APPLICATION, METHODOLOGY, AND THEORY FOR GAUSSIAN PROCESSES

A Dissertation

by

DANIEL S. ZILBER

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Matthias Katzfuss
Co-Chair of Committee,	Debdeep Pati
Committee Members,	Bani Mallick
	Boris Hanin
Head of Department,	Branislav Vidakovic

August 2021

Major Subject: Statistics

Copyright 2021 Daniel S. Zilber

ABSTRACT

Gaussian processes are a powerful and flexible class of nonparametric models that use covariance functions, or kernels, to describe correlations across data. In addition to expressing realistic assumptions, correlation between samples acts as a substitute for larger sample sizes to improve predictions. This is demonstrated with an application to remote sensing, in which key components of airborne spectroscopy measurements are correlated to achieve greater accuracy and realism in predictions of atmospheric quantities.

In applying or developing methodology for GP's, scalability is a primary concern because the manipulation of the covariance matrix incurs a cubic complexity in the sample size. This is addressed for the case of GP inference with exponential family observations by the Vecchia-Laplace approximation method. By imposing sparsity in the posterior precision and a second order approximation to the exponential family likelihood, we achieve tractable inference with linear complexity in the sample size.

Using approximations for scalability raises theoretical questions about the tradeoff between efficiency and accuracy as studied in minimax theory, so it is of interest to know what level of approximation can be applied and still preserve the optimality of an estimator. Our work on truncated kernel ridge regression provides an answer for the case of a supremum norm loss and a finite eigenbasis representation of the kernel function. The result matches similar findings in the literature in which the effective dimension of the estimator determines the minimum level of approximation.

Aside from the use of approximation to improve scalability, a nonstationary field can be approximated with a stationary GP. We define and study the spaces that result from taking linear combinations of stationary Hilbert spaces, taking a step towards understanding nonstationary functions and the efficiency of corresponding estimators.

DEDICATION

To Mom and Pop.

You overcame real challenges so that I could overcome imaginary ones.

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Matthias Katzfuss, for his patience, guidance, and thoroughness in helping me prepare manuscripts. His efforts consistently raised the quality of my work and I will strive to maintain his high standards. I would like to thank my co-advisor, Dr. Debdeep Pati, for his patience, guidance, and willingness to work on challenging problems with me even though I was not as well prepared as others. It was a special opportunity to grow as a scholar and find my place in the broader scientific community.

I would also like to thank the other members of my committee, Drs. Bani Mallick and Boris Hanin, along with the faculty and colleagues in the Texas A&M Department of Statistics, with whom I shared discussions that were always learning experiences for me. Jon Hobbs, David Thompson, Vijay Natraj, and Amy Braverman at NASA JPL also have my gratitude for their assistance during our collaborations.

I would certainly have struggled in my first year, and learned less in the other years, if I did not have help from my classmates: Reza, Biraj, Sandipan, Xiaomeng, Huiling, Yan, Junho, Pallavi, Sean, Alex, Wenlong, Marcin, Se Yoon, Yabo, and others.

Finally I thank my parents and brother for their support. This was a long and difficult road where the biggest obstacle was often a lack of confidence. Stubborn persistence is certainly an asset, but sometimes you need a sympathetic ear, a reminder that you can do it, and a fried chicken cutlet.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a dissertation committee consisting of Professor Matthias Katzfuss as advisor, Professor Debdeep Pati as co-advisor and Professor Bani Mallick of the Department of Statistics and Professor Boris Hanin of the Department of Mathematics.

The data analyzed for Chapter 2 and 3 were collected by the NASA JPL MODIS and NASA JPL AVIRIS-NG instruments as part of their respective missions. The analyses depicted in Chapter 3 were conducted with the ISOFIT software written by David Thompson of NASA JPL that is available on GitHub, and some background explanations for sections 3.1 and 3.2 were written with input from David Thompson and Jon Hobbs. The analysis of Chapter 2 was published in 2021 in the journal Computational Statistics and Data Science.

All other work conducted for the dissertation was completed by the student independently.

Funding Sources

Graduate study was supported by a teaching assistantship, technology teaching assistantship, research assistantship, and teaching position from Texas A&M University, and a research assistantship from NASA JPL.

NOMENCLATURE

GP	Gaussian Process
VL	Vecchia-Laplace, algorithm
MCMC	Markov Chain Monte Carlo
HMC	Hamiltonian Monte Carlo
UQ	Uncertainty Quantification
OE	Optimal Estimation
VSWIR	Visible/ShortWave InfraRed
RTM	Radiative Transfer Model
ATREM	ATmosphere REMoval, algorithm
LibRadTran	Library for Radiative Transfer
MODTRAN	MODerate resolution atmospheric TRANsmission
AVIRIS-NG	Airborne Visible-Infrared Imaging Spectrometer - Next Generation
MODIS	MODerate-resolution Imaging Spectroradiometer
ISOFIT	Imaging Spectrometer Optimal FITting
KRR	Kernel Ridge Regression
RKHS	Reproducing Kernel Hilbert Space
L^p	Space of absolutely Lebesgue integrable functions; i.e., $\int \lvert f \rvert^p < \infty$
$\mathcal{F}(f)$	Fourier transform of f
TAMU	Texas A&M University

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

Gaussian distributions are ubiquitous in science due to the central limit theorem. Many regression models, which are essentially pattern detection algorithms, start with assumptions of independent Gaussian noise and can therefore be solved with simple methods. However, when noise starts to have its own patterns and is no longer independent, we can introduce correlations, so that any collection of noise terms has a Gaussian distribution with correlations. This is the basic idea of a Gaussian process (GP) [2], which describes all the possible curves and functions representing noise terms.

A GP only requires two components to be fully defined: a mean function and a covariance function. Like simpler noise models, the mean is usually assumed to be 0, so the GP is entirely specified by the covariance function, or kernel. Although the kernel has a simple job, which is to describe the covariance between any two inputs, there is an incredible richness and utility to structuring a collection of functions in terms of covariances or correlations between points. In fact, the kernel can be thought of as defining the collection of functions, encapsulating all the assumptions we want to make when looking for a pattern in data.

This thesis presents a collection of research projects about GP's that share a common theme of manipulating covariance kernels for desired effects. These effects include scalability, in which we want to approximate a covariance to have a faster algorithm, and the addition of modeling assumptions, in which we add terms to the covariance to model more complicated behavior we expect in the data. A recurring theme in this thesis and in the GP literature in general is the tradeoff between model accuracy and computational cost.

The first work of this thesis, the Vecchia-Laplace algorithm, describes a method for performing inference on large spatially correlated data sets with observations that follow exponential family distributions, as oppose to the more restrictive special case of Gaussian distributed data. There are two issues our method addresses: the intractability of computing the prior due to the exponential family likelihoods and the cubic computational complexity. The former issue is solved by applying

a Laplace approximation, requiring a Newton-Raphson iteration to compute the posterior mode at which to perform the second order approximation. The latter issue is the great weakness of GP based models: computing any conditional quantity, such as a posterior mean, involves inverting the matrix of all possible pairwise correlations at a cost cubic in the sample size. The spatial statistics literature offers a wide array of approximations, but we use the sparse general Vecchia approximation [3] which offers linear computational complexity and high performance by inducing sparsity in the precision matrix.

The second work of this thesis is an application of GP modeling to remote sensing, in which the observed data consists of intensity measurements of multiple light frequencies measured by a camera. The objective of remote sensing is to use the observed radiance to learn about the target after correcting for any effects such as scattered or diffused light that accumulate between the target and the observation point. For our case, the objective is to use an earth-facing camera mounted in an aircraft to infer the composition of the earths surface while simultaneously estimating atmospheric water vapor and aerosols [4]. Existing Bayesian methods, referred to as optimal estimation [5], invert each radiance measurement individually to get the surface and atmospheric states, ignoring the fact that nearby radiance measurements should have nearly identical atmospheric states. We introduce correlation directly into the model to account for this, resulting in a spatial visible/shortwave remote sensing model that offers more realistic predictions and accurate uncertainty quantification.

The third project establishes minimax optimality for the KRR estimator in the case of a truncated kernel under supremum norm. Intended as first step towards more general approximations, this work explores the tradeoff between kernel approximation and estimator optimality for kernel ridge regression. Optimality is understood in the minimax sense, which defines the best possible performance in terms of minimizing the worst case (over all possible true functions) error between a prediction by an estimator given a particular amount of data and the true function as measured by a loss function (risk) such as expectation over L_2 or supremum norm. The supremum norm is challenging to work under because normed terms do not correspond to integrals, but the resulting bounds are attractive because they hold uniformly. Fortunately, much of the groundwork for supre-

mum norm minimax optimalty has been established in an earlier work [6], so our work can be seen as a minor extension to existing results.

The last project characterizes a nonstationary reproducing kernel Hilbert space. Stationarity is a common simplifying assumption for GP modeling in which a global mean is fixed and correlations only depend on distance, not orientation or location. Natural processes typically are not stationary, though. For example, health outcomes, social or consumer behavior, and ecological patterns may change over time as new information or policy is disseminated or geographic features change. Abrupt changes that induce independence can be thought of as change points, while gradual changes are better represented in terms of slowly vary correlations that are aggregated via convolutions. These two cases form the basis for our characterization, which expresses elements of the nonstationary space as weighted sums of functions taken from stationary spaces,

$$f(x) = \sum_{i=1}^{m} \psi_i(x) f_i(x).$$

Considering different weighting functions leads to different effective kernels, and we show that we can take countably infinite combinations under reasonable assumptions. Our space of nonstationary functions is a relatively simple alternative to more sophisticated spaces such as Besov spaces [7], opening the door to embedding, decomposition, and concentration theorems.

The dissertation concludes with a summary of the major results and some thoughts of the role of the kernels in the broader scientific world.

2. A VECCHIA-LAPLACE APPROXIMATION FOR BIG NON-GAUSSIAN SPATIAL DATA [1]

2.1 Introduction

Dependent non-Gaussian data are ubiquitous in time series, geospatial applications, and more generally in nonparametric regression and classification. A popular model for such data is obtained by combining a latent Gaussian process (GP) with conditionally independent, potentially non-Gaussian likelihoods from the exponential family. This is traditionally referred to as a spatial generalized linear mixed model (SGLMM) in the spatial statistics literature [8], but the same model has more recently also been referred to as a generalized GP (GGP) [9]; we will use the latter, more concise term throughout. GGPs are highly flexible, interpretable, and allow for natural, probabilistic uncertainty quantification. However, inference for GGPs can be analytically intractable, and large datasets pose additional computational challenges due to the inversion of the GP covariance matrix.

Popular techniques to numerically perform the intractable marginalization necessary for inference are, in order of increasing speed: Markov chain Monte Carlo (MCMC), expectation propagation, variational methods, and Laplace approximations. See [10] for a recent review of deterministic techniques and [11] for a comparison of MCMC and expectation propagation. [12] argues that variational methods and expectation propagation suffer from underestimated and overestimated posterior variances, respectively. Here, we consider the Laplace approximation [13, 14], a classic technique for integral evaluation based on second-order Taylor expansion. [15] show numerically that the Laplace approximation can be a practical and accurate method for GGP inference.

It has long been recognized that the computational cost for GPs is cubic in the data size. Much work has been done on GP approximations that address this problem in the context of Gaussian noise [16]. Low-rank approaches [17, 18, 19, 20, 21] have limitations in the presence of fine-scale

[1] This article was published in Computational Statistics & Data Analysis, 153, Daniel Zilber and Matthias Katzfuss, Vecchia–Laplace approximations of generalized Gaussian processes for big non-Gaussian spatial data, 107081, Copyright Elsevier (2021).

structure [22], but they have proved popular due to their simplicity. Approximations relying on sparsity in covariance matrices [23, 24] by definition can only capture local, short-range dependence and cannot guarantee low computation cost. Approaches that take advantage of Toeplitz or Kronecker structure [25, 26, 27] can be extremely efficient but are not as generally applicable. Recently, methods relying on sparsity in precision matrices [28, 29, 30] have gained popularity due to their accuracy and flexibility. In particular, a class of highly promising GP approximations [31, 32, 33, 34, 3, 35] rely on ordered conditional independence that can guarantee linear scaling in the data size while resolving dependence at all scales.

There are also a number of existing methods for large non-Gaussian datasets modeled using GGPs. A popular approach is to combine a low-rank GP with an approximation of the non-Gaussian likelihood, as the dimension reduction inherent in the low-rank approximation carries through even to the non-Gaussian case. [36] estimate parameters using an expectation-maximization algorithm with low-rank and Laplace approximations. [37] use variational inference to obtain the posterior and select a set of conditioning points for their low-rank approximation. Some methods of dimension reduction, including random sketching [38] and projection, offer theoretical guarantees and can be combined with MCMC methods for the analysis of non-Gaussian data [39, 40], but are still subject to the limitations of low-rank methods. [41] develop state-space models for one-dimensional non-Gaussian time series, which can perform inference in linear time and memory using a set of knots in time, a form of low-rank approximation. Alternate priors such as log gamma priors for count data [42] are an interesting but specialized approach to completely avoid the intractability issues with GGPs.

Similar to what we shall propose, some authors have combined a sparse-precision approach with a non-Gaussian approximation. A prominent example is [29], in which an integrated nested Laplace approximation (INLA) is combined with a sparse-precision approximation of the GP using its representation as the solution to a stochastic partial differential equation. [33] proposed to apply the GP approximation of [31] to a latent GP, but did not provide an explicit algorithm for large non-Gaussian data. While both [29] and [33] limit the number of nonzero entries per row or column

in the precision matrix to a small constant, the computational complexity for decomposing this sparse $n \times n$ matrix is not linear in n, but rather $\mathcal{O}(n^{3/2})$ in two dimensions [43, Thm. 6], and at least $\mathcal{O}(n^2)$ in higher dimensions. In the Gaussian setting, this scaling problem can be overcome by applying a Vecchia approximation to the observed data [31] or to the joint distribution of the observed data and the latent GP [3].

To handle both scaling and intractability, we propose a Vecchia-Laplace (VL) approximation for GGPs. The posterior mode necessary for the Laplace approximation is found using the Newton-Raphson algorithm, which can be viewed as iterative GP inference based on Gaussian pseudo-data. At each iteration of our VL algorithm, the joint Gaussian distribution of the pseudo-data and the latent GP realizations is approximated using the general Vecchia framework [3, 35]. By modeling the joint distribution of pseudo-data and GP realizations at each iteration, our VL approach can ensure sparsity and guarantee linear scaling in n for any dimension, overcoming the scaling issues of the sparse-matrix approaches mentioned above.

To our knowledge, we provide the first explicit algorithm extending and applying the highly promising class of general-Vecchia GP approximations to large non-Gaussian data. We believe it to be a useful addition to the literature due to its speed, simplicity, guaranteed numerical performance, and wide applicability (e.g., binary, count, right-skewed, and point-pattern data). In addition, as shown in [3], the general Vecchia approximation includes many popular GP approximations [31, 44, 45, 46, 33, 47, 48] as special cases, and so our VL methodology also directly provides extensions of these GP approximations to non-Gaussian data.

The remainder of this document is organized as follows. In Section 2.2, we review the Laplace approximation and general Vecchia. In Section 2.3, we introduce and examine our VL method, including parameter inference and predictions at unobserved locations. In Sections 2.4 and 2.5, we study and compare the performance of VL on simulated and real data, respectively. Some details are left to the appendix. A separate Supplementary Material document contains Sections A.3–A.8 with additional derivations, simulations, and discussion. The methods and algorithms proposed here are implemented in the `R` package `GPvecchia` [49] with sensible default settings, so that

a wide audience of practitioners can immediately use the code with little background knowledge. Our results and figures can be reproduced using the code and data at `https://github.com/katzfuss-group/GPvecchia-Laplace`.

2.2 Review of existing results

2.2.1 Generalized Gaussian processes

Let $y(\cdot) \sim GP(\mu, K)$ be a latent Gaussian process with mean function μ and kernel or covariance function K on a domain $\mathcal{D} \subset \mathbb{R}^d$, $d \in \mathbb{N}^+$. Observations $\mathbf{z} = (z_1, \ldots, z_n)'$ at locations $\mathbf{s}_i \in \mathcal{D}$ are assumed to be conditionally independent, $z_i|\mathbf{y} \overset{ind.}{\sim} g_i(z_i|y_i)$, where $\mathbf{y} = (y_1, \ldots, y_n)'$ and $y_i = y(\mathbf{s}_i)$. We assume that the observation densities or likelihoods g_i are from the exponential family. Parameters $\boldsymbol{\theta}$ in μ, K, or the g_i will be assumed fixed and known for now; for example, $\boldsymbol{\theta}$ may contain regression coefficients determining the mean function μ, or variance, smoothness, and range parameters determining a Matérn covariance K.

Our goal is to obtain an approximation of the posterior of $\mathbf{y}$, which takes the form

$$p(\mathbf{y}|\mathbf{z}) = \frac{\mathcal{N}_n(\mathbf{y}|\boldsymbol{\mu}, \mathbf{K}) \prod_{i=1}^n g_i(z_i|y_i)}{p(\mathbf{z})}, \tag{2.1}$$

where $\boldsymbol{\mu} = (\mu(\mathbf{s}_1), \ldots, \mu(\mathbf{s}_n))'$, and $\mathbf{K}$ is an $n \times n$ covariance matrix with (i, j) entry $(\mathbf{K})_{i,j} = K(\mathbf{s}_i, \mathbf{s}_j)$. Once an approximation of the posterior (2.1) has been obtained, it is conceptually straightforward to extend this result to other quantities of interest, such as the integrated likelihood for inference on parameters $\boldsymbol{\theta}$ (see Section 2.3.2), and prediction of $y(\cdot)$ at unobserved locations (see Section 2.3.3).

2.2.2 Review of the Laplace approximation

The normalizing constant $p(\mathbf{z})$ in (2.1) is not available in closed form for non-Gaussian likelihoods. A popular approach to this issue is the Laplace approximation [14, 2, e.g.,], which approximates $p(\mathbf{z}) = \int \exp(\log p(\mathbf{z}|\mathbf{y}))p(\mathbf{y})d\mathbf{y}$ via a second-order Taylor expansion of $\log p(\mathbf{z}|\mathbf{y})$ at the mode of the posterior density $p(\mathbf{y}|\mathbf{z})$. As this results in an exponentiated quadratic form in $\mathbf{y}$,

it is equivalent to a Gaussian approximation of the likelihood. The mode of $\log p(\mathbf{y}|\mathbf{z})$ does not depend on the normalizing constant, and so it can be obtained using standard optimization procedures such as the Newton-Raphson algorithm. The crucial observation for our later developments is that each Newton-Raphson update in the GGP setting is equivalent to computing the posterior mean of $\mathbf{y}$ given *Gaussian* pseudo-data [2, Sect. 3.4.1]. Upon convergence of the algorithm, we have a Laplace approximation for the normalizing constant and a Gaussian approximation for the likelihood, which gives us a Gaussian posterior.

We now go into the details of this approximation. Based on the first and second derivative of $\log g_i$, we define

$$u_i(y_i) = \tfrac{\partial}{\partial y_i} \log g_i(z_i|y_i) \qquad \text{and} \qquad d_i(y_i) = -\big(\tfrac{\partial^2}{\partial y_i^2} \log g_i(z_i|y_i)\big)^{-1}, \qquad i = 1, \ldots, n.$$

Stacking these quantities as $\mathbf{u}_\mathbf{y} = \big(u_1(y_1), \ldots, u_n(y_n)\big)'$ and $\mathbf{D}_\mathbf{y} = \operatorname{diag}\big(d_1(y_1), \ldots, d_n(y_n)\big)$, it is easy to see that $\frac{\partial}{\partial \mathbf{y}} \log p(\mathbf{y}|\mathbf{z}) = -\mathbf{K}^{-1}(\mathbf{y} - \boldsymbol{\mu}) + \mathbf{u}_\mathbf{y}$ and $-\frac{\partial^2}{\partial \mathbf{y} \partial \mathbf{y}'} \log p(\mathbf{y}|\mathbf{z}) = \mathbf{K}^{-1} + \mathbf{D}_\mathbf{y}^{-1} =: \mathbf{W}_\mathbf{y}$. When given the posterior mode $\boldsymbol{\alpha} = \arg\max_{\mathbf{y} \in \mathbb{R}^n} \log p(\mathbf{y}|\mathbf{z})$, the combined Gaussian/Laplace approximation of the posterior is

$$\widehat{p}_L(\mathbf{y}|\mathbf{z}) = \mathcal{N}_n(\mathbf{y}|\boldsymbol{\alpha}, \mathbf{W}_{\boldsymbol{\alpha}}^{-1}). \tag{2.2}$$

The subscript $\boldsymbol{\alpha}$ in $\mathbf{W}_{\boldsymbol{\alpha}}^{-1}$ implies evaluation of $\mathbf{W}_\mathbf{y}$ at the mode $\boldsymbol{\alpha}$, rather than at an arbitrary $\mathbf{y}$. To obtain the mode $\boldsymbol{\alpha}$ with the Newton-Raphson algorithm, we start with an initial value $\mathbf{y}^{(0)}$, and update the current guess for $\ell = 0, 1, 2, \ldots$ until convergence as $\mathbf{y}^{(\ell+1)} = \mathbf{h}(\mathbf{y}^{(\ell)})$, where

$$\mathbf{h}(\mathbf{y}) = \mathbf{y} - \big(\tfrac{\partial^2}{\partial \mathbf{y} \partial \mathbf{y}'} \log p(\mathbf{y}|\mathbf{z})\big)^{-1} \big(\tfrac{\partial}{\partial \mathbf{y}} \log p(\mathbf{y}|\mathbf{z})\big). \tag{2.3}$$

This Newton-Raphson update is equivalent to computing the posterior mean of $\mathbf{y}$ given Gaussian pseudo-data $\mathbf{t}_\mathbf{y} = \mathbf{y} + \mathbf{D}_\mathbf{y}\mathbf{u}_\mathbf{y}$ with noise covariance matrix $\mathbf{D}_\mathbf{y}$. Specifically, we can write the

distribution	likelihood $g(z\|y)$	pseudo-data t_y	pseudo-variance $d(y)$
Gaussian	$\mathcal{N}(y, \tau^2)$	z	τ^2
Bernoulli	$\mathcal{B}(\text{logit}^{-1}(y))$	$y + \frac{(1+e^y)^2}{e^y}(z - \frac{e^y}{1+e^y})$	$(1+e^{-y})(1+e^y)$
Poisson	$\mathcal{P}(e^y)$	$y + e^{-y}(z - e^y)$	e^{-y}
Gamma	$\mathcal{G}(a, ae^{-y})$	$y + (1 - z^{-1}e^y)$	$e^y/(az)$

Table 2.1: Examples of popular likelihoods, together with the Gaussian pseudo-data and pseudo-variances implied by the Laplace approximation. The non-canonical logarithmic link function is used for the Gamma likelihood to ensure that the second parameter, ae^{-y}, is positive.

Newton-Raphson update in (2.3) as:

$$\mathbf{h}(\mathbf{y}) = \boldsymbol{\mu} + \mathbf{W}_{\mathbf{y}}^{-1}\mathbf{D}_{\mathbf{y}}^{-1}(\mathbf{t}_{\mathbf{y}} - \boldsymbol{\mu}) = \mathrm{E}(\mathbf{y}|\mathbf{t}_{\mathbf{y}}), \tag{2.4}$$

which is the conditional mean of $\mathbf{y}$ given Gaussian pseudo-data $\mathbf{t}_{\mathbf{y}}|\mathbf{y} \sim \mathcal{N}_n(\mathbf{y}, \mathbf{D}_{\mathbf{y}})$. The derivation of (2.4) is straightforward and included in Appendix A.1 for completeness. This means we can obtain the mode $\boldsymbol{\alpha}$ by iterating between (a) computing pseudo-data $\mathbf{t}_{\mathbf{y}^{(\ell)}}$ with ith entry $y_i^{(\ell)} + d_i(y_i^{(\ell)})u_i(y_i^{(\ell)})$, and (b) obtaining the posterior mean $\mathbf{y}^{(\ell+1)}$ of $\mathbf{y}$ given $\mathbf{t}_{\mathbf{y}^{(\ell)}}$ assuming independent Gaussian noise with variances $d_1(y_1^{(\ell)}), \ldots, d_n(y_n^{(\ell)})$.

Some examples of popular likelihoods and the corresponding pseudo-data and pseudo-variances are summarized in Table 2.1. The Bernoulli and Poisson cases are also illustrated in Figure 2.1.

Once the algorithm has converged (i.e., $\boldsymbol{\alpha} := \mathbf{y}^{(\ell+1)} = \mathbf{y}^{(\ell)}$), we can use the second-order expansion of the loglikelihood at the mode as a Gaussian approximation of the likelihood based on pseudo-data,

$$\widehat{p}_L(\mathbf{z}|\mathbf{y}) = p(\mathbf{t}_{\boldsymbol{\alpha}}|\mathbf{y}) = \mathcal{N}_n(\mathbf{t}_{\boldsymbol{\alpha}}|\mathbf{y}, \mathbf{D}_{\boldsymbol{\alpha}}), \tag{2.5}$$

or combine it with the Laplace approximation to get a Gaussian approximation of the posterior conditional on pseudo-data,

$$\widehat{p}_L(\mathbf{y}|\mathbf{z}) = p(\mathbf{y}|\mathbf{t}_{\boldsymbol{\alpha}}) = \mathcal{N}_n(\mathbf{y}|\boldsymbol{\alpha}, \mathbf{W}_{\boldsymbol{\alpha}}^{-1}). \tag{2.6}$$

For conciseness, we henceforth refer to (2.6) as the “Laplace approximation," rather than the more

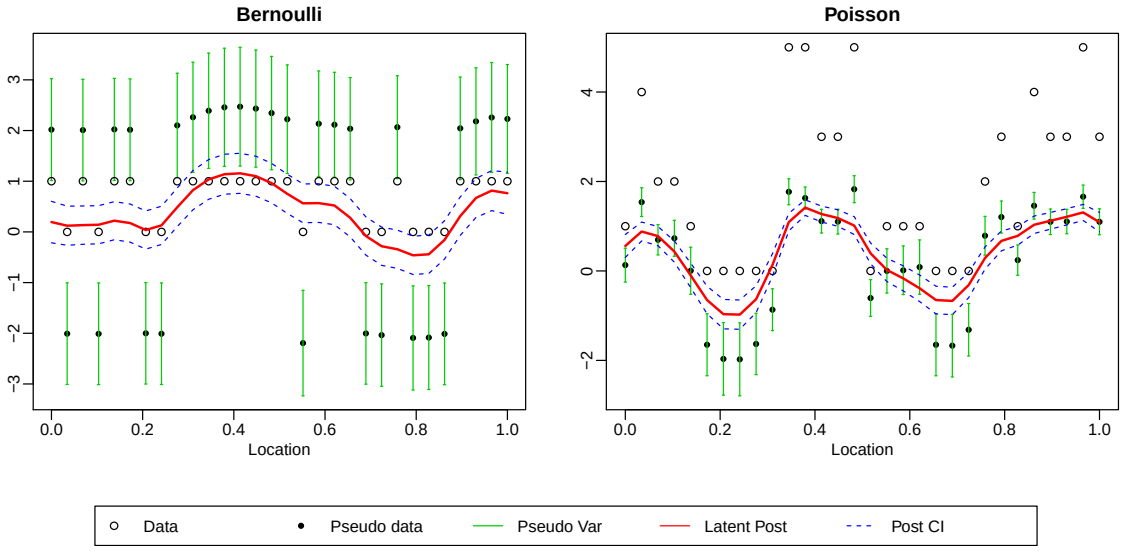


Figure 2.1: Pseudo-data $\mathbf{t}_{\boldsymbol{\alpha}}$ plus or minus half the standard deviation of the pseudo-noise for simulated data $\mathbf{z}$ in one spatial dimension, along with the latent posterior mode $\boldsymbol{\alpha}$ plus or minus half the posterior standard deviation. Note that the data exhibit a different scale than the pseudo-data due to the link function.

precise "combined Gaussian and Laplace approximation."

2.2.3 Review of the general Vecchia approximation

The Laplace approximation described in Section 2.2.2 allows us to deal with non-Gaussian likelihoods, but it still requires decomposing the $n \times n$ matrix $\mathbf{K}$ and thus scales as $\mathcal{O}(n^3)$. To achieve computational feasibility even for data sizes n in the tens of thousands or more, we also apply a general Vecchia approximation [3], which we will briefly review here.

Assume that $\mathbf{y} \sim \mathcal{N}_n(\boldsymbol{\mu}, \mathbf{K})$ is a vector of GP realizations and $\mathbf{t}|\mathbf{y} \sim \mathcal{N}_n(\mathbf{y}, \mathbf{D})$ a vector of noisy data, where $\mathbf{D}$ is diagonal. Then, consider a vector $\mathbf{x} = \mathbf{y} \cup \mathbf{t}$ consisting of the $2n$ elements of $\mathbf{y}$ and $\mathbf{t}$ in some ordering (more details below). It is well known that the density function, $p(\mathbf{x})$, can be factored into a product of univariate conditional densities, $p(\mathbf{x}) = \prod_{i=1}^{2n} p(x_i|\mathbf{x}_{1:i-1})$. The general Vecchia framework extends the approximation of [31] to the vector $\mathbf{x}$ consisting of latent GP realizations and noisy data, resulting in the approximate density

$$\widehat{p}(\mathbf{x}) = \prod_{i=1}^{2n} p(x_i|\mathbf{x}_{c(i)}), \tag{2.7}$$

where $c(i) \subset \{1, \ldots, i-1\}$ is a conditioning index set of size m (or of size $i-1$ for $i \leq m$). A small m can lead to enormous computational savings and good approximations; [50] show that under some settings, the approximation error can be bounded when m increases only polylogarithmically with n. While Vecchia is related to composite likelihood [51, e.g.,], most variants of the latter assume some form of marginal or unordered conditional independence, which may reduce approximation accuracy; for more details and numerical comparisons, see [3, Sect. 3.8 and App. D].

As $y_i = y(\mathbf{s}_i)$ is indexed by location and t_i is the corresponding noisy observation, the ordering within $\mathbf{y}$ and within $\mathbf{t}$ is determined by an ordering of the observed locations, $\mathbf{s}_1, \ldots, \mathbf{s}_n$. We will use a coordinate-based (left-to-right) ordering in one spatial dimension. In higher-dimensional spaces, we recommend a maxmin ordering [34, 52], which sequentially chooses each location in the ordering to maximize the minimum distance to previous locations in the ordering.

By straightforward extension of the proof of Prop. 1 in [3] to the case $\boldsymbol{\mu} \neq \mathbf{0}$, it can be shown that the approximation in (2.7) implies a multivariate normal joint distribution, $\widehat{p}(\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_x, \mathbf{Q}^{-1})$, where $\boldsymbol{\mu}_{x,i} = \mu(\mathbf{s}_j)$ if $x_i = y_j$ or $x_i = t_j$, $\mathbf{Q} = \mathbf{U}\mathbf{U}'$, and $\mathbf{U}$ is the sparse upper triangular Cholesky factor based on a reverse row-column ordering of $\mathbf{Q}$. We write this as $\mathbf{U} = \mathrm{rchol}(\mathbf{Q}) := \mathrm{rev}(\mathrm{chol}(\mathrm{rev}(\mathbf{Q})))$, where $\mathrm{rev}(\cdot)$ reverse-orders the rows and columns of its matrix argument. The nonzero entries of $\mathbf{U}$ are computed directly based on the covariance function K as described in Appendix A.2.

Let $\mathbf{U}_y$ and $\mathbf{U}_t$ be the submatrices of $\mathbf{U}$ consisting of the rows of $\mathbf{U}$ corresponding to $\mathbf{y}$ and $\mathbf{t}$, respectively. Then, the sparse matrix $\mathbf{W} = \mathbf{U}_y\mathbf{U}_y'$ is the general Vecchia approximation to the posterior precision matrix of $\mathbf{y}$ given $\mathbf{t}$. Defining $\mathbf{V} := \mathrm{rchol}(\mathbf{W})$, we can obtain the posterior mean of $\mathbf{y}$ as $\mathrm{E}(\mathbf{y}|\mathbf{t}) = \boldsymbol{\mu} - (\mathbf{V}')^{-1}\mathbf{V}^{-1}\mathbf{U}_y\mathbf{U}_t'(\mathbf{t} - \boldsymbol{\mu})$.

2.2.4 Ordering in Vecchia approximations

We now describe two specific approximations within the general Vecchia framework, which are based on how the elements of $\mathbf{y}$ and $\mathbf{t}$ are ordered in the vector $\mathbf{x}$ in (2.7): Interweaved (IW) ordering and response-first (RF) ordering. While other ordering and conditioning schemes can also

be used in the Vecchia-Laplace methodology to be introduced in Section 2.3, we recommend these specific schemes to achieve high accuracy while ensuring linear complexity.

2.2.4.1 Interweaved (IW) ordering

Vecchia-Interweaved (IW) is the sparse general Vecchia approach proposed for likelihood inference in [3], reviewed briefly here. It is a special case of general Vecchia in (2.7), in which $\mathbf{x} = (y_1, t_1, y_2, t_2, \ldots, y_n, t_n)'$ is specified using an interweaved ordering of latent $\mathbf{y}$ and responses $\mathbf{t}$. We consider the following specific expression for (2.7):

$$\widehat{p}_{IW}(\mathbf{x}) = \prod_{i=1}^{n} p(t_i|y_i)\, p(y_i|\mathbf{y}_{q_y(i)}, \mathbf{t}_{q_t(i)}). \tag{2.8}$$

If $x_j = t_i$, we only condition on y_i, because $\mathbf{D}$ is diagonal and therefore t_i is conditionally independent of all other variables in $\mathbf{y}$ and $\mathbf{t}$ given y_i. If $x_j = y_i$, we condition on $\mathbf{y}_{q_y(i)}$ and $\mathbf{t}_{q_t(i)}$, where $q(i) = q_y(i) \cup q_t(i)$ is the conditioning index vector consisting of the indices of the nearest m locations previous to i in the ordering. For splitting $q(i)$ into $q_y(i)$ and $q_t(i)$, we attempt to maximize $q_y(i)$ while ensuring linear complexity [3]. Specifically, for $i = 1, \ldots, n$, we set $q_y(i) = (k_i) \cup (q_y(k_i) \cap q(i))$, where $k_i \in q(i)$ is the index whose latent-conditioning set has the most overlap with $q(i)$: $k_i = \arg\max_{j \in q(i)} |q_y(j) \cap q(i)|$, choosing the closest k_i in space to $\mathbf{s}_i$ in case of a tie. In one-dimensional space with coordinate ordering, this results in $q_y(i) = q(i) = (\max(1, i-m), \ldots, i-1)$ and $q_z(i) = \emptyset$. In higher-dimensional space, we may not be able to condition entirely on $\mathbf{y}$, so the remaining conditioning indices are assigned to $q_t(i) = q(i) \setminus q_y(i)$. These conditioning rules guarantee that $\mathbf{U}$ and $\mathbf{V}$ are both highly sparse with at most m nonzero off-diagonal elements per column. [3] showed that these matrices, and the resulting posterior mean and precision matrix, can be obtained in $\mathcal{O}(nm^3)$ time.

2.2.4.2 Response-first (RF) ordering

For approximating predictions at observed locations in Algorithm 1 in more than one dimension, we recommend the new RF-full method described in [35], reviewed briefly here. RF-full orders first all response variables, then all latent variables: $\mathbf{x} = (\mathbf{t}', \mathbf{y}')' = (t_1, \ldots, t_n, y_1, \ldots, y_n)'$.

We consider the following specific expression for (2.7):

$$\widehat{p}_{RF}(\mathbf{x}) = \prod_{i=1}^{n} p(t_i)\, p(y_i|\mathbf{y}_{q_y(i)}, \mathbf{t}_{q_t(i)}).$$

The responses t_i do not condition on anything and are considered independent; this implies a poor approximation to $p(\mathbf{t})$, but it does not affect the posterior distribution $p(\mathbf{y}|\mathbf{t})$, which is the relevant quantity for our purposes. We now assume $q(i) = q_y(i) \cup q_t(i)$ to be set of indices corresponding to the m locations closest to $\mathbf{s}_i$ (including $\mathbf{s}_i$), not considering the ordering. For any $j \in q(i)$, we then let y_i condition on y_j if it is ordered previously in $\mathbf{x}$; otherwise, we condition on t_j. More precisely, we set $q_y(i) = \{j \in q(i) : j < i\}$ and $q_t(i) = \{j \in q(i) : j \geq i\}$. Similar to IW, RF-full inference can be carried out in $\mathcal{O}(nm^3)$ time [35].

2.3 Vecchia-Laplace methods

We now introduce our Vecchia-Laplace (VL) approximation, which allows fast inference for large datasets modeled using GGPs, by combining the Laplace and general Vecchia approximations reviewed in Section 2.2.

2.3.1 The VL algorithm

To apply a Laplace approximation, it is first necessary to find the mode of the posterior density of $\mathbf{y}$. Rapid convergence to the mode can be achieved using a Newton-Raphson algorithm, which can be viewed as iteratively computing a new value $\mathbf{y}^{(l+1)}$ as the posterior mean of the latent GP realization $\mathbf{y}$ based on Gaussian pseudo-data $\mathbf{t} = \mathbf{t}_{\mathbf{y}^{(l)}}$, as discussed in Section 2.2.2. Our VL algorithm applies a general Vecchia approximation $\widehat{p}(\mathbf{x})$ to the joint distribution of $\mathbf{x} = \mathbf{y} \cup \mathbf{t}$ at each iteration l, and computes the posterior mean of $\mathbf{y}$ given $\mathbf{t}$ under this approximate distribution. We recommend IW ordering (Section 2.2.4.1) in one spatial dimension, and RF ordering (Section 2.2.4.2) when working in more than one dimension. The resulting VL algorithm is presented as Algorithm 1. After convergence, we obtain the approximation

$$\widehat{p}_{VL}(\mathbf{y}|\mathbf{z}) = \mathcal{N}_n(\mathbf{y}|\boldsymbol{\alpha}_V, \mathbf{W}_V^{-1}). \tag{2.9}$$

Algorithm 1 Vecchia-Laplace (VL)

procedure VECCHIA-SPECIFY($\mathcal{S}, m$) ▷ Define Vecchia Structure
    Order locations $\mathcal{S}$ using coordinate (in 1D) or maxmin ordering (in 2D or higher)
    For VL-IW, determine variable ordering and conditioning as in Sect. 2.2.4.1
    For VL-RF, determine variable ordering and conditioning as in Sect. 2.2.4.2
    return ordering and conditioning info in Vecchia Approximation Object `VAO`
end procedure

procedure VL-INFERENCE($\mathbf{z}$, `VAO`, g_i, $\boldsymbol{\mu}$, K) ▷ Maximize GP Posterior
    Derive $u_i(\cdot) = \frac{\partial}{\partial y} \log g_i\big|_{(\cdot)}$ and $d_i(\cdot) = -\big(\frac{\partial^2}{\partial y^2} \log g_i\big)^{-1}\big|_{(\cdot)}$
    Initialize $\mathbf{y}^{(0)} = \boldsymbol{\mu}$
    for l=0,1,... **do**
        Compute $\mathbf{u} = (u_1(y_1^{(l)}), \ldots, u_n(y_n^{(l)}))'$ and $\mathbf{D} = \mathrm{diag}(d_1(y_1^{(l)}), \ldots, d_n(y_n^{(l)}))$
        Update pseudo-data $\mathbf{t} = \mathbf{y}^{(l)} + \mathbf{D}\mathbf{u}$
        Compute $\mathbf{U}$ (see Appendix A.2) based on $\mathbf{D}$, K, and `VAO`
        Extract submatrices $\mathbf{U}_y$ and $\mathbf{U}_t$
        Compute $\mathbf{W} = \mathbf{U}_y\mathbf{U}_y'$ and $\mathbf{V} = \mathrm{rchol}(\mathbf{W})$
        Compute the new posterior mean: $\mathbf{y}^{(l+1)} = \boldsymbol{\mu} - (\mathbf{V}')^{-1}\mathbf{V}^{-1}\mathbf{U}_y\mathbf{U}_t'(\mathbf{t} - \boldsymbol{\mu})$
        if $\|\mathbf{y}^{(l+1)} - \mathbf{y}^{(l)}\| < \epsilon$ **then**
            return $\boldsymbol{\alpha}_V = \mathbf{y}^{(l+1)}$ and $\mathbf{W}_V = \mathbf{W}$ ▷ Posterior Mode Estimate
        end if
    end for
end procedure

Once the algorithm has converged and the posterior mean $\boldsymbol{\alpha}_V$ and precision $\mathbf{W}_V$ have been obtained, the posterior distribution in (2.9) can be used for estimation of the integrated likelihood (Section 2.3.2) and for prediction at unobserved locations (Section 2.3.3). As we will see in our simulation studies later, even for moderate m, the VL procedure in Algorithm 1 essentially finds the exact mode of the posterior.

2.3.2 Integrated likelihood for parameter inference

In the case of unknown parameters $\boldsymbol{\theta}$ in μ, K, or in the g_i, we would like to carry out parameter inference based on the integrated likelihood,

$$\mathcal{L}(\boldsymbol{\theta}) = p(\mathbf{z}|\boldsymbol{\theta}) = \int p(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}) p(\mathbf{y}|\boldsymbol{\theta}) d\mathbf{y}.$$

However, this quantity is exactly the unknown normalizing constant in the denominator of (2.1), and the integral can generally not be carried out analytically. Instead, we will base parameter inference on the integrated likelihood implied by our VL approximation. In the following, we will again suppress dependence on $\boldsymbol{\theta}$ for ease of notation.

First, rearranging terms in (2.1), we have $p(\mathbf{z}) = p(\mathbf{z}|\mathbf{y})p(\mathbf{y})/p(\mathbf{y}|\mathbf{z})$. The Laplace approximation approximates the posterior in the denominator as $\widehat{p}_L(\mathbf{y}|\mathbf{z}) = p(\mathbf{y}|\mathbf{t}_{\boldsymbol{\alpha}})$ (see (2.6)). Noting that rearranging the definition of a conditional density gives $p(\mathbf{y}) = p(\mathbf{y}, \mathbf{t})/p(\mathbf{t}|\mathbf{y})$, we obtain the Laplace approximation of the integrated likelihood:

$$\mathcal{L}_L(\boldsymbol{\theta}) = \widehat{p}_L(\mathbf{z}) = \frac{p(\mathbf{y}, \mathbf{t})}{p(\mathbf{y}|\mathbf{t})} \cdot \frac{p(\mathbf{z}|\mathbf{y})}{p(\mathbf{t}|\mathbf{y})} = p(\mathbf{t}) \cdot \frac{p(\mathbf{z}|\mathbf{y})}{p(\mathbf{t}|\mathbf{y})}, \tag{2.10}$$

where the terms are evaluated at $\mathbf{y} = \boldsymbol{\alpha}$ and $\mathbf{t} = \mathbf{t}_{\boldsymbol{\alpha}}$. In this form, the approximation of the integrated likelihood of the data $\mathbf{z}$ can be interpreted as a product of the integrated likelihood of the Gaussian pseudo-data $p(\mathbf{t})$, times a correction term given by the ratio of the true likelihood to the Gaussian likelihood of the pseudo-data: $p(\mathbf{z}|\mathbf{y})/p(\mathbf{t}|\mathbf{y}) = \prod_{i=1}^{n} g_i(z_i|y_i)/\mathcal{N}(t_i|y_i, d_i)$.

To achieve scalability, we approximate the density $p(\mathbf{t}) = p(\mathbf{x})/p(\mathbf{y}|\mathbf{t})$ as implied by the IW approximation $\widehat{p}_{IW}(\mathbf{x})$ in (2.8). The resulting expression for $\widehat{p}_{IW}(\mathbf{t})$ is derived in [3] for the case of $\boldsymbol{\mu} = \mathbf{0}$. We show in Section A.3 that the approximate density essentially has the same form if the prior mean is not zero:

$$-2\log\widehat{p}_{IW}(\mathbf{t}) = -2\textstyle\sum_i \log \mathbf{U}_{ii} + 2\sum_i \log \mathbf{V}_{ii} + \tilde{\mathbf{t}}'\tilde{\mathbf{t}} - \breve{\mathbf{t}}'\breve{\mathbf{t}} + n\log(2\pi),$$

where $\tilde{\mathbf{t}} = \mathbf{U}_t'(\mathbf{t} - \boldsymbol{\mu})$ and $\breve{\mathbf{t}} = \mathbf{V}^{-1}\mathbf{U}_y\tilde{\mathbf{t}}$.

Thus, for a specific parameter value $\boldsymbol{\theta}$, we run Algorithm 1 based on $\boldsymbol{\theta}$ to obtain $\boldsymbol{\alpha}_V$, set $\mathbf{y} = \boldsymbol{\alpha}_V$, $\mathbf{t} = \mathbf{t}_{\boldsymbol{\alpha}_V}$, and $d_i = (\mathbf{D}_{\boldsymbol{\alpha}_V})_{ii}$, and then evaluate the VL integrated likelihood as

$$\mathcal{L}_{VL}(\boldsymbol{\theta}) = \widehat{p}_{VL}(\mathbf{z}|\boldsymbol{\theta}) = \widehat{p}_{IW}(\mathbf{t}) \prod_{i=1}^{n} \frac{g_i(z_i|y_i)}{\mathcal{N}(t_i|y_i, d_i)}. \tag{2.11}$$

We can plug $\mathcal{L}_{VL}(\boldsymbol{\theta})$ into any numerical likelihood-based inference procedure, such as numerical optimization for finding the maximum likelihood estimator of $\boldsymbol{\theta}$, or sampling-based algorithms for finding the posterior of $\boldsymbol{\theta}$. In an iterative inference procedure, we recommend initializing $\mathbf{y}^{(0)}$ in Algorithm 1 at the mode $\boldsymbol{\alpha}_V$ obtained for the previous parameter value. Our integrated likelihood can also be used directly to evaluate the posterior of $\boldsymbol{\theta}$ over a grid of high-probability points [12, Sect. 3.1]. An extension to the integrated nested Laplace approximation (INLA) that improves the accuracy of the marginal posteriors of the y_i [12, Sect. 3.2] is straightforward.

2.3.3 Predictions at unobserved locations

We now consider making predictions at $n^\star$ unobserved locations, $\mathcal{S}^\star = \{\mathbf{s}_1^*, \ldots, \mathbf{s}_{n^\star}^*\}$, by obtaining the posterior distribution of $\mathbf{y}^\star = (y_1^\star, \ldots, y_{n^\star}^\star)'$ with $y_i^\star = y(\mathbf{s}_i^\star)$. Using the Laplace approximation as expressed in (2.5), GGP predictions are approximated as GP predictions given Gaussian pseudo-data $\mathbf{t}_{\boldsymbol{\alpha}}$ with noise covariance matrix $\mathbf{D}_{\boldsymbol{\alpha}}$.

Hence, to obtain scalable predictions at unobserved locations, we use the recommended prediction methods in [35] that apply Vecchia approximations to the multivariate normal vector $\widetilde{\mathbf{x}} = \mathbf{t} \cup \mathbf{y} \cup \mathbf{y}^\star$. For one-dimensional space, we use an extension of IW called LF-auto in [35], and for higher-dimensional space we use the RF-full method of [35]. In both cases, the pseudo-data $\mathbf{t} = \mathbf{t}_{\boldsymbol{\alpha}_V}$ and the noise variances $\mathbf{D} = \mathbf{D}_{\boldsymbol{\alpha}_V}$ are evaluated at the approximate mode $\boldsymbol{\alpha}_V$ obtained using Algorithm 1. Based on this approximation, we can compute the implied posterior distribution of $\widetilde{\mathbf{y}} = \mathbf{y} \cup \mathbf{y}^\star$ as described in Section 2.2.3: $\widehat{p}(\widetilde{\mathbf{y}}|\mathbf{t}) \sim \mathcal{N}(\widetilde{\boldsymbol{\mu}}, (\widetilde{\mathbf{V}}\widetilde{\mathbf{V}}')^{-1})$. [35] describe how to efficiently extract quantities of interest from this distribution, including the posterior mean and variances at unobserved locations. Finally, summaries or samples from the posterior of $\widetilde{\mathbf{y}}$ can be transformed to the data scale using the likelihood function $g(z|y)$, if desired. Sometimes it is difficult to compute certain predictive summaries at the data scale analytically, but it is always possible to approximate them via sampling.

Algorithm 5 in Section A.4 provides pseudo-code for maximum-likelihood estimation of parameters and for prediction.

2.3.4 Properties

2.3.4.1 Complexity

Inference for GPs with independent Gaussian noise using the Vecchia approximations considered here requires $\mathcal{O}(nm^3)$ time, where m is the maximum size of the conditioning sets $q(i)$, and can be easily parallelized [3, 35]. Our VL Algorithm 1 iteratively computes the Vecchia approximation multiple times until convergence, only adding $\mathcal{O}(n)$ cost at each iteration for computing the pseudo-data $\mathbf{t}_{\mathbf{y}^{(l)}}$. Hence, the VL algorithm requires $\mathcal{O}(knm^3)$ time, where k, the number of iterations required until convergence, can be very small (often, $k < 10$).

Once $\boldsymbol{\alpha}_V$ has been determined using Algorithm 1, evaluating the integrated likelihood (2.11) for parameter inference requires $\mathcal{O}(nm^3)$ time [3], and prediction at $n^\star$ unobserved locations requires $\mathcal{O}((n + n^\star)m^3)$ time [35]. Thus, all computational costs are linear in n for fixed m.

2.3.4.2 Approximation errors

Our VL approximation $\widehat{p}_{VL}(\mathbf{y}|\mathbf{z}) = \mathcal{N}_n(\mathbf{y}|\boldsymbol{\alpha}_V, \mathbf{W}_V^{-1})$ in (2.9) has two sources of error relative to the true posterior $p(\mathbf{y}|\mathbf{z})$: the Vecchia approximation and the Laplace approximation. Both errors are difficult to quantify in general, but our numerical experiments in Section 2.4 show that our approximation can be very accurate. The error due to the Vecchia approximation can always be reduced by increasing m [35, e.g.,].

The error of the Laplace approximation is known to depend on the likelihood being approximated. Laplace is exact for Gaussian likelihoods, in which case the VL approximation reverts to the general Vecchia approximation. For non-Gaussian spatial data, theoretical error bounds are difficult to obtain [12, Sect. 4.1]. From an empirical point of view, [53] affirm the non-spatial results of [54], showing that INLA, an extension of the Laplace approximation, generally performs well for GGPs, with the exception of some types of binomial data. [15] provide a thorough simulation study comparing Laplace to MCMC methods for parameter estimation in the case of binomial, Poisson, and negative-binomial spatial data; they conclude that the Laplace approximation is "a safe option" that is computationally practical.

2.3.4.3 Convergence

For GGPs as described in Section 2.2.1, the log-posterior in (2.1) is concave under appropriate parameterizations. Existing results show that the Newton-Raphson algorithm used in the Laplace approximation is then theoretically guaranteed to converge to its mode [55, Section 9.5.2]. In our VL Algorithm 1, the distribution $\widehat{p}(\mathbf{y})$ implied by the general Vecchia approximation changes at each iteration, which makes it difficult to theoretically guarantee convergence, except in special cases. Fortunately, empirical testing of Algorithm 1 under different parameter and data settings showed that convergence can always be expected when machine precision is not an issue.

2.4 Simulations and comparisons

We compared our VL approaches to other methods using simulated data. Throughout Section 2.4, unless specified otherwise, we simulated realizations $\mathbf{y}$ on a grid of size $\sqrt{n} \times \sqrt{n}$ on the unit square from a GP with mean zero and a Matérn covariance function with variance 1, smoothness ν, and range parameter $\lambda = 0.05$. Gridded locations allow us to carry out simulations for large n using Fourier methods. The data were then generated conditional on $\mathbf{y}$ using the four likelihoods in Table 2.1, with $a = 2$ in the Gamma case.

As low-rank approximations are very popular for large spatial data, we also considered a fully independent conditional or modified-predictive-process approximation to Laplace with m knots (abbreviated as LowRank here), which is equivalent to VL-IW except that each conditioning set $q_y(i) = (1, \ldots, m)$ simply consists of the first m latent variables in maxmin ordering. This equivalence allowed us to run VL and LowRank using the same code base, thus avoiding differences solely due to programming.

Criteria used for comparison are the run time (on a 2017 MacBook Pro), the relative root mean square error (RRMSE) and the difference in log scores (dLS). Results are averaged over 100 simulated datasets, unless noted otherwise. The RRMSE is the root mean square error of the posterior mean of $\mathbf{y}$ obtained by one of the approximation methods relative to the true simulated $\mathbf{y}$, divided by the RMSE of the Laplace approximation. The log score is computed as the negative

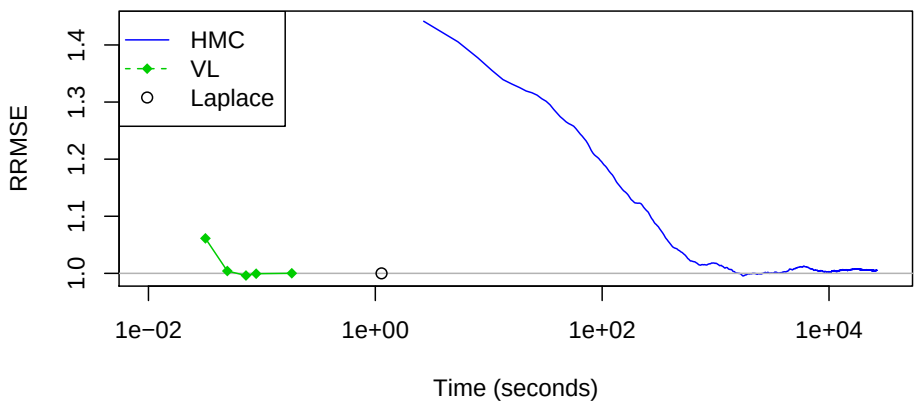


Figure 2.2: RRMSE versus time (on a log scale) for Bernoulli data of size $n = 625$ on the unit square. Laplace is run once until convergence. For VL-RF, we considered $m \in \{1, 5, 10, 20, 40\}$. The number of HMC iterations varies from 10,100 to 1,000,000 in increments of 100, with the first 10,000 considered burn-in.

logarithm of the approximated posterior density of $\mathbf{y}$ evaluated at the true $\mathbf{y}$, with low values corresponding to well calibrated and sharp posterior distributions [56, Sect. 3]. The dLS is the log score of an approximation method minus the log score for the Laplace approximation. When averaged over a sufficient number of simulated data, the dLS can be shown to approximate the difference between the Kullback-Leibler (KL) divergence of the exact posterior distribution and the considered approximation, minus the KL divergence between the exact distribution and the Laplace approximation.

2.4.1 Comparison to MCMC

Non-Gaussian spatial models are often fitted using Markov chain Monte Carlo (MCMC), which under mild regularity conditions is "exact approximate," converging to the true posterior as the number of iterations approaches infinity. For finite computation time and large n, however, MCMC results can be very poor relative to the Laplace approximation. We demonstrate this with a single simulated dataset consisting of $n = 625$ Bernoulli observations based on a GP with smoothness parameter $\nu = 0.5$ on the unit square. We compared Laplace and VL-RF to Hamiltonian Monte Carlo (HMC) [57], a MCMC method well suited to sampling correlated variables. As shown in

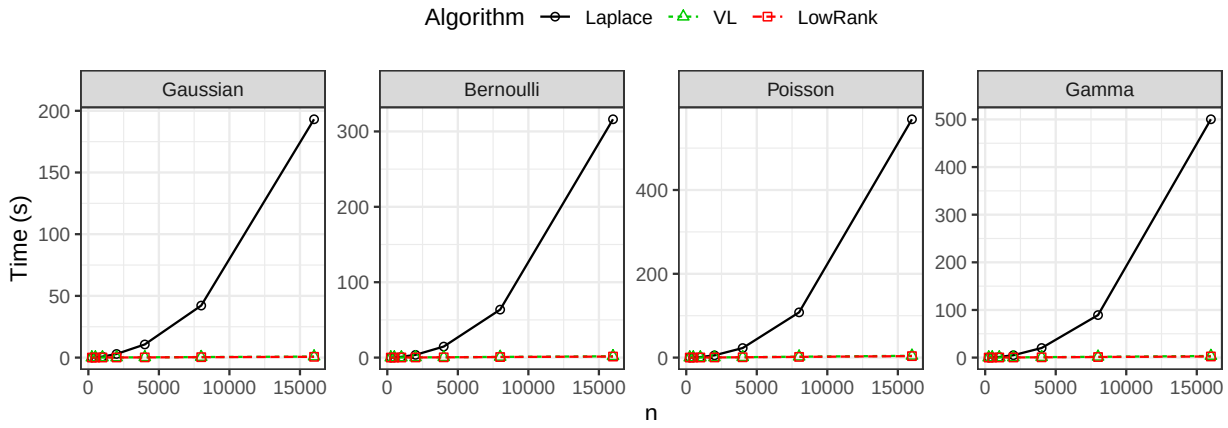


Figure 2.3: For sample size n between 250 and 16,000, computing time for the Laplace approximation based on Newton-Raphson, compared to VL and LowRank using Algorithm 1 with $m = 10$.

Figure 2.2, VL quickly achieved the same accuracy as Laplace as m increased, but at a fraction of the computing time. In contrast, HMC took orders of magnitude longer to achieve similar accuracy. Even with 1 million iterations, the RMSE for HMC was slightly higher than for VL; this is in line with existing simulation studies suggesting that the Laplace approximation error may be negligible in many GGP settings (see Section 2.3.4.2). We expect the relative performance of HMC to degrade further as n increases. More details and results can be found in Section A.5.

2.4.2 Computational scaling of Laplace approximations

While the Laplace approximation is very useful for moderate data sizes n, we now briefly illustrate the computational infeasibility for large n due to its cubic scaling. In Figure 2.3, we show the average computation time for observations with smoothness $\nu = 0.5$ in the setting described later in Section 2.4.4. Clearly, Laplace using Newton-Raphson quickly became infeasibly slow as n increased, while VL and LowRank were much faster.

2.4.3 VL accuracy in one-dimensional space

We now compare the accuracy of the VL and LowRank approximations. Both approaches scale linearly in n for fixed m, and both approaches converge to the Laplace approximation as m increases, with equivalence guaranteed for $m = n - 1$.

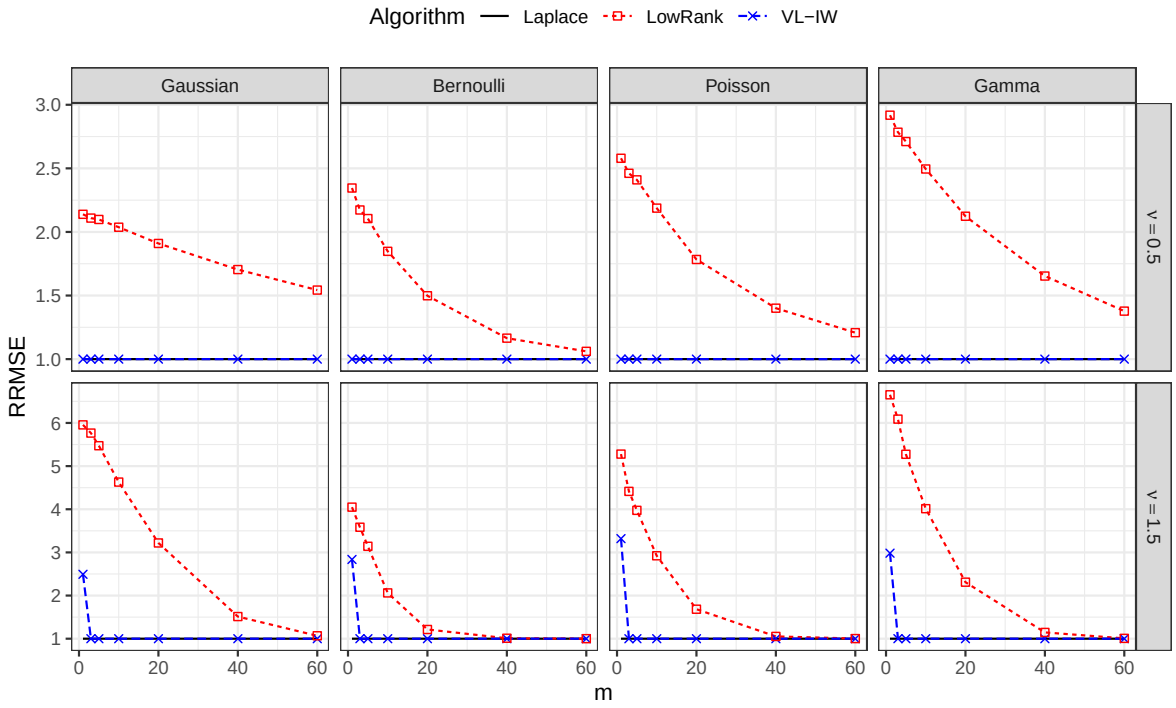
Figure 2.4 shows the average results for 100 simulated datasets of size $n = 2{,}500$ each, on the

unit interval. For the Gaussian likelihood, the noise variance was $\tau^2 = 0.1^2$. Clearly, VL-IW was extremely accurate and delivered essentially equivalent results to the Laplace approximation, even for very small m. For exponential covariance (i.e., Matérn with smoothness $\nu = 0.5$), an exact screening effect holds in one-dimensional space, and so VL-IW is exactly equal to Laplace for any $m \geq 1$. LowRank required much larger m to achieve equivalence to Laplace.

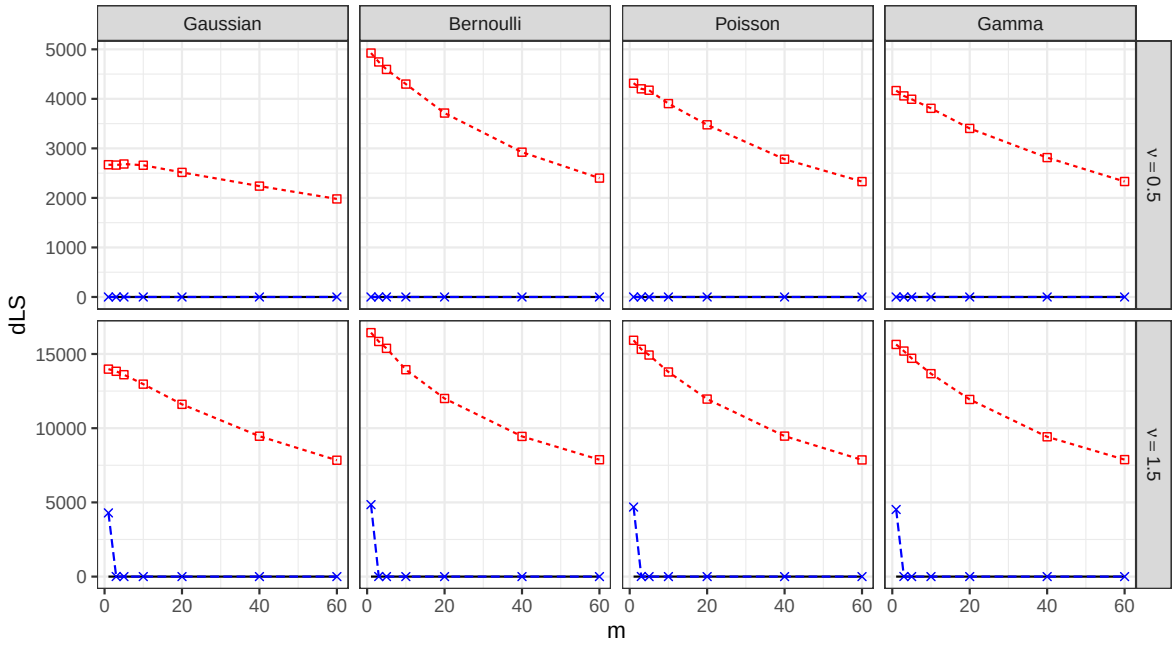
2.4.4 VL accuracy in two-dimensional space

Figure 2.5 shows results for the same simulation study as in Section 2.4.3, except that the data were simulated on the two-dimensional unit square, with noise variance $\tau^2 = 0.1$ for the Gaussian likelihood. While all methods are again equivalent to Laplace for $m = n - 1$, the two-dimensional problem is considerably more difficult, and higher values of m were required for accurate approximations. As we can see, the recommended VL-RF had roughly equivalent performance to Laplace once m reached 20, and it was more accurate than VL-IW for $m > 10$. LowRank performed considerably worse than the VL methods, and further simulations (not shown) showed that in some cases LowRank approached the accuracy of Laplace only when m was almost as large as n. A simulation with larger range parameter $\lambda = 0.2$ is shown in Section A.6.1 of the supplement; while VL-RF was still more accurate than LowRank for all settings, the larger range reduced the amount of fine-scale variation, thus reducing the advantage of VL over LowRank relative to Figure 2.5, especially for logistic regression models. The relative performance of the methods was similar in higher dimensions; plots for 3 and 4 dimensions are shown in Section A.6.2.

For larger n, the differences between LowRank and VL became even more pronounced. Figure 2.6 shows the RMSE for simulations with increasing sample size n but fixed m. VL-RF improved in accuracy under this asymptotic in-fill scenario almost as fast as Laplace, while LowRank failed to improve.

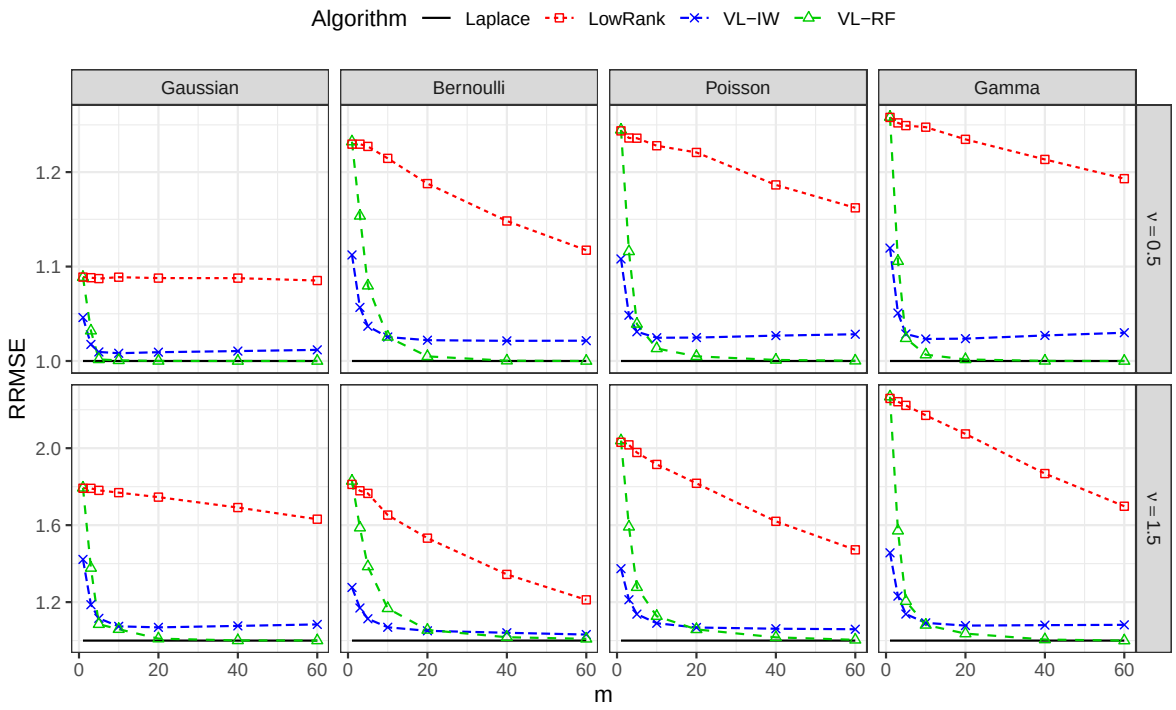


(a) RMSE (relative to Laplace)

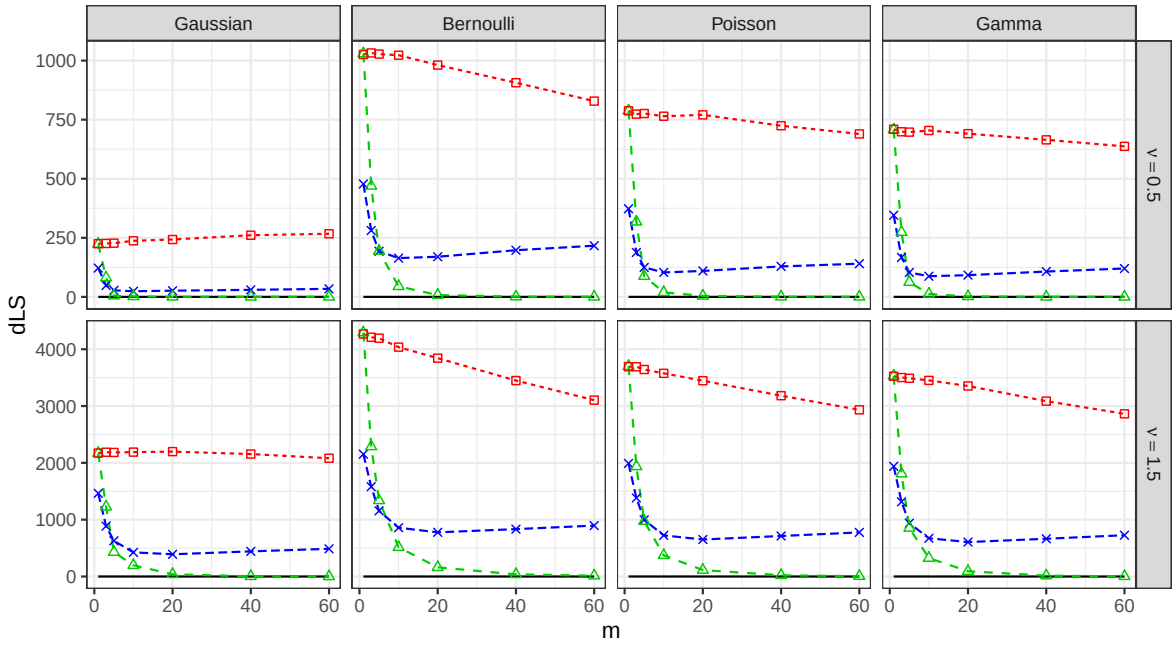


(b) Difference in log score (relative to Laplace)

Figure 2.4: Simulation results for $n = 2,500$ observations on a one-dimensional domain.



(a) RMSE (relative to Laplace)



(b) Difference in log score (relative to Laplace)

Figure 2.5: Simulation results for $n = 2{,}500$ observations on a two-dimensional domain.

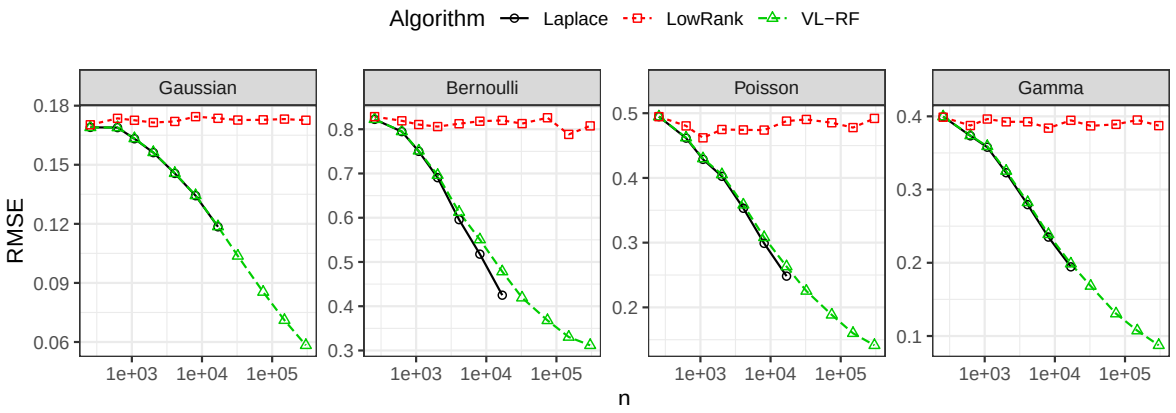


Figure 2.6: On a two-dimensional domain with $\nu = 0.5$ and fixed $m = 10$, RMSE between true $\mathbf{y}$ and posterior mode $\boldsymbol{\alpha}_V$ for increasing sample size n (on a log scale) up to 300,000. Laplace without further approximation becomes prohibitively expensive for large n, so we only computed it up to $n = 16{,}000$.

2.4.5 Simulations for log-Gaussian Cox processes

Point patterns are sets of points or locations $\mathbf{s}_1, \ldots, \mathbf{s}_N$ in a domain $\mathcal{D}$. A popular model for point patterns is the log-Gaussian Cox process (LGCP), a doubly stochastic Poisson process whose intensity function $\lambda(\cdot)$ is modeled as random, $\log \lambda(\cdot) = y(\cdot) \sim GP(\mu, C)$. Inference for LGCPs is difficult due to stochastic integrals.

A natural approximation [58, e.g.,] for LGCPs relies on partitioning the domain $\mathcal{D}$ into n grid cells $A_1, \ldots, A_n$ with center points $\mathbf{a}_1, \ldots, \mathbf{a}_n$, respectively. The number of observed points falling into A_i is treated as the data, $z_i = z(A_i) = \sum_{j=1}^{N} 1_{\mathbf{s}_j \in A_i}$. These gridded data conditionally follow a Poisson distribution, $z_1, \ldots, z_n \,|\, y(\cdot) \overset{ind.}{\sim} \mathcal{P}(\mu(A_i))$, where

$$\mu(A_i) = \textstyle\int_{A_i} \lambda(\mathbf{s}) d\mathbf{s} \approx |A_i| \, \lambda(\mathbf{a}_i) = |A_i| \, e^{y(\mathbf{a}_i)}.$$

This model falls under the GGP framework, so we can apply our VL methods to obtain fast inference for point patterns.

Figure 2.7 shows a LGCP whose log-intensity is modeled as a GP with Matérn covariance with range parameter 2.5 on a spatial domain $\mathcal{D} = [0, 50]^2$, discretized into $n = 2{,}500 = 50 \times 50$ unit-square grid cells. This is equivalent to the simulation in Section 2.4.4, as the domain can be scaled

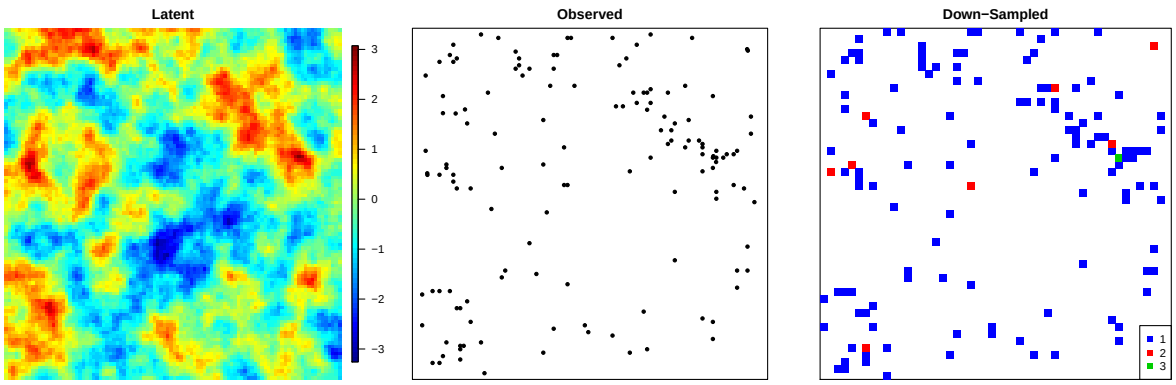


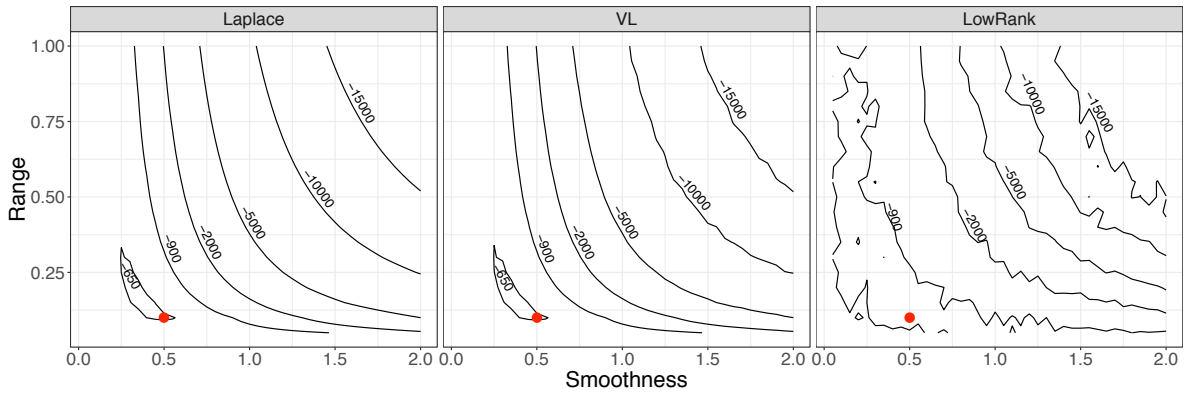
Figure 2.7: Gridding a simulated LGCP point pattern: The latent log-intensity $y(\cdot)$ (left), a corresponding simulated point pattern (center), and the down-sampled Poisson count data used for analysis on a $n = 50 \times 50 = 2{,}500$ grid (right).

to a unit-square domain with range 0.05, but on the original scale the grid induces areal regions with unit area, $|A_i| = 1$, with intensity function $\mu(A_i) = \exp(y(\mathbf{a}_i))$. Thus, the averaged results for fitting repeatedly simulated datasets from this LGCP are equivalent to the Poisson results shown in the third column of Figure 2.5, indicating that VL can be used to obtain virtually equivalent inference to that using a Laplace algorithm, albeit at much lower computational cost for large n.

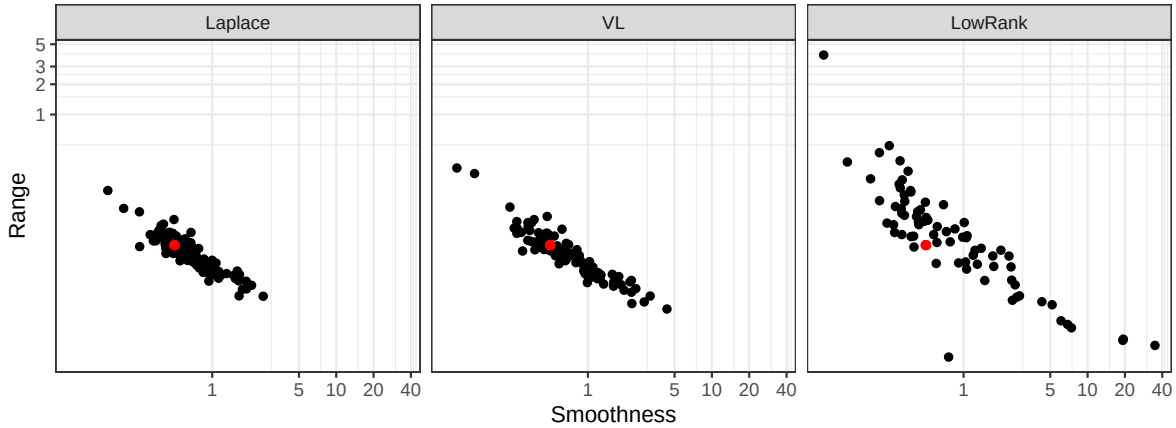
2.4.6 Parameter estimation

We also explored parameter estimation based on each method's integrated-likelihood approximation. Specifically, we considered Poisson data at $n = 625$ locations in the unit square, based on a GP with true smoothness $\nu = 0.5$ and range $\lambda = 0.05$.

First, we simulated a single realization of the spatial data. Holding the variance fixed at the true value of one, we sequentially evaluated the integrated likelihood on a grid of values for the range and smoothness parameters, using the Laplace approximation in (2.10), and the VL-RF approximation with $m = 20$ in (2.11). The exact integrated likelihood is intractable. As shown in Figure 2.8, the integrated likelihoods as approximated by Laplace and by VL were almost identical, while the LowRank approximation was quite poor. These likelihood approximations are equivalent to approximations to the posterior distribution $p(\boldsymbol{\theta}|\mathbf{z})$ assuming flat priors for $\boldsymbol{\theta}$. This indicates that Bayesian inference for GGPs can be carried out quickly and accurately using the VL approxima-



(a) For a single realization of Poisson data, contour lines of the integrated likelihood at $(-650, -900, -2000, -5000, -10000, -15000)$



(b) Parameter estimates obtained by optimizing the integrated likelihood for 100 sample realizations; also see Table 2.2

Figure 2.8: For Poisson data at $n = 625$ locations in the unit square, comparison of different approximations to the integrated likelihood, using conditioning sets of size $m = 20$ for VL and LowRank. Red dots show the true parameter values.

tion.

We then simulated 100 different realizations of the spatial Poisson data and examined the parameter estimates obtained by maximizing the different approximations to the integrated likelihood. The scatter plot in Figure 2.8 shows the parameter estimates, using $m = 20$ conditioning points for VL and LowRank. While the estimates using Laplace and VL were similar, LowRank had significant outliers that increased the RMSE of the parameter estimates (see Table 2.2). The LowRank parameter estimation frequently diverged due to the rough likelihood surface, and for those cases we repeated the optimization with bounds $(0.001, 20)$ for both range and smoothness,

	Range			Smoothness		
	$m = 10$	$m = 20$	$m = 40$	$m = 10$	m=20	m=40
LowRank	0.107	0.407	0.098	9.17	8.40	12.60
VL	0.293	0.040	0.023	1.01	0.78	0.47
Laplace			0.023			0.51

Table 2.2: For 100 simulated Poisson datasets at $n = 625$ locations in the unit square, RMSE for parameter estimates based on different approximations to the integrated likelihood. Both range and smoothness parameters were bounded to the interval $[0.001, 20]$, but LowRank estimation still failed repeatedly.

but LowRank still failed repeatedly.

2.4.7 Interpretation of simulation results

In our simulations, VL provided similar accuracy as Laplace with a considerably smaller number m of conditioning points compared to LowRank. The time required per iteration for VL approaches is $\mathcal{O}(nm^3)$. At the expense of fully parallel computation, LowRank can be carried out in $\mathcal{O}(nm^2)$ time by computing the decomposition of the covariance of the conditioning set once at the beginning of the procedure. However, as VL with any given m, say $m = \widetilde{m}$, was substantially more accurate than LowRank with $m = \widetilde{m}^{3/2}$, we conclude that VL is more computationally efficient than LowRank for a given approximation accuracy, except for very smooth posteriors. The improvement in accuracy for VL relative to LowRank became even more pronounced as we increased the sample size under in-fill asymptotics.

2.5 Application to satellite data

We applied our methodology to a large, spatially correlated, non-Gaussian dataset of column water vapor. These data were collected by NASA's Moderate Resolution Imaging Spectroradiometer (MODIS), which is mounted on the NASA Aqua satellite [59]. We considered a Level-2 near-infrared dataset of total precipitable water at a $1354 \times 2030 = 2{,}746{,}820$ grid of 1km pixels. We used up to 500,000 of these data points for our demonstration. Our dataset was observed between 13:45 and 13:50 on March 28, 2019 over a rectangular region off the coast of west Africa with west, north, east, and south bounding coordinates -42.707, 67.476, 4.443, and 45.126, respectively and was found on the NASA Earthdata website, `https://earthdata.nasa.gov`.

Precipitable water amounts are continuous and strictly positive, with values near 0 corresponding to clear skies and larger values implying more water. Exploratory plots showed a right-skewed density, so we assumed that the data can be modeled using a spatial generalized GP with a Gamma likelihood:

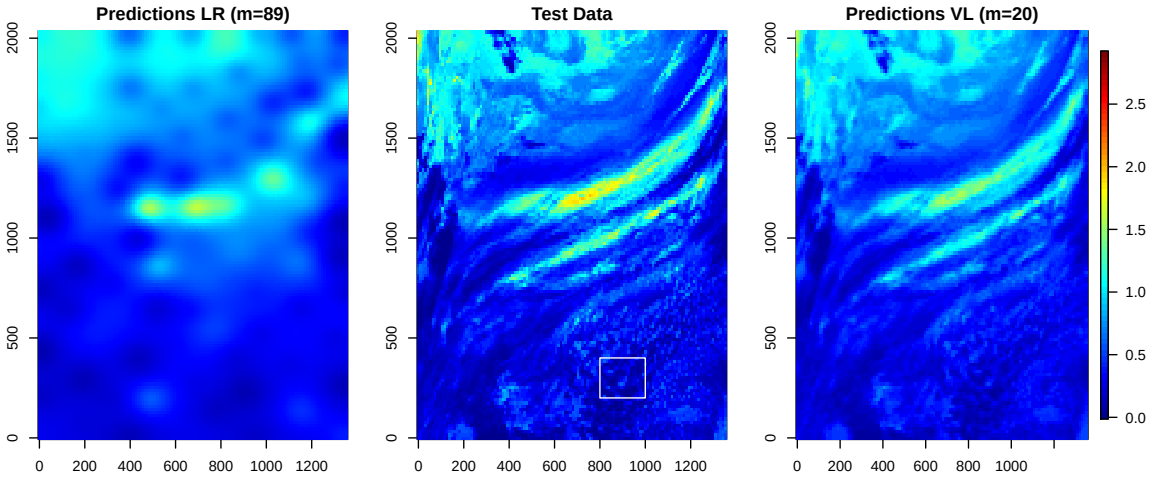
$$z(\mathbf{s}_i)|y(\mathbf{s}_i) \overset{ind.}{\sim} \mathcal{G}(a, ae^{-y(\mathbf{s}_i)}), \qquad y(\cdot) \sim \mathcal{N}(\mu, K),$$

where $E(z(\mathbf{s})|y(\mathbf{s})) = \exp(y(\mathbf{s}))$, $\mu(\mathbf{s}) = \beta_1 + \beta_2 \,\text{lat}(\mathbf{s})$ is a linear trend consisting of an intercept and a latitudinal gradient, and K is an isotropic Matérn covariance function with variance σ^2, smoothness ν, and range parameter ρ. We estimated the parameter values $\beta_1 = -1.515$, $\beta_2 = 0.000766$, $a = 0.89$, $\sigma^2 = .25$, $\rho = 31$km, and $\nu = 3$ as described in Section A.8.

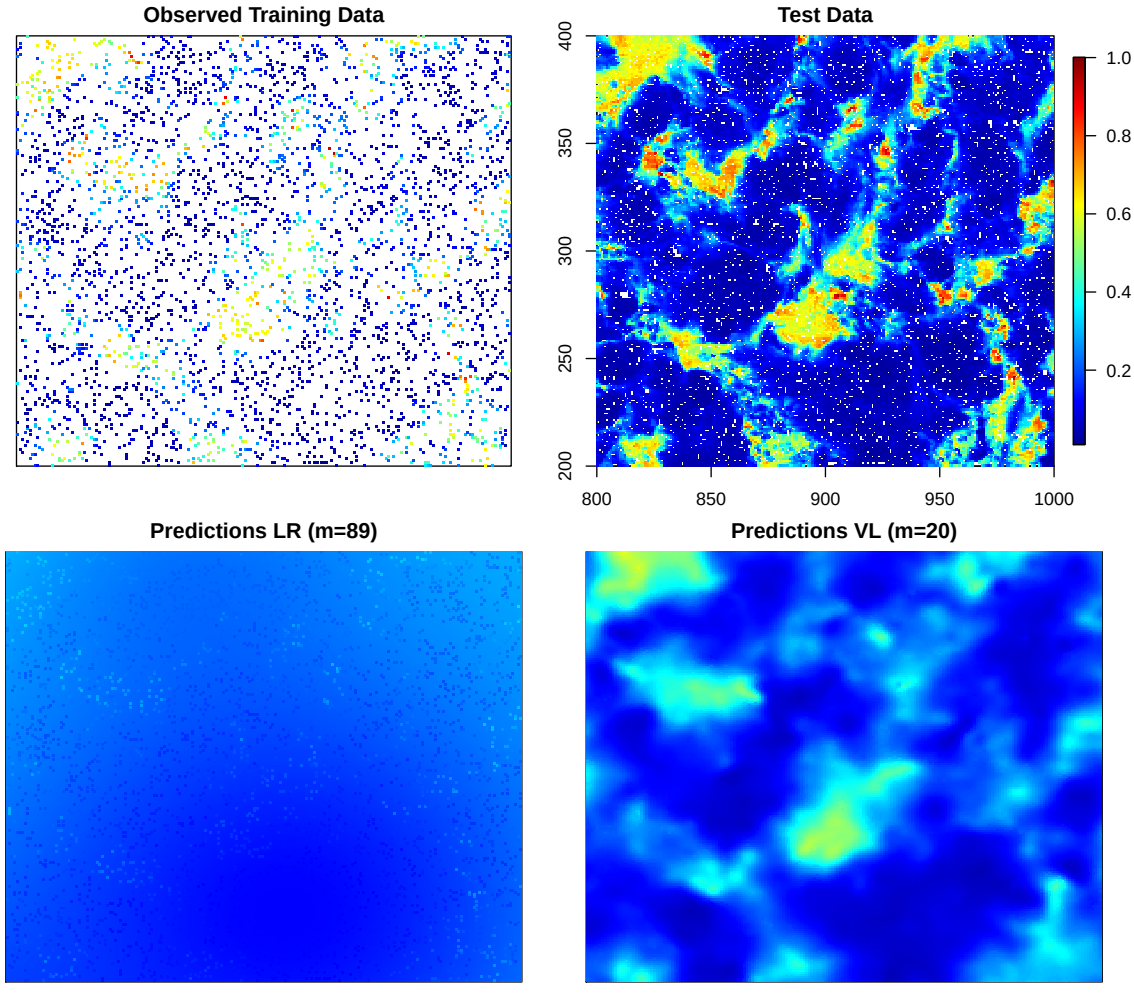
We again compared our VL approach to a LowRank method. We randomly sampled $n = 250{,}000$ observations $\mathbf{z}$ of the full dataset as training data, and $250{,}000$ of the remaining observations as test data $\mathbf{z}^\star$ at locations $\mathcal{S}^\star$. For VL, we set $m = 20$ following our recommendations in Section 2.4.4 and further justified in Section A.8. For LowRank, we used $m = 89 \approx (20)^{3/2}$ for a computationally fair comparison. On an Intel Xeon E5-2690 CPU with 64GB RAM, Algorithm 1 for VL required 10 iterations with a total run time of about 18 minutes (1.8 minutes per iteration). Taking advantage of an implementation that achieves the $\mathcal{O}(nm^2)$ scaling, each iteration for LowRank required 1.3 minutes on average across 6 iterations. Note that, based on our numerical experiments, we estimate that Laplace without further approximation would take months of computing time, while HMC-based approaches would take years to achieve the same accuracy as VL.

Figure 2.9a shows prediction maps of the posterior mean $E(\mathbf{z}^\star|\mathbf{z}) = E(\exp(\mathbf{y}^\star)|\mathbf{z})$ with ith entry $\exp(E(y_i^\star|\mathbf{z}) + var(y_i^\star|\mathbf{z})/2)$. Clearly, much of the fine-scale structure was lost when using LowRank. To further illustrate this issue, we made predictions on a 200×200 grid over a small subregion. As shown in Figure 2.9b, the LowRank predictions were virtually useless at this scale, while VL was able to recover much of the important spatial structure from the noisy and incomplete training data.

Table 2.3 quantifies the improvement in predictions using VL over LowRank. We computed



(a) Entire spatial domain



(b) Zooming into the white square shown in Panel (a)

Figure 2.9: Prediction maps for MODIS data using VL and LowRank (LR).

Method	MSE	CRPS
VL	0.0149	0.144
LowRank	0.0528	0.170
Ratio	3.54×	1.18×

Table 2.3: For the MODIS data, comparison of prediction scores (lower is better) between VL and LowRank.

the MSE based on the posterior mean $E(\mathbf{z}^\star|\mathbf{z})$. To compare the accuracy of the uncertainty quantification, we also computed the continuous ranked probability score (CRPS) [56, e.g.,], which encourages well calibrated and sharp predictive distributions. Table 2.3 shows that VL strongly outperformed LowRank for comparable computational complexity.

2.6 Conclusions and future work

In this work, we presented a novel combination of techniques that allow for efficient analysis of large, spatially correlated, non-Gaussian datasets or point patterns. The key idea is to apply a Vecchia approximation to the Gaussian (and hence tractable) joint distribution of GP realizations and pseudo-data at each iteration of a Newton-Raphson algorithm, leading to a Gaussian Laplace approximation. This allows us to carry out inference for non-Gaussian data by iteratively applying existing Vecchia approximations for Gaussian pseudo-data, which are updated at each iteration. Our Vecchia-Laplace (VL) techniques guarantee linear complexity in the data size while capturing spatial dependence at all scales. Compared to alternative methods such as low-rank approximations or sampling-based approaches, our VL approximations can achieve higher accuracy at a fraction of the computation time.

Vecchia approximations require specification of an ordering of the model variables and of a conditioning set for each variable, and these two issues also play a critical role in the performance of our VL approaches. Through simulation studies, we showed that, in one-dimensional space, interweaving the GP realizations and the pseudo-data [3] can provide results that are virtually indistinguishable from Laplace, even for very small conditioning sets. For two-dimensional space, we recommend the response-first Vecchia approximation [35]. Due to the computational efficiency of our approach, it is also possible to use a VL approximation of the integrated likelihood for

parameter inference, for which we recommend the interweaved ordering in any dimension.

The methods and algorithms proposed here are implemented in the `R` package `GPvecchia` [49]. The default settings of the package functions reflect the recommendations in the previous paragraph. The tuning parameter m, which controls a trade-off between accuracy and computation cost, can be set by the user. In practice, a useful strategy is to start with a relatively small value of m and gradually increase it until the inference converges or the computational resources are exhausted.

Our methods and code are applicable in more than two dimensions, but a thorough investigation of their properties in this context will be carried out in future work. For example, [60] show that Vecchia-based approximations with appropriate extensions can be highly accurate for computer-model emulation in up to ten dimensions; a combination with our VL methods could allow emulation of non-Gaussian computer-model output. Other potential future work includes extending the Laplace approximation in our methods to an integrated nested Laplace approximation (INLA) that improves the accuracy of the marginal posteriors of the latent variables [12, Sect. 3.2]; the use of conjugate-gradient [61] or incomplete-Cholesky [50] methods that allow the computation of the latent posterior mean in linear time even for completely latent Vecchia approximations; or extensions to spatio-temporal filtering using Vecchia approximations based on domain partitioning [62, 63].

3. SPATIAL SURFACE RETRIEVALS FOR VISIBLE/SHORTWAVE INFRARED REMOTE SENSING

3.1 Introduction

Remote Visible/ShortWave InfraRed (VSWIR) imaging spectroscopy is a powerful tool for studying Earth science questions ranging from geology, to the cryosphere, to the composition of terrestrial and aquatic ecosystems [64]. These instruments, such as the Airborne Visible-Infrared Imaging Spectrometer - Next Generation, or AVIRIS-NG [65], measure a full spectrum of reflected solar radiant intensity, from visible wavelengths through the shortwave infrared, at every location in a scene. Such instruments can be a component on an orbiting satellite or mounted in an aircraft to offer greater flexibility over where the data are collected. The physical and chemical composition of the surface induces absorption features which modify the spectral shape of the measured radiance. These radiance shapes indicate what materials are present in the spatial footprint of the spectrum. However, the intervening atmosphere also modifies the radiance with various absorption and scattering processes along the light path from the sun to the ground to the sensor. Consequently, analysts first remove the atmospheric effects to estimate the intrinsic reflectance of the surface [66]. It is the resulting reflectance spectrum, free from atmospheric influence, which is used in all subsequent studies of surface composition.

Estimating surface reflectance requires modeling how the atmosphere contributes to the illumination measured at the sensor. Existing implementations of radiative transfer models such as MODTRAN [67] and LibRadTran [68] model the observation with variety of parameters, including geometric terms like the camera and sun position, and atmospheric terms such as the vertical distribution of water vapor and aerosols. These codes then solve the equations of radiative transfer to predict the radiance that will be measured at the sensor. The radiative transfer model acts as a nonlinear function which predicts the radiance for a given surface and atmospheric state. The challenge then is to invert this nonlinear model to estimate the most probable surface and atmospheric

state variables which might have produced the observation [1].

There are many algorithms for inverting the nonlinear physical model, such as those based on the ATmosphere REMoval algorithm (ATREM) [69, 70]. In all previous imaging spectroscopy literature, the inversion models have operated on each pixel independently. In other words, they have assumed the latent surface and atmosphere states generating the measurements are spatially independent. This is reasonable for the surface variables given that the surface materials change abruptly; for example, there is no reason to assume a tree should have a surface state that correlates with a nearby asphalt road. But atmospheric variables like water vapor are smooth and continuous over space, and so nearby observations will have highly correlated atmospheric states. By ignoring this correlation, preceding works have ignored powerful information that can be used to improve the fidelity of both atmosphere and reflectance estimates.

In this work we demonstrate the first ever joint inversion of multiple locations for imaging spectroscopy, respecting the local correlations in the atmosphere. We focus on qualitative improvements, uncertainty quantification, and scalability aspects of the spatial inversion. Qualitatively, ignoring spatial correlations in atmospheric states is not problematic if the single-pixel atmospheric retrievals are accurate. This is the case in many dry, homogeneous scenes. However, errors can become significant in the case of high aerosol loads or high water vapor content, where systematic retrieval uncertainties dependent on surface type can cause discontinuities in the retrieved atmospheric field. While post-hoc smoothing via spatial prediction or Gaussian process regression [71] can be applied after computing single-pixel retrievals [72], the dependencies introduced by the non-linear forward model are completely ignored. As a result, the reflectances still contain the error of the unsmoothed atmospheric components, making a principled estimate of uncertainty in the state estimates problematic.

Uncertainty quantification (UQ) for surface retrievals has been developed under the label of optimal estimation (OE) [73, 5]. Modeling correlations across push-broom measurements has been shown to improve variance measurements [74], but only recently have the surface and atmosphere states been modeled jointly to decrease error while simultaneously achieving UQ [4]. A similar

approach is used in [75], although with multiband input data that includes multiple angles and polarization rather than a single radiance measurement.

Here we propose to include the spatial correlation in the inversion itself, improving the reflectance retrievals while allowing more appropriate reflectance uncertainties to be propagated downstream. As in previous work, our method relies on a hierarchical model in which the observed radiance is a noisy version of the true radiance, which in turn is a nonlinear function of the state vector. The prior state vector is modeled as a multivariate Gaussian with a covariance structure reflecting how the variables in a state vector for a single location correlate with each other. Uniquely, we extend this covariance into a cross-covariance matrix to represent spatial correlations in the atmospheric terms. This transforms the multivariate Gaussian prior into a multivariate Gaussian process prior, capturing the spatially-smooth behavior of atmospheric fields.

Retrievals for multiple spatial locations have been investigated for other applications under the OE framework. The approach has been implemented for multiple instruments focused on aerosol retrievals from multi-angle observations [76, 75] and for atmospheric trace gas retrievals [77] with a simplified linear model. These applications share the general strategy of exploiting spatial correlation in space for retrieval of atmospheric state variables. In the current setting, the dimension of the surface state is substantially larger and is the primary quantity of interest, requiring additional computational considerations.

The remainder of this article is organized as follows. Section 2 reviews the nonlinear, independent surface retrieval model. Section 3 introduces the spatially correlated version of the model and some considerations for scalability. Section 4 has a simulation study, Section 5 has an application, and Section 6 provides conclusions.

3.2 Optimal estimation of surface reflectance

A representative radiance spectrum, and its associated reflectance, appear in Figure 3.1. The radiance spectrum represents energy incident at the detector per unit wavelength per solid angle per unit area, in units of $\mu\, W nm^{-1} sr^{-1} cm^{-2}$. Sharp dips at 940, 1140, 1380 and 1880 nm represent the influence of absorbing atmospheric gases. The reflectance spectrum at right, showing the spectrum

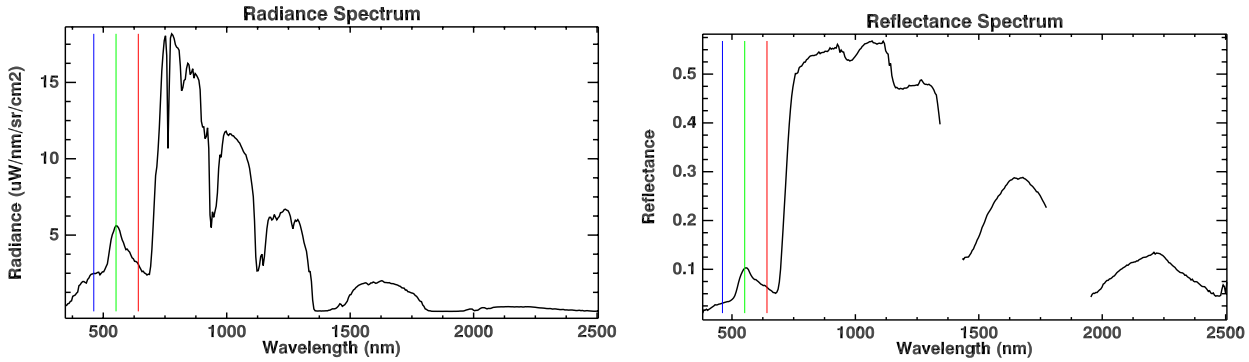


Figure 3.1: Representative radiance and reflectance spectra, adapted from [1]. Red, green, and blue lines indicate visible color channels

.

of a vegetated pixel, is comparatively smooth. Roughly speaking, it represents the ratio of light leaving the target over the light hitting the target, which is an intrinsic property of the surface. The deepest absorption features at 1380 and 1880 nm are not plotted; atmospheric gas absorption in these wavelengths is so strong that the atmosphere is opaque and it is not possible to estimate the surface reflectance. Mathematically, a single radiance observation y is a vector of intensity values corresponding to a set of wavelengths as measured by a remote sensor. The satellite radiance can be expressed as a function of the surface reflectance according to a forward model that takes into account atmospheric and physical effects, $y \approx f(x)$. We denote the surface state x, which combines the reflectance x_s with additional components corresponding to atmospheric conditions x_h. Optimal estimation [5] refers to the inversion of the forward model to compute the surface reflectance x given remotely sensed observations y and a prior assumption on x in a Bayesian context.

In this work, the additional components are aerosol optical depth (AOD) and column water vapor. The deterministic forward model is at the heart of the surface retrieval process and is briefly reviewed before describing the baseline retrievals (Section 3.2.2) and the details of the statistical model that will be relevant for our methodology (Section 3.2.3). This section is a summary of the statistical analysis described in [4], which contains many additional details.

3.2.1 Forward model and uncertainty

The forward model is a nonlinear function describing the processes of absorption and scattering of light by atmospheric gases, particulates and clouds, and reflection by an underlying surface, and is referred to as the radiative transfer model (RTM). The true physical model is complicated, so many simplifying assumptions are used, such as treating surfaces as Lambertian (isotropic) rather than describing them using a bidirectional reflectance distribution function. Since there are many parameters to the RTM, optimization over all possible combinations is infeasible. Instead, a look-up table of optical coefficients is calculated in advance. This table, indexed by the atmospheric state, allows a fast calculation of the forward model in each channel [1]. For a full description of the forward model assumptions, we refer the reader to previous work [78]. Uncertainties in the radiance prediction include instrument-related uncertainty such as measurement noise, as well as errors in atmospheric properties such as aerosol absorption or scattering. In the following experiments, we use the LibRadTran radiative transfer library [68] with the ISOFIT inversion package [4]. This allows us to focus on the specific innovations of this paper, the prior specification and the optimization procedure.

3.2.2 Baseline optimal retrievals

As mentioned in the introduction, the baseline retrieval model assumes that any one observed radiance y with dimension 425 is a nonlinear function of a latent state x of dimension 434, independent of any nearby data: $y = f(x) + \epsilon$. The forward model function $f(\cdot)$ described in the previous section is an approximation to the true physical system with higher-order complexities relegated to a Gaussian error term. The state x is given a Gaussian prior to provide a tractable posterior $x|y$ when combined with a linear approximation (as in the Levenberg-Marquardt algorithm) for the non-linear forward model:

$$p(y|x) \sim N(f(x), S_\epsilon), \quad p(x) \sim N(\mu, S_a),$$

$$p(x|y) \propto p(y|x)p(x).$$

The prior $p(x)$ is discussed in the next section. The likelihood variance term for a single observation S_ϵ can be attributed to instrument noise and unobserved variables.

The optimal state vector $\hat{x}$ is understood to be the retrieved vector that maximizes the posterior density $p(x|y)$, given prior assumptions and observations y. Negating, taking a logarithm, and dropping constants of the posterior yields a minimization problem with respect to a cost function $Q(x) \propto -\log p(x|y) + \text{constant}$:

$$Q(x) = (x-\mu)^\top S_a^{-1}(x-\mu) + (y-f(x))^\top S_\epsilon^{-1}(y-f(x)). \tag{3.1}$$

The optimal estimate for the cost function Q can be found with the Newton-Raphson algorithm, which is an iterative method with update steps

$$x^{(\ell+1)} = x^{(\ell)} - [\nabla_x^2 Q]^{-1}\nabla_x Q. \tag{3.2}$$

As shown in Appendix B.1, the Levenberg-Marquardt variant of Newton-Raphson is an approximation that yields an inexpensive update step of the form

$$\begin{aligned} x^{(\ell+1)} &= \mu + [S_a^{-1} + K^\top S_\epsilon^{-1} K]^{-1}[K^\top S_\epsilon^{-1} K(x^{(\ell)} - \mu) - K^\top S_\epsilon^{-1}(y - f(x^{(\ell)})] \\ &= \mu + \Delta_{LM}. \end{aligned} \tag{3.3}$$

When the iterations converge to some state $x^\star$, the converged value represents the posterior mode, which can also be viewed as the mean of a Gaussian approximation to the posterior at the mode. The uncertainty is approximated with

$$S_\star = [S_a^{-1} + K_\star^\top S_\epsilon^{-1} K_\star]^{-1}. \tag{3.4}$$

The posterior is then approximated with the distribution $N(x_\star, S_\star)$, where the optimal estimate is $x_\star$ with uncertainty $S_\star$.

3.2.3 Prior

In preparation for our spatial methodology, we detail the prior used for the baseline optimal estimation procedure. Recall that the prior state contains a surface state x_s and an atmosphere state x_h. The baseline method inverts each radiance measurement independently, and further assumes that the surface and atmosphere states are independent. This is represented with block diagonal covariances S_s, S_h that make up a prior multivariate normal distribution:

$$N\left(\begin{bmatrix} x_s \\ x_h \end{bmatrix}, \begin{bmatrix} S_s & 0 \\ 0 & S_h \end{bmatrix} \right) = N(\mu, S_a).$$

The surface state components for a single radiance measurement can co-vary, as can the atmospheric state components; two prior states x_i, x_j at different locations are however totally independent. Allowing different pixels to have co-varying atmospheric states will involve a cross-covariance function and is the focus of Section 3.3.

Natural and man-made materials have different reflectance profiles, so there are multiple prior means $\mu_k = [x_{s,k}, x_h]^\top$ and variances $S_{a,k}$, $k = 1, ..., \kappa$ to take this into account. Note that there is a single global prior mean and variance for the atmospheric components. At the first iteration of the optimization routine, a heuristic algebraic inversion is used to estimate the reflectance, and then the closest prior is selected in an ad-hoc way using a Euclidean distance $||x^{(\ell)} - x_{a,k}||$ or Mahalonobis distance:

$$d(k) = ||x^{(\ell)} - \mu_k||^2_{S^{-1}_{a,k}} = (x^{(\ell)} - \mu_k)^\top S^{-1}_{a,k}(x^{(\ell)} - \mu_k). \tag{3.5}$$

This prior is then fixed for subsequent iterations, and the optimization proceeds as outlined in Algorithm 2. For example, if the estimated reflectance at the first iteration is closest by distance to vegetation compared to concrete, water, or mud, a prior representing vegetation is used for computing the posterior until convergence. Although it is possible to update the prior with every iteration, this may prevent convergence. The parameters for the different priors are estimated with

field observations made at Santa Barbara (UCSB) and Hawaii, see [4] for details.

Algorithm 2 Simplified Optimal Spatial Inversion

procedure SPATIAL INVERSION(Radiances $\{y\}$, RTM terms, Spatial parameters $\{\nu, \rho, \sigma^2, (lat_x, long_x)\}$)
  for each block of n radiance value(s) **do**
    Initialize $x^{(0)}$ using an inexpensive guess
    Assign best prior $N(\mu_k, S_{a,k})$ at each pixel, see (3.5)
    Populate cross-correlations in prior covariance $\mathbf{S}_a$
    repeat
      Compute forward estimate $f(x^{(\ell)})$ for each pixel and concatenate
      Compute block error $y - f(x^{(\ell)})$, uncertainty S_ϵ, and Jacobian K
      Perform update step $x^{(\ell+1)} = \mu + \Delta_{LM}$ from (3.3)
    until convergence
    return Predicted reflectances $\{x_\star\}$
  end for
end procedure

3.3 Spatial retrievals

3.3.1 Naive spatial structure

Extending the original model to a spatial model requires working with multiple observations at once. Following the notation earlier, let $y = y_i \in \mathbb{R}^d$ denote a single measurement and $\mathbf{y} \in \mathbb{R}^{nd}$ denote a collection of n concatenated measurements. Likewise for the state vector, let $\mathbf{x} \in \mathbb{R}^{np}$ denote the set of state vectors to be retrieved with prior mean $\boldsymbol{\mu}$. In this notation, the spatial model takes the form

$$\mathbf{y}|\mathbf{x} \sim N_{nd}(f(\mathbf{x}), \mathbf{S}_\epsilon),$$

$$\mathbf{x} \sim N_{np}(\boldsymbol{\mu}, \tilde{\mathbf{S}}_a),$$

where $\mathbf{S}_\epsilon = I_n \otimes S_\epsilon$, $\tilde{\mathbf{S}}_a = I_n \otimes S_a$, $\boldsymbol{\mu} = E_n \otimes \mu$ all represent Kronecker product expansions of their non-spatial counterparts, and $E_n = (1, ..., 1)^\top$ is an n-dimensional column vector of ones. Note that $f(\mathbf{x}) = (f(x_1), f(x_2)...)^\top$ is applying the forward model to each corresponding state term.

Each location may have a different prior for the surface component as described in Section 3.2.3, but for clarity we drop the k index from $\mu_k, S_{a,k}$. As written, the model does not yet have spatial (cross-) correlations and $\tilde{\mathbf{S}}_a$ is block diagonal. We introduce these correlations with off-diagonal elements, illustrated as follows for an example with $n = 3$:

$$\tilde{\mathbf{S}}_a = \left[\begin{array}{cc|cc|cc} S_s & 0 & 0 & 0 & 0 & 0 \\ 0 & S_h & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & S_s & 0 & 0 & 0 \\ 0 & 0 & 0 & S_h & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & S_s & 0 \\ 0 & 0 & 0 & 0 & 0 & S_h \end{array}\right] \rightarrow \mathbf{S}_a = \left[\begin{array}{cc|cc|cc} S_s & 0 & 0 & 0 & 0 & 0 \\ 0 & S_h & 0 & D_{12} & 0 & D_{13} \\ \hline 0 & 0 & S_s & 0 & 0 & 0 \\ 0 & D_{12} & 0 & S_h & 0 & D_{23} \\ \hline 0 & 0 & 0 & 0 & S_s & 0 \\ 0 & D_{13} & 0 & D_{23} & 0 & S_h \end{array}\right].$$

We simplify the model by assuming the off diagonal blocks are all diagonal matrices, $D_{ij} = \text{diag}(C(x_{i,h_1}, x_{j,h_1}), C(x_{i,h_2}, x_{j,h_2}), ...)$ where $C(x_{i,h_1}, x_{j,h_1})$ denotes the covariance of the first atmospheric variable x_{h_1} with itself at locations i and j.

To be precise, let $\mathcal{I}$ denote the set of indices corresponding to the diagonal atmospheric components in the off-diagonal blocks of the prior cross covariance matrix, $\mathbf{S}_a$, so that in our $n = 3$ case,

$$(\mathbf{S}_a)_{\mathcal{I}} = \begin{bmatrix} \mathbf{S}_h & \mathbf{D}_{12} & \mathbf{D}_{13} \\ \mathbf{D}_{12} & \mathbf{S}_h & \mathbf{D}_{23} \\ \mathbf{D}_{13} & \mathbf{D}_{23} & \mathbf{S}_h \end{bmatrix}.$$

We can precisely specify the covariance matrix for a particular atmospheric variable. Denote the covariance for component k at locations i, j as $C(x_{i,h_k}, x_{j,h_k}) = C_{k,ij}$. Then the covariance matrix for the kth atmopsheric component across all locations, $\mathbf{x}_{h_k}$, is

$$(\mathbf{S}_a)_{\mathcal{I}_k} = \begin{bmatrix} C_{k,11} & C_{k,12} & C_{k,13} \\ C_{k,21} & C_{k,22} & C_{k,23} \\ C_{k,31} & C_{k,32} & C_{k,33} \end{bmatrix} = C(\mathbf{x}_{h_k}, \mathbf{x}_{h_k}).$$

In our situation, we only have two spatial atmospheric components, with $(\mathbf{S}_a)_{\mathcal{I}_1} = \mathbf{S}_{H_2O}$ and $(\mathbf{S}_a)_{\mathcal{I}_2} = \mathbf{S}_{AOD}$.

Concatenating the state and observed vectors and performing joint inference on the larger vector is a natural way to spatially extend a model, but may be inefficient for large samples, because we must invert the $nd \times nd$ prior covariance $\mathbf{S}_a$ as shown in (3.4). In the next section, we modify the specification to take advantage of the limited spatial structure.

3.3.2 Efficient implementation

As described in Section 3.3.1, our spatial structure is restrictive in that each spatially correlated component only (spatially) interacts with itself and does not have cross-correlation with any other component. This independence can be exploited for scalability by writing the gradient descent step in terms of the non-spatial surface component for one pixel and the set of all atmospheric components. As before, let $\mathbf{x}$ denote the concatenated version of the latent state vector. For the update step shown in Equation 3.2 with $\alpha \approx [\nabla_x^2 Q]^{-1}$ representing the constant matrix that results from the Levenberg-Marquardt approximation in (B.3), we have

$$\mathbf{x}^{(\ell+1)} = \mathbf{x}^{(\ell)} - \alpha \, \nabla \mathbf{Q}(\mathbf{x}^{(\ell)}) \tag{3.6}$$

with concatenated gradient term

$$\nabla \mathbf{Q}(\mathbf{x}) = \mathbf{S}_a^{-1}(\mathbf{x} - \boldsymbol{\mu}) + \mathbf{K}_{\mathbf{x}}^{\top} \mathbf{S}_{\epsilon}^{-1}(\mathbf{y} - f(\mathbf{x})),$$

where $\mathbf{K}_{\mathbf{x}}$ and $\mathbf{S}_{\epsilon}$ are block diagonal. Hence, for pixel $\mathbf{i}$,

$$(\nabla \mathbf{Q}(\mathbf{x}))_{\mathbf{i}} = (\mathbf{S}_a^{-1}(\mathbf{x} - \boldsymbol{\mu}))_{\mathbf{i}} + K_{x_{\mathbf{i}}}^{\top} \mathbf{S}_{\epsilon\mathbf{i}}^{-1}(y_{\mathbf{i}} - f_{\mathbf{i}}(x)),$$

where $(\mathbf{S}_a^{-1}(\mathbf{x}-\boldsymbol{\mu}))_\mathbf{i}$ denotes the subvector of components corresponding to the $\mathbf{i}$th state vector. A key observation is that this subvector only depends on the $\mathbf{i}$th surface components $S_{s,\mathbf{i}}^{-1}(x_{s,\mathbf{i}}-\mu_{s,\mathbf{i}})$, and atmospheric components $(\mathbf{S}_{H_2O}^{-1}(\mathbf{x}_{H_2O}-\boldsymbol{\mu}_{H_2O}))_i$ and $(\mathbf{S}_{AOD}^{-1}(\mathbf{x}_{AOD}-\boldsymbol{\mu}_{AOD}))_i$. In other words, retrieving the $\mathbf{i}$th state vector under spatial atmospheric effects does not cost much more than a non-spatial retrieval if the number of spatial components is small in comparison to the surface components. Furthermore, the block diagonal approach maintains some parallelizability of the original model. More sophisticated techniques are suggested in the conclusion.

3.4 Simulation study

In this section, we present results of a simulation study, in which individual retrievals are compared to joint spatial retrievals. The simulation procedure consists of three high-level steps:

1. Sample surface reflectance states of vegetation, the most common of the priors described in Section 3.2.3. The atmospheric states are correlated following the technique outlined in Section 3.3.1.

2. Simulate noisy satellite radiance measurements given the sampled surface state using the built-in methods and configuration of the ISOFIT code [4]; the noise model is described in Section 3.2.1

3. Invert the radiance measurements according to the implementation outlined in Section 3.3.2. Setting prior cross-pixel covariances to 0 results in individual retrievals as a special case.

The input pixels are given evenly spaced locations with gaps $1/n$ fixed according to the number of pixels n for the 1D case. The 2D case uses a regular grid of $\sqrt{n}\times\sqrt{n}$ pixels on the unit square. The spatial covariance function was taken to be Matérn with smoothness $\nu = 1.5$ and range parameter values of $\rho_{1D} = 3$ for the 1D case and $\rho_{2D} = 9$ for the 2D case, to account for the greater distance between points. For context, the Matérn covariance generalizes more common choices like the exponential covariance (Matérn $\nu = 0.5$) and squared exponential covariance (Matérn $\nu = \infty$); an intermediate value like $\nu = 1.5$ is more realistic according to our analysis (see Section B.2). The

variance parameters for the atmospheric components are 0.5 g^2cm^{-4} for water vapor and 0.2 for AOD.

Algorithm 3 Simulation Procedure: generate n pixels, compute correlated radiances, and invert. Repeat m_{iter} times.

procedure SIMULATE AND INVERT(m_{iter}, $\{(\mu_k, S_{a,k})\}$, $f(\cdot)$, n)
  for i in $1, 2, ..., m_{iter}$ **do**
    Concatenate n priors and fill cross correlations
    Sample a vector $\mathbf{x} = [x_1, ..., x_n]$ from the concatenated prior
    Simulate n noisy correlated radiance measurements $[y_1, ..., y_n] = f(\mathbf{x})$
    Invert $\{y_j\}_{j=1}^n$ individually or by block using Algorithm 2
  end for
end procedure

While the sampled data were taken from a distribution with a realistic mean and covariance, it is important to note that there was no attempt to measure the realism of the samples themselves. Over a few hundred wavelengths, it is possible that many small variations accumulate to yield a simulated reflectance that is unlike any real surface. Furthermore, a realized latent atmospheric state could correspond to extreme conditions that require unique configuration. As a result, both inversion methods were prone to failing at individual points, adding noise to all of the simulated results. For example, out of five pixels, the second pixel may fail to converge; the resulting total error for the method across the five pixels would be larger, as the retrieved surface reflectance values may diverge for particular wavelengths and atmospheric components concentrate on boundary values. Under a spatial model, this error is then spread to the nearby points. To remedy the issue, we truncated the realizations to realistic values of $[1.5, 2]$ g cm^{-2} for vapor and $[0.01, 0.1]$ for aerosol optical depth. Reducing the variance for the atmospheric components also helped avoid extreme realizations.

Figures 3.2, 3.3 and 3.4 illustrate the qualitative improvements that are possible with a spatial prior. While the independent inversions are at times closer to the truth, they may exhibit large oscillations that are avoided by the spatial retrievals due to the imposed correlation. In this way,

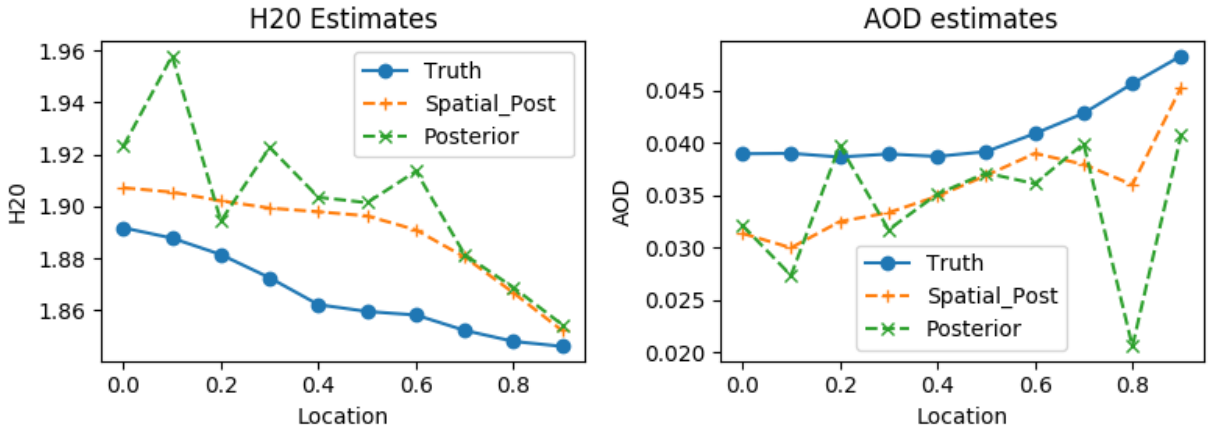


Figure 3.2: Inversions of simulated data showing the water vapor and aerosol optical depth estimates across 10 pixels in 1D. The retrieved fields are more realistic for spatial (Spatial_Post) than for individual retrievals (Posterior).

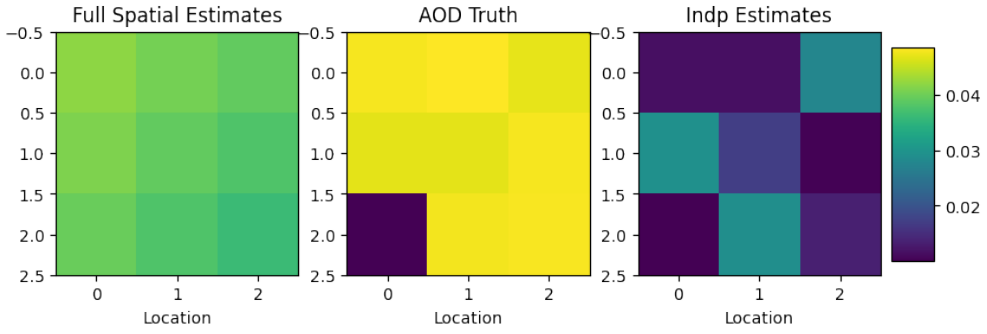


Figure 3.3: Inversions of simulated data showing the aerosol optical depth estimates across 9 pixels on a 3×3 grid. The spatial prior smooths the extremes that appeared in the random realizations.

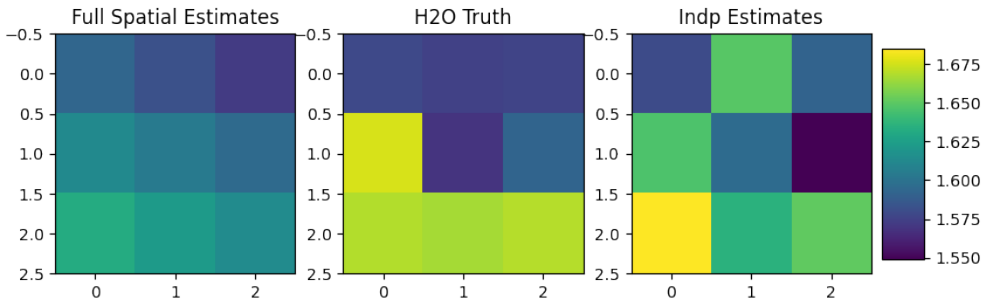


Figure 3.4: Inversions of simulated data showing the water vapor estimates across 9 pixels on a 3×3 grid. The spatial field better represents the truth.

the spatial inversions are more realistic.

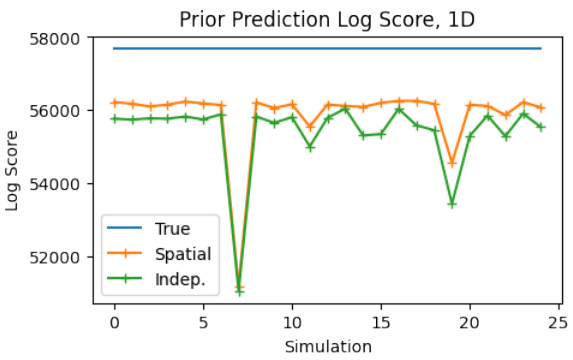
The mean square error is an unreliable indicator for inversion quality in the sense that highly variable components can inflate the MSE. Instead, we measure how closely the posterior mean reflects the true (prior) distribution with an ad-hoc “prior score," and we quantify the predictive performance with the log score. The prior score simply estimates the log likelihood of the posterior mean given the prior, $\log \mathcal{N}(\mathbf{x}_\star|\mu_a, \mathbf{S}_a)$. The log score [79] is a proper score [56, e.g.,] that reflects how likely the simulated true data $\mathbf{x}$ were under the estimated (Gaussian) predictive distribution, $\log \mathcal{N}(\mathbf{x}|\mathbf{x}_\star, \mathbf{S}_\star)$. It is important to note that the atmospheric components make up only two variables compared to the roughly 400 components of the reflectance per pixel inversion, so any improvements in log or prior scores are expected to be relatively small.

Figures 3.5a and 3.5b illustrate the prior score for the $m_{iter} = 25$ simulated realizations each of 1D and 2D pixel arrays. In most cases, the posterior is closer to the prior for the spatial case, resulting in a better prior score and implying that the spatial model better represents the data, as expected. For the 2D case, the difference is smaller, because of the greater inherent variability of a 2D field and the larger maximum distances between points.

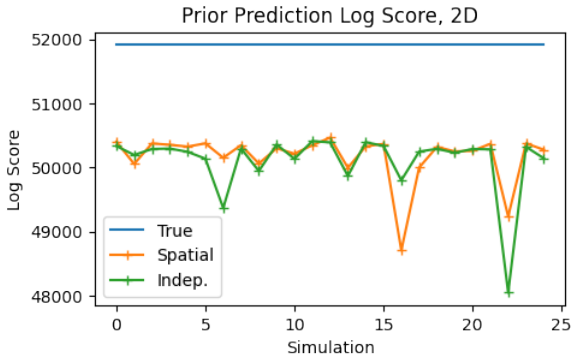
Figure 3.6 illustrates how the spatial inversion usually has better predictive performance.

3.5 Application

We apply the spatial inversion to three sets of remotely sensed data from AVIRIS-NG. The current implementation of the inversion software, ISOFIT, produces pixelwise-independent estimates



(a) Prior score results for a 1D array of 10 pixels.



(b) Prior score results for a 2D grid of 9 pixels.

Figure 3.5: Prior score plots for 25 simulated realizations. The posterior estimates for the spatial model are usually closer to their priors than the independent models. The effect is weaker for the 2D case, suggesting that the improvement tends to be most pronounced with highly correlated data.

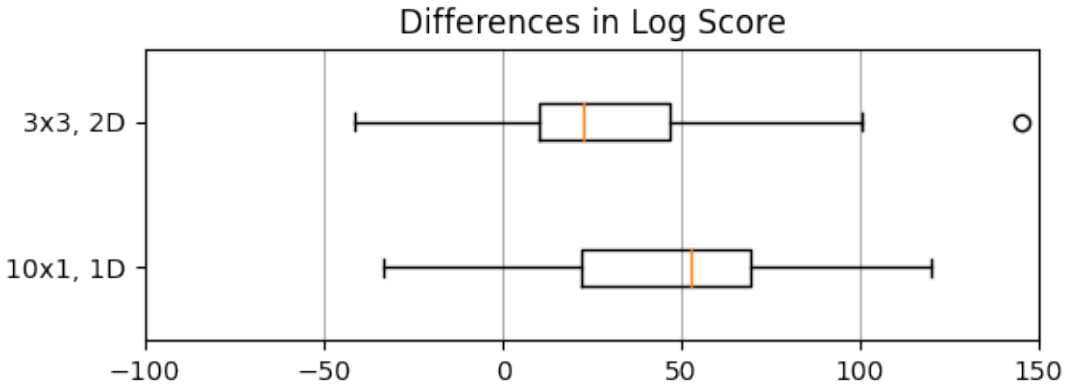


Figure 3.6: Box plots show the difference in log score between the spatial and independent models across 25 simulations. The $(25\%, 50\%, 75\%)$ quantile values for the 1D and 2D cases are $(21.8, 52.8, 69.1)$ and $(10.1, 22.4, 46.8)$, respectively.

of surface reflectance and the two atmospheric components of water column vapor and aerosol optical thickness or depth (AOD), see Section 3.2.2. Measurements were taken by plane from 5 to 10 km altitude and were orthocorrected for plane movement.

Before applying the methodology to real data, we estimate the covariance parameters of the spatial model with a field of water vapor measurements estimated by the independent inversion procedure on an unrelated data set in India. Our chosen parameter values for both water vapor and aerosols were: a range $\rho = 750m$, smoothness $\nu = 1.5$, a nugget effect of 0.001 and variance $\sigma^2 = 1$. The procedure and justification for this choice are presented in Appendix B.2.

The first data set we consider is a validation measurement taken at Ivanpah Playa, CA on March 28, 2017 at about 5:30 PM. Ideally the data set would consist of AVIRIS-NG observations along with multiple simultaneous measurements of in situ aerosols and water vapor over the region, which would allow for validation of the method as in [4]. Since a data set like this does not currently exist, the Ivanpah data set with just a single, area-wide measurement for the aerosols and vapor is the best available alternative. The weather conditions for the measurement are extremely uniform and clear, so we perform a spatial inversion to determine if the noise in the atmospheric components is smoothed.

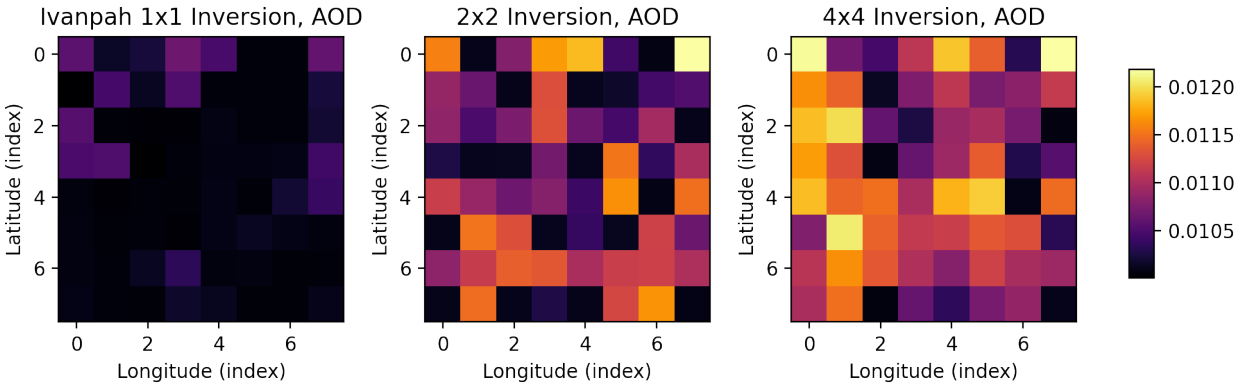


Figure 3.7: The aerosol optical depth prediction for validation data at Ivanpah. The predictions are effectively identical, but the spatial retrievals are closer to the in situ measurement of 0.043.

The results of the validation show that the atmospheric components can have slightly less bias under the spatial model, but the effect is practically insignificant. The in situ measured aerosol optical thickness and water vapor are roughly 0.043 and 0.88, respectively. The estimates for aerosols shown in Figure 3.7 vary from 0.01 to 0.012, which underestimates the in-situ measurement of 0.043, but in practice the difference is negligible as AOD values up to 0.05 correspond to extremely clear skies. The water vapor measurements are nearly identical and uniformly valued at 0.67 for all methods, which also underestimates the in situ measurements of 0.88. Such differences of 0.2 g $cm^{-}2$ are not unrealistic, since the in situ measurement carries its own uncertainty and the optical absorption path of the two instruments is different. Together, this validation study confirms that a spatial model does no harm and can help lower the overall error of the aerosol estimates, but the spatial error for such homogeneous scenes is negligible.

The next data set we explored was measured on June 25, 2014 at roughly 7:30 PM local time over Cuprite Hills, Nevada. Here we have a swath of 50x150 pixels and perform individual, 1x5 pixel inversions, and 2x2 pixel inversions, with the spatial inversions using the same Matern parameters (1.5, 0.75) as the previous data set. We find very little difference in the surface reflectance across pixels shown in Figure 3.8. There is a mild scaling effect that occurs with the spatial versions, which we attribute to the different results for the atmospheric components, but the shape is consistently characteristic of soil with minerals. The results for atmospheric water vapor shown in Figure 3.10 show that the spatial models provide a smoothing effect that reduces the noisy estimates of the independent inversions. The aerosol optical thickness in Figure 3.9 has a similar story, where the spatial values tend to be lower and smoother than the independent inversion, which has stronger gradients between pixels. The fourth subfigure of Figure 3.9 shows reflectance for an arbitrary wavelength and suggests that the aerosols detected by all methods are influenced by the land reflectance, with the independent inversions more strongly influenced compared to the spatial methods.

Our last data set was collected over Yolo, CA on the outskirts of Sacramento, CA on September 7, 2020 at about 7pm. The conditions for this data set were smoky: wildfires had increased the

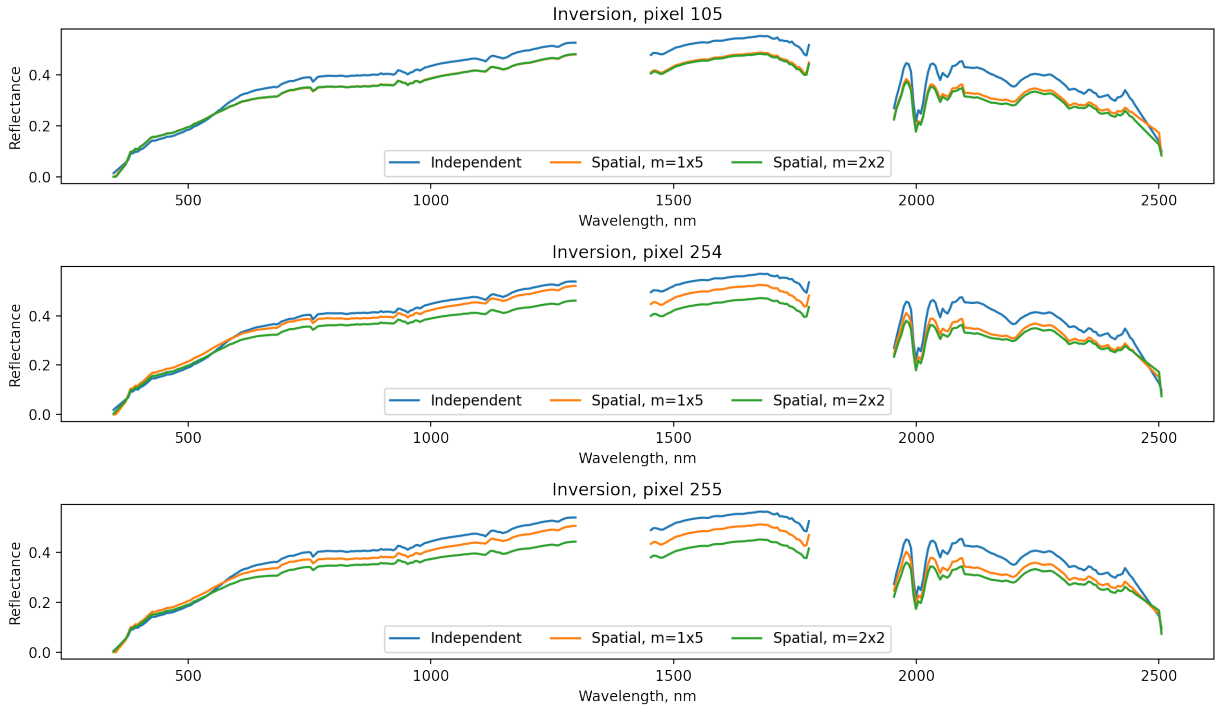


Figure 3.8: The surface reflectance profiles are nearly identical for the Cuprite data, with scaling changes due to the estimation of atmospheric parameters. This suggests that independent inversions may be overestimating reflectance. Pixels 105, 254, and 255 are adjacent and the reflectance can be interpreted as a percent, so at a particular wavelength a reflectance of 0.4 means 40% of the incoming radiant energy is reflected.

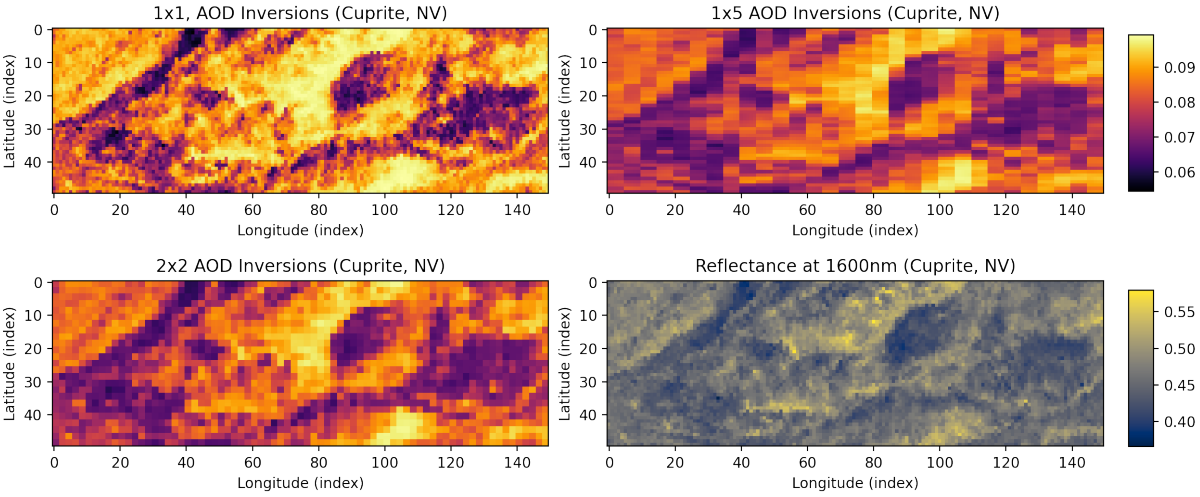


Figure 3.9: For the Cuprite dataset, the aerosol optical depth prediction is susceptible to the surface state prediction (bottom right), but smoothing with a spatial prior decreases the noise.

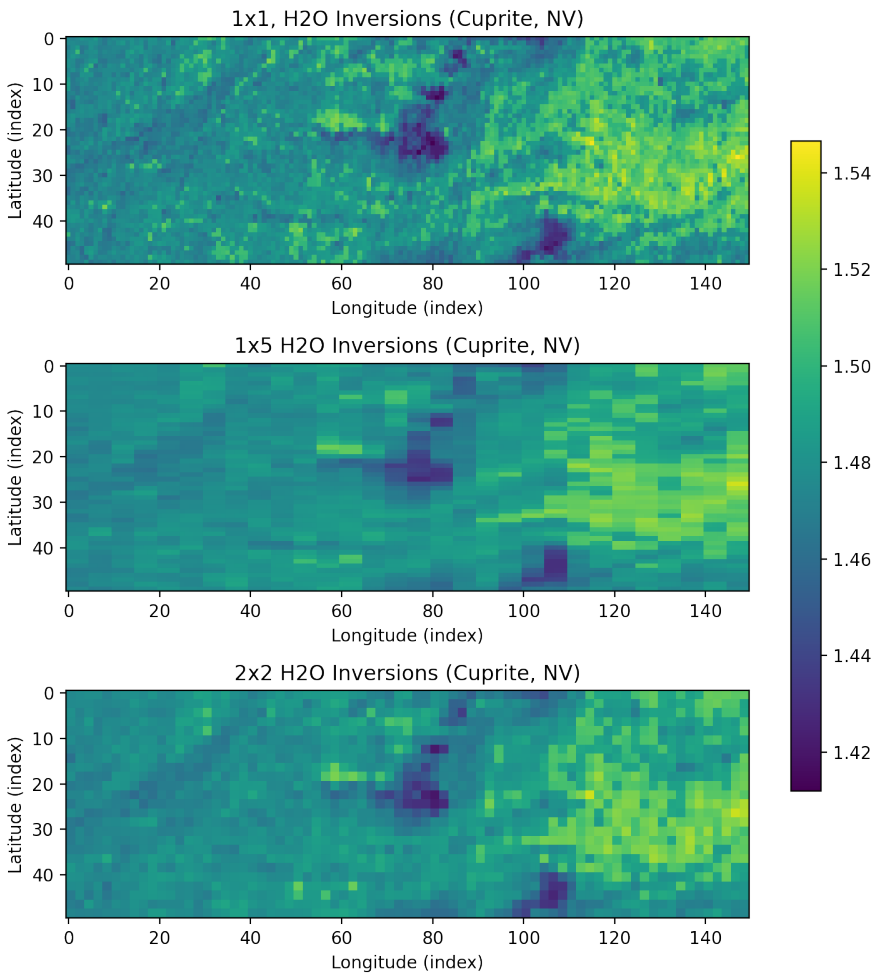


Figure 3.10: The water vapor estimates are noticeably smoother under the spatial models. The predicted fields are qualitatively more realistic and are a principled alternative to post-hoc smoothing.

amount of aerosols in the atmosphere and varying amounts of smoke are visible in the color images of the scene. We invert a coarse grid over the entire scene to see if the recovered aerosol states can capture the smoothly varying field suggested by the imagery. The full swath is about 2500 x 500 pixels, so we subsample every 25th pixel with a buffer from the edges to get 94x16 inversions.

Figure 3.11 shows a comparison of the independent and a 2x2 inversion. While the H_2O predictions were nearly identical, the aerosol field was significantly smoothed. There are a few areas in the spatial model that appear to be outliers but may be explained as the spatial model spreading the effect of large individual pixel values for the aerosols. It is expected that inverting a larger col-

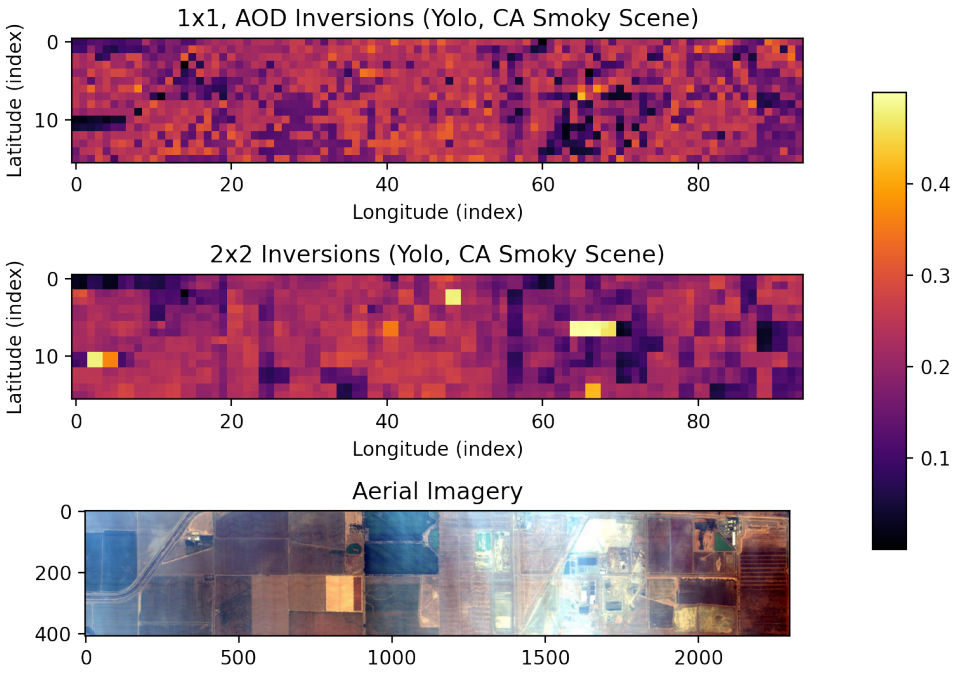


Figure 3.11: A retrieved aerosol field under a spatial model is smoother than the independent retrievals and spreads out large estimates.

lection of pixels simultaneously (for example, 10x10) will result in the large values being spread out even more and higher overall estimates for the aerosol field. Combined with the results of the validation data at Ivanpah, the spatial model may counteract or provide lower bias for atmospheric components compared to independent inversions.

3.6 Conclusion

In this work we showed how to account for spatial correlations in retrievals of surface reflectance from imaging spectroscopic measurements. The standard methodology inverts a single radiance measurement to estimate surface reflectance and atmospheric states of aerosols and water vapor. By directly modeling the physical correlations of the atmospheric components, we can invert multiple measurements simultaneously and borrow strength from nearby locations to get more robust predictions of the non-spatial reflectances. In contrast, kriging or post-processing the fields to create smoothness does not take into account the dependencies between variables induced by the nonlinear model and would result in inaccurate fields.

We illustrated the mathematical details and addressed the basic computation challenges that arise with the introduction of cross-correlation with a Gaussian process prior. The block independent implementation we chose is both simple and allows for straightforward parallelization, but can exhibit a computational complexity that is cubic in cardinality of the block. Our simulations showed that a spatial radiative transfer model offers a better log score when compared to the non-spatial version. With real satellite data, we demonstrated how the spatial model can offer qualitatively improved retrievals with lower perceived error in the the atmospheric components. However, we note that the estimates of surface reflectance were not significantly affected.

Although we do not have spatially varying situ measurements to compute accuracy scores for real data, we showed that the spatial model does provide additional smoothing to the atmospheric components, resulting in more realistic predictions for the atmospheric state across space. We also noticed a consistent trend in which the spatial models show slightly less bias in the atmospheric components.

From a development point of view, a next step is to apply one of the many spatial approximations to allow for efficient, simultaneous inversion of larger data sets. Inducing sparsity in precision matrices [29, 30, 3] or low-rank approaches [18, 19, 21] stand out as the best options. From an application point of view, essentially any inversion that involves smoothly varying components can be extended with this methodology. One special case is exoplanet surface analysis, in which the exoplanet surface is expected to have some type of atmosphere and even a very simple atmospheric model may lead to improved retrievals. Alternately, a spatial model for the local atmosphere offers telluric corrections on upward-looking observation time series of exoplanet spectra from a ground-based spectrometer. The "surface" of interest may be a star, and the local atmosphere can be modeled as a 1-D Gaussian Markov system where belief propagation gives a tractable exact solution. Correlations over the temporal domain can be included as well if there are multiple reflectances measured over time.

In addition to using approximations for the spatial prior, further speed-ups might be obtained by GP emulation of the forward model after dimension reduction via active subspace on the latent

state and functional PCA on the observations. The data model may be improved by considering the radiance measurement as a count, implying a Poisson or generalized linear model where the variance is equal to the mean, rather than a Gaussian model. An alternative is to assume a log Gaussian model for the observations, which would avoid some of the additional computational burden of a Poisson model.

4. FREQUENTIST COVERAGE FOR TRUNCATED KERNEL RIDGE REGRESSION

4.1 Introduction

Gaussian processes as priors on a space of functions are widely used for nonparametric models in a variety of fields such as spatial statistics [80, 81], machine learning [2], and emulation [82]. The GP prior is often used in a Bayesian framework, but it also appears implicitly in kernel ridge regression [83].

There is a strong theoretical foundation for the regularized GP or Kernel Ridge regression (KRR) problem. Frequentist minimax optimality [84] and matching Bayesian posterior contraction rates [85] justify the use of GPs and provide guidance for the regularization weight or prior distribution. As there is no Bernstein von Mises result for nonparametric regression, frequentist coverage is a natural surrogate and shown for inverse problems [86] and Gaussian white noise models [87]. Recent work [6] establishes coverage for GP regression under the supremum norm while also showing minimax optimality under this norm, concluding an avenue of research for un-approximated GP's.

We pick up one of the threads of [6] and seek to extend the coverage and minimax optimality results when the covariance has been approximated. It is well known that a GP model has complexity that is cubic in the sample size n due to the inversion of the $n \times n$ covariance matrix. There is a vast literature on covariance approximation but the most popular and general techniques are low rank and sparsity inducing methods. Sparsity inducing methods include (among spatial statistics) Markov random fields [28, 29], nearest neighbor approximations [31, 88, 89], and multiresolution filters [90, 91]. Low rank methods are often the simplest methods and will therefore be the domain of our theory, despite their known limitations [92].

In this work we focus on a truncated kernel approximation. This specific problem is relatively simple and does not seem to have been addressed, but there is a notable line of work for the case of data subsampling. There are bounds for the error between the random rank-p subset and the

best rank-p approximations [93, 94], including adaptive approaches [95, 96] that can guarantee good performance. Minimax optimality is shown in [97] for the subsampling case with a rank on the order of the effective dimension with constants that depend on kernel complexity. We recover results that match with the existing literature, namely that effective dimension provides a sufficient rank to maintain minimax optimality. We also show that frequentist coverage properties from the un-approximated KRR estimator are maintained.

This chapter is organized as follows. We first review relevant background material such as Hilbert spaces and the equivalence of kernel ridge regression and Bayesian estimation. We then provide a series of results extending the theory of KRR supnorm optimality and coverage in [6] and make some concluding remarks. Proofs are mostly left to the appendix.

4.2 Background

4.2.1 Review of RKHS

We briefly review properties of and set notation for reproducing kernel Hilbert spaces that will be used later. A RKHS $\mathcal{H}$ is a vector space of functions with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and support $\mathcal{X}$. There is an evaluation operator L_t in the dual space with $t \in \mathcal{X}$ that can be expressed as an element of the space $K_t \in \mathcal{H}$, such that

$$\langle f, K_t \rangle_{\mathcal{H}} = f(t)$$

The reproducing element $K_t(\cdot) = K(t, \cdot)$ is a kernel and is positive definite (pd) if $\sum_{i,j=1}^{\infty} a_i K(t_i, t_j) a_j > 0$ for any real a_i, a_j. Mercers theorem states that a positive definite kernel has an eigendecomposition,

$$K(t, t') = \sum_{i=1}^{\infty} u_i \phi_i(t) \phi_i(t')$$

satisfying

$$\int_{\mathcal{X}} K(t, t') \phi_i(t') dt = u_i \phi_i(t).$$

A necessary condition is that the kernel is square integrable over both indices, which leads to $\sum_{i=1}^{\infty} u_i^2 < \infty$, or $(u_i) \in \ell^2$. Using the Mercer eigendecomposition and reproducing property of

K_t, we can compute the inner product explicitly for the elements of the RKHS,

$$\langle f, g \rangle_{\mathcal{H}} = \left\langle \sum_{i=1}^{\infty} f_i \phi_i, \sum_{i=1}^{\infty} g_i \phi_i \right\rangle_{\mathcal{H}} = \sum_{i=1}^{\infty} \frac{f_i g_i}{u_i}$$

This induces a norm on the space which can then be used to define the space, namely

$$\mathcal{H} = \left\{ f : \|f\|_{\mathcal{H}} = \sum_{i=1}^{\infty} \frac{f_i^2}{u_i} < \infty \right\}$$

4.2.2 Notation for kernel regressions

Since we will build on the work of [6], we use their exact notation and summarize the relevant components. The data is represented as $\mathbb{D}_n = (Y_i, X_i)$ where $Y_i \in \mathbb{R}$ and $X_i \in \mathcal{X} \subset \mathbb{R}$ and $i = 1, ..., n$. It is assumed that there is a nonlinear relationship,

$$Y_i = f(X_i) + w_i, \quad w_i \sim N(0, \sigma^2),$$

where the true function is denoted f^*. We take a Gaussian process prior for the unknown function, $f \sim GP(0, \sigma^2(n\lambda)^{-1}K)$, where K is the reproducing kernel for the space and λ represents a penalization weight. The posterior distribution of the Gaussian process regression (GPR) given data is also a GP, denoted $f|\mathbb{D}_n \sim GP(\hat{f}_n, \widetilde{C}_n^B)$ where the mean and variance are written as

$$\hat{f}_n(x) = K(x, \mathbb{X})[n\lambda I + K(\mathbb{X}, \mathbb{X})]^{-1}\mathbb{Y}$$

$$\widetilde{C}_n^B(x, x') = \sigma^2(n\lambda)^{-1}\{K(x, x') - K(x, \mathbb{X})[I + K(\mathbb{X}, \mathbb{X})]^{-1}K(\mathbb{X}, x')\}. \tag{4.1}$$

This is equivalent to the optimal solution to the kernel ridge regression (KRR) [2]. KRR minimizes the squared error of a non-parametric regression over a space of functions from a Hilbert space $\mathcal{H}$, adding a penalty on the Hilbert space norm of the function to penalize "large" function

values:

$$\hat{f}_{n,\lambda} = \arg\min_{f \in \mathcal{H}} \sum_{i=1}^{n} [Y_i - f(X_i)]^2 + \lambda ||f||_{\mathcal{H}}^2 = \arg\min_{f \in \mathcal{H}} \ell_{n,\lambda}(f).$$

Solving for the KRR solution is tractable using the representer theorem [98] and is equal to the posterior as found with GPR.

A key breakthrough of [6] is to represent the variance of the KRR solution as the bias of the solution to another regression problem. This is accomplished by rearranging equation 4.1,

$$\sigma^{-2}(n\lambda)\widetilde{C}_n^B(x, x') = K(x, x') - \hat{K}_x(x')$$

and expressing the quadratic form of $\hat{K}_x(x')$ as the solution to a KRR

$$\hat{K}_x(x') = \arg\min_{g \in \mathcal{H}} \left[\frac{1}{n} \sum_{i=1}^{n} (Z_i^x - g(X_i))^2 + \lambda \|g\|_{\mathcal{H}} \right]. \tag{4.2}$$

The model in this case is noiseless, with observations modeled as $Z_i^x = K(x, X_i)$.

4.2.3 Equivalent kernels

The posterior or KRR solution is difficult to work with due to the matrix inversion in both the mean and variance terms. The equivalent kernel trick [2, Ch 7.1] is used by [6] to work around the inversion and prove frequentist coverage. They define the equivalent kernel for a new Hilbert space related to the KRR and GPR problems using the inner product

$$\langle f, g \rangle_\lambda = \langle f, g \rangle_{L^2(\mathcal{X})} + \lambda \langle f, g \rangle_{\mathcal{H}}$$

Plugging in the formulas for the norms over the eigenbasis from Mercers theorem allows for a simple representation:

$$\langle f, g \rangle_\lambda = \sum_{j=1}^{\infty} f_j g_j + \lambda \sum_{j=1}^{\infty} \frac{f_j g_j}{u_j} = \sum_{j=1}^{\infty} \frac{f_j g_j}{v_j}, \quad v_j = \frac{u_j}{\lambda + u_j}$$

This matches the formula for an RKHS norm and implies that v_i are the eigenbasis for the equivalent kernel,

$$\widetilde{K}(t,t') = \sum_{j=1}^{\infty} v_j \phi_j(t)\phi_j(t')$$

The equivalent kernel makes it possible to express the KRR problem so that the solution does not contain inverted matrices, but involves the definition of two additional operators: the convolution or population level solution,

$$F_\lambda f(t) = \int f(s)\widetilde{K}(s,t)ds$$

and the complement or bias,

$$P_\lambda f(t) = (Id - F_\lambda)f(t).$$

It is shown in [6] using the eigendecomposition that these operators have the property of recovering the two original norms from the equivalent norm:

$$\langle f, F_\lambda g\rangle_\lambda = \langle f, g\rangle_{L^2(\mathcal{X})}, \quad \langle f, P_\lambda g\rangle_\lambda = \lambda \langle f, g\rangle_{\mathcal{H}}$$

The KRR objective can then be rewritten as

$$\ell_{n,\lambda}(f) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \left\langle f, \widetilde{K}_{X_i}\right\rangle_\lambda)^2 + \langle f, P_\lambda f\rangle_\lambda,$$

The score function, or derivative, of the objective can be derived using a Frechet derivative, since the objective maps an infinite dimensional object in the Hilbert space to a real number.

4.2.4 Standing assumptions

The main results of [6] provide supremum norm error bounds for the KRR problem assuming true functions f^* come from Sobolev or Holder spaces,

$$\Theta_S^\alpha(B) = \{f = \sum_{j=1}^{\infty} f_j\phi_j \in L^2(\chi) : \sum_{j=1}^{\infty} j^{2\alpha} f_j^2 \leq B^2\},$$

$$\Theta_H^\alpha(B) = \{f = \sum_{j=1}^{\infty} f_j \phi_j \in L^2(\chi) : \sum_{j=1}^{\infty} j^\alpha |f_j| \leq B\}.$$

The basis functions ϕ_i used across formulas and function spaces so far can be taken to be the Fourier basis, $\phi_{2j}(s) = \cos(\pi j s)$ and $\phi_{2j-1}(s) = \sin(\pi j s)$. This basis satisfies the two additional assumptions made for all subsequent results:

Assumption (B): The eigenfunctions of the kernel $\{\phi_j\}_{j=1}^{\infty}$ are bounded and Lipshitz bounded. There exist C_ϕ and L_ϕ such that $|\phi_j(t)| \leq C_\phi$ and $|\phi_j(t) - \phi_j(s)| \leq L_\phi j|t-s|$ for all $j \geq 1$ and $s, t \in \mathcal{X}$.

The second assumption uses the notation $a \asymp b$, which means that $a \lesssim b$ and $b \lesssim a$, where $\lesssim$ means the inequalities hold up to some constant multiple.

Assumption (E): The kernel K eigenvalues have a decay rate given by $u_j \asymp j^{-2\alpha}$.

These assumptions on the kernel lead to bounds on the trace of the equivalent kernel, since $u_j \asymp j^{-2\alpha}$ implies that $v_j = u_j/(u_j + \lambda) \asymp 1/(1 + \lambda j^{2\alpha})$ and we can bound the sum by its integral, or trace:

$$tr(\widetilde{K}) \asymp \sum 1/(1 + \lambda j^{2\alpha}) \asymp \lambda^{-1/2\alpha}. \tag{4.3}$$

This is sometimes referred to as the effective dimension. The same bound holds for the squared trace and for future convenience we define $\lambda = h^{2\alpha}$. The bound on the trace is an important quantity for this work, as our results show that this trace term can be used to determine the truncation limit p that maintains estimator optimality.

4.3 Results for truncated decomposition

The goal of this section is to demonstrate that kernel truncation maintains minimax optimality under supremum norm for the KRR estimator. Since optimality is measured in terms of risk, and risk is commonly decomposed as bias and variance term, we proceed by first determining general error rates, then quantifying the bias and variance terms, and then determining pointwise convergence and coverage.

4.3.1 Kernel truncation approximation

Kernel truncation refers to the truncation of the Mercer decomposition to obtain a finite dimensional representation,

$$K_p(x, x') = \sum_{j=1}^{p} u_j \phi_j(x)\phi_j(x') \approx \sum_{j=1}^{\infty} u_j \phi_j(x)\phi_j(x'). \tag{4.4}$$

To recover the results of [6] under an approximated kernel, we start with computing the equivalent kernel and associated operators. Given the full rank equivalent kernel $\widetilde{K}(s,t) = \sum_{j=1}^{\infty} \nu_j \phi_j(s)\phi_j(t)$, we denote the finite rank equivalent kernel

$$\widetilde{K}_p(s,t) = \sum_{j=1}^{p} \nu_j \phi_j(s)\phi_j(t).$$

It is easy to see that the convolution operator $F_\lambda f$ takes a reduced rank form of $F_{\lambda,p} f = \sum_{j=1}^{p} \nu_j f_j \phi_j$, due to the orthonormality of the eigenfunction basis. The remainder operator is

$$P_{\lambda,p} f = (Id - F_{\lambda,p}) f = \sum_{j=1}^{p} (1-\nu_i) f_j \phi_j + \sum_{j=p+1}^{\infty} f_j \phi_j, \tag{4.5}$$

The second term contains the tail that has been truncated from the equivalent kernel. We now restate the theorems of [6], making small changes for the truncated case.

4.3.2 Error bounds

We assume for all theorems that the the kernels satisfy the assumptions of Section 4.2.4.

Claim 1 (Sup-norm bounds for the Truncated KRR estimator)**.** *Define coefficients* $\widetilde{A}_n$ *and* γ_n*,*

$$\widetilde{A}_n = 2h^{-\alpha}\big(\|P_{\lambda,p} f^*\| + \sigma\sqrt{\frac{\log n}{nh}}(1 + h(\log n)^2)\big)$$

$$\gamma_n = \left[1 + \sqrt{\frac{\log n}{nh}} + \widetilde{A}_n^{1/(2\alpha)} + \widetilde{A}_n^{1/(\alpha)}\left(\frac{1}{\sqrt{nh}} + n^{1/(2\alpha)-1} h^{-1/(2\alpha)}\right)\right]\sqrt{\frac{\log n}{nh}}.$$

Now if there is a constant c_K that only depends on the kernel and $\gamma_n < c_K$, then with probability at least $1 - n^{-10}$ with respect to the random terms (X_i, w_i) and $p \asymp \lambda^{-1/(2\alpha)}$

$$||\hat{f}^{(p)}_{n,\lambda} - f^\star||_\infty \leq (1 + C\gamma_n)||P_{\lambda,p} f^\star||_\infty + C\sigma\sqrt{\frac{\log n}{nh}}. \tag{4.6}$$

This same probability holds for the bound on a higher-order expansion,

$$||\hat{f}^{(p)}_{n,\lambda} - F_{\lambda,p} f^\star - \frac{1}{n}\sum_i^n w_i \widetilde{K}_{p,X_i}||_\infty \leq C'\gamma_n \left((1 + C\gamma_n)||P_{\lambda,p} f^\star||_\infty + C\sigma\sqrt{\frac{\log n}{nh}} \right). \tag{4.7}$$

The constants C and C' are independent of (n, h, λ, σ).

Proof. This bound is essentially a trivial consequence of theorem 2.1 of [6], with the only change being the kernel. Since the bounds contain the generic remainder operator term $P_{\lambda,p}$, the tail component $\sum_{i=p+1}^{\infty} f_i \phi_i$ from the kernel approximation in equation 4.5 is accounted for and does not affect the proof. Therefore, we only need to check that the truncated kernel satisfies the assumptions of the original theorem to have the original proof go through. But it is immediate that the conditions **(B)** and **(E)** are still valid with truncation, namely $u_j = 0 \leq j^{-2\alpha}$ for $j > p$.

The higher order results hold by a similar argument because the original result relies on a Bernstein type inequality that makes no restriction on the kernel $\widetilde{K}$. The term $||P_\lambda f^\star||$ once again contains the secondary error introduced by approximating kernel. □

The remainder operator $P_{\lambda,p} f^*$ is a general term that can be made more precise with assumptions on the function space of f^*. We show next that minimax optimality over these function classes is preserved under appropriate kernel truncation. The rates given are known in the literature to be minimax optimal.

Theorem 4.3.1. *The following results hold with $p \asymp \lambda^{-1/(2\alpha)}$ and probability at least $1 - n^{-10}$ with respect to the randomness in (X_i, w_i).*

1. *For* $f^* \in \Theta_S^\alpha(B)$, $\alpha > (3+\sqrt{57})/12$, *and* $h = \left(\frac{B^2 n}{\sigma^2 \log n}\right)^{-1/(2\alpha)}$,

$$\|\hat{f}_{n,\lambda}^{(p)} - f^*\|_\infty \lesssim B^{1/(2\alpha)} \left(\frac{\sigma^2 \log n}{n}\right)^{(\alpha-1/2)/2\alpha)}$$

2. *For* $f^* \in \Theta_H^\alpha(B)$, $\alpha > 1/\sqrt{2}$, *and* $h = \left(\frac{B^2 n}{\sigma^2 \log n}\right)^{-1/(2\alpha+1)}$,

$$\|\hat{f}_{n,\lambda}^{(p)} - f^*\|_\infty \lesssim B^{1/(2\alpha+1)} \left(\frac{\sigma^2 \log n}{n}\right)^{\alpha/(2\alpha+1)}$$

Proof in appendix C.1. We observe that the same truncation level p is used for both function classes. When p grows faster than $\lambda^{-1/(2\alpha)}$, the results still hold because the tail component decreases with larger p. In contrast, when p grows slowly, the tail term in the bias dominates the error and we do not recover an optimal rate.

Theorem 4.3.2. *For the case when* $p \asymp \lambda^{-1/(2\alpha_o)}$ *with* α *satisfying the requirements of theorem 4.3.1 and* $\alpha_o > \alpha$, *the excess truncation leads to the following suboptimal rates.*

1. *For* $f^* \in \Theta_S^\alpha(B)$ *and* $h = \left(\frac{B^2 n}{\sigma^2 \log n}\right)^{-1/(2\alpha)}$, *the error is sub-optimal with rate*

$$\|\hat{f}_{n,\lambda}^{(p)} - f^*\|_\infty = \mathcal{O}\left(\frac{\log n}{n}\right)^{\frac{\alpha}{2\alpha_o} - \frac{1}{4\alpha}}$$

2. *For* $f^* \in \Theta_H^\alpha(B)$ *and optimal* $h = \left(\frac{B^2 n}{\sigma^2 \log n}\right)^{-1/(2\alpha+1)}$, *the error is sub-optimal with rate*

$$\|\hat{f}_{n,\lambda}^{(p)} - f^*\|_\infty = \mathcal{O}\left(\frac{\log n}{n}\right)^{\frac{\alpha^2}{\alpha_o(2\alpha+1)}}$$

Proof in appendix C.2. Given the error bounds above, the next step towards optimality in terms of minimax risk is to compute variance bounds.

4.3.3 Posterior variance

We first express that the posterior variance for the truncated case,

$$\widetilde{C}^B_{n,p}(x,x') = \sigma^2(n\lambda)^{-1}\big[K_p(x,x') - K_p(x,\mathbb{X})[K_p(\mathbb{X},\mathbb{X}) + n\lambda I_n]^{-1}K_p(\mathbb{X},x')\big].$$

The quadratic term can be expressed as the solution to a noiseless KRR just as in equation 4.2:

$$\widehat{K}_{p,x} = \underset{g\in H}{\arg\min}\left[\frac{1}{n}\sum_{i=1}^{n}(Z_i^x - g(X_i))^2 + \lambda||g||_H^2\right]$$

The observations Z_i^x correspond to the truncated kernel $K_{p,x}(X_i)$.

We can apply claim 1 to the case of a noiseless KRR simply by setting $\sigma = 0$. Where there was previously a true function $f^*(\cdot)$, there is now a kernel term $K_p(x,\cdot)$. The error bound from equation 4.6 now has supremum norm bound $(1 + C\gamma_n)||P_\lambda K_{p,x}||$ (since $\sigma^2 = 0$). We expand the norm term $P_\lambda K_{p,x}$ using the truncated kernel (the tail is excluded because the KRR problem uses the truncated kernel):

$$P_{\lambda,p}K_{p,x}(\cdot) = \sum_{i=1}^{p}(1-\nu_i)u_i\phi_i(x)\phi(\cdot) = \lambda\widetilde{K}_{p,x}(\cdot), \tag{4.8}$$

since $u_i(1 - v_i) = u_i\lambda/(u_i + \lambda) = \lambda v_i$. Next we apply claim 1 to the KRR estimator for the covariance to get a variance bound for the truncated kernel posterior variance,

$$||\sigma^{-2}n\lambda\widetilde{C}^B_n(x,x')||_\infty = ||K_{p,x}(x') - \widehat{K}_{p,x}(x')||_\infty \le C||P_\lambda K_{p,x}||_\infty \tag{4.9}$$

Using the higher order result of equation 4.7 and plugging in $\sigma = 0$, we recover

$$\|\hat{f}^{(p)}_{n,\lambda} - F_{\lambda,p}f^\star\| = \|f^* - f^{(p)}_{n,\lambda} - P_{\lambda,p}f^\star\|$$

Plugging in the kernel observations $K_p(x,\cdot)$ in place of f^* allows us to write

$$||\sigma^{-2}n\lambda\widetilde{C}^B_{n,p}(x,\cdot) - P_{\lambda,p}K_{p,x}||_\infty \leq C\gamma_n||P_\lambda K_x||_\infty$$

and subsequently

$$||\sigma^{-2}n\widetilde{C}^B_{n,p}(x,\cdot) - \widetilde{K}_{p,x}||_\infty \leq C\gamma_n||\widetilde{K}_{p,x}||_\infty$$

This can be used to approximate the posterior truncated variance with $\widetilde{K}_{p,x}$. We take one more step, denoting $\hat{C}^B_{n,p} = \sigma^2 h\widetilde{K}_{p,x}$. By the previous result, this is close to $\sigma^2 h\sigma^{-2}n\widetilde{C}^B_{n,p} = nh\widetilde{C}^B_{n,p}$ so we get

$$\sup_{x,x'}|nh\widetilde{C}^B_{n,p}(x,x') - \hat{C}^B_{n,p}(x,x')| \leq C\sigma^2 h\gamma_n||\widetilde{K}_{p,x}||_\infty \lesssim \gamma_n \tag{4.10}$$

The bound comes from the trace norm assumptions described in equation 4.3.

Now that we have a bound for the variance, we can study the risk.

4.3.4 Risk bounds

Here we show that the pointwise error of the posterior $f|\mathbb{D}_n \sim GP(\hat{f}_{n,p}, \widetilde{C}^B_{n,p})$ with truncated kernel converges at the same rate as for the original kernel. We again assume that the truncation scales with the sample size, $p \asymp h^{-1} = \lambda^{-1/2\alpha}$.

Theorem 4.3.3. *Let $\lambda^{1/(2\alpha)} = h$ and $p \asymp \lambda^{-1/(2\alpha)}$. For the case $f^* \in \Theta^\alpha_H(B)$, if $h \asymp \{B^2n/(\sigma^2\log n)\}^{-1/(2\alpha+1)}$ and $\alpha \geq (3+\sqrt{57})/12$, we have with probability $1-n^{-10}$ over the randomness of (X_i, w_i)*

$$E[|f(x) - f^\star(x)|^2|\mathbb{D}_n] \leq B^{2/(2\alpha+1)}\left(\frac{\sigma^2\log n}{n}\right)^{\frac{2\alpha}{2\alpha+1}}$$

For the case of $f^* \in \Theta^\alpha_S(B)$, if $h \asymp \{B^2n/(\sigma^2\log n)\}^{-1/(2\alpha)}$ and $\alpha \geq 1/\sqrt{2}$, with the same probability we have

$$E[|f(x) - f^\star(x)|^2|\mathbb{D}_n] \leq B^{1/\alpha}\left(\frac{\sigma^2\log n}{n}\right)^{\frac{2\alpha-1}{2\alpha}}$$

The same bounds hold if we replace the pointwise bound $E[|f(x)-f^\star(x)|^2|\mathbb{D}_n]$ with the supre-

mum norm $E[\|f - f^\star\|_\infty^2 | \mathbb{D}_n]$

Proof in appendix C.3. Like the previous results, these claims are proven by using the existing theorems of [6] and making adjustments where needed for the truncation. With these risk bounds, the next step is to consider the coverage.

4.3.5 Frequentist coverage

The coverage results of [6] hold under truncation if the truncation point grows as before, $p \asymp \lambda^{-1/2\alpha}$.

For convenience, we review the notation needed for the subsequent theorem. The standard Gaussian c.d.f. is denoted Φ and z_γ represents the γ quantile, $\Phi(z_\gamma) = \gamma$. The credible interval for the truncated posterior $f|\mathbb{D}$ is

$$CI_{n,p} = [\hat{f}_{n,p} - l_{n,p}(x;\beta), \hat{f}_{n,p} + l_{n,p}(x;\beta)],$$

which is just the mean $\hat{f}_{n,p}$ with a margin $l_{n,p}$ equal to the posterior variance scaled by the quantile,

$$l_{n,p}(x,\beta) = z_{(1+\beta)/2}\sqrt{\widetilde{C}^B_{n,p}(x,x)}$$

By definition, the probability of the true function f^* appearing in the credible interval is nominally β. The true data generating distribution has probability function $\mathbb{P}_\rho$.

The variance bound of equation 4.10 is useful to the current discussion in that we can define a new process $W_p^B \sim GP(0, \hat{C}^B_{n,p})$ that is close to the centered posterior process $GP(0, \widetilde{C}^B_{n,p})$. This new process is related to the noise process

$$U(\cdot) = \sqrt{h/n}\sum_{i=1}^n w_i \widetilde{K}_{p,X_i}(\cdot) \sim GP(0, \hat{C}_{n,p})$$

which occurs as the higher order error in claim 1 and is a starting point for the proof.

Theorem 4.3.4. *For γ_n as in claim 1, there is a constant C independent of (n,h) such that that*

credible interval $CI_{n,p}(x;\beta)$ *satisfies the following bound for any* $x \in \mathcal{X}$*:*

$$\begin{aligned}\Big|\mathbb{P}_\rho\big[f^*(x) \in CI_{n,p}(x;\beta)\big] - \big[\Phi(u_{n,p}(x;\beta) + b_{n,p}(x)) - \Phi(-u_{n,p}(x;\beta) + b_{n,p}(x))\big]\Big| \\ \leq C\left(\frac{1}{\sqrt{nh}} + \gamma_n + \delta_n\right)\end{aligned} \tag{4.11}$$

where $u_{n,p} = \sqrt{\hat{C}^B_{n,p}/\hat{C}_{n,p}}z_{(1+\beta)/2}$ *and* $b_{n,p} = \{\hat{C}_{n,p}\}^{-1/2}\sqrt{nh}P_{\lambda,p}f^*(x)$ *is a bias.*

Proof in appendix C.4. As mentioned in [6], the remarkable aspect of this result is that the coverage is slightly conservative, in other words giving a slightly larger interval than nominally required.

Theorem 4.3.2 provides the suboptimal rate of convergence for an overtruncated kernel with $\alpha_o > \alpha$, and the resulting coverage has a slower rate of convergence. In particular, for the case that the truncation is fixed to some level p, the coverage is asymptotically 0. This is shown in appendix C.4.1.

Under cases where the smoothness is exactly matched, the coverage is nominal as shown in corollary 3.3 of [6]. In our context, the same situation holds when both smoothness and truncation are matched. In other words, if the true function has a truncated basis decomposition, the credible interval bands have the correct asymptotic coverage. This is shown in appendix C.5.

4.4 Conclusion

In this work, we showed how the supremum norm error bounds and frequentist coverage results for KRR of [6] still hold when the reproducing kernel for the underlying RKHS is approximated by truncating the basis decomposition. With the increasing popularity of Gaussian process models and kernel approximations for scalability with big data, a theoretical understanding of the coverage and minimax optimality is instrumental for guiding practical usage. While not a commonly used approximation, truncation is a simple case that serves as a first step towards the study of other kernel approximations.

Future work can investigate low rank or Nyström models like the predictive processes, in which

the kernel is approximated with a projection onto a finite dimensional subspace. A different direction would be sparse approximations such as nearest neighbor kernels. The key question for all of these cases is how weak the approximation can be without sacrificing estimator optimality under supremum norm.

5. CHARACTERIZATION FOR A NONSTATIONARY REPRODUCING KERNEL HILBERT SPACE

5.1 Introduction

Gaussian process (GP) models [2] account for spatial dependencies and allow for the direct specification of correlations through a kernel function. The relation between the stationary kernel properties and the sample path properties for realizations of a GP has been studied in detail [99], along with concentration [85]. These theoretical results rely on the characterization of the space of functions corresponding to all the possible realizations of the GP with a particular kernel. Given this characterization, it is possible to describe the size or metric entropy of the space, which determines performance in application. An underlying constraint for this theory is that the kernel is stationary.

Stationarity is a simplifying assumption that the correlations only depend on distance, not location or orientation. However, this is often unrealistic. For example, stock market volatility, diffusion of pollution in environments, and classification of images can exhibit correlations that vary over the input space. Moving beyond stationarity essentially reduces to replacing fixed correlations with varying terms.

In this work we build towards a theory for nonstationary Gaussian processes by characterizing the spaces of functions for nonstationary kernels. Our main contribution is a collection of definitions for Hilbert spaces and their reproducing kernels, which are closely related to each other and recover commonly used nonstationary kernels. These spaces are initially defined as finite combinations of spaces, but we prove that their limiting cases remain valid under reasonable conditions. This theory may be relevant beyond GP modeling; for example, neural networks are known to be equivalent to GPs [100] with nonstationary kernels that depend on the activation functions [101].

For inspiration we draw on a few types of nonstationary extensions for standard kernels. The simplest extensions assume piecewise stationarity by partitioning the space and assuming change-

points [102] or completely independent areas [103]. We also consider spectral methods that exploit Bochner's theorem or the Wiener-Khinchin theorem [104] to directly specify the spectral density of the kernel [105, 106] or to specify the dependent-increment process of the sample paths [107]. Warping methods [108, 109, 110] are attractive and powerful but difficult to work with due to the operation of composition.

Although there does not seem to be any literature on Hilbert spaces of nonstationary functions, there is a long history of inhomogeneous or variable exponent spaces. Variable integrability Lebesgue or Sobolev spaces, $L^{p(x)}$ and $W^{k,p(x)}$ respectively [111], are an intuitive starting point and a variety of results and embedding theorems are reviewed in [112]. Besov and Triebel-Lizortkin spaces [7] are frequently used as the prior space for more general inhomogeneous functions, such as those with variable smoothness and variable integrability simultaneously [113]. Besov spaces in particular are used for evaluating spatially adaptive techniques [114]. However, since part of our interest lies in defining reproducing kernels, we cannot directly use the aforementioned spaces.

The outline of this paper is as follows. In Section 5.2 we provide notation and background on Hilbert and Banach spaces and harmonizable functions. In Section 5.3 we characterize the space of functions for the change point kernel while a collection of special cases is described in Section 5.4. We provide simulations in Section 5.5 and next steps in the conclusion.

5.2 Background

5.2.1 Stochastic processes and Banach spaces

Following the notation of [115], let W be the stochastic process with RKHS $\mathcal{H}$. We can assume W maps a probability space (Ω, A, μ) into a Banach space $(B, \|\cdot\|)$ with dual space B^* such that $b^* \in B^*$ maps W to a Gaussian random variable. When we assume that $b^* = \pi_t$ with $t \in T$ is a coordinate projection, we have the relation

$$b^*(W) : \omega \to W_t(\omega)$$

which is the implicit connection between the original probability space with elements ω, the Banach space of realizations $W(\omega)$, and the Gaussian random variable $W_t(\omega)$. The span of the Banach space can be expressed as the range over B^* of the Pettis integral,

$$Sb^* = \int W(\omega) b^*(W(\omega)) d\mu(\omega) = EWb^*W \tag{5.1}$$

For the case of the dual space element being a projection $b^* = \pi_t$, it is possible to show that the Banach space coincides with the Hilbert space from the stochastic process through the relation $S\pi_t = K(t, \cdot)$, see theorem 2.2 of [115].

We show a more explicit form of the Hilbert space $\mathcal{H}$ by assuming the Banach space is separable and fixing a countable dense set of functions $(h_j)_{j=1}^{\infty}$. This set becomes an orthonormal basis by introducing iid Gaussians Z_j so that the process can be represented as

$$W = \sum_{j=1}^{\infty} Z_j h_j \tag{5.2}$$

with corresponding Pettis integral elements

$$Sb^* = EWb^*(W) = E(\sum_{j=1}^{\infty} Z_j h_j)(\sum_{j=1}^{\infty} Z_j b^* h_j) = \sum_{j=1}^{\infty} b^*(h_j) h_j = \sum_{j=1}^{\infty} w_j h_j,$$

so that the Hilbert space product is equivalent to the L_2 inner product of the random variables,

$$\langle Sb^*, Sb^* \rangle_H = b^*(Sb^*) = \sum_{j=1}^{\infty} b^*(h_j)^2 = \sum_{j=1}^{\infty} w_j^2.$$

This implies $w \in \ell_2$. This is the result of theorem 4.2 of [115], where the Hilbert space for stochastic process is represented as $\mathcal{H} = \{f = \sum_{i=1}^{\infty} w_i h_i : \|f\| = \sum w_i^2 \leq \infty\}$.

5.2.2 Harmonizable functions

Following [104] and [107], we may represent the spectral decomposition of a process $X(t)$ as a Fourier Stieltjes integral with random coefficients X_k:

$$X(s) = \sum e^{i\omega_k s} X_k = \int_\Omega e^{i\omega t} dZ(\omega)$$

For $Z(\omega)$ an independent increment process and δ_0 a delta function taking value 1 at 0 and 0 elsewhere, we have by definition that

$$E(Z(\omega_1 + d\omega)Z(\omega_2 + d\omega)) = \delta_{\omega_1 - \omega_2} E(Z(\omega_1 + d\omega)^2) = \delta_{\omega_1 - \omega_2} F(\omega_1 + d\omega) = \delta_{\omega_1 - \omega_2} dF(\omega)$$

The function $F(\omega)$ represents the spectral distribution for the covariance:

$$C(X(s), X(s+\tau)) = E(X(s)X(t)) = \int e^{i\omega(\tau)} dF(\omega).$$

For the case that $Z(\omega)$ is not an independent increment process, the resulting covariance is still well defined [104, Section 26.4] as

$$C(X(s), X(s+\tau)) = E(X(s)X(t)) = \int e^{i\omega_1 s - i\omega_2 (s+\tau)} dF(\omega_1, \omega_2),$$

assuming the distribution is integrable,

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} dF(\omega_1, \omega_2) < \infty.$$

The process $X(s)$ that has this nonstationary covariance is called a harmonizable process. A linear combination of (stationary) processes,

$$\mathcal{X}(s) = \sum_i w_i(s) X_i(s)$$

can be expressed as a harmonizable process [107]. In this work we go a step further and define a Hilbert space of such functions.

5.2.3 Multivariate extension

For notational convenience when working with a sum or mixture of functions, we provide notation for a vector extension of a Banach and Hilbert space. In other words, an element of the Hilbert space is represented as a vector rather than a single term. We explain the two-dimensional case, which easily extends to a finite number of components, and later extend to a countable set.

Take two independent stochastic processes, W_1 and W_2, each mapping into the same Banach space (eg L^2) and having corresponding Pettis integrals, $S_1 b^*$ and $S_2 b^*$, as in Section 5.2.1. Within the dual space, there exists the coordinate projections $\pi_t, t \in T$.

As shown in Section 5.2.1, the Hilbert space for a single stochastic process has an inner product with the relation

$$\langle Sb^*, Sb^* \rangle_H = b^* S b^* = E b^*(W) b^*(W)$$

This is easily extended with the S operator now a matrix.

$$Sb^* = \begin{bmatrix} S_1 & 0 \\ 0 & S_2 \end{bmatrix} \begin{bmatrix} b^\star \\ b^* \end{bmatrix} = \begin{bmatrix} S_1 b^\star \\ S_2 b^* \end{bmatrix} \tag{5.3}$$

The inner product for this space extends from the individual spaces:

$$\begin{aligned} \langle Sb_t^*, Sb_\tau^* \rangle_H &= \left\langle \begin{bmatrix} S_1 b_t^\star \\ S_2 b_t^* \end{bmatrix}, \begin{bmatrix} S_1 b_\tau^\star \\ S_2 b_\tau^* \end{bmatrix} \right\rangle_H \\ &= \begin{bmatrix} b_t^\star \\ b_t^* \end{bmatrix}^\top \begin{bmatrix} S_1 & 0 \\ 0 & S_2 \end{bmatrix} \begin{bmatrix} b_\tau^\star \\ b_\tau^* \end{bmatrix} \\ &= E b_t^*(W_1) b_\tau^*(W_1) + E b_t^*(W_2) b_\tau^*(W_2) \\ &= \langle b_t^*(W_1), b_\tau^*(W_1) \rangle_{L_2(\mu_1)} + \langle b_t^*(W_2), b_\tau^*(W_2) \rangle_{L_2(\mu_2)} \end{aligned} \tag{5.4}$$

The reproducing property is a simple consequence of this inner product:

$$\begin{aligned}\langle Sb_t^*, h\rangle_H =& \langle \begin{bmatrix} S_1 b_t^\star \\ S_2 b_t^* \end{bmatrix}, \begin{bmatrix} h_1 \\ h_2 \end{bmatrix} \rangle_H \\ =& b_t^\star(h_1) + b_t^*(h_2) \\ =& h_1(t) + h_2(t) \\ =& f(t)\end{aligned} \tag{5.5}$$

5.3 Characterization for a general kernel

5.3.1 Kernel description

A nonstationary kernel can be expressed as a convolution of stationary kernels K_i as seen in [107]:

$$C(x, x') = \sum_{i=1}^{m} \psi_i(x)\psi_i(x')K_i(x, x') \tag{5.6}$$

To express the limiting case, we write the stationary kernels K_i as a single kernel with varying parameters, $C_{\theta(s)}$. For example, the parameters could be smoothness, range, and variance of a Matérn kernel, $\theta(s) = (\nu(s), \rho(s), \sigma^2(s))$. Then the limiting kernel takes the form

$$C(x, x') = \int \psi_s(x)\psi_s(x')C_{\theta(s)}(x, x')ds \tag{5.7}$$

Special cases are discussed in Section 5.4 and include finite changepoint kernels, kernel convolution, and spectral mixtures. This kernel can be seen as the limiting covariance of a mixture of functions f_i from independent Hilbert spaces $\mathcal{H}_i$. The mixture is

$$\lim_{m\to\infty} \frac{1}{m} \sum_{i=1}^{m} \psi_i(x)f_i(x) = \int \psi_s(x)f_s(x)ds, \tag{5.8}$$

so that we recover the covariance by the assumption of independent Hilbert spaces:

$$\begin{aligned} Cov(f(x), f(x')) =& Cov\left[\int \psi_s(x) f_s(x) ds\right]\left[\int \psi_{s'}(x') f_{s'}(x') ds'\right] \\ =& \int\int \psi_s(x)\psi_{s'}(x') C(f_{s'}(x'), f_s(x)) ds ds' \\ =& \int \psi_s(x)\psi_s(x') C_s(x, x') ds = C(x, x') \end{aligned} \tag{5.9}$$

The mixture can be shown to induce dependent increments [107, 104]. First express the mixture products as a convolution over a Fourier basis,

$$\psi_i(x) f_i(x) = \mathcal{F}(\hat{\psi}_i * \hat{f}_i) \tag{5.10}$$

Denote the convolution with $\hat{\psi}$ as an integral operator L_ψ, so that

$$\hat{\psi}_i * \hat{f}_i = \int \hat{\psi}_i(\cdot - v)\hat{f}_i(v) dv = L_{\psi_i}\hat{f}_i = \int L_{\psi_i}(\cdot, v)\hat{f}_i(v) dv \tag{5.11}$$

For a Fourier series representation $f_i = \sum_{j=1}^{\infty} f_{ij}\phi_j$ and similar for ψ_i, the convolution operator denoted L_ψ acts on the series coefficients:

$$\psi_i(x) f_i(x) = \sum_{j=1}^{\infty}\sum_{k=1}^{\infty} \psi_{i(j-k)} f_{ik}\phi_k(x) = \sum_{j=1}^{\infty} [L_{\psi_i}\hat{f}]_j \phi_j(x) \tag{5.12}$$

When computing the covariance with respect to the convolved functions, we still have that the Hilbert spaces are independent, but the functions within the Hilbert spaces derive from dependent increment process rather than independent increment processes due to the multiplication of ψ functions. Let $S(w, v)$ represent a spectral density for the covariance:

$$
\begin{aligned}
Cov(\psi_i(x)f_i(x), \psi_i(x')f_i(x')) =& E(\int e^{iwx}(L_\psi \hat{f})(w)dw \int e^{-ivx'}(L_\psi \hat{f})(v)dv \\
=& \int e^{i(wx-vx')} E\left[(L_\psi \hat{f})(w)dw(L_\psi \hat{f})(v)dv\right] \\
=& \int e^{i(wx-vx')} S(w,v)dwdv
\end{aligned} \tag{5.13}
$$

This expression can be proven by properties of Fourier transforms; a direct proof is in the Appendix, see D.1.1.

5.3.2 Linear combinations of Hilbert spaces

We are interested in combinations of functions as shown in equation 5.8:

$$
f(x) = \psi_1(x)f_1(x) + \cdots + \psi_m(x)f_m(x). \tag{5.14}
$$

We first define the Hilbert space with a direct approach using the convolution of Equation 5.10 and later consider an indirect approach.

Theorem 5.3.1. *Let $f_i \in \mathcal{H}_i$ be functions from stationary Hilbert spaces with reproducing kernels K_i all expressed in a common Fourier basis, $\{\phi_i\}$, as $K_i(x,x') = \sum_{j=1}^{\infty} u_{ij}\phi_j(x)\phi_j(x')$. Each $\mathcal{H}_i$ is the closure of the set of functions $f_i(x) = \sum_{j=1}^{\infty} f_{ij}\phi_j(x)$ where $\sum_{j=1}^{\infty} f_{ij}^2/u_{ij} < \infty$. Let $\hat{f}_i$ represent the sequence of coefficients. Suppose ψ_i are any smooth, bounded, positive functions that also have Fourier series representations $\psi_i(x) = \sum_{j=1}^{\infty} \psi_{ij}\phi_j(x)$. Expressing the product as a series convolution as in equation 5.12,*

$$
\psi_i(x)f_i(x) = \sum [L_{\psi_i}\hat{f}_i]_j \phi_j(x),
$$

the Hilbert space is the closure of the span of functions below across all smooth bounded positive

functions ψ_i and functions f_i from their respective Hilbert spaces,

$$\mathcal{H} = \left\{ f = \begin{bmatrix} \sum_{j=1}^{\infty} (L_{\psi_1} \hat{f}_1)_j \phi_j \\ \vdots \\ \sum_{j=1}^{\infty} (L_{\psi_m} \hat{f}_m)_j \phi_j \end{bmatrix} : \|f\| = \sum_{i}^{m} \sum_{j=1}^{\infty} \frac{(L_{\psi_i} \hat{f}_i)[j]^2}{u_{ij}} < \infty \right\} \tag{5.15}$$

with reproducing kernel expressed as a vector,

$$K_x = \begin{bmatrix} K_1(x, \cdot) & \dots & K_m(x, \cdot) \end{bmatrix}^\top . \tag{5.16}$$

Proof in appendix D.2. We remark that as a special case, having weight functions with coefficients $\hat{\psi}_{ij} = u_{\psi j} \phi_j(s)$ implies that $\psi_i(x) = K(s_i, x)$ is a kernel weighting function centered at some s_i.

5.3.3 Spectral diffusion

The spaces described in the previous section rely on explicit scaling functions ψ, which induce dependent increments as mentioned in sections 5.2.2 and near equation 5.10. We propose another space by observing that there is nothing stopping us from introducing a lower triangular correlating operator $L : \ell_\infty \to \ell_\infty$ (an infinite dimensional "Cholesky" matrix) to the process W of equation 5.2:

$$W = (LZ)^\top h = \sum_{j=1}^{\infty} \mathcal{Z}_j h_j \tag{5.17}$$

Now the previously iid Z_j have become correlated $\mathcal{Z}_j$. Let $LL^\top = \Sigma = [s_{ij}]$ so that

$$E\mathcal{Z} = 0, \qquad E(\mathcal{Z}\mathcal{Z}^\top) = \Sigma.$$

Then it can be shown that the process above has dependent increments and is not stationary [104].

Theorem 5.3.2 (Diffused spectrum RKHS)**.** *Assume a separable Banach space B has a sequence of elements $\{h_j\}$ where an ℓ_2 sequence v with $\sum v_j h_j$ converges in B and $v = 0$ if the sum is 0.*

Further, $v = Lw$ for some other sequence $w \in \ell_2$ and a bounded operator L. Now let (Z_j) be a sequence of iid standard Gaussian variables so that the process $W_{NS} = (LZ)^\top h$ converges a.s. in B. Then the nonstationary RKHS for the process W_{NS} is

$$\mathcal{H} = \{f = \sum_{j,k=1}^{\infty} w_j s_{jk} h_k : \|f\|_H = \sum_{j,k=1}^{\infty} s_{jk} w_j w_k < \infty, w \in \ell_2\}$$

The elements of the space can be represented as Pettis integrals,

$$f = S_L b^*.$$

The vector of Pettis integrals with b^* a projection π_t is the reproducing kernel.

This case represents a single mixture component.

Proof in Appendix D.4. For convolutions with an L derived from a smoothing covariance matrix, we observe a localization effect that causes the process covariance to decay. This is described in appendix D.7. With just a bit more notation, we extend the previous result to multiple convolutions as in Theorem 5.3.1 with different contributing spaces.

Corollary 5.3.1 (Finite Spectral Mixture RKHS)**.** *Assume the conditions of theorem 5.3.2. Denote different processes W_i by using a single sequence of basis elements $h = (h_j)$ and a basis scaling sequence $\sigma_i = (\sigma_{ij}) = (\|h_{ij}\|)$ with notation $\sigma_i \odot h$. The nonstationary process is defined as*

$$W_{SM} = \sum_{i}^{m} (L_i Z)^\top (\sigma_i \odot h)$$

The nonstationary RKHS for the process W_{SM} is

$$\mathcal{H} = \left\{ f = \begin{bmatrix} \sum_{j,k=1}^{\infty} w_{1j} s_{jk}^{(1)} \sigma_{1k} h_k \\ \vdots \\ \sum_{j,k=1}^{\infty} w_{mj} s_{jk}^{(m)} \sigma_{mk} h_k \end{bmatrix} : \|f\|_H = \sum_{i=1}^{m} \sum_{j,k=1}^{\infty} s_{jk}^{(i)} w_{ij} w_{ik} < \infty \right\} \tag{5.18}$$

where we now assume $w_{ij} = \sigma_{ij} b^*(h_j)$

There are at least two special cases of the theorem above: one for a single process with multiple correlations L, and another for a single correlator L and multiple processes W_i.

5.3.4 Equivalent spaces

The space we define in Theorem 5.3.1 can be expressed from a Banach space point of view just as the stationary stochastic process RKHS is expressed as a Banach space RKHS in [115]. We also clarify how the stochastic process relates to the dependent increment process of Theorem 5.3.2.

We first express the Banach space RKHS. The nonstationary kernel is linked to a Pettis integral of the Banach space by bringing the scaling functions ψ into the dual space. For the projection $b_s^* = \pi_s(\cdot)$, the scaled projection is written $b_{s,\psi}^* = \psi(s)\pi_s(\cdot)$. The stationary case,

$$\langle K_s, K_t \rangle = K(s,t) = E\pi_s^*(W)\pi_t^*(W) = \langle Sb_s^*, Sb_t^* \rangle$$

becomes

$$\langle L_\psi K, L_\psi K_t \rangle = \psi(s)K(s,t)\psi(t) = E\pi_{s,\psi}^*(W)\pi_{t,\psi}^*(W) = \left\langle Sb_{s,\psi}^*, Sb_{t,\psi}^* \right\rangle$$

Hence, for a particular vector of scaling functions ψ_i, $i = 1, ..., m$, we can express the process W_{NS} as the span over the dual space B^* of Pettis integrals Sb_ψ^*. Following the results of [115], when the dual elements are restricted to projections $\psi(s)\pi_s(\cdot)$, the RKHS and stochastic process coincide.

Lemma 1. *The stochastic process RKHS described in theorem 5.3.1 is equivalent to the following Banach space RKHS, represented as a span over the dual space $b^* \in B^*$:*

$$\mathcal{H} = \left\{ f = \begin{bmatrix} S_1 b_{\psi_1}^* \\ \vdots \\ S_m b_{\psi_m}^* \end{bmatrix} : \|f\| = \sum_i^m \sum_{j=1}^\infty b_{\psi_i}^*(h_j)^2 < \infty \right\}$$

It is assumed that the Banach space random element is in a complete separable subspace of $\ell^\infty(T)$

with the uniform norm.

Proof in appendix D.3.

To relate the combination spaces to the diffusion spaces, we can check the conditions under which they yield the same kernel. The diffusion relies on a matrix multiplication, which can be expressed as a convolution if the columns of the matrix are shifted copies padded with zero:

$$\begin{aligned} S_L b^* =& E(W_L b^*(W_L)) \\ =& \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} h_j s_{jk} b^*(h_k) \\ =& \sum_{j=1}^{\infty} h_j \sum_{k=1}^{\infty} s_{jk} b^*(h_k) \\ =& \sum_{j=1}^{\infty} h_j (s * b^* h)_j \end{aligned}$$

From the last line, we can connect the diffusion case to the combination case if we assume that h_j are a Fourier basis, $b^*(h)$ represent Fourier coefficients for a function, and s is the spectral density of Σ. See appendix D.7 for additional details about the diffusion approach.

5.3.5 Infinite combinations

In the previous section we defined the Hilbert spaces for finite linear combinations. As demonstrated in [107] the convolution kernel of Equation 5.7 is the Monte Carlo limit of the finite case once we add an additional $1/m$ scaling term. If the ψ_i are uniformly bounded, the multiplication by a normalizing term $1/m$ is equivalent to the condition that $\sup_x \psi_i(x) \to 0$ as $m \to \infty$.

Theorem 5.3.3. *Taking the limit $m \to \infty$ for the RKHS specified in Theorem 5.3.1,*

$$\mathcal{H} = \left\{ f = \begin{bmatrix} \sum_{j=1}^{\infty} (L_{\psi_1} \hat{f}_1)_j \phi_j \\ \vdots \\ \sum_{j=1}^{\infty} (L_{\psi_\infty} \hat{f}_2)_j \phi_j \end{bmatrix} : \|f\| = \sum_{i}^{\infty} \sum_{j=1}^{\infty} \frac{(L_{\psi_i} \hat{f}_i)[j]^2}{u_{ij}} < \infty \right\} \tag{5.19}$$

remains a valid RKHS within the Banach space if ψ_i satisfy the conditions of Theorem 5.3.1

and form a uniformly bounded resolution, ie there exists a constant c such that

$$\sum_{i=1}^{\infty} \psi_i(x) \leq c \quad \forall x.$$

Proof in Appendix D.5. Some additional remarks:

- The functions ψ_i are not necessarily in the Banach space (for example, $\psi_i = 1$ is not integrable), but $\psi_i f_i$ are in the space by the assumptions of boundedness, positivity, and smoothness. Supposing the Banach space uses the L2 norm, we have $f_i \in B$ implies $\|f_i\| = \int f_i^2(x)dx < \infty$. Since $0 \leq \psi_i \leq 1$, $\|\phi_i f_i\| = \int \psi_i^2 f_i^2 \leq \|f_i\|$.

- As a special case of the previous result, take a compact domain, say $x \in [0, 1]$, and let the bump functions take the form $\psi_i(x) = 1_{x \in [\frac{i-1}{n}, \frac{i}{n}]}$. In the limit we have $\psi_i(x)\psi_j(y) = \delta_{ij}\delta(x = y)$ implying every point is independent. Since each point has its own kernel, we have a very general heteroskedastic white noise model which can perfectly (over)fit any collection of observations.

The equivalent theorem can be stated for the diffusion perspective.

Theorem 5.3.4. *For a collection of correlating operators L such that $\sum_{i=1}^{\infty} \|L\|_{op} < \infty$, the infinite sum of diffused processes converges to the Hilbert space*

$$\mathcal{H} = \left\{ f = \begin{bmatrix} \sum_{j,k=1}^{\infty} w_{1j} s_{jk}^{(1)} \sigma_{1k} h_k \\ \vdots \\ \sum_{j,k=1}^{\infty} w_{\infty j} s_{jk}^{(\infty)} \sigma_{\infty k} h_k \end{bmatrix} : \|f\|_H = \sum_{i=1}^{\infty} \sum_{j,k=1}^{\infty} s_{jk}^{(i)} w_{ij} w_{ik} < \infty \right\} \tag{5.20}$$

where $w_{ij} = \sigma_{ij} b^*(h_j)$

Proof in Appendix D.6.

5.4 Special cases

In section 5.3.1, we review how a convolution kernel can be expressed as the Monte Carlo limit of a linear combination of kernels. Depending on the weighting functions ψ_i, we recover different

cases. The general case has $\psi_i(x) = K_{\theta(x_i)}(x_i, x)$, so that ψ is itself a kernel where, for example, the range parameter is specified by $\theta(x)$. This is a variable bandwidth kernel as described in [114]. Another basic case assumes the underlying stochastic process are white noise processes,

$$C(W_{\theta(s)}(x), W_{\theta(s')}(x')) = \delta(s - s').$$

This results in a simplified version of kernel convolution,

$$C(f(x), f(x')) = \int K(x, s)K(x', s)ds.$$

We review a few other kernels that do not look like kernel convolution but are nonetheless easy to derive as linear combination with appropriate weighting functions.

5.4.1 Changepoint kernels

A single changepoint kernel [102] can be expressed as a combination of $m = 2$ stationary kernels using a sigmoid weighting function,

$$CP(K_1, K_2) = \psi(x_1)K_1(x_1, x_2)\psi(x_2) + (1 - \psi(x_1))K_2(x_1, x_2)(1 - \psi(x_2)) \tag{5.21}$$

$$\psi(x) = \frac{1}{1 + e^{-x}}$$

This is easily extended to the case of multiple changepoints on a compact domain by using a resolution $\{\psi_i\}$ of bump functions:

$$\psi_{start,width}(x) = \frac{1}{1 + e^{x-\text{start}}} \frac{1}{1 + e^{width-(x-start)}}.$$

In the limiting case, the bump functions approach delta functions and the convolution approaches a heteroskedastic white noise process.

5.4.2 Multiresolution kernel

We can define ψ_i in terms of a tree to get multiresolution interpretations; for example, let ψ_0 have support $[0, 1]$ and correspond to a kernel with a high degree of smoothness or a large range. Then define ψ_1 on $[0, 1/2]$ and ψ_2 on $[1/2, 1]$ with corresponding kernels with half the range of the first level or possible a completely different kernel, such as a periodic or linear one. Subsequent components follow the pattern, for example ψ_3 on $[0, 1/4]$, and so on.

By varying the cardinality of weighting functions at a value in the support, we have the interpretation of a spatially adapted kernel.

5.4.3 Multivariate kernels

For $x \in R^p$, we can express an additive component-wise kernel that is nonstationary for each component in our vector notation:

$$C(x, x') = \sum_{i=1}^{p} \psi_i(x_i)\psi_i(x_i')K_i(x_i, x_i')$$

More generally, a cross-covariance can be accommodated with the vector notation in the sense of Kronecker products. For example, the following expressions fit into our notation by substituting the original vector notation with Kronecker products of identity matrices or column vectors of ones.

$$\psi_i(x)^\top = [\psi_{i1}(x_1), \ldots, \psi_{ip}(x_p)]$$

$$K_i(x, x') = \begin{bmatrix} K_i(x_1, x_1) & K_i(x_1, x_2) & \ldots & K_i(x_1, x_p) \\ K_i(x_1, x_1) & K_i(x_1, x_2) & \ldots & K_i(x_1, x_p) \\ \vdots & \vdots & \ddots & \vdots \\ K_i(x_p, x_1) & K_i(x_p, x_2) & \ldots & K_i(x_p, x_p) \end{bmatrix}$$

$$C(x, x') = \sum_{i=1}^{m} \psi_i(x)^\top K_i(x, x')\psi_i(x')$$

5.4.4 Spectral mixtures

Spectral mixtures as proposed by [105] are a way to fit stationary but nonstandard kernels. The kernel spectrum is approximated with a symmetric mixture of Gaussians along the lines of Bochner's theorem (or Wiener-Khinchin):

$$K(\tau) = \int e^{2\pi i w \tau} S(w) dw$$

Assuming $\phi(s; \mu, \sigma^2) = N(s|\mu, \sigma^2)$, we take a mixture of such ϕ's and force symmetry to get a real-valued spectral density

$$S(s) = \sum_{i=1}^{m} a_i(\phi_i(s; \mu_i, \sigma_i^2) + \phi_i(-s; \mu_i, \sigma_i^2))/2. \tag{5.22}$$

The corresponding kernel is easily shown to be

$$K(\tau) = \sum_{i=1}^{m} a_i \exp(-2\pi^2\tau^2\sigma_i^2)\cos(2\pi\tau\mu_i). \tag{5.23}$$

Using the Gaussian mixture leads to a smooth kernel; this observation is used to extended to more general classes by [106], plugging in a Matern kernel in equation (5.23) instead of the squared exponential. They further extend the mixture strategy to the nonstationary case, using a Hadamard product for the case $x \in R^p$ and hiding any norms in the function $C(\cdot, \cdot)$

$$K_n(x, x') = \sum_{i=1}^{n} a_i C(x \odot \gamma, x' \odot \gamma)\Psi_i(x)^\top \Psi_i(x') \tag{5.24}$$

Here the $\Psi_i(x) = \begin{pmatrix} \cos(2\pi x^\top w_i^1) + \cos(2\pi x^\top w_i^2) \\ \sin(2\pi x^\top w_i^1) + \sin(2\pi x^\top w_i^2) \end{pmatrix}$.

In the appendix we showed how $\sum \psi_i(x) f_i(x)$ corresponds to a spectral density of the form

$$S(w, v) = \sum_{i=1}^{m} \int L(w, s) S_i(s) L(v, s) ds$$

To recover the stationary spectral mixture from convolutions, we would need the correlating matrices to become diagonal,

$$L_i(w, s) = \sqrt{\phi_i(w) + \phi_i(-w)}\delta(w - s).$$

However, our characterization cannot handle the more general case in which the spectrum is modeled with the inverse Fourier transform of the Gibbs kernel, illustrated in [116].

5.5 Simulations

In this section we illustrate a few of the special cases of the nonstationary Hilbert space and use the kernel representations to generate sample paths.

5.5.1 Change point kernels

Change point kernels are straightforward to use for generating sample paths. Using either a grid or a random set of locations, sample paths can be realized by computing the Gram matrix for the locations, taking the Cholesky decomposition, and applying the Cholesky matrix to a random vector of independent standard normal variables.

5.5.2 Spectral smoothing

The spectral method is more involved. We first need to choose or sample a base spectrum, and then spread the spectrum beyond the diagonal using a Cholesky operator, which itself may be computed from some kernel or spectrum. In formulas:

1. Set a grid of frequencies, for example $w_i \in [0, 10]$ with $i = 1 : n, n = 200$ points. Denote the continuous spectrum for some kernel (such as Matern) as $\lambda(w_i)$ for the gridded frequencies w_i.

2. Choose some additional kernel, such as a squared exponential (can add periodic kernel, etc), and compute the Gram matrix $K(w_i, w_j)$ for the gridded frequencies w_i. Compute the Cholesky L and apply to the spectrum $\lambda(w)$ on both sides to get a new spectral density

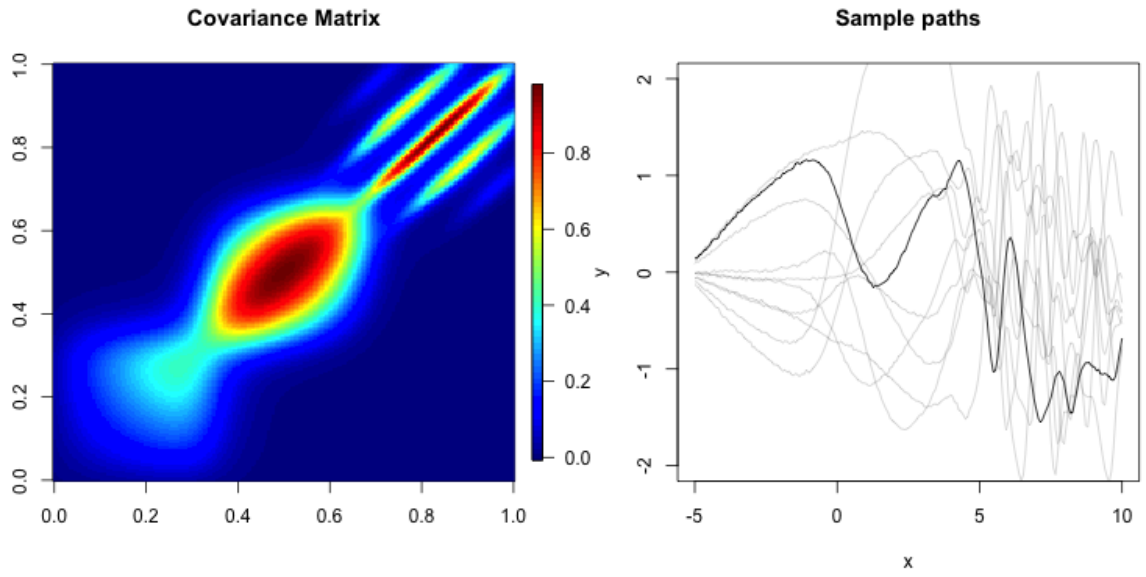


Figure 5.1: Linear, square exponential, and period kernels.

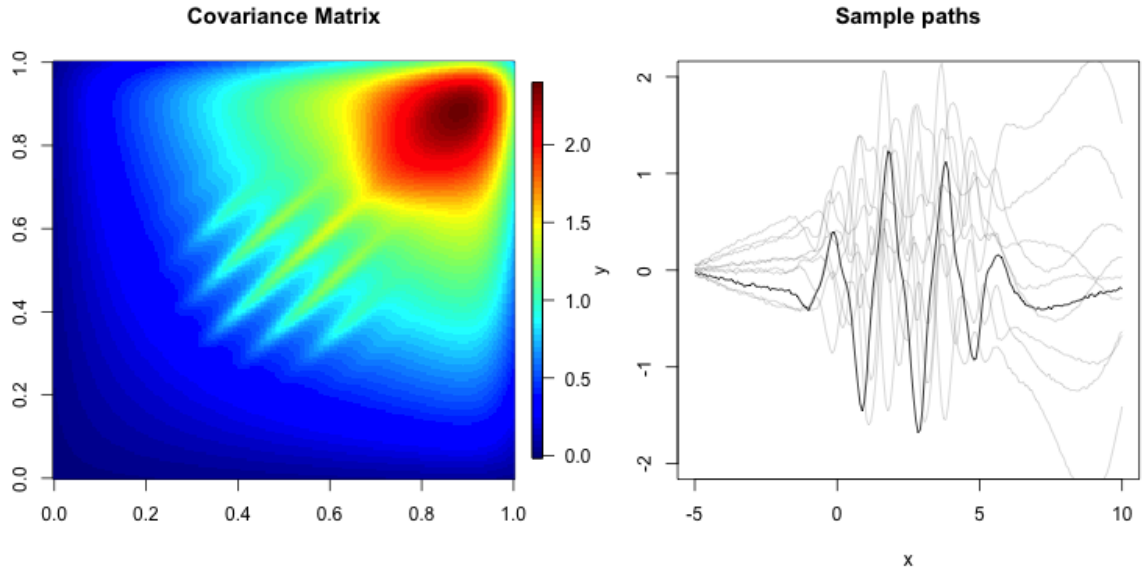


Figure 5.2: A multiresolution kernel. The lowest level is a linear kernel. The center [0,5] adds a periodic effect and the right segment [5,10] is squared exponential, adding a smooth curve to the linear effect. The dips at 10 are a reversion to the mean as the kernels all drop out at that point.

expressed as a matrix,

$$\tilde{f} = L\lambda(w)L^\top, \quad f(w_i, w_j) = \frac{[L\lambda(w)L^\top]_{ij}}{\|\tilde{f}\|_F}$$

This is based on equation D.2. The normalization term is represented as a Frobenius norm of the matrix $\tilde{f}$ and ensures that the f matrix is a probability density.

3. Estimate the covariance between two points x_1 and x_2 by a sum,

$$\sum_i \sum_j \cos(2\pi x_1 w_i) f(w_i, w_j) \cos(2\pi x_2 w_j) + \sin(2\pi x_1 w_i) f(w_i, w_j) \sin(2\pi x_2 w_j)$$

This sum approximates $\int e^{2\pi i(wx_1 - w'x_2)} f(w, w') dw dw'$.

The figures below demonstrate how the smoothed spectral density leads to a decaying kernel and degenerate sample paths.

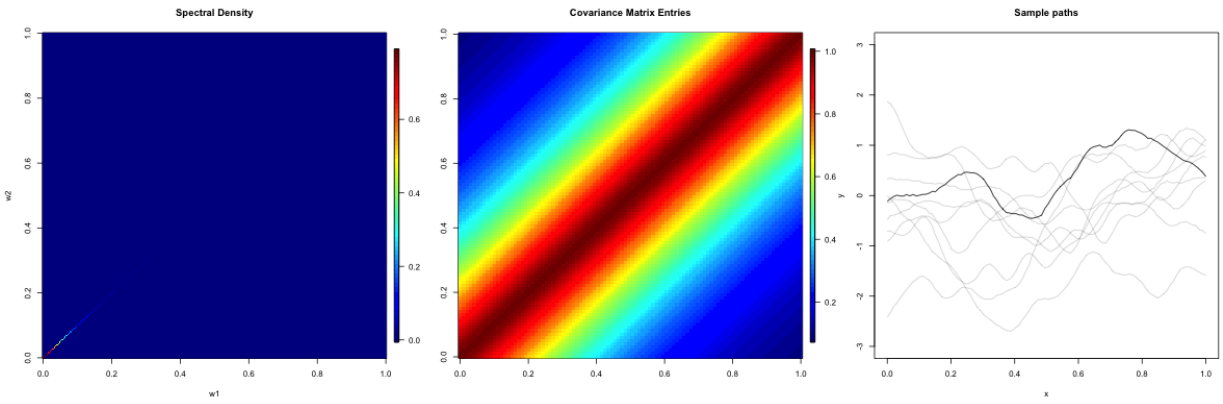


Figure 5.3: Stationary case for comparison. The first plot shows a diagonal matrix representing the spectrum, the second plot shows the corresponding covariance over a grid of points on [0,1], and the third plot has sample paths.

5.6 Conclusion

In this work we provided a characterization of a nonstationary RKHS by taking linear combinations of stationary Hilbert spaces. By considering special cases of the weighting functions

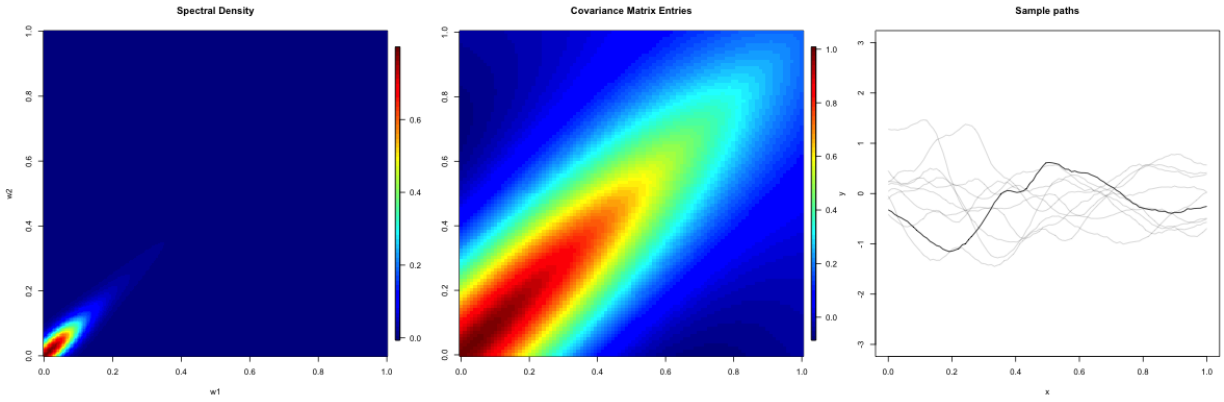


Figure 5.4: Nonstationary case with slightly dependent spectral terms

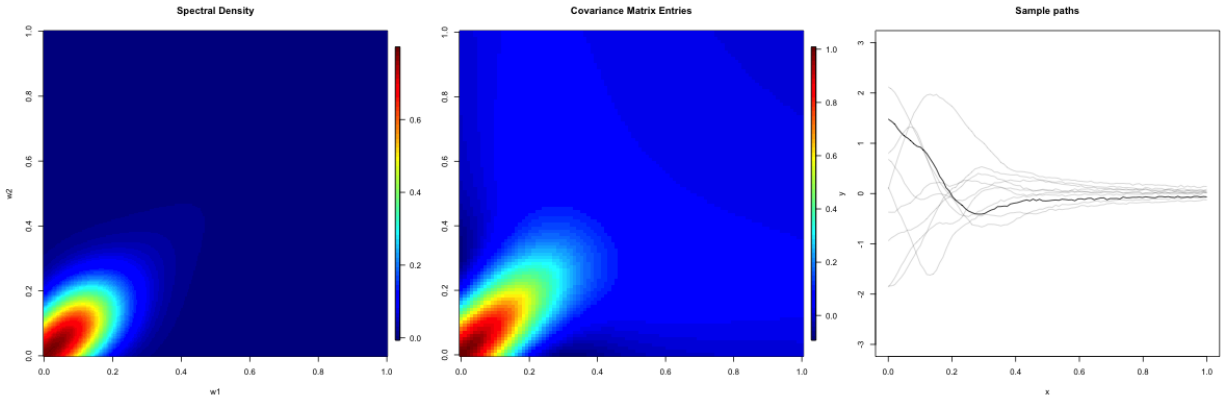


Figure 5.5: Nonstationary case with highly dependent spectral terms. The covariance decays noticeably and the sample paths revert to the mean

in the linear combinations, we can easily recover a few common kernels that are used for modeling nonstationary data, including change-point kernels and spectral mixtures. Under reasonable conditions, we recover a space of functions represented through kernel convolutions.

Future work in this area can be taken in a few directions. An initial motivation was to derive entropy properties which can be used to compute concentration results for nonstationary function estimators. We expect that concentration under estimators that allow for nonstationarity should be faster than more rigid estimators that require large function spaces if additional data is provided, such as the regions of nonstationarity.

An alternative direction for this work is to establish embeddings of our nonstationary space into the existing spaces mentioned earlier, such as variable Lebesgue or Besov spaces. Such embeddings can guide further work for transferring results from other spaces onto our own.

6. SUMMARY AND CONCLUSIONS

This thesis presented four works related to the modeling of spatial correlation with Gaussian processes. While the projects vary in complexity and emphasis on theory versus application, they share a central object of interest, the covariance kernel.

The VL and truncated KRR projects focus on improving efficiency at the expense of accuracy through covariance approximation. In the VL project, we showed that a sparse general Vecchia approximation, which is a type of nearest neighbor approximation, allows for a great improvement in efficiency, from cubic to linear computational complexity in the sample size, with very little cost in accuracy. The most efficient and common competitor, a low rank approximation, did not provide such an advantageous trade. For the truncated KRR project, we showed that a low rank approximation based on a truncated eigendecomposition can actually be optimal and therefore not incur any accuracy penalty, but requires some knowledge about the true function and a number of basis terms that scales with the sample size.

The remote sensing application and nonstationary RKHS projects involve improving model accuracy at the cost of a greater computational burden. When using satellite radiance data collected by cameras such as AVIRIS-NG, we showed that it can help to introduce correlations into the radiative transfer model. The inversion process becomes less efficient, but the resulting estimates for the atmospheric terms can be more accurate and realistic. We consider this same goal of accurate modeling theoretically in the nonstationary RKHS project, where we provide explicit decompositions of nonstationary functions in terms of linear combinations of stationary functions. The corresponding reproducing kernel has a similar decomposition, allowing for a more accurate model of the correlations when there is known nonstationarity compared to using a single stationary kernel.

That kernels or covariance functions are the central theme of this work is no coincidence. Although GP models are often used for environmental applications with small data sets, their utility extends to far more data types and sizes when one takes advantage of the flexibility of the ker-

nel. In fact, the kernel can be seen as a thread that intertwines a variety of seemingly disparate fields. In statistics, as demonstrated in this thesis, the properties of the aforementioned Gaussian stochastic processes are almost entirely specified by a kernel. In functional analysis, the definition for a complete vector space equipped with an inner product, ie a Hilbert space, can be given in terms of a kernel. In the study of differential equations, solutions can be expressed in terms of Green's functions, integral operators that contain kernels. In machine learning, techniques such as neural networks are essentially nonparametric estimators with complicated, but in some cases tractable, kernels. Furthermore, many existing linear data analysis techniques such as support vector machines or linear discriminant analysis are special cases of nonlinear kernel techniques. Even quantum mechanics can be understood through operators on a Hilbert space where probability, and in particular covariance, is generalized to account for particle interference. It is initially surprising that one object can be so broadly applicable, but the generality is natural when considering that so much of the mathematical and physical modeling of the world is simplified down to interactions between pairs of objects. It is this revelation that inspires the author to continue to pursue the study of covariance and kernels.

REFERENCES

[1] D. R. Thompson, L. Guanter, A. Berk, B.-C. Gao, R. Richter, D. Schläpfer, and K. J. Thome, "Retrieval of atmospheric parameters and surface reflectance from visible and shortwave infrared imaging spectroscopy data," *Surveys in Geophysics*, vol. 40, no. 3, pp. 333–360, 2019.

[2] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[3] M. Katzfuss and J. Guinness, "A general framework for Vecchia approximations of Gaussian processes," *Statistical Science*, vol. accepted, 2019.

[4] D. R. Thompson, V. Natraj, R. O. Green, M. C. Helmlinger, B.-C. Gao, and M. L. Eastwood, "Optimal estimation for imaging spectrometer atmospheric correction," *Remote Sensing of Environment*, vol. 216, pp. 355–373, 2018.

[5] C. D. Rodgers, *Inverse methods for atmospheric sounding: theory and practice*, vol. 2. World scientific, 2000.

[6] Y. Yang, A. Bhattacharya, and D. Pati, "Frequentist coverage and sup-norm convergence rate in gaussian process regression," *arXiv preprint arXiv:1708.04753*, 2017.

[7] H. Triebel, "Theory of function spaces ii," *Monographs in mathematics*, vol. 84, pp. 56199–11367, 1992.

[8] P. Diggle, J. Tawn, and R. Moyeed, "Model-based geostatistics," *Journal of the Royal Statistical Society, Series C*, vol. 47, no. 3, pp. 299–350, 1998.

[9] A. B. Chan and D. Dong, "Generalized Gaussian process models.," in *CVPR*, pp. 2681–2688, 2011.

[10] L. Shang and A. B. Chan, "On approximate inference for generalized Gaussian process models," *arXiv:1311.6371*, 2013.

[11] M. Filippone and M. Girolami, "Pseudo-marginal Bayesian inference for Gaussian processes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 11, pp. 2214–2226, 2014.

[12] H. Rue, S. Martino, and N. Chopin, "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations," *Journal of the Royal Statistical Society: Series B*, vol. 71, pp. 319–392, apr 2009.

[13] L. Tierney and J. B. Kadane, "Accurate approximations for posterior moments and marginal densities," *Journal of the American Statistical Association*, vol. 81, no. 393, pp. 82–86, 1986.

[14] C. K. Williams and D. Barber, "Bayesian classification with Gaussian processes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 12, pp. 1342–1351, 1998.

[15] W. H. Bonat and P. J. Ribeiro Jr, "Practical likelihood analysis for spatial generalized linear mixed models," *Environmetrics*, vol. 27, no. 2, pp. 83–89, 2016.

[16] M. J. Heaton, A. Datta, A. O. Finley, R. Furrer, J. Guinness, R. Guhaniyogi, F. Gerber, R. B. Gramacy, D. Hammerling, M. Katzfuss, F. Lindgren, D. W. Nychka, F. Sun, and A. Zammit-Mangion, "A case study competition among methods for analyzing large spatial data," *Journal of Agricultural, Biological, and Environmental Statistics*, vol. 24, no. 3, pp. 398–425, 2019.

[17] D. Higdon, "A process-convolution approach to modelling temperatures in the North Atlantic Ocean," *Environmental and Ecological Statistics*, vol. 5, no. 2, pp. 173–190, 1998.

[18] J. Quiñonero-Candela and C. E. Rasmussen, "A unifying view of sparse approximate Gaussian process regression," *Journal of Machine Learning Research*, vol. 6, pp. 1939–1959, 2005.

[19] 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*, vol. 70, no. 4, pp. 825–848, 2008.

[20] N. Cressie and G. Johannesson, "Fixed rank kriging for very large spatial data sets," *Journal of the Royal Statistical Society, Series B*, vol. 70, no. 1, pp. 209–226, 2008.

[21] M. Katzfuss and N. Cressie, "Spatio-temporal smoothing and EM estimation for massive remote-sensing data sets," *Journal of Time Series Analysis*, vol. 32, no. 4, pp. 430–446, 2011.

[22] M. L. Stein, "Limitations on low rank approximations for covariance matrices of spatial data," *Spatial Statistics*, vol. 8, pp. 1–19, jul 2014.

[23] R. Furrer, M. G. Genton, and D. Nychka, "Covariance tapering for interpolation of large spatial datasets," *Journal of Computational and Graphical Statistics*, vol. 15, no. 3, pp. 502–523, 2006.

[24] 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*, vol. 103, no. 484, pp. 1545–1555, 2008.

[25] C. R. Dietrich and G. N. Newsam, "Fast and exact simulation of stationary Gaussian processes through circulant embedding of the covariance matrix," *SIAM Journal on Scientific Computing*, vol. 18, no. 4, pp. 1088–1107, 1997.

[26] S. Flaxman, A. Wilson, D. Neill, H. Nickisch, and A. Smola, "Fast Kronecker inference in Gaussian processes with non-Gaussian likelihoods," in *International Conference on Machine Learning*, pp. 607–616, 2015.

[27] J. Guinness and M. Fuentes, "Circulant embedding of approximate covariances for inference from Gaussian data on large lattices," *Journal of Computational and Graphical Statistics*, vol. 26, no. 1, pp. 88–97, 2017.

[28] H. Rue and L. Held, *Gaussian Markov Random Fields: Theory and Applications*. CRC press, 2005.

[29] 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*, vol. 73, no. 4, pp. 423–498, 2011.

[30] D. W. Nychka, S. Bandyopadhyay, D. Hammerling, F. Lindgren, and S. R. Sain, "A multiresolution Gaussian process model for the analysis of large spatial data sets," *Journal of Computational and Graphical Statistics*, vol. 24, no. 2, pp. 579–599, 2015.

[31] A. Vecchia, "Estimation and model identification for continuous spatial processes," *Journal of the Royal Statistical Society, Series B*, vol. 50, no. 2, pp. 297–312, 1988.

[32] M. L. Stein, Z. Chi, and L. Welty, "Approximating likelihoods for large spatial data sets," *Journal of the Royal Statistical Society: Series B*, vol. 66, no. 2, pp. 275–296, 2004.

[33] A. Datta, S. Banerjee, A. O. Finley, and A. E. Gelfand, "Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets," *Journal of the American Statistical Association*, vol. 111, no. 514, pp. 800–812, 2016.

[34] J. Guinness, "Permutation methods for sharpening Gaussian process approximations," *Technometrics*, vol. 60, no. 4, pp. 415–429, 2018.

[35] M. Katzfuss, J. Guinness, W. Gong, and D. Zilber, "Vecchia approximations of Gaussian-process predictions," *Journal of Agricultural, Biological, and Environmental Statistics*, vol. 25, no. 3, pp. 383–414, 2020.

[36] A. Sengupta and N. Cressie, "Hierarchical statistical modeling of big spatial datasets using the exponential family of distributions," *Spatial Statistics*, vol. 4, pp. 14–44, 2013.

[37] R. Sheth, Y. Wang, and R. Khardon, "Sparse variational inference for generalized Gaussian Process models," *Proceedings of the 32nd International Conference on Machine Learning*, vol. 37, 2015.

[38] Y. Yang, M. Pilanci, M. J. Wainwright, *et al.*, "Randomized sketches for kernels: Fast and optimal nonparametric regression," *The Annals of Statistics*, vol. 45, no. 3, pp. 991–1023, 2017.

[39] J. Hughes and M. Haran, "Dimension reduction and alleviation of confounding for spatial generalized linear mixed models," *Journal of the Royal Statistical Society: Series B*, vol. 75, no. 1, pp. 139–159, 2013.

[40] Y. Guan and M. Haran, "A computationally efficient projection-based approach for spatial generalized linear mixed models," *Journal of Computational and Graphical Statistics*, vol. 27, no. 4, pp. 701–714, 2018.

[41] H. Nickisch, A. Solin, and A. Grigorievskiy, "State-space Gaussian processes with non-Gaussian likelihood," *arXiv preprint arXiv:1802.04846*, 2018.

[42] J. R. Bradley, S. H. Holan, and C. K. Wikle, "Computationally efficient multivariate spatio-temporal models for high-dimensional count-valued data (with discussion)," *Bayesian Analysis*, vol. 13, no. 1, pp. 253–310, 2018.

[43] R. J. Lipton, D. J. Rose, and R. E. Tarjan, "Generalized nested dissection," *SIAM Journal on Numerical Analysis*, vol. 16, no. 2, pp. 346–358, 1979.

[44] E. Snelson and Z. Ghahramani, "Local and global sparse Gaussian process approximations," in *Artificial Intelligence and Statistics 11 (AISTATS)*, 2007.

[45] A. O. Finley, H. Sang, S. Banerjee, and A. E. Gelfand, "Improving the performance of predictive process modeling for large datasets," *Computational Statistics & Data Analysis*, vol. 53, pp. 2873–2884, jun 2009.

[46] H. Sang, M. Jun, and J. Z. Huang, "Covariance approximation for large multivariate spatial datasets with an application to multiple climate model errors," *Annals of Applied Statistics*, vol. 5, no. 4, pp. 2519–2548, 2011.

[47] M. Katzfuss, "A multi-resolution approximation for massive spatial datasets," *Journal of the American Statistical Association*, vol. 112, pp. 201–214, feb 2017.

[48] M. Katzfuss and W. Gong, "A class of multi-resolution approximations for large spatial datasets," *Statistica Sinica*, vol. accepted, 2019.

[49] M. Katzfuss, M. Jurek, D. Zilber, W. Gong, J. Guinness, J. Zhang, and F. Schäfer, *GPvecchia: Fast Gaussian-process inference using Vecchia approximations*, 2020. R package version 0.1.3.

[50] F. Schäfer, M. Katzfuss, and H. Owhadi, "Sparse Cholesky factorization by Kullback-Leibler minimization," *arXiv:2004.14455*, 2020.

[51] C. Varin, N. Reid, and D. Firth, "An overview of composite likelihood methods," *Statistica Sinica*, pp. 5–42, 2011.

[52] F. Schäfer, T. J. Sullivan, and H. Owhadi, "Compression, inversion, and approximate PCA of dense kernel matrices at near-linear computational complexity," *arXiv:1706.02205*, 2017.

[53] Y. Fong, H. Rue, and J. Wakefield, "Bayesian inference for generalized linear mixed models," *Biostatistics*, vol. 11, no. 3, pp. 397–412, 2010.

[54] N. E. Breslow and D. G. Clayton, "Approximate inference in generalized linear mixed models," *Journal of the American Statistical Association*, vol. 88, no. 421, pp. 9–25, 1993.

[55] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[56] T. Gneiting and M. Katzfuss, "Probabilistic forecasting," *Annual Review of Statistics and Its Application*, vol. 1, pp. 125–151, 1 2014.

[57] R. M. Neal *et al.*, "MCMC using Hamiltonian dynamics," *Handbook of Markov Chain Monte Carlo*, vol. 2, no. 11, 2011.

[58] P. J. Diggle, P. Moraga, B. Rowlingson, and B. M. Taylor, "Spatial and spatio-temporal log-Gaussian Cox processes: Extending the geostatistical paradigm," *Statistical Science*, vol. 28, no. 4, pp. 542–563, 2013.

[59] E. Borbas, P. Menzel, and B. C. Gao, "MODIS Atmosphere L2 Water Vapor Product," 2017.

[60] M. Katzfuss, J. Guinness, and E. Lawrence, "Scaled Vecchia approximation for fast computer-model emulation," *arXiv:2005.00386*, 2020.

[61] L. Zhang, A. Datta, and S. Banerjee, "Practical bayesian modeling and inference for massive spatial data sets on modest computing environments," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 12, no. 3, pp. 197–209, 2019.

[62] M. Jurek and M. Katzfuss, "Multi-resolution filters for massive spatio-temporal data," *arXiv:1810.04200*, 2018.

[63] M. Jurek and M. Katzfuss, "Hierarchical sparse Cholesky decomposition with applications to high-dimensional spatio-temporal filtering," *arXiv:2006.16901*, 2020.

[64] Space Studies Board, National Academies of Sciences, Engineering, and Medicine, and others, *Thriving on our changing planet: A decadal strategy for Earth observation from space*. National Academies Press, 2019.

[65] J. W. Chapman, D. R. Thompson, M. C. Helmlinger, B. D. Bue, R. O. Green, M. L. Eastwood, S. Geier, W. Olson-Duvall, and S. R. Lundeen, "Spectral and radiometric calibration of the next generation airborne visible infrared spectrometer (aviris-ng)," *Remote Sensing*, vol. 11, no. 18, p. 2129, 2019.

[66] G. Schaepman-Strub, M. E. Schaepman, T. H. Painter, S. Dangel, and J. V. Martonchik, "Reflectance quantities in optical remote sensing—definitions and case studies," *Remote sensing of environment*, vol. 103, no. 1, pp. 27–42, 2006.

[67] A. Berk, P. Conforti, R. Kennett, T. Perkins, F. Hawes, and J. Van Den Bosch, "Modtran® 6: A major upgrade of the modtran® radiative transfer code," in *2014 6th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, pp. 1–4, IEEE, 2014.

[68] C. Emde, R. Buras-Schnell, A. Kylling, B. Mayer, J. Gasteiger, U. Hamann, J. Kylling, B. Richter, C. Pause, T. Dowling, *et al.*, "The libradtran software package for radiative trans-

fer calculations (version 2.0. 1)," *Geoscientific Model Development*, vol. 9, no. 5, pp. 1647–1672, 2016.

[69] B.-C. Gao, M. J. Montes, R.-R. Li, H. M. Dierssen, and C. O. Davis, "An atmospheric correction algorithm for remote sensing of bright coastal waters using modis land and ocean channels in the solar spectral region," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 6, pp. 1835–1843, 2007.

[70] D. R. Thompson, B.-C. Gao, R. O. Green, D. A. Roberts, P. E. Dennison, and S. R. Lundeen, "Atmospheric correction for global mapping spectroscopy: Atrem advances for the hyspiri preparatory campaign," *Remote Sensing of Environment*, vol. 167, pp. 64–77, 2015.

[71] N. Cressie, *Statistics for Spatial Data, revised edition*. New York, NY: John Wiley & Sons, 1993.

[72] D. R. Thompson, B. H. Kahn, P. G. Brodrick, M. D. Lebsock, M. Richardson, and R. O. Green, "Spectroscopic imaging of sub-kilometer spatial structure in lower tropospheric water vapor," *Atmospheric Measurement Techniques*, 2021.

[73] C. D. Rodgers, "Retrieval of atmospheric temperature and composition from remote measurements of thermal radiation," *Reviews of Geophysics*, vol. 14, no. 4, pp. 609–624, 1976.

[74] P. Z. Mouroulis, "Spectral and spatial uniformity in pushbroom imaging spectrometers," in *Imaging Spectrometry V*, vol. 3753, pp. 133–141, International Society for Optics and Photonics, 1999.

[75] O. Dubovik, M. Herman, A. Holdak, T. Lapyonok, D. Tanré, J. Deuzé, F. Ducos, A. Sinyuk, and A. Lopatin, "Statistically optimized inversion algorithm for enhanced retrieval of aerosol properties from spectral multi-angle polarimetric satellite observations," *Atmospheric Measurement Techniques*, vol. 4, no. 5, pp. 975–1018, 2011.

[76] F. Xu, D. J. Diner, O. Dubovik, and Y. Schechner, "A correlated multi-pixel inversion approach for aerosol remote sensing," *Remote Sensing*, vol. 11, no. 7, 2019.

[77] J. Hobbs, M. Katzfuss, D. Zilber, J. Brynjarsdóttir, A. Mondal, and V. Berrocal, "Spatial retrievals of atmospheric carbon dioxide from satellite observations," *Remote Sensing*, vol. 13, p. 571, 2021.

[78] D. R. Thompson, A. Braverman, P. G. Brodrick, A. Candela, N. Carmon, R. N. Clark, D. Connelly, R. O. Green, R. F. Kokaly, L. Li, *et al.*, "Quantifying uncertainty for remote spectroscopy of surface composition," *Remote Sensing of Environment*, vol. 247, p. 111898, 2020.

[79] T. Gneiting and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," *Journal of the American Statistical Association*, vol. 102, pp. 359–378, 3 2007.

[80] A. E. Gelfand, P. Diggle, P. Guttorp, and M. Fuentes, *Handbook of spatial statistics*. CRC press, 2010.

[81] S. Banerjee, B. P. Carlin, and A. E. Gelfand, *Hierarchical modeling and analysis for spatial data*. Crc Press, 2014.

[82] M. C. Kennedy and A. O'Hagan, "Bayesian calibration of computer models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 63, no. 3, pp. 425–464, 2001.

[83] G. Wahba, *Spline models for observational data*. SIAM, 1990.

[84] A. Caponnetto and E. De Vito, "Optimal rates for the regularized least-squares algorithm," *Foundations of Computational Mathematics*, vol. 7, no. 3, pp. 331–368, 2007.

[85] A. W. van der Vaart, J. H. van Zanten, *et al.*, "Rates of contraction of posterior distributions based on gaussian process priors," *The Annals of Statistics*, vol. 36, no. 3, pp. 1435–1463, 2008.

[86] B. T. Knapik, A. W. Van Der Vaart, J. H. van Zanten, *et al.*, "Bayesian inverse problems with gaussian priors," *The Annals of Statistics*, vol. 39, no. 5, pp. 2626–2657, 2011.

[87] B. Szabó, A. W. Van Der Vaart, J. van Zanten, *et al.*, "Frequentist coverage of adaptive nonparametric bayesian credible sets," *The Annals of Statistics*, vol. 43, no. 4, pp. 1391–1428, 2015.

[88] A. Datta, S. Banerjee, A. O. Finley, and A. E. Gelfand, "Hierarchical nearest-neighbor gaussian process models for large geostatistical datasets," *Journal of the American Statistical Association*, vol. 111, no. 514, pp. 800–812, 2016.

[89] M. Katzfuss and J. Guinness, "A general framework for vecchia approximations of gaussian processes," *arXiv preprint arXiv:1708.06302*, 2017.

[90] 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*, vol. 24, no. 2, pp. 579–599, 2015.

[91] M. Katzfuss, "A multi-resolution approximation for massive spatial datasets," *Journal of the American Statistical Association*, vol. 112, no. 517, pp. 201–214, 2017.

[92] M. L. Stein, "Limitations on low rank approximations for covariance matrices of spatial data," *Spatial Statistics*, vol. 8, pp. 1–19, 2014.

[93] P. Drineas and M. W. Mahoney, "On the nyström method for approximating a gram matrix for improved kernel-based learning," *journal of machine learning research*, vol. 6, no. Dec, pp. 2153–2175, 2005.

[94] A. Gittens and M. Mahoney, "Revisiting the nystrom method for improved large-scale machine learning," in *International Conference on Machine Learning*, pp. 567–575, PMLR, 2013.

[95] S. T. Tokdar, "Adaptive gaussian predictive process approximation," *arXiv preprint arXiv:1108.0445*, 2011.

[96] A. Banerjee, D. B. Dunson, and S. T. Tokdar, "Efficient gaussian process regression for large datasets," *Biometrika*, vol. 100, no. 1, pp. 75–89, 2013.

[97] F. Bach, "Sharp analysis of low-rank kernel matrix approximations," in *Conference on Learning Theory*, pp. 185–209, 2013.

[98] B. Schölkopf, R. Herbrich, and A. J. Smola, "A generalized representer theorem," in *International conference on computational learning theory*, pp. 416–426, Springer, 2001.

[99] R. J. Adler and J. E. Taylor, *Random fields and geometry*. Springer Science & Business Media, 2009.

[100] R. M. Neal, *Bayesian learning for neural networks*, vol. 118. Springer Science & Business Media, 2012.

[101] A. Jacot, F. Gabriel, and C. Hongler, "Neural tangent kernel: Convergence and generalization in neural networks," *arXiv preprint arXiv:1806.07572*, 2018.

[102] D. Duvenaud, *Automatic model construction with Gaussian processes*. PhD thesis, University of Cambridge, 2014.

[103] R. B. Gramacy and H. K. H. Lee, "Bayesian treed gaussian process models with an application to computer modeling," *Journal of the American Statistical Association*, vol. 103, no. 483, pp. 1119–1130, 2008.

[104] A. M. Yaglom, *Correlation Theory of Stationary and Related Random Functions.*, vol. 526. Springer-Verlag, 1987.

[105] A. Wilson and R. Adams, "Gaussian process kernels for pattern discovery and extrapolation," in *International conference on machine learning*, pp. 1067–1075, PMLR, 2013.

[106] Y.-L. K. Samo and S. Roberts, "Generalized spectral kernels," *arXiv preprint arXiv:1506.02236*, 2015.

[107] M. Fuentes, "Spectral methods for nonstationary spatial processes," *Biometrika*, vol. 89, no. 1, pp. 197–210, 2002.

[108] P. D. Sampson and P. Guttorp, "Nonparametric estimation of nonstationary spatial covariance structure," *Journal of the American Statistical Association*, vol. 87, no. 417, pp. 108–119, 1992.

[109] A. Damianou and N. D. Lawrence, "Deep gaussian processes," in *Artificial intelligence and statistics*, pp. 207–215, PMLR, 2013.

[110] A. Zammit-Mangion, T. L. J. Ng, Q. Vu, and M. Filippone, "Deep compositional spatial models," *Journal of the American Statistical Association*, pp. 1–47, 2021.

[111] O. Kováčik and J. Rákosník, "On spaces $L^{p(x)}$ and $W^{k,p(x)}$," *Czechoslovak mathematical journal*, vol. 41, no. 4, pp. 592–618, 1991.

[112] L. Diening, P. Hästö, and A. Nekvinda, "Open problems in variable exponent lebesgue and sobolev spaces," *FSDONA04 proceedings*, pp. 38–58, 2004.

[113] L. Diening, P. Hästö, and S. Roudenko, "Function spaces of variable smoothness and integrability," *Journal of Functional Analysis*, vol. 256, no. 6, pp. 1731–1768, 2009.

[114] O. V. Lepski, E. Mammen, and V. G. Spokoiny, "Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors," *The Annals of Statistics*, pp. 929–947, 1997.

[115] A. W. van der Vaart, J. H. van Zanten, *et al.*, "Reproducing kernel hilbert spaces of gaussian priors," in *Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh*, pp. 200–222, Institute of Mathematical Statistics, 2008.

[116] S. Remes, M. Heinonen, and S. Kaski, "Non-stationary spectral kernels," *arXiv preprint arXiv:1705.08736*, 2017.

[117] P. McCullagh and J. A. Nelder, *Generalized Linear Models*. CRC press, 1989.

[118] D. Pollard *et al.*, "Asymptotics via empirical processes," *Statistical science*, vol. 4, no. 4, pp. 341–354, 1989.

APPENDIX A

A VECCHIA-LAPLACE APPROXIMATION FOR BIG NON-GAUSSIAN SPATIAL DATA

A.1 Newton-Raphson update using pseudo-data

The desired Newton-Raphson update has the form

$$\mathbf{h}(\mathbf{y}) = \mathbf{y} - \big(\tfrac{\partial^2}{\partial \mathbf{y}\mathbf{y}'} \log p(\mathbf{y}|\mathbf{z})\big)^{-1} \big(\tfrac{\partial}{\partial \mathbf{y}} \log p(\mathbf{y}|\mathbf{z})\big). \tag{A.1}$$

As shown in Section 2.2.2, we have $\frac{\partial}{\partial \mathbf{y}} \log p(\mathbf{y}|\mathbf{z}) = \mathbf{K}^{-1}(\boldsymbol{\mu} - \mathbf{y}) + \mathbf{u}_{\mathbf{y}}$ and $-\frac{\partial^2}{\partial \mathbf{y}\mathbf{y}'} \log p(\mathbf{y}|\mathbf{z}) = \mathbf{K}^{-1} + \mathbf{D}_{\mathbf{y}}^{-1} = \mathbf{W}_{\mathbf{y}}$. Using an idea similar to iterative weighted least squares [117, Section 2.5,], we can premultiply the variable $\mathbf{y}$ by the Hessian to combine terms, and then rearrange and pull out the prior mean. Dropping the iteration subscript of $\mathbf{y}$ for ease of notation, we can write (A.1) as

$$\begin{aligned}
\mathbf{h}(\mathbf{y}) &= \mathbf{y} + \mathbf{W}^{-1}(\mathbf{K}^{-1}(\boldsymbol{\mu} - \mathbf{y}) + \mathbf{u}) \\
&= \mathbf{W}^{-1}\big((\mathbf{K}^{-1} + \mathbf{D}^{-1})\mathbf{y} - \mathbf{K}^{-1}\mathbf{y} + (\mathbf{K}^{-1}\boldsymbol{\mu} + \mathbf{D}^{-1}\boldsymbol{\mu}) - \mathbf{D}^{-1}\boldsymbol{\mu} + \mathbf{D}^{-1}\mathbf{D}\mathbf{u}\big) \\
&= \boldsymbol{\mu} + \mathbf{W}^{-1}\big(\mathbf{D}^{-1}(\mathbf{y} + \mathbf{D}\mathbf{u} - \boldsymbol{\mu})\big) \\
&= \boldsymbol{\mu} + \mathbf{W}^{-1}\mathbf{D}^{-1}(\mathbf{t} - \boldsymbol{\mu}),
\end{aligned}$$

where $\mathbf{t} = \mathbf{y} + \mathbf{D}\mathbf{u}$.

Now consider the posterior mean in the case of a Gaussian likelihood $\mathbf{t}|\mathbf{y} \sim \mathcal{N}_n(\mathbf{y}, \mathbf{D})$ with a conjugate Gaussian prior, $\mathbf{y} \sim \mathcal{N}_n(\boldsymbol{\mu}, \mathbf{K})$. Employing a well-known formula, we have

$$\mathrm{E}(\mathbf{y}|\mathbf{t}) = (\mathbf{K}^{-1} + \mathbf{D}^{-1})^{-1}(\mathbf{K}^{-1}\boldsymbol{\mu} + \mathbf{D}^{-1}\mathbf{t}) = \boldsymbol{\mu} + \mathbf{W}^{-1}\mathbf{D}^{-1}(\mathbf{t} - \boldsymbol{\mu}).$$

Thus, we have $\mathbf{h}(\mathbf{y}) = \mathrm{E}(\mathbf{y}|\mathbf{t})$, the posterior mean under the assumption of Gaussian pseudo-data $\mathbf{t}$.

A.2 Computing U

Consider a general Vecchia approximation of the form (2.7). To obtain $\mathbf{U}$, define $C(x_i, x_j)$ as the covariance between x_i and x_j implied by the true model; that is, $C(y_i, y_j) = C(t_i, y_j) = K(\mathbf{s}_i, \mathbf{s}_j)$ and $C(t_i, t_j) = K(\mathbf{s}_i, \mathbf{s}_j) + \mathbb{1}_{i=j} d_i$. Then, the (j, i)th element of $\mathbf{U}$ can be calculated as

$$\mathbf{U}_{ji} = \begin{cases} r_i^{-1/2}, & i = j, \\ -b_i^{(j)} r_i^{-1/2}, & j \in c(i), \\ 0, & \text{otherwise}, \end{cases} \tag{A.2}$$

where $\mathbf{b}_i' = C(x_i, \mathbf{x}_{c(i)}) C(\mathbf{x}_{c(i)}, \mathbf{x}_{c(i)})^{-1}$, $r_i = C(x_i, x_i) - \mathbf{b}_i' C(\mathbf{x}_{c(i)}, x_i)$, and $b_i^{(j)}$ denotes the kth element of $\mathbf{b}_i$ if j is the kth element in $c(i)$ (i.e., $b_i^{(j)}$ is the element of $\mathbf{b}_i$ corresponding to x_j).

A.3 Vecchia-Laplace likelihood

We follow the approach of [3], replacing the Gaussian observation $\mathbf{z}$ with the pseudo-observation $\mathbf{t}$ and extending to the case $\boldsymbol{\mu} \neq \mathbf{0}$. Note first that $p(\mathbf{x})/p(\mathbf{y}|\mathbf{t}) = p(\mathbf{t})$. From Section 2.2.3, the density of $\mathbf{x}$ resulting from general Vecchia has the form $p(\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_x, \mathbf{Q}^{-1})$. The denominator is given by the posterior and has the form $p(\mathbf{y}|\mathbf{t}) = \mathcal{N}(\boldsymbol{\alpha}, \mathbf{W}_{\boldsymbol{\alpha}}^{-1})$. Thus,

$$\begin{aligned} \log p(\mathbf{t}) = &-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_x)' \mathbf{Q} (\mathbf{x} - \boldsymbol{\mu}_x) + \frac{1}{2} \log |\mathbf{Q}| - \frac{2n}{2} \log(2\pi) \\ &+ \frac{1}{2}(\mathbf{y} - \boldsymbol{\alpha})' \mathbf{W}_{\boldsymbol{\alpha}} (\mathbf{y} - \boldsymbol{\alpha}) - \frac{1}{2} \log |\mathbf{W}_{\boldsymbol{\alpha}}| + \frac{n}{2} \log(2\pi) \end{aligned}$$

By definition $\mathbf{Q} = \mathbf{U}\mathbf{U}'$, so the determinant simplifies according to $\log |\mathbf{Q}| = 2 \log |\mathbf{U}|$. Similarly, $\mathbf{W} = \mathbf{U}_y \mathbf{U}_y' = \mathbf{V}\mathbf{V}'$, where $\mathbf{V} = \text{chol}(\mathbf{W})$, so $\log |\mathbf{W}| = 2 \log |\mathbf{V}|$. Expanding the term $(\mathbf{x} -$

$\boldsymbol{\mu}_x)'\mathbf{Q}(\mathbf{x}-\boldsymbol{\mu}_x)$ yields:

$$
\begin{aligned}
(\mathbf{x}-\boldsymbol{\mu}_x)'\mathbf{Q}(\mathbf{x}-\boldsymbol{\mu}_x) &= \begin{bmatrix}\mathbf{y}-\boldsymbol{\mu}, & \mathbf{t}-\boldsymbol{\mu}_t\end{bmatrix}\begin{bmatrix}\mathbf{U}_y\mathbf{U}_y' & \mathbf{U}_y\mathbf{U}_t' \\ \mathbf{U}_t\mathbf{U}_y' & \mathbf{U}_t\mathbf{U}_t'\end{bmatrix}\begin{bmatrix}\mathbf{y}-\boldsymbol{\mu} \\ \mathbf{t}-\boldsymbol{\mu}_t\end{bmatrix} \\
&= (\mathbf{y}-\boldsymbol{\mu})'\mathbf{W}(\mathbf{y}-\boldsymbol{\mu}) + (\mathbf{t}-\boldsymbol{\mu}_t)'\mathbf{U}_t\mathbf{U}_t'(\mathbf{t}-\boldsymbol{\mu}_t) + 2(\mathbf{y}-\boldsymbol{\mu})'\mathbf{U}_y\mathbf{U}_t'(\mathbf{t}-\boldsymbol{\mu}_t),
\end{aligned}
$$

where $\boldsymbol{\mu}_t = \mathrm{E}(\mathbf{y}) + \mathbf{0} = \boldsymbol{\mu}$ according to our model for the pseudo-data. Subtracting the term $\mathbf{y}'\mathbf{W}\mathbf{y}$ that occurs in the denominator and plugging in for $\mathbf{y}$ the mode found via Vecchia-Laplace iterations, $\boldsymbol{\alpha} = \boldsymbol{\mu} - \mathbf{W}^{-1}\mathbf{U}_y\mathbf{U}_t'(\mathbf{t}-\boldsymbol{\mu})$, we are left with

$$
\begin{aligned}
\log p(\mathbf{t}) =& -\frac{1}{2}(\mathbf{t}-\boldsymbol{\mu}_t)'\mathbf{U}_t\mathbf{U}_t'(\mathbf{t}-\boldsymbol{\mu}) - (\mathbf{y}-\boldsymbol{\mu})'\mathbf{U}_y\mathbf{U}_t'(\mathbf{t}-\boldsymbol{\mu}) - \frac{1}{2}\boldsymbol{\mu}'\mathbf{W}\boldsymbol{\mu} \\
&+ \mathbf{y}'\big[\mathbf{W}\boldsymbol{\mu} - \mathbf{W}(\boldsymbol{\mu} - \mathbf{W}^{-1}\mathbf{U}_y\mathbf{U}_t'(\mathbf{t}-\boldsymbol{\mu}))\big] - \frac{n}{2}\log(2\pi) + \log|\mathbf{U}| - \log|\mathbf{V}| \\
&+ \frac{1}{2}(\boldsymbol{\mu} - \mathbf{W}^{-1}\mathbf{U}_y\mathbf{U}_t'(\mathbf{t}-\boldsymbol{\mu}))'\mathbf{W}(\boldsymbol{\mu} - \mathbf{W}^{-1}\mathbf{U}_y\mathbf{U}_t'(\mathbf{t}-\boldsymbol{\mu})) \\
=& -\frac{1}{2}(\mathbf{t}-\boldsymbol{\mu})'\mathbf{U}_t\mathbf{U}_t'(\mathbf{t}-\boldsymbol{\mu}) + \frac{1}{2}(\mathbf{t}-\boldsymbol{\mu})\mathbf{U}_t\mathbf{U}_y'\mathbf{W}^{-1}\mathbf{U}_y\mathbf{U}_t'(\mathbf{t}-\boldsymbol{\mu}) - (\mathbf{y}-\boldsymbol{\mu})'\mathbf{U}_y\mathbf{U}_t'(\mathbf{t}-\boldsymbol{\mu}) \\
&+ \frac{1}{2}\boldsymbol{\mu}'\mathbf{W}\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{\mu}'\mathbf{W}\boldsymbol{\mu} + \mathbf{y}'\big[\mathbf{U}_y\mathbf{U}_t'(\mathbf{t}-\boldsymbol{\mu})\big] - \boldsymbol{\mu}\mathbf{U}_y\mathbf{U}_t'(\mathbf{t}-\boldsymbol{\mu}) \\
&- \frac{n}{2}\log(2\pi) + \log|\mathbf{U}| - \log|\mathbf{V}| \\
=& -\frac{1}{2}(\mathbf{t}-\boldsymbol{\mu})'(\mathbf{U}_t\mathbf{U}_t' - \mathbf{U}_t\mathbf{U}_y'\mathbf{W}^{-1}\mathbf{U}_y\mathbf{U}_t')(\mathbf{t}-\boldsymbol{\mu}) - \frac{n}{2}\log(2\pi) + \log|\mathbf{U}| - \log|\mathbf{V}| \\
&+ (\mathbf{y}'-\boldsymbol{\mu})\mathbf{U}_y\mathbf{U}_t'(\mathbf{t}-\boldsymbol{\mu}) - (\mathbf{y}-\boldsymbol{\mu})'\mathbf{U}_y\mathbf{U}_t'(\mathbf{t}-\boldsymbol{\mu}) \\
=& -\frac{n}{2}\log(2\pi) + \log|\mathbf{U}| - \log|\mathbf{V}| \\
&- \frac{1}{2}(\mathbf{t}-\boldsymbol{\mu})'\mathbf{U}_t\mathbf{U}_t'(\mathbf{t}-\boldsymbol{\mu}) + \frac{1}{2}(\mathbf{t}-\boldsymbol{\mu})'\mathbf{U}_t\mathbf{U}_y'\mathbf{W}^{-1}\mathbf{U}_y\mathbf{U}_t'(\mathbf{t}-\boldsymbol{\mu}).
\end{aligned}
$$

A.4 Extended algorithmic example

Algorithm 5 provides pseudo-code for VL prediction and parameter estimation.

Algorithm 5 VL Prediction and Parameter Estimation

procedure PARAMETER ESTIMATION($\mathbf{z}, \mathcal{S}, g$)
  Define and initialize parameter vector, e.g., $\boldsymbol{\theta} = (\boldsymbol{\mu}', \nu, \rho, \sigma^2)'$
  Run VECCHIA-SPECIFY($\mathcal{S}, m$) with VL-IW to obtain $\texttt{VAO}_2$
  if dim = 1 **then**
    Set $\texttt{VAO}_1 = \texttt{VAO}_2$
  else
    Run VECCHIA-SPECIFY($\mathcal{S}, m$) with VL-RF to obtain $\texttt{VAO}_1$
  end if
  repeat
    Obtain new value of θ (e.g., using Nelder-Mead)
    Run VL-LIKELIHOOD($\mathbf{z}, \mathcal{S}, \texttt{VAO}_1, \texttt{VAO}_2, g, \boldsymbol{\mu}, K_{\boldsymbol{\theta}}$) to get $\mathcal{L}_{VL}(\boldsymbol{\theta})$
  until convergence
  return $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}$
end procedure

procedure VL-LIKELIHOOD($\mathbf{z}, \mathcal{S}, \texttt{VAO}_1, \texttt{VAO}_2, g, \boldsymbol{\mu}, K$)
  Run VL-INFERENCE using $\texttt{VAO}_1$ to obtain the posterior mode $\boldsymbol{\alpha}_V$ and pseudo-data $\mathbf{t}, \mathbf{D}$
  Evaluate $\mathcal{L}_{VL}(\boldsymbol{\theta})$ in Eqt. (2.11) using the data, $\boldsymbol{\theta}, \mathbf{t}, \mathbf{D}$, based on $\texttt{VAO}_2$
  return $\mathcal{L}_{VL}(\boldsymbol{\theta})$
end procedure

procedure VL-PREDICTION($\mathbf{z}, \mathcal{S}, \mathcal{S}^\star, g, \boldsymbol{\theta}$)
  Run VECCHIA-SPECIFY($\mathcal{S}, m, \mathcal{S}^\star$) to get $\texttt{VAO}$ (Use VL-RF if dim > 1)
  Run VL-INFERENCE with $\texttt{VAO}$ to obtain the posterior mode $\boldsymbol{\alpha}_V$ and pseudo-data $\mathbf{t}, \mathbf{D}$
  Perform latent prediction (Section 2.3.3) with $\boldsymbol{\alpha}_V, \mathbf{t}, \mathbf{D}$ to get $(\mathbf{y}^\star, \mathbf{y})|\mathbf{t} \sim N(\widetilde{\boldsymbol{\mu}}, (\widetilde{\mathbf{V}}\widetilde{\mathbf{V}}')^{-1})$
  If desired, obtain predictive summaries of $\mathbf{z}^\star$ by transforming samples of $\mathbf{y}^\star$ based on $g(z|y)$
  return Predictions and uncertainty measures of $\mathbf{y}^\star$ (and $\mathbf{z}^\star$)
end procedure

A.5 Details for comparison to Hamiltonian Monte Carlo (HMC)

A.5.1 HMC results

As described in Section 2.4.1, we simulated a single dataset consisting of $n = 625$ Bernoulli observations with $\nu = .5$, and compared Laplace and VL methods with $m = 10$ to HMC with path step size of $\epsilon = .001$ and a path step count (leapfrog iteration count) of $L = 50$. To account for finite computing resources and have a fair comparison to VL, we ran HMC for 8,000 iterations (the first 5,000 of which were considered burn-in) and repeated this 20 times to average out randomness. In an attempt to get a close approximation to the exact posterior, we also ran HMC for 1,000,000 iterations (burn-in: 10,000 iterations). The HMC samples were thinned by a factor of 10.

Method	Iterations (k)	Complexity	RMSE	CRPS	Time (s)
HMC	1,000,000	$\mathcal{O}(k(Ln + n^3))$	0.647	0.455	26,438.4
HMC	8,000	$\mathcal{O}(k(Ln + n^3))$	0.929	0.542	276.8
Laplace	<10	$\mathcal{O}(kn^3)$	0.639	0.452	1.2
VL-DL	<10	$\mathcal{O}(kn)$	0.639	0.452	0.1

Table A.1: Comparison to HMC for $n = 625$ simulated Bernoulli data

The comparison results are shown in Table A.1. Timings were acquired on a laptop, and the scores were only based on a single simulated dataset, so the table can only serve as a rough comparison. Even though HMC typically exhibits better mixing than Metropolis-Hastings sampling, HMC with $k = 8{,}000$ iterations was less accurate than Laplace-based methods despite being several orders of magnitude slower. Remarkably, even with 1,000,000 iterations, HMC did not achieve better scores than VL-DL. For larger n, the performance of HMC will likely degrade even further relative to VL, due to its cubic scaling in n for each iteration, and the increased number of required iterations for convergence.

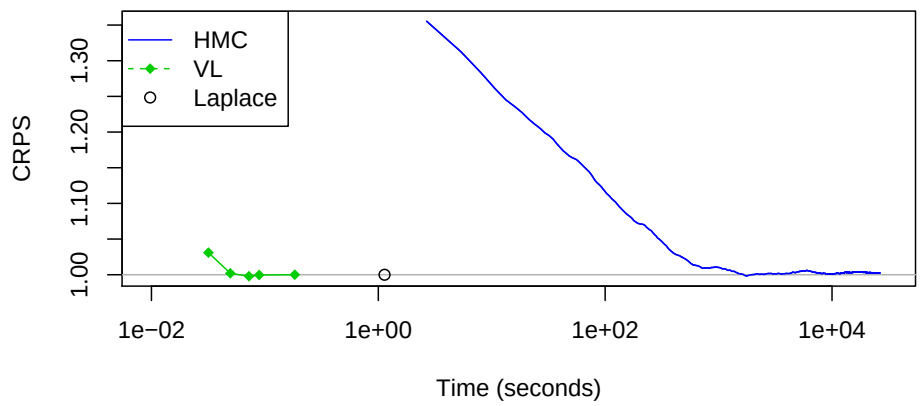


Figure A.1: CRPS (relative to Laplace's CRPS) versus time (on a log scale) for Bernoulli data of size $n = 625$. Laplace is run once until convergence. For VL, we considered $m \in \{1, 5, 10, 20, 40\}$. The number of HMC iterations varies from 10,100 to 1,000,000 in increments of 100, with the first 10,000 considered burn-in.

A.5.2 HMC CRPS comparison

The average continuous rank probability score (CRPS) [56, e.g.,] simultaneously considers calibration and sharpness of the posterior distribution at each location, and thus rewards accuracy of the posterior mean (like the RMSE) and accuracy of the uncertainty quantification. Figure A.1 repeats Figure 2.2 using CRPS (relative to Laplace's CRPS); the results are very similar.

A.5.3 HMC trace plots

Figure A.2 shows a set of the trace plots that result from running Hamiltonian Monte Carlo (HMC). The plots show the path taken by the variable in the specified position, so that the first plot shows 10th latent variable, etc. The plots show 300,000 iterations thinned by a factor of 10, assuming a burn-in of 5,000.

A.6 Additional comparisons between VL and LowRank

A.6.1 Additional simulations for 2D data

We repeated the 2D simulation results of Section 2.4.4 with a significantly larger range parameter, $\lambda = 0.2$ (instead of $\lambda = 0.05$). The results are shown in Figure A.3. For this increased

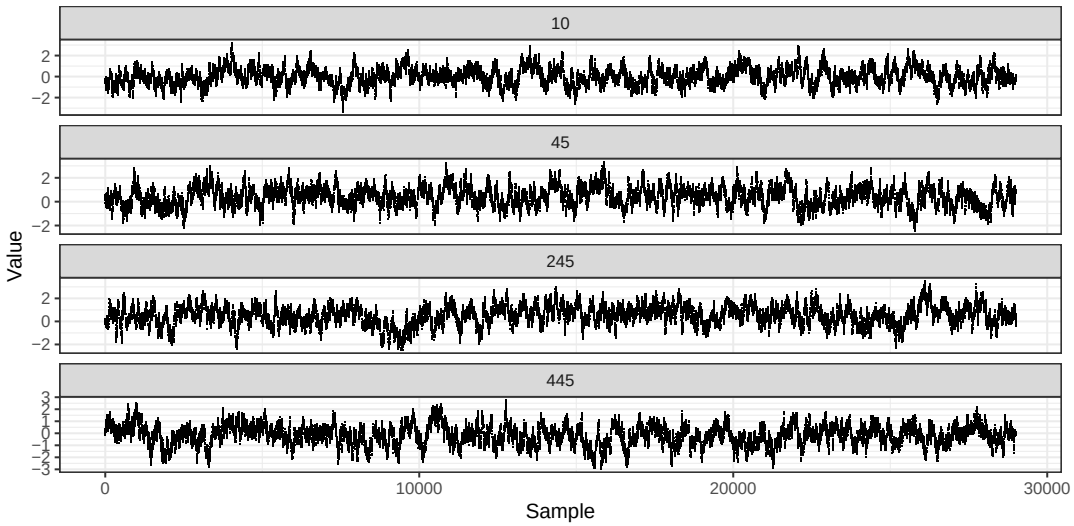


Figure A.2: Trace plot showing sample paths for the first 300,000 iterations for four latent variables for Hamiltonian Monte Carlo

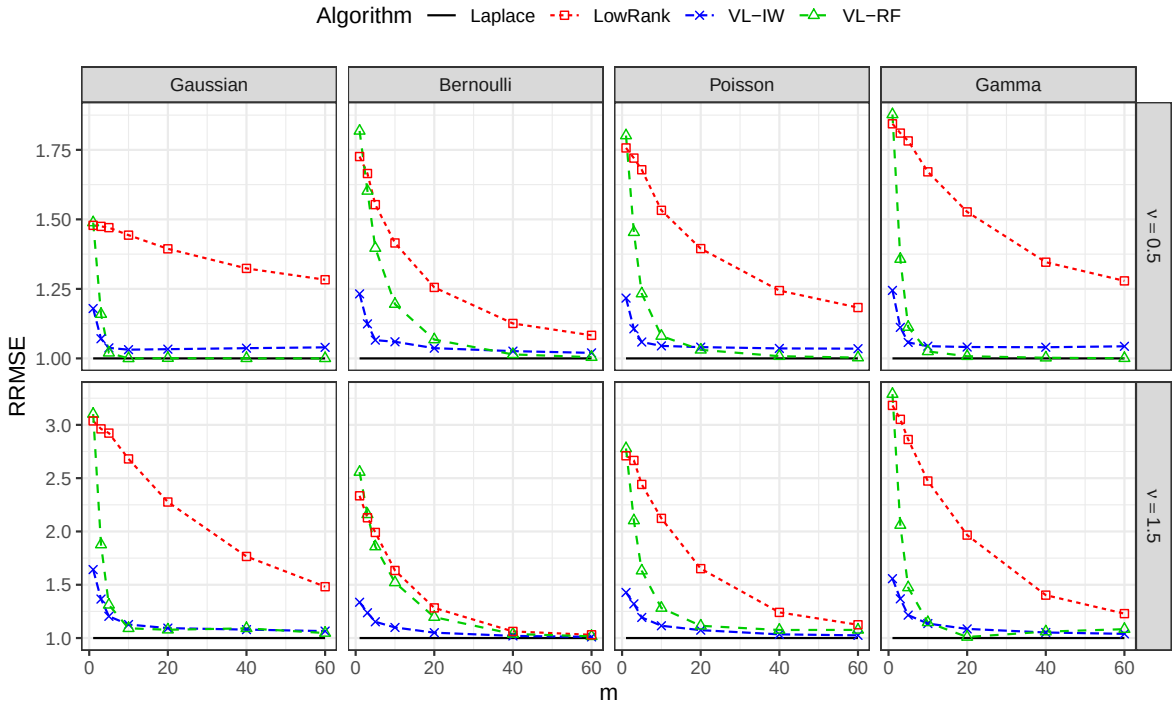
range parameter, the difference between VL and LowRank was less pronounced, but VL-RF was still substantially more accurate for all measures except for the RMSE for logistic regression with smoothness $\nu = 1.5$. The unstable behavior of the LowRank log score for smoothness $\nu = 1.5$ indicates that the parameters were close to the limits of machine precision.

A.6.2 Higher-dimensional simulations

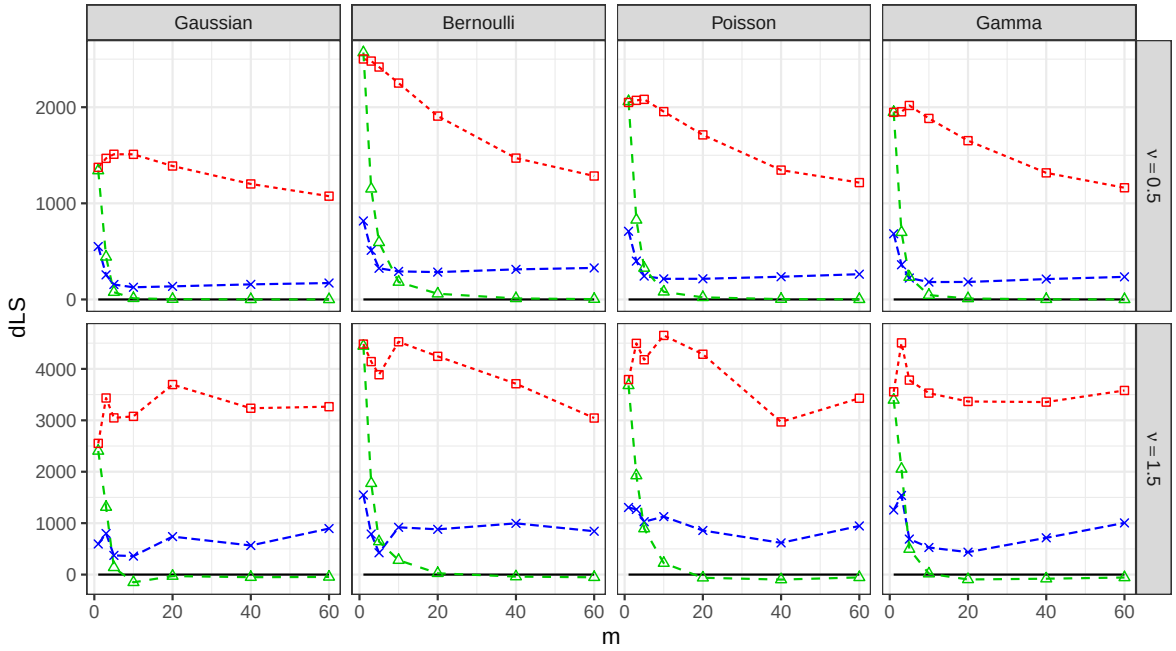
While we have focused on one- and two-dimensional space, we also briefly examined the performance of VL in three and four dimensions using simulation. In Figures A.4 and A.5 below, the sample size was $13^3 = 2197$ and $7^4 = 2401$ for the 3D and 4D demonstrations. Due to the relatively small number of points per axis, the range parameters were set to $\lambda_{3D} = .1$ and $\lambda_{4D} = .2$. The relative performance between VL and LowRank was similar to the 2D scenario, and we expect this to hold in higher dimensions as well.

A.7 Qualitative comparison of predictions in 1D

Here we present a few qualitative advantages of VL over LowRank. Figure A.6a demonstrates the visual difference between the approximations we compared. While the VL approximation was similar to the Laplace approximation, the low-rank approximation exhibited spikes that correspond



(a) RMSE (relative to Laplace)



(b) Difference in log score (relative to Laplace)

Figure A.3: Simulation results for $n = 2{,}500$ observations on a two-dimensional spatial domain with range $\lambda = 0.2$

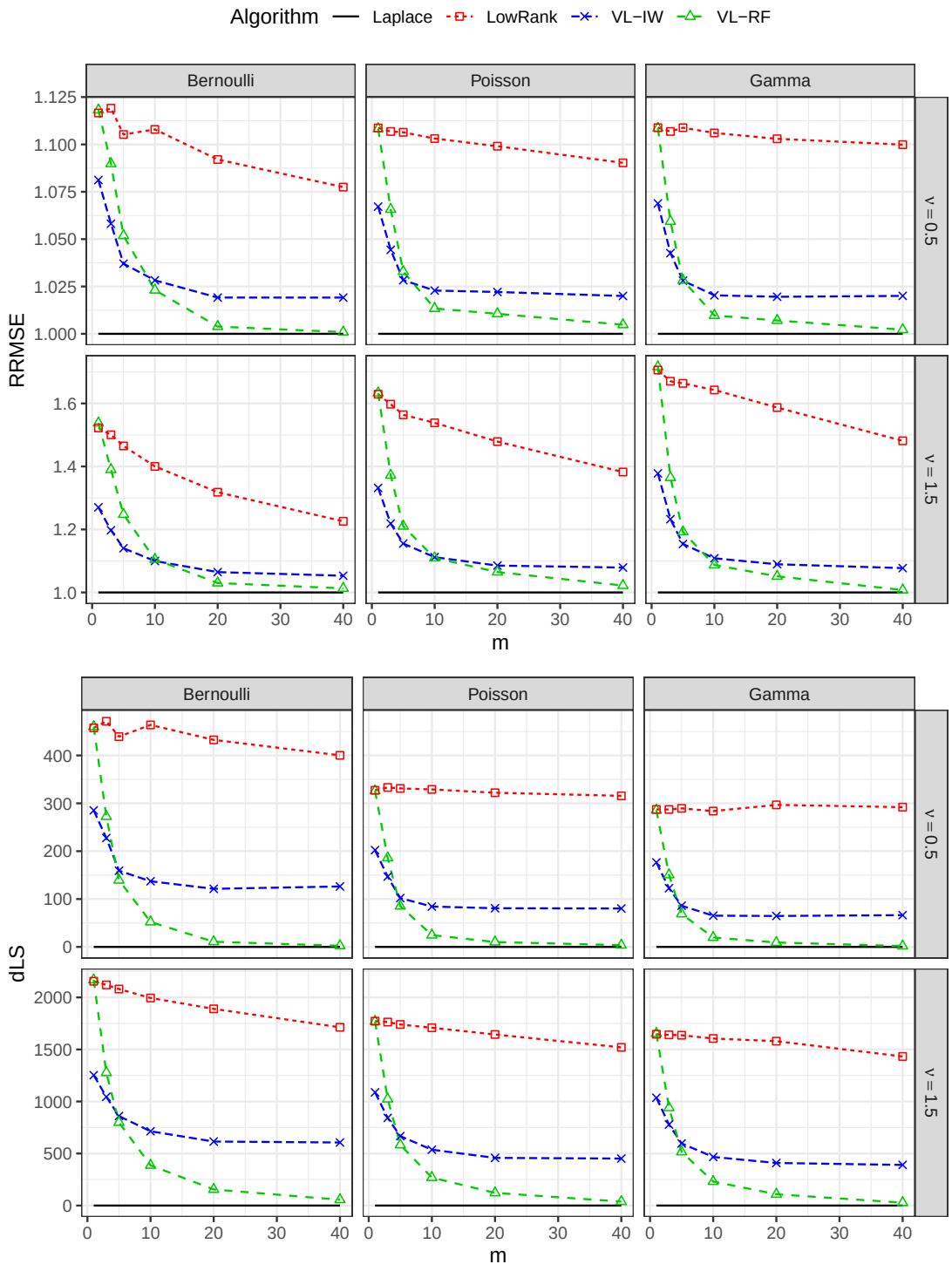


Figure A.4: Relative root mean square error (RRMSE) and Log Score difference from Laplace (dLS) for 3D data

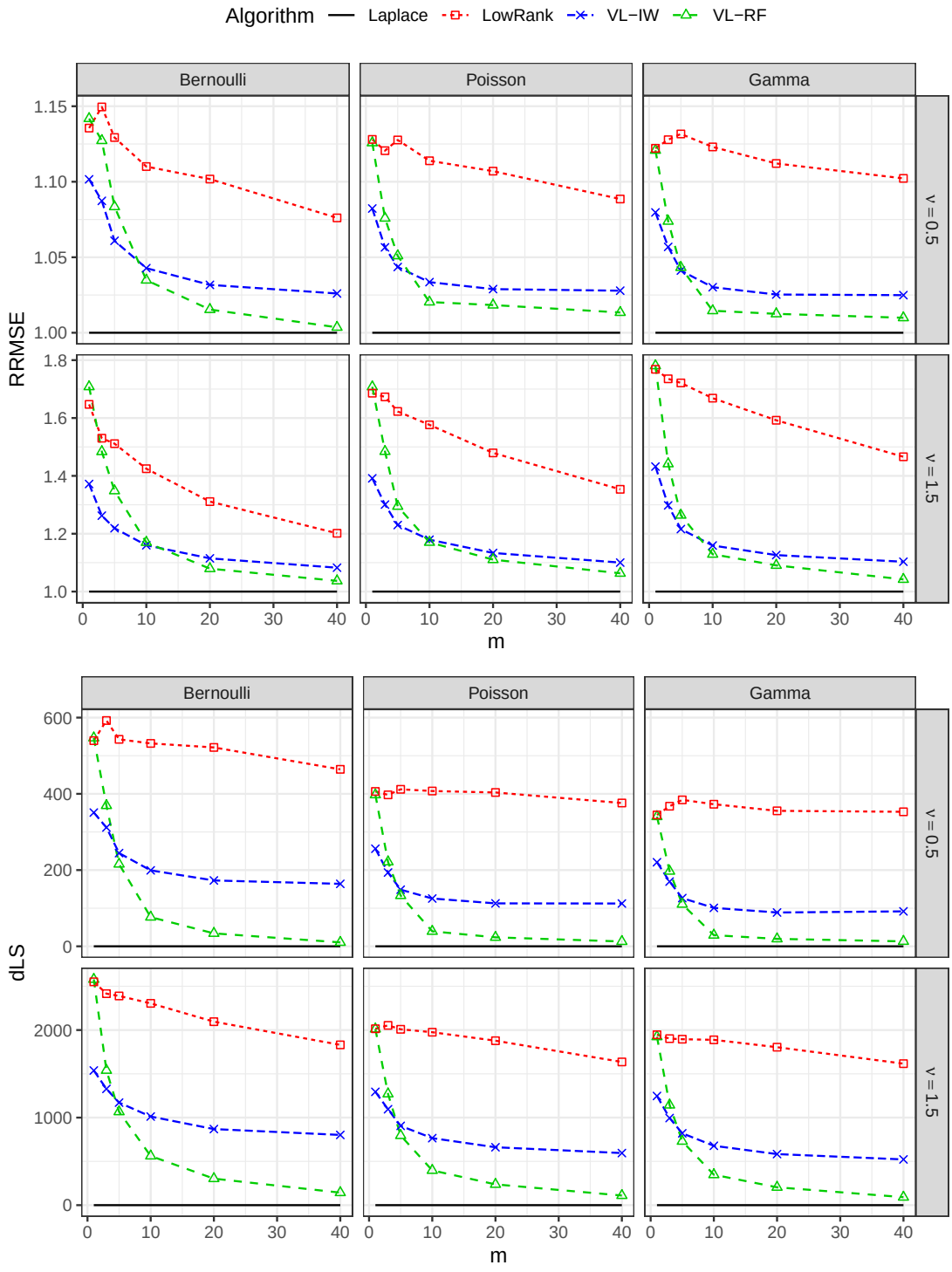
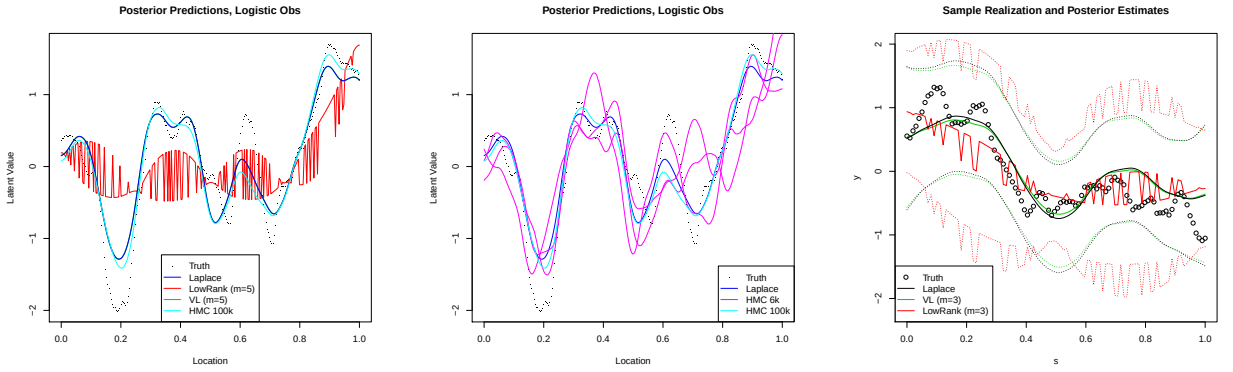


Figure A.5: Relative root mean square error (RRMSE) and Log Score difference from Laplace (dLS) for 4D data

to the correction terms of the modified predictive process. As a prediction location became far from the knots, the correction term increased up to the process variance at the location and resulted in artifacts, as shown in Figure A.6c.



(a) Summary comparison (b) HMC (c) MPP showing correction terms

Figure A.6: Comparison plots showing the posterior estimates of various methods

A.8 Parameter estimation for MODIS data

To apply the spatial Gamma GGP model described in Section 2.5, we needed to estimate several parameters: the Gamma shape parameter a, the trend parameter $\boldsymbol{\beta} = (\beta_1, \beta_2)'$, and the Matérn covariance parameters $\boldsymbol{\theta} = (\sigma^2, \rho, \nu)'$ determining the variance, range and smoothness. Simply estimating all parameters together based on the integrated likelihood was not possible due to identifiability issues.

Our parameter estimation procedure began by estimating the linear trend parameter $\boldsymbol{\beta}$. We temporarily ignored dependence in the residuals and essentially assumed a generalized linear model [117]. Thus, $\boldsymbol{\beta}$ was fitted with the standard technique of iteratively reweighted least squares using a subsample of 1,000,000 data points, yielding the estimated value $\boldsymbol{\beta} = (-1.515, 0.000766)'$.

Then, given $\boldsymbol{\beta}$, we carried out an iterative procedure in which we alternated between optimizing the covariance parameters $\boldsymbol{\theta}$ conditional on the shape parameter a, and vice versa. The covariance parameters $\boldsymbol{\theta}$ were obtained by maximizing the integrated VL likelihood from Section 2.3.2 via the

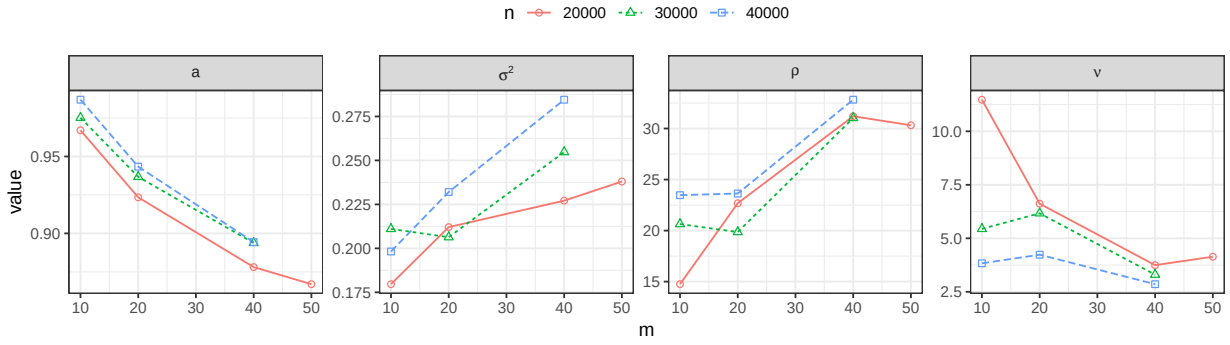


Figure A.7: Results of exploratory parameter estimation for the shape parameter a and covariance parameters (σ^2, ρ, ν) for variance, range, and smoothness. We concluded that $a = .89$, $\sigma^2 = .25$, $\rho = 31$km, and $\nu = 3$ were reasonable values.

Nelder-Mead algorithm, as described by Algorithm 5 in Section A.4. The shape parameter a was estimated by maximizing $p(\mathbf{z}|\mathbf{y} = \boldsymbol{\alpha}_V)$ with $\boldsymbol{\alpha}_V$ obtained using the VL Algorithm 1 based on $\boldsymbol{\beta}$ and the current estimate of $\boldsymbol{\theta}$. We believe that this approach can result in more accurate estimates of a relative to estimates obtained under the assumption of $\mathbf{y} = \boldsymbol{\mu}$ [36, e.g.].

While we found that three iterations of alternating between estimating $\boldsymbol{\theta}$ and a typically sufficed for convergence, in total this procedure still required hundreds of calls to the VL Algorithm 1, which could be quite time-consuming for large sample sizes. Hence, as shown in Figure A.7, we progressively increased the (subsampled) sample size n from 10,000 to 40,000 and conditioning set size m from 10 to 50 until the estimates started to converge. While the individual parameter estimates changed slightly as a function of m, Figure A.8 shows that the integrated VL likelihood for fixed $n = 250,000$ was virtually identical between $m = 20$ and $m = 40$. The integrated likelihood implied by LowRank was considerably worse.

Together, these results led us to conclude that $a = 0.89, \sigma^2 = .25, \rho = 31, \nu = 3$ were reasonable parameter values, and that $m = 20$ was adequate for VL in the prediction comparisons shown in Section 2.5.

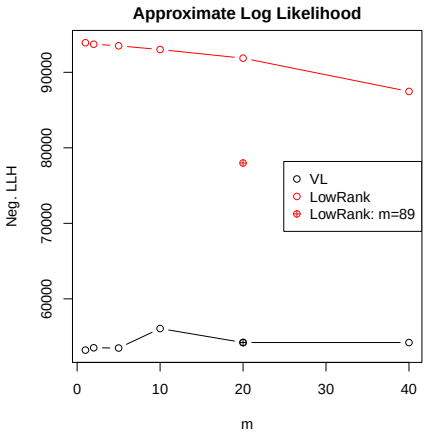


Figure A.8: The integrated VL likelihood was virtually constant between $m = 20$ and $m = 40$, while the LowRank likelihood varied greatly in comparison. The crossed points compare the likelihoods for the m values used in the application in Section 2.5.

APPENDIX B

SPATIAL SURFACE RETRIEVALS FOR VISIBLE/SHORTWAVE INFRARED REMOTE SENSING

B.1 Iterative optimization

Recall the cost function

$$Q(x) = -\log p(x|y) \propto \frac{1}{2}(x - x_a)S_a^{-1}(x - x_a) + \frac{1}{2}(y - f(x))S_\epsilon^{-1}(y - f(x)) + \text{constant.}$$

The gradient with respect to the state x is

$$\nabla_x Q = S_a^{-1}(x - x_a) - K_x^\top S_\epsilon^{-1}(y - f(x)), \tag{B.1}$$

where $K_x = \frac{\partial f(x)}{\partial x} = \frac{\partial y}{\partial x}$. The second derivative or Hessian could be calculated in an analogous way, but the forward model is expensive to differentiate. This problem is avoided by pre-computing a collection of Hessians $K_{x_{a,i}}$ on a set of $d(= 8)$ reference prior means $\{x_{a_i}\}_{i=1}^d$. These are then used with the Levenberg-Marquardt variant of Newton Raphson, which ignores higher order terms by using a linear approximation based on a Taylor expansion: $f(x) \approx f(x_0) + K_0(x - x_0) + R(||x - x_0||^2)$, where $K_0 = \frac{\partial f(x)}{\partial x}|_{x_0}$ and $x_0 \in \{x_{a,i}\}_{i=1}^d$. In effect, the Hessian is approximated by dropping higher-order derivative terms:

$$\nabla_x^2 Q = S_a^{-1} - {K'_x}^\top S_\epsilon^{-1}(y - f(x)) + K_x^\top S_\epsilon^{-1} K_x \approx S_a^{-1} + K_0^\top S_\epsilon^{-1} K_0. \tag{B.2}$$

Putting the gradient and Hessian together leads to an update step

$$x^{(\ell+1)} = x^{(\ell)} - [S_a^{-1} - K_\ell^\top S_\epsilon^{-1} K_\ell]^{-1}[S_a^{-1}(x^{(\ell)} - x_a) - K_\ell^\top S_\epsilon^{-1}(y - f(x^{(\ell)})], \tag{B.3}$$

where $K_\ell = \frac{\partial f(x)}{\partial x}|_{x^{(\ell)}}$. To improve computational stability with real data, the term S_a^{-1} in the

Hessian may be premultiplied by a factor $(I + \gamma)$ and this Hessian term is represented as α in Equation 3.6.

During the LM iterations, the Taylor expansion of the function f at iteration $\ell + 1$ is centered around $x^{(\ell)}$ of the previous iteration. In other words, the step to compute $x^{(\ell+1)}$ uses the approximation $f_{a(\ell)}(x^{(\ell)}) = f(x_{a(\ell)}) + K_{a(\ell)}(x^{(\ell)} - x_{a(\ell)})$ with the closest reference prior mean denoted by the $a(\ell)$. Then the likelihood using the approximation is written as

$$y|x \sim N\left(f(x_{a(\ell)}) + K_{a(\ell)}(x^{(\ell)} - x_{a(\ell)}), S_\epsilon\right).$$

Under this linear model, it is simple to show that each step of the LM algorithm is simply the posterior expectation: $x^{(\ell+1)} = E(x|x^{(\ell)}, y)$, where

$$\pi(x|y, x^{(\ell)}) \propto p(y|x, x^{(\ell)})p(x).$$

We abbreviate the approximated precision matrix with $P = [S_a^{-1} - K_\ell^\top S_\epsilon^{-1} K_\ell]$.

$$\begin{aligned}
x^{(\ell+1)} &= x^{(\ell)} - P^{-1}[S_a^{-1}(x^{(\ell)} - x_{a(\ell)} - K^\top S_\epsilon^{-1}(y - f_{a(\ell)}(x^{(\ell)})] \\
&= x^{(\ell)} - P^{-1}[S_a^{-1}(x^{(\ell)} - x_{a(\ell)}) - K^\top S_\epsilon^{-1}(y - f(x_{a(\ell)}) - K_{a(\ell)}(x^{(\ell)} - x_{a(\ell)}))] \\
&= P^{-1}[Px^{(\ell)} - S_a^{-1}(x^{(\ell)} - x_{a(\ell)}) + K^\top S_\epsilon^{-1}(y - f(x_{a(\ell)})) + K^\top S_\epsilon^{-1} K(x^{(\ell)} - x_{a(\ell)})] \\
&= P^{-1}[Px^{(\ell)} - Px^{(\ell)} + S_a^{-1} x_{a(\ell)} - K^\top S_\epsilon^{-1} K x_{a(\ell)} + K^\top S_\epsilon^{-1}(y - f(x_{a(\ell)}))] \\
&= P^{-1}[Px_{a(\ell)} + K^\top S_\epsilon^{-1}(y - f(x_{a(\ell)}))] \\
&= x_{a(\ell)} + P^{-1}[K^\top S_\epsilon^{-1}(y - f(x_{a(\ell)})] \\
&= E(x|x^{(\ell)}, y).
\end{aligned}$$

To get to the result shown in Section 3.2.2, we can use a Taylor expansion $f(x_{a(\ell)}) \approx f(x^{(\ell)} + K_\ell(x^{(\ell)} - x_{a(\ell)})$, which makes explicit the relation to $x^{(\ell)}$.

The important point of this calculation is to show that we can gain efficiency by choosing a

small collection of prior means $\{x_{a_i}\}_{i=1}^{d}$ for evaluating the forward model and its gradient.

B.2 Parameter estimation

We estimated Matérn smoothness and range parameters for water vapor using measurements collected by AVIRIS-NG over Desalpar in India on March 25, 2018 at roughly 7am. We used a cross-validation-type procedure in which we maximize the predictive likelihood (related to the log score) of a set of test points given a posterior computed from a set of training points. The data set was roughly 3000 by 500 pixels, so we used a training set lattice of 300 by 50 pixels (subsampling every tenth pixel) and a test set defined by offsetting the training set by five pixels. Given the cubic complexity when computing likelihoods for a Gaussian process, we utilize the `GPVecchia` package to perform efficient (linear in sample size, [3]) computation of the likelihood with a nearest-neighbor approximation.

The estimation of both range and smoothness parameters simultaneously was unstable, so we iteratively optimized the parameters one at a time until the change in each parameter value was less than a threshold, one percent in our case. Initializing the procedure with smoothness 1.5, range 5 (measured in pixels), with variance fixed to 1, and a nugget (representing noise) of 0.01, we converged to a range of 146.49 pixels and smoothness of 1.411. The pixel size for this data set is recorded as 5m, hence the range can be interpreted as about 750 meters. Since the Matern covariance has a very efficient form for smoothness values of 1.5, we rounded to that value for all computations. For simplicity, we assumed the aerosol field to have the same spatial covariance parameters.

APPENDIX C

FREQUENTIST COVERAGE FOR TRUNCATED KERNEL RIDGE REGRESSION

Unless specified otherwise, the Hilbert space norm is assumed to be with respect to the equivalent kernel, $\|\cdot\| = \|\cdot\|_\lambda$.

C.1 Proof for Theorem 4.3.1

Given the result of claim 1, the task is to show that the supremum norm bounds for $P_{\lambda,p}f^\star$ are sufficiently small for appropriate p. We recall from [6] that for the Hilbert space norm $\|\cdot\|$, we can use the Cauchy-Schwarz inequality to find

$$||P_\lambda f^\star||_\infty = \sup_x \langle P_\lambda f^\star, K_x\rangle \leq ||P_\lambda f^\star|| \sup_x ||K_x||,$$

Under the Fourier basis, $\sup_x ||K_x|| \lesssim \lambda^{-1/4\alpha}$ since $||K_x||^2 = K(x,x) \lesssim \sum_i \nu_i \asymp \lambda^{-1/2\alpha}$.

For the Holder case, note that the coefficients must satisfy $f_i < j^{-\alpha-1-\delta}$ for $\delta > 0$ for convergence in the sense $\sum j^\alpha |f_i| \leq \sum j^{-(1+\delta)} < \infty$. Then the tail has bound $\sum_{i=p+1}^\infty |f_i| \leq p^{-\alpha}/\alpha \leq B$ using an integration argument. Then if we assume $p \asymp \lambda^{-1/2\alpha} = h^{-1}$, we see

$$\begin{aligned}
||P_{\lambda,p}f^\star||_\infty \leq& \sum_{i=1}^p \frac{\lambda}{\mu_i+\lambda}|f_i| + \sum_{i=p+1}^\infty |f_i| \\
=& \sqrt{\lambda}\sum_{i=1}^p \frac{\sqrt{\lambda\mu_i}}{\mu_i+\lambda}\frac{|f_i|}{\sqrt{\mu_i}} + \sum_{i=p+1}^\infty |f_i| \\
\lesssim& \sqrt{\lambda}\sum_{i=1}^p i^\alpha |f_i| + p^{-\alpha} \\
\lesssim& \sqrt{\lambda}
\end{aligned}$$

In particular, note that a fixed p results in an error term $p^{-\alpha}$ that scales with the smoothness and prevents consistency.

For the Sobolev case, the coefficients must satisfy $f_j^2 \leq j^{-2a-1-\delta}$, so $\sum_{p+1}^\infty f_j^2 \leq p^{-2\alpha}/(2\alpha)$.

Then we bound the norm $||P_\lambda f^\star||^2$

$$\begin{aligned}
||P_{\lambda,p} f^\star||^2 = \langle P_{\lambda,p} f^\star, P_{\lambda,p} f^\star \rangle_\lambda =& \lambda \sum_{i=1}^{p} \frac{\lambda}{\mu_i + \lambda} \frac{f_i^2}{\mu_i} + \sum_{i=p+1}^{\infty} (1 + \lambda/\mu_i) f_i^2 \\
\leq& \lambda \sum_{i=1}^{\infty} \frac{f_i^2}{\mu_i} + \sum_{i=p+1}^{\infty} f_i^2 \\
\lesssim& \lambda \sum_{i=1}^{\infty} i^{2\alpha} f_i^2 + p^{-2\alpha} \\
\lesssim& \lambda
\end{aligned}$$

Plugging in $p \asymp \lambda^{-1/2\alpha}$ in the second to last line provides the appropriate scaling. We can then take the square root and combine with the bound $||K_x|| \lesssim \lambda^{-1/(4\alpha)}$ to reproduce

$$||P_\lambda f^\star||_\infty \lesssim \lambda^{1/2 - 1/4\alpha},$$

recovering the bound from corollary 2.3 of [6] upon plugging in the optimal value for h.

C.2 Proof for Theorem 4.3.2

Using the intermediate steps from theorem C.1, we plug in the alternate rate α_o for $p \asymp \lambda^{-1/(2\alpha_o)}$ for the two cases and then plug back in to the error terms of claim 1.

First, recall that the Sobolev case had a bias decomposition

$$||P_{\lambda,p} f^\star||_\infty \lesssim \lambda^{-1/4\alpha} \left(\lambda B^2 + p^{-2\alpha}\right)^{1/2}$$

We set $p \asymp \lambda^{-1/(2\alpha_o)}$ and plug the bias term into the error bound of claim 1, using the approximation $(1 + C\gamma_n) < 2$. We use the correct smoothness for all other terms, including $\lambda = h^{2\alpha}$, to get the result:

$$
\begin{aligned}
||\hat{f}^{(p)}_{n,\lambda} - f^\star||_\infty \leq & ||P_{\lambda,p} f^\star||_\infty + c\sigma\sqrt{\frac{\log n}{nh}} \\
\lesssim & h^{-1/4\alpha}\left(\lambda B^2 + p^{-2\alpha}\right)^{1/2} + \sigma\sqrt{\frac{\log n}{nh}} \\
\lesssim & h^{-1/2}\left(h^{2\alpha}B^2 + \lambda^{\alpha/\alpha_o}\right)^{1/2} + \sigma\sqrt{\frac{\log n}{nh}} \\
\lesssim & h^{-1/2}\left(h^{2\alpha}B^2 + h^{2\alpha^2/\alpha_o}\right)^{1/2} + \sigma\sqrt{\frac{\log n}{n}}h^{-1/2} \\
= & \mathcal{O}\left\{\left(\frac{\log n}{n}\right)^{-1/(4\alpha)}\left[\frac{\log n}{n} + \left(\frac{\log n}{n}\right)^{\alpha/\alpha_o}\right]^{1/2} + \left(\frac{\log n}{n}\right)^{1/2-1/(4\alpha)}\right\}
\end{aligned}
$$

When $\alpha/\alpha_o \geq 1$, the other terms dominate and the optimal rate is recovered; otherwise, the term $\left(\frac{\log n}{n}\right)^{\alpha/\alpha_o}$ dominates and collecting exponents leads to the result.

A similar proof is used for the Holder case, where $||P_{\lambda,p} f^\star||_\infty \leq \sqrt{\lambda}B + p^{-\alpha}$. Repeating the previous process with the optimal h, we get

$$
\begin{aligned}
||\hat{f}^{(p)}_{n,\lambda} - f^\star||_\infty \leq & ||P_{\lambda,p} f^\star||_\infty + c\sigma\sqrt{\frac{\log n}{nh}} \\
\leq & \sqrt{\lambda}B + p^{-\alpha} + \sigma\sqrt{\frac{\log n}{nh}} \\
\lesssim & h^\alpha + h^{\alpha^2/\alpha_o} + \sqrt{\frac{\log n}{n}}h^{-1/2} \\
= & \mathcal{O}\left\{\left(\frac{\log n}{n}\right)^{\frac{\alpha}{2\alpha+1}} + \left(\frac{\log n}{n}\right)^{\frac{\alpha^2}{\alpha_o(2\alpha+1)}} + \left(\frac{\log n}{n}\right)^{\frac{\alpha}{2\alpha+1}}\right\}
\end{aligned}
$$

As in the Sobolev case, $\alpha_o > \alpha$ leads to a dominating (ie smallest, and therefore slowest rate) exponent of $\frac{\alpha^2}{\alpha_o(2\alpha+1)}$.

C.3 Proof of risk bounds 4.3.3

C.3.1 Pointwise convergence

For both Sobolev and Holder class functions, the pointwise convergence depends on the mean squared error decomposition:

$$E[|f(x) - f^\star(x)|^2|D_n] = E[f(x) - f^\star(x)]^2 + V[f(x)] = [\hat{f}_n(x) - f^\star(x)]^2 + \widetilde{C}_n^B(x)$$

The uniform bounds developed earlier are directly applied:

$$E[|f(x) - f^\star(x)|^2|D_n] \lesssim \sigma^2 \frac{\log n}{nh} + ||P_{\lambda,p} f^\star||_\infty + \frac{C\sigma^2}{n\lambda} ||P_{\lambda,p} K_{p,x}||_\infty$$

For claim 1 and theorem 4.3.1, we showed that for sufficiently large p, the error bounds are the same as those for the full kernel. Specifically, $||P_{\lambda,p} f^\star||_\infty$ and $||P_{\lambda,p} K_{p,x}||_\infty$ are bounded as before, so the pointwise convergence bound for the original KRR holds for the truncated KRR, assuming the truncation scales following $p \asymp \lambda^{-1/(2\alpha)} = h^{-1}$ as specified.

C.3.2 Uniform convergence

To show supremum norm convergence for the truncated kernel problem, we need to bound the second moment of the supremum of the process, which splits into two terms:

$$E[||f(x) - f^\star(x)||_\infty^2|D_n] \leq 2||\hat{f}_{n,\lambda}^{(p)}(x) - f^\star(x)||_\infty^2 + 2E[||f - \hat{f}_{n,\lambda}^{(p)}||_\infty^2]$$

A truncated variance must increase the bias. We can use the results of section 4.3 to bound the bias assuming that p grows fast enough,

$$||\hat{f}_{n,\lambda}^{(p)}(x) - f^\star(x)||_\infty \leq ||P_\lambda f^\star||.$$

The second term is bounded using a chaining argument in [6] that relies on Dudley's inequality

[118, Thm 3.2], which gives the bound

$$||f - \hat{f}_{n,p}||_\infty = \sup_t \widetilde{f}_p^B(t) \leq C \int_0^1 [\log D(x, T, \rho)]^{1/2} dx.$$

In the equation above we have $f - \hat{f}_{n,p} = \widetilde{f}_p^B \sim GP(0, \widetilde{C}_{n,p}^B(x, \cdot))$ and the pseudo metric is taken to be $\rho_p(s,t) = \sqrt{V(\widetilde{f}_p^B(s) - \widetilde{f}_p^B(t)}$. The function $D(x, T, \rho_p)$ represents the packing number for balls of radius x measured with metric ρ_p over the domain $T = [0, 1]$ and is bounded by the covering number $N(x, T, \rho_p)$. Since the truncated kernel yields a smaller covariance, the truncated pseudo-metric will be smaller for points $s, s + \tau$ than the full rank pseudo-metric ρ as defined in the proof of Theorem 3.1 of [6]. Hence the result of their proof of Theorem 3.5, which provides a Euclidean norm bound for the pseudo-norm, still holds:

$$\rho_p(s, s + \tau) \leq \rho(s, s + \tau) \leq h^{-1} \frac{|\tau|^{1/2}}{n^{1/2}}.$$

In particular, the argument involves treating the posterior variance as the solution to a noiseless KRR problem with true data represented as the truncated kernel, so applying the truncated bias operator does not create any tail terms and the results carry over. The metric entropy $\log N(u, T, \rho_p)$ for the truncated pseudo-metric, which bounds the log packing number $\log D(x, T, \rho)$, is then bounded by the log covering number under the Euclidean norm $\log(n/u)$, since the unit ball under the stronger (smaller and truncated) metric will be larger than the unit ball under the Euclidean norm. In other words, the original Gaussian comparison inequality as derived remains valid. Then the entropy bound is the same and the supremum (uniform) norm bound holds for the truncated kernel when $p \asymp \lambda^{-1/2\alpha}$.

C.4 Proof of coverage

This proof follows the original proof in [6] almost verbatim under the assumption of $p \asymp \lambda^{-1/2\alpha}$, but we reproduce it here for convenience.

Recall the distribution

$$U_p = \sqrt{\frac{h}{n}} \sum_i^n w_i \widetilde{K}_{p,X_i} \sim GP(0, \hat{C}_{n,p})$$

where $\hat{C}_{n,p}(x, x') = E(U_p(x)U_p(x'))$. Then the Berry-Esseen theorem allows us to bound the Kolmogorov-Smirnoff distance as

$$||P[\hat{C}_{n,p}^{-1/2}(x,x)\sqrt{h/n}\sum_i^n w_i \widetilde{K}_{p,X_i}(x) \leq u] - \Phi(u)||_\infty \leq \frac{C}{\sqrt{nh}}$$

The next step is based on the higher-order result for the mean in claim 1, which implies that we have asymptotically $\sqrt{n/h}(\hat{f}_{n,p,\lambda} - F_{\lambda,p}f^\star) \sim GP(0, \hat{C}_{n,p})$. Then summarizing the higher order error term as

$$\delta_n = C'\gamma_n \left((1 + C\gamma_n)||P_{\lambda,p}f^\star||_\infty + C\sigma\sqrt{\frac{\log n}{nh}} \right)$$

we get the next step,

$$||P[\hat{C}_{n,p}^{-1/2}(x,x)\sqrt{nh}(f_{n,\lambda}^{(p)} - F_{\lambda,p}f^\star)) \leq u] - \Phi(u)||_\infty \leq C(\frac{1}{\sqrt{nh}} + \delta_n), \forall x \in \mathcal{X}.$$

Next we use equation 4.10 plus a Taylor expansion argument to bound the Kolmogorov distance for the credible intervals compared to the standard Gaussian cdf,

$$||P[\{\hat{C}_{n,p}^B(x,x)\}^{-1/2}\sqrt{nh}(f - f_{n,\lambda}^{(p)})) \leq u|\mathbb{D}_n] - \Phi(u)||_\infty \leq \gamma_n, \forall x \in \mathcal{X}. \tag{C.1}$$

We can use this and the properties of the inverse cdf Φ^{-1} to get a bound for the credible interval width by using another Taylor expansion argument:

$$\left| l_{n,p}(x;\beta) - \sqrt{\frac{\hat{C}_{n,p}^B(x,x)}{nh}} z_{(1+\beta)/2} \right| \leq C\sqrt{\frac{\hat{C}_{n,p}^B(x,x)}{nh}}\gamma_n.$$

Next we express the coverage in terms of the sample quantities,

$$
\begin{aligned}
&\mathbb{P}_\rho[f^*(x) \in CI_{n,p}(x;\beta)] \\
=&\mathbb{P}_\rho[-l_{n,p}(x;\beta) \le f_{n,\lambda}^{(p)} - f^*(x) \le l_{n,p}(x;\beta)] \\
=&\mathbb{P}_\rho[-\{\hat{C}_{n,p}(x,x)\}^{-1/2}\sqrt{nh}l_{n,p}(x;\beta) + \{\hat{C}_{n,p}(x,x)\}^{-1/2}\sqrt{nh}P_{\lambda,p}f^*(x) \\
&\le \{\hat{C}_{n,p}(x,x)\}^{-1/2}\sqrt{nh}(f_{n,\lambda}^{(p)} - F_{\lambda,p}f^*(x)) \\
&\le \{\hat{C}_{n,p}(x,x)\}^{-1/2}\sqrt{nh}l_{n,p}(x;\beta) + \{\hat{C}_{n,p}(x,x)\}^{-1/2}\sqrt{nh}P_{\lambda,p}f^*(x)]
\end{aligned}
$$

Now plugging in the approximation $\sqrt{\frac{\hat{C}_{n,p}^B(x,x)}{nh}}z_{(1+\beta)/2}$ for $l_{n,p}$, we get new terms

$$
\{\hat{C}_{n,p}(x,x)\}^{-1/2}\sqrt{nh}l_{n,p}(x;\beta) = \sqrt{\frac{\hat{C}_{n,p}^B(x,x)}{\hat{C}_{n,p}(x,x)}}z_{(1+\beta)/2}
$$

with a Kolmogorov distance

$$
\left|\mathbb{P}_\rho[f^*(x) \in CI_{n,p}(x;\beta)] - \Phi\left(\sqrt{\frac{\hat{C}_{n,p}^B(x,x)}{\hat{C}_{n,p}(x,x)}}z_{(1+\beta)/2} + \{\hat{C}_{n,p}(x,x)\}^{-1/2}\sqrt{nh}P_{\lambda,p}f^*(x)\right)\right.
$$

$$
\left.+\Phi\left(-\sqrt{\frac{\hat{C}_{n,p}^B(x,x)}{\hat{C}_{n,p}(x,x)}}z_{(1+\beta)/2} + \{\hat{C}_{n,p}(x,x)\}^{-1/2}\sqrt{nh}P_{\lambda,p}f^*(x)\right)\right| \le C(\frac{1}{\sqrt{nh}} + \gamma_n + \delta_n).
$$

This recovers the result of [6].

We remark that we can expand the higher order variance term through the kernel, and since $w_i \sim N(0,1)$ are iid, we get

$$
\hat{C}_n = E\left[(\frac{h}{n}\sum_{i=1}^n w_i \sum_{j=1}^\infty \phi_j(x_i)\phi_j(x)\nu_j)(\sum_{i=1}^n w_i \sum_{j=1}^\infty \phi_j(x_i)\phi_j(x')\nu_j\right]
$$

$$
= \sigma^2 h \sum_{j=1}^\infty \nu_j^2 \phi_j(x)\phi_j(x') = \sigma^2 h F_\lambda \widetilde{K}.
$$

With this observation, note that the coverage interval has an inflation term that can be written

$$\frac{\hat{C}_n^B}{\hat{C}_n} = \frac{\widetilde{K}}{F_\lambda \widetilde{K}} \implies \frac{\widetilde{K}_p}{F_\lambda \widetilde{K}_p}$$

Since the truncated kernel is smaller than the original, the ratio under truncation is closer to 1, implying the inflation of the interval width is slightly smaller than the full rank case of [6].

C.4.1 Loss of coverage with fixed truncation

A fixed truncation for the kernel represents a class of infinitely smooth functions when we assume the eigendecomposition uses the Fourier basis. Therefore, when the data generating function f^* is rough, there will be intervals over which the asymptotic coverage is 0. In other words, we can show that

$$P(f^* \in CI_{n,\beta}) \to 0$$

As described in the appendix section C.1, the bias term $P_{\lambda,p} f^*$ has an extra term corresponding to the tail, $p^{-\alpha}$ or $p^{-2\alpha}$, that does not decay when the truncation is fixed. The coverage result contains the bias term in two places: δ_n from the higher order error bound and directly in the comparison interval as the term $b_{n,p}$. Since we have a fixed bound $\|P_{\lambda,p} f^*\|_\infty \leq p^{-\alpha}$ rather than a limit to 0, we have

$$b_{n,p} \lesssim p^{-\alpha}\sqrt{nh} = \mathcal{O}(p^{-\alpha} \log n)$$

Hence it is possible that $b_{n,p} = \log np^{-\alpha} \to \infty$. In this case, we observe that

$$\Phi(u_{n,p}(x;\beta) + b_{n,p}(x)) - \Phi(-u_{n,p}(x;\beta) + b_{n,p}(x)) \to 0,$$

since $u_{n,p}(x;\beta)$ is bounded. Then the coverage result leads to the bound

$$\left| \mathbb{P}_\rho \big[f^*(x) \in CI_{n,p}(x;\beta) \big] \leq C p^{-\alpha} \gamma_n \to 0 \right. \tag{C.2}$$

In other words, as the posterior sample variance decreases, the relative difference between the true

function and the posterior mean increases until there is no overlap in credible interval.

C.5 Nominal coverage for truncated functions

For the confidence window $\sqrt{\frac{\hat{C}_n^B}{\hat{C}_n}} z_{\frac{1+\beta}{2}} + (\hat{C}_n)^{-1/2}\sqrt{nh} P_\lambda f^\star$, suppose the true function has smoothness α and can be represented in a finite number of terms; $f^\star = \sum_{i=1}^p f_i \phi_i$. If we use a kernel that is degenerate with similar expansion $\sum_{i=1}^p u_i \phi_i \phi_i$ and same smoothness, we find that the bias is of the form

$$|P_\lambda f^\star| = \lambda \sum_{j=1}^p \frac{j^{-2\alpha}}{\lambda - j^{-2\alpha}} \frac{|f_j^\star|}{j^{-2\alpha}} \lesssim \lambda^{1-1/2\alpha},$$

where the tail portion that was previously bounded with order $p^{-\alpha}$ drops out because the coefficients $f_i = 0$ for $i > p$. We are left with a term $\lambda^{1-1/2\alpha} \to 0$ as $n \to \infty$. The variance is still $\hat{C}_n$, which is used for normalizing. The inflation factor for the confidence level from section C.4 has the form

$$\frac{\hat{C}_{n,p}^B}{\hat{C}_{n,p}} = \frac{\widetilde{K}_p}{F_\lambda \widetilde{K}_p} = \frac{\sigma^2 h \sum_{j=1}^p \frac{1}{1+\lambda/\mu_j} \phi_j \phi_j}{\sigma^2 h \sum_{j=1}^p \frac{1}{(1+\lambda/\mu_j)^2} \phi_j \phi_j}$$

As $n \to \infty$, $\lambda \to 0$ so the kernel coefficients and scaling term approach 1. Hence the limiting coverage is nominal for the truncated kernel when the true function $f^\star$ is truncated to an equal (or lesser) degree and kernel smoothness. More generally, the kernel can have smoothness less than or equal to the true function and still recover nominal coverage.

We focus for a moment on the error term $P_\lambda f$. For an approximated kernel, this depends on smoothness it two ways: the inherent smoothness giving $\mu_j \asymp j^{-2\alpha}$ and the truncation rate $p \asymp \lambda^{-1/2\alpha}$. Suppose the truncation rate follows α_2 and the kernel smoothness follows α_1, with true smoothness α_0. Recalling equation 4.5 and assuming orthonormal Fourier eigenfunctions, we have

$$|P_\lambda f^\star| = \sum_{j=1}^p (1 - \nu_j)|f_j^\star| + \sum_{j=p+1}^\infty |f_j^\star|$$

$$\lesssim \lambda \sum_{j=1}^p \frac{1}{\lambda - \mu_j} \frac{\mu_j^\star}{\mu_j^\star} |f_j^\star| + p^{-\alpha_0}$$

$$\lesssim \lambda \sum_{j=1}^{p} \frac{j^{-2\alpha_0}}{\lambda - j^{-2\alpha_1}} \frac{|f_j^\star|}{j^{-2\alpha_0}} + \lambda^{\alpha_0/2\alpha_2}$$

Under smooth match or undersmoothed conditions, this bias term can be shown to go to 0 by studying the function $x \to \frac{x^{-2\alpha_0}}{\lambda - x^{-2\alpha_1}}$, showing the optima are finite valued or negative and replacing the infinite sum with the maxima times the function class bound B. For the truncated case, the additional tail term scales with λ and thus is not an issue.

APPENDIX D

CHARACTERIZATION FOR A NONSTATIONARY REPRODUCING KERNEL HILBERT SPACE

D.1 Spectral densities

Theorem D.1.1 (Spectral Density for convolution)**.** *For a stochastic process W defined in terms of a spectrum w_i, denote the spectral density for the covariance as $S(\cdot)$. For the process with correlated spectrum Lw described in Theorem 5.3.2, the corresponding spectral density for the covariance becomes*

$$S_{NS}(w_1, w_2) = \int_V L(w_1, v)L(w_2, v)S(v)dv$$

Proof. For stationary process $W = \sum_{i=1}^{\infty} Z_i h_i$, a realization at a point is expressed as $W_t = \sum_{i=1}^{\infty} Z_i b_t^*(h_i) = \sum_{i=1}^{\infty} Z_i h_i(t)$. We slightly modify the notation to get an integral form and then introduce the cumulative term $X(w)$

$$W_t = \int Z(w)h_t(w)dw = \int h_t(w)dX(w)$$

$$X(w) = \int_{-\infty}^{w} Z(v)dv$$

Then $Z(w)$ is the derivative $dX(w)/dw$, while $X(w)$ is an independent increment process. The covariance is then

$$\begin{aligned} E(W_t W_s) &= \int\int h_t(w_1)h_s(w_2)E(dX(w_1)dX(w_2)) \\ &= \int h_t(w)h_s(w)dF(w)) \\ &= \int h_t(w)h_s(w)S(w)dw \end{aligned} \tag{D.1}$$

The spectral density has been denoted $S(w)$. We now use the same procedure with the correlated version, $W = (LZ)^\top h$. We express the term LZ as an integral operator,

$$LZ(w) = \mathcal{Z}(w) = \int L(w,u)Z(u)du$$

As before, denote the cumulative term for the correlated $LZ = \mathcal{Z}$ as $\mathcal{X}$, so

$$\mathcal{X}(w) = \int_{-\infty}^{w} \mathcal{Z}(v)dv$$

We repeat the covariance calculation once again:

$$\begin{aligned}
E(W_tW_s) =& E\int_\Omega h_s(w_1)d\mathcal{X}(w)\int_\Omega h_t(w_2)d\mathcal{X}(w_2) \\
=& \int_\Omega\int_\Omega h_s(w_1)h_t(w_2)Ed^2\mathcal{X}(w_1)\mathcal{X}(w_2) \\
=& \int_\Omega\int_\Omega h_s(w_1)h_t(w_2)Ed^2\int_{-\infty}^{w_1}\mathcal{Z}(v)dv\int_{-\infty}^{w_2}\mathcal{Z}(v')dv' \\
=& \int_\Omega\int_\Omega h_s(w_1)h_t(w_2)d^2\int_{-\infty}^{w_1}\int_{-\infty}^{w_2}\int_V\int_{V'}L(u_1,v)L(u_2,v')E(X(v)X(v')dvdv')du_1du_2 \\
=& \int_\Omega\int_\Omega h_s(w_1)h_t(w_2)d^2\int_{-\infty}^{w_1}\int_{-\infty}^{w_2}\int_V L(u_1,v)L(u_2,v)f(v)dvdu_1du_2 \\
=& \int_\Omega\int_\Omega h_s(w_1)h_t(w_2)\left[\frac{d^2\int_V\int_{-\infty}^{w_1}\int_{-\infty}^{w_2}L(u_1,v)L(u_2,v)f(v)du_1du_2dv}{dw_1dw_2}\right]dw_1dw_2 \\
=& \int_\Omega\int_\Omega h_s(w_1)h_t(w_2)\left[\int_V L(w_1,v)L(w_2,v)f(v)dv\right]dw_1dw_2 \\
=& \int_\Omega\int_\Omega h_s(w_1)h_t(w_2)f_{NS}(w_1,w_2)dw_1dw_2
\end{aligned} \tag{D.2}$$

We recover the nonstationary spectral density given a stationary density and the "Cholesky" integral operators,

$$S_{NS}(w_1,w_2) = \int_V L(w_1,v)L(w_2,v)f(v)dv$$

□

D.2 Proof for Theorem 5.3.1

Proof. By the assumption of boundedness, there exists some M such that $\psi_i(x) < M$. Given series expansions for $f_i \in \mathcal{H}_i$ are integrable, the product is integrable since we can write

$$\int_X f_i(x)\psi_i(x)dx \leq M \int_X f_i(x)dx < \infty$$

So there exists a converging sequence that corresponds to the coefficients for the product. The inner product follows the usual definition extended to the multivariate form,

$$\begin{aligned}\langle f, g\rangle_{\mathcal{H}} &= \left\langle \begin{bmatrix} \sum_{j=1}^{\infty}(L_{\psi_1}\hat{f}_1)_j\phi_j \\ \vdots \\ \sum_{j=1}^{\infty}(L_{\psi_m}\hat{f}_2)_j\phi_j \end{bmatrix}, \begin{bmatrix} \sum_{j=1}^{\infty}(L_{\psi_1'}\hat{g}_1)_j\phi_j \\ \vdots \\ \sum_{j=1}^{\infty}(L_{\psi_m'}\hat{g}_2)_j\phi_j \end{bmatrix} \right\rangle \\ &= \sum_{i=1}^{m}\sum_{j=1}^{\infty}\frac{(L_{\psi_i}\hat{f}_i)_j(L_{\psi_i'}\hat{g}_i)_j}{u_{ij}}\end{aligned}$$

By the assumption that each function has bounded norm, we can use the Cauchy-Scwharz inequality to see that the inner product is bounded. The reproducing property is easy to see using the vector reproducing kernel:

$$
\begin{aligned}
\langle f, K_x \rangle &= \left\langle \begin{bmatrix} \sum_{j=1}^{\infty} (L_{\psi_1} \hat{f}_1)_j \phi_j \\ \vdots \\ \sum_{j=1}^{\infty} (L_{\psi_m} \hat{f}_m)_j \phi_j \end{bmatrix}, \begin{bmatrix} K_x^{(1)} \\ \vdots \\ K_x^{(m)} \end{bmatrix} \right\rangle \\
&= \sum_{i=1}^{m} \left\langle \sum_{j=1}^{\infty} (L_{\psi_1} w_1)_j h_j, K_x^{(1)} \right\rangle \\
&= \sum_{i=1}^{m} \sum_{j=1}^{\infty} (L_{\psi_1} w_1)_j h_j(x) \\
&= \sum_{i=1}^{m} \mathcal{F}(L_{\psi} \hat{f}_i)(x) \\
&= \sum_{i=1}^{m} \psi_i(x) f_i(x) = f(x)
\end{aligned}
\tag{D.3}
$$

□

D.3 Proof of Lemma 1

Proof. By theorem 2.1 of [115], a single Banach space RKHS $\mathcal{H}$ coincides with the stochastic process RKHS associated to W by the relation between the Pettis integral for projections and the reproducing kernel,

$$Sb^* = S\pi_t^* = K(t, \cdot).$$

The proof is based on the completeness of the span of projections π_t and the existence of a pseudo-measure for the stochastic process that induces total boundedness of the index space T containing t. If we fix a weighting function ψ, the two Hilbert spaces $\psi\mathcal{H}$ and ψW are still equivalent and the Pettis integral relation is

$$Sb_{\psi}^* = S[\psi \pi_t^*] = \psi(t) K(t, \cdot).$$

For a finite number of linear combination terms, the direct sum of spaces are equivalent by linearity of summation. The argument of [115] can now be repeated using this scaled Pettis integral and kernel relation. □

D.4 Proof for Theorem 5.3.2

Here the Hilbert space norm is represented as an L_2 norm in the Banach space. We remark that the case $s_{ij} = \delta_{ij}$ reduces to the stationary case. The case $s_{ij} = \delta_{ij}[\psi_1\sigma_{i1}, ..., \psi_k\sigma_{ik}]^\top$ reduces to the changepoint kernel mentioned earlier. We can also write

$$\mathcal{H} = \{f = \sum_{ij} w_j s_{ji} h_i : \|f\|_H = \|Lw\|_{L2} < \infty, Lw \in \ell_2\}, \tag{D.4}$$

recalling that Lw represents the "correlated" process coefficients while the process realization is expressed with elements of $[LL^\top]_{ij} = s_{ij}$.

Proof. Since LZ is assumed to be in ℓ_2, the proof mostly follows [115]. The Pettis integral now takes the form

$$\begin{aligned} S_L b^* = EWb^*W =& E[(LZ)^\top b^*(h)]^\top (LZ)^\top h \\ =& b^*(h)^\top LE(ZZ^\top)L^\top h \\ =& b^*(h)^\top LL^\top h \\ =& b^*(h)^\top \Sigma h = \sum b^*(h_j) s_{ij} h_i \end{aligned} \tag{D.5}$$

The inner product for the NS Hilbert space can be defined as the L2 norm of the Banach space.

$$\langle S_L b^*, S_L b^* \rangle_H = b^* S_L b^* = \sum_{ij} b^*(h_j) s_{ij} b^*(h_i) = E b^* W_L b^* W_L = \langle b^* W_L, b^* W_L \rangle_{L2} = \int b^*(W_{L,w})^2 dw \tag{D.6}$$

Defining $w_j = b^*(h_j)$, we recover the norm $\langle S_L b^*, S_L b^* \rangle_H = \sum_{i,j} w_j \sigma_{ij} w_i$. For two different functions $S_L b^*$ and $S_L \bar{b}^*$, we have

$$\left\langle S_L b^*, S_L \bar{b}^* \right\rangle_H = \sum_{i,j} w_j \sigma_{ij} \bar{w}_i$$

where $\bar{w}_i = \bar{b}^*(h_i)$. The reproducing function is also directly given by the Pettis integral when the

dual element is a projection, where for any function h_j, we have

$$\langle S_L b_t^*, h_j \rangle = b_t^*(h_j) = h_j(t) \tag{D.7}$$

□

D.5 Proof for Theorem 5.3.3

Proof. At any point x, the limit $m \to \infty$ converges if the series is Cauchy. Since the scaling functions are a resolution, the functions can be reordered by decreasing value for a particular x with new index order $r_x(i)$,

$$\{\psi_i(x)\} \to \{\psi_{r_x(i)}(x) : \psi_{r_x(i)}(x) > \psi_{r_x(i)+1}(x)\}.$$

Let the brackets $\{\cdot\}_{m1}^{m_2}$ represent a partial sum over the elements $m_1 : m_2$. We see that the Cauchy property holds at any x for the partial sum of mixture terms by the resolution assumption.

To show that $\|[\psi_{m_1} f_{m_1}, \ldots, \psi_{m_2} f_{m_2}]^\top\|_{\mathcal{H}} \to 0$ for $m_1 < m_2$ and $m_1, m_2 \to \infty$, we apply the definition of the operator norm by interpreting the mixture as an operator on the vector of kernels $[K_{m_1}(x, \cdot), \ldots, K_{m_2}(x, \cdot)]$. Let $b = \max_i f_i(x)$.

$$\begin{aligned} \|\{\psi_i f_i\}_{i=m_1}^{m_2}\|^2 =& \sup_{\|x\|} \frac{\sum_{i=m_1}^{m_2} \psi_i(x) f_i(x)}{\sum_{i=m_1}^{m_2} K_i(x, x)} \\ \leq& \frac{b \sum_{i=m_1}^{m_2} \psi_{r_x(i)}(x^*)}{\sum_{i=m_1}^{m_2} K_i(0)} \to 0 \end{aligned} \tag{D.8}$$

The first equality is the definition of operator norm. The left hand side is the result of applying the operator to the vector of kernels, which is the evaluation operation. The second line uses the assumption that the f_i's are bounded as members of their respective RKHS, and for any finite set we can specify a maximum b. The kernels are all stationary so the supremum does not apply. For the mixture functions, we assume that the ordering is retroactively optimized pointwise to take advantage of the resolution property. Hence, the supremum over x for the sum $m_1 : m_2$ is applied to an ordering $r_x(i)$ that has $\psi_{r_x(i)}(x)$ decreasing as $i \to \infty$. This way the supremum goes to 0 for

any x.

□

D.6 Proof for Theorem 5.3.4

Proof. To show that the Hilbert space is non-empty, we show the infinite series over the space index i satisfies the Cauchy property for sufficiently large N and $M > N$ and arbitrary ϵ,

$$\sum_{i=N}^{M} \sum_{j,k=1}^{\infty} s_{jk}^{(i)} w_{ij} w_{ik} \leq \epsilon.$$

First note that the summation $\sum j, k = 1^{\infty} s_{jk}^{(i)} w_{ij} w_{ik}$ can be expressed as a quadratic form, $w_i^\top S^{(i)} w_i = w_i^\top L_i L_i^\top w_i$. For each index i, the sum is bounded since the operator norm is bounded,

$$w_i^\top S^{(i)} w_i = \|L_i^\top w_i\|_2 \leq \|L_i\|_{op} \|w_i\|_2 < \infty$$

The w_i are L_2 integrable by assumption since they were taken as elements of their respective RKHS. To get to the Cauchy property, express the summation of quadratic forms as a vector product,

$$\sum_{i=N}^{M} w_i^\top S^{(i)} w_i = \left\| \begin{bmatrix} L_N^\top w_N \\ \vdots \\ L_M^\top w_M \end{bmatrix} \right\|_2 \leq \left\| \begin{bmatrix} L_N^\top & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & L_M^\top \end{bmatrix} \right\|_{op} \left\| \begin{bmatrix} w_N \\ \vdots \\ w_M \end{bmatrix} \right\|_2 < \infty.$$

The inequality holds because, for arbitrary ϵ, we can choose N sufficiently large such that

$$\left\| \begin{bmatrix} L_N^\top & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & L_M^\top \end{bmatrix} \right\|_{op} \leq \sum_{i=N}^{M} \|L_i\|_{op} \leq \epsilon$$

Hence the elements of the Hilbert space can be expressed as converging series and the space is not empty. □

D.7 Decay of kernel and relation to scaling

Applying the L operator to the independent increment process Z is essentially a convolution. Therefore it is equivalent to applying some scaling function to the process itself, denote $\mathcal{F}$ as a Fourier transform:

$$l(x)W(x) = \mathcal{F}(L * Z)$$

If L is the transform of a function l that concentrates near 0, then it is expected that the convolution operation will lead to a kernel that decays. This is precisely why a spectral mixture is introduced, [106, 105, 107]. If we use a collection of processes, $W_1, ..., W_k$ and allow each to be scaled by a different function $l_1(x), ..., l_k(x)$, the resulting increment process is now a sum of convolutions,

$$\sum_{1=1}^{k} (L_i * Z_i)(w)$$

and the covariance is a sum

$$\sum_{1=1}^{k} L_i(w_1)L_i(w_2) * Z_i$$

$$= \sum_{i=1}^{k} \int_V L_i(w_1, v)L_i(w_2, v)f_i(v)dv$$

Hence it is advantageous to use multiple functions l_i where a single function would have a small effective support.

We can see this another way by rearranging terms in the covariance using Theorem 5.3.2.

$$\begin{aligned} C(s,t) &= \int_{\Omega_1}\int_{\Omega_2} e^{i(w_1 s - w_2 t)} S(w_1, w_2) dw_1 dw_2 \\ &= \int_{\Omega_1}\int_{\Omega_2} e^{i(w_1 s - w_2 t)} \left[\int_V L(w_1, v)S(v)L(v, w_2)dv\right] dw_1 dw_2 \\ &= \int_V \left[\int_{\Omega_1} e^{iw_1 s} L(w_1, v)dw_1\right] S(v) \left[\int_{\Omega_2} e^{-iw_2 t} L(w_2, v)dw_2\right] dv \end{aligned} \tag{D.9}$$

The term $\int_{\Omega_1} e^{iw_1 s} L(w_1, v) dw_1$ is a transform and we can represent the kernel as

$$C(s,t) = \int_V \hat{L}(s,v) S(v) \hat{L}(v,t) dv \tag{D.10}$$

This representation makes it more clear why the covariance will decay for fixed $|s-t|$ as $s, t \to \infty$. If for example L is Gaussian, the transform $\hat{L}$ will also be Gaussian. Interpreting $\hat{L}(s,v)$ as a Gaussian centered at s, the largest values of the spectrum $S(v)$, those near 0, will be lost as $s \to \infty$ and the mass of $\hat{L}(s,v)$ moves away from the origin.