UNIVERSITY OF CALIFORNIA
SANTA CRUZ

A MODELING FRAMEWORK FOR NON-GAUSSIAN SPATIAL AND TEMPORAL PROCESSES

A dissertation submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

STATISTICAL SCIENCE

by

Xiaotian Zheng

September 2022

The Dissertation of Xiaotian Zheng is approved:

Athanasios Kottas, Co-Chair

Bruno Sansó, Co-Chair

Raquel Prado

Paul Parker

Peter Biehl
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

A Modeling Framework for Non-Gaussian Spatial and Temporal Processes

by

Xiaotian Zheng

This dissertation builds a modeling framework for non-Gaussian spatial processes, time series, and point processes, with a Bayesian inference paradigm that provides uncertainty quantification. Our methodological development emphasizes direct modeling of non-Gaussianity, in contrast with traditional approaches that consider data transformations or modeling through functionals of the data probability distribution. We achieve the goal by defining a joint distribution through factorization into a product of univariate conditional distributions according to a directed acyclic graph which implies conditional independence. We model each conditional distribution as a weighted combination of first-order conditionals, with weights that can be locally adaptive, for each one of a given number of parents which correspond to spatial nearest-neighbors or temporal lags. Such a formulation features specification of bivariate distributions that define the first-order conditionals for flexible, parsimonious modeling of multivariate non-Gaussian distributions. We obtain, in time, high-order Markov models with stationary marginals, and point process models for limited memory, dependent renewals, and duration clustering; and in space, nearest-neighbor mixture models for spatial processes. Regarding computation, representing the framework by directed acyclic graphs, with a mixture model formulation for the conditionals, gains efficiency and scalability relative to many non-Gaussian models. We develop Markov chain Monte Carlo algorithms for implementation of posterior inference and prediction, with data illustrations in biological, environmental, and social sciences.

Acknowledgments

First and foremost, I would like to thank my advisors, Athanasios Kottas and Bruno Sansó, for their mentorship. This dissertation would not have been possible without their patience, support, and guidance. Their insights into how to approach statistical problems have deeply influenced me, and their generous advice throughout the entire research process has shaped me to become an independent researcher. I cannot thank them enough.

I would also like to thank: Raquel Prado and Paul Parker for serving on my dissertation reading committee, and providing excellent suggestions that improved my dissertation; Eric Aldrich for guiding me to organize code in a package to make my daily research easier, and for a collaboration on a project involving high-frequency data that partially motivated the work in Chapter 3; Fabrizio Ruggeri and Robert Lund for their useful comments on my dissertation proposal; Tyler McCormick, an associated editor, and three anonymous referees for their helpful reviews of a manuscript version of Chapter 2.

The research for Chapter 2 was supported in part by the National Science Foundation under awards SES 1631963 and SES 2050012. The text of the chapter includes an adapted reprint of the following previously published article:

Zheng, X., Kottas, A., and Sansó, B. (2022), "On Construction and Estimation of Stationary Mixture Transition Distribution Models," *Journal of Computational and Graphical Statistics*, 31, 283–293.

The co-authors listed in this publication supervised the research that forms the basis for the chapter.

Chapter 1

Introduction

1.1 Motivation and Objective

Non-Gaussian data are prevalent in a wide range of disciplines including health, biological, environmental, and social sciences. The term non-Gaussian, in general, refers to either the nature of the data, such as integer and positive values, or any behaviors beyond the modeling capacity of a Gaussian distribution, such as asymmetry and heavy-tailedness. Statistical models that account for non-Gaussian features are essential as they are building blocks for experts to uncover underlying patterns and make scientific decisions.

Current approaches for modeling non-Gaussian data can be roughly classified into two categories: (hierarchical) first-stage Gaussian models and hierarchical first-stage non-Gaussian models. The former, assuming continuous data, exploits representing a non-Gaussian distribution as a location-scale mixture of Gaussian distributions, by choosing appropriate continuous or discrete mixing distributions for the location and/or scale parameter(s). Transforming the data and applying a Gaussian or Gaussian mixture model, loosely speaking, is also regarded as first-stage Gaussian modeling. This approach is straightforward, but the transformation may introduce distortion. Alternatively, one can turn to the class of hierarchical first-stage non-Gaussian models, by specifying non-Gaussian marginal distributions for the data, and including fixed and/or random effects through some functional of the data marginal distributions. For example, when the marginal distribution corresponds to an exponential dispersion family, and the functional is transformed mean, the resulting model is referred to as the generalized linear model (GLM; McCullagh and Nelder 1983). We recognize that our discussion here can by no means include all non-Gaussian models. For example, some non-parametric methods such as tree models do not fall into the two categories. On the other hand, a Dirichlet process mixture model can belong to either one of the categories when the model is written in a hierarchical form, in which the first stage corresponds to the mixture kernel. Again, we only attempt to summarize the general idea for modeling non-Gaussian data to motivate our development of the proposed framework. Comprehensive reviews of non-Gaussian models relevant to the topic of each chapter are provided therein.

When non-Gaussian data are irregularly located in space, the general strategy for modeling such data is to remove the independence assumption among observations, and add to the aforementioned approaches components that can describe spatial variability across locations. Consequently, the Gaussian distribution is replaced with a Gaussian process that offers a rich modeling framework for capturing complex spatial dependence, usually under the expectation that follows Tobler's first law of geography. The issue of distortion when we apply a transformation to non-real-valued data can be exaggerated, as it is generally difficult to guarantee one-to-one correspondence in terms of spatial dependence before and after transformation. It is tempting, then, to consider independent first-stage non-Gaussian likelihoods, conditional on a latent Gaussian process that is associated with some functional of the first-stage marginal distribution. When the Gaussian process is assumed for the spatial random effects through a link function for some parameter(s) of the first-stage distribution, the model is traditionally known as the spatial generalized linear mixture model (SGLMM; Diggle et al. 1998). A spatial copula model specifies the joint distribution over the spatial domain using a multivariate copula with a collection of marginal distributions. This model can be regarded as a hierarchical non-Gaussian first-stage model when augmented with a vector of latent variables that are probability integral transformations of the observations. The second stage consists of a multivariate distribution underlying the chosen copula for the latent variables. A question here is whether modeling non-Gaussian dependence through functionals of the data probability distribution, rather than data themselves, is sufficient.

As for non-Gaussian data that are dependent in time, first, it is important to identify the data structure. We focus on two types of temporal data. The first one concerns data that are collected sequentially with regular intervals, referred to as discrete time series. The second type relates to continuous times that are occurrence times of events. These events take place irregularly over time, and form a point pattern. Stochastic models for point patterns proceed either with modeling the occurrence times or intervals between them, both of which are naturally non-Gaussian since they are positive-valued. We are interested in the latter approach, and for brevity and clarity, hereafter, the interval between event times is referred to as duration. Durations can be independent or dependent based on the assumption for the point pattern. When there is dependence among durations, the durations form a discrete time series, since indices of the durations correspond to event occurrences, which are integers and are, of course, regularly spaced.

For non-Gaussian discrete time series, a dynamic generalized linear model

(DGLM; West et al. 1985; West and Harrison 2006) is in place to model the dynamic. Generally speaking, a DGLM is a hierarchical model that consists of data distributions in the first stage, conditional on a latent process for some parameter(s) of the data distribution, which forms the second stage. Thus, similar to the SGLMM setting in spatial statistics, the DGLM assumes conditional independent, non-Gaussian likelihoods, with the difference being the employment of the latent process to uncover temporal or spatial dependence. This similarity also applies to the class of copula models for time series. Another popular approach for non-Gaussian discrete time series results from the well-known autoregressive moving average (ARMA) models, e.g., the class of integer-valued autoregressive (INAR) models. In general, the additive representation of ARMA models is restrictive for non-Gaussian data. Formulating an additive model for non-Gaussian data needs additional care. For example, defining an INAR model needs the specification of a thinning operator and either the innovations or the marginal distributions.

All previously mentioned models have been successful in solving a lot of problems and applied to a great variety of scientific fields. The objective of this dissertation is to develop methodologies that can handle some problems unsolved by existing models or problems that have unsatisfactory solutions. For example, the DGLM is well suited to modeling nonstationary time series. On the other hand, it does not provide a good solution when the application of interest lies in inference for a non-Gaussian stationary marginal distribution, with a possible assumption that the temporal dependence is high order. With the high-order dependence in mind, autoregressive models formulated as stochastic difference equations hinder the study, which generally requires parameter constraints for the solutions. Copula models allow for constructing non-Gaussian time series given a pre-specified family of stationary marginal distribution, with a multivariate copula to capture the latent temporal dependence. However, inference for this class is generally difficult, especially when the data are discrete-valued. Therefore, the first goal of this dissertation is to provide a framework that allows for constructing non-Gaussian time series given a pre-specified family for the marginal distributions, with both stationarity conditions and model inference are easy to implement.

Similar problems are encountered in point process modeling. A renewal process is typically specified via independent and identically distributed durations. Historically, a particular interest in modeling point processes is motivated by removing the independence assumption in the specification of a renewal process. This, interestingly, coincides with our first goal: building a high-order Markov model with a non-Gaussian stationary marginal distribution. In fact, even keeping only the modeling objective of a non-Gaussian high-order time series is interesting enough for point process modeling. The resulting point process, depending on the shape of the associated conditional intensity function, can be classified as a self-exciting or self-regulating process. A popular class of self-exciting point process model is the Hawkes process (Hawkes, 1971a,b). When one does not seek inference to recover clustering structure, the Hawkes process specified via an additive formulation for the conditional intensity that assumes full history dependence can be computationally unattractive. Moreover, its assumption of dependence on all past events may be redundant. Thus, the second goal of this dissertation is to develop a framework for modeling point processes with high-order memory, which will include various types of point patterns mentioned above.

In non-Gaussian spatial modeling, there are two problems that we consider need further exploration for better solutions. The first one lies in the hierarchical first-stage non-Gaussian models that involve working with a large vector of spatially correlated random effects. These models, in particular, are almost default choices for a wide range of applications where the data are positive-valued, proportions, or discrete. Introducing spatial dependence through functionals of the data probability distribution, however, may fail to capture certain aspects of the complex non-Gaussian dependence. Moreover, inference for the vector of correlated random variables poses great challenges. Non-Gaussian first-stage models do not enjoy the Gaussian model property to work with marginalized likelihood with the spatial random effects integrated out. Thus, under simulation-based inference such as Markov chain Monte Carlo (MCMC) algorithms, it is unavoidable to estimate the spatial random effects which generally requires sampling a large number of highly correlated variables, not to mention the large computational burden. Approximate inference is commonly used as the solution at the cost of possibly inaccurate uncertainty estimation, which may be important for some applications.

The second problem is the lack of a general modeling framework for non-Gaussian spatial data, although most of the non-Gaussian models connect to a Gaussian process in certain ways, either through transformation or using it as a building block. However, simply relying on covariance functions of a Gaussian process may limit the modeling capacity for capturing non-Gaussian dependence. The final goal of this dissertation aims at providing a unified framework with generality that supports constructing spatial processes for general types of non-Gaussian data, and flexibility for capturing non-Gaussian dependence in a way beyond the usual link to a Gaussian process, as well as computationally efficiency that avoids working with a large vector of correlated random variables.

1.2 A Coherent Framework with Directed Acyclic Graphs

We achieve the objective of building a coherent framework for spatially and temporally dependent non-Gaussian data by representing a joint distribution with respect to a directed acyclic graph (DAG). A directed graphical model (DGM; Lauritzen 1996; Jordan 2004; Murphy 2012), also known as a Bayesian network, is a probabilistic model with a DAG that explains the conditional dependence structure among the random variables. DGMs factorize a joint distribution into a product of univariate conditionals that imply conditional independence. Expressing a joint distribution as a product of conditionals facilitates the construction of a multivariate non-Gaussian distribution, by choosing appropriate models for the univariate conditionals, thus achieving flexible, parsimonious modeling, and computational scalability.

Consider a vector of dependent scalar-valued random variables $(Z_1, \dots, Z_n)$, and let z_i be the realization of Z_i, for $i = 1, \dots, n$. We assume all random variables are continuous, and denote by $p(z_1, \dots, z_n)$ the joint density of $(Z_1, \dots, Z_n)$. For brevity, we drop the conditioning of the density on some parameters. We note that the following discussion in general holds for discrete random variables. By the chain rule of probability theory, we can always factorize the joint density $p(z_1, \dots, z_n) = p(z_1)p(z_2 \mid z_1)p(z_3 \mid z_2, z_1) \cdots p(z_n \mid z_{n-1}, \dots, z_1)$, where $p(z_i)$ is the marginal density of Z_i and $p(z_i \mid z_{i-1}, \dots, z_1)$ is the conditional density of Z_i given $(Z_{i-1} = z_{i-1}, \dots, Z_1 = z_1)$.

For certain problems, some variables in the conditioning set of $p(z_i \mid z_{i-1}, \dots, z_1)$, as i gets larger, can be redundant, i.e., conditional on some useful variables, the other ones provide no additional information. An immediate attempt may be to choose a suitable, small subset of the conditioning set that suffices for statistical modeling. For example, if the vector $(z_1, \ldots, z_n)$ corresponds to a realization of a discrete time series, the first-order Markov assumption reduces the joint density to $p(z_1, \ldots, z_n) = p(z_1) \prod_{i=2}^{n} p(z_i \,|\, z_{i-1})$. This assumption corresponds to the notion of conditional independence. It assumes that, for $i \geq 3$, conditional on Z_{i-1}, the random variable Z_i is independent of the random vector $(Z_{i-2}, \ldots, Z_1)$. It turns out that we can generalize this idea to introduce conditional independence through the conditionals of a DGM when building a model for the joint distribution, with the conditional independence structure summarized in a DAG.

Let $\mathcal{G}(\mathcal{V}, \mathcal{E})$ be a graph, where $\mathcal{V} = \{1, \ldots, n\}$ is a set of vertices, and $\mathcal{E} = \{(i, j) : i, j \in \mathcal{V}\}$ a set of edges. We take $G(i, j) = 1$ if $(i, j) \in \mathcal{E}$, i.e., $i \to j$ is a edge in $\mathcal{G}$. A graph is undirected if $G(i, j) = 1$ if and only if $G(j, i) = 1$. Otherwise, the graph is directed. A directed cycle refers to a set of vertices $v_1, \ldots, v_k \in \mathcal{V}$ such that $v_1 = v_k$ and $G(v_j, v_{j+1}) = 1$ for $j = 1, \ldots, k-1$. A DAG is a directed graph with no directed cycles. The set of parents of vertex i in a DAG, denoted as $\mathrm{pa}(i)$, contains vertices j such that $\mathrm{pa}(i) = \{j : G(j, i) = 1\}$. A DAG is numbered in topological order if, for every $(i, j) \in \mathcal{E}$, $i < j$. Without loss of generality, we will assume that the graph is in topological order.

We define a joint density of $(Z_1, \ldots, Z_n)$ with respect to a DAG, $\mathcal{G}(\mathcal{V}, \mathcal{E})$, by associating vertex i of $\mathcal{G}$ with random variable Z_i. The joint density can be expressed as

$$p(z_1, \ldots, z_n) = p(z_1) \prod_{i=2}^{n} p(z_i \,|\, \boldsymbol{z}_{\mathrm{pa}(i)}), \tag{1.1}$$

where $\boldsymbol{z}_{\mathrm{pa}(i)}$ contains elements of $(z_1, \ldots, z_{i-1})$ whose associated vertices are the parents of i, that is, the set of vertices that have directed edges to vertex i.

The expression (1.1) is ready to use for discrete time series or spatial modeling. In time, $\boldsymbol{z}_{\mathrm{pa}(i)}$ can be viewed as a set of temporal lags of z_i at time i. Although it is not necessarily the case, it is common to assume that $\mathrm{pa}(i)$ is, say, a subset of $\{i-1,\ldots,i-p\}$, which corresponds to a p-order Markov model. For spatial data, the parents in set $\mathrm{pa}(i)$ can be interpreted as the nearest neighbors of the vertex i, that is, vertices (or locations) in $\mathrm{pa}(i)$ that have shorter distances to vertex i, compared to other vertices $\{1,\ldots,i-1\} \setminus \mathrm{pa}(i)$. Therefore, the DAG-based joint density in (1.1) allows for the development of a coherent framework for spatially and temporally dependent data. Note that spatial locations are not naturally ordered, so the indices of the random variables implicitly impose an ordering on the locations. More details of the ordering effect can be found in Chapters 4 and 5. Here, we focus on motivating the use of DGMs for our framework.

The key to our framework for modeling non-Gaussian multivariate distribution comes from the modeling approach for the conditionals $p(z_i \mid \boldsymbol{z}_{\mathrm{pa}(i)})$ in (1.1). Note that the dimension of $\mathrm{pa}(i)$ could be large in certain cases, meaning that $p(z_i \mid \boldsymbol{z}_{\mathrm{pa}(i)})$ corresponds to a high-dimensional multivariate distribution $p(z_i, \boldsymbol{z}_{\mathrm{pa}(i)})$. In general, multivariate non-Gaussian distributions are not as tractable as the multivariate Gaussian ones. Developing a framework for modeling multivariate non-Gaussian distribution is challenging. We overcome the challenge by using a structured mixture model for the conditionals, that is,

$$p(z_i \mid \boldsymbol{z}_{\mathrm{pa}(i)}) = \sum_{l=1}^{i_L} w_l(i) f_{il}(z_i \mid z_{\mathrm{pa}(i)}^{(l)}), \tag{1.2}$$

where i_L is the size of the set $\mathrm{pa}(i)$, and $z_{\mathrm{pa}(i)}^{(l)}$ is the lth element in the vector $\boldsymbol{z}_{\mathrm{pa}(i)}$. The elements in $\boldsymbol{z}_{\mathrm{pa}(i)}$ are placed in ascending order with respect to some distance function. If (1.2) corresponds to a time series model, then we may let $z_{\mathrm{pa}(i)}^{(l)} = z_{i-l}$, which is the lth lag of z_i. Note that the weights for the conditional density $p(z_i \mid \boldsymbol{z}_{\mathrm{pa}(i)})$ in (1.2) are allowed to change across vertices i. In other words, each vertex i is associated with a vector of weights $\{w_l(i)\}_{l=1}^{i_L}$. This general definition features local adaptability, which can be crucial for capturing complex dependence. For example, in the geospatial context, vertex i corresponds to location $\boldsymbol{s}_i$. Thus, the weights $w_l(i) = w_l(\boldsymbol{s}_i)$ can vary across space to accommodate potentially different dependence structures in different regions.

Combining (1.1) with (1.2), modeling a multivariate distribution boils down to specification of a collection of bivariate distributions that define the conditionals f_{il} for $l = 1, \ldots, i_L$ and $i = 2, \ldots, n$, and a marginal distribution for Z_1. Compared to working with a multivariate distribution, constructing its bivariate analogue for a non-Gaussian random vector is in general manageable. Various approaches for constructing bivariate distributions are well studied in the literature, providing different perspectives to specify a bivariate distribution, such as using a pair of compatible conditionals (Arnold et al., 1999) and using a bivariate copula (Joe, 2014). These approaches will be discussed and applied throughout this dissertation. Overall, we obtain a general modeling framework for non-Gaussian data, and, in certain conditions, for non-Gaussian processes in time and space.

The class of mixture transition distribution (MTD) models, which is the focus in Chapter 2, is a special case of (1.2). The MTD model was first proposed in Raftery (1985) as a parsimonious model for approximating high-order Markov chains on finite state space, and is later extended in Le et al. (1996) to models supported on continuous state space. Chapter 2 modifies the MTD model in Le et al. (1996). The modified model corresponds to (1.2), by first taking the first-order conditional densities $f_{il}(z_i \,|\, z^{(l)}_{\text{pa}(i)}) = f_l(z_i \,|\, z_{i-l})$, and letting the order $i_L = L$ and weights $w_l(i) = w_l$ for $i > L$. When $2 \leq i \leq L$, we let $i_L = i - 1$, and $w_l(i) = w_l$ for $1 \leq l \leq i - 2$, and $w_{i-1}(i) = 1 - \sum_{r=1}^{i-2} w_r(i)$. This modification facilitates the study of conditions for first-order strict stationarity, which allows for different constructions with either continuous or discrete families for the conditional densities $f_l(z_i \mid z_{i-l})$ given a pre-specified family for the marginal density, and with general forms for the resulting conditional expectations. Apart from a framework to construct MTDs with stationary marginals, we develop a Bayesian framework for posterior inference and prediction, with particular emphasis on flexible, structured priors for the weights. Such priors support high-order dynamic modeling, i.e., under a large value of L, with efficient computation leveraged from the mixture model formulation in (1.2).

We then specify the modified MTD as the conditional density of the duration of a point process, and propose a point process modeling framework in Chapter 3. Using an MTD model for the conditional duration density accommodates high-order, non-Gaussian dynamics, and thus, it enables flexible modeling for temporal point processes with memory. A byproduct of this modeling approach is that the conditional intensity function of the resulting point process admits a representation as a local mixture of first-order hazard functions. This allows for constructing self-exciting or self-regulating point processes by specifying an appropriate family of distributions for the conditional densities of the MTD. The stationarity condition developed in Chapter 2 is applied to construct point processes given a pre-specified family for the marginal density of the duration process. This overcomes the particular challenge in modeling point process via dependent durations as discussed in Section 1.1. The resulting model, interpreted as a dependent renewal process, has identically distributed but high-order, Markov-dependent durations with general shapes for the associated hazard function. We also study extensions to cluster point processes that can describe duration clustering behaviors attributed to different factors, expanding the scope of the modeling framework to a wider range of applications.

Finally, in Chapters 4 and 5, we use (1.2) for spatial modeling. In particular, we replace the weights $w_l(i)$ with $w_l(\boldsymbol{s}_i)$ and the conditional densities $f_{il}(z_i \mid z^{(l)}_{\mathrm{pa}(i)})$ with $f_{\boldsymbol{s}_i,l}(z(\boldsymbol{s}_i) \mid z_{\boldsymbol{s}_{(il)}})$, where $\boldsymbol{s}_{(il)}$ is the lth nearest neighbor of a location $\boldsymbol{s}_i$ with respect to a distance function. Such a formulation is locally adaptive, combined with the approach of building nearest-neighbor processes (Datta et al., 2016a), providing a direct, computationally efficient, probabilistic modeling framework for non-Gaussian spatial processes. It emphasizes the ability to describe spatial dependence at the data level. Introducing spatial dependence at the data level avoids associating some functionals of the data probability distribution with a latent process, thus offering a solution to the first problem in non-Gaussian spatial modeling as discussed earlier in Section 1.1. Our coherent framework that models multivariate non-Gaussian distribution through the combination of (1.1) and (1.2) naturally solves the second issue. To the best of our knowledge, the proposed framework is the first one that provides generality for modeling general types of non-Gaussian spatial data. Another property that distinguishes our framework from the existing non-Gaussian spatial models is that its computation involves no large matrix operations, which are commonly known as the biggest barrier in Gaussian process computations for large data sets.

The rest of the dissertation is organized as follows. Chapters 2 and 3 focus on developing frameworks for non-Gaussian time series and point processes. Chapters 4 and 5 turn to non-Gaussian data indexed in space, providing frameworks for non-Gaussian continuous-valued processes and discrete-valued processes, respectively. In each chapter, we develop posterior simulation methods for model inference and prediction, and illustrate the benefits of the modeling frameworks through both synthetic and real data examples. We conclude with some future perspectives and remarks in Chapter 6. Technical details on proofs and model implementations are provided in Appendices A and B.

Chapter 2

Models for Stationary Non-Gaussian Time Series

2.1 Introduction

MTD models describe a time series $\{X_t : t \in \mathbb{N}\}$, where $X_t \in \mathcal{S} \subseteq \mathbb{R}$ for all t, by specifying the distribution of X_t conditional on the past as

$$F(x_t \mid \boldsymbol{x}^{t-1}) = \sum_{l=1}^{L} w_l \, F_l(x_t \mid x_{t-l}), \tag{2.1}$$

for $t > L$, based on initial values for $(x_1, \ldots, x_L)^\top$. In Equation (2.1), $F(x_t \,|\, \boldsymbol{x}^{t-1})$ is the conditional cumulative distribution function (cdf) of X_t given that $\boldsymbol{X}^{t-1} = \boldsymbol{x}^{t-1}$, and $F_l(x_t \,|\, x_{t-l})$ is the conditional cdf of X_t with respect to the lth transition component given that $X_{t-l} = x_{t-l}$, where $\boldsymbol{X}^{t-1} = \{X_i : i \leq t-1\}$ and $\boldsymbol{x}^{t-1} = \{x_i : i \leq t-1\}$. The parameters $w_l \geq 0$, $l = 1, \ldots, L$, assign weights to the transition components, such that $\sum_{l=1}^{L} w_l = 1$. On a finite state space this model provides a parsimonious approximation of high-order Markov chains (Raftery, 1985; Raftery and Tavaré, 1994; Berchtold, 2001). On a more general space, the model structure can represent time series that depict non-Gaussian features such as burst, outliers, and flat stretches (Le et al., 1996), or change-points (Raftery, 1994). We refer to Berchtold and Raftery (2002) for a review. An MTD model consists of L first-order transition components. The mixture autoregressive model of Wong and Li (2000) is a generalization that allows for each transition component to depend on a different number of lags; Lau and So (2008) consider a Bayesian nonparametric prior for the transition component of such models. There are several related extensions that consider mixtures of autoregressive conditional heteroscedastic terms, including Wong and Li (2001b), Berchtold (2003), Zhu et al. (2010) and Li et al. (2017). Other extensions include multivariate model settings (Hassan and Lii, 2006; Fong et al., 2007; Kalliovirta et al., 2016), time-varying mixture weights (Wong and Li, 2001a; Bartolucci and Farcomeni, 2010; Bolano and Berchtold, 2016), non-linear transition dynamics (Heiner and Kottas, 2022a) and order/lag selection (Khalili et al., 2017; Heiner and Kottas, 2021). Applications of these models appear in many fields such as finance, and the environmental and medical sciences; see, for example, MacDonald and Zucchini (1997); Lanne and Saikkonen (2003); Escarela et al. (2006); Cervone et al. (2014).

Stationarity for MTD models, and their extensions, is generally difficult to attain due to the mixture model structure. This limits the choices of parametric families for the transition components for these models. Families considered in the literature include: Gaussian (Le et al., 1996; Wong and Li, 2000; Kalliovirta et al., 2015); Student-t (Wong et al., 2009; Meitz et al., 2021); Laplace (Nguyen et al., 2016); Weibull (Luo and Qiu, 2009); and Poisson (Zhu et al., 2010). These models are typically parameterized in ways that result in conditional expectations that are linear functions of the lags. This particular parameterization facilitates the study of stationarity, though only in a weak sense, at the cost of reducing model flexibility. Indeed, the conditional expectation of an MTD model has the general form $\sum_{l=1}^{L} w_l \, \mu_l(x_{t-l})$, where $\mu_l(y) = \int x dF_l(x \mid y)$, allowing for non-linear dependence of the mean, conditional on past observations.

The primary goal of this chapter is to develop conditions for first-order strictly stationary MTD models, that is, stationary models with an invariant marginal distribution. We show that a sufficient condition is to assume the same marginal distribution for all the components of the mixture. It turns out that this marginal distribution is also the invariant marginal distribution of the time series. Under this condition, first-order strict stationarity is achieved with respect to any particular parameterization. We thus obtain a rich class of distribution specifications for the model, facilitating the study of component distributions that have not been explored in the literature, and enhancing the modeler's ability to extend beyond high-order linear dependence in the conditional expectation. Although the focus of our methodology is on strictly stationary models, we also study weak stationarity conditions for MTD models with linear conditional expectation.

MTD models are usually built by specifying transition densities $f_{U_l|V_l}$ for each component $l = 1, \ldots, L$. These correspond to conditional densities for random variable U_l given random variable V_l. This specification raises a question of existence of a coherent bivariate density f_{U_l,V_l}. Our second goal is to provide a constructive approach to building MTD models that satisfy our strict stationarity condition under a coherent bivariate density f_{U_l,V_l}. We present two distinct approaches: the bivariate distribution method, which is based on specifying the bivariate distribution of the pair $(U_l, V_l), l = 1, \ldots, L$; and the conditional distribution method, which consists of finding pairs of compatible conditional distributions $f_{U_l|V_l}$ and $f_{V_l|U_l}$ for all (U_l, V_l).

Our final goal is to develop a Bayesian framework for MTD model inference and prediction. We assume that the order of dependence is unknown, but is bounded above by a finite number L. We use an over-specified model with L chosen conservatively, under the expectation that only a few of the lags contribute to the dynamics of the series. We consider two priors for the mixture weights, one based on a truncated stick-breaking process, and the other obtained by discretization of a cdf which is assigned a nonparametric prior. While the former supports stochastically decreasing weights, the latter favors important, but not necessarily consecutive weights.

The rest of the chapter is organized as follows. In Section 2.2 we review the issues related to establishing stationarity conditions for MTD models. We then introduce the invariant condition that yields the class of first-order strictly stationary MTD models, and connect it to weak stationarity. Section 2.3 illustrates two methods to construct such models with many examples. In Section 2.4, we outline the Bayesian approach for model estimation and prediction, followed in Section 2.5 by an illustration of the properties of two structured priors for mixture weights on synthetic data, and applications of the models on two real data sets of different nature. Finally, we conclude with a discussion in Section 2.6.

2.2 First-Order Strict Stationarity

Consider the conditional density specification of the model in Equation (2.1):

$$f(x_t \mid \boldsymbol{x}^{t-1}) = \sum_{l=1}^{L} w_l \, f_l(x_t \mid x_{t-l}). \tag{2.2}$$

Under our modeling framework, each transition component is taken to correspond to the distribution for a random vector (U_l, V_l), for $l = 1, ..., L$, where $f_l \equiv f_{U_l|V_l}$ denotes the associated conditional density.

Earlier work has studied necessary and sufficient conditions for constant first and second moments (Le et al., 1996). In general, such conditions are difficult to establish, especially for the second moment $\int_{\mathcal{S}} x_t^2\, g_t(x_t) dx_t$, where $g_t(x_t) = \sum_{l=1}^{L} w_l \int_{\mathcal{S}} f_l(x_t \,|\, x_{t-l}) g_{t-l}(x_{t-l}) dx_{t-l}$ is the marginal density of the process $\{X_t\}$. This restricts the choices of parametric families for the component transition densities. In particular, those choices result in linear conditional expectations. Even when conditions for time-independent first and second moments can be obtained, the resulting constrained parameter spaces complicate estimation.

The key result for our methodology is given in the following proposition, the proof of which can be found in the Appendix. The result provides the foundation for different constructions of first-order strictly stationary MTD models. Rather than imposing restrictions on the parameter space, the proposition formulates a substantially easier to implement condition on the marginals of the bivariate distributions that define the transition components.

Proposition 2.1. *Consider a set of bivariate random vectors (U_l, V_l) taking values in $\mathcal{S} \times \mathcal{S}$, $\mathcal{S} \subseteq \mathbb{R}$, with conditional densities $f_{U_l|V_l}$, $f_{V_l|U_l}$ and marginal densities f_{U_l}, f_{V_l}, for $l = 1, \dots, L$, and let $w_l \geq 0$, for $l = 1, ..., L$, with $\sum_{l=1}^{L} w_l = 1$. Consider a time series $\{X_t : t \in \mathbb{N}\}$, where $X_t \in \mathcal{S}$, generated from*

$$f(x_t \mid \boldsymbol{x}^{t-1}) = \sum_{l=1}^{L} w_l\, f_{U_l|V_l}(x_t \mid x_{t-l}), \quad t > L, \tag{2.3}$$

and from

$$f(x_t \mid \boldsymbol{x}^{t-1}) = \sum_{l=1}^{t-2} w_l\, f_{U_l|V_l}(x_t \mid x_{t-l}) + \left(1 - \sum_{k=1}^{t-2} w_k\right) f_{U_{t-1}|V_{t-1}}(x_t \mid x_1), \quad 2 \leq t \leq L.$$

This time series is first-order strictly stationary with invariant marginal density f_X if it satisfies the invariant condition: $X_1 \sim f_X$, and $f_X(x) = f_{U_l}(x) = f_{V_l}(x)$, *for all* $x \in \mathcal{S}$*, and for all* l*.*

The two different expressions for the transition density allow us to establish the stationarity condition for the entire time series. The relevant form for inference is the one in Equation (2.3), since we work with the likelihood conditional on the first L time series observations. Proposition 1 applies regardless of X_t being a continuous, discrete or mixed random variable.

Regarding strict stationarity, the literature mostly focuses on existence of a stationary distribution. Exceptions are Kalliovirta et al. (2015) and Meitz et al. (2021), where a stationary marginal distribution for a mixture autoregressive model is obtained, albeit again under constrained parameter spaces, and Mena and Walker (2007) whose approach is the one most closely related to our proposed methods.

Mena and Walker (2007) use the latent variable method proposed in Pitt et al. (2002) to construct the conditional density for each transition component of the MTD. More specifically, $f_l(x_t \,|\, x_{t-l}) = \int h_{X|Z}(x_t \,|\, z) h_{Z|X}(z \,|\, x_{t-l}) dz$, where $h_{X|Z}(x \,|\, z) \propto h_{Z|X}(z \,|\, x) f_X(x)$, and the integral is replaced by a sum if Z is a discrete variable. Then, provided $X_1 \sim f_X$, the MTD model is first-order strictly stationary with invariant density f_X. Under this construction, the invariant density f_X can be viewed as the prior for likelihood $h_{Z|X}$, which is built through latent variable Z. In practice, this restricts the approach to continuous time series, and the choices for the invariant density to cases where f_X is conjugate to $h_{Z|X}$. Even for such cases, the transition component will typically have a complex form. In particular, the example explored in Mena and Walker (2007) involves a gamma invariant distribution, with $h_{Z|X}$ corresponding to a Poisson distribution. In this case, $f_l(x_t \,|\, x_{t-l})$ is a countable sum whose evaluation requires modified Bessel functions of the first kind. Moreover, following Pitt et al. (2002), Mena and Walker (2007) restrict attention to choices of $h_{Z|X}$ that yield linear conditional expectations for the transition components, and thus also for the MTD models.

The key feature of our approach is that it builds from the bivariate distributions, f_{U_l,V_l}, corresponding to the transition components. In the next section, we discuss two approaches to specifying those bivariate distributions, either directly or via compatible conditionals, $f_{U_l|V_l}$ and $f_{V_l|U_l}$. In conjunction with Proposition 1, we obtain a general framework to constructing first-order strictly stationary MTD models that can be applied to both discrete and continuous time series, while allowing for a wide variety of invariant marginal distributions, as well as for both linear and non-linear lag dependence in the conditional expectation.

In general, an explicit expression for the autocorrelation function for general MTD models is difficult to derive. However, a recursive equation can be obtained for a class of linear MTD models. We say the MTD model is linear if $E(U_l \mid V_l = y) = a_l + b_l\, y$ for some $a_l, b_l \in \mathbb{R},\ l = 1, \ldots, L$. Consider a linear MTD model that satisfies the invariant condition of Proposition 1, and assume that the first and second moments of the process, denoted by μ and $\mu^{(2)}$, exist and are finite. Then, for any L and $h \geq L$, we can derive

$$E(X_{t+h}X_t) \;=\; \sum_{l=1}^{L} w_l a_l \mu \;+\; \sum_{l=1}^{L} w_l b_l E(X_{t+h-l}X_t). \tag{2.4}$$

Assuming that, for any $h \geq 1$, $E(X_{t+h}X_t)$ does not depend on time t, let $r(h)$ be the lag-h autocorrelation function. Then,

$$r(h) = \phi + \sum_{l=1}^{L} w_l b_l r(h-l), \quad h \geq L, \tag{2.5}$$

where $\phi = (\sum_{l=1}^{L} w_l a_l \mu - (1 - \sum_{l=1}^{L} w_l b_l)\mu^2)/(\mu^{(2)} - \mu^2)$ is zero if and only if $\mu = 0$ or $a_l = (1 - b_l)\mu$. When $b_l = \rho,\ \rho \in (0, 1)$ and $a_l = (1-\rho)\mu$, for all l, Equation (2.5) reduces to $r(h) = \rho \sum_{l=1}^{L} w_l r(h-l),\ \ h \geq L$, which is the result in Mena and Walker (2007).

In the case of distinct roots, the general solution to Equation (2.5) is

$$r(h) = c_1 z_1^h + \cdots + c_L z_L^h + \phi \left((1 - z_1) \ldots (1 - z_L) \right)^{-1}, \tag{2.6}$$

where $c_1, \ldots, c_L$ are determined by the initial conditions $r(0), \ldots, r(L-1)$ and $z_1, \ldots, z_L$ are the roots of the associated polynomial $z^L - w_1 b_1 z^{L-1} - \cdots - w_L b_L = 0$. It follows that, as $h \to \infty$, $r(h) \to 0$ if and only if: (1) $\phi = 0$; (2) $z_1, \ldots, z_L$ all lie inside the unit circle.

The above discussion provides an approach to obtaining a weakly stationary MTD model based on Equation (2.3), and is summarized in the following proposition the proof.

Proposition 2.2. *The time series defined in Equation (2.3) is weakly stationary if: (1) the invariant condition of Proposition 1 is satisfied with a stationary marginal for which the first two moments exist and are finite; (2) the conditional expectation with respect to $f_{U_l|V_l}$ is $E(U_l \,|\, V_l = y) = a_l + b_l\, y$, for some $a_l, b_l \in \mathbb{R}$, and for all l; (3) Equation* (2.4) *is independent of time t, and the roots of the equation $z^L - w_1 b_1 z^{L-1} - \cdots - w_L b_L = 0$ all lie inside the unit circle.*

Proposition 2 illustrates the construction of a weakly stationary MTD model building from the invariant condition of Proposition 1. We focus on first-order strictly stationary MTD models. Weak stationarity can be further studied if conditions (2) and (3) of Proposition 2 are satisfied.

2.3 Construction of First-Order Strictly Stationary MTD Models

Here, we present two methods to develop first-order strictly stationary MTD models. The bivariate distribution method constructs the transition density given a specific marginal distribution. This method may result in analytically intractable transition densities. The second method, consisting of directly specifying the transition component conditional densities, has estimation advantages, although the analytical form of the marginal density may not be readily available. Thus, the selection among these methods depends on the modeling objectives. In fact, there are special cases where both the transition and marginal densities belong to the same family of distributions.

2.3.1 Bivariate Distribution Method

Under this method, we seek bivariate distributions f_{U_l,V_l} whose marginals f_{U_l} and f_{V_l} are equal to a given f_X, for $l = 1, \ldots, L$. Consequently, the lth transition component density is $f_{U_l|V_l}(u \mid v) = f_{U_l,V_l}(u, v)/f_X(v)$. In contrast to the approach in Mena and Walker (2007), which is practical when the marginal density is a conjugate prior for some likelihood, the bivariate distribution method is applicable to essentially any discrete or continuous marginal invariant density f_X. In fact, for most parametric families, there is a rich literature defining collections of bivariate distributions with a desired marginal distribution, and allowing for a variety of dependence structures. The following examples illustrate the method.

Example 1: Gaussian and continuous mixtures of Gaussians MTD models. Under marginal $f_X(x) = N(x \mid \mu, \sigma^2)$, the Gaussian MTD model can be constructed via the bivariate Gaussian distribution for (U_l, V_l), with mean $(\mu, \mu)^\top$ and covariance matrix $\Sigma = \sigma^2 \left(\begin{smallmatrix} 1 & \rho_l \\ \rho_l & 1 \end{smallmatrix}\right)$, resulting in a Gaussian density for $f_{U_l|V_l}$. In particular,

$$f(x_t \mid \boldsymbol{x}^{t-1}) = \sum_{l=1}^{L} w_l \, N\left(x_t \mid (1-\rho_l)\mu + \rho_l x_{t-l}, \sigma^2(1-\rho_l^2)\right). \tag{2.7}$$

Let $t(x \,|\, \mu, \sigma, \nu) \propto (1 + \nu^{-1}((x-\mu)/\sigma)^2)^{-(\nu+1)/2}$ denote the Student-t density, where μ, σ and ν are respectively location, scale and tail parameters. To construct as a natural extension of the Gaussian MTD model a stationary Student-t MTD model, consider the bivariate Student-t distribution, which can be defined as a scale mixture of a bivariate Gaussian with mean $(\mu, \mu)^\top$ and covariance matrix $q\Sigma$, with Σ as previously defined, mixing on q with respect to an inverse-gamma, $\mathrm{IG}(\nu/2, \nu/2)$, distribution. Under marginal $f_X(x) = t(x \,|\, \mu, \sigma, \nu)$, the conditional density $f(x_t \mid \boldsymbol{x}^{t-1})$ of the Student-t MTD model is given by

$$\sum_{l=1}^{L} w_l \, t\left(x_t \mid (1-\rho_l)\mu + \rho_l x_{t-l}, \sigma^2(1-\rho_l^2)(\nu + d_l)/(\nu+1), \nu+1\right), \tag{2.8}$$

where $d_l = (x_{t-l} - \mu)^2/\sigma^2$. In both the Gaussian and Student-t MTD examples, the transition component densities and the invariant density belong to the same family of distributions.

The Student-t MTD model is an example for building MTD models through bivariate distributions that admit a location-scale mixture representation. Taking an exponential distribution for the scale q yields the bivariate Laplace distribution of Eltoft et al. (2006), thus producing an MTD model with an invariant Laplace marginal density. Scaling both the mean μ and the covariance Σ of the bivariate Gaussian distribution by a unit rate exponential random variable yields the bivariate asymmetric Laplace distribution of Kotz et al. (2012), and thus an MTD model with an asymmetric Laplace distribution as the invariant marginal. We can further elaborate on this approach using appropriate mixing distributions for the Gaussian location and scale to obtain skewed-Gaussian and skewed-t distributions (Azzalini, 2013) for the bivariate component distributions, as well as for the invariant marginal distribution.

Example 2: Poisson and Poisson mixture MTD models. To model time series of counts taking countably infinite values, we can construct an MTD model with a Poisson marginal by considering the bivariate Poisson distribution of Holgate (1964) for the transition components. This choice has been discussed in Berchtold and Raftery (2002), without addressing the stationarity condition. In particular, we consider the latent variable representation of Holgate's bivariate Poisson. Given a Poisson marginal $f_X(x) = \text{Pois}(x \mid \phi)$, we take $(U_l, V_l) \equiv (U, V) = (Q + Z, W + Z)$, for all l, where Q, W and Z are independent Poisson random variables with means λ, λ and γ, respectively. It follows that both U and V are Poisson random variables with rate parameter $\phi = \lambda + \gamma$. Using the latent variable representation, the lth component transition density of the Poisson MTD model can be sampled through $Q_t \sim \text{Pois}(q_t \mid \lambda)$ and $Z_t \mid X_{t-l} = x_{t-l} \sim \text{Bin}(z_t \mid x_{t-l}, \gamma/\phi)$, with $X_t = Q_t + Z_t$ obtained as the realization from the lth component conditional distribution $X_t \mid X_{t-l} = x_{t-l}$. Here, $\text{Bin}(x \mid n, p)$ denotes the binomial distribution with n trials and probability of success p.

A common extension of the Poisson to account for counts that have excess zeros is a mixture of Poisson and a distribution that degenerates at 0. A random variable X is zero-inflated Poisson distributed, denoted as $\text{ZIP}(x \mid \phi, q)$, if its distribution is a mixture of a point mass at zero and a Poisson distribution with parameter ϕ, with respective probabilities $0 < q < 1$ and $(1 - q)$. Given an invariant marginal $f_X(x) = \text{ZIP}(x \mid \phi, q)$, we use the bivariate zero-inflated Poisson distribution of Li et al. (1999) for $(U_l, V_l) \equiv (U, V)$, for all l, given by a mixture of a point mass at $(0,0)$, two univariate Poisson distributions, and a bivariate Poisson distribution; that is $f_{U,V}(u,v) = q_0(0,0) + 0.5q_1(\text{Pois}(u \mid \phi), 0) + 0.5q_1(0, \text{Pois}(v \mid \phi)) + q_2\text{BP}(u, v \mid \phi, \phi)$, where $\sum_{j=0}^{2} q_j = 1$, $q_0 + 0.5q_1 = q$, and $\text{BP}(\cdot, \cdot \mid \phi, \phi)$ denotes Holgate's bivariate Poisson distribution. Although the corresponding component density $f_{U \mid V}(u \mid v) = f_{U,V}(u,v)/f_X(v)$ is complex, this example provides possibilities for modeling stationary zero-inflated count time series.

Exploiting the latent variable representation of Holgate's bivariate Poisson, we can obtain extensions of the Poisson MTD model that allow for more flexible dependence structure and for overdispersion. Following the earlier notation, replace the means λ and γ of the latent Poisson random variables with $\alpha\lambda$ and $\alpha\gamma$, and mix over α with respect to a $\text{Ga}(\alpha \mid k, \eta)$ distribution, where $\text{Ga}(x \mid a, b)$ denotes the gamma distribution with mean a/b. Such mixing yields a bivariate negative binomial distribution after α is marginalized out (Kocherlakota and Kocherlakota, 2006). The conditional distribution of U given $V = v$ admits a convolution representation. Let Z_1 and Z_2 be conditionally independent, given $V = v$, following a $\text{Bin}\,(z_1 \mid v, \gamma/(\lambda+\gamma))$ and $\text{NB}\,(z_2 \mid k+v, 1-\lambda/(2\lambda+\gamma+\eta))$ distribution, respectively, where $\text{NB}(x \mid r, p)$ denotes the negative binomial distribution with r number of successes and probability of success p. Then, $U = Z_1 + Z_2$ is a realization from the conditional distribution $U \mid V = v$. Similar to the Poisson case, we can use this convolution representation to define a stationary MTD model with a negative binomial marginal $f_X(x) = \text{NB}\,(x \mid k, \eta/(\lambda+\gamma+\eta))$.

Example 3: Bernoulli and Binomial MTD models. Assume again $(U_l, V_l) \equiv (U, V)$, for all l, and consider the bivariate Bernoulli distribution with probability mass function $p(u,v) = p_1^{uv} p_2^{u(1-v)+(1-u)v}(1-p_1-2p_2)^{(1-u)(1-v)}$, where $p_1 > 0$, $p_2 > 0$ and $p_1 + 2p_2 < 1$. Then, marginally U and V are both Bernoulli distributed with probability of success $p_1 + p_2$. The conditional distribution of U given $V = v$ is also Bernoulli (Dai et al., 2013) with probability of success $p(1, v)/\left(p(1, v) + p(0, v)\right)$. Using this bivariate Bernoulli distribution, we define a stationary Bernoulli MTD model

$$f(x_t \mid \boldsymbol{x}^{t-1}) = \sum_{l=1}^{L} w_l \operatorname{Ber}\left(x_t \mid p(1, x_{t-l})/(p(1, x_{t-l}) + p(0, x_{t-l}))\right), \tag{2.9}$$

which has a stationary marginal distribution $f_X(x) = \operatorname{Ber}(x \,|\, p_1 + p_2)$.

Sequences of independent bivariate Bernoulli random vectors can be used as building blocks for various bivarate distributions. In particular, a family of bivariate binomial distributions for (U, V) can be constructed by setting $U = \sum_{i=1}^{n} \tilde{U}_i$ and $V = \sum_{i=1}^{n} \tilde{V}_i$, where $(\tilde{U}_i, \tilde{V}_i), i = 1, \ldots, n$, are independent from the bivariate Bernoulli distribution given above (Kocherlakota and Kocherlakota, 2006). The conditional distribution of U given $V = v$ can be defined through the convolution of two conditionally independent, given $V = v$, binomial random variables, one with parameters $n - v$ and $p_2/(1 - p_1 - p_2)$ and the other with parameters v and $p_1/(p_1 + p_2)$. Again, this convolution representation can be used to define a stationary binomial MTD model with marginal $f_X(x) = \operatorname{Bin}(x \,|\, n, p_1 + p_2)$.

Examples 2 and 3 illustrate MTD models for finite/infinite-range discrete-valued time series with high-order dependence, and with stationary marginal distributions belonging to a range of families. These can be used, for example, for classification of time series data, or for time-varying counts that exhibit features such as overdispersion or excess of zero values when compared to a traditional Poisson model. It is worth mentioning that some of our examples induce non-linear conditional expectations. For example, the conditional expectation of the Bernoulli MTD model is $\sum_{l=1}^{L} w_l p(1, x_{t-l})/(p(1, x_{t-l}) + p(0, x_{t-l}))$. Building MTD models like the ones we have proposed using the existing methods in the MTD literature is a formidable task.

2.3.2 Conditional Distribution Method

The strategy here is to use compatible conditional densities, $f_{U_l|V_l}$ and $f_{V_l|U_l}$, to specify the bivariate density of (U_l, V_l) for the lth transition component. Conditional densities $f_{U|V}$ and $f_{V|U}$ are said to be compatible if there exists a bivariate density with its conditionals given by $f_{U|V}$ and $f_{V|U}$; see Arnold et al. (1999) for general conditions under which candidate families of two conditionals are compatible.

We begin with the assumption that $f_{U_l|V_l}$ and $f_{V_l|U_l}$ belong to the same family. This assumption is reasonable, since the invariant condition of Proposition 1 requires that all marginals are the same. Once the family of distributions for the conditionals is chosen, we ensure the conditionals are compatible, as well as that both marginals of the corresponding bivariate density are given by the target invariant density f_X. In some special cases, the marginal densities are in the same family as the compatible conditionals. To demonstrate this method, we use a pair of Lomax conditionals and a pair of gamma conditionals; both cases are considered in Arnold et al. (1999) to identify compatibility restrictions for their parameters.

Example 4: Lomax MTD models. The Lomax distribution is a shifted version of the Pareto Type I distribution such that it is supported on $\mathbb{R}^+$. Denote by $P(x \mid \sigma, \alpha) = \alpha\sigma^{-1} \left(1 + x\sigma^{-1}\right)^{-(\alpha+1)}$ the Lomax density, where $\alpha > 0$ is the shape parameter, and $\sigma > 0$ the scale parameter. The corresponding tail distribution function is $\Pr(X > x) = (1 + x\sigma^{-1})^{-\alpha}$, implying a polynomial tail that supports modeling for time series with high levels of skewness. We consider a pair of compatible Lomax densities for $(U_l, V_l) \equiv (U, V)$, for all l, such that $f_{U\,|\,V}(u \,|\, v) =$

$P\,(u \,|\, (\lambda_0 + \lambda_1 v)/(\lambda_1 + \lambda_2 v), \alpha)$, and $f_{V\,|\,U}(v \,|\, u) = P\,(v \,|\, (\lambda_0 + \lambda_1 u)/(\lambda_1 + \lambda_2 u), \alpha)$, with the restriction that $\lambda_0, \lambda_1, \lambda_2 > 0$ if $\alpha = 1$, $\lambda_0 \geq 0, \lambda_1, \lambda_2 > 0$ if $0 < \alpha < 1$, and $\lambda_0, \lambda_1 > 0, \lambda_2 \geq 0$ if $\alpha > 1$, to guarantee that these are proper densities. Lomax MTD models specified using the conditional distributions above have an invariant marginal $f_X(x) \propto (\lambda_1 + \lambda_2 x)^{-1}(\lambda_0 + \lambda_1 x)^{-\alpha}$. Taking $\alpha > 1$ and $\lambda_2 = 0$ leads to a special case where both the component transition density and the marginal density are Lomax. This particular Lomax MTD model is

$$f(x_t \mid \boldsymbol{x}^{t-1}) = \sum_{l=1}^{L} w_l \, P(x_t \mid \phi + x_{t-l}, \alpha), \tag{2.10}$$

where $\phi = \lambda_0/\lambda_1$, and the invariant marginal is $f_X(x) = P(x \,|\, \phi, \alpha - 1)$.

Example 5: Gamma MTD models. We consider a pair of conditional gamma densities for the random vector $(U_l, V_l) \equiv (U, V)$, for all l, such that $f_{U|V}(u \,|\, v) = \mathrm{Ga}(u \,|\, m_0, m_1 + m_2 v)$, and $f_{V|U}(v \,|\, u) = \mathrm{Ga}(v \,|\, m_0, m_1 + m_2 u)$, where $m_0, m_1, m_2 > 0$. This pair of conditionals is one of six choices discussed in Arnold et al. (1999) in the context of conditional gamma distributions that produce proper bivariate densities for (U, V). The resulting transition density is

$$f(x_t \mid \boldsymbol{x}^{t-1}) = \sum_{l=1}^{L} w_l \, \mathrm{Ga}(x_t \mid m_0, m_1 + m_2 x_{t-l}), \tag{2.11}$$

and the invariant marginal is $f_X(x) \propto x^{m_0 - 1} \exp(-m_1 x)(m_1 + m_2 x)^{-m_0}$.

Examples 4 and 5 present two stationary MTD models with, respectively, polynomial and exponential tail behaviors. They provide alternatives to the existing MTD model literature for positive-valued time series, where the only model that has received attention is based on the Weibull distribution. In addition, the general Lomax MTD model with $\lambda_2 \neq 0$ and the gamma MTD model have non-linear conditional expectations.

2.4 Bayesian Implementation

2.4.1 Hierarchical Model Formulation

Here, we outline an approach to perform posterior inference for the general MTD model, using a likelihood that is conditional on the first L observations of the time series realization $\{x_t\}_{t=1}^n$. We introduce a set of latent variables $\{Z_t\}_{t=L+1}^n$ with Z_t taking values in $\{1, \ldots, L\}$ such that $p(z_t \,|\, \boldsymbol{w}) = \sum_{l=1}^L w_l \delta_l(z_t)$, where $\boldsymbol{w} = (w_1, \ldots, w_L)^\top$, and $\delta_l(z_t) = 1$ if $z_t = l$ and 0 otherwise. Conditioning on the set of latent variables and the first L observations, the hierarchical representation of the model is:

$$\begin{aligned} x_t \mid z_t, \boldsymbol{\theta} &\overset{ind.}{\sim} f_{z_t}(x_t \mid x_{t-z_t}, \boldsymbol{\theta}_{z_t}), \quad z_t \mid \boldsymbol{w} \overset{i.i.d.}{\sim} \sum_{l=1}^{L} w_l \delta_l(z_t), \quad t = L+1, \ldots, n, \\ \boldsymbol{w} &\sim \pi_w(\cdot), \quad \boldsymbol{\theta}_l \overset{ind.}{\sim} \pi_l(\cdot), \quad l = 1, \ldots, L, \end{aligned} \tag{2.12}$$

where $\boldsymbol{\theta}_l$ denotes the transition component parameters, and $\boldsymbol{\theta}$ collects all $\boldsymbol{\theta}_l$. Any MCMC algorithm for finite mixture models is readily adoptable. If the transition density of the model is sampled via a latent process, such as for Example 2 of Section 3, an additional step to sample the latent variables needs to be added in Equation (2.12).

A key component of the Bayesian model formulation is the choice of the prior distribution for the mixture weights. As a point of reference, we consider a uniform Dirichlet prior that assumes equal contribution from each lag, denoted by $\mathrm{Dir}(\cdot \mid \mathbf{1}_L/L)$, where $\mathbf{1}_L$ is a unit vector of length L. We discuss next two priors that assume more structure.

The first prior is a truncated version of the stick-breaking prior, which characterizes the weights for random discrete distributions generated by the Dirichlet process (Sethuraman, 1994). More specifically, the weights are constructed as follows: $w_1 = \zeta_1$, $w_l = \zeta_l \prod_{r=1}^{l-1}(1-\zeta_r)$, $l = 2, \ldots, L-1$, and $w_L = \prod_{l=1}^{L-1}(1-\zeta_l)$, where $\zeta_l \overset{i.i.d.}{\sim} \text{Beta}(1, \alpha_s)$, for $l = 1, \ldots, L-1$. The resulting joint distribution for the mixture weights is a special case of the generalized Dirichlet distribution (Connor and Mosimann, 1969). We denote the truncated stick-breaking prior as $\text{SB}(\cdot \mid \alpha_s)$. For $l = 1, ..., L-1$, $E(w_l) = \alpha_s^*(1-\alpha_s^*)^{l-1}$, where $\alpha_s^* = (1+\alpha_s)^{-1}$. Hence, on average, this prior implies geometrically decreasing weights, with smaller α_s values favoring stronger contributions from recent lags. In certain applications, it may be natural to expect some directionality in the relevance of the weights implied by time, and this prior provides one option to incorporate into the model such a property.

An alternative prior is obtained by assuming that the weights are increments of a cdf G with support on $[0, 1]$; that is, $w_l = G(l/L) - G((l-1)/L)$, for $l = 1, \ldots, L$. We place a Dirichlet process prior on G, denoted as $\text{DP}(\alpha_0, G_0)$, where $G_0 = \text{Beta}(a_0, b_0)$ and $\alpha_0 > 0$ is the precision parameter. From the Dirichlet process definition (Ferguson, 1973), given α_0 and G_0, the vector of mixture weights follows a Dirichlet distribution with shape parameter vector $\alpha_0(a_1, \ldots, a_L)^\top$, where $a_l = G_0(l/L) - G_0((l-1)/L)$, for $l = 1, \ldots, L$. We refer to this prior as the cdf-based prior, and denote it as $\text{CDP}(\cdot \mid \alpha_0, a_0, b_0)$. Under this prior, we have that $E(\boldsymbol{w}) = (a_1, \ldots, a_L)^\top$. The nonparametric prior for G supports general distributional shapes, and thus allows for flexibility in the estimation of the mixture weights. In particular, multimodal distributions G can produce sparse weight vectors, with some/several entries near zero. Hence, this prior may be suitable for scenarios where there are inactive lags between influential lags and the influential lags are not necessarily the most recent lags. Heiner et al. (2019) proposed a different prior for sparse probability vectors, which generally requires a larger number of prior hyperparameters.

Overall, the properties of both structured priors support flexible inference for the mixture weights, enabling our strategy to specify a large value of L, assigning a priori small probabilities to distant lags. The contribution of each lag will be induced by the mixing, with important lags being assigned large weights a posteriori.

2.4.2 Estimation, Model Checking, and Prediction

The posterior distribution of the model parameters, based on the conditional likelihood, is

$$p(\boldsymbol{w}, \boldsymbol{\theta}, \{z_t\}_{t=L+1}^{n} \mid D_n) \propto \pi_w(\boldsymbol{w}) \prod_{l=1}^{L} \pi_l(\boldsymbol{\theta}_l) \prod_{t=L+1}^{n} \left\{ f_{z_t}(x_t \mid x_{t-z_t}, \boldsymbol{\theta}_{z_t}) \sum_{l=1}^{L} w_l \delta_l(z_t) \right\} \tag{2.13}$$

where $D_n = \{x_t\}_{t=L+1}^{n}$, and it can be explored using MCMC posterior simulation.

Conditional on $\boldsymbol{\theta}$ and $\boldsymbol{w}$, the posterior full conditional of each Z_t is a discrete distribution on $\{1, ..., L\}$ with probabilities proportional to $w_l f_l(x_t \,|\, x_{t-l}, \boldsymbol{\theta}_l)$. Conditional on the latent variables and w, the sampling for each θ_l depends on the particular choice of the transition component distributions. Details for the models implemented are given in the appendix. The sampling for $\boldsymbol{w}$, conditional on $\{z_t\}_{t=L+1}^{n}$ and $\boldsymbol{\theta}$, depends only on $M_l = |\{t : z_t = l\}|$, for $l = 1, ..., L$, where $|\{\cdot\}|$ is the cardinality of the set $\{\cdot\}$. Both priors for the mixture weights result in ready updates. The posterior full conditional of $\boldsymbol{w}$ under the truncated stick-breaking prior can be sampled through latent variables ζ_l^*, which are conditionally independent Beta$(1+M_l, \alpha_s+\sum_{r=l+1}^{L} M_r)$, for $l = 1, \ldots, L-1$, such that $w_1 = \zeta_1^*$, $w_l = \zeta_l^* \prod_{r=1}^{l-1}(1-\zeta_r^*)$, for $l = 2, \ldots, L-1$, and $w_L = \prod_{l=1}^{L-1}(1-\zeta_l^*)$. Under the cdf-based prior, the posterior full conditional of $\boldsymbol{w}$ is Dirichlet with parameter vector $(\alpha_0 a_1 + M_1, \ldots, \alpha_0 a_L + M_L)^\top$.

We assess the model's validity using randomized quantile residuals (Dunn and Smyth, 1996; Escarela et al., 2006). Such residuals are calculated by inverting the fitted conditional cdf for the time series. Posterior samples of these quantile sets can then be compared with the standard Gaussian distribution, providing a measure of goodness-of-fit with uncertainty quantification. Specifically, the randomized quantile residual for continuous x_t is defined as $r_t = \Phi^{-1}\left(F(x_t \mid \boldsymbol{x}^{t-1})\right)$ where $\Phi(\cdot)$ is the cdf of the standard Gaussian distribution. If x_t is discrete, $r_t = \Phi^{-1}(u_t)$, where u_t is generated from a uniform distribution on the interval (a_t, b_t) with $a_t = F(x_t - 1 \mid \boldsymbol{x}^{t-1})$ and $b_t = F(x_t \mid \boldsymbol{x}^{t-1})$. If F is correctly specified, the residuals r_t, $t = L+1, \ldots, n$, will be independently and identically distributed as a standard Gaussian distribution.

Finally, we consider prediction for future observations. The posterior predictive density of X_{n+1}, corresponding to the first out-of-sample observation, is obtained by marginalizing the transition density with respect to the posterior distribution of model parameters:

$$p(x_{n+1} \mid D_n) = \int\int \left\{ \sum_{l=1}^{L} w_l \, f_l(x_{n+1} \mid x_{n+1-l}, \boldsymbol{\theta}_l) \right\} p(\boldsymbol{\theta}, \boldsymbol{w} \mid D_n) \, d\boldsymbol{\theta} d\boldsymbol{w}. \qquad (2.14)$$

Exploiting the structure of the conditional distributions of the MTD model, we can sample from the k-step-ahead posterior predictive density using a straightforward extension of Equation (2.14). Note that the k-step-ahead posterior predictive uncertainty incorporates both the uncertainty from the parameter estimation, and the uncertainty from the predictions of the previous $(k-1)$ out-of-sample observations.

2.5 Data Illustrations

2.5.1 First Simulation Experiment

We generated 2000 observations from the Gaussian MTD model specified in Equation (2.7) with $\mu = 10, \sigma^2 = 100$, under two scenarios for the mixture weights, one with exponentially decreasing weights and the other one with an uneven arrangement of the relevant lags. In Scenario 1, we took $\boldsymbol{\rho} = (0.7, 0.3, 0.1, 0.05, 0.05)^\top$ and $w_i \propto \exp(-i), i = 1, \ldots, 5$. In Scenario 2, we took $\boldsymbol{\rho} = (0.4, 0.1, 0.7, 0.1, 0.5)^\top$ and $\boldsymbol{w} = (0.2, 0.05, 0.45, 0.05, 0.25)^\top$. We consider these two scenarios to examine the effectiveness of structured priors for the mixture weights.

We applied the Gaussian MTD model with three different orders $L = 5, 15, 25$. In each case, we considered three priors for the weights: the Dirichlet prior, the truncated stick-breaking prior, and the cdf-based prior. The shape parameter of the Dirichlet prior was $\mathbf{1}_L/L$ for each L. The precision parameter α_s for the truncated stick-breaking prior was taken to be $1, 2, 3$, corresponding to the three L values. For the cdf-based prior, we chose $\alpha_0 = 5$ as the precision parameter, and used as base distribution a beta with shape parameter $a_0 = 1$, and $b_0 = 3, 6, 7$ respectively for the three orders considered. Thus, this prior elicited a decreasing pattern similar to the truncated stick-breaking prior. For all models, the mean μ and the variance σ^2 received conjugate priors $N(\mu \,|\, 0, 100)$ and $\mathrm{IG}(\sigma^2 \,|\, 2, 0.1)$, respectively, and the component-specific correlation coefficient ρ_l was assigned a uniform prior $\mathrm{Unif}(-1, 1)$ independently for all l.

We ran the Gibbs sampler for 165000 iterations, discarding the first 5000 samples as burn-in, and collected samples every 20th iterations. We focus on

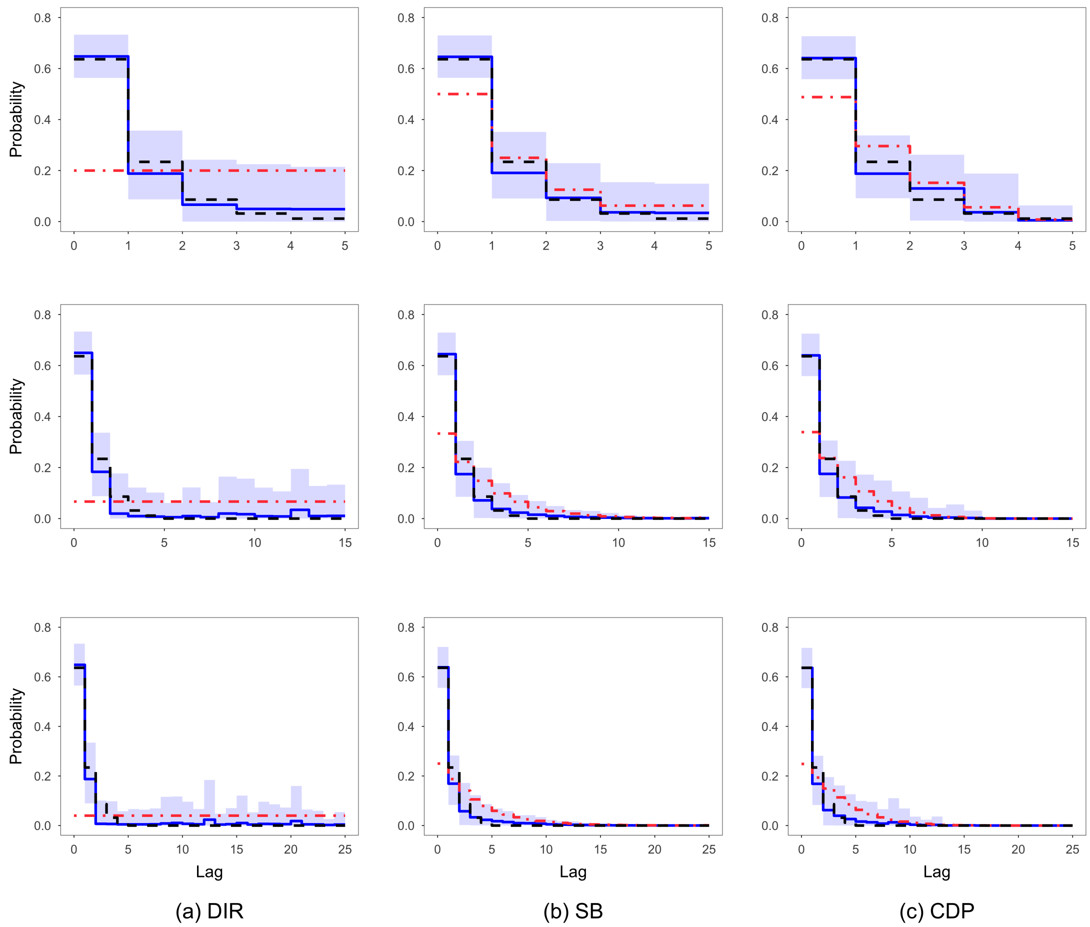


Figure 2.1: Chapter 2 - first simulation data analysis. Inference results for the weights under Scenario 1, based on the Gaussian MTD model, with the Dirichlet (column (a)), the truncated stick-breaking (column (b)), and the cdf-based (column (c)) priors, when $L = 5$ (top), $L = 15$ (middle) and $L = 25$ (bottom). Black dashed lines are true weights, red dot-dashed lines are prior means, blue solid lines are posterior means, and blue polygons are 95% credible intervals.

inference results for the mixture weights. Figures 2.1 and 2.2 provides a visual inspection on the posterior estimates for the mixture weights, respectively. When the order was correctly specified, that is, $L = 5$, all three models provided good estimates. In Scenario 1, all models underestimated the weight for lag 2. Models with the proposed priors produced accurate estimates for the rest of the lags, while the model that used the Dirichlet prior systematically overestimated the weight for the first lag, and underestimated all other weights. In Scenario 2,

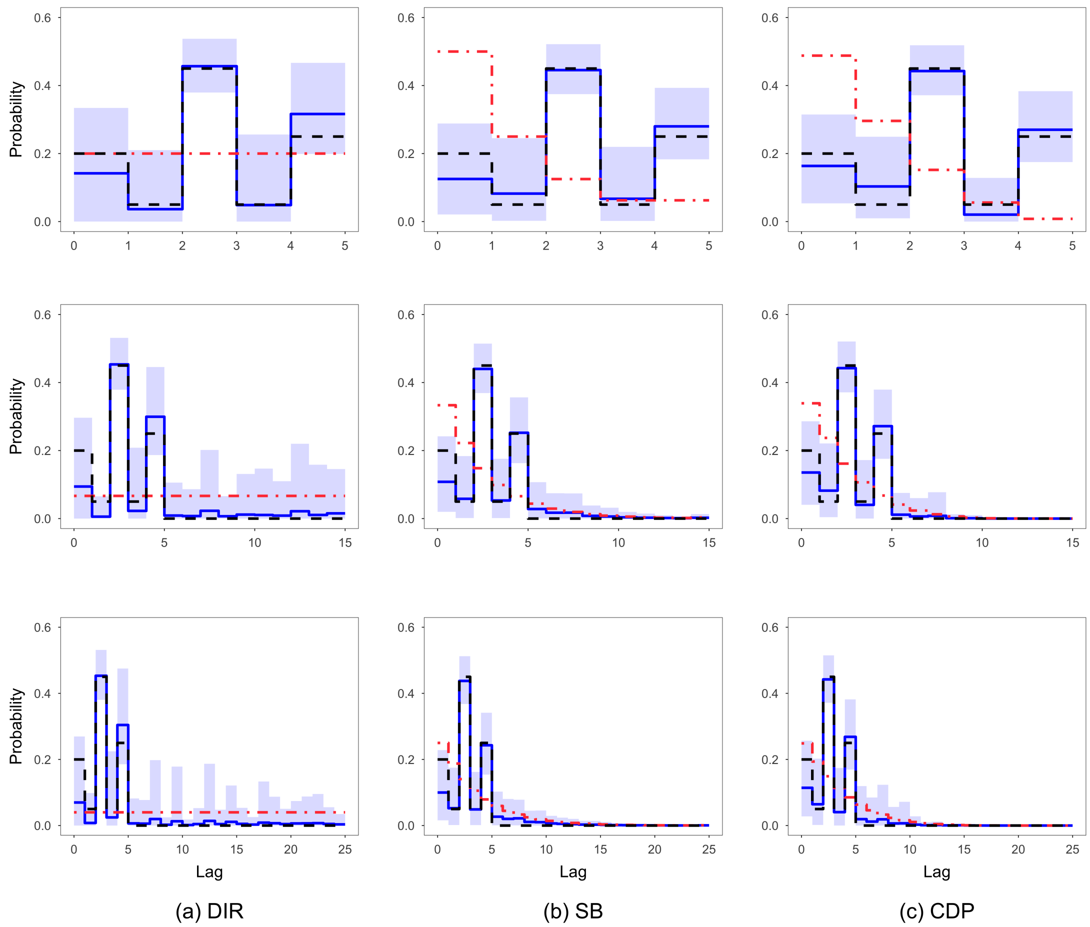


Figure 2.2: Chapter 2 - first simulation data analysis. Inference results for the weights under Scenario 2, based on the Gaussian MTD model, with the Dirichlet (column (a)), the truncated stick-breaking (column (b)), and the cdf-based (column (c)) priors, when $L = 5$ (top), $L = 15$ (middle) and $L = 25$ (bottom). Black dashed lines are true weights, red dot-dashed lines are prior means, blue solid lines are posterior means, and blue polygons are 95% credible intervals.

all models underestimated the weight for the first lag. For the other non-zero weights, the model with the Dirichlet prior tended to underestimate the weights for lag 2, 4 and overestimated the weight for lag 5, while the other two models estimated the weights quite well. In both scenarios, the proposed priors had a parsimonious behavior in that, given the data, distant lags were assigned almost zero probability mass with low posterior uncertainty. Overall, we note that, under an over-specified order L, the proposed priors offer inferential advantages when compared to the Dirichlet prior.

2.5.2 Second Simulation Experiment

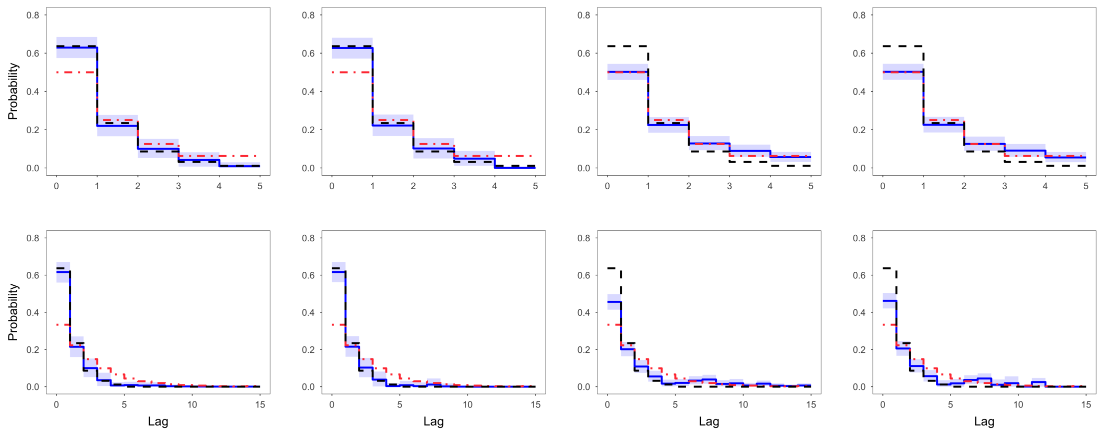
In the second simulation experiment, we demonstrate the ability of the negative binomial MTD (NBMTD) model to accommodate over-dispersed count data, including comparison with the Poisson MTD (PMTD) model. We generated 800 observations from the NBMTD model as follows.

$$\begin{aligned} x_t \mid q_t, x_{t-z_t}, z_t, q_t, \lambda, \gamma &\overset{ind.}{\sim} \text{Bin}(x_t - q_t \mid x_{t-z_t}, \gamma/(\lambda+\gamma), \\ q_t \mid x_{t-z_t}, \kappa, \lambda, \gamma, \eta &\overset{ind.}{\sim} \text{NB}(q_t \mid \kappa + x_{t-z_t}, 1 - \lambda/(2\lambda+\gamma+\eta)), \\ z_t \mid \boldsymbol{w} &\overset{i.i.d.}{\sim} \sum_{l=1}^{L} w_l \delta_l(\cdot), \end{aligned}$$

for $t = L+1, \ldots, n$, given the first $L = 5$ initial values. We took $\lambda = 5, \gamma = 3, \kappa = 3, \eta = 2$, and specified exponentially decreasing weights such that $w_i \propto \exp(-i)$, $i = 1, \ldots, 5$. As a result, the synthetic data was over-dispersed, with empirical mean and variance being 12.95 and 67.46, respectively.

We applied the PMTD and the NBMTD models to the synthetic data. For efficient posterior simulation, we reparameterized both models. In particular, for both models, we used $\theta = \gamma/(\lambda + \gamma)$ as the probability of success of the binomial distribution for X_t. Furthermore, for the negative binomial model, we took $\psi = 1 - \lambda/(2\lambda+\gamma+\eta)$ as the probability of success of the negative binomial distribution for Q_t. Implementation details of the two models are provided in Section 4.

For each model, we chose two different orders, with one correctly specified, $L = 5$, and the other one over-specified, $L = 15$, based on the autocorrelation and partial autocorrelation functions. The priors for θ and ψ were elicited based on



(a) NBMTD-SB (b) NBMTD-CDP (c) PMTD-SB (d) PMTD-CDP

Figure 2.3: Chapter 2 - second simulation data analysis. Inference results for the weight when $L = 5$ (top) and $L = 15$ (bottom). Black dashed lines are true weights, red dot-dashed lines are prior means, blue solid lines are posterior means, and blue polygons are 95% credible intervals.

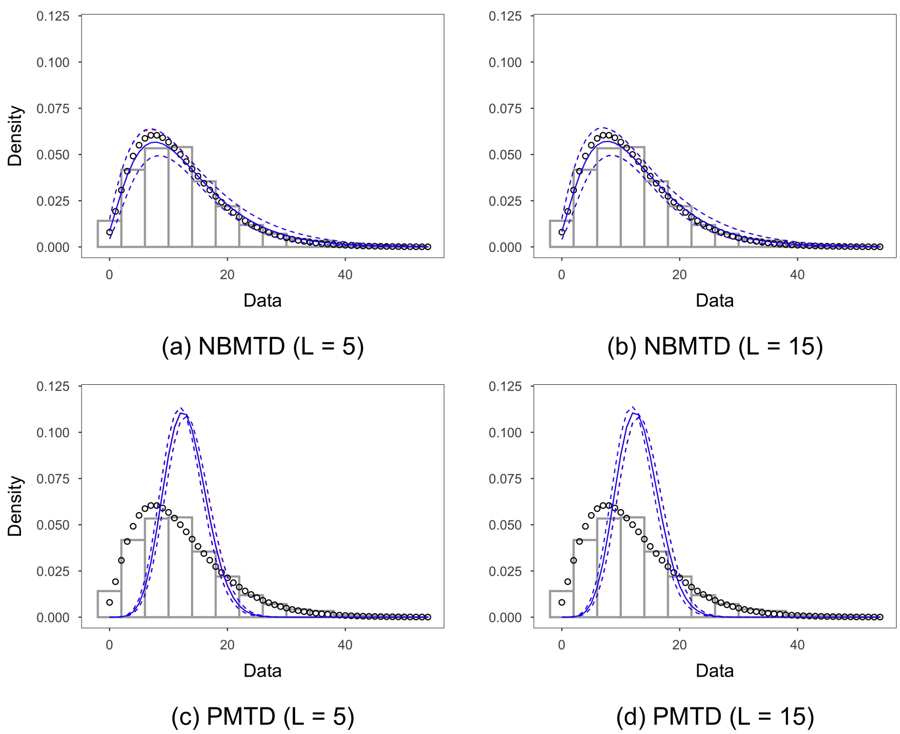


Figure 2.4: Chapter 2 - second simulation data analysis. Inference results for the stationary marginal distributions. White bars are histogram of the data. Circles are probabilities of the true marginal distribution NB$(3, 0.2)$ evaluated at the effective support. Red (blue) solid lines are posterior means from the fitted model with SB (CDP) prior. Red (blue) dashed lines are 95% credible intervals from the fitted model with SB (CDP) prior.

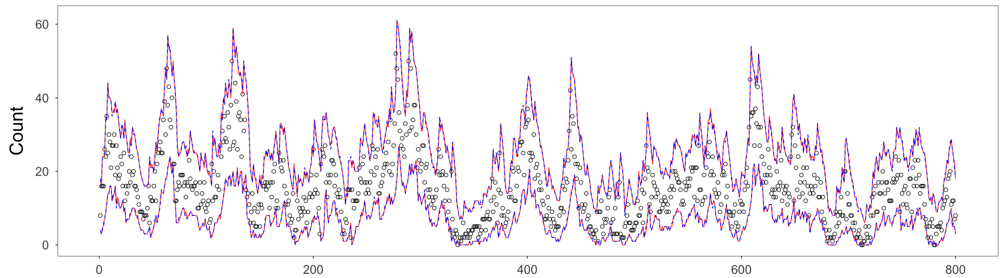
Table 2.1: Chapter 2 - second simulation data analysis. Empirical coverage of the 95% predictive intervals.

	NBMTD-SB	NBMTD-CDP	PMTD-SB	PMTD-CDP
L = 5	0.96	0.96	0.86	0.86
L = 15	0.96	0.96	0.87	0.88

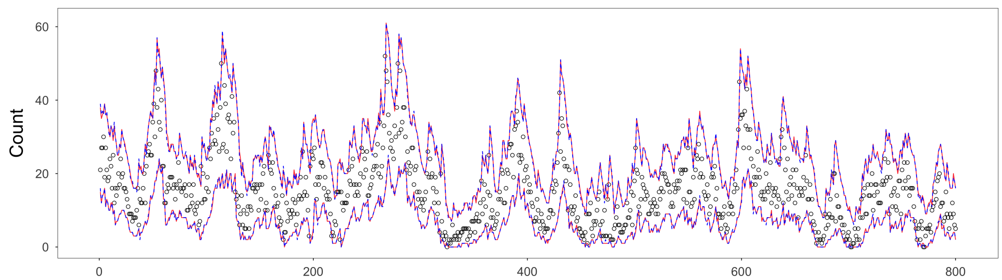
priors for (λ, γ, η). We took Ga$(2, 1)$ for each of (λ, γ, η), which implies that both θ and ψ follow beta distributions Beta$(\theta \mid 2, 2)$ and Beta$(\psi \mid 6, 2)$, respectively. We also assigned Ga$(2, 1)$ to κ. For the weights, we considered both the truncated stick-breaking and cdf-based priors. For the former prior, we took $\alpha_s = 1, 2$ corresponding to the orders, and for the latter one, we chose $\alpha_0 = 5$, $a_0 = 1$, and $b_0 = 3, 6$ respectively for the orders. To obtain the estimates, in each case, we ran a Gibbs sampler for 85000 iterations, discarding the first 5000 samples as burn-in, and collected samples every 10th iterations.

We focus on the results in estimating the weights and the stationary marginal distributions. Figure 2.3 shows that the NBMTD model was able to capture the weights in all cases, while the PMTD model systematically missed the first weight, in terms of the 95% credible interval estimates. Moreover, even when L is correctly specified, the PMTD model missed the last three weights. Figure 2.4 illustrates the stationary marginal estimated by the two models. In each case, the same model with the two proposed priors provided estimates that were almost identical. As expected, the PMTD model was not capable of recovering the marginal, while the NBMTD model provided an accurate estimate.

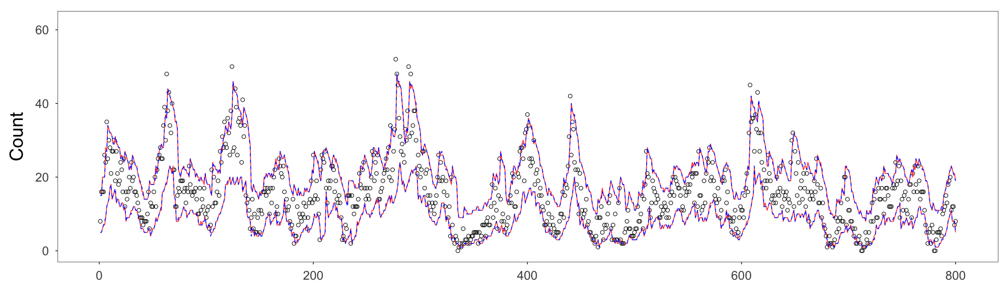
Turning to the predictive performance of the two models, Figure 2.5 shows the one-step ahead 95% posterior predictive intervals for the data. Under a visual examination, we can observe that the predictive intervals estimated by the NBMTD model were able to cover most of the small or large values, while the es-



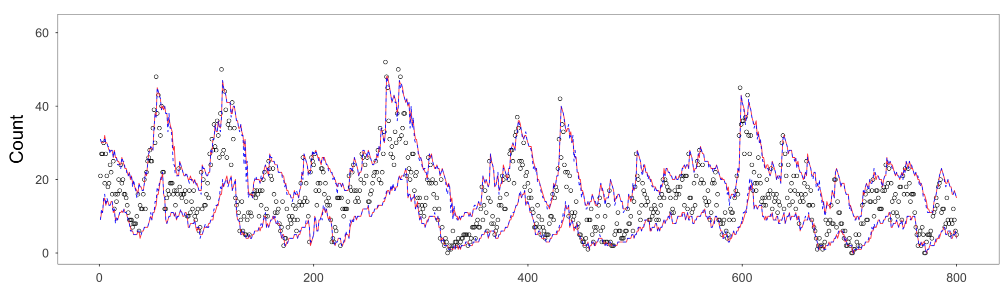
(a) NBMTD (L = 5)



(b) NBMTD (L = 15)



(c) PMTD (L = 5)



(d) PMTD (L = 15)

Figure 2.5: Chapter 2 - second simulation data analysis. 95% one-step ahead posterior predictive intervals. Red (blue) dashed lines are predictive intervals from the fitted model with SB (CDP) prior.

timated predictive interval by the PMTD model missed many such values. Table 2.1 presents the empirical coverage of the 95% posterior predictive intervals. We see that the NBMTD model provided a close estimate, while the PMTD model underestimated the coverage by a large margin.

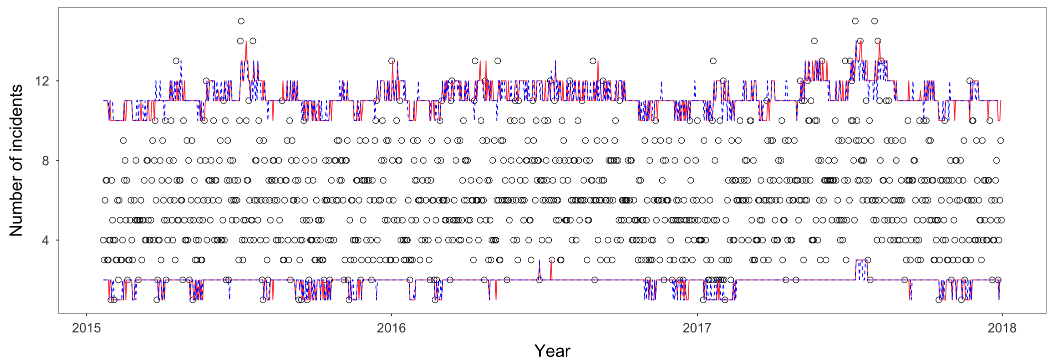
Overall, we note the NBMTD model's ability to account for over-dispersion. Moreover, even when L was over-specified, the model provided estimates that were very close to the ones under the model with L correctly specified.

2.5.3 Chicago Crime Data

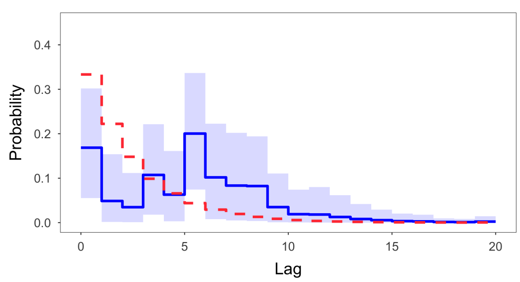
The first real data example involves the 1090 daily reported incidents of domestic-related theft that have occurred in Chicago from 2015 to 2017, extracted online from the Chicago Data Portal (`https://data.cityofchicago.org/`). The data exhibits some flat stretches, without evidence of overdispersion. The empirical mean and variance are 6.05 and 6.39.

We applied the Poisson MTD model discussed in Example 2 of Section 3, with order $L = 20$, selected based on the autocorrelation and partial autocorrelation functions. We reparameterize the model in terms of rate parameter λ, and binomial probability $\theta = \gamma/\phi$ for $Z_t \mid X_{t-l}$. This allows updates for λ and θ with posterior full conditionals available in closed form. The prior for (λ, θ) was taken to be $\mathrm{Ga}(\lambda \mid 2, 1)\mathrm{Beta}(\theta \mid 2, 2)$, implying a $\mathrm{Ga}(4, 1)$ prior for ϕ. Two priors, $\mathrm{SB}(w \mid 2)$ and $\mathrm{CDP}(w \mid 5, 1, 8)$, were considered for the mixture weights. Both models were fitted to the entire data set. After fitting the model, we obtained the one-step posterior predictive distribution at each time t and the corresponding posterior predictive intervals.

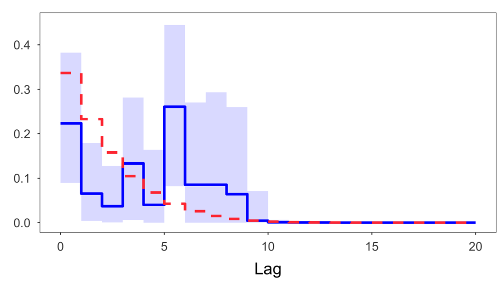
We obtained a thinned sample retaining every 10th iteration, from a total of 85000 samples with the first 5000 as burn-in. The posterior mean and 95%



(a) 95% one-step posterior predictive intervals for the crime data



(b) SB



(c) CDP

Figure 2.6: Chapter 2 - crime data analysis. In panel (a), the circles denote the data, and solid and dashed lines correspond to the model with the SB and CDP prior, respectively. Panels (b) and (c): prior means (dashed line), posterior means (solid line) and 95% credible intervals (polygon) of the weights under the SB and CDP prior, respectively.

interval for ϕ are $6.04\,(5.79, 6.30)$ and $6.05\,(5.82, 6.29)$ for models with $\mathrm{SB}(\boldsymbol{w} \mid 2)$ and $\mathrm{CDP}(\boldsymbol{w} \mid 5, 1, 8)$ priors. This indicates an average of around six incidents of domestic-related theft per day. Multiple influential lags, with gaps in between, are suggested by the results in Figure 2.6(b)-2.6(c). Both models agree on the pattern for the weights, as well as on lags 1, 4, 6 being the most relevant ones. Compared to the truncated stick-breaking prior, the cdf-based prior suggests a weight pattern that decreases slightly faster, and it assigns relatively larger weights to important lags, albeit with higher uncertainty. Figure 2.6(a) shows that both models produce similar one-step predictive intervals.

Randomized quantile residual analysis results were similar for both models as shown in Figure 2.7. The figure shows posterior mean and interval estimates

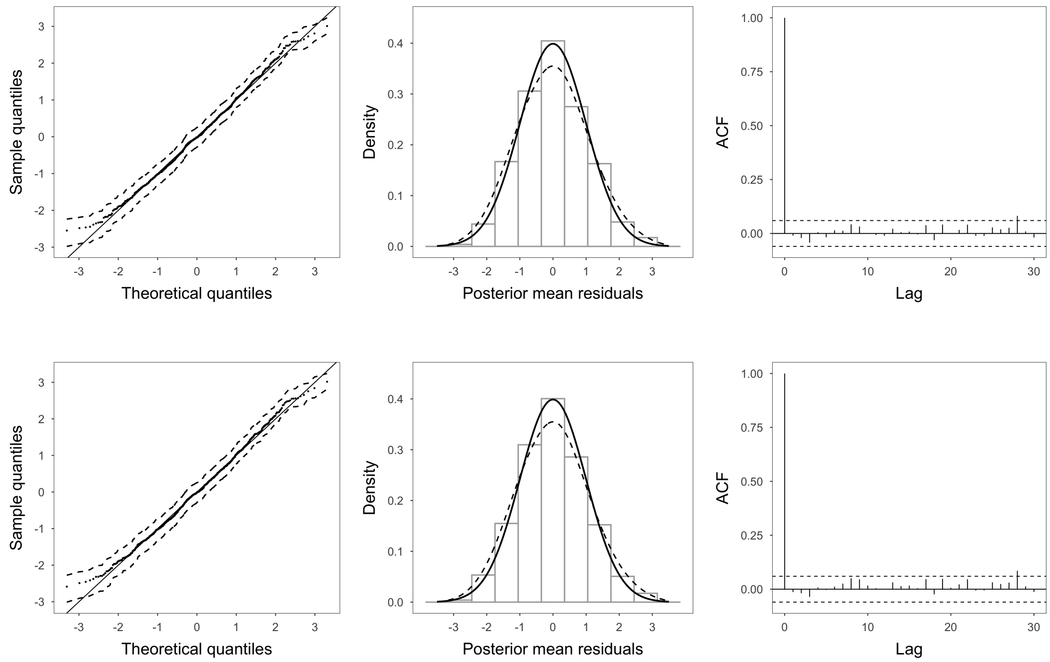


Figure 2.7: Chapter 2 - crime data analysis. Randomized quantile residual analysis for the fitted model with the $\mathrm{SB}(w \mid 2)$ prior (top) and $\mathrm{CDP}(\boldsymbol{w} \mid 5, 1, 8)$ prior (bottom), respectively. In the left column, the circles and dashed lines correspond to the posterior mean and 95% interval bands, respectively. In the middle column, the solid and dashed line are the standard Gaussian density and the kernel density estimate of the posterior means of the residuals, respectively. The right column is based on the posterior means of the residuals.

for the Gaussian quantile-quantile plot, and the histogram and autocorrelation function for the posterior means of the residuals. The results suggest reasonably good model fit, providing an illustration of the flexibility of the proposed MTD model to capture non-Gaussian tails.

2.5.4 Tunkhannock Creek Precipitation Data

Our second real data example involves 22 years of rainfall data from January 1982 to December 2003. The data consists of 1149 mean areal precipitation amounts ranging from 0.01 to 128.87 millimeters, aggregated to a weekly time scale from the daily data for the Tunkhannock Creek near Tunkhannock, Penn-

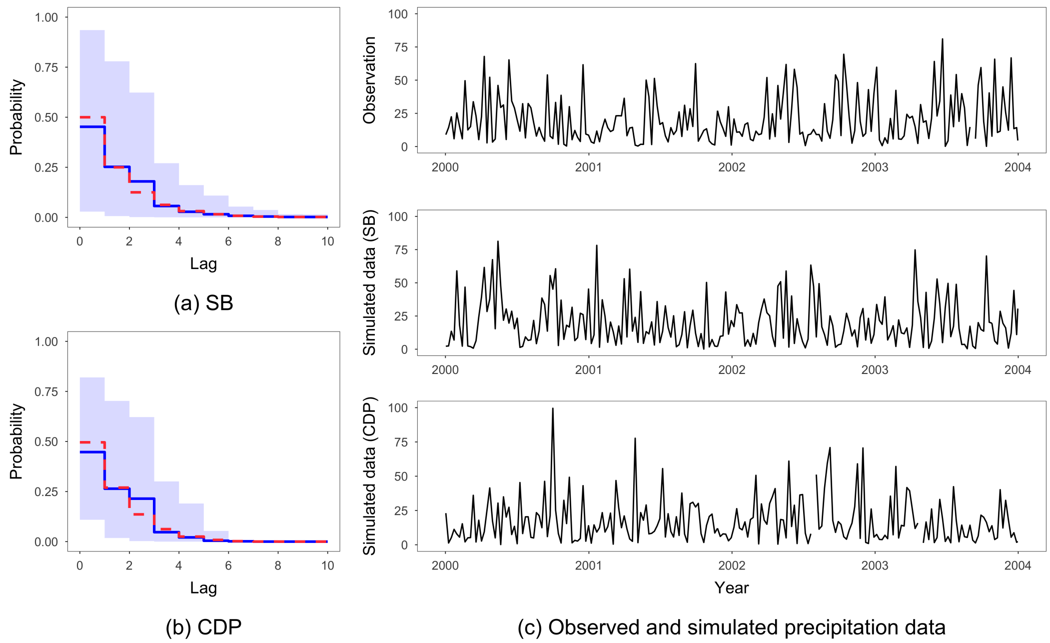


Figure 2.8: Chapter 2 - precipitation data analysis. Panels (a) and (b): prior means (dashed line), posterior means (solid line) and 95% intervals (polygons) of the weights under two priors. The top row of panel (c) plots the observed precipitation amounts from 2000 to 2004, and the middle and bottom rows show sample paths generated from the fitted models with SB and CDP priors, respectively.

sylvania. The data was extracted through R package hddtools (Vitolo, 2017).

We consider a multiplicative model $y_t = \mu_t \epsilon_t$, where μ_t is a seasonal factor and ϵ_t is generated by a Lomax MTD model specified in Equation (2.10), with polynomial tails that can accommodate large precipitation events. More specifically, the model is given by

$$
\begin{aligned}
y_t = \mu_t \epsilon_t, \quad \mu_t = \exp(\boldsymbol{x}_t^\top \boldsymbol{\beta}), \quad t = 1, \dots, n, \\
\epsilon_t \mid \boldsymbol{\epsilon}^{t-1}, w, \phi, \alpha \sim \sum_{l=1}^{L} w_l \, P(\epsilon_t \mid \phi + \epsilon_{t-l}, \alpha), \quad t = L+1, \dots, n,
\end{aligned}
\tag{2.15}
$$

with $\boldsymbol{x}_t = (\cos(\omega t), \sin(\omega t), \cos(2\omega t), \sin(2\omega t), \cos(3\omega t), \sin(3\omega t))^\top$ and $\omega = 2\pi/T$ where $T = 52$ is the period for weekly data. On the basis of the autocorrelation and partial autocorrelation functions, we chose model order $L = 10$. The

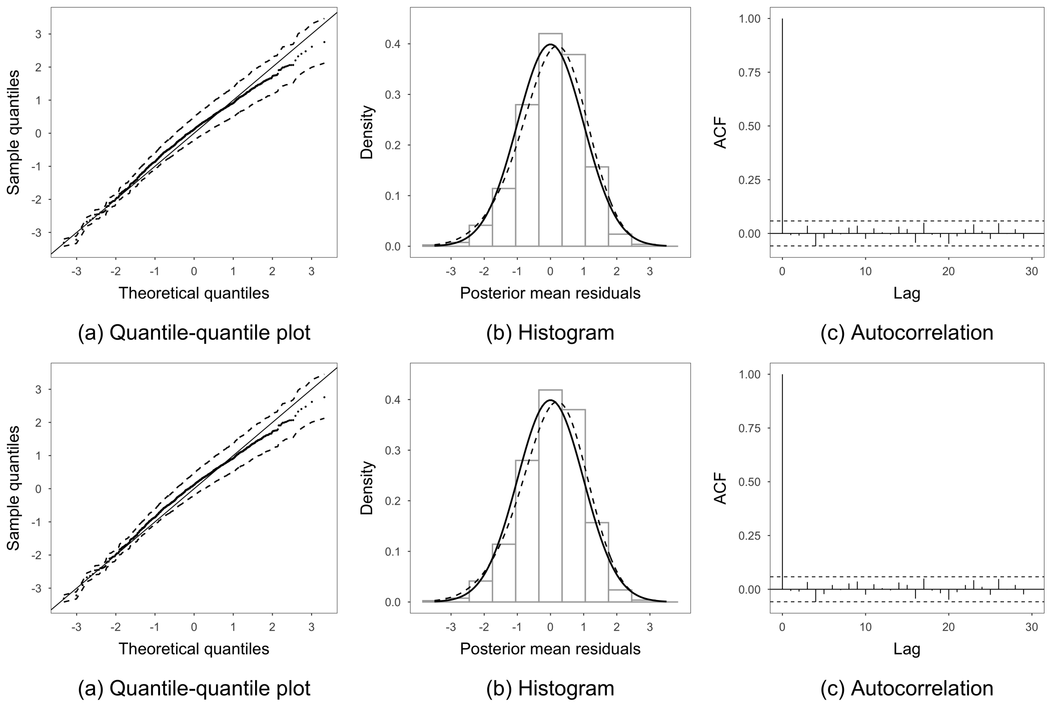


Figure 2.9: Chapter 2 - precipitation data analysis. Randomized quantile residual analysis for the fitted model with the $\mathrm{SB}(w \mid 1)$ prior (top) and $\mathrm{CDP}(\boldsymbol{w} \mid 5, 1, 6.5)$ prior (bottom), respectively. In panels (a), the circles and dashed lines correspond to the posterior mean and 95% interval bands, respectively. In panels (b), the solid and dashed line are the standard Gaussian density and the kernel density estimate of the posterior means of the residuals, respectively. Panels (c) are based on the posterior means of the residuals.

regression coefficients vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_6)^\top$ was assigned a flat prior. The shape parameter α was assigned a $\mathrm{Ga}(\alpha \mid 6, 1)$ prior, and the scale parameter ϕ an $\mathrm{IG}(\phi \mid 3, 20)$ prior. Note that the invariant marginal of the process $\{\epsilon_t\}$ is $P(\epsilon \mid \phi, \alpha - 1)$ and its tail distribution function is $(1 + \epsilon/\phi)^{-(\alpha-1)}$. A small value of α indicates a heavy tail, while a large value of α ensures the existence of finite high moments. Under the priors above, $E(\alpha) = 6$, implying the expectation that the first four moments are finite with respect to both the component and marginal distributions of the Lomax MTD for $\{\epsilon_t\}$. We fit the model with $\mathrm{SB}(\boldsymbol{w} \mid 1)$ and $\mathrm{CDP}(\boldsymbol{w} \mid 5, 1, 6.5)$ priors for the weights.

We ran the algorithm for 85000 iterations and collected samples every 10 iterations after the first 5000 was discarded. The inference results were almost the same for the two models. Here we report the ones under the $\text{SB}(\boldsymbol{w} \mid 1)$ prior. The posterior mean and 95% credible interval of the shape parameter α are $14.80\,(10.30, 20.91)$, indicating a moderately heavy tail. The corresponding estimates for the scale parameter ϕ are $254.33\,(166.36, 370.04)$, indicating substantial dispersion. Among the harmonic component coefficients, the first and the fourth have 95% posterior credible intervals that indicate statistical significance; the estimates are $-0.14\,(-0.23, -0.05)$ for β_1, and $-0.13\,(-0.22, -0.03)$ for β_4, implying the presence of semiannual and annual seasonality in the data. Figure 2.8(a)-2.8(b) show that both models suggest a decreasing weight pattern, with the first three lags being the most influential. As shown in Figure 2.8(c), the sample paths generated from the models resemble the observed precipitation time series.

Randomized quantile residual analysis results were similar for both models as shown in Figure 2.9. The figure shows posterior mean and interval estimates for the Gaussian quantile-quantile plot, and the histogram and autocorrelation function for the posterior means of the residuals. The results suggest reasonably good model fit, providing an illustration of the flexibility of the proposed MTD model to capture non-Gaussian tails.

2.6 Discussion

We have developed a broad class of stationary MTD models focusing on attaining stationarity from the perspective of a distributional formulation. The advantage of our proposed approach over more traditional methods is that no constraints on the parameter space are needed. This facilitates inference for model parameters, as the need for constrained optimization or sampling is avoided. We further proposed structured priors to support flexible inference on the weights, which accommodate non-standard scenarios that a model with a Dirichlet prior may fail to capture.

The proposed constructive framework brings several options for alternative parametric families that were formidable to tackle for the MTD model and its extensions, when stationarity is a desirable property. A limitation of our approach is that, if the stationary marginal distribution shares all the parameters with the bivariate component distribution, the resulting transition component lacks component-varying parameters. One solution is to specify the bivariate distribution using a copula (Joe, 2014), which we regard as a special case of the bivariate distribution method. Given a pre-specified marginal, the construction boils down to the selection of a copula. The copula function, which brings additional component parameters, allows specifying dependence in the bivariate distribution, separately from modeling the marginal distribution. On the other hand, some properties of the resulting model, including the conditional expectation, may be intractable, and the computational cost may increase, especially in the discrete case. We will explore the use of copula functions in the latter chapters.

The class of models proposed in this chapter can be easily extended for non-stationary time series that exhibit trends and seasonality, by incorporating corresponding factors into the model, either multiplicatively or additively. This is illustrated in our second real data example. A similar approach can be applied to incorporate covariates. Therefore, this class of models is quite general, and is useful as an alternative to the existing time series models, especially when traditional models fail to capture non-Gaussian features suggested by the data.

Chapter 3

Models for Temporal Point Processes with Memory

3.1 Introduction

Temporal point processes are stochastic models for sequences of random events that occur in continuous time, with irregular durations (or inter-arrival times) between occurrence times. Data collected in such patterns appear in a wide range of applications, such as earthquake occurrence (Ogata, 1988), recurrent events (Cook et al., 2007), financial high frequency trading and orders (Hautsch, 2011), and neural spike trains (Tang and Li, 2021), to name a few. For many point patterns, it is believed that occurrence of a future event depends on historical ones. This motivates the use of point processes with memory, for example, the Hawkes process (Hawkes, 1971a,b) with full memory. Another commonly-used model is the renewal process in which an event occurs depending on the most recent event. The renewal process has an additional property that between event times are independent and identically distributed (i.i.d.) durations. The goal of this chapter is to propose a modeling framework for point processes with high-order memory, relaxing the independent duration assumption in the renewal process, and including the ability to model duration clustering behaviors that are present in applications such as health care (Yang et al., 2018), climatology (Cowpertwait, 2001), and finance (Easley and O'hara, 1992; O'hara, 1997).

A popular way for modeling point process dependence is by specifying the process conditional intensity, namely the instantaneous event rate conditional on the historical events. Under this approach, the Hawkes process has been used extensively in the literature. The conditional intensity of the Hawkes process is decomposed into a baseline intensity and a triggering component. The triggering component is commonly chosen with an excitation function such as an exponential or power law kernel. As a result, a new event causes a jump in the conditional intensity, and the Hawkes process is said to be a self-exciting point process. We refer to Reinhart (2018) and references therein for a thorough review.

This chapter explores an alternative approach for modeling point processes with memory. Specifically, we aim at models for temporal dependence of the durations which themselves form a discrete time series. Dependent models for the duration process produce conditional densities of the event arrival times, hereafter, referred to as the conditional arrival densities. These conditional arrival densities will uniquely determine the distribution of the resulting point process (Daley and Vere-Jones, 2003). Specifying a point process in this manner results in a valid conditional intensity derived from the conditional arrival densities. This is in contrast with the approach of building models for conditional intensities, which requires mathematical validation of the proposed intensity function, such as whether the function is locally integrable over a finite domain, in order to obtain a well-defined point process model.

Statistical models for durations date back to Wold (1948). In particular, Wold (1948) formulates a first-order Markov chain for the durations with an additive model representation. Subsequent works (Jacobs and Lewis, 1977; Gaver and Lewis, 1980) investigate the stationary marginal distribution of the duration process, interpreting the model as a dependent renewal process, i.e., point processes with dependent, identically distributed durations. Formulation of these models with the desired property generally results in impractical likelihood inference (Engle and Russell, 1998). Besides, the first-order Markov assumption may be too restricted in practical settings.

More recently, there has emerged a large family of models that builds from the autoregressive conditional duration (ACD) model (Engle and Russell, 1998). The ACD model assumes a multiplicative error model for the durations, in which the errors are i.i.d. and each factor is modeled as a linear function of the past factors and durations. Extensions of this class aimed at providing more flexibility through specification on the multiplicative factor or the error distribution; see, e.g., Grammig and Maurer (2000), Bauwens and Veredas (2004), Fernandes and Grammig (2006), Deo et al. (2010), and Brownlees and Vannucci (2013). We refer to Pacurar (2008) and Bhogal and Thekke Variyam (2019) for comprehensive reviews. The ACD models are popular for high-frequency data in financial applications. However, modeling temporal dependence through multiplicative factors in a linear fashion may limit the modeling capacity for non-linear and non-Gaussian dynamics. Moreover, the model structure complicates inference for real applications when the assumption of high-order memory is necessary.

The ACD model and its extensions assume independent durations conditional on a function or process that incorporates temporal dependence. A different approach is to directly model the transition mechanism of the duration process.

Recall that the MTD model describes the transition density of the process as a weighted combination of first-order transition densities for each one of a specified number of lags. Hassan and Lii (2006) propose a bivariate MTD model for the joint conditional distribution of the duration and a continuous mark, where marks are random variables associated with the point events. Hassan and El-Bassiouni (2013) further extend the bivariate model to include a discrete mark. However, point process properties such as stationarity were not investigated in their works. Moreover, their framework involves particular structures that require pairs of certain families of distributions for the duration and mark, which can be practically restrictive. Hassan and Lii (2006) point out that the choice of parameterization is non-trivial in order to ensure model stability and prediction capability.

In this chapter, we introduce a class of temporal point process models that builds on the idea of modeling duration process dynamics with MTD models. Modeling dependent, positive-valued durations poses great challenges to employing traditional high-order autoregressive models without transforming the durations. These challenges include model inference under a constrained, possibly high-dimensional parameter space, e.g., negative coefficients that do not lead to a negative-valued duration, and practical implementation for stationarity conditions. The aforementioned works using MTD models attempt to handle the former issue, albeit under restrictive structures. A major contribution of the present work is the development of an MTD point process (MTDPP) constructive framework that provides flexible modeling of high-order, non-Gaussian dynamics of the duration process without parameter constraints. The framework not only allows for model construction for various types of practically relevant point patterns such as those with self-excitation or self-regulation effects, but also retains efficient implementation for model inference. The MTDPP likelihood evaluation grows linearly with the number of events, thus delivering computational scalability, especially for large point pattern data where the high-order memory assumption suffices.

Within the MTDPP framework, we provide easily-implemented conditions to construct point processes given a pre-specified family for the duration marginal distribution, and obtain a limit result analogous to that in the renewal theory. The resulting class of models with identically distributed, high-order dependent durations can be interpreted as a class of dependent renewal processes, which relaxes the independent duration assumption that may be unrealistic in practice. Besides, the proposed framework features an extension to flexibly describe duration clustering. We achieve the extension by using a two-component mixture model for the conditional duration density. In particular, one component of the mixture corresponds to an independent duration model that accounts for external factors. The other component is an MTDPP that explains self-excitation. Point patterns of this type can be found, for example, in hospital emergency department visits of patients, where long duration was observed between clusters of multiple visits in short bursts (Yang et al., 2018), and in the financial market in which fluctuation can be caused by either external or internal processes. (Filimonov and Sornette, 2012). The proposed extension accounts for the possibility of different factors that can drive the point process dynamics and provides quantification.

The remainder of the chapter is organized as follows. Section 3.2 introduces the MTDPP framework, studies model properties, and presents examples for constructing various types of MTDPPs and extensions. Section 3.3 develops the Bayesian hierarchical model, MCMC algorithms, and point process model validation method. Section 3.4 illustrates the proposed framework with synthetic and real data examples.

3.2 Temporal MTD Point Processes

3.2.1 Background

We consider a temporal point process $N(t)$ defined on the positive half-line $\mathbb{R}^+$, where $N(t) = \sum_{i\geq 1} \mathbf{1}_{\{t_i \leq t\}}$ is a right-continuous integer-valued function, and $t_1, t_2, \cdots \in \mathbb{R}^+$ denote the event (or arrival) times. A temporal point process is usually modeled via specifying its conditional intensity, defined as $\lambda^*(t) \equiv \lambda(t \,|\, \mathcal{H}_t) = \lim_{dt\to 0} E[dN(t) \,|\, \mathcal{H}_t]/dt$, where $dN(t) = N(t+dt) - N(t)$, and $\mathcal{H}_t$ is the history of the events up to but not including t, and dt is an arbitrary small interval. We say a point process has memory if $\mathcal{H}_t$ is not empty. A Poisson process is an example of a memoryless process. A renewal process has the least limited memory, that is, $\mathcal{H}_t = t_{N(t)}$ where $t_{N(t)}$ is the most recent arrival time before t, while evolution of a Hawkes process depends on the entire past. Given an observed point pattern $\{t_i\}_{i=1}^n$ over $(0, T)$, the likelihood of the point pattern using the conditional intensity $\lambda^*(t)$ is expressed as

$$p(t_1, \dots, t_n) = \left(\prod_{i=1}^{n} \lambda^*(t_i) \right) \exp\left(- \int_0^T \lambda^*(t) dt \right). \tag{3.1}$$

An alternative way to characterize the probability structure of a point process is to use the collection of conditional arrival densities, denoted as $p_i^*(t) \equiv p_i(t \,|\, \mathcal{H}_t)$, supported on (t_{i-1}, ∞), with associated conditional survival functions $S_i^*(t) = 1 - \int_{t_{i-1}}^t p_i^*(u) du$, for $i = 2, \dots, n+1$. When $i = 1$, $p_1^*(t) \equiv p_1(t)$ and $S_1^*(t) = 1 - \int_0^t p_1^*(u) du$. The conditional arrival densities p_i^* jointly define the density (if exists) of the Janossy measure that uniquely determines a point process. Additionally, the Janossy density is exactly the likelihood of the point process. We can write the likelihood using the conditional arrival densities $p_i^*(t)$ as

$$p(t_1, \dots, t_n) = \left(\prod_{i=1}^{n} p_i^*(t_i)\right)\left(1 - \int_{t_n}^{T} p_{n+1}^*(u)du\right). \tag{3.2}$$

The last term in (3.2) defines the likelihood normalizing constant, which corresponds to the probability of no events occurring in the interval $(t_n, T]$.

Using the collection of conditional densities p_i^* and conditional survival functions S_i^*, we can define the hazard functions as $\lambda_i^*(t) = p_i^*(t)/S_i^*(t)$, for $i = 1, \dots, n$. The hazard function is naturally interpreted as the conditional instantaneous event rate. Consequently, given the set of arrival times, we can write the conditional intensity of the process as $\lambda^*(t) = \lambda_i^*(t),\ t_{i-1} < t \leq t_i,\ 1 \leq i \leq n$. Observing that $p_i^*(t) = \lambda_i^*(t) \exp(-\int_{t_{i-1}}^{t} \lambda_i^*(u)du)$, we can easily recover the process likelihood in the form of (3.1) from (3.2).

Although there is a one-to-one correspondence between modeling the conditional intensities and the conditional arrival densities for a point process, we note that the model likelihood formulation may result in possibly different computational costs. The logarithm of the likelihood (3.1) involves logarithms of conditional intensities and an integral over $(0, T)$. For the Hawkes process, the conditional intensity involves the sum of the excitation functions over historical points, which poses great challenges to model inference (Veen and Schoenberg, 2008). Moreover, the computation burden to evaluate the Hawkes process likelihood grows quadratically with the number of observed points, making it infeasible for large data sets. Point process models defined using conditional arrival densities typically assume limited memory, with an autoregressive structure on the durations. The resulting likelihood based on (3.2) is similar to that of an autoregressive time series, with an extra likelihood normalizing constant term. The tractability of the normalizing constant depends on particular model formulations. Nevertheless, the likelihood (3.2) overall is simpler than (3.1). Thus, model inference based on (3.2) is easier to implement.

3.2.2 MTD Models for the Conditional Duration Density

Consider an ordered sequence of arrival times $0 = t_0 < t_1 < \cdots < t_n < T$. Let $x_i = t_i - t_{i-1}$ be the durations, for $i = 1, \ldots, n$. We assume the point process evolves with limited memory. The memory of the process is modeled by specifying an MTD model for the conditional densities of x_i for all i. We motivate the construction using a duration x_i for $i > L$. A formal definition of the MTDPP is given in Definition 1. We specify the density of x_i conditional on the past as a weighted combination of first-order transition densities, each of which depends on a specific past duration, that is, $f(x_i \,|\, x_{i-1}, \ldots, x_1) = \sum_{l=1}^{L} w_l f_l(x_i \,|\, x_{i-l})$, where $w_l \geq 0$ for all l and $\sum_{l=1}^{L} w_l = 1$. Transforming the conditional density of x_i to that for the arrival time $t_i = t_{i-1} + x_i$, for every i, creates conditional arrival densities that uniquely determine a point process. The formal definition of the MTDPP is given as follows.

Definition 1. *Let $N(t)$ be a temporal point process defined on $\mathbb{R}^+$ with event arrival times $t_1, t_2, \cdots \in \mathbb{R}^+$. Denote by $f^*(t - t_{N(t)}) \equiv f(t - t_{N(t)} \,|\, \mathcal{H}_t)$ the conditional duration density. Then $N(t)$ is said to be an MTD point process if (i) $t \sim f_0$ for $N(t) = 0$; (ii) for $1 \leq N(t) \leq L - 1$, the conditional duration density*

$$\begin{aligned} f^*(t - t_{N(t)}) = \sum_{l=1}^{N(t)-1} & w_l f_l(t - t_{N(t)} \,|\, t_{N(t)-l+1} - t_{N(t)-l}) + \\ & (1 - \sum_{r=1}^{N(t)-1} w_r) f_{N(t)}(t - t_{N(t)} \,|\, t_1); \end{aligned} \tag{3.3}$$

(iii) for $N(t) \geq L$, the conditional duration density

$$f^*(t - t_{N(t)}) = \sum_{l=1}^{L} w_l f_l(t - t_{N(t)} \,|\, t_{N(t)-l+1} - t_{N(t)-l}). \tag{3.4}$$

In both (3.3) *and* (3.4)*, the weights $w_l \geq 0$ for $l = 1, \dots, L$, with $\sum_{l=1}^{L} w_l = 1$.*

Remark 1. Using the marginal density f_0 and the conditional density $f^*(t - t_{N(t)})$ given in Definition 1, we can define the conditional arrival densities p_i^* for an observed point pattern$\{t_i\}_{i=1}^n$, by taking $p_1^*(t) = f_0$ and $p_i^*(t) = f^*(t - t_{i-1}),\ t_{i-1} < t \leq t_i,\ i = 2, \dots, n$. Thus, specification of the densities f_0 and $f^*(t - t_{N(t)})$ suffices to characterize the probability structure of the resulting MTDPP.

Remark 2. The two different expressions (3.3) and (3.4) for the conditional duration density $f^*(t - t_{N(t)})$ allow us to study stationarity conditions for the MTDPP in Section 3.2.3. As we develop model inferential method based on conditional likelihood, Equation (3.4) is the relevant expression for inference. For brevity, we will use (3.4) to discuss model properties throughout the rest of the chapter.

The specification of the conditional density $f^*(t - t_{N(t)})$ involves building the first-order transition density f_l for $l = 1, \dots, L$. Following Chapter 2, we build the transition density f_l from a bivariate positive-valued random vector (U_l, V_l) with joint density f_{U_l,V_l} and marginals f_{U_l} and f_{V_l}, by taking $f_l = f_{U_l|V_l}$ as the conditional density of U_l given V_l. In general, there are two strategies to find the joint density f_{U_l,V_l}. One is to find f_{U_l,V_l} given a specific marginal distribution, and the other one consists of specifying a pair of compatible conditional densities. The two conditional densities $f_{U_l|V_l}$ and $f_{V_l|U_l}$ are said to be compatible if there exists a bivariate density with its conditionals given by $f_{U_l|V_l}$ and $f_{V_l|U_l}$. We note that each strategy has its own benefit depending on the modeling objective. In Section 3.2.4, we illustrate construction of the transition density f_l with various examples for different goals.

An important consequence of using the MTD model for the conditional duration density is a mixture formulation for the implied conditional intensity $\lambda^*(t) \equiv h^*(t - t_{N(t)}) = f^*(t - t_{N(t)})/S^*(t - t_{N(t)})$, where $h^*(t - t_{N(t)})$ and $S^*(t - t_{N(t)})$ are the hazard and survival functions associated with $f^*(t - t_{N(t)})$, respectively. Let h_l and S_l be the hazard and survival functions associated with f_l, for all l. We can write the conditional intensity $\lambda^*(t)$ as

$$\lambda^*(t) \;=\; \sum_{l=1}^{L} w_l^*(t)\, h_l(t - t_{N(t)} \,|\, t_{N(t)-l+1} - t_{N(t)-l}), \tag{3.5}$$

with time-dependent weights $w_l^*(t)$ given by $w_l^*(t) = w_l S_l(t - t_{N(t)} \,|\, t_{N(t)-l+1} - t_{N(t)-l})/S^*(t - t_{N(t)})$ and $S^*(t - t_{N(t)}) = \sum_{l=1}^{L} w_l S_l(t - t_{N(t)} \,|\, t_{N(t)-l+1} - t_{N(t)-l})$, where $w_l^*(t) \geq 0$ and $\sum_{l=1}^{L} w_l^*(t) = 1$ for all t.

The mixture formulation of $\lambda^*(t)$ not only provides flexibility to accommodate a wide range of intensity shapes, but also guides modeling choice. Each mixture component h_l is a first-order hazard function. If we select f_l such that $h_l \leq C_l$, for constant $C_l > 0$, and for all l, then $\lambda^*(t) \leq \sum_{l=1} w_l^*(t) C_l$, for every t. Similarly, we can find a lower bound for $\lambda^*(t)$. For both cases, if $h_l \to C$ as $t \to \infty$ for all l, we have that $\lambda^*(t) \to C$ as $t \to \infty$. On the other hand, if one of the component hazard functions $h_l \to \infty$ as $t \to \infty$, then $\lambda^*(t) \to \infty$. Moreover, choosing f_l such that h_l in certain shapes allows us to construct particular types of point processes. A point process is said to be self-exciting if a new arrival causes the conditional intensity to jump, and is called self-regulating (or self-correcting) if a new arrival causes the conditional intensity to drop. If we choose f_l such that h_l monotonically decreases for all l, the resulting MTDPP is self-exciting.

3.2.3 Model Properties

We first investigate the stationarity of the MTDPP. There are different forms of stationary definition (Daley and Vere-Jones, 2003). We refer to the first-order strict stationarity, i.e., the MTDPP has a stationary marginal density for the duration process. Let g_i be the marginal density of the duration x_i for all i. Then, for $i \geq 2$, $g_i(x_i) = \int f^*(t_i - t_{i-1})\, p(t_1, \ldots, t_{i-1}) \prod_{r=1}^{i-2} d(t_r)$, where $p(t_1, \ldots, t_{i-1})$ is the joint density of the event times $(t_1, \ldots, t_{i-1})$. The constructive approach to build f_l through the bivariate random vector (U_l, V_l) allows us to obtain a stationary marginal density f_X, using the approach in Chapter 2. We summarize the conditions in the following proposition.

Proposition 3.1. *Consider an MTD point process $N(t)$ with event arrival times $t_1, t_2, \cdots \in \mathbb{R}^+$. Let $\{x_i : i \geq 1\}$ be the duration process, where $x_1 = t_1$, and $x_i = t_i - t_{i-1}$, for $i \geq 2$. The duration process has a stationary marginal density f_X if (i) $t \sim f_X$ for $N(t) = 0$; (ii) the transition density f_l in* (3.3) *and* (3.4) *is taken to be the conditional density $f_{U_l|V_l}$ of a bivariate positive-valued random vector (U_l, V_l) with marginal densities f_{U_l} and f_{V_l}, such that $f_{U_l}(x) = f_{V_l}(x) = f_X(x)$, for all $x \in \mathbb{R}^+$ and for all l.*

We refer to the class of MTDPPs that satisfies the conditions in Proposition 3.1 as the class of stationary MTDPPs. Compared to renewal processes that have i.i.d. durations, the stationary MTDPP can be interpreted as a class of dependent renewal processes, where the durations are identically distributed, and Markov-dependent, up to L-order. In fact, the independence assumption in classical renewal processes can be violated in practice (Coen et al., 2019). For example, in reliability engineering, times to failure between component replacements can be correlated (Modarres et al., 2017). The class of stationary MTDPPs is practically relevant to this type of applications.

In addition to its practical relevance, the class of stationary MTDPPs has a limit result analogous to that of the renewal process. In renewal theory for the above example, it is of interest to learn the rate of the component replacement in the long run. This corresponds to the rate at which $N(t)$ goes to infinity, i.e. $\lim_{t\to\infty} N(t)/t$. The following theorem summarizes the limit result for the stationary MTDPPs.

Theorem 1. *Consider an MTD point process $N(t)$ such that its duration process has stationary marginal density f_X with finite mean $\mu > 0$ and finite variance. It holds that, as $t \to \infty$, $N(t)/t \to 1/\mu$ a.s..*

Theorem 1 for MTDPPs relies on assumptions for the first two moments with respect to the stationary marginal density of the duration process. In practice, these assumptions may not hold. The following proposition establishes an upper bound for the mean-value function $m(t) = E[N(t)]$ for general MTDPPs.

Proposition 3.2. *Consider an MTD point process $N(t)$ with conditional intensity function given by (3.5). Then, for $N(t) \geq L$, $m(t) = E[N(t)]$ satisfies*

$$m(t) \;\leq\; M(t_1, \ldots, t_{N(t)}) + \sum_{l=1}^{L} w_l E\left[\Lambda_l(t - t_{N(t)} \,|\, t_{N(t)-l+1} - t_{N(t)-l})\right],$$

where the first term $M(t_1, \ldots, t_{N(t)}) = \sum_{i=1}^{N(t)} \sum_{l=1}^{L} w_l E\left[\Lambda_l(t_i - t_{i-1} \,|\, t_{i-l} - t_{i-1-l}\right]$, with $\Lambda_l(a - t_k \,|\, t_{k-l+1} - t_{k-l}) = \int_{t_k}^{a} h_l(u - t_k \,|\, t_{k-l+1} - t_{k-l}) du$, and the expectation is taken with respect to the probability distribution $p(t_1, \ldots, t_{N(t)})$ of the point process.

In general, the upper bound is difficult to compute analytically. However, if, for example, $h_l \leq C_l$ for all l, then we have that $\lim_{t\to\infty} m(t)/t \leq \sum_{l=1}^{L} w_l C_l$. If, furthermore, the MTDPP is stationary in the context of Proposition 3.1, then all functions h_l share the same parameters. It follows that $h_l \leq C$ for all l, and we obtain $\lim_{t\to\infty} m(t)/t \leq C$. In other words, the expected average renewal rate in the long run is no larger than the upper bound of the hazard rate.

3.2.4 Construction of the MTD Point Processes

We provide guidance to construct MTDPPs, focusing on the transition density f_l. As discussed in Section 3.2.2, we derive f_l from a bivariate density f_{U_l,V_l}, where f_{U_l,V_l} can either be specified with a pair of compatible conditionals $f_{U_l|V_l}$ and $f_{V_l|U_l}$, or be found/constructed given marginals f_{U_l} and f_{V_l}. The former is particularly useful when the objective is to construct self-exciting or self-regulating MTDPPs, by choosing $f_{U_l|V_l}$ such that its associated hazard function is monotonically decreasing or increasing, respectively. We illustrate this approach in Example 1.

The strategy of constructing MTDPPs given pre-specified families of marginals is natural with the objective of modeling dependent renewal processes, or more generally, modeling point process with the duration distribution as well as the associated hazard function in desired shapes. For example, Grammig and Maurer (2000) point out that it may be more appropriate to consider non-monotonic hazard functions for modeling financial duration processes. We achieve this goal by using bivariate copula functions that introduce dependence between two random variables given their marginals. We illustrate this strategy in Example 2.

Example 1: Self-exciting MTDPPs

We derive MTDPPs based on bivariate Lomax distributions. The Lomax distribution is a shifted version of the Pareto Type I distribution, denoted as $P(u \,|\, b, a) = ab^{-1}(1 + xb^{-1})^{-(a+1)}$, where $a > 0$ is the shape parameter, and $b > 0$ is the scale parameter. In this example, we derive a new pair of compatible Lomax conditionals, based on the pair of Lomax conditionals given in Arnold et al. (1999). The definition is given in the following proposition.

Proposition 3.3. *Consider a bivariate Lomax random vector* (X, Y) *with density* $f_{X,Y}(x, y) \propto (\lambda_0 + \lambda_1 x + \lambda_2 y)^{-(\alpha+1)}$. *Let* $(U, V) = (\alpha X, \alpha Y)$. *Then the bivariate random vector* (U, V) *has conditionals* $f_{U|V}(u|v) = P(u \mid \lambda_1^{-1}(\alpha\lambda_0 + \lambda_2 v), \alpha)$, $f_{V|U}(v|u) = P(v \mid \lambda_2^{-1}(\alpha\lambda_0 + \lambda_1 u), \alpha)$ *and marginals* $f_U(u) = P(u \mid \lambda_1^{-1}\alpha\lambda_0, \alpha - 1)$ *and* $f_V(v) = P(v \mid \lambda_2^{-1}\alpha\lambda_0, \alpha - 1)$.

In Proposition 3.3, the random vector (X, Y) is scaled by α. We refer to the distribution of (U, V) as the bivariate scaled-Lomax distribution, with difference from the original one being that the shape parameter of the scaled-Lomax distribution is part of the scale parameter.

To construct an MTDPP, we first take a set of bivariate scaled-Lomax densities f_{U_l,V_l} with parameters $\alpha_l, \lambda_{0l}, \lambda_{1l}, \lambda_{2l}$, for $l = 1, \ldots, L$. For each f_{U_l,V_l}, we simplify the parameterization by taking $\lambda_l = \lambda_{1l} = \lambda_{2l}$ and letting $\phi_l = \lambda_{0l}/\lambda_l$. Then $f_{U_l|V_l}(u|v) = P(u \mid \alpha_l\phi_l + v, \alpha_l)$. Taking $f_l \equiv f_{U_l|V_l}$ for all l, we obtain the following conditional duration density,

$$f^*(t - t_{N(t)}) = \sum_{l=1}^{L} w_l \, P(t - t_{N(t)} \mid \alpha_l\phi_l + t_{N(t)-l+1} - t_{N(t)-l}, \alpha_l). \qquad (3.6)$$

We complete the construction of a scaled-Lomax MTDPP (SLMTDPP) by letting $f_0(t) = P(t \mid \alpha_1\phi_1, \alpha_1 - 1)$. When $\alpha_l = \alpha$ and $\phi_l = \phi$ for all l, the SLMTDPP has stationary duration density $f_X(t - t_{N(t)}) = P(\alpha\phi, \alpha - 1)$. When $\alpha > 3$, by Theorem 1, as $t \to \infty$, the limit of $N(t)/t$ is $(\alpha - 2)/(\alpha\phi)$.

The following proposition describes the limiting behavior of the SLMTDPP conditional duration distribution $f^*(t - t_{N(t)})$.

Proposition 3.4. *As the shape parameter* $\alpha_l \to \infty$ *for all* l, *the conditional duration distribution* $f^*(t - t_{N(t)})$ *of the SLMTDPP is a finite mixture of exponential distributions in which the* l*th mixture component of* $f^*(t - t_{N(t)})$ *corresponds to* *an exponential distribution with rate parameter ϕ_l^{-1}. If, furthermore, $\phi_l = \phi$ for all l, the limiting conditional duration distribution is an exponential distribution with rate parameter ϕ^{-1}.*

According to (3.5), the conditional intensity of the SLMTDPP can be expressed as $\lambda^*(t) = \sum_{l=1}^{L} w_l^*(t)(\phi_l + \alpha_l^{-1}(t - t_{N(t)} + t_{N(t)-l+1} - t_{N(t)-l}))^{-1}$. Obviously, the lth component of the conditional intensity is bounded above by ϕ_l^{-1}, for all l. so $\lambda^*(t) \leq \sum_{l=1}^{L} w_l^*(t)\phi_l^{-1}$ for every t. By Proposition 3.2, we have that $\lim_{t\to\infty} m(t)/t \leq \sum_{l=1}^{L} w_l \phi_l^{-1}$.

Finally, we note that if we remove α from the scale parameter component in (3.6) such that $f_l = P(t - t_{N(t)} \mid \phi_l + t_{N(t)-l+1} - t_{N(t)-l}, \alpha_l)$ and $f_0(t) = P(t \mid \phi_1, \alpha_1 - 1)$, then f_l corresponds to the bivariate Lomax distribution of Arnold et al. (1999). The resulting point process is referred to as the Lomax MTDPP (LoMTDPP). Since the hazard function of a Lomax distribution is monotonically decreasing, both the SLMTDPP and LoMTDPP are self-exciting point processes. A self-regulating MTDPP can be constructed with a pair of compatible conditionals associated with monotonically increasing hazard functions. See, for example, the pair of gamma conditionals in Arnold et al. (1999).

Example 2: Copula MTDPPs

Recall that we can build the transition density f_l with a bivariate density f_{U_l,V_l} and marginals f_{U_l}, f_{V_l}, by letting $f_l \equiv f_{U_l|V_l} = f_{U_l,V_l}/f_{V_l}$. According to Proposition 3.1, given a stationary density f_X, we take $f_{U_l} = f_{V_l} = f_X$ for all l. Given the desired marginals, it remains to specify the joint density f_{U_l,V_l} to obtain $f_{U_l|V_l}$. In this example, we introduce the idea of specifying a bivariate copula function $C : [0,1]^2 \to [0,1]$ to build f_{U_l,V_l}, which provides a general scheme to construct MTDPPs given a stationary marginal f_X.

Let F_{U_l,V_l} be the joint cdf of the random vector (U_l, V_l), and denote F_{U_l}, F_{V_l} as the corresponding marginal cdfs. Given F_{U_l} and F_{V_l}, by Sklar (1959), there exists a unique copula C_l such that $F_{U_l,V_l}(u,v) = C_l(F_{U_l}(u), F_{V_l}(v))$, and the joint density f_{U_l,V_l} is given by $c_l(u,v)f_{U_l}(u)f_{V_l}(v)$, where $c_l(u,v) = \partial C(F_{U_l}(u), F_{V_l}(v))/(\partial F_{U_l}\partial F_{V_l})$ is the copula density. Then given f_X with a copula C_l, we take $f_l(u) \equiv f_{U_l|V_l}(u \mid v) = c_l(u,v)f_X(u)$. The conditional duration density of the resulting MTDPP is

$$f^*(t - t_{N(t)}) = \sum_{l=1}^{L} w_l c_l(t - t_{N(t)}, t_{N(t)-l+1} - t_{N(t)-l}) f_X(t - t_{N(t)}). \tag{3.7}$$

We refer to this class of models as copula MTDPPs. The conditional intensity of this class written in the form of (3.5) consists of hazard functions $h_l(u \mid v) = f_l(u \mid v)/S_l(u \mid v)$, where $S_l(u \mid v) = 1 - \partial C_l(F_{U_l}(u), F_{V_l}(v))/\partial F_{V_l}$. Existence of a closed-form expression for h_l depends on the selection of the copula function. For example, a Gaussian copula that has intractable integrals for the cdf leads to an analytically intractable h_l.

Using copula with marginals to specify bivariate densities can obtain conditionals that are in the same family as the marginals. Consider a heavy right tail (HRT) copula with Burr marginals. The conditional distribution is also a Burr distribution (Venter et al., 2002). Therefore, we can construct a class of Burr MTDPP (BuMTDPP) with stationary Burr marginals. More specifically, denote the Burr distribution as $\text{Burr}(x \mid \gamma, \lambda, \kappa) = \kappa\gamma x^{\gamma-1}\lambda^{-\gamma}(1 + (x/\lambda)^{\gamma})^{-(\kappa+1)}$, with shape parameters $\gamma > 0, \kappa > 0$, and scale parameter $\lambda > 0$. We let $f_{U_l}(x) = f_{V_l}(x) = f_X(x) = \text{Burr}(x \mid \gamma, \lambda, \kappa - 1)$, for all l, and specify an HRT copula C_l such that $f_{U_l,V_l}(u,v) = c_l(u,v)f_X(u)f_X(v)$. Then the conditional density $f_{U_l|V_l}(u \mid v) = \text{Burr}(u \mid \gamma, \lambda + v, \kappa)$. The resulting BuMTDPP conditional duration density is given by

$$f^*(t - t_{N(t)}) = \sum_{l=1}^{L} w_l \text{Burr}(t - t_{N(t)} \mid \gamma, \lambda + t_{N(t)-l+1} - t_{N(t)-l}, \kappa), \tag{3.8}$$

with stationary marginal $f_X(t-t_{N(t)}) = \text{Burr}(t-t_{N(t)} \mid \gamma, \lambda, \kappa-1)$. The component hazard function h_l of the resulting conditional intensity written as in (3.5) is $\kappa\gamma(t-t_{N(t)})^{\gamma-1}\{(\lambda+t_{N(t)-l+1}-t_{N(t)-l})^{\gamma}+(t-t_{N(t)})^{\gamma}\}^{-1}$. The parameter γ controls the shape of the function h_l. When $0 \leq \gamma \leq 1$, h_l is monotonically decreasing. When $\gamma > 1$, h_l is hump-shaped.

In fact, the stationary BuMTDPP model includes special cases. If $\gamma = 1$, it becomes an LoMTDPP with stationary marginal $P(t-t_{N(t)} \mid \lambda, \kappa-1)$. When $\kappa = 2$, it reduces to a model with stationary log-logistic marginal $LL(t - t_{N(t)}) \mid \gamma, \lambda)$, where $LL(x \mid \gamma, \lambda) = \gamma x^{\gamma-1}\lambda^{-\gamma}(1 + (x/\lambda)^{\gamma})^{-2}$. Moreover, if the Burr distribution is reparameterized to that of Lancaster (1990), according to Grammig and Maurer (2000), it can be shown that as $\kappa \to \infty$, the BuMTDPP turns to a model with a stationary Weibull distribution.

3.2.5 Extension to MTD Cluster Point Processes

A self-exciting MTDPP encourages clustering behaviors. In practice, the clustering behaviors in the duration process may involve different factors. As an example, consider in hydrology the durations being dry spells are classified into two types, corresponding to cyclonic and anticyclonic weather (Cowpertwait, 2001). It is expected that a point process model for the data is able to account for the two weather types as the lengths of the dry spells could be distinctly different. Similar examples can also be found in Li et al. (2021). This section extends the MTDPP to a class of MTD cluster point process (MTDCPP) based on a two-component mixture model. The defintion of the MTDCPP is given as follows.

Definition 2. *Let $N(t)$ be a temporal point process defined on $\mathbb{R}^+$ with event arrival times $t_1, t_2, \cdots \in \mathbb{R}^+$. Let $f^*(t - t_{N(t)})$ be the conditional duration density of a self-exciting MTD point process. Then $N(t)$ is said to be an MTD cluster* *point process if (i) $t \sim f_I$ for $N(t) = 0$; (ii) for $N(t) \geq 1$, the conditional duration density is given by*

$$f_C^*(t - t_{N(t)}) = \pi_0 f_I(t - t_{N(t)}) + (1 - \pi_0) f^*(t - t_{N(t)})), \tag{3.9}$$

where $0 \leq \pi_0 \leq 1$, and f_I is the probability density of a positive-valued random variable.

Similar to the MTDPP, we use densities f_I and $f_C^*(t - t_{N(t)})$ to define the conditional arrival densities p_i^* of event time t_i, for an observed point pattern$\{t_i\}_{i=1}^n$, by taking $p_1^*(t) = f_I$ and $p_i^*(t) = f_C^*(t - t_{i-1}),\ t_{i-1} < t \leq t_i,\ i = 2, \dots, n$, which characterize the probabilistic structure of the point process. The MTDCPP includes several special cases. When $\pi_0 = 1$, the model reduces to a renewal process; furthermore, if f_I corresponds to an exponential distribution, it becomes a Poisson process. When $\pi_0 = 0$, the MTDCPP reduces to an MTDPP. When $0 < \pi_0 < 1$, a new duration is generated from f_I independently from the past durations with probability π_0, and is generated from a self-exciting MTDPP that depend on the past L durations with probability $1 - \pi_0$.

Let h_I be the hazard function associated with f_I. The conditional intensity of the MTDCPP is given by

$$\lambda_C^*(t) = \pi_0(t) h_I(t - t_{N(t)}) + \sum_{l=1}^{L} \pi_l(t) h_l(t - t_{N(t)} \,|\, t_{N(t)-l+1} - t_{N(t)-l}), \tag{3.10}$$

where the weights $\pi_0(t) = \pi_0 S_I(t - t_{N(t)})/S_C^*(t - t_{N(t)}),\ \ \pi_l(t) = (1 - \pi_0) w_l S_l(t - t_{N(t)} \,|\, t_{N(t)-l+1} - t_{N(t)-l})/S_C^*(t - t_{N(t)})$, for $l = 1, \dots, L$, and $S_C^*(t - t_{N(t)}) = \pi_0 S_I(t - t_{N(t)}) + (1 - \pi_0) S^*(t - t_{N(t)})$, and we have that $\pi_l(t) \geq 0$, for $l = 0, \dots, L$, and $\sum_{l=0}^{L} \pi_0(t) = 1$, for all t. Compared to the MTDPP conditional intensity function in (3.5), the MTDCPP conditional intensity has an extra term contributed from the independent component f_I, with time-dependent weights renormalized incorporating the survival function associated with f_I. If we take f_I to be an exponential density with rate μ and the MTDPP to be a stationary LoMTDPP, the MTDCPP condition intensity is $\lambda_C^*(t) = \pi_0(t)\mu + \sum_{l=1}^{L} \pi_l(t)\alpha(\phi + t - t_{N(t)} + t_{N(t)-l+1} - t_{N(t)-l})^{-1}$. We refer to this model as the Lomax MTDCPP. Note that we consider a stationary LoMTDPP instead of a stationary SLMTDPP to avoid potential identifiability problem, since the conditional duration density of the stationary SLMTDPP converges to an exponential density as α tends to infinity.

3.3 Bayesian Implementation

3.3.1 Conditional Likelihood and Prior Specification

We outline the approach to posterior inference for the MTDCPP model based on conditional likelihood. The relevant expression in Definition 1 for inference is (3.4). The MTDPP is regarded as a special case of the MTDCPP with $\pi_0 = 0$.

Let $\{t_i\}_{i=1}^{n}$ be the observed temporal point pattern over the interval $(0, T)$, with durations $x_1 = t_1$ and $x_i = t_i - t_{i-1}$ for $i = 2, \ldots, n$. The process likelihood can be expressed equivalently using $\{t_i\}$ or $\{x_i\}$. For brevity, we use the latter collection, and for convenience of noation, we take $x_{n+1} = T - t_n$. Thus, combining (3.2) and (3.9), the likelihood conditional on $(x_1, \ldots, x_L)$ is given by

$$
\begin{aligned}
& p(x_1, \ldots, x_n \mid \pi_0, \boldsymbol{w}, \boldsymbol{\phi}, \boldsymbol{\theta}) \\
& \approx \prod_{i=L+1}^{n} \left\{ \pi_0 f_I(x_i \mid \boldsymbol{\phi}) + (1 - \pi_0) \sum_{l=1}^{L} w_l f_l(x_i \mid x_{i-l}, \boldsymbol{\theta}_l) \right\} \\
& \times \left(1 - \int_0^{x_n} \left\{ \pi_0 f_I(u \mid \boldsymbol{\phi}) + (1 - \pi_0) \sum_{l=1}^{L} w_l f_l(u \mid x_{n+1-l}, \boldsymbol{\theta}_l) \right\} du \right)
\end{aligned}
\tag{3.11}
$$

where $\boldsymbol{w} = (w_1, \ldots, w_L)^\top$, and the vectors $\boldsymbol{\phi}$ and $\boldsymbol{\theta} = \{\boldsymbol{\theta}_l\}_{l=1}^{L}$ collect the parameters of the independent duration density f_I and the MTDPP component densities, respectively.

The Bayesian model involves prior specifications for the probability π_0, the MTDPP weight vector $\boldsymbol{w}$, and the density parameters $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$. The priors for $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ depend on particular choices of the densities f_I and f_l. For the probability π_0, we consider a beta prior $\mathrm{Beta}(\pi_0 \,|\, u_0, v_0)$. Without further information, we recommend a noninformative prior $\mathrm{Beta}(\pi_0 \,|\, 1, 1)$ which corresponds to a uniform prior with $E(\pi_0) = 0.5$. We take the weights w_l as increments of a cdf G, that is, $w_l = G(l/L) - G((l-1)/L)$, for $l = 1, \ldots, L$, where G has support on the unit interval. Flexible estimation of the weights w_l depends on the shape of G. We consider a DP prior that supports general distributional shape for G, denoted as $\mathrm{DP}(\alpha_0, G_0)$, where $G_0 = \mathrm{Beta}(a_0, b_0)$ is the baseline cdf, and $\alpha_0 > 0$ is the precision parameter. Given G_0 and α_0, the vector of weights $\boldsymbol{w}$ follows a Dirichlet distribution with shape parameter vector $\alpha_0(a_1, \ldots, a_L)^\top$, where $a_l = G_0(l/L) - G_0((l-1)/L)$, for $l = 1, \ldots, L$. We denote this prior for the weights as $\mathrm{CDP}(\cdot \,|\, \alpha_0, a_0, b_0)$, and refer to Chapter 2 for a discussion of its properties.

3.3.2 Bayesian Estimation

We outline an MCMC algorithm to simulate from the joint posterior distribution of the model parameters. To facilitate posterior inference, we first rewrite the transition density of the MTDCPP in (3.11) as

$$f_C^*(x_i) = \sum_{l=0}^{L} \pi_l f_l^c(x_i \,|\, \boldsymbol{\phi}, \boldsymbol{\theta}_l), \quad f_0^c = f_I, \quad f_l^c = f_l, \quad l = 1, \ldots, L, \tag{3.12}$$

where $\pi_l = (1 - \pi_0) w_l$, for $l = 1, \ldots, L$, and $\sum_{l=0}^{L} \pi_l = 1$.

The mixture model formulation in (3.12) allows us to introduce a set of configuration variables ℓ_i, taking values in $\{0, 1, \ldots, L\}$, with a discrete distribution $P(\ell_i = l) = \sum_{l=0}^{L} \pi_l \delta_l(\ell_i)$ where $\delta_l(\ell_i) = 1$ if $\ell_i = l$ and 0 otherwise, for $i = L+1, \ldots, n$. Therefore, $\ell_i = 0$ indicates that the duration x_i is generated from f_I, and $\ell_i = l$ indicates that x_i is generated from the lth component of the MTDPP, for $l = 1, \ldots, L$. Note that the second term that corresponds to the likelihood normalizing constant in (3.11) can be written as $\sum_{l=0}^{L} \pi_l S_l^c(x_{n+1} \mid \boldsymbol{\phi}, \boldsymbol{\theta}_l)$ where $S_0^c = S_I$ and $S_l^c = S_l$ for $l = 1, \ldots, L$. Similarly, we can introduce a configuration variables ℓ_{n+1} to identify the component of the mixture for x_{n+1}. The joint posterior distribution of the model parameters is given by

$$p(\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{w}, \pi_0 \mid x_1, \ldots, x_n) \;\propto\; p(\boldsymbol{\phi}) \times \prod_{l=1}^{L} p(\boldsymbol{\theta}_l) \times \mathrm{Dir}(\boldsymbol{w} \mid \alpha_0 a_1, \ldots, \alpha_0 a_L)$$
$$\times\, \mathrm{Beta}(\pi_0 \mid u_0, v_0) \times \left\{ \prod_{i=L+1}^{n} f_{\ell_i}^c(x_i) \mid \boldsymbol{\phi}, \boldsymbol{\theta}_l) \sum_{l=0}^{L} \pi_l \delta_l(\ell_i) \right\} \left\{ S_{\ell_{n+1}}^c(x_{n+1}) \sum_{l=0}^{L} \pi_l \delta_l(\ell_{n+1}) \right\}$$

The posterior updates for parameters $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ depend on choice of the density f_I and f_l, respectively. The posterior full conditional distribution of each configuration variable ℓ_i is a discrete distribution on $\{0, ..., L\}$ with probabilities proportional to $\pi_l f_l^c(x_i \mid \boldsymbol{\phi}, \boldsymbol{\theta}_l)$. Let $M_l = |\{i : \ell_i = l\}|$, for $l = 0, ..., L$, where $|\{\cdot\}|$ returns the size of the set $\{\cdot\}$. Given the configuration variables, we update the weights $\boldsymbol{w}$ with a Dirichlet posterior full conditional distribution with parameter vector $(\alpha_0 a_1 + M_1, \ldots, \alpha_0 a_L + M_L)^\top$. A beta prior for π_0 yields conjugate posterior full conditional distribution $\mathrm{Beta}(\pi_0 \mid u_0 + M_0, v_0 + \sum_{l=1}^{L} M_l)$.

3.3.3 Inference for Point Process Functionals and Model Checking

Using the MCMC algorithm, we obtain posterior samples of the model parameters that provide full posterior inference for any functional relevant to the MTDPP or MTDCPP. For example, given a posterior draw of the model parameters, we obtain a posterior realization of the conditional intensity function by evaluating (3.5) or (3.10) over a grid of time points, conditional on the observed points $\{t_i\}$. Similarly, for stationary MTDPPs, we can obtain posterior distribution of the stationary marginal distribution.

Turning to model validation. For simplicity, we introduce the approach of model checking for MTDPPs. It can be easily extended for MTDCPPs. The model goodness-of-fit is examined with the random time-change theorem (Daley and Vere-Jones, 2003). Denote by $\Lambda^*(t) = \int_0^t \lambda^*(u)du$ the compensator of a point process. Given a realized point pattern $\{t_i\}_{i=1}^n$, by the theorem, $\{\Lambda^*(t_i)\}_{i=1}^n$ is a realization from a Poisson process with unit rate. It follows that the random variables $U_i^* = 1 - \exp\{-(\Lambda^*(t_i) - \Lambda^*(t_{i-1}))\}$ are independent uniform random variables over the unit interval.

The compensator of the MTDPP event time t_i is $\Lambda^*(t_i) = \sum_{j=1}^{i} \int_{t_{j-1}}^{t_j} h^*(u - t_{j-1})du$. It follows that $\Lambda^*(t_i) - \Lambda^*(t_{i-1}) = -\log(1 - F^*(t_i - t_{i-1}))$, and $U_i^* = F^*(t_i - t_{i-1})$, where $F^*(t_i - t_{i-1}) = \int_{t_{i-1}}^{t_i} f^*(u - t_{i-1})du$, for $i = 1, \ldots, n$. The point process model can be assessed graphically using quantile-quantile plots for the estimated U_i^* with uncertainty quantification using posterior samples. If the model is correctly specified, the random variables U_i^* are independent and identically distributed as uniform distribution over the unit interval.

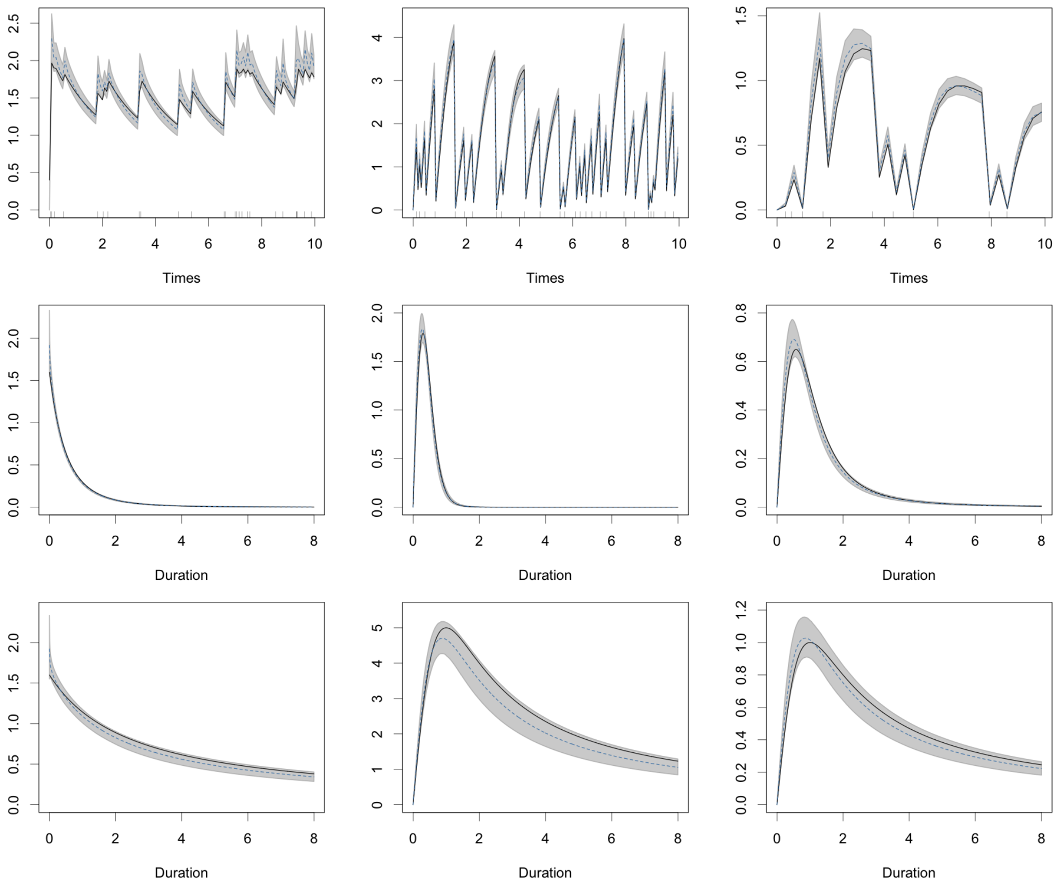
3.4 Data Illustrations

We illustrate the point process modeling framework with two synthetic data examples and a real data analysis. In the first simulation experiment, we investigate the benefit of the proposed framework for modeling dependent point processes with duration hazard functions in different shapes. The goal of the second simulation experiment is to examine the ability of the MTDCPP to recover various clustering behaviors, which is further illustrated in the real data analysis. For each scenario of the synthetic data examples, we chose interval $(0, T)$ such that there were around 2000 event times simulated within the window. Posterior analysis for each data example is based on posterior samples collected every 4 iteration from a Markov chain of 25000 iterations with the first 5000 as burn-in samples. MCMC algorithms for all models were implemented on a computer with a 2-GHz Intel Core i5 processor and 32-GB RAM. For each model implemented in the synthetic data examples, the computing time was around 1.5 minutes. In the real data analysis, the computing time for fitting the model to all 121 point patterns sequentially was around 2.5 hours.

3.4.1 First Simulation Experiment

We generated data from several MTDPP models each of which owns a different family of stationary marginal distribution. In particular, we chose three MTDPP models discussed in Section 3.2.4, with stationary scaled-Lomax, Burr, and log-logistic marginal distributions, respectively. We specified model parameters, $(\phi, \alpha) = (2, 5)$, $(\lambda, \gamma, \kappa) = (1, 2, 6)$, and $(\lambda, \gamma) = (1, 2)$, respectively, such that the associated hazard functions for the durations are decreasing for the scaled-Lomax families and hump-shaped for the last two families. We chose model order $L = 3$ for all simulations, with decaying weights $\boldsymbol{w} = (0.5, 0.3, 0.2)$.

We then applied the BuMTDPP model with $L = 3$ to the three synthetic data sets. Recall that the shape parameter γ of the model controls the shape of the hazard function, and $\gamma = 1$ is the critical value such that Burr hazard function becomes monotonic or non-monotonic when $\gamma \leq 1$ or $\gamma > 1$. Therefore, we assigned a gamma prior $\mathrm{Ga}(1, 1)$ to γ. Besides, the mth moment of the Burr



(a) Scaled-Lomax **(b)** Burr **(c)** Log-logistic

Figure 3.1: Chapter 3 - first simulation data analysis. The first, second, and third rows correspond to the posterior means (blue dashed lines) and 95% credible interval estimates (grey polygons) of the conditional intensity, marginal density, and marginal hazard function. Black solid lines are true values.

distribution exists if $\kappa\gamma > m$. We specified an independent prior $\text{Ga}(5, 1)$ for κ, implying the expectation that the first four moments exist with respect to the component Burr distribution. The scale parameter λ received $\text{Ga}(1, 1)$, and the vector of weights $\boldsymbol{w}$ was assigned $\text{CDP}(\boldsymbol{w} \mid 5, 1, 2)$, indicating decaying weights.

We focus on the inference for the point process conditional intensity, and the corresponding duration process stationary marginal density and its associated hazard function, as shown in Figure 3.1. In particular, the BuMTDPP model was able to distinguish between monotonically decreasing and hump-shaped functions for both the conditional intensity function and the hazard function associated with the stationary marginal density. For the duration hazard function, the point estimates tended to underestimate when the function decreases and overestimate when the function increases. Overall, the model provided reasonably good estimates to these functionals, with uncertainty bands that contain most of the true values. From the figure, we can also observe that the posterior mean of the hazard function is not as closer to the true one, compared to the difference between that of the marginal and true densities. This may result from the hazard function estimation which involves first estimating the survival function. Consequently, the uncertainty estimate includes the uncertainty from the estimation of the survival function. As in Figure 3.1, the posterior 95% credible interval estimates of the hazard functions are wider than those of the density functions.

3.4.2 Second Simulation Experiment

We generate data from a Lomax MTDCPP, that is, with f_I corresponding to an exponential distribution with rate μ and $f^*(t-t_{N(t)})$ a stationary LoMTDPP. We consider four scenarios, with π_0 taking one of the following values, (0.2, 0.5, 0.8, 1). The first three values indicate to certain degree is the process affected by external factors, and the last one results in simply a Poisson process. For all scenarios, we take $\mu = 0.2$, $\alpha = 5$, $\phi = 0.1$ and decaying weights $\boldsymbol{w} = (0.35, 0.25, 0.2, 0.1, 0.1)^\top$.

We applied the Lomax MTDCPP model with $L = 5$ to the synthetic data. We specified a beta prior $\mathrm{Beta}(\pi_0 \,|\, 1, 1)$ for the probability π_0 and a gamma prior $\mathrm{Ga}(\mu \,|\, 1, 1)$ for the rate parameter μ of the exponential distribution f_I. For the stationary LoMTDPP, the shape and scale parameters received gamma priors $\mathrm{Ga}(\alpha \,|\, 5, 1)$ and $\mathrm{Ga}(\phi \,|\, 1, 1)$. Similar to the first simulation experiment, we chose $\mathrm{Ga}(\alpha \,|\, 5, 1)$ with the expectation that the first four moments exist with respect to

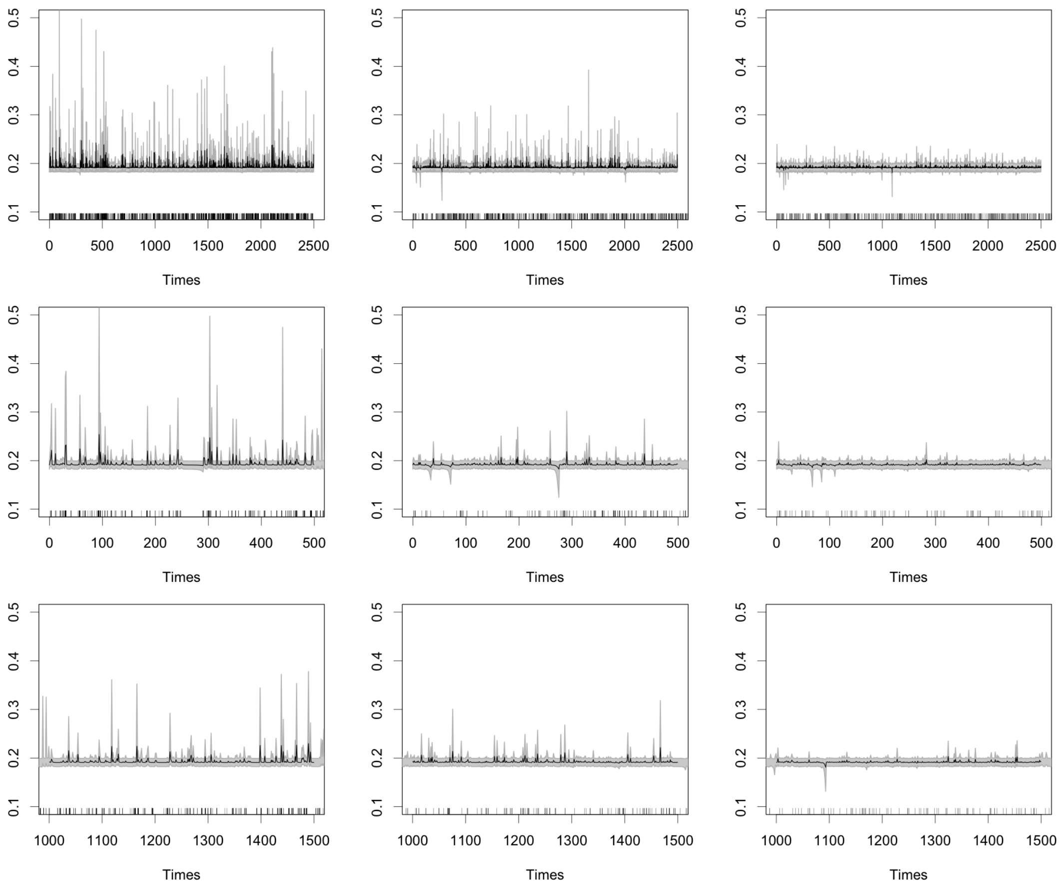
Table 3.1: Chapter 3 - second simulation data analysis. Posterior means and 95% credible interval estimates of the MTDCPP model parameters under different scenarios.

	$\pi_0 = 0.2$	$\pi_0 = 0.5$	$\pi_0 = 0.8$	$\pi_0 = 1$
π_0	0.22 (0.19, 0.24)	0.52 (0.48, 0.56)	0.81 (0.75, 0.85)	0.99 (0.96, 1.00)
μ	0.22 (0.19, 0.24)	0.19 (0.17, 0.20)	0.20 (0.19, 0.21)	0.19 (0.18, 0.20)
ϕ	0.12 (0.09, 0.15)	0.13 (0.09, 0.19)	0.10 (0.02, 0.26)	1.28 (0.05, 4.38)
α	5.34 (4.53, 6.28)	6.25 (4.86, 7.87)	4.46 (2.91, 7.48)	4.53 (1.76, 9.42)

the component Lomax distribution. The vector $\boldsymbol{w}$ was assigned $\mathrm{CDP}(\boldsymbol{w} \mid 5, 1, 3)$, which elicits a decreasing pattern in the weights.

We focus on the inference on the two-component mixture probability π_0 and the component density parameters (μ, ϕ, α). The posterior means and 95% credible interval estimates of the parameters are presented in Table 3.1. The posterior estimates of the parameter π_0 suggest that the model was able to recover the proportion of the point process driven by f_I, even in the extreme case when $\pi_0 = 1$. For other parameters, the model produced estimates close to the true values for all scenarios.

Figure 3.2 illustrates the conditional intensities of the Lomax MTDPP evaluated at three different time windows, under the scenarios of $\pi_0 = 0.2, 0.5, 1$. We omit the scenario with $\pi_0 = 0.8$ for better visualization. When π_0 becomes larger, the process dynamic is driven more by the exponential component. Thus, accordingly, in the figure, we can observe larger gaps between clusters, with less and shorter spikes. When $\pi_0 = 1$, the estimated intensity is overall very close to the true value $\mu = 0.2$.



(a) $\pi_0 = 0.2$ **(b)** $\pi_0 = 0.5$ **(c)** $\pi_0 = 1$

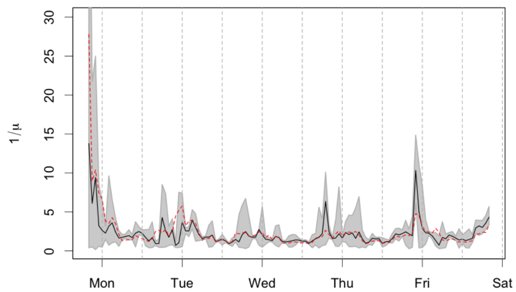
Figure 3.2: Chapter 3 - second simulation data analysis. Posterior means (solid lines) and 95% credible interval estimates (grey polygons) of the conditional intensity of the MTDCPP model evaluated at different time windows under different scenarios.

3.4.3 Mid-Price Changes of the AUD/USD Exchange Rate

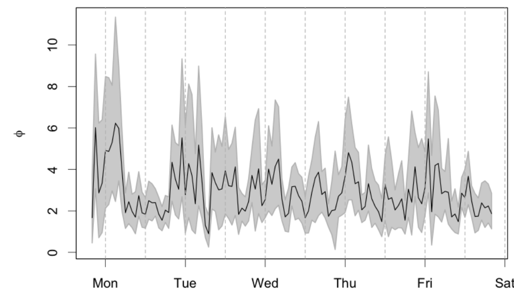
Financial markets involve complex human activities, with both external and internal factors driving market dynamics. Contrary to the "efficient market hypothesis" that expects large price changes occurring with significant news, it is suggested that, for high-frequency financial data, only a small portion of the price movements is caused by external factors such as relevant news releases (Filimonov and Sornette, 2012). Therefore, to better understand the financial market microstructure, is is important to quantify the degree of market reflexivity, measured as the proportion of price movements due to internal processes rather than external processes. Recently, the Hawkes process and its extensions (Filimonov and Sornette, 2012; Wheatley et al., 2016; Chen and Stindl, 2018; Stindl and Chen, 2021) have been introduced to study the market reflexivity, where each price move is considered as an event. The Hawkes process admits a branching structure that allows for separating endogenous events from exogenous events. The model provides a parameter called branching ratio that can be used to quantify the market reflexivity. The goal of this section is to investigate the market reflexivity from the duration clustering perspective using the MTDCPP, where the probability component $(1 - \pi_0)$ in (3.9) can be interpreted as the proportion of price movements due to endogenous interaction. The ability of the MTDCPP to quantify the market reflexivity for this particular data exmaple is also compared with the Hawkes process at the end of the section.

We analyze the price movements of the AUD/USD foreign exchange rate. In particular, a price movement is recorded when there is a mid-price change, where mid-price is defined as the average of the best bid and ask prices. A detailed explanation of using mid-price change as a measure of the price movements can be found in Filimonov and Sornette (2012). The data consist of 121 non-overlapping point patterns, with total number of events ranging from 108 to 3961. Each point pattern corresponds to one hour time window of the trading week from 20:00:00 Greenwich Mean Time (GMT) July 19 to 21:00:00 GMT July 24 in the year of 2015. Analyzing sequences of point patterns within small windows, to some extent, avoids the issue of nonstationarity such as diurnal pattern. We refer to Chen and Stindl (2018) for more details of the data, and the data are available in the R package *RHawkes* (Chen and Stindl, 2022).

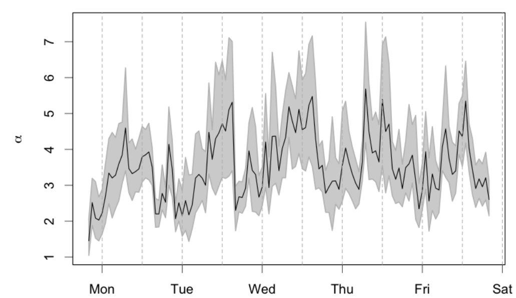
We considered the Lomax MTDCPP model illustrated in the second simulation



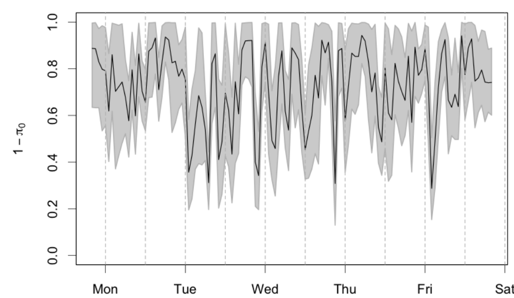
(a) Exponential distribution parameter $1/\mu$



(b) LoMTDPP scale parameter ϕ



(c) LoMTDPP shape parameter α



(d) Endogenous probability $(1 - \pi_0)$

Figure 3.3: Chapter 3 - AUD/USD foreign exchange market data analysis. Time series plots of the posterior means (solid lines) and pointwise 95% credible intervals (grey polygons) of the parameters $1/\mu$, ϕ, $\alpha, (1-\pi_0)$ for the MTDCPP based on the hourly data. Vertical dashed lines correspond to midnight and midday GMT. The red dashed line in Panel (a) corresponds to the averages of the observed durations of the one hour windows.

experiment, and applied the model to the 121 point patterns, with the same prior specification for $(\pi_0, \mu, \alpha, \phi)$ as in the second simulation experiment. Based on the autocorrelation and partial autocorrelation functions of the observed durations, we chose model order $L = 15$ for all point patterns, and the mixture weights were assigned a $\text{CDP}(\boldsymbol{w} \,|\, 5, 1, 6)$.

For each point pattern, we obtained posterior mean and 95% credible interval estimates of the model parameters. Figure 3.3 shows the time series of the point and interval estimates of four parameters: exponential distribution parameter

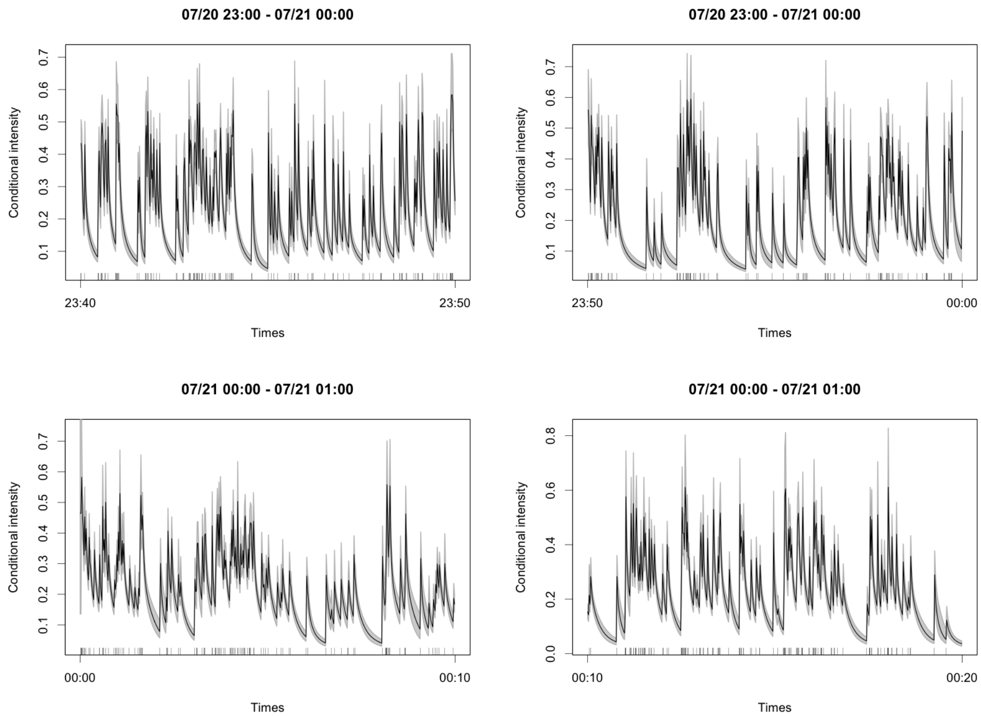


Figure 3.4: Chapter 3 - AUD/USD foreign exchange market data analysis. Posterior means (solid lines) and pointwise 95% credible intervals (grey polygons) of the MTDCPP conditional intensity evaluated at the time windows around midnight of Tuesday, July 21.

$1/\mu$, LoMTDPP scale parameter ϕ, LoMTDPP shape parameter α, the endogenous probability $(1 - \pi_0)$. Specifically, parameters $1/\mu$ and ϕ indicate the mean waiting time between two mid-price changes due to external and internal factors, respectively. The estimates of the mean waiting time $1/\mu$ shows obvious diurnal pattern, with peaks and troughs appearing around midnight and midday, respectively. The posterior estimates of ϕ for all point patterns seem more volatile. As shown in Figure 3.3(c), the posterior estimates of α reflect a pattern that is consistent with the exponential distribution waiting time estimates. Recall that a small value of α indicates a heavy tail. Smaller value of α estimates around midnight suggest larger expected waiting time between events. Figure 3.4 shows the estimated conditional intensity at time windows around the midnight of Tuesday, July 21. We can observe clear clustering pattern especially during the last ten minutes before the midnight.

From Figure 3.3(d), we see that the estimates of the market reflexivity $(1-\pi_0)$ fluctuate heavily over the whole trading week, with most of them greater than 0.5. The posterior means of $(1-\pi_0)$ for the 121 point patterns range from 0.29 to 0.94 with median 0.75 and quartiles $(0.62, 0.86)$, suggesting that the market dynamics are mostly driven by internal processes. A similar conclusion was drawn by Chen and Stindl (2018) in which a renewal Hawkes (RHawkes) process was used to fit the same data set to quantify the market reflexivity. The RHawkes process (Wheatley et al., 2016) extends the Hawkes process to capture dependence between clusters, by replacing the immigrant Poisson process with a renewal process. Despite this feature, it remains the same to use the branching ratio to estimate the market reflexivity as the Hawkes process. The estimated ratios in Chen and Stindl (2018) range from 0.29 to 0.98 with median 0.66 and quartiles $(0.53, 0.80)$. Their results are similar to our findings. This suggests that the Lomax MTDCPP was able to quantify how much the observed dynamics are caused by internal factors versus external influences.

Finally, we would like to remark that, in addition to the computational advantage, using the MTDCPP for the present objective does not require any stationarity assumptions. On the contrary, stationarity is essential for both the Hawkes and RHawkes processes in order to use the branching ratio as an estimator for the market reflexivity. However, as discussed in Filimonov and Sornette (2012), market activities are typically nonstationary. Although seasonality can be addressed by splitting the time window into small intervals, one has to balance the size of the intervals and the number of the events within the interval to ensure reliable estimates produced from the point process models. Moreover, stationarity is not guaranteed for each small interval.

3.5 Discussion

We have developed a modeling framework for constructing various types of point processes, including self-exciting and self-regulating processes, dependent renewal processes, and cluster point processes. The framework relies on specifying an MTD model for the conditional duration density of a point process. Thus, the resulting point process has restricted memory, i.e., the process evolution depends on some recent events. The high-order Markov assumption facilitates efficient computation, but it could be a limitation in practice for some scenarios when the assumption of full history dependence is needed. One solution to ease this issue, within our framework, is to place a large value for the MTD order. The MTD mixture model structure with the structured prior for the weights allows efficient inference with a large order.

Section 3.2.4 illustrates the use of a bivariate copula for the joint cdf of the random vector (U_l, V_l). It is worth mentioning that, alternatively, we can specify the joint survival function of (U_l, V_l) with a copula C_l such that the joint survival probability $S_{U_l,V_l}(u, v) = C_l(S_{U_l}(u), S_{V_l}(v))$ (Georges et al., 2001). It follows that $f_{U_l,V_l}(u, v) = \tilde{c}_l(u, v)f_{U_l}(u)f_{V_l}(v)$ where $\tilde{c}_l(u, v) = \partial C(S_{U_l}(u), S_{V_l}(v))/(\partial S_{U_l}\partial S_{V_l})$. The conditional duration density of the resulting MTDPP can be obtained by replacing the c_l in (3.7) with $\tilde{c}_l$. The component hazard function of the mdoel is $h_l(u \,|\, v) = -d\log S_{U_l|V_l}(u \,|\, v)/du$, where $S_{U_l|V_l}(u \,|\, v) = \partial C(S_{U_l}(u), S_{V_l}(v))/\partial S_{V_l}$. This approach produces a class of copula NNMPs different from those introduced in Section 3.2.4, and can be useful in the topic of survival analysis, e.g., for modeling recurrent event processes.

In many applications, a point process carries information about times of some events that are of particular interest. These events are referred to as marks of the point process, and one is usually interested in the inference for the marks such as their conditional distributions. Without loss of generality, let $\boldsymbol{y}$ be a vector of continuous marks. The proposed framework can be easily extended to incorporate marks through a convenient decomposition of a marked point process intensity (Daley and Vere-Jones, 2003), that is, $\lambda^*(t, \boldsymbol{y}) = \lambda_g^*(t)m^*(\boldsymbol{y} \mid t)$, where $\lambda_g^*(t)$ is the conditional intensity of the point process for the event times, which is referred to as the ground process. Assuming the ground process does not depend on marks, we can mode the ground process with an MTDPP or MTDCPP. Then, modeling a marked point process boils down to specification of the conditional mark distribution $m^*(\boldsymbol{y} \mid t)$; see, for example, in the context of peaks-over-threshold analysis, Herrera and Schipp (2013) use a generalized Pareto distribution for $m^*(\boldsymbol{y} \mid t)$ with an ACD model for the ground process.

Chapter 4

Models for Non-Gaussian Continuous-Valued Spatial Processes

4.1 Introduction

Gaussian processes have been widely used as an underlying structure in model-based analysis of irregularly located spatial data in order to capture short range variability. The fruitfulness of these spatial models owes to the simple characterization of the Gaussian process by a mean and a covariance function, and the optimal prediction it provides that justifies kriging. However, the assumption of Gaussianity is restrictive in many fields where the data exhibits non-Gaussian features, for example, vegetation abundance (Eskelson et al., 2011), precipitation data (Sun et al., 2015), contaminated soil (Paul and Cressie, 2011), temperature data (North et al., 2011), and wind speed data (Bevilacqua et al., 2020). In this chapter, we aim at developing a flexible class of geostatistical models that is customizable to general non-Gaussian distributions, with particular focus on continuous data.

Several approaches have been developed for non-Gaussian geostatistical modeling. A straightforward approach consists of fitting a Gaussian process after transformation of the original data. Possible transformations include Box-Cox (De Oliveira et al., 1997), power (Allcroft and Glasbey, 2003), square-root (Johns et al., 2003), and Tukey g-and-h (Xu and Genton, 2017) transforms, to name a few. Alternative to transformations, we can represent a non-Gaussian distribution as a location-scale mixture of Gaussian distributions. This yields Gaussian process extensions that are able to capture skewness and long tails (Kim and Mallick, 2004; Palacios and Steel, 2006; Zhang and El-Shaarawi, 2010; Mahmoudian, 2017; Morris et al., 2017; Zareifard et al., 2018; Bevilacqua et al., 2021). Beyond methods based on continuous mixtures of Gaussian distributions, Bayesian nonparametric methods have been explored for geostatistical data modeling, starting with the approach in Gelfand et al. (2005) which extends the Dirichlet process (Ferguson, 1973) to a prior model for random spatial surfaces. We refer to Müller et al. (2018) for a review. From a different perspective, a class of non-Gaussian Matérn fields is formulated with stochastic partial differential equations driven by non-Gaussian noise (Bolin, 2014; Wallin and Bolin, 2015; Bolin and Wallin, 2020).

An alternative popular approach involves a hierarchical model structure that assumes conditionally independent non-Gaussian marginals, combined with a latent spatial process that is associated with some functional or link function of the first-stage marginals. Hereafter, we refer to these models as hierarchical first-stage non-Gaussian models. If the latent process is linked through a function of some parameter(s) of the first-stage marginal which belongs to the exponential dispersion family, the approach is known as the spatial generalized linear mixed model and its extensions (Diggle et al., 1998; Chan and Dong, 2011). Non-Gaussian spatial models that build from copulas (Joe, 2014) can also be classified into this category. Copula models assume pre-specified families of marginals for observations, with a multivariate distribution underlying the copula for a vector of latent variables that are probability integral transformations of the observations (Danaher and Smith, 2011). Spatial copula models replace the multivariate distribution with one that corresponds to a spatial process, thus introducing spatial dependence (Bárdossy, 2006; Ghosh and Mallick, 2011; Krupskii et al., 2018; Beck et al., 2020).

The non-Gaussian modeling framework proposed in this chapter is distinctly different from the previously mentioned approaches. Our methodology builds on the class of nearest-neighbor processes obtained by extending a joint density for a reference set of locations to the entire spatial domain. The original joint density is factorized into a product of conditionals with respect to a DAG. Deriving each conditional from a Gaussian process results in the nearest-neighbor Gaussian process (NNGP; Datta et al. 2016a). Models defined over DAGs have received substantial attention; see, e.g., Datta et al. (2016b); Finley et al. (2019); Peruzzi et al. (2020); Peruzzi and Dunson (2022). The class of DAG-based models originates from Vecchia's approach (Vecchia, 1988) that considers nearest-neighbor approximations. Related works that exploited sparsity for approximating an expensive Gaussian likelihood include, e.g., Stein et al. (2004), Gramacy and Apley (2015), Sun and Stein (2016), Stroud et al. (2017), Guinness (2018), and Schäfer et al. (2021). Katzfuss and Guinness (2021) provide a further generalization of the Vecchia approximation framework.

Considerably less attention, however, has been devoted to defining models over a DAG with non-Gaussian distributions for the conditionals of the joint density. This is in general a difficult problem, as each conditional involves, say, a

p-dimensional conditioning set, which requires a coherent model for a $(p+1)$-dimensional non-Gaussian distribution, with p potentially large. In this chapter, we take on the challenging task of developing a computationally efficient, interpretable framework that provides generality for modeling different types of non-Gaussian data and flexibility for complex spatial dependence.

We overcome the aforementioned challenge by modeling each conditional of the joint density as a weighted combination of first-order spatially varying transition kernels, each of which depends on a specific neighbor. This approach produces multivariate non-Gaussian distributions by specification of the bivariate distributions that define the local transition kernels. Thus, it provides generality for modeling different non-Gaussian behaviors, since, relative to the multivariate analogue, constructing bivariate distributions is substantially easier, for instance, using bivariate copulas. Moreover, such a model structure offers the convenience of quantifying multivariate dependence through the collection of bivariate distributions. As an illustration, we study tail dependence properties under appropriate families of bivariate distributions, and provide results that guide modeling choices. The modeling framework achieves flexibility by letting both the weights and transition kernels be spatially dependent, inducing sufficient local dependence to describe a wide range of spatial variability. We refer to the resulting geospatial process as the nearest-neighbor mixture process (NNMP).

An important feature of the model structure is that it facilitates the study of conditions for constructing NNMPs with pre-specified families of marginal distributions. Such conditions are easily implemented without parameter constraints, thus resulting in a general modeling tool to describe spatial data distributions that are skewed, heavy-tailed, positive-valued, or have bounded support, as illustrated through several examples in Section 4.4. The NNMP framework emphasizes direct modeling by introducing spatial dependence at the data level. It avoids the use of transformations that may distort the Gaussian process properties (Wallin and Bolin, 2015). It is fundamentally different from the class of hierarchical first-stage non-Gaussian models that introduce spatial dependence through functionals of the data probability distribution, such as the transformed mean. Regarding computation, NNMP models do not require estimation of potentially large vectors of spatially correlated latent variables, something unavoidable with hierarchical first-stage non-Gaussian models. Approaches for such models typically resort to approximate inference, either directly or combined with a scalable model (Zilber and Katzfuss, 2021). Estimation of NNMPs is analogous to that of a finite mixture model, thus avoiding the need to perform costly matrix operations for large data sets, and allowing for computationally efficient, full simulation-based inference. Overall, the NNMP framework offers a flexible class of models that is able to describe complex spatial dependence, coupled with an efficient computational approach, leveraged from the mixture structure of the model.

The rest of the chapter is organized as follows. In Section 4.2, we formulate the NNMP framework and study model properties. Specific examples of NNMP models illustrate different components of the methodology. Section 4.3 develops the general approach to Bayesian estimation and prediction under NNMP models. In Section 4.4, we demonstrate different NNMP models with synthetic data examples and with an analysis of Mediterranean Sea surface temperature data. Finally, Section 4.5 concludes with a summary and discussion of future work.

4.2 Nearest-Neighbor Mixture Processes

4.2.1 Modeling Framework

Consider a univariate spatial process $\{Z(\boldsymbol{v}) : \boldsymbol{v} \in \mathcal{D}\}$, where $\mathcal{D} \subset \mathbb{R}^p$, for $p \geq 1$. Let $\mathcal{S} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$ be a finite collection of locations in $\mathcal{D}$, referred to as the reference set. Let $\boldsymbol{z}_{\mathcal{S}} = (z(\boldsymbol{s}_1), \ldots, z(\boldsymbol{s}_n))$ be a realization of the random vector $\boldsymbol{Z}_{\mathcal{S}} = (Z(\boldsymbol{s}_1), \ldots, Z(\boldsymbol{s}_n))$. If we regard the locations $\boldsymbol{s}_i$ as vertices of a DAG, we can factorize the joint density $p(\boldsymbol{z}_{\mathcal{S}})$ of $\boldsymbol{Z}_{\mathcal{S}}$ into a product of univariate conditionals as

$$p(\boldsymbol{z}_{\mathcal{S}}) = p(z(\boldsymbol{s}_1)) \prod_{i=2}^{n} p(z(\boldsymbol{s}_i) \,|\, \boldsymbol{z}_{\mathrm{Ne}(\boldsymbol{s}_i)}), \tag{4.1}$$

where the set $\mathrm{Ne}(\boldsymbol{s}_i) \subset \mathcal{S}_i = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_{i-1}\}$ consists of parents of $\boldsymbol{s}_i$. The joint density in (4.1) corresponds to a directed graphical model (Jordan 2004; also known as a Bayesian network), with a DAG that summaries the conditional independence structure between random variables. In particular, conditional on $\boldsymbol{Z}_{\mathrm{Ne}(\boldsymbol{s}_i)}$, $Z(\boldsymbol{s}_i)$ is independent of $Z_{\mathcal{S}_i \setminus \mathrm{Ne}(\boldsymbol{s}_i)}$, for $i = 2, \ldots, n$. Thus, choosing the set $\mathrm{Ne}(\boldsymbol{s}_i)$ creates different DAGs. Our selection is based on the geostatistical distance between $\boldsymbol{s}_i$ and $\boldsymbol{s}_j \in \mathcal{S}_i$, and $\mathrm{Ne}(\boldsymbol{s}_i)$ is referred to as the nearest-neighbor set of $\boldsymbol{s}_i$, having at most L elements with $L \ll n$. The selected locations $\boldsymbol{s}_j$ are placed in ascending order according to the distance, denoted as $\boldsymbol{s}_{(i1)}, \ldots, \boldsymbol{s}_{(i,i_L)}$, where $i_L := (i-1) \wedge L$. We note that the development of the proposed framework holds true for any choice of the neighbor sets. For different ways to choose nearest neighbors in spatial modeling, see, for example, Vecchia (1988), Stein et al. (2004), and Gramacy and Apley (2015).

Constructing a nearest-neighbor process involves specification of the conditional density $p(z(\boldsymbol{s}_i) \,|\, \boldsymbol{z}_{\mathrm{Ne}(\boldsymbol{s}_i)})$ in (4.1), and extension to an arbitrary finite set in $\mathcal{D}$ that is not overlapping with $\mathcal{S}$. We approach the problem of constructing a nearest-neighbor non-Gaussian process following this idea. Specifically, we define the conditional density as

$$p(z(\boldsymbol{s}_i) \,|\, \boldsymbol{z}_{\mathrm{Ne}(\boldsymbol{s}_i)}) = \sum_{l=1}^{i_L} w_l(\boldsymbol{s}_i)\, f_{\boldsymbol{s}_i,l}(z(\boldsymbol{s}_i) \,|\, z(\boldsymbol{s}_{(il)})), \tag{4.2}$$

where $f_{\boldsymbol{s}_i,l}$ is the lth component of the mixture density p, and the weights satisfy $w_l(\boldsymbol{s}_i) \geq 0$, for all l, and $\sum_{l=1}^{i_L} w_l(\boldsymbol{s}_i) = 1$, for every $\boldsymbol{s}_i \in \mathcal{S}$. In a DAG, nearest neighbors in set $\mathrm{Ne}(\boldsymbol{s}_i)$ are vertices that have directed edges pointing to $\boldsymbol{s}_i$. Thus, it is appealing to consider a high-order Markov model in which temporal lags have a similar notion of direction. Our approach to formulating (4.2) is motivated by a class of mixture transition distribution models (Le et al., 1996), which consists of a mixture of first-order transition densities with a vector of common weights. A key feature of the formulation in (4.2) is the decomposition of a non-Gaussian conditional density, with a potentially large conditioning set, into a weighted sum of local conditional densities. This provides flexible, parsimonious modeling of $p(z(\boldsymbol{s}_i) \,|\, \boldsymbol{z}_{\mathrm{Ne}(\boldsymbol{s}_i)})$ through specifying bivariate distributions that define the local conditionals $f_{\boldsymbol{s}_i,l}(z(\boldsymbol{s}_i) \,|\, z(\boldsymbol{s}_{(il)})$. We provide further discussion on this feature for model construction and relevant properties in the following sections.

Spatial dependence characterized by (4.2) is twofold. First, each component $f_{\boldsymbol{s}_i,l}$ is associated with spatially varying parameters indexed at $\boldsymbol{s}_i \in \mathcal{S}$, defined by a probability model or a link function. Secondly, the weights $w_l(\boldsymbol{s}_i)$ are spatially varying. As each component density $f_{\boldsymbol{s}_i,l}$ depends on a specific neighbor, the weights indicate the contribution of each neighbor of $\boldsymbol{s}_i$. Besides, the weights adapt to the change of locations. For two different $\boldsymbol{s}_i, \boldsymbol{s}_j$ in $\mathcal{S}$, the relative locations of the nearest neighbors $\mathrm{Ne}(\boldsymbol{s}_i)$ to $\boldsymbol{s}_i$ are different from that of $\mathrm{Ne}(\boldsymbol{s}_j)$ to $\boldsymbol{s}_j$. If all elements of $\mathrm{Ne}(\boldsymbol{s}_i)$ are very close to $\boldsymbol{s}_i$, then values of $(w_1(\boldsymbol{s}_i), \ldots, w_{i_L}(\boldsymbol{s}_i))^\top$ should be quite even. On the other hand, if, for $\boldsymbol{s}_j$, only a subset of its neighbors in

$\mathrm{Ne}(\boldsymbol{s}_j)$ are close to $\boldsymbol{s}_j$, then the weights corresponding to this subset should receive larger values. We remark that in general, probability models or link functions for the spatially varying parameters should be considered case by case, given different specifications on the components $f_{\boldsymbol{s}_i,l}$. Details of the construction for the component densities and the weights are deferred to later sections.

We obtain the NNMP, a legitimate spatial process, by extending (4.2) to an arbitrary set of non-reference locations $\mathcal{U} = \{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r\}$ where $\mathcal{U} \subset \mathcal{D} \setminus \mathcal{S}$. In particular, we define the conditional density of $\boldsymbol{z}_{\mathcal{U}}$ given $\boldsymbol{z}_{\mathcal{S}}$ as

$$p(\boldsymbol{z}_{\mathcal{U}} \mid \boldsymbol{z}_{\mathcal{S}}) = \prod_{i=1}^{r} p(z(\boldsymbol{u}_i) \mid \boldsymbol{z}_{\mathrm{Ne}(\boldsymbol{u}_i)}) = \prod_{i=1}^{r} \sum_{l=1}^{L} w_l(\boldsymbol{u}_i)\, f_{\boldsymbol{u}_i,l}(z(\boldsymbol{u}_i) \mid z(\boldsymbol{u}_{(il)})), \qquad (4.3)$$

where the specification on $w_l(\boldsymbol{u}_i)$ and $f_{\boldsymbol{u}_i,l}$ for all i and all l is analogous to that for (4.2), except that $\mathrm{Ne}(\boldsymbol{u}_i) = \{\boldsymbol{u}_{(i1)}, \ldots, \boldsymbol{u}_{(iL)}\}$ are the first L locations in $\mathcal{S}$ that are closest to $\boldsymbol{u}_i$ in terms of geostatistical distance. Building the construction of the neighbor sets $\mathrm{Ne}(\boldsymbol{u}_i)$ on the reference set ensures that $p(\boldsymbol{z}_{\mathcal{U}} \mid \boldsymbol{z}_{\mathcal{S}})$ is a proper density.

Given (4.2) and (4.3), we can obtain the joint density $p(\boldsymbol{z}_{\mathcal{V}})$ of a realization $\boldsymbol{z}_{\mathcal{V}}$ over any finite set of locations $\mathcal{V} \subset \mathcal{D}$. When $\mathcal{V} \subset \mathcal{S}$, the joint density $p(\boldsymbol{z}_{\mathcal{V}})$ is directly available as the appropriate marginal of $p(\boldsymbol{z}_{\mathcal{S}})$. Otherwise, we have that $p(\boldsymbol{z}_{\mathcal{V}}) = \int p(\boldsymbol{z}_{\mathcal{U}} \mid \boldsymbol{z}_{\mathcal{S}}) p(\boldsymbol{z}_{\mathcal{S}}) \prod_{\{\boldsymbol{s}_i \in \mathcal{S} \setminus \mathcal{V}\}} dz(\boldsymbol{s}_i)$, where $\mathcal{U} = \mathcal{V} \setminus \mathcal{S}$. If $\mathcal{S} \setminus \mathcal{V}$ is empty, $p(\boldsymbol{z}_{\mathcal{V}})$ is simply $p(\boldsymbol{z}_{\mathcal{U}} \mid \boldsymbol{z}_{\mathcal{S}}) p(\boldsymbol{z}_{\mathcal{S}})$. In general, the joint density $p(\boldsymbol{z}_{\mathcal{V}})$ of an NNMP is intractable. However, since both $p(\boldsymbol{z}_{\mathcal{U}} \mid \boldsymbol{z}_{\mathcal{S}})$ and $p(\boldsymbol{z}_{\mathcal{S}})$ are products of mixtures, we can recognize that $p(\boldsymbol{z}_{\mathcal{V}})$ is a finite mixture, which suggests flexibility of the model to capture complex non-Gaussian dependence over the domain $\mathcal{D}$. Moreover, we show in Section 4.2.3 that for some NNMPs, the joint density $p(\boldsymbol{z}_{\mathcal{V}})$ has a closed-form expression. In the subsequent development of model properties, we will use the conditional density

$$p(z(\boldsymbol{v}) \mid \boldsymbol{z}_{\mathrm{Ne}(\boldsymbol{v})}) = \sum_{l=1}^{L} w_l(\boldsymbol{v}) \, f_{\boldsymbol{v},l}(z(\boldsymbol{v}) \mid z(\boldsymbol{v}_{(l)})), \quad \boldsymbol{v} \in \mathcal{D}, \tag{4.4}$$

to characterize an NNMP, where $\mathrm{Ne}(\boldsymbol{v})$ contains the first L locations that are closest to $\boldsymbol{v}$, selected from locations in $\mathcal{S}$. These locations in $\mathrm{Ne}(\boldsymbol{v})$ are placed in ascending order according to distance, denoted as $\boldsymbol{v}_{(1)}, \ldots, \boldsymbol{v}_{(L)}$.

In comparison to the nearest-neighbor spatial models discussed in Section 4.1, we remark on a conceptual difference between them and our modeling framework. Unlike the nearest-neighbor Gaussian process approach in Datta et al. (2016a), we do not posit a parent process when building our models. Datta et al. (2016a) assume a full Gaussian process over the reference set, and use it to derive the conditional densities $p(z(\boldsymbol{s}_i) \mid \boldsymbol{z}_{\mathrm{Ne}(\boldsymbol{s}_i)})$. Similar ideas underlie the Vecchia approximation framework which considers (4.1) as an approximation to the density of a full Gaussian process realization. In the present work, we directly model the joint density $p(\boldsymbol{z}_{\mathcal{S}})$, utilizing the nearest-neighbor DAG representation with a structured mixture model, which is key to our modeling objective of developing spatial processes for general non-Gaussian data.

Before closing this section, we note that spatial locations are not naturally ordered. Given a distance function, a different ordering on the locations results in different neighbor sets. Therefore, a different DAG with density $p(\boldsymbol{z}_{\mathcal{S}})$ is created accordingly for model inference. Literature that considers nearest-neighbor models for Gaussian data by default orders locations based on sorting coordinates. We refer to Guinness (2018) and references therein for more details. In particular, Guinness (2018) demonstrates that certain orderings such as random orderings can improve model performance when compared with coordinate-based orderings. For the NNMP models illustrated in the data examples, we found through simulation experiments that there were no discernible differences between the inferences based on $p(\boldsymbol{z}_{\mathcal{S}})$, given two different random orderings. As outlined by Datta et al. (2016a), the effectiveness of $p(\boldsymbol{z}_{\mathcal{S}})$ depends on the information borrowed from the neighbors, which is often determined by the size of $\mathrm{Ne}(\boldsymbol{s}_i)$ rather than the ordering. A further remark is that the ordering of the reference set $\mathcal{S}$ is typically reserved for observed data. Thus, the ordering effect lies only in the model estimation based on (4.2) with realization $\boldsymbol{z}_{\mathcal{S}}$. Spatial prediction typically rests on locations outside $\mathcal{S}$ using (4.3), where the ordering effect disappears.

4.2.2 NNMPs with Stationary Marginal Distributions

We develop a sufficient condition to construct NNMPs with general stationary marginal distributions. The key feature of this result is that the condition relies on the bivariate distributions that define the first-order transition kernels in (4.4) without the need to impose restrictions on the parameter space.

Proposition 4.1. *Consider an NNMP for which the component density $f_{\boldsymbol{v},l}$ is specified by the conditional density of $U_{\boldsymbol{v},l}$ given $V_{\boldsymbol{v},l}$, where the random vector $(U_{\boldsymbol{v},l}, V_{\boldsymbol{v},l})$ follows a bivariate distribution with marginal densities $f_{U_{\boldsymbol{v},l}}$ and $f_{V_{\boldsymbol{v},l}}$, for $l = 1, \ldots, L$. The NNMP has stationary marginal density f_Z if it satisfies the invariant condition: $Z(\boldsymbol{s}_1) \sim f_Z$, $\boldsymbol{s}_1 \in \mathcal{S}$, and for every $\boldsymbol{v} \in \mathcal{D}$, $f_Z(z) = f_{U_{\boldsymbol{v},l}}(z) = f_{V_{\boldsymbol{v},l}}(z)$, for all z and for all l.*

This result builds from the one in Chapter 2 where temporal MTD processes with stationary marginal distributions were constructed. It applies regardless of $Z(\boldsymbol{v})$ being a continuous, discrete or mixed random variable, thus allowing for a wide range of non-Gaussian marginal distributions and a general functional form, either linear or non-linear, for the expectation with respect to the conditional density p in (4.4).

As previously discussed, the mixture model formulation for the conditional density in (4.4) induces a finite mixture for the NNMP finite-dimensional distributions. On the other hand, due to the mixture form, an explicit expression for the covariance function is difficult to derive. A recursive equation can be obtained for a class of NNMP models for which the conditional expectation with respect to $(U_{\boldsymbol{v},l}, V_{\boldsymbol{v},l})$ is linear, that is, $E(U_{\boldsymbol{v},l} \,|\, V_{\boldsymbol{v},l} = z) = a_l(\boldsymbol{v}) + b_l(\boldsymbol{v})\, z$ for some $a_l(\boldsymbol{v}), b_l(\boldsymbol{v}) \in \mathbb{R}$, $l = 1, \ldots, L$, and for all $\boldsymbol{v} \in \mathcal{D}$. Suppose the NNMP has a stationary marginal distribution with finite first and second moments. Without loss of generality, we assume the first moment is zero. Then the covariance over any two locations $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathcal{D}$ is

$$
\begin{aligned}
&\mathrm{Cov}(Z(\boldsymbol{v}_1), Z(\boldsymbol{v}_2)) \\
&= \begin{cases}
\sum_{l=1}^{L} w_l(\boldsymbol{s}_i)\, b_l(\boldsymbol{s}_i)\, E(Z(\boldsymbol{s}_j)Z(\boldsymbol{s}_{(il)})), & \boldsymbol{v}_1 \equiv \boldsymbol{s}_i \in \mathcal{S}, \boldsymbol{v}_2 \equiv \boldsymbol{s}_j \in \mathcal{S}, \\
\sum_{l=1}^{L} w_l(\boldsymbol{v}_1)\, b_l(\boldsymbol{v}_1)\, E(Z(\boldsymbol{s}_j)Z(\boldsymbol{v}_{(1l)})), & \boldsymbol{v}_1 \notin \mathcal{S}, \boldsymbol{v}_2 \equiv \boldsymbol{s}_j \in \mathcal{S}, \\
\sum_{l=1}^{L} \sum_{l'=1}^{L} w_{ll'} \left\{ a_{ll'} + b_{ll'} E(Z(\boldsymbol{v}_{(1l)})Z(\boldsymbol{v}_{(2l')})) \right\}, & \boldsymbol{v}_1, \boldsymbol{v}_2 \notin \mathcal{S},
\end{cases}
\end{aligned}
\tag{4.5}
$$

where $w_{ll'} \equiv w_l(\boldsymbol{v}_1) w_{l'}(\boldsymbol{v}_2)$, $a_{ll'} \equiv a_l(\boldsymbol{v}_1) a_{l'}(\boldsymbol{v}_2)$, $b_{ll'} \equiv b_l(\boldsymbol{v}_1) b_{l'}(\boldsymbol{v}_2)$, and without loss of generality, we assume $i > j$. The covariance in (4.5) implies that, even though the process has a stationary marginal distribution, the NNMP is second-order non-stationary.

4.2.3 Construction of NNMP models

The spatially varying conditional densities $f_{\boldsymbol{v},l}$ in (4.4) correspond to a sequence of bivariate distributions indexed at $\boldsymbol{v}$, namely, the distributions of $(U_{\boldsymbol{v},l}, V_{\boldsymbol{v},l})$, for $l = 1, ..., L$. To balance model flexibility and scalability, we build spatially varying distributions by considering the distribution of random vector (U_l, V_l), for

$l = 1, \ldots, L$, and extending some of its parameters to be spatially varying, that is, indexed in $\boldsymbol{v}$. To this end, we use a probability model or a link function. We refer to the random vectors (U_l, V_l) as the set of base random vectors. With a careful choice of the model/function for the spatially varying parameter(s), this construction method reduces significantly the dimension of the parameter space, while preserving the capability of the NNMP model structure to capture spatial dependence.

We illustrate the method with several examples below, starting with a bivariate Gaussian distribution and its continuous mixtures for real-valued data, followed by a general strategy using bivariate copulas that can model data with general support. Before proceeding to the examples, we emphasize that our method allows for general bivariate distributions. One can also consider using a pair of compatible conditionals to specify bivariate distributions (Arnold et al., 1999), for instance, a pair of Lomax conditionals. This is illustrated in Example 4 in Section 4.2.4.

Example 1. *Gaussian and continuous mixture of Gaussian NNMP models.*

For $l = 1, \ldots, L$, take (U_l, V_l) to be a bivariate Gaussian random vector with mean $\mu_l \mathbf{1}_2$ and covariance matrix $\Sigma_l = \sigma_l^2 \left(\begin{smallmatrix} 1 & \rho_l \\ \rho_l & 1 \end{smallmatrix}\right)$, where $\mathbf{1}_2$ is the two-dimensional column vector of ones, resulting in a Gaussian conditional density $f_{U_l|V_l}(u_l \mid v_l) = N(u_l \mid (1-\rho_l)\mu_l + \rho_l v_l, \sigma_l^2(1-\rho_l^2))$. If we extend the correlation parameter to be spatially varying, $\rho_l(\boldsymbol{v}) = k_l(\boldsymbol{v}, \boldsymbol{v}_{(l)})$, for a correlation function k_l, we obtain the spatially varying conditional density $p(z(\boldsymbol{v}) \mid \boldsymbol{z}_{\mathrm{Ne}(\boldsymbol{v})})$ of the model expressed as

$$\sum_{l=1}^{L} w_l(\boldsymbol{v})\, N(z(\boldsymbol{v}) \mid (1-\rho_l(\boldsymbol{v}))\mu_l + \rho_l(\boldsymbol{v}) z(\boldsymbol{v}_{(l)}), \sigma_l^2(1-(\rho_l(\boldsymbol{v}))^2)). \tag{4.6}$$

This NNMP is referred to as the Gaussian NNMP (GNNMP). If we take $Z(\boldsymbol{s}_1) \sim N(z \mid \mu, \sigma^2)$, and set $\mu_l = \mu$ and $\sigma_l^2 = \sigma^2$, for all l, the resulting model satisfies the invariant condition of Proposition 4.1 with stationary marginal given by the $N(\mu, \sigma^2)$ distribution. Moreover, the finite-dimensional distribution of the GNNMP is characterized by the following proposition.

Proposition 4.2. *Consider the GNNMP in* (4.6) *with* $\mu_l = \mu$ *and* $\sigma_l^2 = \sigma^2$*, for all* l*. If* $Z(\boldsymbol{s}_1) \sim N(z \,|\, \mu, \sigma^2)$*, the GNNMP has the* $N(\mu, \sigma^2)$ *stationary marginal distribution, and its finite-dimensional distributions are mixtures of multivariate Gaussian distributions.*

We refer to the model in Proposition 4.2 as the stationary GNNMP. Based on the GNNMP, various NNMP models with different families for (U_l, V_l) can be constructed by exploiting location-scale mixtures of Gaussian distributions. We illustrate the approach with the skew-GNNMP model. Denote by $\mathrm{TN}(\mu, \sigma^2; a, b)$ the Gaussian distribution with mean μ and variance σ^2, truncated at the interval (a, b). Building from the GNNMP, we start with a conditional bivariate Gaussian distribution for (U_l, V_l), given $z_0 \sim \mathrm{TN}(0, 1; 0, \infty)$, where μ_l is replaced with $\mu_l + \lambda_l z_0$. Marginalizing out z_0 yields the bivariate skew-Gaussian distribution for (U_l, V_l) (Azzalini, 2013). Extending again ρ_l to $\rho_l(\boldsymbol{v})$, for all l, we can express the conditional density $p(z(\boldsymbol{v}) \,|\, \boldsymbol{z}_{\mathrm{Ne}(\boldsymbol{v})})$ for the skew-GNNMP model as

$$\sum_{l=1}^{L} w_l(\boldsymbol{v}) \int_0^\infty N(z(\boldsymbol{v}) \,|\, \mu_l(\boldsymbol{v}), \sigma_l^2(\boldsymbol{v}))\, \mathrm{TN}(z_0(\boldsymbol{v}) \,|\, \mu_{0l}(\boldsymbol{v}_{(l)}), \sigma_{0l}^2; 0, \infty) dz_0(\boldsymbol{v}), \quad (4.7)$$

where we have: $\mu_l(\boldsymbol{v}) = \{1 - \rho_l(\boldsymbol{v})\}\{\mu_l + \lambda_l z_0(\boldsymbol{v})\} + \rho_l(\boldsymbol{v}) z(\boldsymbol{v}_{(l)})$; $\sigma_l^2(\boldsymbol{v}) = \sigma_l^2\{1 - (\rho_l(\boldsymbol{v}))^2\}$; $\mu_{0l}(\boldsymbol{v}_{(l)}) = \{z(\boldsymbol{v}_{(l)}) - \mu_l\}\lambda_l/(\sigma_l^2 + \lambda_l^2)$; and $\sigma_{0l}^2 = \sigma_l^2/(\sigma_l^2 + \lambda_l^2)$. Setting $\lambda_l = \lambda$, $\mu_l = \mu$, and $\sigma_l^2 = \sigma^2$, for all l, we obtain the stationary skew-GNNMP model, with skew-Gaussian marginal $f_Z(z) = 2\, N(z \,|\, \mu, \lambda^2 + \sigma^2)\, \Phi((z - \mu)\lambda/(\sigma\sqrt{\lambda^2 + \sigma^2}))$, denoted as $\mathrm{SN}(\mu, \lambda^2 + \sigma^2, \lambda/\sigma)$.

The skew-GNNMP model is an example of a location mixture of Gaussian distributions. Scale mixtures can also be considered to obtain such as the Student-t model. In that case, we replace the covariance matrix Σ_l with $c\Sigma_l$, taking c as a random variable with an appropriate inverse-gamma distribution. Important families that admit a location and/or scale mixture of Gaussians representation include the skew-t, Laplace, and asymmetric Laplace distributions. Using a similar approach to the one for the skew-Gaussian NNMP example, we can construct the corresponding NNMP models.

Example 2. *Copula NNMP models.*

A copula function $C : [0,1]^p \rightarrow [0,1]$ is a function such that, for any multivariate distribution function $F(z_1, \ldots, z_p)$, there exists a copula C for which $F(z_1, \ldots, z_p) = C(F_1(z_1), \ldots, F_p(z_p))$, where F_j is the marginal distribution function of Z_j, $j = 1, \ldots, p$ (Sklar, 1959). If Z_j is continuous for all j, C is unique, and the joint density $f(z_1, \ldots, z_p) = c(z_1, \ldots, z_p) \prod_{j=1}^{p} f_j(z_j)$, where the copula density $c = \partial^p C/(\partial F_1 \ldots \partial F_p)$ and f_j is the density of Z_j. A copula enables us to separate the modeling of the marginal distributions from the dependence. Thus, the invariant condition in Proposition 4.1 can be attained by specifying the stationary distribution F_Z as the marginal distribution of (U_l, V_l) for all l. The copula parameter that determines the dependence of (U_l, V_l) can be modeled as spatially varying to create a sequence of spatially dependent bivariate vectors $(U_{\boldsymbol{v},l}, V_{\boldsymbol{v},l})$. Here, we focus on continuous distributions, although this strategy can be applied for any family of distributions for F_Z. We consider bivariate copulas with a single copula parameter, and illustrate next the construction of a copula NNMP given a stationary marginal density f_Z.

For the bivariate distribution of each (U_l, V_l) with marginals f_{U_l} and f_{V_l}, we consider a copula C_l with parameter η_l, for $l = 1, \ldots, L$. We obtain a spatially varying copula $C_{\boldsymbol{v},l}$ for $(U_{\boldsymbol{v},l}, V_{\boldsymbol{v},l})$ by extending η_l to $\eta_l(\boldsymbol{v})$. The joint density of $(U_{\boldsymbol{v},l}, V_{\boldsymbol{v},l})$ is given by $c_{\boldsymbol{v},l}(z(\boldsymbol{v}), z(\boldsymbol{v}_{(l)}))f_{U_{\boldsymbol{v},l}}(z(\boldsymbol{v}))f_{V_{\boldsymbol{v},l}}(z(\boldsymbol{v}_{(l)}))$, where $c_{\boldsymbol{v},l}$ is the copula density of $C_{\boldsymbol{v},l}$, and $f_{U_{\boldsymbol{v},l}} = f_{U_l}$ and $f_{V_{\boldsymbol{v},l}} = f_{V_l}$ are the marginal densities of $U_{\boldsymbol{v},l}$ and $V_{\boldsymbol{v},l}$, respectively. Given a pre-specified stationary marginal f_Z, we replace both $f_{U_{\boldsymbol{v},l}}$ and $f_{V_{\boldsymbol{v},l}}$ with f_Z, for every $\boldsymbol{v}$ and for all l. We then obtain the conditional density

$$p(z(\boldsymbol{v}) \mid \boldsymbol{z}_{\mathrm{Ne}(\boldsymbol{v})}) = \sum_{l=1}^{L} w_l(\boldsymbol{v})\, c_{\boldsymbol{v},l}(z(\boldsymbol{v}), z(\boldsymbol{v}_{(l)}))\, f_Z(z(\boldsymbol{v})) \tag{4.8}$$

that characterizes the stationary copula NNMP.

Under the copula framework, one strategy to specify the spatially varying parameter is through the Kendall's τ coefficient. The Kendall's τ, taking values in $[-1, 1]$, is a bivariate concordance measure with properties useful for non-Gaussian modeling. In particular, its existence does not require finite second moment and it is invariant under strictly increasing transformations. If (U_l, V_l) is continuous with a copula C_l, its Kendall's τ is $\rho_{\tau,l} = 4\int_{[0,1]^2} C_l dC_l - 1$. Taking $A_l \subset [-1, 1]$ as the range of $\rho_{\tau,l}$, we can construct a composition function $h_l := g_l \circ k_l$ for some link function $g_l : A_l \to H_l$ and kernel function $k_l : \mathcal{D} \times \mathcal{D} \to A_l$, where H_l is the parameter space associated with C_l. The kernel k_l should be specified with caution; k_l must satisfy axioms in the definition of a bivariate concordance measure (Joe 2014, Section 2.12). We illustrate the strategy with the following example.

Example 3. *Spatial Gumbel copula.*

The bivariate Gumbel copula is an asymmetric copula useful for modeling dependence when the marginals are positive and heavy-tailed. The spatial Gumbel copula can be defined as

$$C_{\boldsymbol{v},l} = \exp\Big(- \Big[\Big\{ - \log F_{U_{\boldsymbol{v},l}}(z(\boldsymbol{v})) \Big\}^{\eta_l(\boldsymbol{v})} + \Big\{ - \log F_{V_{\boldsymbol{v},l}}(z(\boldsymbol{v}_{(l)})) \Big\}^{\eta_l(\boldsymbol{v})} \Big]^{1/\eta_l(\boldsymbol{v})} \Big),$$

where $\eta_l(\boldsymbol{v}) \in [1, \infty)$ and perfect dependence is obtained if $\eta_l(\boldsymbol{v}) \to \infty$. The Kendall's τ is $\rho_{\tau,l}(\boldsymbol{v}) = 1 - \eta_l^{-1}(\boldsymbol{v})$, taking values in $[0, 1]$. We define $\rho_{\tau,l}(\boldsymbol{v}) := k_l(||\boldsymbol{v} - \boldsymbol{v}_{(l)}||)$, an isotropic correlation function. Let $g_l(x) = (1-x)^{-1}$. Then, the function $h_l(||\boldsymbol{v} - \boldsymbol{v}_{(l)}||) = g_l \circ k_l(||\boldsymbol{v} - \boldsymbol{v}_{(l)}||) = (1 - k_l(||\boldsymbol{v} - \boldsymbol{v}_{(l)}||))^{-1}$. Thus, the parameter $\eta_l(\boldsymbol{v}) \equiv \eta(||\boldsymbol{v} - \boldsymbol{v}_{(l)}||)$ is given by $h_l(||\boldsymbol{v} - \boldsymbol{v}_{(l)}||)$, and $\eta_l(\boldsymbol{v}) \to \infty$ as $||\boldsymbol{v} - \boldsymbol{v}_{(l)}|| \to 0$.

After we define a spatially varying copula, we obtain a family of copula NNMPs by choosing a desired family of marginal distributions. Examples of NNMPs with different copulas and marginals are illustrated in Section 4.4.

Copula NNMP models offer avenues to capture complex dependence using general bivariate copulas. Traditional spatial copula models specify the finite dimensional distributions of the underlying spatial process with a multivariate copula. However, multivariate copulas need to be used with careful consideration in a spatial setting. For example, it is common to assume that spatial processes exhibit stronger dependence at smaller distances. Thus, copulas such as the multivariate Archimedean copula that induce an exchangeable dependence structure are inappropriate. Though spatial vine copula models (Gräler, 2014) can resolve this restriction, their model structure and computation are substantially more complicated than copula NNMP models.

4.2.4 Mixture Component Specification and Tail Dependence

A benefit of building NNMPs from a set of base random vectors is that specification of the multivariate dependence of $Z(\boldsymbol{v})$ given its neighbors is determined mainly by that of the base random vectors. In this section, we illustrate this attractive property of the model with the establishment of lower bounds for two measures used to assess strength of tail dependence.

The main assumption is that the base random vector (U_l, V_l) has stochastically increasing positive dependence. U_l is said to be stochastically increasing in V_l, if $P\Big(U_l > u_l \,|\, V_l = v_l\Big)$ increases as v_l increases. The definition can be extended to a multivariate random vector $(Z_1, \ldots, Z_p)$. Z_1 is said to be stochastically increasing in $(Z_2, \ldots, Z_p)$ if $P\Big(Z_1 > z_1 \,|\, Z_2 = z_2, \ldots, Z_p = z_p\Big) \leq P\Big(Z_1 > z_1 \,|\, Z_2 = z_2', \ldots, Z_p = z_p'\Big)$, for all $(z_2, \ldots, z_p)$ and $(z_2', \ldots, z_p')$ in the support of $(Z_2, \ldots, Z_p)$, where $z_j \leq z_j'$, for $j = 2, \ldots, p$. The conditional density in (4.4) implies that

$$P\Big(Z(\boldsymbol{v}) > z \,|\, \boldsymbol{Z}_{\mathrm{Ne}(\boldsymbol{v})} = \boldsymbol{z}_{\mathrm{Ne}(\boldsymbol{v})}\Big) = \sum_{l=1}^{L} w_l(\boldsymbol{v})\, P\Big(Z(\boldsymbol{v}) > z \,|\, Z(\boldsymbol{v}_{(l)}) = z(\boldsymbol{v}_{(l)})\Big).$$

Therefore, $Z(\boldsymbol{v})$ is stochastically increasing in $\boldsymbol{Z}_{\mathrm{Ne}(\boldsymbol{v})}$ if $Z(\boldsymbol{v})$ is stochastically increasing in $Z(\boldsymbol{v}_{(l)})$ with respect to $(U_{\boldsymbol{v},l}, V_{\boldsymbol{v},l})$ for all l. If the sequence $(U_{\boldsymbol{v},l}, V_{\boldsymbol{v},l})$ is built from the vector (U_l, V_l), then the set of base random vectors determines the stochastically increasing positive dependence of $Z(\boldsymbol{v})$ given its neighbors.

For a bivariate random vector (U_l, V_l), the upper and lower tail dependence coefficients, denoted as $\lambda_{\mathcal{H},l}$ and $\lambda_{\mathcal{L},l}$, respectively, are $\lambda_{\mathcal{H},l} = \lim_{q \to 1^-} P\Big(U_l > F_{U_l}^{-1}(q) \,|\, V_l > F_{V_l}^{-1}(q)\Big)$ and $\lambda_{\mathcal{L},l} = \lim_{q \to 0^+} P\Big(U_l \leq F_{U_l}^{-1}(q) \,|\, V_l \leq F_{V_l}^{-1}(q)\Big)$. When $\lambda_{\mathcal{H},l} > 0$, we say U_l and V_l have upper tail dependence. When $\lambda_{\mathcal{H},l} = 0$, U_l and V_l are said to be asymptotically independent in the upper tail. Lower tail dependence and asymptotically independence in the lower tail are similarly defined using $\lambda_{\mathcal{L},l}$. Let $F_{Z(\boldsymbol{v})}$ be the marginal distribution function of $Z(\boldsymbol{v})$. Analogously, we can define the upper and lower tail dependence coefficients for $Z(\boldsymbol{v})$ given its nearest neighbors,

$$\lambda_{\mathcal{H}}(\boldsymbol{v}) = \lim_{q \to 1^-} P\Big(Z(\boldsymbol{v}) > F^{-1}_{Z(\boldsymbol{v})}(q) \mid Z(\boldsymbol{v}_{(1)}) > F^{-1}_{Z(\boldsymbol{v}_{(1)})}(q), \ldots, Z(\boldsymbol{v}_{(L)}) > F^{-1}_{Z(\boldsymbol{v}_{(L)})}(q)\Big),$$

$$\lambda_{\mathcal{L}}(\boldsymbol{v}) = \lim_{q \to 0^+} P\Big(Z(\boldsymbol{v}) \leq F^{-1}_{Z(\boldsymbol{v})}(q) \mid Z(\boldsymbol{v}_{(1)}) \leq F^{-1}_{Z(\boldsymbol{v}_{(1)})}(q), \ldots, Z(\boldsymbol{v}_{(L)}) \leq F^{-1}_{Z(\boldsymbol{v}_{(L)})}(q)\Big).$$

The following proposition provides lower bounds for the tail dependence coefficients.

Proposition 4.3. *Consider an NNMP for which the component density $f_{\boldsymbol{v},l}$ is specified by the conditional density of $U_{\boldsymbol{v},l}$ given $V_{\boldsymbol{v},l}$, where the random vector $(U_{\boldsymbol{v},l}, V_{\boldsymbol{v},l})$ follows a bivariate distribution with marginal distribution functions $F_{U_{\boldsymbol{v},l}}$ and $F_{V_{\boldsymbol{v},l}}$, for $l = 1, \ldots, L$. The spatial dependence of random vector $(U_{\boldsymbol{v},l}, V_{\boldsymbol{v},l})$ is built from the base vector (U_l, V_l), which has a bivariate distribution such that U_l is stochastically increasing in V_l, for $l = 1, \ldots, L$. Then, for every $\boldsymbol{v}$, the lower bound for the upper tail dependence coefficient $\lambda_{\mathcal{H}}(\boldsymbol{v})$ is $\sum_{l=1}^{L} w_l(\boldsymbol{v}) \lim_{q \to 1^-} P\Big(Z(\boldsymbol{v}) > F^{-1}_{U_{\boldsymbol{v},l}}(q) \,|\, Z(\boldsymbol{v}_{(l)}) = F^{-1}_{V_{\boldsymbol{v},l}}(q)\Big)$, and the lower bound for the lower tail dependence coefficient $\lambda_{\mathcal{L}}(\boldsymbol{v})$ is $\sum_{l=1}^{L} w_l(\boldsymbol{v}) \lim_{q \to 0^+} P\Big(Z(\boldsymbol{v}) \leq F^{-1}_{U_{\boldsymbol{v},l}}(q) \,|\, Z(\boldsymbol{v}_{(l)}) = F^{-1}_{V_{\boldsymbol{v},l}}(q)\Big)$.*

Proposition 4.3 establishes that the lower and upper tail dependence coefficients are bounded below by a convex combination of, respectively, the limits of the conditional distribution functions and the conditional survival functions. These are fully determined by the dependence structure of the bivariate distribution for (U_l, V_l). The result is best illustrated with an example.

Example 4. *Lomax NNMP models.*

Consider a Lomax NNMP for which the bivariate distributions of the base random vectors correspond to a bivariate Lomax distribution (Arnold et al., 1999), resulting in conditional density,

$$p(z(\boldsymbol{v}) \mid \boldsymbol{z}_{\mathrm{Ne}(\boldsymbol{v})}) = \sum_{l=1}^{L} w_l(\boldsymbol{v}) \, \mathrm{Lo}(z(\boldsymbol{v}) \mid z(\boldsymbol{v}_{(l)}) + \phi_l, \alpha_l(\boldsymbol{v})),$$

where $\mathrm{Lo}(x \mid \phi, \alpha) = \alpha\phi^{-1}(1 + x\phi^{-1})^{-(\alpha+1)}$ denotes the Lomax density, a shifted version of the Pareto Type I density. A small value of α indicates a heavy tail. The component conditional survival function of the Lomax NNMP, expressed in terms of the quantile q, is $\left\{1 + F^{-1}_{U_{\boldsymbol{v},l}}(q) / \left(F^{-1}_{V_{\boldsymbol{v},l}}(q) + \phi_l\right)\right\}^{-\alpha_l(\boldsymbol{v})}$ which converges to $2^{-\alpha_l(\boldsymbol{v})}$ as $q \to 1^-$. Therefore, the lower bound for $\lambda_{\mathcal{H}}(\boldsymbol{v})$ is $\sum_{l=1}^{L} w_l(\boldsymbol{v}) \, 2^{-\alpha_l(\boldsymbol{v})}$. As $\alpha_l(\boldsymbol{v}) \to 0$ for all l, the lower bound for $\lambda_{\mathcal{H}}(\boldsymbol{v})$ tends to one, and hence $\lambda_{\mathcal{H}}(\boldsymbol{v})$ tends to one, since $\lambda_{\mathcal{H}}(\boldsymbol{v}) \leq 1$. As $\alpha_l(\boldsymbol{v}) \to \infty$ for all l, the lower bound tends to zero.

Proposition 4.3 holds for the general framework. If the distribution of (U_l, V_l) with $F_{U_l} = F_{V_l}$ has first-order partial derivatives and exchangeable dependence, namely (U_l, V_l) and (V_l, U_l) have the same joint distribution, the lower bounds of the tail dependence coefficients depend on the component tail dependence coefficients. The result is summarized in the following corollary.

Corollary 1. *Suppose that the base random vector (U_l, V_l) in Proposition 4.3 is exchangeable, and its bivariate distribution with marginals $F_{U_l} = F_{V_l}$ has first-order partial derivatives for all l. The upper and lower tail dependence coefficients $\lambda_{\mathcal{H}}(\boldsymbol{v})$ and $\lambda_{\mathcal{L}}(\boldsymbol{v})$ are bounded below by $\sum_{l=1}^{L} w_l(\boldsymbol{v})\lambda_{\mathcal{H},l}(\boldsymbol{v})/2$ and $\sum_{l=1}^{L} w_l(\boldsymbol{v})\lambda_{\mathcal{L},l}(\boldsymbol{v})/2$, where $\lambda_{\mathcal{H},l}(\boldsymbol{v})$ and $\lambda_{\mathcal{L},l}(\boldsymbol{v})$ are tail dependence coefficients with respect to random vector $(U_{\boldsymbol{v},l}, V_{\boldsymbol{v},l})$.*

Under Corollary 1, if the bivariate distribution of (U_l, V_l) is symmetric, for instance, an elliptically symmetric distribution, the upper and lower tail dependence coefficients coincide, and can simply be denoted as $\lambda(\boldsymbol{v})$. Then, we have that $\lambda(\boldsymbol{v}) \geq \sum_{l=1}^{L} w_l(\boldsymbol{v})\lambda_l(\boldsymbol{v})/2$, where $\lambda_l(\boldsymbol{v})$ is the tail dependence coefficient with respect to $(U_{\boldsymbol{v},l}, V_{\boldsymbol{v},l})$.

Tail dependence can also be quantified using the boundary of the conditional distribution function, as proposed in Hua and Joe (2014) for a bivariate random vector. In particular, the upper tail dependence of (U_l, V_l) is said to have some strength if $F_{U_l|V_l}\Big(F_{U_l}^{-1}(q) \,|\, F_{V_l}^{-1}(1)\Big)$ is positive at $q = 1$. Likewise, a non-zero $F_{U_l|V_l}\Big(F_{U_l}^{-1}(q) \,|\, F_{V_l}^{-1}(0)\Big)$ at $q = 0$ indicates some strength of dependence in the lower tails. The functions $F_{U_l|V_l}\Big(\cdot \mid F_{V_l}^{-1}(0)\Big)$ and $F_{U_l|V_l}\Big(\cdot \mid F_{V_l}^{-1}(1)\Big)$ are referred to as the boundary conditional distribution functions.

We use $F_{1|2}\Big(\cdot \,|\, F_{\boldsymbol{Z}_{\mathrm{Ne}(\boldsymbol{v})}}^{-1}(q)\Big)$ for simpler notation for the conditional distribution function of $Z(\boldsymbol{v})$, $F\Big(\cdot \,|\, Z(\boldsymbol{v}_{(1)}) = F_{Z(\boldsymbol{v}_{(1)})}^{-1}(q), \ldots, Z(\boldsymbol{v}_{(L)}) = F_{Z(\boldsymbol{v}_{(L)})}^{-1}(q)\Big)$. Then $F_{1|2}\Big(\cdot \,|\, F_{\boldsymbol{Z}_{\mathrm{Ne}(\boldsymbol{v})}}^{-1}(0)\Big)$ and $F_{1|2}\Big(\cdot \,|\, F_{\boldsymbol{Z}_{\mathrm{Ne}(\boldsymbol{v})}}^{-1}(1)\Big)$ are the boundary conditional distribution functions for the NNMP model. The upper tail dependence is said to be i) strongest if $F_{1|2}\Big(F_{Z(\boldsymbol{v})}^{-1}(q) \,|\, F_{\boldsymbol{Z}_{\mathrm{Ne}(\boldsymbol{v})}}^{-1}(1)\Big)$ equals 0 for $0 \leq q < 1$ and has a mass of 1 at $q = 1$; ii) intermediate if $F_{1|2}\Big(F_{Z(\boldsymbol{v})}^{-1}(q) \,|\, F_{\boldsymbol{Z}_{\mathrm{Ne}(\boldsymbol{v})}}^{-1}(1)\Big)$ has positive but not unit mass at $q = 1$; iii) weakest if $F_{1|2}\Big(F_{Z(\boldsymbol{v})}^{-1}(q) \,|\, F_{\boldsymbol{Z}_{\mathrm{Ne}(\boldsymbol{v})}}^{-1}(1)\Big)$ has no mass at $q = 1$. The strength of lower tail dependence is defined likewise using $F_{1|2}\Big(F_{Z(\boldsymbol{v})}^{-1}(q) \,|\, F_{\boldsymbol{Z}_{\mathrm{Ne}(\boldsymbol{v})}}^{-1}(0)\Big)$. The following result provides lower bounds for the boundary conditional distribution functions.

Proposition 4.4. *Consider an NNMP for which the component density $f_{\boldsymbol{v},l}$ is specified by the conditional density of $U_{\boldsymbol{v},l}$ given $V_{\boldsymbol{v},l}$. The spatial dependence of random vector $(U_{\boldsymbol{v},l}, V_{\boldsymbol{v},l})$ is built from the base vector (U_l, V_l), which has a bivariate distribution such that U_l is stochastically increasing in V_l, for $l = 1, \ldots, L$. Let $\lambda_{\mathcal{L},l}(\boldsymbol{v})$ and $\lambda_{\mathcal{H},l}(\boldsymbol{v})$ be the lower and upper tail dependence coefficients corresponding to $(U_{\boldsymbol{v},l}, V_{\boldsymbol{v},l})$. If for a given $\boldsymbol{v}$, there exists $\lambda_{\mathcal{L},l}(\boldsymbol{v}) > 0$ for some l, then the conditional distribution function $F_{1|2}\Big(F_{Z(\boldsymbol{v})}^{-1}(q) \,|\, F_{\boldsymbol{Z}_{Ne(\boldsymbol{v})}}^{-1}(0)\Big)$ has strictly* *positive mass* $p_0(\boldsymbol{v})$ *at* $q = 0$ *with* $p_0(\boldsymbol{v}) \geq \sum_{l=1}^{L} w_l(\boldsymbol{v})\lambda_{\mathcal{L},l}(\boldsymbol{v})$. *Similarly, if for a given* $\boldsymbol{v}$, *there exists* $\lambda_{\mathcal{H},l}(\boldsymbol{v}) > 0$ *for some* l, *then the conditional distribution function* $F_{1|2}\Big(F^{-1}_{Z(\boldsymbol{v})}(q) \,|\, F^{-1}_{\boldsymbol{Z}_{Ne(\boldsymbol{v})}}(1)\Big)$ *has strictly positive mass* $p_1(\boldsymbol{v})$ *at* $q = 1$ *with* $p_1(\boldsymbol{v}) \geq \sum_{l=1}^{L} w_l(\boldsymbol{v})\lambda_{\mathcal{H},l}(\boldsymbol{v})$.

Proposition 4.4 complements Proposition 4.3 to assess the strength of the tail dependence. It readily applies for bivariate distributions, especially for copulas which yield explicit expressions for the tail dependence coefficients. In particular, the spatial Gumbel copula $C_{\boldsymbol{v},l}$ in Example 3 has upper tail dependence coefficient $2 - 2^{1/\eta_l(\boldsymbol{v})} > 0$ for $\eta_l(\boldsymbol{v}) > 1$, so the tail dependence of a Gumbel copula NNMP model has some strength if $\eta_l(\boldsymbol{v}) > 1$ for some l. In fact, applying the result in Hua and Joe (2014), with a Gumbel copula, $F_{1|2}\Big(F^{-1}_{Z(\boldsymbol{v})}(q) \,|\, F^{-1}_{\boldsymbol{Z}_{\mathrm{Ne}(\boldsymbol{v})}}(1)\Big)$ degenerates at $q = 1$, implying strongest tail dependence.

4.3 Bayesian Hierarchical Model and Inference

4.3.1 Hierarchical Model Formulation

We introduce the general approach for NNMP Bayesian implementation, treating the observed spatial responses as an NNMP realization. The inferential framework can be easily extended to incorporate model components that may be needed in practical settings, such as covariates and additional error terms. We illustrate the extensions with the real data analysis in Section 4.4, and provide further discussion in Section 4.5.

Our approach for inference is based on a likelihood conditional on the first L elements of the realization $\boldsymbol{z}_{\mathcal{S}} = (z(\boldsymbol{s}_1), \ldots, z(\boldsymbol{s}_n))^\top$ over the reference set $\mathcal{S} \subset \mathcal{D}$. Following a commonly used approach for mixture models fitting, we use data augmentation to facilitate inference. For $z(\boldsymbol{s}_i)$, $i = L+1, \ldots, n$, we introduce a configuration variable ℓ_i, taking values in $\{1, \dots, L\}$, such that $P\Big(\ell_i \,|\, \boldsymbol{w}(\boldsymbol{s}_i)\Big) = \sum_{l=1}^{L} w_l(\boldsymbol{s}_i)\delta_l(\ell_i)$, where $\boldsymbol{w}(\boldsymbol{s}_i) = (w_1(\boldsymbol{s}_i), \dots, w_L(\boldsymbol{s}_i))^\top$, and $\delta_l(\ell_i) = 1$ if $\ell_i = l$ and 0 otherwise. Conditional on the configuration variables and the vector $(z(\boldsymbol{s}_1), \dots, z(\boldsymbol{s}_L))^\top$, the augmented model on $z(\boldsymbol{s}_i)$ is

$$\begin{aligned} z(\boldsymbol{s}_i) \mid z(\boldsymbol{s}_{(i,\ell_i)}), \ell_i, \boldsymbol{\theta} &\overset{ind.}{\sim} f_{\boldsymbol{s}_i,\ell_i}(z(\boldsymbol{s}_i) \,|\, z(\boldsymbol{s}_{(i,\ell_i)}), \boldsymbol{\theta}), \\ \ell_i \mid \boldsymbol{w}(\boldsymbol{s}_i) &\overset{ind.}{\sim} \sum_{l=1}^{L} w_l(\boldsymbol{s}_i)\delta_l(\ell_i), \end{aligned} \tag{4.9}$$

where $\boldsymbol{\theta}$ collects the parameters of the densities $f_{\boldsymbol{s}_i,l}$.

A key component of the Bayesian model formulation is the prior model for the weights. Weights are allowed to vary in space, adjusting to the neighbor structure of different reference locations. We describe the construction for weights corresponding to a point in the reference set. For non-reference points, weights are defined analogously. Consider a collection of spatially dependent distribution functions $\{G_{\boldsymbol{s}_i} : \boldsymbol{s}_i \in \mathcal{S}\}$ supported on $(0, 1)$. For each $\boldsymbol{s}_i$, the weights are defined as the increments of $G_{\boldsymbol{s}_i}$ with cutoff points $r_{\boldsymbol{s}_i,0}, \dots, r_{\boldsymbol{s}_i,L}$. More specifically,

$$w_l(\boldsymbol{s}_i) = \int \mathbb{1}_{(r_{\boldsymbol{s}_i,l-1}, r_{\boldsymbol{s}_i,l})}(t)\, dG_{\boldsymbol{s}_i}(t), \quad l = 1, \dots, L, \tag{4.10}$$

where $\mathbb{1}_A$ denotes the indicator function for set A. The cutoff points $0 = r_{\boldsymbol{s}_i,0} < r_{\boldsymbol{s}_i,1} < \cdots < r_{\boldsymbol{s}_i,L} = 1$ are such that, for $l = 1, \dots, L$, $r_{\boldsymbol{s}_i,l} - r_{\boldsymbol{s}_i,l-1} = k'(\boldsymbol{s}_i, \boldsymbol{s}_{(il)} \,|\, \boldsymbol{\zeta}) \,/ \sum_{l=1}^{L} k'(\boldsymbol{s}_i, \boldsymbol{s}_{(il)} \,|\, \boldsymbol{\zeta})$, where $k' : \mathcal{D} \times \mathcal{D} \to [0, 1]$ is a bounded kernel function with parameters $\boldsymbol{\zeta}$. The kernel and its associated parameters affect the smoothness of the resulting random field. By default we take $G_{\boldsymbol{s}_i}$ as a logit Gaussian distribution, denoted as $G_{\boldsymbol{s}_i}(\cdot \,|\, \mu(\boldsymbol{s}_i), \kappa^2)$, such that the corresponding Gaussian distribution has mean $\mu(\boldsymbol{s}_i)$ and variance κ^2. The spatial dependence across the weights is introduced through the mean $\mu(\boldsymbol{s}_i) = \gamma_0 + \gamma_1 s_{i1} + \gamma_2 s_{i2}$, where $\boldsymbol{s}_i = (s_{i1}, s_{i2})$.

Given the cutoff points and κ^2, a small value of $\mu(\boldsymbol{s}_i)$ favors large weights for the near neighbors of $\boldsymbol{s}_i$. A simpler version of the model in (4.10) is obtained by letting $G_{\boldsymbol{s}_i}$ be the uniform distribution on $(0,1)$. Then the weights become $k'(\boldsymbol{s}_i, \boldsymbol{s}_{(il)} \,|\, \boldsymbol{\zeta}) / \sum_{l=1}^{L} k'(\boldsymbol{s}_i, \boldsymbol{s}_{(il)} \,|\, \boldsymbol{\zeta})$. We notice that Cadonna et al. (2019) use a set of fixed, uniform cutoff points on $[0,1]$, i.e., $r_{\boldsymbol{s}_i,l} - r_{\boldsymbol{s}_i,l-1} = 1/L$, for spectral density estimation, with a collection of logit Gaussian distributions indexed by frequency.

The full Bayesian model is completed with prior specification for parameters $\boldsymbol{\theta}, \boldsymbol{\zeta}, \boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2)^\top$, and κ^2. The priors for $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}$ depend on the choices of the densities $f_{\boldsymbol{s}_i,l}$ and the cutoff point kernel k', respectively. For parameters $\boldsymbol{\gamma}$ and κ^2, we specify $N(\boldsymbol{\gamma} \,|\, \boldsymbol{\mu}_\gamma, \boldsymbol{V}_\gamma)$ and $\mathrm{IG}(\kappa^2 \,|\, u_{\kappa^2}, v_{\kappa^2})$ priors, respectively, where IG denotes the inverse gamma distribution.

Finally, we note that an NNMP model requires selection of the neighborhood size L. This can be done using standard model comparison metrics, scoring rules, or information criteria. In general, a larger L increases computational cost. Datta et al. (2016a) conclude that a moderate value L ($\leq$ 20) typically suffices for NNGPs. On the other hand, it is possible that information from distant neighbors is also important (Stein et al., 2004). Therefore, one may seek a larger L to include more neighbor information for large non-Gaussian data sets with complex dependence. Our model for the weights allows taking a relatively large neighbor set with less computational demand. We assign small probabilities a priori to distant neighbors. The contribution of each neighbor will be induced by the mixing, with important neighbors being assigned large weights a posteriori.

4.3.2 Estimation and Prediction

We implement an MCMC sampler to simulate from the posterior distribution of the model parameters. To allow for efficient simulation of parameters $\boldsymbol{\gamma}$ and

κ^2, we associate each $y(\boldsymbol{s}_i)$ with a latent Gaussian variable t_i with mean $\mu(\boldsymbol{s}_i)$ and variance κ^2, for $i = L+1, \ldots, n$. There is a one-to-one correspondence between the configuration variables ℓ_i and latent variables t_i: $\ell_i = l$ if and only if $t_i \in (r^*_{\boldsymbol{s}_i,l-1}, r^*_{\boldsymbol{s}_i,l})$ where $r^*_{\boldsymbol{s}_i,l} = \log(r_{\boldsymbol{s}_i,l}/(1 - r_{\boldsymbol{s}_i,l}))$, for $l = 1, \ldots, L$. The posterior distribution of the model parameters, based on the new augmented model, is

$$p(\boldsymbol{\theta}, \boldsymbol{\zeta}, \boldsymbol{\gamma}, \kappa^2, \{t_i\}_{i=L+1}^n \,|\, \boldsymbol{z}_{\mathcal{S}}) \propto \; \pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \times \pi_{\boldsymbol{\zeta}}(\boldsymbol{\zeta}) \times N(\boldsymbol{\gamma} \,|\, \boldsymbol{\mu}_\gamma, \boldsymbol{V}_\gamma) \times \mathrm{IG}(\kappa^2 \,|\, u_{\kappa^2}, v_{\kappa^2})$$
$$\times N(\boldsymbol{t} \,|\, \boldsymbol{D}\boldsymbol{\gamma},\, \kappa^2 \mathbf{I}_{n-L}) \times \prod_{i=L+1}^{n} \sum_{l=1}^{L} f_{\boldsymbol{s}_i,l}(z(\boldsymbol{s}_i) \,|\, z(\boldsymbol{s}_{(il)}), \boldsymbol{\theta}) \mathbb{1}_{(r^*_{\boldsymbol{s}_i,l-1}, r^*_{\boldsymbol{s}_i,l})}(t_i),$$

where $\pi_{\boldsymbol{\theta}}$ and $\pi_{\boldsymbol{\zeta}}$ are the priors for $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}$, respectively, $\mathbf{I}_{n-L}$ is an $(n-L)\times(n-L)$ identity matrix, the vector $\boldsymbol{t} = (t_{L+1}, \ldots, t_n)^\top$, and the matrix $\boldsymbol{D}$ is $(n-L) \times 3$ such that the ith row is $(1, s_{L+i,1}, s_{L+i,2})$.

The posterior full conditional distribution of $\boldsymbol{\theta}$ depends on the form of $f_{\boldsymbol{s}_i,l}$. To update $\boldsymbol{\zeta}$, we first marginalize out the latent variables t_i from the joint posterior distribution. We then update $\boldsymbol{\zeta}$ using a random walk Metropolis step with target density $\pi_{\boldsymbol{\zeta}}(\boldsymbol{\zeta}) \prod_{i=L+1}^n \{G_{\boldsymbol{s}_i}(r_{\boldsymbol{s}_i,\ell_i} \,|\, \mu(\boldsymbol{s}_i), \kappa^2) - G_{\boldsymbol{s}_i}(r_{\boldsymbol{s}_i,\ell_i-1} \,|\, \mu(\boldsymbol{s}_i), \kappa^2)\}$. The posterior full conditional distribution of t_i is $\sum_{l=1}^{i_L} q_l(\boldsymbol{s}_i) \mathrm{TN}(t_i \,|\, \mu(\boldsymbol{s}_i), \kappa^2; r^*_{\boldsymbol{s}_i,l-1}, r^*_{\boldsymbol{s}_i,l})$, where $q_l(\boldsymbol{s}_i) \propto w_l(\boldsymbol{s}_i) f_{\boldsymbol{s}_i,l}(z(\boldsymbol{s}_i) \,|\, z(\boldsymbol{s}_{(il)}), \boldsymbol{\theta})$ and $w_l(\boldsymbol{s}_i) = G_{\boldsymbol{s}_i}(r_{\boldsymbol{s}_i,l} \,|\, \mu(\boldsymbol{s}_i), \kappa^2) - G_{\boldsymbol{s}_i}(r_{\boldsymbol{s}_i,l-1} \,|\, \mu(\boldsymbol{s}_i), \kappa^2)$, for $l = 1, ..., L$. Hence, each t_i can be updated by sampling from the l-th truncated Gaussian with probability proportional to $q_l(\boldsymbol{s}_i)$. The posterior full conditional distribution of $\boldsymbol{\gamma}$ is $N(\boldsymbol{\gamma} \,|\, \boldsymbol{\mu}^*_\gamma, \boldsymbol{V}^*_\gamma)$, where $\boldsymbol{V}^*_\gamma = (\boldsymbol{V}_\gamma^{-1} + \kappa^{-2} \boldsymbol{D}^\top \boldsymbol{D})^{-1}$ and $\boldsymbol{\mu}^*_\gamma = \boldsymbol{V}^*_\gamma(\boldsymbol{V}_\gamma^{-1} \boldsymbol{\mu}_\gamma + \kappa^{-2} \boldsymbol{D}^\top \boldsymbol{t})$. The posterior full conditional of κ^2 is $\mathrm{IG}(\kappa^2 \,|\, u_{\kappa^2} + (n-L)/2,\, v_{\kappa^2} + \sum_{i=L+1}^n (t_i - \mu(\boldsymbol{s}_i))^2/2)$.

Turning to the prediction, let $\boldsymbol{v}_0 \in \mathcal{D}$ be a new location of interest. We obtain posterior predictive samples of $z(\boldsymbol{v}_0)$ in the following way. If $\boldsymbol{v}_0 \notin \mathcal{S}$, for each posterior sample of the parameters, we first compute the cutoff points $r_{\boldsymbol{v}_0,l}$ for which $r_{\boldsymbol{v}_0,l} - r_{\boldsymbol{v}_0,l-1} = k'(\boldsymbol{v}_0, \boldsymbol{v}_{(0l)} \,|\, \boldsymbol{\zeta}) / \sum_{l=1}^L k'(\boldsymbol{v}_0, \boldsymbol{v}_{(0l)} \,|\, \boldsymbol{\zeta})$, and obtain the weights

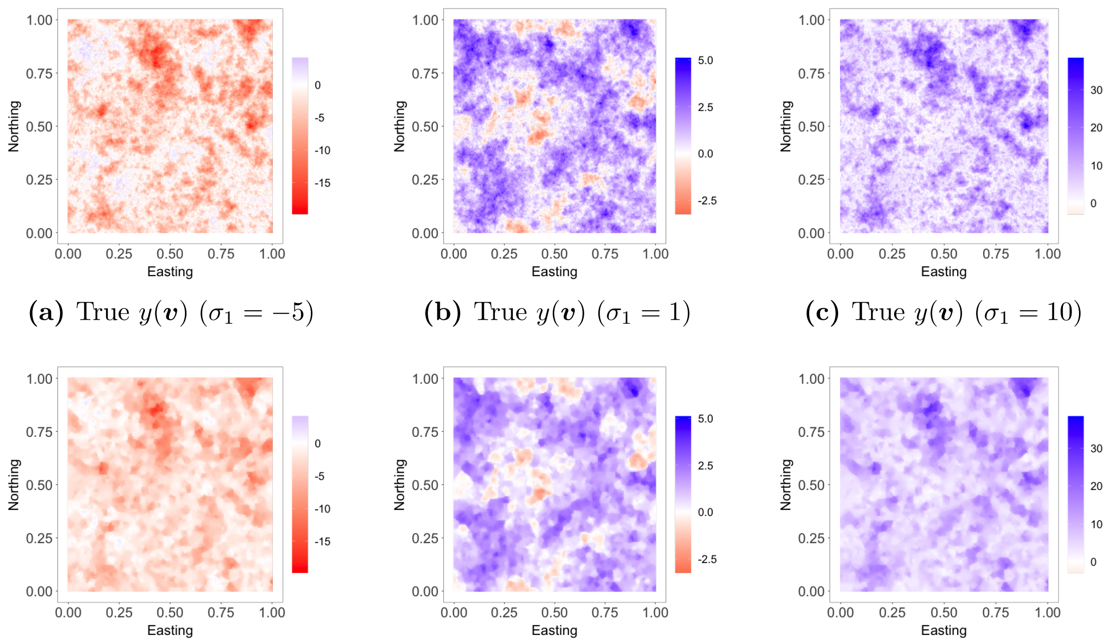
$w_l(\boldsymbol{v}_0) = G_{\boldsymbol{v}_0}(r_{\boldsymbol{v}_0,l} \,|\, \mu(\boldsymbol{v}_0), \kappa^2) - G_{\boldsymbol{v}_0}(r_{\boldsymbol{v}_0,l-1} \,|\, \mu(\boldsymbol{v}_0), \kappa^2)$ for $l = 1, \ldots, L$. We then predict $z(\boldsymbol{v}_0)$ using (4.3). If $\boldsymbol{v}_0 \equiv \boldsymbol{s}_i \in \mathcal{S}$, we generate $z(\boldsymbol{v}_0)$ similar to the earlier case but using posterior samples of the weights collected from the MCMC, and applying (4.2) instead of (4.3) to generate $z(\boldsymbol{v}_0)$.

4.4 Data Illustrations

We present three synthetic data examples and an analysis of the Mediterranean Sea surface temperature data to demonstrate the benefits of the proposed modeling framework. For the simulation experiments, first, we illustrate the ability of the NNMP model to handle skewed data using a skew-GNNMP model. Next, we study inference for tail dependence using copula NNMP models. Finally, we demonstrate the effectiveness of the NNMP model for bounded spatial data.

In each experiment, we created a regular grid of 200×200 resolution on a unit square domain, and generated data ver the grid. We randomly selected 2000 locations as the reference set with a random ordering for model fitting. For the purpose of illustration, we chose neighborhood size $L = 10$ for all cases.

Simulation study results are based on posterior samples collected every 10 iterations from a Markov chain of 30000 iterations, with the first 10000 samples being discarded. The MCMC algorithms were implemented in the R programming language on a computer with a 2-GHz Intel Core i5 processor and 32-GB RAM. We integrated C++ code for the update of latent variables without particular emphasis on optimizing the code. The computing time for the models in the three experiments was around 9, 18, and 18 minutes, respectively.



(a) True $y(\boldsymbol{v})$ ($\sigma_1 = -5$) **(b)** True $y(\boldsymbol{v})$ ($\sigma_1 = 1$) **(c)** True $y(\boldsymbol{v})$ ($\sigma_1 = 10$)

(d) Skew-GNNMP **(e)** Skew-GNNMP **(f)** Skew-GNNMP

Figure 4.1: Chapter 4 - first simulation data analysis. Top panels are interpolated surfaces of $y(\boldsymbol{v})$ generated by the true model. Bottom panels are the posterior median estimates from the skew-GNNMP model.

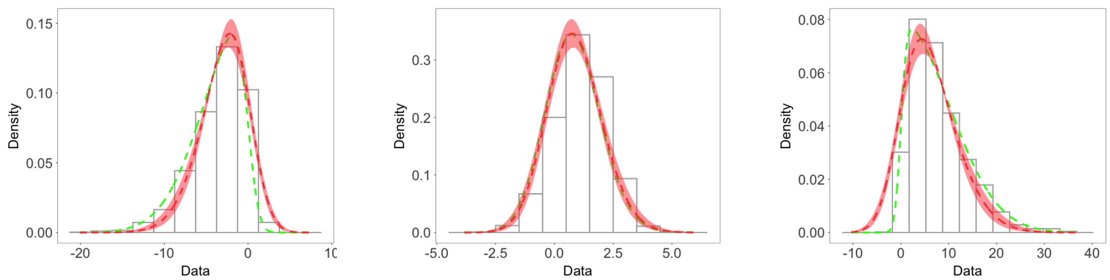
4.4.1 First Simulation Experiment

We generated data from the following skew-Gaussian process (Zhang and El-Shaarawi, 2010),

$$y(\boldsymbol{v}) = \sigma_1 \, |\omega_1(\boldsymbol{v})| + \sigma_2 \, \omega_2(\boldsymbol{v}), \quad \boldsymbol{v} \in \mathcal{D},$$

where $\omega_1(\boldsymbol{v})$ and $\omega_2(\boldsymbol{v})$ are both standard Gaussian processes with correlation matrix specified by an exponential correlation function with range parameter $1/12$. The parameter $\sigma_1 \in \mathbb{R}$ controls the skewness, whereas $\sigma_2 > 0$ is a scale parameter. The model has a stationary skew-Gaussian marginal density $\mathrm{SN}(0, \sigma_1^2 + \sigma_2^2, \sigma_1/\sigma_2)$. We took $\sigma_2 = 1$, and generated data with $\sigma_1 = -5$, 1 and 10, resulting in three random fields with different levels of skewness, as shown in Figure 4.1(a)-4.1(c).

We applied the stationary skew-GNNMP model. The model is obtained as a special case of the skew-GNNMP model in (4.7), taking $\lambda_l = \lambda$, $\mu_l = 0$, and



(a) Estimated marginal ($\sigma_1 = -5$)

(b) Estimated marginal ($\sigma_1 = 1$)

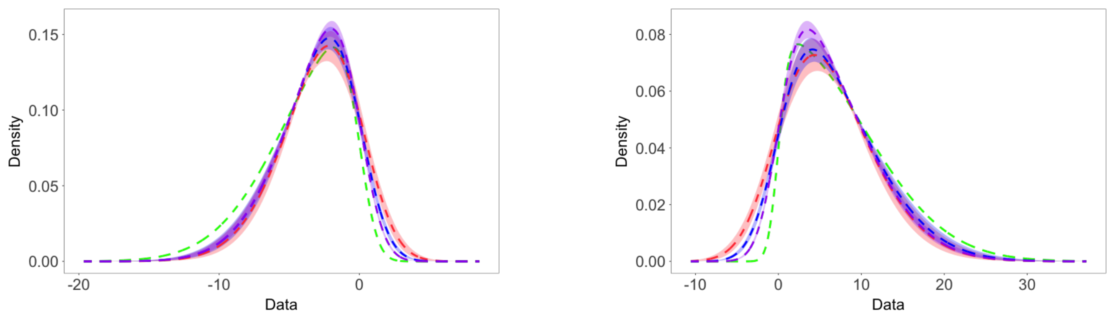
(c) Estimated marginal ($\sigma_1 = 10$)

Figure 4.2: Chapter 4 - first simulation data analysis. Green lines are true marginal densities. Dashed lines and shaded regions are posterior means and 95% credible interval estimates.

$\sigma_l^2 = \sigma^2$, for all l. Here, $\lambda \in \mathbb{R}$ controls the skewness, such that a large positive (negative) value of λ indicates strong positive (negative) skewness. If $\lambda = 0$, the skew-GNNMP model reduces to the GNNMP model. After marginalizing out z_0, we obtain a stationary skew-Gaussian marginal density $\mathrm{SN}(0, \lambda^2 + \sigma^2, \lambda/\sigma)$. We completed the full Bayesian specification for the model, by assigning priors $N(\lambda \,|\, 0, 5)$, $\mathrm{IG}(\sigma^2 \,|\, 2, 1)$, $\mathrm{IG}(\phi \,|\, 3, 1/3)$, $\mathrm{IG}(\zeta \,|\, 3, 0.2)$, $N(\boldsymbol{\gamma} \,|\, (-1.5, 0, 0)^\top, 2\mathbf{I}_3)$, and $\mathrm{IG}(\kappa^2 \,|\, 3, 1)$, where ζ is the range parameter of the exponential correlation function specified for the cutoff point kernel.

We focus on the model performance on capturing skewness. The posterior mean and 95% credible interval estimates of the parameter λ for the three scenarios were $-3.65\,(-4.10, -3.27)$, $1.09\,(0.91, 1.28)$, and $7.69\,(6.88, 8.68)$, respectively, indicating the model's ability to estimate different levels of skewness. The bottom row of Figure 4.1 shows that the posterior median estimates of the surfaces capture well features of the true surfaces, even when the level of skewness is small, thus demonstrating that the model is also able to recover near-Gaussian features.

Figure 4.2 plots the posterior mean and pointwise 95% credible interval for the marginal density, overlaid on the histogram of the simulated data for each of the three cases. In particular, we can observe that the estimates do not align well with



(a) Estimated marginals ($\sigma_1 = -5$) **(b)** Estimated marginals ($\sigma_1 = 1$)

Figure 4.3: Chapter 4 - first simulation data analysis. Dashed lines and shaded regions are posterior means and 95% credible interval estimates, with colors in red, blue and purple corresponding to sample sizes $n = 2000$, 10000, 50000, respectively.

the true density for the scenarios of $\sigma_1 = -5$ and 10. We note that, in general, the skewness characteristic of a distribution/process is difficult to estimate; see, e.g., Liseo and Loperfido (2006) and Liseo and Parisi (2013). Moreover, for this example, the model we used to fit the data is very different from the data generating process, and the data correspond to a single stochastic process realization, which may not provide enough information to accurately estimate the skewness. We did an additional simulation experiment. For the scenarios with $\sigma_1 = -5$ and 10, we generated 10000 and 50000 observations, and fit the model to the data. Figure 4.3 shows the posterior mean and pointwise 95% credible interval for the marginal density, with different sample sizes $n = 2000, 10000, 50000$. The posterior estimate, as shown in the figure, is closer to the true density when the sample size increases. Overall, we believe that this simulation experiment demonstrates the adaptability of the skew-GNNMP model in capturing skewed random fields with different levels of skewness, as well as our versatile framework for modeling different non-Gaussian behaviors.

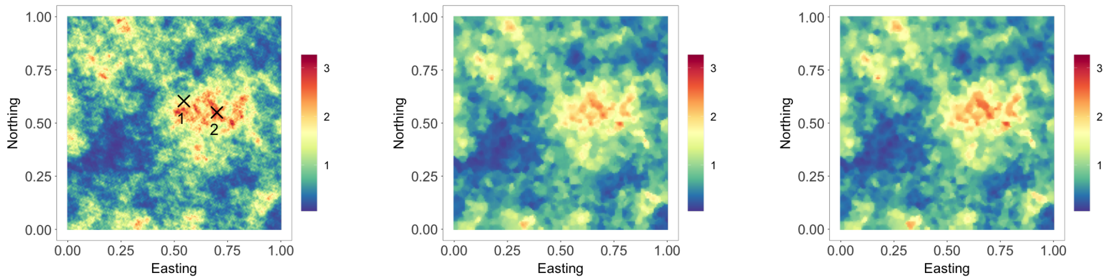
4.4.2 Second Simulation Experiment

The goal of the second experiment is to demonstrate the use of copulas to construct NNMPs for tail dependence modeling. We note that the focus here is to illustrate the flexibility of the NNMPs with copulas for modeling complex dependence structures, but not for extreme value modeling. To this end, we generated data from the random field

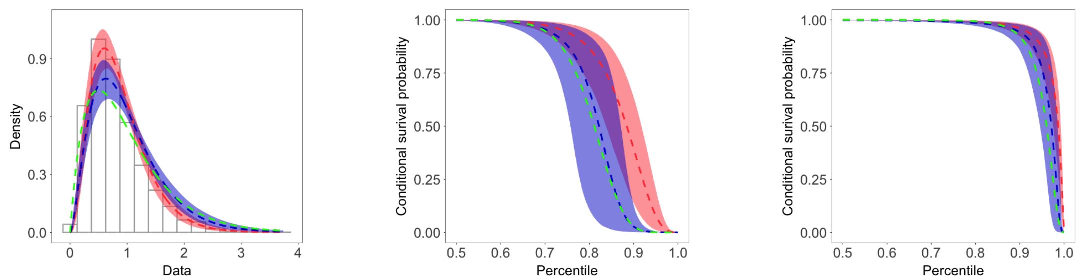
$$y(\boldsymbol{v}) = F^{-1}\Big(T_\nu(\omega(\boldsymbol{v}))\Big), \quad \boldsymbol{v} \in \mathcal{D},$$

where $\omega(\boldsymbol{v})$ is a standard Student-t process with tail parameter ν and scale matrix specified by an exponential correlation function with range parameter ϕ_w. The distribution functions F and T_ν correspond to a gamma distribution $\mathrm{Ga}(2,2)$ and a standard Student-t distribution with tail parameter ν, respectively. For a given pair of locations in $\mathcal{D}$ with correlation $\rho_0 = \exp(-d_0/\phi_w)$, the corresponding tail dependence coefficient of the random field is $\chi_\nu = 2T_{\nu+1}\Big(-\sqrt{(1+\nu)(1-\rho_0)/(1+\rho_0)}\Big)$. We took $\phi_w = 1/12$, and chose $\nu = 10$ so that the synthetic data exhibits moderate tail dependence at close distance, and the dependence decreases rapidly as the distance d_0 becomes larger. When $\rho_0 = 0.05, 0.5, 0.95$, $\chi_{10} = 0.01, 0.08, 0.61$, respectively.

We applied two copula NNMP models. The models are of the form in (4.8) with stationary gamma marginal $\mathrm{Ga}(a,b)$ with mean a/b. In the first model, the component copula density $c_{\boldsymbol{v},l}$ corresponds to a bivariate Gaussian copula that is known to be unsuitable for tail dependence modeling. The correlation parameter of the copula was specified by an exponential correlation function with range parameter ϕ_1. In the second model, we consider a spatially varying Gumbel copula as in Example 3. The spatially varying parameter of the copula density



(a) True $y(\boldsymbol{v})$ **(b)** Gaussian copula NNMP **(c)** Gumbel copula NNMP



(d) Estimated marginals from two models **(e)** Estimated probabilities at site 1 **(f)** Estimated probabilities at site 2

Figure 4.4: Chapter 4 - second simulation data analysis. Top panels are interpolated surfaces of the true field and posterior median estimates from both models. Bottom panels are estimated marginal densities and conditional survival probabilities from the two models. The green dashed lines correspond to the true model. The red (blue) dash lines and shaded regions are the posterior mean and 95% credible interval estimates from the Gaussian (Gumbel) copula NNMP models.

is defined with the link function $\eta_l(\boldsymbol{v}) \equiv \eta_l(||\boldsymbol{v} - \boldsymbol{v}_{(l)}||) = \min\{(1 - \exp(-||\boldsymbol{v} - \boldsymbol{v}_{(l)}||/\phi_2))^{-1}, 50\}$, where the upper bound 50 ensures numerical stability. When $\eta_l(d_0) = 50$, $\exp(-d_0/\phi_2) = 0.98$. With this link function, we assume that given ϕ_2, the strength of the tail dependence with respect to the lth component of the Gumbel model stays the same for any distance smaller than d_0 between two locations. For the cutoff point kernels, we specified an exponential correlation function with range parameters ζ_1 and ζ_2, respectively, for each model. The Bayesian model is fully specified with a $\mathrm{IG}(3, 1/3)$ prior for ϕ_1 and ϕ_2, a $\mathrm{Ga}(1, 1)$ prior for a and b, a $\mathrm{IG}(3, 0.2)$ prior for ζ_1 and ζ_2, $N(\boldsymbol{\gamma} \,|\, (-1.5, 0, 0)^\top, 2\mathbf{I}_3)$, and $\mathrm{IG}(\kappa^2 \,|\, 3, 1)$ priors.

We focus on the performance of the two models with respect to tail depen-

Table 4.1: Chapter 4 - second simulation data analysis. Log-scores for subsets that exceed the c-th percentile of the held-out data

c	0	10	30	50	70	90	95
Gaussian copula	140.009	80.713	30.224	-15.479	-28.806	-20.377	-13.750
Gumbel copula	118.684	63.962	17.242	-24.300	-28.232	-16.952	-11.422

dence inference. Table 4.1 presents the log-scores (Gneiting and Raftery, 2007) for subsets of the held-out data that exceed the c-th percentile of the held-out data. The out-of-sample log-score is the predictive log-likelihood averaging over the model parameters. It reflects the ability of a model to capture dependence structure in the data. We can see that for held-out data that exceed high sample percentiles, the Gumbel copula model gives a higher log-score.

Figure 4.4 shows the random fields, marginals and conditional survival probabilities estimated by the two models. From Figure 4.4(a)-4.4(c), we see that, comparing with the true field, the posterior median estimate by the Gumbel copula model seems to recover the large values better than the Gaussian copula model. Besides, as shown in Figure 4.4(d), the Gumbel copula NNMP model provides a more accurate estimate of the marginal distribution, especially in the tails. We computed the conditional survival probabilities at two different unobserved sites marked in Figure 4.4(a). In particular, Site 1 is surrounded with reference observations with moderate values, while Site 2 is surrounded with large reference observations. We see that the Gumbel copula model provides much closer estimates to the probabilities, indicating that the model captures better the tail dependence structure in the data. Overall, this example demonstrates that the Gumbel copula NNMP model is a useful option for modeling spatial processes with tail dependence.

4.4.3 Third Simulation Experiment

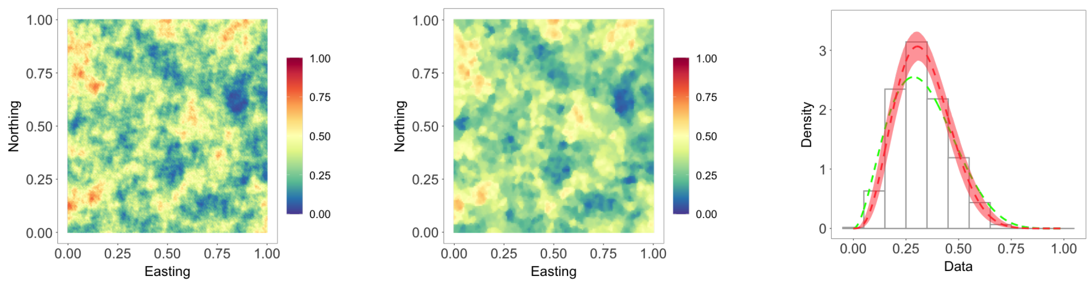
Many spatial processes are measured over a compact interval. As an example, data on proportions are common in ecological applications. In this experiment, we demonstrate the effectiveness of the NNMP model for directly modeling bounded spatial data. In particular, we generated data using the following model

$$y(\boldsymbol{v}) = F^{-1}\Big(\Phi(\omega(\boldsymbol{v}))\Big),$$

where the distribution function F corresponds to a beta distribution, denoted as $\text{Beta}(a_0, b_0)$, and $\omega(\boldsymbol{v})$ is a standard Gaussian process with exponential correlation function with range parameter 0.1. We set $a_0 = 3$, $b_0 = 6$.

We applied a Gaussian copula NNMP model with stationary marginal beta distribution, denoted as $\text{Beta}(a, b)$, with the same spatial Gaussian copula and prior specification used in the second experiment. Figure 4.5(b) shows the estimated random field which captures well the main features of the true field in Figure 4.5(a). The posterior mean and pointwise 95% credible interval of the estimated marginal density in Figure 4.5(c) overlay on the data histogram. These show that the beta NNMP estimation and prediction provide good approximation to the true field.

It is worth mentioning that implementing the beta NNMP model is simpler than fitting existing models for data corresponding to proportions. For example, a spatial Gaussian copula model, that corresponds to the data generating process of this experiment, involves computations for large matrices. Alternatively, if a multivariate non-Gaussian copula is used, the resulting likelihood can be intractable and require certain approximations. Another model that is commonly used in this setting is defined analogously to a spatial generalized linear mixed model.



(a) True $y(\boldsymbol{v})$ **(b)** Beta NNMP **(c)** Estimated marginals

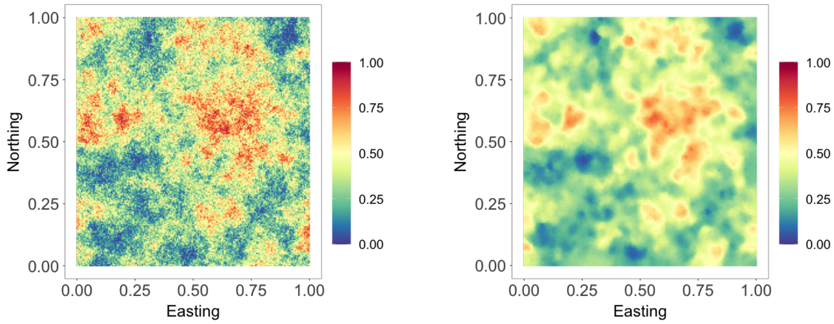
Figure 4.5: Chapter 4 - third simulation data analysis. Panels (a) and (b) are interpolated surfaces of the true field and posterior median estimate from the beta NNMP model, respectively. In Panel (c), the green dotted line corresponds to the true marginal. The red dash line and shaded region are the posterior mean and pointwise 95% credible interval for the estimated marginal.

The spatial element in the model is introduced through the transformed mean of the observations. A sample-based approach to fit such model requires sampling a large number of highly correlated latent variables. We conducted an additional simulation experiment to demonstrate the effectiveness of the beta NNMP to approximate the random field simulated by the link function approach. In particular, we generated bounded data using the following model,

$$
\begin{aligned}
y(\boldsymbol{v}) \mid \mu(\boldsymbol{v}), \psi &\sim \text{Beta}(\mu(\boldsymbol{v})\psi, (1-\mu(\boldsymbol{v}))\psi), \\
\text{logit}(\mu(\boldsymbol{v})) &= \mu_0 + \sigma_0 \omega(\boldsymbol{v}).
\end{aligned}
$$

The above model is analogous to a spatial generalized linear mixed model where the mean $\mu(\boldsymbol{v})$ of the beta distribution is modeled via a logit link function, and $\omega(\boldsymbol{v})$ is a standard Gaussian process with exponential correlation function with range parameter 0.1. We set $\psi = 20$, $\mu_0 = -0.5$ and $\sigma_0 = 0.8$.

Since our purpose is primarily demonstrative, we applied a Gaussian copula NNMP model with a stationary beta marginal $\text{Beta}(a, b)$, referred to as the beta NNMP model. The correlation parameter of the Gaussian copula was specified by an exponential correlation function with range parameter ϕ. We specified an



(a) True $y(\boldsymbol{v})$ **(b)** Beta NNMP

Figure 4.6: Chapter 4 - third simulation data analysis. Interpolated surfaces of the true field and posterior median estimate from the beta NNMP model.

exponential correlation function for the random cutoff points kernel function with range parameter ζ. The Bayesian model is fully specified with a $\text{IG}(3, 1/3)$ prior for ϕ, a $\text{Ga}(1, 1)$ prior for a and b, a $\text{IG}(3, 0.2)$ prior for ζ, $N(\boldsymbol{\gamma} \,|\, (-1.5, 0, 0)^\top, \, 2\mathbf{I}_3)$ and $\text{IG}(\kappa^2 \,|\, 3, 1)$.

We trained the model using 2000 observations. Figure 4.6(a)-(b) shows the interpolated surface of the true field and the predictive field given by the beta NNMP model. Although the beta NNMP's stationary marginal distribution assumption does not align with the true model, we can see that the predictive filed was able to capture the main feature of the true field. Moreover, it is worth mentioning that the MCMC algorithm for the beta NNMP to fit the data set took around 18 minutes with 30000 iterations. This is substantially faster than the MCMC algorithm for fitting the true model which involves sampling a large number of highly correlated latent variables.

4.4.4 Mediterranean Sea Surface Temperature Data

The study of Ocean's dynamics is crucial for understanding climate variability. One of the most valuable sources of information regarding the evolution of the state of the ocean is provided by the centuries-long record of temperature

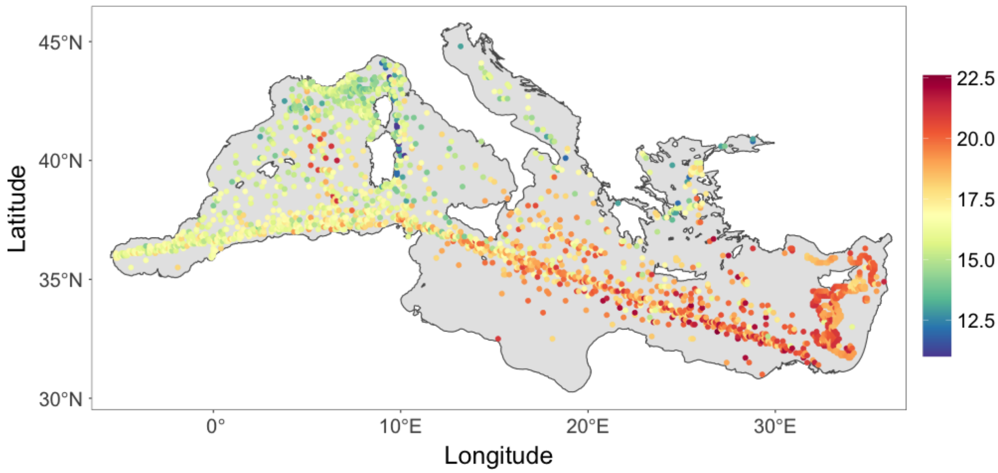


Figure 4.7: Chapter 4 - Mediterranean SST data analysis. Observed SST.

observations recorded from the surface of the oceans. The record of sea surface temperatures (SST) consists of data collected over time at irregularly scattered locations. In this section, we examine the SST from the Mediterranean Sea area during December 2003.

It is well known that the Mediterranean Sea area produces very heterogeneous temperature fields. A goal of the spatial analysis of SST in the area is to generate a spatially continuous field that accounts for the complexity of the surrounding coastlines as well as the non-linear dynamics of the circulation system. An additional source of complexity comes from the data collection process. Historically, SST observations are collected from different types of devices: buckets launched from navigating vessels, readings from the water intake of ships' engine rooms, moored buoys, and drifting buoys (Kirsner and Sansó, 2020). The source of some observations is known, but not all the data are labelled. A thorough case study will be needed to include all this information in order to account for possible heterogeneities due to the different measuring devices. That is beyond the scope of this paper. We will focus on demonstrating the ability of the proposed framework to model non-Gaussian spatial processes that, hopefully, capture the complexities of the physical process and the data collection protocol better than Gaussian processes. We notice that in the original record several sites had multiple observations. In those cases we took the median of the observations, resulting in a total of 1966 observations. The data are shown in Figure 4.7.

We first focus on a limited region that allows us to explore in detail the behavior of the GNNMP. The GNNMP has the same Gaussian marginals as the NNGP, but its finite-dimensional distribution is a mixture of multivariate Gaussian distributions. We compare the GNNMP with the NNGP in a spatially varying regression model, demonstrating the benefit of using a non-Gaussian process to explain the SST variability. We then illustrate the ability of the NNMPs to model non-Gaussian marginals by using an extended skew-GNNMP model to analyze the whole data set.

Regional Analysis

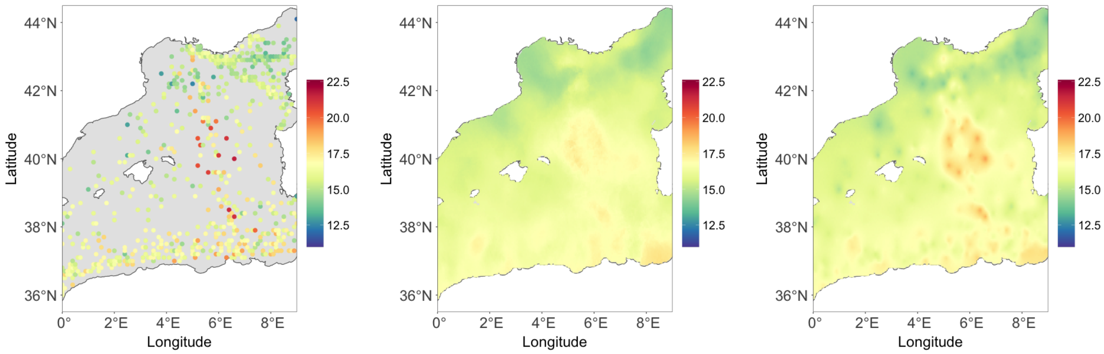
We first focus on SST over an area near the Gulf of Lion, along the islands near the shores of Spain, France, Monaco and Italy, between 0 - 9 E. longitude and 33.5 - 44.5 N. latitude. There are 642 observations in the region. As shown in Figure 4.8(a), the SST observations are very heterogeneous, implying that the short range variability is likely to be non-Gaussian. To capture the variability, we consider the following spatially varying regression model,

$$y(\boldsymbol{v}) = \boldsymbol{x}(\boldsymbol{v})^\top \boldsymbol{\beta} + z(\boldsymbol{v}) + \epsilon(\boldsymbol{v}), \quad \boldsymbol{v} \in \mathcal{D}, \tag{4.11}$$

where $y(\boldsymbol{v})$ is the SST observation, $\boldsymbol{x}(\boldsymbol{v}) = (1, v_1, v_2)^\top$ includes longitude v_1 and latitude v_2 to account for the long range variability in SST with regression parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^\top$, $z(\boldsymbol{v})$ is a spatial process, and $\epsilon(\boldsymbol{v}) \overset{i.i.d.}{\sim} N(0, \tau^2)$ represents the micro-scale variability and/or the measurement error.

We model $z(\boldsymbol{v})$ with the GNNMP defined in (4.6) with $\mu_l = 0$ and $\sigma_l^2 = \sigma^2$, for all l. For comparison, we also applied an NNGP model for $z(\boldsymbol{v})$ with variance σ_0^2 and exponential correlation function with range paramger ϕ_0. For the GNNMP, we used exponential correlation functions with range parameter ϕ and ζ, respectively, for the correlation with respect to the component density, and the cutoff point kernel. For both models, the regression coefficients $\boldsymbol{\beta}$ were assigned flat priors. The variances σ_0^2 and σ^2 received the same inverse gamma prior $\mathrm{IG}(2, 1)$, and τ^2 was assigned $\mathrm{IG}(2, 0.1)$. The range parameter ϕ_0 of the NNGP received a uniform prior $\mathrm{Unif}(1/30, 1/3)$, while the range parameters ϕ and ζ of the GNNMP received inverse gamma priors $\mathrm{IG}(3, 1/3)$ and $\mathrm{IG}(3, 0.2)$, respectively. Regarding the logit Gaussian distribution parameters, $\boldsymbol{\gamma}$ and κ^2, we used $N((-1.5, 0, 0)^\top, 2\mathbf{I}_3)$ and $\mathrm{IG}(3, 1)$ priors, respectively.

To compare models, we use ten-fold cross-validation. More specifically, we first create ten empty groups, followed by randomly assigning each observation into one of the groups based on a uniform distribution on $\{1, \ldots, 10\}$. For each $k = 1, \ldots, 10$, we use observations in group k as validation set, and the remaining ones to train the model, using neighborhood sizes L from 10 to 20. For each k and each L, we calculate the values of the following metrics: root mean squared prediction error (RMSPE), 95% posterior credible interval coverage rate (95% CI cover), 95% posterior credible interval width (95% CI width), deviance information criterion (DIC; Spiegelhalter et al. 2002), posterior predictive loss criterion (PPLC; Gelfand and Ghosh 1998), and continuous ranked probability score (CRPS; Gneiting and Raftery 2007). Thus, for each L, we can obtain ten different values for each metric. We then calculate the average of the ten values for each metric, and use the averages to compare models.



(a) Regional SST **(b)** Predicted SST (GNNMP) **(c)** Predicted SST (NNGP)

Figure 4.8: Chapter 4 - Mediterranean SST data analysis. Panel (a) shows the observations at the selected region. Panels (b) and (c) are SST posterior median estimates by different models.

Table 4.2: Chapter 4 - Mediterranean SST data analysis. Performance metrics of different models.

	RMSPE	CRPS	95% CI cover	95% CI width	PPLC	DIC
NNGP(13)	1.144	0.620	0.937	4.322	742.845	1492.722
GNNMP(13)	1.117	0.595	0.941	4.132	249.152	778.239
GNNMP(20)	1.113	0.593	0.945	4.178	198.986	582.681

Note: the numbers in the parentheses correspond to the neighborhood sizes.

For the NNGP model, we implemented the latent NNGP algorithm from the spNNGP package in R (Finley et al., 2020). In all cases, we ran the MCMC with 120000 iterations, discarding the first 20000 samples, and collected samples every 20 iterations. The computing time for the NNGP model ranged from around 6 minutes to 20 minutes, while that for the GNNMP model ranged from around 11 minutes to 16 minutes.

We report the results for both models. Both models provided similar posterior estimates of the regression parameters, indicating that there was a trend of SST decreasing in the latitude at the selected region. For the error variance τ^2, overall, the GNNMP provided smaller estimates, compared to the estimates given by the NNGP. For example, when $L = 13$, the average of the posterior means of τ^2 of all

Table 4.3: Chapter 4 - Mediterranean SST data analysis. Ten-fold cross validation results for the NNGP and GNNMP models with different neighborhood sizes.

	RMSPE	CRPS	95% CI cover	95% CI width	PPLC	DIC
NNGP(10)	1.147	0.620	0.941	4.323	741.729	1485.777
GNNMP(10)	1.128	0.603	0.936	4.125	287.105	855.934
NNGP(11)	1.146	0.620	0.938	4.321	744.793	1496.990
GNNMP(11)	1.120	0.598	0.937	4.114	265.694	825.431
NNGP(12)	1.145	0.620	0.940	4.322	738.566	1484.625
GNNMP(12)	1.119	0.597	0.938	4.121	258.192	805.944
NNGP(13)	1.144	0.620	0.937	4.322	742.845	1492.722
GNNMP(13)	1.117	0.595	0.941	4.132	249.152	778.239
NNGP(14)	1.145	0.620	0.937	4.322	746.096	1499.596
GNNMP(14)	1.123	0.599	0.942	4.163	231.989	680.071
NNGP(15)	1.144	0.620	0.935	4.323	747.969	1500.092
GNNMP(15)	1.118	0.597	0.936	4.152	215.208	655.841
NNGP(16)	1.144	0.620	0.935	4.321	751.871	1511.061
GNNMP(16)	1.118	0.596	0.937	4.155	209.155	630.992
NNGP(17)	1.144	0.620	0.938	4.321	749.086	1505.191
GNNMP(17)	1.114	0.594	0.939	4.161	204.148	605.677
NNGP(18)	1.144	0.620	0.940	4.321	751.829	1513.302
GNNMP(18)	1.114	0.594	0.940	4.162	203.136	605.900
NNGP(19)	1.144	0.619	0.937	4.321	754.944	1521.234
GNNMP(19)	1.120	0.596	0.940	4.189	207.918	580.755
NNGP(20)	1.144	0.620	0.938	4.322	752.811	1516.361
GNNMP(20)	1.113	0.593	0.945	4.178	198.986	582.681

Note: the numbers in the parentheses correspond to the neighborhood sizes.

groups was 0.21 from the GNNMP and 0.47 from the NNGP.

Regarding the model performance metrics, the ten-fold cross validation results are shown in Table 4.3. For all neighborhood sizes, both the PPLC and DIC suggest that the GNNMP had a better goodness-of-fit than the NNGP. For out-of-sample prediction, the GNNMP produced smaller RMSPE and CRPS than the NNGP. On average, the GNNMP gave slightly better coverage performance of 95% credible intervals with narrower widths. For each model, we selected the optimal and economical neighborhood size that corresponds to the smallest RMSPE. Such a neighborhood size was $L = 13$ for the NNGP and $L = 20$ for the GNNMP. We compared the selected models, together with the GNNMP with $L = 13$, as shown in Table 4.2. Overall, we can see that the GNNMP outperforms the NNGP in this particular data example.

Finally, we fit the two models with $L = 13$ to all observations over the region. Figure 4.8(b)-4.8(c) show the posterior median estimates of the temperature field from both models. We can see that both models yield estimates that resemble the pattern in the observations. The predictive surface produced by the NNGP depicts some very localized, unrealistic features. These are not present in the results from the GNNMP.

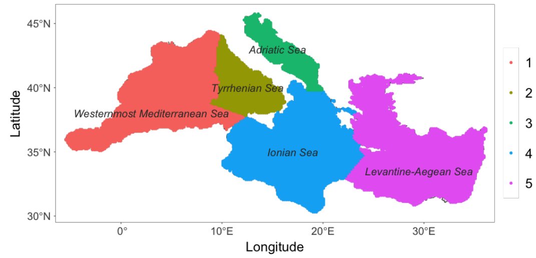
Full Analysis with an extended Skew-GNNMP Model

The regional analysis results suggest that a non-Gaussian process may be a better assumption for the Mediterranean SST spatial variability. In light of the evidence (Pisano et al., 2020) that the spatial patterns of SST are different over Mediterranean sub-basins, as shown in Figure 4.9(a), which are characterized with different dynamics and high variability of surface currents (Bouzaiene et al., 2020), we further investigate the SST over those sub-basins. We fit a non-spatial linear model to all SST data, including longitude and latitude as covariates, and obtained residuals from the linear model. Figure 4.9(b) shows that the histograms of the residuals are asymmetric over the sub-basins, indicating skewness in the marginal SST distribution, with levels of skewness that vary across sub-basins.

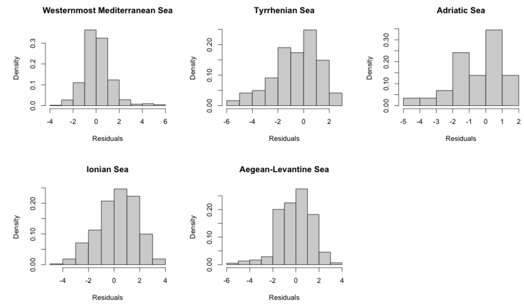
The exploratory data analysis suggests the need for a spatial model that can capture skewness. The symmetric distribution assumption of $\epsilon(\boldsymbol{v})$ in the additive model (4.11) may be inappropriate for modeling skewness. Moreover, the weak identifiability of its variance τ^2 may further undermine estimation of the skewness especially when it is mild. Thus, we analyze the full SST data with an extension of the skew-GNNMP model in (4.7). The new model has two features that extend the skew-GNNMP: (i) it incorporates fixed effect through the location parameter of the mixture component; (ii) it allows the skewness parameter λ to vary in space. More specifically, the spatially varying conditional density $f_{\boldsymbol{v},l}$ builds from a Gaussian random vector with mean $\left(\boldsymbol{x}(\boldsymbol{v})^\top\boldsymbol{\beta} + \lambda(\boldsymbol{v})z_0(\boldsymbol{v}),\, \boldsymbol{x}(\boldsymbol{v}_{(l)})^\top\boldsymbol{\beta} + \lambda(\boldsymbol{v}_{(l)})z_0(\boldsymbol{v})\right)^\top$ and covariance matrix $\sigma^2\left(\begin{smallmatrix} 1 & \rho_l(\boldsymbol{v}) \\ \rho_l(\boldsymbol{v}) & 1 \end{smallmatrix}\right)$, where $\boldsymbol{x}(\boldsymbol{v}) = (1, v_1, v_2)^\top$ and $z_0(\boldsymbol{v}) \sim \mathrm{TN}(0,1)$, for all $\boldsymbol{v}$ and for all l. The associated conditional density $p(y(\boldsymbol{v}) \,|\, \boldsymbol{y}_{\mathrm{Ne}(\boldsymbol{v})})$ of the extended model is

$$\sum_{l=1}^{L} w_l(\boldsymbol{v}) \int_0^\infty N(y(\boldsymbol{v}) \,|\, \mu_l(\boldsymbol{v}), \sigma_l^2(\boldsymbol{v}))\mathrm{TN}(z_0(\boldsymbol{v}) \,|\, \mu_{0l}(\boldsymbol{v}_{(l)}), \sigma_{0l}^2(\boldsymbol{v}_{(l)}))dz_0(\boldsymbol{v}), \quad (4.12)$$

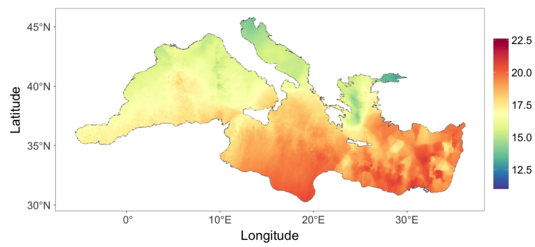
where $\mu_l(\boldsymbol{v}) = \boldsymbol{x}(\boldsymbol{v})^\top\boldsymbol{\beta} + \lambda(\boldsymbol{v})z_0(\boldsymbol{v}) + \rho_l(\boldsymbol{v})\{y(\boldsymbol{v}_{(l)}) - \boldsymbol{x}(\boldsymbol{v}_{(l)})^\top\boldsymbol{\beta} - \lambda(\boldsymbol{v}_{(l)})z_0(\boldsymbol{v})\}$, $\sigma_l^2(\boldsymbol{v}) = \sigma^2\{1-(\rho_l(\boldsymbol{v}))^2\}$, $\mu_{0l}(\boldsymbol{v}_{(l)}) = \{y(\boldsymbol{v}_{(l)}) - \boldsymbol{x}(\boldsymbol{v}_{(l)})^\top\boldsymbol{\beta}\}\lambda(\boldsymbol{v}_{(l)})/\{\sigma^2 + (\lambda(\boldsymbol{v}_{(l)}))^2\}$, and $\sigma_{0l}^2(\boldsymbol{v}_{(l)}) = \sigma^2/\{\sigma^2 + (\lambda(\boldsymbol{v}_{(l)}))^2\}$. After marginalizing out $z_0(\boldsymbol{v})$, we obtain that the marginal distribution of $Y(\boldsymbol{v})$ is $\mathrm{SN}\left(\boldsymbol{x}(\boldsymbol{v})^\top\boldsymbol{\beta}, (\lambda(\boldsymbol{v}))^2 + \sigma^2, \lambda(\boldsymbol{v})/\sigma\right)$, based on the result of Proposition 4.1. We model the spatially varying $\lambda(\boldsymbol{v})$ via a partitioning approach. In particular, we partition the Mediterranean Sea $\mathcal{D}$ according to the sub-basins, i.e., $\mathcal{D} = \cup_{k=1}^{K} P_k$, $P_i \cap P_j = \emptyset$ for $i \neq j$, where $K = 5$. For



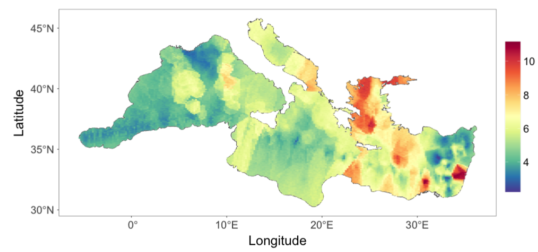
(a) Mediterranean Sea partitions



(b) Histograms of residuals



(c) 50% predicted SST



(d) 95% CI width

Figure 4.9: Chapter 4 - Mediterranean SST data analysis. Panels (a) and (b) are partitions according to Mediterranean sub-basins and histograms of the residuals obtained from a non-spatial linear model. Panels (c) and (d) are posterior median and 95% credible interval estimates of the SST from the extended skew-GNNMP model.

all $\boldsymbol{v} \in P_k$, we take $\lambda(\boldsymbol{v}) = \lambda_k$, for $k = 1, \ldots, K$. The partitions $P_1, \ldots, P_K$ correspond to the sub-basins: Westernmost Mediterranean Sea, Tyrrhenian Sea, Adriatic Sea, Ionian Sea, and Levantine-Aegean Sea, respectively.

We applied the extended skew-GNNMP model (4.12) using the whole data set as reference set with L chosen to be $10, 15$ or 20. The regression parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^\top$ were assigned mean-zero, dispersed normal priors. For the skewness parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_5)$, each element received a $N(0, 5)$ prior. We used the same prior specification for other parameters as in the first simulation experiment. Posterior inference was based on thinned samples retaining every 4th iteration, from a total of 30000 samples with a burn-in of first 10000 samples. The computing time was around 14, 16, and 20 minutes, respectively, for each of the L values.

We focus on the estimation of regression parameters $\boldsymbol{\beta}$ and skewness parameters $\boldsymbol{\lambda}$. We report the estimates for $L = 15$; they were similar for $L = 10$ or 20. The posterior mean and 95% credible interval estimates of β_0, β_1, and β_2 were $30.51\,(28.88, 32.16)$, $0.12\,(0.09, 0.15)$, and $-0.37\,(-0.42, -0.33)$, indicating that there was an increasing trend in SST as longitude increased and latitude decreased. The corresponding posterior estimates of the skewness parameters $\boldsymbol{\lambda}$ were $-0.38\,(-0.94, 0.14)$, $-1.37\,(-2.10, -0.71)$, $-2.44\,(-4.03, -1.14)$, $-1.60\,(-2.54, -0.86)$, and $-2.69\,(-3.95, -1.82)$. These estimates suggest different levels of left skewness over the sub-basins except for the Westernmost Mediterranean Sea.

Figure 4.9(c) provides the posterior median estimate of the SST over a dense grid of locations on the Mediterranean Sea. Compared to Figure 4.7, the estimate overall resembles the observed pattern. The prediction was quite smooth even for areas with few observations. The 95% credible interval width of the SST over the gridded locations, as shown in Figure 4.9(d), demonstrates that the model describes the uncertainty in accordance with the observed data structure; the uncertainty is higher in areas where there are less observations or the observations are volatile.

4.5 Discussion

We have introduced a class of geostatistical models for large, non-Gaussian data sets, based on nearest-neighbor processes. Using an MTD model as the parent process, we have demonstrated the NNMP's flexibility for modeling complex dependence by specification of a collection of bivariate distributions indexed at space. The scope of the methodology has been illustrated through data examples with skewness, heavy tails or compact support.

To incorporate covariates, the NNMP can be embedded in an additive or multiplicative regression model. The former is illustrated in the regional analysis of Section 4.4.4. Under an additive model, the MCMC algorithm requires extra care as it involves sequential updating of the elements in $\boldsymbol{z}_{\mathcal{S}}$. This may induce slow convergence behavior. An alternative strategy for covariate inclusion is to model the weights or some parameter(s) of the spatially varying conditional density as a function of covariates. For example, in the full analysis of Section 4.4.4, we modeled the location parameter of the skew-Gaussian marginal as a linear function of the covariates. Posterior simulation under this approach is easily developed by modifying the update of the relevant parameters discussed in Section 4.3.2 to that of the regression coefficients.

The computation of the NNMP not only bypasses all the potential issues from large matrix operations, but also enhances modeling power. Kernel functions, such as wave covariance functions, that are impractical for Gaussian process-based models due to numerical instability from matrix inversion, can be used as link functions for the spatially varying parameter of the NNMP. One limitation of the NNMP's computation, similar to mixture models, is that the MCMC algorithm may experience slow convergence issues. Further development is needed on efficient algorithms for fast computation, especially when dealing with massive, complex data sets.

It is also interesting to explore the opportunities for the analysis of spatial extremes using the NNMP framework. We developed guidelines in Section 4.2.4 to choose NNMP mixture components based on strength of tail dependence. The results highlight the ability of the NNMP model structure to capture local tail dependence in different levels, controlled by the mixture component bivariate distributions, e.g., with a class of bivariate extreme-value copulas. Moreover, using

NNMPs for spatial extreme modeling allows for efficient computation for implementiation of inference which is typically a challenge with existing approaches (Davison et al., 2012).

Other research directions include extensions to multivariate and spatio-temporal settings. The former extension requires families of high-dimensional multivariate distributions to construct an NNMP. Effective strategies will be needed to define the spatially varying multivariate distributions that balance flexibility and scalability. When it comes to a joint model over time and space, there is large scope for exploring the integration of the time component into the model, including extending the NNMP weights or the NNMP mixture components.

Chapter 5

Models for Discrete-Valued Spatial Processes

5.1 Introduction

Discrete geostatistical data arise in many areas, such as biology, ecology, and forestry. Such data sets consist of observations that take discrete values and are indexed in a continuous spatial domain. As an example, consider observations for counts of a species of interest, commonly used to estimate the species distribution over a geographical domain.

The most common approach to modeling such data is through a spatial generalized linear mixed model (SGLMM, Diggle et al. (1998)), under which an exponential family distribution is specified for the response at a given location, assuming independence between locations, conditional on an underlying spatial process. Such process is specified in the second stage of the SGLMM through a link function that associates the response mean to a set of spatial random effects. A Gaussian process is typically used for the spatial random effects. Thus, SGLMMs provide a general modeling tool for geostatistical discrete data applications; see, for examples, Wikle (2002), Recta et al. (2012), Berrett and Calder (2016), and Zhang et al. (2020).

However, SGLMMs have several properties that can be limited. First, they do not correspond to spatial processes for the observed data. Since the spatial random effects are incorporated into the transformed mean, SGLMMs model spatial structure on a function of the response means, not the observations directly. Thus, the model may impose a strong correlation between means over locations that are close, even though the corresponding observations may not be strongly correlated. In addition, the SGLMM specification poses computational challenges. Unlike Gaussian geostatistical models, the spatial random effects cannot be marginalized out. Under simulation-based inference, estimating the spatial random effects generally requires sampling a large number of highly correlated parameters within an MCMC algorithm, which is likely to produce slow convergence, and a large memory footprint. Although efficient computational strategies have been explored in the literature (e.g., Zhang 2002; Christensen and Waagepetersen 2002; Christensen et al. 2006; Sengupta and Cressie 2013; Sengupta et al. 2016; Guan and Haran 2018), the computational challenge is unavoidable, especially for large spatial datasets.

An alternative to SGLMMs involves Gaussian copula models which construct random fields given a pre-specified family of marginal distributions. Here, the joint cdf of the spatial responses is characterized by a Gaussian copula corresponding to an underlying Gaussian process; see, e.g., Madsen (2009), Kazianka and Pilz (2010), and Han and De Oliveira (2016). Gaussian copulas provide simplicity in specifying spatial dependence, and flexibility in selecting discrete marginal distributions. However, the evaluation of the resulting likelihood requires efficient approximations of high-dimensional multivariate Gaussian integrals, limiting the applicability of this class of models.

The goal of this chapter is threefold. First, we develop a discrete analogue of the NNMP introduced in Chapter 4, referred to as the discrete NNMP, with particular focus on using bivariate copulas to define the spatially varying conditional probability mass functions (pmfs) for the structured mixture that gives rise to the joint distribution. We show that the joint pmf of the discrete copula NNMP can be further decomposed into a collection of bivariate copulas, providing interpretability for model construction using different families of copulas. In fact, our approach allows for the use of arbitrary bivariate copula families, which enhances model flexibility and enables the description of complex spatial dependencies. We demonstrate with a simulation study the impact of using different copula families, exploring alternatives to the traditional Gaussian copula for spatial modeling. Secondly, we extend the first-order strict stationarity result in Chapter 4. The extension is key for discrete NNMPs, providing a constructive approach to develop models with spatially varying marginal pmfs. This can be used, for example, to incorporate either continuous or discrete covariates, which is practically important in the context of regression modeling for discrete-valued spatial responses. Finally, utilizing the stationarity extension result, we develop a Bayesian hierarchical framework that consists of using uniform random variables to transform discrete variables into continuous ones. The proposed approach leverages the properties of copulas for continuous random vectors, thus facilitating the use of different copulas as well as efficient computation. We show through a simulation study that, compared with popular SGLMM methods, this approach yields reliable posterior inference at a much lower computational cost.

The rest of the chapter is organized as follows. In Section 5.2, we introduce

NNMPs for discrete data, with copula-based discrete NNMPs developed in Section 5.3. Section 5.4 presents the Bayesian model formulation for inference, validation and prediction, followed by illustration with synthetic and real datasets in Section 5.5. Finally, Section 5.6 concludes with a summary and discussion.

5.2 NNMPs for Discrete Data

5.2.1 Modeling Framework

Consider a univariate discrete-valued spatial process $Y(\boldsymbol{v})$ indexed by $\boldsymbol{v} \in \mathcal{D} \subset \mathbb{R}^p$, for $p \geq 1$. Let $\boldsymbol{y}_{\mathcal{S}} = (y(\boldsymbol{s}_1), \ldots, y(\boldsymbol{s}_n))^\top$ be a realization of the process $Y(\boldsymbol{v})$, where $\mathcal{S} = (\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n)$ denotes the reference set. We introduce NNMPs for discrete-valued spatial processes, referred to as discrete NNMPs. Similar to the steps in Section 4.2.1, we first build a valid joint pmf over $\mathcal{S}$ with a weighted combination of conditional pmfs:

$$p(y(\boldsymbol{s}_i) \,|\, \boldsymbol{y}_{\mathrm{Ne}(\boldsymbol{s}_i)}) = \sum_{l=1}^{i_L} w_l(\boldsymbol{s}_i)\, f_{\boldsymbol{s}_i,l}(y(\boldsymbol{s}_i) \,|\, y(\boldsymbol{s}_{(il)})), \tag{5.1}$$

where $w_l(\boldsymbol{s}_i) \geq 0$ for every $\boldsymbol{s}_i \in \mathcal{S}$ and for all l, and $\sum_{l=1}^{i_L} w_l(\boldsymbol{s}_i) = 1$.

There are two model elements in (5.1) that describe spatial variability: the mixture component pmfs $f_{\boldsymbol{s}_i,l}$, and the weights $w_l(\boldsymbol{s}_i)$. We defer the specification of the pmfs $f_{\boldsymbol{s}_i,l}$ to the next section. Following Chapter 4, we define the weights as increments of a logit Gaussian cdf $G_{\boldsymbol{s}_i}$, i.e., $w_l(\boldsymbol{s}_i) = G_{\boldsymbol{s}_i}(r_{\boldsymbol{s}_i,l}) - G_{\boldsymbol{s}_i}(r_{\boldsymbol{s}_i,l-1})$, for $l = 1, \ldots, i_L$. Here, $0 = r_{\boldsymbol{s}_i,0} < r_{\boldsymbol{s}_i,1} < \cdots < r_{\boldsymbol{s}_i,i_L-1} < r_{\boldsymbol{s}_i,i_L} = 1$ are random cutoff points such that $r_{\boldsymbol{s}_i,l} - r_{\boldsymbol{s}_i,l-1} = k'(\boldsymbol{s}_i, \boldsymbol{s}_{(il)}) / \sum_{l=1}^{i_L} k'(\boldsymbol{s}_i, \boldsymbol{s}_{(il)})$, for some bounded kernel $k' : \mathcal{D} \times \mathcal{D} \to [0, 1]$. Convenient choices for k' are kernels that compute the correlation between two points. The underlying Gaussian distribution for $G_{\boldsymbol{s}_i}$

has mean $\mu(\boldsymbol{s}_i) = \gamma_0 + \gamma_1 s_{i1} + \gamma_2 s_{i2}$, and variance κ^2, with $\boldsymbol{s}_i = (s_{i1}, s_{i2})$ where s_{i1} and s_{i2} correspond to the $x-$ and $y-$ coordinates of location $\boldsymbol{s}_i$. This formulation allows for spatial dependence among the weights through $\mu(\boldsymbol{s}_i)$. Also, the random cutoff points can flexibly reflect the neighbor structure of $\boldsymbol{s}_i$.

The second step completes the construction of a valid stochastic process over $\mathcal{D}$ by extending (5.1) to an arbitrary finite set of locations outside $\mathcal{S}$, denoted as $\mathcal{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r)$, where $\mathcal{U} \subset \mathcal{D} \setminus \mathcal{S}$. In particular, we define the pmf of $\boldsymbol{y}_{\mathcal{U}}$ conditional on $\boldsymbol{y}_{\mathcal{S}}$ as

$$p(\boldsymbol{y}_{\mathcal{U}} \,|\, \boldsymbol{y}_{\mathcal{S}}) = \prod_{i=1}^{r} p(y(\boldsymbol{u}_i) \,|\, \boldsymbol{y}_{\mathrm{Ne}(\boldsymbol{u}_i)}) = \prod_{i=1}^{r} \sum_{l=1}^{L} w_l(\boldsymbol{u}_i)\, f_{\boldsymbol{u}_i,l}(y(\boldsymbol{u}_i) \,|\, y(\boldsymbol{u}_{(il)})), \qquad (5.2)$$

where the weights and conditional pmfs are defined analogously to Equation (5.1), and the points $(\boldsymbol{u}_{(i1)}, \ldots, \boldsymbol{u}_{(iL)})$ in $\mathrm{Ne}(\boldsymbol{u}_i)$ are the first L locations in $\mathcal{S}$ that are closest to $\boldsymbol{u}_i$.

Given (5.1) and (5.2), a discrete-valued spatial process over $\mathcal{D}$ is well defined. For any finite set $\mathcal{V} \subset \mathcal{D}$ that is not a subset of $\mathcal{S}$, the joint pmf over $\mathcal{V}$ is obtained by marginalizing $p(\boldsymbol{y}_{\mathcal{U}} \,|\, \boldsymbol{y}_{\mathcal{S}})p(\boldsymbol{y}_{\mathcal{S}})$ over $\boldsymbol{y}_{\mathcal{S} \setminus \mathcal{V}}$, where $\mathcal{U} = \mathcal{V} \setminus \mathcal{S}$. Practically, Equations (5.1) and (5.2) serve different purposes. The reference set $\mathcal{S}$ is often reserved for observed data, so model estimation is based on (5.1), while spatial prediction at new locations outside the reference set relies on (5.2). Henceforth, we use

$$p(y(\boldsymbol{v}) \mid \boldsymbol{y}_{\mathrm{Ne}(\boldsymbol{v})}) = \sum_{l=1}^{L} w_l(\boldsymbol{v})\, f_{\boldsymbol{v},l}(y(\boldsymbol{v}) \,|\, y(\boldsymbol{v}_{(l)})) \qquad (5.3)$$

to characterize discrete NNMPs, where $\boldsymbol{v}$ is a generic location in $\mathcal{D}$. The neighbor set $\mathrm{Ne}(\boldsymbol{v})$ contains the first L locations in $\mathcal{S}$ that are closest to $\boldsymbol{v}$. We place these locations in ascending order according to distance, denoted as $\mathrm{Ne}(\boldsymbol{v}) = (\boldsymbol{v}_{(1)}, \ldots, \boldsymbol{v}_{(L)})$.

We note that the discrete NNMP involves selecting the neighborhood size L. Our prior model for the spatially varying weights supports the strategy of using an over-specified L that gives a large neighbor set, with important neighbors assigned large weights a posteriori. For specific data examples, a sensitivity analysis for L can be further carried out to find an optimal L according to standard model comparison metrics or scoring rules. This is illustrated with the real data application in Section 5.5.

The discrete NNMP formulation implies two distinct features that set it apart from SGLMMs. In a SGLMM, responses $y(\boldsymbol{v})$ are conditionally independent with distribution $f(y(\boldsymbol{v}) \,|\, z(\boldsymbol{v}), \boldsymbol{\beta}, r) = a(y(\boldsymbol{v}), r) \exp\left(r\{y(\boldsymbol{v})\eta(\boldsymbol{v}) - \psi(\eta(\boldsymbol{v}))\}\right)$, where $z(\boldsymbol{v})$ is a spatial random effect, $\boldsymbol{\beta}$ are regression parameters, r is a dispersion parameter, and $h(\eta(\boldsymbol{v})) = \boldsymbol{x}(\boldsymbol{v})^\top \boldsymbol{\beta} + z(\boldsymbol{v})$ for some link function h. The joint distribution of observations $(y(\boldsymbol{s}_1), \ldots, y(\boldsymbol{s}_n))$ involves integrating out the spatial random effects, i.e., $\int \{\prod_{i=1}^n f(y(\boldsymbol{s}_i) \,|\, z(\boldsymbol{s}_i), \boldsymbol{\beta}, r)\} p(\boldsymbol{z}_{\mathcal{S}}) d\boldsymbol{z}_{\mathcal{S}}$, where $\boldsymbol{z}_{\mathcal{S}} = (z(\boldsymbol{s}_1), \ldots, z(\boldsymbol{s}_n))^\top$. This restricts the choice of $z(\boldsymbol{v})$ to stochastic processes for which the corresponding joint densities are easy to work with, limiting the range of spatial variability the SGLMM can describe over the domain. In practice, $z(\boldsymbol{v})$ is commonly assumed to be a Gaussian process. This limitation, however, does not affect discrete NNMPs, as the spatial dependence is introduced at the data level. The joint pmf of a discrete NNMP is fully specified through (5.1) and (5.2), which is a finite mixture of generic spatial components that can flexibly capture spatial variability. In addition, the mixture model structure of discrete NNMPs allows for efficient implementation, using inference approaches for mixtures.

5.2.2 Model Construction with Spatially Varying Marginals

The key ingredient in constructing discrete NNMPs lies in the specification of the mixture component conditional pmfs $f_{\boldsymbol{v},l}$. There are many avenues to specify $f_{\boldsymbol{v},l}$. As each conditional pmf corresponds to a bivariate random vector, say $(U_{\boldsymbol{v},l}, V_{\boldsymbol{v},l})$, our strategy is to model $f_{\boldsymbol{v},l}$ through its bivariate pmf, denoted as $f_{U_{\boldsymbol{v},l},V_{\boldsymbol{v},l}}$. Let $f_{U_{\boldsymbol{v},l}}$ and $f_{V_{\boldsymbol{v},l}}$ be the marginal pmfs of $(U_{\boldsymbol{v},l}, V_{\boldsymbol{v},l})$, such that $f_{\boldsymbol{v},l} \equiv f_{U_{\boldsymbol{v},l}|V_{\boldsymbol{v},l}} = f_{U_{\boldsymbol{v},l},V_{\boldsymbol{v},l}}/f_{V_{\boldsymbol{v},l}}$. The benefits of this strategy are twofold. First, it simplifies the multivariate dependence specification by focusing on the bivariate random vectors $(U_{\boldsymbol{v},l}, V_{\boldsymbol{v},l})$. The multivariate dependence will be induced by bivariate distributions through the model's mixture formulation. Second, the strategy allows for the construction of models with a pre-specified family of marginal distributions, facilitating the study of local variability. For example, it is common in discrete geostatistical data modeling to include covariates through the (transformed) mean of the marginal distribution.

The second feature of this strategy relies on an extension of the first-order strict stationarity result from Chapter 4. Based on that result, an NNMP has stationary marginal pmf f_Y if $f_{U_{\boldsymbol{v},l}} = f_{V_{\boldsymbol{v},l}} = f_Y$, for all $\boldsymbol{v}$ and all l. Here, we generalize the result such that discrete NNMPs can be built from pre-specified spatially varying marginal pmfs $g_{\boldsymbol{v}}$, where $g_{\boldsymbol{v}}$ is the marginal pmf of $Y(\boldsymbol{v})$. The generalization of the stationarity proposition applies to all NNMPs. For the interest of this chapter, we summarize the result in the following proposition for discrete NNMPs.

Proposition 5.1. *Consider a discrete NNMP model for spatial process $\{Y(\boldsymbol{v}) : \boldsymbol{v} \in \mathcal{D}\}$, and a collection of spatially varying pmfs $\{g_{\boldsymbol{v}} : \boldsymbol{v} \in \mathcal{D}\}$. If, for each $\boldsymbol{v}$, the marginal pmfs of the mixture component bivariate distributions are such that $f_{U_{\boldsymbol{v},l}} = g_{\boldsymbol{v}}$ and $f_{V_{\boldsymbol{v},l}} = g_{\boldsymbol{v}_{(l)}}$, the discrete NNMP has marginal pmf $g_{\boldsymbol{v}}$ for $Y(\boldsymbol{v})$, for every $\boldsymbol{v} \in \mathcal{D}$.*

A natural example for $\{g_{\boldsymbol{v}} : \boldsymbol{v} \in \mathcal{D}\}$ is a family of distributions with (at least) one of its parameters indexed in space, i.e., $g_{\boldsymbol{v}}(\cdot) = g(\cdot \mid \theta(\boldsymbol{v}), \boldsymbol{\xi})$, in particular, through spatially varying covariates. Using a link function for $\theta(\boldsymbol{v})$, we can include such covariates that provide additional spatially referenced information. A more general example involves partitioning the domain into several regions, where in each region, $g_{\boldsymbol{v}}$ is associated with a different family of marginal distributions. A relevant application is estimation of the abundance of a species that shows overdispersion in most areas, but underdispersion in areas where the species is less prevalent (Wu et al., 2015). Overall, Proposition 5.1 provides flexibility for construction of discrete-valued spatial models with specific marginal pmfs.

We develop next a key component of the methodology, that is, discrete copula NNMP model construction and inference. Given a family of marginal pmfs $g_{\boldsymbol{v}}$, we create spatial copulas for random vectors $(U_{\boldsymbol{v},l}, V_{\boldsymbol{v},l})$. We begin with copulas for a set of base random vectors (U_l, V_l), and extend them to be spatially dependent by modeling the copula parameter that controls the dependence structure as spatially varying. Together with Proposition 5.1, this strategy allows for construction of discrete NNMPs with marginal pmfs in general families.

5.3 Discrete Copula NNMPs

5.3.1 Copula Functions

A bivariate copula function $C : [0,1]^2 \to [0,1]$ is a distribution function whose marginals are uniform distributions on $[0,1]$. Following Sklar (1959), given a random vector (Z_1, Z_2) with joint probability distribution F and marginals F_1 and F_2, there exists a copula function C such that $F(z_1, z_2) = C(F_1(z_1), F_2(z_2))$. If F_1 and F_2 are continuous, C is unique. In this case, the copula density is $c(z_1, z_2) = \partial C(F_1(z_1), F_2(z_2))/(\partial F_1 \partial F_2)$, and the joint density is $f(z_1, z_2) = c(z_1, z_2) f_1(z_1) f_2(z_2)$, where f_1 and f_2 are the densities of F_1 and F_2, respectively.

If both marginals are discrete, the copula C is only unique on the set $\text{Ran}(F_1) \times \text{Ran}(F_2)$, where $\text{Ran}(F_j)$ consists of all possible values of F_j, $j = 1, 2$ (Joe, 2014). Nevertheless, if C is a copula and F_1 and F_2 are discrete distribution functions, then $F(z_1, z_2) = C(F_1(z_1), F_2(z_2))$ is a valid joint distribution; in practice, we select a parametric family for C (Smith and Khaled, 2012). Note that, in contrast with the continuous case, when the marginals are discrete, some popular dependence measures, such as Kendall's τ, will depend on the marginals (Denuit and Lambert, 2005; Genest and Nešlehová, 2007). Consequently, the Kendall's τ of the random vector (Z_1, Z_2) will not be equivalent to the Kendall's τ of the copula. Without loss of generality, hereafter, we assume the bivariate copula carries a single parameter.

5.3.2 Copula NNMPs for Discrete Geostatistical Data

Here, we introduce copula NNMPs with discrete marginals, with focus on using copulas to specify the bivariate distributions of the mixture components. Dropping the dependence on l for clarity, consider a random vector (U, V) with discrete marginal distributions F_U, F_V, and marginal pmfs f_U, f_V. Let $a_u = F_U(u^-)$ and $b_u = F_U(u)$, where $F_U(u^-)$ denotes the left limit of F_U at u. If U is ordinal, $F_U(u^-) = F_U(u-1)$. Analogous definitions of a_v and b_v apply for V. The joint pmf $f_{U,V}$ of (U, V) is obtained by finite differences,

$$f_{U,V}(u, v) = C(b_u, b_v) - C(b_u, a_v) - C(a_u, b_v) + C(a_u, a_v). \tag{5.4}$$

Let $c(u, v) = f_{U,V}(u, v)/(f_U(u) f_V(v))$, such that $f_{U,V}(u, v) = c(u, v) f_U(u) f_V(v)$, using a notation that is analogous to that of the joint density when (U, V) is continuous. Therefore, the conditional pmf, $f_{U|V}(u \,|\, v) = c(u, v) f_U(u)$.

To specify the distribution of base random vector (U_l, V_l), we use copula C_l with parameter η_l. For a parsimonious location-dependent model, we create spatially varying copulas $C_{\boldsymbol{v},l}$ on $(U_{\boldsymbol{v},l}, V_{\boldsymbol{v},l})$ by extending η_l to $\eta_l(\boldsymbol{v})$. In practice, we associate $\eta_l(\boldsymbol{v})$ to a spatial kernel that depends on $\boldsymbol{v} \in \mathcal{D}$ through a link function. Using Proposition 5.1 with a family of marginal pmfs $g_{\boldsymbol{v}}$, the joint pmf on $(U_{\boldsymbol{v},l}, V_{\boldsymbol{v},l})$ is $f_{U_{\boldsymbol{v},l},V_{\boldsymbol{v},l}}(u, v) = c_{\boldsymbol{v},l}(u, v) f_{U_{\boldsymbol{v},l}}(u) f_{V_{\boldsymbol{v},l}}(v)$, where $f_{U_{\boldsymbol{v},l}} = g_{\boldsymbol{v}}$ and $f_{V_{\boldsymbol{v},l}} = g_{\boldsymbol{v}_{(l)}}$, and the conditional pmf is $f_{\boldsymbol{v},l}(u \,|\, v) = c_{\boldsymbol{v},l}(u, v) g_{\boldsymbol{v}}(u)$. Finally, the conditional pmf of the discrete copula NNMP model is given by

$$p(y(\boldsymbol{v}) \mid \boldsymbol{y}_{\mathrm{Ne}(\boldsymbol{v})}) = \sum_{l=1}^{L} w_l(\boldsymbol{v})\, c_{\boldsymbol{v},l}(y(\boldsymbol{v}), y(\boldsymbol{v}_{(l)}))\, g_{\boldsymbol{v}}(y(\boldsymbol{v})), \tag{5.5}$$

where the marginal pmf for $Y(\boldsymbol{v})$ is $g_{\boldsymbol{v}}$.

Recall that an NNMP model involves two sets of locations, the reference set $\mathcal{S}$ and nonreference set $\mathcal{U}$. As done in practice, we take the reference set $\mathcal{S}$ to correspond to the observed locations, and consider a generic finite set $\mathcal{U}$ such that $\mathcal{S} \cap \mathcal{U} = \emptyset$. Then, the joint pmf $p(\boldsymbol{y}_{\mathcal{V}})$ over set $\mathcal{V} = \mathcal{S} \cup \mathcal{U}$ describes the NNMP distribution over any finite set of locations that includes the observed locations. In general, for a discrete NNMP, an explicit expression for $p(\boldsymbol{y}_{\mathcal{V}})$ is not available, since working with a bivariate discrete distribution and its conditional pmf is difficult. However, using copulas to specify the bivariate mixture component yields a structured conditional pmf and allows for the study of the joint pmf. The following proposition provides an explicit expression for $p(\boldsymbol{y}_{\mathcal{V}})$ under a discrete copula NNMP.

Proposition 5.2. *Consider a discrete copula NNMP model for spatial process $\{Y(\boldsymbol{v}) : \boldsymbol{v} \in \mathcal{D}\}$, with $\mathcal{S} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$ and $\mathcal{U} = \{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m\}$, where $n \geq 2$, $m \geq 1$, and $\mathcal{S} \cap \mathcal{U} = \emptyset$. Take $\mathcal{V} = \mathcal{S} \cup \mathcal{U}$ and $\boldsymbol{y}_{\mathcal{V}} = (y(\boldsymbol{s}_1), \ldots, y(\boldsymbol{s}_n), y(\boldsymbol{u}_1), \ldots, y(\boldsymbol{u}_m))^\top$.*

Then the joint pmf of $\boldsymbol{y}_{\mathcal{V}}$ *is* $p(\boldsymbol{y}_{\mathcal{V}}) = p(\boldsymbol{y}_{\mathcal{U}} \,|\, \boldsymbol{y}_{\mathcal{S}}) p(\boldsymbol{y}_{\mathcal{S}})$*, where*

$$\begin{aligned} p(\boldsymbol{y}_{\mathcal{S}}) &= \prod_{i=1}^{n} g_{\boldsymbol{s}_i}(y(\boldsymbol{s}_i)) \sum_{l_n=1}^{n_L} \cdots \sum_{l_2=1}^{2_L} w_{\boldsymbol{s}_n,l_n} \ldots w_{\boldsymbol{s}_2,l_2} c_{\boldsymbol{s}_n,l_n} \ldots c_{\boldsymbol{s}_2,l_2}, \\ p(\boldsymbol{y}_{\mathcal{U}} \,|\, \boldsymbol{y}_{\mathcal{S}}) &= \prod_{i=1}^{m} g_{\boldsymbol{u}_i}(y(\boldsymbol{u}_i)) \sum_{l_m=1}^{L} \cdots \sum_{l_1=1}^{L} w_{\boldsymbol{u}_m,l_m} \ldots w_{\boldsymbol{u}_1,l_1} c_{\boldsymbol{u}_m,l_m} \ldots c_{\boldsymbol{u}_1,l_1}. \end{aligned} \tag{5.6}$$

where $w_{\boldsymbol{s}_i,l_i} \equiv w_{l_i}(\boldsymbol{s}_i)$ *and* $c_{\boldsymbol{s}_i,l_i} \equiv c_{\boldsymbol{s}_i,l_i}(y(\boldsymbol{s}_i), y(\boldsymbol{s}_{(i,l_i)}))$*, for* $l_i = 1, \ldots, i_L$*,* $i = 2, \ldots, n$*, and* $w_{\boldsymbol{u}_i,l_i} \equiv w_{l_i}(\boldsymbol{u}_i)$ *and* $c_{\boldsymbol{u}_i,l_i} \equiv c_{\boldsymbol{u}_i,l_i}(y(\boldsymbol{u}_i), y(\boldsymbol{u}_{(i,l_i)}))$*, for* $l_i = 1, \ldots, L$*,* $i = 1, \ldots, m$*.*

We note that Proposition 5.2 also applies when $\boldsymbol{y}_{\mathcal{V}}$ is continuous. It indicates that, given the sequence of pmfs $g_{\boldsymbol{v}}$, the joint pmf of $\boldsymbol{y}_{\mathcal{V}}$ is determined by the collection of bivariate copulas, motivating the use of different copula families to construct discrete NNMPs. To balance flexibility and scalability, our strategy is to take all copulas C_l in one family with the same link function for the copula parameters. Table 5.1 presents three examples with copula parameters modeled via a link function $k : \mathcal{D} \times \mathcal{D} \to [0, 1]$. In particular, the Gumbel and Clayton copulas are asymmetric. They exhibit greater dependence in the positive and negative tails, respectively. In the first simulation example, we demonstrate that when the underlying spatial dependence is non-Gaussian, it may be appropriate to choose asymmetric copulas. We present next an example of a discrete copula NNMP construction.

Example 1. *Gaussian copula NNMP with negative binomial marginals.* For the family of marginal pmfs $g_{\boldsymbol{v}}$, consider the negative binomial distribution with mean $\alpha(\boldsymbol{v})$ and dispersion parameter r, denoted as $\mathrm{NB}(\alpha(\boldsymbol{v}), r)$. Therefore, $g_{\boldsymbol{v}}(y) = \binom{y+r-1}{y} (p(\boldsymbol{v}))^r (1 - p(\boldsymbol{v}))^y$, with $p(\boldsymbol{v}) = r/(\alpha(\boldsymbol{v}) + r)$. To include a vector of covariates $\boldsymbol{x}(\boldsymbol{v})$, we take a log-link function for $\alpha(\boldsymbol{v})$ such that $\log(\alpha(\boldsymbol{v})) = \boldsymbol{x}(\boldsymbol{v})^\top \boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is a vector of regression parameters. We first specify Gaussian copulas

Table 5.1: Chapter 5 - spatial copula functions. Examples of spatial copulas $C_{\boldsymbol{v},l}$ and corresponding link functions, $k : \mathcal{D} \times \mathcal{D} \to [0, 1]$.

	$C_{\boldsymbol{v},l}(z_1, z_2)$	link function
Gaussian	$\Phi_2(\Phi^{-1}(z_1),\, \Phi^{-1}(z_2))$	$\rho_l(\boldsymbol{v}) = k(\boldsymbol{v}, \boldsymbol{v}_{(l)})$
Gumbel	$\exp\{-[(-\log z_1)^{\eta_l(\boldsymbol{v})} + (-\log z_2)^{\eta_l(\boldsymbol{v})}]^{1/\eta_l(\boldsymbol{v})}\}$	$\eta_l(\boldsymbol{v}) = (1 - k(\boldsymbol{v}, \boldsymbol{v}_{(l)}))^{-1}$
Clayton	$(z_1^{-\delta_l(\boldsymbol{v})} + z_2^{-\delta_l(\boldsymbol{v})} - 1)^{-1/\delta_l(\boldsymbol{v})}$	$\delta_l(\boldsymbol{v}) = 2k(\boldsymbol{v}, \boldsymbol{v}_{(l)})/(1 - k(\boldsymbol{v}, \boldsymbol{v}_{(l)}))$

Note: the bivariate cdf Φ_2 corresponds to the standard bivariate Gaussian distribution with correlation $\rho \in (0, 1)$, and the cdf Φ corresponds to the standard univariate Gaussian distribution.

C_l with correlation parameters ρ_l for the base random vectors (U_l, V_l). We then modify the correlation parameters ρ_l using a correlation function k for all l such that $\rho_l(\boldsymbol{v}) := k(\boldsymbol{v}, \boldsymbol{v}_{(l)})$, creating a sequence of spatially varying copulas $C_{\boldsymbol{v},l}$. The resulting model is given by (5.5) with $g_{\boldsymbol{v}} = \text{NB}(\alpha(\boldsymbol{v}), r)$.

5.3.3 Inference for Discrete Copula NNMPs

A traditional copula model for an n-variate discrete-valued vector involves evaluating 2^n terms of n-dimensional copulas. Unless n is very small, the computation is infeasible. Notable exceptions are discrete vine copula models (Panagiotelis et al., 2012) that decompose a multivariate pmf into bivariate copulas and marginals under a set of trees. The computations for likelihood evaluations grow quadratically in n. Discrete copula NNMPs compare favorably with discrete vine models, as the structured mixture formulation results in only $4nL$ bivariate copula function evaluations for the likelihood, providing linear growth in n.

Here, we develop a framework for discrete copula NNMP inference, based on transforming the discrete random variables to continuous ones by adding auxiliary variables, using the continuous extension (CE) approach in Denuit and Lambert (2005). Working with continuous marginals improves computational efficiency and stability: the likelihood requires only nL bivariate copula density evaluations;

and, computing the conditional pmf using the finite differences in (5.4) is bypassed, thus avoiding numerical instability especially for copulas that are not analytically available, such as the Gaussian copula. Moreover, this framework makes more efficiency the key task of spatial prediction over unobserved sites by avoiding computation that involves inverting the conditional cdf based on (5.4).

We associate each $Y(\boldsymbol{v})$ with a continuous random variable $Y^*(\boldsymbol{v})$, such that $Y^*(\boldsymbol{v}) = Y(\boldsymbol{v}) - O(\boldsymbol{v})$, where $O(\boldsymbol{v})$ is a continuous uniform random variable on $(0, 1)$, independent of $Y(\boldsymbol{v})$ and of $O(\boldsymbol{v}')$, for $\boldsymbol{v}' \neq \boldsymbol{v}$. We refer to $Y^*(\boldsymbol{v})$ as the continued $Y(\boldsymbol{v})$ by $O(\boldsymbol{v})$. Let $Q_{\boldsymbol{v}}$ and $g_{\boldsymbol{v}}$ be the marginal cdf and pmf of $Y(\boldsymbol{v})$, respectively. Then, the marginal cdf and density of $Y^*(\boldsymbol{v})$ are $Q^*_{\boldsymbol{v}}(y^*(\boldsymbol{v})) = Q_{\boldsymbol{v}}([y^*(\boldsymbol{v})]) + (y^*(\boldsymbol{v}) - [y^*(\boldsymbol{v})])g_{\boldsymbol{v}}([y^*(\boldsymbol{v}) + 1])$, and $g^*_{\boldsymbol{v}}(y^*(\boldsymbol{v})) = g_{\boldsymbol{v}}([y^*(\boldsymbol{v}) + 1])$, respectively, where $[x]$ denotes the integer part of x.

Based on marginal densities $g^*_{\boldsymbol{v}}$, we take spatial copulas $C^*_{\boldsymbol{v},l} = C_{\boldsymbol{v},l}$ for continuous random vectors $(U^*_{\boldsymbol{v},l}, V^*_{\boldsymbol{v},l})$, with marginals $f_{U^*_{\boldsymbol{v},l}} = g^*_{\boldsymbol{v}}$ and $f_{V^*_{\boldsymbol{v},l}} = g^*_{\boldsymbol{v}_{(l)}}$, using copulas $C_{\boldsymbol{v},l}$ from the original NNMP model. The joint density on $(U^*_{\boldsymbol{v},l}, V^*_{\boldsymbol{v},l})$ is $f_{U^*_{\boldsymbol{v},l},V^*_{\boldsymbol{v},l}}(u, v) = c^*_{\boldsymbol{v},l}(u, v)g^*_{\boldsymbol{v}}(u)g^*_{\boldsymbol{v}_{(l)}}(v)$, and the conditional density is $f^*_{\boldsymbol{v},l}(u \mid v) = c^*_{\boldsymbol{v},l}(u, v)g^*_{\boldsymbol{v}}(u)$, where $c^*_{\boldsymbol{v},l}$ is the copula density. Denote by $\boldsymbol{y}^*_{\mathrm{Ne}(\boldsymbol{v})}$ the vector that contains the continued elements of $\boldsymbol{y}_{\mathrm{Ne}(\boldsymbol{v})}$, and $\boldsymbol{o}_{\mathrm{Ne}(\boldsymbol{v})}$ the vector of auxiliary variables for elements of $\boldsymbol{y}_{\mathrm{Ne}(\boldsymbol{v})}$. Then, the implied model on $y^*(\boldsymbol{v})$ is

$$p(y^*(\boldsymbol{v}) \mid D^*(\boldsymbol{v})) = \sum_{l=1}^{L} w_l(\boldsymbol{v})\, c^*_{\boldsymbol{v},l}(y^*(\boldsymbol{v}), y^*(\boldsymbol{v}_{(l)}))\, g^*_{\boldsymbol{v}}(y^*(\boldsymbol{v})) \tag{5.7}$$

where $y^*(\boldsymbol{v}) = y(\boldsymbol{v}) - o(\boldsymbol{v})$, and $D^*(\boldsymbol{v}) = \{\boldsymbol{y}^*_{\mathrm{Ne}(\boldsymbol{v})}, o(\boldsymbol{v}), \boldsymbol{o}_{\mathrm{Ne}(\boldsymbol{v})}\}$. Based on Proposition 5.1, model (5.7) has marginal density $g^*_{\boldsymbol{v}}$ for $Y^*(\boldsymbol{v})$. To recover $y(\boldsymbol{v})$, we first generate $y^*(\boldsymbol{v})$ from the extended model, and then set $y(\boldsymbol{v}) = [y^*(\boldsymbol{v}) + 1]$.

Regarding the existing literature, statistical inference for spatial copula models based on the CE approach is typically conducted by maximizing the expected likelihood with respect to the auxiliary variables (Madsen, 2009; Hughes, 2015). We develop inferential methods under the Bayesian framework. Posterior simulation based on (5.7) takes advantage of copula properties for continuous random variables, thus providing efficient computation for both model estimation and prediction.

5.4 Bayesian Implementation

5.4.1 Hierarchical Model Formulation

Assume that $\boldsymbol{y}_{\mathcal{S}} = (y(\boldsymbol{s}_1), \ldots, y(\boldsymbol{s}_n))^\top$ is a realization of a discrete copula NNMP with spatially varying marginal pmfs through spatially dependent covariates, $g_{\boldsymbol{s}_i}(y(\boldsymbol{s}_i)) \equiv g(y(\boldsymbol{s}_i) \,|\, \boldsymbol{\beta}, \boldsymbol{\xi})$. Here, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$, where β_0 is an intercept and $(\beta_1, \ldots, \beta_p)^\top$ is the regression parameter vector for covariates $\boldsymbol{x}(\boldsymbol{s}_i)$, and $\boldsymbol{\xi}$ collects all other parameters of g. The copula parameter is modeled through a link function k with parameter(s) $\boldsymbol{\phi}$. We use the CE approach associating each $y(\boldsymbol{s}_i)$ with $y^*(\boldsymbol{s}_i)$, such that $y^*(\boldsymbol{s}_i) = y(\boldsymbol{s}_i) - o_i$, where $o_i \equiv o(\boldsymbol{s}_i)$ is uniformly distributed on $(0, 1)$, independent of $y(\boldsymbol{s}_i)$ and of o_j, for $j \neq i$. Moreover, denote by $\boldsymbol{\zeta}$ the parameter of the cutoff point kernel for the mixture weights, defined in Section 5.2.1.

The formulation of the mixture weights allows us to augment the model with a sequence of auxiliary variables, $\{t_i : i = 3, \ldots, n\}$, where t_i is a Gaussian random variable with mean $\mu(\boldsymbol{s}_i)$ and variance κ^2. The augmented model for the data can be expressed as

$$
\begin{aligned}
&y(\boldsymbol{s}_i) = y^*(\boldsymbol{s}_i) + o_i, \;\; o_i \overset{i.i.d.}{\sim} \text{Unif}(0,1), \; i = 1, \dots, n, \\
&y^*(\boldsymbol{s}_1) \mid \boldsymbol{\beta}, \boldsymbol{\xi} \sim g^*_{\boldsymbol{s}_1}(y^*(\boldsymbol{s}_1)), \;\; y^*(\boldsymbol{s}_2) \mid y^*(\boldsymbol{s}_1), \boldsymbol{\phi}, \boldsymbol{\beta}, \boldsymbol{\xi} \sim f^*_{\boldsymbol{s}_2,1}(y^*(\boldsymbol{s}_2) \,|\, y^*(\boldsymbol{s}_1)), \\
&y^*(\boldsymbol{s}_i) \mid \{y^*(\boldsymbol{s}_{(il)})\}_{l=1}^{i_L}, t_i, \boldsymbol{\phi}, \boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\zeta} \overset{ind.}{\sim} \sum_{l=1}^{i_L} f^*_{\boldsymbol{s}_i,l}(y^*(\boldsymbol{s}_i) \,|\, y^*(\boldsymbol{s}_{(il)})) \, \mathbb{1}_{(r^*_{\boldsymbol{s}_i,l-1}, r^*_{\boldsymbol{s}_i,l})}(t_i), \\
&t_i \mid \boldsymbol{\gamma}, \kappa^2 \overset{ind.}{\sim} N(t_i \mid \gamma_0 + \gamma_1 s_{i1} + \gamma_2 s_{i2}, \kappa^2), \; i = 3, \dots, n,
\end{aligned}
$$

where $r^*_{\boldsymbol{s}_i,l} = \log\{r_{\boldsymbol{s}_i,l}/(1 - r_{\boldsymbol{s}_i,l})\}$, and $f^*_{\boldsymbol{s}_i,l}(y^*(\boldsymbol{s}_i) \,|\, y^*(\boldsymbol{s}_{(il)})) = c^*_{\boldsymbol{s}_i,l}(y^*(\boldsymbol{s}_i), y^*(\boldsymbol{s}_{(il)}))$ $g^*_{\boldsymbol{s}_i}(y^*(\boldsymbol{s}_i))$, for $l = 1, \dots, i_L$. The full Bayesian model is completed with prior specification for parameters $\boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\phi}, \boldsymbol{\zeta}, \boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2)^\top$ and κ^2. The priors for $\boldsymbol{\xi}$, $\boldsymbol{\phi}$, and $\boldsymbol{\zeta}$ depend on the choices of the pmf $g_{\boldsymbol{s}_i}$, the copula $C^*_{\boldsymbol{s}_i,l}$, and the kernel k', respectively. For parameters $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and κ^2, we consider $N(\boldsymbol{\beta} \,|\, \mu_\beta, \boldsymbol{V}_\beta)$, $N(\boldsymbol{\gamma} \,|\, \mu_\gamma, \boldsymbol{V}_\gamma)$, and $\text{IG}(\kappa^2 \,|\, u_{\kappa^2}, v_{\kappa^2})$ priors, where IG denotes the inverse gamma distribution.

5.4.2 Model Estimation, Validation, and Prediction

We outline the MCMC sampler for parameters $(\boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\phi}, \boldsymbol{\zeta}, \boldsymbol{\gamma}, \kappa^2)$, and latent variables $\{t_i\}_{i=3}^n$ and $\{o_i\}_{i=1}^n$. We note that there is a set of configuration variables $\{\ell_i\}_{i=3}^n$ in one-to-one correspondence with t_i, i.e., $\ell_i = l$ if and only if $t_i \in (r^*_{\boldsymbol{s}_i,l-1}, r^*_{\boldsymbol{s}_i,l})$, for $l = 1, \dots, i_L$.

The updates for parameters $\boldsymbol{\beta}$, $\boldsymbol{\xi}$ and $\boldsymbol{\phi}$ require Metropolis steps, since they enter in copula densities $c^*_{\boldsymbol{s}_i,l}$. We use a Metropolis step also for kernel k' parameter $\boldsymbol{\zeta}$, which is involved in the definition of the mixture weights. Let $\boldsymbol{D}$ be the $(n-2) \times 3$ matrix with ith row $(1, s_{i+2,1}, s_{i+2,2})$. The posterior full conditional of $\boldsymbol{\gamma}$ is $N(\boldsymbol{\gamma} \,|\, \boldsymbol{\mu}^*_\gamma, \boldsymbol{V}^*_\gamma)$, where $\boldsymbol{V}^*_\gamma = (\boldsymbol{V}_\gamma^{-1} + \kappa^{-2} \boldsymbol{D}^\top \boldsymbol{D})^{-1}$ and $\boldsymbol{\mu}^*_\gamma = \boldsymbol{V}^*_\gamma (\boldsymbol{V}_\gamma^{-1} \boldsymbol{\mu}_\gamma + \kappa^{-2} \boldsymbol{D}^\top \boldsymbol{t})$, with the vector $\boldsymbol{t} = (t_3, \dots, t_n)^\top$. The posterior full conditional distribution of κ^2

is $\mathrm{IG}(\kappa^2 \mid u_{\kappa^2} + (n-2)/2, v_{\kappa^2} + \sum_{i=3}^{n}(t_i - \mu(\boldsymbol{s}_i))^2/2)$.

The posterior full conditional distribution for each latent variable t_i, $i = 3, \dots, n$, can be expressed as $\sum_{l=1}^{i_L} q_l(\boldsymbol{s}_i)\,\mathrm{TN}(t_i \mid \mu(\boldsymbol{s}_i), \kappa^2; r^*_{\boldsymbol{s}_i,l-1} < t_i \leq r^*_{\boldsymbol{s}_i,l})$, where TN denotes the truncated normal distribution over the indicated interval, and $q_l(\boldsymbol{s}_i) \propto w_l(\boldsymbol{s}_i)\, c^*_{\boldsymbol{s}_i,l}(y^*(\boldsymbol{s}_i), y^*(\boldsymbol{s}_{(il)}))$, for $l = 1, ..., i_L$. Hence, each t_i can be readily updated by sampling from the l-th truncated normal with probability proportional to $q_l(\boldsymbol{s}_i)$. For auxiliary variables o_i, the posterior full conditional of o_1 is proportional to $\prod_{\{j:\boldsymbol{s}_{(j,\ell_j)}=\boldsymbol{s}_1\}} c^*_{\boldsymbol{s}_j,\ell_j}(y(\boldsymbol{s}_j) - o_j, y(\boldsymbol{s}_1) - o_1)$, and that of o_i, $i \geq 2$, is proportional to $c^*_{\boldsymbol{s}_i,\ell_i}(y(\boldsymbol{s}_i) - o_i, y(\boldsymbol{s}_{(i,\ell_i)}) - o_{(i,\ell_i)}) \prod_{\{j:\boldsymbol{s}_{(j,\ell_j)}=\boldsymbol{s}_i\}} c^*_{\boldsymbol{s}_j,\ell_j}(y(\boldsymbol{s}_j) - o_j, y(\boldsymbol{s}_i) - o_i)$, where $\ell_2 = 1$ and $o_{(i,\ell_i)} \equiv o(\boldsymbol{s}_{(i,\ell_i)})$. We update each latent variable o_i with an independent Metropolis step with a $\mathrm{Unif}(0,1)$ proposal distribution.

The continued model likelihood has the form $g_{\boldsymbol{s}_1}(y^*(\boldsymbol{s}_1)) \prod_{i=2}^{n} p(y^*(\boldsymbol{s}_i) \mid D^*(\boldsymbol{s}_i))$. The product formulation allows for model validation, using a generalization of the randomized quantile residuals (Dunn and Smyth, 1996) for independent data. Specifically, we define the marginal quantile residual, $r_1 = \Phi^{-1}(Q^*_{\boldsymbol{s}_1}(y^*(\boldsymbol{s}_1)))$, and the ith conditional quantile residual, $r_i = \Phi^{-1}(F(y^*(\boldsymbol{s}_i) \mid D^*(\boldsymbol{s}_i)))$, $i = 2, \dots, n$, where F is the conditional cdf of $y^*(\boldsymbol{s}_i)$. If the model is correctly specified, the residuals r_i, $i = 1, \dots, n$, would be independent and identically distributed as a standard Gaussian distribution.

Finally, we turn to posterior predictive inference at a new location $\boldsymbol{v}_0$. If $\boldsymbol{v}_0 \notin \mathcal{S}$, for each posterior sample, we first compute the cutoff points $r_{\boldsymbol{v}_0,l}$, such that $r_{\boldsymbol{v}_0,l} - r_{\boldsymbol{v}_0,l-1} = k'(\boldsymbol{v}_0, \boldsymbol{v}_{(0l)}) / \sum_{l=1}^{L} k'(\boldsymbol{v}_0, \boldsymbol{v}_{(0l)})$, and the weights $w_l(\boldsymbol{v}_0) = G_{\boldsymbol{v}_0}(r_{\boldsymbol{v}_0,l}) - G_{\boldsymbol{v}_0}(r_{\boldsymbol{v}_0,l-1})$, for $l = 1, \dots, L$. We then generate $y^*(\boldsymbol{v}_0)$ based on (5.7), and set $y(\boldsymbol{v}_0) = [y^*(\boldsymbol{v}_0) + 1]$. If $\boldsymbol{v}_0 \equiv \boldsymbol{s}_i \in \mathcal{S}$, we generate $y(\boldsymbol{v}_0)$ similarly, the difference being that we now use the posterior samples for the mixture weights obtained from the MCMC algorithm.

5.5 Data Illustrations

To illustrate the proposed methodology, we present two synthetic data examples and a real data analysis. The goal of the first simulation experiment is to investigate the flexibility of discrete copula NNMPs, using different copula functions to define the NNMP mixture components. In the second experiment, we demonstrate the inferential and computational advantages of our approach for count data modeling, compared to SGLMMs. Implementation details for the models are provided in the Appendix. Since our purpose is primarily demonstrative, we took $L = 10$ for the simulation experiments. A comprehensive sensitivity analysis for L was conducted for the real data application.

In both simulated data examples, we ran the MCMC algorithm for each copula NNMP model for 20000 iterations, discarding the first 4000 iterations, and collecting posterior samples every four iterations. The SGLMM models were implemented using the spBayes package in R (Finley et al., 2007); we ran the algorithm for 40000 iterations and collected posterior samples every five iterations, with the first 20000 as burn-in.

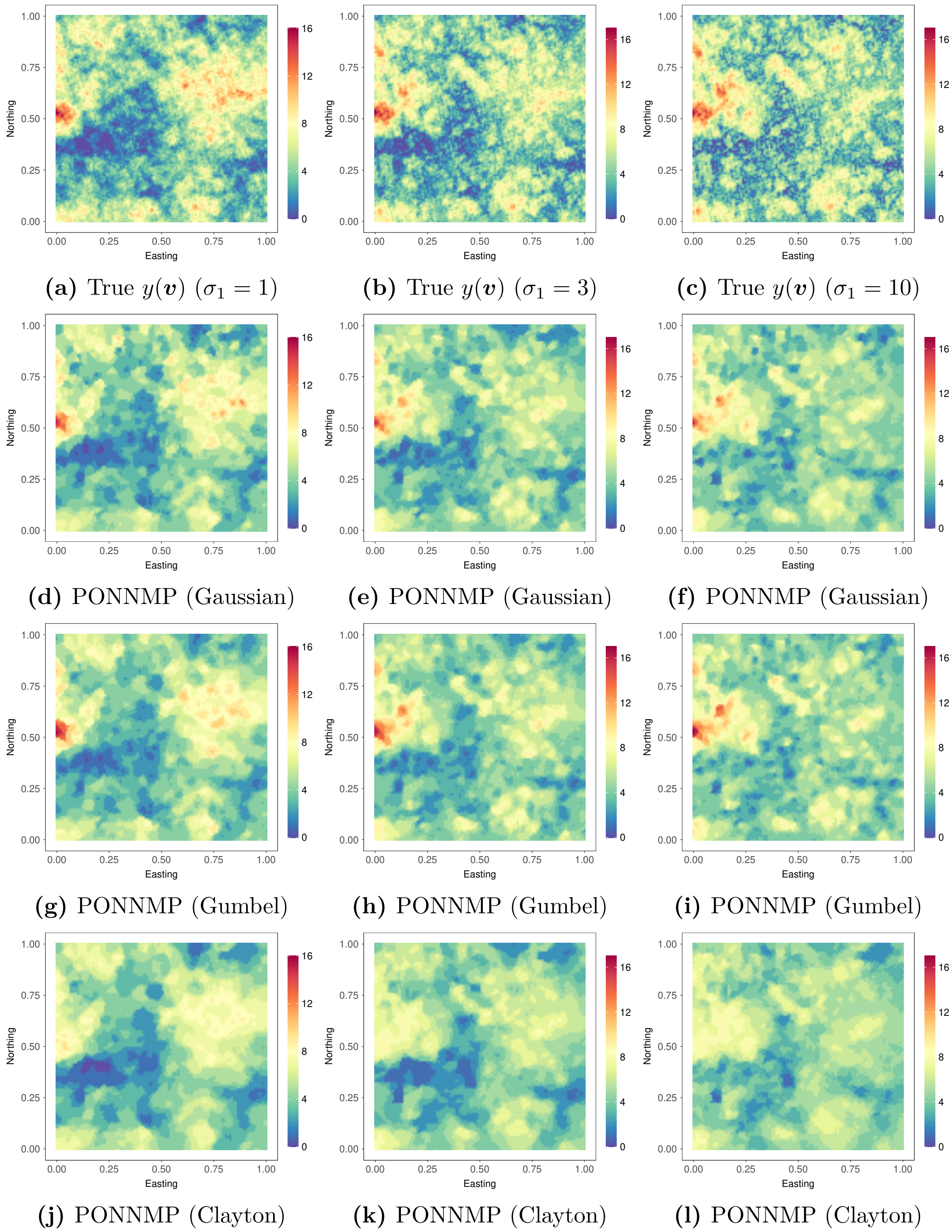
We compare models based on parameter estimates, root mean squared prediction error (RMSPE), 95% credible interval width (95% CI width), 95% credible interval coverage rate (95% CI cover), continuous ranked probability score (CRPS; Gneiting and Raftery 2007), energy score (ES; Gneiting and Raftery 2007), and variogram score of order one (VS; Scheuerer and Hamill 2015). The energy score is a multivariate extension of the CRPS, while the variogram score examines pairwise differences of the components of the multivariate quantity. Both the ES and VS allow for comparison of model predictive performance with respect to dependence structure.

5.5.1 First Simulation Experiment

We first generated sites over a regular grid of 120×120 resolution on a unit square domain, and then simulated data from $y(\boldsymbol{v}) = F_Y^{-1}\Big(F_Z(z(\boldsymbol{v}))\Big)$, where F_Y corresponds to the Poisson distribution with rate parameter $\lambda_0 = 5$, and $z(\boldsymbol{v})$ is the skew-Gaussian random field from Zhang and El-Shaarawi (2010) with stationary marginal distribution F_Z. More specifically, $z(\boldsymbol{v}) = \sigma_1 \, |\omega_1(\boldsymbol{v})| + \sigma_2 \, \omega_2(\boldsymbol{v})$, where both $\omega_1(\boldsymbol{v})$ and $\omega_2(\boldsymbol{v})$ are standard Gaussian processes with exponential correlation function based on range parameter 0.1. The density of F_Z is $f_Z(z) = 2 \, N(z \,|\, 0, \sigma_1^2 + \sigma_2^2) \, \Phi(\sigma_1 z / (\sigma_2 \sqrt{\sigma_1^2 + \sigma_2^2}))$, where $\sigma_1 \in \mathbb{R}$ controls the skewness, and $\sigma_2 > 0$ is a scale parameter. We took $\sigma_2 = 1$, and $\sigma_1 = 1, 3, 10$, which corresponds to positive weak, moderate, and strong skewness.

We considered three discrete copula NNMPs with stationary Poisson marginals, i.e., $g_{\boldsymbol{v}} = f_Y$, for all $\boldsymbol{v}$, where f_Y is the Poisson pmf with rate λ. The three models correspond to the copulas in Table 5.1, with the link function k given by an exponential correlation function with range parameter denoted by ϕ_1, ϕ_2, and ϕ_3 for the Gaussian, Gumbel, and Clayton copula models, respectively. We specified the cutoff point kernel through an exponential correlation function with range parameter ζ_1, ζ_2, and ζ_3 for the Gaussian, Gumbel, and Clayton copula models, respectively. The Bayesian models are fully specified with an $\mathrm{IG}(3, 1)$ prior for the ϕ and ζ parameters, and with $N(\boldsymbol{\gamma} \,|\, (-1.5, 0, 0)^\top, \, 2\mathbf{I}_3)$ and $\mathrm{IG}(\kappa^2 \,|\, 3, 1)$ priors. Finally, the prior for the rate parameter λ was taken as $\mathrm{Ga}(1, 1)$, where $\mathrm{Ga}(a, b)$ denotes the gamma distribution with mean a/b. We simulated 1000 responses and used 800 of them to fit the three NNMP models. The remaining 200 observations were used for model comparison.

Table 5.2 provides estimates for the rate parameter λ of the Poisson marginal distribution, and predictive performance metrics. For all three cases for $\sigma_1 =$



(a) True $y(\boldsymbol{v})$ $(\sigma_1 = 1)$ (b) True $y(\boldsymbol{v})$ $(\sigma_1 = 3)$ (c) True $y(\boldsymbol{v})$ $(\sigma_1 = 10)$

(d) PONNMP (Gaussian) (e) PONNMP (Gaussian) (f) PONNMP (Gaussian)

(g) PONNMP (Gumbel) (h) PONNMP (Gumbel) (i) PONNMP (Gumbel)

(j) PONNMP (Clayton) (k) PONNMP (Clayton) (l) PONNMP (Clayton)

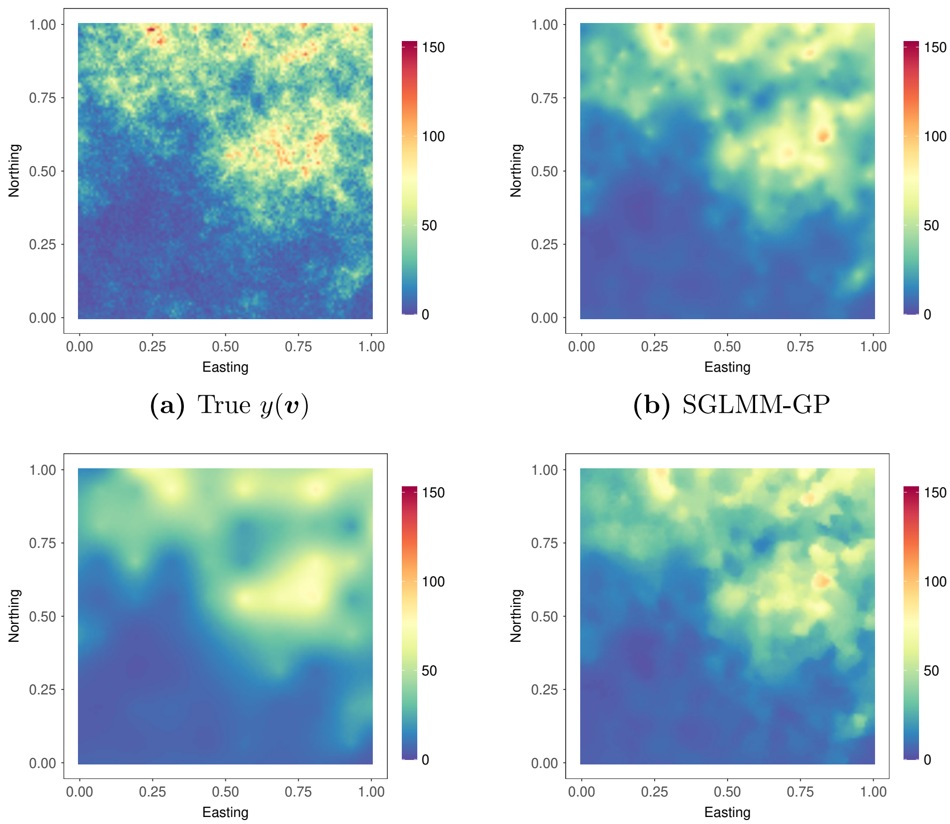
Figure 5.1: Chapter 5 - first simulation data analysis. Interpolated surfaces of the true model (first row), and posterior median estimates of the Poisson NNMP (PONNMP) models using Gaussian (second row), Gumbel (third row), and Clayton (fourth row) copulas. Columns from left to right correspond to scenarios with $\sigma_1 = 1, 3, 10$, respectively.

Table 5.2: Chapter 5 - first simulation data analysis. Posterior mean and 95% credible interval estimates for the rate parameter λ of the Poisson NNMP marginal distribution, and scores for comparison of Gaussian, Gumbel and Clayton copula NNMP models, under each of the three simulation scenarios for σ_1.

	$\sigma_1 = 1$	$\sigma_1 = 3$	$\sigma_1 = 10$
	λ	λ	λ
Gaussian	4.55 (4.16, 4.94)	4.71 (4.37, 5.07)	4.88 (4.55, 5.22)
Gumbel	4.78 (4.39, 5.21)	4.88 (4.56, 5.24)	4.94 (4.66, 5.23)
Clayton	5.33 (4.99, 5.68)	5.25 (4.96, 5.56)	5.36 (5.08, 5.65)

	$\sigma_1 = 1$			$\sigma_1 = 3$			$\sigma_1 = 10$		
	CRPS	ES	VS	CRPS	ES	VS	CRPS	ES	VS
Gaussian	0.69	12.77	94855	0.85	15.54	124893	0.93	16.98	138592
Gumbel	0.69	12.58	92278	0.85	15.32	120932	0.92	16.71	134774
Clayton	0.75	14.34	125800	0.90	17.36	164148	1.00	18.70	174123

$1, 3, 10$, the Gumbel model yields the more accurate estimates for λ. In particular, the Gumbel model's 95% CIs include the true parameter value, whereas those of the Gaussian and Clayton models failed to cover it when $\sigma_1 = 1$ and $\sigma_1 = 10$, respectively. Regarding predictive performance, the Gumbel model outperforms to a smaller or larger extent the other two models across different scenarios. Predictive random fields under the three models are provided in Figure 5.1. We found that prediction by the Clayton model was not able to recover large values. Compared to the Gaussian model, the Gumbel model recovered large values slightly better. Overall, this example demonstrates that, when the underlying spatial dependence is driven by non-Gaussian processes, it is practically useful to consider copulas from asymmetric families, including use of appropriate model comparison tools.



(a) True $y(\boldsymbol{v})$ **(b)** SGLMM-GP **(c)** SGLMM-GPP **(d)** NBNNMP

Figure 5.2: Chapter 5 - second simulation data analysis. Interpolated surfaces of the true model and posterior median estimates of the SGLMM-GP, SGLMM-GPP and NBNNMP models.

5.5.2 Second Simulation Experiment

We generated data over a grid of sites with 120×120 resolution, uniformly on the square $[0,1] \times [0,1]$, using a Poisson SGLMM with $y(\boldsymbol{v}) \,|\, \eta(\boldsymbol{v}) \sim \text{Pois}(\eta(\boldsymbol{v}))$, and $\log(\eta(\boldsymbol{v})) = \beta_0 + v_1\beta_1 + v_2\beta_2 + z(\boldsymbol{v})$, where $\boldsymbol{v} = (v_1, v_2)$, and $z(\boldsymbol{v})$ is a zero-centered Gaussian process (GP) with variance parameter $\sigma^2 = 0.2$ and an exponential correlation function with range parameter $\phi_0 = 1/12$. We set the regression coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^\top = (1.5, 1, 2)^\top$, resulting in a random field with a trend, as shown in Figure 5.2(a).

We considered three models. The first is the negative binomial NNMP model (NBNNMP) with a Gaussian copula, as discussed in Example 1. The second model

Table 5.3: Chapter 5 - second simulation data analysis. posterior mean and 95% credible interval estimates for the regression parameters, performance metrics, and computing time, under the NBNNMP model and the two SGLMM models.

	True	NBNNMP	SGLMM-GP	SGLMM-GPP
β_0	1.5	1.61 (1.29, 1.97)	1.53 (1.22, 1.81)	1.41 (1.02, 1.73)
β_1	1	0.90 (0.51, 1.31)	0.70 (0.25, 1.15)	0.91 (0.43, 1.34)
β_2	2	1.94 (1.51, 2.32)	2.18 (1.91, 2.53)	2.25 (1.81, 2.84)
RMSPE	-	9.06	8.88	10.00
95% CI cover	-	0.98	0.97	0.78
95% CI width	-	37.02	32.24	19.02
CRPS	-	4.58	4.52	5.37
ES	-	92.07	91.41	107.46
VS	-	5175591	5199629	6378263
Time (mins)	-	11.18	935.02	11.68

(SGLMM-GP) is a Poisson SGLMM with a GP prior assigned to $z(\boldsymbol{v})$. For the last model (SGLMM-GPP), we considered a Poisson SGLMM with spatial random effects $z(\boldsymbol{v})$ corresponding to a Gaussian predictive process (GPP, Banerjee et al. 2008), with 10×10 knots placed on a grid over the domain. We chose the number of knots such that the computing times for the SGLMM-GPP and NBNNMP models are similar. As in the first simulation example, all models were fit to 800 observations and compared on the basis of 200 additional observations.

The regression coefficients for all models were assigned mean-zero, dispersed normal priors. We worked with an exponential correlation function for all models, used for $\rho_l(\boldsymbol{v})$ of the Gaussian copula in the NBNNMP model, and as the correlation function for the GP and GPP in the SGLMMs. The range parameter was assigned an inverse gamma prior $\mathrm{IG}(3, 1)$ for the NBNNMP model, and a uniform prior $\mathrm{Unif}(1/30, 1/3)$ for the other two models. The cutoff point kernel of the NBNNMP was also specified an exponential correlation function, with an $\mathrm{IG}(3, 1)$ prior for the range parameter. The variance parameter for the SGLMM models was assigned an inverse gamma prior $\mathrm{IG}(2,1)$. For the logit Gaussian distribution parameters $\boldsymbol{\gamma}$ and κ^2 of the NBNNMP, we used $N((-1.5,0,0)^\top, 2\mathbf{I}_3)$ and $\mathrm{IG}(3,1)$ priors, respectively. Finally, we placed a $\mathrm{Ga}(1,1)$ prior on the NBNNMP dispersion parameter r.

Estimates of the regression parameters and performance metrics for out-of-sample prediction are provided in Table 5.3. We observe that, overall, the NBNNMP model provided the more accurate estimation for $\boldsymbol{\beta}$. Regarding predictive performance, the NBNNMP model outperformed the SGLMM-GPP model by a large margin, and was comparable to the SGLMM-GP model, which corresponds to the data generating process for this simulation experiment. Moreover, the last row of the table highlights the NBNNMP model's huge gains in computing time compared to the SGLMM-GP model.

Figure 5.2(b)-5.2(d) plots the posterior median estimates of the random field for the three models. The SGLMM-GPP yields an overly smooth estimate, whereas the SGLMM-GP and NBNNMP provide similar estimates that approximate well the true surface. Overall, this example illustrates the inferential and computational advantages of discrete copula NNMPs for modeling count data.

5.5.3 North American Breeding Bird Survey Data

The primary source of information on population evolution for birds is count data surveys. Since 1966, the North American Breeding Bird Survey (BBS) has been conducted to monitor bird population change. There are over 4000 sampling units in the survey, each with a 24.5-mile roadside route. Along each route, volunteer observers count the number of birds by sight or sound, in a 3-min period at each of 50 stops (Pardieck et al., 2020). The BBS data are often used to determine temporal or geographical patterns of relative abundance. Spatial maps of relative abundance are crucial for ecological studies.

We are interested in the Northern Cardinal, a bird species that is prevalent in Eastern United States. Figure 5.5(a) shows the number of birds observed in 2019, with the sizes of the circle radii proportional to the number of birds at each sampling location. The dataset was extracted with the help of the R package *bbsAssistant* (Burnett et al., 2019); it contains 1515 irregular sampling locations. From Figure 5.5(a) we observe that the counts tend to increase as latitude decreases, and we thus take latitude as a covariate to account for the long range variability in the population.

Analysis of neighborhood sizes

We considered the Gaussian copula NBNNMP model defined in Example 1, with spatially varying marginal $\mathrm{NB}(\exp(\boldsymbol{x}(\boldsymbol{v})^\top \boldsymbol{\beta}), r)$, where $\boldsymbol{\beta} = (\beta_0, \beta_1)^\top$. We used the same link functions and prior specifications as in Section 5.5.2. We first examined model performance under different values of L. We applied the Gaussian copula NBNNMP model to the whole data set with $L = 5, 10, 15, 20$. For each L, we ran the MCMC algorithm for 30000 iterations, discarding the first 10000 iterations, and collecting posterior samples every 5th iteration.

Table 5.4 provides the posterior means and 95% credible interval estimates of the model parameters. They were quite robust across different values of L, except for those of ϕ and ζ, even though the different credible intervals have substantial overlap. Note that ϕ and ζ are the range parameters of the exponential correlation functions for the Gaussian copula correlation and for the cutoff point kernel, respectively. Since a model with a large value of L includes more distant neighbors, ϕ and ζ should be larger as they indicate effective ranges.

To examine the model performance on estimating the weights, we randomly selected ten locations $(\boldsymbol{s}_{j_1}, \ldots, \boldsymbol{s}_{j_{10}})$ such that $21 \leq j_k \leq 200$ for $k = 1, \ldots, 5$ and

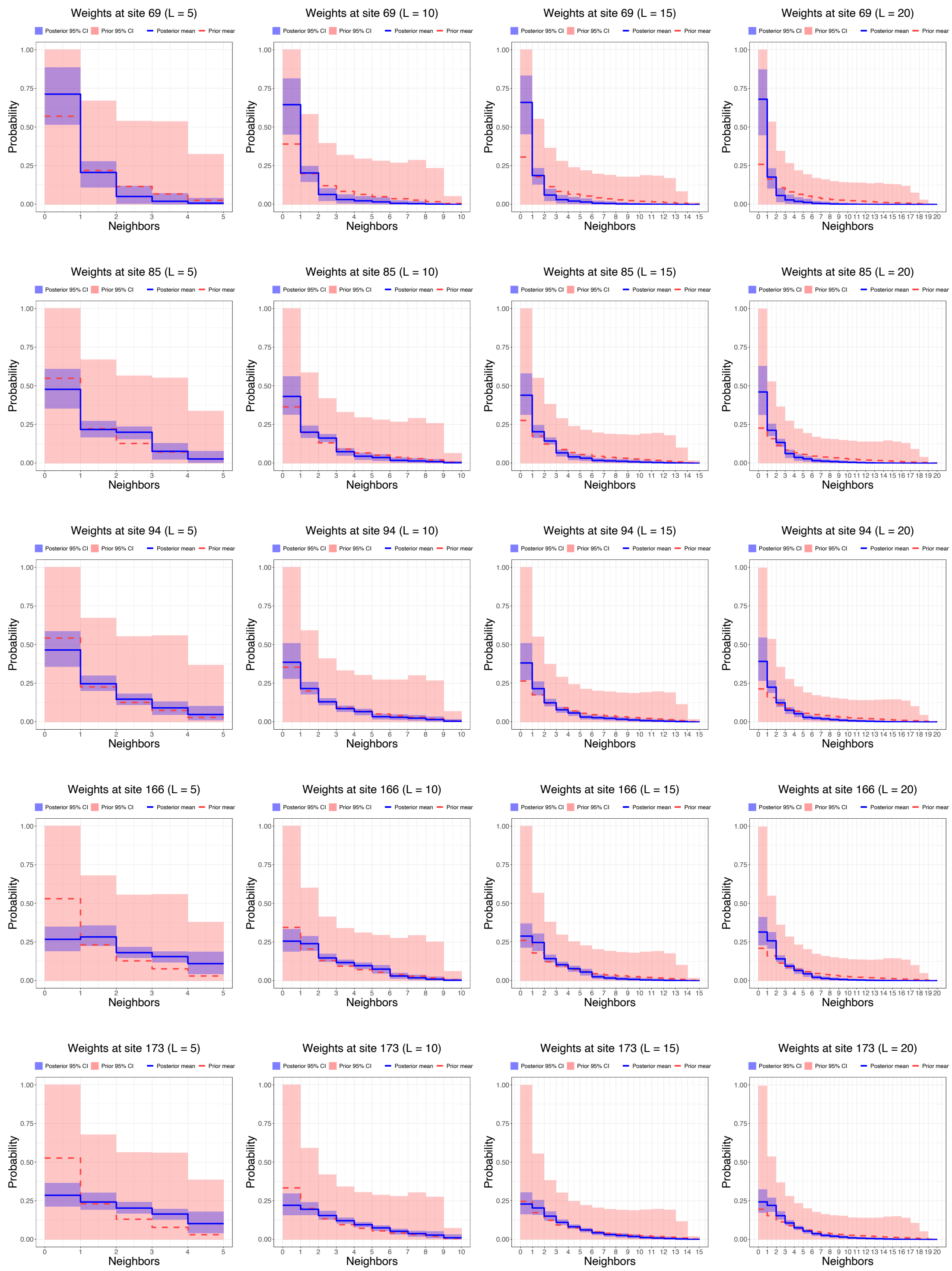


Figure 5.3: Chapter 5 - North American BBS data analysis. Posterior means and 95% credible interval estimates of the weights of the first five locations.

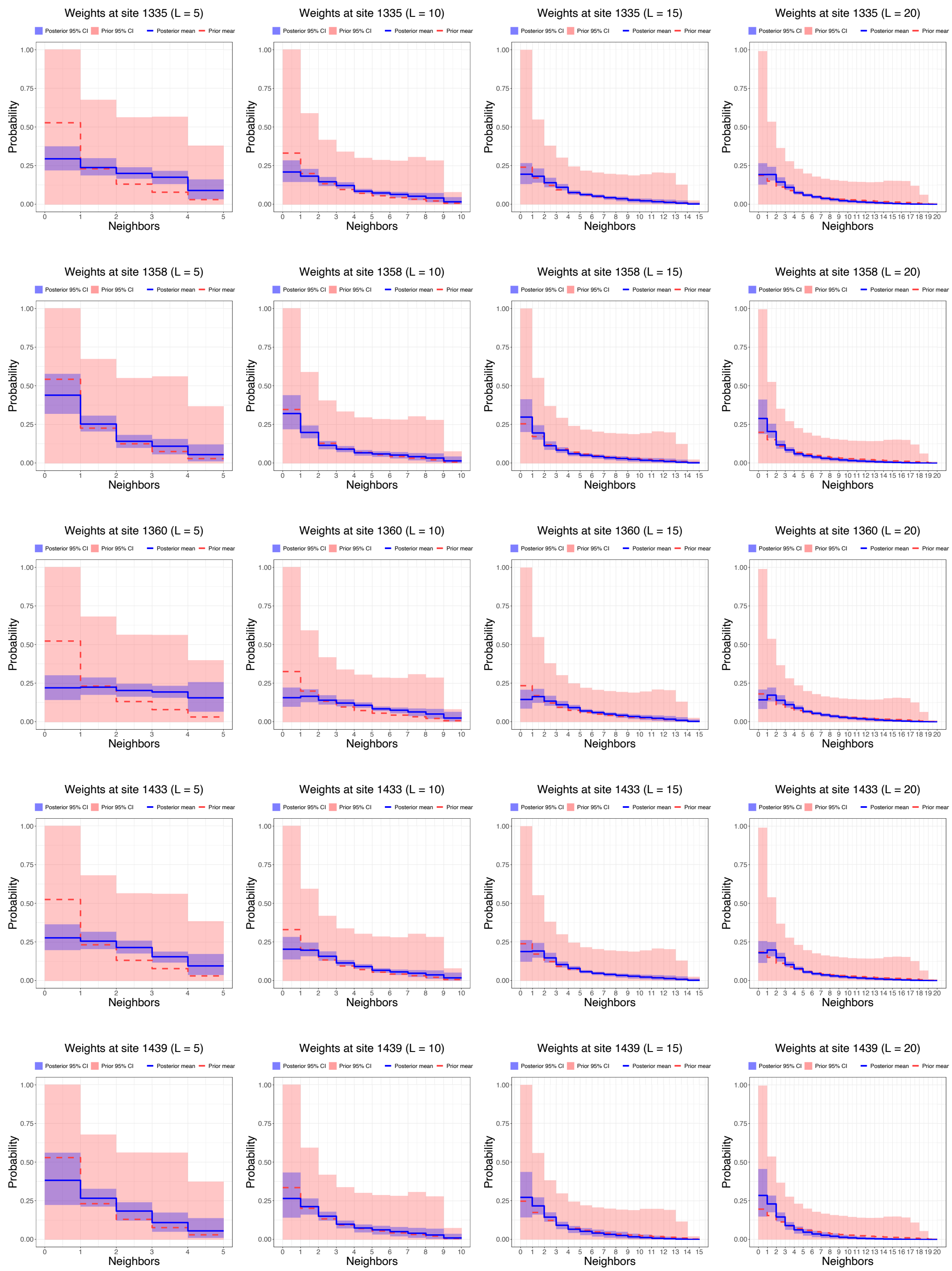


Figure 5.4: Chapter 5 - North American BBS data analysis. Posterior means and 95% credible interval estimates of the weights of the last five locations.

Table 5.4: Chapter 5 - North American BBS data analysis. Posterior means and 95% credible interval estimates for the parameters and computing time, under the Gaussian copula NBNNMP models with different values of L.

	L = 5	L = 10	L = 15	L = 20
β_0	6.52 (5.88, 7.33)	6.56 (5.69, 7.22)	6.48 (5.72, 7.28)	6.48 (5.62, 7.29)
β_1	-0.09 (-0.11, -0.07)	-0.09 (-0.11, -0.06)	-0.09 (-0.11, -0.07)	-0.09 (-0.11, -0.06)
ϕ	1.61 (1.26, 2.04)	2.51 (1.80, 3.47)	2.65 (1.93, 3.59)	2.62 (1.81, 3.68)
ζ	0.82 (0.45, 1.82)	1.10 (0.63, 2.15)	1.37 (0.77, 2.70)	1.71 (0.87, 3.80)
r	1.94 (1.65, 2.22)	1.86 (1.51, 2.19)	1.87 (1.54, 2.21)	1.88 (1.53, 2.22)
γ_0	-1.28 (-3.60, 0.96)	-1.29 (-3.49, 1.01)	-1.51 (-3.77, 0.66)	-1.69 (-3.85, 0.41)
γ_1	0.00 (-0.02, 0.03)	0.00 (-0.02, 0.02)	0.00 (-0.02, 0.02)	0.00 (-0.02, 0.02)
γ_2	0.03 (-0.01, 0.08)	0.02 (-0.02, 0.06)	0.01 (-0.02, 0.06)	0.01 (-0.02, 0.05)
κ^2	2.39 (1.48, 3.65)	2.23 (1.46, 3.31)	1.93 (1.24, 2.95)	1.63 (1.09, 2.30)
Time (mins)	29.17	32.71	38.49	50.91

$1312 \leq j_k \leq 1512$ for $k = 6, \ldots, 10$. Since we used random ordering to assign indices to the locations, the neighbors of $\boldsymbol{s}_{j_k}$, $k = 1, \ldots, 5$, may consist of distant locations, whereas the neighbors of $\boldsymbol{s}_{j_k}$, $k = 6, \ldots, 10$, were expected to be all nearby. Figures 5.3 and 5.4 illustrate the posterior means and 95% credible interval estimates of the weights at these ten locations. From the figures, we see that the model provided estimates of the weights that adjust to different neighborhood structures. The effective number of neighbors varied across locations. In addition, the estimates of the weights were quite robust as the value of L increased. We can observe that the model was able to penalize irrelevant neighbors by assigning very small probabilities. While $L = 5$ seems too small to work as an upper bound, we observe that when L ranged from 10 to 20, the effective number of weights for each location was quite consistent.

Finally, a sensitivity analysis was carried out to study the impact of L on the model performance. We randomly split the data into two sets, a training set with 1212 observations and a testing set with 300 observations. We then applied

Table 5.5: Chapter 5 - North American BBS data analysis. Performance metrics of the Gaussian copula NBNNMP models with different values of L.

	RMSPE	95% CI cover	95% CI width	CRPS	ES	VS
L = 5	19.90	0.93	66.07	9.79	235.34	39759593
L = 6	19.82	0.94	65.91	9.75	234.50	39446330
L = 7	19.83	0.94	66.04	9.75	234.73	39464801
L = 8	19.80	0.94	66.19	9.75	234.36	39345232
L = 9	19.75	0.94	66.33	9.72	233.42	39073447
L = 10	19.72	0.94	66.27	9.72	233.50	39066501
L = 11	19.74	0.94	66.40	9.73	233.75	39179711
L = 12	19.73	0.95	66.67	9.70	233.10	38919544
L = 13	19.73	0.94	66.50	9.71	233.29	38978258
L = 14	19.70	0.95	66.69	9.71	233.20	38920854
L = 15	19.72	0.95	66.70	9.71	233.26	38865662
L = 16	19.73	0.94	66.70	9.72	233.50	38998533
L = 17	19.72	0.95	66.67	9.72	233.55	38982480
L = 18	19.72	0.94	66.80	9.72	233.63	39013058
L = 19	19.74	0.94	66.67	9.72	233.94	39111633
L = 20	19.79	0.94	66.75	9.74	234.30	39194713

the Gaussian copula NBNNMP with L from 5 to 20, and evaluated the model performance based on out-of-sample predictive performance as shown in Table 5.5. There were no discernible differences among the models with L between 9 and 20. The conclusion from the robustness analysis of the choice of L is that $L = 20$ works as a reasonable upper bound for this particular data example.

Comparison of three copula NBNNMP models

We compare three discrete copula NBNNMP models with $L = 20$. Each model used either the spatial Gaussian, Gumbel or Clayton copulas, with negative binomial marginals $\mathrm{NB}(\exp(\boldsymbol{x}(\boldsymbol{v})^\top \boldsymbol{\beta}), r)$. We used the same link functions and prior specifications for copulas as in Section 5.5.1 and the same priors for other pa-

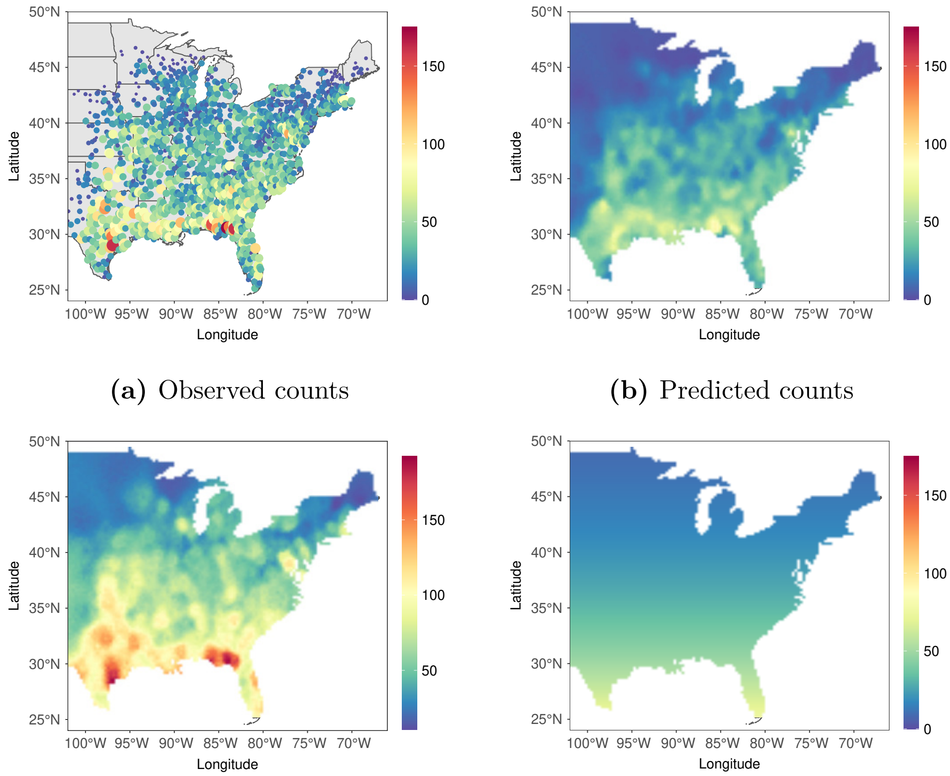
Table 5.6: Chapter 5 - North American BBS data analysis. Performance metrics for NBNNMPs based on different copulas.

	RMSPE	95%CI cover	95%CI width	CRPS	ES	VS
Gaussian	19.75	0.94	66.62	9.72	233.91	39136486
Gumbel	19.71	0.96	68.77	9.81	236.18	39665090
Clayton	19.97	0.93	71.51	9.91	237.21	39566563

rameters as in Section 5.5.2. We fitted the models to 1215 randomly selected observations and used the remaining 300 for model comparison. For each model, we ran the MCMC algorithm for 30000 iterations, discarding the first 10000 iterations, and collected posterior samples every 5th iteration. Table 5.6 shows the comparison based on out-of-sample predictive performance. Overall, the Gaussian copula outperformed the other two.

North American BBS data analysis

We proceeded to analyze the BBS data with the Gaussian copula NBNNMP model with $L = 20$. The posterior mean and 95% credible interval estimates of the regression parameters β_0 and β_1 are $6.53\,(5.61, 7.38)$ and $-0.09\,(-0.11, -0.06)$, respectively, suggesting an increasing trend in the Northern Cardinal counts as the latitude decreases. The corresponding estimates of the dispersion parameter r are $1.88\,(1.55, 2.22)$, indicating that there is overdispersion over the domain. Figure 5.5(b) and 5.5(c) show the posterior predictive median of the counts and the 95% posterior predictive CI width, respectively. Figure 5.5(b) displays the domain's spatial variability. The estimated uncertainty, as shown in Figure 5.5(c), is meaningful, as areas with high uncertainty correspond to those where the observed counts are quite heterogeneous. Figure 5.5(d) provides a spatial map of the mean of the negative binomial marginals, which depicts a North–South trend. Model checking results are shown in Figure 5.6, including a posterior summary



(a) Observed counts

(b) Predicted counts

(c) 95% CI widths

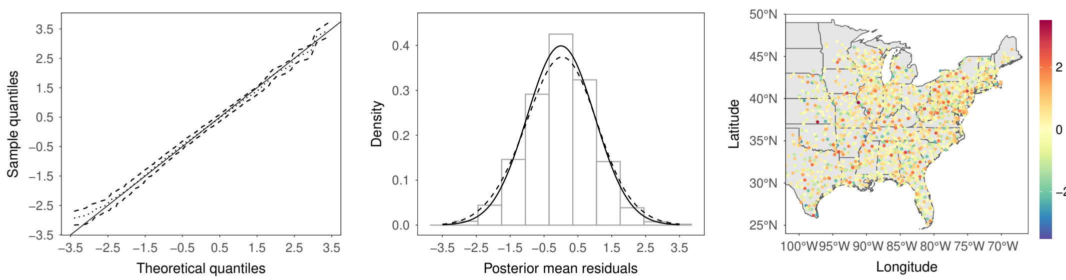
(d) Posterior mean of $\exp(\boldsymbol{x}(\boldsymbol{v})^\top \boldsymbol{\beta})$

Figure 5.5: Chapter 5 - North American BBS data analysis: (a) observed counts for 2019 North American BBS of Northern Cardinal, with circle radius proportional to the observed counts; (b) median of the posterior predictive distribution for Northern Cardinal count; (c) 95% CI widths of the posterior predictive distribution for Northern Cardinal count; (d) posterior mean of $\exp(\boldsymbol{x}(\boldsymbol{v})^\top \boldsymbol{\beta})$.

of the Gaussian quantile-quantile plot, and the histogram and spatial plot of the posterior means of the residuals. The results suggest good model fit.

Comparison with the SGLMM method

Finally, we assessed the model performance by comparison with the SGLMM-GP model. Again, we randomly split the data into a training set with 1212 observations and a testing set with 300 observations. We ran the MCMC algorithm for the Gaussian copula NBNNMP ($L = 20$) for 30000 iterations, discarding the first 10000 iterations, and collecting posterior samples every 5th iteration. Since



(a) Quantile-quantile plot **(b)** Histogram **(c)** Posterior mean residuals

Figure 5.6: Chapter 5 - North American BBS data analysis. Randomized quantile residual analysis: (a) dotted and dashed lines correspond to the posterior mean and 95% credible interval estimates, respectively; (b) solid and dashed lines are the standard Gaussian density and the kernel density estimate of the posterior means of the residuals, respectively; (c) spatial plot of the posterior means of the residuals.

the MCMC for SGLMM-GP involves sampling the spatial random effects, we ran the algorithm for 50000 iterations and collected posterior samples every five iterations, with the first 30000 as burn-in.

Table 5.7 shows the parameter estimates and predictive performances by the two models. The parameter estimates of $\boldsymbol{\beta}$ were quite close under the two models. Both models indicate an increasing trend in the counts as the latitude decreases. On the other hand, the NBNNMP model resulted in better out-of-sample predictive performance, and, notably, it was substantially more efficient to implement, with computing time 110 times faster than that for the SGLMM-GP model.

5.6 Discussion

We have introduced a new class of models for discrete geostatistical data, with particular focus on using different families of bivariate copulas to build modeling and inference. Compared to traditional SGLMM methods, the proposed class of models is scalable, and is able to accommodate complex dependence structures.

In general, multivariate discrete distributions are not as tractable as certain

Table 5.7: Chapter 5 - North American BBS data analysis. Parameter estimates and performance metrics of the Gaussian copula NBNNMP and the SGLMM-GP models.

	NBNNMP	SGLMM-GP
β_0	6.57 (5.83, 7.19)	6.67 (6.55, 6.81)
β_1	-0.09 (-0.10, -0.07)	-0.10 (-0.10, -0.09)
RMSPE	19.79	20.41
95% CI cover	0.94	0.94
95% CI width	66.56	76.56
CRPS	9.74	10.10
ES	234.22	239.02
VS	39204378.76	40185343.15
Time (mins)	37.56	4208.33

families of multivariate continuous distributions, in particular, the Gaussian family. This is the fundamental difficulty of process-based modeling for discrete geostatistical data. Our methodology overcomes this difficulty through a structured mixture model formulation, reducing the specification of a multivariate pmf to that of bivariate copulas that define the mixture components. This formulation yields models for spatial processes that provide flexibility and deliver computational scalability.

In this chapter, we explored the strategy of using a single copula family for all bivariate distributions. Exploring the alternative which builds from different copula families for the bivariate distributions remains an interesting question to investigate. We can cast this as a model selection problem and develop algorithms to select models; see examples in Panagiotelis et al. (2017) and Gruber and Czado (2018) in the context of regular vine copula models. Different copula families for bivariate distributions yield more flexibility for the model to capture complex dependence, albeit at the cost of computational scalability. If the main purpose of the application is prediction, rather than model selection, one could explore calibrating the prediction using all candidate copula families. This could be done, for example, with the pseudo Bayesian model averaging approach, where the weight for each model is estimated based on stacking (Yao et al., 2018).

We conduct inference for the discrete copula NNMPs based on the continuous extension approach. Apart from the aforementioned benefits, this approach may allow discrete copula NNMPs to make use of alternative algorithms for faster computation. Moreover, with the CE approach, it is possible to develop a class of NNMPs for a multivariate response that consists of both continuous and discrete components, while at the same time retaining computational efficiency.

Chapter 6

Conclusions

We conclude with a discussion of some possible extensions of the proposed methodologies in this dissertation. We have presented a DAG-based framework for non-Gaussian processes in time and space, using the following mixture densities for the vertexes of the DAG,

$$p(z_i \mid \boldsymbol{z}_{\mathrm{pa}(i)}) = \sum_{l=1}^{i_L} w_l(i) f_{il}(z_i \mid z_{\mathrm{pa}(i)}^{(l)}), \tag{6.1}$$

which forms the joint density of interest $p(z_1, \ldots, z_n)$. The key to (6.1) is the specification of a collection of bivariate distributions that define the conditional densities f_{il}, which provides generality for modeling non-Gaussian data.

Methodological extensions can be explored according to different components of (6.1), e.g., the weights $w_l(i)$, and the conditional densities f_{il}. For example, if we define the weights $w_l(i) = \pi_l h_l(z_{\mathrm{pa}(i)}^{(l)}) / \sum_{r=1}^{i_L} \pi_r h_r(z_{\mathrm{pa}(i)}^{(r)})$, and construct f_{il} from a bivariate distribution with marginals $f_Z(z) = \sum_{l=1}^{L} \pi_l h_l(z)$. It is possible to obtain a stationary marginal f_Z after we impose some conditions relevant to the first L elements $(z_1, \ldots, z_L)$. For density estimation problems, mixture models with appropriate kernels can well handle non-Gaussianity. Thus, this extension improves modeling power for marginal non-Gaussian behaviors. When the kernel h_l is a Gaussian distribution, the finite mixture model for f_Z can be extended to an infinite mixture one without much effort, by using a Dirichlet process Gaussian mixture model for the bivariate mixture components. The resulting conditional density f_{il} is similar to those in Antoniano-Villalobos and Walker (2016) and DeYoreo and Kottas (2017).

Including one of the given parents in each mixture component of (6.1) facilitates the study of model properties, and provides generality for modeling non-Gaussian data based on specifying bivariate distributions. However, the resulting models are limited in capturing behaviors such as high-order interactions among parents. In the temporal/spatial context, high-order interactions mean the joint effects of lags/nearest neighbors on the target. If only second-order interactions are of interest, an economical extension, i.e., without fundamentally changing the model structure, is to add mixture components with densities f_{il} conditional on two parents, which requires trivariate distributions to build such conditionals. When we expect extensions to k-order interactions ($k > 2$), this modeling strategy becomes infeasible as the number of mixture components will be tremendously large even with a small k. Alternatively, we can expand the conditioning set to include all parents, defining the density as

$$p(z_i \,|\, \boldsymbol{z}_{\mathrm{pa}(i)}) = \sum_{l=1}^{i_L} w_l(i) f_{il}(z_i \,|\, \boldsymbol{z}_{\mathrm{pa}(i)}). \tag{6.2}$$

Although expanding the conditioning set defeats the objective of modeling general non-Gaussian data, it can be useful for particular data examples. Suppose $\{Z_i\}$ is a time series and the parents are temporal lags, that is, $\boldsymbol{z}_{\mathrm{pa}(i)} = \boldsymbol{z}_{i_L,i-1} = (z_{i-1}, \dots, z_{i-i_L})$. Take $w_l(i) = w_l$, and let $f_{il} = f_l$ be a conditional Gaussian distribution, for all i. Then the mixture density $p(z_i \,|\, \boldsymbol{z}_{\mathrm{pa}(i)})$ in (6.2) corresponds to the mixture autoregressive (MAR) model (Wong and Li, 2000). Letting i_L go to infinity, the mixture density can be regarded as the infinite mixture representation of a Bayesian nonparametric model, that is, $p(z_i \mid \boldsymbol{z}_{\mathrm{pa}(i)}) = \int f(z_i \mid \boldsymbol{z}_{\mathrm{pa}(i)}, \boldsymbol{\theta})\, dG(\boldsymbol{\theta})$, where G is assigned a Dirichlet process prior. In particular, Lau and So (2008) study this formulation as a nonparametric extension of the MAR model. From a different perspective, Heiner and Kottas (2022b) build a nonparametric conditional density by placing a prior for the joint density $p(z_i, \boldsymbol{z}_{i_L, i-1})$. Their modeling approach yields localized weights that are functions of temporal lags, in contrast with the static weights in Lau and So (2008). Nevertheless, nonparametric modeling of $p(z_i \mid \boldsymbol{z}_{\mathrm{pa}(i)})$ adds more flexibility to accommodate complex dependence. While much effort has been pursued for modeling temporal data, little has been done for spatial statistics. It will be interesting to explore the spatial counterparts of these Bayesian nonparametric models within the DAG-based framework.

The presented work provides a collection of flexible modeling tools for non-Gaussian dependent data, without emphasis on forms of the conditionals and related marginals in (6.1). On the other hand, we believe that the proposed framework would contribute new statistical methodologies to a variety of particular problems, by placing focus on specific families of the conditionals and related marginals. With this in mind, many examples that are discussed briefly in the previous chapters can be further explored. As an illustration, consider the NNMP extension to extreme value analysis. Utilizing Propositions 4.1 and 5.1, we can construct spatial processes given a set of spatially varying marginal distributions. This coincides with the idea of specifying a family of generalized extreme value (GEV) distributions in modeling extremes. It is tempting, then, to extend the approach illustrated in Section 4.2.4, investigating strength of tail dependence with measures implied by both the dependence structure of the bivariate distributions and their marginals, in particular, in the GEV family. An example of such measures is conditional tail expectation (CTE), e.g., $E(Z_i \mid Z_{\mathrm{pa}(i)} = z_0)$, or $E(Z_i \mid Z_{\mathrm{pa}(i)} > z_0)$. When the marginals are GEVs, examining the rate at which the CTE grows as $z_0 \to \infty$ could be interesting to help understand the model's tail properties. In a spatial setting, some multivariate distributions such as a skew-t distribution have the long-range dependence property (Morris et al., 2017). That is, observations at two locations are asymptotically dependent regardless of the distance between them. This property is undesirable if only local dependence is expected. Studying the long-range dependence feature of the NNMP would benefit the formulation of the modeling framework for spatial extremes.

This dissertation centers on developing applied methodologies. Their practical importance has been demonstrated in several data examples. It is no doubt that these methods would find opportunities in applications from diverse areas. Incidentally, many fields have witnessed technological advancements that have led to the increasing prevalence of large-scale data. Although we have proposed models for non-Gaussian data that balance flexibility and scalability, it remains a challenging task to scale up statistical inference in the presence of massive data sets. MCMC that delivers full Bayesian inference has been the focus throughout the dissertation. However, posterior simulation with MCMC is in general computationally expensive. To retain MCMC as an attractive choice in the proposed framework for "big data" problems, extensions can be pursued in two avenues, both of which use the idea of "subsetting". The first one explores scalable MCMC algorithms that employ data subsampling in each MCMC iteration, while the other obtains a pseudo posterior distribution in a distributed manner, by combining sub-posteriors estimated using MCMC for all subsets of the data. We refer to Quiroz et al. (2018) and Guhaniyogi et al. (2022), respectively, for more details regarding the two avenues; see also a related work (Grenier and Sansó, 2021) that develops a distributed method for the NNGP models.

Alternative to MCMC, the most popular algorithm that solves the problem in scaling up is variational inference (VI). VI seeks a variational density, say $q^*(\boldsymbol{\theta})$, to approximate the intractable joint posterior distribution, formulating Bayesian estimation as an optimization problem. The density $q^*(\boldsymbol{\theta})$ is commonly factorized over a partition of $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_J)$, namely, $q^*(\boldsymbol{\theta}) = \prod_{j=1}^{J} q_j^*(\boldsymbol{\theta}_j)$, referred to as the mean-field variational family. Our current framework defines the DAG conditionals in the form of (6.1), leading to an estimation scheme similar to that of finite mixture models. However, unlike many finite mixture models that fall within the category of conditionally conjugate models, mean-field VI for the current framework does not enjoy the conjugate property, i.e., closed-form expressions for updates of most $q_j^*(\theta_j)$ are not available in the coordinate ascent algorithm (Bishop, 2006). Options to resolve this issue include implementing approximation methods, e.g., Laplace approximation (Wang and Blei, 2013) and importance sampling (Ren et al., 2011; Barata et al., 2022), or developing more general VI algorithms based on those such as the black box VI (Ranganath et al., 2014) and the automatic differentiation VI (Kucukelbir et al., 2017).

We close this section with remarks on the potential of our DAG-based framework. In Chapters 4 and 5, we propose the NNMP, a whole new class of spatial processes that is flexible for modeling general non-Gaussian data. It is computationally attractive relative to many non-Gaussian models as it is constructed via a nearest-neighbor DAG representation that achieves sparsity. The NNMP has the potential of practical interest in fields such as health and environmental sciences. This will be achieved with continuous work on expanding the NNMP framework for relevant problems, combined with the development of publicly available software, e.g., R packages. In Chapter 3, we propose the MTDPP, a broad class of point process models that can explain various types of point patterns. I believe that the MTDPP, coupling with the stationarity conditions developed in Chapter 2, would find its place in point process modeling of practical interest, either as an extension of the renewal processes when the independence assumption is unrealistic, or as a computationally efficient model for large-scale point patterns with dependence between points.

Appendix A

Proofs

Proof of Proposition 2.1. Without loss of generality, we consider the case where X_t has a continuous distribution for all t. Moreover, for the argument that follows to apply to any $t \geq 2$, we express the transition density as $f(x_t \mid x^{t-1}) = \sum_{l=1}^{t_L} w_l^* f_{U_l|V_l}(x_t \mid x_{t-l})$, for $t \geq 2$, where $t_L = \min\{t-1, L\}$. When $t > L$, $w_l^* \equiv w_l$, for $l = 1, ..., L$, whereas for $2 \leq t \leq L$, $w_l^* = w_l$, for $l = 1, ..., t_L - 1$, and $w_{t_L}^* = 1 - \sum_{k=1}^{t_L - 1} w_k$. With this notational convention, we have $\sum_{l=1}^{t_L} w_l^* = 1$.

Using the proposition assumptions,

$$g_2(x_2) = \int_{\mathcal{S}} f(x_2 \mid x_1) f_X(x_1) dx_1 = \int_{\mathcal{S}} f_{U_1|V_1}(x_2 \mid x_1) f_{V_1}(x_1) dx_1 = f_{U_1}(x_2) = f_X(x_2)$$

and thus the result is valid for $t = 2$. To prove the proposition by induction, assume the result holds true for generic $t-1$, that is, $g_{t'}(x_{t'}) = f_X(x_{t'})$, for all $x_{t'} \in \mathcal{S}$, and for all $t' \leq t-1$. Denote by $p(x_1, \ldots, x_{t-1})$ and $p(x_{t-t_L}, \ldots, x_{t-1})$ the joint density for random vector $(X_1, \ldots, X_{t-1})$ and $(X_{t-t_L}, \ldots, X_{t-1})$, respectively.

Then, the marginal density for X_t can be derived as follows:

$$\begin{aligned}
g_t(x_t) &= \int_{\mathcal{S}^{t-1}} f(x_t \mid x^{t-1})\, p(x_1, \dots, x_{t-1})\, dx_1 \dots dx_{t-1} \\
&= \sum_{l=1}^{t_L} w_l^* \int_{\mathcal{S}^{t_L}} f_{U_l|V_l}(x_t \mid x_{t-l})\, p(x_{t-t_L}, \dots, x_{t-1})\, dx_{t-t_L} \dots dx_{t-1} \\
&= \sum_{l=1}^{t_L} w_l^* \int_{\mathcal{S}} f_{U_l|V_l}(x_t \mid x_{t-l})\, g_{t-l}(x_{t-l})\, dx_{t-l} \\
&= \sum_{l=1}^{t_L} w_l^* \int_{\mathcal{S}} f_{U_l|V_l}(x_t \mid x_{t-l})\, f_{V_l}(x_{t-l})\, dx_{t-l} \\
&= f_X(x_t),
\end{aligned}$$

where for the second-to-last equation we used $g_{t-l} = f_X$, for $l = 1, ..., t_L$, obtained from the induction argument, as well as the proposition assumption, $f_X = f_{V_l}$, for all l. Finally, the last equation is based on the proposition assumption that $f_{U_l} = f_X$, for all l. □

Proof of Proposition 2.2. We refer to the definition of weak stationarity from Brockwell and Davis (1991). A time series $\{X_t : t \in \mathbb{N}\}$, with index set $\mathbb{N} = \{1, 2, \dots\}$, is said to be weakly stationary if i) $E(X_t^2) < \infty$ for all $t \in \mathbb{N}$; ii) $E(X_t) = m$ for some finite m and for all $t \in \mathbb{N}$; iii) $\mathrm{Cov}(X_{t+h}, X_t) = \gamma(h)$ for all $t, h \in \mathbb{N}$. Under condition (1) of Proposition 2, if an MTD time series has a stationary marginal distribution such that its corresponding first and second moments exist and are finite, then $\mu = E(X_t)$ and $\mu^{(2)} = E(X_t^2)$ are finite for all $t \in \mathbb{N}$. Thus, the weak stationarity conditions (i) and (ii) are satisfied.

Under condition (2) of Proposition 2, the cross moment

$$\begin{aligned}
E(X_{t+h}X_t) &= E(X_t E(X_{t+h} \,|, X_{t+h-1}, \dots, X_{t+h-L})) \\
&= E(X_t \sum_{l=1}^{L} w_l(a_l + b_l X_{t+h-l})) = \sum_{l=1}^{L} w_l a_l \mu + \sum_{l=1}^{L} w_l b_l E(X_{t+h-l}X_t),
\end{aligned}$$

for all $t \in \mathbb{N}$ and $h \geq L$. Assuming that the cross moment is independent of t for $h \geq 1$, we can obtain the following non-homogeneous difference equation for the autocovariance function:

$$\gamma(h) = E(X_{t+h}X_t) - \mu^2 = \sum_{l=1}^{L} w_l a_l \mu - (1 - \sum_{l=1}^{L} w_l b_l)\mu^2 + \sum_{l=1}^{L} w_l b_l \gamma(h-l), \quad h \geq L.$$

With regard to the autocorrelation function, we have $r(h) = \gamma(h)/(\mu^{(2)} - \mu^2) = \phi + \sum_{l=1}^{L} w_l b_l r(h-l), \quad h \geq L$, where $\phi = (\sum_{l=1}^{L} w_l a_l \mu - (1 - \sum_{l=1}^{L} w_l b_l)\mu^2)/(\mu^{(2)} - \mu^2)$.

The necessary and sufficient condition for the non-homogeneous difference equation $r(h)$ to have a stable solution is that the roots $z_1, \ldots, z_L$ of the equation $z^L - w_1 b_1 z^{L-1} - \cdots - w_L b_L = 0$ all lie inside the unit circle. This condition, with the assumption that the cross moment is independent of t, forms condition (3) of Proposition 2. Under condition (3), the weak stationarity condition (iii) is satisfied. $\square$

Proof of Theorem 1. Consider a stationary MTD point process, i.e., the corresponding duration process has a stationary marginal distribution. Thus, the durations $X_1, \ldots, X_{N(t)}$ are a collection of identically distributed but dependent random variables. Let $T_{N(t)} = \sum_{i=1}^{N(t)} X_i$ be the last arrival time prior to time t. We have $T_{N(t)} < t < T_{N(t)+1}$, and

$$\frac{T_{N(t)}}{N(t)} < \frac{t}{N(t)} < \frac{T_{N(t)+1}}{N(t)},$$

for $N(t) \geq 1$. Note that $T_{N(t)}/N(t)$ is the average of durations $X_1, \ldots, X_{N(t)}$. Assume the stationary marginal distribution corresponds to a finite mean $\mu > 0$ and finite variance. By the strong law of large numbers for dependent non-negative random variables (Korchevsky and Petrov 2010, Theorem 4), we have that $T_{N(t)}/N(t) \to \mu$ *a.s.* $(t \to \infty)$, since as $t \to \infty$, $N(t) \to \infty$. Observing that $T_{N(t)+1}/N(t) = \{T_{N(t)+1}/(N(t)+1)\}\{(N(t)+1)/N(t)\}$, where the first term $T_{N(t)+1}/(N(t)+1) \to \mu$ *a.s.*, and the second term $(N(t)+1)/N(t) \to 1$, we can conclude that $N(t)/t \to 1/\mu$ *a.s.* by the squeeze theorem for limits. □

Proof of Proposition 3.2. The MTD point process $N(t)$ can be decomposed as $N(t) = M(t) + \Lambda(t)$ (Daley and Vere-Jones, 2003), where $M(t)$ is a zero-mean martingale, and $\Lambda(t) = \int_0^t \lambda^*(u)du$. The decomposition implies that $m(t) = E[N(t)] = E[\Lambda(t)]$. Without loss of generality, suppose $N(t) > L$. We have that

$$\begin{aligned}\int_0^t \lambda^*(u)du &= \int_0^{t_1} h^*(u)du + \cdots + \int_{t_{N(t)-1}}^{t_{N(t)}} h^*(u - t_{N(t)-1})du + \int_{t_{N(t)}}^t h^*(u - t_{N(t)})du \\ &= \sum_{i=1}^{N(t)} \int_{t_{i-1}}^{t_i} h^*(u - t_{i-1})du + \int_{t_{N(t)}}^t h^*(u - t_{N(t)})du \\ &= \sum_{i=1}^{N(t)} (-\log\{S^*(t_i - t_{i-1})\}) - \log\{S^*(t - t_{N(t)})\},\end{aligned}$$

where $t_0 = 0$. For $i = 1, \ldots, N(t)$, by Jensen's inequality, we have that

$$\begin{aligned}-\log\{S^*(t_i - t_{i-1})\} &= -\log\left\{\sum_{l=1}^{L} w_l S_l(t_i - t_{i-1} \,|\, t_{i-l} - t_{i-1-l})\right\} \\ &\le \sum_{l=1}^{L} w_l \left(-\log\{S_l(t_i - t_{i-1} \,|\, t_{i-l} - t_{i-1-l})\}\right) \\ &= \sum_{l=1}^{L} w_l \int_{t_{i-1}}^{t_i} h_l(u - t_{i-1} \,|\, t_{i-l} - t_{i-1-l})du\end{aligned}$$

Similarly, we apply the Jensen's inequality for $-\log\{S^*(t - t_{N(t)})\}$. It follows that

$$\begin{aligned}\int_0^t \lambda^*(u)du \;\le\; &\sum_{i=1}^{N(t)} \sum_{l=1}^{L} w_l \int_{t_{i-1}}^{t_i} h_l(u - t_{i-1} \,|\, t_{i-l} - t_{i-1-l})du + \\ &\sum_{l=1}^{L} w_l \int_{t_{N(t)}}^t h_l(u - t_{N(t)} \,|\, t_{N(t)-l+1} - t_{N(t)-l})du\end{aligned}$$

Then, we have that the mean-value function

$$m(t) = E\left[\int_0^t \lambda^*(u)du\right] \leq M(t_1,\ldots,t_{N(t)}) + \sum_{l=1}^{L} w_l E\left[\Lambda_l(t - t_{N(t)})\right],$$

where $M(t_1,\ldots,t_{N(t)}) = \sum_{i=1}^{N(t)}\sum_{l=1}^{L} w_l E\left[\Lambda_l(t_i - t_{i-1})\right]$, and $\Lambda_l(a - t_k) = \int_{t_k}^{a} h_l(u - t_k \mid t_{k-l+1} - t_{k-l})du$.

$\square$

Proof of Proposition 3.3. Let $(U,V) = (\alpha X, \alpha Y)$, where the joint density of (X,Y) is $f_{X,Y}(x,y) \propto (\lambda_0 + \lambda_1 x + \lambda_2 y)^{-(\alpha+1)}$. By change of variable, we obtain the joint density of (U,V), namely, $f_{U,V}(u,v) \propto (\lambda_0 + \lambda_1 u/\alpha + \lambda_2 v/\alpha)^{-(\alpha+1)}$, with normalizing constant $C = \int_0^\infty\int_0^\infty(\lambda_0+\lambda_1 u/\alpha+\lambda_2 v/\alpha)^{-(\alpha+1)}dudv = \alpha\lambda_0^{-(\alpha-1)}\{(\alpha-1)\lambda_1\lambda_2\}^{-1}$. It follows that the marginal density of U is $f_U(u) = C^{-1}\int_0^\infty \alpha^{-2}(\lambda_0 + \lambda_1 u/\alpha+\lambda_2 v/\alpha)^{-(\alpha+1)}dv = (\alpha-1)(\lambda_0\alpha)^{-1}\lambda_1\{1+(\lambda_0\alpha)^{-1}\lambda_1 u\}^{-\alpha}$. Since u and v are symmetric in the joint density, the marginal density $f_V(v) = (\alpha-1)(\lambda_0\alpha)^{-1}\lambda_2\{1+(\lambda_0\alpha)^{-1}\lambda_2 v\}^{-\alpha}$. By definition, $f_{U|V}(u \mid v) = f_{U,V}(u,v)/f_V(v) = \alpha\lambda_1(\alpha\lambda_0 + \lambda_2 v)^{-1}\{1 + \lambda_1 u(\alpha\lambda_0 + \lambda_2 v)^{-1}\}^{-(\alpha+1)}$. $\square$

Proof of Proposition 3.4. Let $S^*(t - t_{N(t)})$ be the conditional survival function corresponding to the conditional duration distribution of the scaled-Lomax MTDPP. Without loss of generality, suppose $N(t) > L$. Then we have that

$$\begin{aligned}
S^*(t - t_{N(t)}) &= \sum_{l=1}^{L} w_l \left(1 + \frac{t - t_{N(t)}}{\alpha_l\phi + t_{N(t)-l+1} - t_{N(t)-l}}\right)^{-\alpha_l} \\
&= \sum_{l=1}^{L} w_l \left\{\left(1 + \frac{t - t_{N(t)}}{\alpha_l\phi + t_{N(t)-l+1} - t_{N(t)-l}}\right)^{-(\alpha_l\phi_l + t_{N(t)-l+1} - t_{N(t)-l})}\right\}^{1/\phi_l} \\
&\quad \times \left(1 + \frac{t - t_{N(t)}}{\alpha_l\phi_l + t_{N(t)-l+1} - t_{N(t)-l}}\right)^{(t_{N(t)-l+1} - t_{N(t)-l})/\phi_l}.
\end{aligned}$$

As $\alpha_l \to \infty$ for all l, for each component, the limits of the first term and the second term are $\exp(-(t - t_{N(t)})\phi_l^{-1})$ and 1, respectively. Then we have that in the limit of all α_l, the conditional duration distribution of the scaled-Lomax MTDPP is $\sum_{l=1}^{L} w_l \phi_l^{-1} \exp(-(t - t_{N(t)})\phi_l^{-1})$ for $N(t) > L$. If $\phi_l = \phi$ for all l, the limiting distribution becomes $\phi^{-1} \exp(-(t - t_{N(t)})\phi^{-1})$, which is an exponential distribution with rate ϕ^{-1}. $\square$

Proof of Proposition 4.1. We consider a univariate spatial process $\{Z(\boldsymbol{v}), \boldsymbol{v} \in \mathcal{D}\}$, where $Z(\boldsymbol{v})$ takes values in $\mathcal{X} \subseteq \mathbb{R}$, and $\mathcal{D} \subset \mathbb{R}^p, p \geq 1$. Let $\mathcal{S} \subset \mathcal{D}$ be a reference set. Without loss of generality, we consider the continuous case, i.e., $Z(\boldsymbol{v})$ has a continuous distribution for which its density exists, for all $\boldsymbol{v} \in \mathcal{D}$. To verify the proposition, we partition the domain $\mathcal{D}$ into the reference set $\mathcal{S}$ and the nonreference set $\mathcal{U}$.

Given any $\boldsymbol{v} \in \mathcal{D}$, consider a bivariate random vector indexed at $\boldsymbol{v}$, denoted as $(U_{\boldsymbol{v},l}, V_{\boldsymbol{v},l})$ taking values in $\mathcal{X} \times \mathcal{X}$. We denote $f_{\boldsymbol{v},l}$ as the conditional density of $U_{\boldsymbol{v},l}$ given $V_{\boldsymbol{v},l}$, and $f_{U_{\boldsymbol{v},l}}, f_{V_{\boldsymbol{v},l}}$ as the marginal densities of $U_{\boldsymbol{v},l}$, $V_{\boldsymbol{v},l}$, respectively. Using the proposition assumption that $f_Z = f_{U_{\boldsymbol{v},l}} = f_{V_{\boldsymbol{v},l}}$, we have that

$$\int_{\mathcal{X}} f_{\boldsymbol{v},l}(u \mid v) f_Z(v) dv = \int_{\mathcal{X}} f_{\boldsymbol{v},l}(u \mid v) f_{V_{\boldsymbol{v},l}}(v) dv = f_{U_{\boldsymbol{v},l}}(u) = f_Z(u), \tag{A.1}$$

for every $\boldsymbol{v} \in \mathcal{D}$ and for all l.

We first prove the result for the reference set $\mathcal{S}$. By the model assumption, locations in $\mathcal{S}$ are ordered. In this regard, using the proposition assumptions, we can show that $Z(\boldsymbol{s}) \sim f_Z$ for all $\boldsymbol{s} \in \mathcal{S}$ by applying Proposition 2.1.

Turning to the nonreference set $\mathcal{U}$. Let $g_{\boldsymbol{u}}(z(\boldsymbol{u}))$ be the marginal density of $Z(\boldsymbol{u})$ for every $\boldsymbol{u} \in \mathcal{U}$. Denote by $p(\boldsymbol{z}_{\mathrm{Ne}(\boldsymbol{u})})$ the joint density for the random vector $\boldsymbol{z}_{\mathrm{Ne}(\boldsymbol{u})}$ where $\mathrm{Ne}(\boldsymbol{u}) = \{\boldsymbol{u}_{(1)}, \ldots, \boldsymbol{u}_{(L)}\} \subset \mathcal{S}$, so every element of $\boldsymbol{Z}_{\mathrm{Ne}}$ has marginal density f_Z. Then, the marginal density for $Z(\boldsymbol{u})$ is given by:

$$
\begin{aligned}
g_{\boldsymbol{u}}(z(\boldsymbol{u})) &= \int_{\mathcal{X}^L} p(z(\boldsymbol{u}) \mid \boldsymbol{z}_{\mathrm{Ne}(\boldsymbol{u})}) p(\boldsymbol{z}_{\mathrm{Ne}(\boldsymbol{u})}) \prod_{\{\boldsymbol{s}_i \in \mathrm{Ne}(\boldsymbol{u})\}} d(z(\boldsymbol{s}_i)) \\
&= \sum_{l=1}^{L} w_l(\boldsymbol{v}) \int_{\mathcal{X}^L} f_{\boldsymbol{v},l}(z(\boldsymbol{u}) \mid z(\boldsymbol{u}_{(l)})) p(\boldsymbol{z}_{\mathrm{Ne}(\boldsymbol{u})}) \prod_{\{\boldsymbol{s}_i \in \mathrm{Ne}(\boldsymbol{u}), \boldsymbol{s}_i \neq \boldsymbol{u}_{(l)}\}} d(z(\boldsymbol{s}_i)) \\
&= \sum_{l=1}^{L} w_l(\boldsymbol{v}) \int_{\mathcal{X}} f_{\boldsymbol{v},l}(z(\boldsymbol{u}) \mid z(\boldsymbol{u}_{(l)})) f_Z(z(\boldsymbol{u}_{(l)})) d(z(\boldsymbol{u}_{(l)})) \\
&= f_Z(z(\boldsymbol{u})),
\end{aligned}
$$

where the second-to-last equality holds by the result that $Z(\boldsymbol{s}) \sim f_Z$ for all $\boldsymbol{s} \in \mathcal{S}$ and $\mathrm{Ne}(\boldsymbol{u}) \subset \mathcal{S}$ for every $\boldsymbol{u} \in \mathcal{U}$. The last equality follows from (A.1).

$\square$

Proof of Proposition 4.2. We verify the proposition by partitioning the domain $\mathcal{D}$ into the reference set $\mathcal{S}$ and the nonreference set $\mathcal{U}$. We first prove by induction the result for the joint density $p(\boldsymbol{z}_{\mathcal{S}})$ over $\mathcal{S}$. Then to complete the proof, it suffices to show that for every location $\boldsymbol{u} \in \mathcal{U}$, the joint density $p(\boldsymbol{z}_{\mathcal{U}_1})$ is a mixture of multivariate Gaussian distributions, where $\mathcal{U}_1 = \mathcal{S} \cup \{\boldsymbol{u}\}$.

Without loss of generality, we assume $\mu = 0$ for the stationary GNNMP, i.e., the GNNMP has invariant marginal $f_Z(z) = N(z \,|\, 0, \sigma^2)$. The conditional density for the reference set is

$$
p(z(\boldsymbol{s}_i) | \boldsymbol{z}_{\mathrm{Ne}(\boldsymbol{s}_i)}) = \sum_{l=1}^{i_L} w_l(\boldsymbol{s}_i) N(z(\boldsymbol{s}_i) | \rho_l(\boldsymbol{s}_i) z(\boldsymbol{s}_{(il)}), \sigma^2(1 - (\rho_l(\boldsymbol{s}_i))^2)),
$$

where for, $i = 2, \ldots, L$, $i_L = i - 1$, and for $i > L$, $i_L = L$. For each i, we denote as $\{w_{i,l_i}\}_{l_i=1}^{i_L}$ the vector of mixture weights, as $\{\rho_{i,l_i}\}_{l_i=1}^{i_L}$ the vector of the correlation coefficients, and as $\{z_{i,l_i}\}_{l_i=1}^{i_L}$ the vector of the nearest neighbors of $z_i \equiv z(\boldsymbol{s}_i)$, for $i \geq 2$, where $w_{i,l_i} \equiv w_{l_i}(\boldsymbol{s}_i)$, $\rho_{i,l_i} \equiv \rho_{l_i}(\boldsymbol{s}_i)$, $z_{i,l_i} \equiv z(\boldsymbol{s}_{(i,l_i)})$.

Let $z_1 \equiv z(\boldsymbol{s}_1)$. We denote by $\boldsymbol{z}_{1:k} = (z(\boldsymbol{s}_1), \ldots, z(\boldsymbol{s}_k))$ the realization of $Z(\boldsymbol{s})$ over locations $(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_k)^\top$ for $k \geq 2$, and use $\boldsymbol{z}_{1:k}^{-z_j}$ to denote the random vector $\boldsymbol{z}_{1:k}$ with element z_j removed, $1 \leq j \leq k$. In the following, for a vector $\boldsymbol{a} = (a_1, \ldots, a_m)^\top$, we have that $\boldsymbol{a}c = (a_1 c, \ldots, a_m c)^\top$, where c is a scalar.

Take $Z_1 \sim N(z_1 \,|\, 0, \sigma^2)$. When $i = 2$, $i_L = 1$ and $w_{2,1} = 1$. The joint density of $\boldsymbol{z}_{1:2}$ is $p(\boldsymbol{z}_{1:2}) = N(z_2|\rho_{2,1}z_1), \sigma^2(1-\rho_{2,1}^2))N(z_1|0,\sigma^2) = N(\boldsymbol{z}_{1:2}|\boldsymbol{0}, \sigma^2\boldsymbol{\Omega}_{2,1})$, where $\boldsymbol{\Omega}_{2,1} = \left(\begin{smallmatrix} 1 & \rho_{2,1} \\ \rho_{2,1} & 1 \end{smallmatrix}\right)$. The joint density of $\boldsymbol{z}_{1:3}$ is

$$
\begin{aligned}
p(\boldsymbol{z}_{1:3}) &= p_3(z_3|\boldsymbol{z}_{1:2})p(\boldsymbol{z}_{1:2}) \\
&= \sum_{l_3=1}^{2} w_{3,l_3} N(z_3|\rho_{3,l_3} z_{3,l_3}, \sigma^2(1-\rho_{3,l_3}^2)) N(\boldsymbol{z}_{1:2}|\boldsymbol{0}, \sigma^2\boldsymbol{\Omega}_{2,1}) \\
&= \sum_{l_3=1}^{2} w_{3,l_3} N(z_3|\rho_{3,l_3} z_{3,l_3}, \sigma^2(1-\rho_{3,l_3}^2)) N(\boldsymbol{z}_{1:2}^{-z_{3,l_3}}|\rho_{2,l_2} z_{3,l_3}, \sigma^2(1-\rho_{2,1}^2)) N(z_{3,l_3}|0,\sigma^2) \\
&= \sum_{l_3=1}^{2} w_{3,l_3} N((z_3, \boldsymbol{z}_{1:2}^{-z_{3,l_3}})^\top|\boldsymbol{m}_{3,l_3} z_{3,l_3}, \boldsymbol{V}_{3,l_3}) N(z_{3,l_3}|0,\sigma^2)
\end{aligned}
$$

where $\boldsymbol{m}_{3,l_3} = (\rho_{3,l_3}, \rho_{2,1})^\top$, and $\boldsymbol{V}_{3,l_3} = \left(\begin{smallmatrix} \sigma^2(1-\rho_{3,l_3}^2) & 0 \\ 0 & \sigma^2(1-\rho_{2,1}^2) \end{smallmatrix}\right)$. The last equality follows from the fact that a product of conditionally independent Gaussian densities is a Gaussian density. By the properties of the Gaussian distribution and the property of the model that has a stationary marginal $N(0, \sigma^2)$, for each l_3, we have that $N(\tilde{\boldsymbol{z}}_{1:3,l_3}|\boldsymbol{0}, \sigma^2\boldsymbol{R}_{3,l_3}) = N((z_3, \boldsymbol{z}_{1:2}^{-z_{3,l_3}})^\top|\boldsymbol{m}_{3,l_3} z_{3,l_3}, \boldsymbol{V}_{3,l_3}) N(z_{3,l_3}|0,\sigma^2)$, where $\tilde{\boldsymbol{z}}_{1:3,l_3} = (z_3, \boldsymbol{z}_{1:2}^{-z_{3,l_3}}, z_{3,l_3})^\top$, with the following partition relevant to the vector $\tilde{\boldsymbol{z}}_{1:3,l_3}$, $\tilde{\boldsymbol{z}}_{1:3,l_3} = \begin{pmatrix} (z_3, \boldsymbol{z}_{1:2}^{-z_{3,l_3}})^\top \\ z_{3,l_3} \end{pmatrix}$, $E(\tilde{\boldsymbol{z}}_{1:3,l_3}) = \begin{pmatrix} \boldsymbol{0} \\ 0 \end{pmatrix}$, $\boldsymbol{R}_{3,l_3} = \begin{pmatrix} \boldsymbol{R}_{3,l_3}^{(11)} & \boldsymbol{R}_{3,l_3}^{(12)} \\ \boldsymbol{R}_{3,l_3}^{(21)} & \boldsymbol{R}_{3,l_3}^{(22)} \end{pmatrix}$, where $\boldsymbol{R}_{3,l_3}^{(22)} = 1$ corresponds to z_{3,l_3}. It follows that

$$
\begin{aligned}
\boldsymbol{m}_{3,l_3} z_{3,l_3} &= E((Z_3, \tilde{\boldsymbol{Z}}_{1:2}^{-Z_{3,l_3}}) \,|\, Z_{3,l_3} = z_{3,l_3}) = \boldsymbol{R}_{3,l_3}^{(12)} z_{3,l_3}, \\
\boldsymbol{V}_{3,l_3} &= \sigma^2(\boldsymbol{R}_{3,l_3}^{(11)} - \boldsymbol{R}_{3,l_3}^{(12)}\boldsymbol{R}_{3,l_3}^{(21)}).
\end{aligned} \tag{A.2}
$$

From (A.2), we obtain $\boldsymbol{m}_{3,l_3} = \boldsymbol{R}^{(12)}_{3,l_3}$ and $\boldsymbol{R}_{3,l_3} = \begin{pmatrix} 1 & \rho_{2,1}\rho_{3,l_3} & \rho_{3,l_3} \\ \rho_{2,1}\rho_{3,l_3} & 1 & \rho_{2,1} \\ \rho_{3,l_3} & \rho_{2,1} & 1 \end{pmatrix}$ for $l_3 = 1, 2$. Then we reorder $\tilde{\boldsymbol{z}}_{1:3,l_3}$ with a matrix $\boldsymbol{B}_{3,l_3}$ such that $\boldsymbol{z}_{1:3} = \boldsymbol{B}_{3,l_3}\tilde{\boldsymbol{z}}_{1:3,l_3}$. It follows that $\boldsymbol{\Omega}_{3,l_31} = \boldsymbol{B}_{3,l_3}\boldsymbol{R}_{3,l_3}\boldsymbol{B}^T_{3,l_3}$. The joint density is $p(\boldsymbol{z}_{1:3}) = \sum_{l_3=1}^{2} w_{3,l_3} N(\boldsymbol{z}_{1:3}|\mathbf{0}, \sigma^2\boldsymbol{\Omega}_{3,l_31})$. Similarly, the joint density of $\boldsymbol{z}_{1:4}$ is given by

$$
\begin{aligned}
p(\boldsymbol{z}_{1:4}) &= p_4(z_4|\boldsymbol{z}_{1:3})p(\boldsymbol{z}_{1:3}) \\
&= \sum_{l_4=1}^{3} w_{4,l_4} N(z_4|\rho_{4,l_4} z_{4,l_4}, \sigma^2(1-\rho^2_{4,l_4})) \sum_{l_3=1}^{2} w_{3,l_3} N(\boldsymbol{z}_{1:3}|\mathbf{0}, \sigma^2\boldsymbol{\Omega}_{3,l_31}) \\
&= \sum_{l_4=1}^{3}\sum_{l_3=1}^{2} w_{4,l_4} w_{3,l_3} N(z_4|\rho_{4,l_4} z_{4,l_4}, \sigma^2(1-\rho^2_{4,l_4})) \\
&\qquad N((\boldsymbol{z}_{1:3}^{-z_{4,l_4}})^\top \mid \tilde{\boldsymbol{\Omega}}^{(12)}_{3,l_31} z_{4,l_4}, \sigma^2(\tilde{\boldsymbol{\Omega}}^{(11)}_{3,l_31} - \tilde{\boldsymbol{\Omega}}^{(12)}_{3,l_31}\tilde{\boldsymbol{\Omega}}^{(21)}_{3,l_31})) N(z_{4,l_4}|0, \sigma^2) \\
&= \sum_{l_4=1}^{3}\sum_{l_3=1}^{2} w_{4,l_4} w_{3,l_3} N((z_4, \boldsymbol{z}_{1:3}^{-z_{4,l_4}})^\top|\boldsymbol{m}_{4,l_4l_3} z_{4,l_4}, \boldsymbol{V}_{4,l_4l_3}) N(z_{4,l_4}|0, \sigma^2),
\end{aligned}
$$

where $\tilde{\boldsymbol{\Omega}}_{3,l_31} = \tilde{\boldsymbol{B}}_{4,l_4}\boldsymbol{\Omega}_{3,l_31}\tilde{\boldsymbol{B}}^\top_{4,l_4}$, and $\tilde{\boldsymbol{B}}_{4,l_4}$ is a rotation matrix such that the vector $(\boldsymbol{z}_{1:3}^{-z_{4,l_4}}, z_{4,l_4})^\top = \tilde{\boldsymbol{B}}_{4,l_4}\boldsymbol{z}_{1:3}$. We partition the matrix $\tilde{\boldsymbol{\Omega}}_{3,l_31}$ such that $\tilde{\boldsymbol{\Omega}}^{(11)}_{3,l_41}$ and $\tilde{\boldsymbol{\Omega}}^{(22)}_{3,l_41}$ correspond to $\boldsymbol{z}_{1:3}^{-z_{4,l_4}}$ and z_{4,l_4}, respectively. We have that for $l_3 = 1, 2,\ l_4 = 1, 2, 3$,

$$
N(\tilde{\boldsymbol{z}}_{1:4,l_4}|\mathbf{0}, \sigma^2\boldsymbol{R}_{4,l_4l_3}) = N((z_4, \boldsymbol{z}_{1:3}^{-z_{4,l_4}})^\top|\boldsymbol{m}_{4,l_4l_3} z_{4,l_4}, \boldsymbol{V}_{4,l_4l_3}) N(z_{4,l_4}|0, \sigma^2),
$$

where $\tilde{\boldsymbol{z}}_{1:4,l_4} = (z_4, \boldsymbol{z}_{1:3}^{-z_{4,l_4}}, z_{4,l_4})^\top$, $\boldsymbol{m}_{4,l_4l_3} = (\rho_{4,l_4}, (\tilde{\boldsymbol{\Omega}}^{(12)}_{3,l_31})^\top)^\top$, and

$$
\begin{aligned}
\boldsymbol{V}_{4,l_4l_3} &= \begin{pmatrix} \sigma^2(1-\rho^2_{4,l_4}) & \mathbf{0}^T \\ \mathbf{0} & \sigma^2(\tilde{\boldsymbol{\Omega}}^{(11)}_{3,l_31} - \tilde{\boldsymbol{\Omega}}^{(12)}_{3,l_31}\tilde{\boldsymbol{\Omega}}^{(21)}_{3,l_31}) \end{pmatrix}, \\
\boldsymbol{R}^{(12)}_{4,l_4l_3} &= (\boldsymbol{R}^{(21)}_{4,l_4l_3})^\top = \boldsymbol{m}_{4,l_4l_3}, \quad \boldsymbol{R}^{(11)}_{4,l_4l_3} = \boldsymbol{V}_{4,l_4l_3}/\sigma^2 + \boldsymbol{m}_{4,l_4l_3}\boldsymbol{m}^T_{4,l_4l_3}.
\end{aligned}
$$

We reorder $\tilde{\boldsymbol{z}}_{1:4,l_4}$ with a matrix $\boldsymbol{B}_{4,l_4}$ such that $\boldsymbol{z}_{1:4} = \boldsymbol{B}_{4,l_4}\tilde{\boldsymbol{z}}_{1:4,l_4}$ and let $\boldsymbol{\Omega}_{4,l_4l_31} = \boldsymbol{B}_{4,l_4}\boldsymbol{R}_{4,l_4l_3}\boldsymbol{B}^T_{4,l_3}$. We obtain the joint density $p(\boldsymbol{z}_{1:4}) = \sum_{l_4=1}^{3}\sum_{l_3=1}^{2} w_{4,l_4} w_{3,l_3} N(\boldsymbol{z}_{1:4}|\mathbf{0},$

$\sigma^2\boldsymbol{\Omega}_{4,l_4l_31})$. Applying the above technique iteratively for $p(\boldsymbol{z}_{1:j})$ for $5 \leq j \leq k$, we obtain the joint density $p(\boldsymbol{z}_{1:k}) \equiv p(\boldsymbol{z}_{\mathcal{S}})$, for $k \geq 2$, namely,

$$p(\boldsymbol{z}_{1:k}) = \sum_{l_k=1}^{k_L} \cdots \sum_{l_2=1}^{2_L} w_{k,l_k} \dots w_{3,l_3} w_{2,l_2} N(\boldsymbol{z}_{1:k}|\boldsymbol{0}, \sigma^2\boldsymbol{\Omega}_{k,l_k\dots l_3l_2})$$

where $k_L := (k-1) \wedge L$, $w_{2,1} = 1$, and for $k \geq 3$,

$$\begin{aligned}
\tilde{\boldsymbol{\Omega}}_{k-1,l_{k-1}\dots l_31} &= \tilde{\boldsymbol{B}}_{k,l_{k-1}} \boldsymbol{\Omega}_{k-1,l_{k-1}\dots l_31} \tilde{\boldsymbol{B}}_{k,l_k}, \quad \boldsymbol{m}_{k,l_k\dots l_1} = (\rho_{k,l_k}, (\tilde{\boldsymbol{\Omega}}^{(12)}_{k-1,l_{k-1}\dots l_31})^\top)^\top, \\
\boldsymbol{V}_{k,l_k\dots l_3} &= \begin{pmatrix} \sigma^2(1-\rho^2_{k,l_k}) & \boldsymbol{0} \\ \boldsymbol{0}^T & \sigma^2(\tilde{\boldsymbol{\Omega}}^{(11)}_{k-1,l_{k-1}\dots l_31} - \tilde{\boldsymbol{\Omega}}^{(12)}_{k-1,l_{k-1}\dots l_31}\tilde{\boldsymbol{\Omega}}^{(21)}_{k-1,l_{k-1}\dots l_31}) \end{pmatrix}, \\
\boldsymbol{R}^{(12)}_{k,l_k\dots l_3} &= (\boldsymbol{R}^{(21)}_{k,l_k\dots l_3})^\top = \boldsymbol{m}_{k,l_k\dots l_3}, \quad \boldsymbol{R}^{(11)}_{k,l_k\dots l_3} = \boldsymbol{V}_{k,l_k\dots l_3}/\sigma^2 + \boldsymbol{m}_{k,l_k\dots l_3}\boldsymbol{m}^\top_{k,l_k\dots l_3}, \\
\boldsymbol{\Omega}_{k,l_k\dots l_31} &= \boldsymbol{B}_{k,l_k}\boldsymbol{R}_{k,l_k\dots l_3}\boldsymbol{B}^T_{k,l_k},
\end{aligned}$$

where $\tilde{\boldsymbol{B}}_{k,l_k}$ is the rotation matrix such that $(\boldsymbol{z}^{-z_{k,l_k}}_{1:(k-1)}, z_{k,l_k})^\top = \tilde{\boldsymbol{B}}_{k,l_k}\boldsymbol{z}_{1:(k-1)}$, and $\boldsymbol{B}_{k,l_k}$ is the rotation matrix such that the vector $\boldsymbol{z}_{1:k} = \boldsymbol{B}_{k,l_k}\tilde{\boldsymbol{z}}_{1:k,l_k}$, where $\tilde{\boldsymbol{z}}_{1:k,l_k} = (z_k, \boldsymbol{z}^{-z_{k,l_k}}_{1:(k-1)}, z_{k,l_k})^\top$.

To complete the proof, what remains to be shown is that the density $p(\boldsymbol{z}_{\mathcal{U}_1})$ is a mixture of multivariate Gaussian distributions, where $\mathcal{U}_1 = \mathcal{S} \cup \{\boldsymbol{u}\}$. We have that $p(\boldsymbol{z}_{\mathcal{U}_1}) = \sum_{l=1}^{L} w_l(\boldsymbol{u}) N(z(\boldsymbol{u}) \mid \rho_l(\boldsymbol{u}) z(\boldsymbol{u}_{(l)}), \sigma^2(1-(\rho_l(\boldsymbol{u}))^2))\, p(\boldsymbol{z}_{1:k})$, where $z(\boldsymbol{u}_{(l)})$ is an element of $\boldsymbol{z}_{1:k}$, for $l = 1, \dots, L$. We can express each component density $N(\boldsymbol{z}_{1:k} \,|\, \boldsymbol{0}, \sigma^2\Omega_{k,l_k\dots l_31})$ of the joint density $p(\boldsymbol{z}_{1:k})$ as the product of a Gaussian density of $\boldsymbol{Z}^{-Z(\boldsymbol{u}_{(l)})}_{1:k}$ conditional on $Z(\boldsymbol{u}_{(l)}) = z(\boldsymbol{u}_{(l)})$ and a Gaussian density of $Z(\boldsymbol{u}_{(l)})$. Using the approach in deriving the joint density over $\mathcal{S}$, we can show that $p(\boldsymbol{z}_{\mathcal{U}_1})$ is a mixture of multivariate Gaussian distributions.

□

Proof of Proposition 4.3. For an NNMP $Z(\boldsymbol{v})$, The conditional probability that $Z(\boldsymbol{v})$ is greater than z given its neighbors $\boldsymbol{Z}_{\mathrm{Ne}(\boldsymbol{v})} = \boldsymbol{z}_{\mathrm{Ne}(\boldsymbol{v})}$, where $\boldsymbol{z}_{\mathrm{Ne}(\boldsymbol{v})} =$

$(z_{\boldsymbol{v}_{(1)}}, \ldots, z_{\boldsymbol{v}_{(L)}})$, is

$$P(Z(\boldsymbol{v}) > z \,|\, \boldsymbol{Z}_{\mathrm{Ne}(\boldsymbol{v})} = \boldsymbol{z}_{\mathrm{Ne}(\boldsymbol{v})}) = \sum_{l=1}^{L} w_l(\boldsymbol{v}) P(Z(\boldsymbol{v}) > z \,|\, Z(\boldsymbol{v}_{(l)}) = z(\boldsymbol{v}_{(l)})),$$

where the conditional probability $P(Z(\boldsymbol{v}) > z \,|\, Z(\boldsymbol{v}_{(l)}) = z(\boldsymbol{v}_{(l)}))$ corresponds to the bivariate random vector $(U_{\boldsymbol{v},l}, V_{\boldsymbol{v},l})$. If U_l is stochastically increasing in V_l for all l, by the assumption that the sequence $(U_{\boldsymbol{v},l}, V_{\boldsymbol{v},l})$ is built from the base random vectors (U_l, V_l) for all l, we have that $Z(\boldsymbol{v})$ is stochastically increasing in $\boldsymbol{Z}_{\mathrm{Ne}(\boldsymbol{v})}$ for every $\boldsymbol{v} \in \mathcal{D}$, i.e. $P(Z(\boldsymbol{v}) > z \,|\, \boldsymbol{Z}_{\mathrm{Ne}(\boldsymbol{v})} = \boldsymbol{z}_{\mathrm{Ne}(\boldsymbol{v})}) \leq P(Z(\boldsymbol{v}) > z \,|\, \boldsymbol{Z}_{\mathrm{Ne}(\boldsymbol{v})} = \boldsymbol{z}'_{\mathrm{Ne}(\boldsymbol{v})})$ for all $\boldsymbol{z}_{\mathrm{Ne}(\boldsymbol{v})}$ and $\boldsymbol{z}'_{\mathrm{Ne}(\boldsymbol{v})}$ in the support of $\boldsymbol{Z}_{\mathrm{Ne}(\boldsymbol{v})}$, such that $z_{\boldsymbol{v}_{(l)}} \leq z'_{\boldsymbol{v}_{(l)}}$ for all l.

Let $F_{Z(\boldsymbol{v})}$ and $F_{Z(\boldsymbol{v}_{(1)}),\ldots,Z(\boldsymbol{v}_{(L)})}$ be the distribution functions of $Z(\boldsymbol{v})$ and $\boldsymbol{Z}_{\mathrm{Ne}(\boldsymbol{v})}$. Denote by $S_{Z(\boldsymbol{v}_{(1)}),\ldots,Z(\boldsymbol{v}_{(L)})}(z_1, \ldots, z_L) = P(Z(\boldsymbol{v}_{(1)}) > z_1, \ldots, Z(\boldsymbol{v}_{(L)}) > z_L)$ the joint survival probability. Then for every $\boldsymbol{v} \in \mathcal{D}$ and $q \in (0,1)$,

$$\begin{aligned}
&P(Z(\boldsymbol{v}) > F^{-1}_{Z(\boldsymbol{v})}(q) \mid Z(\boldsymbol{v}_{(1)}) > F^{-1}_{Z(\boldsymbol{v}_{(1)})}(q), \ldots, Z(\boldsymbol{v}_{(L)}) > F^{-1}_{Z(\boldsymbol{v}_{(L)})}(q)) \\
&= \Bigg\{ \int_{F^{-1}_{Z(\boldsymbol{v}_{(1)})}(q)}^{\infty} \cdots \int_{F^{-1}_{Z(\boldsymbol{v}_{(L)})}(q)}^{\infty} P(Z(\boldsymbol{v}) > F^{-1}_{Z(\boldsymbol{v})}(q) \mid Z(\boldsymbol{v}_{(1)}) = z_1, \ldots, Z(\boldsymbol{v}_{(L)}) = z_L) \\
&\quad dF_{Z(\boldsymbol{v}_{(1)}),\ldots,Z(\boldsymbol{v}_{(L)})}(z_1, \ldots, z_L) \Bigg\} \Big/ C \\
&\geq \Bigg\{ \int_{F^{-1}_{Z(\boldsymbol{v}_{(1)})}(q)}^{\infty} \cdots \int_{F^{-1}_{Z(\boldsymbol{v}_{(L)})}(q)}^{\infty} P(Z(\boldsymbol{v}) > F^{-1}_{Z(\boldsymbol{v})}(q) \mid Z(\boldsymbol{v}_{(1)}) = F^{-1}_{Z(\boldsymbol{v}_{(1)})}(q), \ldots, \\
&\quad Z(\boldsymbol{v}_{(L)}) = F^{-1}_{Z(\boldsymbol{v}_{(L)})}(q))\, dF_{Z(\boldsymbol{v}_{(1)}),\ldots,Z(\boldsymbol{v}_{(L)})}(z_1, \ldots, z_L) \Bigg\} \Big/ C \\
&= P(Z(\boldsymbol{v}) > F^{-1}_{Z(\boldsymbol{v})}(q) \mid Z(\boldsymbol{v}_{(1)}) = F^{-1}_{Z(\boldsymbol{v}_{(1)})}(q), \ldots, Z(\boldsymbol{v}_{(L)}) = F^{-1}_{Z(\boldsymbol{v}_{(L)})}(q)) \\
&= \sum_{l=1}^{L} w_l(\boldsymbol{v}) P(Z(\boldsymbol{v}) > F^{-1}_{U_{\boldsymbol{v},l}}(q) \mid Z(\boldsymbol{v}_{(l)}) = F^{-1}_{V_{\boldsymbol{v},l}}(q)),
\end{aligned} \tag{A.3}$$

where $C = S_{Z(\boldsymbol{v}_{(1)}),\ldots,Z(\boldsymbol{v}_{(L)})}(F^{-1}_{Z(\boldsymbol{v}_{(1)})}(q), \ldots, F^{-1}_{Z(\boldsymbol{v}_{(L)})}(q))$, and the first inequality follows from the stochastically increasing positive dependence of $Z(\boldsymbol{v})$ given $\boldsymbol{Z}_{\mathrm{Ne}(\boldsymbol{v})}$.

Taking $q \to 1^-$ on both sides of (A.3), we obtain

$$\lambda_{\mathcal{H}}(\boldsymbol{v}) \geq \sum_{l=1}^{L} w_l(\boldsymbol{v}) \lim_{q \to 1^-} P(Z(\boldsymbol{v}) > F_{U_{\boldsymbol{v},l}}^{-1}(q) \mid Z(\boldsymbol{v}_{(l)}) = F_{V_{\boldsymbol{v},l}}^{-1}(q)),$$

where $F_{U_{\boldsymbol{v},l}}$ and $F_{V_{\boldsymbol{v},l}}$ are the marginal distribution functions of $(U_{\boldsymbol{v},l}, V_{\boldsymbol{v},l})$.

Similarly, we can obtain the lower bound for $\lambda_{\mathcal{L}}(\boldsymbol{v})$.

□

Proof of Corollary 1. We prove the result for $\lambda_{\mathcal{L}}(\boldsymbol{v})$. The result for $\lambda_{\mathcal{H}}(\boldsymbol{v})$ is obtained in a similar way.

Consider a bivariate distribution F_{U_l,V_l} for random vector (U_l, V_l), with marginal distributions $F_{U_l} = F_{V_l} = F_l$ and marginal densities $f_{U_l} = f_{V_l} = f_l$, for all l. The lower tail dependence coefficient is expressed as $\lambda_{\mathcal{L},l} = \lim_{q \to 0^+} \frac{F_{U_l,V_l}(F_l^{-1}(q), F_l^{-1}(q))}{F_{U_l,V_l}(F_l^{-1}(q))}$ with $q \in [0,1]$. If F_{U_l,V_l} has first order partial derivatives, applying the L'Hopital's rule, we obtain

$$\begin{aligned}
\lambda_{\mathcal{L},l} &= \lim_{q \to 0^+} \frac{\partial F_{U_l,V_l}/\partial V_l(F_l^{-1}(q), F_l^{-1}(q)) + \partial F_{U_l,V_l}/\partial U_l(F_l^{-1}(q), F_l^{-1}(q))}{f_l(F_l^{-1}(q))} \\
&= \lim_{q \to 0^+} P(U_l \leq F_l^{-1}(q) \mid V_l = F_l^{-1}(q)) + \lim_{q \to 0^+} P(V_l \leq F_l^{-1}(q) \mid U_l = F_l^{-1}(q)).
\end{aligned}$$

The above is a reproduced result from Theorem 8.57 of Joe (2014). If (U_l, V_l) is exchangeable, we have $\lambda_{\mathcal{L},l} = 2 \lim_{q \to 0^+} P(U_l \leq F_l^{-1}(q) \,|\, V_l = F_l^{-1}(q))$. If the sequences $(U_{\boldsymbol{v},l}, V_{\boldsymbol{v},l})$ of an NNMP model are built from the base random vectors (U_l, V_l). By our assumption that $F_{U_l} = F_{V_l}$ for all l, the marginal distributions of $(U_{\boldsymbol{v},l}, V_{\boldsymbol{v},l})$ extended from (U_l, V_l) are $F_{\boldsymbol{v},l} = F_{U_{\boldsymbol{v},l}} = F_{V_{\boldsymbol{v},l}}$ for all $\boldsymbol{v}$ and all l. Then we have $\lambda_{\mathcal{L},l}(\boldsymbol{v}) = 2 \lim_{q \to 0^+} P(U_{\boldsymbol{v},l} \leq F_{\boldsymbol{v},l}^{-1}(q) \,|\, V_{\boldsymbol{v},l} = F_{\boldsymbol{v},l}^{-1}(q))$. Using the result of

Proposition 3, we obtain

$$\lambda_{\mathcal{L}}(\boldsymbol{v}) \geq \sum_{l=1}^{L} w_l(\boldsymbol{v}) \lim_{q \to 0^+} P(U_{\boldsymbol{v},l} \leq F_{\boldsymbol{v},l}^{-1}(q) \mid V_{\boldsymbol{v},l} = F_{\boldsymbol{v},l}^{-1}(q)) = \sum_{l=1}^{L} w_l(\boldsymbol{v}) \lambda_{\mathcal{L},l}(\boldsymbol{v})/2.$$

□

Proof of Proposition 4.4. By the assumption that U_l is stochastically increasing in V_l and that $(U_{\boldsymbol{v},l}, V_{\boldsymbol{v},l})$ is constructed based on (U_l, V_l), $U_{\boldsymbol{v},l}$ is stochastically increasing in $V_{\boldsymbol{v},l}$ for all $\boldsymbol{v} \in \mathcal{D}$ and for all l. Then for $(Z(\boldsymbol{v}), Z(\boldsymbol{v}_{(l)}))$ with respect to the bivariate distribution of $(U_{\boldsymbol{v},l}, V_{\boldsymbol{v},l})$ with marginal distributions $F_{U_{\boldsymbol{v},l}}$ and $F_{V_{\boldsymbol{v},l}}$, we have that

$$\begin{aligned}
&P(Z(\boldsymbol{v}) \leq F_{U_{\boldsymbol{v},l}}^{-1}(q) \mid Z(\boldsymbol{v}_{(l)}) \leq F_{V_{\boldsymbol{v},l}}^{-1}(q)) \\
&= \int_{F_{V_{\boldsymbol{v},l}}^{-1}(0)}^{F_{V_{\boldsymbol{v},l}}^{-1}(q)} P(Z(\boldsymbol{v}) \leq F_{U_{\boldsymbol{v},l}}^{-1}(q) \mid Z(\boldsymbol{v}_{(l)}) = z_l) dF_{V_{\boldsymbol{v},l}}(z_l) \Big/ \int_{F_{V_{\boldsymbol{v},l}}^{-1}(0)}^{F_{V_{\boldsymbol{v},l}}^{-1}(q)} dF_{V_{\boldsymbol{v},l}} \\
&\leq \int_{F_{V_{\boldsymbol{v},l}}^{-1}(0)}^{F_{V_{\boldsymbol{v},l}}^{-1}(q)} P(Z(\boldsymbol{v}) \leq F_{U_{\boldsymbol{v},l}}^{-1}(q) \mid Z(\boldsymbol{v}_{(l)}) = F_{V_{\boldsymbol{v},l}}^{-1}(0)) dF_{V_{\boldsymbol{v},l}}(z_l) \Big/ \int_{F_{V_{\boldsymbol{v},l}}^{-1}(0)}^{F_{V_{\boldsymbol{v},l}}^{-1}(q)} dF_{V_{\boldsymbol{v},l}} \\
&= P(Z(\boldsymbol{v}) \leq F_{U_{\boldsymbol{v},l}}^{-1}(q) \mid Z(\boldsymbol{v}_{(l)}) = F_{V_{\boldsymbol{v},l}}^{-1}(0)).
\end{aligned}$$

It follows that the boundary cdf of the NNMP model

$$\begin{aligned}
&F_{1|2}(F_{Z(\boldsymbol{v})}^{-1}(q) \mid F_{\boldsymbol{Z}_{\mathrm{Ne}(\boldsymbol{v})}}^{-1}(0)) \\
&= P(Z(\boldsymbol{v}) \leq F_{Z(\boldsymbol{v})}^{-1}(q) \mid Z(\boldsymbol{v}_{(1)}) = F_{Z(\boldsymbol{v}_{(1)})}^{-1}(0), \ldots, Z(\boldsymbol{v}_{(L)}) = F_{Z(\boldsymbol{v}_{(L)})}^{-1}(0)) \\
&= \sum_{l=1}^{L} w_l(\boldsymbol{v}) P(Z(\boldsymbol{v}) \leq F_{U_{\boldsymbol{v},l}}^{-1}(q) \mid Z(\boldsymbol{v}_{(l)}) = F_{V_{\boldsymbol{v},l}}^{-1}(0)) \\
&\geq \sum_{l=1}^{L} w_l(\boldsymbol{v}) P(Z(\boldsymbol{v}) \leq F_{U_{\boldsymbol{v},l}}^{-1}(q) \mid Z(\boldsymbol{v}_{(l)}) \leq F_{V_{\boldsymbol{v},l}}^{-1}(q)),
\end{aligned} \tag{A.4}$$

Taking $q \to 0^+$ on both sides of (A.4), we obtain $F_{1|2}(F_{Z(\boldsymbol{v})}^{-1}(0) \mid F_{\boldsymbol{Z}_{\mathrm{Ne}(\boldsymbol{v})}}^{-1}(0)) \geq \sum_{l=1}^{L} w_l(\boldsymbol{v}) \lambda_{\mathcal{L},l}(\boldsymbol{v})$. Hence, if there exists some l such that $\lambda_{\mathcal{L},l}(\boldsymbol{v}) > 0$, the conditional cdf $F_{1|2}(F^{-1}_{Z(\boldsymbol{v})}(q) \,|\, F^{-1}_{\boldsymbol{Z}_{\mathrm{Ne}(\boldsymbol{v})}}(0))$ has strictly positive mass at $q = 0$. We can prove the result for $F_{1|2}(F^{-1}_{Z(\boldsymbol{v})}(q) \mid F^{-1}_{\boldsymbol{Z}_{\mathrm{Ne}(\boldsymbol{v})}}(1))$ in a similar way. $\square$

Proof of Proposition 5.2. Consider a discrete copula NNMP characterized by $p(y(\boldsymbol{v}) \,|\, \boldsymbol{y}_{\mathrm{Ne}(\boldsymbol{v})}) = \sum_{l=1}^{L} w_l(\boldsymbol{v})\, c_{\boldsymbol{v},l}(y(\boldsymbol{v}), y(\boldsymbol{v}_{(l)}))g_{\boldsymbol{v}}(y(\boldsymbol{v}))$, where $g_{\boldsymbol{v}}$ is the marginal pmf of $Y(\boldsymbol{v})$.

Let $\boldsymbol{y}_{\mathcal{V}} = (y(\boldsymbol{s}_1), \ldots, y(\boldsymbol{s}_n), y(\boldsymbol{u}_1), \ldots, y(\boldsymbol{u}_m))^\top$ for $n \geq 2$ and $m \geq 1$, where $\mathcal{V} = \mathcal{S} \cup \mathcal{U}$, $\mathcal{S} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$, $\mathcal{U} = \{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m\}$, and $\mathcal{S} \cap \mathcal{U} = \emptyset$. The joint pmf of $\boldsymbol{y}_{\mathcal{V}}$ can be written as $p(\boldsymbol{y}_{\mathcal{V}}) = p(\boldsymbol{y}_{\mathcal{U}} \,|\, \boldsymbol{y}_{\mathcal{S}})p(\boldsymbol{y}_{\mathcal{S}})$. We will first derive the joint pmf $p(\boldsymbol{y}_{\mathcal{S}}) = p(y(\boldsymbol{s}_1), \ldots, y(\boldsymbol{s}_n))$ and then $p(\boldsymbol{y}_{\mathcal{U}} \,|\, \boldsymbol{y}_{\mathcal{S}})$, where $\boldsymbol{y}_{\mathcal{U}} = (y(\boldsymbol{u}_1), \ldots, y(\boldsymbol{u}_m))^\top$.

Let $c_{\boldsymbol{s}_i,l_i} \equiv c_{\boldsymbol{s}_i,l_i}(y(\boldsymbol{s}_i), y(\boldsymbol{s}_{(i,l_i)}))$ and $w_{\boldsymbol{s}_i,l_i} \equiv w_{l_i}(\boldsymbol{s}_i)$ with $l_i = 1, \ldots, i_L$ and $i_L = (i-1) \wedge L$, for all i. Then $p(y(\boldsymbol{s}_1), y(\boldsymbol{s}_2)) = p(y(\boldsymbol{s}_2) \,|\, y(\boldsymbol{s}_1))g_{\boldsymbol{s}_1}(y(\boldsymbol{s}_1)) = c_{\boldsymbol{s}_2,1}g_{\boldsymbol{s}_2}(y(\boldsymbol{s}_2))g_{\boldsymbol{s}_1}(y(\boldsymbol{s}_1))$. By definition of the discrete NNMP, $w_{\boldsymbol{s}_2,1} = 1$. Then

$$
\begin{aligned}
&p(y(\boldsymbol{s}_1), y(\boldsymbol{s}_2), y(\boldsymbol{s}_3)) = p(y(\boldsymbol{s}_3) \,|\, y(\boldsymbol{s}_1), y(\boldsymbol{s}_2))p(y(\boldsymbol{s}_1), y(\boldsymbol{s}_2)) \\
&\quad = \left(\sum_{l_3=1}^{2} w_{\boldsymbol{s}_3,l_3} c_{\boldsymbol{s}_3,l_3} g_{\boldsymbol{s}_3}(y(\boldsymbol{s}_3))\right) c_{\boldsymbol{s}_2,1} g_{\boldsymbol{s}_2}(y(\boldsymbol{s}_2)) g_{\boldsymbol{s}_1}(y(\boldsymbol{s}_1)) \\
&\quad = \prod_{i=1}^{3} g_{\boldsymbol{s}_i}(y(\boldsymbol{s}_i)) \sum_{l_3=1}^{2} w_{\boldsymbol{s}_3,l_3} c_{\boldsymbol{s}_3,l_3} c_{\boldsymbol{s}_2,1} = \prod_{i=1}^{3} g_{\boldsymbol{s}_i}(y(\boldsymbol{s}_i)) \sum_{l_3=1}^{2} \sum_{l_2=1}^{1} w_{\boldsymbol{s}_3,l_3} w_{\boldsymbol{s}_2,l_2} c_{\boldsymbol{s}_3,l_3} c_{\boldsymbol{s}_2,l_2}.
\end{aligned}
$$

Similarly, for $4 \leq n \leq L$, the joint pmf is

$$
\begin{aligned}
&p(y(\boldsymbol{s}_1), \ldots, y(\boldsymbol{s}_n)) = p(y(\boldsymbol{s}_n) \,|\, \boldsymbol{y}_{\mathrm{Ne}(\boldsymbol{s}_n)})p(y(\boldsymbol{s}_1), \ldots, y(\boldsymbol{s}_{n-1})) \\
&= \left(\sum_{l_n=1}^{n-1} w_{\boldsymbol{s}_n,l_n} c_{\boldsymbol{s}_n,l_n} g_{\boldsymbol{s}_n}(y(\boldsymbol{s}_n))\right) \\
&\quad \left(\prod_{i=1}^{n-1} g_{\boldsymbol{s}_i}(y(\boldsymbol{s}_i)) \sum_{l_{n-1}=1}^{n-2} \cdots \sum_{l_2=1}^{1} w_{\boldsymbol{s}_{n-1},l_{n-1}} \ldots w_{\boldsymbol{s}_2,l_2} c_{\boldsymbol{s}_{n-1},l_{n-1}} \ldots c_{\boldsymbol{s}_2,l_2}\right) \\
&= \prod_{i=1}^{n} g_{\boldsymbol{s}_i}(y(\boldsymbol{s}_i)) \sum_{l_n=1}^{n-1} \cdots \sum_{l_2=1}^{1} w_{\boldsymbol{s}_n,l_n} \ldots w_{\boldsymbol{s}_2,l_2} c_{\boldsymbol{s}_n,l_n} \ldots c_{\boldsymbol{s}_2,l_2}.
\end{aligned}
$$

Finally, for $n > L$, it is easy to show that the joint pmf is

$$
\begin{aligned}
&p(y(\boldsymbol{s}_1), \ldots, y(\boldsymbol{s}_n)) = p(y(\boldsymbol{s}_n) \,|\, \boldsymbol{y}_{\mathrm{Ne}(\boldsymbol{s}_n)}) p(y(\boldsymbol{s}_1), \ldots, y(\boldsymbol{s}_{n-1})) \\
&= \left(\sum_{l_n=1}^{L} w_{\boldsymbol{s}_n, l_n} c_{\boldsymbol{s}_n, l_n} g_{\boldsymbol{s}_n}(y(\boldsymbol{s}_n)) \right) \\
&\quad \prod_{i=1}^{n-1} g_{\boldsymbol{s}_i}(y(\boldsymbol{s}_i)) \sum_{l_{n-1}=1}^{L} \cdots \sum_{l_{L+1}=1}^{L} \sum_{l_L=1}^{L-1} \cdots \sum_{l_2=1}^{1} w_{\boldsymbol{s}_{n-1}, l_{n-1}} \ldots w_{\boldsymbol{s}_2, l_2} c_{\boldsymbol{s}_{n-1}, l_{n-1}} \ldots c_{\boldsymbol{s}_2, l_2} \\
&= \prod_{i=1}^{n} g_{\boldsymbol{s}_i}(y(\boldsymbol{s}_i)) \sum_{l_n=1}^{L} \cdots \sum_{l_{L+1}=1}^{L} \sum_{l_L=1}^{L-1} \cdots \sum_{l_2=1}^{1} w_{\boldsymbol{s}_n, l_n} \ldots w_{\boldsymbol{s}_2, l_2} c_{\boldsymbol{s}_n, l_n} \ldots c_{\boldsymbol{s}_2, l_2}.
\end{aligned}
$$

Therefore, we have that, for $n \geq 2$, the joint pmf

$$
p(\boldsymbol{y}_{\mathcal{S}}) = p(y(\boldsymbol{s}_1), \ldots, y(\boldsymbol{s}_n)) = \prod_{i=1}^{n} g_{\boldsymbol{s}_i}(y(\boldsymbol{s}_i)) \sum_{l_n=1}^{n_L} \cdots \sum_{l_2=1}^{2_L} w_{\boldsymbol{s}_n, l_n} \ldots w_{\boldsymbol{s}_2, l_2} c_{\boldsymbol{s}_n, l_n} \ldots c_{\boldsymbol{s}_2, l_2}.
$$

Turning to the non-reference set $\mathcal{U}$. Let $c_{\boldsymbol{u}_i, l_i} \equiv c_{\boldsymbol{u}_i, l_i}(y(\boldsymbol{u}_i), y(\boldsymbol{u}_{(i,l_i)}))$ and $w_{\boldsymbol{u}_i, l_i} \equiv w_{l_i}(\boldsymbol{u}_i)$ with $l_i = 1, \ldots, L$. When $m = 1$, $p(\boldsymbol{y}_{\mathcal{U}} \,|\, \boldsymbol{y}_{\mathcal{S}}) = p(y(\boldsymbol{u}_1) \,|\, \boldsymbol{y}_{\mathrm{Ne}(\boldsymbol{u}_1)})$. When $m \geq 2$, without loss of generality, we consider the case of $m = 2$, i.e., we take $\mathcal{U} = \{\boldsymbol{u}_1, \boldsymbol{u}_2\}$. Then we have that

$$
\begin{aligned}
p(\boldsymbol{y}_{\mathcal{U}} \,|\, \boldsymbol{y}_{\mathcal{S}}) &= p(y(\boldsymbol{u}_1) \,|\, \boldsymbol{y}_{\mathrm{Ne}(\boldsymbol{u}_1)}) p(y(\boldsymbol{u}_2) \,|\, \boldsymbol{y}_{\mathrm{Ne}(\boldsymbol{u}_2)}) \\
&= \left(\sum_{l_1=1}^{L} w_{\boldsymbol{u}_1, l_1} c_{\boldsymbol{u}_1, l_1} g_{\boldsymbol{u}_1}(y(\boldsymbol{u}_1)) \right) \left(\sum_{l_2=1}^{L} w_{\boldsymbol{u}_2, l_2} c_{\boldsymbol{u}_2, l_2} g_{\boldsymbol{u}_2}(y(\boldsymbol{u}_2)) \right) \\
&= \prod_{i=1}^{2} g_{\boldsymbol{u}_i}(y(\boldsymbol{u}_i)) \sum_{l_2=1}^{L} \sum_{l_1=1}^{L} w_{\boldsymbol{u}_2, l_2} w_{\boldsymbol{u}_1, l_1} c_{\boldsymbol{u}_2, l_2} c_{\boldsymbol{u}_1, l_1}.
\end{aligned}
$$

Obviously, the result is easily generalized for $\mathcal{U} = \{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m\}$ for any $m > 2$. $\square$

Appendix B

Implementation Details

B.1 Copulas

We introduce Gaussian, Gumbel and Clayton copulas with properties relevant to model estimation and prediction. For more details we refer to Joe (2014). Consider a bivariate vector (X_1, X_2) with marginal cumulative distribution functions (cdfs) such that $F_1(x_1) = t_1$ and $F_2(x_2) = t_2$.

Gaussian copula A Gaussian copula with correlation $\rho \in (0, 1)$ for (X_1, X_2) is $C(F_1(x_1), F_2(x_2) \mid \rho) = C(t_1, t_2 \mid \rho) = \Phi_2(\Phi^{-1}(t_1),\, \Phi^{-1}(t_2) \mid \rho)$. If both X_1 and X_2 are continuous random variables, the copula has density

$$\frac{1}{\sqrt{1-\rho^2}} \exp\left(\frac{2\rho\Phi^{-1}(t_1)\Phi^{-1}(t_2) - \rho^2\{(\Phi^{-1}(t_1))^2 + (\Phi^{-1}(t_2))^2\}}{2(1-\rho^2)} \right). \tag{B.1}$$

The conditional cdf of T_1 given $T_2 = t_2$, denoted as $C_{1|2}(t_1 \mid t_2)$, is given by $C_{1|2}(t_1 \mid t_2) = \frac{\partial C(t_1,t_2)}{\partial t_2} = \Phi\left(\frac{\Phi^{-1}(t_1) - \rho\Phi^{-1}(t_2)}{\sqrt{1-\rho^2}}\right)$. To simulate X_1 given $X_2 = x_2$, we first compute $t_2 = F_2(x_2)$. We then generate a random number z from a uniform distribution on $[0, 1]$, and compute $t_1 = C_{1|2}^{-1}(z \mid t_2)$ where $C_{1|2}^{-1}(z \mid t_2) =$

$\Phi\left(\sqrt{(1-\rho^2)}\Phi^{-1}(z)+\rho\Phi^{-1}(t_2)\right)$ is the inverse of $C_{1|2}(t_1 \mid t_2)$. Finally, we obtain x_1 from the inverse cdf $F_1^{-1}(t_1)$.

Gumbel copula A Gumbel copula with parameter $\eta \in [1,\infty)$ for (X_1, X_2) is $C(F_1(x_1), F_2(x_2) \mid \eta) = C(t_1, t_2 \mid \eta) = \exp(-((-\log(t_1))^\eta + (-\log(t_2))^\eta)^{1/\eta})$. Let $u_1 = -\log(t_1)$ and $u_2 = -\log(t_2)$. If both X_1 and X_2 are continuous random variables, the copula has density

$$\exp(-(u_1^\eta + u_2^\eta)^{1/\eta})((u_1^\eta + u_2^\eta)^{1/\eta} + \eta - 1)(u_1^\eta + u_2^\eta)^{1/\eta - 2}(u_1 u_2)^{\eta-1}(t_1 t_2)^{-1}. \quad \text{(B.2)}$$

The conditional cdf of T_1 given $T_2 = t_2$ is $C_{1|2}(t_1 \mid t_2) = \overline{C}_{1|2}(u_1 \mid u_2) = t_2^{-1}\exp(-(u_1^\eta + u_2^\eta)^{1/\eta})(1 + (u_1/u_2)^\eta)^{1/\eta - 1}$, where the conditional cdf $\overline{C}_{1|2}(u_1 \mid u_2)$ corresponds to the copula $\overline{C}(u_1, u_2 \mid \eta) = \exp(-(u_1^\eta + u_2^\eta)^{1/\eta})$ which is a bivariate exponential survival function, with marginals corresponding to a unit rate exponential distribution. The inverse conditional cdf $C_{1|2}^{-1}(\cdot \mid t_2)$ does not have a closed form. To generate X_1 given $X_2 = x_2$, following Joe (2014), we first define $y = (u_1^\eta + u_2^\eta)^{1/\eta}$. Then we have a realization of X_1, say $x_1 = (y_0^\eta - u_2^\eta)^{1/\eta}$, where y_0 is the root of $h(y) = y + (\eta - 1)\log(y) - (u_2 + (\eta - 1)\log(u_2) - \log z) = 0$, where $y \geq u_2$, and z is a random number generated from a uniform distribution on $[0, 1]$.

Clayton copula A Clayton copula with parameter $\delta \in [0,\infty)$ for (X_1, X_2) is $C(F_1(x_1), F_2(x_2) \mid \delta) = C(t_1, t_2 \mid \delta) = (t_1^{-\delta} + t_2^{-\delta} - 1)^{-1/\delta}$. If both X_1 and X_2 are continuous random variables, the copula has density $(1+\delta)(t_1 t_2)^{-\delta-1}(t_1^{-\delta} + t_2^{-\delta} - 1)^{-2-1/\delta}$. The conditional cdf of T_1 given $T_2 = t_2$ is $C_{1|2}(t_1 \mid t_2) = (1 + t_2^\delta(t_1^{-\delta} - 1))^{-1-1/\delta}$. To simulate X_1 given X_2, we first compute $t_2 = F_2(x_2)$, and generate a uniform random number z on $[0, 1]$. Then we compute $t_1 = C_{1|2}^{-1}(z \mid t_2)$ where $C_{1|2}^{-1}(z \mid t_2) = ((z^{-\delta/(1+\delta)} - 1)t_2^{-\delta} + 1)^{-1/\delta}$. Finally, we obtain x_1 from the inverse cdf $F_1^{-1}(t_1)$.

B.2 MTD Models

We provide necessary details of the posterior simulation for the Gaussian, Poisson, negative binomial and Lomax MTD models.

B.2.1 GMTD Models

For the Gaussian MTD model, we consider the following prior $p(\{\rho_l\}_{l=1}^L, \mu, \sigma^2) = \prod_{l=1}^L \text{Unif}(\rho_l \mid -1, 1) N(\mu \mid \mu_0, \sigma_0^2) \text{IG}(\sigma^2 \mid u_0, v_0)$. The full conditional distribution of μ is $N(\mu \mid \mu_1, \sigma_1^2)$, where $\mu_1 = \sigma_1^2 \left(\mu_0/\sigma_0^2 + c/\sigma^2\right)$, and $\sigma_1^2 = \left(1/\sigma_0^2 + b/\sigma^2\right)^{-1}$ with $b = \sum_{t=L+1}^n (1-\rho_{z_t})^2/(1-\rho_{z_t}^2)$ and $c = \sum_{t=L+1}^n (1-\rho_{z_t})(x_t - \rho_{z_t} x_{t-z_t})/(1-\rho_{z_t}^2)$. The inverse gamma prior for σ^2 yields a conjugate full conditional distribution $\text{IG}(\sigma^2 \mid u_1, v_1)$ where $u_1 = u_0 + (n-L)/2$ and $v_1 = v_0 + \sum_{t=L+1}^L (x_t - \rho_{z_t} x_{t-z_t} - (1-\rho_{z_t})\mu)^2/(2(1-\rho_{z_t}^2))$. Finally, we update each ρ_l using a slice sampler with target density $\text{Unif}(\rho_l \mid -1, 1) \prod_{t: z_t = l} N(x_t \mid (1-\rho_l)\mu + \rho_l x_{t-l}, \sigma^2(1-\rho_l^2))$, for $l = 1, \ldots, L$. For each time t, $t = L+1, \ldots, n$, the posterior probability of $z_t = l$ is proportional to $w_l N(x_t \mid (1-\rho_l)\mu + \rho_l x_{t-l}, (1-\rho_l^2)\sigma^2)$.

B.2.2 Poisson and Negative Binomial MTD Models

For the Poisson MTD, we reparameterize the model such that the lth component transition density of the model is sampled through $x_t \mid q_t, x_{t-l}, \theta \sim \text{Bin}(x_t - q_t \mid x_{t-l}, \theta)$ and $q_t \mid \lambda \sim \text{Pois}(q_t \mid \lambda)$. We consider conjugate prior $p(\lambda, \theta) = \text{Ga}(\lambda \mid u_\lambda, v_\lambda)\text{Beta}(\theta \mid u_\theta, v_\theta)$. The posterior full conditional distribution of λ is gamma distribution with shape parameter $u_\lambda + \sum_{t=L+1}^n q_t$ and rate parameter $v_\lambda + n - L$. The posterior full conditional distribution of θ is a beta distribution $\text{Beta}(\theta \mid u_\theta +$

$\sum_{t=L+1}^{n}(x_t - q_t), v_\theta + \sum_{t=L+1}^{n}(x_{t-z_t} - x_t + q_t))$. We update q_t with an independent Metropolis step with target density $\mathrm{Bin}(x_t - q_t \,|\, x_{t-z_t}, \theta)\mathrm{Pois}(q_t \,|\, \lambda)$ and proposal distribution being a discrete uniform distribution over the interval $[0 \vee (x_t - x_{t-z_t}), x_t]$, for $t = L+1, \ldots, n$. For each time t, $t = L+1, \ldots, n$, the posterior probability of $z_t = l$ is proportional to $w_l \mathrm{Bin}(x_t - q_t \,|\, x_{t-l}, \theta)$.

Similar to the Poisson model, we reparameterize the negative binomial MTD to facilitate posterior simulation. In particular, the lth component transition density of the model is sampled through $x_t \,|\, q_t, x_{t-l}, \theta \sim \mathrm{Bin}(x_t - q_t \,|\, x_{t-l}, \theta)$ and $q_t \,|\, x_{t-l}, \kappa, \psi \sim \mathrm{NB}(q_t \,|\, \kappa + x_{t-l}, \psi)$, with $p(\theta, \psi, \kappa) = \mathrm{Beta}(\theta \,|\, u_\theta, v_\theta)\mathrm{Beta}(\psi \,|\, u_\psi, v_\psi)$ $\mathrm{Ga}(\kappa \,|\, u_\kappa, v_\kappa)$. The beta priors for θ and ψ yield conjugate posterior full conditional distributions. They are $\mathrm{Beta}(\theta \,|\, u_\theta + \sum_{t=L+1}^{n}(x_t - q_t), v_\theta + \sum_{t=L+1}^{n}(x_{t-z_t} - x_t + q_t))$ and $\mathrm{Beta}(\psi \,|\, u_\psi + (n-L)\kappa + \sum_{t=L+1}^{n} x_{t-z_t}, v_\psi + \sum_{t=L+1}^{n} q_t)$. The posterior full conditional of κ is proportional to $\mathrm{Ga}(\kappa \,|\, u_\kappa, v_\kappa) \prod_{t=L+1}^{n} \mathrm{NB}(q_t \,|\, \kappa + x_{t-z_t}, \psi)$. We update κ on its log scale using a random-walk Metropolis step. We update q_t with an independent Metropolis step with target density $\mathrm{Bin}(x_t - q_t \,|\, x_{t-z_t}, \theta)\mathrm{NB}(q_t \,|\, \kappa + x_{t-z_t}, \psi)$ and proposal distribution being a discrete uniform distribution over the interval $[0 \vee (x_t - x_{t-z_t}), x_t]$, for $t = L+1, \ldots, n$. For each time t, $t = L+1, \ldots, n$, the posterior probability of $z_t = l$ is proportional to $w_l \mathrm{Bin}(x_t - q_t \,|\, x_{t-l}, \theta)\mathrm{NB}(q_t \,|\, \kappa + x_{t-l}, \psi)$.

B.2.3 Lomax MTD Models

For the Lomax MTD, we consider prior $p(\alpha, \phi, \boldsymbol{\beta}) \propto \mathrm{Ga}(\alpha \,|\, u_\alpha, v_\alpha)\mathrm{IG}(\phi \,|\, u_\phi, v_\phi)$. The posterior full conditional distribution of α is $\mathrm{Ga}\,(\alpha \,|\, u_\alpha + n - L, v'_\alpha)$, where $v'_\alpha = v_\alpha + \sum_{t=L+1}^{n} \log\Big(1 + y_t \exp(-\boldsymbol{x}_t^\top \boldsymbol{\beta})/(\phi + y_{t-z_t} \exp(-\boldsymbol{x}_{t-z_t}^\top \boldsymbol{\beta}))\Big)$. To improve mixing, we integrated out α from the posterior full conditionals of ϕ and that of $\boldsymbol{\beta}$. The resulting posterior full conditional distributions of ϕ and $\boldsymbol{\beta}$ are proportional to $\mathrm{IG}(\phi \,|\, u_\phi, v_\phi) g(\{y_t\}_{t=L+1}^n, \phi, \boldsymbol{\beta})$ and $\prod_{t=L+1}^n \exp(-\boldsymbol{x}_t^\top \boldsymbol{\beta}) g(\{y_t\}_{t=L+1}^n, \phi, \boldsymbol{\beta})$, respectively, where

$$g(\{y_t\}_{t=L+1}^n, \phi, \boldsymbol{\beta}) = \left\{ \prod_{t=L+1}^n \left(\phi + y_{t-z_t} \exp(-\boldsymbol{x}_{t-z_t}^\top \boldsymbol{\beta}) + y_t \exp(-\boldsymbol{x}_t^\top \boldsymbol{\beta}) \right)^{-1} \right\} \left(v_\alpha + \sum_{t=L+1}^n \log(1 + y_t \exp(-\boldsymbol{x}_t^\top \boldsymbol{\beta}) / (\phi + y_{t-z_t} \exp(-\boldsymbol{x}_{t-z_t}^\top \boldsymbol{\beta}))) \right)^{-(u_\alpha + n - L)}.$$

We use random walk Metropolis steps to update ϕ and $\boldsymbol{\beta}$ on their log scales, respectively. Take $\epsilon_t = y_t \exp(-\boldsymbol{x}_t^\top \boldsymbol{\beta})$. For each time t, $t = L+1, \dots, n$, the posterior probability of $z_t = l$ is proportional to $w_l P(\epsilon_t \,|\, \phi + \epsilon_{t-l}, \alpha)$.

B.3 MTDPP Models

We introduce necessary posterior simulation steps for the Burr MTDPP and the Lomax MTDCPP models illustrated in the data examples. Given an observed point pattern $\{t_i\}_{i=1}^n$ over the time window $(0, T)$, we let $x_1 = t_1$ and $x_i = t_i - t_{i-1}$ for $i = 2, \dots, n$. For convenience of notation, we take $x_{n+1} = T - t_n$. The posterior inference is based on a likelihood, conditional on $(x_1, \dots, x_L)$.

B.3.1 Burr MTDPP

We associate each x_i with a latent variable ℓ_i such that $P(\ell_i = l) = \sum_{l=1}^L w_l \delta_l(\ell_i)$, for $i = L+1, \dots, n+1$. With customary priors for parameters $(\gamma, \lambda, \kappa)$, we obtain the joint distribution

$$\begin{aligned}
&\mathrm{Ga}(\lambda \,|\, u_\lambda, v_\lambda) \times \mathrm{Ga}(\gamma \,|\, u_\gamma, v_\gamma) \times \mathrm{Ga}(\kappa \,|\, u_\kappa, v_\kappa) \times \mathrm{Dir}(\boldsymbol{w} \,|\, \alpha_0 a_1, \ldots, \alpha_0 a_L) \\
&\times \left\{ \prod_{i=L+1}^{n} \mathrm{Burr}\Big(x_i \,|\, \gamma, \lambda + x_{i-\ell_i}, \kappa\Big) \sum_{l=1}^{L} w_l \delta_l(\ell_i) \right\} \\
&\times \left\{ S\Big(x_{n+1} \,|\, \gamma, \lambda + x_{n+1-\ell_{n+1}}, \kappa\Big) \sum_{l=1}^{L} w_l \delta_l(\ell_{n+1}) \right\},
\end{aligned}$$

where $S\Big(x_{n+1} \,|\, \gamma, \lambda + x_{n+1-\ell_i}, \kappa\Big)$ is the survival function associated the Burr distribution. In particular, $S\Big(x_{n+1} \,|\, \gamma, \lambda + x_{n+1-\ell_i}, \kappa\Big) = (1 + \{x_{n+1}/(\lambda + x_{n+1-\ell_i})\}^\gamma)^{-\kappa}$.

We focus on the posterior updates for parameters $(\lambda, \gamma, \kappa)$. Let $p(\{x_i\}_{i=L+1}^{n+1}, \gamma, \lambda, \kappa, \{\ell_i\}_{i=L+1}^{n+1}) = \Big\{\prod_{i=L+1}^{n} \mathrm{Burr}(x_i \,|\, \gamma, \lambda + x_{i-l}, \kappa)\Big\} S(x_{n+1} \,|\, \gamma, \lambda + x_{n+1-\ell_{n+1}}, \kappa)$. We use random walk Metropolis steps to update parameters λ and γ with target densities $\mathrm{Ga}(\lambda \,|\, u_\lambda, v_\lambda) p(\{x_i\}_{i=L+1}^{n+1}, \gamma, \lambda, \kappa, \{\ell_i\}_{i=L+1}^{n+1})$ and $\mathrm{Ga}(\gamma \,|\, u_\gamma, v_\gamma) p(\{x_i\}_{i=L+1}^{n+1}, \gamma, \lambda, \kappa, \{\ell_i\}_{i=L+1}^{n+1})$, respectively. The posterior full conditional distribution of κ is a gamma distribution with shape parameter $u_\kappa + n - L$ and rate parameter $v_\kappa + \sum_{i=L+1}^{n+1} \log(1 + \{x_i/(\lambda + x_{i-\ell_i})\}^\gamma)$.

B.3.2 Lomax MTDCPP

The conditional duration density of the Lomax MTDCPP, for $i > L$, can be written as $f_C^*(x_i) = \sum_{l=0}^{L} \pi_l f_l^c(x_i \,|\, \mu, \phi, \alpha)$, where $f_0^c \equiv \mu \exp(-\mu x_i)$, $f_l^c \equiv P(x_i \,|\, \phi + x_{i-l}, \alpha)$, and $\pi_l = (1 - \pi_0) w_l$, $l = 1, \ldots, L$. Let S_0^c and S_l^c be the survival functions associated with f_0^c and f_l^c, respectively, for $l = 1, \ldots, L$. With customary priors for parameters (μ, ϕ, α), the joint distribution is

$$\begin{aligned}&\mathrm{Ga}(\mu \,|\, u_\mu, v_\mu) \times \mathrm{Ga}(\phi \,|\, u_\phi, v_\phi) \times \mathrm{Ga}(\alpha \,|\, u_\alpha, v_\alpha) \times \mathrm{Dir}(\boldsymbol{w} \,|\, \alpha_0 a_1, \ldots, \alpha_0 a_L)\\ &\times \mathrm{Beta}(\pi_0 \,|\, u_0, v_0) \times \left\{ \prod_{i=L+1}^{n} f^c_{\ell_i}(x_i \,|\, \mu, \phi, \alpha) \sum_{l=0}^{L} \pi_l \delta_l(\ell_i) \right\}\\ &\times \left\{ S^c_{\ell_{n+1}}(x_{n+1} \,|\, \mu, \phi, \alpha) \sum_{l=0}^{L} \pi_l \delta_l(\ell_{n+1}) \right\}.\end{aligned}$$

We focus one the posterior updates for parameters (μ, ϕ, α). Let $M_l = |\{i : \ell_i = l\}|$, for $l = 0, \ldots, L$. A gamma prior for μ yields conjugate full conditional distribution $\mathrm{Ga}(\mu \,|\, u_\mu + M_0 - \delta_0(\ell_{n+1}), v_\mu + \sum_{i:\ell_i=0} x_i)$. The posterior full conditional distribution of α is gamma distribution with shape parameter $u_\alpha + \sum_{l=1}^{L} M_l - 1 + \delta_0(\ell_{n+1})$ and rate parameter $v_\alpha + \sum_{i:\ell_i \neq 0} \log(1 + x_i/(\phi + x_{i-\ell_i}))$. Let $B_0^n = \{i : 1 \leq i \leq n \wedge \ell_i \neq 0\}$, and $p(\{x_i\}_{i=L+1}^{n+1}, \mu, \phi, \alpha, \{\ell_i\}_{i=L+1}^{n+1}) = \prod_{i \in B_0^n} f^c_{\ell_i}(x_i \,|\, \mu, \phi, \alpha) \left\{ S^c_{\ell_{n+1}}(x_{n+1} \,|\, \mu, \phi, \alpha) \right\}^{1-\delta_0(\ell_{n+1})}$. We update ϕ using a random walk Metropolis step with target density $\mathrm{Ga}(\phi \,|\, u_\phi, v_\phi)\, p(\{x_i\}_{i=L+1}^{n+1}, \mu, \phi, \alpha, \{\ell_i\}_{i=L+1}^{n+1})$.

B.4 NNMP Models

B.4.1 GNNMP Models

Consider a univariate response $y(\boldsymbol{v})$, at location $\boldsymbol{v} \in \mathcal{D}$, in a spatially varying regression model, $y(\boldsymbol{v}) = \boldsymbol{x}(\boldsymbol{v})^\top \boldsymbol{\beta} + z(\boldsymbol{v}) + \epsilon(\boldsymbol{v}), \;\; \boldsymbol{v} \in \mathcal{D}$ where $\epsilon(\boldsymbol{v}) \overset{i.i.d.}{\sim} N(0, \tau^2)$, and the spatial random effect $z(\boldsymbol{v})$ follows a stationary GNNMP model. The associated conditional density of the GNNMP is

$$p(z(\boldsymbol{v}) \,|\, \boldsymbol{z}_{\mathrm{Ne}(\boldsymbol{v})}) = \sum_{l=1}^{L} w_l(\boldsymbol{v}) N(z(\boldsymbol{v}) \,|\, \rho_l(\boldsymbol{v}) z(\boldsymbol{v}_{(l)}), \sigma^2(1 - (\rho_l(\boldsymbol{v}))^2)), \tag{B.3}$$

where $\rho_l(\boldsymbol{v}) \equiv \rho_l(||\boldsymbol{v}-\boldsymbol{v}_{(l)}||) = \exp(-||\boldsymbol{v}-\boldsymbol{v}_{(l)}||/\phi)$. For the weights, we consider an exponential correlation function with range parameter ζ for the kernel function that defines the random cutoff points. The Bayesian model is completed with priors $N(\boldsymbol{\beta} \mid \boldsymbol{\mu}_\beta, \boldsymbol{V}_\beta)$, $\mathrm{IG}(\sigma^2 \mid u_{\sigma^2}, v_{\sigma^2})$, $\mathrm{IG}(\tau^2 \mid u_{\tau^2}, v_{\tau^2})$, $\mathrm{IG}(\phi \mid u_\phi, v_\phi)$, $\mathrm{IG}(\zeta \mid u_\zeta, v_\zeta)$, $N(\boldsymbol{\gamma} \mid \boldsymbol{\mu}_\gamma, \boldsymbol{V}_\gamma)$, $\mathrm{IG}(\kappa^2 \mid u_{\kappa^2}, v_{\kappa^2})$.

Let $y(\boldsymbol{s}_i)$, $i = 1, \ldots, n$, be the observations over reference set $\mathcal{S} = (\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n)$. We introduce the MCMC sampler for the spatially varying regression model in which the GNNMP is used as a prior for the latent spatial random effect. The MCMC sampler involves sampling the latent variables $z(\boldsymbol{s}_i)$, but it is easily developed based on the sampler described in the Chapter 4.

For each $z(\boldsymbol{s}_i)$, $i = 3, \ldots, n$, we introduce a configuration variable ℓ_i, taking values in $\{1, \ldots, i_L\}$ where $i_L = (i-1) \wedge L$, such that $\Pr(\ell_i \mid \boldsymbol{w}(\boldsymbol{s}_i)) = \sum_{l=1}^{i_L} w_l(\boldsymbol{s}_i)\delta_l(\ell_i)$, where $\boldsymbol{w}(\boldsymbol{s}_i) = (w_1(\boldsymbol{s}_i), \ldots, w_{i_L}(\boldsymbol{s}_i))^\top$ and $\delta_l(\ell_i) = 1$ if $\ell_i = l$ and 0 otherwise. To allow for efficient simulation of parameters $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2)^\top$ and κ^2 for the weights, we associate each $z(\boldsymbol{s}_i)$ with a latent Gaussian variable with mean $\mu(\boldsymbol{s}_i) = \gamma_0 + s_{i1}\gamma_1 + s_{i2}\gamma_2$ and variance κ^2, where $\boldsymbol{s}_i = (s_{i1}, s_{i2})$, for $i = 3, \ldots, n$. There is a one-to-one correspondence between the configuration variables ℓ_i and latent variables t_i: $\ell_i = l$ if and only if $t_i \in (r^*_{\boldsymbol{s}_i,l-1}, r^*_{\boldsymbol{s}_i,l})$ where $r^*_{\boldsymbol{s}_i,l} = \log(r_{\boldsymbol{s}_i,l}/(1 - r_{\boldsymbol{s}_i,l}))$, for $l = 1, \ldots, i_L$. The posterior distribution of the model parameters, based on the latent variables t_i, is proportional to

$$
\begin{aligned}
& N(\boldsymbol{\beta} \mid \boldsymbol{\mu}_\beta, \boldsymbol{V}_\beta) \times \mathrm{IG}(\tau^2 \mid u_{\tau^2}, v_{\tau^2}) \times \mathrm{IG}(\sigma^2 \mid u_{\sigma^2}, v_{\sigma^2}) \times \mathrm{IG}(\phi \mid u_\phi, v_\phi) \times \mathrm{IG}(\zeta \mid u_\zeta, v_\zeta) \\
& \times N(\boldsymbol{\gamma} \mid \boldsymbol{\mu}_\gamma, \boldsymbol{V}_\gamma) \times \mathrm{IG}(\kappa^2 \mid u_{\kappa^2}, v_{\kappa^2}) \times \prod_{i=1}^{n} N(y(\boldsymbol{s}_i) \mid \boldsymbol{x}(\boldsymbol{s}_i)^\top \boldsymbol{\beta} + z(\boldsymbol{s}_i), \tau^2) \\
& \times N(\boldsymbol{t} \mid \boldsymbol{D}\boldsymbol{\gamma},\, \kappa^2 \mathbf{I}_{n-2}) \times N(z(\boldsymbol{s}_1) \mid 0, \sigma^2) \times N(z(\boldsymbol{s}_2) \mid \rho_1(\boldsymbol{s}_2) z(\boldsymbol{s}_1), \sigma^2(1 - (\rho_1(\boldsymbol{s}_2))^2)) \\
& \times \prod_{i=3}^{n} \sum_{l=1}^{i_L} N(z(\boldsymbol{s}) \mid \rho_l(\boldsymbol{s}_i) z(\boldsymbol{s}_{(il)}), \sigma^2(1 - (\rho_l(\boldsymbol{s}_i))^2)) \mathbb{1}_{(r^*_{\boldsymbol{s}_i,l-1}, r^*_{\boldsymbol{s}_i,l})}(t_i),
\end{aligned}
$$

where the vector $\boldsymbol{t} = (t_3, \ldots, t_n)^\top$, and the matrix $\boldsymbol{D}$ is $(n-2) \times 3$ such that the ith row is $(1, s_{2+i,1}, s_{2+i,2})$.

We describe the MCMC sampler to simulate from the posterior distribution of model parameters $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2, \phi, \zeta, \tau^2, \kappa^2)$ and latent variables $\{t_i\}_{i=3}^n, \{z(\boldsymbol{s}_i)\}_{i=1}^n$. Denote by $\boldsymbol{y}_{\mathcal{S}} = (y(\boldsymbol{s}_1), \ldots, y(\boldsymbol{s}_n))^\top$, $\boldsymbol{z}_{\mathcal{S}} = (z(\boldsymbol{s}_1), \ldots, z(\boldsymbol{s}_n))^\top$, and let $\boldsymbol{X}$ be the covariate matrix with the ith row being $\boldsymbol{x}(\boldsymbol{s}_i)^\top$. The posterior full conditional distribution for $\boldsymbol{\beta}$ is $N(\boldsymbol{\beta} \,|\, \boldsymbol{\mu}_\beta^*, \boldsymbol{V}_\beta^*)$ where $\boldsymbol{V}_\beta^* = (\boldsymbol{V}_\beta^{-1} + \tau^{-2}\boldsymbol{X}^\top\boldsymbol{X})^{-1}$ and $\boldsymbol{\mu}_\beta^* = \boldsymbol{V}_\beta^*(\boldsymbol{V}_\beta^{-1}\boldsymbol{\mu}_\beta + \tau^{-2}\boldsymbol{X}^\top(\boldsymbol{y}_{\mathcal{S}} - \boldsymbol{z}_{\mathcal{S}}))$. An inverse gamma prior for τ^2 yields an $\mathrm{IG}(\tau^2 \,|\, u_{\tau^2} + n/2, v_{\tau^2} + \sum_{i=1}^n e_i^2/2)$ posterior full conditional, where $e_i = y(\boldsymbol{s}_i) - \boldsymbol{x}(\boldsymbol{s}_i)^\top\boldsymbol{\beta} - z(\boldsymbol{s}_i)$.

Given the latent variable t_i, we have the configuration variable $\ell_i = l$ if $t_i \in (r^*_{\boldsymbol{s}_i,l-1}, r^*_{\boldsymbol{s}_i,l})$, for $i = 3, \ldots, n$. To update ζ, we first marginalize out the latent variables t_i from the joint posterior distribution. The posterior full conditional distribution of ζ is proportional to $\mathrm{IG}(\zeta \,|\, u_\zeta, v_\zeta)\prod_{i=3}^n\{G_{\boldsymbol{s}_i}(r_{\boldsymbol{s}_i,\ell_i} \,|\, \mu(\boldsymbol{s}_i), \kappa^2) - G_{\boldsymbol{s}_i}(r_{\boldsymbol{s}_i,\ell_i-1} \,|\, \mu(\boldsymbol{s}_i), \kappa^2)\}$. We update ζ on its log scale using a random walk Metropolis step. The posterior full conditional distribution of t_i is $\sum_{l=1}^{i_L} q_l(\boldsymbol{s}_i)\mathrm{TN}(t_i \,|\, \mu(\boldsymbol{s}_i), \kappa^2; r^*_{\boldsymbol{s}_i,l-1}, r^*_{\boldsymbol{s}_i,l})$, where $q_l(\boldsymbol{s}_i) \propto w_l(\boldsymbol{s}_i) f_{\boldsymbol{s}_i,l}(z(\boldsymbol{s}_i) \,|\, z(\boldsymbol{s}_{(il)}), \boldsymbol{\theta})$, and $w_l(\boldsymbol{s}_i) = G_{\boldsymbol{s}_i}(r_{\boldsymbol{s}_i,l} \,|\, \mu(\boldsymbol{s}_i), \kappa^2) - G_{\boldsymbol{s}_i}(r_{\boldsymbol{s}_i,l-1} \,|\, \mu(\boldsymbol{s}_i), \kappa^2)$, for $l = 1, ..., L$. Hence, each t_i can be readily updated by sampling from the l-th truncated Gaussian with probability proportional to $q_l(\boldsymbol{s}_i)$. The posterior full conditional distribution of $\boldsymbol{\gamma}$ is $N(\boldsymbol{\gamma} \,|\, \boldsymbol{\mu}_\gamma^*, \boldsymbol{V}_\gamma^*)$ where $\boldsymbol{V}_\gamma^* = (\boldsymbol{V}_\gamma^{-1} + \kappa^{-2}\boldsymbol{D}^\top\boldsymbol{D})^{-1}$ and $\boldsymbol{\mu}_\gamma^* = \boldsymbol{V}_\gamma^*(\boldsymbol{V}_\gamma^{-1}\boldsymbol{\mu}_\gamma + \kappa^{-2}\boldsymbol{D}^\top\boldsymbol{t})$. The posterior full conditional distribution of κ^2 is $\mathrm{IG}(\kappa^2 \,|\, u_{\kappa^2} + (n-2)/2, v_{\kappa^2} + \sum_{i=3}^n(t_i - \mu(\boldsymbol{s}_i))^2/2)$.

Since $\mathrm{Ne}(\boldsymbol{s}_2) = z(\boldsymbol{s}_1)$, we take $\ell_2 = 1$. To make expressions more compact, we let $\ell_1 = 0$, $\rho_0(\boldsymbol{s}_1) = 0$, and $z(\boldsymbol{s}_{(1,0)}) = 0$. The posterior full conditional distribution of σ^2 is $\mathrm{IG}(\sigma^2 \,|\, u_{\sigma^2} + n/2, v_{\sigma^2} + \sum_{i=1}^n(z(\boldsymbol{s}_i) - \rho_{\ell_i}(\boldsymbol{s}_i)z(\boldsymbol{s}_{(i,\ell_i)}))^2/\{2(1 - (\rho_{\ell_i}(\boldsymbol{s}_i))^2)\})$. The posterior full conditional distribution of ϕ is proportional to

$\mathrm{IG}(\phi \mid u_\phi, v_\phi) \prod_{i=2}^n N(z(\boldsymbol{s}_i) \mid \rho_{\ell_i}(\boldsymbol{s}_i) z(\boldsymbol{s}_{(i,\ell_i)}), \sigma_l^2(1-(\rho_{\ell_i}(\boldsymbol{s}_i))^2))$. We update ϕ on its log scale with a random walk Metropolis step. Denote by $\boldsymbol{A}_j^{(i)} = \{j : z(\boldsymbol{s}_{(j,\ell_j)}) = z(\boldsymbol{s}_i)\}$. The posterior full conditional distribution of the latent spatial random effects $z(\boldsymbol{s}_i)$, for $i = 1, \ldots, n$, is $N(z(\boldsymbol{s}_i) \mid \tilde{\sigma}_i^2 \tilde{\mu}_i, \tilde{\sigma}_i^2)$, where $\tilde{\sigma}_i^2 = \Big(\tau^{-2} + \sigma^{-2}(1 - (\rho_{\ell_i}(\boldsymbol{s}_i))^2)^{-1} + \sum_{j:j\in \boldsymbol{A}_j^{(i)}} \tilde{s}_{ij}^{-2}\Big)^{-1}$, and $\tilde{\mu}_i = \tau^{-2}(y(\boldsymbol{s}_i) - \boldsymbol{x}(\boldsymbol{s}_i)^\top \boldsymbol{\beta}) + \sigma^{-2}(1-(\rho_{\ell_i}(\boldsymbol{s}_i))^2)^{-1} \rho_{\ell_i}(\boldsymbol{s}_i) z(\boldsymbol{s}_{(i,\ell_i)}) + \sum_{j:j\in \boldsymbol{A}_j^{(i)}} z(\boldsymbol{s}_j)(\rho_{\ell_j}(\boldsymbol{s}_j))^{-1} \tilde{s}_{ij}^{-2}$ with $\tilde{s}_{ij}^2 = \sigma^2(1-(\rho_{\ell_j}(\boldsymbol{s}_j))^2)/(\rho_{\ell_j}(\boldsymbol{s}_j))^2$.

B.4.2 Skew-GNNMP Models

Bivariate skew-Gaussian distribution

Exploiting the location mixture representation of the skew-Gaussian distribution (Azzalini and Valle, 1996) for random vector $\boldsymbol{Z} = (U, V)$, we write

$$f(\boldsymbol{z} \mid z_0) \sim N\left(\begin{pmatrix} \xi_u + \lambda_u z_0 \\ \xi_v + \lambda_v z_0 \end{pmatrix}, \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right), \quad z_0 \sim N(z_0 \mid 0, 1) I(z_0 \geq 0). \quad \text{(B.4)}$$

It follows that, conditional on $Z_0 = z_0$, the marginal densities of $\boldsymbol{Z}$ are $N(u \mid \xi_u + \lambda_u z_0, \sigma^2)$ and $N(v \mid \xi_v + \lambda_v z_0, \sigma^2)$, respectively. Then the conditional density of Z_0 given $V = v$ is $p(z_0 \mid v) \propto N(z_0 \mid (v - \xi_v)/\lambda_v, \sigma^2/\lambda_v^2) N(z_0 \mid 0, 1) I(z_0 \geq 0)$. Therefore, the conditional density $p(z_0 \mid v)$ is a Gaussian distribution with mean $\mu_0(v) = (v - \xi_v)\lambda_v/(\sigma^2 + \lambda_v^2)$ and variance $\tau_0^2(v) = \sigma^2/(\sigma^2 + \lambda_v^2)$, truncated at $[0, \infty)$, denoted as $\mathrm{TN}_0(z_0 \mid \mu_0(v), \tau_0^2(v))$.

Then the conditional distribution of U given V can be written as

$$f_{U|V}(u \mid v) = \int_0^\infty N(u \mid \mu_u + \rho(v - \mu_v), \sigma^2(1-\rho^2)) \mathrm{TN}(z_0 \mid \mu_0(v), \tau_0^2(v)) dz_0, \quad \text{(B.5)}$$

where $\mu_u = \xi_u + \lambda_u z_0$, $\mu_v = \xi_v + \lambda_v z_0$.

Let $\boldsymbol{\xi} = (\xi_u, \xi_v)^\top$ and $\boldsymbol{\lambda} = (\lambda_u, \lambda_v)$. After marginalizing out z_0, the joint density of $\boldsymbol{Z}$ is given by $f(\boldsymbol{z}) = 2N(\boldsymbol{z} \,|\, \boldsymbol{\xi}, \boldsymbol{\Sigma})\, \Phi((1 - \boldsymbol{\lambda}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\lambda})^{-1/2} \boldsymbol{\lambda}^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{z} - \boldsymbol{\xi}))$, where $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{R} + \boldsymbol{\lambda}\boldsymbol{\lambda}^\top$, $\boldsymbol{R} = \left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right)$, and Φ is the cdf of a standard Gaussian distribution. The marginal density of U is $f_U(u) = 2N(u \,|\, \xi_u, \omega_u^2)\, \Phi(\alpha_u(u - \xi_u)/\omega_u)$, where $\omega_u^2 = \lambda_u^2 + \sigma^2$ and $\alpha_u = \lambda_u/\sigma$. We denote $f_U(u)$ as $\mathrm{SN}(u \,|\, \xi_u, \omega_u^2, \alpha_u)$. Similarly, the marginal density of V is $f_V(v) = \mathrm{SN}(\xi_v, \omega_v^2, \alpha_v)$. It follows that the conditional density of U given $V = v$ is

$$
\begin{aligned}
f_{U|V}(u \mid v) = N(u \mid \xi_u + \gamma(v - \xi_v), \tilde{\omega}^2) \\
\times \Phi(\alpha_1(u - \xi_u) + \alpha_2(v - \xi_v))/\Phi(\alpha_v(v - \xi_v)/\omega_v),
\end{aligned}
\tag{B.6}
$$

with $\gamma = (\rho\sigma^2 + \lambda_u\lambda_v)/(\sigma^2 + \lambda_v^2)$, $\tilde{\omega}^2 = \sigma^2 + \lambda_u^2 - (\rho\sigma^2 + \lambda_u\lambda_v)^2/(\sigma^2 + \lambda_v^2)$, $\alpha_1 = (\lambda_u - \rho\lambda_v)/m$, $\alpha_2 = (\lambda_v - \rho\lambda_u)/m$, and $m = \sqrt{(1-\rho)s^2}\sqrt{(1-\rho)s^2 + \lambda_u^2 + \lambda_v^2 - 2\rho\lambda_u\lambda_v}$.

In the special case where $\xi_u = \xi_v = 0$ and $\lambda_u = \lambda_v = \lambda$, let $\omega^2 = \lambda^2 + \sigma^2$ and $\alpha = \lambda/\sigma$. The joint density of $\boldsymbol{Z}$ can be written as $f(\boldsymbol{z}) = 2N(\boldsymbol{z} \mid \boldsymbol{0}, \boldsymbol{\Sigma})\, \Phi(\lambda(1 - \lambda^2 \boldsymbol{1}_2^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{1}_2)^{-1/2} \boldsymbol{1}_2^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{z})$, where the marginal density of $\boldsymbol{Z}$ is $\mathrm{SN}(x \,|\, 0, \omega^2, \alpha)$. The conditional density of U given $V = v$ is then given by

$$
f_{U|V}(u \,|\, v) = N(u \,|\, \tilde{\rho}v, \omega^2(1 - \tilde{\rho}^2))\, \Phi(\alpha'(u + v)/\omega')/\Phi(\alpha v/\omega), \tag{B.7}
$$

where $\tilde{\rho} = (\rho\sigma^2 + \lambda^2)/(\sigma^2 + \lambda^2)$, $\alpha' = \lambda/s$, $\omega'^2 = s^2 + 2\lambda^2$, and $s^2 = (1 + \rho)\sigma^2$.

Stationary skew-GNNMP models

We take a set of base random vectors $(U_l, V_l) \equiv (U, V)$ for all l, where (U, V) is a bivariate skew-Gaussian vector with distribution given by (B.4), and take $\xi_u = \xi_v = 0$, $\lambda_u = \lambda_v = \lambda$. We then extend (U_l, V_l) to $(U_{\boldsymbol{v},l}, V_{\boldsymbol{v},l})$ by extending ρ to $\rho_l(\boldsymbol{v})$ using an exponential correlation function such that $\rho_l(\boldsymbol{v}) \equiv \rho(||\boldsymbol{v} - \boldsymbol{v}_{(l)}||) = \exp(-||\boldsymbol{v} - \boldsymbol{v}_{(l)}||/\phi)$, for $l = 1, \ldots, L$. Using the resulting bivariate distribution for $(U_{\boldsymbol{v},l}, V_{\boldsymbol{v},l})$, we define the spatially varying density $f_{\boldsymbol{v},l} = f_{U_{\boldsymbol{v},l}|V_{\boldsymbol{v},l}}$ based on the formulation in (B.5). The resulting associated conditional density of the stationary skew-GNNMP is

$$p(y(\boldsymbol{v}) \,|\, \boldsymbol{y}_{\mathrm{Ne}(\boldsymbol{v})}) = \sum_{l=1}^{L} w_l(\boldsymbol{v}) \int_0^\infty N(y(\boldsymbol{v}) \,|\, \mu_l(\boldsymbol{v}), \sigma_l^2(\boldsymbol{v}))\mathrm{TN}(z_0(\boldsymbol{v}) \,|\, \mu_0(\boldsymbol{v}_{(l)}), \sigma_0^2)dz_0(\boldsymbol{v}), \tag{B.8}$$

where $\mu_l(\boldsymbol{v}) = (1-\rho_l(\boldsymbol{v}))\lambda z_0(\boldsymbol{v}) + \rho_l(\boldsymbol{v})y(\boldsymbol{v}_{(l)})$, $\sigma^2(\boldsymbol{v}) = \sigma^2(1-(\rho_l(\boldsymbol{v}))^2)$, $\mu_0(\boldsymbol{v}_{(l)}) = y(\boldsymbol{v}_{(l)})\lambda/(\sigma^2+\lambda^2)$, and $\sigma_0^2 = \sigma^2/(\sigma^2+\lambda^2)$.

The component conditional density in (B.8) is sampled via a latent variable $z_0(\boldsymbol{v})$. We marginalize out $z_0(\boldsymbol{v})$ to facilitate computation. Based on (B.7), we obtain the associated conditional density of the skew-GNNMP as

$$p(y(\boldsymbol{v}) \mid \boldsymbol{y}_{\mathrm{Ne}(\boldsymbol{v})}) = \sum_{l=1}^{L} w_l(\boldsymbol{v})\, b_l(\boldsymbol{v}) N(y(\boldsymbol{v}) \mid \tilde{\rho}_l(\boldsymbol{v})y(\boldsymbol{v}_{(l)}), \omega^2(1-(\tilde{\rho}_l(\boldsymbol{v}))^2)), \tag{B.9}$$

where $b_l(\boldsymbol{v}) = \Phi(\alpha_l'(\boldsymbol{v})(y(\boldsymbol{v}) + y(\boldsymbol{v}_{(l)}))/\omega_l'(\boldsymbol{v}))/\Phi(\alpha y(\boldsymbol{v}_{(l)})/\omega)$, $\tilde{\rho}_l(\boldsymbol{v}) = (\rho_l(\boldsymbol{v})\sigma^2 + \lambda^2)/(\sigma^2+\lambda^2)$, $\alpha_l'(\boldsymbol{v}) = \lambda/\sqrt{(1+\rho_l(\boldsymbol{v}))\sigma^2}$, $\omega_l'^2(\boldsymbol{v}) = (1+\rho_l(\boldsymbol{v}))\sigma^2 + 2\lambda^2$, $\alpha = \lambda/\sigma$, and $\omega^2 = \sigma^2 + \lambda^2$. For the weights $w_l(\boldsymbol{v})$, we use an exponential correlation function with range parameter ζ for the kernel functions of the random cutoff points. The Bayesian model is completed with prior specifications for $\lambda, \sigma^2, \phi, \zeta, \boldsymbol{\gamma}, \kappa^2$. In particular, we consider priors $N(\lambda \,|\, \mu_\lambda, \sigma_\lambda^2)$, $\mathrm{IG}(\sigma^2 \,|\, u_{\sigma^2}, v_{\sigma^2})$, $\mathrm{IG}(\phi \,|\, u_\phi, v_\phi)$, $\mathrm{IG}(\zeta \,|\, u_\zeta, v_\zeta)$, $N(\boldsymbol{\gamma} \,|\, \boldsymbol{\mu}_{\boldsymbol{\gamma}}, \boldsymbol{V}_{\boldsymbol{\gamma}})$ and $\mathrm{IG}(\kappa^2 \,|\, u_{\kappa^2}, v_{\kappa^2})$.

Given observations $y(\boldsymbol{s}_i)$, $i = 1, \ldots, n$, over reference set $\mathcal{S} = (\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n)$, we perform Bayesian inference based a likelihood conditional on the first L observations. The posterior distribution of the model parameters, given the conditional likelihood, is proportional to

$$N(\lambda \,|\, \mu_\lambda, \sigma_\lambda^2) \times \mathrm{IG}(\sigma^2 \,|\, u_{\sigma^2}, v_{\sigma^2}) \times \mathrm{IG}(\phi \,|\, u_\phi, v_\phi) \times \mathrm{IG}(\zeta \,|\, u_\zeta, v_\zeta)$$
$$\times\, N(\boldsymbol{\gamma} \,|\, \boldsymbol{\mu}_\gamma, \boldsymbol{V}_\gamma) \times \mathrm{IG}(\kappa^2 \,|\, u_{\kappa^2}, v_{\kappa^2}) \times N(\boldsymbol{t} \,|\, \boldsymbol{D}\boldsymbol{\gamma},\, \kappa^2 \mathbf{I}_{n-L})$$
$$\times \prod_{i=L+1}^{n} \sum_{l=1}^{L} b_l(\boldsymbol{s}_i) N(y(\boldsymbol{s}_i) \mid \tilde{\rho}_l(\boldsymbol{s}_i) y(\boldsymbol{s}_{(il)}), \omega^2(1-(\tilde{\rho}_l(\boldsymbol{s}_i))^2)) \mathbb{1}_{(r^*_{\boldsymbol{s}_i,l-1}, r^*_{\boldsymbol{s}_i,l})}(t_i),$$

where the vector $\boldsymbol{t} = (t_{L+1}, \ldots, t_n)^\top$, and the matrix $\boldsymbol{D}$ is $(n-L) \times 3$ such that the ith row is $(1, s_{L+i,1}, s_{L+i,2})$.

The MCMC sampler to obtain samples from the joint posterior distribution is described in Chapter 4. We present the posterior updates of λ, σ^2 and ϕ. Note that the configuration variables ℓ_i are such that $\ell_i = l$ if $t_i \in (r^*_{\boldsymbol{s}_i,l-1}, r^*_{\boldsymbol{s}_i,l})$ for $i \geq L+1$. Denote by $f_{\boldsymbol{s}_i,l} = b_l(\boldsymbol{s}_i) N(y(\boldsymbol{s}_i) \,|\, \tilde{\rho}_l(\boldsymbol{s}_i) y(\boldsymbol{s}_{(il)}), \omega^2(1-(\tilde{\rho}_l(\boldsymbol{s}_i))^2))$. We use a random walk Metropolis step to update parameter λ with target density $N(\lambda \,|\, \mu_\lambda, \sigma_\lambda^2) \prod_{i=L+1}^n f_{\boldsymbol{s}_i,\ell_i}$. The posterior full conditional distributions of σ^2 and ϕ are proportional to $\mathrm{IG}(\sigma^2 \,|\, u_{\sigma^2}, v_{\sigma^2}) \prod_{i=L+1}^n f_{\boldsymbol{s}_i,\ell_i}$, and $\mathrm{IG}(\phi \,|\, u_\phi, v_\phi) \prod_{i=L+1}^n f_{\boldsymbol{s}_i,\ell_i}$, respectively. For each parameter, we update it on its log scale with a random walk Metropolis step.

Extended skew-GNNMP models

Again, we take a set of base random vectors $(U_l, V_l) \equiv (U, V)$ for all l, where (U, V) is a bivariate skew-Gaussian vector with distribution given by (B.4). We extend (U_l, V_l) to $(U_{\boldsymbol{v},l}, V_{\boldsymbol{v},l})$ by extending ρ to $\rho_l(\boldsymbol{v})$ using an exponential correlation function such that $\rho_l(\boldsymbol{v}) \equiv \rho(||\boldsymbol{v} - \boldsymbol{v}_{(l)}||) = \exp(-||\boldsymbol{v} - \boldsymbol{v}_{(l)}||/\phi)$, and extending $\xi_u = \boldsymbol{x}(\boldsymbol{v})^\top \boldsymbol{\beta}$, $\xi_v = \boldsymbol{x}(\boldsymbol{v}_{(l)})^\top \boldsymbol{\beta}$, λ_u to $\lambda(\boldsymbol{v})$, and λ_v to $\lambda(\boldsymbol{v}_{(l)})$, for $l = 1, \ldots, L$. Using the resulting bivariate distribution for $(U_{\boldsymbol{v},l}, V_{\boldsymbol{v},l})$, we define the spatially varying density $f_{\boldsymbol{v},l} = f_{U_{\boldsymbol{v},l}|V_{\boldsymbol{v},l}}$ based on the formulation in (B.5). The resulting associated conditional density $p(y(\boldsymbol{v}) \,|\, \boldsymbol{y}_{\mathrm{Ne}(\boldsymbol{v})})$ of the extended skew-GNNMP is

$$\sum_{l=1}^{L} w_l(\boldsymbol{v}) \int_0^\infty N(y(\boldsymbol{v}) \,|\, \mu_l(\boldsymbol{v}), \sigma_l^2(\boldsymbol{v}))\mathrm{TN}(z_0(\boldsymbol{v}) \,|\, \mu_{0l}(\boldsymbol{v}_{(l)}), \sigma_{0l}^2(\boldsymbol{v}_{(l)}))dz_0(\boldsymbol{v}), \quad \text{(B.10)}$$

where $\mu_l(\boldsymbol{v}) = \boldsymbol{x}(\boldsymbol{v})^\top \boldsymbol{\beta} + \lambda(\boldsymbol{v})z_0(\boldsymbol{v}) + \rho_l(\boldsymbol{v})(y(\boldsymbol{v}_{(l)}) - \boldsymbol{x}(\boldsymbol{v}_{(l)})^\top \boldsymbol{\beta} - \lambda(\boldsymbol{v}_{(l)})z_0(\boldsymbol{v}))$, $\sigma_l^2(\boldsymbol{v}) = \sigma^2(1 - (\rho_l(\boldsymbol{v}))^2)$, $\mu_{0l}(\boldsymbol{v}_{(l)}) = (y(\boldsymbol{v}_{(l)}) - \boldsymbol{x}(\boldsymbol{v}_{(l)})^\top \boldsymbol{\beta})\lambda(\boldsymbol{v}_{(l)})/(\sigma^2 + (\lambda(\boldsymbol{v}_{(l)}))^2)$, and $\sigma_{0l}^2(\boldsymbol{v}_{(l)}) = \sigma^2/(\sigma^2 + (\lambda(\boldsymbol{v}_{(l)}))^2)$.

After marginalizing out $z_0(\boldsymbol{v})$, the conditional density (B.10) based on formulation (B.6) can be written as

$$p(y(\boldsymbol{v}) \mid \boldsymbol{y}_{\mathrm{Ne}(\boldsymbol{v})}) = \sum_{l=1}^{L} w_l(\boldsymbol{v})\, \tilde{b}_l(\boldsymbol{v}) N(y(\boldsymbol{v}) \mid \tilde{\mu}_l(\boldsymbol{v}), \tilde{\omega}_l^2(\boldsymbol{v})), \qquad \text{(B.11)}$$

with $\tilde{\mu}_l(\boldsymbol{v}) = \boldsymbol{x}(\boldsymbol{v})^\top \boldsymbol{\beta} + \tilde{\gamma}_l(\boldsymbol{v})(y(\boldsymbol{v}_{(l)}) - \boldsymbol{x}(\boldsymbol{v}_{(l)})^\top \boldsymbol{\beta})$, $s_l^2(\boldsymbol{v}) = (1 + \rho_l(\boldsymbol{v}))\sigma^2$, $\tilde{\gamma}_l(\boldsymbol{v}) = (\rho_l(\boldsymbol{v})\sigma^2 + \lambda(\boldsymbol{v})\lambda(\boldsymbol{v}_{(l)}))/\omega^2(\boldsymbol{v}_{(l)})$, $\tilde{\omega}_l^2(\boldsymbol{v}) = \omega(\boldsymbol{v})^2 - (\rho_l(\boldsymbol{v})\sigma^2 + \lambda(\boldsymbol{v})\lambda(\boldsymbol{v}_{(l)}))^2/\omega^2(\boldsymbol{v}_{(l)})$, $\alpha(\boldsymbol{v}) = \lambda(\boldsymbol{v})/\sigma$, $(\omega(\boldsymbol{v}))^2 = \lambda(\boldsymbol{v})^2 + \sigma^2$, and

$$\begin{aligned}
\tilde{b}_l(\boldsymbol{v}) &= \frac{\Phi(\alpha_1(\boldsymbol{v}, \boldsymbol{v}_{(l)})(y(\boldsymbol{v}) - \boldsymbol{x}(\boldsymbol{v})^\top \boldsymbol{\beta}) + \alpha_2(\boldsymbol{v}, \boldsymbol{v}_{(l)})(y(\boldsymbol{v}_{(l)}) - \boldsymbol{x}(\boldsymbol{v}_{(l)})^\top \boldsymbol{\beta}))}{\Phi(\alpha(\boldsymbol{v}_{(l)})(y(\boldsymbol{v}_{(l)}) - \boldsymbol{x}(\boldsymbol{v}_{(l)})^\top \boldsymbol{\beta})/\omega(\boldsymbol{v}_{(l)}))}, \\
\alpha_1(\boldsymbol{v}, \boldsymbol{v}_{(l)}) &= (\lambda(\boldsymbol{v}) - \rho_l(\boldsymbol{v})\lambda(\boldsymbol{v}_{(l)}))/m(\boldsymbol{v}), \\
\alpha_2(\boldsymbol{v}, \boldsymbol{v}_{(l)}) &= (\lambda(\boldsymbol{v}_{(l)}) - \rho_l(\boldsymbol{v})\lambda(\boldsymbol{v}))/m(\boldsymbol{v}), \\
m(\boldsymbol{v}) &= \sqrt{(1 - \rho_l(\boldsymbol{v}))s_l^2(\boldsymbol{v})} \\
&\quad \times \sqrt{(1 - \rho_l(\boldsymbol{v}))s_l^2(\boldsymbol{v}) + (\lambda(\boldsymbol{v}))^2 + (\lambda(\boldsymbol{v}_{(l)}))^2 - 2\rho_l(\boldsymbol{v})\lambda(\boldsymbol{v})\lambda(\boldsymbol{v}_{(l)})}.
\end{aligned}$$

We model the spatially varying $\lambda(\boldsymbol{v})$ via a partitioning approach. In particular, we partition the domain $\mathcal{D}$ such that $\mathcal{D} = \cup_{k=1}^{K} P_k$, $P_i \cap P_j = \emptyset$ for $i \neq j$. For all $\boldsymbol{v} \in P_k$, we take $\lambda(\boldsymbol{v}) = \lambda_k$, for $k = 1, \ldots, K$. For the weights $w_l(\boldsymbol{v})$, we use an exponential correlation function with range parameter ζ for the kernel function of the random cutoff points. The Bayesian model is completed with prior specifications for $\boldsymbol{\beta}, \boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K), \sigma^2, \phi, \zeta, \boldsymbol{\gamma}, \kappa^2$. We assign a $N(\boldsymbol{\beta}, \mid \boldsymbol{\mu}_\beta, \boldsymbol{V}_\beta)$ to the regression parameter $\boldsymbol{\beta}$ and $N(\lambda \mid \mu_{\lambda k}, \sigma^2_{\lambda k})$ to λ_k, for $k = 1, \ldots, K$. For other parameters, we take $\mathrm{IG}(\sigma^2 \mid u_{\sigma^2}, v_{\sigma^2})$, $\mathrm{IG}(\phi \mid u_\phi, v_\phi)$, $\mathrm{IG}(\zeta \mid u_\zeta, v_\zeta)$, $N(\boldsymbol{\gamma} \mid \boldsymbol{\mu}_\gamma, \boldsymbol{V}_\gamma)$ and $\mathrm{IG}(\kappa^2 \mid u_{\kappa^2}, v_{\kappa^2})$.

Given observations $y(\boldsymbol{s}_i)$, $i = 1, \ldots, n$, over reference set $\mathcal{S} = (\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n)$, we perform Bayesian inference based a likelihood conditional on the first L observations. The posterior distribution of the model parameters, given the conditional likelihood, is proportional to

$$
\begin{aligned}
& N(\boldsymbol{\beta} \mid \boldsymbol{\mu}_\beta, \boldsymbol{V}_\beta) \times \prod_{k=1}^{K} N(\lambda_k \mid \mu_{\lambda k}, \sigma^2_{\lambda k}) \times \mathrm{IG}(\sigma^2 \mid u_{\sigma^2}, v_{\sigma^2}) \times \mathrm{IG}(\phi \mid u_\phi, v_\phi) \\
& \times \mathrm{IG}(\zeta \mid u_\zeta, v_\zeta) \times N(\boldsymbol{\gamma} \mid \boldsymbol{\mu}_\gamma, \boldsymbol{V}_\gamma) \times \mathrm{IG}(\kappa^2 \mid u_{\kappa^2}, v_{\kappa^2}) \times N(\boldsymbol{t} \mid \boldsymbol{D}\boldsymbol{\gamma},\, \kappa^2 \mathbf{I}_{n-L}) \\
& \times \prod_{i=L+1}^{n} \sum_{l=1}^{L} \tilde{b}_l(\boldsymbol{s}_i) N(y(\boldsymbol{s}_i) \mid \tilde{\mu}_l(\boldsymbol{s}_i), \tilde{\omega}_l^2(\boldsymbol{s}_i)) \mathbb{1}_{(r^*_{\boldsymbol{s}_i,l-1}, r^*_{\boldsymbol{s}_i,l})}(t_i),
\end{aligned}
$$

where the vector $\boldsymbol{t} = (t_{L+1}, \ldots, t_n)^\top$, and the matrix $\boldsymbol{D}$ is $(n-L) \times 3$ such that the ith row is $(1, s_{L+i,1}, s_{L+i,2})$.

The MCMC sampler to obtain samples from the joint posterior distribution is described in Chapter 4. We present the posterior updates of $\boldsymbol{\beta}, \boldsymbol{\lambda}, \sigma^2$ and ϕ. Note that the configuration variables ℓ_i are such that $\ell_i = l$ if $t_i \in (r^*_{\boldsymbol{s}_i,l-1}, r^*_{\boldsymbol{s}_i,l})$ for $i \geq L+1$. Denote by $f_{\boldsymbol{s}_i,l} = \tilde{b}_l(\boldsymbol{s}_i) N(y(\boldsymbol{s}_i) \mid \tilde{\mu}_l(\boldsymbol{s}_i), \tilde{\omega}_l^2(\boldsymbol{s}_i))$. We use a random walk Metropolis step to update $\boldsymbol{\beta}$ with target density $N(\boldsymbol{\beta} \mid \boldsymbol{\mu}_\beta, \boldsymbol{V}_\beta) \prod_{i=L+1}^n f_{\boldsymbol{s}_i,\ell_i}$. Let $B_k = \{i : \boldsymbol{s}_i \in P_k\} \cup \{i : \boldsymbol{s}_{(i,\ell_i)} \in P_k\}$. The posterior full conditional of λ_k is proportional to $N(\lambda_k \mid \mu_{\lambda k}, \sigma^2_{\lambda k}) \prod_{i: i \in B_k} f_{\boldsymbol{s}_i,\ell_i}$, and we use a random walk Metropolis step to sample λ_k, for $k = 1, \ldots, K$. The posterior full conditional distributions of σ^2 and ϕ are proportional to $\mathrm{IG}(\sigma^2 \mid u_{\sigma^2}, v_{\sigma^2}) \prod_{i=L+1}^n f_{\boldsymbol{s}_i,\ell_i}$, and $\mathrm{IG}(\phi \mid u_\phi, v_\phi) \prod_{i=L+1}^n f_{\boldsymbol{s}_i,\ell_i}$, respectively. For each parameter, we update it on its log scale with a random walk Metropolis step.

B.4.3 Copula NNMP Models

We take a set of base random vectors $(U_l, V_l) \equiv (U, V)$ where its bivariate distribution is specified by a Gaussian copula with correlation parameter ρ. We extend (U_l, V_l) to $(U_{\boldsymbol{v},l}, V_{\boldsymbol{v},l})$ by extending ρ to $\rho_l(\boldsymbol{v}) \equiv \rho_l(||\boldsymbol{v} - \boldsymbol{v}_{(l)}||) = \exp(-||\boldsymbol{v} - \boldsymbol{v}_{(l)}||/\phi)$, creating a sequence of spatially varying Gaussian copula $C_{\boldsymbol{v},l}$ for $(U_{\boldsymbol{v},l}, V_{\boldsymbol{v},l})$ with marginal cdfs $F_{U_{\boldsymbol{v},l}} = F_{V_{\boldsymbol{v},l}} = F_Y$ for all $\boldsymbol{v}$ and all l. The cdf F_Y corresponds to the stationary marginal distribution of the model. The associated conditional density of the Gaussian copula NNMP is

$$p(y(\boldsymbol{v}) \mid \boldsymbol{y}_{\mathrm{Ne}(\boldsymbol{v})}) = \sum_{l=1}^{L} w_l(\boldsymbol{v}) \, c_{\boldsymbol{v},l}(y(\boldsymbol{v}), y(\boldsymbol{v}_{(l)})) f_Y(y(\boldsymbol{v})), \tag{B.12}$$

where $c_{\boldsymbol{v},l}(y(\boldsymbol{v}), y(\boldsymbol{v}_{(l)}))$ is the Gaussian copula density obtained by replacing ρ in (B.1) with $\rho_l(\boldsymbol{v})$, and f_Y is the density of F_Y.

Similarly, we can obtain the Gumbel copula NNMP model using a collection of spatially varying Gumbel copulas by extending η in (B.2) to $\eta_l(\boldsymbol{v}) \equiv \eta_l(||\boldsymbol{v} - \boldsymbol{v}_{(l)}||) = \min\{(1 - \exp(-||\boldsymbol{v} - \boldsymbol{v}_{(l)}||/\phi))^{-1}, 50\}$, where the upper bound 50 ensures numerical stability. We discuss the inferential approach for the Gaussian copula NNMP; the approach for the Gumbel copula NNMP model is similar. Assume the stationary marginal density is a gamma density, denoted as $f_Y = \mathrm{Ga}(a, b)$, with mean $E(Y) = a/b$. For the weights $w_l(\boldsymbol{v})$, we use an exponential correlation function with range parameter ζ to define the random cutoff points. The Bayesian model is completed with prior specifications for $a, b, \phi, \zeta, \boldsymbol{\gamma}, \kappa^2$. In particular, we consider priors $\mathrm{Ga}(u_a, v_a)$, $\mathrm{Ga}(u_b, v_b)$, $\mathrm{IG}(\phi \mid u_\phi, v_\phi)$, $\mathrm{IG}(\zeta \mid u_\zeta, v_\zeta)$, $N(\boldsymbol{\gamma} \mid \boldsymbol{\mu}_\gamma, \boldsymbol{V}_\gamma)$ and $\mathrm{IG}(\kappa^2 \mid u_{\kappa^2}, v_{\kappa^2})$.

Given observations $y(\boldsymbol{s}_i), i = 1, \ldots, n$, over reference set $\mathcal{S} = (\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n)$, we perform Bayesian inference using a likelihood conditional on $(y(\boldsymbol{s}_1), \ldots, y(\boldsymbol{s}_L))$.

The posterior distribution of the model parameters, given the conditional likelihood, is proportional to

$$\begin{aligned}
&\mathrm{Ga}(u_a, v_a) \times \mathrm{Ga}(u_b, v_b) \times \mathrm{IG}(\phi \,|\, u_\phi, v_\phi) \times \mathrm{IG}(\zeta \,|\, u_\zeta, v_\zeta) \\
&\times N(\boldsymbol{\gamma} \,|\, \boldsymbol{\mu}_\gamma, \boldsymbol{V}_\gamma) \times \mathrm{IG}(\kappa^2 \,|\, u_{\kappa^2}, v_{\kappa^2}) \times N(\boldsymbol{t} \,|\, \boldsymbol{D}\boldsymbol{\gamma},\, \kappa^2 \mathbf{I}_{n-L}) \\
&\times \prod_{i=L+1}^{n} \sum_{l=1}^{L} c_{\boldsymbol{s}_i,l}(y(\boldsymbol{s}_i), y(\boldsymbol{s}_{(il)})) f_Y(y(\boldsymbol{s}_i)) \mathbb{1}_{(r^*_{\boldsymbol{s}_i,l-1}, r^*_{\boldsymbol{s}_i,l})}(t_i),
\end{aligned}$$

where the vector $\boldsymbol{t} = (t_{L+1}, \ldots, t_n)^\top$, and the matrix $\boldsymbol{D}$ is $(n-L) \times 3$ such that the ith row is $(1, s_{L+i,1}, s_{L+i,2})$.

We provide the updates for parameters (a, b, ϕ). Note that the configuration variables ℓ_i are such that $\ell_i = l$ if $t_i \in (r^*_{\boldsymbol{s}_i,l-1}, r^*_{\boldsymbol{s}_i,l})$ for $i \geq L+1$. Denote by $f_{\boldsymbol{s}_i,l} = c_{\boldsymbol{s}_i,l}(y(\boldsymbol{s}_i), y(\boldsymbol{s}_{(il)})) f_Y(y(\boldsymbol{s}_i))$. The posterior full conditional distributions for parameters a, b and ϕ are proportional to $\mathrm{IG}(a \,|\, u_a, v_a) \prod_{i=L+1}^{n} f_{\boldsymbol{s}_i,\ell_i}$, $\mathrm{IG}(b \,|\, u_b, v_b) \prod_{i=L+1}^{n} f_{\boldsymbol{s}_i,\ell_i}$, and $\mathrm{IG}(\phi \,|\, u_\phi, v_\phi) \prod_{i=L+1}^{n} c_{\boldsymbol{s}_i,l}(y(\boldsymbol{s}_i), y(\boldsymbol{s}_{(il)}))$, respectively. Each parameter is updated on its log scale with a random walk Metropolis step.

B.5 DNNMP Models

In this section, we introduce necessary posterior simulation steps for the Poisson NNMP (PONNMP) and negative binomial NNMP (NBNNMP) models illustrated in the data examples. For both models, we use an exponential correlation function with range parameter ϕ to create spatial copulas. More specifically, given two different sites $\boldsymbol{v} \neq \boldsymbol{v}'$, the link functions for parameters of the Gaussian, Gumbel and Clayton copulas, respectively, are

$$\rho(||\boldsymbol{v} - \boldsymbol{v}'||) = \exp(-||\boldsymbol{v} - \boldsymbol{v}'||/\phi),$$
$$\eta(||\boldsymbol{v} - \boldsymbol{v}'||) = \min\{(1 - \exp(-||\boldsymbol{v} - \boldsymbol{v}'||/\phi))^{-1}, 50\},$$
$$\delta(||\boldsymbol{v} - \boldsymbol{v}'||) = \min\{2\exp(-||\boldsymbol{v} - \boldsymbol{v}'||/\phi)/(1 - \exp(-||\boldsymbol{v} - \boldsymbol{v}'||/\phi)), 98\},$$

where the upper bounds 50 and 98 for Gumbel and Clayton copulas are chosen for numerical stability. When $\eta(d_0) = 50$, $\exp(-d_0/\phi) = 0.98$. Similarly, when $\delta(d_0) = 98$, $\exp(-d_0/\phi) = 0.98$. Both link functions imply that given ϕ, the dependence implied by the copulas stays the same for any distance between $\boldsymbol{v}$ and $\boldsymbol{v}'$ smaller than d_0.

We assume that $\boldsymbol{y}_{\mathcal{S}} = (y(\boldsymbol{s}_1), \ldots, y(\boldsymbol{s}_n))^\top$ is a vector of observations, where $\mathcal{S} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$ is the reference set. Each $y(\boldsymbol{s}_i)$ is associated with $y^*(\boldsymbol{s}_i)$ such that $y^*(\boldsymbol{s}_i) = y(\boldsymbol{s}_i) - o_i$, where $o_i \equiv o(\boldsymbol{s}_i)$, $o(\boldsymbol{s}_i) \overset{i.i.d.}{\sim} \text{Unif}(0,1)$, for $i = 1, \ldots, n$. The auxiliary variables o_i is independent of $y(\boldsymbol{s}_i)$ and of o_j for $j \neq i$. Let $\boldsymbol{y}^*_{\text{Ne}(\boldsymbol{s}_i)} = (y^*(\boldsymbol{s}_{(i1)}), \ldots, y^*(\boldsymbol{s}_{(i,i_L)}))^\top$ and $\boldsymbol{o}_{\text{Ne}(\boldsymbol{s}_i)} = (o(\boldsymbol{s}_{(i1)}), \ldots, o(\boldsymbol{s}_{(i,i_L)}))^\top$, for $i = 2, \ldots, n$.

B.5.1 Poisson NNMP Models

The conditional density of the continued Poisson NNMP (PONNMP) over the reference set is given by

$$p(y^*(\boldsymbol{s}_i) \,|\, \boldsymbol{y}^*_{\text{Ne}(\boldsymbol{s}_i)}, o(\boldsymbol{s}_i), \boldsymbol{o}_{\text{Ne}(\boldsymbol{s}_i)}) = \sum_{l=1}^{i_L} w_l(\boldsymbol{s}_i)\, c^*_{\boldsymbol{s}_i,l}(y^*(\boldsymbol{s}_i), y^*(\boldsymbol{s}_{(il)}))f^*_Y(y^*(\boldsymbol{s}_i)),$$

for $i = 2, \ldots, n$, where $f^*_Y(y^*(\boldsymbol{s}_i)) = f_Y([y^*(\boldsymbol{s}_i) + 1])$, and f_Y is a Poisson distribution with rate parameter λ. The component $c^*_{\boldsymbol{s}_i,l}$ is the copula density of a spatial copula. We will illustrate the posterior inference using the Gaussian case as an example. The copula density $c^*_{\boldsymbol{s}_i,l}(y^*(\boldsymbol{s}_i), y^*(\boldsymbol{s}_{(il)}))$ of the spatial Gaussian copula is given by

$$\frac{1}{\sqrt{1-(\rho_l(\boldsymbol{s}_i))^2}}\exp\left(\frac{2\rho(\boldsymbol{s}_i)\Phi^{-1}(q_i)\Phi^{-1}(q_{il})-(\rho_l(\boldsymbol{s}_i))^2\{(\Phi^{-1}(q_i))^2+(\Phi^{-1}(q_{il}))^2\}}{2(1-(\rho_l(\boldsymbol{s}_i))^2)}\right),$$

where $\rho_l(\boldsymbol{s}_i)\equiv\rho(||\boldsymbol{s}_i-\boldsymbol{s}'_{(il)}||)=\exp(-||\boldsymbol{s}_i-\boldsymbol{s}_{(il)}||/\phi)$, $q_i=F_Y^*(y^*(\boldsymbol{s}_i))$, $q_{il}=F_Y^*(y^*(\boldsymbol{s}_{(il)}))$, and F_Y^* is the cdf of f_Y^*.

The formulation of the mixture weights allows us to augment the model with a sequence of auxiliary variables t_i, $i=3,\ldots,n$, where t_i is a Gaussian random variable with mean $\mu(\boldsymbol{s}_i)=\gamma_0+s_{i1}\gamma_1+s_{i2}\gamma_2$ and variance κ^2. The conditional density of the augmented model on $y^*(\boldsymbol{s}_i)$ is

$$p(y^*(\boldsymbol{s}_i)\,|\,\boldsymbol{y}^*_{\mathrm{Ne}(\boldsymbol{s}_i)},o(\boldsymbol{s}_i),\boldsymbol{o}_{\mathrm{Ne}(\boldsymbol{s}_i)})=\sum_{l=1}^{i_L}c^*_{\boldsymbol{s}_i,l}(y^*(\boldsymbol{s}_i),y^*(\boldsymbol{s}_{(il)}))f_Y^*(y^*(\boldsymbol{s}_i))\mathbb{1}_{(r^*_{\boldsymbol{s}_i,l-1},r^*_{\boldsymbol{s}_i,l})}(t_i),$$

for $i=3,\ldots,n$, where $r^*_{\boldsymbol{s}_i,l}=\log(r_{\boldsymbol{s}_i,l}/(1-r_{\boldsymbol{s}_i,l}))$ for $l=1,\ldots,i_L$. The random cutoff points $r_{\boldsymbol{s}_i,l}$ is defined such that $r_{\boldsymbol{s}_i,l}-r_{\boldsymbol{s}_i,l-1}=k'(\boldsymbol{s}_i,\boldsymbol{s}_{(il)})/\sum_{l=1}^{i_L}k'(\boldsymbol{s}_i,\boldsymbol{s}_{(il)})$, where $k'(\boldsymbol{s}_i,\boldsymbol{s}_{(il)})=\exp(||\boldsymbol{s}_i-\boldsymbol{s}_{(il)}||/\zeta)$.

Let $\boldsymbol{\gamma}=(\gamma_0,\gamma_1,\gamma_2)$. The Bayesian model is completed with prior specifications for $(\lambda,\phi,\zeta,\boldsymbol{\gamma},\kappa^2)$. Let $f^*_{\boldsymbol{s}_i,l}(y^*(\boldsymbol{s}_i)\,|\,y^*(\boldsymbol{s}_{(il)}))=c^*_{\boldsymbol{s}_i,l}(y^*(\boldsymbol{s}_i),y^*(\boldsymbol{s}_{(il)}))f_Y^*(y^*(\boldsymbol{s}_i))$. With customary prior specifications, the joint distribution can be written as

$$\begin{aligned}&\mathrm{Ga}(\lambda\,|\,u_\lambda,v_\lambda)\times\mathrm{IG}(\phi\,|\,u_\phi,v_\phi)\times\mathrm{IG}(\zeta\,|\,u_\zeta,v_\zeta)\times N(\boldsymbol{\gamma}\,|\,\boldsymbol{\mu}_\gamma,\boldsymbol{V}_\gamma)\times\mathrm{IG}(\kappa^2\,|\,u_{\kappa^2},v_{\kappa^2})\\&\times N(\boldsymbol{t}\,|\,\boldsymbol{D\gamma},\,\kappa^2\mathbf{I}_{n-2}))\times f_Y^*(y(\boldsymbol{s}_1)-o_1\,|\,\lambda)\times f^*_{\boldsymbol{s}_2,1}(y(\boldsymbol{s}_2)-o_2\,|\,y(\boldsymbol{s}_1)-o_1,\lambda,\phi)\\&\times\prod_{i=1}^{n}\mathrm{Unif}(o_i\,|\,0,1)\times\prod_{i=3}^{n}\sum_{l=1}^{i_L}f^*_{\boldsymbol{s}_i,l}(y(\boldsymbol{s}_i)-o_i\,|\,y(\boldsymbol{s}_{(il)})-o_{(il)},\lambda,\phi)\mathbb{1}_{(r^*_{\boldsymbol{s}_i,l-1},r^*_{\boldsymbol{s}_i,l})}(t_i),\end{aligned}$$

where $o_{(il)}\equiv o(\boldsymbol{s}_{(il)})$, the vector $\boldsymbol{t}=(t_3,\ldots,t_n)^\top$, and the matrix $\boldsymbol{D}$ is $(n-2)\times 3$ such that the ith row is $(1,s_{2+i,1},s_{2+i,2})$.

The MCMC algorithm to obtain posterior samples consists of updates from the posterior full conditional distribution of each of $(\lambda, \phi, \zeta, \boldsymbol{\gamma}, \kappa^2, \{t_i\}_{i=3}^n, \{o_i\}_{i=1}^n)$. The posterior full conditional distributions of each of $(\boldsymbol{\gamma}, \kappa^2, \{t_i\}_{i=3}^n, \{o_i\}_{i=1}^n)$ are described in Chapter 5. We focus on the posterior updates for (λ, ϕ, ζ). Note that there is a set of configuration variables $\{\ell_i\}_{i=3}^n$ in one-to-one correspondence with t_i, i.e., $\ell_i = l$ if and only if $t_i \in (r^*_{\boldsymbol{s}_i,l-1}, r^*_{\boldsymbol{s}_i,l})$, for $l = 1, \ldots, i_L$. We take $\ell_2 = 1$. The posterior full conditional distributions of λ and ϕ are proportional to $\mathrm{Ga}(\lambda \,|\, u_\lambda, v_\lambda) f^*_Y(y(\boldsymbol{s}_1) - o_1) \prod_{i=2}^n f^*_{\boldsymbol{s}_i,\ell_i}(y(\boldsymbol{s}_i) - o_i \,|\, y(\boldsymbol{s}_{(i,\ell_i)}) - o_{(i,\ell_i)})$ and $\mathrm{IG}(\phi \,|\, u_\phi, v_\phi) \prod_{i=2}^n c^*_{\boldsymbol{s}_i,\ell_i}(y(\boldsymbol{s}_i) - o_i, y(\boldsymbol{s}_{(i,\ell_i)}) - o_{(i,\ell_i)})$, respectively. For each parameter, we update it on its log scale with a random walk Metropolis step. To update ζ, we first marginalize out the latent variables t_i from the joint posterior distribution. The posterior full conditional distribution of ζ is proportional to $\mathrm{IG}(\zeta \,|\, u_\zeta, v_\zeta) \prod_{i=3}^n \{G_{\boldsymbol{s}_i}(r_{\boldsymbol{s}_i,\ell_i} \,|\, \mu(\boldsymbol{s}_i), \kappa^2) - G_{\boldsymbol{s}_i}(r_{\boldsymbol{s}_i,\ell_i-1} \,|\, \mu(\boldsymbol{s}_i), \kappa^2)\}$. We update ζ on its log scale with a random walk Metropolis step.

B.5.2 Negative Binomial NNMP Models

The conditional density of the continued negative binomial NNMP (NBNNMP) over the reference set is given by

$$p(y^*(\boldsymbol{s}_i) \,|\, \boldsymbol{y}^*_{\mathrm{Ne}(\boldsymbol{s}_i)}, o(\boldsymbol{s}_i), \boldsymbol{o}_{\mathrm{Ne}(\boldsymbol{s}_i)}) = \sum_{l=1}^{i_L} w_l(\boldsymbol{s}_i) c^*_{\boldsymbol{s}_i,l}(y^*(\boldsymbol{s}_i), y^*(\boldsymbol{s}_{(il)})) g_{\boldsymbol{s}_i}(y^*(\boldsymbol{s}_i)),$$

for $i = 2, \ldots, n$, where $g^*_{\boldsymbol{s}_i}(y^*(\boldsymbol{s}_i)) = g_{\boldsymbol{s}_i}([y^*(\boldsymbol{s}_i) + 1])$, and $g_{\boldsymbol{s}_i}$ is a negative binomial distribution with mean $\alpha(\boldsymbol{s}_i) = \exp(\boldsymbol{x}(\boldsymbol{s}_i)^\top \boldsymbol{\beta})$ and dispersion parameter r. Similar to the Poisson case, we illustrate the posterior inference using a spatial Gaussian copula with copula density $c^*_{\boldsymbol{s}_i,l}(y^*(\boldsymbol{s}_i), y^*(\boldsymbol{s}_{(il)}))$ given by

$$\frac{1}{\sqrt{1 - (\rho_l(\boldsymbol{s}_i))^2}} \exp\left(\frac{2\rho(\boldsymbol{s}_i)\Phi^{-1}(q_i)\Phi^{-1}(q_{il}) - (\rho_l(\boldsymbol{s}_i))^2 \{(\Phi^{-1}(q_i))^2 + (\Phi^{-1}(q_{(il)}))^2\}}{2(1 - (\rho_l(\boldsymbol{s}_i))^2)} \right),$$

where $\rho_l(\boldsymbol{s}_i) \equiv \rho(||\boldsymbol{s}_i - \boldsymbol{s}'_{(il)}||) = \exp(-||\boldsymbol{s}_i - \boldsymbol{s}_{(il)}||/\phi)$, $q_i = Q^*_{\boldsymbol{s}_i}(y^*(\boldsymbol{s}_i))$, $q_{il} = Q^*_{\boldsymbol{s}_{(il)}}(y^*(\boldsymbol{s}_{(il)}))$, and $Q^*_{\boldsymbol{s}_i}$ is the cdf of $g^*_{\boldsymbol{s}_i}$ for all $\boldsymbol{s}_i$.

Similarly, we use an exponential correlation function for the cutoff point kernel k', and augment the model with a set of Gaussian random variables t_i with mean $\mu(\boldsymbol{s}_i)$ and κ^2. Let $f^*_{\boldsymbol{s}_i,l}(y^*(\boldsymbol{s}_i) \mid y^*(\boldsymbol{s}_{(il)})) = c^*_{\boldsymbol{s}_i,l}(y^*(\boldsymbol{s}_i), y^*(\boldsymbol{s}_{(il)}))g^*_{\boldsymbol{s}_i}(y^*(\boldsymbol{s}_i))$. With customary prior specifications, we obtain the joint distribution

$$
\begin{aligned}
&N(\boldsymbol{\beta} \mid \boldsymbol{\mu}_\beta, \boldsymbol{V}_\beta) \times \mathrm{IG}(r \mid u_r, v_r) \times \mathrm{IG}(\phi \mid u_\phi, v_\phi) \times \mathrm{IG}(\zeta \mid u_\zeta, v_\zeta) \times N(\boldsymbol{\gamma} \mid \boldsymbol{\mu}_\gamma, \boldsymbol{V}_\gamma) \\
&\times \mathrm{IG}(\kappa^2 \mid u_{\kappa^2}, v_{\kappa^2}) \times N(\boldsymbol{t} \mid \boldsymbol{D}\boldsymbol{\gamma},\ \kappa^2 \mathbf{I}_{n-2})) \times \prod_{i=1}^{n} \mathrm{Unif}(o_i \mid 0, 1) \\
&\times g^*_{\boldsymbol{s}_1}(y(\boldsymbol{s}_1) - o_1 \mid \boldsymbol{\beta}, r) \times f^*_{\boldsymbol{s}_2,1}(y(\boldsymbol{s}_2) - o_2 \mid y(\boldsymbol{s}_1) - o_1, \boldsymbol{\beta}, r, \phi) \\
&\times \prod_{i=3}^{n} \sum_{l=1}^{i_L} f^*_{\boldsymbol{s}_i,l}(y(\boldsymbol{s}_i) - o_i \mid y(\boldsymbol{s}_{(il)}) - o_{(il)}, \boldsymbol{\beta}, r, \phi) \mathbb{1}_{(r^*_{\boldsymbol{s}_i,l-1}, r^*_{\boldsymbol{s}_i,l})}(t_i),
\end{aligned}
$$

where $o_{(il)} \equiv o(\boldsymbol{s}_{(il)})$, the vector $\boldsymbol{t} = (t_3, \ldots, t_n)^\top$, and the matrix $\boldsymbol{D}$ is $(n-2) \times 3$ such that the ith row is $(1, s_{2+i,1}, s_{2+i,2})$.

The MCMC algorithm to obtain posterior samples consists of updates from the posterior full conditional distribution of each of $(\boldsymbol{\beta}, r, \phi, \zeta, \boldsymbol{\gamma}, \kappa^2, \{t_i\}_{i=3}^n, \{o_i\}_{i=1}^n)$. The posterior full conditional distributions of each of $(\boldsymbol{\gamma}, \kappa^2, \{t_i\}_{i=3}^n, \{o_i\}_{i=1}^n)$ are described in Chapter 5. We focus on the posterior updates for $(\boldsymbol{\beta}, r, \phi, \zeta)$. Note that there is a set of configuration variables $\{\ell_i\}_{i=3}^n$ in one-to-one correspondence with t_i, i.e., $\ell_i = l$ if and only if $t_i \in (r^*_{\boldsymbol{s}_i,l-1}, r^*_{\boldsymbol{s}_i,l})$, for $l = 1, \ldots, i_L$. We take $\ell_2 = 1$. The posterior full conditional distributions of $\boldsymbol{\beta}$ and r are proportional to $N(\boldsymbol{\beta} \mid \boldsymbol{\mu}_\beta, \boldsymbol{V}_\beta) g^*_{\boldsymbol{s}_1}(y(\boldsymbol{s}_1) - o_1) \prod_{i=2}^n f^*_{\boldsymbol{s}_i,\ell_i}(y(\boldsymbol{s}_i) - o_i \mid y(\boldsymbol{s}_{(i,\ell_i)}) - o_{(i,\ell_i)})$ and $\mathrm{IG}(r \mid u_r, v_r) g^*_{\boldsymbol{s}_1}(y(\boldsymbol{s}_1) - o_1) \prod_{i=2}^n f^*_{\boldsymbol{s}_i,\ell_i}(y(\boldsymbol{s}_i) - o_i \mid y(\boldsymbol{s}_{(i,\ell_i)}) - o_{(i,\ell_i)})$, respectively. We use a random walk Metropolis step to update $\boldsymbol{\beta}$ and r on its log scale, respectively. The posterior full conditional distribution of ϕ is proportional to $\mathrm{IG}(\phi \mid u_\phi, v_\phi) \prod_{i=2}^n c^*_{\boldsymbol{s}_i,\ell_i}(y(\boldsymbol{s}_i) - o_i, y(\boldsymbol{s}_{(i,\ell_i)}) - o_{(i,\ell_i)})$. We update ϕ on its log scale with a random walk Metropolis step. To update ζ, we first marginalize out the latent variables t_i from the joint posterior distribution. The posterior full conditional distribution of ζ is proportional to $\mathrm{IG}(\zeta \mid u_\zeta, v_\zeta) \prod_{i=3}^{n} \{G_{\boldsymbol{s}_i}(r_{\boldsymbol{s}_i,\ell_i} \mid \mu(\boldsymbol{s}_i), \kappa^2) - G_{\boldsymbol{s}_i}(r_{\boldsymbol{s}_i,\ell_i - 1} \mid \mu(\boldsymbol{s}_i), \kappa^2)\}$. We update ζ on its log scale with a random walk Metropolis step.

Bibliography

Allcroft, D. J. and Glasbey, C. A. (2003), "A latent Gaussian Markov random-field model for spatiotemporal rainfall disaggregation," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 52, 487–498.

Antoniano-Villalobos, I. and Walker, S. G. (2016), "A nonparametric model for stationary time series," *Journal of Time Series Analysis*, 37, 126–142.

Arnold, B. C., Castillo, E., and Sarabia, J. M. (1999), *Conditional specification of statistical models*, New York: Springer.

Azzalini, A. (2013), *The skew-normal and related families*, Cambridge, UK: Cambridge University Press.

Azzalini, A. and Valle, A. D. (1996), "The multivariate skew-normal distribution," *Biometrika*, 83, 715–726.

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008), "Gaussian predictive process models for large spatial data sets," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70, 825–848.

Barata, R., Prado, R., and Sansó, B. (2022), "Fast inference for time-varying quantiles via flexible dynamic models with application to the characterization of atmospheric rivers," *The Annals of Applied Statistics*, 16, 247–271.

Bárdossy, A. (2006), “Copula-based geostatistical models for groundwater quality parameters,” *Water Resources Research*, 42.

Bartolucci, F. and Farcomeni, A. (2010), “A note on the mixture transition distribution and hidden Markov models,” *Journal of Time Series Analysis*, 31, 132–138.

Bauwens, L. and Veredas, D. (2004), “The stochastic conditional duration model: a latent variable model for the analysis of financial durations,” *Journal of Econometrics*, 119, 381–412.

Beck, N., Genest, C., Jalbert, J., and Mailhot, M. (2020), “Predicting extreme surges from sparse data using a copula-based hierarchical Bayesian spatial model,” *Environmetrics*, 31, e2616.

Berchtold, A. (2001), “Estimation in the mixture transition distribution model,” *Journal of Time Series Analysis*, 22, 379–397.

— (2003), “Mixture transition distribution (MTD) modeling of heteroscedastic time series,” *Computational statistics & data analysis*, 41, 399–411.

Berchtold, A. and Raftery, A. (2002), “The mixture transition distribution model for high-order Markov chains and non-Gaussian time series,” *Statistical Science*, 17, 328–356.

Berrett, C. and Calder, C. A. (2016), “Bayesian spatial binary classification,” *Spatial Statistics*, 16, 72–102.

Bevilacqua, M., Caamaño-Carrillo, C., Arellano-Valle, R. B., and Morales-Oñate, V. (2021), “Non-Gaussian geostatistical modeling using (skew) t processes,” *Scandinavian Journal of Statistics*, 48, 212–245.

Bevilacqua, M., Caamaño-Carrillo, C., and Gaetan, C. (2020), "On modeling positive continuous data with spatiotemporal dependence," *Environmetrics*, 31, e2632.

Bhogal, S. K. and Thekke Variyam, R. (2019), "Conditional duration models for high-frequency data: a review on recent developments," *Journal of Economic Surveys*, 33, 252–273.

Bishop, C. M. (2006), *Pattern Recognition and Machine Learning*, New York:Springer.

Bolano, D. and Berchtold, A. (2016), "General framework and model building in the class of hidden mixture transition distribution models," *Computational Statistics & Data Analysis*, 93, 131–145.

Bolin, D. (2014), "Spatial Matérn fields driven by non-Gaussian noise," *Scandinavian Journal of Statistics*, 41, 557–579.

Bolin, D. and Wallin, J. (2020), "Multivariate type G Matérn stochastic partial differential equation random fields," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82, 215–239.

Bouzaiene, M., Menna, M., Poulain, P.-M., Bussani, A., and Elhmaidi, D. (2020), "Analysis of the surface dispersion in the Mediterranean sub-basins," *Frontiers in Marine Science*, 7, 486.

Brockwell, P. J. and Davis, R. A. (1991), *Time series: theory and methods*, New York: Springer.

Brownlees, C. T. and Vannucci, M. (2013), "A Bayesian approach for capturing daily heterogeneity in intra-daily durations time series," *Studies in Nonlinear Dynamics and Econometrics*, 17, 21–46.

Burnett, J. L., Wszola, L., and Palomo-Muñoz, G. (2019), "bbsAssistant: An R package for downloading and handling data and information from the North American Breeding Bird Survey," *Journal of Open Source Software*, 4, 1768.

Cadonna, A., Kottas, A., and Prado, R. (2019), "Bayesian spectral modeling for multiple time series," *Journal of the American Statistical Association*, 114, 1838–1853.

Cervone, D., Pillai, N. S., Pati, D., Berbeco, R., Lewis, J. H., et al. (2014), "A location-mixture autoregressive model for online forecasting of lung tumor motion," *The Annals of Applied Statistics*, 8, 1341–1371.

Chan, A. B. and Dong, D. (2011), "Generalized Gaussian process models." in *CVPR*.

Chen, F. and Stindl, T. (2018), "Direct likelihood evaluation for the renewal Hawkes process," *Journal of Computational and Graphical Statistics*, 27, 119–131.

— (2022), *RHawkes: Renewal Hawkes Process*, URL `https://cran.r-project.org/package=RHawkes`. R package version 1.0.

Christensen, O. F., Roberts, G. O., and Sköld, M. (2006), "Robust Markov chain Monte Carlo methods for spatial generalized linear mixed models," *Journal of Computational and Graphical Statistics*, 15, 1–17.

Christensen, O. F. and Waagepetersen, R. (2002), "Bayesian prediction of spatial count data using generalized linear mixed models," *Biometrics*, 58, 280–286.

Coen, A., Gutiérrez, L., and Mena, R. H. (2019), "Modelling failures times with dependent renewal type models via exchangeability," *Statistics*, 53, 1112–1130.

Connor, R. J. and Mosimann, J. E. (1969), "Concepts of independence for proportions with a generalization of the Dirichlet distribution," *Journal of the American Statistical Association*, 64, 194–206.

Cook, R. J., Lawless, J. F., et al. (2007), *The statistical analysis of recurrent events*, Springer.

Cowpertwait, P. S. (2001), "A renewal cluster model for the inter-arrival times of rainfall events," *International Journal of Climatology*, 21, 49–61.

Dai, B., Ding, S., Wahba, G., et al. (2013), "Multivariate Bernoulli distribution," *Bernoulli*, 19, 1465–1483.

Daley, D. J. and Vere-Jones, D. (2003), "An introduction to the theory of point processes. Vol. I. Probability and its Applications," .

Danaher, P. J. and Smith, M. S. (2011), "Modeling multivariate distributions using copulas: Applications in marketing," *Marketing Science*, 30, 4–21.

Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016a), "Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets," *Journal of the American Statistical Association*, 111, 800–812.

Datta, A., Banerjee, S., Finley, A. O., Hamm, N. A., and Schaap, M. (2016b), "Nonseparable dynamic nearest neighbor Gaussian process models for large spatio-temporal data with an application to particulate matter analysis," *The Annals of Applied Statistics*, 10, 1286.

Davison, A. C., Padoan, S. A., and Ribatet, M. (2012), "Statistical modeling of spatial extremes," *Statistical science*, 27, 161–186.

De Oliveira, V., Kedem, B., and Short, D. A. (1997), "Bayesian prediction of transformed Gaussian random fields," *Journal of the American Statistical Association*, 92, 1422–1433.

Denuit, M. and Lambert, P. (2005), "Constraints on concordance measures in bivariate discrete data," *Journal of Multivariate Analysis*, 93, 40–57.

Deo, R., Hsieh, M., and Hurvich, C. M. (2010), "Long memory in intertrade durations, counts and realized volatility of NYSE stocks," *Journal of Statistical Planning and Inference*, 140, 3715–3733.

DeYoreo, M. and Kottas, A. (2017), "A Bayesian nonparametric Markovian model for non-stationary time series," *Statistics and Computing*, 27, 1525–1538.

Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998), "Model-based geostatistics," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47, 299–350.

Dunn, P. K. and Smyth, G. K. (1996), "Randomized quantile residuals," *Journal of Computational and Graphical Statistics*, 5, 236–244.

Easley, D. and O'hara, M. (1992), "Time and the process of security price adjustment," *The Journal of finance*, 47, 577–605.

Eltoft, T., Kim, T., and Lee, T.-W. (2006), "On the multivariate Laplace distribution," *IEEE Signal Processing Letters*, 13, 300–303.

Engle, R. F. and Russell, J. R. (1998), "Autoregressive conditional duration: a new model for irregularly spaced transaction data," *Econometrica*, 1127–1162.

Escarela, G., Mena, R. H., and Castillo-Morales, A. (2006), "A flexible class of parametric transition regression models based on copulas: application to poliomyelitis incidence," *Statistical Methods in Medical Research*, 15, 593–609.

Eskelson, B. N., Madsen, L., Hagar, J. C., and Temesgen, H. (2011), "Estimating riparian understory vegetation cover with beta regression and copula models," *Forest Science*, 57, 212–221.

Ferguson, T. S. (1973), "A Bayesian analysis of some nonparametric problems," *The Annals of Statistics*, 1, 209–230.

Fernandes, M. and Grammig, J. (2006), "A family of autoregressive conditional duration models," *Journal of Econometrics*, 130, 1–23.

Filimonov, V. and Sornette, D. (2012), "Quantifying reflexivity in financial markets: Toward a prediction of flash crashes," *Physical Review E*, 85, 056108.

Finley, A., Datta, A., and Banerjee, S. (2020), *spNNGP: Spatial Regression Models for Large Datasets using Nearest Neighbor Gaussian Processes*. R package version 0.1.4.

Finley, A. O., Banerjee, S., and Carlin, B. P. (2007), "spBayes: An R Package for Univariate and Multivariate Hierarchical Point-Referenced Spatial Models," *Journal of Statistical Software*, 19, 1–24.

Finley, A. O., Datta, A., Cook, B. D., Morton, D. C., Andersen, H. E., and Banerjee, S. (2019), "Efficient algorithms for Bayesian nearest neighbor Gaussian processes," *Journal of Computational and Graphical Statistics*, 28, 401–414.

Fong, P. W., Li, W. K., Yau, C., and Wong, C. (2007), "On a mixture vector autoregressive model," *Canadian Journal of Statistics*, 35, 135–150.

Gaver, D. P. and Lewis, P. (1980), "First-order autoregressive gamma sequences and point processes," *Advances in Applied Probability*, 12, 727–745.

Gelfand, A. E. and Ghosh, S. K. (1998), "Model choice: a minimum posterior predictive loss approach," *Biometrika*, 85, 1–11.

Gelfand, A. E., Kottas, A., and MacEachern, S. N. (2005), "Bayesian nonparametric spatial modeling with Dirichlet process mixing," *Journal of the American Statistical Association*, 100, 1021–1035.

Genest, C. and Nešlehová, J. (2007), "A primer on copulas for count data," *ASTIN Bulletin: The Journal of the IAA*, 37, 475–515.

Georges, P., Lamy, A.-G., Nicolas, E., Quibel, G., and Roncalli, T. (2001), "Multivariate survival modelling: a unified approach with copulas," *Available at SSRN 1032559*.

Ghosh, S. and Mallick, B. K. (2011), "A hierarchical Bayesian spatio-temporal model for extreme precipitation events," *Environmetrics*, 22, 192–204.

Gneiting, T. and Raftery, A. E. (2007), "Strictly proper scoring rules, prediction, and estimation," *Journal of the American Statistical Association*, 102, 359–378.

Gräler, B. (2014), "Modelling skewed spatial random fields through the spatial vine copula," *Spatial Statistics*, 10, 87–102.

Gramacy, R. B. and Apley, D. W. (2015), "Local Gaussian process approximation for large computer experiments," *Journal of Computational and Graphical Statistics*, 24, 561–578.

Grammig, J. and Maurer, K.-O. (2000), "Non-monotonic hazard functions and the autoregressive conditional duration model," *The Econometrics Journal*, 3, 16–38.

Grenier, I. and Sansó, B. (2021), "Distributed nearest-neighbor Gaussian processes," *Communications in Statistics-Simulation and Computation*, 1–13.

Gruber, L. F. and Czado, C. (2018), "Bayesian model selection of regular vine copulas," *Bayesian Analysis*, 13, 1111–1135.

Guan, Y. and Haran, M. (2018), "A computationally efficient projection-based approach for spatial generalized linear mixed models," *Journal of Computational and Graphical Statistics*, 27, 701–714.

Guhaniyogi, R., Li, C., Savitsky, T., and Srivastava, S. (2022), "Distributed Bayesian Inference in Massive Spatial Data," *UCSC Technical Reports*.

Guinness, J. (2018), "Permutation and grouping methods for sharpening Gaussian process approximations," *Technometrics*, 60, 415–429.

Han, Z. and De Oliveira, V. (2016), "On the correlation structure of Gaussian copula models for geostatistical count data," *Australian & New Zealand Journal of Statistics*, 58, 47–69.

Hassan, M. and El-Bassiouni, M. (2013), "Modelling Poisson marked point processes using bivariate mixture transition distributions," *Journal of Statistical Computation and Simulation*, 83, 1440–1452.

Hassan, M. Y. and Lii, K.-S. (2006), "Modeling marked point processes via bivariate mixture transition distribution models," *Journal of the American Statistical Association*, 101, 1241–1252.

Hautsch, N. (2011), *Econometrics of financial high-frequency data*, Springer Science & Business Media.

Hawkes, A. G. (1971a), "Point spectra of some mutually exciting point processes," *Journal of the Royal Statistical Society: Series B (Methodological)*, 33, 438–443.

— (1971b), "Spectra of some self-exciting and mutually exciting point processes," *Biometrika*, 58, 83–90.

Heiner, M. and Kottas, A. (2021), "Estimation and selection for high-order Markov chains with Bayesian mixture transition distribution models," *Journal of Computational and Graphical Statistics*, 1–13.

— (2022a), "Autoregressive density modeling with the Gaussian process mixture transition distribution," *Journal of Time Series Analysis*, 43, 157–177.

— (2022b), "Bayesian Nonparametric Density Autoregression with Lag Selection," *Bayesian Analysis*, 1, 1–29.

Heiner, M., Kottas, A., and Munch, S. (2019), "Structured priors for sparse probability vectors with application to model selection in Markov chains," *Statistics and Computing*, 29, 1077–1093.

Herrera, R. and Schipp, B. (2013), "Value at risk forecasts by extreme value models in a conditional duration framework," *Journal of Empirical Finance*, 23, 33–47.

Holgate, P. (1964), "Estimation for the bivariate Poisson distribution," *Biometrika*, 51, 241–287.

Hua, L. and Joe, H. (2014), "Strength of tail dependence based on conditional tail expectation," *Journal of Multivariate Analysis*, 123, 143–159.

Hughes, J. (2015), "copCAR: A flexible regression model for areal data," *Journal of Computational and Graphical Statistics*, 24, 733–755.

Jacobs, P. and Lewis, P. (1977), "A mixed autoregressive-moving average exponential sequence and point process (EARMA 1, 1)," *Advances in Applied Probability*, 9, 87–104.

Joe, H. (2014), *Dependence modeling with copulas*, Boca Raton, FL: CRC Press.

Johns, C. J., Nychka, D., Kittel, T. G. F., and Daly, C. (2003), "Infilling sparse records of spatial fields," *Journal of the American Statistical Association*, 98, 796–806.

Jordan, M. I. (2004), "Graphical models," *Statistical science*, 19, 140–155.

Kalliovirta, L., Meitz, M., and Saikkonen, P. (2015), "A Gaussian mixture autoregressive model for univariate time series," *Journal of Time Series Analysis*, 36, 247–266.

— (2016), "Gaussian mixture vector autoregression," *Journal of Econometrics*, 192, 485–498.

Katzfuss, M. and Guinness, J. (2021), "A general framework for Vecchia approximations of Gaussian processes," *Statistical Science*, 36, 124–141.

Kazianka, H. and Pilz, J. (2010), "Copula-based geostatistical modeling of continuous and discrete data including covariates," *Stochastic Environmental Research and Risk Assessment*, 24, 661–673.

Khalili, A., Chen, J., and Stephens, D. A. (2017), "Regularization and selection in Gaussian mixture of autoregressive models," *Canadian Journal of Statistics*, 45, 356–374.

Kim, H.-M. and Mallick, B. K. (2004), "A Bayesian prediction using the skew Gaussian distribution," *Journal of Statistical Planning and Inference*, 120, 85–101.

Kirsner, D. and Sansó, B. (2020), "Multi-scale shotgun stochastic search for large spatial datasets," *Computational Statistics & Data Analysis*, 146, 106931.

Kocherlakota, S. and Kocherlakota, K. (2006), "Bivariate discrete distributions," *Encyclopedia of Statistical Sciences*.

Korchevsky, V. and Petrov, V. (2010), "On the strong law of large numbers for sequences of dependent random variables," *Vestnik St. Petersburg University: Mathematics*, 43, 143–147.

Kotz, S., Kozubowski, T., and Podgorski, K. (2012), *The Laplace distribution and generalizations: a revisit with applications to communications, economics, engineering, and finance*, Springer Science & Business Media.

Krupskii, P., Huser, R., and Genton, M. G. (2018), "Factor copula models for replicated spatial data," *Journal of the American Statistical Association*, 113, 467–479.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. M. (2017), "Automatic differentiation variational inference," *Journal of machine learning research*.

Lancaster, T. (1990), *The econometric analysis of transition data*, Cambridge, UK: Cambridge University Press.

Lanne, M. and Saikkonen, P. (2003), "Modeling the US short-term interest rate by mixture autoregressive processes," *Journal of Financial Econometrics*, 1, 96–125.

Lau, J. W. and So, M. K. (2008), “Bayesian mixture of autoregressive models,” *Computational Statistics & Data Analysis*, 53, 38–60.

Lauritzen, S. L. (1996), *Graphical models*, Oxford, UK: Clarendon Press.

Le, N. D., Martin, R. D., and Raftery, A. E. (1996), “Modeling flat stretches, bursts outliers in time series using mixture transition distribution models,” *Journal of the American Statistical Association*, 91, 1504–1515.

Li, C.-S., Lu, J.-C., Park, J., Kim, K., Brinkley, P. A., and Peterson, J. P. (1999), “Multivariate zero-inflated Poisson models and their applications,” *Technometrics*, 41, 29–38.

Li, G., Zhu, Q., Liu, Z., and Li, W. K. (2017), “On mixture double autoregressive time series models,” *Journal of Business & Economic Statistics*, 35, 306–317.

Li, X., Genest, C., and Jalbert, J. (2021), “A self-exciting marked point process model for drought analysis,” *Environmetrics*, 32, e2697.

Liseo, B. and Loperfido, N. (2006), “A note on reference priors for the scalar skew-normal distribution,” *Journal of Statistical planning and inference*, 136, 373–389.

Liseo, B. and Parisi, A. (2013), “Bayesian inference for the multivariate skew-normal model: A population Monte Carlo approach,” *Computational Statistics & Data Analysis*, 63, 125–138.

Luo, J. and Qiu, H.-b. (2009), “Parameter estimation of the WMTD model,” *Applied Mathematics-A Journal of Chinese Universities*, 24, 379.

MacDonald, I. L. and Zucchini, W. (1997), *Hidden Markov and other models for discrete-valued time series*, volume 110, CRC Press.

Madsen, L. (2009), "Maximum likelihood estimation of regression parameters with spatially dependent discrete data," *Journal of Agricultural, Biological, and Environmental Statistics*, 14, 375–391.

Mahmoudian, B. (2017), "A skewed and heavy-tailed latent random field model for spatial extremes," *Journal of Computational and Graphical Statistics*, 26, 658–670.

McCullagh, P. and Nelder, J. A. (1983), *Generalized linear models*, London: Chapman and Hall.

Meitz, M., Preve, D., and Saikkonen, P. (2021), "A mixture autoregressive model based on Student's t–distribution," *Communications in Statistics-Theory and Methods*, 1–76.

Mena, R. H. and Walker, S. G. (2007), "Stationary mixture transition distribution (MTD) models via predictive distributions," *Journal of Statistical Planning and Inference*, 137, 3103–3112.

Modarres, M., Kaminskiy, M. P., and Krivtsov, V. (2017), *Reliability engineering and risk analysis: a practical guide*, Boca Raton: CRC Press.

Morris, S. A., Reich, B. J., Thibaud, E., and Cooley, D. (2017), "A space-time skew-t model for threshold exceedances," *Biometrics*, 73, 749–758.

Müller, P., Quintana, F. A., and Page, G. (2018), "Nonparametric Bayesian inference in applications," *Statistical Methods & Applications*, 27, 175–206.

Murphy, K. P. (2012), *Machine learning: a probabilistic perspective*, MIT press.

Nguyen, H. D., McLachlan, G. J., Ullmann, J. F., and Janke, A. L. (2016), "Laplace mixture autoregressive models," *Statistics & Probability Letters*, 110, 18–24.

North, G. R., Wang, J., and Genton, M. G. (2011), "Correlation models for temperature fields," *Journal of Climate*, 24, 5850–5862.

Ogata, Y. (1988), "Statistical models for earthquake occurrences and residual analysis for point processes," *Journal of the American Statistical association*, 83, 9–27.

O'hara, M. (1997), *Market microstructure theory*, Wiley.

Pacurar, M. (2008), "Autoregressive conditional duration models in finance: a survey of the theoretical and empirical literature," *Journal of Economic Surveys*, 22, 711–751.

Palacios, M. B. and Steel, M. F. J. (2006), "Non-Gaussian Bayesian geostatistical modeling," *Journal of the American Statistical Association*, 101, 604–618.

Panagiotelis, A., Czado, C., and Joe, H. (2012), "Pair copula constructions for multivariate discrete data," *Journal of the American Statistical Association*, 107, 1063–1072.

Panagiotelis, A., Czado, C., Joe, H., and Stöber, J. (2017), "Model selection for discrete regular vine copulas," *Computational Statistics & Data Analysis*, 106, 138–152.

Pardieck, K., Ziolkowski Jr, D., Lutmerding, M., Aponte, V., and Hudson, M. (2020), "North American Breeding Bird Survey Dataset 1966–2019: US Geological Survey data release," .

Paul, R. and Cressie, N. (2011), “Lognormal block kriging for contaminated soil,” *European journal of soil science*, 62, 337–345.

Peruzzi, M., Banerjee, S., and Finley, A. O. (2020), “Highly scalable Bayesian geostatistical modeling via meshed Gaussian processes on partitioned domains,” *Journal of the American Statistical Association*, 0, 1–14.

Peruzzi, M. and Dunson, D. B. (2022), “Spatial Multivariate Trees for Big Data Bayesian Regression.” *Journal of Machine Learning Research*, 23, 17–1.

Pisano, A., Marullo, S., Artale, V., Falcini, F., Yang, C., Leonelli, F. E., Santoleri, R., and Buongiorno Nardelli, B. (2020), “New evidence of Mediterranean climate change and variability from sea surface temperature observations,” *Remote Sensing*, 12, 132.

Pitt, M. K., Chatfield, C., and Walker, S. G. (2002), “Constructing first order stationary autoregressive models via latent processes,” *Scandinavian Journal of Statistics*, 29, 657–663.

Quiroz, M., Kohn, R., Villani, M., and Tran, M.-N. (2018), “Speeding up MCMC by efficient data subsampling,” *Journal of the American Statistical Association*.

Raftery, A. and Tavaré, S. (1994), “Estimation and modelling repeated patterns in high order Markov chains with the mixture transition distribution model,” *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 43, 179–199.

Raftery, A. E. (1985), “A model for high-order Markov chains,” *Journal of the Royal Statistical Society: Series B (Methodological)*, 47, 528–539.

— (1994), “Change point and change curve modeling in stochastic processes and spatial statistics,” *Journal of Applied Statistical Science*, 1, 403–423.

Ranganath, R., Gerrish, S., and Blei, D. (2014), "Black box variational inference," in *Artificial intelligence and statistics*, PMLR.

Recta, V., Haran, M., and Rosenberger, J. L. (2012), "A two-stage model for incidence and prevalence in point-level spatial count data," *Environmetrics*, 23, 162–174.

Reinhart, A. (2018), "A review of self-exciting spatio-temporal point processes and their applications," *Statistical Science*, 33, 299–318.

Ren, Q., Banerjee, S., Finley, A. O., and Hodges, J. S. (2011), "Variational Bayesian methods for spatial data analysis," *Computational statistics & data analysis*, 55, 3197–3217.

Schäfer, F., Katzfuss, M., and Owhadi, H. (2021), "Sparse Cholesky Factorization by Kullback–Leibler Minimization," *SIAM Journal on Scientific Computing*, 43, A2019–A2046.

Scheuerer, M. and Hamill, T. M. (2015), "Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities," *Monthly Weather Review*, 143, 1321–1334.

Sengupta, A. and Cressie, N. (2013), "Hierarchical statistical modeling of big spatial datasets using the exponential family of distributions," *Spatial Statistics*, 4, 14–44.

Sengupta, A., Cressie, N., Kahn, B. H., and Frey, R. (2016), "Predictive inference for big, spatial, non-Gaussian data: MODIS cloud data and its change-of-support," *Australian & New Zealand Journal of Statistics*, 58, 15–45.

Sethuraman, J. (1994), "A constructive definition of Dirichlet priors," *Statistica Sinica*, 4, 639–650.

Sklar, M. (1959), "Fonctions de repartition an dimensions et leurs marges," *Publications de l'Institut de Statistique de L'Université de Paris*, 8, 229–231.

Smith, M. S. and Khaled, M. A. (2012), "Estimation of copula models with discrete margins via Bayesian data augmentation," *Journal of the American Statistical Association*, 107, 290–303.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002), "Bayesian measures of model complexity and fit," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64, 583–639.

Stein, M. L., Chi, Z., and Welty, L. J. (2004), "Approximating likelihoods for large spatial data sets," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66, 275–296.

Stindl, T. and Chen, F. (2021), "Accelerating the estimation of renewal Hawkes self-exciting point processes," *Statistics and Computing*, 31, 1–17.

Stroud, J. R., Stein, M. L., and Lysen, S. (2017), "Bayesian and maximum likelihood estimation for Gaussian processes on an incomplete lattice," *Journal of computational and Graphical Statistics*, 26, 108–120.

Sun, Y. and Stein, M. L. (2016), "Statistically and computationally efficient estimating equations for large spatial datasets," *Journal of Computational and Graphical Statistics*, 25, 187–208.

Sun, Y., Stein, M. L., et al. (2015), "A stochastic space-time model for intermittent precipitation occurrences," *The Annals of Applied Statistics*, 9, 2110–2132.

Tang, X. and Li, L. (2021), "Multivariate temporal point process regression," *Journal of the American Statistical Association*, 1–16.

Vecchia, A. V. (1988), “Estimation and model identification for continuous spatial processes,” *Journal of the Royal Statistical Society: Series B (Methodological)*, 50, 297–312.

Veen, A. and Schoenberg, F. P. (2008), “Estimation of space–time branching process models in seismology using an EM–type algorithm,” *Journal of the American Statistical Association*, 103, 614–624.

Venter, G. G. et al. (2002), “Tails of copulas,” in *Proceedings of the Casualty Actuarial Society*, volume 89.

Vitolo, C. (2017), “hddtools: Hydrological data discovery tools,” *Journal of Open Source Software*, 2, 56.

Wallin, J. and Bolin, D. (2015), “Geostatistical modelling using non-Gaussian Matérn fields,” *Scandinavian Journal of Statistics*, 42, 872–890.

Wang, C. and Blei, D. M. (2013), “Variational Inference in Nonconjugate Models.” *Journal of Machine Learning Research*, 14.

West, M. and Harrison, J. (2006), *Bayesian forecasting and dynamic models*, Springer Science & Business Media.

West, M., Harrison, P. J., and Migon, H. S. (1985), “Dynamic generalized linear models and Bayesian forecasting,” *Journal of the American Statistical Association*, 80, 73–83.

Wheatley, S., Filimonov, V., and Sornette, D. (2016), “The Hawkes process with renewal immigration & its estimation with an EM algorithm,” *Computational Statistics & Data Analysis*, 94, 120–135.

Wikle, C. K. (2002), "Spatial modeling of count data: A case study in modelling breeding bird survey data on large spatial domains," *Spatial Cluster Modelling*, 199, 209.

Wold, H. (1948), "On stationary point processes and Markov chains," *Scandinavian Actuarial Journal*, 1948, 229–240.

Wong, C., Chan, W., and Kam, P. (2009), "A Student t-mixture autoregressive model with applications to heavy-tailed financial data," *Biometrika*, 96, 751–760.

Wong, C. S. and Li, W. K. (2000), "On a mixture autoregressive model," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62, 95–115.

— (2001a), "On a logistic mixture autoregressive model," *Biometrika*, 88, 833–846.

— (2001b), "On a mixture autoregressive conditional heteroscedastic model," *Journal of the American Statistical Association*, 96, 982–995.

Wu, G., Holan, S. H., Nilon, C. H., and Wikle, C. K. (2015), "Bayesian binomial mixture models for estimating abundance in ecological monitoring studies," *The Annals of Applied Statistics*, 9, 1–26.

Xu, G. and Genton, M. G. (2017), "Tukey g-and-h random fields," *Journal of the American Statistical Association*, 112, 1236–1249.

Yang, C., Delcher, C., Shenkman, E., and Ranka, S. (2018), "Clustering inter-arrival time of health care encounters for high utilizers," in *2018 IEEE 20th International Conference on e-Health Networking, Applications and Services (Healthcom)*, IEEE.

Yao, Y., Vehtari, A., Simpson, D., and Gelman, A. (2018), "Using stacking to average Bayesian predictive distributions (with discussion)," *Bayesian Analysis*, 13, 917–1007.

Zareifard, H., Khaledi, M. J., Rivaz, F., Vahidi-Asl, M. Q., et al. (2018), "Modeling skewed spatial data using a convolution of Gaussian and log-Gaussian processes," *Bayesian Analysis*, 13, 531–557.

Zhang, B., Cressie, N., et al. (2020), "Bayesian inference of spatio-temporal changes of Arctic Sea ice," *Bayesian Analysis*, 15, 605–631.

Zhang, H. (2002), "On estimation and prediction for spatial generalized linear mixed models," *Biometrics*, 58, 129–136.

Zhang, H. and El-Shaarawi, A. (2010), "On spatial skew-Gaussian processes and applications," *Environmetrics*, 21, 33–47.

Zhu, F., Li, Q., and Wang, D. (2010), "A mixture integer-valued ARCH model," *Journal of Statistical Planning and inference*, 140, 2025–2036.

Zilber, D. and Katzfuss, M. (2021), "Vecchia–Laplace approximations of generalized Gaussian processes for big non-Gaussian spatial data," *Computational Statistics & Data Analysis*, 153, 107081.