The Pennsylvania State University

The Graduate School

STATISTICAL METHODS FOR SPATIAL AND MULTIVARIATE SPATIAL EXTREME VALUES

A Dissertation in

Statistics

by

Mauricio Fernandes do Nascimento Junior

Submitted in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

August 2020

The dissertation of Mauricio Fernandes do Nascimento Junior was reviewed and approved by the following:

Benjamin A. Shaby

Assistant Professor of Statistics

Dissertation Advisor

Chair of Committee

Ephraim Hanks

Associate Professor of Statistics

Chair of Graduate Studies

Lynn Lin

Assistant Professor of Statistics

Chris E Forest

Professor of Climate Dynamics

Abstract

Chapter 1.
We analyze the joint tail of two variables related to fire threat associated with Santa Ana Winds in Southern California. To do this, we apply a flexible model for the joint tail of asymptotically dependent multivariate distributions, when samples are taken at several locations across space. We use a spatial prior on the underlying multivariate extremal dependence structure, which enables us to borrow strength across space while still allowing for different joint tail distributions at different spatial locations, and to predict the joint tail of the distribution at un-observed locations. A simulation shows that this model is able to capture complex dependence structures well.

Chapter 2
We introduce an approach to quickly and accurately approximate the cumulative distribution function of multivariate Gaussian distributions arising from spatial Gaussian processes. This approximation is trivially parallelizeable and simple to implement using standard software. We demonstrate its accuracy and computational efficiency in a series of simulation experiments, and apply it to analyzing the joint tail of a large precipitation dataset using a recently-proposed scale mixture model for spatial extremes. This dataset is many times larger than what was previously considered possible to fit using preferred inferential techniques.

Chapter 3
We analyze the the remaining life time of simulated engines from the NASA Commercial MModular Aero-Propulsion System Simulation. We apply a variety of regression methods, including models with change point detection, to accommodate features of the data. We demonstrate the accuracy of this method using MSE and compare it with competing methods from previous published papers. We conclude that the change point model has better results than some of the previous works and that performance increases when looking at smaller values of remaining life time.

Contents

List of Figures

List of Tables

Acknowledgments

This research was supported in part by NSF grant DMS-1752280.

Computations for this research were performed on the Pennsylvania State University's Institute for Computational and Data Sciences Advanced CyberInfrastructure (ICDS-ACI).

Disclaimer, the findings and conclusions of this thesis do not necessarily reflect the view of the NSF.

Chapter 1

Spatial Semi-parametric Spectral Density Estimation for Multivariate Extremes, with Application to Fire Threat

1.1 Introduction

We present a flexible model for the joint tail of asymptotically dependent multivariate distributions, when samples are taken at several locations across space. We use a spatial prior on the underlying multivariate extremal dependence structure, which enables us to borrow strength across space while still allowing for different joint tail distributions at different spatial locations, and to predict the joint tail of the distribution

at un-observed locations. We apply our multivariate spatial model to joint extremes of two variables that, taken together, are informative about extreme wildfire threat.

California suffers from fires that burn more than 172,000 acres of land annually. These fires translate into loss of habitat, infrastructure, and life. In 2015, for example, wildfires caused damage estimated to be more than $3 billion, with the biggest losses caused by fires started from downed electrical power lines (CalFire 2015). As a consequence, it is important to understand the conditions that allow wildfires to ignite and spread. The California Department of Forestry and Fire Protection (CalFire) is responsible, among other things, for responding to active fires and for forming risk mitigation strategies in the state. In order to efficiently allocate resources and formulate preventative policy proposals, CalFire needs to know where conditions conducive to fire ignition and spread are likely to occur. Of particular interest are the conditions that represent the most extreme threat.

Conditions that are conducive to wildfire ignition and spread are the subject of intense study and have drawn keen interest of regulators and infrastructure planners. This interest has reached new levels in the wake of the worst wildfire season in California history in 2018. Furthermore, the tails of the distribution of fire threat characteristics warrant particular attention because fire size and damage distributions have been shown to be sensitive to underlying tail assumptions (Bowman et al. 2009, Moritz et al. 2005). Here, we study the joint tail of two variables that represent the risk due to meteorological conditions of wildfire ignition and spread potential.

One phenomenon that plays a key role in wildfire threat in California is the Santa

Ana winds. Santa Ana winds are an atmospheric condition in Southern California that produces hot, dry, high-speed winds that originate in inland desert regions. This phenomenon is particularly active during the period of October through March. Because of these characteristics, and because they occur in a region of high development and population density, these winds are responsible for initiating particularly damaging wildfires (Westerling et al. 2004). We therefore focus our attention to the region of Southern California that experiences Santa Ana winds.

A common ignition mechanism in the Southern California is electrical arcing resulting from utility poles downed in high winds. We therefore take wind speed to be the key meteorological variable that serves as a proxy for fire ignition potential. The composite Fosberg Fire Weather Index (FFWI) combines relative humidity, temperature, and wind measurements into a single measure of fire weather that determines fire spread potential (Fosberg 1978). Moritz et al. (2010) reconstructed the weather at South California during the Santa Ana wind season, and found that high level of FFWI was associated with higher frequency of wild fires. We therefore study the joint tail of wind speed and FFWI, which represents the region in the space of meteorological variables where the probability of ignition and the potential for spread is concurrently extreme.

There have been previous approaches to modeling multivariate extreme observations. Coles & Tawn (1991) described 6 parametric distributions to model multivariate extremal dependence: the asymmetric and negative asymmetric logistic models, the Dirichlet model, the biologistic model, the nested logistic model, and the time series

logistic model. More recent work includes Cooley et al. (2010), who proposed the pairwise beta distribution, which defines a full joint distribution for multivariate extremes through their pairwise relationships, similar to a covariance matrix for Gaussian data. Vettori et al. (2017) used Bayesian model averaging with the nested logistic model defined in Coles & Tawn (1991) to probabilistically create clusters of exchangeable dependent variables, with dependence allowed to differ among clusters. They used reversible jump methods to average across the random number of clusters.

Several nonparametric approaches have been proposed as well, mostly limited to the bivariate case. Most focus on estimating the spectral measure, which we define in Section 1.2. Einmahl et al. (2001) estimated the bivariate spectral measure nonparmetrically using the ranks of the data, and proved consistency and asymptotic normality. Einmahl & Segers (2009) also estimated the bivariate spectral measure, this time using an empirical likelihood approach. de Carvalho et al. (2013) proposed a simplified version of the Einmahl & Segers (2009) bivariate estimator using Euclidean likelihood approach, and showed that it has the same asymptotic behavior. Guillotte et al. (2011) proposed a nonparametric Bayesian scheme to estimate the bivariate spectral measure. The infinite-dimensional prior on the spectral measure was shown to be dense in the space of valid spectral measures. Inference was performed using a trans-dimensional metropolis Hastings algorithm.

The problem is considerably more difficult in dimension greater than two. Marcon et al. (2017) used Bernstein polynomials to estimate the Pickands dependence function, which is an alternative way of characterizing an extreme-value dependence. This

approach in principle applies to any dimension, but is difficult to scale to more than two dimensions. Most closely related to the current work is Boldi & Davison (2007), who proposed a nonparametric mixture of Dirichlet distributions to model the dependence structure between arbitrarily many variables through the spectral distribution, after standardization of the margins to unit Fréchet. A critical moment restriction resulted in very poor mixing of the MCMC, however, and limited the practicality of the approach. Sabourin & Naveau (2014) fit an identical model using a re-parametrization of the Dirichlet mixture that avoids the awkward constraints and associated mixing problems.

An important aspect of our model is that it borrows strength across space to improve estimation of the multivariate tail dependence. However, we make no attempt to model the spatial dependence of the extreme events themselves. In this sense, our model inherits from Cooley et al. (2007), which considers a univariate response and places spatial Gaussian process priors on marginal generalized Pareto parameters. In contrast, a great deal of recent effort has been made to model the spatial dependence in the extreme events themselves (Davison et al. 2013). Models that attempt to simultaneous capture both multivariate dependence and spatial dependence of the extreme events include multivariate max-stable processes (Genton et al. 2015, Reich & Shaby 2018). These models are more realistic than our model in the sense that when limiting spatial dependence in the response variables is strong, multivariate max-stable processes can account for it. However, the dependence that they allow between variables is much more restrictive than our model.

This work will describe a model for multiple variables while taking into account the spatial dependence in the multivariate relationships. We describe the model in the general case of d variables, although we later restrict our attention to the special case of $d = 2$ for wind speed and FFWI. Section 1.2 describes the model originally proposed by Boldi & Davison (2007) and re-parameterized by Sabourin & Naveau (2014), and our extension from a single multivariate sample to a collection of multivariate samples observed at several spatial locations. Section 1.3 describes our simulation study to assess the performance of our model. Finally, Section 1.4 contains the joint analysis of wind speed and FFWI in southern California.

1.2 Model

Our model for the joint tail of wind and FFWI builds upon classical extreme value theory. We will define the model for d-dimensional random vectors, and proceed with the analysis for the special case of $d = 2$, which pertains to our application. Let $\mathbf{Y}$ be a random vector of observations in $\mathbb{R}^d$ with joint distribution function $\mathbf{F}$ and marginal distribution F_i for $1 \leq i \leq d$. Define a new variable of transformed observations $\mathbf{X}$ as

$$\mathbf{X} = (-1/F_1(Y_1), ..., -1/F_d(Y_d))$$

so that each X_i has a unit Fréchet marginal distribution, i.e. $P(X_i \leq x) = e^{-1/x}$. This transformation allows all variables to have a common marginal distribution that is

convenient for defining the dependence structure among them in the far joint tail. To model the dependence in the far tail of a random vector whose margins are unit Fréchet, it is useful make a further transformation from the original coordinates to pseudo-polar coordinates. First, define the radial component to be

$$R = \sum_{i=1}^{d} X_i \tag{1.1}$$

and the angular component to be

$$\mathbf{W} = \frac{\mathbf{X}}{R}, \tag{1.2}$$

where $\mathbf{W} \in \mathbb{S}_d$ and $\mathbb{S}_d$ is the unit simplex defined by $\left\{\mathbf{w} : w_i \geq 0, \sum_{i=1}^{d} x_i = 1\right\}$. This transformation us useful because, as long as $\mathbf{F}$ is in the multivariate maximum domain of attraction of a max-stable random vector, $\mathbf{F}$ can be expressed in terms of its angular and radial components, which are independent in the limit (Resnick 1987). That is, for a large radial threshold r_0,

$$P(R > r, \mathbf{W} \in A \mid R > r_0) = \frac{r}{r_0} H(A)$$

as $r_0 \to \infty$, for $r > r_0$, $A \in \mathbb{S}_d$, and H a probability distribution usually referred to as the *angular distribution* or the *angular measure*. This decomposition makes it clear that the dependence structure of the random vector is completely described by the angular probability measure H, independent of the radial component. As a consequence, to

define the model for dependence in the joint tail, we need only to consider the angular measure H, independently of the distribution of the radial component.

A probability measure H on $\mathbb{S}_d$ is a valid angular probability measure if and only if it satisfies the moment condition

$$\int_{\mathbb{S}_d} w_i dH(\mathbf{w}) = \frac{1}{d} \qquad \text{for all } i = 1, \ldots, d. \tag{1.3}$$

Several parametric families have been proposed that can be used to model H (Cooley et al. 2012, e.g.), but in general, no parametric family includes the entire space of valid angular measures because, as long as the constraint (1.3) is satisfied, H is otherwise un-restricted.

Tail dependence between a pair of variables is often succinctly summarized using the quantity known as χ, defined as a limit as the quantile q approaches 1 as

$$\chi = \lim_{q \to 1} P\left(Y_i > F_i^{-1}(q) \mid Y_j > F_j^{-1}(q)\right).$$

When the tail dependence parameter χ has a limiting value greater than zero, we say the two variables are asymptotically dependent. The interpretation of asymptotic dependence is that regardless of how far in the tails of the distribution components of the random vector are, the components remain dependent on one another. When the probability mass of the angular distribution H is concentrated in the interior of the simplex, the result is asymptotic dependence among the variables (Resnick 1987). Conversely, when the limiting value of χ is zero, we say the two variables are

asymptotically independent, with the interpretation that if one goes far enough in the tails of the distribution, components become independent of one another. If any probability mass of H lies on the vertices or edges of the simplex, the corresponding variables are asymptotically independent.

In the context of wind and FFWI, asymptotic dependence would imply that far in the tail of he distribution, extreme values wind speed and extreme FFWI could occur simultaneously. In contrast, assuming asymptotic independence would mean that in the far in the tail, extreme values of wind speed and extreme values of FFWI would never occur concurrently. For the scope of this thesis we will assume asymptotic dependence between wind speed and FFWI. This assumption is natural because wind speed is one of the ingredients of FFWI, so we would expect that extreme values of wind speed would be strongly associated with extreme values of FFWI. We therefore model the dependence in the joint tail by considering angular measures that have mass concentrated in the interior of the simplex.

1.2.1 Semi-parametric representation of multivariate dependence

Since multivariate dependence in the far tail is entirely described by a distribution H on the unit simplex, we focus our efforts of specifying a flexible representation for H that can be extended to the multivariate spatial context. The best-known distribution with support on the unit simplex is the Dirichlet distribution. A sensible approach then to constructing flexible models for angular distributions H would be to use Dirichlet distributions as building blocks. Boldi & Davison (2007) proposed a semi-parametric

model for H using a mixture of Dirichlet distributions. The density of this mixture model is defined as

$$h(\boldsymbol{w}) = \sum_{k=1}^{K} p_k \frac{\Gamma(\nu_m)}{\prod_{j=1}^{J} \Gamma(\nu_m \mu_j^{(m)})} \prod_{j=1}^{J} w_j^{\nu\mu_j - 1}, \tag{1.4}$$

where p_k is the mixture weight of the kth component, with $p_k \geq 0, \sum_{k=1}^{K} p_k = 1$, $\boldsymbol{\mu}_k \in \mathbb{S}_d$ is the location parameter of of the kth component, and $\nu_k \in \mathbb{R}^+$ is the concentration parameter of the kth component. This model is attractive because it is very flexible compared to other parametric models. The modeler can choose the number of components K to be as large as needed estimate the angular density $h(\boldsymbol{w})$ well, including shapes of $h(\boldsymbol{w})$ that are not symmetric and have otherwise unusual shapes. In fact, as $K \to \infty$, this family of distributions is dense in the space of distributions on the interior of the simplex, which in a sense makes this model as flexible as one would ever want. The difficulty with this model arises from the moment restriction (1.3). This restriction is enforced by requiring that

$$\sum_{i=1}^{k} p_i \mu_i = (1/d, ..., 1/d). \tag{1.5}$$

This constraint on the parameters forces the model to satisfy (1.3), but because it is not a hyper-rectangle in the parameter space, it makes estimation difficult. Specifically, when estimating the parameters using Markov Chain Monte Carlo (MCMC) methods, the constraint (1.5) makes the components of the chain highly dependent, leading to

catastrophically poor mixing (Boldi & Davison 2007).

1.2.1.1 Re-parametrization of the mixture model

To mitigate the difficulties caused by the constraint (1.5), Sabourin & Naveau (2014) re-parametrized the mixture model (1.4) in a way that replaces the dependent constraint (1.5) with independent box constraints on a transformed set of parameters. The transformation is made from the original parameters to the new parameters such that $T(p_1, \ldots, p_k, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K) \Longrightarrow (\epsilon_1, \ldots, \epsilon_{K-1}, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{K-1})$, where $\epsilon_m \in (0, 1), m = 1, \ldots, K-1$ (See Appendix for details). This transformation replaces the original mixture weights $p_1, \ldots, p_K$ and the Kth location parameter $\boldsymbol{\mu}_K$ and replaces them with a collection of $K-1$ "eccentricity" parameters $\epsilon_1, \ldots, \epsilon_{K-1}$. The last location parameter $\boldsymbol{\mu}_K$ may be recovered as a deterministic function of the first $K-1$ locations and the eccentricities, with the reduced degree of freedom taking the place of the awkward constraint (1.5). The key is that under the Sabourin & Naveau (2014) parameterization, the moment constraint (1.3) is satisfied even though all parameters ϵ_i and $\boldsymbol{\mu}_i, 1 \leq i \leq K-1$, are independent of each other in the prior. Hence, the parameter space becomes a hyper-rectangle, and the complicated cross dependencies that hindered MCMC mixing in the original parameterization are completely removed.

1.2.2 Spatial prior

Our main goal is to use spatial information to both improve estimation of the flexible multivariate tail dependence model and be able to predict the dependence structure at

locations where no observations are sampled. Since the tail dependence is completely described by the mixture defined by equation (1.4), allowing the (transformed) parameters in (1.4) to vary smoothly in space is equivalent to allowing the dependence structure to vary smoothly in space. We therefore allow each parameter in the mixture to follow a smooth process so that locations close together will have similar values, while locations farther apart will have values that are independent from each other.

To induce this spatial smoothing in the underlying dependence parameters, we assign them Gaussian process priors according to

$$\begin{aligned} \log(\nu_k(s)) &\sim \text{GP}(\boldsymbol{X}(s)\boldsymbol{\beta}_{\nu_k}, C_{\nu_k}) \quad \text{for } k \in 1, ..., K \\ \text{logit}(\epsilon_k(s)) &\sim \text{GP}(\boldsymbol{X}(s)\boldsymbol{\beta}_{\epsilon_k}, C_{\epsilon_k}) \quad \text{for } k \in 1, ..., K-1 \\ \boldsymbol{\mu}_k &\sim \text{Dirichlet}(\boldsymbol{\mu}_0, \nu_0) \quad \text{for } k \in 1, ..., K-1, \end{aligned} \tag{1.6}$$

where the notation $\text{GP}(m(s), C)$ refers a Gaussian process with mean function $m(s)$ and covariance function C, and $\boldsymbol{X}(s)$ is a collection of spatially-varying covariates. The prior specification in (1.6) also includes link functions necessary to transform the support of the parameters to the real line. For example, each concentration parameter $\nu(s_0)$ at a location s_0 is a value in $\mathbb{R}^+$, so taking its logarithm allows us to assign it a Gaussian process prior. Similarly, we transform each eccentricity parameter $\epsilon(s_0)$ at location s_0 from $(0, 1)$ to the real line using the logit function. Finally, the first $K-1$ Dirichlet location parameters $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{K-1}$ are shared across spatial locations to avoid over-parameterization (see Lock & Dunson 2015), and have independent Dirichlet priors. Because of the lost degree of freedom from the Sabourin & Naveau (2014)

parameterization, the remaining Dirichlet location parameter $\boldsymbol{\mu}_K(s)$ is a function of the other location parameters and the spatially-varying eccentricities, so it is itself spatially varying, even though it is not assigned a prior distribution.

For simplicity, all covariance functions are assumed to be stationary and isotropic, with $C_{\nu_k}(s_i, s_j) = \sigma^2_{\nu_k} \exp\{-\|s_i - s_j\|/\rho_{\nu_k}\}$ independently for $k \in 1, ..., K$ for the log concentration parameters and $C_{\epsilon_k}(s_i, s_j) = \sigma^2_{\epsilon_k} \exp\{-\|s_i - s_j\|/\rho_{\epsilon_k}\}$ independently for $k \in 1, ..., K-1$ for the logit eccentricity parameters. To complete the model, the scale and range parameters $\sigma_{\nu_1}, \ldots, \sigma_{\nu_K}, \rho_{\nu_1}, \ldots, \rho_{\nu_K}, \sigma_{\epsilon_1}, \ldots, \sigma_{\epsilon_{K-1}}$, and $\rho_{\epsilon_1}, \ldots, \rho_{\epsilon_{K-1}}$ are all assigned vague zero-centered positive half normal priors, and the regression coefficients $\boldsymbol{\beta}_{\nu_1}, \ldots, \boldsymbol{\beta}_{\nu_K}$ and $\boldsymbol{\beta}_{\epsilon_1}, \ldots, \boldsymbol{\beta}_{\epsilon_{K-1}}$ are assigned vague zero-centered normal priors.

1.2.3 MCMC

We fit the model using Markov Chain Monte Carlo (MCMC), for a fixed number of mixture components K, and then use model diagnostics to choose the value of K (see Section 1.3). All parameter updates are performed using Metropolis-Hastings (MH), except the regression coefficients, which have closed-form full conditional distributions. All MH proposals are random walk normal proposals, except for $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{K-1}$, which live on the unit simplex and get Dirichlet proposals.

1.3 Simulation

We performed a simulation study to assess the performance of the model. We created a grid of 20 spatial locations on the unit square and sampled 100 replications from our model at those locations. We constructed $h(\boldsymbol{w})$ according to (1.4) using three mixture components. The spatial range parameters $\rho_{\nu_1}, \rho_{\nu_2}, \rho_{\nu_3}, \rho_{\epsilon_1}$ and ρ_{ϵ_2} were all set to 0.2, which results in moderate spatial correlation for the Dirichlet parameters across the spatial domain. The spatially varying covariate matrix $\boldsymbol{X}(s)$ was a column vector of ones for simplicity, specifying an intercept-only model. The regression coefficients $\boldsymbol{\beta}$ corresponding to the concentration parameters were set to 3, and those corresponding to the eccentricity parameters were set to -1.791 for the first component and 0.619 for the second component. The first 2 Dirichlet location parameters $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ were set to $(0.4, 0.6)$ and $(0.2, 0.8)$, respectively.

We fit the model four times, with the number of components ranging from 2 to 5. For comparison with a parametric model for the angular distribution, we also fitted a logistic model (Coles 2001) to the data, using a spatial Gaussian process prior for the logistic dependence parameter α. Since mixture models are susceptible to label switching, convergence diagnostics based on MCMC parameter trace plots are ineffective. Instead, we monitored convergence by plotting the posterior log likelihood of the data. To choose the number of components, we used 5-fold cross validation, using the log likelihood as the criterion. To do this, we randomly divided the 20 locations into 5 groups of 4 and ran the model 5 times, each time using 16 locations for

model fitting and the remaining 4 locations for evaluation. The average log likelihood across the 5 test sets, for the varying number of model components as well as the logistic model, are shown in table 1.1.

Table 1.1: Cross validation results for the dependence model with 2 to 5 components, as well as the parametric logistic model. The cross validation procedure chose 2 mixture components, although the data was simulated using 3 components.

# Components	CV log likelihood
Logistic	-105.18
2	52.53
3	-92.49
4	-181.77
5	-176.29

Even though the data was simulated using 3 mixture components, the cross validation procedure found that the model with 2 components to be the best fit. Figure 1.1 shows the true simulated angular density $h(\boldsymbol{w})$ in black at four arbitrarily-selected locations, along with the predicted posterior angular density using the best-fitting spatial Dirichlet mixture with 2 components (blue curve) and corresponding pointwise 95% credible region (blue region). The fitted spatial logistic model and corresponding pointwise 95% credible region is shown in red. It is evident from Figure 1.1 that the simple parametric model is not capable of accurately representing the somewhat complex tail dependence in the simulated data, as evidenced by the credible region failing to include the true curve over much of the domain. In contrast, the mixture model with 2 components does a qualitatively good job capturing the important features, even though the true density was simulated with three components. From Figure 1.1 then, it is not surprising that cross validation selected 2 components rather than 3.

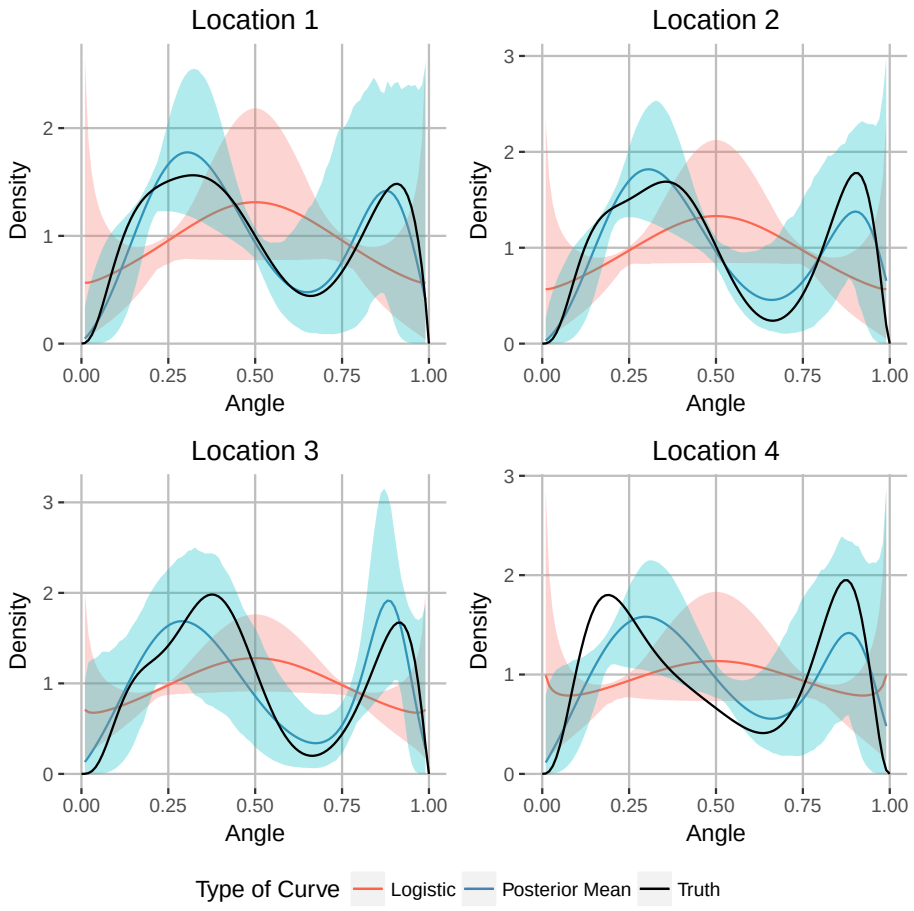


Figure 1.1: Predicted and true simulated angular densities for 4 arbitrarily-chosen spatial locations. The solid black curve is the true angular density $h(\boldsymbol{w})$. The solid red curve is the posterior mean angular density estimated using the logistic model, and the red band is the corresponding pointwise 95% credible region. The blue curve is the posterior mean angular density estimated using the mixture of Dirichlets from (1.4), and the blue band is the corresponding pointwise 95% credible region.

1.4 Analysis of concurrent extreme wind speed and FFWI

The data we used for this analysis consists of daily maximum wind speeds, along with other variables needed to construct the Fosberg Fire Weather Index, from 20 weather

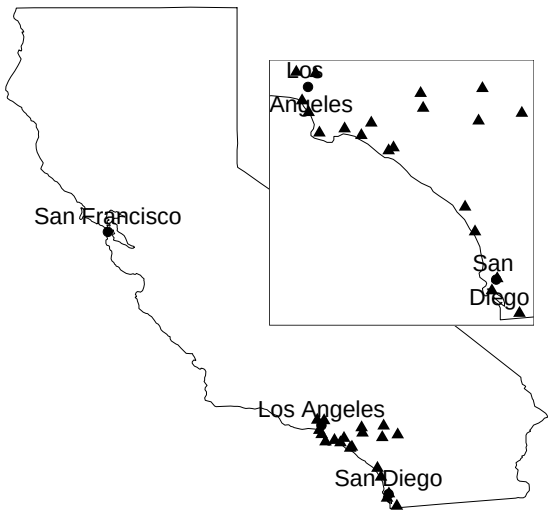


Figure 1.2: Location of weather stations in California.

stations in a region of Southern California that is susceptible to the Santa Ana Winds. The data was downloaded from the Hadley Center website (available at `http://www.metoffice.gov.uk/hadobs/hadisd/index.html`). We extracted the data from October through March, when the Santa Ana Winds are most active, from 1973–2015, and calculated FFWI using the formula in Fosberg (1978). Figure 1.2 shows the locations of the weather stations we considered, which lie inside a 22,500 km^2 area in Southern California which includes large population centers like Los Angeles and San Diego. This location is of particular interest because because of its vulnerability to large wildfires and its proximity to so much at-risk population and infrastructure.

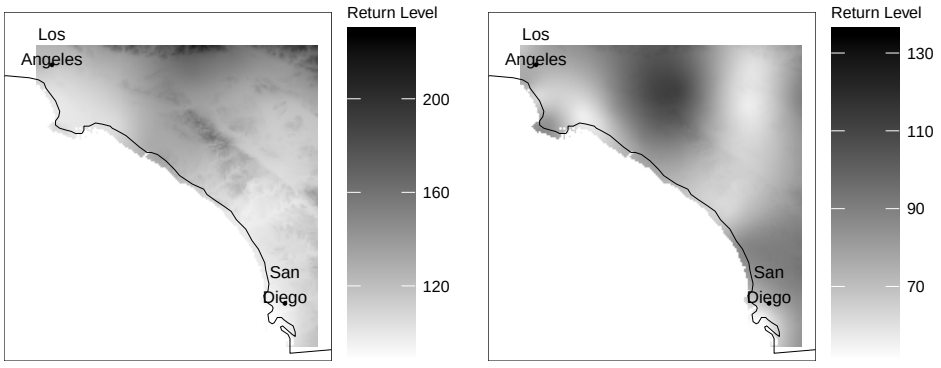
1.4.1 Marginal model

Our tail dependence model assumes unit Fréchet marginal distributions for exceedances of high thresholds, for all components. We therefore have to choose thresholds and

transform wind speed and FFWI to unit Fréchet. The standard technique for transforming to unit Fréchet is to use a nonparametric rank transformation. We did not do this for two reasons. First, we wanted to use spatial information to borrow strength in estimating the marginal distributions. Second, applying a rank transformation to data at observation locations would have left us no way to predict at un-observed locations. For these reasons, we used a spatial model to estimate the marginal distributions. To do this, we fit spatial generalized extreme value (GEV) distribution models separately to annual maxima of the wind speed and FFWI variables. This model uses Gaussian process priors for GEV location and (log) scale parameters, and assumes conditionally independent responses, similar to the model in Cooley et al. (2007). We included latitude, longitude and elevation as covariates in the marginal models for both wind speed and FFWI. We fit this model using the implementation in the R package SpatialExtremes (Ribatet 2018). Figure 1.3 shows the 50 and 100-year return level surface calculated from the fitted models for the wind speed (Figure 1.3 (b) and 1.3 (d)) and FFWI (Figure 1.3 (a) and 1.3 (c)) marginal components.

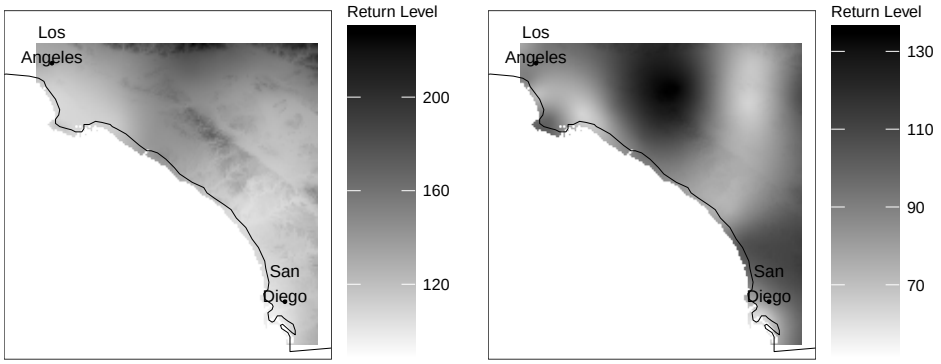
Return values of FFWI are smaller near the coast compared to values in the northeastern portion of the study region. In contrast, areas along the coast tend to have the strongest extreme wind, with the northeastern portion tending to have less extreme wind speeds. Over almost the entire region, the 100-year return level is greater than 100, which is sometimes considered the largest feasible value of FFWI.

We selected days when both FFWI and wind speed exceeded their marginal 90% empirical quantiles (35.15 and 18.34mph for FFWI and wind speed, respectively),



(a) FFWI, 50 year return level.

(b) Wind speed (miles/hour), 50 year return level.



(c) FFWI, 100 year return level.

(d) Wind speed (miles/hour), 100 year return level.

Figure 1.3: Posterior mean 50 and 100 year return levels for FFWI and wind speed components.

where the quantiles were computed by pooling data across all observed locations. Out of the original 127,802 and 143,162 daily observations from FFWI and wind speed, respectively, 10,484 and 12,027 exceeded their respective marginal thresholds, and 6,228 did so simultaneously. Setting the threshold at 90% empirical quantiles was a compromise between being far enough in the tail for the asymptotic model to be a

good approximation on one hand, and on the other hand having enough data to fit the dependence model.

Based on the posterior sample of the marginal GEV parameter fields obtained from the spatial GEV model, we marginally transformed the exceedances to unit Fréchet using the posterior mean GEV parameters at the observation locations.

1.4.2 Dependence model

After transforming the exceedances to unit Fréchet, we further transformed them into pseudo-polar coordinates using equations (1.1) and (1.2). We then fit our spatial multivariate extremal dependence model to the derived angles.

In order to choose the number of components to include, we used cross validation, randomly partitioning the data into 5 folds by location, just as in the simulation in Section 1.3. This scheme of cross validation selects for the ability to predict the joint tail at un-observed locations, preventing over-fitting at any individual location, which helps to regularize the estimation of the flexible tail model. Table 1.2 shows the cross validation results for the FFWI and wind speed data. The cross validation suggests that the 4-component model provides the best fit. Therefore, we proceed with this analysis using 4 mixture components. For comparison, we also evaluated a spatial logistic model the same cross validation scheme. As Table 1.2 shows, the logistic model is not competitive with any of the mixture models, and is unable to fit the dependence well.

Proceeding with a 4-component mixture, we re-fit the model using the complete

Table 1.2: Cross validation results for the extremal dependence between FFWI and wind speed, for a spatial logistic model and spatial mixture models with 2 to 5 mixture components. The righ-hand column shows predictive log likelihoods when all replications at 80% of the locations are used for fitting and the remaining 20% of the locations are used for testing. Larger log likelihoods are preferred.

# Components	CV log likelihood
Logistic	-6195.596
2	-6075.282
3	-5537.285
4	-4794.236
5	-5075.750

data. Figure 1.4 shows the fitted angular densities for 4 locations in the study region. Solid lines represent pointwise posterior means, and shaded regions represent pointwise 95% credible regions. Observed angles are shown as ticks on the x-axes. These four locations show considerable heterogeneity in their joint tail characteristics between FFWI and wind speed. At all observed locations, the tick marks coincide with high posterior density, which suggests that the model is doing a good job capturing the dependence, and that imposing the restriction that the Dirichlet location parameters must be shared across space was not too restrictive. In each angular density plot, the angular mass is concentrated far in the interior of the simplex, suggesting that the two variables are indeed asymptotically dependent in this geographical region.

In addition to the spatial prior enabling borrowing strength across locations, it also enables spatial prediction of the joint tail at un-observed locations. Figure 1.5 (a) shows the predicted angular density at one such un-observed location, which was burned by the disastrous Lilac Fire in 2017. At this particular location, the model predicts strong dependence between extreme FFWI and wind speed, with the highest

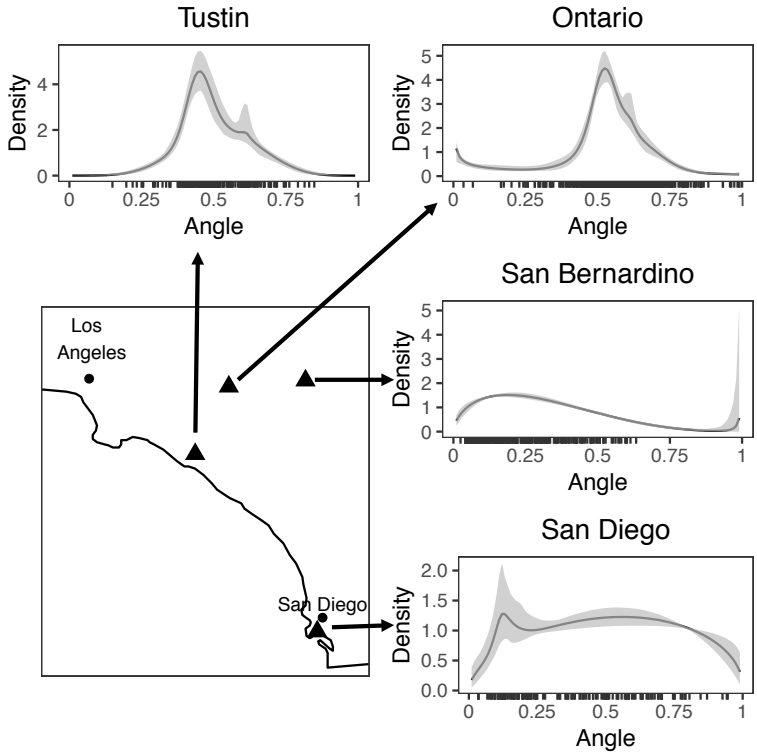
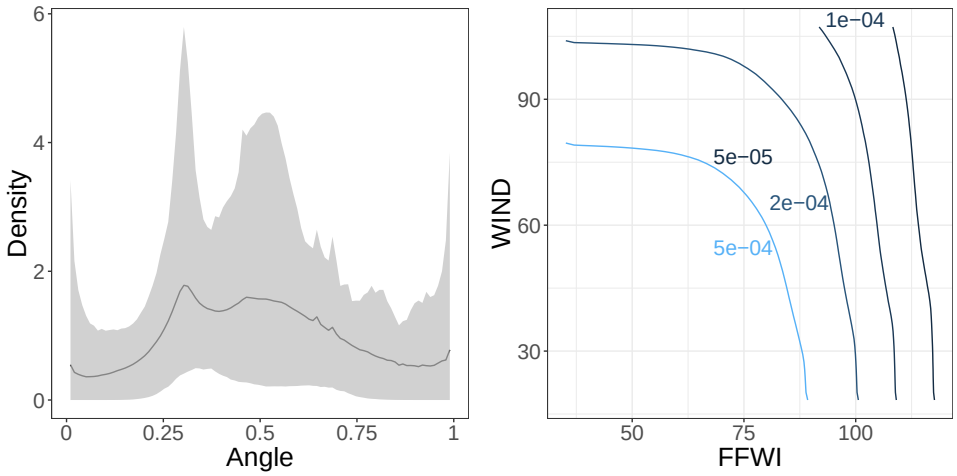


Figure 1.4: Estimated posterior densities for 4 locations across Southern California. The black curves are the pointwise posterior mean densities, the gray bands are the 95% pointwise credible bands, and the ticks on the x-axes are the observed angles.

posterior density at around 0.3. This result indicates that joint extremes of FFWI and wind speed are likely to occur simultaneously, exacerbating fire threat due to each variable being individually extreme.

To make this result more interpretable, we back-transformed the two variables from pseudo-polar coordinates to the unit Fréchet scale and then to the scale of the original data, and calculated joint exceedance probabilities implied by the angular density shown in Figure 1.5 (a). The tail of the joint survivor function of FFWI and wind speed predicted at this location is shown in Figure 1.5 (b). Each contour represents a set of constant joint exceedance probability. That is, at every (y_1, y_2) point along the



(a) Predicted angular density

(b) Estimated joint survivor function of FFWI and wind speed

Figure 1.5: Panel (a) shows the predicted angular density at an un-observed location which was near the center of the Lilac Fire in 2017. The solid line represents the pointwise posterior mean, and the shaded region represents the pointwise 95% credible region. Panel (b) shows the estimated joint survivor function of FFWI and wind speed, calculated by transforming pseudo-polar coordinates with the angular density in panel (a) first to unit Fréchet vectors and then to the original scale of the data. Each contour represents a set of constant joint exceedance probability. For example, at every (y_1, y_2) point along the contour labeled $5\text{e}{-04}$, the joint probability of FFWI exceeding y_1 and wind speed exceeding y_2 is $5\text{e}{-04}$.

contour of constant probability p, the joint probability of FFWI exceeding y_1 and wind speed exceeding y_2 is p. The joint survivor function might be of direct use for risk management. For example, if safety standards require consideration of events that have probability 0.00005 (i.e. events which occur on average once every 100 years, since our study period is about 200 days in each year), these events in the joint space of FFWI and wind speed are described by the contour in Figure 1.5 (b) labeled $5\text{e}{-05}$.

1.5 Conclusion

We analyzed the joint tail of two variables related to fire threat associated with Santa Ana Winds in Southern California. We wanted a flexible model that could accommodate the complicated shape of the angular distribution that determines extremal dependence between the two variables. We used spatial priors to regularize estimation of the flexible angular distribution model, which also enabled us to predict the joint tail at un-observed locations. To do this, we used a mixture of Dirichlet distributions, combined with the new parametrization proposed by Sabourin & Naveau (2014). The re-parametrization enabled the mixture parameters to be independent in the prior for each location, which enabled specification of Gaussian process priors to borrow strength between locations. Our simulation analysis showed that the model is capable of recovering complicated shapes of tail dependence, and 5-fold fold cross validation did a reasonable job of choosing a good number of mixture components. Although we restricted our attention to the bivariate case due to the structure of the fire threat application, the spatial mixture of Dirichlet distributions is straightforwardly generalizable to higher dimensions.

The joint analysis of extreme Fosberg Fire Weather Index and wind speed was able to capture the tail dependence between the two variables, as well as the spatial correlation in the angular densities across locations. One possible drawback to our modeling approach is that it is unable to capture spatial tail dependence in the extreme events themselves (as opposed to spatial dependence in the underlying joint tail

distribution), so that if one wished to evaluate areal joint exceedance probabilities, that would be impossible. For example, if practitioners needed to know the probability that an area of at least size A would simultaneously experience a joint exceedance of X FFWI and Y mph wind, they would not be able to do that with our model.

A second drawback of our approach is that we estimated marginal surfaces and transformed marginally to unit Fréchet as a pre-processing step. Done this way, the uncertainty in the marginal transformation is not propagated to the final analysis. It might be possible to do both simultaneously, as Sabourin (2015) did in the non-spatial context. We made a sustained attempt to adapt the Sabourin (2015) model to the spatial context so that we could properly account for uncertainty in the marginal estimation, but we were unable to make it work.

One limitation of our result is that FFWI only takes into account wind speed, temperature, and humidity, and it assumes that the fuels are extremely fine with high moisture of extinction, a condition most suited to grasslands. In particular, FFWI does not take into account precipitation. This means that the index assumes a constant fuel moisture and equilibrium moisture content. This is a known limitation for operational applications. A modified version of FFWI was proposed by Goodrick (2002) to include information about precipitation and fuel availability, and could be used instead of the original FFWI in a similar analysis to the one presented here.

Chapter 2

A Vecchia Approximation for High-Dimensional Gaussian Cumulative Distribution Functions Arising from Spatial Data

2.1 Introduction

We introduce a trivially parallelizable approach to quickly and accurately approximate the cumulative distribution function (*cdf*) of multivariate Gaussian distributions with highly structured covariance matrices, such as those arising from spatial Gaussian processes. The multivariate Gaussian distribution is by far the most widely used for modeling multivariate and spatial data. To a large degree, its near universal adoption is

the result of its simplicity; it is concisely and intuitively parametrized by a mean vector and pairwise dependence in the form of a covariance matrix. Prominent examples of its use include time series models like autoregressive and moving average models, which consider the joint distribution of the observations observed at discrete time points to be multivariate Gaussian, as well as geostatistics models, which consider spatially-indexed observations to be realizations of a Gaussian process, usually with a parsimoniously parametrized covariance structure. Even multivariate models that do not assume Gaussian responses often represent dependence using some kind of latent multivariate Gaussian distribution.

In most situations, likelihood-based inference on popular models just requires calculation of the joint density of all observations. The probability density function (*pdf*) for a multivariate Gaussian random variable is

$$f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \equiv \phi_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\mathbf{x}) = (2\pi)^{-k/2} \det(\boldsymbol{\Sigma})^{-1/2} \exp\Big[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\Big], \qquad (2.1)$$

where $\boldsymbol{\mu}$ is the mean vector of length D and $\boldsymbol{\Sigma}$ is the $D \times D$ covariance matrix.

In principle, there is nothing difficult about calculating this density; it simply requires commonplace operations like calculating an exponent, matrix determinant, matrix multiplication, and matrix inversion. However this is not an easy task in practice when the dimension D is large. The complexity of calculating the determinant and inverse of a $D \times D$ matrix is typically $O(D^3)$ for algorithms in common use. This means that for large values of D, the calculation of the *pdf* becomes prohibitive.

Computing the Gaussian *cdf*, which is a much more difficult problem, has received much less attention. The problem has increased in prominence recently with advances in spatial modeling of extreme events. State-of-the-art approaches for spatial extremes like Wadsworth & Tawn (2014), Thibaud et al. (2016), de Fondeville & Davison (2018), and Huser & Wadsworth (2019) all require high-dimensional Gaussian *cdf*s for inference. This turns out to be the dominant computational bottleneck, and all but de Fondeville & Davison (2018) restricted their analyses to fewer than 20 spatial locations because larger datasets are computationally intractable using widely-used techniques for computing the Gaussian *cdf*. In real-world spatial applications, one should expect to see many more spatial locations, and existing approaches are not equipped to handle datasets of even moderate size.

Multivariate Gaussian *cdf*s appear in other contexts as well; for example the density of multivariate skewed Gaussian and t random variables are functions of the multivariate Gaussian *cdf*(Arellano-Valle & Azzalini 2006). Here, we will focus on the case of spatial extremes. To make things concrete, we will use the example of the Gaussian scale mixture model from Huser et al. (2017), although our computational strategy would work equally well in any context with highly-structured covariance matrices.

Most approaches to calculating multivariate Gaussian probabilities are intended for problems of small or moderate dimension. Genz (1992) proposed a transformation from the original integral over $\mathbb{R}^D$ to an integral over a unit hypercube. Transforming to a finite region then allows the use any standard numerical integration method.

Genz (2004) derived formulas to calculate bivariate and trivariate Gaussian *cdf*s with high precision using Gauss-Legendre numerical integration. The calculations are fast and precise but do not apply in higher dimensions. Miwa et al. (2003) proposed a two-stage recursive approach to estimate the Gaussian *cdf*. Their approach does not scale to high dimensions because it requires a sum over a combinatorially exploding (in D) number of terms.

The most popular approach for approximating Gaussian *cdf*s in moderate dimensions was proposed by Genz & Bretz (2009). They describe the use of Monte Carlo (MC) and quasi-Monte Carlo (QM) methods to estimate the joint *cdf*. Their QM methods have smaller asymptotic errors than the MC versions, and hence are the more widely used.

More recently, Genton et al. (2018) sped up the Genz & Bretz (2009) QM algorithm by performing matrix computations with fast hierarchical matrix libraries (Hackbusch 2015). As a follow-up, Cao et al. (2019) combined hierarchical matrix computations with a blocking technique to further speed up computations. These approaches are much faster than their predecessors and work for Gaussian random variables with arbitrary covariance structures. They lean heavily on linking to specialized libraries for matrix operations. Our approach achieves speedups using a fundamentally different strategy, by specifically leveraging the properties of highly-structured covariance forms arising from, for example, time series or spatial data. It requires no exotic software, and is trivially parallelizable using simple tools in `R`.

2.2 A Vecchia Approximation for the Multivariate Gaussian Distribution Function

The multivariate Gaussian *cdf* that we wish to calculate is simply the integral of the *pdf* (2.1),

$$P(\mathbf{X} < \mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \equiv \Phi_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\mathbf{x}) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_k} \phi_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\mathbf{y})\, dy_1 \ldots dy_D. \tag{2.2}$$

To calculate the integral (2.2), one must resort to numerical techniques, as it is well-known that no closed form exists, even in a single dimension. In high dimensions, numerical integration is very difficult simply due to geometry and the curse of dimensionality. The difficulty is compounded in the case of the Gaussian *cdf* because while the curse of dimensionality requires an exponentially (in D) increasing number of evaluations of the integrand, the cost of each evaluation of the integration itself grows as D^3. We seek a technique that simultaneously 1) reduces the effective dimension of the integral and 2) reduces the dimension of the *pdf* in the integrand.

2.2.1 Vecchia Approximation for the Gaussian *pdf*

Vecchia (1988) introduced a way to approximate high-dimensional Gaussian *pdf*s arising from spatial data, which is particularly amenable to modification for our purposes. The starting point of the Vecchia (1988) approximation is to write the joint

density as a product of cascading conditional densities,

$$f(\mathbf{x}) = f(x_1) \prod_{i=2}^{D} f(x_i \,|\, \mathbf{x}_{1:i-1}). \tag{2.3}$$

Here, $f(x_1)$ is the univariate Gaussian density with mean μ_1 and variance Σ_{11}, and, for $i = 2, \ldots, k$, the conditional density $f(x_i \,|\, \mathbf{x}_{1:i-1})$ is the univariate Gaussian density with mean $\mu_i + \boldsymbol{\Sigma}_{[i,1:i-1]} \boldsymbol{\Sigma}^{-1}_{[1:i-1,1:i-1]} (\mathbf{x}_{1:i-1} - \boldsymbol{\mu}_{1:i-1})$ and variance $\Sigma_{i,i} - \boldsymbol{\Sigma}_{[i,1:i-1]} \boldsymbol{\Sigma}^{-1}_{[1:i-1,1:i-1]} \boldsymbol{\Sigma}_{[1:i-1,i]}$. The leading terms in this product are fast to calculate, but for terms corresponding to large i, the computations are nearly as burdensome as those of the original representation (2.1).

To help solve this problem, Vecchia (1988) proposed an approximation to the full joint distribution, in the setting where the random vector $\mathbf{X}$ is observed from a spatial Gaussian process. He modified the cascading conditional representation (2.3) by replacing the conditioning on high-dimensional vectors $\mathbf{x}_{1:i-1}$ with conditioning on well-chosen vectors that have much smaller dimension. By limiting the conditioning sets to vectors of length $m << D$, this strategy replaces expensive $\mathcal{O}(D^3)$ matrix operations with much faster $\mathcal{O}(m^3)$ matrix operations. The approximation of the joint density is then

$$f(\mathbf{x}) \approx f(x_1) \prod_{i=2}^{D} f(x_i \,|\, \mathbf{x}_{\mathcal{N}_i}), \tag{2.4}$$

where $\mathcal{N}_i$ is the conditioning set of size m (more precisely, $\min(m, i-1)$) chosen for the component x_i, for $i = 2, \ldots, D$.

A good choice for a conditioning set to approximate the complete conditional density of each x_i might be the m components that are most correlated with x_i. In the context where the random vector $\mathbf{X} = (X(\mathbf{s}_1), \ldots, X(\mathbf{s}_k))^{\mathrm{T}}$ arises from a stationary spatial Gaussian process observed at locations $\mathbf{s}_1, \ldots, \mathbf{s}_k$, the components most correlated with $X_i \equiv X(\mathbf{s}_i)$ will be those observed at locations that are the m nearest neighbors to $\mathbf{s}_i$ (under covariance models in common use). Other strategies for constructing conditioning sets have also been explored (Guinness 2018, Stein et al. 2004).

Vecchia (1988)'s approximation has been found to be quite accurate under many covariance models and sampling scenarios relevant to analysis of spatial Gaussian processes (Guinness 2018). Moreover, it is very fast to compute, even using the most naive implementation. However, its power is fully realized when the D components of the product are computed in parallel, which is trivially easy to implement using standard tools in `R`.

2.2.2 Extending the Vecchia Approximation for the Gaussian *cdf*

Our approach to approximating the high-dimensional Gaussian *cdf* is to re-write the joint *cdf* as a telescoping product of conditional *cdf*s, analogously to (2.3), and then to approximate each complete conditional *cdf* with *cdf* that conditions on a smaller collection of components, analogously to (2.4). In the case of the *pdf*, this strategy of choosing smaller conditioning sets eliminates the need to compute high-dimensional matrix computations required by (2.1), whereas in the case of the *cdf*, this strategy eliminates the need to compute the high-dimensional integral required by (2.2).

Specifically, we can re-write any joint *cdf* as

$$F(\mathbf{x}) = P(\mathbf{X} < \mathbf{x}) = P(X_1 < x_1) \prod_{i=2}^{D} P(X_i < x_i \,|\, X_1 < x_1, \ldots, X_{i-1} < x_{i-1})$$
$$= P(X_1 < x_1) \prod_{i=2}^{D} P(X_i < x_i \,|\, \mathbf{X}_{1:i-1} < \mathbf{x}_{1:i-1}) \tag{2.5}$$

Then, just as in the approximation to the *pdf* (2.4), in the *cdf* (2.5) each conditional probability in the product can be approximated by reducing the size of the conditioning set to at most m components. Thus, our Vecchia approximation for the Gaussian *cdf* is

$$F(\mathbf{x}) \approx P(X_1 < x_1) \prod_{i=2}^{D} P(X_i < x_i \,|\, \mathbf{X}_{\mathcal{N}_i} < \mathbf{x}_{\mathcal{N}_i})$$
$$= P(X_1 < x_1) \prod_{i=2}^{D} \frac{P(X_i < x_i, \mathbf{X}_{\mathcal{N}_i} < \mathbf{x}_{\mathcal{N}_i})}{P(\mathbf{X}_{\mathcal{N}_i} < \mathbf{x}_{\mathcal{N}_i})}$$
$$= \Phi(x_1) \prod_{i=2}^{D} \frac{\Phi(\mathbf{x}_{\{i,\mathcal{N}_i\}})}{\Phi(\mathbf{x}_{\mathcal{N}_i})}, \tag{2.6}$$

where again $\mathcal{N}_i$ is the conditioning set of size $\min(m, i-1)$ chosen for the component x_i, for $i = 2, \ldots, D$.

The approximation given by (2.6) reduces computational costs by replacing the D-dimensional integral in (2.2) with a series of much simpler integrals of dimension $m+1$ and m, for $m << D$. Furthermore, all of the elements in the product can be computed in parallel.

The multivariate *cdf*s in (2.6) still have to be evaluated numerically. For all but the smallest possible choices of m, best practices suggest using a QM method like that of Genz & Bretz (2009) to approximate the numerator and denominator.

Similarly to the original Vecchia (1988) approximation to the Gaussian *pdf*, choosing the conditioning sets involves a trade-off; choose m too small and the accuracy of the approximation will suffer, but choose m too large and the computational benefits will diminish.

2.3 Simulation Study

To assess the accuracy and speed of this approximation, and to explore the trade-off inherent in the choice of m, we conduct a simulation study. Since the true value of the *cdf* is not available, the best we can do to check for accuracy is to see whether it is consistent with results obtained from direct use of the Genz & Bretz (2009) QM approach. We simulate a Gaussian process observed on equally spaced grids of five different sizes, 15×15, 30×30, 50×50, 75×75 and 100×100. We try two different covariance functions for the Gaussian process to see whether this has an impact on the *cdf* estimation: an exponential model with range parameter 1 and and exponential model with range parameter 5, each with unit variance. This makes a total of 10 different scenarios. For each scenario, we used four different sizes of conditioning sets, choosing $m = 5$, 10, 30 and 50 closest neighbors. For comparison, we computed the Genz & Bretz (2009) QM method using 499 and 3,607 sample points. We use the implementation of the Genz & Bretz (2009) algorithm in the `mvPot` package (de Fondeville & Belzile 2018) for `R`. In principle, the accuracy and computational requirements of the QM grows with the number of sample points (which, here, must

be a prime number). Since the algorithms are stochastic, we repeated each calculation five times and plotted each replication as a dot in Figures 2.1, 2.2, 2.3, and 2.4.

Figure 2.1 shows the value of the estimated log *cdf* for all grid sizes and all estimation methods for the simulated Gaussian process with range parameter 1. The log *cdf* estimated with the Vecchia approximation increases with the number of neighbors until it stabilizes for 30 neighbors, after which it is consistent with the two QM approximations. This suggests that, under this scenario, it is advisable to use at least 30 neighbors in order to estimate the log *cdf*. For the two smaller grids, it appears that the Vecchia approximation has a similar variance to the QM approximation using 499 sample points, but a higher variance that the QM approximation using 3,607 sample points. For the larger grids, the Vecchia approximation appears to have a lower variance than both QM approximations. Figure 2.2 shows the same as Figure 2.1, but for exponential Gaussian processes with range parameter 5. The story is similar to the case with the shorter range process, except it appears that 50 neighbors may be necessary in order to stabilize the estimated log *cdf*. It may be the case that the number of neighbors necessary to accurately approximate the log *cdf* increases with length of the dependence of the Gaussian process. Intuitively, this may occur because for processes with longer-range dependence, a smaller proportion of the information in data may be captured by local approximations.

Figures 2.3 and 2.4 show the time required to approximate the log *cdf*, on a single core, for Gaussian processes with range parameters of 1 and 5, respectively. The computation time is influenced by both the number of observations and number of

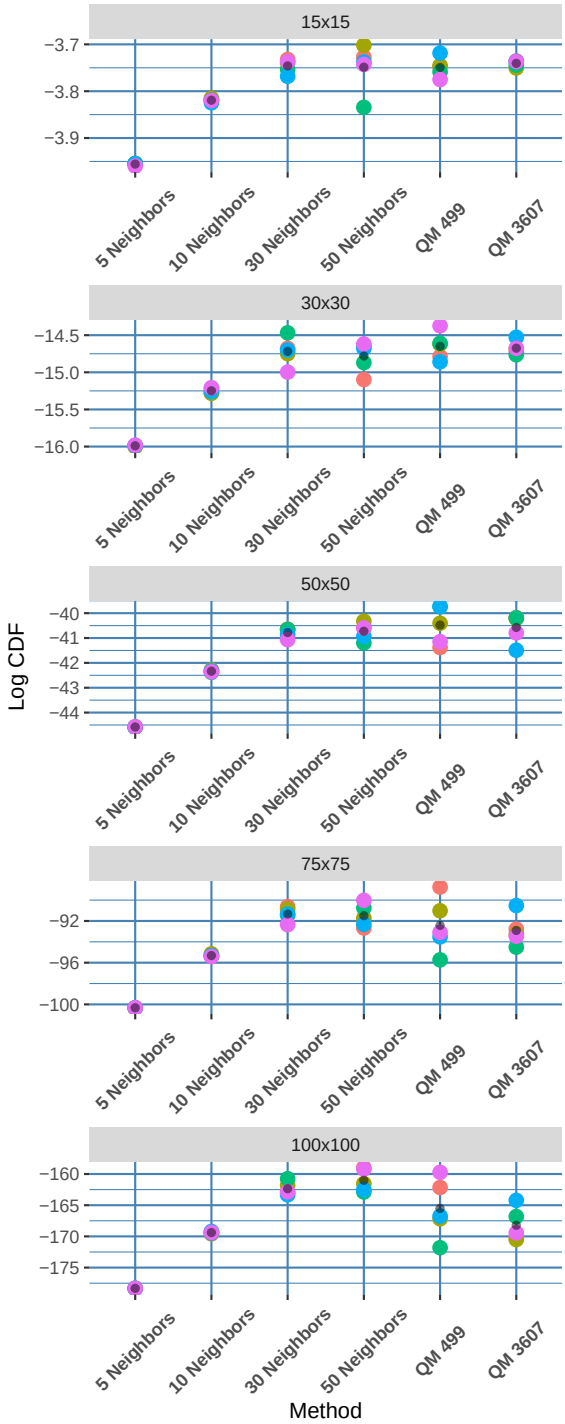


Figure 2.1: Estimated log *cdf* for exponential Gaussian processes with range parameter $\rho = 1$. The x-axis represents the different methods used for the *cdf* computation and the y-axis is the log *cdf*. Each point is an independent estimate of the log *cdf*, and each black point is the average over the replications. The Vecchia approximation seems to stabilize when at least 30 neighbors are used, and results in values that are consistent with the QM approximations.

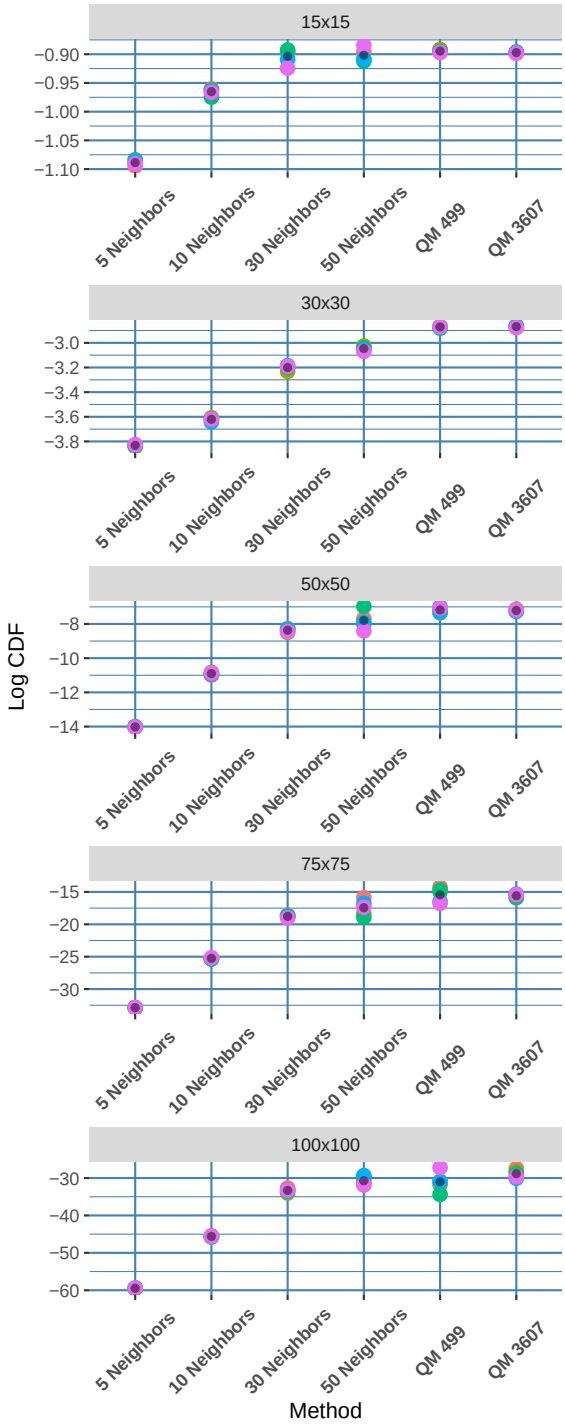


Figure 2.2: Estimated log *cdf* for exponential Gaussian processes with range parameter $\rho = 5$. The x-axis represents the different methods used for the *cdf* computation and the y-axis is the log *cdf*. Each point is an independent estimate of the log *cdf*, and each black point is the average over the replications. For this process with longer-range dependence, the Vecchia approximation may not stabilize until at least 50 neighbors are used, when results become consistent with the QM approximations.

neighbors used in the Vecchia approximation. Computational costs increase with the number of observations, for both the Vecchia and QM approximation methods, and also increase with the number of neighbors in the conditioning set. Oddly, the empirical computation time did not increase for the QM approximation with the larger set of sample points. For smaller grid sizes, the QM methods are faster than the Vecchia approximations, except when the size of the conditioning set very small. For grids of size 50×50 and larger, computation time of the approximation using 30 neighbors was as fast as or faster than the QM method. When the number of observations is extremely large, in the case of the 100×100 grid, the computation time was much lower for the Vecchia approximation compared to the QM approximation. This suggests that for high-dimensional datasets the use of the Vecchia approximation is preferable QM method, even if computations are done sequentially.

2.3.1 Parallel Computing

Since each term of the Vecchia *cdf* approximation (2.6) is independent of every term, it is trivial to parallelize the computations. In practice, we compute all of the required low-dimensional Gaussian *cdf*s on the log scale, and then sum them at the end. In principal, the speedup should be linear in the number of cores used for the calculation. To explore this relationship, we compute the *cdf* approximation based on a Gaussian process observed at 10,000 locations, varying the number of compute cores used between 5 and 40. For each setup, we repeat the computation 15 times. Figure 2.5 shows time required to compute the log *cdf* approximation. The computing

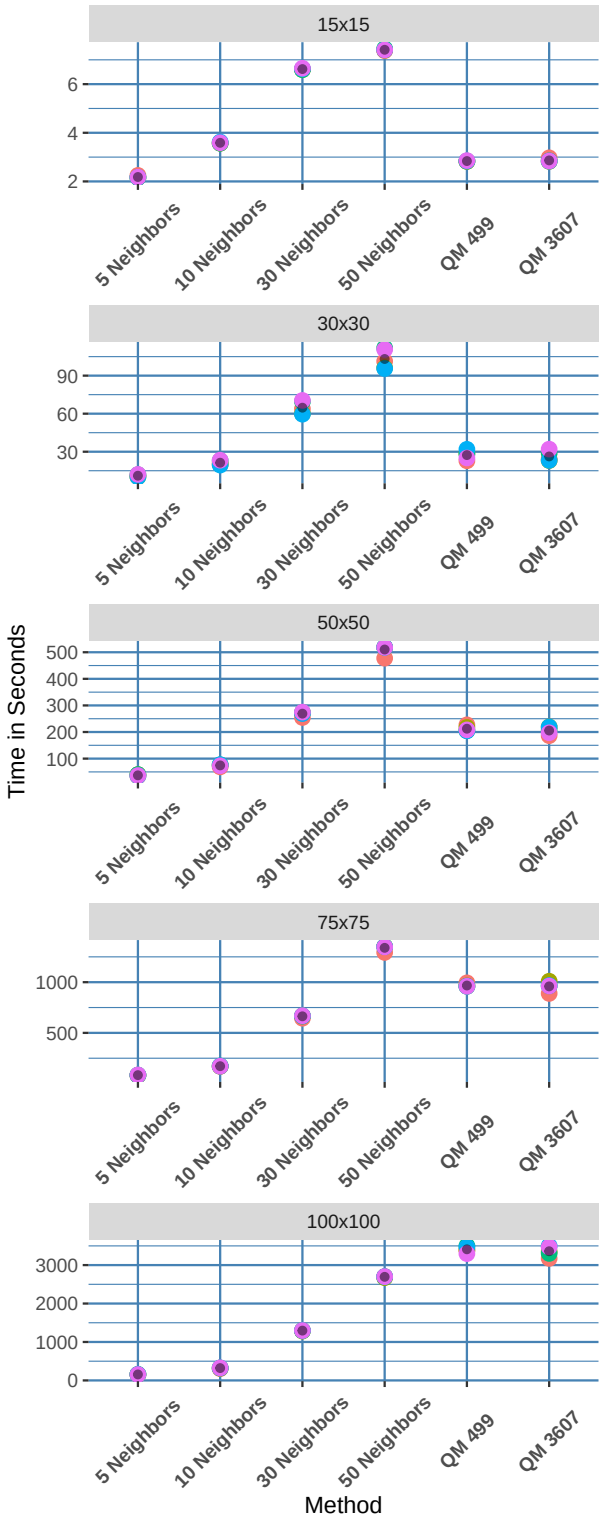


Figure 2.3: Time to estimate the *cdf* approximation for an exponential Gaussian process with range parameter $\rho = 1$. The x-axis represents the different approximation methods, and the y-axis is the computation time. Each point is an independent replication of the procedure, and the black point is the average over the replications.

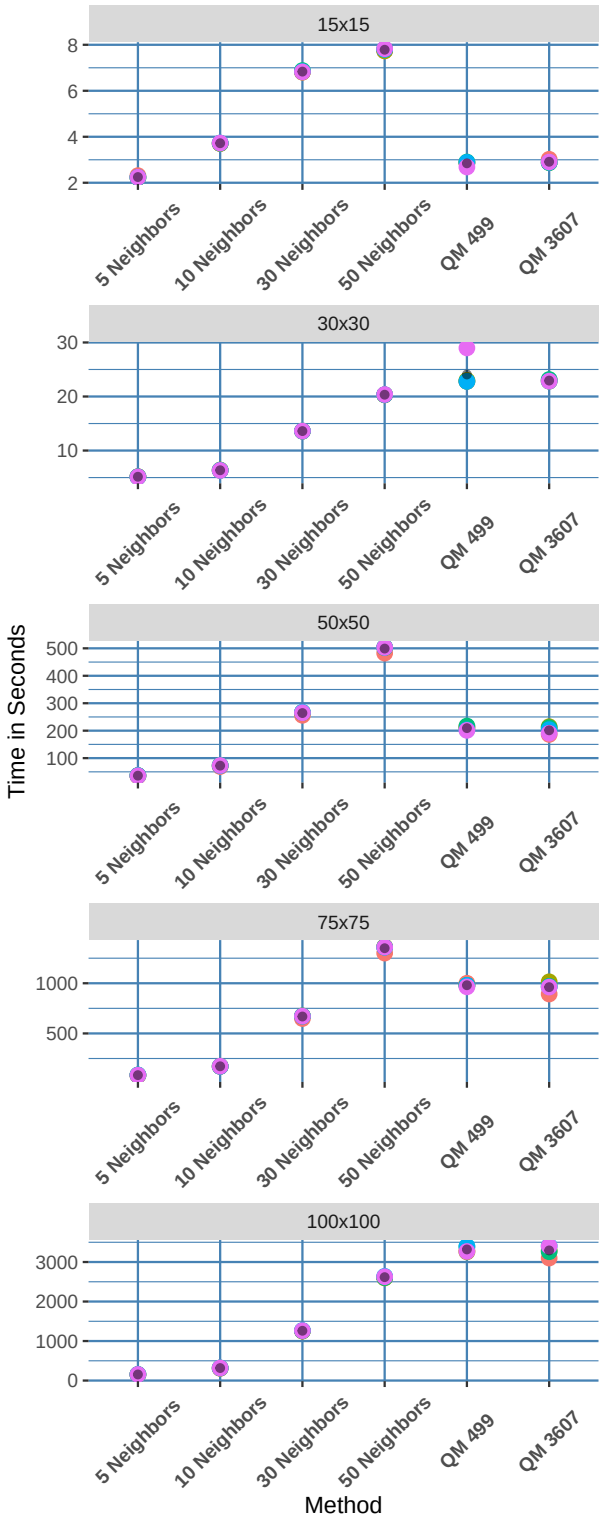


Figure 2.4: Time to estimate the *cdf* approximation for an exponential Gaussian process with range parameter $\rho = 5$. The x-axis represents the different approximation methods, and the y-axis is the computation time. Each point is an independent replication of the procedure, and the black point is the average over the replications.

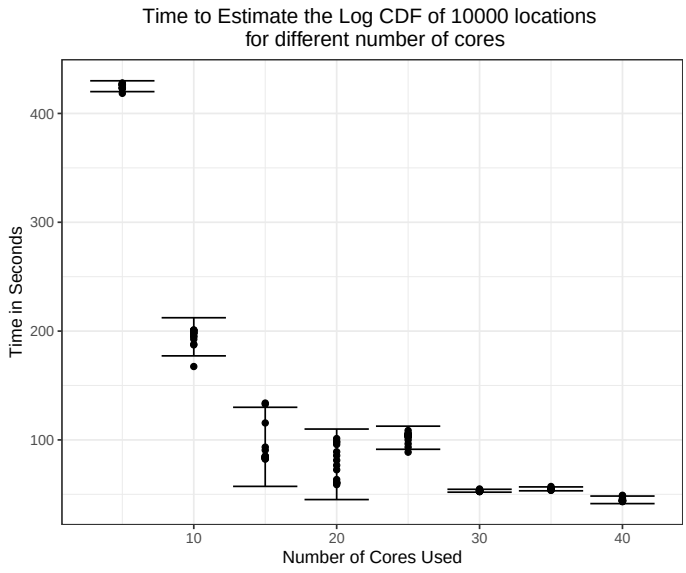


Figure 2.5: Time to compute log *cdf* approximation parallelized across different numbers of computing cores.

time decreases with the number of cores. We observe roughly the expected linear relationship up to 20 cores, when a jump occurs before again decreasing. We suspect that this is behavior a result of the particular hardware configuration we used, which consists of networked 20-core processors. That is, we guess that once an additional physical processor is engaged, which occurs beyond 20 cores, overhead costs increase and attenuate the expected computational gains. When 40 cores were used, it took less than 1 minute to compute the log *cdf* approximation for 10,000 observations. There are clearly some diminishing returns due to communication overhead, but in principle, this approximation could be made arbitrarily fast with a big enough computing system.

2.3.2 Effect of Neighbor Selection and Joint Estimation

The representation defined by equation (2.5) and its approximation (2.6) calculates the joint probability as the product of univariate conditional distributions. However it is also possible to write the full joint *cdf* as a cascading product of multivariate, rather than univariate, conditional *cdf*s. Under equation 2.6, it is necessary to calculate the n univariate conditional probabilities, each of which requires a $m + 1$-dimensional *cdf* calculation. If instead we divide the components into q groups of p joint observations, such that $q \times p = n$, we would only need to calculate the product of q conditional probabilities. However, doing so would make the dimensionality of each individual Gaussian *cdf* calculation in (2.6) between $m + p$ and $pm + p$. So it would trade the cost of computing higher-dimensional *cdf* terms for the benefit of computing fewer terms. Such a trade-off could affect both the accuracy and computational efficiency of the approximation. Guinness (2018) explored this possibility in the context of *pdf*s and found that it can be advantageous to consider multivariate conditional densities in the Vecchia density approximation. To explore the effect of calculating higher dimensional conditional probabilities, we calculate the log *cdf* approximation based on groupings of observations of different sizes.

An additional consideration that could effect the accuracy and speed of the approximation is the construction of the conditioning sets. Using the nearest neighbors, as we have done above, requires the additional step of ordering the components by distance, which could be slow. Choosing randomly-selected conditioning sets could potentially

speed up the computation by avoiding this sorting step.

Figures 2.6 and 2.7 show the estimated log *cdf* and time (in seconds) to compute the approximated log *cdf*, using the 100×100 grid. We used approximations based on joint conditional *cdf*s of dimension 2, 5, 10, 20, 30, and 50. For each grouping size, we constructed conditioning sets using 3 different methods. The first method conditions on the m most correlated observations (in this case simply the nearest neighbors) for each observation in the joint grouping, resulting in a conditioning set of size pm. The second method simply conditions on m random observations. The third method conditions on m random observations per element of the multivariate conditional calculation, again resulting of a conditioning set of size pm.

From Figure 2.6 it is clear that simply conditioning on m random observations fails to yield an acceptable approximation. Performance can be improved by conditioning on more random observations, which is what the third method does. Method 3 shows somewhat improved behavior, however it was only able to perform acceptably when both the dimensionality p of the joint conditional probability and the size of the conditioning set pm were both large. It is clear from Figure 2.6 that conditioning on random neighbors is much less accurate than conditioning on the most highly correlated neighbors. For conditioning sets consisting of small numbers m of neighbors per element in the joint conditional probability, the use of a large group p of joint observations had a better result, probably simply due to fact that the total number pm of neighbors in the conditioning set was larger. However, when the number m of correlated neighbors gets large enough the number p of joint observations does not

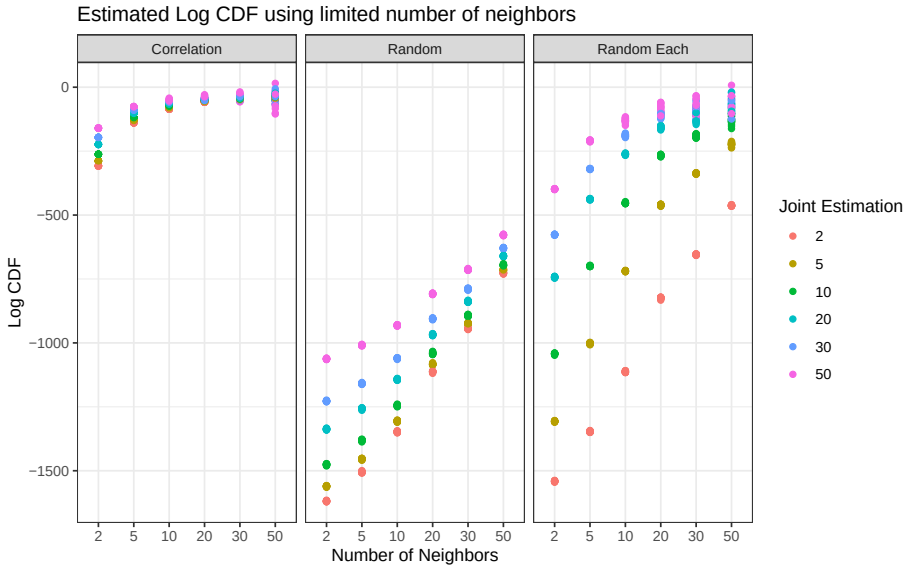


Figure 2.6: Estimated log *cdf* based on observations from an exponential Gaussian process with range parameter $\rho = 1$, using 3 different methods to select conditioning sets, and different dimensionalities of joint conditional observations.

seem to affect result of the approximation.

Figure 2.7 shows the computation time required for all of the approximation schemes depicted in Figure 2.6. The clear trend is that choosing a small conditioning set of random observations is very fast (middle panel), using higher-dimensional joint conditional *cdf*s is slower than using lower-dimensional joint conditional *cdf*s (all panels), and for the same size conditioning set, the time required to find the nearest neighbors is not a major bottleneck (right and left panels). This conclusion is different from exploration of the same issues, in the context of the *pdf*, found in Guinness (2018). There, using higher-dimensional joint conditional calculations was found to be beneficial, and the time required to find nearest neighbors was substantial enough to warrant the use of a fast approximate ordering algorithm. In the case of the *cdf* approximation, code profiling confirmed that the time required

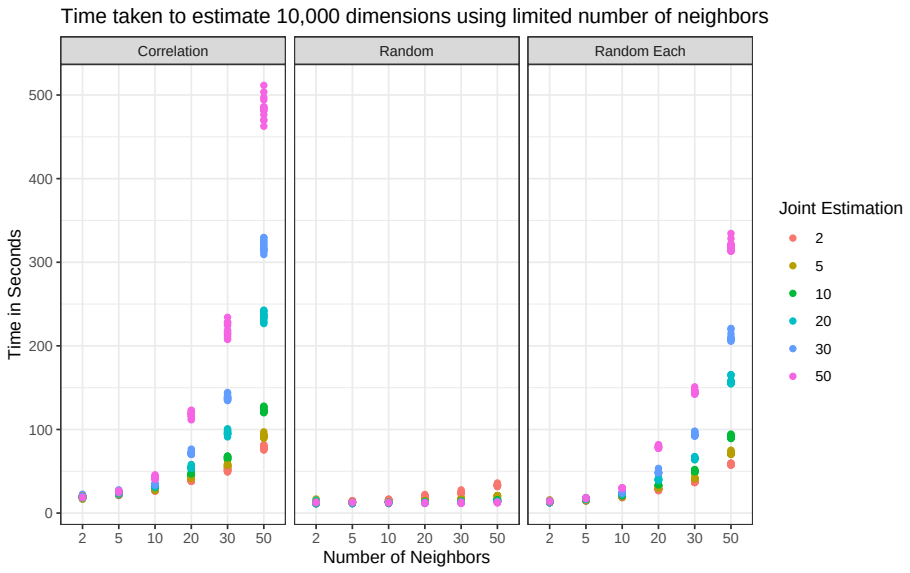


Figure 2.7: Time to estimate the log CDF with dependence parameter $\rho = 1$ using 3 different methods to select neighbors, multiple number of neighbors and multiple number of joint observations.

to order the observations was insignificant, with the overwhelming majority of the computation time being used in calculating the lower-dimensional joint *cdf*s using the QM technique.

2.4 Example: A Gaussian Scale Mixture for Spatial Extremes

Recent advances in the statistics of extremal spatial phenomena have produced models that are flexible enough to accommodate both strong and weak spatial dependence in the far joint tails. One prominent strategy for achieving this is to construct scale mixtures of Gaussian processes, where the mixing distribution is chosen carefully so as to produce the desired tail dependence characteristics (Huser et al. 2017, Huser &

Wadsworth 2019, Morris et al. 2017, Opitz 2016). The preferred flavor of maximum likelihood inference for these models requires computing a Gaussian *cdf* whose dimension is roughly equal to the number of spatial locations in the dataset. Other state-of-the-art models for spatial extremes also rely on high-dimensional Gaussian *cdf*s (de Fondeville & Davison 2018, Thibaud et al. 2016, Wadsworth & Tawn 2014). To show the usefulness of our *cdf* approximation, we analyze data from precipitation gauges in Europe using the Gaussian scale mixture model from Huser et al. (2017), which we describe below.

The class of scale mixtures of Gaussian processes is defined generically by

$$\begin{aligned} X(\mathbf{s}) &= R \times W(\mathbf{s}) \\ R &\sim F_R \perp\!\!\!\perp W(\mathbf{s}). \end{aligned} \tag{2.7}$$

Here, $W(\mathbf{s})$ is a standard Gaussian process (i.e. with unit variance) on some domain $\mathcal{D}$ indexed by $\mathbf{s} \in \mathcal{D}$. For a collection of k observations, the finite dimensional distribution of the Gaussian component is $\mathbf{W} \sim N_k(0, \mathbf{\Sigma}(\theta))$, where $\mathbf{\Sigma}(\theta)$ is a $D \times D$ covariance matrix constructed using a chosen covariance model that is indexed by parameter θ.

The random scaling R comes from distribution F_R. The choice of F_R is critical and determines the strength of the tail dependence in the resulting model (Engelke et al. 2019). A key quantity for summarizing the strength of tail dependence is the conditional probability $\chi_u(\mathbf{s}_i, \mathbf{s}_j) = P\{X(\mathbf{s}_i) > u \,|\, X(\mathbf{s}_j > u\}$, for spatial locations $\mathbf{s}_i$ and $\mathbf{s}_j$. If $\lim_{u\to\infty} \chi_u(\mathbf{s}_i, \mathbf{s}_j) = 0$ for all $\mathbf{s}_i, \mathbf{s}_j \in \mathcal{D}$, we say that $X(\mathbf{s})$ is *asymptotically indepen-*

dent, while if $\lim_{u\to\infty} \chi_u(\mathbf{s}_i, \mathbf{s}_j) > 0$ for all $\mathbf{s}_i, \mathbf{s}_j \in \mathcal{D}$, we say that $X(\mathbf{s})$ is *asymptotically dependent*.

While many choices are available for the mixing distribution F_R, Huser et al. (2017) suggest the parametric model defined by equation (2.8). When $\beta > 0$, the mixture process $X(\mathbf{s})$ is asymptotically independent, and when $\beta = 0$, $X(\mathbf{s})$ is asymptotically dependent. Therefore, this class of scale mixtures is rich enough to include both asymptotic independence and asymptotic dependence as nontrivial sub-models.

$$F_R(r) = \begin{cases} 1 - \exp\{-\gamma(r^\beta - 1)/\beta\}, & \text{for } \beta > 0 \\ 1 - r^\gamma, & \text{for } \beta = 0. \end{cases} \tag{2.8}$$

To construct the likelihood for maximum likelihood estimation, we must integrate out R from the model (2.7). Equations (2.9) and (2.10) show the marginal multivariate *cdf* and *pdf*, respectively, for a finite collection of observations $\mathbf{X}$ from $X(\mathbf{s})$ defined in (2.7). Here Φ_D represents the D-dimensional multivariate *cdf* from a Gaussian distribution with mean vector 0 and covariance matrix $\mathbf{\Sigma}(\theta)$, and ϕ_D represents the D-dimensional multivariate *pdf* from a Gaussian distribution with mean 0 and covariance matrix $\mathbf{\Sigma}(\theta)$. There are no closed forms for these expressions, so it is necessary to use numerical methods to evaluate the (one-dimensional) integrals.

$$G(\mathbf{x}) = \int_0^\infty \Phi_D(\mathbf{x}/r; \mathbf{\Sigma}) f_R(r) dr \tag{2.9}$$

$$g(\mathbf{x}) = \int_0^\infty \phi_D(\mathbf{x}/r; \mathbf{\Sigma}) r^{-D} f_R(r) dr. \tag{2.10}$$

The preferred strategy for maximum likelihood estimation of extremal dependence models is to treat all observations falling below a high threshold as left censored (Huser et al. 2016). This leads to a favorable balance between using the data as efficiently as possible, while not allowing data in the bulk of the distribution to have a large effect on dependence estimation. The censored likelihood for each temporal replicate is obtained by taking one partial derivative of (2.9) for every observation that falls above the threshold. Thus, (2.10) is the relevant likelihood when all observations, at one particular temporal replicate, are above the threshold, so nothing is censored. However, since the threshold is chosen to be a high quantile to prioritize inference on the tail, most observations are usually censored for any temporal replicate. When all observations fall below the threshold, the relevant likelihood is (2.9).

Most often, in any temporal replicate, there will be a mixture of observations above and below the threshold. In this case, the relevant joint likelihood of $\mathbf{x}$ is defined by equation (2.11), which results from taking partial derivatives of (2.9) with respect to only the un-censored observations. If we let I be the set of points above the threshold and I^c be the points below, then

$$\begin{aligned} G_I(\mathbf{x}) &:= \frac{\partial^{|I|}}{\partial \mathbf{x}_I} G(\mathbf{x}) = \int_0^\infty \frac{\partial^{|I|}}{\partial \mathbf{x}_I} \Phi_k(\mathbf{x}/r; \mathbf{\Sigma}) f_R(r) dr \\ &= \int_0^\infty \Phi_{|I^c|}\{(\mathbf{x}_{I^c} - \mathbf{\Sigma}_{I^c;I}\mathbf{\Sigma}_{I;I}^{-1}\mathbf{x}_I)/r; \mathbf{\Sigma}_{I^c|I}\} \phi_{|I|}(\mathbf{x}_I/r; \mathbf{\Sigma}_{I;I}) r^{-|I|} f_R(r) dr, \end{aligned} \tag{2.11}$$

where dependence of the covariance matrices on θ is suppressed for brevity, and the notation $\mathbf{\Sigma}_{A;A}$ refers to rows and columns of $\mathbf{\Sigma}$ pertinent to the points in A. The matrix $\mathbf{\Sigma}_{I^c|I} = \mathbf{\Sigma}_{I^c;I^c} - \mathbf{\Sigma}I^c; I\mathbf{\Sigma}_{I;I}^{-1}\mathbf{\Sigma}_{I;I^c}$ is the covariance matrix of the conditional normal distribution of the censored observations given the un-censored observations.

The computational issue arises because the integrand (2.11) contains a Gaussian *cdf* of dimension $|I^c|$, the number of censored observations in a temporal replicate. Again, for most replicates, this number $|I^c|$ is close to the total number of observation locations D because the censoring threshold is chosen to be high, such that most observations fall below the threshold and are therefore censored.

2.4.1 Precipitation Over Europe

Our dataset consists of weekly maximum precipitation observations between January, 2000 and April, 2019, in the western and central region of continental Europe, north of the mountain ranges the Pyrenees, Alps, and Carpathians. The 6 countries we consider are Germany, Poland, Netherlands, Belgium, Czech Republic, and France. Figure 2.8 shows the locations of the observation stations distributed over Europe. This dataset consists of 1,006 weekly maxima from $D = 528$ weather stations. For context, the computational bottleneck from the Gaussian *cdf* limited the analysis in Huser et al. (2017) to a dataset of $D = 12$ locations, even though analysis was performed on a large high-performance computing cluster. We use the weekly maximum daily accumulations at each location to break temporal dependence that might arise from storms that persist for more than one day. Out of the 531,168 total observations, 32.6%

were missing values. For each weekly maximum, only the available data was used for estimation, and all missing observations were disregarded.

Figure 2.8: $D = 528$ weather stations located over 6 European countries

The covariance model we use for the underlying Gaussian processes is an anisotropic exponential, $\Sigma_{ij}(\theta) = \exp\{-h_{ij}/\rho\}$, where ρ is the range parameter and h_{ij} is the Mahalanobis distance between locations $\mathbf{s}_i$ and $\mathbf{s}_j$. The Mahalanois distance is parametrized as

$$h_{ij}^2 = \Omega^{\mathrm{T}}\Omega, \quad \text{where } \Omega = (\mathbf{s}_i - \mathbf{s}_j)^{\mathrm{T}} \begin{pmatrix} \cos(\phi) & -\sin(\phi) \\ \sin(\phi) & \cos(\phi) \end{pmatrix}^{-1} \begin{pmatrix} 1 & 0 \\ 0 & A \end{pmatrix},$$

for rotation angle $\phi \in [0, \pi)$ and and aspect ratio $A > 1$. Thus, after fixing the mixing

parameter γ at 1, as it plays a much less significant role than the parameter β in determining tail dependence characteristics, we arrive at a total of 4 parameters to estimate, $\psi = (\beta, \rho, \phi, A)^{\mathrm{T}}$.

The first step in estimating the dependence is to transform the observations to be on the same marginal scale. To do this, we start by applying a rank transformation to standard uniform, independently for each station. That is, for each station $k = 1, \ldots, D$ and each time point $t = 1, \ldots T$, the observation X_{kt} on the uniform scale is

$$U_{kt} = \frac{\mathrm{rank}(X_{kt})}{T+1}.$$

We next choose a high threshold to be the 0.95 marginal empirical quantile at each location. Then, denoting the marginal *cdf* and *pdf* of each X_{kt}, respectively, as $G_M(x) = \int_0^\infty \Phi(x/r) f(r) dr$ and $g_M(x) = \int_0^\infty \phi(x/r) r^{-1} f(r) dr$ (we assume stationarity, so the marginal distribution is assumed to be the same at each location), and letting the vector $\mathbf{v}_t = (\max\{u_{1t}, 0.95\}, \ldots, \max\{u_{Dt}, 0.95\})^{\mathrm{T}}$, the copula censored likelihood for each time replicate k is

$$L(\psi; \mathbf{v}_t) = \begin{cases} G\{G_M^{-1}(v_{1t}), \ldots, G_M^{-1}(v_{Dt})\} & \text{if all obs. are below the threshold} \\ \frac{g\{G_M^{-1}(v_{1t}), \ldots, G_M^{-1}(v_{Dt})\}}{\prod_{k=1}^{D} g_M\{G_M^{-1}(v_{kt})\}} & \text{if all obs. are above the threshold} \\ \frac{G_{I_t}\{G_M^{-1}(v_{1t}), \ldots, G_M^{-1}(v_{Dt})\}}{\prod_{k \in I_i}^{D} g_m\{G_M^{-1}(v_{kt})\}} & \text{if some obs. are above and some below the threshold} \end{cases}$$

Finally, the log likelihood across all time points t for the parameter vector ψ is

$$l(\psi; \mathbf{v}) = \sum_{t=1}^{T} \log(L(\psi; \mathbf{v}_t)).$$

We found the maximum likelihood estimator (MLE) by applying the Nelder-Mead numerical optimizer in the `R` function `optim`. MLEs are shown in Table 2.1. The MLE for the mixing parameter β is 0.82, which in this context is fairly far away from zero—far enough to strongly suggest that the process is asymptotically independent. The MLEs for the anisotropy parameters suggest pronounced eccentricity. To interpret and visualize the estimated dependence model implied by the MLEs shown in Table 2.1, we plot level curves in the resulting χ_u function for $u = 0.95$ on the quantile scale, shown in Figure 2.8. Each ellipse represents a constant value of $\chi_{u=0.95}(\mathbf{s}) = P\{F_M[X(\mathbf{s})] > 0.95 \,|\, F_M[X(\mathbf{s}_0)] > 0.95\}$, for an arbitrarily-chosen reference point $\mathbf{s}_0$ near the center of the map. The level curves are ellipses due to the anisotropic construction, with the major axis roughly along a northeast-southwest orientation, and joint exceedances more likely with decreasing distance from $\mathbf{s}_0$.

Parameter	MLE
ρ	1.31
β	0.82
ϕ	1.10
A	2.29

Table 2.1: Maximum likelihood estimates of dependence parameters

2.5 Discussion

The main objective of this work was to propose fast approximation to high-dimensional Gaussian *cdf*s that arise from spatial Gaussian processes. We modified Vecchia (1988)'s approximation for Gaussian *pdf*s to the context of Gaussian *cdf*s. Simulations showed that for large numbers of locations and relatively small conditioning sets, this approximation gives results consistent with state-of-the-art QM methods, and reduces computational time considerably, even when computations are performed sequentially. Furthermore, the approximation is trivially easy to code in parallel using standard `R` packages, and requiring no linking to specialized software libraries.

We demonstrated the utility of our fast *cdf* approximation by using it to find maximum censored likelihood estimates for the scale mixture model of Huser et al. (2017). This model is attractive because of its flexible tail dependence characteristics, but is hampered by computational difficulties arising from the need to compute high-dimensional Gaussian *cdf*s during inference. We fit this model to a precipitation dataset consisting over 500 spatial locations, whereas previous efforts using conventional QM techniques were limited to just 12 locations.

One drawback that we noticed during the data analysis is that conventional optimization routines had trouble converging, due to the stochastic nature of the likelihood objective function. For future studies, one possible approach to circumventing this problem is to use stochastic optimization algorithms, which may be better suited to optimizing random objective functions.

Chapter 3

Report on: Methods to Predict Remaining Life Time of Jet Engines In the Presence of a Change Point

3.1 Introduction

One important question the Department of Defense has is "When should we bring our jet back in order to perform maintenance?" This questions have many ramifications. If a jet is brought back too early, maintenance will be performed before it is necessary, which would generate cumulative costs and decrease the available number of units for missions. If a jet is brought back to late, the cost of maintenance would be higher

than necessary, and the jet would risk malfunction during missions. In order to solve this problem, it is important to be able to decide when is the correct time to call back their jets, minimizing the cost associated with this action.

The minimization of this cost will depend heavily on estimation of the remaining useful life time of the jet's engines. Data from the jets' various sensors can be used to determine an advantageous time to return for maintenance. Due to restrictions on military data, obtaining data from operational jet engines is complicated, however NASA as made available the NASA C-MAPSS (Commercial Modular Aero-Propulsion System Simulation) dataset. This dataset consists of synthetic data produced using a model-based simulation program C-MASPSS. Developed by NASA, this program generates simulated jet engines based on controlling variables. Each engine is simulated from the beginning of its life time until failure, where for each time point we know the remaining life time and simulated readings on the state of the system at each moment.

The main object of this report is to analyze and offer options for methods to estimate the remaining life time of those simulated jet engines based on the state of the system at each given time point. The report is divided in the following structure: Section 3.2 will describe the data and present exploratory data analysis in order to understand the general behaviour of the explanatory variables. Section 3.3 will present the methodology used on this paper to estimate the remaining useful life time, including models and different approaches used. Section 3.4 will analyze one of the available datasets. Lastly, section 3.5 will give our conclusion about our models.

	FD001	FD002	FD003	FD004
Training Trajectories	100	260	100	249
Training Total Number of Observations	20631	53759	24720	61249
Testing Trajectories	100	259	100	248
Testing Total Number of Observations	13096	33991	16596	41214
Operating Conditions	1	6	1	6
Fault Conditions	1	1	2	2

Table 3.1: Basic summary of the C-MAPSS datasets.

3.2 Exploratory Data Analysis

The dataset we use is the NASA C-MAPSS. This dataset has been divided into 4 different sub-datasets, representing 4 different scenarios and conditions. Table 3.1 summarizes the differences between these 4 datasets.

Each dataset consists of 26 columns containing the response variable, a unique identifier, and 24 state variables to use as potential covariates. Each dataset contains a different number of rows, with each row containing an observation of the response and associated state variables. The response variable is *Remaining Useful Lifetime* (RUL), which is the time until failure for each simulated jet trajectory. The exact type of measurements that the covariates represent are obscured from us, so they are referred to generically with names like V_{XX} and OS_X. Of the potential covariates, we use the 21 columns labeled V_{XX}. They represent measurements of 21 observed system variables for each unit through time. An additional set of 3 data columns labeled OS_X contains variables that govern the system, however these variables are constant over time and are therefore not useful for dynamically predicting RUL. Our main objective then is to

use the V_{XX} covariates to predict RUL.

The first dataset, $FD001$, consists of a single operating condition, with all OS_X variables are constant for all 100 simulated trajectories/engines, and a single fault condition. The dataset $FD002$ consists of 6 different conditions, corresponding to 6 different combinations of the OS_X variables. The dataset $FD003$ has a single operating condition, like $FD001$, however there are 2 fault conditions. Lastly $FD004$ consists of 6 operating conditions with 2 fault conditions. The fault condition common to all datasets is high pressure compressor degradation while for $FD003$ and $FD004$ fan degradation was added as the second fault.

The first step of the analysis was to do an explanatory data analysis of the data. The 21 covariates can be divided in 4 different categories, depending on their qualitative behavior across time. To illustrate each of these categories, we present a plot where the x-axis is the RUL and y-axis is a V_{XX}. This choice of plotting is unconventional (i.e. the response variable is on the x-axis) was done in order to emphasize how each covariate changes when the RUL gets close to 0. Each color in the plot represents a different simulated trajectory, to facilitate comparison across different trajectories.

1. The first category of explanatory variables represents variables that do not change much with time. Those variables are V_1, V_5,V_6, V_{10}, V_{16}, V_{18}, and V_{19}. Since those variables do not change, they cannot be used to predict RUL, and hence will not be included in the analysis.

2. The general behaviour of the second category of variables can be seen in Figure 3.1. This figure represents the covariate V_2 in the y-axis and the RUL in the

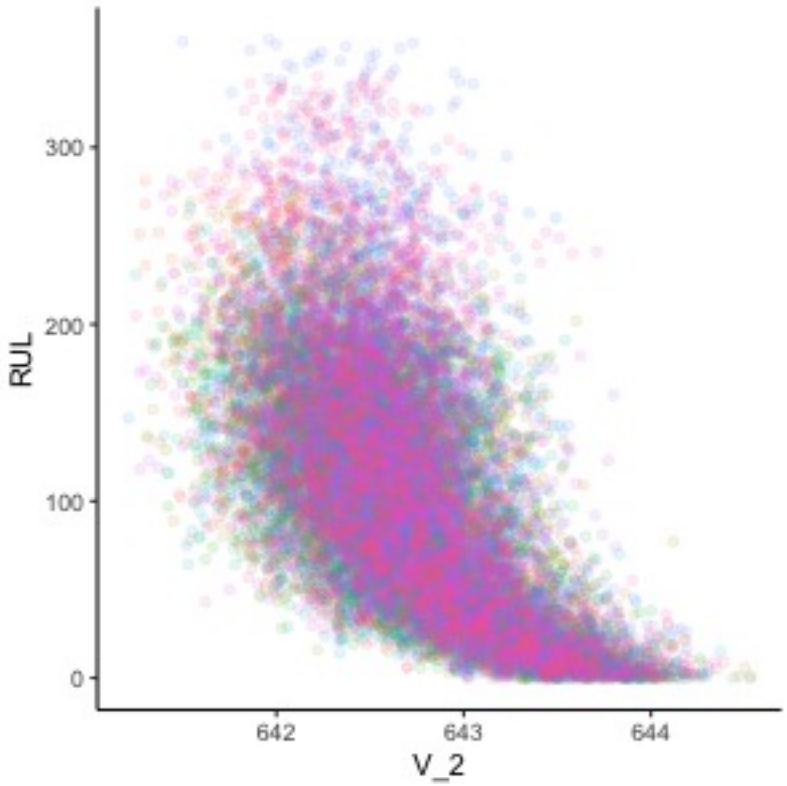


Figure 3.1: Variable V_2 on the x-axis and RUL on the y-axis

x-axis. The general behaviour in this group is constant in the beginning of the simulation, which means the system is working correctly, however after one point the average behaviour changes and variables start to increase. The variables that follow this behavior are V_2, V_3, V_4, V_8, V_{11}, V_{13}, V_{15}, and V_{17}.

3. Figure 3.2 represents the third type of variable in this dataset. This figure represents the variable V_7 in the y-axis and RUL in the x-axis. The variable behaves in a similar way to V_2, constant in the beginning of the simulation, and then the system starts to degenerate after some time. The difference is that instead of a increasing behavior when the RUL gets close to 0, it decreases. The variables that follow this behavior are V_7, V_{12} V_{20}, and V_{21}.

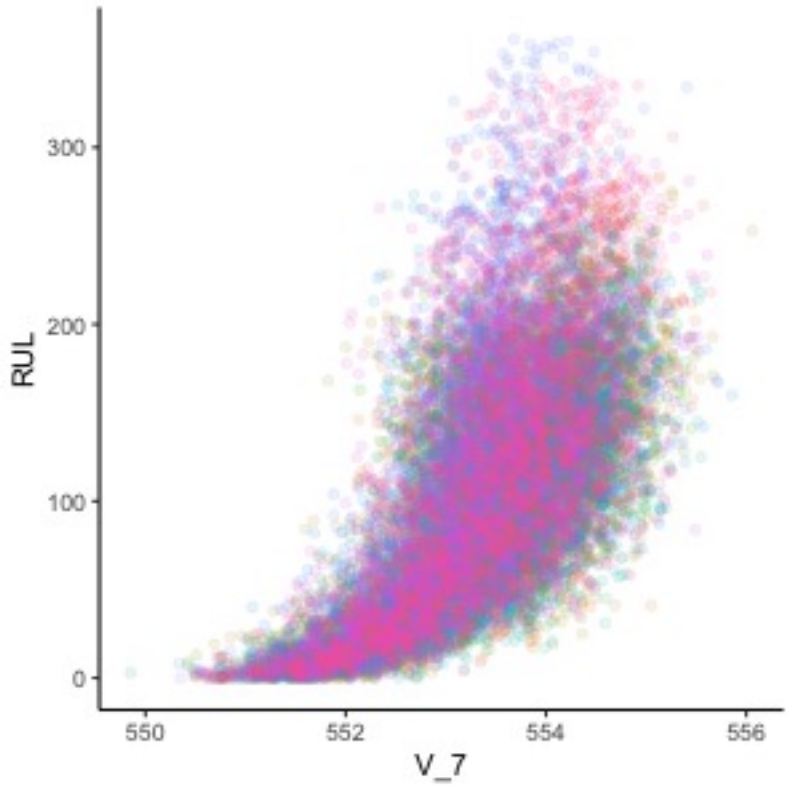


Figure 3.2: Variable V_7 on the x-axis and RUL on the y-axis

4. The last type of variable can be seen in Figure 3.3. The variable represented here is variable V_{14} in the y-axis and RUL in the x-axis. This type of variable is a mixture of the previous 2 cases. Just like before, the variables are constant in the beginning of the simulation but they start to change after one specific point in time. This change corresponds to the change on average behavior, where some of the simulated trajectories start to increase the value of the covariate when RUL gets close to 0, and others decrease the value of the covariate when RUL gets close to 0. The variables that follow this behavior are V_9 and V_{14}.

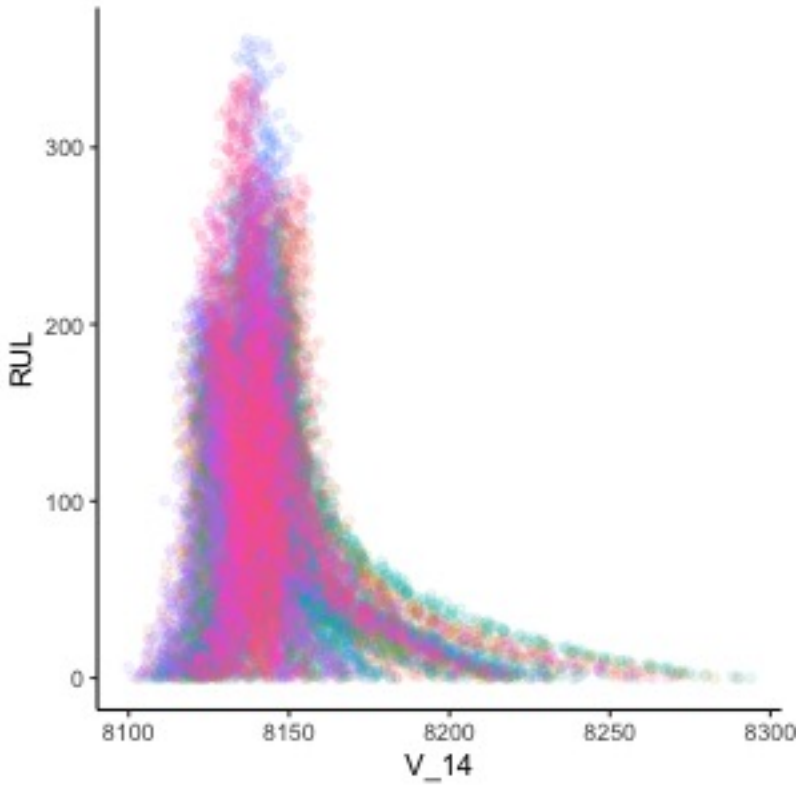


Figure 3.3: Variable V_{14} on the x-axis and RUL on the y-axis

3.3 Methodology

Our main objective is to predict the RUL based on the original 21 V_{XX} covariates. The main idea proposed here is to use regression analysis where the RUL is the response variable and the 21 V_{XX} variables are the explanatory variables. Combined with these regression modeling techniques, three different approaches are used to deal with the nonlinear behavior that the variables exhibit.

3.3.1 Models

We consider four different modeling approaches to analyze the data: multiple linear regression, linear random effects models, generalized linear regression models, and random forests.

3.3.1.1 Multiple Linear Regression

Multiple linear regression (see McCulloch & Searle 2001) is a technique that allows prediction of a response variable based on a linear function of a group of explanatory variables. This method is described by Equation 3.1. In this general framework, the response variable is considered a linear combination of k explanatory variables and some independent error. The response variable is the matrix $\mathbf{Y}$ of $n \times 1$ dimensions, and the explanatory variable is described by $\mathbf{X}$, which has $n \times k$ dimensions. Each column is a different variable, the weight of each variable is described by β, which is of dimension $k \times 1$. The vector $\boldsymbol{\epsilon}$, which is of dimension $n \times 1$, is the matrix of errors. The errors are considered to be random with $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, where $\mathbf{I}$ is the identity matrix.

$$\mathbf{Y} = \beta\mathbf{X} + \boldsymbol{\epsilon} \tag{3.1}$$

3.3.1.2 Random Effects Models

The random effects model (see McCulloch & Searle 2001) is an extension of multiple linear regression. Just like subsection 3.3.1.1, the dependent variable $\mathbf{Y}$ is represented

by a linear combination of k independent variables. The main difference is that the observations are not conditionally independent of each other given the covariates, due to the fact that multiple observations come from the same individual. The way to deal with this dependence is by considering that given the individual, the observations are independent. This allows different individuals to have different behaviors, as exemplified by equation 3.2. This shows the model for a single individual i and observation j. In our case, each individual i is one simulated trajectory and each observation j is a 22-dimensional vector, where the response is the RUL and the 21 V_{XX} covariates are the independent variables. In this case, for example, we can assume that the intercept and X_3 are considered to be dependent on the individual. The intercept is then written as $\beta_0 + u_{0i}$ and the coefficient for V_3 is $\beta_3 + u_{3i}$. In this way, each trajectory will have its own intercept and its own slope for V_3. The full model is then written as

$$Y_{ij} = \beta_0 + u_{0i} + \beta_1 X_{1ij} + \beta_2 X_{2ij} + (\beta_3 + u_{3i}) X_{3ij} + ... + \beta_k X_{kij} + \epsilon_{ij}, \qquad (3.2)$$

where again all of the ϵ_{ij} are independent.

In this framework there are n total observations, I different trajectories, J_i observations per trajectory, and k independent variables. The general framework of regression with random effects is represented in equation 3.3. $\mathbf{Y}$, $\mathbf{X}$, $\boldsymbol{\beta}$, and $\boldsymbol{\epsilon}$ are the same as subsection 3.3.1.1. The coefficient vector $\boldsymbol{\beta}$ is called the fixed effect, and is common to

all individuals. The random effects come from $\mathbf{Zu}$. In our case, $\mathbf{Z}$ is block diagonal matrix, where each block is defined according to which variables are considered to have a different behavior for different individuals, and $u_{ik} \sim N(0, \sigma_k^2)$.

In matrix form, the random effects model is written

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Zu} + \boldsymbol{\epsilon} \tag{3.3}$$

We propose a random effects model like this because looking at figures 3.1, 3.2, and 3.3, we can see that different trajectories present different slopes after a certain point.

3.3.1.3 Generalized Linear Models

Just like section 3.3.1.2, generalized linear models (glm) (see Fahrmeir et al. 2013, Chapter-5) build on the idea of multiple linear regression. In the case of glm, the response variable Y will be connected to the explanatory variables through a nonlinear model. In a general glm framework we have that the response variable will follow some distribution F and will depend on some parameter θ_i. The expected value of the response is written as a transformation of a linear combination of the explanatory variables.

$$Y_i \sim F(\theta_i) \tag{3.4}$$

$$E(Y_i) = h(\theta_i) = g^{-1}(\nu)$$

$$\nu = \mathbf{X}_i\boldsymbol{\beta}, \tag{3.5}$$

where $g(\cdot)$ is called a link function.

In our case, the response is the number of days until failure, which take values in the non-negative integers. A standard distribution that takes values in the non-negative integers is the Poisson distribution. The Poisson distribution depends on a single parameter λ, which can assume values in the positive real line. In this case, our model is written as:

$$Y_i \sim Poisson(\lambda_i) \tag{3.6}$$

$$E(Y_i) = \lambda_i = \exp(\nu_i), \tag{3.7}$$

$$\nu_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_k X_{ik}, \tag{3.8}$$

where the distribution of Y_i is Poisson and the link function is the exponential function.

3.3.1.4 Random Forest

The random forest algorithm (Breiman 2001) is a extension of regression tree learning (Breiman et al. 1984). A regression tree is a data mining algorithm that recursively creates partitions of the data based on the values of the explanatory variables. After the partitions are created, the observations that are grouped together are called the leaves, and the average value of the response variable is considered the predicted value of the leaf. Each partition is created with the objective of minimizing the sum of squared errors, i.e. the squared difference between the predicted value and response variable.

The random forest will take into account multiple regression trees created using multiple random samples from the original data. At each candidate split, a random subset of the features is used. This consists of selecting a smaller random sample of the explanatory variables to create each of the individuals regression trees. The final predicted value of each observation is the average of the predicted value of each tree.

3.3.2 Dealing with Non-Linearity

We discussed in section 3.2 that variables in categories 2 to 4 do not have a linear relationship between the variables and the response variable, RUL. Models like multiple linear regression and random effects models assume linearity between explanatory and response variables. In order to deal with this non-linear behavior, we proposed 3 different approaches.

3.3.2.1 Log Transform

The first approach consists of taking the log of the response variable. Figure 3.1 shows the relationship between the variables, with the remaining life time on the y-axis and V_2 on the x-axis. We can see that there seems to be a exponential relationship between the variables, so taking the log is one way to turn it into a linear relationship. For models 3.3.1.1, 3.3.1.2, and 3.3.1.4 in place of using Y_i we would be using $\log(Y_i)$

3.3.2.2 Splines

The second approach consists of using splines (James et al. 2014, Chapter-7). Splines allow us to approximate a non-linear relationship with a basis expansion of the original variables. In place of using X as we would in a linear model, splines use a set of variables calculated from X, $b_1(X), b_2(X), ..., b_K(X)$. Instead of fitting a model for Y based on X we fit the model

$$Y_i = \beta_0 + \beta_1 b_1(X_i) + \beta_2 b_2(X_i) + ... + \beta_K b_K(X_i) + \epsilon_i, \tag{3.9}$$

which is just a linear model in the expanded basis. Several basis expansions are available, producing different behaviors of the response (e.g. different degrees of smoothness, monotone behavior, periodic behavior, etc.). The basis selected for this study was a B-spline basis for polynomial splines(De Boor 2001). This allowed us to fit piecewise third degree polynomials.

3.3.2.3 Change Point

The third approach consists of using a change point method together with the original variables. From figures 3.1, 3.2, and 3.3 we can see that the variables are roughly constant until one point and start to behave differently after that. The main idea of using a change point is to first detect where the change occurs, and then use this information to change the linear relationship between the explanatory and response variable. This change in relationship is achieved by including an interaction component between

the original variables and a new variable describing when the change in behaviour happened.

In the change point model, each of the original terms $\beta_k X_{ik}$ are replaced by $\beta_k X_{ik} + \beta_{k*} X_{ik*} + \beta^*_{k*} X_{ik} X_{ik*}$, where X_{ik*} is a indicate variable where $X_{ik*} = 0$ if it is before the change point and $X_{ik*} = 1$ if it is after the change point. This means that $\beta_k X_{ik} + \beta_{k*} X_{ik*} + \beta^*_{k*} X_{ik} X_{ik*} = \beta_k X_{ik}$ before the change point and $\beta_k X_{ik} + \beta_{k*} X_{ik*} + \beta^*_{k*} X_{ik} X_{ik*}$ after that. This allows the slope of the variable change from β_k to $\beta_k + \beta^*_{k*}$ which captures the change in behavior.

3.3.2.3.1 At Most One Change In order to detect the change point in the data we used the "At Most One Change" (AMOC) approach (Hinkley 1970). This method consists of estimating the point in the time series where the mean before that point is different from the mean after. It is assumed that the change in mean occurs at time point τ_{XX} and an observation $V_{t,XX}$ can be written using equation 3.10.

$$V_{t,XX} = \begin{cases} \theta_{0,XX} + \epsilon_t & \text{if } t \leq \tau_{XX} \\ \theta_{1,XX} + \epsilon_t & \text{if } t > \tau_{XX} \end{cases} \tag{3.10}$$

Here each observation will have a different constant depending on whether the observation is before of after the change point, and $\epsilon \sim N(0, \sigma^2)$. In this case τ, $\theta_{0,XX}$, and $\theta_{1,XX}$ are estimated by maximizing the log likelihood defined by 3.11.

$$L(\theta_{0,XX}, \theta_{1,XX}, \tau_{XX}) = -\frac{1}{2}\{\sum_{i=1}^{\tau_{XX}} (V_{i,XX} - \theta_{0,XX})^2 + \sum_{i=\tau+1}^{T} (V_{i,XX} - \theta_{1,XX})^2\} \tag{3.11}$$

3.4 Analysis of data set FD001

The data analysis consists of fitting the four different models: linear model, random effects model, generalized linear model, and random forest, together with the 3 different approaches to dealing with the non-linearity, totaling 12 different models.

The variables were not used in their original form, due to the difference in magnitude of the different variables. Rather, each variable was re-scaled to have mean 0 and variance 1. This allows us to have all variables in the same scale, avoiding problems with differences in magnitude. The mean and variance of each variable was calculated using the first 50 values of the simulation, before the changing point. These values were estimated using the available data, however they can alternatively be estimated based on knowledge of the system; the engineer responsible for installing the system can determinate what would be the correct mean and variance when the system is working correctly.

In addition, we applied another transformation to the variables in the fourth category (those that behaved like Figure 3.3). For those variables we decided to not only re-scale, but also to take the absolute value. This is done with the belief that it does not matter in which direction the original variable is deviating, it only matters that it is deviating from its normal state. The absolute value would allow us to make all deviations from the usual behaviour to be only positive.

In order to compare the different models and different approaches to non-linearity, we used 5-fold cross validation and compared the mean squared error (mse) using the

Approach	RUL	LM	Random Effects	GLM	Random Forest
Log	All	2297.7557	3274.8766	2096.0180	1982.9519
	100 Days	1094.9692	1408.1689	1169.8653	841.6268
	50 Days	270.3813	219.4846	472.1143	367.8810
Splines	All	2008.1313	2436.3308	1965.5309	1936.6271
	100 Days	1293.0918	1259.6572	1193.8300	1253.4910
	50 Days	725.4347	471.0816	541.9237	544.7049
Change Point	All	1757.4260	2512.4551	1718.8193	286.3189
	100 Days	1097.9798	1363.0753	1007.4822	167.5617
	50 Days	323.5048	404.5767	212.7128	22.2637

Table 3.2: Mean squared error comparison between all combinations of modeling methods and different approaches to deal with non-linearity. The comparison is made in 3 situations: looking at the entire dataset, restricting to when the remaining life time is smaller than 100 days, and restricting to when the remaining life time is smaller than 50 days.

entire dataset, the mse restricted to when the remaining life time is smaller than 100 days, and the mse restricted to when the remaining life time is smaller than 50 days. Table 3.2 shows the estimated mse for all situations. The mse in these 3 situations indicated that our model accuracy is greater for lower values of RUL. This is due to fact that there is no relationship between RUL and other variables for large values of RUL, since the RUL is decreasing but the values of the variables are mostly constant. We can see dependence between RUL and the other variables only when the system values starts to deviate from the constant behavior.

Among the different models and approaches, the best combination was the change point paired with the random forest. This had the smallest mse, 22.2637. The second best combination was using change point paired with Poisson glm, with an mse of 212.7128. This combination had a higher mse compared to the best, however the computational time required to fit a random forest is much higher than glm.

The mse when using the linear model without any of the three approaches is 2140.090 using all observations, 1452.904 for RUL smaller than 100, and 1062.729 for RUL smaller than 50. Random forest and glm had a better result using all non-linearity approaches when compared to a simple linear model. However all models and approaches had better results when looking at RUL smaller than 100 and 50 days than on the full dataset.

3.4.1 Comparing with Previous Works

Gugulothu et al. (2017) compared the performance of a model they proposed, a recurrent Neural Network to generate embeddings for multivariate time series, against 9 other published models. In order to compare these different methods they used two measures: the mean square error and the timeliness score (S). The timeliness score is defined as

$$S = \sum_{i=1}^{N} \left(e^{\lambda_i |\hat{R}_i - R_i|} - 1 \right),$$

where $\hat{R}_i$ is the estimated remaining life time, R_i it the actual life time, and λ_1 is equal to $1/r_1$ if $\hat{R}_i - R_i < 0$ and $1/r_2$ otherwise. In order to penalize late predictions, r_1 is taken to be bigger than r_2. Table 3.3 presents the mse and timeliness score of the previous methods when analysing the test dataset for FD001. From this table their proposed method had better results, obtaining the smallest mse and second smallest timeliness score.

Approach	S	mse
ESN-KF (Peng et al. 2012)	NR	4026
DeepCNN (Sateesh Babu et al. 2016)	1287	340
SOM^3 (Macmann et al. 2016)	NR	297
Shapelet (Khelif et al. 2014)	652	NR
SVR (Khelif et al. 2017)	449	NR
Deep LSTM (Zheng et al. 2017)	338	260
MODBNE (Zhang et al. 2017)	334	226
$LR\text{-}ED_2$ (Malhotra et al. 2016)	256	164
RULCLIPPER (Ramasso 2014)	216	176
$Embed\text{-}LR_1$ (Gugulothu et al. 2017)	219	155

Table 3.3: Performance of various approaches on the NASA C-MAPSS. Reprinted from Gugulothu et al. (2017).

Approach	RUL	LM	Random Effects	GLM	Random Forest
Change Point	All	1069	1632	1099	1060
	100 Days	1031	1102	1003	1003
	50 Days	258	475	190	122

Table 3.4: Mean squared error of the test dataset from FD001. The comparison is made in 2 situations, looking at the entire dataset, only when the remaining life time is smaller than 100 days, and only when the remaining life time is smaller than 50 days.

In order to compare our analysis with previous work, we trained the change point method using the training set and calculated the mse of the training set. The mse can be seen on table 3.4. When looking at the entire test set, the combination of modeling and change point had better result than the ESN-KF model from Peng et al. (2012), but worse results than the other models. When looking at only remaining time smaller than 50 days, our model becomes more accurate, and the mse is smaller than the previous models. We do not have access to the performance of the previous models for different conditions, so we were not able to compare the performance of this approach for different values of RUL against the other models.

3.5 Conclusion

The main objective of this study was to propose methods that would allow us to estimate the remaining life time of a system based on the system variables only, not using time as a covariate. We proposed four different models with three possible approaches to non-linearity for each. When looking at our region of interest, i.e. lower values of remaining life time, all methods had better results than simply using linear regression.

The simplest method consisted in taking the log of the response variable. The use of the natural logarithm function seemed to improve our prediction compared to simple linear model. This improvement can be seen when the remaining life time is 100 days or less.

The use of splines showed a good improvement for all models compared to the original linear model. This method does not require any estimation as a pre-processing step, which makes it an attractive method compared to change point analysis.

The last approach, the change point analysis, was the one with generally better results compared to the other ones. Being able to include the point where the behavior changes in the explanatory variable greatly improves predictions. One problem with this method is that it requires an estimate of the location of the change point. This can be done in our training data, but it can take a long time to run in real life since it will need enough observations after the change point to estimate it. The change point detection requires enough data to be collected after the change point before the

difference in means can be recognized.

When comparing with previous works, the change point method had better results than ESN-KF method from Peng et al. (2012). However our method has better results when the remaining life time is low. We cannot compare the result for specific ranges of remaining life time, but when the remaining life time is smaller than 50 days our method had better results than the previous one.

One advantage our proposed method has over previous work is the easy interpretation. The linear regression, random effects, and generalized linear models are easily interpretable and can be easily explained to people responsible for evaluating the condition of the engines. The previous methods used neural networks, (Gugulothu et al. 2017, Macmann et al. 2016, Sateesh Babu et al. 2016, Zheng et al. 2017), similarity instance (Khelif et al. 2014), support vector regression (Khelif et al. 2017), or deep belief networks (Zhang et al. 2017). While these models can accurately predict the remaining life time, they are complex models with difficult interpretation.

Appendix

Parameter transformation

The main goal of the transformation proposed by Sabourin & Naveau (2014) is to remove the constraint (1.5) which induces strong prior dependence in the parameters and causes disastrously poor MCMC mixing. After transformation, new parameter space becomes a rectangular subset of $\mathbb{S}_d^{k-1} \times (0,1)^{K-1} \times (\mathbb{R}^+)^K$, removing the prior dependence and dramatically improving MCMC mixing. Intuitively, we replace the weight vector $\mathbf{p}$ and the last mean vector $\boldsymbol{\mu}_K$ by new parameters $\epsilon_1, \ldots, \epsilon_{K-1}$. Where each $\epsilon_m \in (0,1)$ is referred to as an *eccentricity*, defined to indicate departure from centrality induced by decreasing the subsets of mixture components. The idea is that we decrease the number of parameters by one, with the decreased degrees of freedom playing the role of the constraint in the original parameterization. The transformation is defined by a series of recursive equations.

First, given the original parameters $p_1, \ldots, p_K$ and $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K$, each eccentricity

parameter ϵ_m is defined as

$$\epsilon_m = \frac{\|\boldsymbol{\gamma}_m - \boldsymbol{\gamma}_{m-1}\|}{\|\boldsymbol{I}_m - \boldsymbol{\gamma}_{m-1}\|}. \tag{.1}$$

This transformation uses intermediate variables $\boldsymbol{I}_m$, $\boldsymbol{\gamma}_m$, and $\boldsymbol{\gamma}_{m-1}$. Each $\boldsymbol{\gamma}_m$ can be interpreted as the the center of mass of the mth mixture component. Each $\boldsymbol{\gamma}_m$ is then defined as

$$\boldsymbol{\gamma}_m = \rho_m^{-1} \sum_{j=m+1}^{K} p_j \boldsymbol{\mu}_j, \tag{.2}$$

which depends on a collection of variables $\rho_0, \ldots, \rho_{K-1}$. The first $\rho_0 = 1$ and the remaining $\rho_1, \ldots, \rho_{K-1}$ are defined recursively as

$$\rho_m = \rho_{m-1} - p_m. \tag{.3}$$

Finally, each $\mathbf{I}_m$ in (.1) is set according to

$$\mathbf{I}_m = \boldsymbol{\gamma}_{m-1} + T_m(\boldsymbol{\gamma}_{m-1} - \boldsymbol{\mu}_m), \tag{.4}$$

where $T_m = \sup\{t \geq 0 : \boldsymbol{\gamma}_{m-1} + t(\boldsymbol{\gamma}_{m-1} - \boldsymbol{\mu}_m) \in \mathbb{S}_d\}$

The result is that the transformed parameters live in rectangular region and still result in mixtures that satisfy the moment constraint (1.3). See Sabourin & Naveau (2014) for additional details.

Bibliography

Arellano-Valle, R. B. & Azzalini, A. (2006), 'On the unification of families of skew-normal distributions', *Scand. J. Statist.* **33**(3), 561–574.
URL: *https://doi.org/10.1111/j.1467-9469.2006.00503.x*

Boldi, M.-O. & Davison, A. C. (2007), 'A mixture model for multivariate extremes', *J. R. Stat. Soc. Ser. B Stat. Methodol.* **69**(2), 217–229.
URL: *https://doi.org/10.1111/j.1467-9868.2007.00585.x*

Bowman, D. M. J. S., Balch, J. K., Artaxo, P., Bond, W. J., Carlson, J. M., Cochrane, M. A., D'Antonio, C. M., DeFries, R. S., Doyle, J. C., Harrison, S. P., Johnston, F. H., Keeley, J. E., Krawchuk, M. A., Kull, C. A., Marston, J. B., Moritz, M. A., Prentice, I. C., Roos, C. I., Scott, A. C., Swetnam, T. W., van der Werf, G. R. & Pyne, S. J. (2009), 'Fire in the earth system', *Science* **324**(5926), 481–484.

Breiman, L. (2001), 'Random forests', *Machine Learning* **45**(1), 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. (1984), *Classification and regression trees*, Wadsworth Statistics/Probability Series, Wadsworth Advanced Books and Software, Belmont, CA.

CalFire (2015), 2015 wildfire activity statistics, Technical report, CA: California Department of Forestry and Fire Prevention and Office of the State Fire Marshal.

Cao, J., Genton, M. G., Keyes, D. E. & Turkiyyah, G. M. (2019), 'Hierarchical-block conditioning approximations for high-dimensional multivariate normal probabilities', *Stat. Comput.* **29**(3), 585–598.
URL: *https://doi.org/10.1007/s11222-018-9825-3*

Coles, S. (2001), *An introduction to statistical modeling of extreme values*, Springer Series in Statistics, Springer-Verlag London, Ltd., London.
URL: *https://doi.org/10.1007/978-1-4471-3675-0*

Coles, S. G. & Tawn, J. A. (1991), 'Modelling extreme multivariate events', *J. Roy. Statist. Soc. Ser. B* **53**(2), 377–392.
URL: *http://links.jstor.org/sici?sici=0035-9246(1991)53:2<377:MEME>2.0.CO;2-4origin=MSN*

Cooley, D., Davis, R. A. & Naveau, P. (2010), 'The pairwise beta distribution: a flexible parametric multivariate model for extremes', *J. Multivariate Anal.* **101**(9), 2103–2117.
URL: *https://doi.org/10.1016/j.jmva.2010.04.007*

Cooley, D., Davis, R. A. & Naveau, P. (2012), 'Approximating the conditional density given large observed values via a multivariate extremes framework, with application to environmental data', *Ann. Appl. Stat.* **6**(4), 1406–1429.
URL: *https://doi.org/10.1214/12-AOAS554*

Cooley, D., Nychka, D. & Naveau, P. (2007), 'Bayesian spatial modeling of extreme precipitation return levels', *J. Amer. Statist. Assoc.* **102**(479), 824–840.
URL: *https://doi.org/10.1198/016214506000000780*

Davison, A. C., Huser, R. & Thibaud, E. (2013), 'Geostatistics of dependent and asymptotically independent extremes', *Math. Geosci.* **45**(5), 511–529.
URL: *https://doi.org/10.1007/s11004-013-9469-y*

De Boor, C. (2001), *A practical guide to splines: with 32 figures*, Vol. 27;27.;, rev. edn, Springer, New York.

de Carvalho, M., Oumow, B., Segers, J. & Warchoł, M. (2013), 'A Euclidean likelihood estimator for bivariate tail dependence', *Comm. Statist. Theory Methods* **42**(7), 1176–1192.
URL: *https://doi.org/10.1080/03610926.2012.709905*

de Fondeville, R. & Belzile, L. (2018), *mvPot: Multivariate Peaks-over-Threshold Modelling for Spatial Extreme Events*. R package version 0.1.4.
URL: *https://CRAN.R-project.org/package=mvPot*

de Fondeville, R. & Davison, A. C. (2018), 'High-dimensional peaks-over-threshold inference', *Biometrika* **105**(3), 575–592.
URL: *https://doi.org/10.1093/biomet/asy026*

Einmahl, J. H. J., de Haan, L. & Piterbarg, V. I. (2001), 'Nonparametric estimation of the spectral measure of an extreme value distribution', *Ann. Statist.* **29**(5), 1401–1423.
URL: *https://doi.org/10.1214/aos/1013203459*

Einmahl, J. H. J. & Segers, J. (2009), 'Maximum empirical likelihood estimation of the spectral measure of an extreme-value distribution', *Ann. Statist.* **37**(5B), 2953–2989.
URL: *https://doi.org/10.1214/08-AOS677*

Engelke, S., Opitz, T. & Wadsworth, J. (2019), 'Extremal dependence of random scale constructions', *Extremes* **22**(4), 623–666.
URL: *https://doi.org/10.1007/s10687-019-00353-3*

Fahrmeir, L., Kneib, T., Lang, S., Marx, B. & service), S. O. (2013), *Regression: Models, Methods and Applications*, Springer Berlin Heidelberg, Berlin, Heidelberg.

Fosberg, M. A. (1978), Weather in wildland fire management: the fire weather index, *in* 'Proceedings of the Conference on Sierra Nevada Meteorology', American Meteorological Society, Lake Tahoe, California, USA, pp. 1–4.

Genton, M. G., Keyes, D. E. & Turkiyyah, G. (2018), 'Hierarchical decompositions for the computation of high-dimensional multivariate normal probabilities', *J. Comput. Graph. Statist.* **27**(2), 268–277.
URL: *https://doi.org/10.1080/10618600.2017.1375936*

Genton, M. G., Padoan, S. A. & Sang, H. (2015), 'Multivariate max-stable spatial processes', *Biometrika* **102**(1), 215–230.
URL: *https://doi.org/10.1093/biomet/asu066*

Genz, A. (1992), 'Numerical computation of multivariate normal probabilities', *Journal of Computational and Graphical Statistics* **1**(2), 141–149.

Genz, A. (2004), 'Numerical computation of rectangular bivariate and trivariate normal and t probabilities', *Stat. Comput.* **14**(3), 251–260.
URL: *https://doi.org/10.1023/B:STCO.0000035304.20635.31*

Genz, A. & Bretz, F. (2009), *Computation of multivariate normal and t probabilities*, Vol. 195 of *Lecture Notes in Statistics*, Springer, Dordrecht.
URL: *https://doi.org/10.1007/978-3-642-01689-9*

Goodrick, S. L. (2002), 'Modification of the fosberg fire weather index to include drought', *International Journal of Wildland Fire* **11**(4), 205–211.

Gugulothu, N., TV, V., Malhotra, P., Vig, L., Agarwal, P. & Shroff, G. (2017), 'Predicting remaining useful life using time series embeddings based on recurrent neural networks'.

Guillotte, S., Perron, F. & Segers, J. (2011), 'Non-parametric Bayesian inference on bivariate extremes', *J. R. Stat. Soc. Ser. B Stat. Methodol.* **73**(3), 377–406.
URL: *https://doi.org/10.1111/j.1467-9868.2010.00770.x*

Guinness, J. (2018), 'Permutation and grouping methods for sharpening Gaussian process approximations', *Technometrics* **60**(4), 415–429.
URL: *https://doi.org/10.1080/00401706.2018.1437476*

Hackbusch, W. (2015), *Hierarchical matrices: algorithms and analysis*, Vol. 49 of *Springer Series in Computational Mathematics*, Springer, Heidelberg.
URL: *https://doi.org/10.1007/978-3-662-47324-5*

Hinkley, D. V. (1970), 'Inference about the change-point in a sequence of random variables', *Biometrika* **57**(1), 1–17.

Huser, R., Davison, A. C. & Genton, M. G. (2016), 'Likelihood estimators for multivariate extremes', *Extremes* **19**(1), 79–103.
URL: *https://doi.org/10.1007/s10687-015-0230-4*

Huser, R., Opitz, T. & Thibaud, E. (2017), 'Bridging asymptotic independence and dependence in spatial extremes using Gaussian scale mixtures', *Spat. Stat.* **21**(part A), 166–186.
URL: *https://doi.org/10.1016/j.spasta.2017.06.004*

Huser, R. & Wadsworth, J. L. (2019), 'Modeling spatial processes with unknown extremal dependence class', *J. Amer. Statist. Assoc.* **114**(525), 434–444.
URL: *https://doi.org/10.1080/01621459.2017.1411813*

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2014), *An introduction to statistical learning: with applications in R*, Vol. 103, correct at 4 printing 2014. edn, Springer, New York.

Khelif, R., Chebel-Morello, B., Malinowski, S., Laajili, E., Fnaiech, F. & Zerhouni, N. (2017), 'Direct remaining useful life estimation based on support vector regression', *IEEE Transactions on Industrial Electronics* **64**(3), 2276–2285.

Khelif, R., Malinowski, S., Chebel-Morello, B. & Zerhouni, N. (2014), Rul prediction based on a new similarity-instance based approach, IEEE, pp. 2463–2468.

Lock, E. F. & Dunson, D. B. (2015), 'Shared kernel Bayesian screening', *Biometrika* **102**(4), 829–842.
URL: *https://doi.org/10.1093/biomet/asv032*

Macmann, O. B., Seitz, T. M., Behbahani, A. R. & Cohen, K. (2016), *Performing Diagnostics Prognostics On Simulated Engine Failures Using Neural Networks*.

Malhotra, P., TV, V., Ramakrishnan, A., Anand, G., Vig, L., Agarwal, P. & Shroff, G. (2016), 'Multi-sensor prognostics using an unsupervised health index based on lstm encoder-decoder'.

Marcon, G., Padoan, S. A., Naveau, P., Muliere, P. & Segers, J. (2017), 'Multivariate nonparametric estimation of the Pickands dependence function using Bernstein polynomials', *J. Statist. Plann. Inference* **183**, 1–17.
URL: *https://doi.org/10.1016/j.jspi.2016.10.004*

McCulloch, C. E. & Searle, S. R. (2001), *Generalized, linear, and mixed models*, Wiley Series in Probability and Statistics: Texts, References, and Pocketbooks Section, Wiley-Interscience [John Wiley & Sons], New York.

Miwa, T., Hayter, A. J. & Kuriki, S. (2003), 'The evaluation of general non-centred orthant probabilities', *J. R. Stat. Soc. Ser. B Stat. Methodol.* **65**(1), 223–234.
URL: *https://doi.org/10.1111/1467-9868.00382*

Moritz, M. A., Moody, T. J., Krawchuk, M. A., Hughes, M. & Hall, A. (2010), 'Spatial variation in extreme winds predicts large wildfire locations in chaparral ecosystems', *Geophysical Research Letters* **37**(4), n/a.

Moritz, M. A., Morais, M. E., Summerell, L. A., Carlson, J. & Doyle, J. (2005), 'Wildfires, complexity, and highly optimized tolerance', *Proc. Nat. Acad. Science* **102**(50), 17912–17917.
URL: *https://doi.org/10.1073/pnas.0508985102*

Morris, S. A., Riech, B. J., Thibaud, E. & Cooley, D. (2017), 'A space-time skew-t model for threshold exceedances', *Biometrics* **73**(3), 749–758.
URL: *https://doi.org/10.1111/biom.12644*

Opitz, T. (2016), 'Modeling asymptotically independent spatial extremes based on Laplace random fields', *Spat. Stat.* **16**, 1–18.
URL: *https://doi.org/10.1016/j.spasta.2016.01.001*

Peng, Y., Wang, H., Wang, J., Liu, D. & Peng, X. (2012), A modified echo state network based remaining useful life estimation approach, IEEE, pp. 1–7.

Ramasso, E. (2014), 'Investigating computational geometry for failure prognostics', *International Journal of Prognostics and Health Management* **5**(5), 1–18.

Reich, B. J. & Shaby, B. A. (2018), 'Modeling of multivariate spatial extremes', *RESEARCHERS.ONE* .
URL: *https://www.researchers.one/article/2018-09-12*

Resnick, S. I. (1987), *Extreme values, regular variation, and point processes*, Vol. 4 of *Applied Probability. A Series of the Applied Probability Trust*, Springer-Verlag, New York.
URL: *https://doi.org/10.1007/978-0-387-75953-1*

Ribatet, M. (2018), *SpatialExtremes: Modelling Spatial Extremes*. R package version 2.0-7.
URL: *https://CRAN.R-project.org/package=SpatialExtremes*

Sabourin, A. (2015), 'Semi-parametric modeling of excesses above high multivariate thresholds with censored data', *J. Multivariate Anal.* **136**, 126–146.
URL: *https://doi.org/10.1016/j.jmva.2015.01.014*

Sabourin, A. & Naveau, P. (2014), 'Bayesian Dirichlet mixture model for multivariate extremes: a re-parametrization', *Comput. Statist. Data Anal.* **71**, 542–567.
URL: *https://doi.org/10.1016/j.csda.2013.04.021*

Sateesh Babu, G., Zhao, P. & Li, X.-L. (2016), Deep convolutional neural network based regression approach for estimation of remaining useful life, *in* S. B. Navathe, W. Wu, S. Shekhar, X. Du, X. S. Wang & H. Xiong, eds, 'Database Systems for Advanced Applications', Springer International Publishing, Cham, pp. 214–228.

Stein, M. L., Chi, Z. & Welty, L. J. (2004), 'Approximating likelihoods for large spatial data sets', *J. R. Stat. Soc. Ser. B Stat. Methodol.* **66**(2), 275–296.

Thibaud, E., Aalto, J., Cooley, D. S., Davison, A. C. & Heikkinen, J. (2016), 'Bayesian inference for the Brown-Resnick process, with an application to extreme low temperatures', *Ann. Appl. Stat.* **10**(4), 2303–2324.
URL: *https://doi.org/10.1214/16-AOAS980*

Vecchia, A. V. (1988), 'Estimation and model identification for continuous spatial processes', *Journal of the Royal Statistical Society. Series B (Methodological)* **50**(2), 297–312.

Vettori, S., Huser, R., Segers, J. & Genton, M. G. (2017), 'Bayesian model averaging over tree-based dependence structures for multivariate extremes', *arXiv preprint arXiv:1705.10488* .

Wadsworth, J. L. & Tawn, J. A. (2014), 'Efficient inference for spatial extreme value processes associated to log-Gaussian random functions', *Biometrika* **101**(1), 1–15.
URL: *https://doi.org/10.1093/biomet/ast042*

Westerling, A. L., Cayan, D. R., Brown, T. J., Hall, B. L. & Riddle, L. G. (2004), 'Climate, santa ana winds and autumn wildfires in southern california', *Eos, Transactions American Geophysical Union* **85**(31), 289.
URL: *https://doi.org/10.1029/2004EO310001*

Zhang, C., Lim, P., Qin, A. K. & Tan, K. C. (2017), 'Multiobjective deep belief networks ensemble for remaining useful life estimation in prognostics', *IEEE Transactions on Neural Networks and Learning Systems* **28**(10), 2306–2318.

Zheng, S., Ristovski, K., Farahat, A. & Gupta, C. (2017), Long short-term memory network for remaining useful life estimation, IEEE, pp. 88–95.

Vita

Mauricio Fernandes do Nascimento Junior

EDUCATION

Ph.D. Statistics (2020) The Pennsylvania State University, University Park, PA **Bachelor of Statistics** (2012) Federal University of Parana, Curitiba, PR, Brazil

TEACHING EXPERIENCE

Primary Instructor

- Applied Statistics in Science (2019) Pennsylvania State University
- Elementary Statistics Online (2018) Pennsylvania State University
- Statistical Concepts and Reasoning (2017) Pennsylvania State University
- Elementary Statistics (2016) Pennsylvania State University

Graduate Assistant

- Elementary Statistics (2016,2018) Pennsylvania State University
- Introduction to Mathematical Statistics (2014) Pennsylvania State University
- Applied Time Series Analysis (2015,2016,2017,2019) Pennsylvania State University
- Regression Methods (2017) Pennsylvania State University
- Applied Time Series Analysis (2015) Pennsylvania State University

AWARDS

- 2018 Award for support of pedagogy in undergraduate instruction

PUBLICATIONS

- Mauricio Nascimento and Benjamin A. Shaby. Spatial semi-parametric spectral density estimation for multivariate extremes, with application to fire threat. Journal of Environmental Statistics, 9(3), 9 2019. ISSN 1945-1296.
- Yung-Chen Jen Chiu, Hsiao-Ying Vicki Chang, Ann Johnston, Mauricio Nascimento, James T. Herbert and Xiaoyue Maggie Niu. Impact of Disability Services on Academic Achievement among College Students with Disabilities. Journal of Postsecondary Education and Disability 32 (3), 2019.

WORKSHOPS PRESENTED

- STAT 592 Teaching Statistics panel (2017, 2019) Pennsylvania State University
- Use of GGPLOT (2018, 2019) Pennsylvania State University
- Estatística e a Pesquisa Científica (2011) 10^{th} Encontro de Extensão e Cultura da Semana Integrada de Ensino, Pesquisa e Extensào