Theory and Methods for Large Spatial Data

by

Z. Mullen

B.A., Pomona College, 2010

M.S., University of Colorado, 2015

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Applied Mathematics

2018

This thesis entitled:
Theory and Methods for Large Spatial Data
written by Z. Mullen
has been approved for the Department of Applied Mathematics

_____

Prof. William Kleiber

_____

Prof. Balaji Rajagopalan

_____

Prof. Eric Vance

Date _____

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

Mullen, Z. (Ph.D., Applied Mathematics)

Theory and Methods for Large Spatial Data

Thesis directed by Prof. William Kleiber

Correlated Gaussian processes are of central importance to the study of time series, spatial statistics, computer experiments, and many machine learning models. Large spatially or temporally indexed datasets bring with them a host of computational and mathematical challenges.

Parameter estimation of these processes often relies on maximum likelihood, which for Gaussian processes involves manipulations of the covariance matrix including solving systems of equations and determinant calculations. The score function, on the other hand, avoids direct calculation of the determinant, but still requires solving a large number of linear equations. We propose an equivalent kernel approximation to the score function of a stationary Gaussian process. A nugget effect is required for the approximation. We suggest two approximations, and for large sample sizes, our proposals are fast, accurate, and compare well against existing approaches.

We then present a method for simulating time series of high frequency wind data calibrated by real data. The method provides and fits a parametric model for local wind directions by embedding them into the angular projection of a bivariate normal. Incorporating a temporal autocorrelation structure in that normal induces a continuous angular correlation over time in the simulated wind directions. The final joint model for speed and direction can be decomposed into the simulation of a single multivariate normal and a series of transformations thereof, allowing for fast and easy repeated generations of long time series. This is compared to a state of the art approach for simulating angular time series of swapping between discrete regimes of wind direction, a method that does not fully translate to high frequency data.

Acknowledgements

Thanks to my advisor, Will Kleiber for regular discussion and mentoring.

Thanks to my girlfriend, my family, and all of my friends in the Boulder area for keeping me sane.

Thanks Doug Nychka for helpful discussions during the development of the equivalent kernel work. I was supported by NSF DMS-1406536.

Contents

Chapter

Tables

Table

Figures

Figure

Chapter 1

Introduction

The Gaussian process sits at the center of a large number of techniques and results in statistics, ranging from the Central Limit Theorem to the assumptions underlying ANOVAs and most linear model hypothesis testing. Even as computational power has increased access to many non-parametric methods, the Gaussian still sits at the center of many of these methods. The Gaussian cumulative density function is well enough known and efficient to simulate that Gaussian models are often be adapted to non-Gaussian situations via transformations such as copulas or by incorporating marginal Gaussian distributions in larger joint models.

In machine learning, a variety of methods can be approximated or solved directly using Gaussian processes. For example, estimation techniques in neural networks may tend to Gaussian processes as the model size grows. Clustering and classification algorithms are often couched in terms of underlying Gaussianity for ease of computation and estimation. In many forms of Bayesian estimation on linear models, Gaussian priors on weights lend themselves to Gaussian posteriors [65]. Gaussianity can also be applied in determine experimental design techniques, particularly in computer experiments to explore relationships where a detailed power analysis and/or repeated sampling may be largely infeasible [69].

In the case of environmental and spatial statistics, the Gaussian process is one of the most common ways to model dependence structures that describe covariances between observations. The two most common forms of the Gaussian process in spatial statistics are to model observations as jointly distributed normals with a functional covariance or to consider the observations as arising

from a specific class of stochastic partial differential equation generating Gaussian Markov Random Field. The estimation and specification of continuous covariance functions is the problem solving process followed in both analysis of autoregressive processes in time series and that of continuous spatial variance in classical geostatistics. Much of the literature with roots in the continuous spatial problem discusses kriging estimators, the best linear unbiased predictor of a Gaussian process (both in and out of sample) with a known covariance matrix between points.

The kriging estimator is typically constructed via a parameterized form of the covariance which allows for parametric specification of the shape and rate of correlation decay as observations become further apart. The resulting function - which may or may not vary over space or time - determines the entries of the Gaussian process covariance matrix, with entries for each of the pairs of points in the sample. As data sets grow with the availability of continuous temporal data streams or highly resolved satellite images, the mathematics and computational burdens of these covariances grows.

Given their heavy use, mitigating and understanding the computational costs of Gaussian processes is a growing concern in many of the applications listed. In this thesis, we will explore some of the methods and techniques for Gaussian processes on data sets in the tens of thousands to millions of observations, a scale at loading and manipulating the data itself is within reason for personal computers, but evaluation methods that are forced to populate and invert covariance matrices may not succeed to both memory allocation and computational time constraints. At this size, computational techniques must balance the relative ease of simulation from Gaussian processes with highly structured covariances and the often costlier techniques in estimation and prediction.

1.1 Approximate Score Functions for Gridded Gaussian Processes

The first project undertaken in this thesis provides the underlying theory and methodology for the evaluation of Gaussian process score functions under proper formulation of the covariance matrices. In particular, for covariances of isotropic and stationary processes, an approximation to the inverse of the covariance matrix has been established via the equivalent kernel approximation to

the classical kriging estimator. The equivalent kernel solves a spline smoothing problem via a linear function of data given a pre-specified inner product and general smooth parameter. The weights implied by such a linear combination are equivalent to the linear combination of observations that the kriging estimator predicts as the best linear unbiased predictor for an underlying Gaussian process. Such an approximation has only been explored in the process of smoothing, and has not yet been applied to estimating parameters of the covariance function.

We take the matrix equation implied by the equivalence of the spline smoother and kriging smoother and apply it to the derivative of the likelihood function of the multivariate normal: these score functions of Gaussian processes provide an avenue to find local extrema of the likelihood function, which under many ordinary circumstances will exactly align with the global maximum likelihood. Unlike the kriging estimator, however, the spline solution requires only regular evaluations of either a closed form function in special cases or more generally a numerical integral. As a result, there is no need to compute the matrix inverses within the Gaussian likelihood, and we can realize according savings to both computational memory allotment and processor time.

The resulting method scales to as fast as linearly in the size of the data for the trace of the score function, and with proper use of the fast Fourier transform can scale with an additional logarithmic computational cost on the quadratic form term. This allows for fast approximation maximum likelihood wherever gridded data applies, whether in a two dimensional spatial setting or in time series. In addition, it allows for identical parameter interpretations to those used in smaller sample spatial problems, whereas some alternative approached opt for modifications of the underlying covariance.

We compare our method to the most similar existing method in score function solutions to general Gaussian processes, and conclude that when the data is both of sufficient number of observations and sufficiently noisy (i.e. the nonspatial variation provides a significant component of the overall variation), the method produces to similar maximum likelihood solutions in considerably less time. Our approximation's applicability to the trace term of the score function is particularly favorable, as it generates the largest time savings and appears to apply in smaller sample sizes than

are required for use in the quadratic form. A brief data analysis on a daily time series is included to demonstrate the speed of the full method in one dimension, and simulation studies are included for both one and two dimensional analysis.

1.2 Joint Angular-Speed Simulation for High Frequency Wind Data

The second project herein uses an underlying Gaussian structure to jointly model the non-Gaussian joint structure of wind speed and wind directions. In general, wind modeling is broken up into two objectives: simulation and prediction. Predictive methods include the goal of forecasting, wherein a model is measured by its ability to predict the wind direction and speed given its previous set of states. Such models often include covariates for prediction that bridge the statistical properties of wind observations with the dynamics suggested by underlying meteorological dynamics. In unconditional simulation, the goal is to construct an underlying stochastic process whose realizations mimic the statistical behavior of observed wind fields. Emphasis is placed on the model being generative: it should be repeatable and relatively efficient to sample from, as the simulation approach attempts to answer questions about joint wind densities over space and time via the properties of a large simulated sample. As such, this model will likely include fewer regressive covariates, as any such inclusions would also be required in the generative model.

Naturally, many simulative models incorporate a Gaussian processes. We explore two methods each couched in normality in this chapter. First, we describe a model found in existing literature that creates a discrete set of directional wind regimes, and builds a Markov process with Gaussian autoregressive correlation over time as it switching between these regimes. We compare this model to one of our own construction, which begin with a bivariate normal and projects simulations of it into angles via the inverse tangent of their components. This projected normal is in fact a flexible distribution on the unit circle, and can be used both as a single bivariate normal or a set of mixtures thereof to capture complex behavior on the unit circle. For the purposes of wind models, the projected normal is highly appealing due to its ability to capture asymmetries and up to two directional modes that may or may not be antipodal. Our model fits a projected normal to an observed data

set in Oregon, USA and embeds a temporal autocorrelation structure into the underlying bivariate Gaussian process, allowing the resulting circular process to inherit its own relationship structure from the underlying Gaussian process.

The projected normal allows for the same type of smooth variation over time exhibited in the classical geostatistics problem, whereas the regime-based approach requires some amount of manipulation to balance the added challenges of regime classification. We focus on the case of observations occurring at high-frequency in time here, as it both channels the importance of computational efficiency and may be a shortcoming of the discretization implied by a regime model. In particular, higher frequency data sets will more closely explore the behavior of wind as shifts direction from mode to mode, and the clustering of those observations poses a challenge to the regime-based models that is well handled by the continuous variation of the projected normal.

We then construct a model for wind speed that is motivated again by the ease of Gaussian process simulation. Theoretical concerns with positive definiteness suggest that hierarchical modeling of correlated and cross-correlated random variables like wind speed and direction is challenging. In particular, if we constructed a stochastic model for wind speed fully conditioned on our existent model for direction, the resulting cross-correlation structures would be very unlikely to represent a valid correlation structure for the entire model. If instead a correlated process for wind speed is included in the underlying Gaussian process as a third variable in simulation along with the bivariate components of the projected normal, this concern can be more carefully managed. Our model calls on the intuition of a hierarchical modeling approach, but embeds the final simulation into a compact Gaussian structure. Once a proper full Gaussian process is simulated, we provide a set transformations and projections to reduce the now trivariate time series into one of projected angles and speeds, with our fit for wind speed based on moment-matching the empirical conditional density function for wind speed given wind direction at the data locations.

Chapter 2

Equivalent Kernel Approximations of Gaussian Score Functions

2.1 Motivation and Existing Methods in Massive Spatial Processes

As datasets in remote sensing and global climate model outputs become more prevalent, techniques for dealing with tens of thousands or more correlated data points must be refined and implemented. For example, maximum likelihood estimation for correlated Gaussian processes becomes infeasible on personal computers with more than a few thousand observations without severe modeling restrictions.

When estimating parameters of a correlated Gaussian process, the primary obstacle on a big dataset lies with the covariance matrix. This matrix and its associated memory allocation scale in size with the square of the number of data locations, and a direct implementation of maximum likelihood estimation will scale computationally with the cube of the number of locations. A variety of techniques exist to simplify the computational difficulties of working with the covariance matrix, whose inverse occurs as a quadratic form in the likelihood and score functions of a Gaussian random field and as a linear term in the kriging estimator. Many of these techniques enforce sparsity on the covariance matrix to allow for faster computation with the precision matrix. Covariance tapering [24, 41] demands that a compactly supported covariance function be used. Fixed rank kriging represents the spatial covariance model through a relatively small - compared to the number of data locations - set of basis functions [17], and relies on the Woodbury formula for matrix solves. Multiresolution specifications use a weighted basis function representation that can exploit sparsity [60, 39]. Approximations can also be applied directly to the Cholesky decomposition of

the covariance matrix, as in Vecchia approximations [81, 40]. Another approach is to allow for an arbitrary covariance model, but to exploit applied mathematical ideas for computational savings. For instance, [76] uses conjugate gradient methods and a stochastic Hutchinson trace estimator. More recent ideas consider unbiased estimating equations [78, 6], which can be coupled with fast Fourier methods for quick computation.

In this work, we exploit the notion of an equivalent kernel approximation to the kriging weight function. In particular, it is well known that the solution to a spline smoothing problem is a linear function of data, and whose weights behave like kernel functions [73]. Spline smoothing, on the other hand, is equivalent to kriging under a variational formulation [82]. Some recent work has demonstrated the computational efficiency of equivalent kriging [47], and the idea has previously been used to examine asymptotic convergence rates for smoothing under kriging [23]. Such an approximation has never been adapted to estimating parameters of the covariance function in a maximum likelihood framework, however, which we explore herein.

2.1.1 Score Functions

Given observations $\mathbf{Y} = [Y(s_1), Y(s_2), \ldots, Y(s_n)]^T$ at set of n locations $s \in \mathbb{R}^d$, we consider a model of the form

$$Y(s) = Z(s) + \varepsilon(s) \tag{2.1}$$

where $Z(s)$ and $\varepsilon(s)$ are mean zero correlated and uncorrelated (respectively) Gaussian processes. Let $\text{Var}(\varepsilon(s)) = \tau^2$ denote the nugget effect, and let the covariance function of $Z(s)$ be denoted by $\text{Cov}(Z(s_1), Z(s_2)) = k(s_1, s_2)$ where k depends on a set of parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^{\mathrm{T}} \in \mathbb{R}^m$.

The classical geostatistics problem smooths $\mathbf{Y}$ onto any location s_0, which may or may not be in the initial $s_1, \ldots, s_n$. For model (2.1), the most common smoother is the kriging predictor, given by

$$\hat{Z}(s_0) = \Sigma_0^T \Sigma^{-1} \mathbf{Y}$$

where $\Sigma_{i,j} = k(s_i, s_j) + \tau^2 \mathbb{1}_{\{i=j\}}$ and $\Sigma_0 = [k(s_0, s_1), k(s_0, s_2), \ldots, k(s_0, s_n)]^T$.

The maximum likelihood estimates for $\boldsymbol{\theta}$ are those values that maximize the log-likelihood

$$L(\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\mathbf{Y}^T\left[\Sigma(\boldsymbol{\theta}) + \tau^2 I\right]^{-1}\mathbf{Y} - \frac{1}{2}\text{logdet}\left(\Sigma(\boldsymbol{\theta}) + \tau^2 I\right) \tag{2.2}$$

for positive-definite covariance matrix $(\Sigma(\boldsymbol{\theta}))_{i,j} = k(s_i, s_j)$. Both terms of (2.2) present numerical difficulties. The determinant is difficult to evaluate for dense Σ and is sensitive to approximations, leading to many maximum likelihood approaches considering the zeros of associated score functions $L_i(\boldsymbol{\theta}) = \partial L/\partial\theta_i$. Differentiation of the log-likelihood gives the score functions for θ_i:

$$L_i(\boldsymbol{\theta}) = -\frac{1}{2}\mathbf{Y}^T\left(\Sigma + \tau^2 I\right)^{-1}\frac{\partial}{\partial\theta_i}\left[\Sigma + \tau^2 I\right]\left(\Sigma + \tau^2 I\right)^{-1}\mathbf{Y} - \frac{1}{2}\text{tr}\left[\left(\Sigma + \tau^2 I\right)^{-1}\frac{\partial}{\partial\theta_i}\left[\Sigma + \tau^2 I\right]\right].$$

The zero of the score functions correspond to local extrema of L.

[76] suggest a framework for solving these equations via a conjugate gradient on the quadratic form term and a stochastic Hutchinson trace estimator on the trace term. In particular, the estimating equations

$$g_i(\boldsymbol{\theta}, N) = \frac{1}{2}\mathbf{Y}^T\left(\Sigma + \tau^2 I\right)^{-1}\Sigma_i\left(\Sigma + \tau^2 I\right)^{-1}\mathbf{Y} - \frac{1}{2N}\sum_{j=1}^{N} U_j^T\left(\Sigma + \tau^2 I\right)^{-1}\Sigma_i U_j = 0$$

are unbiased for the score functions, where $U_1, \ldots, U_N$ are length n vectors with entries ± 1 simulated as independent symmetric Bernoulli random variables and where $\frac{\partial\left(\Sigma+\tau^2 I\right)}{\partial\theta_i} := \Sigma_i$. With some experimental design on the trace estimator and preconditioning on the conjugate gradient solver, considerable time savings can be realized. However, such a solver may require up to 100 random vectors of length n in the trace estimator as well as dozens of linear solves per computation of $\Sigma^{-1}\mathbf{Y}$. For dense covariance matrices without exploitable structure, multipole methods for matrix vector multiplication will not be as efficient, adding considerable costs per score function evaluation [76]. [78] use unbiased estimating equations to generate a sparse approximation for Σ^{-1}; such an approach yields promising results on irregularly spaced data at the cost of a handful of ad-hoc decisions including the choice of nearest neighbors.

2.2 The Equivalent Kernel and Kriging

It is convenient to note that the kriging estimator is just a linear predictor, and can alternatively be written

$$\hat{Z}(s_0) = \frac{1}{n}\sum_{i=1}^{n} w_n(s_0, s_i)Y(s_i) \tag{2.3}$$

for weight function $w_n : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$.

A key insight we use is that the kriging weight function, w_n, can be approximated by its equivalent kernel, denoted $G_\lambda(s,t)$, where $\lambda = \tau^2/n$. To exactly define the equivalent kernel, we suppose the n sample locations $s_1, \ldots, s_n$ are distributed on a subdomain $\mathcal{D} \in \mathbb{R}^d$ according to an empirical cumulative distribution function (cdf) F_n such that $F_n \to F$ as $n \to \infty$, where F is some limiting cdf on $\mathcal{D}$. If $\langle \cdot, \cdot \rangle$ denotes the inner product on the reproducing kernel Hilbert space (RKHS) whose reproducing kernel is k, then G_λ is the reproducing kernel for the Hilbert space of functions with inner product

$$\langle h_1, h_2 \rangle_\lambda = \int_{\mathcal{D}} h_1(s)h_2(s)dF(s) + \lambda \langle h_1, h_2 \rangle.$$

The equivalent kernel approximates w_n in the sense that

$$w_n(s,t) = G_\lambda(s,t) + \sum_{j=1}^{\infty} (\mathcal{R}_n^j G_\lambda(\cdot, t))(s) \tag{2.4}$$

where

$$(\mathcal{R}_n h)(s) = \int_{\mathcal{D}} G_\lambda(s,t)h(t)d(F - F_n)(t)$$

and $\mathcal{R}_n^j$ denoting the jth power of $\mathcal{R}_n$ [59, 47]. Under common formulations of the spatial model Z on $\mathcal{D} = \mathbb{R}^d$, the associated equivalent kernel is accessible. In general, for weakly stationary covariance k with spectral density $f : \mathbb{R}^d \to \mathbb{R}^+$, the equivalent kernel G_λ satisfies

$$\mathcal{F}(G_\lambda)(\omega) = \frac{f(\omega)}{f(\omega) + \lambda}$$

where $\mathcal{F}$ is the Fourier transform operator and $\omega \in \mathbb{R}^d$.

The most popular covariance function in geostatistics is the isotropic Matérn, given as a correlation by

$$k(s_1, s_2) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(a\|s_1 - s_2\|\right)^{\nu} K_{\nu}(a\|s_1 - s_2\|)$$

with K_ν a modified Bessel function of the second kind of order ν and $\|\cdot\|$ denoting Euclidean distance. We call ν the spatial smoothness parameter, a the spatial scale (with $1/a$ the spatial range). For half-integer ν, the Matérn correlation reduces to a product of a polynomial and an exponential [29].

If $d = 1$ and $\nu = 0.5$, the equivalent kernel is available in closed form as

$$G_\lambda(r) = \frac{a}{\lambda\pi}\left(\sqrt{\frac{a}{\lambda\pi} + a^2}\right)^{-1} \exp\left(-r\sqrt{\frac{a}{\lambda\pi} + a^2}\right). \tag{2.5}$$

Under this scenario, if spatial locations are equally spaced (i.e., gridded) then the observational data according to (2.1) are noisy observations of an autoregressive process of order 1. Without the measurement error process $\varepsilon(s)$ then standard Markov conditions can be used to easily calculate the likelihood function, but with the nugget effect the Markov effect is lost and likelihood-based inference can be difficult for large n. It is worth noting the exponential decay of G_λ; although not a focus of this work, exponential decay ends up being a crucial component of the theory supporting the kernel approximation [59]. Another closed form is available for $d = 1$ and $\nu = 1.5$, wherein

$$G_\lambda(r) = \sqrt{\frac{\pi}{2\lambda\pi a + 4}} \exp\left(-ra\right)\left(B\cos(rB) + A\sin(rB)\right) \tag{2.6}$$

for $2A^2 = \sqrt{a^4 + 2a^3/(\lambda\pi)} + a^2 \quad 2B^2 = \sqrt{a^4 + 2a^3/(\lambda\pi)} - a^2$. Closed forms for other values of d and ν are not apparently available.

A second very general class of models suitable for equivalent kernel approximation are basis representation models. For L known basis functions $\{\phi\}_{i=1}^{L}$, we represent

$$Z(s) = \sum_{i=1}^{L} c_i \phi_i(s) \tag{2.7}$$

where the coefficients $\mathbf{c} = (c_1, c_2, \ldots, c_L)^T$ are stochastic. If we group $\phi(s) = (\phi_1(s), \phi_2(s), \ldots, \phi_L(s))^T$ and denote Q as the precision matrix of $\mathbf{c}$, then $k(s_1, s_2) = \phi(s_1)^T Q^{-1} \phi(s_2)$.

Use of a fixed and small number of basis functions to model Gaussian processes is the basis of fixed rank kriging [17]. The inclusion of stochastic coefficients $\mathbf{c}$ allow for appeals to the stochastic partial differential equations framework discussed in [49] and [68]. Multiresolution representations have been favored recently by [60] and [39] largely due to their ability to capture both large-scale and small-scale variations in spatial structure.

A basis is particularly convenient for using the equivalent kernel, because the kernel is available in closed form. For an $L \times L$ inner product matrix P given by $P_{i,j} = \int \phi_i(s)\phi_j(s)dF(s)$, the equivalent kernel for k is given by

$$G_\lambda(s_1, s_2) = \phi(s_1)^T(P + \lambda Q)^{-1}\phi(s_2). \tag{2.8}$$

Moreover, the remainder terms in (2.4) take the form

$$(\mathcal{R}_n^j G_\lambda(\cdot, s_2))(s_1) = \phi(s_1)^T(P + \lambda Q)^{-1}\left[(P - P_n)(P + \lambda Q)^{-1}\right]^j \phi(s_2), \tag{2.9}$$

where $(P_n)_{i,j} = \int \phi_i(s)\phi_j(s)dF_n(s)$, is just a finite sum approximation to P, based on the observation locations.

2.2.1 Remainder Terms and The Nugget

[47] showed the convergence conditions for the equivalent kernel to the kriging weights. However, the bounds on G_λ and its first and second derivatives are scaled by $1/\lambda^2$, $1/\lambda^4$, and $1/\lambda^6$, respectively. The effective uses of the approximation as a function of the nugget are not well understood. To explore this, we first consider a basis representation model due to the ease of adding additional remainder terms.

Here we consider a special case of bases representations given by a multiresolution stochastic process. Under this model, the bases ϕ exist at multiple resolutions across the spatial domain, with increasing fine partitions of the spatial domain housing increasingly compactly-supported basis functions with increasingly small weights in the precision matrix Q. [61] suggest the use of Wendland basis functions for their LatticeKrig models, the compact support of which implies

sparsity in P. We follow their suggestion and allow for Wendland polynomials with support widths of exactly 2.5 times the grid spacing, allowing for sufficient overlap to capture spatial nuance. Nodes for the Wendland bases extend past the edge of the observation domain $\mathcal{D}$ to allow for adequate description of boundary behavior. We also follow their heuristic of doubling the number of basis nodes for each level of the process. Note that more advanced forms of basis selection exist, including iterative algorithms informed by predictive process models as in [10, 39].

Under LatticeKrig, sparsity in the precision matrix Q results from the properties of the GMRF, allowing for sparse matrix methods in the computation of (14). Under both [61, 49], diagonal B is specified to satisfy a spatial autoregressive model, where $(1/\tau^2)B^TB$ is the covariance matrix for the stochastic coefficients c. This ensures an explicit connection between the Matérn and underlying SPDE theory, but is not necessary in the construction of a valid covariance model (14). The scaling terms ρ_l are fixed for any level of the process, but decrease as we move from level to level, typically of the form $\rho_l \propto e^{-\nu l}$, normalized such that $\sum_{l=1}^{L} \rho_l = 1$.

Our simplest basis model fixes the precision matrix Q in (14) as a diagonal matrix, including parameterization only through ρ. Figure 3.1 shows the difference between the kriging weights and the equivalent kernel at 0, and 1, remainder term. These correspond to a 3-level basis representation of 1000 observations on $[0, 1]$, with the first basis level containing nodes at every $1/8$. This results in 13 splines at this level; 7 on the interior of $\mathcal{D}$, 2 on the endpoints, and 2 on each side of $[0, 1]$ for the boundaries. Shown in figure 2 are the kriging weights and equivalent kernel weights corresponding to the 2nd and 500th observation locations for this model.

The equivalent kernel weights well approximate the kriging weights near the midpoints of the data for all tested nuggets and observation sizes. The surprising results occur at lower sample sizes and lower nugget effects. For example, in the $n = 100$ and $\tau = 0.01$, including the remainder terms R_1 and R_3 into the equivalent kernel makes the weights progressively worse, even as including the full geometric series R_∞ still converges within machine precision to the kriging weights. Even for the same parameters, the midpoint weights appear to converge well. However, the resulting krige is highly divergent as the mis-weighted endpoint are included, as seen in figure 3 with simulated

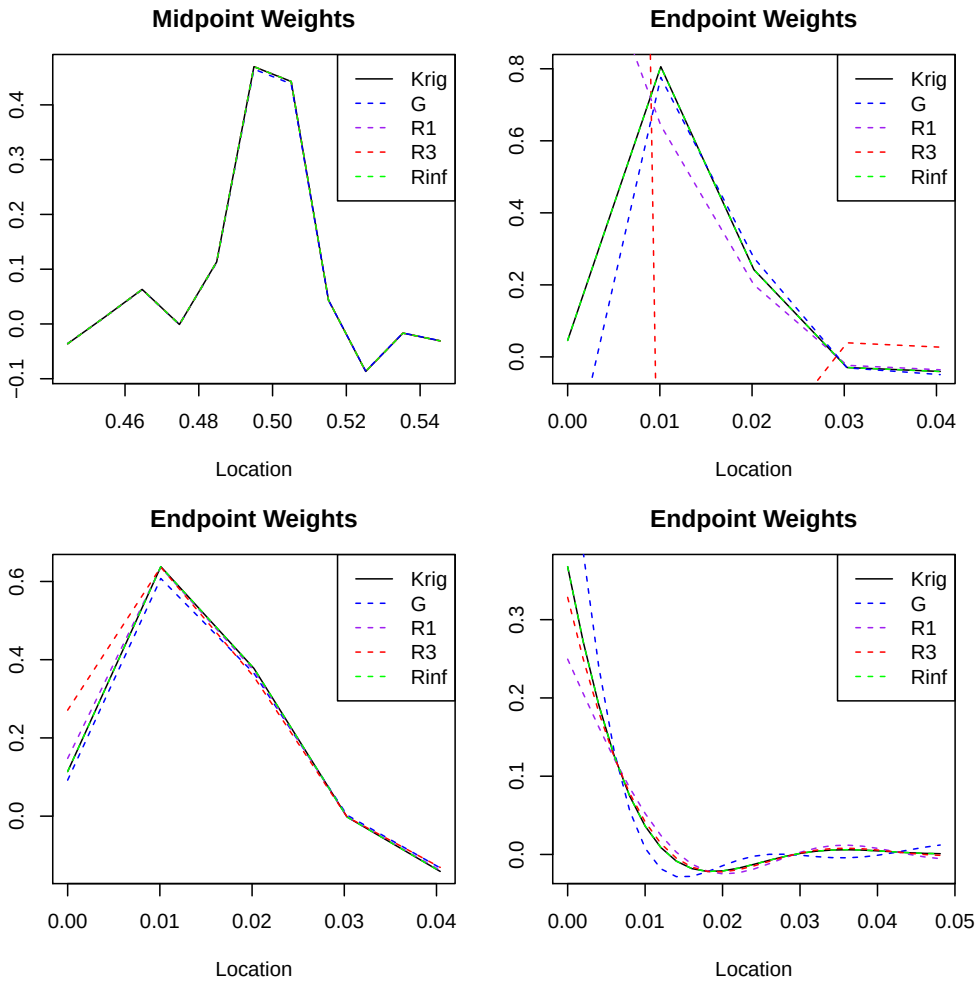

Figure 2.1: Clockwise from top left: Midpoint weights for $\tau = 0.01$, $n = 100$; Near-endpoint weights for $\tau = 0.01$, $n = 100$; $\tau = 0.1$, $n = 100$; $\tau = 0.01$; $n = 500$

exponential process data satisfying $k(r) = e^{-2r}$ included.

Again, R_∞ accurately follows the true krige even as successively adding remainder terms causes a worse and worse approximation. Remainder terms improve the approximation as n reaches 500 (in our simulations, it appeared to stabilize around $n = 200$), as well as when $\tau = 0.1$ for smaller samples. In practice, a standard deviation of $\tau = 0.1$ implies a variance nugget-to-sill or noise-to-signal ratio of $1/100$, so many data sets in practice should not be affected by the observed behavior.

The asymptotic properties of the equivalent kernel also mitigate some of the concerns about the effects of a small nugget. Figure 4 displays the ratio of the L^1 matrix norms of the difference

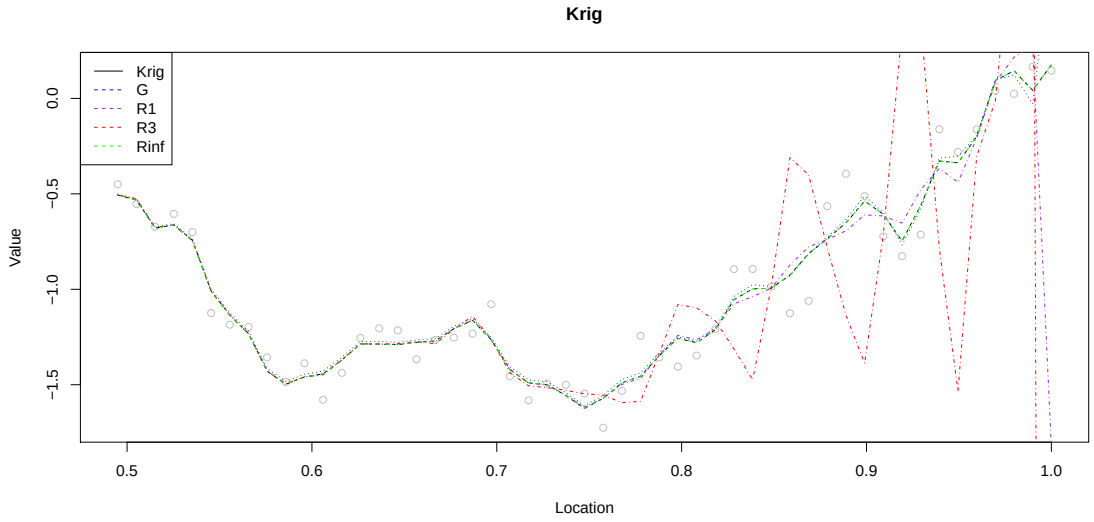

Figure 2.2: Sample krige for $\tau = 0.01$; $n = 100$.

between the kriging weights matrix and the equivalent kernel matrix

$$\frac{|\Sigma_0(\Sigma + \tau^2 I)^{-1} - \boldsymbol{G}|}{|\Sigma_0(\Sigma + \tau^2 I)^{-1}|}.$$

Included are varying amounts of included remainders for a one dimensional process with a basis covariance representation spanning three levels, as above. For almost no nugget when $\tau = 0.01$, smaller samples of both $n = 100$ and $n = 300$ exhibit an equivalent kernel estimation diverging in the L^1 norm from the kriging weights as finitely many remainder terms are added, although their infinite series still converges. This behavior is observed at $\tau = 0.1$ and $n = 100$ as well, but in general either a larger sample size or a larger nugget improves both the equivalent kernel approximation, and the efficacy of added remainder terms. Even at very low nuggets, once we reach a sample of $n = 1000$, addition of remainder terms improves the matrix approximation at all tested τ^2 values.

2.2.2 Joint Estimation Techniques

Discussion on the importance of the nugget invites a broader discussion of joint estimation techniques in spatial statistics. Use of a quasi-Newton algorithm can provide a direct solution to the

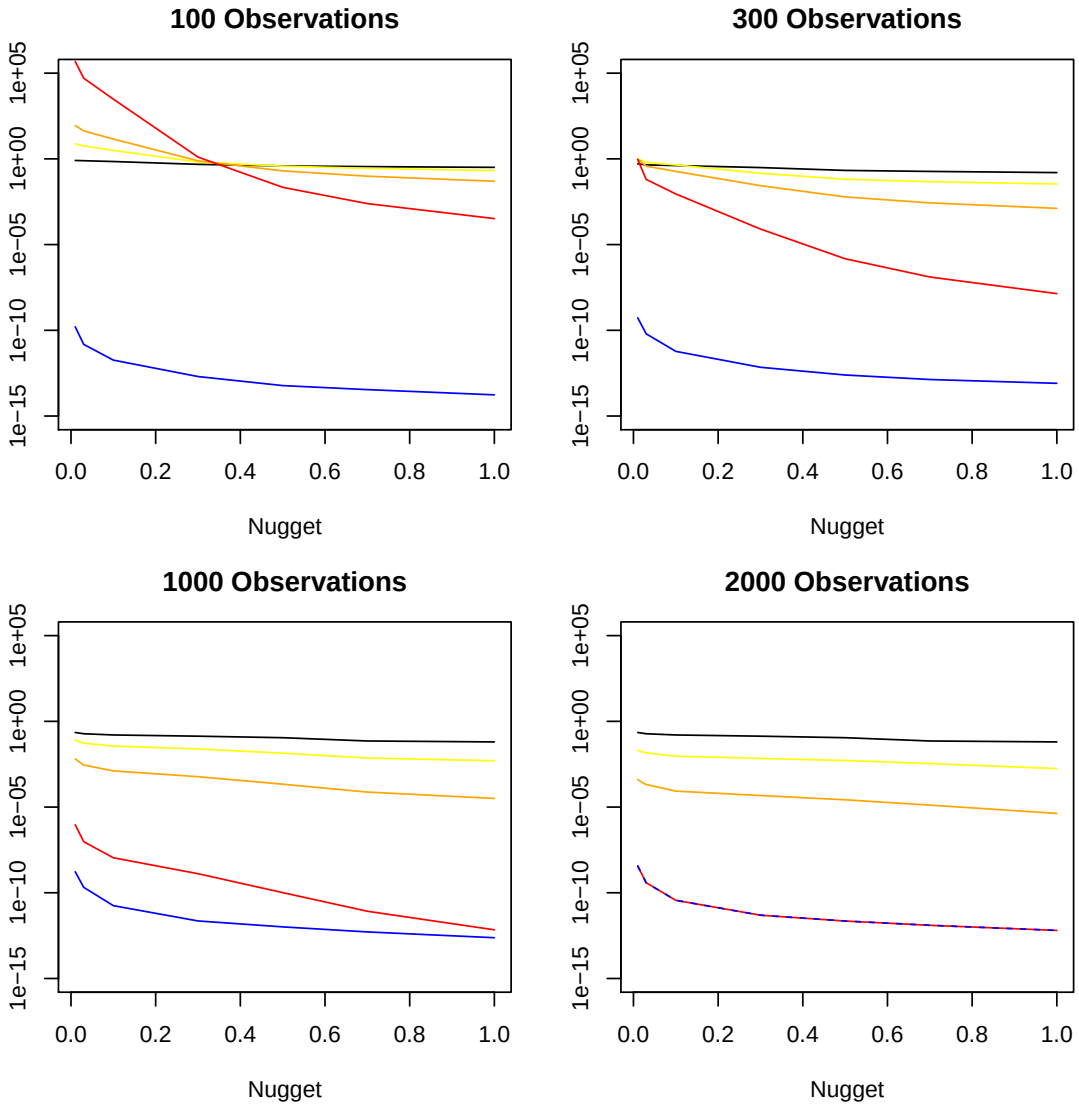

Figure 2.3: Relative L^1 errors. Black: No remainders; Yellow: 1 remainder; Orange: 3 remainders; Red: 10 remainders; Blue: infinite remainders.

maximum likelihood problem [7] or provide the rootfinding algorithm for approximate score function evaluations. In general, the challenge of rootfinding in spatial maximum likelihood is highlighted by the ridge-like shapes of the profile likelihood function with respect to spatial parameters. In particular, for the Matérn, the likelihood $L(a, \nu)$ commonly exhibits a sharp ridge along increasing a and ν - a short range and smooth process oft has a similar likelihood to a long range and non-smooth process. This allows for easy optimization of the marginal maxima of ν given a and vice

versa, but provides challenges for joint optimization. Further, a score function approach will find any solution where the score functions (18) are zero; if there are multiple such solutions there is no guarantee that a global maximum as opposed to a local one was found. [22] use a nonlinear solver combining bisection and quadratic interpolation in the FORTRAN package zeroin and the MATLAB package fzero, which can outperform Newtownian approaches when estimating along ridges.

Even in the presence of accurate score function evaluations and an appropriate rootfinding algorithm, there are difficulties with joint estimation of multiple parameters in (17). If the spatial domain is allowed to grow with the number of observations - increasing domain asymptotics - then the maximum likelihood estimates (MLEs) for the spatial covariance parameters are consistent and asymptotically normal [52]. Under infill asymptotics, as the observed locations $n \to \infty$, the spatial domain $\mathcal{D}$ remains fixed. Under infill asymptotics, For the $d = 1, \nu = 0.5$ exponential case, the range and sill parameters are indistinguishable up to the product $\sigma^2 a$ owing to the absolute continuity of their probability measures [36]. [87] generalized this result to point out the absence of a weakly consistent joint estimator of the Matérn ν, a, and σ, emphasizing the importance instead of the product $\sigma^2 a^{2\nu}$. [15] added concerns about the estimation of the nugget, showing that the estimation of the sill σ^2 falls from order $n^{-1/2}$ to order $n^{-1/4}$ in the presence of non-zero nugget τ^2, even given fixed and known scale parameter a. [88] derived the sampling distributions for the MLEs of nugget, sill, and range of the 2-dimensional exponential, finding both signs of bias and skew in the MLEs that increased with existence of a nugget. Despite these concerns with the lack of a consistency, [42] showed that joint estimation of range and sill empirically outperformed treating the range as fixed, and in particular the joint estimation performed best when the effective range of spatial correlation was small relative to the domain.

To date, no work has attempted to use the equivalent kernel framework in parameter estimation for Gaussian processes. In this chapter, we propose using a new approach to finding zeros of the score function of a Gaussian process by relying on equivalent kernel approximations. Our numerical experiments below suggest the magnitude of the nugget term plays an important role.

We then discuss and explore two alternate approaches to approximating the score function, and end with an application to a large temperature time series dataset.

2.3 The Equivalent Kernel and Score Functions

2.3.1 Score Function Approximations

We introduce two approximations to the score function of n observations from (2.1) where z is a mean zero Gaussian process, both based on the equivalent kernel characterization. We use the notation $\mathbf{G}$ to denote the equivalent kernel matrix with (i,j)th entry $G_\lambda(s_i, s_j)$.

When differentiating with respect to parameters of k, differentiation of the kriging weight itself gives the following, that we term the Differential Approximation (DA).

Differential Approximation (DA)

$$\mathbf{Y}^T \left(\Sigma + \tau^2 I\right)^{-1} \frac{\partial}{\partial \theta_i} \Sigma \left(\Sigma + \tau^2 I\right)^{-1} \mathbf{Y} \approx \frac{1}{\tau^2} \mathbf{Y}^T \frac{\partial \mathbf{G}}{\partial \theta_i} \mathbf{Y}. \tag{2.10}$$

The following Identity Difference Approximation (IDA) exploits a useful trick for rewriting a matrix-plus-ridge inverse.

Identity Difference Approximation (IDA)

$$(\Sigma + \tau^2 I)^{-1} \approx \frac{1}{\tau^2}(I - \mathbf{G}). \tag{2.11}$$

The IDA can be used in both the quadratic form and the trace terms of the score function:

$$L_i \approx \frac{1}{\tau^4} \mathbf{Y}^T (I - \mathbf{G})^T \Sigma_i (I - \mathbf{G}) \mathbf{Y} - \mathrm{tr}(\Sigma_i (I - \mathbf{G}))$$

where the first term can be evaluated using either the identity difference or the differential approximation. Derivation of both estimates can be found in the Appendix.

Figure 2.4 demonstrates the intuition of the IDA trace approximation. Each of the n terms of the trace summand is a dot product of the ith row of $I - G$ and the ith column of Σ_i. As n increases, G approaches a delta function, and very little of the product appears near the boundaries.

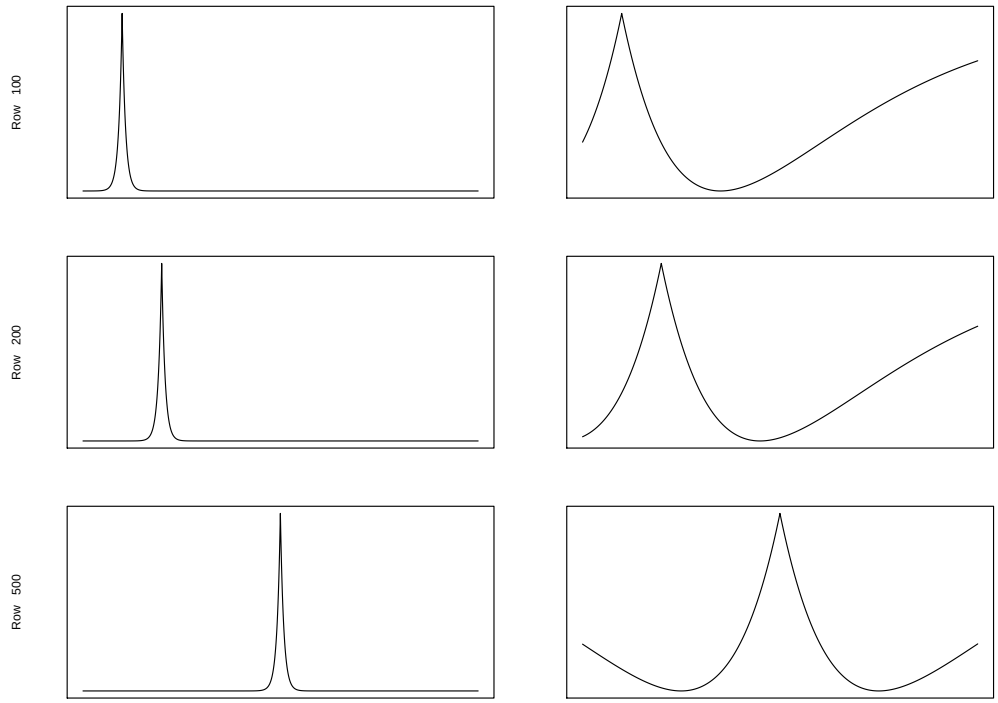

Figure 2.4: Left: Rows 100, 200, 500 of G Right: Columns 100, 200, 500 of Σ_a

Ignoring this effect by taking n copies of the dot product at the center of the observation domain gives rise to 2.11.

Each of these methods avoid the computation of matrix inverses, but involve matrix-matrix products and many matrix-vector multiplications, where even storing matrices on the order of the covariance matrix can be prohibitive. For gridded data, multiple techniques can further improve the time and storage requirements for approximate score function evaluations. In particular, the quadratic form evaluation can be done via a single fast Fourier transform (FFT) and a dot product in the differential case, and a pair of FFTs and a dot product in the identity difference case. The trace term can be approximated on larger sample sizes by ignoring the boundary effects of the equivalent kernel approximation and treating G as a stationary kernel function over the entire

domain. This allows for a single vector multiplication to approximate the trace term as

$$\begin{aligned}\frac{1}{\tau^2}\mathrm{tr}(\Sigma_i(I-\boldsymbol{G})) &= \frac{1}{\tau^2}\left(\mathrm{tr}(\Sigma_i) - \mathrm{tr}(\Sigma_i\boldsymbol{G})\right) \\ &\approx \frac{1}{\tau^2}\left(n\frac{\partial k}{\partial\theta_i}(0) - n\sum_{l=1}^{n} G(|s_l - s_j|)\cdot\frac{\partial k}{\partial\theta_i}(|s_l - s_j|)\right) \end{aligned} \tag{2.12}$$

for a fixed data location s_j in the middle of the domain $\mathcal{D}$ to avoid boundary approximation errors. This reduces the order of the trace term computation down to a single $\mathcal{O}(n)$ vector-vector multiplication. The following lemma describes the adequacy of this approximation as a function of the convergence of F_n to F and the kriging smoothing parameter λ.

Lemma 1. *The jth term of the approximate trace in (2.12) satisfies*

$$\left|(\Sigma_i\boldsymbol{G})_{j,j} - \sum_{l=1}^{n} G(|s_l - s_j|)\cdot\frac{\partial k}{\partial\theta_i}(|s_l - s_j|)\right| \leq K\left|\frac{\delta_n}{(1-\delta_n)\rho}\right|$$

where $K > 0$ is a constant, $\delta_n \propto D_n/\rho$, $\rho = \lambda^{\xi}$ for some $\xi > 0$ and $D_n = \sup|F_n - F|$.

This guarantees elementwise convergence of the diagonal of the approximate trace term in (2.12), but a sum of n of these terms implies possible unbounded error in the approximation. However, our empirical results suggest that even an exponential covariance is sufficiently smooth for this trace approximation to apply to larger one dimensional data sets. The way to interpret ξ is that it is related to the smoothness of the process. For instance, if kriging is performed using an mth order smoothing spline, then $\xi = 1/(2m)$.

2.3.2 Simulation Study Asymptotics

The equivalent kernel approximation to the kriging weight function requires a nugget effect, but also depends on the sample size. Our experience suggests the nugget effect cannot be trivially small for the approximation to work well unless the sample size is very large. The results of Lemma 1 also suggest sensitivity to smoothness, so for empirical results on the trace estimator we begin with a relatively coarse process on a uniformly spaced one-dimensional mesh following

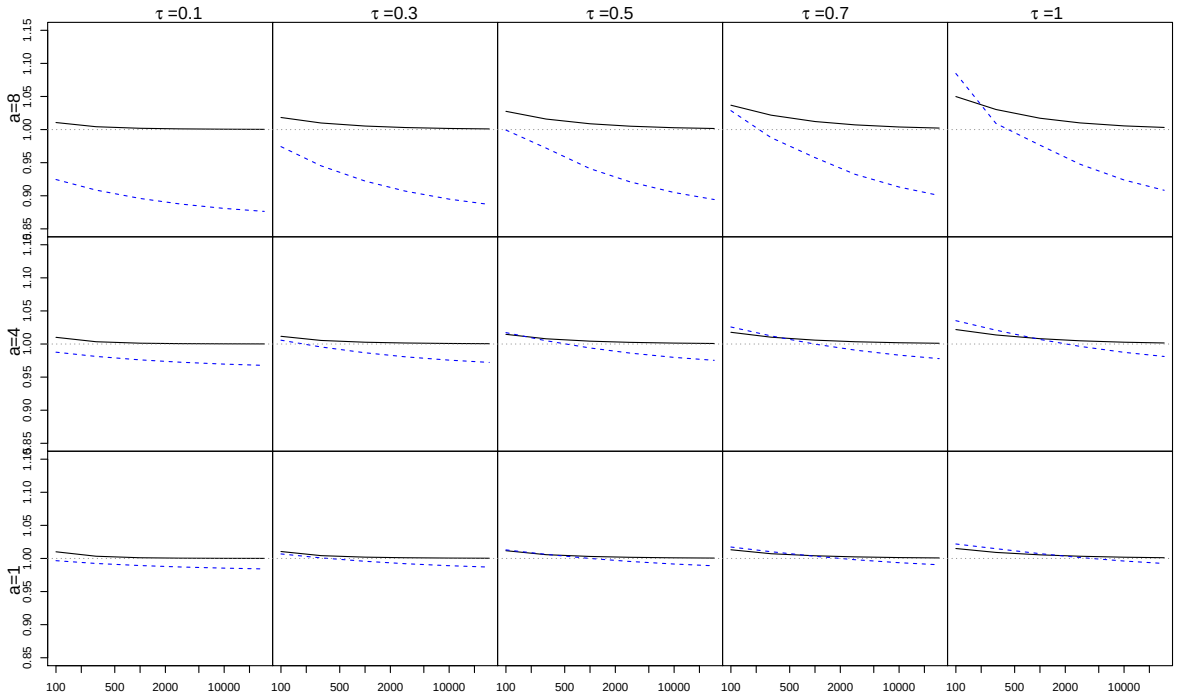

Figure 2.5: Ratios of approximate IDA trace term to exact for $\nu = 0.5$ (solid black) and $\nu = 1.5$ (dashed blue)

an exponential covariance function $k(s,t) = \exp(-|s-t|/a)$. For a factorial design with scale parameters of $a = 1, 4, 8$ and nuggets of $\tau = 0.1, 0.3, 0.5, 0.7, 1$, we compare the exact values of $\frac{1}{\tau^2}\text{tr}(\Sigma_i(I - \boldsymbol{G}))$ to the approximates in (2.12). Figure 2.5 shows the ratio of the approximation to the true value split across tested τ and a parameters for sample sizes ranging from $n = 100$ to $n = 30000$.

The approximate trace has small bias for each parameter tested, and typically this bias decreases rapidly to under one percent by a sample size of $n = 1000$, and for the exponential covariance in the worst case tested was within 5% of the true term at $n = 100$ observations. The error of this approximation tended to increase slightly with increases in τ and with decreases in a, although these effects were minor overall. The higher smoothness process, however, had poorer results for the highest tested spatial range of $a = 1$. While this is not uncommon in similar analyses, it is worth noting that longer range and smoother processes lead to high counts of observations with correlations near 1, which increases the condition number of Σ and causes numerical instability in

many spatial methods, including the FFTs exploited in any matrix-vector multiplication herein.

On sample sizes similar to those in Figure 2.5, use of the IDA on the score function's quadratic form instead of conjugate gradient had mixed results. However, as the sample sizes increase to 10^5 or 10^6, the full $\mathcal{O}(n)$ method implied by usage of (2.10) and (2.11) can be utilized. Results for approximate univariate MLEs $\hat{a}$ for the range parameters of exponential covariance processes on regular grids of 10^5 and 10^6 are shown in Figure 2.6. Each boxplot consists of the ratios $\hat{a}/a$ over 12 data sets each of $a = 2, 5, 12$. While the results are promisingly accurate and quite fast on a personal desktop – the mean time per score function evaluation increased from 0.72s to 5.05s as the sample size increased from 10^5 to 10^6 – the IDA was sometimes quite inaccurate at evaluating score functions at smaller a values. In particular, for values even as close to the true parameter as $0.1a$, the approximate score function would exhibit a second zero as the approximation diverged. The approximate univariate MLEs for the a parameters were computed via the UNIROOT command in `R`, based on the ZEROIN algorithm from FORTRAN. When UNIROOT was given a window from around $a/3$ to $3a$, all of the sets converged, but wider windows would occasionally require a grid search or a search decreasing from large to small a to identify the proper zero.

2.3.3 Comparison to Existing Methods

For larger samples, direct evaluation of the score function is infeasible, so we compare our DA and IDA methods against the combination of conjugate gradient and the Hutchison trace estimator proposed by [76]. For these experiments, we again use the exponential covariance $k(s,t) = \exp(-a|s-t|)$ on uniformly gridded simulated data on $[0, 2\pi]$. For these experiments, a known marginal variance of 1 is included to avoid the difficulties and inconsistencies of joint estimation of the range and marginal variance [87]. A full factorial experiment with sample sizes of $n = 500, 1000, 3000, 5000, 10000, 30000$, scale parameters of $a = 1, 4, 8$ and nuggets standard deviations of $\tau = 0.1, 0.3, 0.5, 0.7, 1$ was performed. For each of the 90 possible choices of parameters, 12 data sets were generated.

Our results suggest that the equivalent kernel does not perform well in the quadratic form at

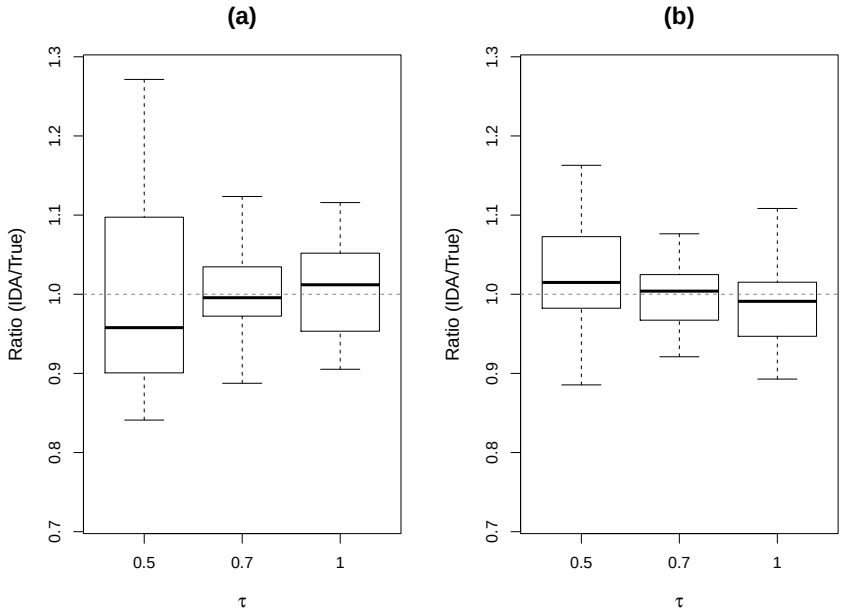

Figure 2.6: Box plot ratios of Approximate MLEs for $\hat{a}/a$ using the IDA on both terms for: (a) 10^5 observations; (b) 10^6 observations

smaller sample sizes than in the previous section: the element-wise convergence of G to the kriging weights does not guarantee convergence through the propagation of error in multiple matrix-vector multiplications, and the resulting score functions from both the IDA and the DA could be strictly positive or strictly negative, resulting in failed algorithms. As a result, we compare the zeros of

$$L_{a,CG} = \frac{1}{2}\mathbf{Y}^T \left(\Sigma + \tau^2 I\right)^{-1} \Sigma_a \left(\Sigma + \tau^2 I\right)^{-1} \mathbf{Y} - \frac{1}{2N}\sum_{j=1}^{N} U_j^T \left(\Sigma + \tau^2 I\right)^{-1} \Sigma_a U_j = 0$$

where the entries of the U vector are symmetric Bernoulli random variables with entries ± 1 and

$$L_{a,IDA} = \frac{1}{2}\mathbf{Y}^T \left(\Sigma + \tau^2 I\right)^{-1} \Sigma_a \left(\Sigma + \tau^2 I\right)^{-1} \mathbf{Y} - \frac{n}{2\tau^2}\sum_{l=1}^{n} G(|s_l - s_j|) \cdot \frac{\partial k}{\partial a}(|s_l - s_j|)$$

for data location s_j in the middle of the domain $\mathcal{D}$.

The Hutchison trace estimator is taken with $N = 50$ independent Bernoullis, and for each method the quadratic form was calculated using conjugate gradient for $\left(\Sigma + \tau^2 I\right)^{-1}\mathbf{Y}$ and the fast Fourier transform with circulant embedding for all matrix-vector computations. Similar to the results in Section 3.1, we observe a slight bias in the IDA solutions that decreases as n and τ

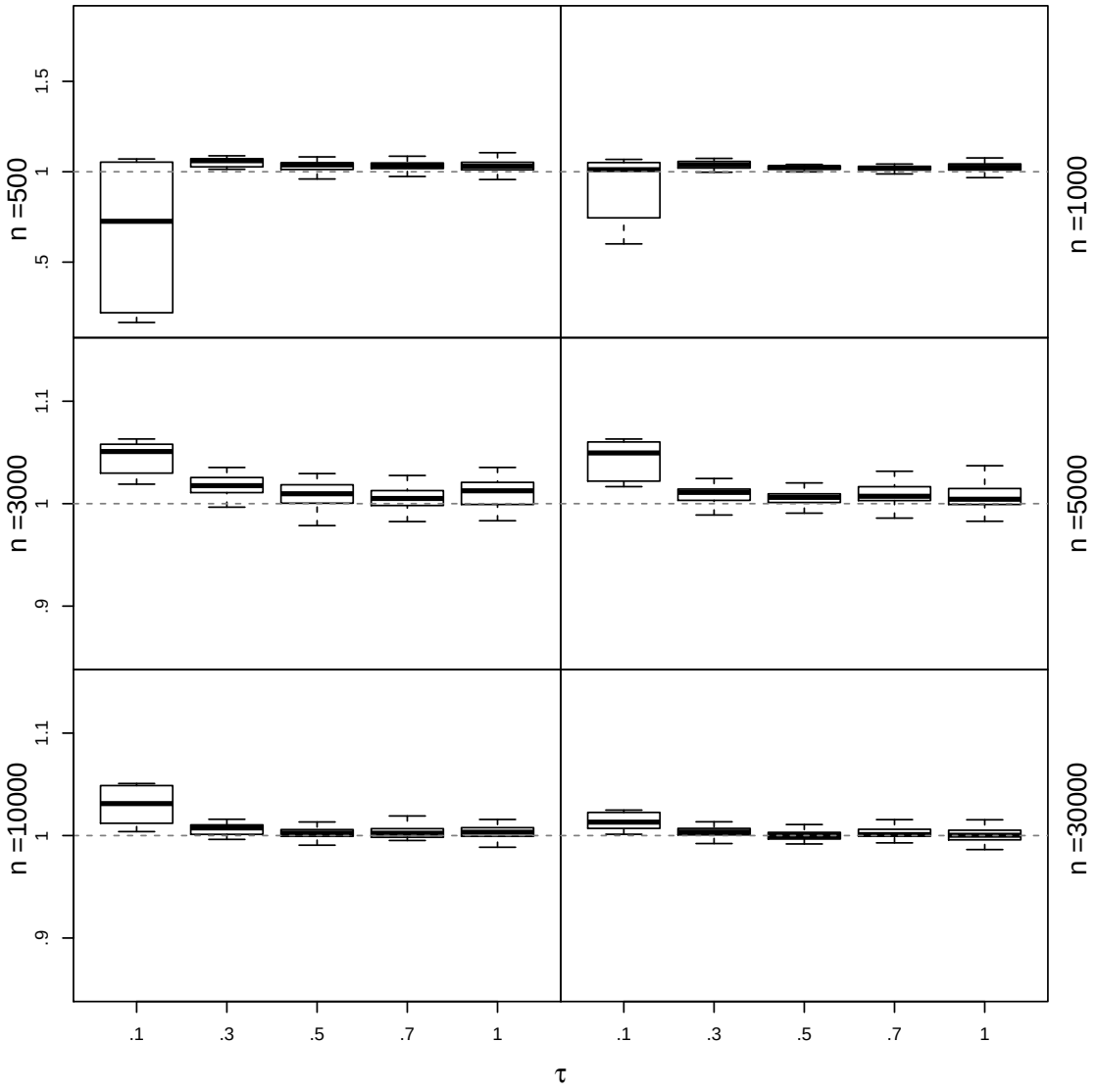

Figure 2.7: Ratio of zeroes of $L_{a,IDA}$ to zeroes of $L_{a,CG}$

increase. In particular, Figure 2.7 shows that for smaller samples and the smallest tested nuggets, the zeros of $L_{a,IDA}$ sometimes underestimate the zeroes of $L_{a,CG}$. On the other hand, for all other tested nuggets the zeroes of $L_{a,IDA}$ were slight overestimates of the zeroes of the Hutchison trace method, but this bias had converged to within 1% by $n = 5000$ for all tested $\tau > 0.1$, and this bias was within 5% for $n = 10000$ and 30000 at the smallest nugget size.

The payoff of using the IDA trace estimator is in the timing, where performing one instance of conjugate gradient per score function evaluation instead of 51 reflected in timing savings of nearly

50 times. Pooling all score function evaluations - for each data set as well as the 8-12 iterations typically needed for UNIROOT - these time savings were consistent across all tested sample sizes. Using the IDA trace term instead of the Hutchison trace estimator took between 41 and 47 times less duration at all tested sample sizes, reflecting that the single iteration of conjugate gradient in the quadratic form was still the gateway. For these experiments the timing for each method was approximately $\mathcal{O}(n^2)$: in moving from 10000 to 30000 samples the time per evaluation of $L_{a,CG}$ increased 7.27 times from 2.58 s to 18.84 s on the tested personal laptop in the R programming language.

2.3.4 The IDA in 2D

In the previously tested cases of the exponential and $\nu = 1.5$ covariance functions, the closed form for the equivalent kernel $G(r)$ was available. In general, the integral in (12) is a form of Hankel transform typical to the $d = 2$ case of isotropic covariance functions, and the long-term oscillation of J_0 poses numerical difficulties for many quadrature techniques, especially for small r [47]. In studying the equivalent kernel approximation to the kriging weight, [75] apply a Taylor expansion of a trail-truncated equivalent kernel, concluding that such an approximation deteriorates depending on the data's signal-to-noise ratio. We opt for a more exact computation to allow for sensitivity to this relative size of the noise term τ^2. Hankel transformation integrals are heavily considered by electromagnetic models in geophysics [16], and can be evaluated quickly by iterative application of Shank's algorithm for alternating series, which compared favorably to other popular algorithms [43]. Shank's transformation [71] is a non-linear series convergence accelerator, and requires reducing the infinite domain integral in (12) into an infinite series whose partial sums are easily calculable. In particular, consider an alternating sequence $\{a\}_{i=1}^{\infty}$ with partial sums $S_n = \sum_{i=1}^{n} a_i$. If the limit $\lim_{n\to\infty} S_n$ exists, then the transformed series given by $R_n(S) = \dfrac{S_{n+1}S_{n-1} - S_n^2}{S_{n+1} - 2S_n + S_{n-1}}$ converges more rapidly than S. The iterative sequences $R_2(S), R_3(S), \ldots$ accelerate the process further. In practice, the convergence is guaranteed by enforcing the a_i sequence is alternating and decreasing

in magnitude by integrating

$$a_i = \int_{z_i}^{z_{i+1}} g(a, \nu, \rho) J_0(r\rho) d\rho$$

over the zeroes of J_0, as g is decreasing and strictly positive under typical transformation. After the first handful, the zeros of J_0 appear almost exactly π apart, which allows for numerically stable integration over finite intervals combined with rapid convergence compared to many default packaged infinite integration techniques.

While requiring numerical integration of $G(r)$ instead of direct evaluation of $k(r)$ adds computational time, all possible distances implied in Σ need not be integrated over. Instead, we evaluate G on a vector of distances of length $j << n$, and fit $G(r)$ to our data set by an interpolating spline, maintaining added computational time on order jn. For our tested experiments, we fix $j = 1000$ for all sample sizes.

In one dimension, our results in Section 4.1 suggest that by around $n = 100$, the boundaryless IDA trace estimator is a good approximation of the true trace. In two dimensions we use gridded data on a $n \times n$ grid; our results suggest that adequate approximation requires grids of dimension 100×100 or more. The left half of Figure 2.8 shows the convergence of the IDA estimator to estimates generated by a Hutchison trace estimator for a Mátern process with $a = 10$, $\nu = 1.5$, $\tau = 0.4$ on $[2\pi, 2\pi]$. For small samples, the IDA approximation performed quite poorly: at $n = 30$ the approximation returned the wrong sign and does not fit on the figure; at $n = 40$ and $n = 50$ the bias is more than 30%. By $n = 100$, this bias is again under 5%. As before, the ratio given by $\text{Trace}_{IDA}/\text{Trace}_{Hutchison}$ is plotted.

The right half of Figure 2.8 shows the 2D estimator at $n = 1000$ per side for a few parameter samples. Three sets of a, ν parameters were tested: $a = 20, \nu = 2.5$; $a = 10, \nu = 2.5$; and $a = 4, \nu = 1.5$. Each set was tested at both $\tau = 0.2$ and $\tau = 0.6$, with results for latter in dashed lines. Across these varying ranges and smoothnesses, the IDA trace estimator performs nearly the same, with a bias of underestimating relative to a Hutchison trace estimator decreasing from around 2-3% to 1% as samples increase from a 200×200 grid to a 1000×1000 grid.

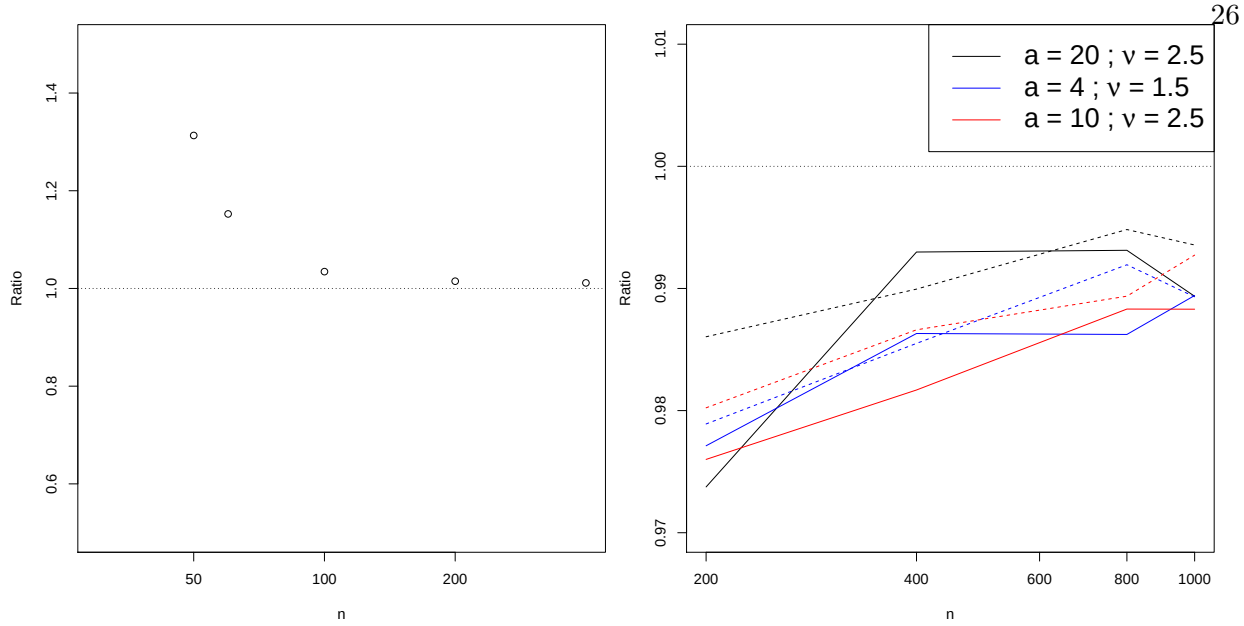

Figure 2.8: Left: 2-dimensional IDA trace approximation ratio on a grid of $n \times n$; Right: the ratio split across across tested pairs of a and ν ($\tau = 0.2$ as solid, $\tau = 0.6$ as dashed).

2.3.5 Data Example

We consider a time series of daily maximum temperature data from the United States Historical Climatology Network (USHCN) which are commonly used to build stochastic weather generators for hydrology and climate studies [45, 46]. These data are freely available online, and represent high-quality, quality-controlled directly observed temperatures at various locations throughout the United States. In particular, we consider uninterrupted daily maximum temperatures in Crosbyton, Texas for the years 1988-2014, yielding a total of 9855 data points. The data is demeaned by regressing out seasonal signals of the form $c_i \cos(k \times 2\pi d/365)$ with both annual and biannual sine and cosine terms ($k = 12, k = 24$). The residuals are heteroskedastic, exhibiting differing variances in summer and winter, so the residual time series is homogenized in magnitude by scaling each point by a rolling standard deviation of the ± 20 adjacent values. The result is an overall approximately unit variance stationary process.

We consider a 1-D exponential covariance of the form $k(s,t) = \exp(-a|s-t|)$ with white noise τ^2. Using UNIROOT in R, we can find the joint zero of the score functions for τ^2 and a. Under a loose

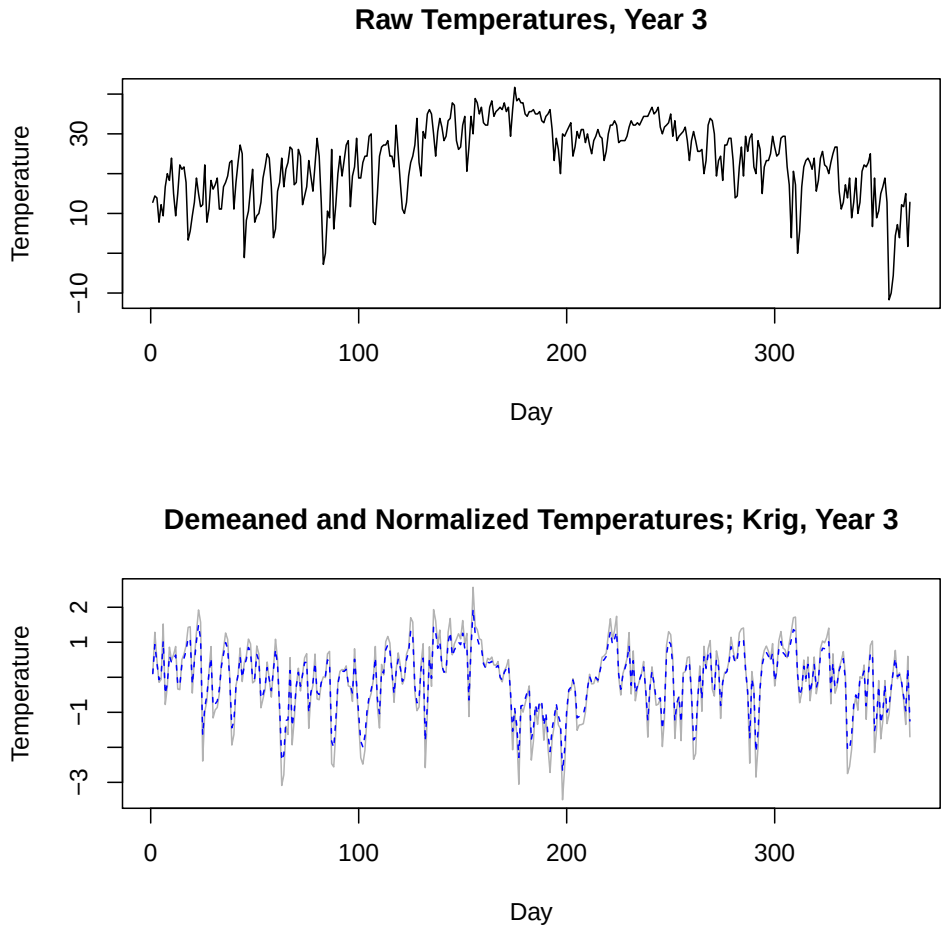

Figure 2.9: Top: the raw time series in Celsius; Bottom: the transformed and normalized series and post-estimation Krige (only year 3 shown for each)

initial search of $\tau^2 \in (0.2, 1.5)$ and $a \in (1/10, 10)$ (in days^{-1}), the algorithm converges in 140 joint score function evaluations to an estimate of $(\hat{a}, \hat{\tau}^2) = (1.328, 0.6176)$. This optimization took a total of 36 seconds on a Lenovo Yoga laptop/tablet hybrid. The ability to perform over 3 approximate score function evaluations per second on a data set of the scale of 10^4 observations on a computer with little computational power seems promising. In practice, the proposed approximations suggest even small personal computers can scale to considerably larger data sets. Repeating the data set in alternating forward and backwards order (to maintain continuity) 5 times for a full sample of approximately 10^5 observations took the same laptop/tablet 5.75 seconds to evaluate L_a and L_τ

once each at their estimated values; for 10^6 observations generated the same way this increased to 35.78 seconds.

2.4 Discussion

This chapter introduces an equivalent kernel-based approximation to the score function of a Gaussian process. Our experimental results suggest the approximations are sensitive to both the amount of noise of the spatial data and the size of the sample, but perform well with moderately-sized datasets exhibiting a small amount of noise, and are computationally fast. While the theory holds for one-dimensional processes, future work may focus on extending theoretical and experimental results to higher dimensions and estimation on other covariance functions.

Chapter 3

High-Frequency Time Series Methods for Wind Modeling

3.1 Motivation

Statistical analysis of wind is loosely broken into two related goals: accurate simulation of wind distributions at given locations or predictive wind forecasting of wind given covariates. Where each posits a model for the evolution of wind over time, forecasting models typically write models whose validation is contingent on the quality of predicting the distribution at a time t given the wind observations and possible covariates at times prior to t. On the other hand, when removing the goal of next-time-step prediction, a weather simulator or stochastic weather generator may instead form a more general generating process and validate their precision on asymptotic behavior and fit of marginal distributions on wind behavior. This generating process and the associated variability over space and time is central to wind resource assessment, as such variations largely determine short-scale wind farm generation. The difference in predictive and generative models extends to their treatment of covariates, as a fully generative model would be required to also posit a simulating distribution for any such covariates in the model.

In either case, wind analysis is of great interest in meteorological and climatological studies, including erosion [74], oceanography [66], and insurance [53], as well as implicating significant financial decisions in wind turbine management [64, 34]. In power grid analysis, weather-oriented forecasting approaches directly impact short horizon power prediction and turbine orientation, whereas the climate-based generative models focus on longer-term descriptions of spatial and temporal variance in wind patterns, and impact longer-term turbine construction decisions. A broader

discussion of the role of stochastic weather generators and current modeling techniques can be found in [1].

The first typical choice in wind analysis is the coordinate system within which to describe observations. Speed and direction can be jointly modeled either observations on $\mathbb{R}^2$, where a model on the u and v components of an observed wind speed implies a direction. Alternatively, wind speed and direction can be treated as a polar coordinate pair (S, θ), whose circular-linear joint distribution models the overall behavior. While a u and v model allows for more traditional techniques in multivariate analysis, the directional model often maps better to the intuition of the problem, and calls on a recently growing field of literature in both parametric and nonparametric methods in circular data.

Local wind direction distributions are typically neither unimodal nor symmetric. This puts an inherent constraint on the possible models available. Variations on mixture models of unimodal circular or planar distributions are common mechanisms to account for this multimodality. Where a simple linear mixture of distributions may account for proper cumulative wind distributions, temporal autocorrelation is often encoded into the mixture via use of a Markov process to switch between modes [55, 56, 34]. Machine learning techniques and algorithms have also been explored in the wind context, including attempting to classify temporal autocorrelation through k-means clustering on the band depth of wind time series [80] and neural network learning for spatiotemporal prediction [89]. This work will explore an alternative continuous approach to regimes, one using correlated a Gaussian process on $\mathbb{R}^2$ to induce autocorrelation in the implied wind direction over time $\theta(t)$, with the intention of more accurately capturing the behavior in the vicinity of regime switches in the Markov models. In addition, embedding the temporal autocorrelation into a projected multivariate normal leverages the considerable literature on correlated Gaussian processes, which can provide fast computational methods for both simulation and parameter estimation.

Once a suitable correlated model for wind direction has been validated, we consider the conditional model of wind speed S given the directions. For any given bandwidth of wind direction, the distribution of wind speed may vary significantly in both mean and variance. Further, wind

speed is often modeled a right-skewed distribution, suggesting significant transformations or use of longer tailed distributions would be necessary to posit a parametric model. Instead, we will turn to a nonparametric model for the conditional mean and variance, and combine these moment estimates with the parametric assumption of Gaussian autocorrelation in the residuals.

3.1.1 Data

We consider 20 Oregon and Washington locations along the Columbia River gorge, at which the Bonneville Power Administration (BPA) calculates and reports 5-minute averages of wind speed (in meters per second), wind direction, pressure, and temperature. This data set is appealing in a variety of ways, as the region has been analyzed in some depth from both a meteorological angle due to its local topology and pressure gradients [11, 86, 72], and from a statistical angle due to disparate local multimodal structures for wind direction. These sites in particular were also considered by Hering et. al, 2015 from a stochastic weather generation standpoint.

Treatment of seasonality varies from analysis to analysis, and rather than fully explore the efficacy of a regressive model for season we choose to explore only the spring months, and examine the around 25000 time series recordings in March, April, and May 2013. The data as available from the BPA requires minor processing, as many locations either contain a significant amount of missing data, or have full reporting but some questionable time periods where a single observation is repeated over multiple reporting periods. Only around half of the locations reporting missing information, so we constrained our analysis to those 10 with full reporting, focusing on 3 cases with differing shapes in marginal wind direction distribution. When wind observations were repeated, the data points were removed from the analysis. These 3 and 10 locations are shown in Figure 3.2.

Figure 3.2 shows the resulting data, with the top two rows corresponding to the Kennewick location and the bottom two to the Wasco location. For each site clockwise from the top-left, the full set of u and v observations, the marginal distribution of wind direction θ, the marginal distribution of wind speed S, and the first 15 days of θ directions against time are displayed. We note the two modes of θ, the non-normality of u, v, and S, and the high amount of temporal auto-correlation as

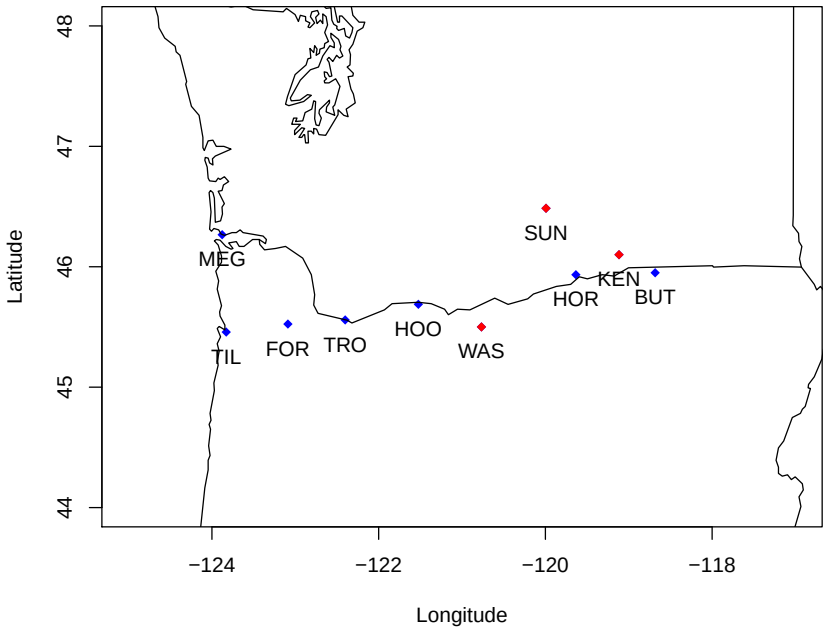

Figure 3.1: The 10 sites in Oregon/Washington without missing data; discussed sites in red.

challenges posed. Our approach begins with a model for the wind direction at each location and its associated autocorrelation, so we proceed with a discussion of techniques in modeling directional data.

3.2 Distributions in Correlated Circular Data

The analysis of correlated circular data requires defining alternative metrics to classical measures of covariance. We will define circular-linear correlation [37] and circular-circular correlation [21] such that for length n circular vectors Φ, Θ with entries on $[0, 2\pi]$ and real-valued linear vector X, we have

$$\rho_h = \mathrm{Cor}(X, \Theta) = \frac{\rho^2(\cos\Theta, X) + \rho^2(\sin\Theta, X) - 2\rho(\cos\Theta, X)\,\rho(\sin\Theta, X)\,\rho(\cos\Theta, \sin\Theta)}{1 - \rho(\cos\Theta, \sin\Theta)} \tag{3.1}$$

and

$$\rho_c = \mathrm{Cor}(\Theta, \Phi) = \frac{E\left(\sin(\Theta - \bar{\Theta})\sin(\Phi - \bar{\Phi})\right)}{\sqrt{E\left(\sin^2(\Theta - \bar{\Theta})\right) E\left(\sin^2(\Phi - \bar{\Phi})\right)}} \tag{3.2}$$

respectively, where ρ denotes the standard Pearson correlation coefficient and e.g. $\cos\Theta$ denotes element-wise cosine. $\bar{\Theta}, \bar{\Phi}$ denote the circular mean [50], given by

$$\bar{\Theta} = \mathrm{atan2}\left(\sum_{j=1}^{n} \sin\theta_j, \sum_{j=1}^{n} \cos\theta_j\right) \tag{3.3}$$

Here the circular-circular correlation measure is defined on $[-1, 1]$, and the circular-linear correlation is supported on $[0, 1]$. atan2 is formally defined as the full circular range extension of $\tan^{-1}$:

$$\mathrm{atan2}(X, Y) = \begin{cases} \tan^{-1}\left(\frac{Y}{X}\right) & \text{if } X > 0,\ Y \geq 0; \\ \tan^{-1}\left(\frac{Y}{X}\right) + \pi & \text{if } X < 0 \\ \tan^{-1}\left(\frac{Y}{X}\right) + 2\pi & \text{if } X \geq 0,\ Y < 0; \\ \text{undefined} & \text{if } X = 0,\ Y = 0; \end{cases}$$

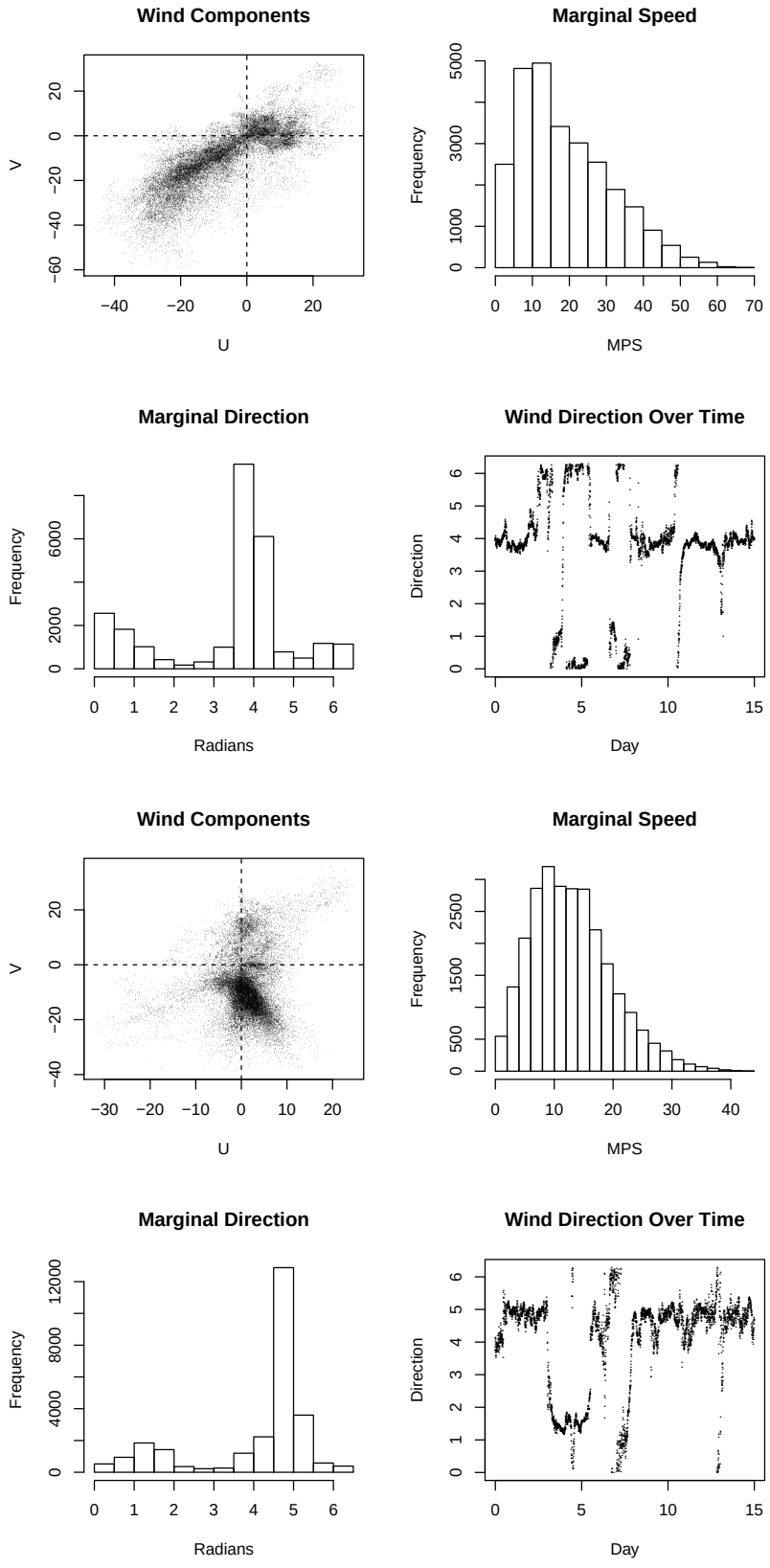

Figure 3.2: Wind Data for Kennewick (top 4) and Wasco

There are three general mechanisms to induce these forms of circular correlation onto data. A first option is to regress the parameters of a circular distribution onto other autocorrelated variables, allowing the mean and/or dispersions of the circular distribution to shift according to the covariates available in the problem. As we are here motivated in the context of stochastic weather generation, this approach will not be heavily employed. A second technique writes the correlation of the model within discrete states, and then allows for some Markov-switching behavior to toggle between states. Within-state similarity will induce some level of autocorrelation, and passing the transitions into an autoregressive process may further add short-term correlation, as in [34].

The option approached here builds a separable multivariate covariance directly onto the circular distribution of the observed angles. This process has been applied to a variety of circular distributions, including the Von Mises [48], wrapped Gaussian [38], and projected Gaussian [84, 56]. Some elaboration on common circular distributions and the techniques employed to induce correlation onto them is appropriate.

3.2.1 Regressive Models

Circular distributions can loosely be broken down into a few categories. First are models unique to the unit circle, such as the Von Mises or cardioid distributions. The Von Mises is a two parameter distribution with density given by

$$f(\theta) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa \cos(\theta - \mu)}$$

with mean parameter $\mu \in [0, 2\pi]$ and concentration parameter $\kappa > 0$ where I_0 denotes the Bessel function of the first kind of order 0. The Von Mises is symmetric about $\theta = \mu$, can be simulated quickly, and robust theory on hypothesis testing has been discussed and developed [51]. For models on directional data, the Von Mises is a common model for the variance or residuals of the process [20]. Correlation can be included in a Multivariate Von Mises by including a symmetric matrix of parameters measuring pairwise conditional dependence of angles analogous to the standard Gaussian covariance matrix Σ. Typical simulation algorithms for the Multivariate Von Mises are

MCMC Gibbs samplers, and in simulation studies require large sample sizes to recover accurate parameter estimations [48]. The Von Mises also lends itself to non-symmetric models in either its generalized form generated by conditioning an isotropic bivariate Gaussian to the unit circle [27, 58], or via mixtures of standard Von Mises random variables.

In addition to circle-specific densities, most densities on $\mathbb{R}$ can be extended to the unit circle by being taken modulus 2π. In particular, the wrapped Gaussian and wrapped Cauchy are symmetric such distributions. The wrapped Gaussian has been used in modeling both wave directions [38] and in a conditional autoregressive model for joint wind speed and direction over a discretized spatial domain [57]. Typically, direct evaluation of the wrapped normal involves estimation of an infinite sum over all the possible modulus numbers, which pushes these approaches towards MCMC samplers.

Traditional bivariate densities on $\mathbb{R}^2$ can also be projected onto the unit circle, wherein the angular component of the polar transformation of the underlying density projects to the circle. The most popular of these is the projected normal, which projects a bivariate normal onto the circle, and planar autocorrelation in the bivariate normal lends itself to a circular autocorrelation as in 3.2. Unlike the prior densities, the projected normal is typically asymmetric, and may exhibit a second mode for more heavily correlated underlying bivariate normals. Further, mixtures of projected normals are dense in among the set of all directional densities, lending itself to a considerable amount of flexibility [84]. The projected normal has been explored for a small set of spatiotemporal data sets, including joint circular-linear wave direction and height [85]. Most work on the projected normal to date has presented modest sample sizes, as estimation is performed with an MCMC slice sampler that requires augmenting the data with the unobserved radial component of the polar coordinate projection for a more tractable joint density [83, 35]. An alternative formulation for conditional parameter estimation of the temporal correlation function after the estimating marginal parameters of the bivariate normal is explored in this work.

As an alternative to projecting a bivariate real-valued density, analysis on wind often constrains itself to modeling the u and v components of the wind observations, which lends itself

naturally to joint estimation of direction and speed. These models may utilize typical mixed model multivariate regression techniques, but have to be quite responsive to non-normality, as wind speeds are typically quite skewed. As a result, transformations to Gaussianity via copulas and nonparametric kernel estimators are techniques we will consider when evaluating either the u and v components of wind or the marginal distribution of wind speed given direction.

Our preliminary explorations in uncorrelated models for the Oregon wind data followed the component-wise models for u and v. The Bonneville data includes a handful of covariates, including temperature and pressure recordings at each station. These allow for a pair of options in directional prediction: either an approximate geostrophic wind estimate, or a regional pressure gradient estimate. Unfortunately, neither approach yielded much success: the pressure observations p are often repeated and of coarse resolution, so numerical estimates for partial derivatives of p with respect to either space or time were numerically unstable and provided little predictive power in regression models. Geostrophic wind is wind flowing along pressure isobars - so orthogonal to the implied pressure gradient - caused by Coriolis. Geostrophic wind can be calculated from the hydrostatic equation for geopotential height Z given by

$$Z = Z_i + \frac{R\bar{T}}{g_0} \ln\left(\frac{p_i}{p_{ref}}\right)$$

where Z_i is the geopotential height of the barometer, R, g_0, and p_{ref} are physical constants, p_i is the observed pressure at the barometer, and $\bar{T}$ is the average temperature from the barometer to the reference pressure p_{ref} (approximated by the observed temperature). Studies in areas of sparse topology have found it among the strongest predictors of ground wind speeds, but the Colombia River gorge's strong topological influences mitigated its impacts here [91]. In the context of simulative weather generators, incorporating a cohesive model for pressure, temperature, wind direction, and wind speed is beyond the scope of this project, so regressions including geostrophic wind were not considered in greater detail.

3.2.2 General Circular-Linear Modeling

We will consider in more detail the hierarchical approach of describing in full the density and correlation structure of wind direction before approaching wind speed. Regression techniques for the expected value of a linear response given a directional predictor began with an appeal to standard linear regression techniques where instead the sine and cosine of the directional variable were used as covariates [50]. In wind modeling, the Trigonometric Direction Diurnal model is one example of such a conditional predictor of wind speed given wind direction via sine and cosine transformations [33][90]. In general, circular-linear models can choose whether to model the variance of the model in the circular domain (e.g. Von Mises) or in the linear domain (e.g. Gaussian), as inverse regression techniques allow for interchangeability of the two to respond to asymptotic concerns to least-squares fits on circular distances [44].

Non-parametric methods to account for the multi-modality and asymmetry of circular variables extend linear kernel density estimation smoothers and additive regression models onto circular-linear prediction. Univariate directional kernel density estimators easily extend to the circular domain [9]. Some care must be given to the bandwidth selection, as rule of thumb [79], cross-validation [31] and bootstrapping [62] can have limitations on data of particular shapes or smaller sample sizes. Alternative forms of the kernel density estimator specific to circular data include those generated from the wrapped Cauchy distribution [14] or from trigonometric polynomials [19].

In the bivariate or circular-linear case kernel estimates provide either a circular-circular or circular-linear form of prediction. Methods have been based on Bernstein copulas [12] and a conditional rule-of-thumb bandwidths on the typical linear kernel [63]. Bivariate circular-linear kernel densities can also be constructed via a transformation of marginal kernel densities [26]. Through use of copulas, this technique has been applied to estimate interdependence of wind directions and chemical concentrations [25], but little work has included (parametric or non-parametric) spatiotemporal correlation while simultaneously couching the linear-circular dependence or marginal densities in terms of kernel smoothers.

In a few cases, joint distributions between projected normal variables and linear random variables have been explored. [85] explored a joint model for wave heights and wind directions, including dually separable correlations over space and time. They focus on developing a directional kriging method on moderately sized sample where slice sampling is efficient with a goal of next-time step forecasting of the full joint model.[56] jointly models wind speeds and directions segmented across a handful of months, otherwise including no formal inclusion of a temporal correlation; they instead opt for a Markov process that swaps regimes based on the mode of the wind direction. A broader discussion of such regime-based methods follows.

3.2.3 Regime-Switching models

Regime-switching models are particularly appealing in wind analysis, as they can easily be scaled to adjust in regimes according to the modes of the joint observed wind speed and directions. Markov-switching autoregressive models were introduced to include a Hidden Markov Model (HMM) in econometric time series [32], and have since been adapted into wind time series modeling, particularly for the purpose of stochastic weather generation. [4] introduced the first such a Markov-Switching Autoregressive model for wind time series. In general, such models follow a similar structure: there exists a latent variable for wind regimes, and within each regime, wind distributions evolve as an AR(p) process with Gaussian innovations. Gamma distributed innovation structures have also been explored, but may often pose computational difficulties which are not mirrored in the Gaussian case [3]. In the case of a joint prediction of speed and direction, the Markov-Switching Vector Autoregressive models (MSVAR) create a joint model for the u and v components of observed wind speeds and couch the within the HMM framework. Both [34] and [2] use such a model, with the former preceding their analysis with a GAM regressive model for seasonal variability and the latter relying on within-state regressions to handle mean behavior within December-January.

In the MSVAR models, short-term temporal correlation is handled by the AR-process with Gaussian innovations, and long-term temporal variability is handled by a probability transition

matrix between wind regimes. Estimation of the number of wind regimes can be handled in a few ways, including directly from wind direction [34, 30] or from a prior model fitting [2]; estimation of autoregressive parameters can be modeled in maximum likelihood [2] or as linear models [34]. In both cases the assignment of observations into clusters has been done with Gaussian Mixture models (e.g. k-means algorithm).

The HMM has also been used to describe regimes on underlying projected normal wind distributions by [55], who create joint estimates for log-transformed wind speeds and directions. The fully Bayesian approach to the HMM estimation requires a fairly slow-mixing sampler [5], and we are particularly concerned about it's efficacy on the higher dimension of high frequency time series.

To our knowledge, HMM models have only been applied on half-hourly [55], hourly [34], and 6-hour [3] wind time series. High frequency series will tend to have more observations in between the modes of directional series; capturing this behavior as well as possible is one intent of this paper. We will provide an alternate approach by using the projected normal without regimes, attempting to create a model that maintains the accuracy in marginal distributions of wind and speed while improving on describing the observed wind directions during transitions between "regimes." We proceed with a contrast between autocorrelation structures in a MSVAR regime model and those implied by a correlated projected normal without regimes.

3.2.4 The Projected Normal

Let $\boldsymbol{Y} = [Y_1, Y_2]'$ be bivariate normal with mean $\mu = [\mu_1, \mu_2]'$, and covariance matrix

$$T = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

Define $\theta := \text{atan2}(Y_1, Y_2)$. Then we say $\theta \sim PN2(\mu, \Sigma)$. Consider a sample $\boldsymbol{\theta} = [\theta_1, \ldots \theta_n]'$ generated from projected normals with components $\boldsymbol{Y}_1 = [Y_1(t_1), Y_1(t_2), \ldots Y_1(t_n)]'$, $\boldsymbol{Y}_2 = [Y_2(t_1), Y_2(t_2), \ldots Y_2(t_n)]$ taken at real-valued times $\boldsymbol{t} = [t_1 \ldots t_n]'$, so $\theta_i = \theta(t_i) = \text{atan2}(Y_1(t_i), Y_2(t_i))$. Denote $\boldsymbol{Y}(t_i) = [Y_1(t_i), Y_2(t_i)]'$.

An isotropic and stationary covariance on the unprojected bivariate normal can be included to induce either spatial or temporal auto-correlation onto $\boldsymbol{\theta}$ by directly correlating the underlying bivariate normals. A simple such covariance function is the fully separable exponential introduced to the projected normal by [84]. In this case, denote $\theta(t) = \text{atan2}\left(Y_1(t), Y_2(t)\right)$, and let $\text{Cor}\left(Y_i(t_1), Y_j(t_2)\right) = k(|t_1 - t_2|)$ for $i, j = 1, 2$ be an isotropic stationary temporal correlation function.

Then for our time series circular points $\boldsymbol{\theta}$,

$$\text{Cov}\left(\boldsymbol{Y}(t_i), \boldsymbol{Y}(t_j)\right) = k(|t_i - t_j|) \cdot T \tag{3.4}$$

or the joint correlation matrix for $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ over the full sample is given by the Kronecker product $T \otimes D$:

$$\Sigma = \begin{bmatrix} C(\boldsymbol{Y}_1, \boldsymbol{Y}_1) & C(\boldsymbol{Y}_1, \boldsymbol{Y}_2) \\ C(\boldsymbol{Y}_2, \boldsymbol{Y}_1) & C(\boldsymbol{Y}_2, \boldsymbol{Y}_2) \end{bmatrix} = \begin{bmatrix} \tau^2 D & \rho\tau D \\ \rho\tau D & 1D \end{bmatrix}$$

for $D_{(i,j)} = k(|t_i - t_j|)$.

The resulting circular-circular correlation does not exactly match the exponential correlation given by an $\mathbb{R}^2$ bivariate model. [84] found that instead, the observed circular-circular correlations are strictly lower than the linear correlations in the Gaussian process. However, the general shape of the curve and falloff with distance remain, as well as the general interpretation of the range parameter, allowing us to call on the considerable existing work on Gaussian process regression and kriging. In addition, the exponential covariance function exhibits a steeper slope of the covariance function near $|t_i - t_j| = 0$ for the circular-circular field, a result we will consider in evaluating the smoothness of the spatial correlation between observed wind regimes.

3.3 The Projected Normal on High-Frequency Data

The $PN2$ has a density given by

$$p(\theta|\mu, T) = \left(\frac{1}{2\pi A(\theta)}\right) |T|^{-1/2} \exp(C) \left[1 + \frac{B(\theta)}{\sqrt{A(\theta)}} \frac{\Phi\left(\frac{B(\theta)}{\sqrt{A(\theta)}}\right)}{\varphi\left(\frac{B(\theta)}{\sqrt{A(\theta)}}\right)}\right] I_{[0,2\pi]}(\theta) \tag{3.5}$$

with $u^T = (\cos\theta, \sin\theta)$, $A(\theta) = u^T T^{-1} u$, $B(\theta) = u^T T^{-1} \mu$, $C(\theta) = -\frac{1}{2}\mu^T T^{-1} u$, and Φ, φ the standard normal distribution and density functions, respectively.

Most work on the subject has found this formulation unwieldy, and instead work with the joint distribution of r, θ, where r is the radial distance $\sqrt{Y_1^2 + Y_2^2}$. This has the advantage of scaling well into larger spherical dimensions [35], and is typically accessible via Gibbs Samplers, as with conjugate priors on T and μ the marginal distributions for the sampler are all normal or inverse-gamma [83].

For parameter identifiability we typically assume covariance matrix for Y_1, Y_2 of the form

$$T = \begin{bmatrix} \tau^2 & \rho\tau \\ \rho\tau & 1 \end{bmatrix},$$

as θ is only unique up to $\boldsymbol{Y}/|\boldsymbol{Y}|$, leading to the following process parameters: μ_1, μ_2, τ, ρ. Any covariates in the process can easily be included in μ_1 and μ_2.

We begin our analysis by positing a model for wind direction based on the projected normal. As empirical circular autocorrelations for the data vary from location to location, the first step is a within-location model for wind direction. For time t in days, diurnal variation is incorporated into as a both day-long and half-day period sinusoids, adding 4 parameters to fit within each mean. Specifically, we have

$$\theta(t) \sim PN2\left([\mu_1(t), \mu_2(t)]', T\right)$$

with

$$\mu_j(t) = \beta_{0,j} + \beta_{1,j}\cos\left(\frac{t}{2\pi}\right) + \beta_{2,j}\sin\left(\frac{t}{2\pi}\right) + \beta_{3,j}\cos\left(\frac{2t}{2\pi}\right) + \beta_{4,j}\sin\left(\frac{2t}{2\pi}\right) \tag{3.6}$$

for $j = 1, 2$ and

$$T = \begin{bmatrix} \tau^2 & \rho\tau \\ \rho\tau & 1 \end{bmatrix} \tag{3.7}$$

is stationary with respect to time.

Properly fitted, this model for θ captures the overall distribution of the wind at any location where a $PN2$ is appropriate, and note that the projected normal formulation can easily be accommodated to locations with any number of modes by including a mixture of projected normals [84, 35].

3.3.1 Stepwise Parameter Estimation

For our purposes, the high frequency of the time observations makes the slice sampling Gibbs samplers suggested by [35] and [54] computationally costly. We found that direct evaluation of the full log-likelihood implied by equation (3.5) was not a significant bottleneck when coupled with optimization from the quasi-Newtonian box-constrained BFGS algorithm in the R `optim` function. L-BFGS-B resolves the MLE of μ_1, μ_2, τ, ρ for a single location of 25,000 observations in 17 likelihood evaluations averaging 0.18 seconds each on a personal laptop, where constraints were only put to ensure $-1 < \rho < 1$. Including the regressors in (3.6) increases the time-per-likelihood evaluation to around 3 seconds, and the MLE optimization requires between 400-500 such evaluations to resolve the 12 parameters in that model for each of the 5 tested locations.

The resulting $PN2$ model includes no direct autocorrelation over time beyond that implied by diurnal variation, so after estimating the overall distribution of wind over time, we appeal to the model in (3.4) to include the separable Gaussian model for temporal autocorrelation. Because the $Y_1(t)$, $Y_2(t)$ underlying $\theta(t)$ are Gaussian and the temporal observations are uniformly spaced, FFTs and methods to exploit gridded autocorrelation in Gaussian processes are available to quickly operate on the $n \times n$ covariance matrices implied by a temporally autocorrelated model. The R package `RandomFields` includes many such methods, with a handful of models to specify bivariate autocorrelation. In general a variety of commonly used valid cross-correlation structures exist. The bivariate Matérn is one such model, generalizing the often used univariate Matérn correlation

structure into bivariate form. It's full specification is given in [29] by:

$$C_{11}(h) = \sigma_1^2 k(h|\nu_1,\, a_1) \tag{3.8}$$

$$C_{22}(h) = \sigma_2^2 k(h|\nu_2,\, a_2) \tag{3.9}$$

$$C_{12}(h) = C_{21} = \rho_{12}\sigma_1\sigma_2 k(h|\nu_{12},\, a_{12}) \tag{3.10}$$

where C_{11} and C_{22} are the covariances between time observations of $C(\boldsymbol{Y}_1(t), \boldsymbol{Y}_1(t))$ and $C(\boldsymbol{Y}_2(t), \boldsymbol{Y}_2(t))$ respectively and $C_{12} = C_{21}$ is the cross-covariance $C(\boldsymbol{Y}_1(t), \boldsymbol{Y}_2(t))$. In our case, σ_2 is constrained to be unity, and we take k to be the Matérn correlation function given by

$$k(d) = \frac{2^{1-\nu}}{\Gamma(\nu)} (ad)^{\nu} K_{\nu}(ad) \tag{3.11}$$

In general, there are constraints on ρ and the cross-correlation terms in C_{12} for (3.8) to represent a valid covariance model. For our purposes, there is no reason to suspect differing correlation parameters between the Y_1 and Y_2 components of the $PN2$, suggesting a model with range coefficient $a = a_1 = a_2 = a_{12}$ and smoothness coefficient $\nu = \nu_1 = \nu_2 = \nu_{12}$. This simplification is one resolution of the necessary constraints on cross-correlation that also reduces the parameter structure down to a pair of interpretable coefficients. For the purposes of circular auto-correlation, we can think about very high frequency variability being captured largely in ν behavior, whereas the longer-term temporal range a helps capture length-of-storm or time-within-regime of the wind observations.

For simulation of the full process, we are in fact simulating $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ from a process with a full covariance matrix given by the product $D \otimes T_{PN}$ for $D_{(i,j)} = k(|t_i - t_j|)$ and T_{PN} the marginal covariance matrix for the projected normal in Equation (3.7),

$$\begin{bmatrix} \tau^2 D & \rho\tau D \\ \rho\tau D & 1D \end{bmatrix}. \tag{3.12}$$

We estimate a and ν for each location via a grid search, considering smoothness values from 0.2 to 2 and spatial ranges from $a = 15$ to $a = 50$. Our measure for similarity in correlation

structure is couched in the circular-circular correlation definition in equation (3.2). In particular, we can construct a circular autocorrelation function by calculating $\rho_c(l, \mathbf{\Theta}) = \text{Cor}(\mathbf{\Theta}, \mathbf{\Theta}_l)$ where $\mathbf{\Theta}_l$ is the time series for Θ right-shifted by l indices/time-steps.

Because there is significant variability in the observed circular autocorrelation functions (CACF) for simulated data sets, for each a, ν pair we simulate 10 data sets and consider the ℓ_2 difference to the first n^* entries of the circular autocorrelation function of the observed data. For each simulated data set $\mathbf{\Phi}$ and true data set $\mathbf{\Theta}$, this is:

$$|\mathbf{\Theta} - \mathbf{\Phi}|_{cacf} = \sqrt{\sum_{l=1}^{n^*} \left(\rho_c(l, \mathbf{\Theta}) - \rho_c(l, \mathbf{\Phi})\right)^2}$$

Taking the average over those ℓ_2 differences at $n^* = 300$ for each parameter pair and smoothing slightly for better visualizing gives us a surface like that in Figure 3.3. This has a ridge defining the minimum values that decreases in a as θ increases, with a general shape mimicking the behavior discussed in Section 2.1.2. We choose the parameter pair corresponding to the minimum value on the smoothed surface for our simulations, but acknowledge that additional criteria could be included to create a more nuanced difference between short-term and long-term autocorrelation.

3.3.2 Single-location validation

We compare our wind direction analysis to that generated by [34], as like ours, their analysis focuses on the overall distribution and autocorrelation at the sampled locations and not forward-time prediction. For single-location validation, [34] suggest that in addition to overall distributions of wind direction and speed, models can be tuned based on temporal autocorrelation of the u and v components of the wind vectors, proper modeling of diurnal variability, the joint distribution of speed and direction, and direct correlation of the u and v components. Because we begin with an attempt to capture direction independent of speed, our model can not directly compare on u, v, and we instead rely on measures of circular autocorrelation as in (3.2).

A summary of the single-location Markov-Switching vector autoregressive model proposed by [34] (the "HKK" model hereafter) and our implementation of it is as follows:

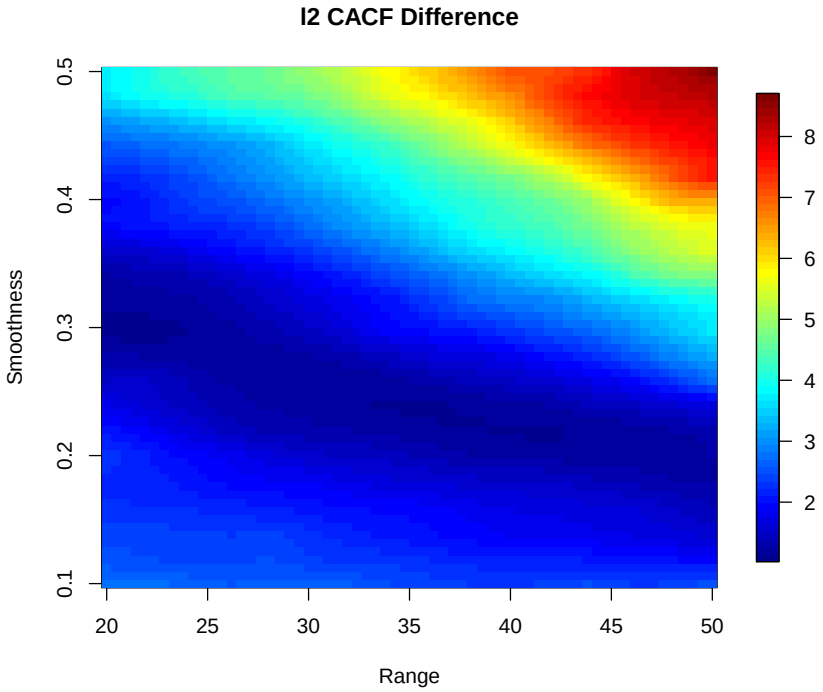

Figure 3.3: Grid Search of Matérn Parameters

(1) Take the u, v components of the observed winds over time, and independently transform them into normality via a Gaussian copula. This involves calculating the empirical cumulative density functions of each component $\hat{F}_u(u)$, $\hat{F}_v(v)$, and then generating normally distributed u', v' via $u' = \Phi^{-1}(\hat{F}_u(u))$ and $v' = \Phi^{-1}(\hat{F}_v(v))$, where $\Phi^{-1}(\cdot)$ is the standard normal inverse cdf.

(2) Create a regressive model for diurnal variability. HKK models the entire year with a GAM of the form $u' = \beta_0 + s(m_t) + s(d_t) + s(h_t) + \varepsilon$, including coefficients for month, day, and hour. Since we only consider a 3 month spring interval, we eschew the longer terms here and apply the HKK and our own with an identical diurnal model of the form in (3.6). Denote the detrended residuals $\hat{u}, \hat{v}$.

(3) Estimate the number of regimes R for Markov-switching, defined as the modes of the joint distribution of speed and direction. The paper does not specify the angular resolution of the wind roses used, so we use 15° windows.

(4) Use an unconstrained GMM cluster in the R package `mclust` to classify each of the u', v' into a corresponding regime. We were unable to replicate this optimization in a true unconstrained fashion. In particular, when the observed wind rose had modes that were nearly antipodal (as in e.g. Kennewick), the GMM would map a single ellipse to cover both modes and apply the second regime orthogonal to it to capture any tertiary behavior.

This suggests some ad hoc classification may be necessary to accuracy capture the intuition behind a regime-switching model based on unconstrained Gaussian mixtures, as the default behavior in `mclust` maps to neither an appropriate autocorrelation structure nor to the underlying physics of the problem. As a first workaround, we applied the GMM onto a Cartesian projection of the observed wind speed and direction, incorporating a phase shift to put the observed least likely direction on the 0 axis and effectively forcing that angle to demarcate a regime shift. We also included strong priors on the mean directions of each observed directional mode. Even with these changes, the clustering model often behaves

poorly, as displayed in Figure 3.4. In order to maintain as accurate of a classification for our comparisons, subsequent analysis instead used an angular mixture model based on Von Mises distributions. For this model, we also used an unconstrained mixture with components equal to the number of observed modes as implemented in the R package `BAMBI`, see [13] for more details.

(5) Estimate parameters for the detrended model $[\hat{u}(t_i), \hat{v}_(t_i)] = A_r[\hat{u}(t_{i-1}), \hat{v}(t_{i-1})] + \boldsymbol{\varepsilon}_r$, where A_r is a 2×2 matrix with unique coefficients for each regime $r = 1, 2, \ldots, R$ and $\boldsymbol{\varepsilon}_r \sim N(0, \Sigma_r)$ is an innovation matrix for each regime.

(6) Estimate an empirical transition probability matrix from the identified clusters.

(7) Simulate a detrended time series u', v' from the parameters and transition probabilities implied in (5) and (6).

(8) Add back in trend from (2), then transform back to non-normality according to the ecdf in (1) via $\hat{F}_u^{-1}(\Phi(u)); \hat{F}_v^{-1}(\Phi(v))$.

The uncorrelated $PN2$ resulting from an initial `L-BFGS-B` optimization can capture the overall distribution of wind directions quite well, as long as the number of modes does not exceed 2. In Figure 3.5, we compared simulated projected normal angular distributions to the observed wind directions at 3 locations. For the Wasco and Kennewick locations, the PN2 does quite well. However, in the Sunnyside location, the PN2 fails to capture the third mode, and as a result misses a significant amount of behavior. Such a location could be captured by including a mixture of projected normals, but this is not explored in more detail here.

We can formalize the quality of the fit if we bin the observed and simulated θ every 15 degrees. An estimate in the difference between distributions is available by summing the squared differences in relative frequencies within bin, or $D_{\ell^2} = \sum_{i=0}^{23}(P(\theta_{sim} \in \Omega_i) - P(\theta_{true} \in \Omega_i)^2)$ for $\Omega_i = [0 + i \cdot 15, 15 + i \cdot 15]$ in degrees. For the three locations shown in Figure 3.5, this gives quality of fit differences shown in Table 3.3.2.

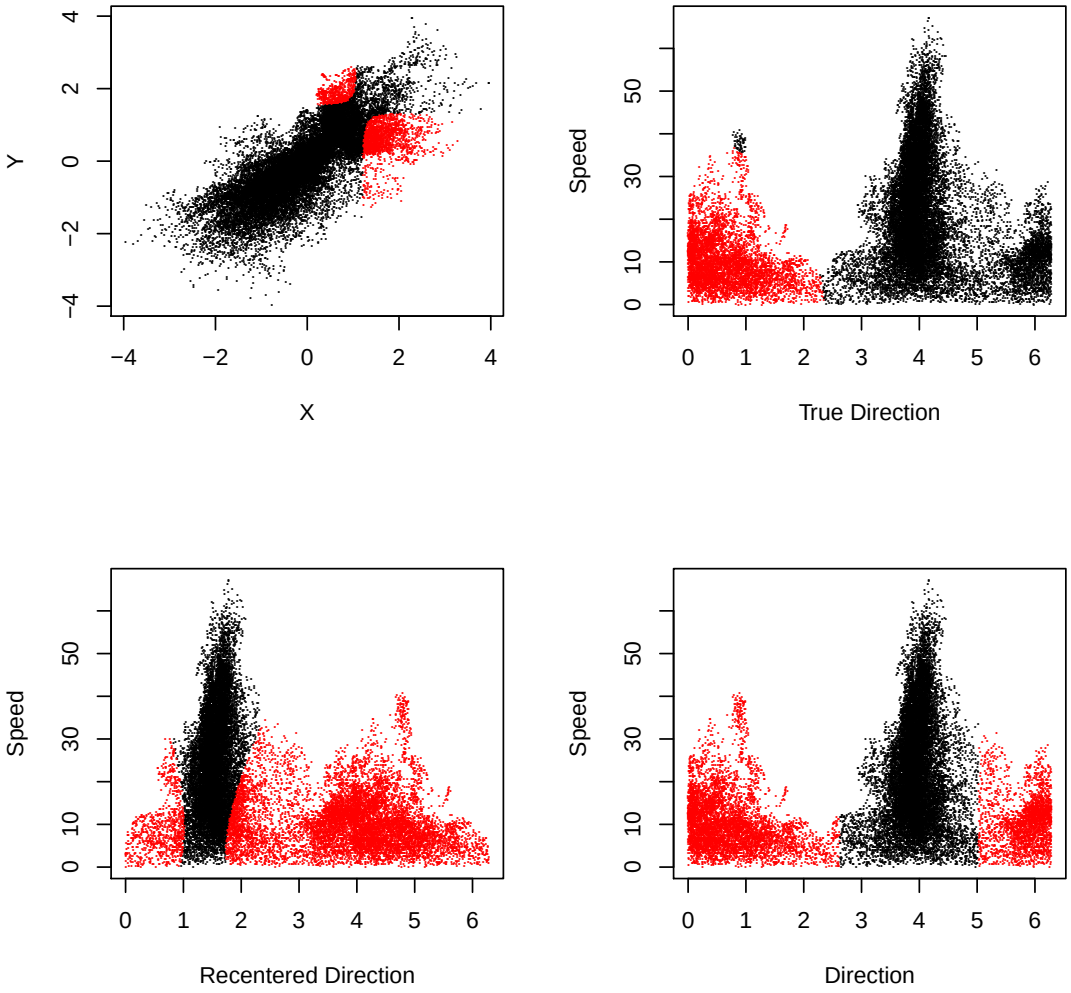

Figure 3.4: MSVAR Classifications for Kennewick Wind Regimes. Clockwise from top-left: unconstrained GMM on u, v; unconstrained GMM on s, θ; unconstrained GMM with priors on both s and shifted θ; unconstrained Von Mises mixture

Location	PN fit	MSVAR fit
Kennewick	0.0028	0.0574
Wasco	0.0024	0.0032
Sunnyside	0.0113	NA

Table 3.1: Root Squared Wind Direction Fits

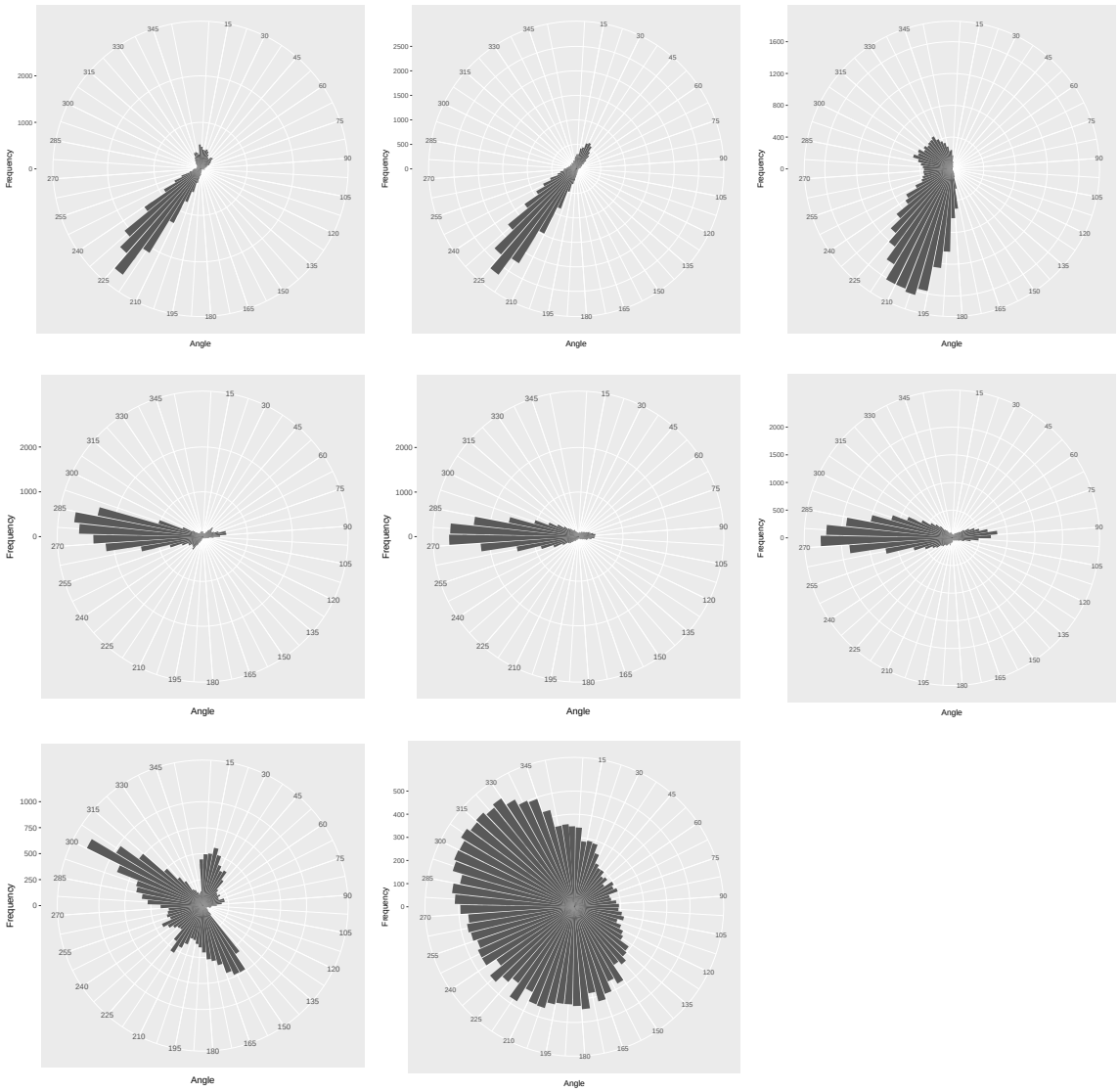

Figure 3.5: Left: True wind directions for entire time period; Center: Simulated via PN; Right: Simulated via MSVAR; Top to bottom: Kennewick, Wasco, Sunnyside

In the Kennewick case, our MSVAR simulations repeatedly generated too wide of angular variance at each mode, leading to a relatively a poor fit by the ℓ^2 histogram measure: we believe this to largely be a shortcoming of the model on the higher frequency data. MSVAR was not tested on Sunnyside, as without the mixture model necessary for the PN model to capture a third mode, we are sure MSVAR would outperform the PN. We leave the specification of covariances on a mixture of projected normals to future work.

We continue to explore Kennewick: a location with an asymmetric, bimodal marginal distribution of wind directions successfully captured by the PN2. Once the correlation structure in (3.4) is included, we can characterize the autocorrelation goodness-of-fit using the summed squared difference in circular autocorrelation functions given by (3.2) for each time lag. Doing so over the first 600 time lags (around 2 days) gives the plots in Figure 3.6, where each functional box plot [77] contains 20 replicates of simulated time series.

The MSVAR and PN2 perform similarly at the shortest scales of time lag. Up to around 3-4 hours of observations, each is capturing the observed data within the interquartile band. After that point, the MSVAR simulations tend to underestimate the true autocorrelation in the data. We speculate that the clustering strategy tends to swap regimes too often, as consecutive points in the time series will be clustered differently solely based on their angular measures. This behavior is visible in the 15 days of simulated data shown in Figure 3.7. The PN2s method of temporal autocorrelation encourages a smoother and more continuous movement from mode to mode, which results in the better fit for time lags from four hours up to the plotted CACF maximum of around a day. Beyond that time point, both models perform similarly, as a large portion of the long term circular autocorrelation is accountable to correlation on the diurnal scale, and we initialized each model with the same diurnal regression model.

As an added benefit, the PN2 also simulates faster after parameter estimation than the MSVAR: instead of having to sequentially calculate each value as the autoregressive parameters adjust from regime to regime, the temporally correlated projected normal can generate a full wind direction sample from a single multivariate normal simulation. The equal spacing of measure-

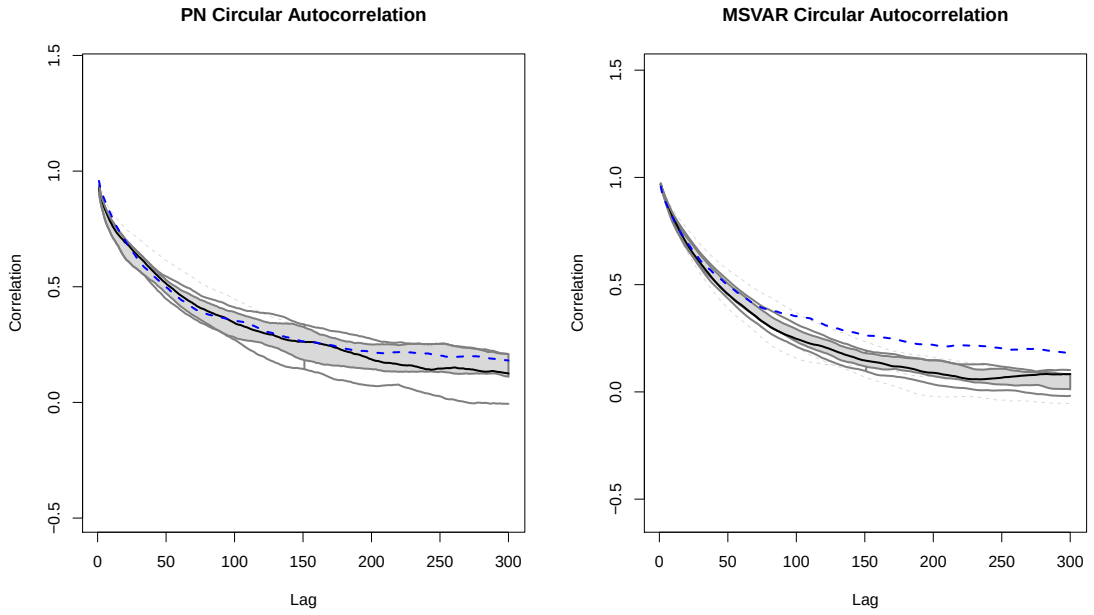

Figure 3.6: Functional Box Plots of Circular ACFs for Kennewick. Left: PN2 with Matérn; Right: MSVAR. Dashed blue line is the true data CACF; interquartile band shaded, and median CACF actualization in black.

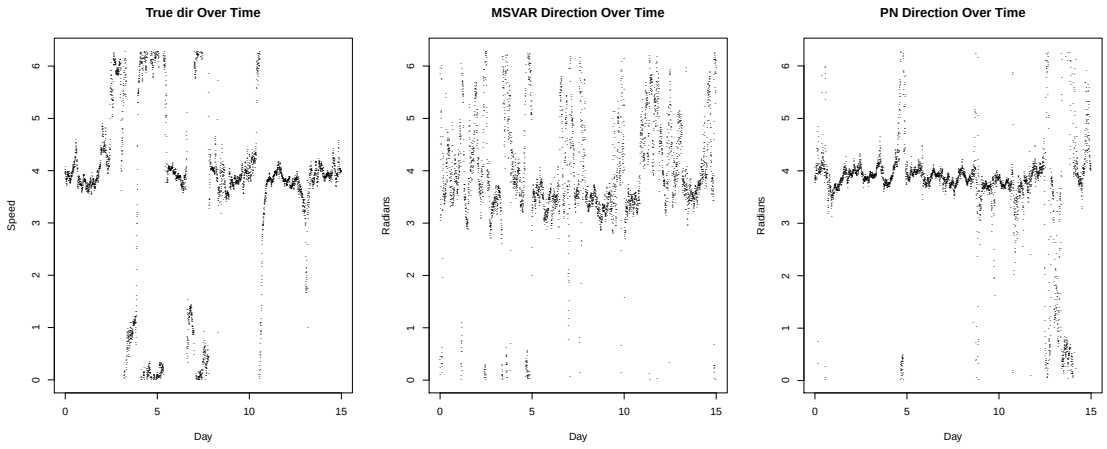

Figure 3.7: Wind directions over time for: true Kennewick data, MSVAR, Matérn PN2.

ments over time allow access to fast computational methods in Gaussian process simulation; the `RandomFields` package in R includes many such options [70].

3.4 A conditional model for Wind Speed

With a suitable fit for both marginal density and correlation structure for wind direction, we begin our discussion of wind speeds with a brief discussion of conditional covariance models. The Matérn cross-covariances used in e.g. Equation (3.8) are part of a broader class of bivariate correlation structures. [67] introduce the idea of joint prediction of multiple covarying processes in a hierarchical framework with a cokriging model for temperature and ozone, in which modeling each component as a linear function of the other allows for valid estimation of their marginal and joint covariances. Typically, the challenge for a hierarchal model where each variable displays autocorrelation is ensuring positive-definiteness in the resulting joint-covariance matrices. The study of valid multivariate covariance functions has been generalized to describe the set of cross-correlation smoothness parameters for a valid multivariate Matérn process of any dimension [29, 28]. For a multivariate Matérn two important bounds exist on the cross-correlation between variables i and j. First, the smoothness $\nu_{i,j}$ of the cross-covariance function is bounded below by the average of the $\nu_{i,i}$ and $\nu_{j,j}$ unless i and j are uncorrelated. The model is most easily generated if both this bound holds and the square of the cross-covariance scale $a^2_{i,j}$ is also bounded below by the average of the squares of the scale parameters for the i and j processes. Given these bounds, the cross-correlation $\rho_{i,j}$ is then bounded above by a function of these parameters; see [29] for a full formulation. A more general set of conditions allows for a broader set of range parameters for 3 or more processes under the "flexible Matérn," but is not implemented here [8].

The conditional framework has been recently explored in more detail by [18]. For a multivariate process linking responses Y_1 and Y_2, we will condition $[Y_1(\cdot), Y_2(\cdot)] = [Y_2(\cdot)|Y_1(\cdot)][Y_1(\cdot)]$. Let the conditional moments take the form

$$E[Y_2(s)|Y_1(\cdot)] = \int_D b(s,v)Y_1(v)dv \tag{3.13}$$

$$\text{cov}[Y_2(s), Y_2(v)|Y_1(\cdot)] = C_{2|1}(s,u) \tag{3.14}$$

for $s \in D$, and b is any integrable function from $\mathbb{R}^d \times \mathbb{R}^d$ into $\mathbb{R}$. Then if $C_{2|1}$ and $C_{11} := \text{cov}[Y_1(s), Y_1(s)]$ are nonnegative definite covariance functions, the model in Equation (3.14) forms a valid cross-covariance model with

$$C_{22}(s,u) = \int_D \int_D b(s,v)C_{11}(v,w)b(u,w)dvdw + C_{2|1}(s,u)$$
$$C_{12}(s,u) = \text{cov}[Y_1(s), E\{Y_2(u)|Y_1(\cdot)\}] = \int_D C_{11}(s,w)b(u,w)dw.$$

For jointly modeling circular-linear variables in such a fashion, [85] condition their wave heights on wind direction in a similar manner, but opt for a simple linear regression on sine and cosine of their wind direction an isotropic and stationary mean zero Gaussian process included in the process covariance.

Equation (3.14) poses a particular challenge for our joint estimation, as it implies a correlation in structure in wind that does not depend on angle. However, the joint distributions in Figure 3.2 clearly demonstrate some conditional dependence in the variance of wind speed on the angle θ. We begin by characterizing this dependence and consider later whether it is possible to recover a valid cross-covariance model. In Figure 3.8 we examine the conditional moments of the wind speed S given θ for the true data at the Kennewick location. We observe both an increase in mean and variance of the wind speed near the two modes of θ, and conclude that under this formulation a $C_{S|\theta}$ satisfying Equation (3.14) to have no functional dependence on θ may not be feasible. However, such a formulation may not be necessary. If we instead couch the cross-covariance between S and θ as a trivariate Matérn on Y_1, Y_2, S for the Y_1, Y_2 of the underlying projected normal, we can ensure valid cross-correlation matrices and later transform the associated random variables into S, θ.

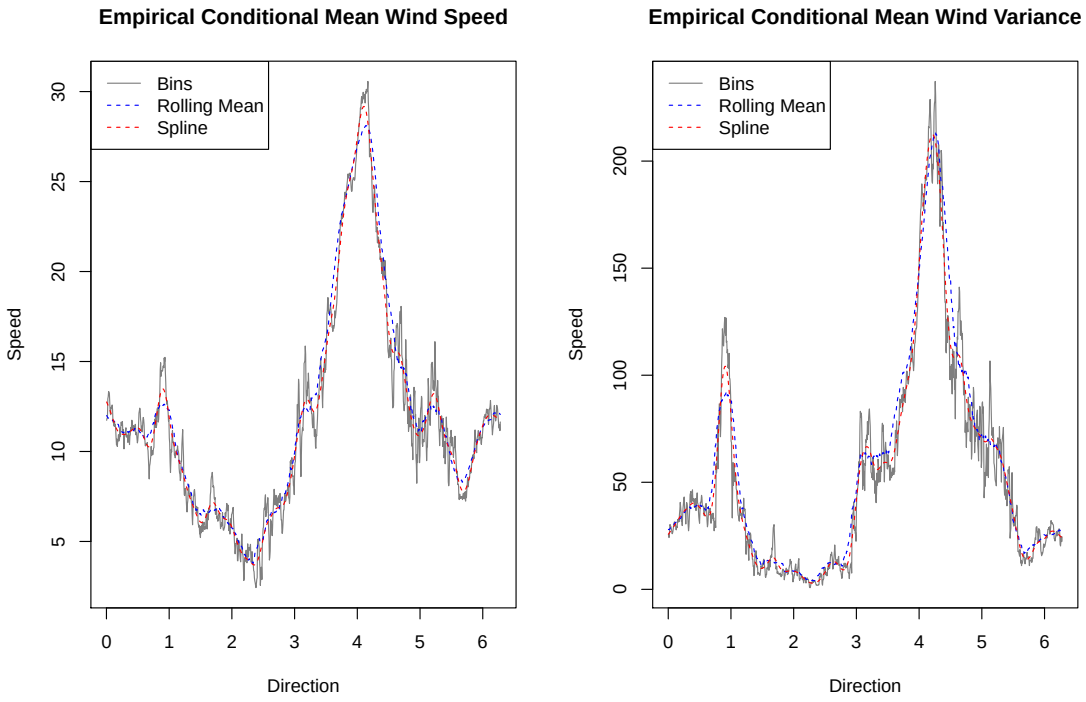

Figure 3.8: Smoothing options for conditionally describing $E[S|\theta]$, $\text{Var}[S|\theta]$.

We use the marginal distributions in Figure 3.8 to create a model for the mean and variance of $S|\theta$ by moment-matching. The plotted black line corresponds to the means and variances of the true data by binning S observations every $2\pi/1000$ radians and calculating the within-bin mean and variance. Due to small samples in some bins, a small amount of smoothing is necessary to provide a reasonable model that we believe describes the process; options of taking a rolling mean over nearby bins (blue) and a smoothing spline over the data repeated in reverse before 0 and after 2π to account for edge effects (red) are considered. The ability of a smoothing spline to interpolate to observations within bins fits the problem well, so this was the model we considered moving forward, using the `smooth.spline` function in the R `fields` package with a smoothing parameter of 0.5.

This moment-matched non-parametric estimate for $S(t)|\theta(t)$ allows us to create a time series for $\boldsymbol{S} = [S(t_1), S(t_2) \dots S(t_n)]'$ satisfying

$$E[S(t)|\theta] = SS_{\mu,\lambda}(\theta(t)) \tag{3.15}$$

$$Var[S(t)|\theta] = SS_{\sigma^2,\lambda}(\theta(t)) \tag{3.16}$$

where $SS_{\mu,\lambda}$ is the smoothing spline with smoothing parameter λ generated on the binned means of the true Kennewick data and $SS_{\sigma^2,\lambda}$ is the corresponding spline on the binned variances. The model

$$S(t) = E[S(t)|\theta] + \sqrt{Var[S(t)|\theta]}\varepsilon(t) \tag{3.17}$$

then defines a time series on S with the appropriate marginal first and second moments.

These transformed data have inherited some autocorrelation from the correlation in θ without formally specifying a cross-covariance, and the errors $\varepsilon(t)$ represent a possible source of autocorrelation in speed that can be estimated independently of direction. For comparison, Figure 3.9 depicts in the left and mid panes the process given by $E[S(t)|\theta]$ compared to the true wind speeds at Kennewick, and visual inspection notes the need for an additional layer of autocorrelation beyond that implied by the mean model. The right pane denotes the model given in Equation (3.17)

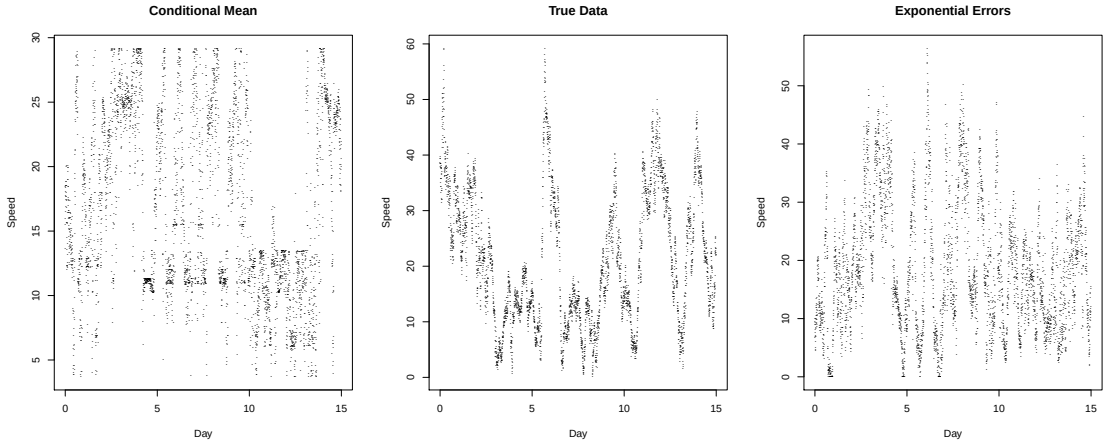

Figure 3.9: Left to Right: Wind Speed (m/s) Time Series for S from $E[S(t)|\theta]$, true data, and $E[S(t)|\theta] + \sqrt{Var[S(t)|\theta]}\varepsilon(t)$

when the residuals are taken to be from a mean zero Gaussian process with exponential covariance satisfying $k(t_i, t_j) = e^{-a|i-j|}$.

3.4.1 Recouching the full model

While the fit in Figure 3.9 provides an appropriate marginal correlation structure on S, the covariances in (3.17) have an implied dependence on θ from the empirical variance of the process, which may in turn imply a non-positive definite covariance structure. We can remedy this by embedding our fit for a into the multivariate Matérn model used to generate the directional time series. Specifically, consider the multivariate Matérn given by

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_X \\ \mu_Y \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} & 0 \\ \Sigma_{YX} & \Sigma_{YY} & 0 \\ 0 & 0 & \Sigma_{ZZ} \end{bmatrix} \right) \tag{3.18}$$

where

$$\begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix} \tag{3.19}$$

is the same matrix given in Equation (3.12) and $\Sigma_{ZZ}(i,j) = e^{-a|i-j|}$. Then set

$$\theta = \text{atan2}\left(\frac{Y}{X}\right) \tag{3.20}$$

and

$$S = SS_{\mu,\lambda}(\theta) + Z\sqrt{SS_{\sigma^2,\lambda}(\theta)} \tag{3.21}$$

and equations (3.18)-(3.21) define our full model for S, θ. A simulation from this process is shown in Figures 3.11 and 3.12. We note that it not only well captures the marginal densities in the top row of each figure, the relationship over time is well captured in the second row, a process we formalized with the CACF for θ in Figure 3.6 and display in the form of a standard ACF in the last row of 3.11. Figure 3.10 displays the ACFs of the true and simulated data, demonstrating in particular that the correlateion in wind speed of the MSVAR model is not captured at longer time lags. The last row of Figure 3.12 shows the composite model's empirical joint density of S, θ compared to that of the true data.

Revisiting the whole algorithm, our process can be replicated as follows:

(1) Given observations S, θ estimate the Projected Normal parameters in equations (3.6) and (3.7) by maximum likelihood.

(2) Simulate Projected Normals from the distribution in (1) and estimate the Matérn correlation parameters a_θ, ν_θ for the temporal variability in θ by optimizing over the difference between the CACFs of the simulated data sets and the CACF of θ.

(3) Create the smoothed functions $SS_{\mu,\lambda}, SS_{\sigma^2,\lambda}$ for the empirical mean and variances of $S|\theta$.

(4) Estimate by maximum likelihood or fix the correlation parameter ν_S corresponding to the time series of S ($\nu_S = 0.5$ represents the exponential used above).

(5) Simulate the full process S, θ from (3.18)-(3.21).

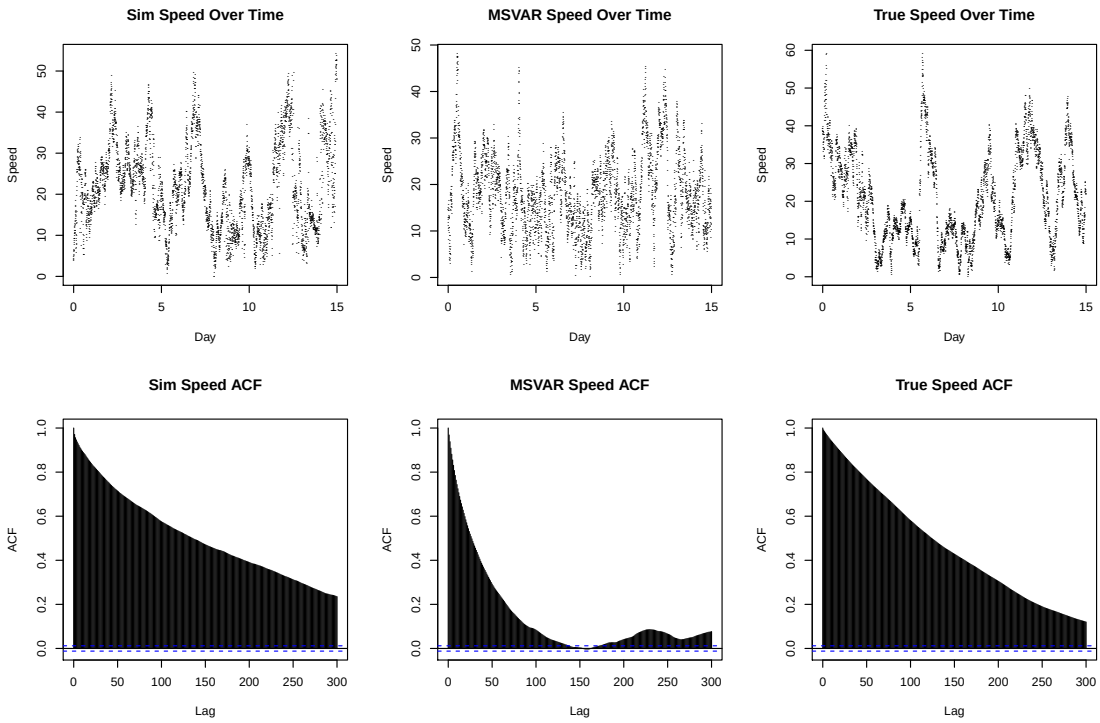

Figure 3.10: Left: Our Model; Middle: MSVAR Right: True. The first 300 lags of wind speed (m/s) ACFs for the full 3 months.

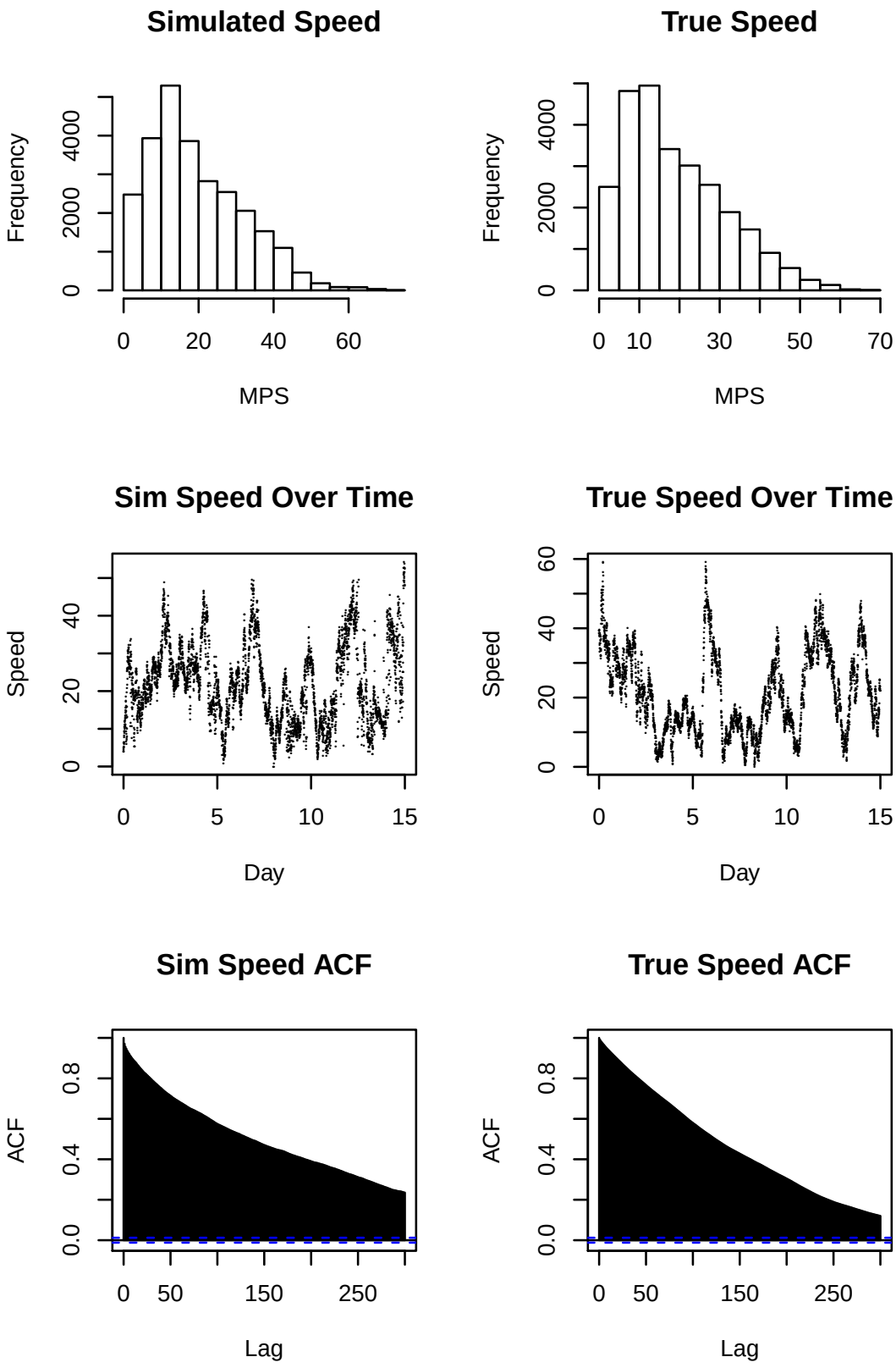

Figure 3.11: Left: Simulated; Right: True; Top to bottom: marginal wind speed, 15 days of speed/time, ACFs S.

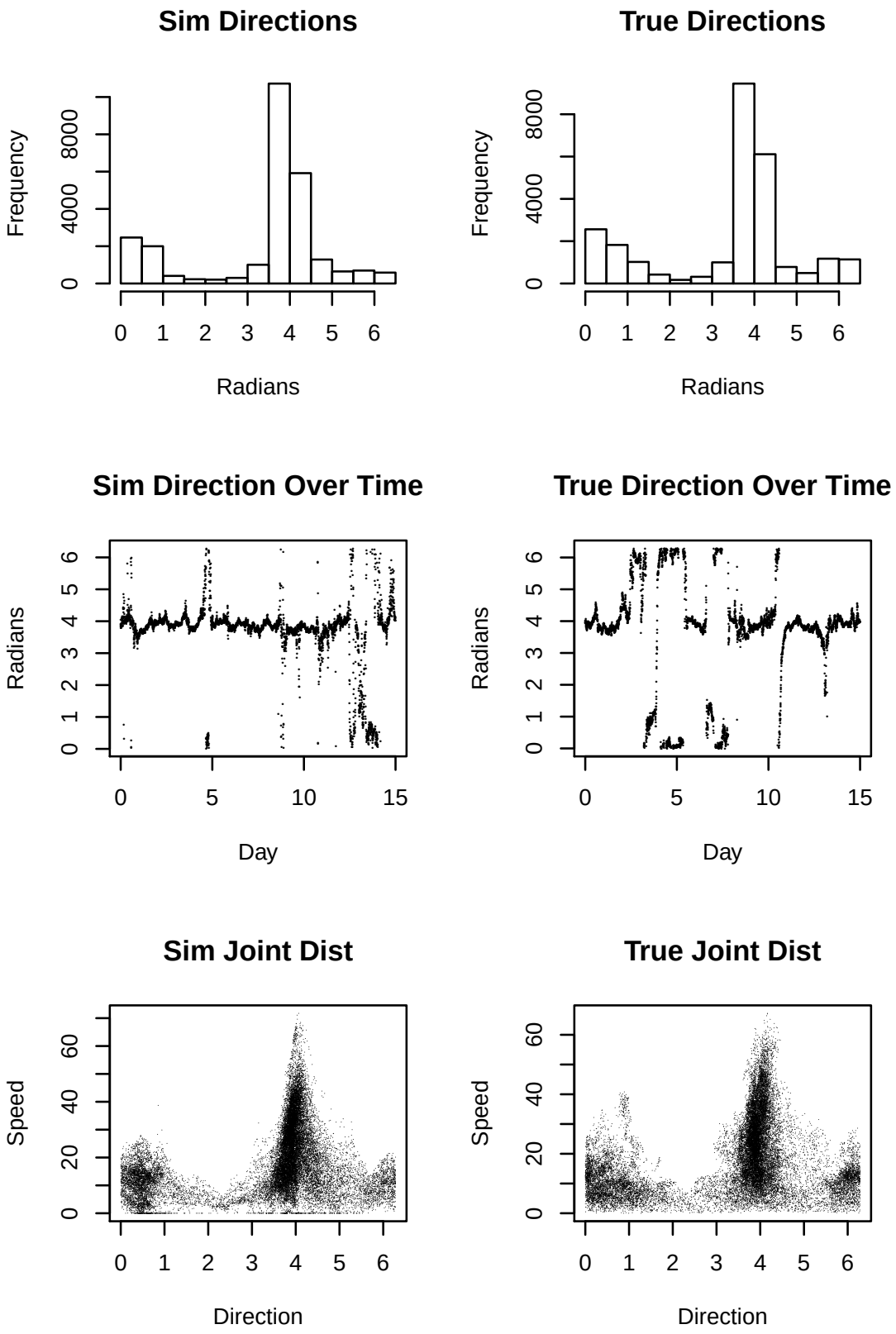

Figure 3.12: Left: Simulated; Right: True; Top to bottom: marginal wind directions, 15 days of direction/time, joint distributions of S, θ.

3.5 Discussion

We have demonstrated a model for joint wind speed and direction at a single location that relies on easy-to-compute marginal parameter estimations and jointly simulates wind speed and directions from the resulting estimation The final model is a set of transformations on a single trivariate parsimonious Matérn model, ensuring valid cross-correlations in the final model despite the hierarchical estimation procedures. This model is fast to simulate through use of `RandomFields`, and the regularly spaced measurements over time invite further computational innovation such as those discussed in Chapter 2. The parsimonious Matérn allows for specifications of varying smoothnesses between processes, but in its simplest form requires the range parameter a be the same between processes.

For future projects, we wish to consider incorporating spatial autocorrelation at multiple locations into the overall model. We believe that a spatial correlation structure could be included directly into the Gaussian process underlying projected normals at each location. In addition, incorporating autocorrelation into mixtures of projected normals is not a well developed theoretically, as some amount of correlation would also have to be encoded into the uniform variable that typically underlies component selection of a mixture model. We speculate that allowing for a model to transition between mixtures at areas of overlapping density would recreate the system appropriately, but do not create such a simulation algorithm here. This understanding would be crucial to generalizing this model to locations with more than 2 directional modes for wind, as the projected normal fit is unable to accurately capture this behavior.

We also suspect that the choice of maximum likelihood for the mean model of the projected normals is only suitable with the many thousands of observations present on a data set such as this. While high frequency observations often imply an abundance of data, it's not clear how well our estimation methods would generalize in such cases, and would perhaps have to turn to Gibbs sampling.

Chapter 4

Conclusion and Discussion

4.1 Equivalent Score Functions

In this project we explored two elements of the computational Gaussian process. The equivalent kernel project provided a set of approximate score functions to estimate the maximum likelihood for the covariance parameters of a Gaussian process. The equivalent kriging spline approximation in general tends to become fairly inaccurate if there are large gaps in data observations, so we utilized only gridded data, which also opened up the computational benefits of the fast Fourier transform. The resulting method could generate considerable time savings on both one and two dimensional data, although its asymptotic accuracy required a considerable sample size. The sample size constraints appeared to be per-dimension, suggesting that the slower to converge approximation on the quadratic form should only be used on data sets with as many as 10^5 gridlines per spatial dimension. However, the boundary-less trace term was less dependent on sample size, and provided accurate estimates of both the true trace term and the resulting MLEs at sample sizes under 1000 per spatial dimension. We look forward to seeing future researchers attempt to use these estimators when applicable.

Future work directly on the equivalent kernel could take a few directions. Our technique scaled well to estimations of the spatial range, nugget, and sill, but did not include any calculations of the sample smoothness. In many cases, it's acceptable to fix the smoothness of the underlying process as half-integer values, but as we demonstrated in the wind directions, at times it is appropriate to include such a term. The derivative of the Bessel functions within the Matérn covariance with

respect to smoothness can be algebraically tedious, but a full estimation of the four covariance parameters would be an excellent follow-up work.

The underlying theory of the equivalent kernel in multiple dimensions is also not fully developed. The theorems granting proper bounds on the elementwise convergence of the equivalent kernel matrix to the kriging weights have not been fully demonstrated in two dimensions, and a theoretical derivation of the empirical results in this project would be welcome. More generally on the case of multiple dimensions, the equivalent kernel has only been considered in the case of a univariate response variable. We are unaware of whether or not theoretical results exist to bridge the spline variational problem to Gaussian processes exhibiting cross-covariance, but if so it would be an excellent progression in the equivalent kernel literature. Our univariate explorations were also only applied in either space or time; a spatiotemporal equivalent kernel would be an powerful development given the added sample size of such data.

In the univariate case, our brief explorations with basis representations also open some doors for closer inspection. The Gaussian Markov Random Field formulation is one with appeal due to its ability to capture non-stationary behavior through the varying stochastic coefficients. We concluded that use of the equivalent kernel on these problem was an equivalent form to fixed rank kriging, but are interested in whether or not other options to use the equivalent kernel to capture non-stationarity exist.

4.2 The Autocorrelated Projected Normal

We created a hierarchical model for estimating and simulating from the joint densities of wind speed and direction. The final model was fast to simulate, as it required only simulation of a trivariate Matérn random variable with length equal to the time observations required. Afterwards, a series of transformations according to the estimated diurnal mean model and conditional wind speed given wind direction yielded a joint time series for the variables of interest. The resulting process appeared much more well suited for high-frequency data than the current standard in stochastic wind generators, as the regime-switching model we implemented struggled both to

properly cluster the additional numbers of points away from the wind modes and to incorporate them well into its transition matrix.

Our method, however, was reliant on a single projected normal, which includes two shortcomings that other models may not. First, it is not entirely clear how to adjust the underlying projected normal to include multimodal data, as we observed at the Sunnyside location. A mixture of projected normals could perfectly capture the marginal density of wind speed, as such mixtures are dense in the set of continuous circular densities. However, a typical mixture model is simulated by simulating a uniform and choosing an underlying mixture to simulate from as a result of that uniform. Such a procedure does not has any clear way to incorporate temporal correlation into the underlying structure, as pivoting between mixtures could lead to odd jumps in the resulting time series. We speculate that a method that allows for transitions between mixtures at areas overlapping density could make sense, but this both requires more exploration and may carry a considerably computational cost, as it may require simulating the data one time step at a time (as the HKK MSVAR does) instead of simulating a single multivariate Gaussian process with autocorrelation over the entire sample explicitly included.

We also only covered the case of a simple spatial location. The projected normal has been used to cover jointly separable temporal and spatial autocorrelation in other works, but not at the frequency of our data. We suspect that the Gibbs sampling schemes used in those works could be used in our own, but some careful thought should be given to the spatial autocorrelation structures. It may make sense to encode spatial covariance directly into the bivariate normals that underly projection, but this may not be fully appropriate to the dynamics of a wind problem. In particular, for nearby sites, wind covariance would often be thought of as a joint rotation, but the projected normal approach loses this interpretation, as an equivalent unit shift in the components of the bivariate normals at two locations will almost certainly not represent an equivalent rotational shift. We believe such a model might be best incorporated conditionally and written in terms of the angular distributions, but the solution to the hierarchical modeling difficulties we discuss is not apparent.

Bibliography

[1] P. Ailliot, D. Allard, V. Monbet, and P. Naveau. Stochastic weather generators: an overview of weather type models. J. de la Société Française de Statistique, 156, 2015.

[2] P. Ailliot, J.. Bessac, V. Monbet, and F. Péne. Non-homogeneous hidden markov-switching models for wind time series. Journal of Statistical Planning and Inference, 160, 12 2014.

[3] P. Ailliot and V. Monbet. Markov-switching autoregressive models for wind time series. Environmental Modelling & Software, 30:92 – 101, 2012.

[4] P. Ailliot, C. Thompson, and P. Thomson. Space-time modelling of precipitation by using a hidden Markov model and censored Gaussian distributions. Applied Statistics, 58:405–426, 2009.

[5] C. Andrieu, A. Doucet, and R. Holenstein. Particle markov chain monte carlo methods. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 72(3):269–342, 2010.

[6] M. Anitescu, J. Chen, and M. Stein. An Inversion-Free Estimation Equation Approach for Gaussian Process Models. Journal of Computational and Graphical Statistics, 26:98–107, 2017.

[7] M. Anitescu, J. Chen, and L. Wang. A matrix-free approach for solving the parametric gaussian process maximum likelihood problem. SIAM Journal on Scientific Computing, 34:A240–A262, 2012.

[8] T. V. Apanasovich, M. G. Genton, and Y. Sun. A valid Matérn class of cross-covariance functions for multivariate random fields with any number of components. Journal of the American Statistical Association, 107:180–193, 2012.

[9] Z. D. Bai, C. Radhakrishna Rao, and L .C. Zhao. Kernel estimators of density function of directional data. Journal of Multivariate Analysis, 27(1):24 – 39, 1988.

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

[11] D. Cameron. Easterly gales in the columbia river gorge during the winter of 1930-1931 - some of their causes and effects. Monthly Weather Review, 1931.

[12] J. A. Carnicero, M. Wiper, and C. Ausín. Density estimation of circular data with bernstein polynomials. Hacettepe University Bulletin of Natural Sciences and Engineering Series B: Mathematics and Statistics, 47, 2018.

[13] S. Chakraborty and S. W. K. Wong. BAMBI: An R package for Fitting Bivariate Angular Mixture Models. arXiv:1708.07804, 2016.

[14] Y. Chaubey. Smooth kernel estimation of a circular density function: A connection to orthogonal polynomials on the unit circle. Journal of Probability and Statistics, 2018, 01 2016.

[15] H. Chen, D. Simpson, and Z. Ying. Infill Asymptotics for a Stochastic Process Model with Measurement Error. Statistica Sinica, 10(1):141–156, 2000.

[16] M. J. Cree and P. J. Bones. Algorithms to numerically evaluate the Hankel transform. Computers & Mathematics with Applications, 26:1–12, 1993.

[17] N. Cressie and G. Johannesson. Fixed rank kriging for very large spatial data sets. Journal of the Royal Statistical Society, Series B, 70:209–226, 2008.

[18] N. Cressie and A. Zammit-Mangion. Multivariate spatial covariance models: A conditional approach. Biometrika, 103, 04 2015.

[19] M. Di Marzio, S. Fensore, A. Panzera, and C. Taylor. Nonparametric estimating equations for circular probability density functions and their derivatives. Electronic Journal of Statistics, 11, 07 2017.

[20] T. D. Downs and K. V. Mardia. Circular regression. Biometrika, 89(3):683–697, 2002.

[21] N. I. Fisher. Statistical Analysis of Circular Data. Cambridge University Press, 1996.

[22] G. Forsythe, M. Malcolm, and C. Moler. Multi-fidelity optimization via surrogate modelling. Computer Methods for Mathematical Computations, 1977.

[23] E. M. Furrer and D. W. Nychka. A framework to understand the asymptotic properties of kriging and splines. Journal of the Korean Statistical Society, 36:57–76, 2007.

[24] R. Furrer, M. G. Genton, and D. Nychka. Covariance tapering for interpolation of large datasets. Journal of Computational and Graphical Statistics, 15:502–523, 2006.

[25] E. García-Portugués, R. Crujeiras, and W. Gonzãlez Manteiga. Exploring wind direction and so2 concentration by circular-linear density estimation. Stochastic Environmental Research and Risk Assessment, 27:1055–1067, 08 2012.

[26] E. García-Portugués, R. Crujeiras, and W. Gonzãlez Manteiga. Kernel density estimation for directional-linear data. Journal of Multivariate Analysis, 10 2012.

[27] R. Gatto and S. Jammalamadaka. The generalized von mises distribution. Statistical Methodology, 4:341–353, 07 2007.

[28] M. G. Genton and W. Kleiber. Cross-covariance functions for multivariate geostatistics. Statistical Science, 30:147–163, 2015.

[29] T. Gneiting, W. Kleiber, and M. Schlather. Matérn cross-covariance functions for multivariate random fields. Journal of the American Statistical Association, 105:1167–1177, 2010.

[30] T. Gneiting, A. E. Raftery, A. H. Westveld III, and T. Goldman. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. Monthly Weather Review, 133:1098–1118, 2005.

[31] P. Hall, G. S. Watson, and J. Cabrera. Kernel density estimation with spherical data. Biometrika, 74(4):751–762, 1987.

[32] J. D. Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. Econometrica, 57(2):357–384, 1989.

[33] A. S. Hering and M. G. Genton. Powering up with space-time wind forecasting. Journal of the American Statistical Association, 105:92–104, 2010.

[34] A. S. Hering, K. Kazor, and W. Kleiber. A Markov-switching vector autoregressive stochastic wind generator for multiple spatial and temporal scales. Resources, 4:70–92, 2015.

[35] D. Hernandez-Stumpfhauser, F. J. Breidt, and M. J. van der Woerd. The general projected normal distribution of arbitrary dimension: Modeling and bayesian inference. Bayesian Anal., 12(1):113–133, 03 2017.

[36] I.. Ibragimov and Y. Rozanov, editors. Gaussian Random Processes. Springer, New York, 1977.

[37] S. Jammalamadaka and Y. Sarma. A Correlation Coefficienct for Angular Variables. In K. Matusita, editor, Statistical Theory and Data Analysis II: Proceedings of the Second Pacific Area Statistical Conference, chapter 10. Elsevier Science Publishers, 1988.

[38] G. Jona-Lasinio, A. Gelfand, and M. Jona-Lasinio. Spatial analysis of wave direction data using wrapped gaussian processes. Ann. Appl. Stat., 6(4):1478–1498, 12 2012.

[39] M. Katzfuss. A multi-resolution approximation for massive spatial datasets. Journal of the American Statistical Association, in press, 2016.

[40] M. Katzfuss and J. Guiness. A general framework for Vecchia approximations of Gaussian processes. arXiv preprint, 1708.06302, 2017.

[41] C. G. Kaufman, M. J. Schervish, and D. W. Nychka. Covariance tapering for likelihood-based estimation in large spatial data sets. Journal of the American Statistical Association, 103:1545–1555, 2008.

[42] C. G. Kaufman and B. A. Shaby. The role of the range parameter for estimation and prediction in geostatistics. Biometrika, 100:473–484, 2013.

[43] K. Key. Is the fast Hankel transform faster than quadrature? Geophysics, 77:F21–F30, 2012.

[44] S. Kim and A. SenGupta. Inverse circular–linear/linear–circular regression. Communications in Statistics - Theory and Methods, 44:4772–4782, 11 2015.

[45] W. Kleiber, R. W. Katz, and B. Rajagopalan. Daily spatiotemporal precipitation simulation using latent and transformed Gaussian processes. Water Resources Research, 48, 2012.

[46] W. Kleiber, R. W. Katz, and B. Rajagopalan. Daily minimum and maximum temperature simulation over complex terrain. Annals of Applied Statistics, 7:588–612, 2013.

[47] W. Kleiber and D. W. Nychka. Equivalent kriging. Spatial Statistics, 12:31–49, 2015.

[48] F. Lagona. Regression analysis of correlated circular data based on the multivariate von mises distribution. Environmental and Ecological Statistics, 23, 10 2015.

[49] F. Lindgren, H. Rue, and J. Lindström. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. Journal of the Royal Statistical Society, Series B, 73:423–498, 2011.

[50] K. V. Mardia. Linear-circular correlation coefficients and rhythmometry. Biometrika, 63(2), 1976.

[51] K. V. Mardia and P. Jupp. Directional Statistics. John Wiley & Sons, Inc., 1999.

[52] K. V. Mardia and R. J. Marshall. Maximum likelihood estimation of models for residual covariance in spatial regression. Biometrika, 71:135–146, 1984.

[53] G. Masala. Wind time series simulation with underlying semi-Markov model: An application to weather derivatives. J. Stat. Manag. Syst, 17, 2014.

[54] G. Mastrantonio. The joint projected normal and skew-normal: A distribution for poly-cylindrical data. Journal of Multivariate Analysis, 165, 11 2017.

[55] G. Mastrantonio, A. Maruotti, and G. Jona Lasinio. Bayesian hidden Markov modelling using circular-linear general projected normal distribution. Environmetrics, 26, 2015.

[56] G. Mastrantonio, A. Pollice, and F. Fedele. Distributions-oriented wind forecast verification by a hidden Markov model for multivariate circular-linear data. Stoch Environ Res Risk Assess, 32, 2018.

[57] D. Modlin, M. Fuentes, and B. M. Reich. Circular conditional autoregressive modeling of vector fields. Environmetrics, 23 1, 2012.

[58] A. Navarro, J. Frellsen, and R. Turner. The multivariate generalised von mises: Inference and applications. arXiv:1602.05003, 02 2016.

[59] D. Nychka. Splines as local smoothers. Annals of Statistics, 23:1175–1197, 1995.

[60] D. Nychka, S. Bandyopadhyay, D. Hammerling, F. Lindgren, and S. Sain. A multiresolution gaussian process model for the analysis of large spatial datasets. Journal of Computational and Graphical Statistics, 24(2):579–599, 2015.

[61] D. Nychka, D. Hammerling, S. Sain, and N. Lenssen. LatticeKrig: Multiresolution Kriging Based on Markov Random Fields, 2016. R package version 5.4-5.

[62] M. Oliveira, R.M. Crujeiras, and A. Rodríguez-Casal. A plug-in rule for bandwidth selection in circular density estimation. Computational Statistics & Data Analysis, 56(12):3898 – 3908, 2012.

[63] X. Qin, J. Zhang, and X. Yan. A nonparametric circular–linear multivariate regression model with a rule-of-thumb bandwidth selector. Computers & Mathematics with Applications, 62(8):3048 – 3055, 2011.

[64] F. Raischel, T. Scholz, V. Lopes, and P. Lind. Uncovering wind turbine properties through two-dimensional stochastic modeling of wind dynamics. Phys. Rev. E, 88, 2013.

[65] C. Rasmussen and C. Williams. Gaussian Processes for Machine Learning. MIT Press, 2005.

[66] B. J. Reich and M. Fuentes. A multivariate semiparametric Bayesian spatial modeling framework for hurricane surface wind fields. Annals of Applied Statistics, 1:249–264, 2007.

[67] A. Royle and L. M. Berliner. A hierarchical approach to multivariate spatial modeling and prediction. Journal of Agricultural, Biological and Environmental Statistics, 4:1–28, 1999.

[68] H. Rue and L. Held. Gaussian Markov Random Fields: Theory and Applications. Boca Raton: Chapman & Hall/CRC, 2005.

[69] T. Santner, B. Williams, and W. Notz. The Design and Analysis of Computer Experiments. Springer Verlag, New York, 2003.

[70] M. Schlather, A. Malinowski, P. Menck, M. Oesting, and K. Strokorb. Analysis, simulation and prediction of multivariate random fields with package randomfields. Journal of Statistical Software, Articles, 63(8), 2015.

[71] D. Shanks. Nonlinear transformations of divergent and slowly converging sequences . Journal of Mathematical Physics, 34:1–42, 1955.

[72] J. Sharp and C. Mass. Columbia gorge gap winds: Their climatological influence and synoptic evolution. Weather and Forecasting, 19, 2004.

[73] B. W. Silverman. Spline smoothing: the equivalent variable kernel method. Annals of Statistics, 12:898–916, 1984.

[74] E. L. Skidmore and J. Tatarko. Stochastic wind simulation for erosion modeling. Trans. ASAE, 33, 1990.

[75] P. Sollich and C. K. I. Williams. Using the equivalent kernel to understand Gaussian process regression. In L. K. Saul, Y. Weiss, and L. Bottou, editors, Advances in Neural Information Processing Systems, pages 1313–1320. MIT Press: Cambridge, MA, 2005.

[76] M. L. Stein, J. Chen, and M. Anitescu. Stochastic approximation of score functions for Gaussian processes. Annals of Applied Statistics, 7:1162–1191, 2013.

[77] Y. Sun and M. G. Genton. Functional boxplots. Journal of Computational and Graphical Statistics, 20:316–334, 2011.

[78] Y. Sun and M. Stein. Statistically and Computationally Efficient Estimating Equations for Large Spatial Datasets. Journal of Computational and Graphical Statistics, 25:187–208, 2016.

[79] C. C. Taylor. Automatic bandwidth selection for circular density estimation. Computational Statistics and Data Analysis, 52(7):3493–3500, 2008.

[80] L. Tupper, D. Matteson, C. Anderson, and L. Zephyr. Band Depth Clustering for Nonstationary Time Series and Wind Speed Behavior. Technometrics, 2017.

[81] A. V. Vecchia. Estimation and model identification for continuous spatial processes. Journal of the Royal Statistical Society, Series B, 50:297–312, 1988.

[82] G. Wahba. Spline Models for Observational Data. Society for Industrial and Applied Mathematics, Philadelphia, PA, 1990.

[83] F. Wang and A. E. E Gelfand. Directional data analysis under the general projected normal distribution. Statistical methodology, 10:113–127, 07 2013.

[84] F. Wang and A. E. Gelfand. Modeling space and space-time directional data using projected gaussian processes. Journal of the American Statistical Association, 109(508):1565–1580, 2014.

[85] F. Wang, A. E. Gelfand, and G. Jona-Lasinio. Joint spatio-temporal analysis of a linear and a directional variable: Space-time modeling of wave heights and wave directions in the adriatic sea. Statistica Sinica, 25(1):25–39, 2015.

[86] W. Wantz and R. Sinclair. Distribution of extreme winds in the bonneville power administration service area. Journal of Applied Meteorology, 20, 1981.

[87] H. Zhang. Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. Journal of the American Statistical Association, 99:250–261, 2004.

[88] H. Zhang and D. L. Zimmerman. Towards reconciling two asymptotic frameworks in spatial statistics. Biometrika, 92:921–936, 2005.

[89] Q. Zhu, J. Chen, L. Zhu, X. Duan, and Y. Liu. Wind speed prediction with spatio–temporal correlation: A deep learning approach. Energies, 11(4), 2018.

[90] X. Zhu, M. G. Genton, Y. Gu, and L. Xie. Space-time wind speed forecasting for improved power system dispatch. TEST, 23(1):1–25, Mar 2014.

[91] X. Zhu, K. P. Bowman, and M. Genton. Incorporating geostrophic wind information for improved space-time short-term wind speed forecasting. The Annals of Applied Statistics, 8, 12 2014.

Appendix A

Proofs and Derivations

Derivation of the IDA. An alternative way to approximate the score function is to consider using the Sherman-Morrison Woodbury equations on $\left(\Sigma + \tau^2 I\right)^{-1}$:

$$\begin{aligned}
\left(\Sigma + \tau^2 I\right)^{-1} &= \frac{1}{\tau^2}\left(I - \Sigma[\Sigma + \tau^2 I]^{-1}\right) \\
\implies \left(\Sigma + \tau^2 I\right)^{-1}\mathbf{Y} &= \frac{1}{\tau^2}\left(I - \Sigma[\Sigma + \tau^2 I]^{-1}\right)\mathbf{Y} \\
&= \frac{\mathbf{Y}}{\tau^2} - \frac{1}{\tau^2}\Sigma[\Sigma + \tau^2 I]^{-1}\mathbf{Y} \\
&\approx \frac{\mathbf{Y}}{\tau^2} - \frac{1}{\tau^2}\mathbf{G}^T\mathbf{Y}
\end{aligned}$$

so the score function can be approximated $L_i = \frac{1}{(\tau^2)^2}\mathbf{Y}^T(I - \mathbf{G})^T\Sigma_i(I - \mathbf{G})\mathbf{Y}$. Remainder terms for $\mathbf{G}$ can be included to make the approximation exact. □

Derivation of the DA. Differentiation of the IDA approximation followed by both left-hand and right-hand multiplication by the data yields this approximation. Again, note that in all cases the approximation is an exact equality when instead $\mathbf{G}$ is instead replaced by

$$\mathbf{G} + \sum_{j=1}^{\infty}\mathcal{R}^j\mathbf{G}.$$

□

Proof of Lemma 1. The (j, j)th entry of the decomposed trace matrix product $\Sigma_i\mathbf{G}$ is given by

$$\frac{1}{n}\sum_{k=1}^{n} G(|s_k - s_j|) \cdot \frac{\partial k}{\partial \theta_i}\left(|s_k - s_j|\right)$$

which approximates the full kriging weights

$$\frac{1}{n}\sum_{k=1}^{n} w_n(|s_k - s_j|) \cdot \frac{\partial k}{\partial \theta_i}(|s_k - s_j|).$$

From Theorem 2.1 in [59],

$$|w_n(|s_k - s_j|) - G(|s_k - s_j|)| \leq \frac{\delta_n K}{(1-\delta_n)\rho}\exp\left(-\alpha\frac{|s_k - s_j|}{\rho}\right).$$

Then, denoting $\frac{\partial k}{\partial \theta_i} := f$ for notational simplicity,

$$\begin{aligned}
\frac{1}{n}\left|\sum_{k=1}^{n} w_{kj}f_{kj} - \sum_{k=1}^{n} G_{kj}f_{kj}\right| &\leq \frac{1}{n}\sum_{k=1}^{n} |f_{kj}|\,|w_{kj} - G_{kj}| \\
&\leq \frac{1}{n}\sum_{k=1}^{n} |f_{kj}|\,\frac{\delta_n K}{(1-\delta_n)\rho}\exp\left(-\alpha\frac{|s_k - s_j|}{\rho}\right) \\
&< \left|\frac{\delta_n K_1}{(1-\delta_n)\rho}\right| \frac{1}{n}\sum_{k=1}^{n}\exp\left(-\alpha\frac{|s_k - s_j|}{\rho}\right) \\
&< \hat{K}\left|\frac{\delta_n}{(1-\delta_n)\rho}\right|
\end{aligned}$$

where $\hat{K}$ includes that f is bounded. □