Modeling Latent Correlation Structures with Application to Agricultural and Environmental Science

by

Luke Bornn

B.Sc. Mathematics and Statistics, University of the Fraser Valley, 2006

M.Sc. Statistics, The University of British Columbia, 2008

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

in

THE FACULTY OF GRADUATE STUDIES

(Statistics)

The University Of British Columbia

(Vancouver)

July 2012

Abstract

In this thesis, we explore the issue of latent correlation structure in spatial and other correlated systems. Firstly, we propose a class of prior distributions on decomposable graphs, allowing for improved modeling flexibility. While existing methods solely penalize the number of edges, the proposed work empowers practitioners to control clustering, level of separation, and other features of the graph. Emphasis is placed on a particular prior distribution which derives its motivation from the class of product partition models; the properties of this prior relative to existing priors is examined through theory and simulation. We then demonstrate the use of graphical models in the field of agriculture, showing how the proposed prior distribution alleviates the inflexibility of previous approaches in properly modeling the interactions between the yield of different crop varieties.

Secondly, we describe how spatial dependence can be incorporated into statistical models for crop yield along with the dangers of ignoring it. In particular, approaches that ignore this dependence suffer in their ability to capture (and predict) the underlying phenomena. Prior distributions are developed to accommodate the spatial non-stationarity arising from distinct between-region differences in agricultural policy and practice. As a result, the model developed has improved

prediction performance relative to existing models, and allows for straightforward interpretation of climatic effects on the model's output.

Lastly, we propose a novel approach to modeling nonstationary spatial fields. The proposed method works by expanding the geographic plane over which these processes evolve into higher dimensional spaces, transforming and clarifying complex patterns in the physical plane. By combining aspects of multi-dimensional scaling, group lasso, and latent variable models, a dimensionally sparse projection is found in which the originally nonstationary field exhibits stationarity. Following a comparison with existing methods in a simulated environment, dimension expansion is studied on a classic test-bed data set historically used to study nonstationary models. Following this, we explore the use of dimension expansion in modeling air pollution in the United Kingdom, a process known to be strongly influenced by rural/urban effects, amongst others, which gives rise to a nonstationary field.

Preface

A version of Chapter 2 has been published. Bornn, L., Caron, F. (2011) Bayesian Clustering in Decomposable Graphs. Bayesian Analysis. Vol. 6, No. 4, Pages 829-846. I conducted all computation and numerical work, and hence generation of figures. The idea was jointly developed by myself and Dr. Caron, and I conducted the majority of writing and editing.

A version of Chapter 4 has been published. Bornn, L., Zidek, J. (2012) Efficient Stabilization of Crop Yield Prediction in the Canadian Prairies. Agricultural and Forest Meteorology. Vol. 152, Pages 223-232. I conducted all computation and numerical work, and hence generation of figures. The idea was jointly developed by myself and Dr. Zidek, and I conducted the majority of writing and editing.

A version of Chapter 5 has been published. Bornn, L., Shaddick, G., Zidek, J. (2012) Modelling Nonstationary Processes Through Dimension Expansion. Journal of the American Statistical Association. Vol. 107, Pages 281-289. I conducted all computation and numerical work, and hence generation of figures. The idea was jointly developed by myself, Dr. Shaddick, and Dr. Zidek, and I conducted the majority of writing and editing.

Table of Contents

List of Tables

List of Figures

Acknowledgments

I gratefully acknowledge the tremendous support of Arnaud Doucet, Raphael Gottardo, and Jim Zidek, who as supervisors over my time at UBC filled me with an insatiable passion for research and a discerning eye towards quality. I am forever in debt for their continual mentorship, insights, and wisdom in helping me establish an independent research programme and lay the groundwork for an academic career. I am honored to have been able to work with them, and am proud to call them friends.

I had the tremendous pleasure to work with Marian Anghel, Sezer Atamturktur, François Caron, Julien Cornebise, Pierre Del Moral, Chuck Farrar, Todd Graves, Dave Higdon, Roman Holenstein, Pierre Jacob, Gyuhae Park, Gavin Shaddick, Ingo Steinwart, and Aline Tabet throughout my PhD studies. The friendships and collaborations developed with this fine group have not only improved my research, but also enriched my life. I also thank my supervisory committee, Claudia Tebaldi, Paul Gustafson, and Alexandre Bouchard-Côté, without whom you would not be reading this.

I wish to thank Charles Serele, Harvey Hill, and Nathaniel Newlands, of Agriculture and Agri-foods Canada for providing the data used in Chapter 4 as well as

useful suggestions throughout the chapter's development. Additionally, we thank numerous referees and editors for valuable input in the published works which drive this thesis. This work was funded by an NSERC PGS-D as well as a Michael Smith Research Trainee Award.

Most importantly, I am forever grateful to my family for their unceasing love and support.

Dedication

To Katie

Chapter 1

Introduction

While introductory statistics courses often begin with the topic of summarizing and visualizing single variables, attention inevitably turns to discussion of the relationship between two or more variables. Starting with correlation, focus quickly shifts to regression, classification, and other techniques. All of these tasks, however, rely on having several variables which are related in such a way that we can express their relationship mathematically. Take for example weight and height. If we are told a person's height we can not guess their weight exactly, but it allows us to better estimate a person's weight. In this way we say that while there is no perfect formula relating height to weight, the variables are correlated. Formally, we define the correlation (or more formally, Pearson's correlation coefficient) between two random variables X and Y with expected values μ_X and μ_Y and standard deviations σ_X and σ_Y, respectively, as:

$$\rho_{X,Y} = corr(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E(X-\mu_X)E(Y-\mu_Y)}{\sigma_X \sigma_Y}.$$

Intuitively, if high values of X correspond to high values of Y, the two variables will be highly correlated, with a value close to 1. If high values of X correspond to low values of Y, the two variables will be negatively correlated, with a value close to -1. If there is no relation between values of X and Y, the correlation coefficient will be 0.

We may calculate an empirical version of Pearson's correlation coefficient, often notated r, by substituting empirical means and standard deviations for the population versions above.

$$r_{X,Y} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{s_X} \right) \left(\frac{Y_i - \bar{Y}}{s_Y}. \right)$$

Correlation is closely related to the concept of independence, which intuitively means that the outcome of one event does not influence the outcome of a second event. In fact, if two random variables X and Y are independent, then $\rho_{X,Y} = 0$. Note that while independence implies zero correlation, the converse is not true, as Pearson's correlation coefficient only captures linear relationships. While the explicit definitions above may seem overly pedantic, they are worth stating simply for the realization that a large portion of methodological development in the field of statistics (and this thesis!) is focused on this simple notion of correlation, extended to complex systems.

When dealing with data which arrives sequentially in time, or across a spatial domain, observations will not be independent. For example, the average temperature tomorrow is largely influenced by the temperature today. Likewise in space, we expect the amount of precipitation which falls at the University of British Columbia to be related to the amount which falls in downtown Vancouver, as the

two locations are only $10km$ apart. As such, observations which are obtained spatially or temporally must be modeled appropriately, as they will not satisfy the independence assumption required by many methods such as standard linear regression.

1.1 Structure in Correlated Systems

The way in which correlation arises in space and time is structured, a phenomena which is intuitively obvious. Using again the example of average temperature, the value tomorrow is likely to be largely influenced by the value today. However, the temperature tomorrow is likely less influenced by the temperature a year earlier. In this way, we intuitively understand that the correlation between values in time decreases as more time separates the observations. In contrast, sometimes patterns are cyclical. For example, whether a house is decorated with Christmas lights is uncorrelated with whether that house had Christmas lights a month earlier. However, it would be highly correlated with whether the house had Christmas lights 365 days earlier. Similarly, whether a person attends a church service on Sunday is only weakly correlated with whether that person attended church the day before, on Saturday. A person's attendance of a Sunday church service would be more highly correlated with their presence or absence in church a week previous.

Similarly in space, we expect locations which are closer together to have more similar values. For instance, the temperatures in Seattle, WA and Vancouver, BC, or Boston, MA and Providence, RI are likely to be quite similar on a given day. In contrast the temperatures in Seattle, WA and Boston, MA are less likely to be close. Similarly for Vancouver, BC, and Providence, RI. Likewise with time, correlation is not always perfectly connected to the amount of spatial separation (as we will

detail later in our discussions of nonstationarity). For example, we would expect the proportion of people living in apartments to be more alike in cities separated by hundreds or thousands of kilometers than a city and one if its neighboring (lower-density) suburbs.

Several methods exist for mathematically encoding this correlation structure. The first, through graphical models, allows us to specify which variables (locations, or time points for instance) are correlated, and which are not. Through a graphical model we can encode, for example, the knowledge that given all neighboring counties the poverty rate in one county is independent of all other counties. Another method for mathematically expressing structured correlation is by assigning it functional form. This is the technique often taken in time series, for example in AR models, which mathematically express the distribution of the current value given the set of previous values. The two techniques of graphical models and correlation functions actually go hand in hand. Graphical models encode the structure of zero/non-zero correlation between variables, whereas correlation functions can express the value of correlation between variables which are correlated.

1.1.1 Graphical Models

Graphical models are a convenient way to encode independence relationships between variables in a multivariate distribution. This convenience derives from clear visualization techniques, as well as straight-forward computation of conditional and marginal densities [55]. Undirected graphical models, or Markov random fields, have a simple interpretation regarding independence: given a set C, nodes A and B are conditionally independent if all paths between A and B pass through the

set C. Formally,

$$P(A \cap B|C) = P(A|C)P(B|C)$$

This class of graphical models is often used in spatial statistics, imaging, and physics to represent for example the independence relationships between regions, pixels, and molecules, respectively. Undirected graphical models used in this way have a direct correspondence with the inverse covariance matrix, Σ^{-1}, in that a lack of an edge between these variables in the graphical model (indicating conditional independence) leads to a zero in the precision matrix Σ^{-1}.

Graphical models may also have directed edges, which allows for the modeling of more advanced relationship, including causation and deterministic relationships. For our purposes, however, we focus on undirected graphical models.

We now introduce some notation; see, for example Dawid and Lauritzen [23], Lauritzen [55]. Let $\mathscr{G} = (V,E)$ be a graphical model with vertices or nodes $V = \{1,\ldots,n\}$ and pairwise edges E. The pair of nodes $\{i,j\} \in V$ are adjacent or neighbors if $(i,j) \in E$, and a subset $C \subset V$ is said to be complete if all its elements are adjacent to each other. A cycle is a sequence of unique adjacent nodes, where the first node is connected to the last node. In this way a cycle represents the path starting at a given node, through a sequence of unique nodes, and returning to the starting location. A cycle is chordless if all pairs of non-adjacent nodes in the cycle are not neighbors. In other words, a chordless cycle has no chord, or shortcut. A graph with no chordless cycles of length greater than two is said to be chordal, or decomposable. As will be shown in Chapter 2, decomposable graphs have several unique properties, one of which is that the likelihood of the graphical

model may be factorized in a computationally efficient way. Specifically, if P satisfies the conditional independencies implied by a decomposable graph $\mathscr{G}$, then the likelihood of the graphical model specified by P can be factorized according to the graph's cliques and separators

$$p(y|\mathscr{G},\theta) = \frac{\prod_{i=1}^{n_c} p(y_{C_i}|\theta_{C_i})}{\prod_{j=1}^{n_s} p(y_{S_j}|\theta_{S_j})}$$

where θ is a quantity parameterizing the graphical model P over the graph $\mathscr{G}$ and satisfying some consistency conditions with respect to $\mathscr{G}$ ([23]).

Undirected graphical models often arise in spatial statistics, where one is interested in modeling counts, proportions, or some other quantities in neighboring regions. Each region, or areal unit, has a measurement attached to it, and correlation is typically modeled through the neighborhood structure. Specifically, regions which are physically neighbors are also neighbors in the undirected graph, or Markov random field. Hence given its neighbors, a given region is independent of all other regions.

When conducting Bayesian inference on undirected graphical models, one must assign a prior distribution over the space of all possible graphs $\mathscr{G}$ for a given set of variables, or nodes V. For a handful of variables, one may assign a prior probability to each individual graph. However, as the number of graphs scales exponentially with the number of nodes, automated methods must be found. Current approaches have been limited in their ability to accommodate varying forms of prior information on the graph. For instance, in an effort to encourage interpretable graphs, the standard approach has been to penalize the number of edges (conditional dependencies) in the graph. However, many situations exist where one might expect

variables to be clustered together and the graph to exhibit block structure. At the moment no such prior distribution exists to handle this problem. In Chapter 2, we propose a class of prior distributions motivated from the class of product partition models which will allow improved flexibility in the specification of prior information on the graph.

1.1.2 Correlation Functions

We may also mathematically encode correlation structure through functional forms. For example, consider a sequence of observations $X_1, X_2, \ldots, X_T$, modeled with the class of autoregressive (AR) models. A simplified AR(p) model may be written as

$$X_t = \sum_{i=1}^{p} \beta_i X_{t-i} + \varepsilon_t$$

where $\beta_1, \beta_2, \ldots, \beta_p$ are parameters of the model, and ε_t is the noise, or error, term. Through this form, we have implicitly defined a correlation function. For an AR(1) process, it is

$$corr(X_t, X_{t-k}) = \beta_1^k.$$

Correlation functions describe the correlation between random variables at different points in space or time. Typically, as with the AR example above, the correlation decreases with space or time. Or, more generally, the correlation is a function of distance in space or time, although in space it can be also modeled as a function of the vector between locations.

In both time-series and spatial modeling, the value of the correlation modeling becomes particularly obvious when one turns to prediction. Because nearby obser-

vations in space or time are similar to each other, one may use the values at nearby points to predict at the unobserved location or time. In Chapter 4, we discuss a model for crop yield which uses this basic idea to stabilize prediction. Subsequently, Chapter 5 explores more complex correlation functions where modeling correlation as a function of spatial distance is no longer justified.

Chapter 2

Bayesian Clustering in Decomposable Graphs[1]

2.1 Introduction

This chapter is concerned with the inference of the conditional independence graph $\mathscr{G}$ of a multivariate random vector Y of dimension n, a problem sometimes referred to as structure learning. We focus here on undirected decomposable graphs, whose popularity is mainly due to the tractable factorization they allow for the likelihood ([23, 55]); related work for directed graphical models can be found in [53]. Learning the conditional independence graph $\mathscr{G}$ is an onerous task due to the large number of graphs on a set of n nodes, or variables. It is possible using optimization methods to find the graph which best fits the data according to some metric [31, 61, 88]; alternatively Bayesian model averaging may be used to

[1] A version of Chapter 2 has been published. Bornn, L., Caron, F. (2011) Bayesian Clustering in Decomposable Graphs. Bayesian Analysis. Vol. 6, No. 4, pages 829-846. [9]

accommodate uncertainty in the estimated graph, or maximum a posteriori estimation may be used to select a given model from the posterior over graphs. Such an approach relies on a prior distribution $\pi(\mathscr{G})$ over the set of decomposable graphs of a given size; through Bayes theorem, this prior is updated based on the data to give an a posteriori estimate of the distribution over graphs.

Current approaches have been limited in their ability to accommodate varying forms of prior information on the graph. For instance, in an effort to encourage interpretable graphs, the standard approach has been to penalize the number of edges (conditional dependencies) in the graph. However, many situations exist where one might expect variables to be clustered together and the graph to exhibit block structure. At the moment no such prior distribution exists to handle this problem. Our contribution in this thesis is to propose a class of prior distributions motivated from the class of product partition models, which will allow improved flexibility in the specification of prior information on the graph.

The field of agriculture is particularly susceptible to the application of graphical models. Due to large spatial domains as well as multifarious crop varieties, it is valuable to have models which both handle the complexity of the biophysical process as well as allow straightforward interpretation. In particular, one might examine the set of zero/non-zero correlations between crop varieties' yields, using the presence or absence of edges to make decisions regarding crop management, marketing, and insurance policies. In addition, due to small sample sizes in many agricultural applications, the choice of prior distribution becomes particularly important.

2.2 Bayesian Inference on Decomposable Graphs

We begin with a brief overview of graphical models, following the exposition in [23]; see also [55] for further details on graphical models. Let $\mathscr{G} = (V,E)$ be a graphical model with vertices $V = \{1,\ldots,n\}$ and pairwise edges E. The pair of nodes $\{i,j\} \in V$ are adjacent if $(i,j) \in E$, and a subset $C \subset V$ is said to be complete if all its elements are adjacent to each other. A complete subgraph that is maximal (i.e. not contained within another complete subgraph) is called a clique. An ordering of the cliques of an undirected graph, $(C_1,\ldots,C_{n_c})$ is said to be perfect if the vertices of each clique C_i also contained in any previous clique $C_1,\ldots,C_{i-1}$ are all members of one previous clique; that is, for $i = 2,3,\ldots,n_c$

$$H_i = C_i \cap \cup_{j=1}^{i-1} C_j \subseteq C_h$$

for some $h \in \{1,2,\ldots,i-1\}$. The sets H_i, $i = 1,\ldots,n_c - 1$ are called separators. We write $S_1,\ldots,S_{n_s}$ for the non-empty separators (some might appear multiple times). If an undirected graph admits a perfect ordering it is said to be decomposable.

We associate to each vertex i a random variable Y_i, with realizations y_i. For $A \subseteq V$, let $Y_A = \{Y_i | i \in A\}$. A distribution P over V is Markov with respect to $\mathscr{G}$ if, for any decomposition (A,B) of $\mathscr{G}$, X_A is independent of X_B given $X_{A\cap B}$. The widespread use of decomposable models is due to the resulting factorization of densities. Specifically, if P satisfies the conditional independencies implied by a decomposable graph $\mathscr{G}$, then the likelihood of the graphical model specified by P can be factorized according to the graph's cliques and separators

$$p(y|\mathscr{G},\theta) = \frac{\prod_{i=1}^{n_c} p(y_{C_i}|\theta_{C_i})}{\prod_{j=1}^{n_s} p(y_{S_j}|\theta_{S_j})} \tag{2.1}$$

where θ is a quantity parameterizing the graphical model P over the graph $\mathscr{G}$ and satisfying some consistency conditions with respect to $\mathscr{G}$ ([23]).

Traditionally, focus has been on Gaussian graphical models, also known as covariance selection models ([26]) where $P = N_n(\mu, \Sigma)$ is a n-dimensional multivariate Gaussian distribution and θ is the $n \times n$ covariance matrix Σ. Conditional independence structure is represented by the precision matrix Σ^{-1}. If the edge $(i, j) \notin E$, then the variables Y_i and Y_j are conditionally independent given the remaining variables, and $\Sigma^{-1}_{(i,j)} = \Sigma^{-1}_{(j,i)} = 0$. As such, the Gaussian graphical model may be factorized as (2.1) with the covariance Σ replacing θ, and the corresponding likelihood terms written as

$$p(y_B|\Sigma_B) = (2\pi)^{-|B|/2} \det(\Sigma_B)^{-|B|/2} \exp[-\frac{1}{2} tr(S_B(\Sigma_B)^{-1})] \tag{2.2}$$

for each complete set B, where $|B|$ denotes the cardinality of B and S_B is the empirical covariance matrix of y_B.

From a Bayesian perspective, we are interested in the posterior distribution $p(\theta, \mathscr{G}|y) \propto p(y|\theta, \mathscr{G}) p(\theta|\mathscr{G}) \pi(\mathscr{G})$. Much work has been dedicated to specifying proper priors $p(\theta|\mathscr{G})$, see e.g. ([23, 36]). The main focus of this chapter is the specification of a prior distribution $\pi(\mathscr{G})$ over the space of decomposable graphs. As this space is very large compared to the number of observations, it is crucial to add as much prior information as possible on the structure of the unknown graph $\mathscr{G}$. Moreover, we are generally interested in obtaining sparse graph estimates for needs of interpretation and prediction. Up until now, the specification of $\pi(\mathscr{G})$ has been limited to the uniform distribution, or priors which penalize the complexity as measured by the number of edges. This brings us to the focus of this work, namely

a class of prior distributions $\pi(\mathscr{G})$ which subsumes control over the structure and features of $\mathscr{G}$.

2.3 Priors on Decomposable Graphs

2.3.1 Previous work

While early work on inference in decomposable models often assumed a uniform prior over graphs (i.e. [36]), such priors put considerable mass on models of intermediate size. In an effort to put more weight on smaller graphs, several authors have proposed using a binomial prior distribution with parameter ρ on the number of edges r in the graph. This yields priors of the type [27, 50]

$$\pi(\mathscr{G}) \propto \rho^r (1-\rho)^{m-r} \tag{2.3}$$

where $m = \frac{n(n-1)}{2}$ is the maximal number of possible edges on n nodes. When $\rho = 1/2$, it reduces to the forementioned uniform prior over graphs. [50] suggest the use of $\rho = 2/(n-1)$, motivated from the resulting density's peak at n edges in the unconstrained graph. Some authors also consider adding a hierarchical Beta prior $\rho \sim \mathtt{Be}(a,b)$ ([19]), giving the marginal prior on the graph as

$$\pi(\mathscr{G}) = \int_0^1 \pi(\mathscr{G}|\rho)\pi(\rho)d\rho \propto \frac{\beta(a+k, b+m-k)}{\beta(a,b)}$$

where $\beta(\cdot,\cdot)$ is the beta function. [19] suggest a default choice of $a = b = 1$, implying a uniform prior on ρ. Interestingly, the resulting prior on $\mathscr{G}$ is

$$\pi(\mathscr{G}) = \frac{1}{m+1}\binom{m}{r}^{-1},$$

which penalizes medium-sized graphs as desired. Such a prior weights each graph according to the number of graphs in the unrestricted space with the same number of edges. However, as shown by [3], the space of decomposable graphs can be considerably different than the unrestricted space. To address this, [3] have proposed a uniform prior on decomposable graphs given the number of edges. However, calculating the number of decomposable graphs of a given size is an arduous task: there exists no list in the literature of decomposable graphs and their breakdown in terms of number of edges, nor are there straightforward ways of computing such quantities. As a result, [3] proposes an MCMC estimation scheme, testing its accuracy up to 12 nodes, although such a scheme will likely become prohibitive in higher dimensions.

While the priors in the above references allow one to control the size of the resulting graphs through the number of edges, often doing so results in undesirable graph structures, namely those with a high number of separators and long strings of nodes. Figure 2.1(top) shows random samples from a binomial prior over 20-node graphs with $\rho = 0.1$ (closely echoing the choice of [50], namely $\rho = 2/(n-1) \approx 0.1$) and $\rho = 0.5$ (the uniform prior). We see from this plot that there is no clustering of the cliques, making interpretation difficult. In addition, the long strings/trees seen for $\rho = 0.1$ do not mesh with reality in most cases. Clearly such a class of priors is not suitable if one suspects clustering amongst the variables, clique sizes to be upper (or lower) bounded, or nearly full separation between cliques. Our focus therefore is on moving beyond priors which focus on the number of edges to priors which focus on graph (clique and separator) structure.

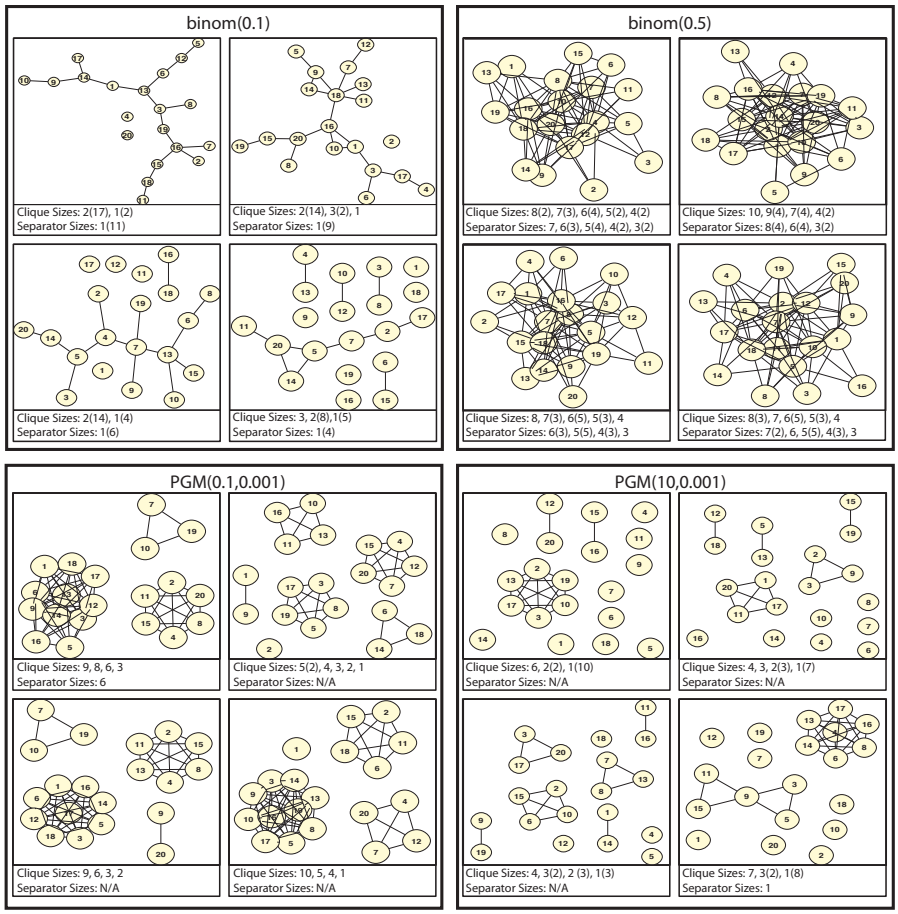


Figure 2.1: Four random samples from binomial and product graphical model (PGM) priors. Clique and separator sizes for each graph are also shown ("Clique Sizes: $2(3)$" implies 3 cliques of size 2). 4 million samples were generated using Markov chain Monte Carlo, and every millionth is shown. While the binomial is characterized by large strings and many separators, the product graphical model allows one to induce clustering by setting b small.

2.3.2 A new prior distribution on decomposable graphs

Motivated from the class of product partition models ([7, 8, 42]), we consider prior distributions of the form

$$\pi(\mathcal{G}) \propto \frac{\prod_{j=1}^{n_c} \psi_C(C_j)}{\prod_{j=1}^{n_s} \psi_S(S_j)} \tag{2.4}$$

where ψ_C and ψ_S are respectively called the clique/separator cohesion functions, with the convention that $\psi_S(\emptyset) = 1$. Evidently one could choose to penalize only cliques or separators by setting ψ_C or ψ_S to constant values. Alternatively, one could simply penalize clique sizes by setting $\psi_B = a|B|$. Motivated from the class of product partition models, consider the cohesion functions $\psi_C(B) = a(|B|-1)!$ and $\psi_S(B) = \frac{1}{b}(|B|-1)!$, $a > 0$, $b > 0$, hence

$$\pi(\mathcal{G}) \propto a^{n_c} b^{n_s} \frac{\prod_{j=1}^{n_c} (|C_j|-1)!}{\prod_{j=1}^{n_s} (|S_j|-1)!} \tag{2.5}$$

The factorial terms result in predilection towards large cliques and small separators – a desirable trait in terms of interpretability of the resulting graph. For instance, even if $a = b = 1$ with 20 nodes, the completely connected graph would be preferred over the complete independence graph by a factor of 20!. The parameters a and b respectively tune the number of cliques and separators in the decomposable graph. For a small, the prior will favour a small number of large cliques. Likewise for b, with small values favouring fewer separators. Figure 2.1 (bottom) shows samples from this prior. Because of its relation to product partition models (described later), we term this prior the product graphical model prior. To clearly demonstrate the control the product graphical model prior (2.5) gives relative to the binomial prior, we set $b = 1/1000$, highly penalizing the number of separators and

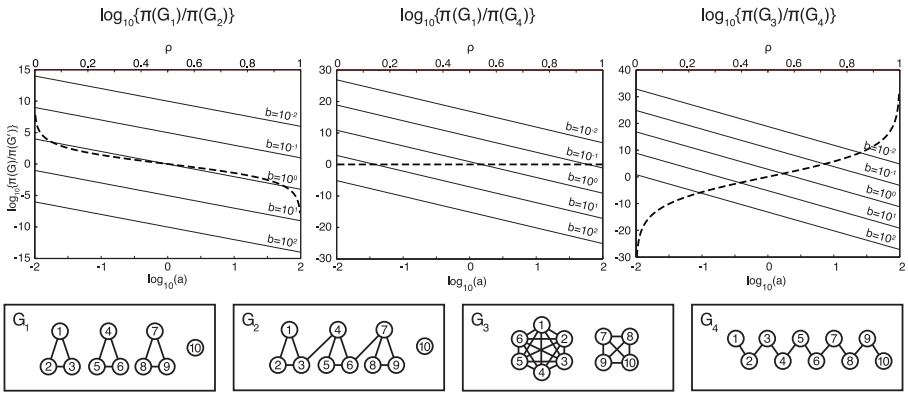


Figure 2.2: Log ratio of priors over two graphs for product graphical model prior for various a, b (solid, bottom axis) and binomial prior for various ρ (dashed, top axis). While the binomial prior allows one to control the number of edges, for instance choosing $\mathscr{G}_1$ over $\mathscr{G}_2$, the same parameter would seldom choose $\mathscr{G}_3$ over $\mathscr{G}_4$, despite $\mathscr{G}_3$ having a sparse covariance matrix, and $\mathscr{G}_4$ having a saturated covariance matrix.

hence resulting in highly separated cliques. In addition, we look at two different values for a; $a = 0.1$, resulting in fewer and larger cliques, and $a = 10$, resulting in more (but smaller) cliques. Fig. 2.1 demonstrates the ability of the prior to induce clustering of the cliques, and therefore sparsity in correlation.

We have seen some general properties of the prior (2.5), namely the ability to control the number of cliques and separators. Figure 2.2 shows $\log_{10}(\pi(\mathscr{G})/\pi(\mathscr{G}')))$ for different graphs $\mathscr{G}, \mathscr{G}'$. Specifically, decreases in b result in increased prior probability on models with few separators; in addition we see that as a is increased, more mass is put on models with many cliques. In contrast, we also plot the same ratio for the binomial prior (2.3). From this, one can see the limited control such a prior gives, favouring small models in terms of number of edges, but putting very little mass on models, for example, which feature clusters of fully-connected nodes (as in $\mathscr{G}_3$) and therefore have sparse covariance matrices.

Selecting the appropriate cohesion functions in equation (2.4) is a difficult problem, but one for which we may gain insight from the existing literature on product partition models ([21, 75, 76]). For instance, one may use Figure 2.2 to select a and b to best fit with prior intuition regarding the features of the graph, then verify the choice through generation of Monte Carlo samples from the prior as in Figure 2.1. Alternatively, cross-validation or related methods may be used to select a and b; due to the potential computational cost of such methods, sequential Monte Carlo approaches may be used to speed up prior distribution selection ([12]).

Given that the likelihood decomposes as (2.1) and the prior is of the form (2.4), the posterior will also be of the form (2.4) with cohesions $\psi_C(C_j)p(y_{C_j})$ and $\psi_S(S_j)p(y_{S_j})$. The prior admits several other attractive properties and connections with well-known clustering methods as well. If $\psi_S(S_j) \to \infty$ for all $S_j \neq \emptyset$, then Equation (2.4) reduces to the following model

$$\pi(\mathscr{G}) \propto \prod_{j=1}^{n_c} \psi_C(C_j)$$

if $n_s = 0$ and 0 otherwise. The resulting prior puts only positive mass on graphs with no separators. It has been introduced as a prior over partitions by [42] and [7, 8] under the name of *product partition models*. In the particular case of (2.4) with $b \to 0$, the prior over $\mathscr{G}$ reduces to

$$\pi(\mathscr{G}) = \frac{a^{n_c}\Gamma(a)}{\Gamma(a+n)} \prod_{j=1}^{n_c} (|C_j| - 1)!$$

As shown by [76] (see also [75]), this is the distribution over partitions induced by

a Dirichlet Process [1, 30]. We also have

$$\mathbb{E}(n_c) = \sum_{i=0}^{n-1} \frac{a}{a+i} \simeq a\log(1+n/a)+\gamma, \quad var(n_c) = \sum_{i=1}^{n-1} \frac{ai}{(a+i)^2}$$

where γ is Euler's constant and

$$\mathrm{pr}(n_c = k) = s(n,k)a^k\Gamma(a)/\Gamma(a+n)$$

where the coefficients $s(n,k)$ are the absolute values of Stirling numbers of the first kind [1]. In this limiting case, the number of cliques increases logarithmically with the number of nodes.

2.3.3 Extensions

Motivated by the larger class of exchangeable partition functions [54, 70], we can also consider four-parameter models, allowing more control over the relative sizes of the cliques/separators

$$\pi(\mathscr{G}) \propto \frac{\prod_{j=1}^{n_c}(a_2+a_1(j-1))\frac{\Gamma(|C_j|-a_1)}{\Gamma(1-a_1)}}{\prod_{j=1}^{n_s}(b_2+b_1(j-1))\frac{\Gamma(|S_j|-b_1)}{\Gamma(1-b_1)}}$$

where $a_2 > -a_1, 0 \le a_1 < 1$, likewise for b_1, b_2. The above model reduces to (2.5) when $a_1 = b_1 = 0$. We can also consider models that control the maximal number of cliques/separators

$$\pi(\mathscr{G}) \propto \frac{\prod_{j=1}^{n_c}(c_1-j+1)\frac{\Gamma(c_2+|C_j|)}{\Gamma(c_2)}}{\prod_{j=1}^{n_s}(d_1-j+1)\frac{\Gamma(d_2+|C_j|)}{\Gamma(d_2)}}$$

where $c_1, c_2, d_1, d_2 > 0$, and $c_1 > d_1$ are the maximal number of cliques/separators. These two models respectively admit as limiting cases the distribution over partitions induced by the two-parameter Poisson-Dirichlet distribution and the finite Dirichlet-multinomial distribution, see e.g. [54] for further details on these distributions. Using such extensions, one is able to both extend the product graphical model prior to control relative sizes and the maximal number of cliques and separators, as well as borrow from the wealth of literature on Dirichlet and related distributions to gain insight into the prior distribution's characteristics.

2.4 Example: Modeling Agricultural Output of Different Species

Determining agricultural policies to govern crop production, harvesting, and export is a challenge fraught with high variability both temporally and spatially. Enabling effective crop management, handling, and marketing techniques thus requires accurate understanding of crop yield to account for and explain these variations. While much effort has been made in developing models for predicting single crops ([10, 72]), little effort has been made in understanding statistically the relationship between crop yield of different crop varieties.

Understanding the connection between yields of different crop varieties is valuable for a multitude of reasons. Firstly, because crops are planted and harvested at different times, the management of one crop might benefit from knowledge obtained from harvesting a similar crop earlier in the year. Additionally, by accounting for correlation between different crops, insurers might better cover themselves against extreme events and better control insurance rates for farmers. Lastly, farmers themselves might wish to ensure some level of stability in their income, and

therefore might prefer to plant crops which are uncorrelated in yield. Through such a practice, a farmer would be proactive in preventing disasters across his entire crop portfolio. Simply by looking at the resulting undirected graph, a farmer could select two crops which do not have a path connecting them, and are therefore uncorrelated.

We examine the total production (in thousands of bushels) of 24 crops in the state of California from the years 1990 to 2009 (20 years). The data is compiled from the U.S Department of Agriculture website, where a considerable database is available for viewing and analysis. The 24 crops include, for example, several varieties of wheat, rice, and beans. We use the now-standard Gaussian hyper-inverse Wishart model: the likelihood of yield is given in (2.1) and (2.2), and the prior for the covariance matrix Σ is hyper-inverse Wishart, which factorizes similarly to (2.1), as a ratio of inverse Wishart distributions over cliques and separators ([37]). To be specific, for each clique C (or, equivalently, separator S), the covariance over the clique is distributed as $\Sigma_C \sim IW(b, D_C)$, with parameters b and D_C and density

$$\pi(\Sigma_C|b, D_C) \propto |\Sigma_C|^{-(b+2|C|)/2} \exp\left\{-\frac{1}{2} tr(\Sigma_C^{-1} D_C)\right\}.$$

See [19] for some alternative marginal likelihoods based on fractional Bayes factors which can help to induce parsimony. The parameters chosen for the hyper-inverse Wishart distribution are as described in [50]; we focus on the specification of $\pi(\mathscr{G})$. Looking at the list of crops, one would expect that there will be clustering of the yields according to crop characteristics. For instance, it would be reasonable to expect the yield of beans to be correlated with each other. We also seek an interpretable graph, namely one with small complexity (in terms of number of

edges and/or separators). The first such prior we examine is the binomial prior of [50] with $\rho = 2/(n-1)$, chosen due to its prevalence in the literature. While such a prior allows for penalization on the number of edges, no control is available over clustering. In contrast, by using the prior (2.5), we can set $b = .01$ to put strong penalization on the number of separators (and hence induce separation of the cliques and therefore sparsity in the correlation matrix), and set $a = .01$ to encourage a small number of cliques in the pursuit of simplicity in the resulting graph.

We run MCMC of length 10 million over the space of decomposable graphs ([36]) for both the binomial and product graphical model priors, thinning to every 100 samples. With both priors, one may save computational resources by making local moves, merging and splitting cliques within the Markov chain. As a result, one need not re-determine the structure of the entire graph at each move.

Figure 2.3 shows the 4 graphs with highest posterior probabilities from each prior. The product graphical model prior results in the top 4 graphs having posterior density values in the range 0.11 to 0.49, whereas for the binomial the range is 0.04 to 0.06, indicating that the binomial prior spreads mass much more evenly across distributions relative to the product graphical model prior with $a = b = 0.01$. Immediately evident from the figure are the different forms resulting from each prior. Specifically, the binomial prior induces long strings of nodes with many separators, whereas the product graphical model posterior reflects our prior beliefs that variables will cluster together, resulting in sparsity in the correlations between variables. A commercial farmer desiring to plant two plots with uncorrelated crops to minimize the risk of loss might reach quite different conclusions from each prior. Specifically, the large strings of nodes from the Binomial prior suggest correlation

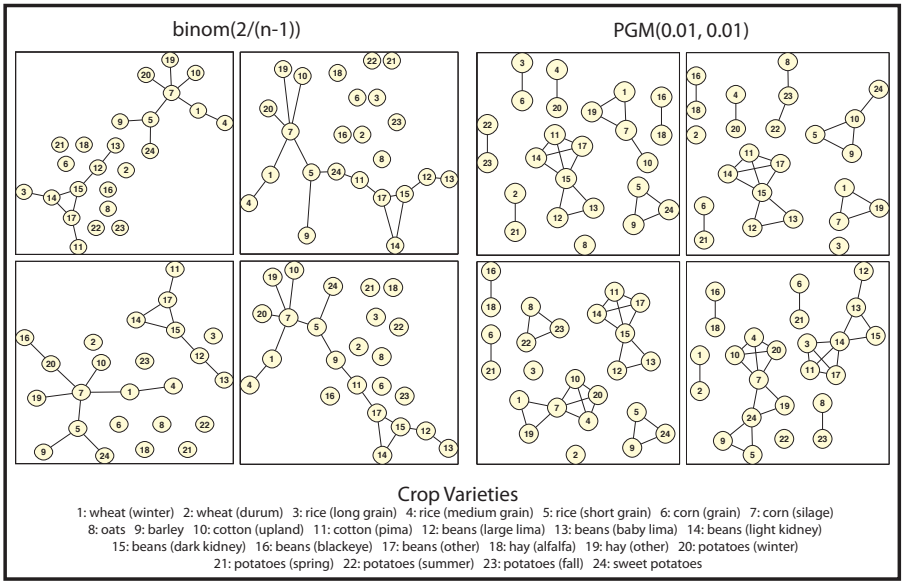


Figure 2.3: Four samples with highest posterior probability from crop yield model using binomial and product graphical model (PGM) priors. We see the bean yields (nodes 12 through 17) seem to cluster together, as do summer and fall potatoes (nodes 22 and 23). We also observe that the product graphical model prior induces separated cliques, whereas the binomial prior results in long strings and trees of connected variables. As a result, the product graphical model prior will induce sparsity in the resulting posterior covariance.

between the majority of crops. The farmer might not plant winter wheat (planted in late fall) and a strain of beans (harvested in early fall) on his two plots, despite their very different growing seasons, due to their connection in two of the highest posterior probability graphs in Figure 2.3. In contrast, the separation of cliques from the product graphical model prior (2.5) would allow these crops to be planted together. Such decisions could be made from the highest posterior graph, or by conducting Bayesian model averaging to obtain the expected utility of a given decision.

To gain an understanding of the product graphical model prior's prediction

Table 2.1: Log predictive density evaluated on test data using various binomial priors

Distribution:	Binomial	
Parameters:	$2/(24-1)$	0.5
Avg. Log Predictive:	-688	-707
Avg. Number of Edges:	17.8	29.6

Table 2.2: Log predictive density evaluated on test data using various PGM priors

Distribution:	PGM		
Parameters:	$(0.01, 0.01)$	$(0.1, 0.1)$	$(1, 1)$
Avg. Log Predictive:	-675	-677	-686
Avg. Number of Edges:	18.1	16.4	20.3

performance, we split the data into a training set (first 12 years) and testing set (last 8 years). After simulating from the posterior distribution arising from the binomial and product graphical model priors, we use Bayesian model averaging via the marginal likelihood evaluated on the test data to judge the model's prediction performance. We evaluate the resulting posterior predictive evaluated on the test set in Tables 2.1, 2.2; indeed, the product graphical model prior provides better prediction in this example, even over a variety of parameter choices. We also show the number of edges for each model, indicating that sparsity in terms of edges alone is not responsible for the improved prediction.

2.5 Example: Modeling 20th Century American Voting Patterns

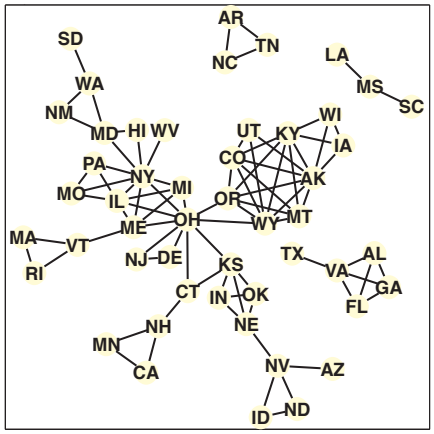
In an effort to demonstrate the product graphical model prior in higher dimensions, we now turn to the modeling of American voting data by state. For each federal

election from 1904 to 1976, occurring every four years, we measure the proportion of votes for the republican party in each of the 50 states ([18]). Our goal is to model and visualize correlation in voting pattern changes over the last century. Some immediate questions come to mind: "Do certain states have an important role in determining election outcomes?", "Are there groups of states which vote together, operating independently from the US as a whole?"

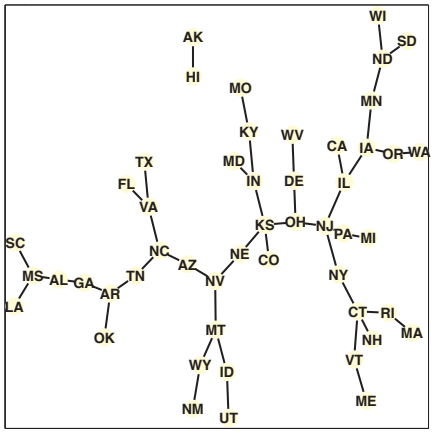
We proceed by exploring the posterior distribution resulting from the binomial prior with edge probability 0.1, and the product graphical model prior with parameters $a = 10, b = 10^{-3}$, in an effort to make the overall number of edges resulting from each model comparable. Figure 2.4 shows the two graphs with highest posterior density from each model. As expected, the binomial graphs contain long strings of variables, while the product graphical model prior demonstrates clustering and grouping of variables. While the binomial prior results in similar variables placed along the same string, the grouping from the product graphical model allows for clearer interpretation. For instance, we immediately observe that the southern states (SC, MS, LA, AL, GA, TX, VA, FL) generally vote in a group. Other patterns of interest also arise, including a close connection between AR, NC, and TN. Also, notice that NY and KS are consistently the single node connecting clusters of variables. As such, these states might be considered as key indicators of voting behavior.

2.6 Discussion

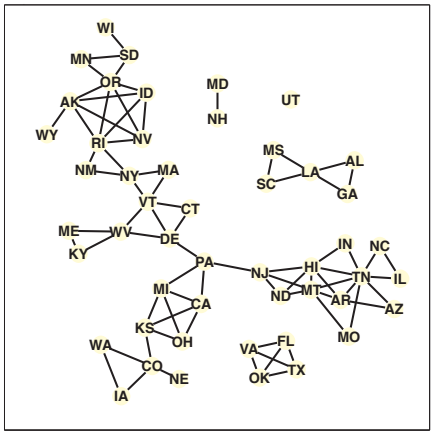
The product graphical model's implementation relies on Markov chain Monte Carlo, whereby samples of graphs from the posterior distribution are generated and inference is made from their empirical distribution. Because of the sheer number of



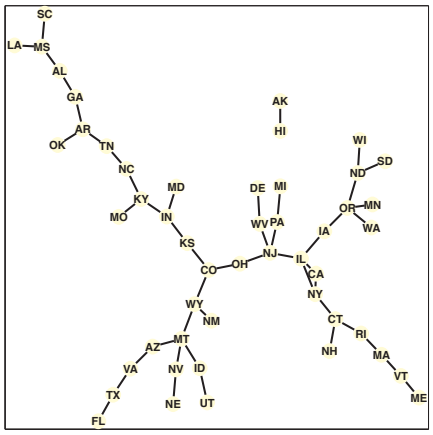
(a) PGM($10, 10^{-3}$): HPD Graph 1



(b) Binom(0.01): HPD Graph 1



(c) PGM($10, 10^{-3}$): HPD Graph 2



(d) Binom(0.01): HPD Graph 2

Figure 2.4: Voting example: two graphs with highest posterior density (HPD) from binomial and product graphical model priors.

graphs, one must be careful in interpreting the results of this approximation, hence our use of model averaging (for prediction) as well as studying a suite of graphs with high posterior probability. While fast and easily implementable software for the above models has been generously released by [50], one must still conduct checks for convergence of the Markov chain. This is exacerbated by the complexity of the space of decomposable models, whereby two highly probable models might be separated by a series of models of low probability. This is particularly true for the product graphical model prior; if one chooses a prior which highly penalizes the number of separators (say, using b very small), transversing the space of graphs may be difficult, as the fully disconnected graph and fully connected graph both have no separators (and therefore high probability), but necessarily have a considerable number of graphs with many separators between them.

While we have focused on the Bayesian approach to covariance selection, significant work has also been done in a non-Bayesian framework. A common approach involves placing an ℓ_1 penalization on the precision matrix Σ^{-1}, which leads to sparse estimates ([31, 61, 88]). Closer to the heart of this chapter, [60] examine the case of estimating $\mathscr{G}$ when clustering is expected, and therefore Σ^{-1} exhibits block structure. However, these models are neither decomposable nor generative.

While we have focused in this article on Gaussian graphical models, the prior defined in this article is far more general and can be used with any type of model for handling discrete or mixed data, see e.g. [55, 57]. We have also considered the hyperparameters a and b to be known constants. Estimating them within the MCMC sampler would require one to compute the normalizing constant in (2.5), which is in general not tractable. An exception of interest is the case $b \to 0$, where we can assign a gamma prior to a and use the data augmentation algorithm described

in [86] to update a given the other variables.

In conclusion, the proposed product graphical model prior improves flexibility in modeling decomposable graphical models and borrows strength from the immense literature on product partition and related models. The product graphical model prior allows one to encourage (or discourage) clustering of the graphs, and therefore can induce sparsity in the correlation matrix through clique separation; consequently, the product graphical model empowers practitioners to encapsulate their true prior beliefs to build a model more attuned to the problem at hand.

Chapter 3

Spatial Statistics and Nonstationarity

In spatial statistics, the role of correlation plays a particularly crucial role. Here observations are tied together through coordinates of geographic systems. Applications range from modeling astrophysical systems by way of their locations within galaxies, to modeling agricultural output or air pollution over large and diverse regions. The defining characteristic of spatial statistics is that each entity under study has attached to it a spatial location; how that location is defined, whether latitude/-longitude or arbitrary coordinates, varies depending on the example at hand.

Our focus is on point-referenced geostatistics and lattice, or areal, data. The latter are typically modeled using Markov random fields, or undirected graphical models, as discussed earlier. In contrast, the former is employed to model spatial processes over continuous domains.

3.1 Geostatistics

The origins of geostatistics is Kriging, where the goal is to interpolate the value of a random field at an unobserved location using observations from nearby locations. Kriging works by computing the best linear unbiased estimator at the unobserved location, using information about the correlation present in the system. The Kriging estimator is a linear combination of the observed values $Z(x_1),\ldots,Z(x_s)$ with weights $w_1,\ldots,w_s$,

$$\hat{Z(x_0)} = \sum_{i=1}^{s} w_i(x_0)Z(x_i).$$

The weights $w_i, i = 1,\ldots,s$ are chosen to minimize the Kriging variance

$$Var\left(\hat{Z}(x_0) - Z(x_0)\right)$$

while maintaining the unbiasedness condition,

$$E\left(\hat{Z}(x) - Z(x)\right) = \sum_{i=1}^{s} w_i(x_0)\mu(x_i) - \mu(x_0) = 0$$

Classically, the correlation in the system was modeled through variograms, which describe the spatial dependence in a spatial random field through the variance of the difference between field values at two locations x_i, x_j:

$$2\gamma(x_i,x_j) = var(Z(x_i) - Z(x_j))$$

or if the field has a constant mean,

$$2\gamma(x_i, x_j) = E(|Z(x_i) - Z(x_j)|^2).$$

For those more familiar with the notation of Gaussian processes, where covariance functions are used, variograms may be related as follows:

$$2\gamma(x_i, x_j) = C(x_i, x_i) + C(x_j, x_j) - 2C(x_i, x_j) + (E(Z(x_i)) - E(Z(x_j)))^2$$

For observation $z_1, \ldots, z_s$, the empirical variogram is calculated as

$$\hat{\gamma}(h) = \frac{1}{|N(h)|} \sum_{i,j \in N(h)} |z_i - z_j|^2$$

where $N(h)$ is the set of observations within a small tolerance of the distance h. It is straightforward to see that the empirical variogram provides an unbiased estimate of the theoretical variogram.

$$\begin{aligned} E\left[\hat{\gamma}(h)\right] &= \frac{1}{2|N(h)|} \sum_{i,j \in N(h)} E\left[Z(x_i) - Z(x_j)\right]^2 \\ &= \frac{1}{2|N(h)|} \sum_{i,j \in N(h)} 2\gamma(x_j - x_i) \\ &= \frac{2|N(h)|}{2|N(h)|} \gamma(h) \end{aligned}$$

Many variants on Kriging exist. Simple Kriging assumes a constant trend, $\mu(x) = 0$. Ordinary Kriging assumes an unknown, but constant, trend $\mu(x) = \mu$, which is extended by Universal Kriging to polynomial trend models.

By assigning a parameteric form to the variogram and learning the parameters,

Kriging is able to provide the best linear unbiased estimate of the field at unobserved locations. While the shape of the parametric variogram may be controlled by many parameters, the primary parameters of interest are the nugget, which describes the measurement error, the sill, which describes the limit as the variogram tends to infinite lag distances, and the range, which describes where the variogram has nearly reached the sill.

3.2 Lattice Data

The primary goal of areal, or lattice, models is to detect spatial patterns. For example, are areas closer to eachother more similar than areas which are far apart? An additional goal can be to smooth the data. For example, if modeling health counts, we may have low population in some regions, and as such counts may be unreliable. As a result, we may wish to smooth observations using counts from (higher population) neighboring regions. Lastly, we sometimes want to change the boundaries of areas, or extend the region to include new areas. In this setting, we'd like to extend the results from the current areas to predict value in the new region.

One of the first steps in areal data is defining the neighborhood structure. This may be done straightforwardly using graphical models, as discussed in the previous chapter. Alternatively, or more classically, we may think of a weight matrix W consisting of elements w_{ij}. Typically, $w_{ii} = 0$ and $w_{ij} = 1$ if areas i and j are neighbors. From this weight matrix, we can then calculate the areal analogue of a variogram, called Geary's C:

$$C = \frac{(n-1)\sum_i \sum_j w_{ij}(Y_i - Y_j)^2}{\sum w_{ij} \sum_i (Y_i - \overline{Y})^2}.$$

For the purposes of the subsequent chapter, however, we model the areal data as point-referenced data, with the spatial reference point defined from the centroid of each region.

3.3 Nonstationarity

Informally, if a spatial (or temporal) random process has the same joint probability distribution at all points in space (or time), it is said to be stationary. Specifically, suppose $\{Z(x_i)\}$ is a stochastic process defined at points $x_i, i = 1, \ldots, s$ with cumulative distribution function $F(Z(x_1), \ldots, Z(x_s))$. If for all δ,

$$F(Z(x_1), \ldots, Z(x_s)) = F(Z(x_1 + \delta), \ldots, Z(x_s + \delta)).$$

As such, in these situations the field's mean and variance do not change over time or space. In fact, if only the first and second moments of the field are constant over space and time, we say that the process is weak-sense stationary. If this holds, the semi-variogram (γ) may be represented as a function,

$$\gamma(x_i, x_j) = \gamma(x_i - x_j).$$

Further, if the process is isotropic, then the variogram may be represented solely as a function of the distance between locations,

$$\gamma(x_i, x_j) = \gamma(||x_i - x_j||).$$

If the process is stationary, the relationship between variogram and covariance function is also simplified:

$$2\gamma(x_i, x_j) = C(x_i, x_i) + C(x_j, x_j) - 2C(x_i, x_j)$$

In the following chapter, Chapter 4, we model crop yields in the Canadian Prairies, and discover that differences between provinces leads to an inherently nonstationary spatial field. Subsequently, in Chapter 5, we propose a new methodology for handling nonstationary spatial fields, whereby nonstationary fields are expanded into a higher-dimensional space where stationarity holds.

Chapter 4

Efficient Stabilization of Crop Yield Prediction in the Canadian Prairies[1]

4.1 Introduction

This chapter presents a method for forecasting wheat crop yields in the Canadian Prairie Provinces – a challenging task due to dramatic variability in yield over space and time. Its importance, however, should not be understated: wheat is one of Canada's primary exports, accounting for 12 percent of wheat and barley traded in the world market. Thus variation in yield has considerable impact both within and beyond Canadian borders ([81]). Enabling effective crop management, handling, and marketing thus requires accurate predictions of crop yield that account

[1] A version of Chapter 4 has been published. Bornn, L., Zidek, J. (2012) Efficient Stabilization of Crop Yield Prediction in the Canadian Prairies. Agricultural and Forest Meteorology. Vol. 152, Pages 223-232. [10]

for and explain these variations. For example, these forecasts are helpful in setting insurance premiums and futures prices as well as in managing grain transport. Since spatial and temporal climate variability affect crop yields ([84]; [72]), a crop yield forecasting method must include climate as an essential component if it is to be successful.

Several process–based models have been successfully used for crop yield prediction including the Agricultural Production Systems Simulator (APSIM) in Australia ([51]) as well as a web-based tool developed by the United States' Southeast Climate Consortium ([49]). These process–based models typically employ tunable and user adjustable deterministic and stochastic models to simulate biological and physical processes related to crop yield. While these models use knowledge pertaining to the individual processes, they often require significant input from the user, including a wide range of meteorological and environmental variables which may be difficult or expensive to obtain.

In contrast to the above, traditional statistical techniques are purely empirical. While these methods may result in accurate predictions, they typically lack the interpretability of process–based models ([6]). As a result of this criticism, recent years have seen the development of statistical models that also provide interpretation of the underlying biophysical process (see, for example, [83], [41]). One such process knowledge-based approach involves water stress indices ([71], [72]; [73], [74]), the result of which has been of tremendous use and benefit to stakeholders, allowing for prediction and understanding of crop yield anomalies. While these models have improved the prediction of crop yield, there exists scope for improvement through a) providing an efficient dimension reduction of explanatory variables; b) accounting for uncertainty in the estimated technology trend; c)

modelling spatial correlation between regions.

This chapter describes the results of a project coordinated by Agriculture and Agri-foods Canada to develop a model that explains and predicts wheat yield and its relation to climatic variables. With plans for an online implementation in the future, efficiency was required as a feature of the model, as was the ability to stabilize the effects of noisy measurements. Building on earlier work, we employ a crop water stress index (SI) to provide explanatory power for a new crop yield predictor ([25]). To improve prediction over existing approaches, we extract a sensitive yet low-dimensional summary of this stress index, comparing various alternatives and bases before ultimately selecting principal components. We then demonstrate its improved prediction performance compared to currently used windowed average approaches. In contrast to previous work which models each agricultural region separately, we create a unified model that allows strength to be borrowed from adjacent and nearby regions, thus stabilizing both inference and prediction. By employing a spatially-motivated context-specific prior distribution on the parameters of interest, we account for and use spatial correlation between sites while smoothing and consequently improving predictions.

Following this introduction, Section 2 describes the crop yield forecasting problem and available data. This section works through a series of successively improved models, eventually leading to a Bayesian model in Section 3 which jointly models all regions simultaneously. Model testing and diagnostics are explored in Section 4. Lastly, Section 5 concludes the work.

4.2 Materials and Methods

This chapter models crop yield in the Canadian Prairies as a function of climate-related explanatory variables. The data include annual wheat yields (in bushels per acre) along with associated measurements of a crop water stress index and growing degree day (both described later) for 40 agricultural regions (plotted in Figure 4.1) across the Canadian Prairies from 1976-2006. The agricultural regions are those used in the 2006 Canadian Census of Agriculture, through which the data are also obtained, and are determined from climate and soil information. For each of the 31 years and 40 regions, yield is an aggregated average across the region. Likewise, stress index and growing degree day are calculated regionally, but on a daily basis throughout the growing season (April 1 to September 30).

4.2.1 Incorporating soil water

The well recognized influence of soil water on crop yields dictates its inclusion in any yield prediction model ([25]). However, due to the time-consuming and costly process of measuring soil water content, in practice its effects must be inferred from more widely available environmental variables such as precipitation, temperature, and easily measured crop and soil-related factors. A suite of models have been developed which attempt to understand soil water availability in the context of these environmental variables. Beginning with simple water balance approaches that balance precipitation and soil water storage with evapotranspiration and water runoff, these models have increased in their complexity over the years ([25, 85]). For the reasons given below we focus on budget models, which build on the premise that above a certain threshold (called the 'field capacity'), soil cannot absorb any more water and therefore any additional water is drained off through

runoff or drainage. Also, if the soil water fails to be replenished through precipitation, irrigation, or other sources, the soil reaches a point where plant roots are no longer capable of uptaking water. This stage is known as the 'wilting point'.

Evapotranspiration, which describes the sum of evaporation and plant transpiration, measures the water lost from plants, soil, and other land surfaces into the atmosphere. There are two key components in the budget model, potential evapotranspiration (PET) and actual evapotranspiration (AET). PET represents the atmospheric demand for evapotranspiration; specifically, it accounts for the energy available to evaporate water and transport it into the lower atmosphere. AET is the actual water content available for evaporation and transpiration, and relies on plant physiology and soil characteristics for its calculation. When the soil has ample water, the actual evapotranspiration (AET) can equal the PET. However when the soil is not at its field capacity, AET will be less than PET. More details on these concepts and soil science in general may be found in [16].

Budget models are straightforward to implement since they require a minimum of meteorological data as well as soil field capacities and wilting points. While more advanced models have been built which include soil hydraulic characteristics and more complex relationships between soil, plant, and meteorological systems, these models require considerably more information from the user, including detailed soil and plant characteristics. Because of the additional variables required by these models, we employ a budget model in the remainder of this work. Our model uses crop water stress index (SI) over agricultural land, defined as 1-AET/PET ([73, 74]). This quantity will be near 0 when water is plentiful in the soil and near 1 when the plant is stressed by a lack of available moisture. Intuition might suggest directly including precipitation, temperature, soil and plant information into

the model. However, doing so would add a large number of variables, especially considering that many of these variables are observed for every day of the growing season. Using the SI instead provides an economical reduction in the dimensionality of the description space in a way that respects the biophysical processes involved in soil water movement and availability.

Predicting yield with SI

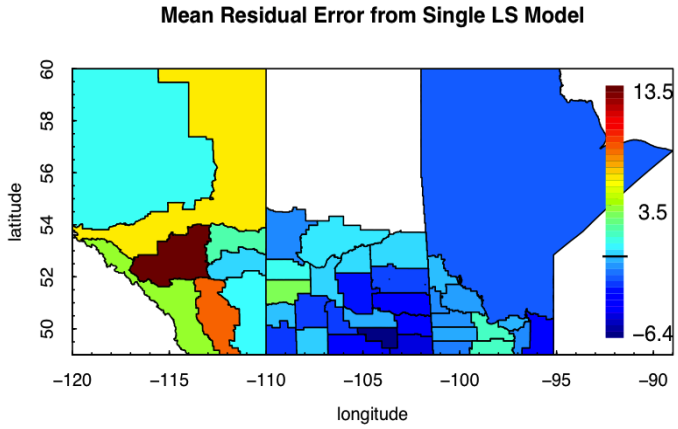
We begin by detailing the process of fitting a regression model to crop yield using least squares (LS). First let $y_{j,t},\ j = 1,\ldots,40$ be the yield from region j for years $t = 1976,\ldots,2006$. Since SI is a daily value, we create an annual average for each year and region; let $\overline{si}_{j,t}$ denote the vector of these means in year t for each region j. We begin by fitting a common regression model to all regions, specifically

$$y_{j,t} = \beta_0 + \beta_1 t + \beta_2 \overline{si}_{j,t} + \varepsilon_{j,t}. \tag{4.1}$$

Here $\varepsilon_{j,t}$ for year t and region j represents a combination of model and measurement error. While previously developed statistical models for crop yield account for a technology trend by first fitting a regression on time and then modelling the residuals, such approaches yield little understanding about the uncertainty associated with forecasting. In particular, while forecasts that use detrended data may be similar, their associated variances will be biased as uncertainty in the trend model is ignored. In fact, to properly account for all sources of variability the technology trend should be an integral part of any forecasting model.

To begin, note that the simple model in Equation 4.1 relies on only 3 parameters – all regions are described by the same equation. The validity of inference for

Figure 4.1: Mean residuals from model (1). We observe that the model residuals are spatially correlated.



such a model relies on assumptions including for instance that the errors $\varepsilon_{j,t}$ are stochastically independent for all j,t. To test this assumption, we plot the mean residual (averaged over the 31 years) for each of the 40 stations in Figure 4.1. This figure makes it clear that the residuals are spatially correlated. For instance, the residuals in Alberta (the western-most Prairie Province) are much larger than the other two provinces, highlighting the fact that the model is biased, particularly in central Alberta. Considering the mean and standard deviation of crop yield across the prairies are 30.9 and 8.2 respectively, the average residual value of 13.5 in this region indicates that the model is consistently underestimating the crop yield there.

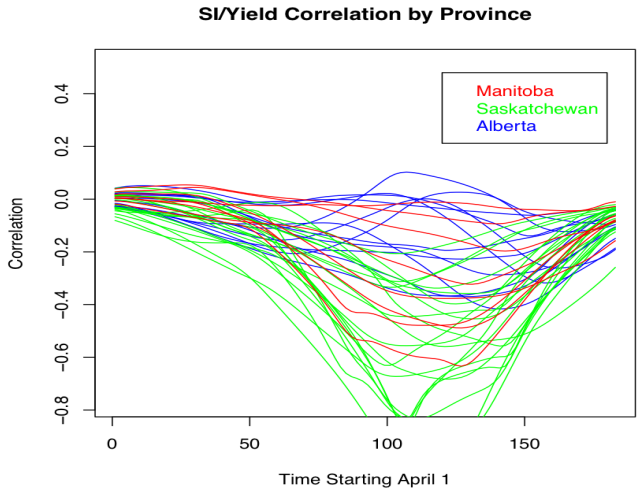
To gain descriptive power, researchers have expanded the above model by fitting a different regression model to each region, specifically

$$y_{j,t} = \beta_{0,j} + \beta_{1,j}t + \beta_{2,j}\overline{si}_{j,t} + \varepsilon_{j,t}. \tag{4.2}$$

The expanded model now accounts for 61% of crop yield variation, compared to 33% for (1), albeit at the expense of additional parameters in its mean structure. In fact, by assigning a unique parameter to each region, this expanded model has $3 \times 40 = 120$ parameters. By using such models, albeit with potentially modified/additional explanatory variables, several authors have been able to create fairly accurate predictions of crop yield ([71]; [74]). It is important to note that the large number of predictor variables (120) makes this model prone to overfitting; while some authors have used cross-validation to prevent this (i.e. [74]), others have sought to further improve model fit by conducting extensive calibration to tune the explanatory variables (i.e. [71]). It is well understood that smoothed, or penalized, models have better prediction properties than larger, more variable models ([43]). This leads us to prefer the most parsimonious model yielding accurate forecasts and to select explanatory variables which provide optimal prediction power for crop yield. While earlier models have been examined with regards to their model fit (as measured through R^2), a much preferred metric is model prediction performance (as measured through cross-validation).

While the availability of SIs for every day of the growing season (in our case April 1 to September 30) means its vector of measured values is of very large dimension, good modelling practice requires that this dimension be reduced before introducing the vector into the regression model. At one extreme, we could do what we did previously, and use just the mean of these daily SI values over the growing season, a one-dimensional feature, as our explanatory variable. However, that would oversimplify the SI's role, since plant growth is influenced more at certain times than others during the growing season. As an extreme example, if the crop is harvested in early September, the SI values in late September would

Figure 4.2: Correlation of SI and yield over time. Correlations smoothed with Lowess smoothing. From this we see that SI is most correlated with yield in an intermediate part of the season, namely days 80 through 160.



aid little in predicting crop yield. To find a low-dimensional feature that provides good predictive power for crop yield, we could average over a reduced window, that is, exclude SI values early and late in the season ([74]). This reflects the point just made that SIs early and late in the season may not be correlated with crop yield. Figure 4.2 shows this correlation between SI and crop yield for each day in the summer, organized by province. This figure suggests we average over days 80 through 160, rather than the entire growing season. However, this produces only a modest improvement, 60.72% of crop yield's variability now being explained instead of 60.56% using the average over the entire season as before. This plot also reveals large spatial variability, particularly between provinces. We explore this issue in more detail later.

There exists considerable scope for tuning this window; for instance [71] select

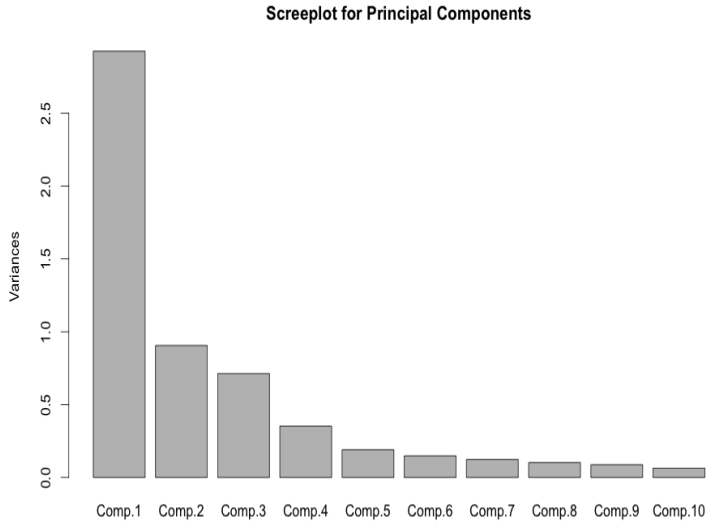
unique window start and end points for each region to achieve an excellent fit – over 75% of variation explained. However such tuning entails much attention to detail. On top of the upper and lower limits for the averaging to take place, [71] calibrate potential available soil water capacities, the maximum number of sowings and the rainfall amount triggering planting in each region. In other words, in addition to the corresponding regression coefficients, this tuning in effect adds 5 additional parameters per region, which in our case would increase the number of parameters being fitted in (2) from 120 to 320, leading most likely to serious over-fitting when considering that such an approach still uses an average SI over the growing season, not accounting for temporally-varying impact of SI. To quote John von Neumann:

> "With four parameters I can fit an elephant and with five I can make him wiggle his trunk."

As such, a preferred alternative would be a lower dimensional feature which captures the key components of the stress index. In addition, we would like to include information which allows for the impact of SI on yield to vary over the growing season.

To capture more information from the SI values than would be available from simple averaging, we extract the principal components and hence main sources of variation from the stress index. To be more precise, after subtracting the average SI from each day, the first principal component is the linear combination of growing season SI values that accounts for the most variability in the SI values. The second, which is orthogonal to the first, explains the next largest amount of variation, and so on. Each observation, in this case each region – year combination, also has a set of loadings that, when multiplied by the corresponding principal components,

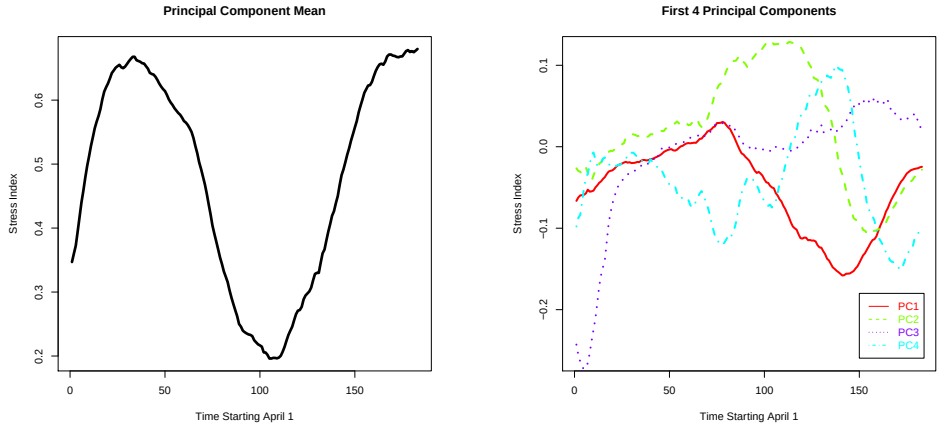
Figure 4.3: Screeplot showing the principal component variances in decreasing order. From this, we see that the first 4 principal components explain much of the variation in SI.



return the original observation. Figure 4.4 shows the subtracted mean process as well as the first four principal components that together show the SIs history over the growing season of our study. Four principal components are chosen as a result of examining a screeplot of the principal components, see Figure 4.3.

Figure 4.4(a) reveals firstly the primary shape of the stress index, showing that initially – from April 1 – the stress is moderate, increasing until May, followed by a gradual decline until it bottoms out in July. It then returns to its highest values by the end of September. The first component (Figure 4.4(b)), which describes 46.9% of the variation in SI, captures a valley in the SI cycle around late August. The second component, which accounts for 14.6% of the variation, shows SI's decline into its July valley followed by its rise to its early September peak. The orthog-

Figure 4.4: Principal components and mean for SI. This figure depicts the major patterns in the variation of the stress index (unitless) over the growing season. Observe how the first four principal components pick up deviations from the overall pattern in (a), and reveal the peaks and valleys of the stress cycle over the course of the summer induced by things like patterns in precipitation and temperature. Together these four components capture most of the variation in stress in a very economical way and eliminate the need for the high dimensional vector of daily SI values.



(a) (b)

onality of the first two components is apparent from (b). The third component of SI's variation captures its low April start. Altogether, the first 4 principal components account for 78.5% of the variation in SI over the growing season. Beyond 4 principal components we observed little in terms of improved modelling fit and a reduction in prediction performance, as shown in figures 4.5 and 4.6. Thus by including the loadings for these 4 principal components as explanatory variables, we have created a 4-dimensional feature which accounts for a large proportion of variation in the stress index.

Note that the first SI principal components aren't necessarily the best predictors

of yield. However, LASSO – a penalized least squares variable selection method – in fact selects these same four principal components as the best four ([43]). This choice of feature also has a natural biophysical interpretation. For instance, a large and positive regression coefficient for the loadings corresponding to principal component 3 would imply that a reduction in stress in early April is highly connected with increased crop yield. By using this approach, the explained variance of the regression model increases from 60.56% from averaging SI over the growing season to 70.06%. In addition, as discussed above, the inclusion of principal components allows the user to gain intuition about the effect different seasonal patterns in the stress index will have on crop yield in a way averaging across the season can not. Using these principal component loadings, our new model is

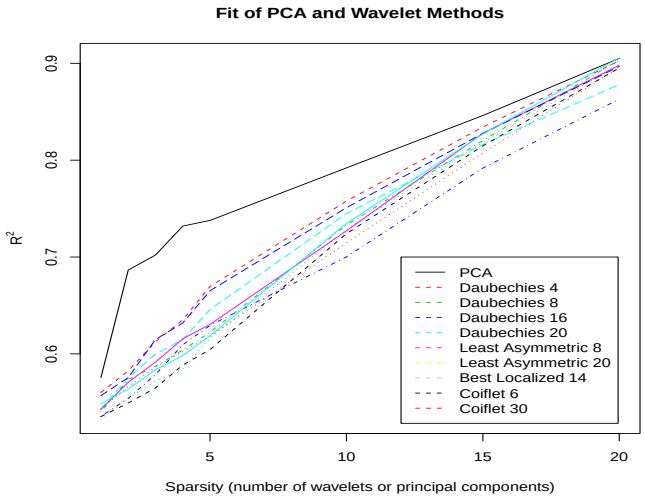
$$y_{j,t} = \beta_{0,j} + \beta_{1,j}t + \beta_{2,j}PC1_{j,t} + \cdots + \beta_{5,j}PC4_{j,t} + \varepsilon_{j,t}. \tag{4.3}$$

where $PC1_{j,t}$ indicates the loading for principal component 1 in region j and year t.

Alternative bases and levels of sparsity

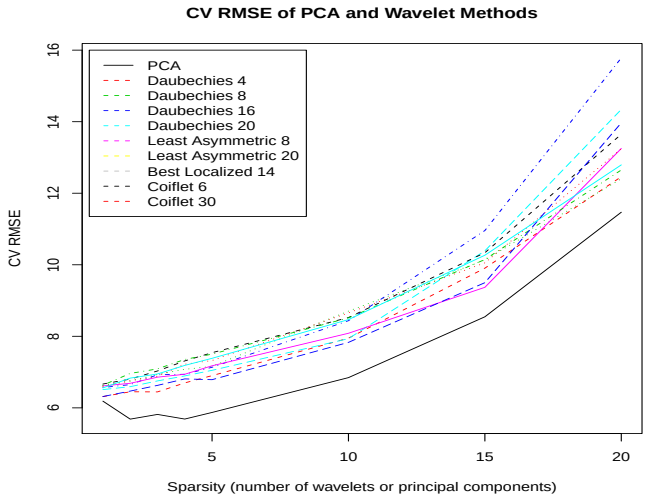
Because of their widely documented ability to model complex nonlinear signals while maintaining sparsity, we briefly explore wavelet bases as an alternative to principal components ([58]). Specifically, we examine a variety of different wavelet bases and levels of sparsity both in terms of cross-validated prediction error as well as R^2. Because SI is inherently piece-wise smooth due to the impact of precipitation resulting in changepoints in the SI values, wavelets provide a valuable alternative to principle components. Figure 4.5 plots R^2 of the yield model for

Figure 4.5: R^2 of the crop yield model for a range of bases and sparsity levels. From this we notice that principal components (PCA) provide better model fit for all sparsity levels.



various bases and levels of sparsity. From this plot we see that principal components dominate in terms of model fit. While R^2 measures how well a model fits to data, it is not a good indicator of a model's prediction abilities. As such, Figure 4.6 plots cross-validation root mean squared error (in bushels per acre) for each basis and sparsity level. Once again we observe that principal components outperform wavelets. From these figures we conclude that principal components lead to a model with better fit and prediction performance. This example highlights the need to be selective in the choice of basis to represent stress index and other variables in such a model. While wavelets excel at representing piece-wise smooth models in a very sparse way (requiring the storage of only 1 vector – the mother wavelet – as well as a series of indices), this is also their downfall in some circumstance such as this one which require a richer representation.

Figure 4.6: Cross-validation RMSE (bushels per acre) of the crop yield model for a range of bases and sparsity levels. From this we notice that principal components (PCA) provide better prediction than the wavelet bases for all sparsity levels.



4.2.2 Incorporating temperature

Temperature affects a plant's development and growth in a variety of ways, in particular its photosynthesis and respiration. In general, temperature affects plant functioning through its action on enzymatic reactions. At low temperatures, enzyme proteins are not sufficiently flexible to complete the conformation necessary for enzymatic reaction. Conversely, high temperatures can coagulate the enzyme leading to similar barriers to the reaction. Alongside a minimum and maximum temperature to allow growth, most plants have an optimum temperature to encourage growth. For instance, [77] conclude that the minimum and optimum temperatures for wheat are respectively 0 and 20-25 degrees celsius. As a result of temperature's influence on plant development, we suspect that its inclusion into the model will result in prediction performance gains. In addition, by directly in-

cluding temperature effects interpreting impacts of climate on yield is made more straightforward.

Growing degree day

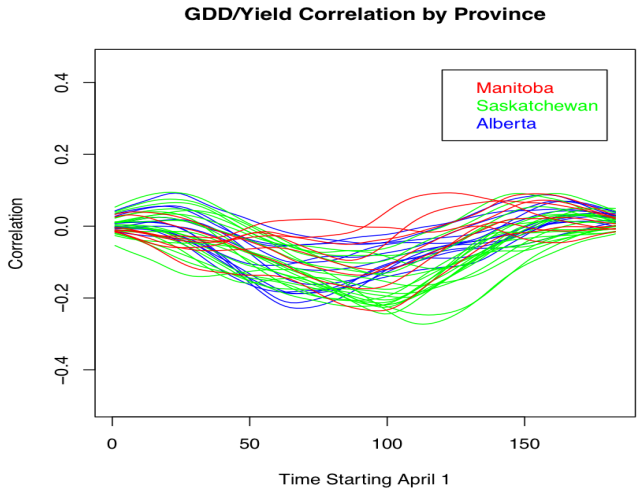
While temperature could go directly into the model, its measurement in hourly or smaller increments creates a considerable amount of data. As a result, some dimension reduction is needed to limit the number of explanatory variables. One could do this using just the maximum and minimum daily temperatures or better still, a one dimensional summary that combines the two. Thus 'growing degree day' (GDD) measures the heat accumulation in a region based on local weather by taking an average of the daily minimum and maximum and subtracting a base temperature as follows:

$$GDD = \max\left(0, \frac{T_{max} + T_{min}}{2} - T_{base}\right).$$

Thus the GDD measures the daily average temperature but in a way that reflects the extremes more sensitively. The base temperature represents the physiological temperature below which development would be zero.

A day with a high and low of 30 and 15 degrees celsius and a base temperature of 10 degrees would have a GDD value of 12.5 degrees celsius. Thus GDD is a simple, single-dimensional summary for describing the plant's exposure to heat. While GDD is a simple heuristic, it is commonly used by horticulturists to estimate the stages of a plant's growth. As an example, the maturation of wheat corresponds to about 1600 GDDs ([28]). Thus GDD provides us with a simple low-dimension summary of temperature, allowing for comparison of the thermal time available in

Figure 4.7: Correlation of GDD and Yield over Time. Smoothed with lowess smoothing. From this we see that a reduced window average may be appropriate.



different climatic zones.

While SI gives scientific insight into the moisture available for plant growth, it says little directly about the heat available to the crop. Thus to improve our model we can also include GDD, which up until now has been used primarily in this context for tuning the explanatory variables ([72]). Like SI, GDD is a daily value, and hence can be treated similarly. Thus through the correlations plotted in Figure 4.7 we look at the time of season where GDD is most correlated with yield. This figure tells us that an appropriate window would be the one bounded by days 50 through 160. Using a cumulative average over the whole season, the explained variation in yield increases from 70.06% to 73.20%, with the shortened window performing similarly. As emphasized by others (see, for instance, [72]), including GDD accounts for the biophysical phenology of the crop as well as improves

interpretability of the model. Hence while the prediction improvements are minimal, the variable's inclusion is an important step in the development of a crop yield model. In addition to averaging over the whole season or a shorter window, we can also use principal components as we did for SI above. While using the first 4 principal components only increases this to 76.48%, the additional 120 variables result in reduced cross-validation prediction performance, hence we prefer using just the windowed average. We emphasize that this is a user and case-specific choice, and for alternate purposes different choices may be preferred. The expanded LS model 4.3 then becomes

$$y_{j,t} = \beta_{0,j} + \beta_{1,j}t + \beta_{2,j}PC1_{j,t} + \cdots + \beta_{5,j}PC4_{j,t} + \beta_{6,j}\overline{GDD}_{j,t} + \varepsilon_{j,t}, \qquad (4.4)$$

where $\overline{GDD}_{j,t}$ is the windowed average of GDD in region j, year t. It is worth noting that temperature is a component of SI; however, the addition of GDD into the model improves both model fit and prediction, thereby eliminating concerns about the deleterious effects that collinearity between the two covariates might introduce.

We compare the previous models as well as those developed later in the chapter in Table 4.1, showing the features and performance of each successive model.

The traditional regression models represented in Table 4.1, fitted for each region separately, ignore a considerable amount of information. Specifically, because of the close spatial proximity of the regions, considerable strength may be gained by exploiting the correlation among regions. For instance, use of neighbouring SI values can help stabilize predictions based on SI values, since the latter come from a small set of regional monitoring stations and hence can be fairly noisy. The

Table 4.1: Features of various models. We see that while model 4 has the best fit to the data ($R^2 = .73$), the Bayesian model gives the best prediction performance in terms of cross-validated root mean squared error (in bushels per acre). Effective parameters is defined as $tr(S)$, where $\hat{y} = Sy$, and may be cosidered a measure of model complexity ([43]).

Model	Parameters	Effective Parameters	R^2	CV RMSE
1: Single LS	3	3	.33	6.83
2: LS (SI)	120	120	.61	5.79
3: LS (PCA)	240	240	.70	5.72
4: LS (PCA+GDD)	280	280	.73	5.69
5: Bayes	283	139	.70	5.35

amount of borrowed strength can be considerable when the correlation between stations is high. In addition, modelling all stations jointly while incorporating spatial information allows us to continue to make predictions even in the presence of missing or noisy data. If a measuring station goes out of operation temporarily, its missing values may be inferred from data collected at nearby regions to yield accurate forecasts. This idea leads into our next section, which focuses on spatial models that look at all regions together in a unified manner.

4.2.3 A context-specific spatial Bayesian approach

Classical regression methods rely on the assumption that their model residuals are uncorrelated. Indeed violation of that assumption can have very serious deleterious effects on parameter estimates compared, for example, to violations of the assumption that those residuals have a Gaussian distribution ([24]). In our case the residuals are most certainly spatially dependent and thus the actual amount of information in the data can be much less than the assumptions underlying (1) would suggest. The unwary analyst would then be led to make overconfident forecasts

with parameter estimates which vary considerably from one region to the next, yet have unduly small standard errors.

One work-around would model the regions separately. However, this wastes the benefits spatial dependence provides by borrowing strength, telegraphing information across the regions through the wires of correlation for the mutual improvement of all their forecasts. This progression naturally leads us to a Bayesian framework for handling this problem, one which jointly models all regions simultaneously while accounting for their spatial dependence. Thus we move from the frequency paradigm of classical statistics to the Bayesian paradigm of modern statistics.

These two paradigms, which tend to give similar inferences at least for fairly large datasets, are very different in concept. Frequentists see data as being generated by a system governed by some true but unknown parameters. They commonly seek to estimate these true parameters well in some sense, for a variety of inferential purposes such as forecasting. The central tenet of their theory is repeated sampling – in the long run the parameters can be estimated to arbitrarily high levels of precision if the system producing the data were unperturbed. However, Bayesian statisticians reject the notion of repeated sampling as a fundamental construct in their theory, recognizing realistically that most systems cannot remain unperturbed and pump out replicate data over an extended sequence of trials. Although their models involve uncertain parameters, these parameters like all uncertain objects such as future data values, are characterized by a probability distribution. Initially that distribution, called a prior, simply reflects the Bayesian's own knowledge. An abundance of such knowledge would mean a prior concentrated around a single point and a state of near certainty. The information in the data adds to the state

of knowledge through the celebrated Bayes theorem. The latter relies on the likelihood function of the uncertain parameters which captures all the information in the data. A likelihood tightly concentrated around a single value would mean the data has eliminated much of the uncertainty about the parameters. However generally, Bayes rule needs to be applied to get the combined effect of data and prior knowledge; this yields the Bayesian's updated prior, or the so-called posterior distribution. Due to its adaptability and ease of use, Bayesian inference has become a prominent fixture in modern spatial statistics, and in particular the modelling of random spatio-temporal fields ([5]; [56]).

Available prior information

Consider, for example, the spatial structure discussed above. Even before estimating the parameters in equation 4.3, we expect parameters in adjacent regions to be similar. Thus we would be surprised if the parameters relating GDD to yield had completely opposite signs in two neighbouring regions. This reflects our prior beliefs about those parameters, namely that knowledge of one would tell us something about the other. More simply, we would see them as stochastically dependent in the language of the probability distribution that characterizes our beliefs about them. We might even have some idea of their approximate magnitudes. For instance, a magnitude of 100 (bushels per acre/degree celsius) for the coefficients $\beta_{6,j}$ for GDD would be completely untenable, since it would mean that changing one cold day to a warm one (adding, say, 10 GDD over the entire cumulative season), would increase the yield by roughly 10 bushels per acre. Thus even without formalizing our beliefs in a prior distribution, loose bounds on parameters are almost always apparent.

Application of the Bayesian approach starts by characterizing our beliefs about the parameters in the form of a prior distribution. In the regression models introduced above, this would amount to a joint prior distribution on each β to account for our belief in their dependence (similarity) for adjoining regions. For simplicity, stack all of the coefficients into a vector $\boldsymbol{\beta}$, the first 7 coefficients being for all variables in region 1, the next 7 for region 2, and so on. Assuming a Gaussian distribution as a convenient prior form, we can explicitly write the prior as follows:

$$\boldsymbol{\beta} \sim N\left(0, \Sigma_0 \otimes g\Omega\right). \tag{4.5}$$

By using such a Kronecker structure, Σ_0 models the correlation within a given coefficient across space, while $g\Omega$ corresponds to Zellner's g-prior ([89]) with Ω the 7×7 empirical covariance between explanatory variables. As such, this choice of prior fits within the empirical Bayes paradigm. While specifying the coefficients (particularly the intercept) to have zero-mean seems restrictive, we note that a priori it does not seem unreasonable to assume that in the presence of full stress (signaling the complete absence of water) the crop yield would be zero. We now specify Σ_0, the correlation between regions, as

$$\Sigma_0 = \exp(-D/\phi), \tag{4.6}$$

with a slight abuse of notation where D is the matrix with element (i, j) the Euclidean distance between regions i and j (as measured from the center of the region). Here ϕ is a parameter controlling the rate of decay of correlation as distance increases. In this way, ϕ controls how spatially smooth the coefficients are, while

g controls how tight around zero the coefficients are. While we are not convinced that wheat is planted in more than the southern section of the three northern-most regions, including the region centers rather than some more southerly geographic location is conservative, as the true geographic center is further removed from adjacent regions, and therefore the correlation expressed through Σ_0 is decreased. Others have proposed more complex spatially-varying models which rely on Markov chain Monte Carlo for inference (e.g. [5, 35]); however, our goal of an online implementation restricts us to models with analytic tractability.

While we suspect neighbouring regions to be similar, Figure 4.2 highlights the differences between provinces. While one could include an indicator variable to allow for provincial effects, accounting for provincially-varying coefficients would involve a considerable number of interaction terms. In fact, the varying irrigation and technology policies in each province result in a sharp boundary between provinces for several of the mean parameters. As such, it is not entirely logical to use a stationary prior ([20]) which assigns correlation between regions solely based on distance without any respect for political boundaries. As a result, we adjust our prior distribution to have reduced correlation between regions in different provinces. While the obvious approach is to scale down the prior correlation between regions in different provinces with a constant value, this may lead to non-positive definiteness of Σ_0; alternative methods which do not suffer from this problem are therefore needed. We accomplish this task by deforming the physical space, in effect pushing neighbouring provinces apart. Motivated by [78], this artificial distortion of the space results in a stationary prior in the deformed space, yet a nonstationary one in the original space. The distance d (measured in degrees latitude/longitude) by which the provinces are pushed apart in the artificial space

is selected through cross-validation. Searching over the integers from 1 to 10, we find $d = 4$ to give the best prediction performance (CV RMSE of 5.35 vs 5.39 for $d = 0$), intuitively meaning that Alberta and Manitoba are pushed respectively west and east from Saskatchewan by 4 degrees longitude in the artificial space. The end result is a reduction in the off-diagonal elements of Σ_0 corresponding to between-province regions while maintaining positive definiteness. Note that the prior parameters ϕ, g, and d result in an additional 3 parameters in the Bayesian model, as reflected in Table 4.1.

Likelihood and posterior distributions

We begin by employing the likelihood corresponding to (4.4), namely

$$y_{j,t} \sim N\left(\beta_{0,j} + \beta_{1,j}t + \beta_{2,j}PC1_{j,t} + \cdots + \beta_{5,j}PC4_{j,t} + \beta_{6,j}\overline{GDD}_{j,t}, \sigma^2\right). \quad (4.7)$$

In keeping with common practice, we also assign an Inverse-Gamma prior distribution on σ^2 with shape and scale parameters a and b set to be highly noniformative. Before proceeding, we introduce the notation y, the column vector of stacked y_j, and X, the $(31 \times 40) \times 240$ block-diagonal matrix of explanatory variables. Using Bayes theorem to combine our initial knowledge (in the form of prior distributions) and the information provided by the data (in the form of the likelihood), we can obtain the posterior distribution of the parameters. Specifically, for the regression coefficients $\boldsymbol{\beta}$, the marginal posterior is obtained using Bayes Theorem as follows:

$$\pi(\boldsymbol{\beta}|X,y) \propto \int \pi(y|X,\boldsymbol{\beta},\Sigma)\pi(\boldsymbol{\beta}|\Sigma)\pi(\Sigma)d\Sigma. \quad (4.8)$$

Due to the conjugate nature of the prior and likelihood, we are able to analytically complete this integral. The resulting distribution is a multivariate Student-T,

$$\boldsymbol{\beta} \sim T\left(\boldsymbol{\beta}_f, \Psi, n+2a\right) \tag{4.9}$$

where

$$\boldsymbol{\beta}_f = (X^T X + (\Sigma_0 \otimes g\Omega)^{-1})^{-1}(X^T y)$$

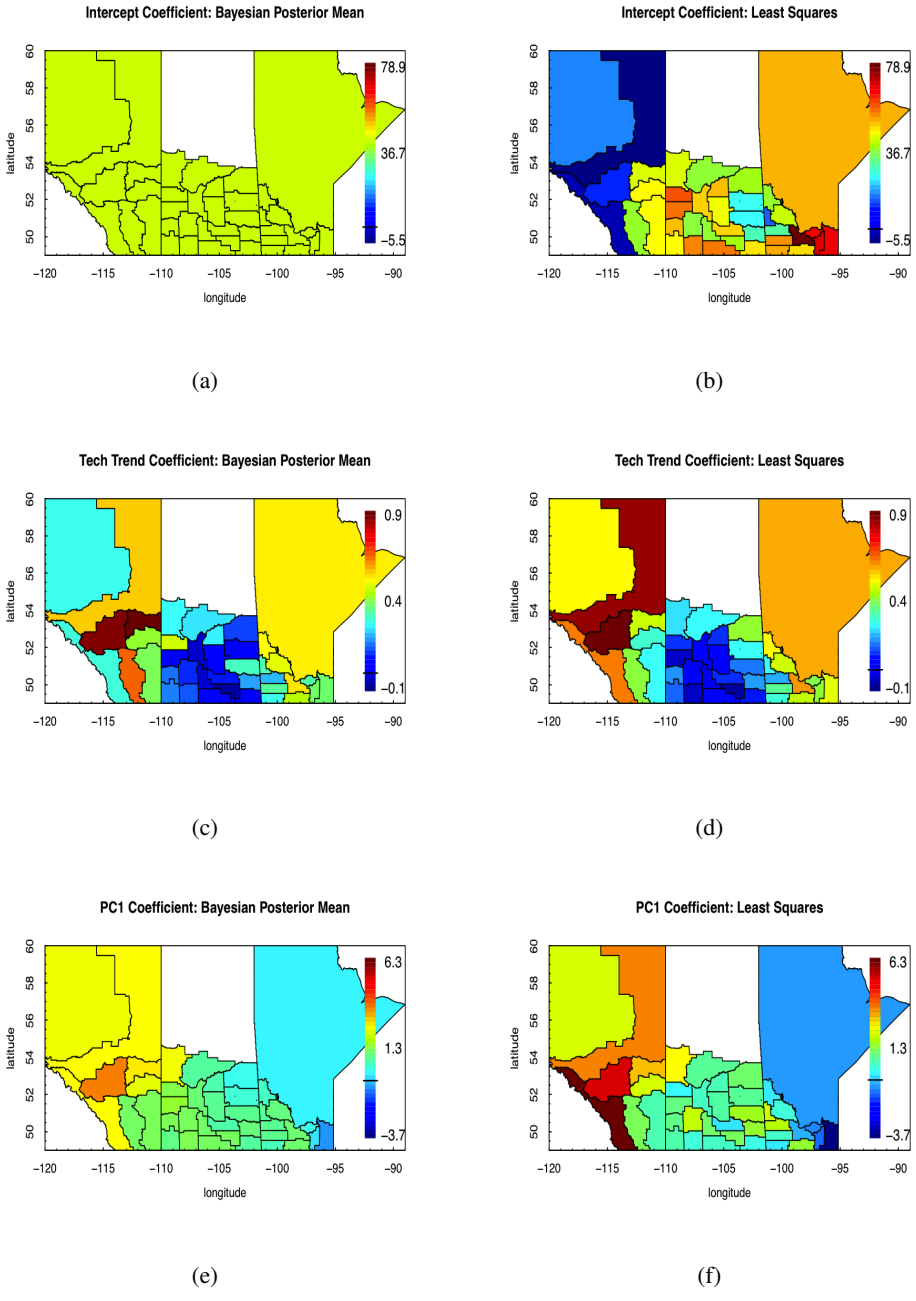
$$\Psi = (X^T X + (\Sigma_0 \otimes g\Omega)^{-1})(SS + 2b)/(n+2a)$$

$$SS = y^T y - \boldsymbol{\beta}_f^T (X^T X + (\Sigma_0 \otimes g\Omega)^{-1})\boldsymbol{\beta}_f^T.$$

From this last expression, we get the posterior mean $\boldsymbol{\beta}_f$, which may be used as a simple estimator for $\boldsymbol{\beta}$. In fact, comparing $\boldsymbol{\beta}_f = (X^T X + (\Sigma_0 \otimes g\Omega)^{-1})^{-1}(X^T y)$ to the LS estimate $(X^T X)^{-1}(X^T y)$, we readily see how the prior covariance affects the parameter estimates. In particular, a diffuse prior distribution adjusts the estimate little, whereas an informative prior distribution – one that is fairly tightly concentrated around zero – shrinks the posterior estimate considerably.

Setting $g = 10$ and $\phi = 10^6$, we obtain coefficient estimates as shown in Figure 4.8, which also shows the corresponding least squares estimate using (4). We see that the spatial information used in the Bayesian model causes the coefficients to be more correlated across space. In addition, the zero-mean prior distribution leads to some shrinkage in the coefficient estimates. Interestingly, we notice little shrinkage in the estimated coefficient for technology trend, suggesting that the data contains considerable information on this quantity.

Figure 4.8: Coefficient surfaces for intercept, technology trend, and PC1. Other coefficients are similarly smoothed.



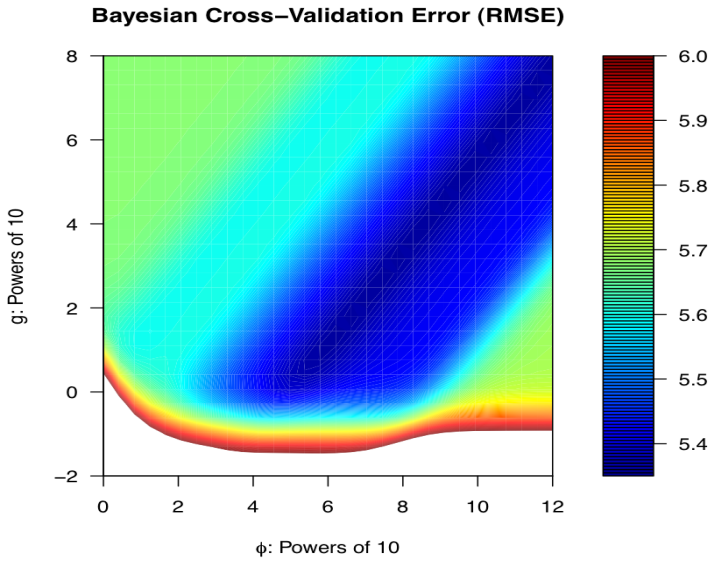
4.3 Results

We proceed by comparing the prediction performance of the least squares and Bayesian methods. To accomplish this we use leave-one-out cross-validation, removing years one at a time in succession to compare each model's predictive ability. More specifically, we successively remove each year in turn, using the remaining years to find the posterior mean, notated $\hat{\boldsymbol{\beta}}^{i}$ if year i is removed. This posterior mean is then used to perform prediction on the removed year. From this the root mean squared error (RMSE) is calculated as the square root of the sum of squared prediction errors for each year and region.

$$RMSE = \sqrt{\sum_{i=1}^{31}\sum_{j=1}^{40}(y_{i,j} - X_{i,j}\hat{\boldsymbol{\beta}}^{i})^2/(31\times 40)}. \tag{4.10}$$

Figure 4.9 shows the cross-validation root mean squared error (RMSE) of the posterior mean estimate for various settings of g and ϕ. As $g \to \infty$ and $\phi \to 0$, the Bayesian model converges to the least squares solution, as evidenced by converging cross-validation errors. However, if g is too small, the prior on the regression coefficients is too informative towards zero, and hence the resulting posterior means are overly shrunken, resulting in poor prediction (RMSE > 6). While one could assign prior distributions to these parameters, we prefer finding them through cross-validation for computational efficiency. Specifically, given the optimal parameters, the model is conjugate, and hence sequential updating and prediction is analytic and therefore nearly instant. It is very interesting to note that the optimal prediction error for the Bayesian model is less than for the least squares model, indicating that prediction is improved with regularization (provided by the zero-

Figure 4.9: Cross-validation errors using Bayesian posterior mean. For comparison, the least squares error is 5.69. From this, we observe a ridge of excellent prediction. Hence there is some tradeoff between the two parameters to be tuned.



mean prior and/or correlation). The area of lowest prediction error occurs along a diagonal of g and ϕ and has value approximately 5.35. This is likely due to the fact that an increase in g results in a more diffuse posterior which regularizes less, while increases in ϕ result in increased correlation between regions and hence more regularization. Hence the optimal prediction seems to occur for moderate amounts of regularization.

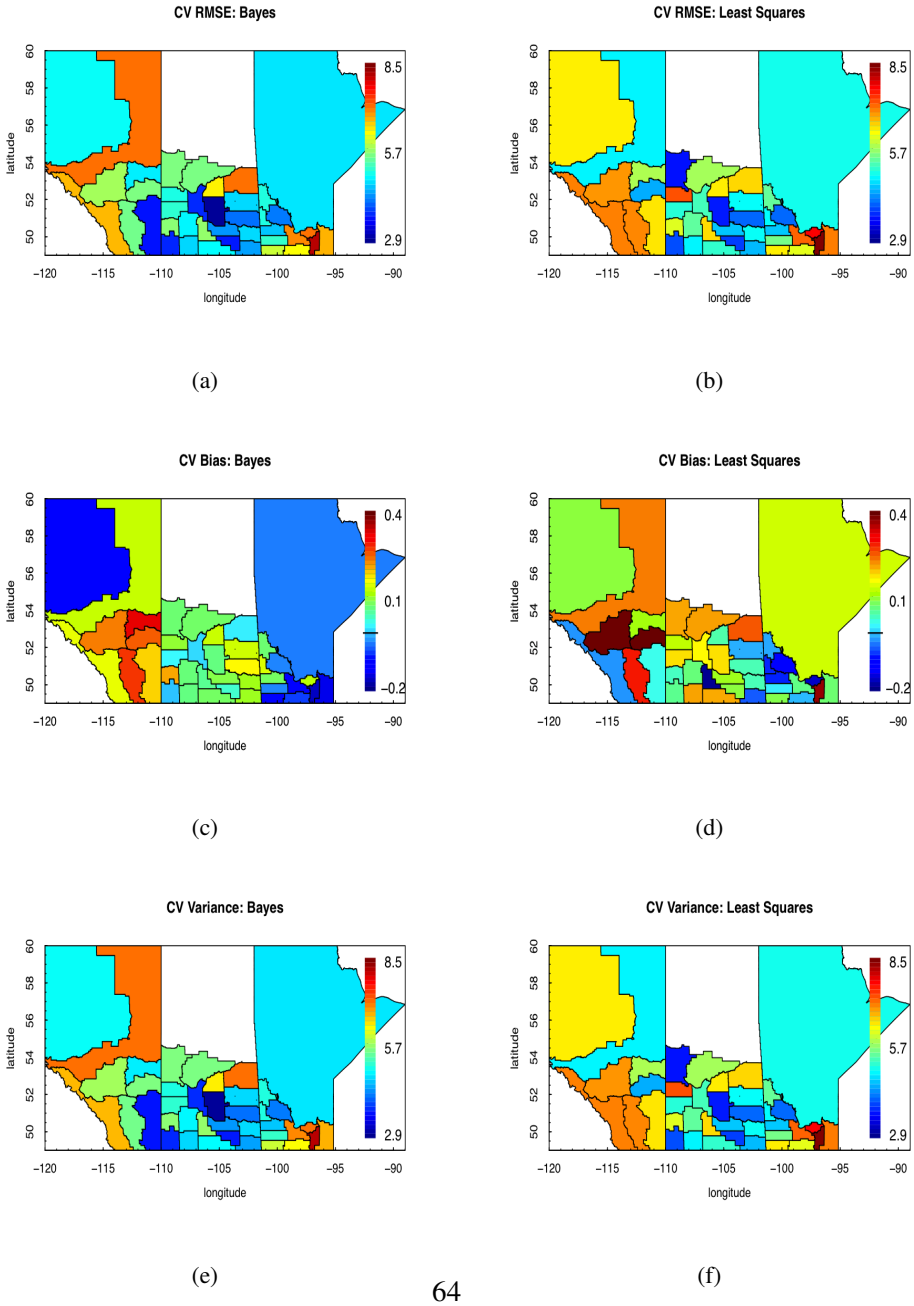
The cross-validation RMSE can also be calculated for each region by summing only over years. In this way we can gain an improved perspective on the model's prediction performance. However, while cross-validation RMSE gives an idea of the prediction performance of a model, it does little to tell of a model's bias. To

do this we decompose the RMSE into the model's prediction bias and variance. Doing this for each region, we obtain Figure 4.10 detailing the prediction RMSE, bias, and variance of the Bayesian and LS models in each region. From this figure we observe that, with the exception of one or two individual regions, the Bayesian model improves RMSE in all areas except for southern Manitoba. Digging deeper, we see a negative bias in this area. Thus the regularization of the model is perhaps not useful in this region due to some systematic differences in this area. Specifically, this section of southern Manitoba is known to use significant irrigation ([34]). As a result, further model development might be explored in this area to account for irrigation.

4.4 Conclusion

In this chapter we have examined the role of SI in predicting crop yields, emphasizing the need to create a judicious low-dimensional summary in order to improve prediction. Simply averaging SI over the entire season is inefficient, as yield may be insensitive to stress in certain parts of the summer. The traditional solution to this problem is to average over a reduced window of data, hence cutting out those areas lacking in sensitivity from the analysis. However, this one dimensional feature is not particularly sensitive to changes in stress indices within that window. For example, a region which has low SI in June but high SI in July might ultimately have the same averaged value as another region which had just the opposite trend. To address this issue, we have implemented principal components analysis to create a set of flexible summary statistics which better describe the variations in SI, and as a result improve prediction considerably. We also demonstrated principal components' improved performance over wavelet bases.

Figure 4.10: Cross-Validation results by region. The Bayesian model improves prediction by all standards in the majority of regions. We plot root mean squared error (RMSE) as well as the breakdown to bias and variance for the models.



(a) (b)

(c) (d)

(e) (f)

We have also shown the importance of incorporating spatial correlation into crop yield models; ignoring this information can lead to bias both in model identification and prediction. Specifically, we observed that a common least squares fit of crop yield on some explanatory variables over the entire region resulted in biased residual errors, and hence violated the assumptions of the model. To avoid this problem, we could fit each agricultural region with its own model. The problem, however, is that this ignores information between crop regions, and as such we observed reduced prediction power and model identifiability. We addressed this issue through the use of a Bayesian model which modelled all regions together, yet accounted for spatial correlation. This model smooths and stabilizes prediction and also allows for analytic and therefore efficient updating and prediction. In addition, we created a non-stationary prior distribution to address the issue of province to province variability resulting from provincial differences in policy, management, and other factors affecting yield. Through cross-validation, we demonstrated this model to achieve improved prediction performance in modelling Canadian wheat yield over the least squares model which ignores spatial dependence, and hope that others will attempt to replicate our findings in other contexts based on the promise seen in this application.

Chapter 5

Nonstationary Modeling Through Dimension Expansion[1]

5.1 Introduction

Recently there has been great interest in using spatial statistical methods to model environmental processes, with the aim of both gaining an improved understanding of underlying processes and making predictions at locations where measurements of a process are not available. The majority of such methods make the assumption that the underlying process is stationary ([20]) which, for many environmental systems, may be untenable.

In this chapter, we focus on accurately explicating the nonstationary structure that often arises in measurements of atmospheric, agricultural, and other environmental systems. If these systems share one underlying theme it is complex spatial

[1] A version of Chapter 5 has been published. Bornn, L., Shaddick, G., Zidek, J. (2012) Modelling Nonstationary Processes Through Dimension Expansion. Journal of the American Statistical Association. Vol. 107, Pages 281-289. [15]

structures, being influenced by such features as topography, weather, and other environmental factors. For example, the air quality characteristics of cities are likely to be more similar than that of rural areas irrespective of their geographic proximity. Ideally we might model these effects directly; however, information on the underlying causes is often not routinely available. Hence when modeling environmental systems there exists a need for a class of models that are more complex than those which rely on the assumption of stationarity.

In the field of atmospheric science, empirical orthogonal functions have been used to model a nonstationary process as the sum of a stationary isotropic process and a set of basis functions with random coefficients representing departures from nonstationarity ([66, 67]). Current approaches to modeling nonstationary processes in the statistical literature broadly comprise those that (i) combine locally stationary processes to create an overall nonstationary process and (ii) build upon the framework of 'image warping'.

A number of approaches for handling nonstationarity assume that over small enough spatial domains, the effects of nonstationarity are negligible, and hence locally stationary models may be used. This concept is the basis of kernel approaches, early examples of which can be found in [39, 40]. The process–convolution approach ([45, 46]) relies on the notion that a wide range of stationary Gaussian processes may be expressed as a kernel convolved with a Gaussian white noise process, with the kernel being allowed to vary spatially to account for nonstationarity. The form of the kernel allows for a broad expression of potential covariance functions, with a Gaussian kernel corresponding to a Gaussian covariance function and other choices of kernel resulting in other correlation structures. Similarly, [32] suggested modeling the field as a weighted average of local stationary processes

within a set of regions, an idea which was later extended to include a continuous convolution of stationary processes ([33]). Various difficulties still remain in this class of models, including the lack of a complete and easily interpretable global model and the choice of local regions and details of the weight structures. An alternative approach proposed by [78] is that of "image warping", the central idea of which is that a nonstationary process may be stationary in a deformed, or warped, version of geographic space. Multi-dimensional scaling (or related methods) can be used to find the deformed locations with a mapping between the original and deformed space found using, for example, a thin plate spline.

The principal idea underlying the proposed method is that of embedding the original field in a space of higher dimension where it can be more straightforwardly described and modelled. Specifically, we shift the dimensionality of the problem from 2 or 3 dimensions to 4, 5, or more in order to recover stationarity in the process; we term our methodology "dimension expansion." Our starting point is that nonstationary systems may be represented as low-dimensional projections of high-dimensional stationary systems (see, for example, [68]). The method is superficially similar to that of image warping; however, it differs fundamentally in that here the locations in the geographic space are retained, with added flexibility obtained through the extra dimensions. Additionally, it addresses one of the major issues with the image warping approach, namely folding of the space. This occurs in image warping if the estimate of the function that transforms from geographical to deformed space is not injective. As a result of folding, two geographically distinct locations may be mapped to the same location, meaning the variation between them will be incorrectly treated as measurement error and small scale variation (i.e. the nugget), which is expressly appropriate only for collocated and other proximal

monitoring sites. In such cases, mapping quantities such as prediction intervals becomes particularly challenging both in terms of implementation and interpretation.

The remainder of the chapter is organised as follows: Section 2 introduces the dimension expansion framework proposed here, including an illustrative example to demonstrate the fundamental concepts behind the approach. This example is then used to draw comparisons to image warping. In Section 3, dimension expansion is applied to two real life examples. First, the solar radiation dataset originally used in [78] and used as a test-bed in various more recent image warping papers is studied. Second, we study air pollution from seventy-seven monitoring locations in the United Kingdom which show clear signs of nonstationarity. We highlight the ability of dimension expansion to accurately model such data as measured through cross-validated prediction error. Finally, Section 4 provides a discussion and suggestions for future developments.

5.2 Dimension Expansion

While early work ([20]) dealt primarily with stationary models, it is now generally recognized that many spatial processes $\{Y(\boldsymbol{x}) : \boldsymbol{x} \in \mathscr{S}\}$, $(\mathscr{S} \in \mathscr{R}^d)$ fail to satisfy this assumption. Environmental systems might exhibit behaviour that looks locally stationary, yet when considered over large and heterogenous domains they very often exhibit nonstationarity. For ease of notation, we consider $Y(\boldsymbol{x})$ to be a (potentially nonstationary) mean-zero Gaussian process and place our emphasis on representing the nonstationary structure.

A principal task in spatial statistics is estimating a variogram model (or correlation function) to explain spatial dependencies. The theoretical variogram, defined

as

$$2\gamma(\boldsymbol{x_i},\boldsymbol{x_j}) = E\left(|Y(\boldsymbol{x_i}) - Y(\boldsymbol{x_j})|^2\right)$$

is typically modeled using a parametric stationary variogram $\gamma_\phi(\boldsymbol{h})$ depending only on $\boldsymbol{h} = \boldsymbol{x_i} - \boldsymbol{x_j}$, the difference vector between locations, and the parameter(s) ϕ. If the field is nonstationary, such a model will be a misspecification. In response, we transform the set of observed spatial locations $\mathscr{S} \in \mathscr{R}^d$ into one of higher dimension $\mathscr{S}' \in \mathscr{R}^{d+p}$, where $p > 0$ and $\mathscr{S}$ is a subset of the dimensions of $\mathscr{S}'$. Put plainly, such an approach amounts to allowing extra dimensions for the observed locations $\boldsymbol{x_1},\ldots,\boldsymbol{x_s}$, notated as $\boldsymbol{z_1},\ldots,\boldsymbol{z_s}$ such that the field $Y([\boldsymbol{x},\boldsymbol{z}])$ is stationary with a variogram model $\gamma_\phi([\boldsymbol{x_i},\boldsymbol{z_i}] - [\boldsymbol{x_j},\boldsymbol{z_j}])$. Here $[\boldsymbol{x},\boldsymbol{z}]$ is the concatenation of the dimensions $\boldsymbol{x}$ and $\boldsymbol{z}$.

Perrin and Meiring [68] explore this idea in the particular case where both the covariance function and the expansion from $\boldsymbol{x}$ to [$\boldsymbol{x}$, $\boldsymbol{z}$] are known. In their motivating example, they consider the following stationary covariance on the plane:

$$cov(Y([\boldsymbol{x_i},\boldsymbol{z_i}]),Y([\boldsymbol{x_j},\boldsymbol{z_j}])) = \exp(-|\boldsymbol{x_i} - \boldsymbol{x_j}| - |\boldsymbol{z_i} - \boldsymbol{z_j}|).$$

By restricting to the set $\boldsymbol{z} = \boldsymbol{x}^2$ and defining $Y'(\boldsymbol{x}) = Y([\boldsymbol{x},\boldsymbol{x}^2])$, the resulting covariance function on this reduced-dimension field is nonstationary, namely

$$cov(Y'(\boldsymbol{x_i}),Y'(\boldsymbol{x_j})) = \exp(-|\boldsymbol{x_i} - \boldsymbol{x_j}|)\exp(1 + |\boldsymbol{x_i} + \boldsymbol{x_j}|).$$

Perrin and Meiring [68] then consider the reverse problem, proving that a nonstationary random field indexed by $\mathscr{R}^d$ (with moments of order greater than 2)

can always be represented as second-order stationary in $\mathscr{R}^{2d}$. It is not, however, necessary to move from $\mathscr{R}^d$ to $\mathscr{R}^{2d}$ to obtain the existence of a stationary field. Consider a recent result of Perrin and Schlather [69], which states that a Gaussian random vector can always be interpreted as a realization of a stationary field in $\mathscr{R}^p, p \geq 2$, subject to moment constraints on the vector, namely that all components have equal expectation with the covariance matrix having identical components on the diagonal. From this it is straightforward to state that, similarly, a realization of a Gaussian process in $\mathscr{R}^d$ may be interpreted as a realization of a stationary field in $\mathscr{R}^{d+p}, p \geq 2$ (similarly, subject to moment constraints), with the covariance function ignoring the initial d dimensions.

The above results show the existence of higher-dimensional stationary representations for nonstationary fields, yet in the vast majority of situations neither a nonstationary variogram, nor an analytic mapping to higher dimensions, is known. Here we construct a framework for using higher-dimensional representations to model nonstationary systems, with the goal of learning the latent dimensions nonparametrically from information contained within the data.

To learn the expanded, or latent, dimensions $\boldsymbol{z_1}, \ldots, \boldsymbol{z_s}$ we propose

$$\hat{\phi}, \boldsymbol{Z} = \underset{\phi, \boldsymbol{Z'}}{\operatorname{argmin}} \sum_{i<j} (v^*_{i,j} - \gamma_\phi(d_{i,j}([\boldsymbol{X}, \boldsymbol{Z'}])))^2 \tag{5.1}$$

where v^*_{ij} estimates the spatial dispersion between sites i and j, for example

$$v^*_{ij} = \frac{1}{|\tau|} \sum_{\tau} |Y(\boldsymbol{x}_i) - Y(\boldsymbol{x}_j)|^2,$$

with $\tau > 1$ indexing multiple observations of the system, the handling of which

is considered in the discussion, and $d_{i,j}([\boldsymbol{X},\boldsymbol{Z}])$ is the i,j^{th} element of the distance matrix of the (augmented) locations $[\boldsymbol{X},\boldsymbol{Z}]$. Once the matrix $\boldsymbol{Z} \in \mathscr{R}^s \times \mathscr{R}^p$ is found, a function f is built such that $f(\boldsymbol{X}) \approx \boldsymbol{Z}$. While a wide range of options exist, we follow Sampson and Guttorp [78] in using thin plate splines, here one for each dimension of $\boldsymbol{Z}$. Thin plate splines are a mapping from one space to another such that the integral of the squared second order derivatives of the mapping function is minimized. For $d = 2$, this corresponds to minimizing

$$\sum_{i=1}^{s} ||\boldsymbol{z}_i - f(\boldsymbol{x}_i)||^2 + \lambda_2 \int_{\mathscr{R}^2} \left[\left(\frac{\partial^2 f}{\partial^2 x_1} \right)^2 + 2 \left(\frac{\partial^2 f}{\partial x_1 \partial x_2} \right)^2 + \left(\frac{\partial^2 f}{\partial^2 x_2} \right)^2 \right] dx_1 dx_2,$$

and therefore the smoothing parameter λ_2 is analogous to λ_{IW}, the thin plate spline parameter in the image warping framework. Setting $\lambda_2 = 0$ results in an interpolating spline, whereas $\lambda_2 \to \infty$ results in the linear least squares fit. The nonlinear functions f are therefore linear combinations of basis functions centered at the locations $\mathscr{S} \in \mathscr{R}^d$. Once a model is built in the expanded space, f^{-1} will bring us from the manifold in $\mathscr{R}^{d+p}$ defined by $(\boldsymbol{X}, f(\boldsymbol{X})), \boldsymbol{X} \in \mathscr{R}^d$ back to the original space.

Due to our unique formulation, we have $f^{-1}(\boldsymbol{Z}) = \boldsymbol{X}$, and we need not be concerned with the difficulties associated with ensuring that f is bijective as in earlier approaches. Thus we may view the originally observed locations $\boldsymbol{X}$ as a projection from a manifold within a higher dimensional space, $[\boldsymbol{X},\boldsymbol{Z}]$, in which the process is stationary. As an obvious (and direct) example, a process which is stationary given both geographical location and elevation may result in a nonstationary field given only longitude and latitude. Learning the latent dimensions (whether or not they have a physical meaning, such as elevation) means that a stationary model may be

used in the expanded space.

In many situations, it is unclear how many additional dimensions are needed to accurately model the spatial field. One could use cross-validation or a model selection technique to determine the dimensionality of $\boldsymbol{Z}$; however, recognizing that (5.1) might result in overfitting the spatial dispersions, we would also like to regularize the estimation of $\boldsymbol{Z}$. As a result, we modify (5.1) by including a group lasso penalty term on $\boldsymbol{Z}$, where the groups are the dimensions of $\boldsymbol{Z}$ ([87]). The resulting objective function is

$$\hat{\phi}, \boldsymbol{Z} = \underset{\phi, \boldsymbol{Z}'}{\operatorname{argmin}} \sum_{i<j} (v^*_{i,j} - \gamma_\phi(d_{i,j}([\boldsymbol{X}, \boldsymbol{Z}'])))^2 + \lambda_1 \sum_{k=1}^{p} ||\boldsymbol{Z}'_{\cdot,k}||_1. \tag{5.2}$$

where $\boldsymbol{Z}'_{\cdot,k}$ is the k-th column (dimension) of $\boldsymbol{Z}'$. As a consequence of this revised objection function, one need only determine a maximum number of dimensions p and the shrinkage parameter λ_1, whereupon the learned augmented dimensions $\boldsymbol{Z}$ will be both regularized towards zero and sparse in dimension. Hence λ_1 can be viewed as regularizing the estimation of $\boldsymbol{Z}$ and determining the dimension of the problem, whereas λ_2 controls the smoothness of the augmented dimensions; we suggest learning both through cross-validation, although other model fit diagnostics or prior information may be used as well.

It is relatively straightforward to solve (5.2) using the gradient projection method of Kim et al. [52], which conducts block-wise updates for group lasso with general loss functions. Here the blocks are the dimensions of $\boldsymbol{Z}$, and hence the optimization is efficient even for a large number of spatial locations. Optimization details are

given in the appendix. For ease of exposition we use an exponential variogram,

$$\gamma_\phi(\boldsymbol{x_1}, \boldsymbol{x_2}) = \phi_1(1 - \exp(-||\boldsymbol{x_1} - \boldsymbol{x_2}||/\phi_2)) + \phi_3,$$

which works well in the examples that follow, although the method applies analogously to other variograms.

5.2.1 Illustrative example

We now present an illustrative example to help explain the concepts behind this proposed dimension expansion approach, as well as demonstrate the inability of image warping to handle complex nonstationarity. Specifically, we simulate a Gaussian process with $s = 100$ locations on a 3-dimensional half-ellipsoid centered at $(0,0,0)$ such that the projection to the first 2 dimensions is a disk centered at the origin. Here, as throughout, distances are Euclidean. Figure 5.1 plots the empirical variograms for the original 3-dimensional space as well as the 2-dimensional projection, the latter of which results in a highly noisy empirical variogram cloud. Our goal is to recover the lost dimension through dimension expansion by optimizing (5.2) with $\lambda_1 = 0.5$, chosen to induce $\boldsymbol{Z}$ to have one dimension. Here we calculate the matrix of empirical dispersions v^*_{ij} using 1000 realizations of the Gaussian process. Figure 5.2 shows the resulting learned locations as well as the corresponding empirical variogram, where we see that dimension expansion is capable of recovering the lost dimension, resulting in a variogram that closely reproduces the original.

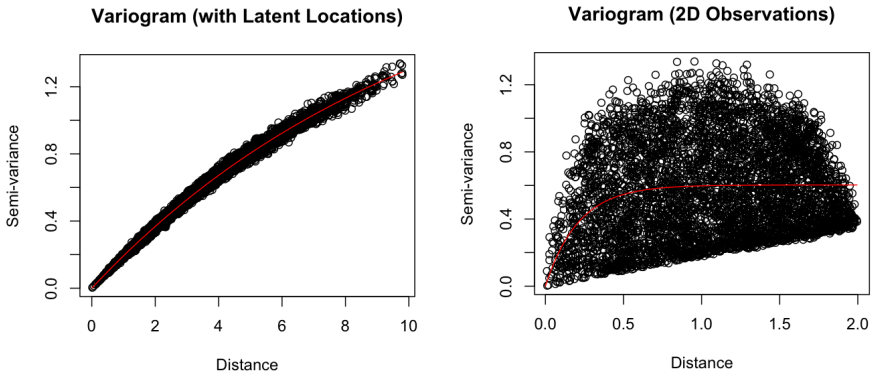


Figure 5.1: Empirical variograms from the original process (left) as well as a 2-D projection (right) on the illustrative ellipsoid example. A fitted exponential variogram is shown by the solid line.

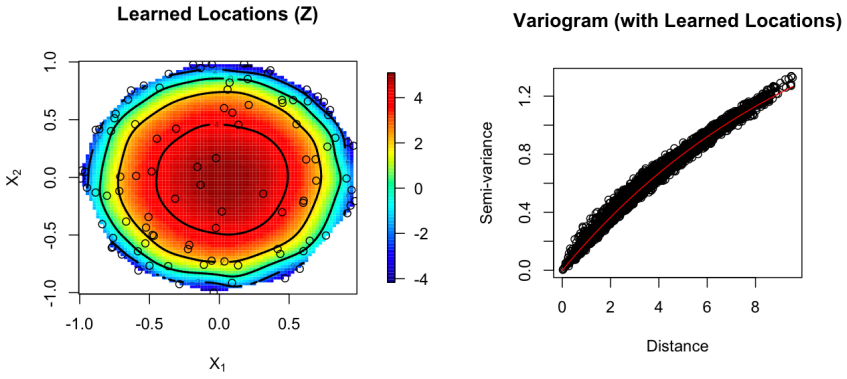


Figure 5.2: Learned latent locations (left) using $\lambda_2 = 10^{-4}$ as well as corresponding empirical variogram (right) after dimension expansion is applied. The fitted exponential variogram is shown by the solid line.

5.2.2 Image warping and folding

In the image warping approach, Sampson and Guttorp [78] employ non-metric multidimensional scaling to move the locations along the geographic space, followed by fitting of the variogram γ_ϕ using traditional variogram fitting methods. From this, a function f is found to go from the original to the warped locations, and back via f^{-1}. A number of adaptations of this approach have been proposed. [82] proposed modeling the covariance function as a linear combination of radial basis functions using maximum likelihood (as suggested by [59]). [63] and [64] noted that the multi-stage algorithm of Sampson and Guttorp does not correspond to a unified optimization problem and instead propose finding the locations and fitting the variogram using a single objective function, an approach also pursued by [62]. It is worth noting that [63] also explore mappings from $\mathscr{R}^2$ to $\mathscr{R}^3$ in the context of analyzing acid rain data, as the same-dimension mapping was incapable of describing the nonstationarity arising in the observed field. In a similar vein, [48] propose using simulated annealing to fit the spatial deformation model. Rather than imposing smoothness on the deformation through thin plate splines, they use Delauney triangulation to constrain the mapping f from folding on itself. In order to acknowledge the uncertainty associated with the deformed locations, [22], [79], and recently [80] have proposed Bayesian implementations of this approach, the latter additionally using observed covariate information to warp into higher dimensions.

As described in the introduction, the image warping framework can suffer from problems of folding, namely of f not being bijective (See Zidek et al. [90] for a particularly extreme example of folding). Considering the illustrative example of

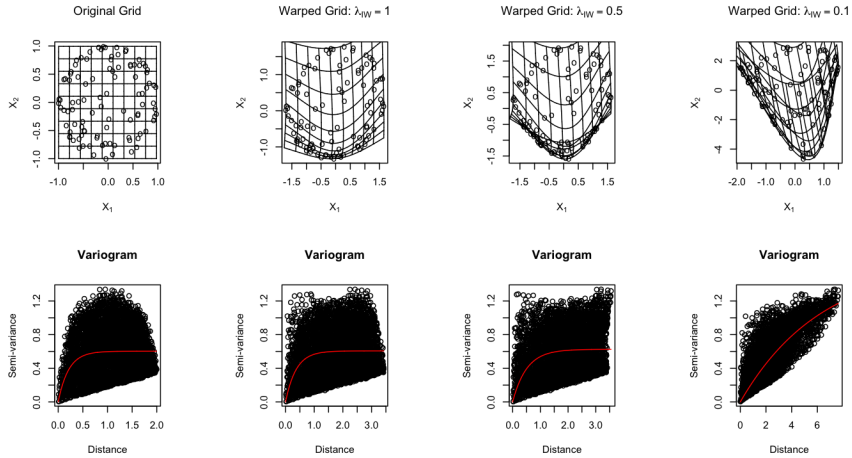


Figure 5.3: Warped grid of locations (top) and corresponding variograms (bottom) for various settings of the thin plate spline parameter λ_{IW} using the image warping technique of Sampson and Guttorp [78].

Section 2.1, admittedly designed to be illustrative of such folding, we apply the image warping technique ([78]) with f modeled as a thin plate spline. Because the image warping framework contains no term similar to λ_1 to regularize the warped locations, smoothing must be done through the thin plate spline parameter λ_{IW} (analogous to λ_2 in the proposed dimension expansion framework). Figure 5.3 shows the warped grids and resulting empirical variograms for various settings of λ_{IW} applied to the ellipsoid example introduced in Section 2.1. We observe immediately that for a highly penalized f (corresponding to large λ_{IW}) the space does not fold; however, the variogram fit is very poor. As λ_{IW} is relaxed to improve the fit, the space begins to fold, highlighting a potentially serious problem with the image warping framework – an issue which is addressed in the dimension expansion paradigm proposed here.

Also related to the proposed dimension expansion method are latent space

models such as that proposed by Hoff et al. [47]. Here, latent dimensions are used to help learn a network of relationships between individuals. Recent work in the field of spatial statistics has also exploited latent dimensions to ensure valid cross-covariance functions in multivariate fields. Specifically, Apanasovich and Genton [2] use latent dimensions for the different variables in order to build a class of valid cross-covariance functions.

5.3 Applications

We now present two applications of dimension expansion applied to the modeling of nonstationary processes using two real data sets. The first uses the solar radiation data ([44]) studied in the original image warping paper of Sampson and Guttorp [78]. The second consists of measurements from a network of air pollution (black smoke) monitoring sites in the UK, further details of which can be found in [29].

5.3.1 Solar radiation

The data of Hay [44] includes measurements of solar radiation from 12 stations in the area surrounding Vancouver. Due to the location and elevation of station 1 (Grouse mountain), the field is inherently nonstationary, as exhibited by the sample variogram (Figure 5.4). This figure shows the original and warped locations using Sampson and Guttorp's image warping approach with corresponding variogram plot. Image warping moves the locations (in particular the station at Grouse mountain, which is the northernmost location) to achieve something closer to stationarity. It is worth noting that overfitting may be controlled through the parameter λ_{IW} of the thin plate spline.

Figures 5.4 and 5.5 show the results of applying the dimension expansion ap-

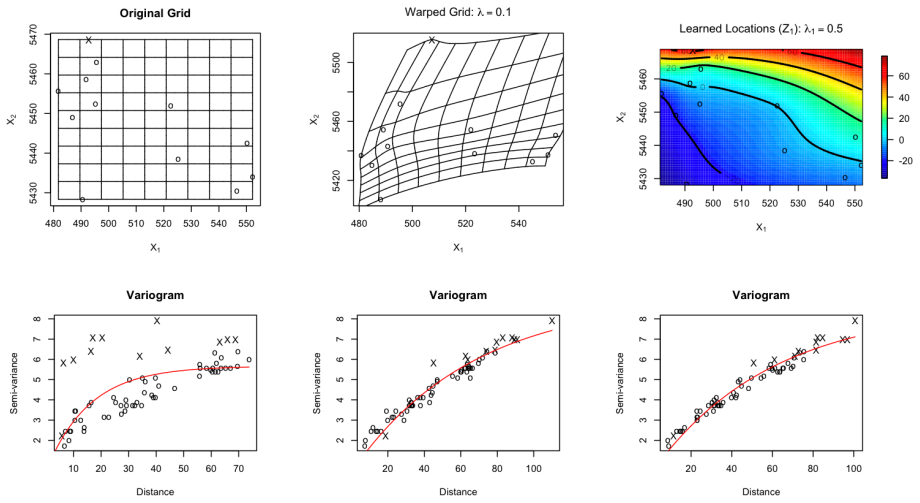


Figure 5.4: Original locations and empirical variogram for the solar radiation data (left); warped locations and associated empirical variogram using image warping with $\lambda_{IW} = 0.1$ (centre); learned locations with associated empirical variogram using dimension expansion with $\lambda_1 = 0.5, \lambda_2 = 10^{-4}$ (right). The units for the semi-variance are $(MJm^{-2}\text{day}^{-1})^2$, and for distance are *km* (UTM coordinates, divided by 1000). The fitted variogram is shown by a solid line, and points associated with Grouse mountain (station 1) are highlighted using an "X".

proach using $\lambda_1 = 0.5$ and $\lambda_1 = 0.2$, respectively, using a maximum number of dimensions of $p = 5$. The original locations are shown, as well as the added dimensions ($\boldsymbol{Z}$). With $\lambda_1 = 0.5$ (Figure 5.4, right), dimension expansion adds one additional dimension which primarily serves to push Grouse mountain out of the plane, reflecting the a priori suggestion that it is elevation that leads to the station's spurious correlation pattern. Interestingly, the contours of the learned dimension closely resemble the elevation contours of the mountains surrounding the Vancouver area. With $\lambda_1 = 0.2$ (Figure 5.5), 2 extra dimensions are used, and the fit of the parametric variogram improves marginally. We can therefore see how λ_1 in-

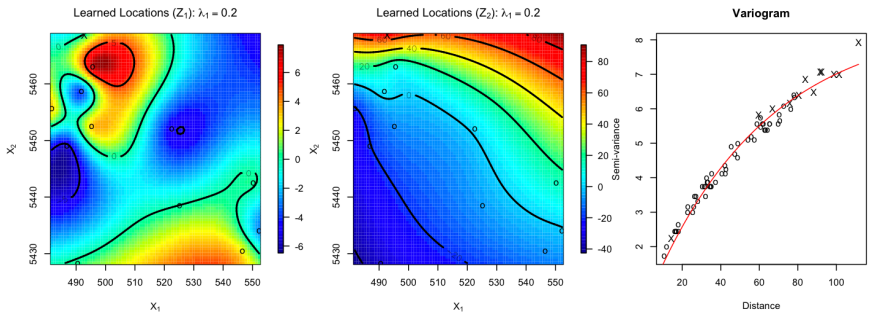


Figure 5.5: Dimension expansion of the solar radiation surface using $\lambda_1 = 0.2, \lambda_2 = 10^{-4}$. Z here is 5 dimensions, with Z_3, Z_4, and Z_5 set to zero as a result of the sparsity-inducing penalization. The first two panels show the learned locations, and the right-most panel shows the associated empirical variogram (fitted variogram shown in red). The units for the semi-variance are $(MJm^{-2}\text{day}^{-1})^2$, and for distance are *km*. Points associated with Grouse mountain (station 1) highlighted using an "X".

fluences the number of extra dimensions used, as well as the shrinkage in each dimension, in order to control the level of fit.

5.3.2 Air pollution

The data consists of annual average concentrations of black smoke (μgm^{-3}) over a period of sixteen years from 77 locations within the UK operating between April 1978 and March 1993 (inclusive) and was obtained from the Great Britain air quality archive (www.airquality.co.uk). Sites were selected in areas defined wholly or partially residential and measurements were aggregated to ward level (based on the 1991 census) using a geographical information system ([29]). The majority of wards contained a single site, but where there were more than one, records were either joined together if the time periods did not overlap or averaged if time periods of operation were simultaneous. Due to similarities in levels of air pollution in urban locations, even if they are not geographically close, the field is known

to be nonstationary. Specifically, we see in Figure 5.8 reduced empirical dispersions for distances around $280 - 290km$ (the distance between London and Liverpool/Manchester), indicating that these urban centers are more similar than their distances would suggest. Our goal is to uncover and explore this nonstationarity through the dimension expansion framework.

We begin with cross-validation to learn the optimal setting of the parameters λ_1, λ_2 using (5.2) as described in Section 2. Figure 5.6 shows the resulting cross-validation RMSE for various parameter settings. We can see that moderate values of both λ_1 and λ_2 result in the best prediction performance. As λ_1 increases to its highest value ($10^{4.5}$), no dimension expansion occurs, and hence λ_2 has no impact. From this it is straightforward to see that the use of the original geographic space is a special case of the dimension expansion framework.

Using these parameter values, the dimensionally sparse optimization (5.2) used by dimension expansion leaves all but one dimension of $\boldsymbol{Z}$ set to zero. This dimension is shown in Figure 5.7, where we see a strong ridge connecting London, Birmingham, Liverpool, and Manchester. Hence in the extra dimension major cities are moved closer together while rural areas are pushed further away. The variograms before and after the dimension expansion are shown in Figure 5.8, where we see no indications of nonstationarity after dimension expansion is applied.

5.4 Discussion

By augmenting the dimensionality of the underlying geographic space, we have developed a highly flexible approach for handling the nonstationarity that often arises in environmental systems. While ostensibly similar to image warping, the proposed method avoids the issue of folding and allows one to model much more

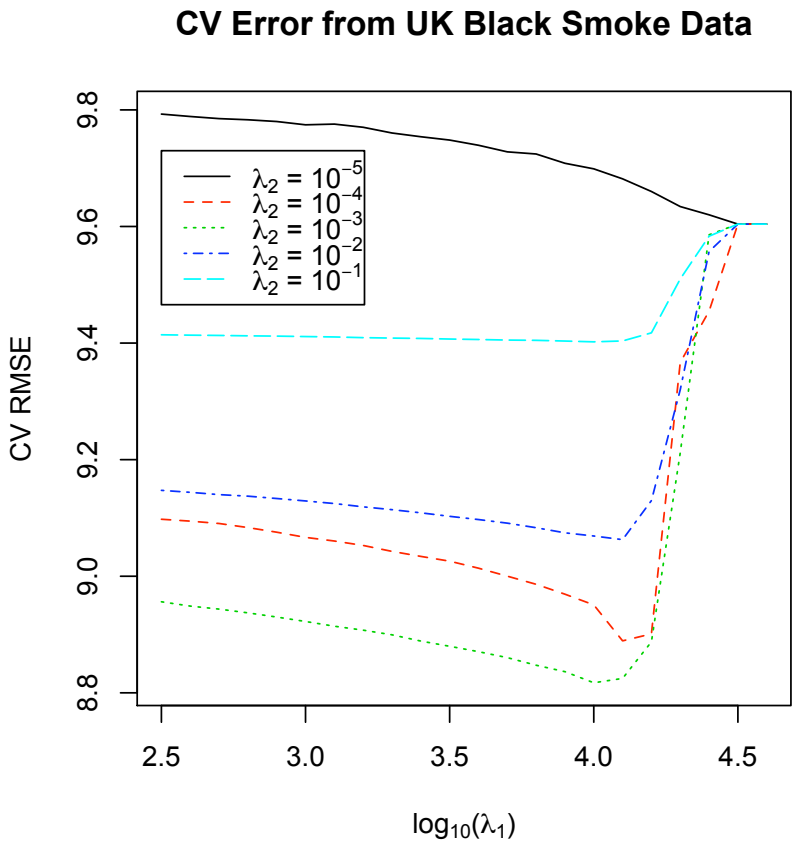


Figure 5.6: Leave-10-out cross-validated prediction error of dimension expansion applied to the UK black smoke data. Here we see optimal prediction for moderate values of both λ_1 and λ_2.

complex nonstationarity patterns through interdimensional expansions, allowing the user to perform nonparametric learning of the mapping function. In addition, through the use of a group lasso penalty, we are able to estimate the number of augmented dimensions, as well as regularize the optimization problem. Lastly, we have highlighted the practical application of the dimension expansion approach in three examples, two of which use data from observed environmental processes. It

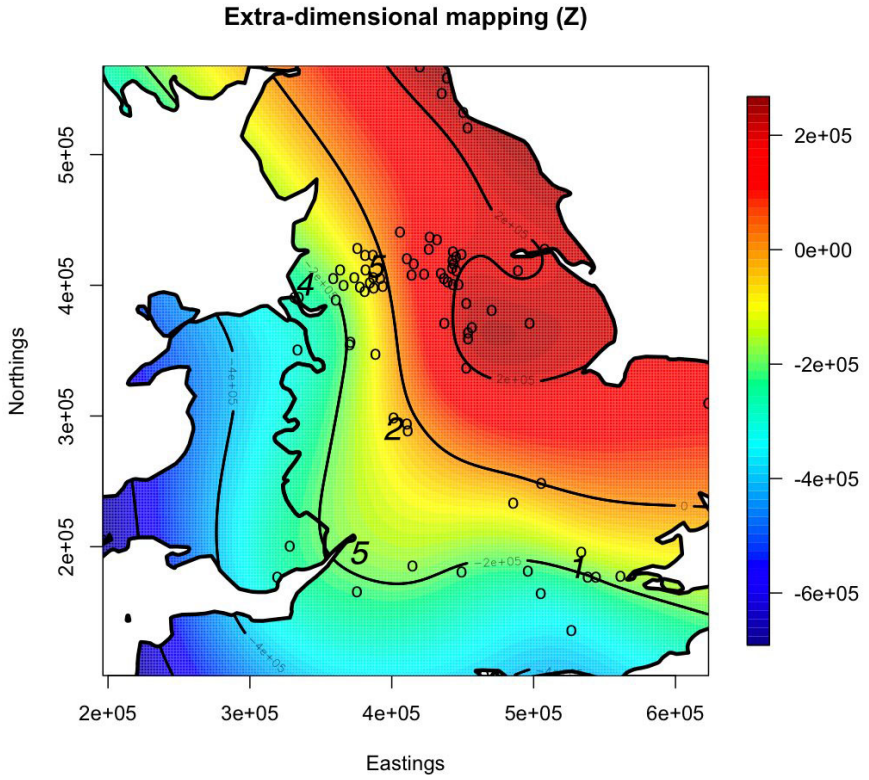


Figure 5.7: Coordinate surface of the learned dimension following dimension expansion. A strong ridge is visible connecting major cities indicating closer correlation between these locations than would be suggested in geographic space. The locations of a selection of major cities are shown; (1) London, (2) Birmingham, (3) Manchester, (4) Liverpool and (5) Bristol.

is worth noting that while we have developed the spatial model in terms of variograms, it could alternatively be expressed in terms of covariances; see, for example, Gneiting et al. [38] for a thorough comparison of the two approaches.

In general, models will comprise a spatial mean or trend term together with spatial covariance for deviations from this trend. It is desirable to maximally reflect the

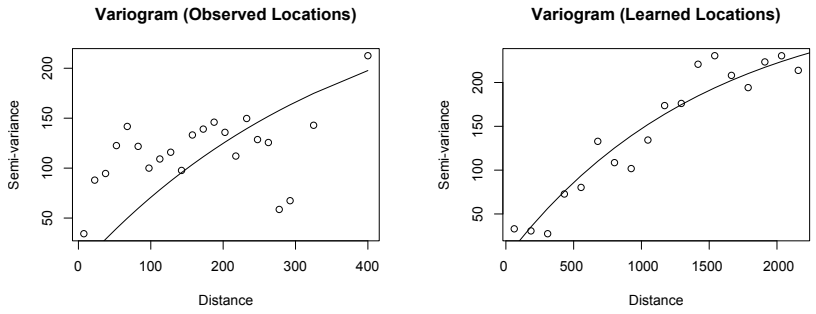


Figure 5.8: Binned empirical and fitted (solid line) variograms on the UK black smoke data, following dimension expansion. In the original geographic space, we see a dip in the empirical variogram at roughly 280km, corresponding to the distance between London and Liverpool/-Manchester. After dimension expansion is applied, the ridge between London and Liverpool/Manchester removes this effect of nonstationarity. The units for the semi-variance are $(\mu \text{gm}^{-3})^2$ and for distance are km.

variation in the response using the mean function and thus known covariates, but inevitably the mean function will not be able to capture all of the spatial variation and thus residual spatial variation must be modeled specifically. When all relevant covariates are included in the mean term, it is commonly assumed that the resulting spatial term is stationary. However, as the Karhunen-Loeve expansion shows, the modeling of spatial trend and covariance are inseparable and misspecification of the former will induce a second order distortion in the latter, thus violating any assumptions of stationarity in many cases. Due to the complexity of environmentally processes, misspecification is inevitable because all relevant covariates can never be known or, even if known, observed. In the air pollution example presented here, concentrations in cities appeared to be more similar than that of rural areas irrespective of their geographic proximity. If available, it would be possible to in-

corporate a measure of rurality in the mean function, possibly produced using a geographical information system based on population density data. However, even if such information were available, stationarity would still not be guaranteed and so there is a need for methods such as that proposed here to allow nonstationary models to be constructed for the spatial process.

5.5 Optimization of Equation (5.2)

As with traditional multi-dimensional scaling, penalization functions of the form (5.1) do not have a unique maximum. However, the learned locations are unique up to rotation, scaling, and sign. The optimization problem (5.2) is more regularized, however, due to the presence of the l_1-norm. Specifically, not all rotations and scalings of the learned locations will have the same l_1-norm, and hence the resulting optimization is unique only up to sign and indices of zero/non-zero dimensions. For example, consider finding the location of a single latent location in 2 latent dimensions. The locations $(1,0)$ and $\left(1/\sqrt{2}, 1\sqrt{2}\right)$ fit the objection function (1) equally well, yet their l_1-norms are 1 and $\sqrt{2}$, respectively. Note that since the end goal is not to learn the dimensions, but rather to find an expanded space in which the process is stationary, the existence of multiple possible expansions is not important, so long as one of the projections in the equivalence class is found.

In our experience, traditional optimization procedures such as Nelder-Mead or the BroydenFletcherGoldfarbShanno method ([65]) work well for a moderate amount of locations ($s < 100$) and dimensions ($p < 3$). For larger problems, it may be necessary to use purpose-built optimization routines intended for generalized group lasso. Let $\Omega(\boldsymbol{U})$ be the first term in equation 5.2, where $\boldsymbol{U} = [\boldsymbol{X}, \boldsymbol{Z}]$. Then

column k of the gradient matrix is

$$\nabla_k \Omega(\boldsymbol{U}) = \frac{2}{p} \boldsymbol{\Gamma} \circ \left(\boldsymbol{U}_{\cdot,k} \mathbf{1}_{p \times p} - \mathbf{1}_{p \times p} \boldsymbol{U}_{\cdot,k}^T \right) \mathbf{1}_{p \times 1}$$

where

$$\boldsymbol{\Gamma}_{i,j} = \left(\gamma_\phi(d_{ij}(\boldsymbol{U})) - \nu_{ij}^* \right) \frac{\partial \gamma_\phi}{\partial d_{ij}}.$$

Using this gradient information, the gradient projection algorithm of Kim et al. [52] may be used to optimize (5.2). The algorithm proceeds as follows:

Initialize : $\boldsymbol{U}^0 = \mathbf{0}$, α : sufficiently small positive constant

for $t = 1, \ldots, T$ **do**

  Set $\boldsymbol{u} = \boldsymbol{U}^{t-1} - \alpha \nabla \Omega(\boldsymbol{U}^{t-1})$ and $\eta = \{1, \ldots, p\}$

  while $M_j > 0 \quad \forall j$ **do**

    For $j = 1, \ldots, p$

$$M_j = I(j \in \eta) \times \left(||\boldsymbol{u}_j|| + \frac{M - \sum_{j \in \eta} ||\boldsymbol{u}_j||}{|\eta|} \right)$$

    Set $\eta = \{j : M_j > 0\}$

  end

  Set $\boldsymbol{U}_{\cdot,j}^{t-1} = \boldsymbol{u}_j \frac{M_j}{||\boldsymbol{u}_j||}$ for $j = 1, \ldots, p$

end

Return $\boldsymbol{U}^T$

From this, one can alternate between optimizing the parameters of the variogram and the latent locations. Further algorithmic details, such as the tuning of M and the setting of the algorithmic parameter α, can be found in Kim et al. [52].

Chapter 6

Conclusion

6.1 Summary

In this thesis, we have expanded the scientific base on latent correlation structures. Firstly, the proposed product graphical model prior improves flexibility in modeling decomposable graphical models, borrowing strength from the immense literature on product partition and related models. The product graphical model prior allows one to encourage (or discourage) clustering of the graphs, and therefore can induce sparsity in the correlation matrix through clique separation; consequently, the product graphical model empowers practitioners to encapsulate their true prior beliefs to build a model more attuned to the problem at hand.

Secondly, we have shown the importance of incorporating spatial correlation into crop yield models; ignoring this information can lead to bias both in model identification and prediction. Specifically, we observed that a common least squares fit of crop yield on some explanatory variables over the entire region resulted in biased residual errors, and hence violated the assumptions of the model. To avoid this

problem, we could fit each agricultural region with its own model. The problem, however, is that this ignores information between crop regions, and as such we observed reduced prediction power and model identifiability. We addressed this issue through the use of a Bayesian model which modelled all regions together, yet accounted for spatial correlation. This model smooths and stabilizes prediction and also allows for analytic and therefore efficient updating and prediction. In addition, we created a non-stationary prior distribution to address the issue of province to province variability resulting from provincial differences in policy, management, and other factors affecting yield. Through cross-validation, we demonstrated this model to achieve improved prediction performance in modelling Canadian wheat yield over the least squares model which ignores spatial dependence, and hope that others will attempt to replicate our findings in other contexts based on the promise seen in this application.

Lastly, by augmenting the dimensionality of the underlying geographic space, we have developed a highly flexible approach for handling the nonstationarity that often arises in environmental systems. While ostensibly similar to image warping, the proposed method avoids the issue of folding and allows one to model much more complex nonstationarity patterns through interdimensional expansions, allowing the user to perform nonparametric learning of the mapping function. In addition, through the use of a group lasso penalty, we are able to estimate the number of augmented dimensions, as well as regularize the optimization problem. Lastly, we have highlighted the practical application of the dimension expansion approach in three examples, two of which use data from observed environmental processes.

6.2 Future Work

6.2.1 Nonstationarity

A Bayesian image warping approach which allows covariate effects to be included in the correlation structure has recently been suggested by Schmidt et al. [80]. By treating covariates as analogous to geographic coordinates, they warp the combined location-covariate space into a deformed space of the same dimension. To achieve computational efficiency, they consider a special case which restricts the form of the possible mapping function and assumes the spatial process to be a 2D manifold with covariates treated as separate values at each location.

In practice, environmental data will often take the form of a number of measurements made over time at each location rather than true spatial replications per se. In order to try and isolate the purely spatial part of the process, the mean function may incorporate a temporal component, modelling underlying temporal patterns and allowing the possibility of time-varying covariates, or even space-time interactions. In the absence of such covariate information, it would be possible to consider the notion of time-varying nonstationarity structure. For instance, if one wants to study the changing impact of cities and industrial areas on air pollution levels, examining changes in stationarity over time would be a valuable way to understand these changes. The dimension expansion framework is also amenable to multivariate extensions. We are currently exploring a scenario whereby the dimension expansion functions and locations have a hierarchical structure, allowing the dimension expansion to vary for different variables, yet be tied together through the hierarchy.

We have demonstrated how dimension expansion can be used to perform pre-

dictions in the transformed, stationary space and mapped back to the original space. At present the choice of the mapping, learning of latent locations, and prediction are performed in isolation. As the Sampson and Guttorp (1992) approach was set within a Bayesian framework by Damian et al. [22] and Schmidt and O'Hagan [79], setting the current algorithmic approach within such an inferential framework would allow the inherent uncertainty to be accurately reflected in resulting inferences and this is the goal of current research.

6.2.2 Sparsity and group correlation

In addition to further studies of nonstationarity, we are also developing new work for modeling sparse multivariate time-series, where dependence is present in the correlation between successive values, between neighboring variables, and in the shared sparsity patterns. We propose an original class of flexible Bayesian linear models for dynamic sparsity modelling. The proposed class of models expands upon the existing Bayesian literature on sparse regression using generalized multivariate hyperbolic distributions. The properties of the models are explored through both analytic results and simulation studies. We demonstrate the model on a financial application where it is shown that it accurately represents the patterns seen in the analysis of stock and derivative data, and is able to detect major events by filtering an artificial portfolio of assets. Early work has been released as a technical report ([17]).

6.2.3 Structural health monitoring

The use of latent structure in modeling correlation is important in a wide variety of fields, including structural health monitoring (SHM). One of the principal goals

of SHM is to monitor structures of various sorts for damage. By placing vibration sensors on a structure and inducing vibration, we can learn the resonant frequencies of the structure in its damaged and undamaged states. We have developed a set of tools based on support vector regression ([4, 11, 13]), and are currently using Bayesian graphical models to address the same problem while simultaneously incorporating knowledge of the correlation structure between vibration sensors.

6.2.4 Monte Carlo

The use of latent structure in modeling often results in partially or fully nonidentifiable models, for which Monte Carlo becomes very difficult due to ridges and multiple modes in the posterior distribution of interest. In addition to developing Monte Carlo methods for prior sensitivity and cross-validation ([12]), we are also developing a suite of tools for automated Monte Carlo analysis in complex models, such as those which attempt to model latent structure. Built on the Wang-Landau algorithm, we have recently developed a fully automated density exploration algorithm ([14]) and are currently conducting further work to provide practitioners with a black-box tool for automated density exploration.

Bibliography

[1] C. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics*, 2:1152–1174, 1974. → pages 19

[2] T. Apanasovich and M. Genton. Cross-covariance functions for multivariate random fields based on latent dimensions. *Biometrika*, 97(1):15, 2010. → pages 78

[3] H. Armstrong, C. Carter, K. Wong, and R. Kohn. Bayesian covariance matrix estimation using a mixture of decomposable graphical models. *Statistics and Computing*, 19(3):303–316, 2009. doi:10.1007/s11222-008-9093-8. → pages 14

[4] S. Atamturktur, L. Bornn, and F. Hemez. Vibration characteristics of vaulted masonry monuments undergoing differential support settlement. *Engineering Structures*, 33(9):2472–2484, 2011. → pages 91

[5] S. Banerjee, B. Carlin, and A. Gelfand. *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall/CRC, 2004. → pages 55, 57

[6] V. Barnett. *Environmental statistics: methods and applications*. John Wiley and Sons, 2004. → pages 36

[7] D. Barry and J. Hartigan. Product partition models for change point problems. *Annals of Statistics*, 20:260–279, 1992. doi:doi:10.1214/aos/1176348521. → pages 16, 18

[8] D. Barry and J. Hartigan. A Bayesian analysis for change point problems. *Journal of the American Statistical Association*, 88:309–319, 1993. → pages 16, 18

[9] L. Bornn and F. Caron. Bayesian clustering in decomposable graphs. *Bayesian Analysis*, 6(4):829–846, 2011. → pages 9

[10] L. Bornn and J. Zidek. Efficient stabilization of crop yield prediction in the canadian prairies. *Agricultural and Forest Meteorology*, 152:223–232, 2012. → pages 20, 35

[11] L. Bornn, C. Farrar, G. Park, and K. Farinholt. Structural health monitoring with autoregressive support vector machines. *Journal of Vibration and Acoustics*, 131:021004, 2009. → pages 91

[12] L. Bornn, A. Doucet, and R. Gottardo. An efficient computational approach for prior sensitivity analysis and cross-validation. *Canadian Journal of Statistics*, 38(1):47–64, 2010. → pages 18, 91

[13] L. Bornn, C. Farrar, and G. Park. Damage detection in initially nonlinear systems. *International Journal of Engineering Science*, 48(10):909–920, 2010. → pages 91

[14] L. Bornn, P. Jacob, P. Del Moral, and A. Doucet. An adaptive interacting wang-landau algorithm for automatic density exploration. *To appear in the Journal of Computational and Graphical Statistics*, 2012. → pages 91

[15] L. Bornn, G. Shaddick, and J. Zidek. Modeling non-stationary processes through dimension expansion. *to appear in the Journal of the American Statistical Association*, 2012. → pages 66

[16] N. Brady, R. Weil, and R. Weil. *The nature and properties of soils*. Prentice Hall Upper Saddle River, NJ, 1999. → pages 39

[17] F. Caron, L. Bornn, and A. Doucet. Sparsity-promoting bayesian dynamic linear models. *Arxiv preprint arXiv:1203.0106*, 2012. → pages 90

[18] A. Carr. Presidential elections 1789-2000 (http://psephos.adam-carr.net/countries/u/usa/pres.shtml), 2005. URL http://psephos.adam-carr.net/countries/u/usa/pres.shtml. → pages 25

[19] C. Carvalho and J. Scott. Objective Bayesian model selection in Gaussian graphical models. *Biometrika*, 96:1–16, 2009. → pages 13, 21

[20] N. Cressie. *Statistics for spatial data*. John Wiley & Sons, New York, 1993. → pages 57, 66, 69

[21] E. Crowley. Product Partition Models for Normal Means. *Journal of the American Statistical Association*, 92(437), 1997. → pages 18

[22] D. Damian, P. Sampson, and P. Guttorp. Bayesian estimation of semi-parametric non-stationary spatial covariance structures. *Environmetrics*, 12(2):161–178, 2001. → pages 76, 90

[23] A. Dawid and S. Lauritzen. Hyper-Markov laws in the statistical analysis of decomposable graphical models. *Annals of Statistics*, 3:1272–1317, 1993. → pages 5, 6, 9, 11, 12

[24] R. Day. Probability distributions of field crop yields. *Journal of Farm Economics*, 47(3):713–741, 1965. → pages 53

[25] R. De Jong and A. Bootsma. Review of recent developments in soil water simulation models. *Canadian Journal of Soil Science*, 76(3):263–273, 1996. → pages 37, 38

[26] A. Dempster. Covariance selection. *Biometrics*, 28(1):157–175, 1972. → pages 12

[27] A. Dobra, C. Hans, B. Jones, J. Nevins, G. Yao, and M. West. Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*, 90(1):196–212, 2004. → pages 13

[28] K. Dolan, R. Lammers, and C. Vöröosmarty. Pan-Arctic temperatization: A preliminary study of future climate impacts on agriculture opportunities in the Pan-Arctic drainage system. *Eos, Trans. Amer. Geophys. Union*, 87, 2006. → pages 50

[29] P. Elliott, G. Shaddick, J. Wakefield, C. Hoogh, and D. Briggs. Long-term associations of outdoor air pollution with mortality in Great Britain. *Thorax*, 62(12):1088, 2007. → pages 78, 80

[30] T. Ferguson. A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1:209–230, 1973. → pages 19

[31] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008. → pages 9, 27

[32] M. Fuentes. A high frequency kriging approach for non-stationary environmental processes. *Environmetrics*, 12(5):469–483, 2001. → pages 67

[33] M. Fuentes and R. Smith. A new class of nonstationary spatial models. Technical Report Institute of Statistics Mimeo Series 2534, North Carolina State University, Raleigh, NC, 2001. → pages 68

[34] Gaia Consulting Limited. 2006 manitoba irrigation survey. 2007. → pages 63

[35] A. Gelfand, K. HJ, C. Sirmans, and S. Banerjee. Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Association*, 98(462):387–396, 2003. → pages 57

[36] P. Giudici and P. J. Green. Decomposable graphical Gaussian model determination. *Biometrika*, 84(4):785–801, 1999. → pages 12, 13, 22

[37] P. Giudici and C. Tarantola. Global prior distributions for the analysis of discrete graphical models. *Statistical Methods and Applications*, 5(1): 129–147, 1996. → pages 21

[38] T. Gneiting, Z. Sasvári, and M. Schlather. Analogies and correspondences between variograms and covariance functions. *Advances in Applied Probability*, pages 617–630, 2001. → pages 83

[39] T. Haas. Kriging and automated variogram modeling within a moving window. *Atmospheric Environment. Part A. General Topics*, 24(7): 1759–1769, 1990. → pages 67

[40] T. Haas. Lognormal and moving window methods of estimating acid deposition. *Journal of the American Statistical Association*, 85(412): 950–963, 1990. → pages 67

[41] P. Hansen, J. Jørgensen, and A. Thomsen. Predicting grain yield and protein content in winter wheat and spring barley using repeated canopy reflectance measurements and partial least squares regression. *The Journal of Agricultural Science*, 139(03):307–318, 2002. → pages 36

[42] J. Hartigan. Partition models. *Communications in statistics. Theory and methods*, 19:2745 – 2756, 1990. doi:10.1080/03610929008830345. → pages 16, 18

[43] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2009. → pages 42, 47, 53

[44] J. Hay. An assessment of the mesoscale variability of solar radiation at the earth's surface. *Solar Energy*, 32(3):425–434, 1984. → pages 78

[45] D. Higdon. A process-convolution approach to modelling temperatures in the North Atlantic Ocean. *Environmental and Ecological Statistics*, 5(2): 173–190, 1998. → pages 67

[46] D. Higdon, J. Swall, and J. Kern. Non-stationary spatial modeling. *Bayesian statistics*, 6:761–768, 1999. → pages 67

[47] P. Hoff, A. Raftery, and M. Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460): 1090–1098, 2002. → pages 78

[48] S. Iovleff and O. Perrin. Estimating a nonstationary spatial structure using simulated annealing. *Journal of Computational and Graphical Statistics*, 13 (1):90–105, 2004. → pages 76

[49] S. Jagtap, J. Jones, P. Hildebrand, D. Letson, J. O'Brien, G. Podestá, D. Zierden, and F. Zazueta. Responding to stakeholder's demands for climate information: from research to applications in Florida. *Agricultural systems*, 74(3):415–430, 2002. → pages 36

[50] B. Jones, C. Carvalho, A. Dobra, C. Hans, C. Carter, and M. West. Experiments in stochastic computation for high-dimensional graphical models. *Statistical Science*, 20(4):388–400, 2005. → pages 13, 14, 21, 22, 27

[51] B. Keating, P. Carberry, G. Hammer, M. Probert, M. Robertson, D. Holzworth, N. Huth, J. Hargreaves, H. Meinke, Z. Hochman, et al. An overview of APSIM, a model designed for farming systems simulation. *European Journal of Agronomy*, 18(3-4):267–288, 2003. → pages 36

[52] Y. Kim, J. Kim, and Y. Kim. Blockwise sparse regression. *Statistica Sinica*, 16(2):375, 2006. → pages 73, 86

[53] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press, 2009. → pages 9

[54] J. Lau and P. Green. Bayesian model-based clustering procedures. *Journal of Computational and Graphical Statistics*, 16:526–558, 2007. → pages 19, 20

[55] S. Lauritzen. *Graphical Models*. Oxford University Press, 1996. → pages 4, 5, 9, 11, 27

[56] N. Le and J. Zidek. *Statistical analysis of environmental space-time processes*. Springer Verlag, 2006. → pages 55

[57] D. Madigan and J. York. Bayesian graphical models for discrete data. *International Statistical Review*, 63:215–232, 1995. → pages 27

[58] S. Mallat. *A wavelet tour of signal processing: the sparse way*. Academic Pr, 2009. → pages 47

[59] K. Mardia and C. Goodall. Spatial-temporal analysis of multivariate environmental monitoring data. *Multivariate Environmental Statistics*, 6: 347–385, 1993. → pages 76

[60] B. Marlin and K. Murphy. Sparse Gaussian graphical models with unknown block structure. In *Proceedings of the 26th Annual Conference on Machine Learning*, 2009. → pages 27

[61] N. Meinshausen and P. Buhlmann. High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(3):1436, 2006. → pages 9, 27

[62] W. Meiring, P. Monestiez, P. Sampson, and P. Guttorp. Developments in the modelling of nonstationary spatial covariance structure from space-time monitoring data. *Geostatistics Wollongong*, 96(1):162–173, 1997. → pages 76

[63] P. Monestiez and P. Switzer. Semiparametric estimation of nonstationary spatial covariance models by metric multidimensional scaling, 1991. → pages 76

[64] P. Monestiez, P. Sampson, and P. Guttorp. Modelling of heterogeneous spatial correlation structure by spatial deformation. *Cahiers de Geostatistique*, 3:1–12, 1993. → pages 76

[65] J. Nocedal and S. Wright. *Numerical optimization*. Springer verlag, 1999. ISBN 0387987932. → pages 85

[66] D. Nychka and N. Saltzman. Design of air quality monitoring networks. *Case studies in environmental statistics*, 132:51–76, 1998. → pages 67

[67] D. Nychka, C. Wikle, and J. Royle. Multiresolution models for nonstationary spatial covariance functions. *Statistical Modelling*, 2(4):315, 2002. → pages 67

[68] O. Perrin and W. Meiring. Nonstationarity in R n Is Second-Order Stationarity in R 2n. *Journal of applied probability*, 40(3):815–820, 2003. ISSN 0021-9002. → pages 68, 70

[69] O. Perrin and M. Schlather. Can any multivariate gaussian vector be interpreted as a sample from a stationary random process? *Statistics & Probability Letters*, 77(9):881–884, 2007. → pages 71

[70] J. Pitman. Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields*, 102:145–158, 1995. → pages 19

[71] A. Potgieter, G. Hammer, A. Doherty, and P. De Voil. A simple regional-scale model for forecasting sorghum yield across North-Eastern Australia. *Agricultural and Forest Meteorology*, 132(1-2):143–153, 2005. → pages 36, 42, 43, 44

[72] A. Potgieter, G. Hammer, and A. Doherty. Oz-wheat: a regional-scale crop yield simulation model for Australian wheat. *Queensland Department of Primary Industries & Fisheries, Information Series No. QI06033, Brisbane, Qld (ISSN 0727-6273)*, 2006. → pages 20, 36, 51

[73] B. Qian, R. De Jong, and S. Gameda. Multivariate analysis of water-related agroclimatic factors limiting spring wheat yields on the Canadian prairies. *European Journal of Agronomy*, 30(2):140–150, 2009. → pages 36, 39

[74] B. Qian, R. De Jong, R. Warren, A. Chipanshi, and H. Hill. Statistical spring wheat yield forecasting for the Canadian prairie provinces. *Agricultural and Forest Meteorology*, 149(6-7):1022–1031, 2009. → pages 36, 39, 42, 43

[75] F. Quintana. A predictive view of Bayesian clustering. *Journal of Statistical Planning and Inference*, 136:2407–2429, 2006. → pages 18

[76] F. Quintana and P. Iglesias. Bayesian clustering and product partition models. *Journal of the Royal Statistical Society B*, 65:557–574, 2003. doi:10.1111/1467-9868.00402. → pages 18

[77] J. Ritchie and D. NeSmith. Temperature and crop development. *Modeling plant and soil systems*, pages 5–29, 1991. → pages 49

[78] P. Sampson and P. Guttorp. Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association*, 87(417):108–119, 1992. → pages 57, 68, 69, 72, 76, 77, 78

[79] A. Schmidt and A. O'Hagan. Bayesian inference for non-stationary spatial covariance structure via spatial deformations. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 65(3):743–758, 2003. → pages 76, 90

[80] A. Schmidt, P. Guttorp, and A. O'Hagan. Considering covariates in the covariance structure of spatial processes. *Environmetrics*, 22(4):487–500, 2011. → pages 76, 89

[81] A. Schmitz and W. Furtan. *The Canadian Wheat Board: marketing in the new millennium*. Canadian Plains Research Center, 2000. → pages 35

[82] R. Smith. Estimating nonstationary spatial correlations. *Preprint, University of North Carolina*, 1996. → pages 76

[83] D. Stephens. *Crop yield forecasting over large areas in Australia*. PhD thesis, Murdoch University, 1995. → pages 36

[84] R. Stone and H. Meinke. Operational seasonal forecasting of crop performance. *Philosophical Transactions B*, 360(1463):2109, 2005. → pages 36

[85] C. Thornthwaite. An approach toward a rational classification of climate. *Geographical review*, 38(1):55–94, 1948. → pages 38

[86] M. West. Hyperparameter estimation in Dirichlet process mixture model. ISDS discussion paper no. 92-03, Duke University, Durham, NC, 1992. → pages 28

[87] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006. ISSN 1467-9868. → pages 73

[88] M. Yuan and Y. Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94:19–35, 2007. → pages 9, 27

[89] A. Zellner. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In P. Goel and A. Zellner, editors, *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno do Dinetti*, pages 233–243. North-Holland, 1986. → pages 56

[90] J. Zidek, W. Sun, and N. Le. Designing and integrating composite networks for monitoring multivariate gaussian pollution fields. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 49(1):63–79, 2000. ISSN 1467-9876. → pages 76