a Daubechies level 4 basis. We conducted the Conditional Particle Filtering with Ancestor Resampling under static parameter estimation (Table 4.13) by considering the static parameter vector $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_8)^\top$ each with a truncated normal prior

on $[-1, 1]$ centered at zero with variance equal to 1. The number of particles was again N = 500 with the Gibbs iterations being M = 10^4 with a burn-in period of i = 5000 under the same convergence diagnostics.

A few findings are in order. Firstly, our model provided good estimations for the states $\boldsymbol{\alpha}_t$ (Figure 4.5). However, as the intensities $\boldsymbol{\lambda}_t$ are including the autoregressive component $\boldsymbol{\mu}_t$, the estimation of high discontinuities is more challenging. Surprisingly, we still managed to capture most of the discontinuities, especially for the second, fifth and seventh location (Figure 4.6). However, we failed to capture the magnitude of some discontinuities, for instance, in the sixth location we observed a very high peak compared to the rest time points where our model succeeded into estimating that discontinuity. Additionally, the estimation of $\boldsymbol{\mu}_t$ (Figure 4.9) shows us that probably we needed more particles for a better estimation, however, due to limited time and computational hindrances we were unable to run the model for more particles and Gibbs iterations. Furthermore, it is notable that the autocorrelation parameters $\boldsymbol{\psi}$ (Figure 4.10 and Figure 4.11) are estimated on average under their true value, however, we would expect under a static parameter estimation to have a much better estimation as we reach the final points. This can be again justified to the fact that we need more Gibbs iterations for these parameters to reach their true value.

Considering the coefficients $\boldsymbol{B}$ from the posterior densities (Figure 4.7), we encountered the same problem as in Chapter 3, however, the estimations are improved due to not having a signal-to-noise ratio affecting the Spike and Slab inference. This consequently provided us with fair estimates for the weight function $\boldsymbol{w}$ as seen in Figure 4.8 with high concentration around the true values of the weights.

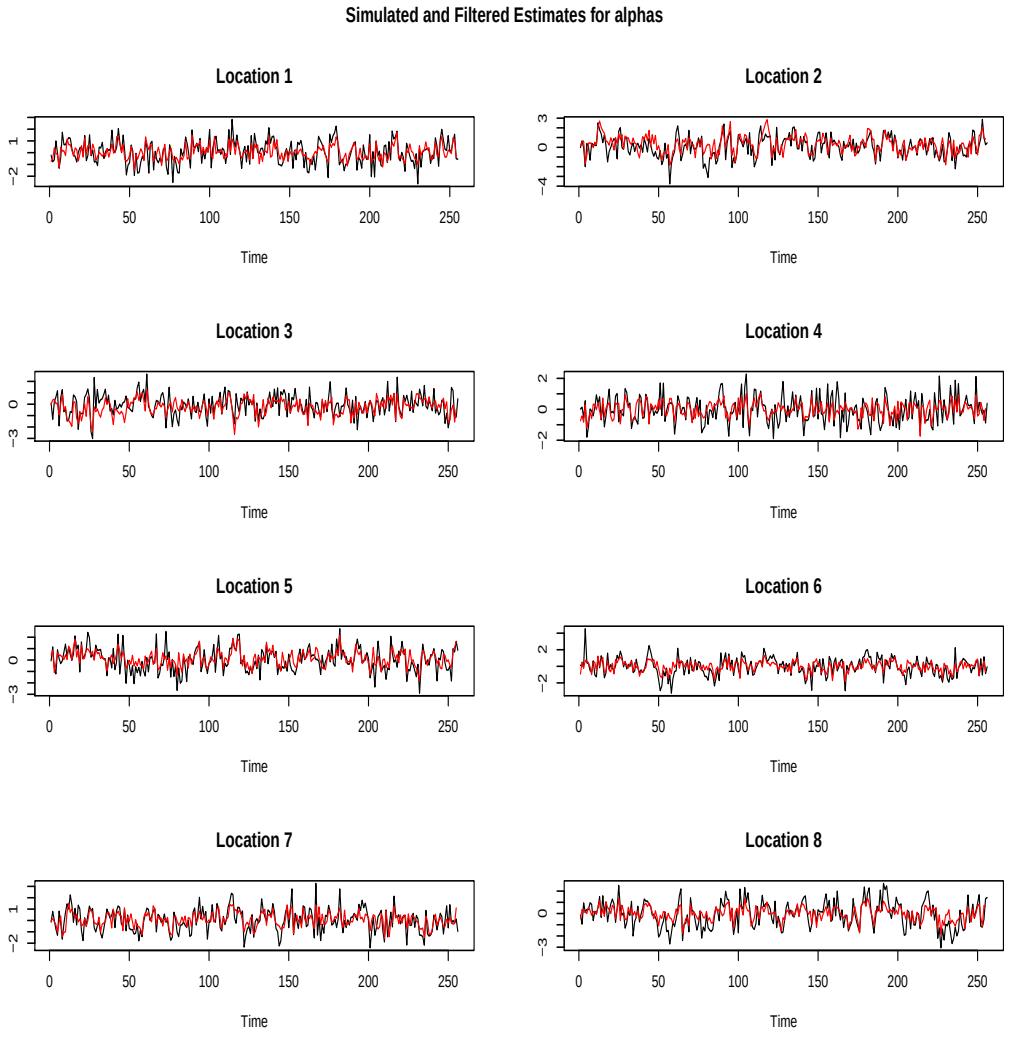

Figure 4.5: Time series plots of simulated states $\boldsymbol{\alpha}_t$ for the locations (black) and the estimated filtered ones (red) for N = 800, M = 10^4, $T = 256$, $n = 8$ under a burn-in period of i = 5000.

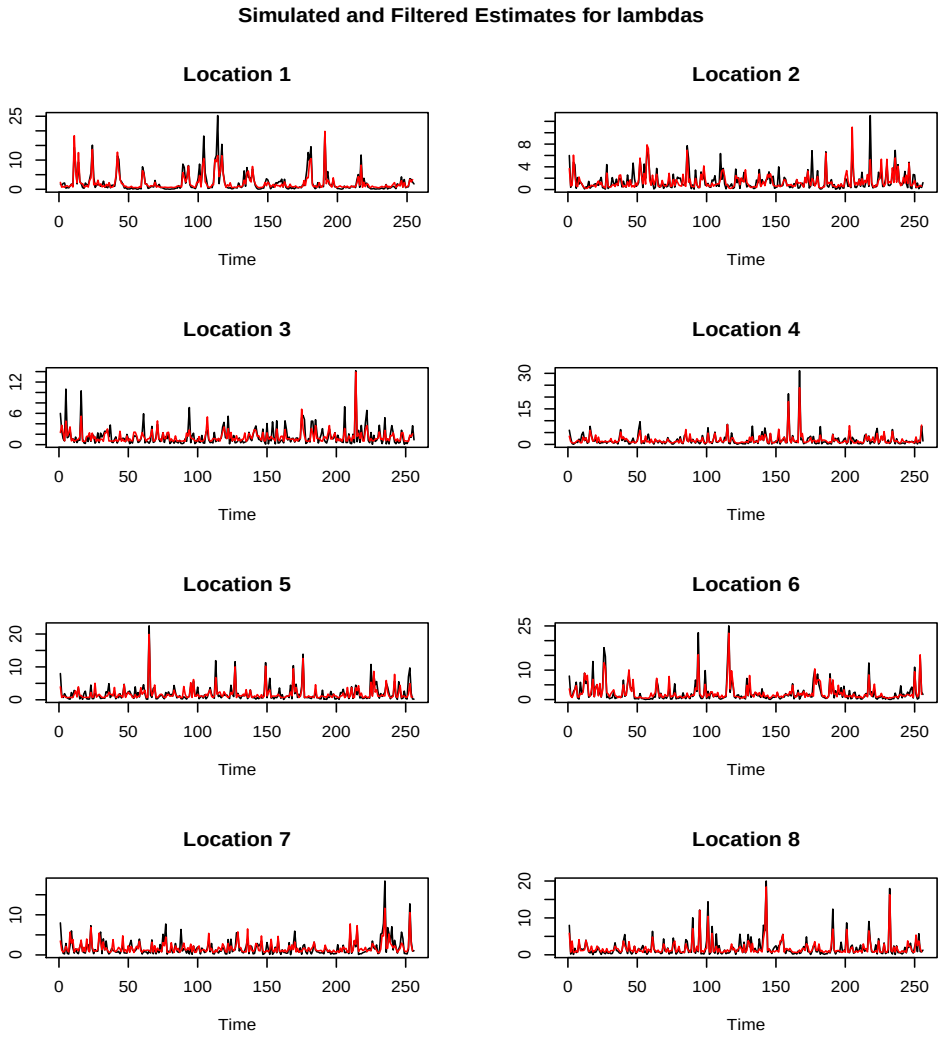

Figure 4.6: Time series plots of the simulated mean count process $\boldsymbol{\lambda}_t$ for the locations (black) and the estimated filtered one (red) for N = 500, M = 10^4, $T = 256$, $n = 8$ under a burn-in period of i = 5000.

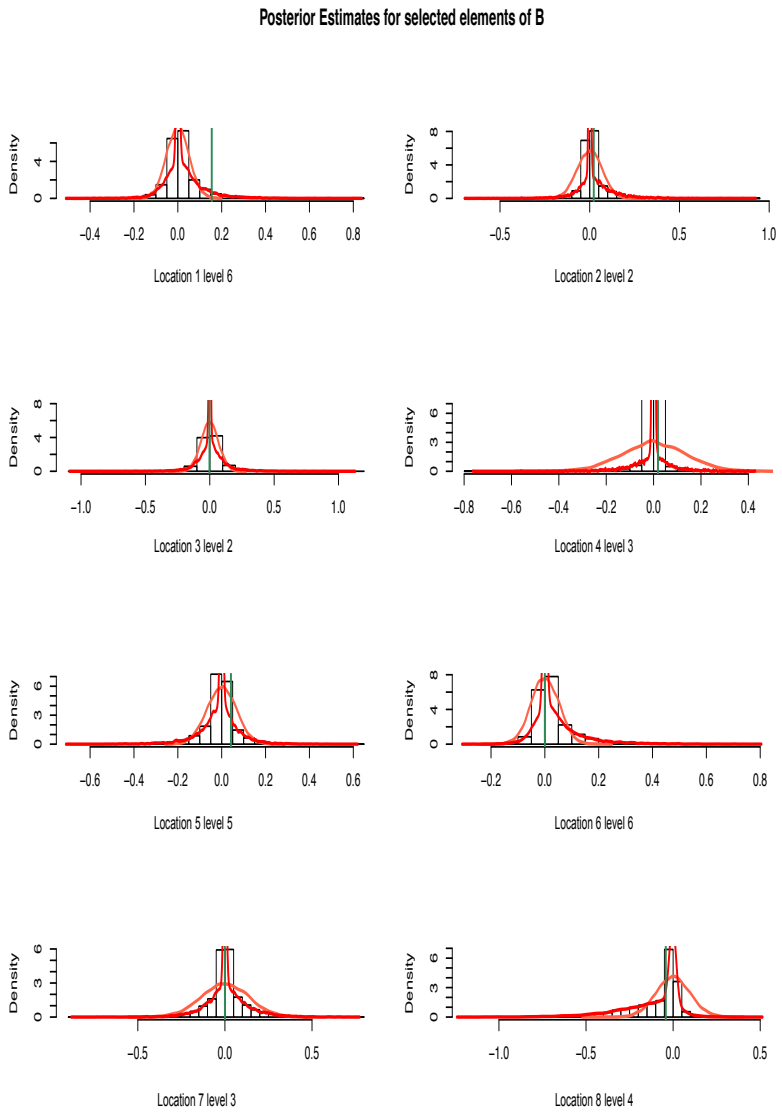

Figure 4.7: Posterior distribution for selected elements of $\boldsymbol{B}$, density posterior estimation is marked with red, the prior distribution with orange and the real value with a green vertical line, for N = 500, M = 10^4, $T = 256$, $n = 8$ under a burn-in period of i = 5000. The hyperparameter values were set to $v_0 = 0.05$, $\omega_1 = 2$ and $\omega_2 = 20$.

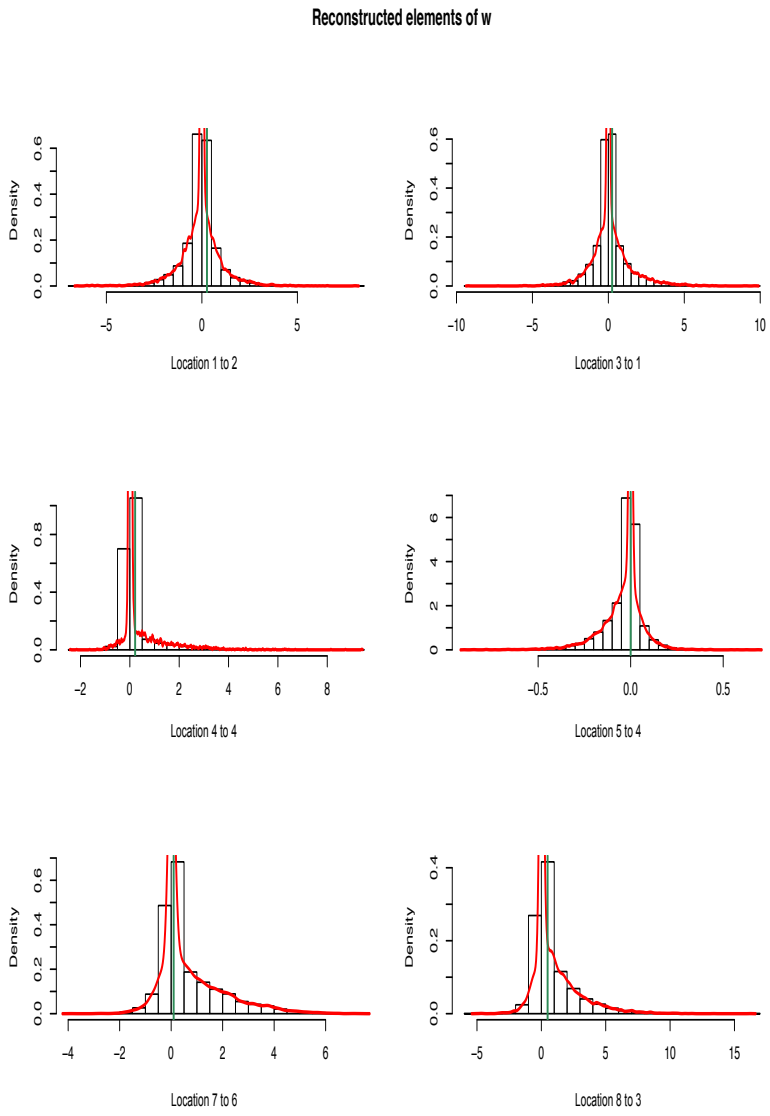

Figure 4.8: Selected reconstructed elements of $\boldsymbol{w}$. The red line indicates the empirical density estimate and the green vertical line indicates the real value for N = 500, M = 10^4, $T = 256$, $n = 8$ under a burn-in period of i = 5000.

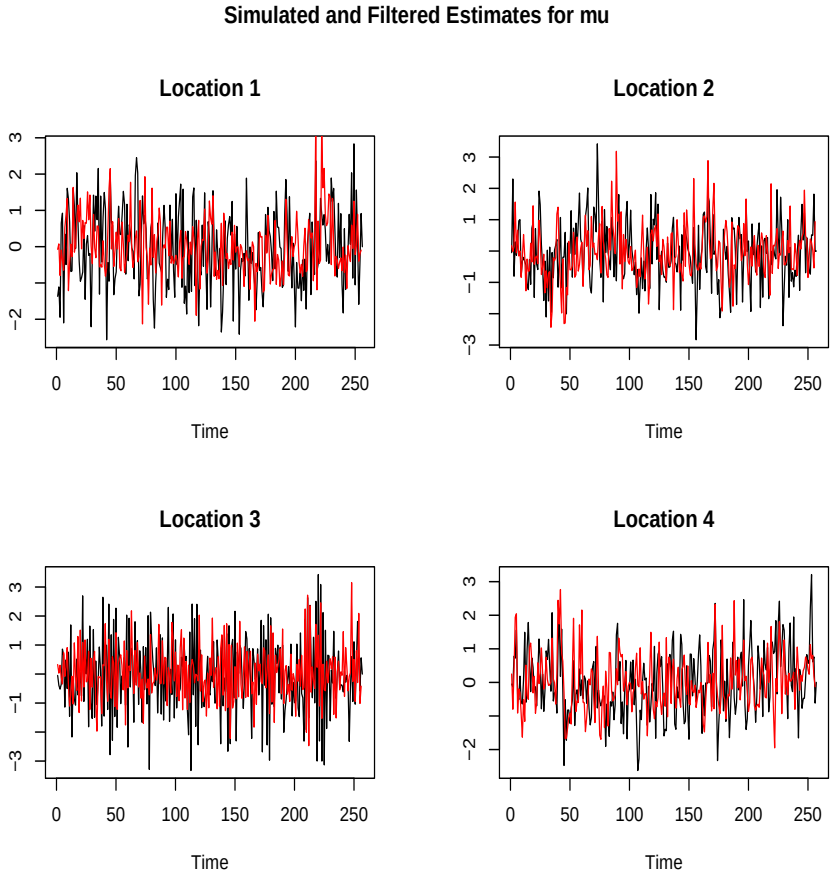

Figure 4.9: Time series plots of the simulated autoregressive mean effect $\boldsymbol{\mu}_t$ for the first four locations (black) and the estimated filtered one (red) for N = 500, M = 10^4, $T = 256$, $n = 8$ under a burn-in period of i = 5000.

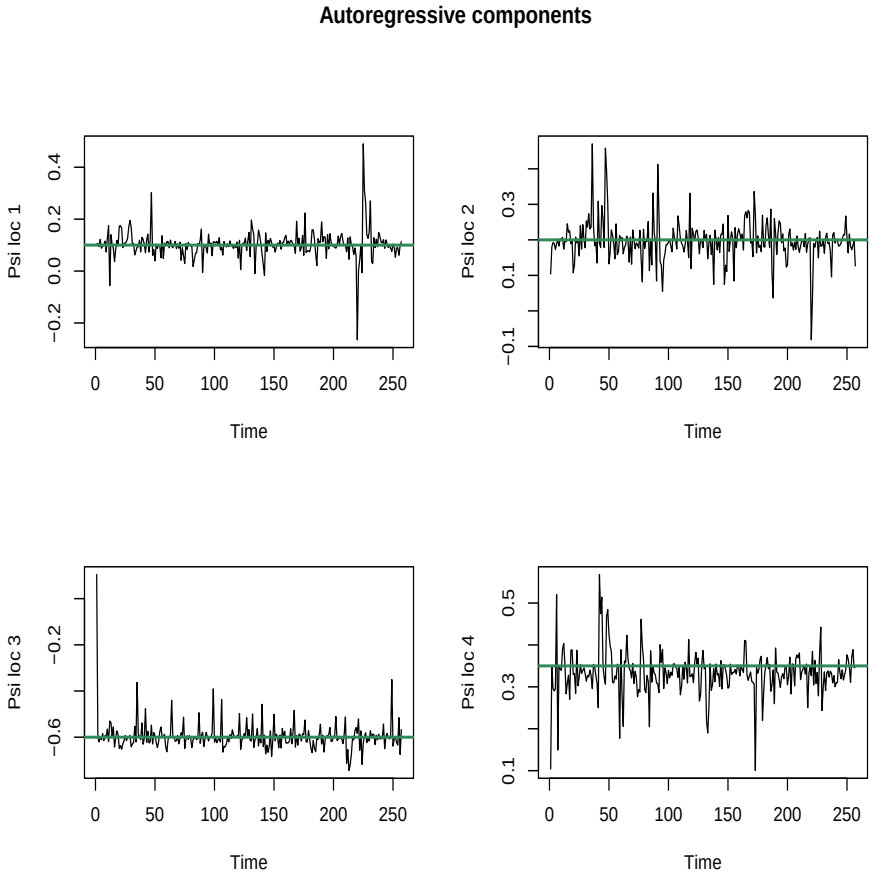

Figure 4.10: Time series plots for the autoregressive parameters $\boldsymbol{\psi}$ for the first four locations for N = 500, M = $2 * 10^4$, $T = 256$, $n = 8$ under a burn-in period of i = 5000. The horizontal green line indicates the real value of the parameter.

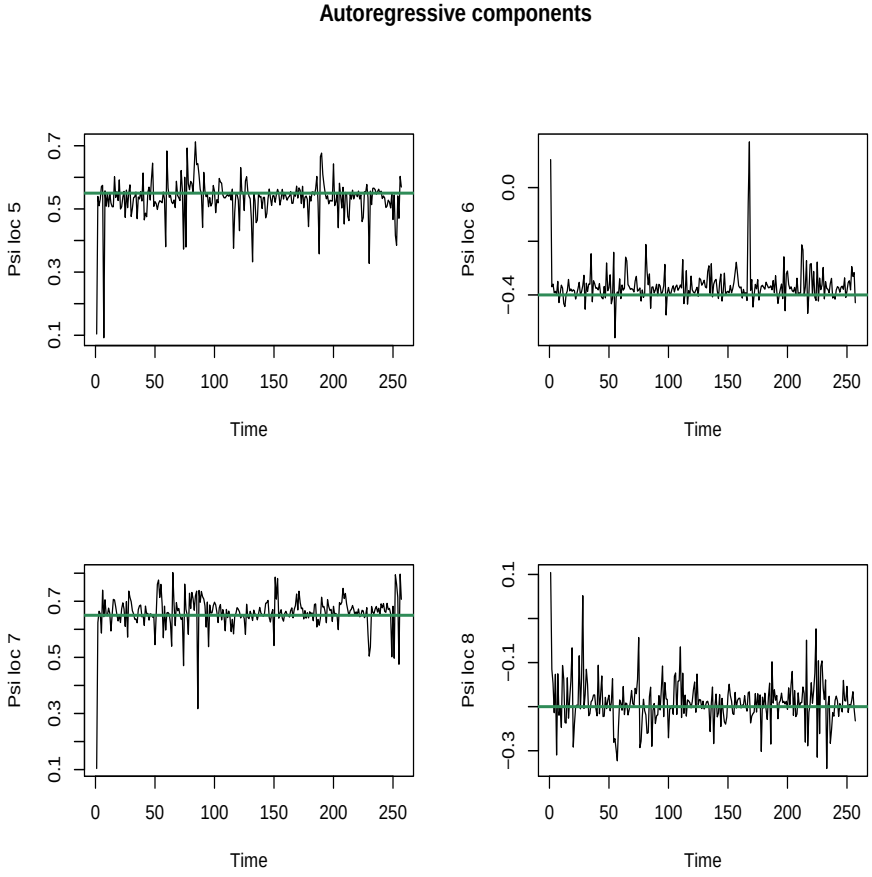

Figure 4.11: Time series plot for the autoregressive parameters $\boldsymbol{\psi}$ for the last four locations for N = 500, M = $2 * 10^4$, $T = 256$, $n = 8$ under a burn-in period of i = 5000. The horizontal green line indicates the real value of the parameter.

4.6.4 No discontinuity in weight function $w_s(u)$, autoregressive $\boldsymbol{\mu}$, autocorrelation and covariance inference

In this simulation scenario we considered the model in (4.5) with the autoregressive equation (4.4) for $n = 8$ locations and $T = 128$ in the 1-D space $[0, 1]$. Moreover, an exponential kernel with rate parameter being equal to 0.2 was considered,i.e.,

$$w_s(u) = \begin{cases} \exp((\|s-u\|^2)|0.2) & \text{if } s, u \in [0,1] \\ 0 & \text{otherwise} \end{cases} \tag{4.29}$$

which means we would expect all locations to contribute almost equally to the spatial dilation of the rest. The spatial mean effect is considered to be spatially vary-

ing evolving autoregressively and thus we consider the autocorrelation components $\boldsymbol{\Psi} = (-0.1, 0, -0.05, 0.1, 0, 0.4, -0.3, 0.12)^\top \cdot \boldsymbol{I}$ as static parameters subject to estimation while σ_ζ is considered to be known and equal to 0.5. Finally, the temporal covariance matrix is inferred under an inverse gamma prior with shape and scale parameters being 10 and 2 respectively while the true matrix was set under a diagonal structure, i.e., $\boldsymbol{\Sigma}_\eta = 1 \cdot \boldsymbol{I}$. Additionally, the reason that we considered $T = 128$ time points was to test our model's performance. Particle Filters tend to provide worse estimations when the time points are low. Finally, the wavelet basis that was used for the decomposition of the weight function $w_s(u)$ was a Daubechies level 6 basis. We conducted the Conditional Particle Filtering with Ancestor Resampling under static parameter estimation (Table 4.12) by considering the static parameter vector $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_8)^\top$ each with a truncated normal prior on $[-1, 1]$ centered at zero with variance equal to 1. The number of particles was again N = 500 with the Gibbs iterations being M = 10^4 with a burn-in period of i = 5000 under the same convergence diagnostics.

A few comments are in order. Considering that we have lower time points and more parameters in the model, under the same number of particles, the approximation of $\boldsymbol{\lambda}_t$ is very good for all locations (Figure 4.12). Specifically, all the high peaks are estimated pretty well with the only exception being an early peak at the eighth location. Furthermore, the reconstruction of the weight function provides us with estimations mostly gathered around in the real simulated values of the process (Figure 4.13). Additionally, while we provided a highly informative prior for low values of σ_η, the posterior distribution (Figure 4.14) actually provided us with posterior mean and median equal to 1.004 and 0.983 respectively which are actually very good point estimates close to the actual value.

Finally, considering the effective sample size (N_{eff}) in Figure 4.14, in this case we have a similar effect of many resampling steps across the inferential procedure. However, it has again a tendency to have high values and avoid resampling in each iterative step.

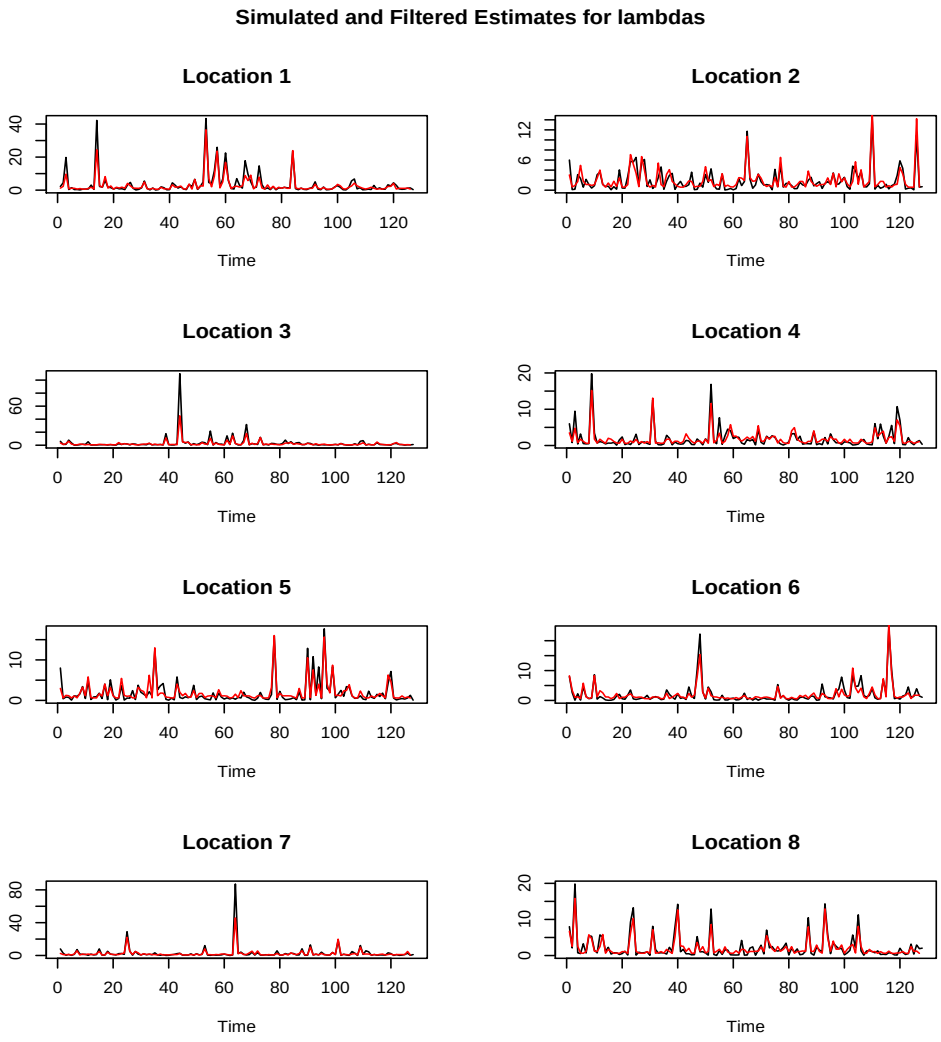

Figure 4.12: Time series plots of the simulated mean count process $\boldsymbol{\lambda}_t$ for the first four locations (black) and the estimated filtered one (red) for N = 500, M = 10^4, $T = 128$, $n = 8$ under a burn-in period of i = 5000.

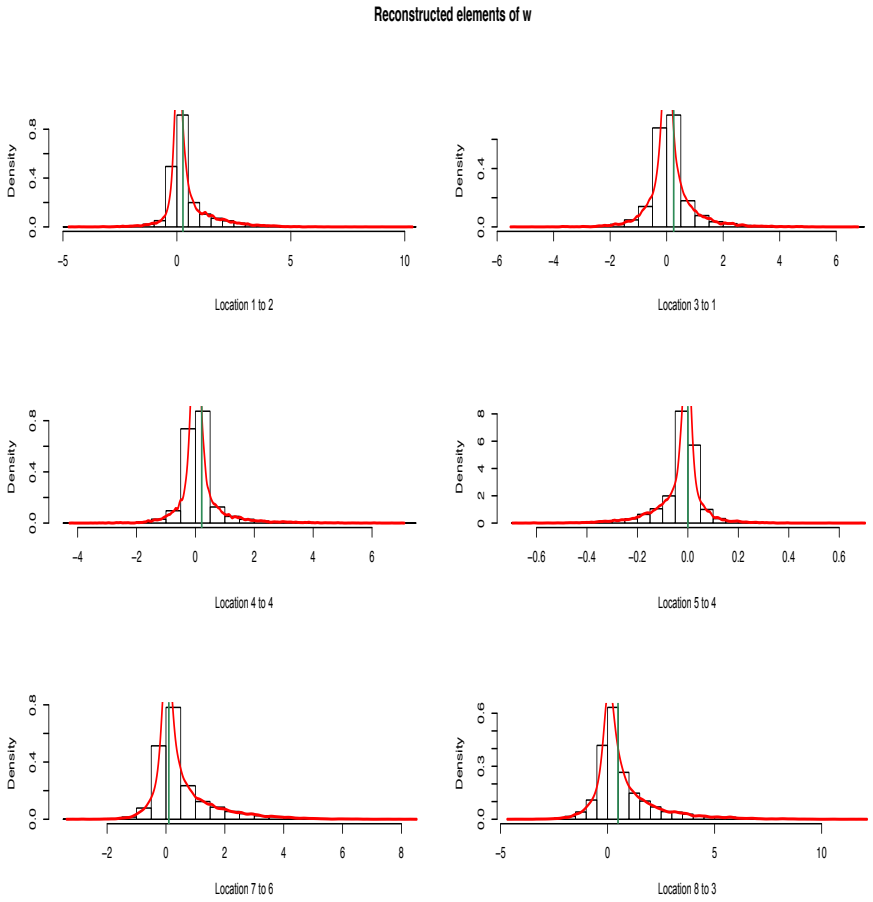

Figure 4.13: Selected reconstructed elements of $\boldsymbol{w}$. The red line indicates the empirical density estimate and the green vertical line indicates the real value for N = 500, M = 10^4, $T = 256$, $n = 8$ under a burn-in period of i = 5000.

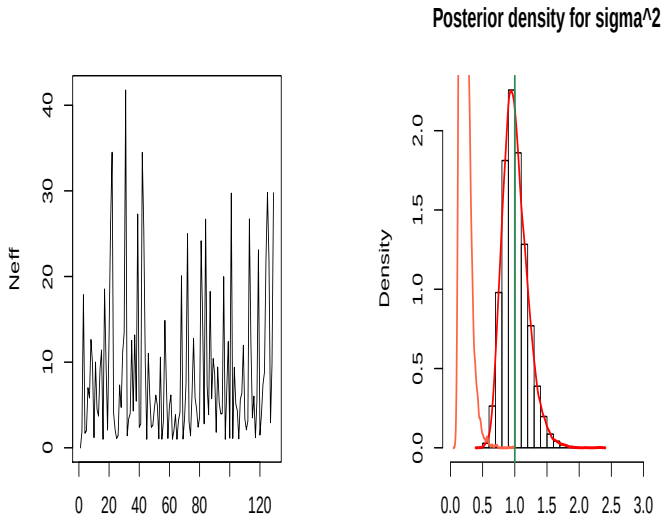

Figure 4.14: On the right: Posterior density estimate for the variance parameter σ_η^2. Red line indicates the empirical distribution of the estimates. The prior that was an inverse gamma with $\delta_0 = 10$ and $\xi_0 = 2$ being the shape and scale parameters. On the left: Effective sample size of the final Gibbs iteration M under particle filtering. The analysis was conducted for N = 800, M = 10^4, $T = 128$, $n = 8$ under a burn-in period of i = 5000.

4.7 Application on Traffic Flow Data

The Department of Transport (DfT) collects data on the amount of traffic by using manual traffic counts and a network of 180 automatic traffic counters (ATCs). Specifically, around 10,000 manual counts are being held every year on both major and minor roads. In this application we will focus on the manual counts that where conducted in the major road M6. This motorway is the first to be built by the British government. The starting point is in Rugby, Warwickshire and it terminates in Carlisle, Cumbria. As of 2016, M6, as well as combining with the length of A14 from Brampton (Cambridgeshire) from junction with A1(M), A74(M) and M74 to the junction with M8 in Glasgow, forms the longest non-stop motorway and one of the busiest in the United Kingdom. It incorporates the Preston by-pass, which as well is known to be one of the most dangerous segment with many accidents per year. Using this traffic flow dataset would provide us with insights on the traffic flows happening throughout the motorway, which could eventually be related on the accidents.

Two datasets are provided by DfT, one is the Annual Average Daily Traffic Flows (AADF) in which the number of vehicles that drive on each stretch of road on an average day of the year is provided. The second dataset that we analyse, provides the total volume of traffic on the stretch of road for the whole year, and that is calculated by multiplying the AADF by the corresponding length of road and by the number of days in the years. A quarterly time period was chosen from 2000 up to 2015 which provides us with $T = 64$ total time points. A Bayesian hierarchical spatio-temporal analysis for both datasets extracted for the area of Leeds has been conducted in Chalk (2014). Specifically, the model of Miaou and Lord (2003) is used to estimate the accident rate per million vehicles per mile.

In this thesis, we extracted the measurements for M6 motorway which were segmented based on municipalities. Due to some road segments being small in length or due to segments being under the same jurisdiction, it was decided to combine these segments and in Table 4.13 we provide the final segmentation along with their respective length in miles. Furthermore, the segmentation is visualised in Figure 4.15 where the map of United Kingdom is provided with the M6 motorway and the respective stretches

Segment	*Municipalities Incorporated*	*Length of M6 in Miles*
1	Warwickshire & Solihull	25
2	Sandwell & Birmingham & Wallsall	32
3	Staffordshire	23
4	Cheshire & East Cheshire	29
5	Warrington	30
6	Wigan & St. Helen's	20
7	Lancashire	34
8	Cumbria	39

Table 4.14: Segmentation of M6 with respective length in miles.

Figure 4.15: Map of the United Kingdom. The thick blue line represents the M6 motorway. The thick black lines indicate the segments along with their associate number of municipality.

Additionally, the data consisted of a fair amount of missing values, which were imputed via Kalman Filtering on their logarithm prior to the analysis. An example of the imputed missing values can be seen in Figure 4.16 for the Warwickshire municipality.

In this analysis we are considering our CPF-AS framework of the Poisson Dimension Reduced DSTM under a Haar wavelet basis decomposition with the model in (4.5) under Algorithm 4.11 in Table 4.12 under the Spike and Slab prior with $v_0 = 0.05$. The hyperparameters for τ_k^{-2} under a gamma prior were set to be both equal to 1000.

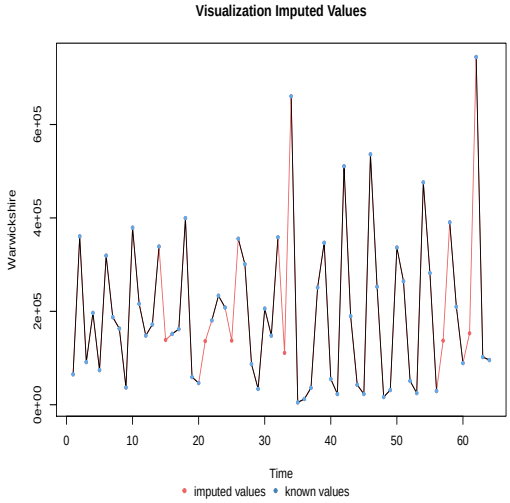

Figure 4.16: Time series plots of the observed mean count process $\boldsymbol{\lambda}_t$ for Warwickshire (1st Segment) and the missing imputed ones.

Moreover, some segments are more volatile than others but also since they are all connected to lead into Glasgow we suspect that there is a covariance structure as well. Therefore a full covariance structure is considered for $\boldsymbol{\Sigma}_\eta$, with an inverse Wishart prior with $\nu = 8$ and $\boldsymbol{Q} = 10 \cdot \boldsymbol{I}$ respectively. Finally, as we are modelling average counts, we considered σ_ϵ to be estimated and on both σ_ϵ^2 and σ_ζ^2 inverse gamma priors with both shape and scale parameters being equal to 0.01 were considered.

A few comments are in order. The proposed methodology captures fairly well the peaks in each location (Figure 4.16). Particle filters tend to be more sensitive when a lower amount of time points exists. However, the mean intensity process for most of the locations seems to be estimated well for the processes that did not have a lot of missing values. Cumbria had a total of 20 missing points. That means that these points where imputed and we had to conduct inference based on them. Furthermore, we need to emphasize that due to not having intense computational resources, more particles and Gibbs iterations were needed, which consequently would provide us better estimation.

Additionally, regarding the estimates of the spatio-temporal mean effect μ_t (Figure 4.17), it seems that after the first quarter of 2010 the mean effect for Warwickshire, Birmingham and Lancashire has more downward oscillations. Furthermore, the model suggests that Cumbria, Staffordshire and Warrington between years 2000 and 2003

there was a lower mean effect affecting the intensity process $\boldsymbol{\lambda}_t$ with Warrington having a much lower impact among all the segments. Furthermore, for all locations excluding Birmingham, during the third quarter of 2012, the mean effect $\boldsymbol{\mu}_t$ drops dramatically while for Birmingham increases. However, the model indicates strong positive autocorrelations within the segments (Figure 4.18) and all being between 0.6 and 0.8. This indicates that the mean effect for each location throughout time is affected quarterly in the same way for each segment which indicates as a possible a autocorrelation possible structure $\boldsymbol{\Psi} = \psi \cdot \boldsymbol{I}$. Although the autocorrelations under static parameter estimation do not seem stabilised at the final time points of the inference. The inference of static parameters can be sensitive when there is a low number time points.

Regarding the reconstructed elements of $\boldsymbol{w}$ a few comments are in order. Firstly, there is a negative causal effect from Warwickshire segment to Wigan and Warrington segments. This can be observed from Figure 4.16 as Warwickshire in specific time points, such as $t = 20$ (Q1 of 2005), has an increasing traffic flow while Wigan has a decreasing one. Similarly, at time point $t = 18$ (Q1 of 2004) Warwickshire has an increase in traffic flow while Warrington a decrease. This indicates that during winter months, M6 has more traffic activity on the southern part which indicates as well that in cold months drivers avoid travelling in the northern parts.

Furthermore, there is a slight positive causal relationship from Warwickshire to Birmingham. These segments are neighbouring ones but also Warwickshire is the starting point. This indicates that the amount of drivers that are using M6 would tend to be similar in these two segments in order to reach the rest of the segments. Furthermore, it is known that between Warwickshire and Birmingham there are toll roads for connections to other areas and that is consequently affecting these two segments. Additionally, the model suggests a slight causal relationship from Cheshire to Wigan, which is explained from the fact that as Cheshire is a rural county and holiday a destination for the British, while Wigan and St. Helen's are both cities, the first located in Manchester and the latter a Metropolitan one. The spatio-temporal mean effect for these two locations is modeled as contrasting for these two segments in the final half of the time points.

Considering one segment to itself we did not find any relationship which indicates that the traffic flow to a segment is not affected by the flow observed on the previous quarter (Figure 4.20).

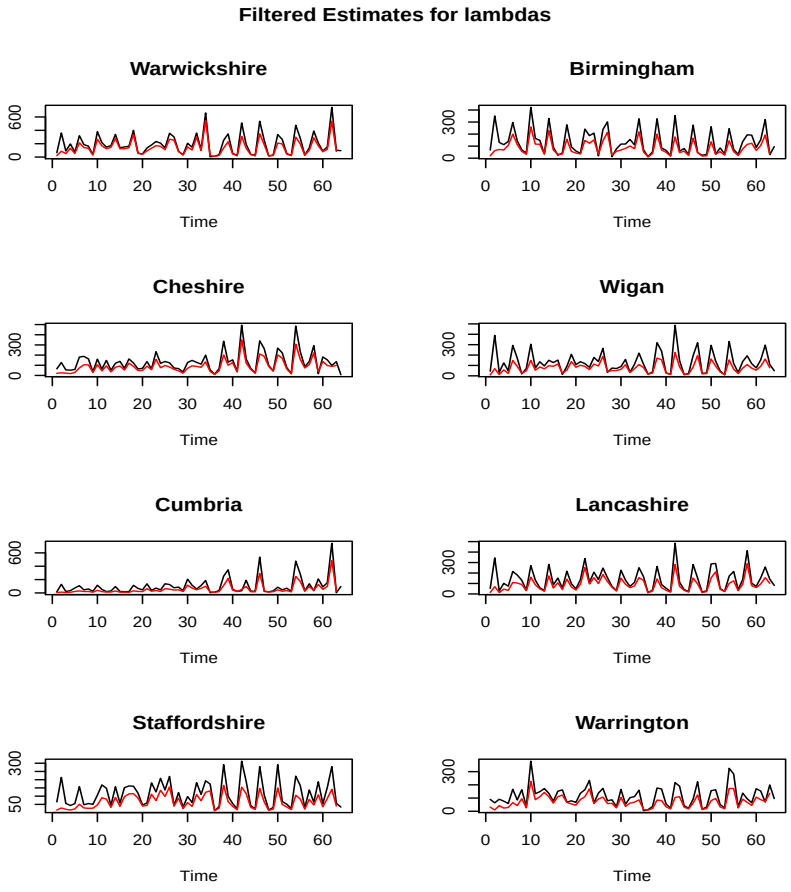

Figure 4.17: Time series plots of the observed mean count process $\boldsymbol{\lambda}_t$ for all segments and the estimated filtered one (red) for N = 500, M = 10^4, $T = 64$, $n = 8$ under a burn-in period of i = 5000.

For the rest of weight function elements, we observed that there is a negative causal relationship from the end point, Cumbria, to Warrington and Wigan. This is due to Cumbria being a non-metropolitan county and a holiday destination incorporating the Lake District. On the other hand, Warrington and Wigan are urban cities close to Liverpool and Manchester consisted of lot of traffic between them. Finally, the model suggests that Lancashire, which is a rural county and far metropolitan areas such as Birmingham do not have any causal effect to each other.

For the covariance inference a few comments are in order. During the inferential part we considered a full covariance structure for the covariance matrix $\boldsymbol{\Sigma}_\eta$. The model suggests that the posterior temporal variances for all segments are similar (Figure 4.21).

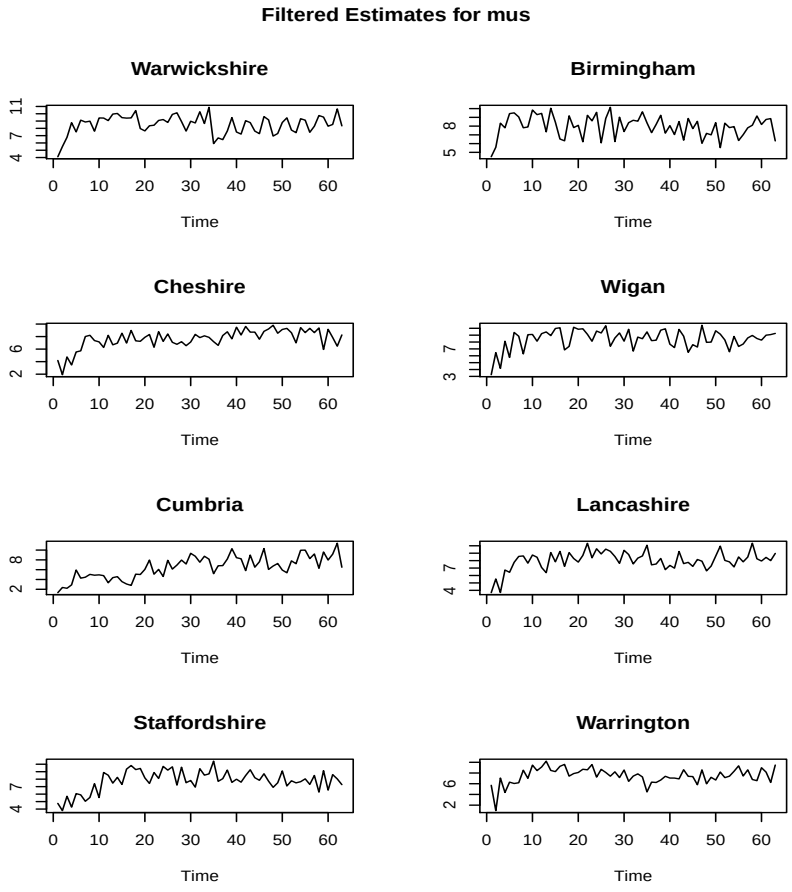

Figure 4.18: Time series plots of the estimated filtered spatio-temporal mean effect $\boldsymbol{\mu}_t$ for all segments for N = 500, M = 10^4, $T = 64$, $n = 8$ under a burn-in period of i = 5000.

Specifically, the posterior means for most segments is between 2.5 and 2.8 only with Warwickshire and Chestershire differing with posterior median variances of 3.85 and 3.45 respectively. Considering the covariance structure, there was no indication of temporal correlation between segments with the only exception of Cheshire and Wigan negatively covaring (Figure 4.22). Finally, the posterior mode error variances σ^2_ϵ and σ^2_μ were estimated as being 129.77 and 34.43 respectively (Figure 4.23). The higher error to spatial variance is due to the high variability of the number of traffic flows for all locations which is mostly incorporated and estimated through σ^2_ϵ. Furthermore, the spatial effect vector $\boldsymbol{\mu}_t$ through time does not show high variability compared to the highly variable count process $\boldsymbol{\lambda}_t$. The diagnostic tools that were used for convergence indicate that more Gibbs iterations were needed in order for the samples to converge.

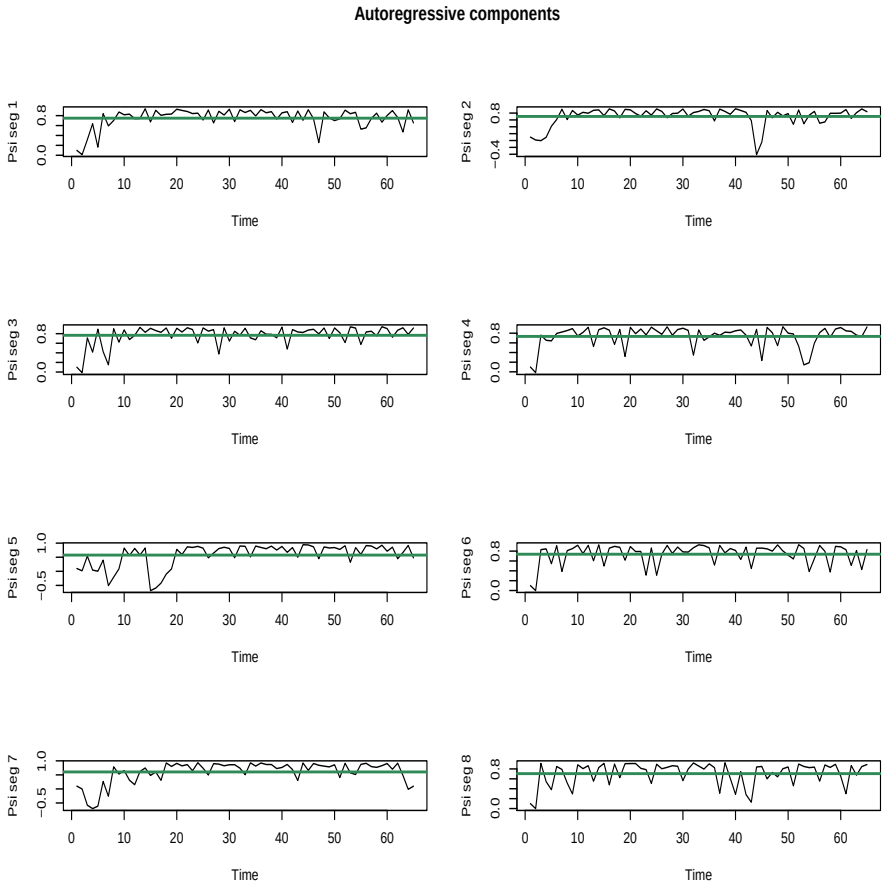

Figure 4.19: Time series plots for the autoregressive parameters $\boldsymbol{\psi}$ for all segments for N = 500, M = 10^4, $T = 64$, $n = 8$ under a burn-in period of i = 5000. The horizontal green line indicates the mean value of the parameter.

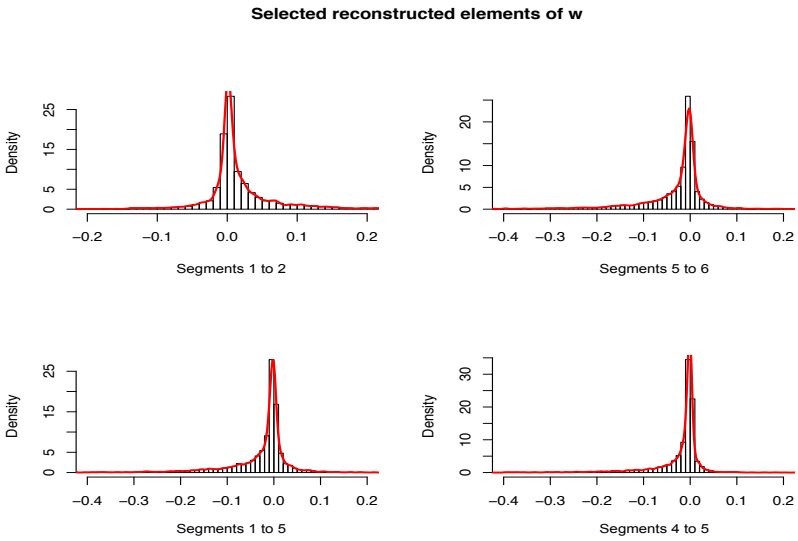

Figure 4.20: Selected reconstructed elements of $\boldsymbol{w}$. The red line indicates the empirical density estimate and the green vertical line indicates the real value for N = 500, M = 10^4, $T = 64$, $n = 8$ under a burn-in period of i = 5000.

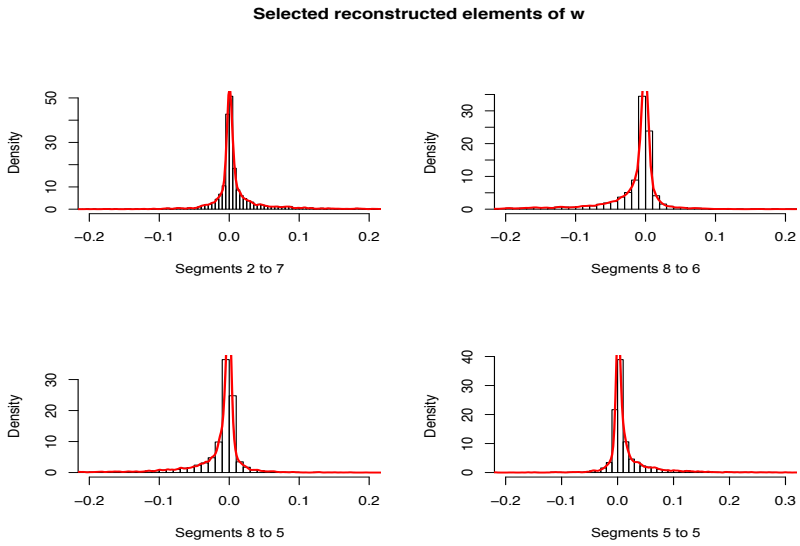

Figure 4.21: Selected reconstructed elements of $\boldsymbol{w}$. The red line indicates the empirical density estimate and the green vertical line indicates the real value for N = 500, M = 10^4, $T = 64$, $n = 8$ under a burn-in period of i = 5000.

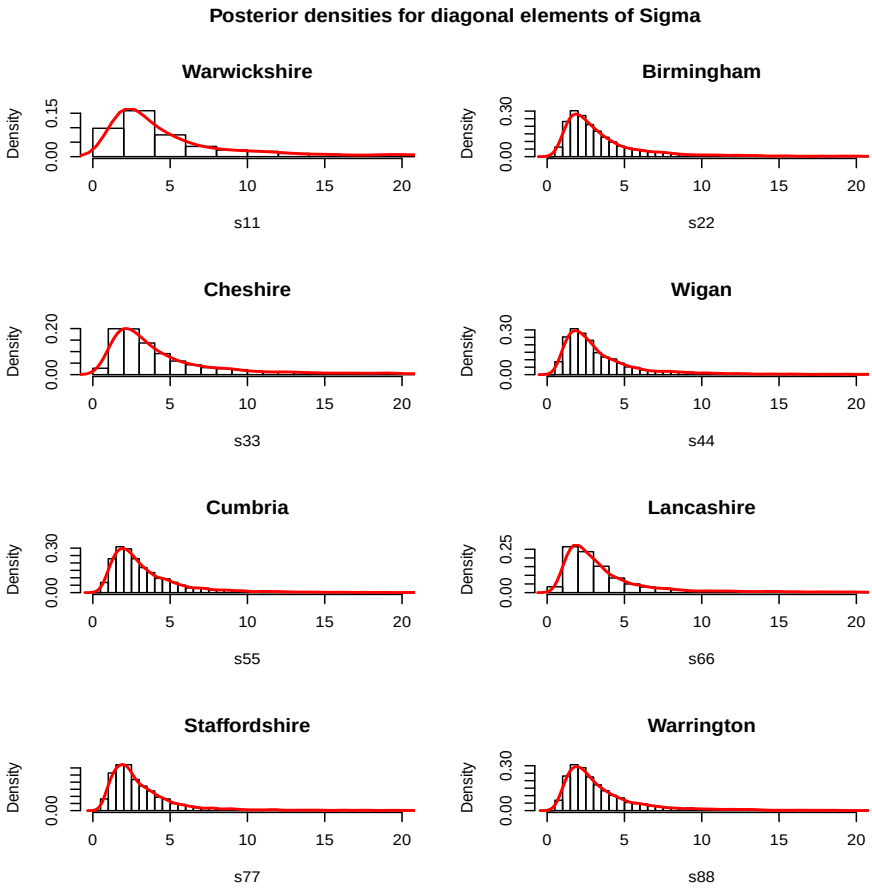

Figure 4.22: Posterior density estimates for the diagonal elements (variances) of Σ_η. Red line indicates the empirical distribution of the estimates. The prior that was an inverse Wishart with $\nu_0 = 8$ and $Q_0 = 100 \cdot I$ being the shape and scale parameters for N = 500, M = 10^4, $T = 64$, $n = 8$ under a burn-in period of i = 5000.

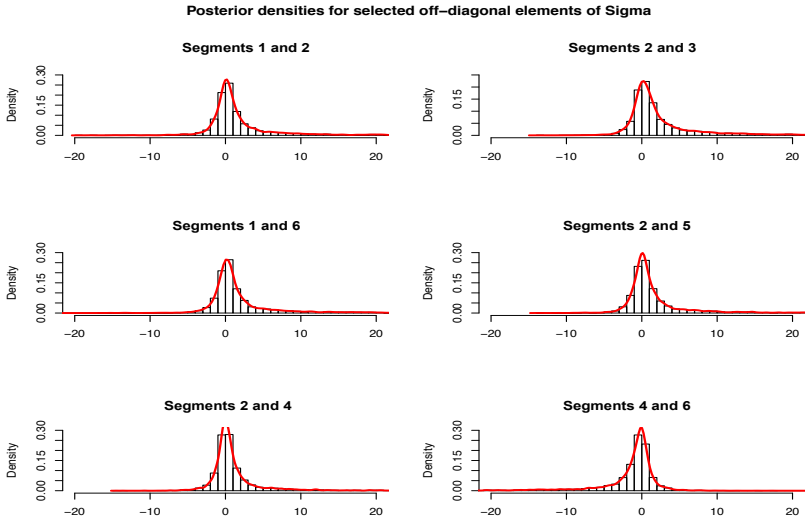

Figure 4.23: Posterior density estimates for selected off-diagonal elements (covariances) of Σ_η. Red line indicates the empirical distribution of the estimates. The prior that was an inverse Wishart with $\nu_0 = 8$ and $Q_0 = 100 \cdot I$ being the shape and scale parameters for N = 500, M = 10^4, $T = 64$, $n = 8$ under a burn-in period of i = 5000.

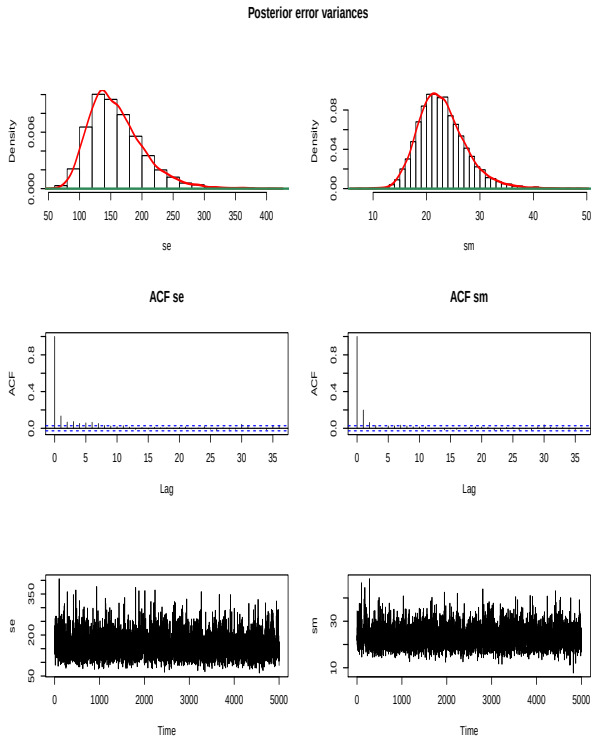

Figure 4.24: On the left: Posterior density estimate for the variance parameter σ_ϵ^2. On the right: Posterior density estimate for the variance parameter σ_μ^2. Red line indicates the empirical distribution of the estimates. The prior distribution was an inverse gamma with $\delta_0 = 10^-2$ and $\xi_0 = 10^2$ being the shape and scale parameters. The analysis was conducted for N = 500, M = 10^4, $T = 64$, $n = 8$ under a burn-in period of i = 5000.

Missing Value Treatment Analogously to Chapter 3, the missing imputation was conducted under a non-Bayesian setting. One approach that we could consider under the particle filtering framework is the idea of multiple imputations. Multiple imputations (Rubin, 2004) consist of creating multiple complete data sets imputing m values for each missing datum so that sampling variability around the actual values is incorporated for performing valid inferences. Under this idea, the Multiple Imputation Particle Filter is an extension of the Particle Filter (Housfater et al., 2006) that incorporates multiple imputation steps for the cases where measurement data is not available, so that the algorithm can assimilate for the corresponding uncertainty in the inferential process.

Thus, let us partition the vector of observations, i.e., $\boldsymbol{Y}_t = (\boldsymbol{Y}_t^{mis}, \boldsymbol{Y}_t^{obs})$ and let us consider the missing value index $j = 1, \ldots, m$. Then, an imputation model can be

expressed as a probability distribution in order to sample the m samples subject to imputation, i.e.,

$$\boldsymbol{Y}_{jt}^{mis} \sim p(\boldsymbol{Y}_t^{mis}|\boldsymbol{Y}_{1:t}^{obs}) \tag{4.30}$$

Similarly to to importance sampling, we can assign a weight v_t^j to each imputation with $\sum_{j=1}^m v_t^j = 1$. Based on Kong et al. (1994), by considering $u_t^j = (\boldsymbol{Y}_{jt}^{mis}, \boldsymbol{Y}_t^{obs})$ to be the complete data sets formed from imputed values, then the filtering posterior distribution is given as:

$$p(\boldsymbol{\alpha}_t|\boldsymbol{Y}_{1:t}^{obs}) = \int p(\boldsymbol{\alpha}_t|\boldsymbol{u}_{1:t-1}, \boldsymbol{Y}_t^{obs}) p(\boldsymbol{Y}_t^{mis}|\boldsymbol{Y}_{1:t}^{obs}) d\boldsymbol{Y}_t^{mis}, \tag{4.31}$$

and through Monte Carlo approximation we get

$$p(\boldsymbol{\alpha}_t|\boldsymbol{Y}_{1:t}^{obs}) \approxeq \sum_{j=1}^{m} v_t^j p(\boldsymbol{\alpha}_t|\boldsymbol{u}_{1:t-1}\boldsymbol{Y}_t^{obs}, u_t^j). \tag{4.32}$$

Additionally, for each of the complete data sets yields

$$p(\boldsymbol{\alpha}_t|u_{1:t-1}, u_t^j) = \sum_{i=1}^{N} w_t^{(i,j)} \delta(\boldsymbol{\alpha}_t - \boldsymbol{\alpha}_t^{(i,j)}), \tag{4.33}$$

where the indexes i and j indicate the particle and the imputation, respectively. Thus, an approximation of the desired posterior distribution is

$$p(\boldsymbol{\alpha}_t|\boldsymbol{Y}_{1:t}^{obs}) \approx \sum_{j=1}^{m} \sum_{i=1}^{N} v_t^j w^{(i,j)} \delta(\boldsymbol{\alpha}_t - \boldsymbol{\alpha}_t^{(i,j)}). \tag{4.34}$$

Thus, estimating the missing responses $\boldsymbol{Y}_t^{mis}$ is equivalent to a posterior prediction from the model fitted to the observed data, however with a filtering estimate for the missing observations as well. Obviously, if we consider the PF-AS for static parameter estimation and PMH, the steps above can be easily be incorporated in the algorithms with no change in the theoretical convergence.

4.8 Conclusion

We have introduced two modeling approaches for Poisson Dimension Reduced DSTMs under an adaptive Conditional Particle Filter procedure with static parameter estimation. The two approaches consider a spatially varying and a spatio-temporal mean effect for the intensity processes respectively. These two models provides us with flex-

ibility of modeling spatio-temporal count processes with different spatial and spatio-temporal features. Additionally, we make use of the truncation proposed in Chapter 3 under the efficient sparse wavelet decomposition and the inference of spatial coefficients through the Spike and Slab prior. Furthermore, an efficient combination of filtering non-Gaussian processes under a Bayesian framework is used for the estimation of the temporal wavelet coefficients. Lastly, a flexible static estimation is provided for the static autocorrelation parameters in order to increase the efficiency of the model. Last but not least, a simulation scheme was introduced for the two Poisson Dimension Reduced DSTM models under wavelet basis decomposition.

Firstly, simulations for both a small number of locations have proved the effectiveness of our methodology on approximating an underlying spatio-temporal intensity with complex dynamics. Moreover, the proposed methodology successfully approximates a Poisson observed spatio-temporal process with spatial discontinuities (Section 4.6.2). Additionally, the reconstruction of the weight function, which is a challenge in practice to be estimated, was predicted fairly well. As for the inferential procedure of the covariance structure the estimation deemed successful. Furthermore, we tested our methodology on real traffic flow data under segmentation of the known motorway M6 after imputing missing values. Our model managed to capture fairly well the intensity processes. Furthermore, we achieved to produce spatial causality between locations. These spatial causalities seemed reasonable based on the distance and the type of counties. Finally, we observed that all all segments vary similarly and own strong quarterly autocorrelations.

However, we encountered hindrances during our estimation. Firstly, the low amount of time points and missing values on the traffic flow application provided us with worse estimations for a few locations of the intensity process but also a lack of convergence for the autocorrelation parameters. Furthermore, the lack of computational infrastructure did not provide us with more Gibbs and particle iterations which are strongly suggested for this application. Finally, the application on real data showed as that the model is sensitive in approximating the intensity process when a large amount of missing data exists.

Finally, it was observed that the more we increase the number of locations, the more computational power we need for the model to run and is more inefficient than the MCMC procedure proposed in Chapter 3. Added to that, an implementation through

Stan probabilistic programming cannot be considered in this aspect as Stan does conduct any sampling from Auxiliary Particle Filtering methods, but only through Hamiltonian Monte Carlo, No-U-turn Sampler and automatic variational inference. A sampling through these methods would be an equivalent to sample from a non-linear state space model.

Chapter 5

Multinomial Reduced-dimension Dynamic Spatio-Temporal Models

5.1 Introduction

In the previous chapter we focused on a proposed extension of Wikle et al. (2001) and Wikle and Cressie (1999) for Poisson distributed observations with the consideration of Wikle (2002) while we provided an adaptive Conditional Particle Filtering under static parameter estimation framework with the help of wavelet basis decomposition. That model proved to be efficient for processes of this nature which consequently brings us into investigating a similar approach for Multinomial distributed data. In real life applications there is a plethora of datasets where certain phenomena can be expressed in proportions. For instance, as discussed briefly in the previous chapter, if we would like to model the cancer rates per location, then we have Multinomial proportions, or else, if we categorise the severity of accidents occurring in a road segment, then we again conclude on Multinomial distributed observations. Therefore, the implementation of Gaussian DSTMs or Poisson DSTMs in these datasets is not appropriate.

Based on the aforementioned arguments, in this chapter we are going to focus on the reduced-dimension DSTM under Multinomial distributed observations. Specifically, we introduce a modelling procedure where the proportions evolve through time and space while the state vector $\boldsymbol{\alpha}_t$ will now be a matrix, i.e., $\boldsymbol{A}_t$ based on the number of cate-

gories. Furthermore, we consider a complex covariance structure for the state matrix $\boldsymbol{A}_t$ which is a combination of spatial and categorical variation. Additionally, we will consider again in this chapter the wavelet basis decompositions due to their efficiency and good localisation properties. The inferential procedure will be based on Particle Filtering (PF) methods that were used for the Poisson DSTM model and we will provide thorough explanation in the latter sections.

Additionally, simulation implementations were conducted for the proposed modelling framework. Our findings are promising for processes of proportions and/or counts. Specifically, the spatio-temporal varying cell probability processes were captured fairly well even during high oscillations, which is in general a challenge in practice for these kind of processes. Furthermore, the weight function was successfully reconstructed, however, updating the elements of $\boldsymbol{B}$ is computationally demanding.

Finally, at the end of this chapter we offer a real life application by revisiting the traffic flow data in Chapter 4 with each category consisting of a different type of vehicle. Under the proposed methodology, our findings include...

5.2 Approaches on modeling Multinomial spatio-temporal processes

There have been recent developments on modelling spatio-temporal multinomial distributed processes. Spatial multinomial modelling has been widely used in the literature in numerous applications; such as spatial multinomial logit models on understaning urban features (Zhou and Kockelman, 2008), classifying vegetation spreads (Augustin et al., 2008)) or identifying determinants of land use changes (Chakir and Parent, 2009).

Only in the last decades spatio-temporal approaches have been developed and applied. Such models were developed for mapping disease rates (Waller et al. (1997) and Yang et al. (2005)), for clustering avalanche counts in the Alps (Lavigne et al., 2012), for modelling unemployment rates (Pereira et al., 2017) and modeling of land-use change (Tepe and Guldmann, 2018). However, for the dimension reduced DSTMs approaches for multinomial distributed data have not been developed. Cressie and Wikle (2015) provide a thorough review on non-linear dynamic spatio-temporal models, however, the complexity of those models surpass the non-Gaussianity of observations which we can

still link them linearly in this thesis.

5.3 Proposed Methodology

In the following sections we propose a modelling framework for modelling multinomial distributed spatio-temporal processes under Particle Filtering approaches. Specifically, we used a *logit* link function to bring the model into a reduced-dimension DSTM. Then, we combine the Conditional Particle Filter (CPF) (Lindsten et al., 2014) and Particle Metropolis-Hastings (PMH) (Andrieu et al., 2010) algorithms for the parameters to be estimated. The formulation of the model is provided on section 5.3.1 while the inferential procedure to be followed is described in section 5.3.2.

5.3.1 Model Formulation

Consider $k \geq 2$ categories and let $Y_j(s,t)$ denoting a spatio-temporal process of the count or the total measurement of a quality characteristic observed at location s and category $j = 1, \ldots, J$ at time $t = 1, \ldots, T$. Moreover, denote with $\pi_j(s,t)$ the cell probability that the random variable $Y_j(s,t)$ is equal to the observed count. If we fix $\lambda(s,t) = Y_1(s,t) + \cdots + Y_j(s,t)$ to be the total count at location s and time t, $\pi_j(s,t) + \cdots + \pi_J(s,t) = 1$ for some known positive integer $\lambda(s,t)$ and define the vector $\boldsymbol{\pi}(s,t) = [\pi_1(s,t), \ldots, \pi_J s, t)]^\top$, then the joint pmf of $\boldsymbol{Y}(s,t) = [Y_1(s,t), \ldots, Y_J(s,t)]^\top$ is

$$p(\boldsymbol{Y}(s,t)|\boldsymbol{\pi}(s,t)) = \frac{\lambda(t,s)!}{\prod_{j=1}^{J} Y_j(s,t)!} \prod_{j=1}^{J} \pi_j(s,t)^{Y_j(s,t)} \tag{5.1}$$

which defines the multinomial distribution for location s at time point t.

What we would like to model is the measurements of a quality characteristic for all locations $s = 1, \ldots, n$ at time point t. We then define the vector of total counts $\boldsymbol{\lambda}_t = (\lambda(1,t), \ldots, \lambda(n,t))^\top$, the vector of cell probabilities at time t to be $\boldsymbol{\pi}_t = [\boldsymbol{\pi}(1,t), \ldots, \boldsymbol{\pi}(n,t)]^\top$ and the observed measurement vector of quality characteristics at time t and location s to be $\boldsymbol{Y}(s,t) = [\boldsymbol{Y}(1,t), \ldots, \boldsymbol{Y}(n,t)]^\top$. Furthermore, let us consider the state equation (2.11). If we like to bring (5.1) into a linear framework in order to predict the cell probabilities for the quality characteristic for each location s at time point t, then analogously to the Poisson case in Chapter 4, a linear link for the predictor should be

considered. The pmf in (5.1) can be written for each location as

$$\begin{aligned} p(\boldsymbol{Y}(s,t)|\boldsymbol{\pi}(s,t)) &= \exp\big(\sum_{j=1}^{J} Y_j(s,t)\log(\pi_j(s,t)) + \lambda(s,t) - \sum_{j=1}^{J} Y_j(s,t)\big) \\ &\times \log(1-\sum_{j=1}^{J}\pi_j(s,t))\big)\frac{\lambda(s,t)!}{\prod_{j=1}^{J} Y_j(s,t)!} \\ &= \exp\big(\lambda(s,t)[\sum_{j=1}^{J} Y_j(s,t)\log\frac{\pi_j(s,t)}{1-\sum_{j=1}^{J}\pi_j(s,t)} \\ &+ \log(1-\sum_{j=1}^{J}\pi_j(s,t))]\big)\frac{\lambda(s,t)!}{\prod_{j=1}^{J} Y_j(s,t)!} \end{aligned} \tag{5.2}$$

which is the form of a multivariate exponential family. For more details on the multinomial distribution and its relationship to the exponential family refer to Chapter 3, Fahrmeir and Tutz (2013). If we define the $n\times(J-1)$ matrix $\boldsymbol{Y}_t = (\boldsymbol{Y}(1,t),\ldots,\boldsymbol{Y}(n,t))^\top$ with the J being reference category then we can proceed on a generalised linear formulation for the modeling framework. Thus, we consider an $n\times(J-1)$ state matrix $\boldsymbol{A}_t$ the $n\times n$ and $(J-1)\times(J-1)$ covariance matrices $\boldsymbol{\Sigma}_\eta$ and $\boldsymbol{R}$. These provide us with the $n\times(J-1)$ dimensional linear predictor matrix $\boldsymbol{Z}_t$ which maps the $n\times(J-1)$ mean matrix $E(\boldsymbol{Y}_t)$ via the canonical link and a matrix formulation of the transition equation (2.11) :

$$\boldsymbol{Z}_t = \begin{bmatrix} \log\frac{\boldsymbol{\pi}_{1t}}{1-\sum_{j=1}^{J-1}\boldsymbol{\pi}_{jt}} \\ \vdots \\ \log\frac{\boldsymbol{\pi}_{Jt}}{1-\sum_{j=1}^{J-1}\boldsymbol{\pi}_{jt}} \end{bmatrix} = \boldsymbol{\Phi}\boldsymbol{A}_t$$

$$\boldsymbol{A}_t = \boldsymbol{\Phi}^\top\boldsymbol{B}\boldsymbol{A}_{t-1} + \boldsymbol{\eta}_t \quad \text{with} \quad \boldsymbol{\eta}_t \sim \mathrm{N}(0, (\boldsymbol{\Phi}^\top\boldsymbol{\Sigma}_\eta\boldsymbol{\Phi})\otimes\boldsymbol{R}) \tag{5.3}$$

A few comments are in order. Firstly, the processes of approximation are the log odds of one category in each location having more counts than the rest. Furthermore, the log of odds for the last category J is complimentary to the rest as we defined that all probabilities $\pi(s,t)$ sum to one. Therefore, we can reconstruct the actual proportions $\pi(s,t)$ for each location s at time t via the inverse link function but only for the final J-th one, we subtract the rest of quantities from one.

The state equation in this case is a matrix where we expect each category to tran-

sition differently but also for each location to vary all categories will vary differently and a covariance structure is considered. In the Gaussian and Poisson cases the state vector $\boldsymbol{\alpha}_t$ which was approximating the processes of interest was multivariate normally distributed with $\boldsymbol{\Phi}^\top \boldsymbol{\Sigma}_\eta \boldsymbol{\Phi}$ being the covariance matrix under the wavelet decomposition. As now the parameters of interest are the cell probabilities $\boldsymbol{\pi}_t$, we expect that the dynamical coefficients of the wavelet matrix $\boldsymbol{\Phi}$ will oscillate differently in each category. Therefore, this kind of variation should be incorporated by choosing this matrix structure of $\boldsymbol{A}_t$ with an extra inclusion of a covariance matrix $\boldsymbol{R}$ which signifies the within cell variation. Specifically, we consider that the spatial and categorical variation can all be incorporated through the covariance matrices $\boldsymbol{\Phi}^\top \boldsymbol{\Sigma}_\eta \boldsymbol{\Phi}$ and $\boldsymbol{R}$ respectively. This means that the structure of $\boldsymbol{\Sigma}_\eta$ will be chosen similarly as in the Gaussian and Poisson cases but the structure of $\boldsymbol{R}$ will provide us with different variation in each category but also possible correlation between them. Additionally the Kronecker product of these two matrices, i.e., $\boldsymbol{H} = (\boldsymbol{\Phi}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \boldsymbol{\Phi}) \otimes \boldsymbol{R}$, provides us with cross correlation for the categories between locations but also possible association between different categories between the locations if a full structure for both matrices will be considered. Thus, the structure of the state matrix $\boldsymbol{A}_t$ is still autoregressive but with a more complicated covariance structure which provides us with a matrix normal distribution which was reviewed in Chapter 3. Consequently, this gives us a matrix normal distribution for $\boldsymbol{\eta}_t$ with the scale matrices being $\boldsymbol{\Phi}^\top \boldsymbol{\Sigma}_\eta \boldsymbol{\Phi}$ (due to the wavelet decomposition for each $\boldsymbol{A}_j$ where j is the $j-th$ column of the state matrix $\boldsymbol{A}_t$) and $\boldsymbol{R}$ respectively.

Furthermore, as the arguments $\boldsymbol{\lambda}_t$ and $\boldsymbol{\pi}_t$ are directly affecting each other, we would expect the underlying spatial characteristics to be explained by this modelling framework, while again the approximation of the weight function under the matrix $\boldsymbol{B}$ will provide us with the spatial diffusion dynamics. Finally, we note that for two categories $s = 2$ the multinomial model (5.3) reduces to a binomial model.

Calculation of $\boldsymbol{\pi}_t$ based on Odds Ratio Consider the model in (5.3). In order to estimate the model parameters, we will consider a reference category, specifically the $J-th$ one. This will automatically provide us with an $n \times (J-1)$ state matrix $\boldsymbol{A}_t$, a canonical link $n \times (J-1)$ matrix $\boldsymbol{Z}_t$ and scale matrices being $n \times n$ and $(J-1) \times (J-1)$ for $\boldsymbol{\Sigma}_\eta$ and $\boldsymbol{R}$ respectively.

We then define the $(J-1) \times 1$ vector of odds ratios for $J-1$ categories being $\boldsymbol{O}_t = (O_{1,t}, \ldots, O_{J-1,t})^\top$ which through matrix $\boldsymbol{A}_t$ becomes available via Inverse Discrete

Wavelet Transform (IDWT). If we consider the reference category to be the $J-th$ one, then the probability vector for k at time point t is calculated as $\boldsymbol{\pi}_{J,t} = 1/(\sum_{j=1}^{J-1} \boldsymbol{O}_{j,t})$. Then, the probabilities $\boldsymbol{\pi}_t$ can be calculated for each category $j = 1, \ldots, J-1$ as the product of the $J-th$ probability at time t and the odds ratio of the respective j category, i.e., $\pi_{j,t} = \boldsymbol{\pi}_{J,t} \cdot O_{j,t}$.

5.3.2 Multinomial spatio-temporal processes' connection to a Poisson Reduced-dimension DSTM

It is plausible that likelihoods for the multinomial case can be related to those for the Poisson, however, in terms of inference a prior matching should be done. Thus, consider the spatio-temporal process for the j category $Y_j(s,t) \sim \text{Poi}(\lambda_j(s,t))$. Then, the distribution of all categories conditional on $\sum_{j=1}^{J} Y_j(s,t) = m_t$ is multinomial distributed, i.e., $(Y_1(s,t), \ldots, Y_J(s,t)) \sim \text{Multinomial}(m_t, \xi_t)$ where $\xi_t = \boldsymbol{\lambda}_t / \sum_{j=1}^{J} \boldsymbol{\lambda}_{jt}$.

By writing the joint distribution as a product of i.i.d. Poisson distributed variables conditioned on m_t we can derive that if $\sum_{j=1}^{J} Y_j(s,t) = m_t$ then:

$$\begin{aligned} p(Y_j(s,t), = y_j(s,t), j = 1, \ldots, J | \sum_{j=1}^{J} Y_j(s,t) = m_t) &= \frac{\prod_{j=1}^{J} e^{-\lambda_j(s,t)} \lambda_j(s,t)^{Y_j(s,t)} / Y_j(s,t)!}{e^{-\sum_{j=1}^{J} \lambda_j(s,t)} (\sum_{j=1}^{J} \lambda_j(s,t))_t^m / m_t!} \\ &= \frac{m_t!}{\prod_{j=1}^{J} Y_j(s,t)} \prod_{j=1}^{J} \xi_j(s,t)^{Y_j(s,t)} \end{aligned} \quad (5.4)$$

and zero otherwise which is the required multinomial. By assuming that n_t for each location is the same and a positive integer and assume the probabilities $\pi_j(s,t)$ with $\pi_1(s,t) + \cdots + \pi_J(s,t) = 1$ then the likelihood at time t can be written as

$$\boldsymbol{L}(\boldsymbol{\pi}_t : \boldsymbol{Y}_t) = \frac{n_t}{\prod_{j,s} Y_j(s,t)} \prod_{j,s} \pi_j(s,t)^{Y_j(s,t)} \quad (5.5)$$

where $\sum_{j,s} Y_j(s,t) = n_t$. Whereas, for a spatial varying n_t, i.e., $\boldsymbol{n}_t = (n(s_1,t), \ldots, n(s_n,))$, then if each $n(s,t)$ is a positive integer and probabilities $\pi_1(s,t) + \cdots + \pi_J(s,t) = 1$ and the vectors are independent over the rows or else the locations then the likelihood at time t can be written as

$$\boldsymbol{L}(\boldsymbol{\pi}_t : \boldsymbol{Y}_t) = \prod_{s} \frac{n_(s,t)}{\prod_j Y_j(s,t)} \prod_{j} \pi_j(s,t)^{Y_j(s,t)} \quad (5.6)$$

were $\sum_j Y_j s, t = n(s,t)$ for each s and gives us the likelihood of a product multinomial.

5.3.3 Inference

As discussed, due to the non-linearity of the observations, MCMC methods are inefficient for the sampling of the states $\boldsymbol{A}_t$. Therefore, analogously to the Poisson DSTM, inference through Particle Filtering (PF) is deemed appropriate.

Thus, given the observed data $\boldsymbol{Y}_t$ and $\boldsymbol{\lambda}_t$ we can conduct inference of the states $\boldsymbol{A}_t$ via Particle Filtering (PF) by assuming the variance matrices known based on the approximation of the posterior through particle filtering with N random sampled particles—or trajectories— and via the logit transform in calculating $\boldsymbol{\pi}_t$ as well. The procedure follows a similar notion to Chapter 4, however, now the density of the observed data changes to a multinomial distribution instead of a Poisson and consequently the link function changes.

For the estimation of the spatial matrix $\boldsymbol{B}$ the full conditional posterior distributions of the Spike and Slab are the same as in Chapter 3 (equations (3.4) and (3.8)-(3.10) are considered); whereas, for the full conditional distribution of $\text{vec}(\boldsymbol{B})|\boldsymbol{A}_t, \boldsymbol{A}_{t-1}, \Sigma_\eta, \boldsymbol{R}$ we have an analogous form to the Gaussian and Poisson cases with the exception of proving the full conditional distribution in terms of the state matrix $\boldsymbol{A}_t$ instead of a vector. Based on preferable covariance structures of the scale matrices $\boldsymbol{\Sigma}_\eta$ and $\boldsymbol{R}$ Particle Metropolis Hastings steps will be incorporated.

5.3.4 Summary of the Modelling Framework

In this part, we provide a summary of the proposed approach. Specifically, in Table 5.1 we give a summary of the framework of the proposed methodology, the model, the parameters to be estimated and their relative priors, followed where the parameters will be updated via Particle Metropolis Hastings (PMH) and Conditional Particle Filtering with Ancestor Resampling (CPF-AS), along with the deterministic steps that are used for the estimation of the underlying process and the weighting function. In order to fit the model in table 5.1, it is required to update the parameters in an iterative procedure based upon their conditional distributions while conducting within Gibbs sampling a particle filtering step. In each section we will explain each of the following updates in more detail.

- Update $(\boldsymbol{A}_t, \boldsymbol{\pi}_t)|\boldsymbol{B}, \boldsymbol{\Sigma}_\eta, \boldsymbol{R}, \boldsymbol{Y}_t$ through Particle Filtering
- Update $\text{vec}(\boldsymbol{B})|\boldsymbol{A}_t, \boldsymbol{A}_{t-1}, \boldsymbol{\Sigma}_\eta, \boldsymbol{R}$ through Conditional Particle Filtering with Ancestor Resampling (CPF-AS)
- Update $(\boldsymbol{\Sigma}_\eta, \boldsymbol{R})|\boldsymbol{A}_t, \boldsymbol{A}_{t-1}, \boldsymbol{B}$ of the temporal components $\boldsymbol{A}_t$ through Particle Metropolis Hastings (PMH).
- Pseudo code based on a combination of CPF-AS and PMH for all parameters.

Data:	
Spatio-temporal process counts:	$\mathbf{Y}$, $T \times n \times J$ array
Cell Probability:	$\boldsymbol{\pi}$, $T \times n \times J$ array
Total count:	$\boldsymbol{\lambda}$, $T \times n$ vector
Linear predictor:	$\boldsymbol{Z}$, $T \times n \times J$ array
Redistribution kernel:	$\mathbf{w}$, $n \times n$ matrix
Approximations:	
$\boldsymbol{Z}_t = \boldsymbol{\Phi}\boldsymbol{A}_t$,	$\boldsymbol{A}_{T\times n\times J}$, coefficients of Wavelet matrix $\Phi_{n\times J}$
$\boldsymbol{w}_s = \boldsymbol{B}\Phi$,	$\boldsymbol{B}_{J\times n}$, coefficients of Wavelet matrix $\Phi_{n\times J}$
Model:	
$\boldsymbol{\pi}_t = \boldsymbol{A}_t\Phi^\top$	$\boldsymbol{\eta}_t \sim \text{N}(0, (\boldsymbol{\Phi}^\top\boldsymbol{\Sigma}_\eta\boldsymbol{\Phi}) \otimes \boldsymbol{R})$
$\boldsymbol{A}_t = \boldsymbol{\Phi}^\top\boldsymbol{B}\boldsymbol{A}_{t-1} + \boldsymbol{\eta}_t$	$\boldsymbol{Z}_t = \begin{bmatrix} \log\frac{\pi(1,t)}{1-\sum_{s=1}^{n-1}\pi(s,t)} \\ \vdots \\ \log\frac{\pi(n-1,t)}{1-\sum_{s=1}^{n-1}\pi(s,t)} \end{bmatrix}$
Parameters and Prior distributions:	
$\text{vec}(\boldsymbol{A}_0) \sim \text{N}(\boldsymbol{m}_0, \boldsymbol{P}_{10} \otimes \boldsymbol{P}_{20})$	$\boldsymbol{m}_0$, $\boldsymbol{P}_{10}, \boldsymbol{P}_{20}$ prior mean and covariances respectively.
$\text{vec}(\boldsymbol{B})|\boldsymbol{\Gamma} \sim \text{N}(\mathbf{0}, \boldsymbol{\Gamma} \otimes \boldsymbol{I})$	$\boldsymbol{\Gamma} = \text{diag}\{\gamma_1, \ldots, \gamma_k\}$, $\gamma_k = \rho_k\tau_k^2$
$\rho_k|v_0, q \sim (1-q)\delta_{v_0}(.) + q\delta_1(.)$	$q \sim \text{U}(0,1)$
$\tau_k^{-2}|\omega_1, \omega_2 \sim \text{G}(\omega_1, \omega_2)$	$\boldsymbol{\beta}_k \sim \text{N}(\mathbf{0}, \gamma_k\boldsymbol{I})$
$\boldsymbol{\Sigma}_\eta \sim \text{IW}(\nu_1, \boldsymbol{Q}_1)$	$\boldsymbol{R} \sim \text{IW}(\nu_2, \boldsymbol{Q}_2)$

Table 5.1: Framework of the model

5.4 Updating the parameters

5.4.1 Updating $\boldsymbol{A}_t, \boldsymbol{\pi}_t | \boldsymbol{B}, \boldsymbol{\Sigma}_\eta, \boldsymbol{R}, \boldsymbol{Y}_t$

Given the observed data $\boldsymbol{Y}_t$ and the total counts $\boldsymbol{\lambda}_t$ we can conduct inference of the state matrix $\boldsymbol{A}_t$ by assuming the same procedure as in Chapter 4. The importance density $q(\boldsymbol{A}_0|\boldsymbol{Y}_0)$ will be considered the same as the prior distribution $p(\boldsymbol{A}_0)$ which brings us into the bootstrap particle filtering approach. The weights of the particles will be updated analogously as in via (4.8) and then normalised, while multinomial resampling will be conducted in the case of low efficient sample size N_{eff}. Furthermore, for the calculation of the particle weights, the multinomial density in (5.1) will be used.

Initial step:
Simulate N particles $\boldsymbol{A}_0^{(1)}, \ldots, \boldsymbol{A}_0^{(N)}$ from $p(\boldsymbol{A}_0)$
Calculate $\boldsymbol{\pi}_0^{(1)}, \ldots, \boldsymbol{\pi}_0^{(N)}$
Set $w_0^{(i)} 1/N$, $i = 1, \ldots, N$
Particle Sampling:
For $t = 1, \ldots, T$:
Sample $\boldsymbol{A}_t^{(1)}, \ldots, \boldsymbol{A}_t^{(N)}$ from the importance function $g(\boldsymbol{A}_t|\boldsymbol{A}_{t-1}^{(i)}, Y_t)$
Calculate $\boldsymbol{\pi}_t^{(i)} = \frac{\exp(\boldsymbol{\Phi}\boldsymbol{A}_t^{(i)})}{(1+\exp(\boldsymbol{\Phi}\boldsymbol{A}_t^{(i)}))}$
Calculate the weights $\tilde{w}_t^{(i)}$ from (4.9)
Normalise the weights $w_t^{(i)} = \frac{\tilde{w}_t^{(i)}}{\sum_{i=1}^N \tilde{w}_t^{(i)}}$
Resampling step: Multinomial Resampling
Calculate the effective sample size $N_{eff} = (\sum_{i=1}^N (w_t^{(i)})^2)^{-1}$
Draw N indices $i_1, \ldots, i_N$ from the discrete distribution $P(\boldsymbol{A}_t = \boldsymbol{A}_t^{(i)}) = w_t^{(i)}$
Relabel the sample $\boldsymbol{A}_t^{(i)} = \boldsymbol{A}_t^{(i_j)}$, for $i = 1, 2, \ldots, N$
Update to equal weights by $w_t^{(i)} = 1/N$
Posterior Estimation Approximate the posteriors
$\hat{p}(\boldsymbol{A}_t|\boldsymbol{Y}_{1:t}) = \sum_{i=1}^N w_t^{(i)} \delta(\boldsymbol{A}_t - \hat{\boldsymbol{A}}_t)$, where $\hat{\boldsymbol{A}}_t = \sum_{i=1}^N w_t^{(i)} \boldsymbol{A}_t^{(i)}$
$\hat{p}(\boldsymbol{\lambda}_t|\boldsymbol{Y}_{1:t}, \boldsymbol{A}_t) = \sum_{i=1}^N w_t^{(i)} \delta(\boldsymbol{\pi}_t - \hat{\boldsymbol{\pi}}_t)$, where $\hat{\boldsymbol{\pi}}_t = \sum_{i=1}^N w_t^{(i)} \boldsymbol{\pi}_t^{(i)}$

Table 5.2: Bootstrap Particle Filtering Pseudo Code for $\boldsymbol{A}_t$ and $\boldsymbol{\lambda}_t$ for the Multinomial DSTM (5.3) for known $\boldsymbol{B}$, $\boldsymbol{\Sigma}_\eta$ and $\boldsymbol{R}$.

Finally, since $\boldsymbol{\pi}_t$ and $\boldsymbol{A}_t$ are related, during the sampling of the N particles of $\boldsymbol{A}_t$, the cell probabilities $\boldsymbol{\pi}_t$ will be calculated so that we can derive the approximate filtered

estimates, i.e., $\hat{\boldsymbol{\pi}}_t|\boldsymbol{A}_t = \sum_{i=1}^{N} w^{(i)}\boldsymbol{\pi}_t^{(i)}$. The general particle filter algorithm for $\boldsymbol{A}_t$ and $\boldsymbol{\pi}_t$ is summarised in Table 5.2 by considering that the parameters $\boldsymbol{B}$, $\boldsymbol{\Sigma}_\eta$ and $\boldsymbol{R}$ are known.

5.4.2 Updating $\text{vec}(\boldsymbol{B})|\boldsymbol{A}_t, \boldsymbol{A}_{t-1}, \Sigma_\eta, \boldsymbol{R}$

In order to reduce the complexity of calculations we will rewrite the conditional likelihood of $\boldsymbol{A}_t$ (Shumway and Stoffer, 2000) in a vectorised form. Thus, by setting $\boldsymbol{H} = (\boldsymbol{\Phi}^\top\boldsymbol{\Sigma}_\eta\boldsymbol{\Phi}) \otimes \boldsymbol{R}$ and by taking into advantage of the conditional independence of $\boldsymbol{A}_t$ we can write

$$
\begin{aligned}
p(\text{vec}(\boldsymbol{A}_{1:T})|\boldsymbol{Y}_{1:T}, \boldsymbol{B}, \boldsymbol{J}) &\propto p(\text{vec}(\boldsymbol{A}_T)|\boldsymbol{Y}_T, \boldsymbol{B}, \boldsymbol{J}) p(\text{vec}(\boldsymbol{A}_{1:T-1})|\boldsymbol{Y}_{1:T-1}, \boldsymbol{B}, \boldsymbol{H}) \\
&\propto \prod_{t=2}^{T} \exp\Big[\big(\text{vec}(\boldsymbol{A}_t) - (\boldsymbol{A}_{t-1}^\top \otimes \boldsymbol{\Phi})\text{vec}(\boldsymbol{B})\big)^\top \boldsymbol{H}^{-1} \\
&\quad \times \big(\text{vec}(\boldsymbol{A}_t) - (\boldsymbol{A}_{t-1}^\top \otimes \boldsymbol{\Phi})\text{vec}(\boldsymbol{B})\big) \Big] \\
&= \exp\Bigg[\sum_{t=2}^{T} \Big(\text{vec}(\boldsymbol{A}_t)^\top \boldsymbol{H}^{-1}\text{vec}(\boldsymbol{A}_t) - 2\text{vec}(\boldsymbol{A}_t)^\top \boldsymbol{H}^{-1}(\boldsymbol{A}_{t-1} \otimes \boldsymbol{\Phi})\text{vec}(\boldsymbol{B}) \\
&\quad + \text{vec}(\boldsymbol{B})^\top (\boldsymbol{A}_{t-1} \otimes \boldsymbol{\Phi}^\top)\boldsymbol{H}^{-1}(\boldsymbol{A}_{t-1} \otimes \boldsymbol{\Phi})\text{vec}(\boldsymbol{B}) + \text{vec}(\boldsymbol{A}_t)^\top \boldsymbol{H}\text{vec}(\boldsymbol{A}_t) \Big) \Bigg] \\
&\propto \exp\Bigg[\sum_{t=2}^{T} \Big(\text{vec}(\boldsymbol{A}_t)^\top \boldsymbol{H}^{-1}\text{vec}(\boldsymbol{A}_t) - 2\text{vec}(\boldsymbol{A}_t)^\top \boldsymbol{H}^{-1}(\boldsymbol{A}_{t-1} \otimes \boldsymbol{\Phi})\text{vec}(\boldsymbol{B}) \\
&\quad + \text{vec}(\boldsymbol{B})^\top (\boldsymbol{A}_{t-1} \otimes \boldsymbol{\Phi}^\top)\boldsymbol{H}^{-1}(\boldsymbol{A}_{t-1} \otimes \boldsymbol{\Phi})\text{vec}(\boldsymbol{B}) \Big) \Bigg] \qquad (5.7)
\end{aligned}
$$

which will help us in the update of the coefficient matrix $\boldsymbol{B}$.

Updating $\boldsymbol{B}|\boldsymbol{A}_t, \boldsymbol{\Sigma}_\eta, \boldsymbol{R}, \boldsymbol{\Gamma}$ Under the multivariate normal prior set on $\text{vec}(\boldsymbol{B})$ with a mean zero vector, i.e., $\text{vec}(\boldsymbol{B}) \sim \text{N}(\boldsymbol{0}, \boldsymbol{\Gamma} \otimes \boldsymbol{I})$, with $\boldsymbol{\Gamma} = \text{diag}(\gamma_1, \ldots, \gamma_K)$ and $\gamma_k = \rho_k\tau_k^2$, the posterior distribution is derived as the product of the likelihood function (5.7) and the prior $p(\text{vec}(\boldsymbol{B})|\boldsymbol{\Gamma})$, i.e.,

$$
\begin{aligned}
p(\boldsymbol{B}|\boldsymbol{\Sigma}_\eta, \boldsymbol{A}_{1:T}, Y_{1:T}) &\propto p(\text{vec}(\boldsymbol{A}_{1:T})|\boldsymbol{Y}_{1:T}, \boldsymbol{B}, \boldsymbol{\Sigma}_\eta, \boldsymbol{R}) p(\text{vec}(\boldsymbol{B})|\boldsymbol{\Gamma}) \\
&= \exp\Bigg[\sum_{t=2}^{T}\Bigg(-2\text{vec}(\boldsymbol{A}_t)^\top \boldsymbol{H}^{-1}(\boldsymbol{A}_{t-1}\otimes\boldsymbol{\Phi})\text{vec}(\boldsymbol{B}) \\
&\quad +\text{vec}(\boldsymbol{B})^\top(\boldsymbol{A}_{t-1}\otimes\boldsymbol{\Phi}^\top)\boldsymbol{H}^{-1}(\boldsymbol{A}_{t-1}\otimes\boldsymbol{\Phi})\text{vec}(\boldsymbol{B})\Bigg) \\
&\quad +\text{vec}(\boldsymbol{B})^\top(\boldsymbol{\Gamma}^{-1}\otimes\boldsymbol{I})\text{vec}(\boldsymbol{B})\Bigg] \\
&= \exp\Bigg[\text{vec}(\boldsymbol{B})^\top\Bigg(\sum_{t=2}^{T}\big((\boldsymbol{A}_{t-1}\otimes\boldsymbol{\Phi}^\top)\boldsymbol{H}^{-1}(\boldsymbol{A}_{t-1}\otimes\boldsymbol{\Phi}+(\boldsymbol{\Gamma}\otimes\boldsymbol{I})^{-1}-2\boldsymbol{C}\big)\text{vec}(\boldsymbol{B})\Bigg]
\end{aligned}
\tag{5.8}
$$

where $\boldsymbol{C} = \text{vec}(\boldsymbol{A}_t)^\top \boldsymbol{H}^{-1}(\boldsymbol{A}_{t-1}\otimes\boldsymbol{\Phi})$ and $\boldsymbol{H} = (\boldsymbol{\Phi}^\top\boldsymbol{\Sigma}_\eta\boldsymbol{\Phi})\otimes\boldsymbol{R}$. Thus, (5.8) is the exponential part of a multivariate Normal distribution with a mean vector and covariance matrix dependent on $\boldsymbol{A}_t$, i.e., $\boldsymbol{B}|\boldsymbol{A}_{1:T}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}_\eta, \boldsymbol{R} \sim \text{N}(\tilde{\boldsymbol{\mu}}, \boldsymbol{D})$ where $\tilde{\boldsymbol{\mu}} = \boldsymbol{\mu}_1 * \boldsymbol{D}$ with $\boldsymbol{\mu}_1 = \sum_{t=2}^{T}\big((\boldsymbol{A}_{t-1}\otimes\boldsymbol{\Phi}^\top)^{-1}(\boldsymbol{A}_{t-1}\otimes\boldsymbol{\Phi})\big)$ and $\boldsymbol{D} = (\boldsymbol{\mu}_1 + (\boldsymbol{\Gamma}\otimes\boldsymbol{I})^{-1})^{-1}$.

The mean vector $\tilde{\boldsymbol{\mu}}$ indicates that the contribution of one location to another will be affected and expanded by the state matrix at time t and $t-1$. That means that the categories of multinomial distribution have a spatial effect on the propagation of the processes. Finally, the magnitude of scaling will be affected by the variation of the categories as well.

5.4.3 Updating $\boldsymbol{\Sigma}_\eta, \boldsymbol{R}|\boldsymbol{A}_t, \boldsymbol{A}_{t-1}, \boldsymbol{B}$

As we are unaware of the scale matrices structure under this model, full structure on both scale matrices will be assumed. Specifically, we consider independent Inverse Wishart priors, i.e., $\boldsymbol{\Sigma}_\eta \sim \text{IW}(\nu_1, \boldsymbol{Q}_1)$ and $\boldsymbol{R} \sim \text{IW}(\nu_2, \boldsymbol{Q}_2)$. The state matrix $\boldsymbol{A}_t$ is a matrix normal distributed variable, or else, $\text{vec}(\boldsymbol{A}_t) \sim \text{N}(\boldsymbol{\Phi}^\top\boldsymbol{B}\text{vec}(\boldsymbol{A}_{t-1}), (\boldsymbol{\Phi}^\top\boldsymbol{\Sigma}_\eta\boldsymbol{\Phi})\otimes\boldsymbol{R}))$. Thus, the full conditional posterior of the joint vector $(\boldsymbol{\Sigma}_\eta, \boldsymbol{R})|\boldsymbol{A}, \boldsymbol{B}$ can be written as the product of the priors and the likelihood of $\boldsymbol{A}$. By defining again $= (\boldsymbol{\Phi}^\top\boldsymbol{\Sigma}_\eta\boldsymbol{\Phi})\otimes\boldsymbol{R}$ we can then write:

$$
\begin{aligned}
p(\boldsymbol{\Sigma}_\eta, \boldsymbol{R}|\boldsymbol{A}, \boldsymbol{B}) &\propto \prod_{t=2}^{T} p(\boldsymbol{A}_t|\boldsymbol{A}_{t-1}, \boldsymbol{B}, \boldsymbol{\Sigma}_\eta, \boldsymbol{R}) \times p(\boldsymbol{\Sigma}_\eta, \boldsymbol{R}) \\
&\propto \prod_{t=2}^{T} p(\boldsymbol{A}_t|\boldsymbol{A}_{t-1}, \boldsymbol{B}, \boldsymbol{\Sigma}_\eta, \boldsymbol{R}) \times p(\boldsymbol{\Sigma}_\eta) \times p(\boldsymbol{R}) \\
&= |\boldsymbol{H}|^{-(T-1)/2} |\boldsymbol{\Sigma}_\eta|^{(\nu_1+n+1)/2} |\boldsymbol{R}|^{(\nu_2+n+1)/2} \\
&\times \exp\big(-\frac{1}{2} tr(\boldsymbol{Q}_1 \boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{-1} + \boldsymbol{Q}_2 \boldsymbol{R}^{-1})\big) \\
&\times \exp\left(-\frac{1}{2} \sum_{t=2}^{T} \mathrm{tr}\big[(\boldsymbol{A}_t - \boldsymbol{\Phi}^\top \boldsymbol{B} \boldsymbol{A}_{t-1})^\top \boldsymbol{H}^{-1} (\boldsymbol{A}_t - \boldsymbol{\Phi}^\top \boldsymbol{B} \boldsymbol{A}_{t-1})\big]\right)
\end{aligned}
\tag{5.9}
$$

The quantity in (5.9) cannot derive a known distribution and therefore the use of Metropolis-Hastings steps will have to be included in the particle filtering algorithm. Furthermore, the unbiased estimation of the likelihood is needed in order to approximate the posterior distribution of the parameter. Under a Random Walk Metropolis-Hastings and by considering a symmetric proposal in each iterative step the likelihood can be estimated as

$$
\log \hat{p}_\theta^N(\boldsymbol{A}_{1:t}^*) = \log \hat{p}_\theta^N(\boldsymbol{A}_{1:t-1}[m]) + \{w_{max} + \sum_{i=1}^{N} w_t^i - \log N\} \tag{5.10}
$$

where $\boldsymbol{A}_{1:t}[m]$ indicates the sampled matrix from the particle filtering at iteration m and therefore can be used in order to calculate the acceptance ratio.

5.4.4 Summary of the model and pseudo code

In this part, we provide a summary of the proposed approach. Firstly, in Table 5.1 a summary of the framework of the proposed methodology along with the model, the parameters to be estimated and their relative priors are provided. Then, in Table 5.2 the CPF-AS combined with PMH procedure of the proposed methodology is described along with the deterministic steps that are used for the estimation of the underlying process and the weighting function.

To sum up, the inferential stage is consisted of an adaptive CPF-AS for the Spike and Slab hierarchy parameters through the posterior densities (3.8) to (3.10) for the

parameters ρ_k, τ_k and Q with the final update on the matrix $\boldsymbol{B}$ through the posterior (3.4). In the meantime, Metropolis Hastings steps are conducted for the estimation of the scale matrices $\boldsymbol{\Sigma}_\eta$ and $\boldsymbol{R}$ with the posterior being (5.9). Then, the weight function $w_s(u)$ is calculated deterministically from the Inverse Discrete Wavelet Transform (IDWT). Furthermore, the underlying process' coefficient matrix $\boldsymbol{A}_t$ is inferred in each Gibbs iteration through PF. Finally, the underlying cell probabilities $\boldsymbol{\pi}_t$ are again deterministically calculated under the IDWT framework.

Initial step:
Initialise at $m = 1$:
Set $\text{vec}(\boldsymbol{B})[1]$ arbitrarily or through the prior $p(\text{vec}(\boldsymbol{B}))$
Set $(\boldsymbol{\Sigma}_\eta, \boldsymbol{R})$ arbitrarily or through the prior $p(\boldsymbol{\Sigma}_\eta, \boldsymbol{R})$
Run Algorithm 1 and estimate the likelihood (5.10)
Particle Sampling:
For $m = 2, \ldots, M$:
Implement Algorithm 5.1 with $\boldsymbol{B} = \boldsymbol{B}[m-1]$
and $(\boldsymbol{\Sigma}_\eta, \boldsymbol{R})_{prop} \vert (\boldsymbol{\Sigma}_\eta, \boldsymbol{R})[m-1]$ $\boldsymbol{\Sigma}_\eta$
Draw $\omega \sim \text{C}(\{w_T\}_{i=1}^N)$ and output the trajectory $\boldsymbol{A}_{0:T}^{\omega}[m]$
Extract the likelihood (5.10)
Gibbs sampling step:
Draw $\boldsymbol{B}[m]$ from $\text{vec}(\boldsymbol{B}) \sim p(\text{vec}(\boldsymbol{B}) \vert \boldsymbol{\alpha}_{0:T}^{\omega}[m], \boldsymbol{\pi}_{0:T}^{\omega}[m])$
Calculate $\boldsymbol{w}$ through $\boldsymbol{B}[m]$ via IDWT
Calculate $\boldsymbol{\pi}_{0:T}^{\omega}[m] \vert \boldsymbol{\alpha}_{0:T}^{\omega}[m]$ via IDWT
Metropolis-Hastings Acceptance step:
Calculate the log-likelihood difference between $(\boldsymbol{\Sigma}_\eta, \boldsymbol{R})_{prop}$ and $(\boldsymbol{\Sigma}_\eta, \boldsymbol{R})[m-1]$
Sample $u \sim \text{U}(0, 1)$:
if $u <$ acceptance probability then update $(\boldsymbol{\Sigma}_\eta, \boldsymbol{R})[m] = (\boldsymbol{\Sigma}_\eta, \boldsymbol{R})_{prop}$
else $(\boldsymbol{\Sigma}_\eta, \boldsymbol{R})[m] = (\boldsymbol{\Sigma}_\eta, \boldsymbol{R})[m-1]$

Table 5.3: Conditional Particle Filtering Pseudo Code for $\boldsymbol{\alpha}_t$ and $\boldsymbol{\pi}_t$ and $\boldsymbol{B}$ for the Multinomial DSTM (5.3).

5.5 Posterior Predictive Distribution

Fore the predictive distribution under the Multinomial DSTM, consider the temporal ℓ-step ahead forecast at the monitoring locations. Then, by defining the $n \times J$ matrix $\boldsymbol{D}_t = \{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_t\}$ and have obtained the samples from the Particle Filtering scheme, then, for any positive integer ℓ, the ℓ-step ahead predictive distribution for the model in (5.3) is

$$p(\boldsymbol{Y}_{t+\ell}|\boldsymbol{D}_t) = \int p(\boldsymbol{Y}_{t+\ell}|\boldsymbol{A}_{t+\ell})p(\boldsymbol{A}_{t+\ell}|\boldsymbol{D}_t)d\boldsymbol{A}_{t+\ell} \tag{5.11}$$

and is approximated by

$$\hat{p}(\boldsymbol{Y}_{t+\ell}|\boldsymbol{D}_t) = \sum_{i=1}^{N} p(\boldsymbol{Y}_{t+\ell}|\boldsymbol{A}_{t+\ell}^{(i)})\boldsymbol{w}_t^{(i)}, \tag{5.12}$$

where by writing recurrently the evolution of $\boldsymbol{A}_{t+\ell}$ as $\boldsymbol{A}_{t+\ell} = (\boldsymbol{\Phi}^\top \boldsymbol{B})^\ell \boldsymbol{A}_t + \sum_{h=1}^{\ell} \boldsymbol{\eta}_{t+h}$, we use $\boldsymbol{A}_{t+\ell}^{(i)} = (\boldsymbol{\Phi}^\top \boldsymbol{B})^\ell \boldsymbol{A}_t^{(i)}$, for $i = 1, \ldots, N$.

Therefore, in each MCMC m-th iteration we acquire the samples of the particle estimates for $\boldsymbol{A}_{t+\ell}$ sampling of non-dynamic unknown parameters conducted in the previous iteration. During the Particle Filtering steps, we calculate $\boldsymbol{\pi}_{t+\ell}^{(i)}$ and thus sample $\boldsymbol{Y}_{t+\ell}$ where its distribution is approximated by the summation stated above.

By considering now the new ungauged spatial vector of length ℓ, at an observed time point $t \in T$, i.e., $\tilde{\boldsymbol{Y}}_t = (\tilde{Y}_t(s_1), \ldots, \tilde{Y}_t(s_\ell))^\top$. In this case we shall use the posterior predictive distribution of the vector of link *logit* of probability vectors $\boldsymbol{\pi}_t$, i.e., $\tilde{\boldsymbol{Z}}_t$ which is written as

$$p(\tilde{\boldsymbol{Z}}_t|\boldsymbol{Y})) = \int_{\boldsymbol{\theta}} p(\log(\tilde{\boldsymbol{Z}}_t)|\boldsymbol{Y}, \boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{Y})d\boldsymbol{\theta} \tag{5.13}$$

where again $\boldsymbol{\theta}$ is the parameter vector associated to the link equation in (5.3). Thus, the distribution in (5.13) can be written as

$$p(\tilde{\boldsymbol{Z}}_t|\boldsymbol{Y}) = \int \cdots \int p(\tilde{\boldsymbol{Z}}_t|\boldsymbol{Y}, \boldsymbol{A})p(\boldsymbol{A}, |\boldsymbol{Y})d\boldsymbol{A}. \tag{5.14}$$

where again, a spatial interpolation can be achieved under a different basis setting.

5.6 Prior matching between a Poisson and a Multinomial Reduced-dimension DSTM

In order for the posterior distributions between the Multinomial Reduced-dimension DSTM and the modelling framework of i.i.d. Poisson distributed spatio-temporal processes to be equivalent, some prior matching should be done.

Consider the prior distribution of the $n \times J$ state matrix at time $t = 0$, i.e., $\boldsymbol{A}_0 \sim \mathrm{N}_{n\times J}(\boldsymbol{M}_0, \boldsymbol{\Sigma}_0, \boldsymbol{R}_0)$ with $\boldsymbol{M}_0$ being the prior mean matrix and $\boldsymbol{\Sigma}_0$ and $\boldsymbol{R}_0$ being the prior scale matrices for the covariance between the locations and categories respectively. This can be written in vectorised form which leads to a multivariate normal prior, i.e.

$$p(\mathrm{vec}(\boldsymbol{A}_0)) \propto \exp\left[(\mathrm{vec}(\boldsymbol{A}_0) - \mathrm{vec}(\boldsymbol{M}_0))^\top (\boldsymbol{\Sigma}_0 \otimes \boldsymbol{R}_0)^{-1} (\mathrm{vec}(\boldsymbol{A}_0) - \mathrm{vec}(\boldsymbol{M}_0))\right] \quad (5.15)$$

where $\mathrm{vec}(\boldsymbol{A}_0)$ is the $nJ \times 1$ vector.

If we now consider the i.i.d. Poisson case, then the prior of each vector state at time $t = 0$ for a category j is multivariate normally distributed. That is, for each category $\boldsymbol{\alpha}_{j0} \sim \mathrm{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ where $\boldsymbol{\alpha}_{j0}$ is a $n \times 1$ vector. Conversely, if we want to define the prior matrix normal distribution at time $t = 0$ for the $n \times J$ matrix $\boldsymbol{A}_0$ we obtain that $\boldsymbol{A}_0 \sim \mathrm{N}_{n\times J}(\boldsymbol{M}_0, \boldsymbol{\Sigma}_0, \boldsymbol{V}_0)$ where each row of $\boldsymbol{M}_0$ is equal to $\boldsymbol{\mu}_0$, i.e., $\boldsymbol{M}_0 = \mathbf{1}_{n\times n} \times \boldsymbol{\mu}_0$, $\boldsymbol{\Sigma}_0$ is the prior covariance matrix between the locations of the transition equation and $\boldsymbol{V}_0$ is the identity matrix which means the rows are independent to each other.

Thus, (5.15) is matching or else identical to the matrix normal prior distribution of the i.i.d. Poisson distributed observations only when the J categories are independent, i.e., $\boldsymbol{R}_0 = \boldsymbol{V}_0 = \boldsymbol{I}_{n\times n}$ and thus can be incorporated to the particle filtering algorithms without the estimation of $\boldsymbol{R}$.

5.7 Simulation Study

Similarly to the Gaussian and Poisson Dimension-reduced DSTM in the previous chapters, we developed an appropriate simulation by considering a Multinomial Dimension-reduced DSTM under wavelet decomposition based on the complete model in (5.3) which is described in 5.3.1. Then we are examining two cases where in both the most

important matrix of interest, $\boldsymbol{B}$, is always considered unknown. Due to our limited computational power, a few comments will be in order as in some cases we needed more iterations and thus computational power to reach convergence. Thus, we test our method by considering a discontinuity in weight function $w_s(u)$ and keeping a known covariance structure— we wish to show that our method can adapt to discontinuities and we can estimate fairly well the underlying processes.

In section 5.7.1 we introduce the simulation scheme of a Multinomial Reduced-Dimension DSTM under wavelet basis decomposition. Furthermore, as in previous chapters instead of simulating the matrix $\boldsymbol{B}$ through the Spike and Slab prior, a kernel is chosen for $w_s(u)$ and through that $\boldsymbol{B}$ is calculated through DWT. Additionally, as mentioned above, in sections 5.7.2 we conduct inference on the processes' parameters simulated under the simulation scheme in section 5.7.1.

5.7.1 Simulation of a Multinomial DSTM under Wavelet decomposition

1 Start by considering a number of equally spaced n locations in an interval $[c_1, c_2] \in D \subset \mathbb{R}$ and T time points, a Wavelet matrix $\Phi_{n\times n}$ and covariance matrices Σ_η and $\boldsymbol{R}$.

2 Consider an autocorrelation diagonal matrix $\boldsymbol{\Psi}$ and a variance σ_ψ

3 Building the weight matrix

- For each of the locations calculate d, where d is the Euclidean distance between the locations $\boldsymbol{s}$
- Choose weight function $\boldsymbol{w}$ (discontinuous or continuous) to calculate the spatial contribution
- Spatial stationary weights:

$$w^*_{i,j} = 0.9 * \frac{w_{i,j}}{\sum_{j=1}^{n} w_{i,j}}$$

4 For $t = 1$

- Calculate the coefficient matrix $\boldsymbol{B} = \boldsymbol{w}^*\boldsymbol{\Phi}^{-1}$

- Initialise $\boldsymbol{Y}_1$ and through that $\boldsymbol{\lambda}_1$
- Initialise $\boldsymbol{\pi}_1$ and normalise them
- Calculate $\boldsymbol{A}_1 = \boldsymbol{\Phi}^\top \log(\boldsymbol{O}_1)$

5 For $t \geq 2$

- $\boldsymbol{A}_t = \boldsymbol{\Phi}^\top \boldsymbol{B} \boldsymbol{A}_{t-1} + \boldsymbol{\Phi}^\top \eta_t$, $\eta_t \sim N(0, \boldsymbol{\Sigma}_\eta \otimes \boldsymbol{R})$
- Perform IDWT on $\boldsymbol{O}_t$: $\boldsymbol{O}_t = e^{\boldsymbol{\Phi} \boldsymbol{A}_t}$
- Calculate $\boldsymbol{\pi}_t$ through $\boldsymbol{O}_t$ and normalise
- Sample $\boldsymbol{\lambda}_t \sim N_{\mathbb{R}^+}(\boldsymbol{\Psi} \boldsymbol{\lambda}_{t-1}, \sigma_\psi \boldsymbol{I})$
- Simulate i.i.d $\boldsymbol{Y}_t \sim \text{Multinomial}(\boldsymbol{\lambda}_t, \boldsymbol{\pi}_t)$

A few comments are in order. High values of $\boldsymbol{A}_t$ will affect the exponential part which produces the Odds Ratios $\boldsymbol{O}_t$ and that can resort to failure on calculating and simulating the rest of the values. Thus, the parameter values should be selected with care.

5.7.2 Discontinuity in weight function $w_s(u)$

In this simulation we considered the model in (5.3) for $n = 8$ locations, $T = 64$ and $k = 3$ categories in the 1-D space $[0, 5]$. Two different kernels were considered for the weight function $w_s(u)$. Specifically, for the locations lying in $[0, 2.5]$ we considered a Gaussian kernel with mean and variance being 1 and 4 respectively, while those lying in $[2.5, 5]$ a Laplace kernel was considered with mean and rate parameters being 0 and 1 respectively, i.e.,

$$w_s(u) = \begin{cases} \text{N}(\|s-u\|^2 \,|1, 4) & \text{if } s, u \in [0, 0.25] \\ \text{Laplace}(\|s-u\|^2 \,|0, 1) & \text{if } s, u \in (2.5, 5] \\ 0 & \text{otherwise} \end{cases} \tag{5.16}$$

The temporal covariance matrix was set to be of a simple diagonal structure, i.e., $\Sigma_\eta = 2 \cdot \boldsymbol{I}$ and the categorical variation matrix was set to be

$$\boldsymbol{R} = \begin{bmatrix} 0.5 & -0.05 \\ -0.05 & 0.5 \end{bmatrix}$$

suggesting equal spread for the first two categories and a strong correlation between them. Finally, the wavelet basis that was used for the decomposition of the weight function $w_s(u)$ was a Haar basis. We conducted the combination of the Conditional Particle Filtering with Ancestor Resampling and Particle Metropolis-Hastings of Table 5.3 with N = 500 particles and M = 10^4 Gibbs iterations with a burn-in period of i = 5000 which was decided through traceplots and autocorrelation plots as diagnostic criteria. However, we are inclined that the parameters need more iterations to converge. We encountered computational difficulties as the model itself is high dimensional to run in a commercial machine. Thus, it is suggested that this simulation should be conducted for more iterations in a cluster server.

For the inferential part, we considered the Spike and Slab hyperparameters $v_0 = 0.01$ and $\omega_1 = 2$, $\omega_2 = 200$ for the point mass and variance components respectively.

By comparing the posterior mode of the state processes $\boldsymbol{A}_t$ for the first two categories and the first four locations to the simulated (real) ones in Figure 5.1 it can observed that they are being estimated fairly well. Obviously, for such a high dimensional problem, considering $N = 500$ particles does not provide us with the perfect estimations that we could get with more particles instead. Thus, we have again to note down that we need more particles and Gibbs iterations combined. This can be consecutively seen in the estimation of the cell probability processes $\boldsymbol{\pi}_t$ (Figure 5.3 and Figure 5.4) as some of the peaks and the oscillations are captured fairly well. However, there are several peaks going upwards (See Location 1, Categories 1 and 2 in Figure 5.3) that the filtered estimates own downward peaks. It is notable though that for all locations the estimated mode of all cell probabilities is around the simulated ones, with one difficulty being observed on the third category of the first location. For that reason, we investigated the behaviour of the filtered estimated for $N = 500$ and $N = 10^4$ (Figure 5.2) and we observe that the specific cell probability, even for a known matrix $\boldsymbol{B}$ is difficult to be captured. The reason behind it is that while the third category oscillates around a smaller mean cell probability than the rest of the categories (around 0.25), its peaks go up to 0.8. These high oscillations have been affected from the high variances that we chose for $\boldsymbol{R}$. The two *logit* cell probabilities share a slight negative correlation of -0.1 while each of the logarithm of odds evolves with a variance of $\text{Var}(\log \boldsymbol{O}_t) = 0.5 \cdot \boldsymbol{I}$ which that through the inverse *logit* transform provides us with a value of around $\text{Var}(\boldsymbol{\pi}_t) = 0.6 \cdot \boldsymbol{I}$ indicating a high variance for the cell probabilities.

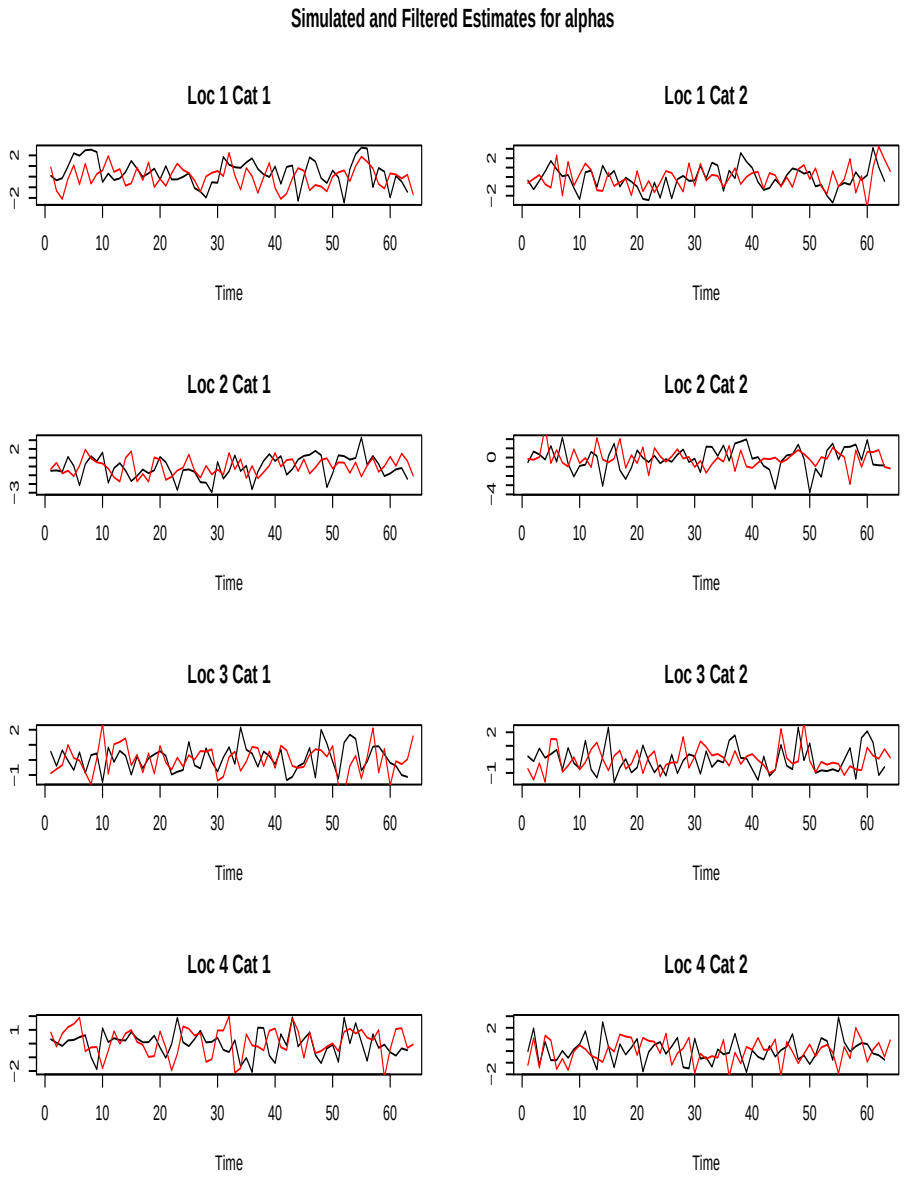

Figure 5.1: Time series plots of simulated states $\boldsymbol{A}_t$ for the first four locations (black) and the estimated filtered ones (red) for N = 500, M = 10^4, $T = 64$, $n = 8$ under a burn-in period of i = 5000.

Furthermore, Particle Filters tend to have a difficulty estimating dynamic processes with high variances. However, due to the limitation of time and computational power we were not able to investigate this matter with further simulations.

Finally, the spatial wavelet coefficients under our Spike and Slab hierarchy were estimated fairly well, with both the close to zero and non zero real values being very close to the posterior modes (Figure 5.5). Although divergence diagnostics indicated us that for most elements of $\boldsymbol{B}$ we did not reach convergence. That can be seen from the histograms of the elements of selected weight function reconstructed elements. For instance, for the weight function of location 3 affecting location 4, there is s left skewness with the mode being centered around the real value. It is believed that with more iterations the posterior estimates would have converged around the real values of the weight function and thus we would have gotten fair estimates for the spatial contribution between the locations.

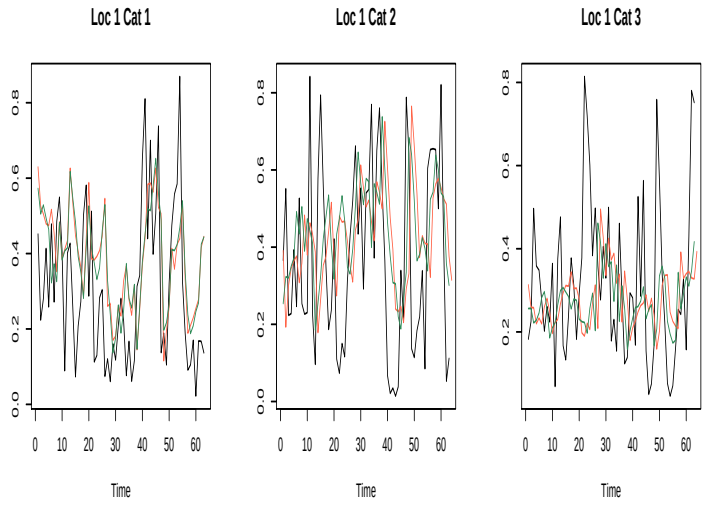

Figure 5.2: Time series plots of the simulated states $\boldsymbol{A}_t$ for the first location (black) and the estimated filtered ones under known $\boldsymbol{B}$ for N = 500 (green) and N = 10000 (orange).

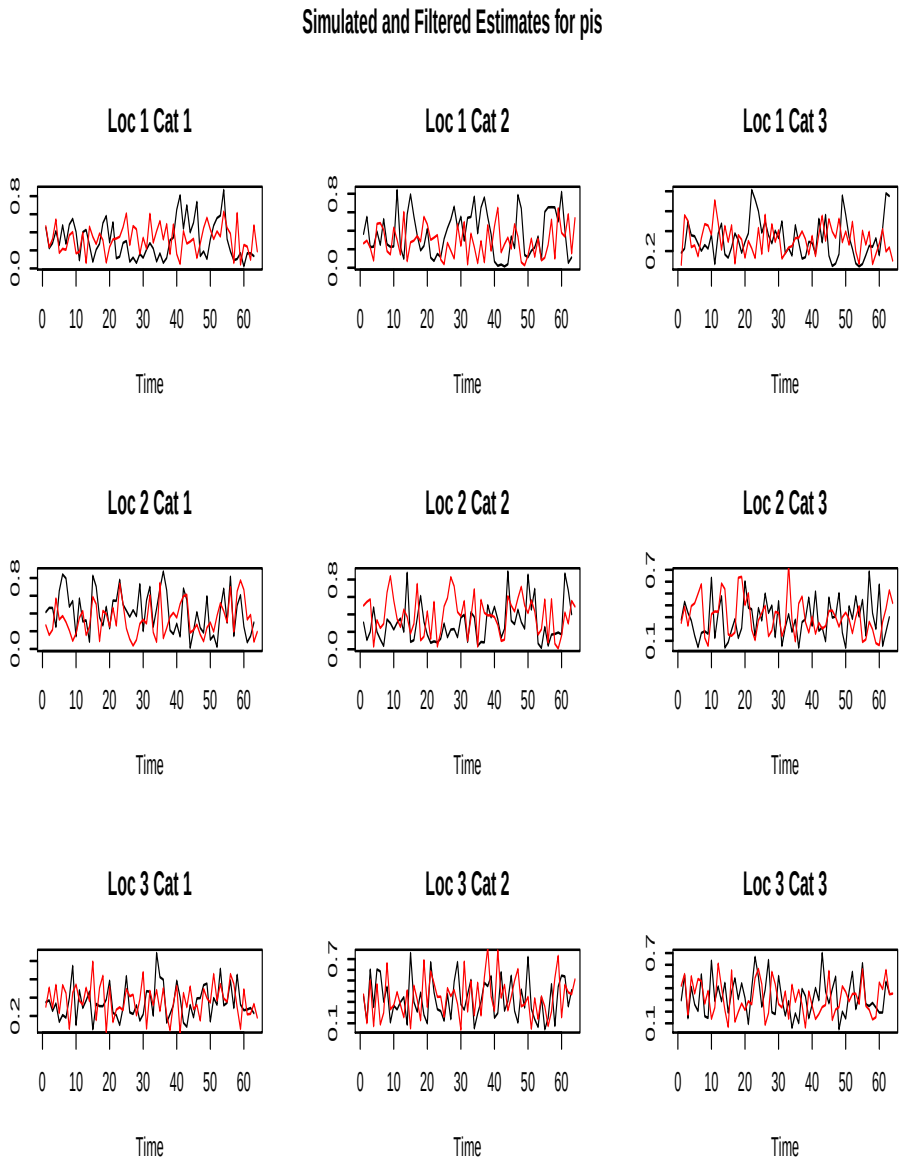

Figure 5.3: Time series plots of the simulated cell probability processes $\boldsymbol{\pi}_t$ for the first three locations (black) and the estimated filtered one (red) for N = 500, M = 10^4, $T = 64$, $n = 8$ under a burn-in period of i = 5000.

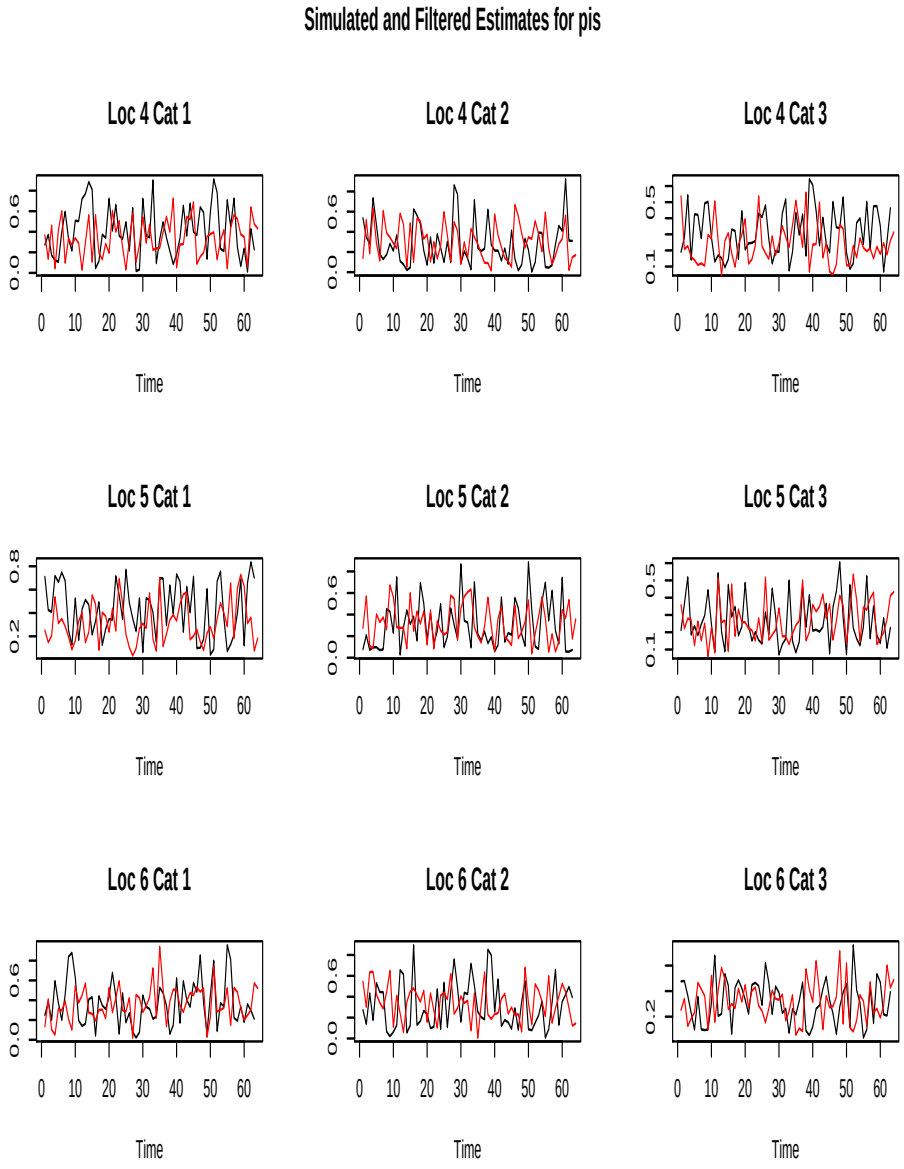

Figure 5.4: Time series plots of the simulated cell probability processes $\boldsymbol{\pi}_t$ for locations 4 to 6 (black) and the estimated filtered one (red) for N = 500, M = 10^4, $T = 64$, $n = 8$ under a burn-in period of i = 5000.

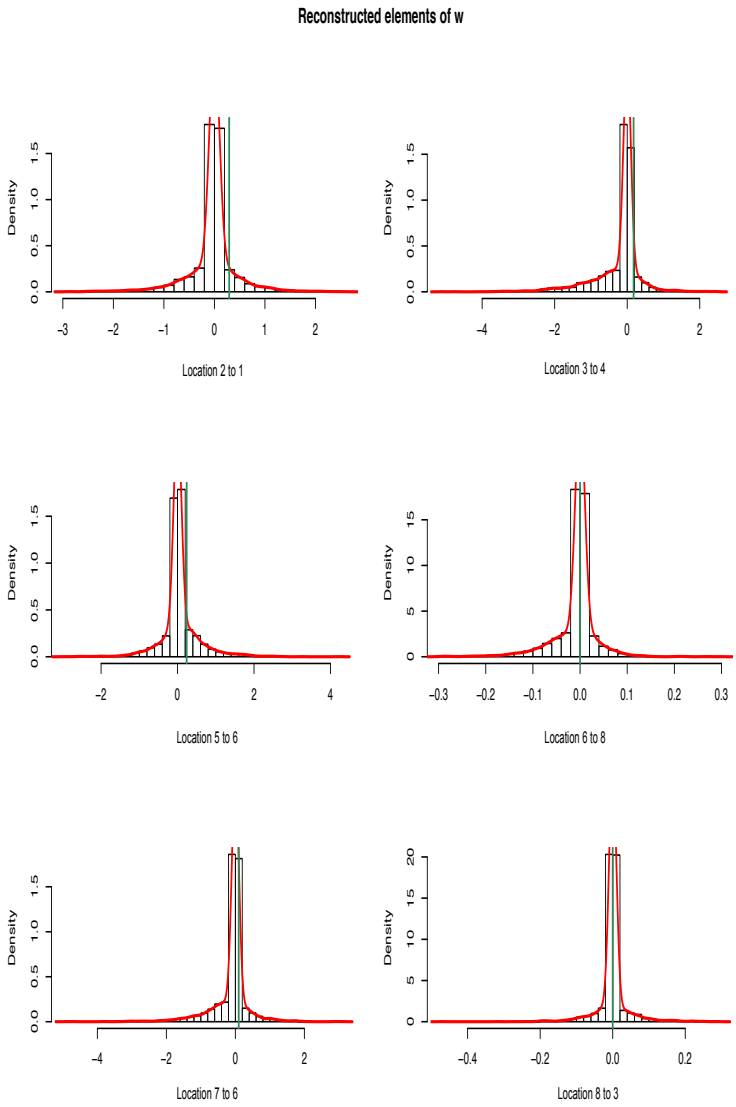

Figure 5.5: Posterior estimates for selected elements for selected reconstructed elements of $\boldsymbol{w}$. The red line indicates the empirical density and the green vertical line indicates the real value.

5.8 Application on Traffic Flows Revisited

By revisiting the traffic flow dataset, we acquired the traffic flows for the same period in M6 for all segments consisting of cars, buses and large goods vehicles (LGV). What we are interested to investigate is in each segment, dynamically, how the proportion of

cars, buses and LGVs changes over time in one segment but also how is that spatially associated to the traffic flow with the rest of segments. That would give us insight if we like to make a further analysis that will incorporate accidents based on civilian, transport and trading vehicles. Additionally, the bus and LGV data consisted as well a fair amount of missing values, which were imputed via Kalman Filtering on their logarithm prior to the analysis similarly to Chapter 4 for cars.

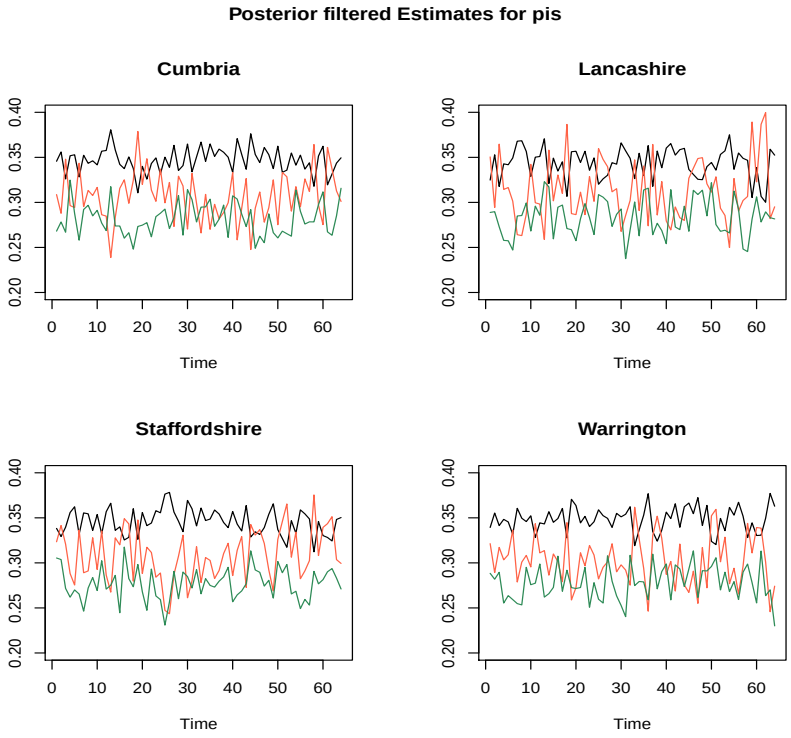

Figure 5.6: Time series plots of posterior filtered cell probabilities $\boldsymbol{\pi}_t$ for four segments. Black line signifies the cell probability for cars, orange line for LGVs and green line for buses. The estimation was conducetd for N = 500, M = 10^4, $T = 64$, $n = 8$ under a burn-in period of i = 5000.

In this analysis we are considering the inferential procedure of Table 5.2 framework of the Multinomial Dimension Reduced DSTM under a Haar wavelet basis decomposition with the model in (5.3) under the Spike and Slab prior with $v_0 = 0.1$. The hyperparameters for τ_k^{-2} under a gamma prior were set to be both equal to 2 and 100 for the shape and scale parameters respectively. Furthermore, we considered a full structure on both the scale matrices $\boldsymbol{\Sigma}_\eta$ and $\boldsymbol{R}$ with priors being $\boldsymbol{\Sigma}_\eta \sim \text{IW}(8, 2 \cdot \boldsymbol{I})$ and $\boldsymbol{R} \sim \text{IW}(3, \boldsymbol{I})$ respectively. The proposal distributions consisted of a tuning parameter $r = 100$ and

each parameter was updated through $\boldsymbol{\Sigma}_\eta[prop]|\boldsymbol{\Sigma}_\eta[m-1] \sim \text{IW}(r+8, r\boldsymbol{\Sigma}_\eta[m-1])$ and $\boldsymbol{R}[prop]|\boldsymbol{R}[m-1] \sim \text{IW}(r+3, r\boldsymbol{R}[m-1])$ which provided us with an acceptance ratio of $\approx 30\%$. In Figure 5.8 we illustrate that the spatial coefficients $\boldsymbol{B}$ have converged, however, that does not apply for the covariance parameters (Figure 5.15). Unfortunately, due to low computational power in a commercial laptop and low memory we were not able to run the model for more iterations. However, our findings, even with caution, are interesting.

Parameter	*Posterior Mode*
$\boldsymbol{\Sigma}_{\eta 11}$	0.128
$\boldsymbol{\Sigma}_{\eta 22}$	0.088
$\boldsymbol{\Sigma}_{\eta 33}$	0.086
$\boldsymbol{\Sigma}_{\eta 44}$	0.072
$\boldsymbol{\Sigma}_{\eta 55}$	0.096
$\boldsymbol{\Sigma}_{\eta 66}$	0.114
$\boldsymbol{\Sigma}_{\eta 77}$	0.106
$\boldsymbol{\Sigma}_{\eta 88}$	0.104
$\boldsymbol{R}_{11}$	0.108
$\boldsymbol{R}_{22}$	0.132
$\boldsymbol{R}_{12}$	0.058

Table 5.4: Posterior Mode estimates for the covariance elements of $\boldsymbol{R}$ and the diagonal elements of $\boldsymbol{\Sigma}_\eta$.

Considering the posterior filtered estimates of $\boldsymbol{\pi}_t$, consistently for all locations, the cell probability for buses lies lower than the cell probabilities of cars and LGVs through time (Figure 5.6 and 5.7) whereas cars share the highest cell probability. It was suspected that we would acquire a lower estimate for buses as there is a much lower number of buses in the motorway across M6. This is due to the buses being used only for mass transportation from one city or county to another. On the contrary, cars which own the highest cell probability across locations are used by civilians for transportation between home and workplace, holiday trips, excursions and so on. LGVs on the other hand, which show lesser probability than cars but more than buses, are based mostly on trading, post deliveries or removal services.

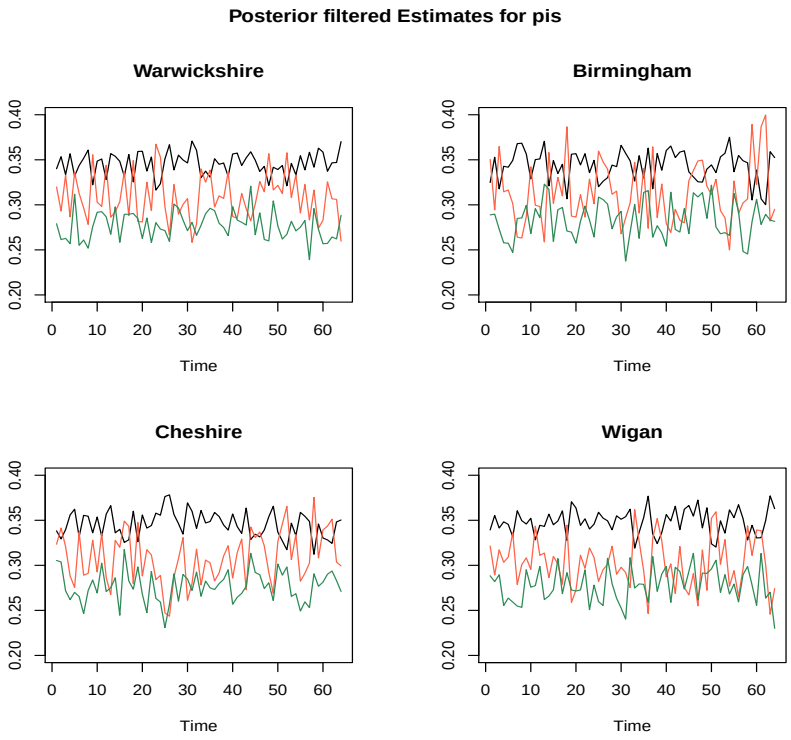

Figure 5.7: Time series plots of posterior filtered cell probabilities $\boldsymbol{\pi}_t$ for four segments. Black line signifies the cell probability for cars, orange line for LGVs and green line for buses. The estimation was conducted for N = 500, M = 10^4, $T = 64$, $n = 8$ under a burn-in period of i = 5000.

Furthermore, for all segments, the cell probabilities for cars through time lie roughly around 0.35, for LGVs around 0.3 and buses around 0.25. For specific segments, such us Birmingham, Cheshire and Lancashire, the cell probabilities for cars show a higher variability while for Lancashire and Birmingham the cell probabilities for LGVs show rapid fluctuations. The posterior modes of the categorical variances show that there is more variability for buses than cars while also they are correlated (Table 5.4).

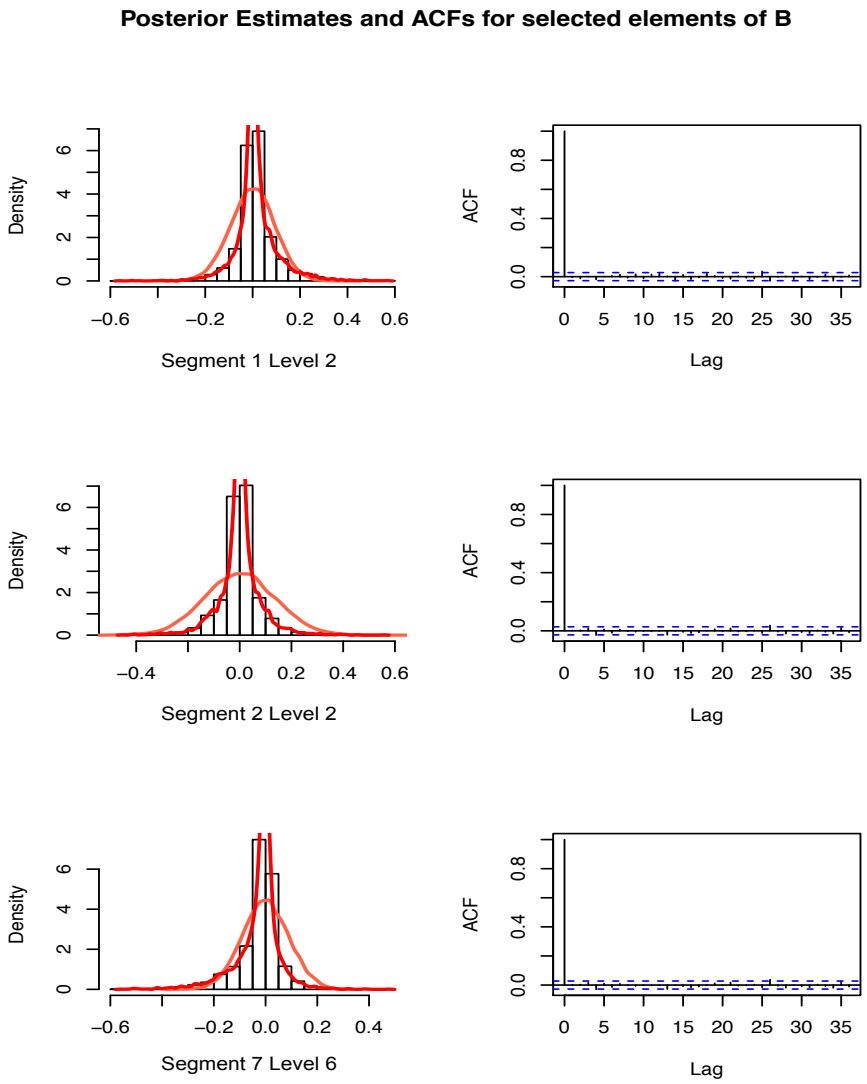

Figure 5.8: Posterior estimates for selected elements for $\boldsymbol{B}$ on the left hand side with their respective Autocorrelation Function (AFC) on the right side. The red line indicates the empirical density and the orange line indicates the prior distribution. We show that the posterior elements fo $\boldsymbol{B}$ have converged. The estimation was conducted for N = 500, M = 10^4, $T = 64$, $n = 8$ under a burn-in period of i = 5000.

The posterior elements $\boldsymbol{B}$ which have also converged (Figure 5.8) provide us with interesting findings for the weight function in terms of spatial contribution between the segments (Figure 5.9 and Figure 5.10). Specifically, there is a slight positive contribution to the traffic flows of Segment 2 and Segment 4 (Cheshire) to Segment 7 (Lancashire).

Interestingly, the application of Chapter 4 provided us too with Birmingham positively contributing to the traffic flows of motorway segment in Lancashire. That means that even if we categorise the traffic flow based on three different types on vehicles, still the flows from Birmingham will affect positively the ones in Lancashire. Cheshire and Lancashire are relatively neighbouring rural counties with Lancashire compromising more residents. That means that if civilian, public transport or trading vehicles are passing by Cheshire, they are probably leading towards Lancashire.

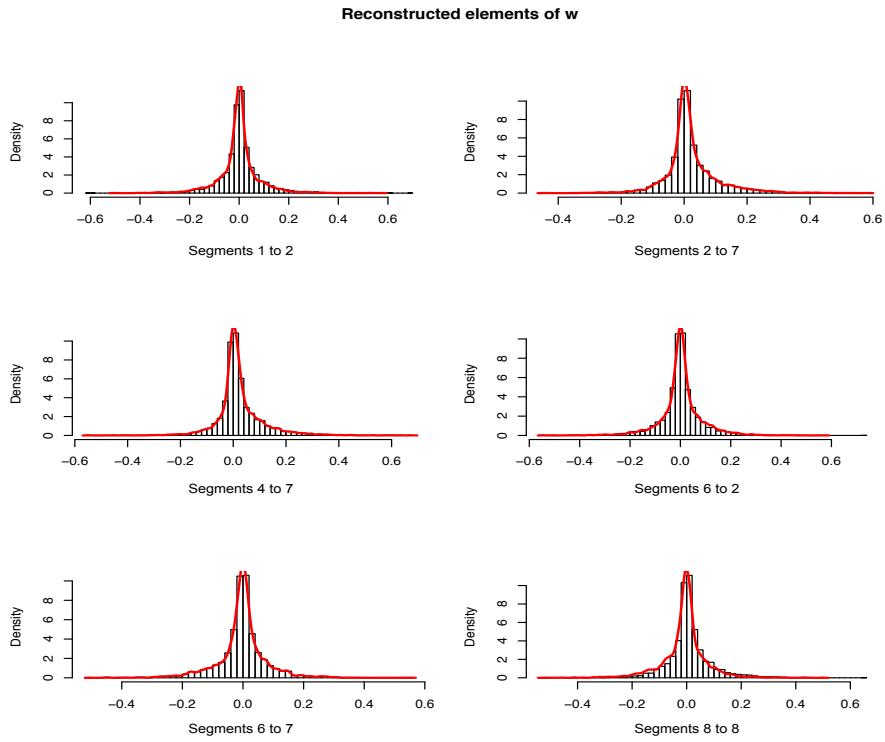

Figure 5.9: Selected reconstructed elements of $\boldsymbol{w}$. The red line indicates the empirical density estimate. The estimation was conducted for N = 500, M = 10^4, $T = 64$, $n = 8$ under a burn-in period of i = 5000.

Additionally, in Chapter 4, we observed that Segment 8 (Cumbria) was affecting negatively the flows in Segment 5 (Warrington). However, now that we consider the public transport and trading vehicles, it seems that there is no association between these two segments. Furthermore, there is a negative contribution in the traffic flows from Segment 6 (Wigan) to Segment 7 (Lancashire). Therefore, if there was a high traffic flow activity in Wigan, we would expect a lower in Lancashire. On the contrary, there is

a slight positive contribution from Segment 7 (Lancashire) to Segment 6 (Wigan) and this shows us a contrast between these two segments based on a quarterly basis.

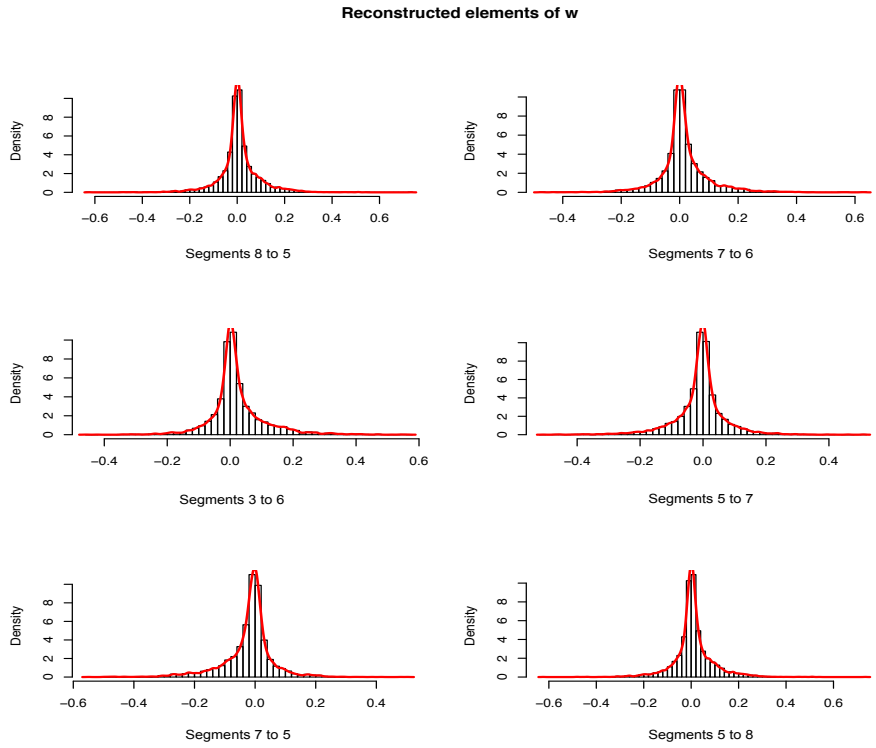

Figure 5.10: Selected reconstructed elements of $\boldsymbol{w}$. The red line indicates the empirical density estimate. The estimation was conducted for N = 500, M = 10^4, $T = 64$, $n = 8$ under a burn-in period of i = 5000.

Finally, there is a slight positive contribution from Segment 5 (Warrington) to Segment 8 (Cumbria) while in Chapter 4 we have shown that there is a strong negative contribution instead. With the addition of traffic flows for public transport and trading

vehicles we suspect that this contribution shows us the travelers and traders that wish to reach the north of England and even reach Scotland.

Regarding the posterior covariance elements of the categorical scale matrix ($\boldsymbol{R}$) in Figure 5.11, there is a higher variance for buses than cars regarding traffic flows from one quarter to another. Furthermore, there is a slight correlation between these two categories. Convergence diagnostics through the autocorrelation functions for the covariance elements suggest us that we needed more iterations for the covariance parameters to converge (Figure 5.15).

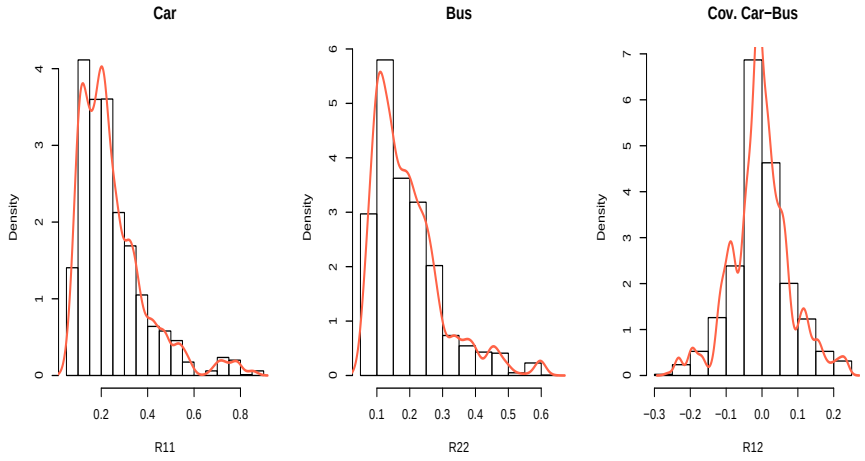

Figure 5.11: Posterior density estimates for the elements (variances and covariance) of $\boldsymbol{R}$. Red line indicates the empirical distribution of the estimates. The prior that was an inverse Wishart with $\nu_0 = 8$ and $Q_0 = 100 \cdot I$ being the shape and scale parameters for N = 500, M = 10^4, $T = 64$, $n = 8$ under a burn-in period of i = 5000. The model suggests that the two categories are correlated. Buses show a larger spread than cars.

We will proceed with caution to provide interpretation of the covariance findings. From Figure 5.12 and Table 5.4 it is suggested that Warwickshire, Wigan, Lancaster and Cumbria own a higher variability in traffic flows than the rest of segments. Indeed, from the estimates of cell probabilities in Figure 5.6 and Figure 5.7 we can observe that all three categories fluctuate way more than the rest of segments. In Chapter 4 it was suggested that the segments vary similarly, however, with the extra consideration of buses and LGVs this does not hold anymore. Furthermore, from Figures 5.14 and 5.15, it is suggested that there exist slight negative correlations between segments based on

the flows at one quarter to the next one. Specifically, there is a negative correlation between Warrington and Birmingham, Warrington and Cheshire and Warrington and Lancaster. This indicates that the traffic flows are affected by neighbouring counties, given that Warrington, Cheshire and Lancaster are next to each other. Finally, there is a slight negative correlation between Wigan and Cumbria and Lancaster and Stafford.

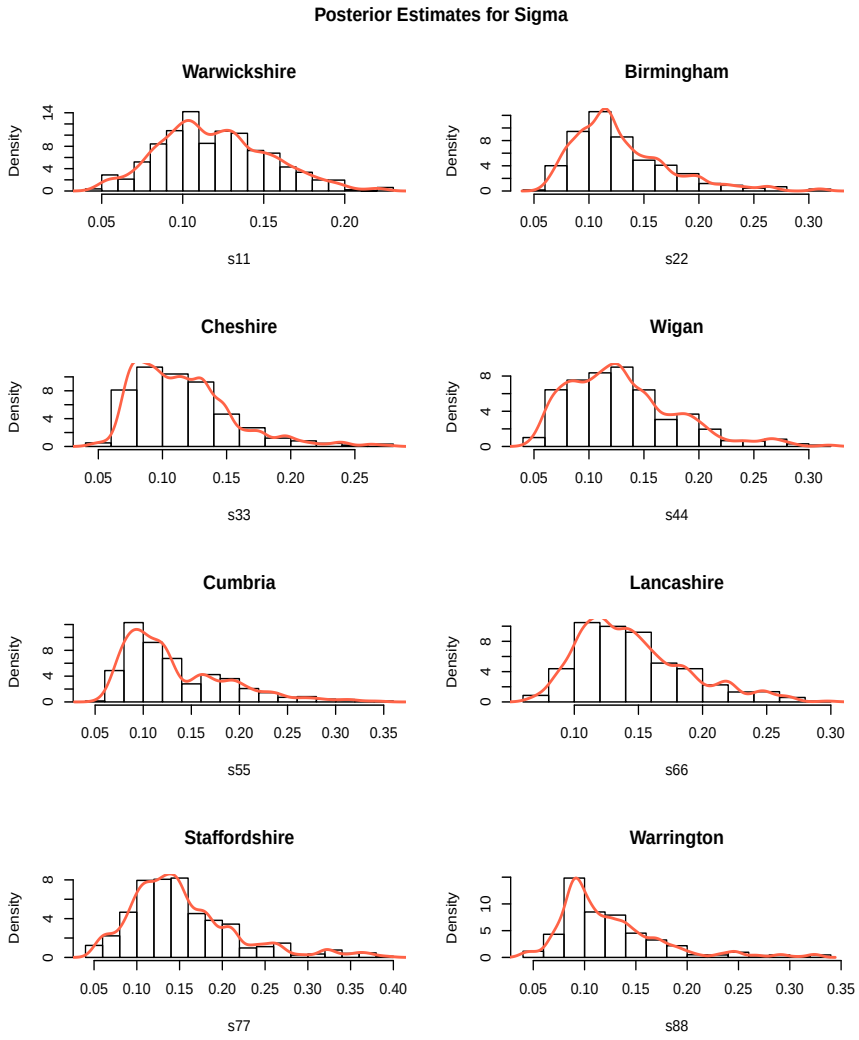

Figure 5.12: Posterior density estimates for the diagonal elements (variances) of $\mathbf{\Sigma}_\eta$. Red line indicates the empirical distribution of the estimates. The prior that was an inverse Wishart with $\nu_0 = 8$ and $Q_0 = 100 \cdot I$ being the shape and scale parameters for N = 500, M = 10^4, $T = 64$, $n = 8$ under a burn-in period of i = 5000.

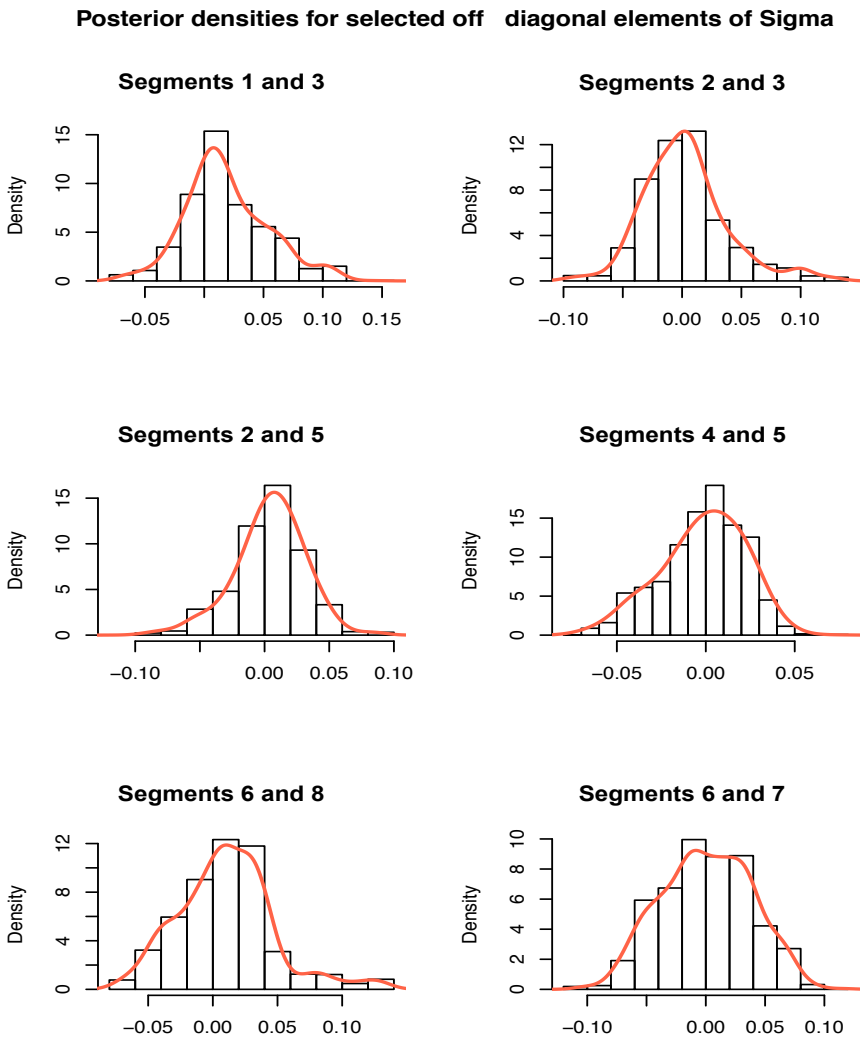

Figure 5.13: Posterior density estimates for selected off-diagonal elements (covariances) of Σ_η. Red line indicates the empirical distribution of the estimates. The prior that was an inverse Wishart with $\nu_0 = 8$ and $Q_0 = 100 \cdot I$ being the shape and scale parameters for N = 500, M = 10^4, $T = 64$, $n = 8$ under a burn-in period of i = 5000.

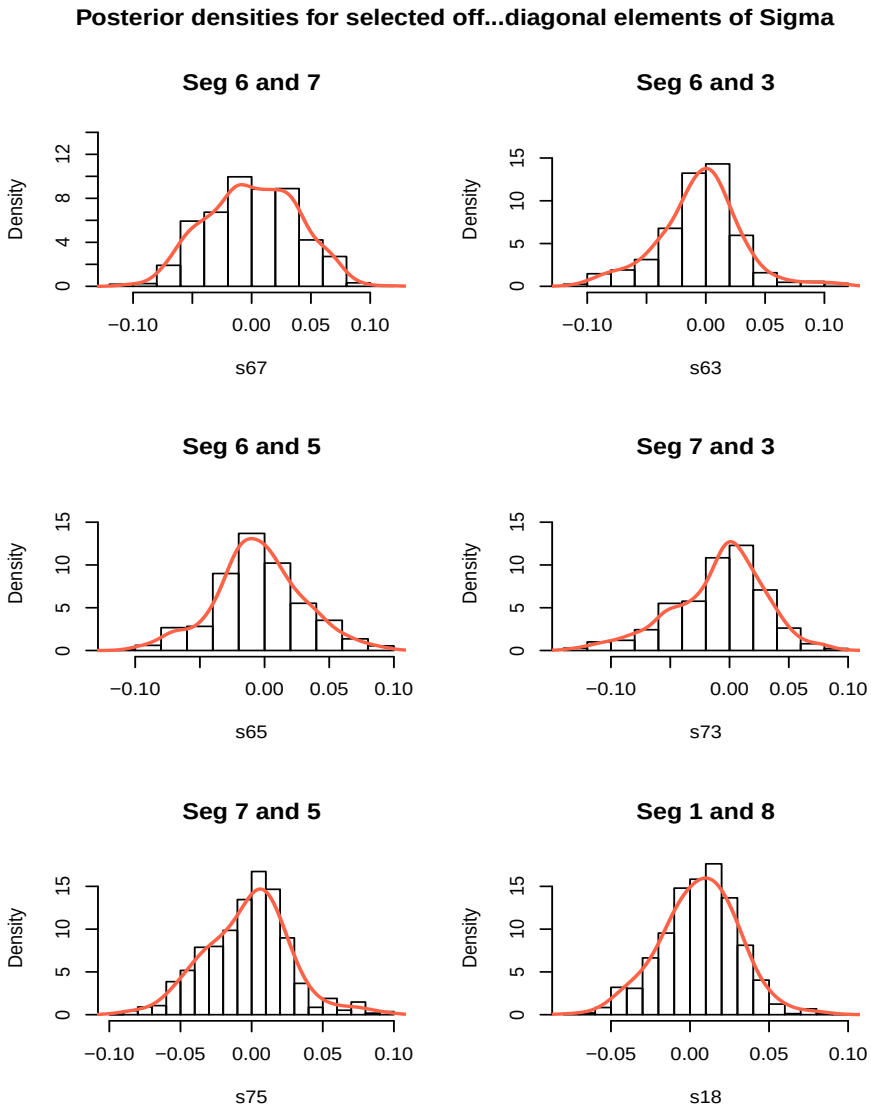

Figure 5.14: Posterior density estimates for selected off-diagonal elements (covariances) of Σ_η. Red line indicates the empirical distribution of the estimates. The prior that was an inverse Wishart with $\nu_0 = 8$ and $Q_0 = 100 \cdot I$ being the shape and scale parameters for N = 500, M = 10^4, $T = 64$, $n = 8$ under a burn-in period of i = 5000.

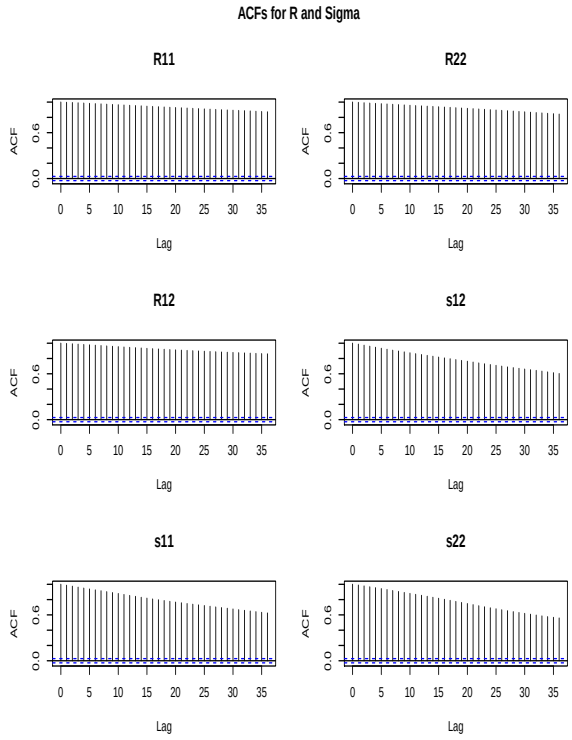

Figure 5.15: Autocorrelation functions for the covariance elements of $\boldsymbol{R}$ and selected covariance elements of $\boldsymbol{\Sigma}_\eta$. They suggest that the covariance parameters have not reached convergence. The estimation was conducted for N = 500, M = 10^4, $T = 64$, $n = 8$ under a burn-in period of i = 5000.

Missing Value Treatment The missing imputation was conducted under a non-Bayesian setting for all categories and segments. Thus, similarly as we mentioned in Chapter 4 on the missing value treatment under a Particle Filtering framework, the Multiple Imputation Particle Filter (Housfater et al., 2006) can be considered for the prediction of the missing values.

Thus, let us partition the vector of observations, i.e., $\boldsymbol{Y}_t = (\boldsymbol{Y}_t^{mis}, \boldsymbol{Y}_t^{obs})$ and let us consider the missing value index $h = 1, \ldots, m$. Then, an imputation model can be expressed as a probability distribution in order to sample the m samples subject to imputation, i.e.,

$$\boldsymbol{Y}_{ht}^{mis} \sim p(\boldsymbol{Y}_t^{mis} | \boldsymbol{Y}_{1:t}^{obs}) \tag{5.17}$$

We again assign a weight $\boldsymbol{v}_t^h$ to each imputation with $\sum_{h=1}^m v_t^h = 1$. Thus, by considering $u_t^j = (\boldsymbol{Y}_{jt}^{mis}, \boldsymbol{Y}_t^{obs})$ to be the complete data sets formed from imputed values, then

the filtering posterior distribution is given as

$$p(\boldsymbol{A}_t|\boldsymbol{Y}_{1:t}^{obs}) = \int p(\boldsymbol{A}_t|\boldsymbol{u}_{1:t-1}, \boldsymbol{Y}_t^{obs}) p(\boldsymbol{Y}_t^{mis}|\boldsymbol{Y}_{1:t}^{obs}) d\boldsymbol{Y}_t^{mis}, \tag{5.18}$$

and through Monte Carlo approximation we get

$$p(\boldsymbol{A}_t|\boldsymbol{Y}_{1:t}^{obs}) \approxeq \sum_{j=1}^{m} v_t^j p(\boldsymbol{A}_t|\boldsymbol{u}_{1:t-1}\boldsymbol{Y}_t^{obs}, u_t^h). \tag{5.19}$$

Additionally, for each of the complete data sets yields

$$p(\boldsymbol{A}_t|u_{1:t-1}, u_t^h) = \sum_{i=1}^{N} w_t^{(i,h)} \delta(\boldsymbol{A}_t - \boldsymbol{A}_t^{(i,h)}), \tag{5.20}$$

where the indexes i and h indicate the particle and imputation, respectively. Thus, an approximation of the desired posterior distribution is

$$p(\boldsymbol{A}_t|\boldsymbol{Y}_{1:t}^{obs}) \approx \sum_{j=1}^{m} \sum_{i=1}^{N} v_t^h w^{(i,h)} \delta(\boldsymbol{A}_t - \boldsymbol{A}_t^{(i,h)}). \tag{5.21}$$

Thus, estimating the missing responses $\boldsymbol{Y}_t^{mis}$ is equivalent to a posterior prediction from the model fitted to the observed data with the inclusion of a filtering estimate for the missing observations.

5.9 Conclusion

We have introduced a modeling approach for Multinomial Dimension-Reduced DSTMs under an adaptive Conditional Particle Filter procedure with Metropolis-Hastings steps. This approach is built around a dynamic generalised linear modeling framework where the cell probabilities $\boldsymbol{\pi}_t$ are predicted through a *logit link* function. Additionally, we introduce a spatio-temporal cross correlation between categories based on two scale matrices which brings the dynamic components of the wavelet coefficients into a matrix normal distribution framework. Additionally, we use an efficient sparse wavelet decomposition with the inference of spatial coefficients being conducted through the Spike and Slab prior under the help of the now state matrix $\boldsymbol{A}_t$. Furthermore, the Conditional Particle Filtering framework provides us with posterior sampled filtered estimates for the state matrix $\boldsymbol{A}_t$, the cell probabilities $\boldsymbol{\pi}_t$ and posterior samples for $\boldsymbol{B}$. Lastly,

Metropolis-Hastings steps provide us with a joint estimation for the scale matrices $\boldsymbol{\Sigma}_\eta$ and $\boldsymbol{R}$ in order to estimate the spatio-temporal variation between and within categories.

Firstly, the simulation study on a small number of locations has proved that our methodology can approximate fairly well an underlying spatio-temporal cell probability vector under spatial discontinuities (Section 5.7.2). Additionally, the reconstruction of the weight function, was predicted fairly well. Furthermore, we tested our methodology on real traffic flow data of cars, buses and LGVs under segmentation of the known motorway M6 after imputing missing values. Specifically, we have predicted that for all segments cars own the highest cell probability across time and buses the lowest. Furthermore, for specific segments there is a higher variation between categories which is produced through the combination of the categorical and temporal covariance elements of $\boldsymbol{R}$ and $\boldsymbol{\Sigma}_\eta$. Additionally, we achieved to produce spatial causality between locations. These causalities were mostly based on the distance and type of counties and provided us with further findings than Chapter 4 after the inclusion of buses and LGVs.

However, we encountered difficulties during our estimation. The lack of computational infrastructure did not provide us with more Gibbs and particle iterations which are necessarily suggested for both the simulation study and the application as for the latter convergence criteria showed that even if the wavelet coefficients parameters converge, the covariance elements did not and thus further iterations were needed. One suggestion in this case would have been to consider $\boldsymbol{\Sigma}_\eta$ and $\boldsymbol{R}$ to be proportional to the identity matrix, i.e, $\boldsymbol{\Sigma}_\eta \propto \boldsymbol{I}$ and $\boldsymbol{R} \propto \boldsymbol{I}$, however this is not a good option. We would suggest to resort to more computing power instead, but failing to do that for this thesis we recommend a consideration of $\boldsymbol{\Sigma}_\eta \propto \boldsymbol{I}$ and $\boldsymbol{R} \propto \boldsymbol{I}$ instead.

Chapter 6

Conclusions

6.1 Concluding Remarks

In this thesis, we have proposed an adaptive Bayesian procedure for the Dimension-reduced DSTMs. Specifically, we took advantage of an efficient sparse wavelet decomposition where its spatial coefficients were inferred through a Spike and Slab prior.

For the Gaussian case we proposed an efficient filtering and smoothing framework for the sampling of the temporal wavelet coefficients which resulted into good estimates of the underlying process. Furthermore, we suggested a flexible covariance estimation approach under a Bayesian setting. The simulation studies provided us with proof that the proposed model can produce good estimates for the underlying spatio-temporal process but also for the spatial weight function. Furthermore, the covariance inference proved to be effective under the proposed methodology. However, due to the signal-to-noise ratio we encountered some overestimations and underestimations for a few elements of the spatial wavelet coefficients.

Overall, the performance of the proposed methodology is promising under the simulation studies even though it is computationally intensive. Finally, considering the application to pollution data, our proposed methodology performs well for non-detrended processes as it provides us with good approximations and captures the trends. Additionally, we have derived causal effects for the spatial weight function between locations which provided us with rational findings based on the distance between the locations but also their geographical and socioeconomic structure. Specifically we noticed that

neghbouring areas or areas of the same structure (urban, industrial or solely residential) in terms of pollution had either positive or negative effect for the Nitrogen Oxide pollutant.

After proving the proposed methodology's effectiveness for Gaussian spatio-temporal processes, we proceeded on creating an extension for Poisson distributed spatio-temporal processes. This again required the use of a wavelet basis decomposition along with the use of a Spike and Slab prior to incorporate a parsimonious spatial propagation. Furthermore, we suggested two different modelling settings according to the application; the first incorporated a spatially varying mean effect where it would affect the mean intensity process; the second incorporated a spatially varying but autoregressive mean effect which assumes that the mean effect affects the intensity process at time t given on how it affected the process at time $t-1$. However, due to the observed measurements being Poisson distributed, the sampling of the temporal wavelet coefficients could not be conducted under a fully MCMC setting. Therefore, we resorted to Particle Filtering techniques where we suggested several algorithms which ranged in terms of efficiency. The simulation studies showed that the proposed methodology provides us with good estimates similarly to the Gaussian case. However, we noticed that the computational complexity is even more difficult in this case and one has to resort to better computational power or maybe use parallel programming.

Additionally, the application on traffic flow data on the segmentation of M6 motorway based on the counties offered us several insights. Firstly, the traffic flow varies in all segments in a similar way as the posterior inference showed us that the variances for all segments are similar. Furthermore, we noticed that only a few segments are correlated to each other and those ones are neighboring segments. Similarly, the reconstruction of the spatial weight function provided us with causal effects between segments which were based on neighbouring segments. However convergence diagnostics showed that the implementation needed more iterations for the error variance to converge.

Finally, we have proposed a Multinomial Dimension-reduced DSTM under the Particle Filtering framework. Specifically, we used a *logit link* to model the cell count probabilities of a spatio-temporal count process. This resulted into the consideration of a state matrix for the temporal wavelet coefficients and thus, a much complex model. We considered two scale matrices for the temporal matrix. That provided us with cross

correlation between the locations and categories which again is a flexible modelling procedure to find possible spatio-temporal patterns. A simulation study has shown that the cell probabilities and the weight function can be successfully captured which is a challenge in practice. However, due to the computational complexity we did not manage to run more iterations. Finally, we revisited the traffic flow data to recognise patterns between the segments for cars, buses and large goods vehicles (LGV). The implementation of our methodology showed us that there is a variation between categories for each segment but also we extracted causal spatial effects. With the inclusion of two extra categories of vehicles we observed a contrast of between the findings of Chapter 4. However, both the simulation and application needed more iterations to converge.

Overall, the proposed modeling approaches are general and can easily be applied to other types of spatio-temporal processes (house prices, accidents, epidemics, public health). The Spike and Slab prior belief combined with the wavelet decomposition allows us through sparsity to integrate and combine different sources of spatial information. The approximations of the underlying processes and the weight functions are not perfect but spatio-temporal processes include underlying patterns which are difficult to be estimated, especially with the presence of a large parameter space. However, we have to emphasize the limitation of those proposed approaches. The algorithms are computationally demanding and the user should resort to either cloud computing, or parallel programming for the implementation in higher dimensions or more iterations. Although, nowadays, most of the big data applications are indeed computationally demanding and we live in th era of cloud computing. These low dimensional examples are based on the limited resources that were provided and are provided for implementation purposes.

6.2 Extensions and future work

Extensions of the proposed approaches are possible. In our study, we investigated the Gaussian, Poisson and Multinomial distributed spatio-temporal processes. However, we could possibly extend it to more complex distributed processes. One consideration is the Multivariate Gamma which was introduced by Ramabhadran (1951) but we are interested in the one discussed by Mathai and Moschopoulos (1991). The interesting part of this approach is the positive correlation between the variables. Interestingly,

there are spatio-temporal processes, such as pollutants or actuarial datasets which are Multivariate Gamma distributed. One limitation of this issue is that the likelihood is not of closed form, thus sampling from that distribution or conducting MCMC is a challenge in practice. However, under Particle Filtering techniques the likelihood can be approximated through the particle weights.

Similarly, we could extend the Poisson Reduced-dimension DSTM for the multivariate case with correlation structure of Kawamura (1979). The fully structured multivariate Poisson is a complex probability function and the dependencies are tracked through recurrence relationships. Bayesian techniques have been implemented in Karlis and Meligkotsidou (2005) and Bermúdez and Karlis (2011). An application on multivariate earthquake counts with a covariance structure under integer valued autoregressive models have been seen in Pedeli and Karlis (2013) . A full correlation structure could provide us with more insight in a spatio-temporal setting, however, due to its complexity the inference will be limited and computationally intensive.

Additionally, we may be interested in capturing trending or seasonal effects. These can be easily expressed by state space models and be incorporated into the Kalman Filter. For instance, we can consider Fourier seasonal componets (Harrison (1965) and more recently discussed in Harvey (1990) and West and Harrison (1997)) or a combination of trend and seasonal components subject to estimation.

Furthermore, these current models can incorporate regression explanatory variables with and time-varying regression components. For instance, regarding the prediction of the Nitrogen Oxide (NO) in Athens, we can use another pollutant, such as Carbon Dioxide (CO_2) or Ozone (O_3). We could then apply basis decomposition to the regressors as well and infer on the coefficients.

Moreover, the Multinomial Dimension-reduced DSTM could be used for clustering locations or areas based on count characteristics. Under a spatial statistics framework, Lavigne et al. (2012) used a multinomial *probit* model to cluster regions in the Alps based on avalanche counts, however, no further applications for clustering spatio-temporal characteristics based on the Multinomial distribution have been conducted.

Appendix A

Kronocker product and vec operator properties

A.1 Kronecker operator properties

- Associativity: $\boldsymbol{A} \otimes (\boldsymbol{B} \otimes \boldsymbol{C}) = (\boldsymbol{A} \otimes \boldsymbol{B}) \otimes \boldsymbol{C}$
- Distributivity: $\boldsymbol{A} \otimes (\boldsymbol{B} + \boldsymbol{C}) = (\boldsymbol{A} \otimes \boldsymbol{B}) + (\boldsymbol{A} \otimes \boldsymbol{C})$ and $(\boldsymbol{A} + \boldsymbol{B}) \otimes \boldsymbol{C}) = (\boldsymbol{A} \otimes \boldsymbol{C}) + (\boldsymbol{B} \otimes \boldsymbol{C})$
- For some scalars a and b: $a\boldsymbol{A} \otimes b\boldsymbol{B} = ab\boldsymbol{A} \otimes \boldsymbol{B}$
- For some matrices with right dimensions: $(\boldsymbol{A} \otimes \boldsymbol{B})(\boldsymbol{C} \otimes \boldsymbol{D}) = \boldsymbol{AC} \otimes \boldsymbol{BD}$
- Transposition: $(\boldsymbol{A} \otimes \boldsymbol{B})^\top = \boldsymbol{A}^\top \otimes \boldsymbol{B}^\top$
- Trace: $\mathrm{tr}(\boldsymbol{A} \otimes \boldsymbol{B}) = \mathrm{tr}(\boldsymbol{A})\mathrm{tr}(\boldsymbol{B})$
- Rank: $\mathrm{rank}(\boldsymbol{A} \otimes \boldsymbol{B}) = \mathrm{rank}(\boldsymbol{A} \otimes \boldsymbol{B}) = \mathrm{rank}(\boldsymbol{A})\mathrm{rank}(\boldsymbol{B})$
- Deeterminant: $\det(\boldsymbol{A} \otimes \boldsymbol{B}) = \det(\boldsymbol{A})^n \det(\boldsymbol{B})^m$, where $\boldsymbol{A}$ and $\boldsymbol{B}$ are respectively $m \times m$ and $n \times n$ matrices
- Inverse: $(\boldsymbol{A} \otimes \boldsymbol{B})^{-1} = \boldsymbol{A}^{-1} \otimes \boldsymbol{B}^{-1}$

A.2 vec operator

Theorem 1. Let $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{X}$ be 3 matrices of conforming sizes. Then

$$\mathrm{vec}(\boldsymbol{AXB}) = (\boldsymbol{B}^\top \otimes \boldsymbol{A})\mathrm{vec}(\boldsymbol{X})$$

Proof. Let $\boldsymbol{B} = [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n]$, $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]$. The $k-th$ column of $\boldsymbol{ABX}$ is

$$\begin{aligned}
(\boldsymbol{ABX})_{.k} = & \quad \boldsymbol{AXb}_k = \boldsymbol{A}\sum_{i=1}^{m} \boldsymbol{x}_i b_{ik} \\
= & \quad [b_{1k}\boldsymbol{A} \cdots_{mk} \boldsymbol{A}] \begin{pmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_m \end{pmatrix} \\
= & \quad ([b_{1k}, \ldots, b_{mk}] \otimes \boldsymbol{A})\text{vec}(\boldsymbol{X}) = (\boldsymbol{b}_k^\top \otimes \boldsymbol{A})\text{vec}(\boldsymbol{X})
\end{aligned}$$

Then by stacking columns below one another we get

$$\text{vec}(\boldsymbol{AXB}) = \begin{pmatrix} \boldsymbol{AXB}_{..1} \\ \vdots \\ \boldsymbol{AXB}_{..n} \end{pmatrix} = \begin{pmatrix} \boldsymbol{b}_1^\top \otimes \boldsymbol{A} \\ \vdots \\ \boldsymbol{b}_n^\top \otimes \boldsymbol{A} \end{pmatrix} \text{vec}(\boldsymbol{X}) = (\boldsymbol{B}^\top \otimes \boldsymbol{A})\text{vec}(\boldsymbol{X})$$

Corollary 1.1.

$$\begin{aligned}
\text{vec}(\boldsymbol{AB}) = & \quad (\boldsymbol{B}^\top \otimes \boldsymbol{A})\text{vec}(\boldsymbol{I}) && \text{(A.1)} \\
= & \quad (\boldsymbol{B}^\top \otimes \boldsymbol{I})\text{vec}(\boldsymbol{A}) && \text{(A.2)} \\
= & \quad (\boldsymbol{I} \otimes \boldsymbol{A})\text{vec}(\boldsymbol{B}) && \text{(A.3)}
\end{aligned}$$

Property:

$$\text{tr}(\boldsymbol{AB}) = \text{vec}(\boldsymbol{A}^\top)^\top \text{vec}(\boldsymbol{B})$$

The proof is straightforward by writing the formula of the trace, using the epxression of the matrices coefficients.

Appendix B

Distribution Theory

B.1 Normal Distribution

A random variable $\boldsymbol{X}$ follows a normal distribution with mean μ and standard deviation σ, written as $X \sim N(\mu, \sigma^2)$, if and only if its density is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

The normal distribution with mean $\mu = 0$ and standard deviation$\sigma = 1$ is called the standard normal distribution.

B.2 Truncated normal distribution

The truncated normal distribution is the probability distribution of a normally distributed random variable, whose values are restricted to lie between an interval $[a, b]$ in the case of a two-tailed truncation, or higher than a or lower than b in the case of an one-tailed truncation. If X follows a truncated normal distribution $N(\mu, \sigma^2)$ between a and b, its density function is:

$$\frac{\frac{1}{\sigma}\phi(\frac{x-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})}$$

for $x \in [a, b]$, where ϕ and Φ respectively denote the probability density function and cumulative distribution function of the standard normal distribution. In the case on

one-tailed truncation $x \geq a$ we write $b = \infty$ and $\Phi(\frac{b-\mu}{\sigma} = 1$, then the density becomes:

$$\frac{\frac{1}{\sigma}\phi(\frac{x-\mu}{\sigma})}{1-\Phi(\frac{a-\mu}{\sigma})} \tag{B.1}$$

Properties

- Expected value: $\mathrm{E}(X|x \geq a) = \mu + \sigma\lambda(a)$, with $\lambda(a) = \phi(a)/[1-\Phi(a)]$
- $\mathrm{V}(X|x \geq a) = \sigma^2[1-\delta(a)]$, with $\delta(a) = \lambda(a)[\lambda(a)-a]$.

B.3 Multivariate normal distribution

A random vector $\boldsymbol{X}$ of size p is said to have a multivariate normal distribution with mean vector μ and covariance matrix Σ, written as $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ or $\boldsymbol{X} \sim \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, when its density function is:

$$(2\pi)^{-\frac{p}{2}}|\boldsymbol{\Sigma}|^{-1/2}\exp\left(-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})\right)$$

where $|\boldsymbol{\Sigma}|$ denotes the determinant of $\boldsymbol{\Sigma}$.

B.4 Matrix-variate normal distribution

A $n \times p$ random matrix is said to have a matrix variate normal distribution with mean matrix $\boldsymbol{M}$, $n \times n$ among-row covariance matrix $\boldsymbol{U}$, $p \times p$ among-column covariance matrix $\boldsymbol{V}$, written as $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{M}, \boldsymbol{U}, \boldsymbol{V})$, or $\boldsymbol{X} \sim \mathrm{N}_{n,p}(\boldsymbol{M}, \boldsymbol{U}, \boldsymbol{V})$ if its density is:

$$\frac{\exp\left(-\frac{1}{2}\mathrm{tr}\left[\boldsymbol{V}^{-1}(\boldsymbol{X}-\boldsymbol{M})^\top\boldsymbol{U}^{-1}(\boldsymbol{X}-\boldsymbol{M})\right]\right)}{(2\pi)^{\frac{np}{2}}|\boldsymbol{V}|^{n/2}|\boldsymbol{U}|^{p/2}}$$

The matrix variate normal distribution is related to the multivariate normal distribution by the following equivalence:

$$\boldsymbol{X} \sim N_{n,p}(\boldsymbol{M}, \boldsymbol{U}, \boldsymbol{V}) \quad \equiv \quad \mathrm{vec}(\boldsymbol{X}) \sim N_{np}(\mathrm{vec}\boldsymbol{M}, \boldsymbol{V} \otimes \boldsymbol{U}).$$

This equivalence can be proved by using properties of the trace, vec operator and kronecker product. More details can be found in Gupta and Nagar (2018).

Property: If $\boldsymbol{X} \sim \mathrm{N}_{n,p}(\boldsymbol{M}, \boldsymbol{U}, \boldsymbol{V})$, then, assuming matrices $\boldsymbol{D}$ and $\boldsymbol{C}$ of appropriate dimensions and of full rank:

$$\boldsymbol{DXC} \sim N_{n,p}(\boldsymbol{BMC}, \boldsymbol{DUD}^\top, \boldsymbol{C}^\top \boldsymbol{VC})/$$

A proof of that property can be found in Gupta and Nagar (2018).

B.5 Exponential distribution

A random variable $x > 0$ follows an exponential distribution with rate pa rameter $\lambda > 0$, denoted by $X \sim \mathrm{Exp}(\lambda)$, if and only if its density is:

$$f(x) = \lambda \exp(-\lambda x)$$

B.6 Gamma distribution

A random variable $x > 0$ has a gamma distribution with shape parameter $\alpha > 0$ and scale parameter $\beta > 0$, denoted by $x \sim \mathrm{G}(\alpha, \beta)$ if and only if its density is:

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} \exp(-x/\beta)$$

where Γ is the gamma function.

Properties:

- Expected value: $\mathrm{E}(x) = \alpha\beta$
- Variance: $\mathrm{Var}(x) = \alpha\beta^2$
- If $\alpha = 1$, then x has an exponential distribution with parameter $1/\beta$

B.7 Inverse-gamma distribution

A random variable $x > 0$ has a gamma distribution with shape parameter $\alpha > 0$ and rate parameter $\beta > 0$, denoted by $x \sim \mathrm{IG}(\alpha, \beta)$ if and only if its density is:

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)}(1/x)^{\alpha+} \exp(-\beta/x)$$

where Γ is the gamma function.

Properties

- Expected value: $\mathrm{E} = \frac{\beta}{\alpha-1}$
- Variance: $\mathrm{V}(x) = \frac{\beta^2}{(a-1)(a-2)}$
- If $x \sim \mathrm{IG}(\alpha, \beta)$, then $x^{-1} \sim \mathrm{G}(\alpha, 1/\beta)$.

B.8 Wishart distribution

A $p \times p$ random symmetric positive definite matrix $\boldsymbol{V}$ is said to have a Wishart distribution with parameters ν degrees of freedom, and scale matrix $\boldsymbol{S}$, written as $\boldsymbol{V} \sim \mathrm{W}_p(\nu, \boldsymbol{S})$, if its density is:

$$\frac{1}{2^{\frac{\nu p}{2}}|\boldsymbol{S}|^{\nu/2}\Gamma_p(\frac{\nu}{2})}|\boldsymbol{V}|^{\frac{\nu-p-1}{2}} \exp\left(-\frac{\mathrm{tr}(\boldsymbol{S}^{-1}\boldsymbol{V})}{2}\right)$$

where the scale matrix $\boldsymbol{S}$ is $p \times p$ positive definite matrix and Γ_p is the multivariate gamma function/

Properties

- Expected value: $\mathrm{E}(\boldsymbol{V}) = \nu\boldsymbol{S}$
- Mode: $\mathrm{Mode}(\boldsymbol{V}) = (\nu - p - 1)\boldsymbol{S}$
- Variance: $\mathrm{Var}(V_{ij}) = \nu(s_{ij}^2 + s_{ii}s_{jj})$

B.9 Inverse-Wishart distribution

A $p \times p$ random symmetric positive definite matrix $\boldsymbol{V}$ is said to have an Inverse Wishart distribution with parameters ν degrees of freedom, and scale matrix $\boldsymbol{\Psi}$, written as $\boldsymbol{V} \sim \mathrm{W}_p(\nu, \boldsymbol{\Psi})$, if its density is:

$$\frac{|\boldsymbol{\Psi}|^{\nu/2}}{2^{\frac{\nu p}{2}}\Gamma_p(\frac{\nu}{2})}|\boldsymbol{X}|^{\frac{\nu+p+1}{2}}\exp\left(-\frac{\mathrm{tr}(\boldsymbol{\Psi}\boldsymbol{X}^{-1})}{2}\right)$$

Properties

- Expected value: $\mathrm{E}(\boldsymbol{X}) = \frac{\boldsymbol{\Psi}}{\nu-p-1}$
- Mode: $\mathrm{Mode}(\boldsymbol{X}) = \frac{\boldsymbol{\Psi}}{\nu+p+1}$
- Variance: $\mathrm{Var}(X_{ij}) = \frac{(\nu-p+1)\psi_{ij}+(\nu-p-1)\psi_{ii}\psi jj}{(\nu-p)(\nu-p-1)^2(\nu-p-3)}$
- If $\boldsymbol{X} \sim \mathrm{IW}(\nu, \boldsymbol{\Psi})$, then $\boldsymbol{X}^{-1} \sim \mathrm{IW}(\nu, \boldsymbol{\Psi}^{-1})$

Bibliography

Abrahamsen, P. (1997). A review on Gaussian random fields and correlation functions, Norsk Regnesentral/Norwegian Computing Center Oslo.

Abramovich, F. and Benjamini, Y. (1996). Adaptive thresholding of wavelet coefficients, *Computational Statistics & Data Analysis*, **22 (4)**: 351–361.

Abramovich, F., Benjamini, Y., Donoho, D. L., Johnstone, I. M. et al. (2006). Adapting to unknown sparsity by controlling the false discovery rate, *The Annals of Statistics*, **34 (2)**: 584–653.

Abramovich, F., Sapatinas, T. and Silverman, B. W. (1998). Wavelet thresholding via a bayesian approach, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **60 (4)**: 725–749.

Andrieu, C., Doucet, A. and Holenstein, R. (2010). Particle markov chain monte carlo methods, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **72 (3)**: 269–342.

Angus, J. E. (1992). *Forecasting, Structural Time Series and the Kalman Filter*, Taylor & Francis.

Antonini, M., Barlaud, M., Mathieu, P. and Daubechies, I. (1992). Image coding using wavelet transform, *IEEE Transactions on image processing*, **1 (2)**: 205–220.

Augustin, N. H., Beevers, L. and Sloan, W. T. (2008). Predicting river flows for future climates using an autoregressive multinomial logit model, *Water resources research*, **44 (7)**.

Banerjee, S., Carlin, B. P. and Gelfand, A. E. (2014). *Hierarchical modeling and analysis for spatial data*, CRC press.

Berliner, L. M., Wikle, C. K. and Cressie, N. (2000). Long-lead prediction of pacific ssts via bayesian dynamic modeling, *Journal of climate*, **13 (22)**: 3953–3968.

Bermúdez, L. and Karlis, D. (2011). Bayesian multivariate poisson models for insurance ratemaking, *Insurance: Mathematics and Economics*, **48 (2)**: 226–236.

Bersimis, S., Sgora, A. and Psarakis, S. (2018). The application of multivariate statistical process monitoring in non-industrial processes, *Quality Technology & Quantitative Management*, **15 (4)**: 526–549.

Bigot, J., Biscay, R. J., Loubes, J.-M. and Muñiz-Alvarez, L. (2011). Group lasso estimation of high-dimensional covariance matrices, *Journal of Machine Learning Research*, **12 (Nov)**: 3187–3225.

Brix, A. and Diggle, P. J. (2001). Spatiotemporal prediction for log-gaussian cox processes, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **63 (4)**: 823–841.

Brockwell, P. and Davis, R. (1991). *Time series: data analysis and theory*, New York: Springer.

Brown, P. E., Diggle, P. J., Lord, M. E. and Young, P. C. (2001). Space–time calibration of radar rainfall data, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **50 (2)**: 221–241.

Cappé, O., Godsill, S. J. and Moulines, E. (2007). An overview of existing methods and recent advances in sequential monte carlo, *Proceedings of the IEEE*, **95 (5)**: 899–924.

Carlin, B. P., Polson, N. G. and Stoffer, D. S. (1992). A Monte Carlo approach to nonnormal and nonlinear state-space modeling, *Journal of the American Statistical Association*, **87 (418)**: 493–500.

Carter, C. K. and Kohn, R. (1994). On Gibbs sampling for state space models, *Biometrika*, **81 (3)**: 541–553.

Catlin, D. E. (2012). *Estimation, control, and the discrete Kalman filter*, vol. 71, Springer Science & Business Media.

Chakir, R. and Parent, O. (2009). Determinants of land use changes: A spatial multinomial probit approach, *Papers in Regional Science*, **88 (2)**: 327–344.

Chalk, A. (2014). A spatio-temporal analysis of road traffic accidents, The University of Sheffield, unpublished MSc dissertation.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm, *The American Statistician*, **49 (4)**: 327–335.

Chipman, H. A., Kolaczyk, E. D. and McCulloch, R. E. (1997). Adaptive Bayesian wavelet shrinkage, *Journal of the American Statistical Association*, **92 (440)**: 1413–1421.

Clyde, M., Desimone, H. and Parmigiani, G. (1996). Prediction via orthogonalized model mixing, *Journal of the American Statistical Association*, **91 (435)**: 1197–1208.

Clyde, M., Parmigiani, G. and Vidakovic, B. (1998). Multiple shrinkage and subset selection in wavelets, *Biometrika*, **85 (2)**: 391–401.

Cohen, A. and Jones, R. H. (1969). Regression on a random field, *Journal of the American Statistical Association*, **64 (328)**: 1172–1182.

Cressie, N. (1992). Statistics for spatial data, *Terra Nova*, **4 (5)**: 613–617.

Cressie, N., Calder, C. A., Clark, J. S., Hoef, J. M. V. and Wikle, C. K. (2009). Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling, *Ecological Applications*, **19 (3)**: 553–570.

Cressie, N. and Wikle, C. K. (2015). *Statistics for spatio-temporal data*, John Wiley & Sons.

Creutin, J. and Obled, C. (1982). Objective analyses and mapping techniques for rainfall fields: an objective comparison, *Water resources research*, **18 (2)**: 413–431.

Daubechies, I. (1992). *Ten lectures on wavelets*, vol. 61, Siam.

Daubechies, I. and Sweldens, W. (1998). Factoring wavelet transforms into lifting steps, *Journal of Fourier analysis and applications*, **4 (3)**: 247–269.

De Jong, P. (1989). Smoothing and interpolation with the state-space model, *Journal of the American Statistical Association*, **84 (408)**: 1085–1088.

Demiralp, T., Yordanova, J., Kolev, V., Ademoglu, A., Devrim, M. and Samar, V. J. (1999). Time–frequency analysis of single-sweep event-related potentials by means of fast wavelet transform, *Brain and Language*, **66 (1)**: 129–145.

Diggle, P. J., Ribeiro, P. J. and Christensen, O. F. (2003). An introduction to model-based geostatistics, in *Spatial Statistics and Computational Methods*, pp. 43–86, Springer.

Diggle, P. J., Tawn, J. and Moyeed, R. (1998). Model-based geostatistics, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **47 (3)**: 299–350.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1996). Density estimation by wavelet thresholding, *The Annals of Statistics*, pp. 508–539.

Donoho, D. L. and Johnstone, J. M. (1994). Ideal spatial adaptation by wavelet shrinkage, *Biometrika*, **81 (3)**: 425–455.

Durbin, J. and Koopman, S. J. (2012). *Time series analysis by state space methods*, vol. 38, Oxford University Press.

Fahrmeir, L. and Tutz, G. (2013). *Multivariate statistical modelling based on generalized linear models*, Springer Science & Business Media.

Farge, M. (1992). Wavelet transforms and their applications to turbulence, *Annual Review of Fluid Mechanics*, **24 (1)**: 395–458.

Fearnhead, P. (2002). Markov chain monte carlo, sufficient statistics, and particle filters, *Journal of Computational and Graphical Statistics*, **11 (4)**: 848–862.

Frühwirth-Schnatter, S. (1994). Applied state space modelling of non-gaussian time series using integration-based Kalman filtering, *Statistics and Computing*, **4 (4)**: 259–269.

Frühwirth-Schnatter, S. (2001). Markov chain monte carlo estimation of classical and dynamic switching and mixture models, *Journal of the American Statistical Association*, **96 (453)**: 194–209.

Furrer, R., Genton, M. G. and Nychka, D. (2006). Covariance tapering for interpolation of large spatial datasets, *Journal of Computational and Graphical Statistics*, **15 (3)**: 502–523.

George, E. and McCulloch, R. (1994). Fast bayes variable selection, *University of Texas CSS Technical Report*, pp. 94–01.

Goodall, C. and Mardia, K. V. (1994). Challenges in multivariate spatio-temporal modeling, in *Proceedings of the XVIIth International Biometric Conference*, pp. 1–17.

Gupta, A. K. and Nagar, D. K. (2018). *Matrix variate distributions*, Chapman and Hall/CRC.

Handcock, M. S. and Wallis, J. R. (1994). An approach to statistical spatial-temporal modeling of meteorological fields, *Journal of the American Statistical Association*, **89 (426)**: 368–378.

Harrison, P. J. (1965). Short-term sales forecasting, *Applied Statistics*, pp. 102–139.

Harvey, A. C. (1990). *Forecasting, structural time series models and the Kalman filter*, Cambridge university press.

Hastie, T. and Tibshirani, R. (1990). Exploring the nature of covariate effects in the proportional hazards model, *Biometrics*, **46**: 1005–1016.

Heaton, T. and Silverman, B. (2008). A wavelet-or lifting-scheme-based imputation method, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **70 (3)**: 567–587.

Hooten, M. B. and Wikle, C. K. (2008). A hierarchical bayesian non-linear spatio-temporal model for the spread of invasive species with application to the eurasian collared-dove, *Environmental and Ecological Statistics*, **15 (1)**: 59–70.

Hooten, M. B., Wikle, C. K., Dorazio, R. M. and Royle, J. A. (2007). Hierarchical spatiotemporal matrix models for characterizing invasions, *Biometrics*, **63 (2)**: 558–567.

Housfater, A. S., Zhang, X.-P. and Zhou, Y. (2006). Nonlinear fusion of multiple sensors with missing data, in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 4, pp. IV–IV, IEEE.

Hsu, H.-H., Weng, C.-H. and Wu, C.-H. (2004). Contrasting characteristics between the northward and eastward propagation of the intraseasonal oscillation during the boreal summer, *Journal of climate*, **17 (4)**: 727–743.

Huerta, G., Sansó, B. and Stroud, J. R. (2004). A spatiotemporal model for mexico city ozone levels, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **53 (2)**: 231–248.

Ishwaran, H. and Rao, J. S. (2005). Spike and slab variable selection: Frequentist and bayesian strategies, *Annals of Statistics*, **33**: 730–773.

Jansen, M., Nason, G. P. and Silverman, B. W. (2001). Scattered data smoothing by empirical bayesian shrinkage of second-generation wavelet coefficients, in *Wavelets: Applications in Signal and Image Processing IX*, vol. 4478, pp. 87–98, International Society for Optics and Photonics.

Jansen, M., Nason, G. P. and Silverman, B. W. (2009). Multiscale methods for data on graphs and irregular multidimensional situations, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **71 (1)**: 97–125.

Jemai, O., Zaied, M., Amar, C. B. and Alimi, M. A. (2011). Fast learning algorithm of wavelet network based on fast wavelet transform, *International Journal of Pattern Recognition and Artificial Intelligence*, **25 (08)**: 1297–1319.

Johannesson, G., Cressie, N. and Huang, H.-C. (2007). Dynamic multi-resolution spatial models, *Environmental and Ecological Statistics*, **14 (1)**: 5–25.

Jorgensen, P. E. (2006). *Analysis and probability: wavelets, signals, fractals*, vol. 234, Springer Science & Business Media.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems, *Journal of basic Engineering*, **82 (1)**: 35–45.

Karl, T. R., Koscielny, A. J. and Diaz, H. F. (1982). Potential errors in the application of principal component (eigenvector) analysis to geophysical data, *Journal of Applied Meteorology*, **21 (8)**: 1183–1186.

Karlis, D. and Meligkotsidou, L. (2005). Multivariate poisson regression with covariance structure, *Statistics and Computing*, **15 (4)**: 255–265.

Kawamura, K. (1979). The structure of multivariate poisson distribution, *Kodai Mathematical Journal*, **2 (3)**: 337–345.

Kong, A., Liu, J. S. and Wong, W. H. (1994). Sequential imputations and bayesian missing data problems, *Journal of the American statistical association*, **89 (425)**: 278–288.

Lavigne, A., Bel, L., Parent, E. and Eckert, N. (2012). A model for spatio-temporal clustering using multinomial probit regression: application to avalanche counts, *Environmetrics*, **23 (6)**: 522–534.

Leung, A. K.-m., Chau, F.-t. and Gao, J.-b. (1998). A review on applications of wavelet transform techniques in chemical analysis: 1989–1997, *Chemometrics and Intelligent Laboratory Systems*, **43 (1-2)**: 165–184.

Lindsey, J. K. (2004). *Statistical analysis of stochastic processes in time*, vol. 14, Cambridge University Press.

Lindsten, F., Jordan, M. I. and Schön, T. B. (2014). Particle gibbs with ancestor sampling, *The Journal of Machine Learning Research*, **15 (1)**: 2145–2184.

MacKay, D. J. (1998). Introduction to gaussian processes, *NATO ASI Series F Computer and Systems Sciences*, **168**: 133–166.

Mardia, K. V., Goodall, C., Redfern, E. J. and Alonso, F. J. (1998). The kriged kalman filter, *Test*, **7 (2)**: 217–282.

Matérn, B. (1960). *Spatial Variation: Stachastic Models and Their Application to Some Problems in Forest Surveys and Other Sampling Investigations*, Statens skogsforskningsinstitut.

Mathai, A. M. and Moschopoulos, P. G. (1991). On a multivariate gamma, *Journal of Multivariate Analysis*, **39 (1)**: 135–153.

Miaou, S.-P. and Lord, D. (2003). Modeling traffic crash-flow relationships for intersections: dispersion parameter, functional form, and Bayes versus empirical Bayes methods, *Transportation Research Record: Journal of the Transportation Research Board*, **(1840)**: 31–40.

Nason, G. (2010). *Wavelet methods in statistics with R*, Springer Science & Business Media.

Nason, G. P. (1996). Wavelet shrinkage using cross-validation, *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**: 463–479.

Nason, G. P. (2002). Choice of wavelet smoothness, primary resolution and threshold in wavelet shrinkage, *Statistics and Computing*, **12 (3)**: 219–227.

Park, T. and Casella, G. (2008). The bayesian lasso, *Journal of the American Statistical Association*, **103 (482)**: 681–686.

Pedeli, X. and Karlis, D. (2013). Some properties of multivariate INAR (1) processes, *Computational Statistics & Data Analysis*, **67**: 213–225.

Pereira, S., Turkman, F. and Correia, L. (2017). Spatio-temporal analysis of regional unemployment rates: A comparison of model based approaches, *arXiv preprint arXiv:1704.05767*.

Petris, G., Petrone, S. and Campagnoli, P. (2009). Dynamic linear models, in *Dynamic Linear Models with R*, Springer.

Ramabhadran, V. (1951). A multivariate gamma-type distribution, *Sankhyā: The Indian Journal of Statistics*, pp. 45–46.

Ramsay, J. O. and Silverman, B. W. (2007). *Applied functional data analysis: methods and case studies*, Springer.

Riccio, A., Barone, G., Chianese, E. and Giunta, G. (2006). A hierarchical bayesian approach to the spatio-temporal modeling of air quality data, *Atmospheric Environment*, **40 (3)**: 554–566.

Roberts, G. O. and Rosenthal, J. S. (1998). Optimal scaling of discrete approximations to langevin diffusions, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **60 (1)**: 255–268.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*, vol. 81, John Wiley & Sons.

Sahu, S. K. and Bakar, K. S. (2012). Hierarchical Bayesian autoregressive models for large space–time data with applications to ozone concentration modelling, *Applied Stochastic Models in Business and Industry*, **28 (5)**: 395–415.

Sahu, S. K., Gelfand, A. E. and Holland, D. M. (2007). High-resolution space–time ozone modeling for assessing trends, *Journal of the American Statistical Association*, **102 (480)**: 1221–1234.

Sahu, S. K., Yip, S. and Holland, D. M. (2009). Improved space–time forecasting of next day ozone concentrations in the eastern us, *Atmospheric Environment*, **43 (3)**: 494–501.

Sandwell, D. T. (1987). Biharmonic spline interpolation of geos-3 and seasat altimeter data, *Geophysical Research Letters*, **14 (2)**: 139–142.

Sansó, B. and Guenni, L. (1999). Venezuelan rainfall data analysed by using a bayesian space–time model, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **48 (3)**: 345–362.

Schmidt, A. M. and Gelfand, A. E. (2003). A bayesian coregionalization approach for multivariate pollutant data, *Journal of Geophysical Research: Atmospheres*, **108 (D24)**.

Shephard, N. (1994). Partial non-Gaussian state space, *Biometrika*, **81 (1)**: 115–131.

Shumway, R. H. and Stoffer, D. S. (2000). Time series analysis and its applications, *Studies In Informatics And Control*, **9 (4)**: 375–376.

Storvik, G. (2002). Particle filters for state-space models with the presence of unknown static parameters, *IEEE Transactions on signal Processing*, **50 (2)**: 281–289.

Stroud, J. R., Müller, P. and Sansó, B. (2001). Dynamic models for spatiotemporal data, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **63 (4)**: 673–689.

Svensson, A., Schön, T. B. and Kok, M. (2015). Nonlinear state space smoothing using the conditional particle filter, *arXiv preprint arXiv:1502.03697*.

Sweldens, W. (1996a). The lifting scheme: A custom-design construction of biorthogonal wavelets, *Applied and Computational Harmonic Analysis*, **3 (2)**: 186–200.

Sweldens, W. (1996b). Wavelets and the lifting scheme: A 5 minute tour, *ZAMM-Zeitschrift fur Angewandte Mathematik und Mechanik*, **76 (2)**: 41–44.

Sweldens, W. (1998). The lifting scheme: A construction of second generation wavelets, *SIAM Journal on Mathematical Analysis*, **29 (2)**: 511–546.

Tepe, E. and Guldmann, J.-M. (2018). Spatio-temporal multinomial autologistic modeling of land-use change: A parcel-level approach, *Environment and Planning B: Urban Analytics and City Science*, p. 2399808318786511.

Vidakovic, B. (2009). *Statistical modeling by wavelets*, vol. 503, John Wiley & Sons.

Waller, L. A., Carlin, B. P., Xia, H. and Gelfand, A. E. (1997). Hierarchical spatio-temporal mapping of disease rates, *Journal of the American Statistical association*, **92 (438)**: 607–617.

Wendt, D. A., Irwin, M. E. and Cressie, N. (2004). Waypoint analysis for command and control, *Naval Research Logistics (NRL)*, **51 (8)**: 1045–1067.

West, M. and Harrison, J. (1997). *Bayesian forecasting and dynamic models*, Springer Series in Statistics.

Wikle, C. K. (1996). Spatio-temporal statistical models with application to atmospheric processes, Digital Repository@ Iowa State University, http://lib. dr. iastate. edu/.

Wikle, C. K. (2002). A kernel-based spectral model for non-gaussian spatio-temporal processes, *Statistical Modelling*, **2 (4)**: 299–314.

Wikle, C. K. (2003). Hierarchical models in environmental science, *International Statistical Review*, **71 (2)**: 181–199.

Wikle, C. K., Berliner, L. M. and Cressie, N. (1998). Hierarchical bayesian space-time models, *Environmental and Ecological Statistics*, **5 (2)**: 117–154.

Wikle, C. K. and Cressie, N. (1999). A dimension-reduced approach to space-time kalman filtering, *Biometrika*, **86 (4)**: 815–829.

Wikle, C. K. and Holan, S. H. (2011). Polynomial nonlinear spatio-temporal integro-difference equation models, *Journal of Time Series Analysis*, **32 (4)**: 339–350.

Wikle, C. K. and Hooten, M. B. (2006). Hierarchical bayesian spatio-temporal models for population spread, *Applications of computational statistics in the environmental sciences: hierarchical Bayes and MCMC methods*, **145169**.

Wikle, C. K., Milliff, R. F., Nychka, D. and Berliner, L. M. (2001). Spatiotemporal hierarchical bayesian modeling tropical ocean surface winds, *Journal of the American Statistical Association*, **96 (454)**: 382–397.

Xu, K. and Wikle, C. K. (2007). Estimation of parameterized spatio-temporal dynamic models, *Journal of Statistical Planning and Inference*, **137 (2)**: 567–588.

Xu, K., Wikle, C. K. and Fox, N. I. (2005). A kernel-based spatio-temporal dynamical model for nowcasting weather radar reflectivities, *Journal of the American statistical Association*, **100 (472)**: 1133–1144.

Xue, J.-Z., Zhang, H., Zheng, C.-X. and Yan, X.-G. (2003). Wavelet packet transform for feature extraction of eeg during mental tasks, in *Machine learning and cybernetics, 2003 international conference on*, vol. 1, pp. 360–363, IEEE.

Yaglom, A. (1987). Introduction, in *Correlation theory of stationary and related random functions*, pp. 1–13, Springer.

Yang, G.-J., Vounatsou, P., Zhou, X.-N., Tanner, M. and Utzinger, J. (2005). A bayesian-based approach for spatio-temporal modeling of county level prevalence

of schistosoma japonicum infection in jiangsu province, china, *International Journal for Parasitology*, **35 (2)**: 155–162.

Zhou, B. and Kockelman, K. M. (2008). Neighborhood impacts on land use change: a multinomial logit model of spatial relationships, *The Annals of Regional Science*, **42 (2)**: 321–340.