Missing Data Problems

Dissertation Advisor:
Gary Chamberlain

Author:
Guillaume Pouliot

Missing Data Problems

Abstract

Missing data problems are often best tackled by taking into consideration specificities of the data structure and data generating process. In this doctoral dissertation, I present a thorough study of two specific problems. The first problem is one of regression analysis with misaligned data; that is, when the geographic location of the dependent variable and that of some independent variable do not coincide. The misaligned independent variable is rainfall, and it can be successfully modeled as a Gaussian random field, which makes identification possible. In the second problem, the missing independent variable a categorical. In that case, I am able to train a machine learning algorithm which predicts the missing variable. A common theme throughout is the tension between efficiency and robustness. Both missing data problems studied herein arise from the merging of separate sources of data.

Contents

List of Tables

List of Figures

Acknowledgments

First and foremost, I would like to thank my advisors. The committee chair, Gary Chamberlain, whose constant support has made this project possible, and whom, through always profoundly enlightening conversations, has allowed me to grasp and tackle the research problems presented herein. Edward Glaeser, whose advice has been inestimable, and who taught me how to write an economics paper. Neil Shephard, whose seemingly infinite knowledge of econometrics and statistics has, far more times than I am proud to say, revived projects that seemed entirely stuck. And Elie Tamer, whose unfailing instincts have guided this project always in the better direction, and whose help and advice have been absolutely invaluable.

I am likewise indebted for their insightful comments to Alberto Abadie, Isaiah Andrews, Alexander Bloemendal, Paul Bourgade, Kirill Borusyak, Peter Ganong, Joe Guinness, Guido Imbens, Simon Jäger, Eric Janofsky, José Luis Montiel Olea, Mikkel Plagborg-Møller, Daniel Pollmann, Andrew Poppick, Suhasini Subba Rao, Martin Rotemberg, Jann Spiess, Michael Stein, Bryce Millett Steinberg and Alexander Volfovsky.

To my parents, who raised me to believe that no dream was too big.

To my advisors, who are thanked individually in the acknowledgements.

That teaching which shapes students' understanding of the world and develops in them an instinctive grasp of fundamental concepts can only arise from extraordinary efforts. For that reason, I wish to thank those who taught me relentlessly: Alain Lachapelle, Elizabeth Townsend Milicevic, Michael Stein, and Gary Chamberlain.

Introduction

This doctoral dissertation studies methods for regression analysis when missing data problems arise from merging separate sources of data. In applied economics, it is often the case that different variables which the economist wishes to compare are collected in separate datasets. Therefore, in order to compare them, the economist has to merge the datasets. The resulting, combined dataset may be nonstandard, and sophisticated methods will be required to analyze it.

For example, to study the impact of rainfall on health and other socio-economic outcomes (Maccini and Yang, 2009; Chapters I and II below), economists have merged publicly available rainfall data with the Indonesian Family Life Survey (IFLS). Likewise, in order to look at the impact of flexible work hours on voter turnout using CPS data (Chapter III), economists need to merge two supplementary CPS surveys collecting observations of these variables.

A common occurrence when merging datasets is that the resulting, combined dataset will be incomplete; i.e., there will be missing data. This is typically because data was collected for different locations or subjects in the different datasets.

For example, in the application of Chapters I and II, the rainfall is measured at the location of the rainfall measurement stations, and we use as the location of the surveyed individuals the coordinates of their village. However, the location of the villages and that of the rainfall measurement stations do not coincide (whence we speak of misaligned data). Of course, this makes regression analysis nontrivial because the economist wishes to estimate the impact of rainfall shocks at the location of the individual. Similarly, CPS supplementary surveys (Chapter III) are non-overlapping. That is, the same subject cannot appear in both surveys.

Consequently, one cannot compare the flexible work hours status of any one individual directly with his own voting status. Standard regression methods therefore cannot be applied without modifications.

In order to carry out empirical work when merged datasets bring about missing data problems, economists will want regression methods that exploit the specificities of the data structure or data generating process and, as such, which will need to be developed on a case-by-case basis.

For example, the rainfall data can be fruitfully modeled as a Gaussian random field so to allow for reliable interpolation of the value of rainfall at the location of the surveyed individuals' villages. Likewise, CPS supplementary surveys include all the variables of the core CPS survey, which may allow for imputation; if one can build a predictor of the flexible work hours status as a function of the core CPS variables and train it on the flexible work hours supplementary survey, then one may be able to impute the flexible work hours status in the voter turnout supplementary survey.

The overarching statistical challenge is to provide methods which, in spite of imputing (implicitly or explicitly) a missing regressor, will provide consistent regression coefficient estimates and will optimally tradeoff the efficiency of the estimate against the strength of assumptions.

Chapter I details two methods for spatial, misaligned data which we refer to as one-step methods. With these methods, the imputation of the missing rainfall values and the estimation of the regression coefficient of rainfall on, say, health status are carried out jointly for a more efficient use of information.

Chapter II presents a two-step method for dealing with misaligned data. The missing rainfall value is imputed in a first step, and in a second step the regression of health status on the imputed rainfall value is carried out. It is detailed therein how such a two-step method may deliver consistent regression coefficient estimates, as well as standard errors that take into account uncertainty due to estimation in the first stage.

Chapters I and II are complementary. First, they share an application, the analysis of the

impact of rainfall on socio-economic outcomes (such as health status) using misaligned data, which is presented in Chapter II and is a reanalysis of the data presented in Maccini and Yang (2009). Second, the three methods presented in these two chapters can be thought of as making up a complete toolkit. If one is confident in the specification of the first- and second-order moments (crucially, if one has a correct model for the covariance of the regression errors), then the maximum likelihood method presented in Chapter I will deliver consistent estimates which will furthermore be efficient (in the Cramér-Rao sense) if the distributional assumptions are satisfied. The two-step method presented in Chapter II is easy-to-implement and does not require specification of the regression error covariance, but is less efficient than the maximum likelihood estimator. As an in-between, a minimum distance estimator is presented in Chapter I; it does not require specification of the covariance of regression errors for point estimation, but does require it for inference. Hence, we have three methods trading off efficiency against robustness, and the economist is able to pick that which delivers the most efficient estimates available, given the assumptions afforded by the application at hand.

Chapter III takes up the exercise carried out in Chapters I and II, developing both a joint modeling approach and a robust two-step approach, but when the missing independent variable is categorical. In particular, a full Bayesian implementation of k-nearest neighbors is developed, as well as a two-step approach requiring neither the computation of a normalizing constant or of a bias correction term.

The two-step methods in Chapters II and III stand as examples that in regression analysis with merged datasets and a missing independent variable of interest, one may find an easy-to-implement, consistent two-step method which will allow for robust inference, and will be useful as long as the signal to detect is strong enough that more efficient methods are not required in order to detect it in small samples.

Chapter 1

One-Step Methods for Misaligned Regression

1.1 Introduction

Spatial data analysis has become increasingly popular in the social sciences. In many applications, data sets providing the specific location of households, firms, villages, or other economic units are matched by location to data with geographic features such as rainfall, temperature, soil quality, ruggedness, or air pollution in order to analyze the impact of such environmental variables on economic outcomes.[1] Such data underpins important economic research; for instance, it informs policy responses to events such as droughts, smog outbreaks, poor harvests, etc. A typical issue is that the matched data sets will be *misaligned*. That is, the respective geographical locations of the observations in the matched data sets do not generally coincide. For instance, a researcher might observe crop outputs from a sample of farms in a large area, as well as measurements of rainfall collected over the same area from several weather stations. The locations of the weather stations and the farms will generally

[1]Maccini and Yang (2009), Miguel et al. (2004), and Shah and Steinberg (2013) study the impact of rainfall. Dell et al. (2014) survey applications using weather data. Kremer et al. (2015) use measurements of soil nutrients in some locations to make fertilizer recommendations in others. Nunn and Puga (2012) uses terrain ruggedness for identification, and Chay and Greenstone (1999) study the impact of air pollution.

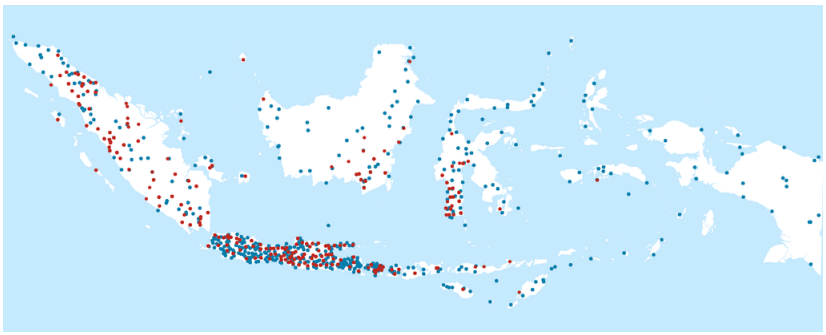

Figure 1.1: *Map of Indonesia*

Geographic location of rainfall measurements (blue) and survey data (red) merged in the Maccini and Yang (2009) analysis.

not coincide, which yields a misaligned data set.

The approaches commonly used in social sciences to address the misalignment problem yield inefficient estimates and incorrect confidence intervals. Popular approaches for analyzing such data sets involve imputing a level of the spatial regressor for each misaligned observation of the outcome variable, and then proceeding with standard regression analysis. It is common to impute using either the value of the nearest location (Acemoglu et al., 2001), the value of the nearest location instrumented with more distant locations (Maccini and Yang, 2009), or a distance-weighted average of nearby locations (Shah and Steinberg, 2013). Because these methods impute the regressors in an initial step before considering the outcome data, I refer to them as "two-step", or "plug-in", methods. These estimates are inefficient because they do not use all the relevant information (see Subsection 2.1). In addition, confidence intervals are typically erroneous because they do not account for the fact the some covariates have been estimated.

Gaussian maximum likelihood is a more efficient, "one-step" method which produces accurate point estimates. It is however known that, when erroneously assumed to be correctly specified, it yields unreliable confidence intervals (Madsen et al., 2008). A natural approach is then to conduct inference using the Gaussian maximum likelihood estimator, but to dispense with the assumption that the data follows a Gaussian distribution. The resulting estimator

is called the quasi-maximum likelihood estimate (QMLE). Even though it allows for both efficient estimates and correct standard errors, the QMLE is not established as the standard method for regression analysis with misaligned data for three reasons. First, the necessary distribution theory is, to the best of my knowledge, unavailable. Second, the computations involved are costly and difficult. Finally, others have been reluctant to model the covariance of the regression errors, believing it to be either too difficult or too restrictive.

This paper addresses each of these issues. I obtain a new central limit theorem for the QMLE with spatial data. In particular, I present a new central limit theorem for quadratic forms in mixing variables, which may have value in different applications. I develop computational strategies to readily compute the QMLE as well as its variance, and I assess their performance. As a robust companion method, I suggest a minimum-distance estimator that does not require specification of the regression error covariance. In order to conduct inference with the robust method, I obtain a new central limit theorem for nonparametric spatial covariance function estimators, itself of independent interest. Simulations strongly suggest that the recommended methods outperform common approaches in the literature.

I reproduce the cross-validation exercise of Madsen et al. (2008), who compare the performance for inference of maximum likelihood with Krig-and-Regress. I find, as they did, that maximum likelihood outperforms the two-step method, but that its standard errors (using the asymptotic variance formula for correctly specified maximum likelihood) are unreliable. I find, however, that the robust standard errors I obtain are very reliable. I reanalyze the influential data set of Maccini and Yang (2009) and find that their general conclusions hold up: rainfall shocks in the first year of life impact adult socio-economic outcomes. However, the analysis nevertheless benefits from the use of my methods, as those yield some statistically and economically significant changes in the value of key parameter estimates.

It is worth noting that, even though many of the two-step methods used in the literature are inconsistent,[2] imputation with the best linear predictor in the first stage makes for a

[2]All of the nearest neighbor (Acemoglu et al., 2001), distance-weighted average (Shah and Steinberg, 2013), and instrumental variables (Maccini and Yang, 2009) approaches produce inconsistent regression coefficient estimates, unless one assumes the data is getting infinitely dense asymptotically.

consistent estimate of the regression coefficient in the second stage. However, correct inference with this procedure requires specification of the regression error covariance. This leaves the researcher wanting a simple, if inefficient, two-step method.

I propose and analyze a simple two-step Bayesian bootstrap method which, by relying on survey sampling of the economic data in Maccini and Yang (2009), allows for a two-step method for which point estimation and standard errors (which account for imputation uncertainty) obtain without requiring specification of the regression error covariance.

1.1.1 Problem Set-Up

I now specify the general misaligned regression problem at the heart of the present inquiry. I am interested in the regression coefficient β in the spatial regression problem

$$\mathbf{Y} = \mathbf{R}_{\text{true}}\beta + F\gamma + \epsilon, \tag{1.1}$$

where $\mathbf{Y} = Y(x)$ is an N-tuple $\begin{bmatrix} Y(x_1) & \dots & Y(x_N) \end{bmatrix}^T$ and

$$\mathbf{R}_{\text{true}} = R(x) = \begin{bmatrix} R(x_1) & \dots & R(x_N) \end{bmatrix}^T \tag{1.2}$$

is drawn from a stationary Gaussian random field[3] with mean function $m(\cdot)$ and covariance function $K_\theta(\cdot,\cdot)$, where θ indexes a parametric model for the covariance function. The geographic locations are $x_i \in \mathscr{D} \subset \mathbb{R}^2$, $i = 1, ..., N$. $\mathbf{Y}$ and F are observed, but not $\mathbf{R}_{\text{true}}$. However, the M-tuple

$$\mathbf{R}^* = R(x^*) = \begin{bmatrix} R(x_1^*) & \dots & R(x_M^*) \end{bmatrix}^T, \tag{1.3}$$

with $x_i^* \in \mathscr{D} \subset \mathbb{R}^2$, $i = 1, ..., M$, is observed. That is, although the outcome variable data $Y(x)$ (e.g. crop yields at farm locations) is not sampled at the same locations as the independent variable data $R(x^*)$ (e.g. rain measured at fixed weather stations), it is R at the same locations

[3]We say that $\{R(x) : x \in \mathbb{R}^d\}$ is a Gaussian random field if for any choice of vector of locations $(x_1, ..., x_n)$, the random vector $(R(x_1), ..., R(x_n))$ is distributed multivariate normal. The practical usefulness of Gaussian random fields to model rainfall data has long been established, see in particular Phillips et al. (1992) and Tabios and Salas (1985).

as that of the outcome variable, that is $R(x)$, which enters the regression function.

The marginal density is then

$$\begin{pmatrix} \mathbf{Y} \\ \mathbf{R}^* \end{pmatrix} \sim N_{N+M}\left(\begin{pmatrix} m(x)\beta + F\gamma \\ m(x^*) \end{pmatrix}, \begin{pmatrix} \beta^2\mathbf{K} + \Sigma & \beta\bar{\mathbf{K}} \\ \beta\bar{\mathbf{K}}^T & \mathbf{K}^* \end{pmatrix}\right). \tag{1.4}$$

where $\mathbf{K} = K_\theta(x,x) = V_\theta(R(x))$, $\bar{\mathbf{K}} = K_\theta(x,x^*) = Cov_\theta(R(x), R(x^*))$ and $\mathbf{K}^* = K_\theta(x^*,x^*) = V_\theta(R(x^*))$ for some $\theta \in \Theta$.

In the absence of rainfall measurements at the locations of outcomes, the identifying assumption is that I have a parametrized covariance function, and know $Cov_\theta(\mathbf{R}_{\text{true}}, \mathbf{R}^*)$ up to the value of a small-dimensional parameter vector θ, which I estimate consistently. This allows, for instance, the construction of a best linear unbiased predictor for unobserved rainfall.

For my purposes, it will generally be the case that m is a constant, and thus the mean parameter of the random field R can be absorbed in the constant vector (for the intercept) in F. Hence, I am concerned with the marginal

$$\begin{pmatrix} \mathbf{Y} \\ \mathbf{R}^* \end{pmatrix} \sim N_{N+M}\left(\begin{pmatrix} F\gamma \\ \mathbf{m}^* \end{pmatrix}, \begin{pmatrix} \beta^2\mathbf{K} + \Sigma & \beta\bar{\mathbf{K}} \\ \beta\bar{\mathbf{K}}^T & \mathbf{K}^* \end{pmatrix}\right), \tag{1.5}$$

where the coefficient of interest, β, only appears in the covariance. Minus twice the log likelihood is then

$$l = \log\left((2\pi)^{n+m}\left|\Omega\right|\right) + \begin{pmatrix} \mathbf{Y} - F\gamma \\ \mathbf{R}^* - \mathbf{m}^* \end{pmatrix}^T \begin{pmatrix} \beta^2\mathbf{K} + \Sigma & \beta\bar{\mathbf{K}} \\ \beta\bar{\mathbf{K}}^T & \mathbf{K}^* \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{Y} - F\gamma \\ \mathbf{R}^* - \mathbf{m}^* \end{pmatrix}. \tag{1.6}$$

If p/N is sizable, where p is the width of F, it may not be advisable to use the maximum likelihood estimate directly. The coefficient I am interested in is a covariance parameter in the likelihood (6), and if a non-negligible share of the degrees of freedom is used to estimate the mean parameters, then the estimate of the covariance may be non-negligibly biased. As explained in Subsection 2.3, a natural solution to this problem is to use a residual maximum likelihood (REML) approach.

1.1.2 Related Literature

My paper contributes to several segments of the literature. First and foremost, I provide methods[4] for applied researchers. As detailed above, even careful applied work (Maccini and Yang, 2009; Shah and Steinberg, 2013) relied on *ad hoc* approaches because econometric methodology had not caught up to the needs of applied economists. My paper fills that gap. It speaks to a well-established literature on generated regressors (Pagan, 1984; Murphy and Topel, 1985), and looks at problems in which the law of the regressor to impute is well described by a Gaussian random field, which makes for a rich structure. Gaussian random fields are well studied in geostatistics (Gelfand et al., 2010; Stein, 1999) where, however, interest is concentrated on interpolation. My paper relates the two literatures and leverages geostatistical methodology and results to provide robust and efficient methods for economists carrying out regression analysis with misaligned data.

The asymptotic distribution I obtain for the QMLE extends available limit distribution theory for maximum likelihood in the spatial setting. Watkins and Al-Bouthiahi (1990) had given the asymptotic distribution of the MLE when only the covariance function was misspecified. I give the distribution of the MLE when both the normal distribution and the covariance function are misspecified.

The central limit theorem I obtain for inference with my minimum-distance estimator builds on results in spatial statistics. Cressie et al. (2002) gave a central limit theorem for empirical covariance estimators when the data is on a lattice. I leverage results from Lahiri (2003), who gives a family of central limit theorems for spatial statistics, in order to extend the asymptotic theory for the empirical covariance estimators to the case of irregular spaced data.

1.1.3 Outline

The remainder of the article is divided as follows. Section 2 presents and discusses key concepts and assumptions for the analysis. Section 3 develops the limit distribution theory and computational strategies for the QMLE. Section 4 introduces a robust companion method,

[4] And soon, statistical packages.

which circumvents the need to model the covariance of the regression errors, and develops limit distribution theory for the estimator. Section 5 presents the two-step Bayesian bootstrap estimator and studies its performance in a simulation study. Section 6 studies the comparative performance of the considered estimators in the cross-validation exercise of Madsen et al. (2008). Section 7 reanalyzes the misaligned data set in Maccini and Yang (2009). Section 8 discusses and concludes. Technical proofs are deferred to the appendix.

1.2 Key Concepts and Background Material

I introduce a few concepts playing a pivotal role in the study. First, I develop on the difference between one-step and two-step methods, and give an intuitive explanation of the efficiency gain one obtains from the latter. Second, I define and characterize the quasi-maximum likelihood estimator (QMLE). Finally, I define and explain the residual maximum likelihood estimator (REML).

1.2.1 One-step and Two-step Methods

Two-step methods for regression analysis of misaligned data consist in first predicting the misaligned covariates at the outcome locations where they are not observed, thus generating an aligned data set, and then proceeding to standard spatial regression analysis with this generated data set.[5] The first step, which consists of predicting the missing independent variables, requires the choice of an interpolation method. Nonparametric methods, such as approximation by the average of a given number of nearest neighbors, may be used. However, when the misaligned variable can be modeled following, or approximately following, the law of a Gaussian random field, Kriging generally affords the researcher more accurate interpolated variables (Gelfand et al., 2010).

Kriging (named after the South African mining engineer D. G. Krige) consists in using the estimated best linear unbiased predictor for interpolation. It can be developed as follows

[5]Note that Maccini and Yang (2009) do suggest a covariance estimator which accounts for the interpolation step by using 2SLS standard errors.

(Stein, 1999). The random field of interest, R, is assumed to follow the model

$$R(x) = m(x)^T\gamma + \varepsilon(x),$$

$x \in \mathscr{D} \subset \mathbb{R}^2$, where ε is a mean zero random field, m is a known function with values in $\mathbb{R}^p$ and γ is a vector of p coefficients. We observe R at locations $x_1, x_2, ..., x_N$. That is, we observe $\mathbf{R}^* = (R(x_1), R(x_2), ..., R(x_N))$ and need to predict $R(x_0)$. With γ known, the best linear predictor (BLP) is

$$m(x_0)^T\gamma + \mathbf{k}^T\mathbf{K}^{-1}(\mathbf{R}^* - \mathbf{m}^*\gamma),$$

where $\mathbf{k} = Cov(\mathbf{R}^*, R(x_0))$, $\mathbf{K} = Cov(\mathbf{R}^*, \mathbf{R}^{*T})$ and $\mathbf{m}^* = (m(x_1), m(x_2), ..., m(x_N))^T$. Of course, the mean parameter is in general unknown. If γ is replaced by its generalized least-squares estimator, $\hat{\gamma} = \left(\mathbf{M}^T\mathbf{K}^{-1}\mathbf{M}\right)^{-1}\mathbf{M}^T\mathbf{K}^{-\mathbf{1}}\mathbf{R}^*$ (under the assumption that $\mathbf{K}$ and $\mathbf{M}$ are of full rank), we obtain the best linear unbiased predictor (BLUP) for $R(x_0)$. Again, in general, the covariance structure will be unknown, and $\mathbf{k}$ and $\mathbf{K}$ will be replaced by estimates $\hat{\mathbf{k}}$ and $\hat{\mathbf{K}}$. The resulting plug-in estimator will be called the estimated BLUP (EBLUP). Prediction with the BLUP and EBLUP are both referred to as Kriging. As far as this article is concerned, the covariance structures will always be *a priori* unknown, and Kriging will refer to prediction with the EBLUP.

There are many choices for the covariance functions (Gelfand et al., 2010), and I present three of them: the exponential covariance function

$$K_{\text{exp}}(d) = \theta_1 \exp\left(-d/\theta_2\right),$$

the Gaussian covariance function

$$K_{\text{Gaussian}}(d) = \theta_1 \exp\left(-d^2/\theta_2^2\right),$$

and the Matérn covariance function

$$K_{\text{Matérn}}(d) = \theta_1 \frac{(d/\theta_2)^\nu \mathcal{K}_\nu (d/\theta_2)}{2^{\nu-1}\Gamma(\nu)},$$

where $\mathcal{K}_\nu$ is the modified Bessel function of the second kind of order ν (sec. 9.6, Abramowitz and Stegun, 1965). All functions have positive parameters θ_1 and θ_2, the sill and range, respectively. The sill parameter should be thought of as controlling the scale of the covariance, and the range should be thought of as controlling how fast the covariance decays over distance. The Matérn function has an additional parameter ν, which controls smoothness.

Jiang (1997) and Stein (1999) present an alternative derivation of the BLUP as the best predictor, under normality, based on all error contrasts. An excellent theoretical treatment of the topic can be found in Stein (1999). Cressie (1993) and Diggle and Ribeiro (2007) offer a more applied treatment of the topic. Matheron (1962) is a classic reference.

The information gain incurred by using a one-step method instead of a consistent two-step method such as Krig-and-Regress is best displayed with a general example. Suppose the vector of observations $\mathbf{Y} \in \mathbb{R}^n$ is observed, and is assumed to follow the simple regression model

$$\mathbf{Y} = \beta_0 + \mathbf{R}_{\text{true}}\beta + \epsilon, \tag{1.7}$$

where $\mathbf{R}_{\text{true}}, \epsilon \in \mathbb{R}^n$ and β_0, β are unknown parameters. Further assume that, although $\mathbf{R}_{\text{true}}$ is unobserved, a vector sampled from the same random field but at different locations, $\mathbf{R}^* \in \mathbb{R}^m$, is observed. In this set-up, a two-step method is minimizing

$$\mathcal{L}_1 \left(f(\mathbf{R}^*) - \mathbf{R}_{\text{true}} \right),$$

in f, where $\mathcal{L}_1$ is some loss function, to obtain $\hat{\mathbf{R}} = \hat{f}(\mathbf{R}^*)$, and then minimizing

$$\mathcal{L}_2 \left(\mathbf{Y} - \hat{\mathbf{R}}\beta \right),$$

where $\mathcal{L}_2$ is some loss function, in β to get $\hat{\beta}$. A one-step method instead minimizes

$$\mathcal{L}_3 \left(f(\mathbf{R}^*) - \mathbf{R}_{\text{true}}, \mathbf{Y} - f(\mathbf{R}^*)\beta \right),$$

where $\mathcal{L}_3$ is some loss function, jointly in f and β. That is, in the one-step method but not in the two-step method, variation in $\mathbf{Y}$ will inform the choice of $\hat{\mathbf{R}}$ (by "pushing towards" an $\hat{\mathbf{R}}$ which minimizes regression error).

It may be tempting to conclude that the two-step method with Kriging, by plugging in a guess of the correct dependent variable, induces attenuation bias (also known as regression dilution). It should be clear that, in the absence of measurement error, this is not the case. For instance, consider Kriging then regressing with known covariance structure and known mean $\mathbf{m} \equiv 0$. The first step estimate is $\hat{R} = \mathbf{R}^* Cov(\mathbf{R}^*, \mathbf{R}^*)^{-1} Cov(\mathbf{R}^*, R)$, and the probability limit (in n) of the two-step estimator is

$$\begin{aligned}
\text{plim}\hat{\beta} &= \frac{Cov\left(Y, \hat{R}\right)}{Cov\left(\hat{R}, \hat{R}\right)} \\
&= \beta \frac{Cov(R, \mathbf{R}^*) Cov(\mathbf{R}^*, \mathbf{R}^*)^{-1} Cov(\mathbf{R}^*, R)}{Cov(R, \mathbf{R}^*) Cov(\mathbf{R}^*, \mathbf{R}^*)^{-1} Cov(\mathbf{R}^*, \mathbf{R}^*) Cov(\mathbf{R}^*, \mathbf{R}^*)^{-1} Cov(\mathbf{R}^*, R)} \\
&= \beta.
\end{aligned}$$

That is, the estimator is consistent. In general, m as well as the covariance structure are unknown and have to be estimated. This will make the estimate less precise or even biased, but it does not generally introduce *attenuation* bias.

To be sure, although two-step methods commonly used in the literature (e.g. nearest neighbors, IV) are not in general consistent, Krig-and-Regress is a consistent two-step method. However, Krig-and-Regress remains inefficient as it omits information which is relevant to the estimation of β and thus allows for a more statistically efficient estimate. Since economists often estimate small effects (e.g. the impact of rainfall in infancy on adult socio-economic outcomes), it may be crucial for estimation, even with large data sets, that they have access to a one-step, efficient method.

The natural candidate for a one-step method is maximum likelihood. A Gaussian likelihood function may be preconized as it is computationally as well as analytically tractable, and in particular allows trivial marginalization of the aligned, but unobserved variables (e.g. rainfall

at the locations of the farms). Furthermore, the likelihood approach will take into account covariance parameter uncertainty, and can thus be expected to generate more accurate standard errors.

I am interested in the comparative performance of the maximum likelihood with two-step methods, in particular when using Kriging in the first step. Madsen et al. (2008) study the problem of misaligned regression and compare the maximum likelihood approach with the two-step approach consisting in first Kriging to obtain good estimates for the missing, aligned values, and then carrying out the regression with the interpolated quantities as "plug-in" covariates. They find that maximum likelihood yields confidence intervals which are too narrow, and thus recommend the two-step approach as better suited for inference. I find this to be an unpalatable conclusion. If the regression model is taken seriously, then a two-step approach incurs efficiency losses, and there ought to be a serious effort to salvage the maximum likelihood approach.

This begs the question of how a quasi-maximum likelihood estimator with a sandwich covariance matrix (i.e. the asymptotic covariance obtained without relying on the information equality, which does not hold for misspecified models) would perform.

1.2.2 Quasi-Maximum Likelihood

In practice, even though one has recognized (perhaps transformed) data as being approximately Gaussian, one often disbelieves the assumption of exact Gaussianity. Likewise, one may worry that a low-dimensional covariance function is not flexible enough to capture the true covariance. Thus I consider the properties of the maximizer of the likelihood function as a quasi-maximum likelihood estimator. That is, I will be interested in the statistical properties of the maximizer of a Gaussian likelihood function when the data entering the likelihood is not necessarily Gaussian, or there isn't necessarily a parameter $\theta \in \Theta$ such that $\mathcal{K}_\theta$ gives the true covariance. I refer to this as a quasi-maximum likelihood estimator (QMLE). Immediate questions, which are crucial for a cogent use of quasi-maximum likelihood are: what is then the interpretation of the target parameter? If the covariance function is correctly specified, are the true covariance

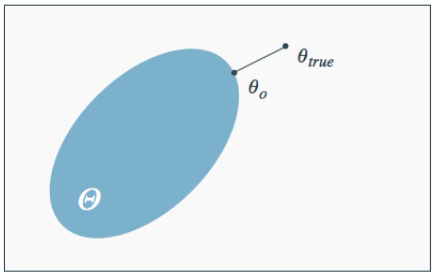

Figure 1.2: *Target of QMLE*

When the underlying data is not Gaussian but the mean and covariance are well specified, the QMLE $\hat{\theta}_n$ is consistent. But even when the first two moments are not well specified, the QMLE has a reasonable target θ_0, the parameter in the constraint set which minimizes the Kullback-Leibler distance to the truth.

parameters identified? Does a central limit theorem obtain? Can we readily compute the asymptotic covariance matrix?

The last two questions will be central subjects of our inquiry, and are answered in Section 3. The first two questions can be readily answered.

For a misspecified procedure to be reasonable, we need to establish the distribution which is in fact being estimated and decide whether or not we find it a desirable target. In many cases, the QMLE is a consistent estimator of a very natural target. In the case of independent observations, it is well-known (White, 1982; van der Vaart, 1998) and easy to see that the limiting value of the QMLE corresponds to the distribution within the constraint set which minimizes the Kullback–Leibler divergence with the true distribution. As detailed below, this observation applies to dependent data as well.

In order to lighten notation, I use $\mathbf{W} = (W(x_1), ..., W(x_n))^T$ as the full vector of observed data. I assume that

$$E[\mathbf{W}] = 0 \text{ and } Cov(\mathbf{W}) = V_*. \tag{1.8}$$

I am interested in the QMLE $\hat{\theta}(n)$ maximizing the Gaussian likelihood

$$l(\mathbf{W};\theta) = -\frac{1}{2}\log|V(\theta)| - \frac{1}{2}\mathbf{W}^T V^{-1}(\theta)\mathbf{W}, \tag{1.9}$$

over the constraint set Θ, where the $V(\theta) = [\sigma(x_i - x_j;\theta)]_{i,j}$ with explicit covariance function σ.

To specify the target, I observe as in Watkins and Al-Boutiahi (1990) that $\hat{\theta}(n) = \theta_0 + o_p(1)$, with θ_0 the limiting value of the sequence $\{\theta_0(n)\}$ where $\theta_0(n)$ minimizes the Kullback-Leibler distance

$$\frac{1}{2}\log|V(\theta)| + \frac{1}{2}\mathrm{tr}\left(V^{-1}(\theta)V_*\right) - \frac{1}{2}\log|V_*| - \frac{n}{2}, \tag{1.10}$$

thus making θ_0 a meaningful target. Figure 2 illustrates this fact.

Now consider the case in which, although the full distribution is not correctly specified, the covariance function is. Then the estimated parameter will converge to the true covariance parameter. This follows from a direct application of the information inequality to (10), and is collected as a fact.

Fact 1 *Suppose there exists $\theta_* \in \Theta$ such that $V(\theta_*) = V_*$. Then (10) must be minimized at $\theta_0 = \theta_*$.*

Operationally, computing the MLE and QMLE are identical tasks. However, the inverse Fisher information matrix does not remain a correct expression of the covariance matrix of the QMLE. To see why that is the case, notice that under model misspecification the information equality fails to hold. Indeed, a classical way[6] of deriving central limit theorems is to use a Taylor series expansion for the score, and obtain the sandwich covariance matrix,

$$\mathcal{H}^{-1}\mathcal{I}\mathcal{H}^{-1}, \tag{1.11}$$

where $\mathcal{H} = E_Q\left[\frac{\partial^2}{\partial\theta^2}\log f(\mathbf{W};\theta)\right]$ and $\mathcal{I} = E_Q\left[\left(\frac{\partial}{\partial\theta}\log f(\mathbf{W};\theta)\right)^2\right]$, as the asymptotic covariance when the true underlying distribution function is Q. Then the information equality is

[6]Note, however, that under general conditions but with a well specified model, the asymptotic distribution may be derived without the use of the information equality. See, in particular, the elegant article by Sweeting (1980).

used to simplify (11) to the inverse Hessian $\mathcal{H}^{-1}$ (Ferguson, 1996). In the misspecified case, this cancelation does not obtain because the information equality does not hold in general.

That is, although using the maximum likelihood estimator when the likelihood is misspecified may give a reasonable estimate, the standard errors obtained from the inverse Hessian provide incorrect inferential statements. This motivates inquiring into the performance of the maximum likelihood estimate for inference when standard errors are computed using a sandwich covariance matrix.

1.2.3 Residual Maximum Likelihood

In the expository example (7), as well as in Madsen et al. (2008), there is only one covariate, the misaligned one. However, in many applications, such as Maccini and Yang (2009), there may be many other covariates, the coefficients of which enter the mean of the distribution only. In such instances, it will often be advisable to use residual maximum likelihood (REML) to obtain more accurate covariance estimates.

Indeed, in economic applications, the regression component of the model may contain many control variables. For instance, one may be using survey data in which many characteristics (gender, age, income, district dummies, etc) are available and used as controls for each individual. If the number of controls is large relative to the number of observations, this can induce tangible bias in the maximum likelihood covariance estimates.

This is best illustrated with a simple example. In the standard Gaussian linear regression model with homoskedastic variance σ^2 (conditional on X),

$$\mathbf{W} \sim N(X\beta, I_n\sigma^2), \tag{1.12}$$

with n observations and p control variables, the instantiation of this phenomenon is explicit. The maximum likelihood estimator for the variance, $\hat{\sigma}^2$, is biased;

$$E\left[\hat{\sigma}^2\right] = \sigma^2 \frac{n-p}{n}.$$

The reason, intuitively, is that p degrees of freedom are "spent" to first estimate the regression

mean, and the variance is then estimated with the residuals, which are left with $n-p$ degrees of freedom worth of information. In the special case of the Gaussian linear model with homoskedastic variance, the problem is readily remedied by applying a correction factor to obtain the unbiased adjusted estimator $s^2 = \frac{n}{n-p}\hat{\sigma}^2$. However, such an immediate correction is not available in general.

A way to capture the idea that the residuals have $n-p$ degrees of freedom is to note that there are exactly $n-p$ linearly independent error contrasts $\hat{\eta}_j$ (i.e. linear transformations of Y with mean 0). In this sense, the residuals live in an $n-p$ dimensional space. Even more to the point, in the canonical model reformulation of (12), the contrasts $\hat{\eta}_j$ are obtained via the QR decomposition, and I can explicitly express the unbiased estimator as a sum of $n-p$ random variables

$$s^2 = \frac{1}{n-p}\sum_{j=1}^{n-p}\hat{\eta}_j^2.$$

The general way to proceed with REML estimation is to obtain estimates by maximizing the likelihood of the error contrasts. The key difference about this approach is that it does not involve the existence and discovery of a correcting coefficient, and is thus applicable for our purposes.

Consider the Gaussian linear model,

$$\mathbf{W} \sim N\left(X\alpha, V(\theta)\right),$$

where Y is an $n \times 1$ data vector, X is an $n \times p$ matrix of covariates, α is an unknown p-long vector of coefficients, and $V(\theta)$ is an $n \times n$ positive-definite matrix known up to a vector of covariance parameters $\theta \in \Theta$ open in $\mathbb{R}^k$. Twice the negative log-likelihood is proportional to

$$\log\left(|V(\theta)|\right) + \left(\mathbf{W} - X\alpha\right)^T V^{-1}(\theta)\left(\mathbf{W} - X\alpha\right).$$

For REML, I am instead interested in maximizing the likelihood of a (in fact, *any*) vector of $n-p$ linearly independent error contrasts. That is, I instead maximize the log-likelihood of $U = \Psi^T Y$ for some $n \times (n-p)$ matrix Ψ satisfying $\Psi^T X = 0$. Clearly, $U \sim N(0, \Psi^T V(\theta)\Psi)$,

and twice the negative log-likelihood function based on U is proportional to

$$\log\left(\left|\Psi^T V(\theta)\Psi\right|\right) + U^T\left(\Psi^T V(\theta)\Psi\right)^{-1} U. \tag{1.13}$$

It is important to notice that there is no notion (or loss) of efficiency depending on which projection on the space of error contrasts is used. Specifically, any choice of $n-p$ linearly independent error contrasts yields an equivalent maximum likelihood problem. Indeed, take any $n \times (n-p)$ matrix Λ satisfying $\Lambda^T X = 0$. Then there exists a nonsingular $(n-p)\times(n-p)$ matrix D such that $\Lambda^T = D\Psi^T$. Consequently, $\tilde{U} = \Lambda^T Y$ has log-likelihood

$$\begin{aligned}
&\log\left(\left|\Lambda^T V(\theta)\Lambda\right|\right) + \tilde{U}^T\left(\Lambda^T V(\theta)\Lambda\right)^{-1}\tilde{U} \\
&= \log\left(\left|D\Psi^T V(\theta)\Psi D^T\right|\right) + U^T D^T\left(D\Psi^T V(\theta)\Psi D^T\right)^{-1} DU \\
&= \log\left(\left|\Psi^T V(\theta)\Psi\right|\right) + U^T\left(\Psi^T V(\theta)\Psi\right)^{-1} U + C,
\end{aligned} \tag{1.14}$$

where $C = 2\log(|D|)$ is a constant which depends on D but not on θ, and hence does not affect the optimization. That is, (13) and (14) yield equivalent optimization problems in θ.

Cressie and Lahiri (1993) give a brief Bayesian justification for REML, drawing on Harville (1974). They remark that if one assumes a noninformative prior for β, which is statistically independent of θ, one obtains that the marginal posterior density for θ is proportional to (13), multiplied by the prior for θ. Then, if that prior is flat, the REML estimate is equivalent to the maximum *a posteriori* estimate.

Jiang (1997) points at another interesting connection. He observes that, under normality, the best linear unbiased predictor for the random effects can be derived as the best predictor based on the error contrasts whilst, as we have seen, REML estimates are maximum likelihood estimates for the covariance parameters based on the error contrasts.

1.3 Quasi-Maximum Likelihood Estimation

My recommendation, for the estimation problem described in (1)-(4), is to use the quasi-maximum likelihood estimator (QMLE). In order to conduct inference with the QMLE, I

obtain its asymptotic distribution. Foregoing the Gaussianity assumption, I need instead some form of assumption guaranteeing that as the sample gets bigger, more information is collected. I will be using mixing assumptions, which formalize the idea that observations farther away from each other are "more independent". Technical details regarding mixing are treated in Subsection A1.

In this Section, I lay out the method, theory, and computational aspects of quasi-maximum likelihood with spatial data. Although my focus remains regression analysis with misaligned data, remark that the central limit theory developed in Subsection 3.1 applies generally to, and is novel for, quasi-maximum likelihood estimation with mixing data.

1.3.1 Limit Distribution Theory

I derive the asymptotic distribution of the maximum likelihood estimator of the Gaussian density under assumptions of weak dependence on the true underlying distribution function Q. I consider the general case (9), which encompasses (6) as well as REML. The main result of this Section, given additional technical conditions detailed in Subsection A.1, is :

> *Assume* $W = (W_1, W_2, ..., W_n)$ *is strongly mixing of size -1 with* $E[W] = 0$ *and* $Cov(W) = V_*$. *The random function* $l : \Theta \to \mathbb{R}$ *is the Gaussian likelihood function*
>
> $$l(W;\theta) = -\frac{1}{2}\log|V(\theta)| - \frac{1}{2}W^T V^{-1}(\theta)W.$$
>
> *Let* $\hat{\theta}$ *denote its maximand, the quasi-maximum log-likelihood estimator, and define the target parameter* θ_0 *to be the minimizer in (10). Then*
>
> $$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} N\left(0, H^{-1}\mathcal{I}H^{-1}\right),$$
>
> *where* $H = E_Q\left[\frac{\partial^2 l}{\partial\theta\partial\theta^T}(\theta_0)\right]$ *and* $\mathcal{I} = V_Q\left(\frac{\partial l}{\partial\theta}(\theta_0)\right)$, *in which* E_Q *and* V_Q *stand for the expectation and variance taken with respect to the true underlying distribution function Q.*

The proof strategy is similar to that of Jiang (1996) and relies on a useful decomposition for the quadratic form in the score $\partial l/\partial\theta$. For Jiang, who deals with quadratic forms in a vector of independent data, the decomposition gives a martingale, which he shows satisfies the sufficient conditions of a martingale central limit theorem (Theorem 3.2, Hall and Heyde,

1980). Correspondingly, I have quadratic forms in a vector of weakly dependent data, and will show that they can be represented in the same decomposition as mixingales, which satisfy the sufficient conditions of a mixingale central limit theorem (de Jong, 1997).

The score for the likelihood of $\theta \in \Theta \subset \mathbb{R}^q$ given in (9) is

$$\frac{\partial l}{\partial \theta} = -\frac{1}{2}\text{tr}\left(V^{-1}V_{\theta_i}\right)_i - \frac{1}{2}\left(W^T V^{\theta_i} W\right)_i$$

where $(a_i)_i = (a_1, ..., a_q)^T$, and V_{θ_i} and V^{θ_i} are the derivatives with respect to θ_i of V and V^{-1}, respectively. Assuming $\hat{\theta}(n) \xrightarrow{P} \theta_0$, Taylor series expansion yields

$$0 = \frac{\partial l}{\partial \theta}(\theta_0) + (\hat{\theta}(n) - \theta_0)^T \frac{\partial^2 l}{\partial \theta \partial \theta^T}(\theta_0) + O_p(1),$$

which implies

$$\hat{\theta}(n) = \theta_0 - \left(E\left[\frac{\partial^2 l}{\partial \theta \partial \theta^T}(\theta_0)\right]\right)^{-1} \frac{\partial l}{\partial \theta}(\theta_0) + o_p(n^{-1/2}).$$

Assume the posited covariance function is well behaved enough that

$$\left(E\left[\frac{\partial^2 l}{\partial \theta \partial \theta^T}(\theta_0)\right]\right)^{-1} V_* \left(E\left[\frac{\partial^2 l}{\partial \theta \partial \theta^T}(\theta_0)\right]\right)^{-1} = O(n^{-1}),$$

where $V_* = Cov(W)$. Then in order to obtain the limiting distribution for $\hat{\theta}(n)$, I need to obtain the limit distribution for the score $\frac{\partial l}{\partial \theta}(\theta_0)$. Note that $E\left[\frac{\partial l}{\partial \theta}(\theta_0)\right] = 0$ and thus

$$-2\frac{\partial l}{\partial \theta}(\theta_0) = \left(W^T V^{\theta_i} W - EW^T V^{\theta_i} W\right)_i.$$

Asymptotic normality of multivariate random variables can be demonstrated using the Cramér-Wold device. For a given $t \in \mathbb{R}^q \backslash \{0\}$, letting $Z = V_*^{-1/2} W$,

$$\begin{aligned}
-t^T \frac{\partial l}{\partial \theta}(\theta_0) &= t^T \left(W^T V^{\theta_l} W - EW^T V^{\theta_l} W\right)_l \\
&= t^T \left(Z^T A^l Z - EZ^T A^l Z\right)_l \\
&= \sum_i \left(a_{ii}\left(Z_i^2 - E\left[Z_i^2\right]\right) + 2\sum_{j<i} a_{ij}\left(Z_i Z_j - E\left[Z_i Z_j\right]\right)\right),
\end{aligned}$$

where $A^l = V_*^{1/2} V^{\theta_l} V_*^{1/2}$ and $a_{ij} = \sum_{l=1}^{q} t_l A_{ij}^l$, $\forall\ i, j$ with A^l. Letting

$$X_{n,i} = a_{ii}\left(Z_i^2 - E\left[Z_i^2\right]\right) + 2\sum_{j<i} a_{ij}\left(Z_i Z_j - E\left[Z_i Z_j\right]\right), \tag{1.15}$$

I can write the quadratic form as

$$-t^T \frac{\partial l}{\partial \theta}(\theta_0) = \sum_{i=1}^{n} X_{n,i}.$$

Defining

$$\mathcal{F}_{n,i} = \sigma\left(Z_{n1}, ..., Z_{n,i}\right), 1 \leq i \leq k_n, \tag{1.16}$$

I show that $\{X_{n,i}, \mathcal{F}_{n,i}\}$, $1 \leq i \leq k_n$ is an array of mixingales. This is an important observation because it will allow me to use a mixingale central limit theorem to derive the limit distribution of the score, and thus of the quasi-maximum likelihood estimator.

Definition 1 *A triangular array $\{X_{n,i}, \mathcal{F}_{n,i}\}$ is an L_2-mixingale of size $-\lambda$ if for $m \geq 0$*

$$\left\|X_{n,i} - E\left[X_{n,i} \middle| \mathcal{F}_{n,i+m}\right]\right\|_2 \leq a_{n,i}\psi(m+1),$$

$$\left\|E\left[X_{n,i} \middle| \mathcal{F}_{n,i-m}\right]\right\|_2 \leq a_{n,i}\psi(m),$$

and $\psi(m) = O(e^{-\lambda-\varepsilon})$ for some $\varepsilon > 0$. We call the $a_{n,i}$ mixingale indices.

The mixingale CLT requires that the mixingale be of size $-1/2$, which in turn implies a requirement on the rate of decay of the dependence for variables in the random field.

Lemma 2 *Suppose that $\{Z_{n,i}, \mathcal{F}_{n,i}\}$ is an α-mixing sequence of size -1 and that all $Z_{n,i}$'s have a fourth moment. Then the triangular array $\{X_{n,i}, \mathcal{F}_{n,i}\}$ defined in (15) and (16) is an L_2-mixingale of size $-1/2$.*

The proof is deferred to the appendix.

1.3.2 Computational Aspects

When computing the sandwich covariance formula $H^{-1}\mathcal{I}H^{-1}$ obtained in Subsection 3.1, some challenges need to be dealt with. Estimation of H^{-1} is standard and must be done in the well-specified case. It is in fact an even lighter task when the covariance functions are modeled since a closed form for H immediately obtains. It is the estimation of $\mathcal{I}$ which raises new challenges.

Indeed, computing the variance of the score may at first seem daunting, even a motivation for sticking with the Hessian as a covariance estimate. In particular, it involves the estimation of fourth-order moments. The difficulty arises from the computation of the covariance of entries of the score corresponding to covariance model parameters. For two such parameters θ_l and θ_m (say, the sill of the rain covariance function and the range of the regression error covariance function) we have

$$E\left[\frac{dl}{d\theta_l}(\theta_0)\frac{dl}{d\theta_m}(\theta_0)\right] = \frac{1}{4}E\left[\left(W^TV^{\theta_l}W - E\left[W^TV^{\theta_l}W\right]\right)\left(W^TV^{\theta_m}W - E\left[W^TV^{\theta_m}W\right]\right)\right]$$

$$= \frac{1}{4}\sum_{i,j}\sum_{i',j'}V^{\theta_l}_{i,j}V^{\theta_m}_{i',j'}\left(E\left[W_iW_jW_{i'}W_{j'}\right] - E\left[W_iW_j\right]E\left[W_{i'}W_{j'}\right]\right),$$

where the dependence on fourth-order moments is made explicit.

The problem of approximating and estimating the asymptotic variance is divided into two cases, depending on the robustness concerns. If the main concern is that the covariance function is misspecified, then a fruitful approach is to keep with the quasi-Gaussian mindset, and estimate the fourth-order moments using the Gaussian formula. The more important case, however, is that in which the researcher is not confident that the Gaussian assumption is a tenable one. Then robustness concerns call for estimation of the fourth-order moments. In order to accomplish that, I provide a shrinkage estimator inspired from the finance literature.

Approximation Using Gaussian Formula

The researcher may be comfortable with the Gaussianity assumption, and principally concerned with robustness to covariance function misspecification. This motivates the use of a "quasi"

estimate of the covariance, whose estimation difficulty is equivalent to that of the Hessian's, and only requires the estimation of second-order moments, which furthermore are already modeled.

The normal distribution is entirely determined by the first two moments, hence so are the higher-order moments of normal random variables. Indeed, their explicit expression is given by Isserlis' theorem (also known as Wick's formula). For fourth-order moments, we have that if W_1, W_2, W_3, W_4 are drawn from a multivariate normal, then

$$E\left[W_1 W_2 W_3 W_4\right] = E[W_1 W_2]E[W_3 W_4] + E[W_1 W_3]E[W_2 W_4] + E[W_1 W_4]E[W_2 W_3].$$

Using Isserlis' theorem, the expression for the covariance simplifies to

$$\begin{aligned} E\left[\frac{dl}{d\theta_l}(\theta_0)\frac{dl}{d\theta_m}(\theta_0)\right] &= \frac{1}{4}\sum_{i,j}\sum_{i',j'} V_{i,j}^{\theta_l} V_{i',j'}^{\theta_m}\left(V_{ij}V_{i'j'} + V_{ii'}V_{jj'} + V_{ij'}V_{ji'} - V_{ij}V_{i'j'}\right) \quad (1.17) \\ &= \frac{1}{4}\sum_{i,j}\sum_{i',j'} V_{i,j}^{\theta_l} V_{i',j'}^{\theta_m}\left(V_{ii'}V_{jj'} + V_{ij'}V_{ji'}\right), \quad (1.18) \end{aligned}$$

and is now in terms of second-order moments only.

It is natural to ask in what sense being "close to Gaussian" makes for a good approximation of the high-order moments using Wick's theorem. The QMLE, by construction, minimizes the Kullback-Leibler divergence with the true distribution. Hence, we may think that if our approximations are reasonable, this divergence will be small. Theorem 5 below and its proof show that the behavior of the tails also matters for the approximation of fourth-order moments with Wick's formula to be reasonable. One way to speak of the tail behavior of a random variable is to compare it to that of Gaussian random variables. In particular, we will say that a random variable is sub-Gaussian if its tails are dominated by those of some Gaussian random variable.

Definition *A mean-zero random variable X is sub-Gaussian with parameter ν^2 if, for all $\lambda \in \mathbb{R}$,*

$$E\left[e^{\lambda X}\right] \leq \exp\left(\frac{\lambda^2\nu^2}{2}\right).$$

I now give a theorem characterizing conditions under which the moment approximation is tenable.

Theorem 5 *Consider a sequence of pairs of distributions* $\{(P_i, F_i)\}$ *for which*

$$D_{KL}(F_i||P_i) \to 0, \tag{1.19}$$

as $i \to \infty$*, where* $D_{KL}(F||P) = \int \ln \frac{dF}{dP} dF$ *is the Kullback-Leibler divergence. Further suppose that the* P_i*,* $i = 1, 2, ...$ *are sub-Gaussian of some given parameter. Let the* N*-tuple* $Y = Y(n)$ *follow distribution* P_n *and the* N*-tuple* $X = X(n)$ *be Gaussian with distribution* F_n*. Then*

$$|E[Y_i Y_j Y_k Y_l] - (E[X_i X_j]E[X_k X_l] + E[X_i X_k]E[X_j X_l] + E[X_i X_l]E[X_j X_k])| \to 0, \tag{1.20}$$

$1 \leq i, j, k, l \leq N$*, as* $n \to \infty$*.*

The proof is given in the appendix, and consists in getting a multivariate Pinsker inequality via a residue theorem. The sub-Gaussian assumption can be replaced by an assumption of uniform integrability on the moments of order $4 + \epsilon$, for some $\epsilon > 0$.

What comes out of the proof of Theorem 5 is that a small Kullback-Leibler distance takes care of the bulk of the density, but we must also assume that the tails vanish in order for the fourth moments to be close. This is particularly nice because these are two features of the data which the researcher would inspect on a QQ-plot.

Of course, submitting this quadruple sum as a loop is prohibitively inefficient, and should not be done as such. Using the indexation, I can rearrange the formula in a computationally economical matrix form. Simply observe that

$$\sum_{i,j} \sum_{i',j'} V_{i,j}^{\theta_l} V_{i',j'}^{\theta_m} \left(V_{ii'} V_{jj'} + V_{ij'} V_{ji'} \right) = \sum_{i,j} \sum_{i',j'} V_{i,j}^{\theta_l} V_{i',j'}^{\theta_m} V_{ii'} V_{jj'} + \sum_{i,j} \sum_{i',j'} V_{i,j}^{\theta_l} V_{i',j'}^{\theta_m} V_{ij'} V_{ji'}$$

$$= \left\| V^{\theta_l} \circ \left(V V^{\theta_m} V \right) \right\|_{\text{sum}} + \left\| V^{\theta_l} \circ \left(V \left(V^{\theta_m} \right)^T V \right) \right\|_{\text{sum}},$$

where $\circ$ is the element-wise multiplication (i.e. for $A = (a_{ij})$ the matrix with (i, j) entry a_{ij}

and $B = (b_{ij})$ defined likewise, $A \circ B = (a_{ij}b_{ij})$) and $\|A\|_{\text{sum}} = \sum_{i,j} a_{ij}$.

To be sure, such an approach is particularly well-suited to the case of covariance misspecification. If the main concern is distribution misspecification, fourth-order moments ought to be estimated.

Estimating Fourth-Order Moments

A key case is that in which first and second moments are well specified but Gaussianity is suspect. In that case, the estimated coefficient is consistent (e.g. the target parameter for $\hat{\beta}$ is the true value β), making the QMLE particularly attractive. However, in that case, the Gaussian formula provides an untenable approximation for the fourth-order moments, and one is well-advised to estimate them.

Estimation of fourth-order moments is difficult. Additional assumptions and more sophisticated methods need to be called upon in order to carry out this task. Some early approaches (Elton & Gruber, 1973) have allowed for less volatile estimation, but at the cost of potentially heavy misspecification.

Recent strides in covariance matrix estimation have consisted in attempts to trade off variance and bias under some optimality criterion. In particular, Ledoit and Wolf (2003) suggested a shrinking estimator, and gave a theoretically founded optimal shrinkage parameter which has an empirical analog. Martellini and Ziemann (2010) have extended these results to higher-order moment tensors. Both articles had their applications in finance.

I extend the method displayed in Martellini and Ziemann (2010), and like them rely on the theory developed in Ledoit and Wolf (2003) to suggest a shrinkage estimator for fourth-order moments.

In order to get some level of replication with spatial data, I must make stationarity assumptions. Here I assume fourth-order stationarity; $E\left[W(x_1)W(x_2)W(x_3)W(x_4)\right]$ only depends on the relative positions of x_1,...,x_4, i.e.

$$E\left[W(x_1)W(x_2)W(x_3)W(x_4)\right] = f(x_2 - x_1, x_3 - x_1, x_4 - x_1) \tag{1.21}$$

for some f and will be approximated (still with centered random variables) by

$$\frac{1}{|N_{d_1 d_2 d_3}|} \sum_{(i,j,k,l) \in N_{d_1 d_2 d_3}} W_i W_j W_k W_l$$

for an appropriate choice of bin $N_{d_1 d_2 d_3}$, which is a set of quadruples (x_i, x_j, x_k, x_l) for which $(x_i - x_j, x_i - x_k, x_i - x_l) \approx (d_1, d_2, d_3)$, under some approximation criteria.

Since the data is irregularly spaced, some approximation measure must be taken to obtain replications of estimated moments and thus leverage the fourth-order stationarity assumption. One approach is to emulate the construction of the empirical variogram and to count as replicates of a given fourth-order product moment other such moments the configuration of whose points is approximately the same. This approach requires delicate choices when defining what "approximately the same" should mean, which corresponds to the choice of bins (or even kernels). Using a polar partition of the lag space (i.e. binning by distance from, and radial degrees around, a center point), I can get bins which will be separated in terms of the distances between points and the angles between pairs of points.

An easier approach is to "gridify" the data. That is, I lay an equally spaced grid over the sampled space and, for each square of the grid, I consider the point at its center to be an observation whose value is the average over the square. This creates a data set on a grid which approximates the original, irregular space data set. The reason for doing this is that nonparametric estimation of fourth-order moments is then straightforward (in particular, the bins are defined by an equality $(x_i - x_j, x_i - x_k, x_i - x_l) = (d_1, d_2, d_3)$) and the approximation assumptions are transparent.

Note that, as opposed to the case of variogram estimation, I do not fit a fourth-order moment function to this nonparametric estimate. Hence, given the location of four random variables, the fitted value for their product moment given by this estimator is simply the average over the bin it corresponds to.

I call the resulting nonparametric estimator S. I want to stabilize the estimator S. A stable but potentially misspecified fourth-order moment tensor is obtained by replacing each fourth-order moment by its Gaussian expression (in terms of second-order moments). I call

this tensor the Isserlis tensor, and label it Λ. The shrinkage estimator I propose is

$$\alpha\Lambda + (1-\alpha)S.$$

The critical step is then the choice of the tuning parameter α.

Ledoit and Wolf (2003) give a theoretically founded approach for picking α when dealing with covariance matrices, and Martellini and Ziemann (2010) extend the approach to deal with tensors of higher order moments. I follow their approach.

The suggested parameter is

$$\hat{\alpha} = \frac{1}{N}\frac{\hat{\pi}}{\hat{\gamma}},$$

where $\hat{\pi} = \sum_{ijkl} \hat{\pi}_{ijkl}$ and $\hat{\gamma} = \sum_{ijkl} \hat{\gamma}_{ijkl}$, with

$$\hat{\pi}_{ijkl} = \frac{1}{|N_{d_1 d_2 d_3}|} \sum_{(i,j,k,l)\in N_{d_1 d_2 d_3}} \left(W_i W_j W_k W_l - S_{ijkl}\right)^2$$

as the sample variance of the tensor, and

$$\hat{\gamma}_{ijkl} = \left(\Lambda_{ijkl} - S_{ijkl}\right)^2$$

is the sample squared error of the structured estimator.

Martellini and Ziemann (2010) suggest accounting for the covariance between entries of S and Λ by adding a covariance estimate term in the numerator. However, this is a difficult quantity to estimate, and since Λ is an order of magnitude less variable than S, I feel comfortable omitting that term so to propose a much simpler procedure.

An approximation which I recommend considering is the use of the Gaussian formula for the $\hat{\pi}_{ijkl}$'s. It gives weights in terms of a sensible function of the configuration, and avoids the rather hopeless reliance on estimated eighth moments (especially since the motivation for using shrinkage in the first place is the high variance of the fourth moments).

The next order of business is to assess the performance of the nonparametric and shrinkage estimators. I assess the performance of the estimators by considering cases in which I can

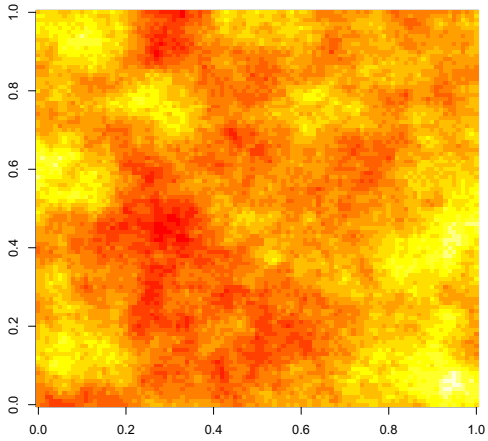

Figure 1.3: *Simulated Gaussian Random Field*

Simulation of a Gaussian random Field with range parameter 0.15.

compare estimated fourth-order moments with their correct values. For Gaussian random variables, the value of fourth-order moments is obtained in closed form using Wick's formula (since I have closed forms for the covariances) and thus the comparison is readily made. Figure 3 presents a heat map of the simulated Gaussian random field.

To visualize and compare the estimated fourth moments with the true ones, I present plots constructed as follows. Each plot presents the complete estimation domain (e.g. the geographical area under study). I fix the location of three of the four points whose fourth-order product moment I estimate (and indicate the location of those three points on the plot, some may overlap). The value at each location on the plot is that of the fourth-order product moment (true or estimated, depending on the plot) whose fourth point is at that location.

From inspection of Figures 4 and 5, we can see that the approximation is good near the location of the three fixed points, and deteriorates farther away from them. This further suggests that, although it is beneficial to leverage the nonparametric estimator for its robustness, embedding it in a shrinkage estimator may make for a more satisfactory method.

We can see from Figure 5 that shrinkage helps attenuate the more erratic behavior of the

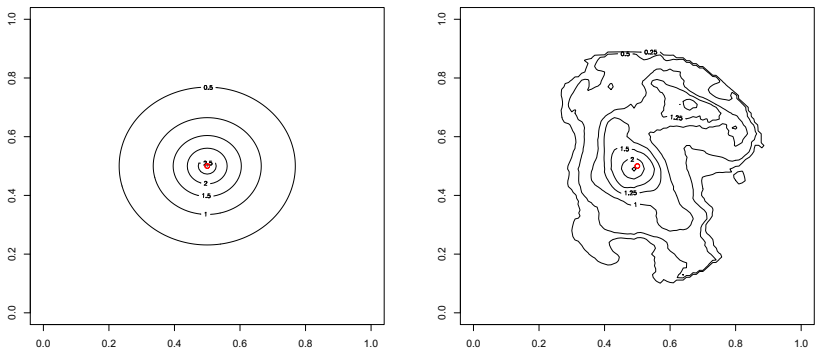

Figure 1.4: *Spatial Fourth Order Moments*

Plot of fourth-order moment $E[W(x_1)W(x_2)W(x_3)W(x_4)]$ as a function of the position x_4 with $x_1 = x_2 = x_3 = (0.5, 0.5)$ and range parameter 0.15. The left-hand side figure gives the contour plot of the true moment. The right-hand side figure gives the contour plot of the nonparametric estimates.

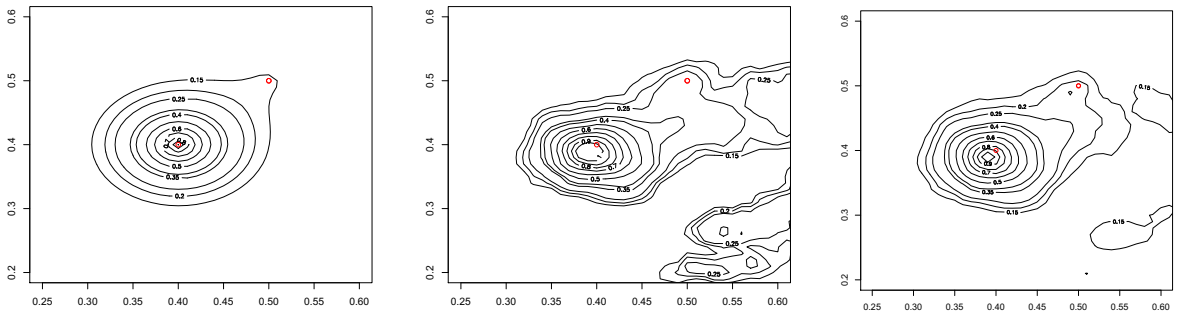

Figure 1.5: *Estimating Fourth Order Moments*

Plot of fourth-order moment $E[W(x_1)W(x_2)W(x_3)W(x_4)]$ as a function of the position x_4, with $x_1 = x_2 = (0.5, 0.5)$, $x_3 = (0.4, 0.4)$ and range parameter 0.05. The left-hand side figure gives the contour plot of the true moments. The middle figure gives the contour plot of the nonparametric estimates. The right-hand side plot gives the contour plot of the estimates using shrinkage.

nonparametric fourth-moment estimator. The undesired local maxima are repressed, and the contours around the true local maxima are better behaved.

1.4 Robust Estimation

Specification of the covariance model for the regression errors is arguably the most onerous assumption the researcher needs to make. One may worry that the regression error covariance does not have the nice structure common to geophysical variables, e.g. since we ineluctably omit variables in the regression function. This motivates an estimation scheme relieved of that specification burden.

Upon inspection of the marginal density (5), we find that β is identified from $V_{R^*} = K^*$ and $V_{YR^*} = \beta \bar{K}^T$ alone, and hence identification does not require modeling the covariance structure of the regression errors. Specifically, using rainfall data only, one obtains an estimate of the covariance parameter θ, and thus has an estimate of $\bar{K}$. Since $\beta\bar{K}$ is directly identified from the covariance between the outcome data and the observed rain, β is identified. This naturally invites a procedure which will rely on this identification observation to yield a robust estimate of β. In this Section, I lay out such a procedure.

I develop a robust estimation method for the regression coefficient. In order to conduct inference, I develop limit distribution for the robust estimator. The limit distribution theory (in particular the central limit theorem for the moment vector with irregular spaced data and increasing domain asymptotics), however, is useful and novel in the general case.

1.4.1 Methodology

Let $\gamma_{R^*}(d;\theta) = V_\theta(R(x) - R(x+d))$ and $\gamma_{YR^*}(d;\theta) = V_\theta(R(x) - Y(x+d))$ be the variogram of R^* and the covariogram of Y with R^*, respectively. Note that $\gamma_{YR^*}(d;\theta) = \beta\gamma_{R^*}(d;\theta)$. Let $\gamma^*_{R^*}(d)$ and $\gamma^*_{YR^*}(d)$ be generic, nonparametric estimators of $\gamma_{R^*}(d;\theta)$ and $\gamma_{YR^*}(d;\theta)$, respectively.

Let $\{c_1, ..., c_{K_{R^*}}\}$ and $\{d_1, ..., d_{K_{YR^*}}\}$ be finite sets of lags such that $\gamma^*_{R^*}(d_i)$ is defined for

all $i = 1, ..., K_{R^*}$ and $\gamma^*_{YR^*}(c_j)$ is defined for all $j = 1, ..., K_{YR^*}$.[7] Let $\phi = (\beta, \theta)$ and

$$g_n(\theta) = 2\left(\gamma^*_{YR^*}(d_1) - \gamma_{YR^*}(d_1;\theta), \cdots, \gamma^*_{YR^*}(d_K) - \gamma_{YR^*}(d_{K_{YR^*}};\theta),\right.$$

$$\left.\gamma^*_{R^*}(c_1) - \gamma_{R^*}(c_1;\theta), \ldots, \gamma^*_{R^*}(c_{K_{R^*}}) - \gamma_{R^*}(c_{K_{R^*}};\theta)\right).$$

For some positive-definite weighting matrix B_n, define

$$\hat{\theta}_{\text{Robust}} = \arg\min_{\theta\in\Theta}\left\{g_n(\theta)^T B_n g_n(\theta)\right\}.$$

Then $\hat{\beta}_{\text{Robust}}$, the estimate of β, does not depend on any specification of the covariance structure of the regression errors. Different choices of B_n will correspond to different traditional estimators; $B_n = I$ yields the ordinary least squares estimator, $B_n = \text{diag}(b_{n,1}(\theta), ..., b_{n,K_{YR^*}+K_{R^*}}(\theta))$, for some choice of weights $b_{n,i}$, $i = 1, ..., K$, gives the weighted least squares estimator, and $B_n(\theta) = \Sigma_\gamma^{-1}(\theta)$, where $\Sigma_\gamma(\theta)$ is the asymptotic covariance matrix of $2\left(\gamma^*_{YR^*}(c_1), \cdots, \gamma^*_{R^*}(d_{K_{YR^*}})\right)^T$ is the generalized least-square version of the minimum-distance estimator.

Another attractive feature of this estimator is its flexibility. The vector of moments can be extended to accommodate other conditions, perhaps motivated by economic theory.

1.4.2 Limit Distribution Theory

In order to carry out inference using the proposed minimum distance estimator, I need asymptotic distribution theory for the statistic, that is the empirical variogram (defined below). Cressie et al. (2002) give such a theorem in the case in which the data is on a lattice, and even give the asymptotic distribution of the minimum distance estimator as a corollary. Lahiri (2003) proves a series of very useful central limit theorems for spatial statistics, one of which can be leveraged to extend the asymptotic theory for the empirical variogram to the case of irregular data. I give this result in the present Section.

[7] The lags can be the default lags of the empirical variogram estimator from a geostatistical package, such as `gstat`.

Let the empirical variogram (p.34, Gelfand et al., 2010) be defined

$$\hat{\gamma}(h) = \frac{1}{|N_n(h)|} \sum_{(s_i,s_j)\in N_n(h)} \left(\hat{\varepsilon}(s_i) - \hat{\varepsilon}(s_j)\right)^2,$$

where the bin $N_n(h)$ is the set of pairs of observations separated by a distance close to h, which I have chosen to approximate $\gamma(h, \theta_0)$.

In spatial econometrics and statistics, limit distribution theory depends on the asymptotic domain chosen by the analyst. One may consider the pure-increasing domain (more and more points, always at a minimum distance from each other), infill asymptotics (more and more points in a fixed, finite area), or a mix of the two (points get denser and over a wider area as their number increases). The data sets analyzed herein have many observations, generally distant from each other (measured both by geographic distance and correlation), which is typical for social science applications (as opposed to mining or medical imaging). Hence, to best capture these features of the data, all asymptotics in this article are done in the pure-increasing domain.

The limit distribution theory can be obtained in the pure-increasing domain, with the so-called stochastic design (Lahiri, 2003). Explicitly, the sampling region is denoted $\mathcal{R}_n$, and is for each n a multiple of a prototype region $\mathcal{R}_0$, defined as follows. The prototype region satisfies $\mathcal{R}_0^* \subset \mathcal{R}_0 \subset \bar{\mathcal{R}}_0^*$, where $\mathcal{R}_0^*$ is an open connected subset of $(-1/2, 1/2]^d$ containing the origin. Let $\{\lambda_n\}_{n\in\mathbb{N}}$ be a sequence of positive real numbers such that $n^\epsilon/\lambda_n \to 0$ as $n \to \infty$ for some $\epsilon > 0$. Then the sampling region is defined as

$$\mathcal{R}_n = \lambda_n \mathcal{R}_0.$$

To avoid pathological cases, I will assume that the boundary of $\mathcal{R}_0$ is delineated by a smooth function. This assumption can be modified and weakened to adapt to other domains (e.g. star-shaped ones), see Lahiri (2003).

Furthermore, we speak of a stochastic design because the data is not placed on a lattice, and observation locations must be modeled otherwise. They are modeled as follows. Let $f(x)$ be a continuous, everywhere positive density on $\mathcal{R}_0$, and let $\{X_n\}_n$ be a sequence of independent

and identically distributed draws from f. Let $x_1, ..., x_n$ be realizations of $X_1, ..., X_n$, and define the locations $s_1, ..., s_n$ of the observed data in $\mathcal{R}_n$ as

$$s_i = \lambda_n x_i, \ i = 1, ..., n.$$

In the stochastic design, pure-increasing asymptotics require that $n/\lambda_n^2 \to C$ for some $C \in (0, \infty)$ as $n \to \infty$.

First, I obtain a central limit theorem for the statistic entering the minimum distance objective function. For simplicity, take g_n as defined above but let $\{d_1, ..., d_K\}$, $K \in \mathbb{N}$, be the full set of lags. Recall that m and $\hat{m}$ are the mean and estimated mean, respectively, of the random field (e.g. of rainfall).

Theorem 6 *Suppose that $f(x)$ is continuous and everywhere positive on R_0 and that $\lambda_n^2 \left\| \hat{m} - m \right\|_2^4 = o_p(1)$. Further assume that $E\left|R(0)\right|^{4+\delta} < \infty$ and for all $a \geq 1$, $b \geq 1$, $\alpha(a,b) \leq Ca^{-\tau_1}b^{\tau_2}$ for some $0 < \delta \leq 4$, $C > 0$, $\tau_1 > (4+d)/\delta$, and $0 \leq \tau_2 < \tau_1/d$. Then if $n/\lambda_n^2 \to C_1$,*

$$n^{1/2} g_n(\theta_0) \to N\left(0, \Sigma_g(\theta_0)\right),$$

where the i,j entry of the covariance matrix is $\left[\Sigma_g(\theta_0)\right]_{i,j} = \sigma_{ij}(0) + Q \cdot C_1 \cdot \int_{\mathbb{R}^2} \sigma_{ij}(x)x$, with $Q = \int_{R_n} f^2(s)ds$ and $\sigma_{ij}(x) = Cov_{\theta_0}\left(\left(R(0) - R(d_i)\right)^2, \left(R(s) - R(x + d_j)\right)^2\right)$.

With the limit distribution of the statistic in hand, the central limit distribution of the minimum distance estimator obtains.

Theorem 7 *Suppose that $f(x)$ is continuous and everywhere positive on R_0 and that $\lambda_n^2 \left\| \hat{m} - m \right\|_2^4 = o_p(1)$. Further assume that $E\left|R(0)\right|^{4+\delta} < \infty$ and for all $a \geq 1$, $b \geq 1$, $\alpha(a,b) \leq Ca^{-\tau_1}b^{\tau_2}$ for some $0 < \delta \leq 4$, $C > 0$, $\tau_1 > (4+d)/\delta$, and $0 \leq \tau_2 < \tau_1/d$. Then under the conditions cited in Assumption A.1, if $n/\lambda_n^2 \to C_1$ and the matrix of partial derivatives $\Gamma(\theta_0) = -2\left(g_1(\theta_0); ...; g_K(\theta_0)\right)$ is full rank,*

$$n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N\left(0, \Sigma(\theta_0)\right),$$

where $\Sigma(\theta_0) = A(\theta_0)\Gamma(\theta_0)^T B(\theta_0)\Sigma_g(\theta_0)B(\theta_0)\Gamma(\theta_0)A(\theta_0)$, $A(\theta_0) = \left(\Gamma(\theta_0)^T B(\theta_0)\Gamma(\theta_0)\right)^{-1}$.

The density of the observation locations f has an intuitive impact on the asymptotic covariance. As one would expect, if the observations are well spread geographically then this makes for a lower variance because $Q = \int_{R_n} f^2(x)dx$ is smaller. Correspondingly, cluttered data arranged as a few clusters provides worse information, and the variance is greater for it.

Comparison of the methods detailed in Sections 3 and 4 illustrates tradeoffs facing the researcher when estimating β. At the cost of ignoring the information about β in the covariance of the outcome variable Y, the researcher obtains a less efficient estimator, but one which does not require specification of the regression error covariance Σ.

In many economic applications, the estimated effect can be weak and the efficient method will be preferred. In that case, the minimum distance estimator of this Section provides a useful robustness check.

The methods developed in Chapter I are used to reanalyze the data set of Maccini and Yang (2009) in Chapter II, in which a third, more robust method is presented.

Chapter 2

Two-Step Method for Misaligned Regression

2.1 Introduction

The main motivation for resorting to a two-step method such as Krig-and-Regress is the desire to avoid specifying a model for the covariance of the regression errors, Σ. However, correct inference requires standard errors which nevertheless take into account the uncertainty brought about by the imputation of the missing covariates. That is, one needs a two-step estimator for which neither estimation nor inference requires knowledge of Σ, and whose standard errors account for the uncertainty in the first step.

2.2 Two-Step Bayesian Bootstrap

Let $\mathscr{D} \subset \mathbb{R}^2$ be the geographic domain under study. Let R and Y be the random fields for rainfall and the outcome variable (say height), respectively. Let X^* be the (fixed) locations of the rainfall measurement stations. Let $\hat{R} = E^*[R|R^*]$, where E^* is the best linear prediction operator and $R^* = R(X^*)$ is the observed rainfall. Let X be the (random) locations of surveyed households.

The key observation is that if all the uncertainty in the second step arises from the resam-

pling (with replacement) of the surveyed households, then the observations are independently distributed, conditional on the realization of R and Y. That is, conditional on the realization of the random field of rainfall and the outcome variable at all households (but unconditional on which household is randomly drawn with replacement from the population), the observations are independent and identically distributed.

I believe the assumption of resampling with replacement is innocuous. Certainly, sampling without replacement describes more accurately the sampling protocol of the survey. Nevertheless, the survey size is so small compared to the population size that both sampling methods (with and without replacement) yield the same sample with overwhelming probability.

2.2.1 Inference

The researcher may conclude that modeling the covariance of the regression errors is too restrictive. For instance, specification of different but equally credible covariance structures may yield tangibly different estimators. The robust estimation method of Section 4 then remains a viable alternative. However, some researchers may prefer employing a two-step method.

In that case, correct inference requires standard errors which will take into account the variation brought about by the estimation of the imputed regressor.

Versions of this problem have come up in the literature under many guises (see for instance Pagan (1984)). A very general case was addressed by Murphy and Topel (1985), who provided an asymptotic covariance formula with a positive-definite correction term accounting for the variation due to the estimation of the imputed regressors.

Maccini and Yang (2009) take into account the variation due to imputation by performing inference using two-stage least-squares. However, identification under this approach is hard to argue, and estimates of the regression coefficients need not be consistent (see Subsection 7.2).

Madsen et al. (2008) work out standard errors for the regression coefficient estimated using the Krig-and-Regress method. They provide a protocol for estimating the unconditional (on the realization of the random field for the misaligned regressor) variance of the regression

coefficient. They find, in their application, that the produced unconditional standard errors differ only mildly from the OLS standard errors. Crucially, the standard errors they provide do not account for the estimation of the covariance parameter for rainfall.

My concern, in contrast with that of Madsen et al. (2008), is not to produce variance estimates for the unconditional distribution of the regression coefficient β. My concern is to provide confidence intervals that take into consideration the uncertainty due to the estimation of the imputed regressor. Note that in the Krig-and-Regress method, accounting for the variation due to the estimation of the imputed regressor is tantamount to accounting for the variance due to the estimation of the covariance and mean parameters of the random field of rainfall.

Since the motivation for using a two-step method is to avoid the modeling burden of specifying the covariance structure for the regression errors, I ought to produce standard errors that do not require evaluating Σ. In particular, the standard errors equivalent to those provided by Murphy and Topel (1985) would require such a specification.

It is plausible that the residual errors of the best linear predictor

$$Y(X) - E^*[Y(X)|R(X)],$$

where E^* is the best linear prediction operator, are spatially correlated (for instance, through an omitted variable such as pollution, which concentrates differently in different areas). Uncertainty assessments of the the Krig-and-Regress coefficient estimates, if we do not condition on the realization of R and Y, are thus bound to rely on the estimation of Σ.[1]

2.2.2 Point Estimation

To be sure, estimation of the regression error covariance Σ may be required for different reasons. If one wants to use the maximum likelihood one-step estimator, then one must specify the covariance structure for the regression errors in order to specify the likelihood and proceed with estimation. Reluctance to specify such a covariance structure may strike some researchers

[1] Unless done with standard errors so conservative that they are useless in practice.

as an encouragement to stick with a two-step method, in particular with a Krig-and-Regress approach. However, under some data generating processes, efficiency concerns may urge the specification of Σ, even for implementation of a two-step method such as Krig-and-Regress. Indeed, in some instances (such as Madsen et al. (2008), see Section 6), the small sample improvements from using a feasible generalized least-squares approach (rather than ordinary least-squares) in the second step can be compelling. That is, even point estimation using the two-step method Krig-and-Regress may require estimation of Σ.

However, if the original motivation for using a two-step estimator is to not specify Σ, a different two-step method must be considered.

2.2.3 Identification Under Survey Sampling

Importantly, one can construct a different estimator which exploits differently the sources of randomness and provides correct inference without relying on knowledge of Σ. Both QML and Krig-and-Regress consistently estimate $E^*[Y(X)|R(X), X]$, where uncertainty in the regression coefficients is thought to capture the variation arising from repeated samples of (Y, R, X). However, I can instead treat the problem as one of survey sampling and estimate the linear regression coefficient of $Y(\mathcal{X})$ on $R(\mathcal{X})$ over the whole population (say, of Indonesian households) for the given realizations of the random fields Y and R, where $\mathcal{X}$ is the vector of locations of all (Indonesian) households. The variation comes from the random selection of the survey households, whose locations are collected in X. If the surveyed households are drawn with replacement from the full population, then the corresponding observations (Y_i, R_i) sampled with replacement from $(Y(\mathcal{X}), R(\mathcal{X}))$ will be independent and identically distributed. In particular, under survey sampling, estimation of the regression coefficient will not rely on the specification of Σ.

2.2.4 Two-step Bayesian Bootstrap

I present an estimator which consistently estimates the (Indonesian) population regression coefficient and does not require specification of Σ for point estimation or inference.

Let $\pi(\theta)$ be the prior distribution of the rainfall covariance coefficient θ (e.g. a Gaussian or an improper uniform prior) and consider its posterior $\bar{\pi}(\theta) \propto f(R^*|\theta)\pi(\theta)$.

Given θ, and conditional on the realization of the random field of rainfall as well as the outcome variable for each household of the population, $\hat{\beta}$ only depends on which households are drawn (randomly, with replacement) to be part of the survey. This variation is captured by resampling using the Bayesian bootstrap. This naturally suggests a two-step Bayesian bootstrap procedure in which θ is first drawn from its posterior to determine $\hat{R}(\theta)$, thus capturing the uncertainty due to the estimation of θ.

The full procedure is described in pseudocode as follows:

- For each $j = 1, ..., J$
  - Draw $\theta^{(j)} \sim \bar{\pi}(\theta)$
  - Compute $\hat{R}^{(j)} = \hat{R}(\theta^{(j)})$
  - Generate $V_1, ..., V_n$ as n independent draws from an exponential with scale parameter 1
  - Compute $\beta^{(j)}$ by regressing $(\sqrt{V_i}Y_i)_{i=1,...,n}$ on $(\sqrt{V_i}\hat{R}_i^{(j)})_{i=1,...,n}$.

Quantiles from the set of posterior draws $\{\beta^{(j)}\}_{j=1,...,J}$ can be used to form credible intervals, and the posterior mode, mean or median can be used to give a point estimate.

2.2.5 Simulation

I implement the two-step Bayesian bootstrap estimator in a simulation exercise. I simulate a small data set of 80 observations ($M = 40,\ N = 40$). The data has been generated such that the (random) location of a household is correlated with its outcome variable. Hence, obtaining the unconditional sampling distribution of the Krig-and-Regress estimator would require careful modeling of the covariance structure of the regression errors. The true value for the regression coefficient is $\beta = 5$.

I want to assess the impact of incorporating the uncertainty due to the estimation of the imputed variables on the credible intervals for β. In order to do that, I compare the two-step

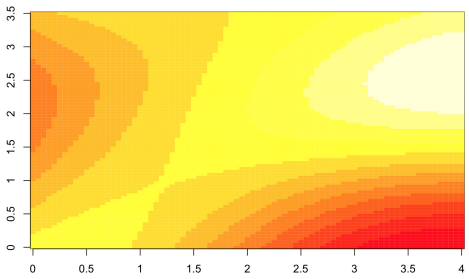

Figure 2.1: *Simulated Random Field of Rainfall*

Heat map of realization of random field of rainfall. The range parameter is 3.

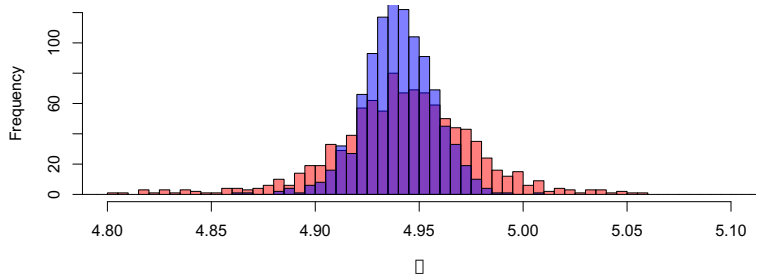

Figure 2.2: *Posterior Distributions for β*

Posterior distributions for β for $\theta = \hat{\theta}$ (blue) and $\theta \sim \bar{\pi}(\theta)$ (red).

Bayesian bootstrap estimator with one which is conditional on θ having the value $\hat{\theta}$, obtained from maximum likelihood estimation. That is, the procedure is the same as described in Subsection 5.4, with the exception that $\theta^{(j)} = \hat{\theta}$ for $j = 1, ..., J$.

We can see from Figure 2 that both posterior distributions have their modes slightly to the left of the true value for β. More to the point, the posterior distribution for the estimator which does account for the uncertainty in θ is less concentrated around its mean, and has thicker tails.

The variance, when accounting for the the uncertainty in the imputed regressor, increases fourfold. The credible intervals change more for greater probability of "coverage", because they depend more on the tail behavior.

assumption for θ	90%	95%	99%
$\theta = \hat{\theta}$	$(4.910, 4.970)$	$(4.906, 4.971)$	$(4.891, 4.980)$
$\theta \sim \bar{\pi}(\theta)$	$(4.883, 5.013)$	$(4.860, 5.022)$	$(4.810, 5.043)$

Table 2.1: *Credible Intervals for β with Short Range*

Credible intervals for β where the Gaussian random field for rainfall has range parameter value 3. The true value is $\beta = 5$.

assumption for θ	90%	95%	99%
$\theta = \hat{\theta}$	$(4.912, 5.073)$	$(4.898, 5.089)$	$(4.867, 5.113)$
$\theta \sim \bar{\pi}(\theta)$	$(4.855, 5.164)$	$(4.830, 5.188)$	$(4.768, 5.231)$

Table 2.2: *Credible Intervals for β with Long Range*

Credible intervals for β where the Gaussian random field for rainfall has range parameter value 5. The true value is $\beta = 5$.

We can see from Table 1 that the difference is quite important. For instance, the 95% credible interval is 2.5 times wider when accounting for the uncertainty in θ. From a decision perspective, when fixing $\theta = \hat{\theta}$, none of the three credible intervals in Table 1 include the true value $\beta = 5$, although all of the credible intervals for the procedure drawing θ from $\bar{\pi}(\theta)$ do include the true value.

If I increase the range covariance parameter from 3 to 5, thus reducing the effective sample size, all credible intervals widen, and the difference in credible intervals with and without accounting for variability in θ remains as important.

As expected, accounting for the estimation of the imputed regressors matters tangibly for inference, and presenting confidence or credible intervals which are in fact conditional on the estimated covariance parameter $\hat{\theta}$ can be misleading, especially if they are not clearly presented as such.

It should be highlighted that this estimation procedure is particularly easy. Relying on standard Bernstein-von Mises results, the posterior $\bar{\pi}(\theta)$ can be approximated by a normal with mean and covariance estimated by maximum or quasi-maximum likelihood. Then sampling $\theta^{(j)} \sim \bar{\pi}(\theta)$ is a single line of code. The second step of the Bayesian bootstrap is then done by

simply multiplying the outcome data and interpolated data by the square root of exponential random variables, which likewise only requires one additional line of code.

A frequentist interpretation can be obtained by relying on increasing domain asymptotics. With decaying dependence, in an arbitrarily large domain, and under ergodicity assumptions, a single realization of a random field with decaying dependence will allow for a consistent estimation of population parameters θ and β. The intuition is that one could take many, say, rectangular subsets from the random field, all distant enough from each other to be considered mutually independent, and use them as repeated realizations of the underlying random field.

2.3 Revisiting Madsen, Rupert, and Altman (2008)

I apply and compare the methods under study using the cross-validation exercise of Madsen et al. (2008). I use the same data set[2] as in their article. As explained therein, and further detailed in Herlihy et al. (1998), the data is a subset of the sample obtained for the Environmental Monitoring and Assessment Program of the Environmental Protection Agency. All samples are from rivers and streams in the American Mid-Atlantic region, and the analysis objective was to relate stream characteristics with land use variables. There are 558 observations over an area of 400,000 squared kilometers. The outcome variables $Y(x)$ are the logarithm of chloride concentration at locations x, and the dependent variables $R(x^*)$ are the logit transformations of the per cent of watershed in forest at locations x^*.

The reference value is obtained by doing generalized least-squares on the full, aligned data set weighting with an estimate of the regression error covariance: I obtain $\hat{\beta}_{\text{full}} = -0.38$. The simulation is run as follows; for each run a randomly chosen half of the independent variables, the R_i's, are "hidden", and the outcome variable, the Y_i's, are "hidden" for the other half of the data set, thus creating a misaligned data set. For each round of the cross-validation exercise, β is estimated with each method, and estimates are recorded.

There are two ways to carry out the second stage regression. The first is to simply do

[2]I would like to thank the authors for kindly providing their data for replication.

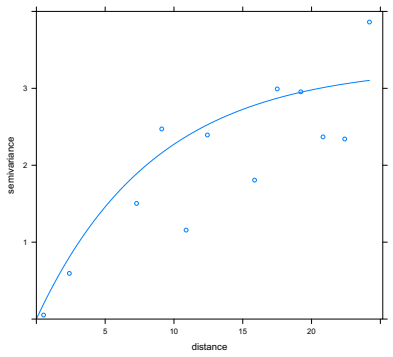

Figure 2.3: *Empirical and Fitted Variogram of R*

ordinary least-square (OLS), and the second is to use feasible generalized least-squares (GLS). For the latter approach, one obtains the weighting matrix by first fitting OLS to obtain residuals, and then fitting a covariance function to the residuals (in this case, an exponential covariance function), thus yielding an estimate of their covariance to use as weighting matrix. With respect to the data set of Madsen et al. (2008), adjusting the inner products with an estimate of the covariance of the regression errors turns out to be crucial for obtaining reliable estimates. This is a very important point and a key motivation for adopting the survey sampling target; under resampling of Y and R, even two-step estimation of the best linear predictor $E^*[Y(X)|R(X)]$ may require specification of Σ, the covariance of the regression errors.

We see from the empirical variogram displayed in Figure 8 that neighboring dependent variables do covary, thus allowing for useful interpolation as a first step. The quality of the fit of the variogram function is however suspect, thus warranting the use of robust standard errors.

Inspection of Table 3 reveals that the OLS estimates are inaccurate, and the regression output yields standard errors which are much too conservative in comparison with the variability observed in simulation. The GLS estimate is on average much more accurate in this small sample, pointing at the importance of correcting for the non-spherical variance of the regression errors in the inner product. However, it appears that reweighting with an estimated covariance

matrix introduces much variability, as the variance of the GLS estimate in simulation is one and a half times that of the OLS estimate. The standard errors from the regression output for GLS are more accurate but still much too conservative.

As seen in Table 3, the maximum likelihood approach yields accurate estimates. The standard errors obtained from the inverse Fisher information matrix are, as observed in Madsen et al. (2008), much smaller than those obtained in simulation. However, as predicted by theory, using instead the sandwich covariance matrix yields more accurate standard errors. In this instance, they are very good. I computed the variance of the score as detailed above, with the "quasi-Gaussian" approach relying on Wick's formula. The Hessian was computed in closed form (using output from the nonparametric and parametric variogram estimation), but it was observed that "off-the-shelf" numerical estimates performed reasonably well.

Intuitively, the success of both the Krig-and-Regress and ML approaches hinge upon capturing a sufficient part of the explanatory variation in $R(x)$ through $R(x^*)$. As detailed in Subsection 2.1, the one-step approach uses more information so to better accomplish this task. I can quantify the extent to which maximum likelihood accomplishes this by comparing the variance of the interpolated dependent variable vector, $\hat{R}(x)$, in the first stage, as well as the R-squared of the second stage regression, with their implicit analogues from the maximum likelihood estimation. In the two-step implementation, the sample variance of the interpolated vector is $\left\|\hat{R}(x) - \overline{\hat{R}(x)}\right\|_2^2 = 0.8$, compared with 4.3 for the original data. The variance of the implicitly interpolated quantity in the ML implementation is $\left(\left\|Y(x) - \overline{Y(x)}\right\|_2^2 - \hat{\sigma}_\varepsilon^2\right)/\hat{\beta}_1^2 = 4.8$, which suggests a much greater use of information,[3] as well as some overfitting. Analogously, the R-squared for the regression step of Krig-and-Regress was 0.09, and the implicit R-squared for the ML implementation was 0.69, which is an order of magnitude greater.

Inspection of Figure 4 makes tangible the identification strength of the robust approach. The pattern in the bottom-left plot of $\sigma^*_{YR^*}$ against distance shows the (decaying) negative

[3]Note that this is different from $\left\|\hat{R}_{\hat{\theta}}(x) - \overline{\hat{R}_{\hat{\theta}}(x)}\right\|_2^2$, where $\hat{R}_{\hat{\theta}}(x)$ is the best linear predictor for $R(x)$ using the maximum likelihood estimate for the covariance parameters, which does not account for the "variation in $\hat{R}$" explained by variation in Y.

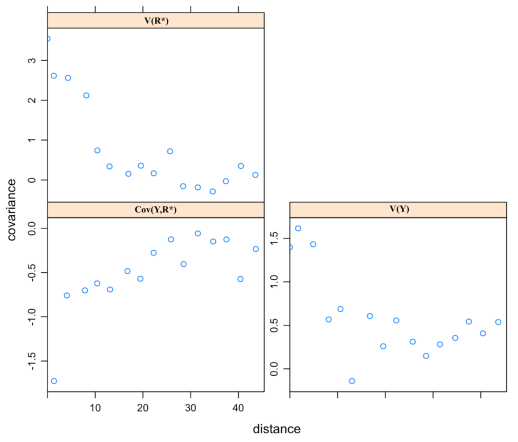

Figure 2.4: *Nonparametric Estimates of Variance and Covariance for Y and R^* as a Function of Distance*

correlation between (neighboring) $Y(x)$ and $R^*(x')$. It suggests that we are capturing sufficient variation to confirm an impact of R^* on Y through R and thus, heuristically speaking, the existence of a "significant" β. The scale of β can be assessed by comparing the magnitudes of $\sigma^*_{YR^*}$ and $\sigma^*_{R^*}$.

If $\hat{\beta}$ is obtained from the ratio of the covariance of R^* and Y with the variance of R^*, how is this approach conceptually different from the ordinary least-squares approach? The key difference is that it is using the information that the covariance function of R^* with Y and the variance of R^* are the same up to rescaling by β, and it estimates β and θ jointly.

2.4 Reanalysis of Maccini and Yang (2009)

This section reanalyzes the data set investigated in Maccini and Yang (2009). In that paper, the authors estimate the effect of a rainfall shock in infancy on adult socio-economic outcomes such as education or health. The paper merges a rainfall measurements data set with survey data, both of which have geographically located observations. The data sets are misaligned, as can be seen from Figure 1.1, which plots the respective locations.

The authors do point out a host of potential issues arising from misalignment and convincingly tackle them using linear regression methods. They use a two-stage least-squares

	Mean	Sim. Strd Errors	Reg. Output Strd Error	Coverage
OLS	-0.44	0.045	0.080	0.98
GLS	-0.33	0.071	0.099	0.97
MLE	-0.39	0.049	0.032	0.75
QMLE	-0.39	0.049	0.050	0.95
2-step BB	-0.40	0.069	0.061	0.93
Robust	-0.43	0.095	0.117	0.94

Table 2.3: *Cross-Validation Output*

The mean is the average over all simulation draws. The simulation standard errors are the standard deviations of the simulation draws. The output standard errors are obtained from the inverse Hessian evaluated at the estimated value of the parameter for MLE, and from the sandwich formula using Wick's formula to compute fourth moments for QMLE. For the two-step Bayesian bootstrap, the mean is the average posterior mean, the regression output standard errors are given by the average over simulation runs of the posterior standard deviation. The coverage corresponds to the fraction of times the confidence interval, computed with the output standard errors of the current run of the simulation, covered -0.38.

approach which, in principle, is easy to compute and produces standard errors that account for uncertainty in the first stage. I compare their approach with the more principled methods developed herein. I find that their general conclusions hold up, but that the analysis nevertheless benefits from the use of my methods as they yield some statistically and economically significant changes in the value of parameter estimates.

2.4.1 Data Description

The data set for the regression is obtained by merging two, misaligned data sets. The first data set contains rainfall measurements from measuring stations across Indonesia. The whole rainfall data spans the years from 1864 to 2004. The yearly count of active rainfall measurement stations is presented in Figure 5.

Only the years 1953-1975 are used to identify yearly variation; the other years are used to estimate long term averages. As can been seen from Figure 5, in almost every year I am using, more than 300 hundred rainfall measurement stations are active. The rainfall data comes from the Global Historical Climatology Network (GHCN), which is a publicly available data set.[4]

[4]Available online at http://www.ncdc.noaa.gov/oa/climate/research/ghcn/ghcn.html.

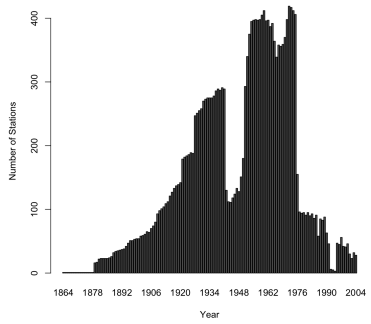

Figure 2.5: *Yearly Count of Active Rainfall Measurement Stations*

The second data set is the third wave of the Indonesian Family Life Survey (IFLS3), which includes each surveyed individual's year and location of birth. It consists of 4,615 women and 4,277 men born outside large cities between 1953 and 1974. Tables A.1 and A.2 provide extensive summary statistics.

The locations of the birthplaces and rainfall stations are given in Figure 1.1. We can see that the data sets are misaligned. We can also see from the figure that most birthplaces are situated fairly close to one or more rainfall stations; the median distance to the closest rainfall station is 11.43 km, and the third quartile distance is 90.28 km. The median distance to the fifth closest station is 30.24 km, and the third quartile distance to the fifth closest station is 317.10 km. We will find in Subsection 4.2 that, at those distances, the rainfall measurements are still correlated, thus allowing for useful interpolation.

The rainfall data on a given month has a point mass at zero, and smooth distribution over strictly positive values; a typical plot is given in Figure A.2. However, I do not use the raw rainfall data, but a transformation of it. To obtain results that are comparable to those of Maccini and Yang (2009) I use, as they did, the log of the ratio of the yearly rainfall with the long run average yearly rainfall. The authors of the original analysis use this transformation because the subject of the study is the impact of rainfall shocks, which are captured by departures from the long run average. We find in Figure 6 that this has the added benefit of eliminating the point mass at zero, and making the distribution "closer to Gaussian". In particular, although there are entirely dry months, there is always a month with nonzero

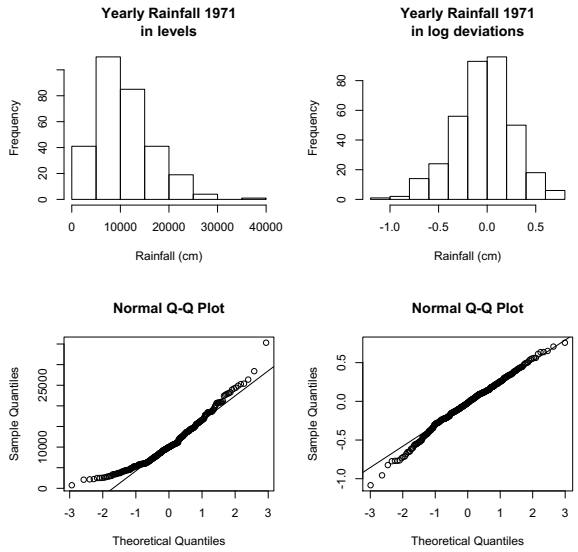

Figure 2.6: *1971 Rainfall*

Histograms and Gaussian QQ-plots of 1971 rainfall in levels (left) and in logarithm of ratio of yearly sum to long run yearly average (right).

rainfall over any given, entire year.

Inspection of the histograms and QQ-plots in Figure 6 as well as the variograms in Figure 7 shows features of the data that speak to robustness concerns. It is clear from Figure 6 that transforming the data helped make more plausible the Gaussian assumption. However, there remains some skewness in the histogram, also detectable in the QQ-plot, leaving the assumption of exact Gaussianity somewhat suspect. The yearly histograms and QQ-plot of Figure 6 are typical of this data set. The variogram fit of Figure 7 suggests a good fit of the variogram and corroborates the assumption of a Gaussian covariance function. However, Stein (1999) warns against relying on such plots to draw definitive conclusions. Furthermore, some years (not shown) have worse fit. Altogether, this calls for inference with methods that are robust to covariance model misspecification.

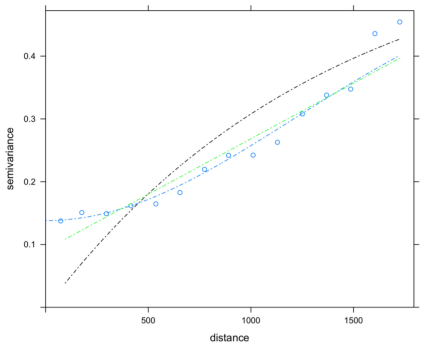

Figure 2.7: *Semivariogram*

Plots of empirical and fitted semivariogram using the exponential (black), Gaussian (blue) and linear (green) models. The data is from the year 1971.

2.4.2 Regression Analysis

The first order of business is to obtain a two-step estimator. I use the Krig-and-Regress method, which is a consistent two-step method. It would be a natural first inclination to want to perform Kriging on the monthly data, and then aggregate the monthly predictions over the year to obtain the transformed data points. Indeed, we can see from Figure A.2 that on a typical year, the covariance functions are quite different for each month. I find however that Kriging on the log ratio of yearly rainfall to long run yearly average is preferable. First, it seems to follow a well-behaved covariance function for which I have twelve times more data; second, the transformed data seems closer to Gaussian, and has no point mass at zero, as can be seen from Figure 6; third, it makes more immediate the comparison with the regression output of Maccini and Yang (2009) and with the QML output. Most importantly, if the imputed covariate is a nonlinear transformation of the interpolated quantity, the estimated coefficients may not be consistent estimators of the best linear predictor coefficients.

A potential concern is the isotropy assumption, i.e., that the direction of the vector giving the difference between the locations of two points in the random field does not matter. That is, for the variogram to only be a function of the distance between its arguments, and not the direction of the vector from one to the other, we must assume that the direction does not impact the covariance. One way to assess this is to plot and inspect the directional variograms,

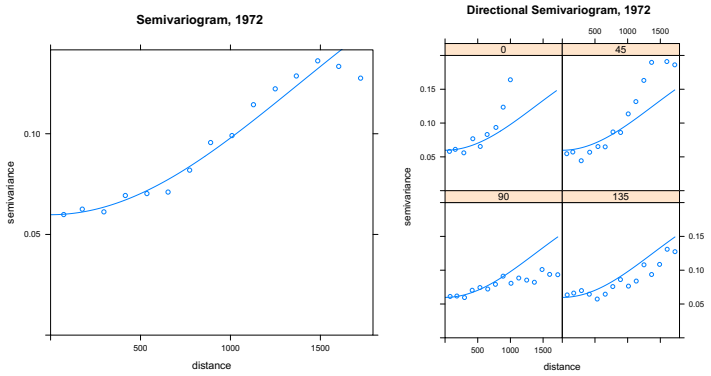

Figure 2.8: *Variogram and Directional Variograms (data from 1972)*

shown in Figure 8.

The directional variograms are reassuring. The empirical directional variograms seem to align well with the fitted isotropic variogram up to at least 500 km. Beyond that distance very little data goes into estimating each dot figuring in the plots; hence they are quite variable. The appearance of a trend away from the fitted isotropic variogram can be due to the fact that these points are highly correlated.

I am interested in fitting the model

$$Y_i = \delta_{\text{boy}} 1\{i \in \mathcal{B}\} + \delta_{\text{girl}} 1\{i \in \mathcal{G}\} + R_{\text{true},i} \left(\beta_{\text{boy}} 1\{i \in \mathcal{B}\} + \beta_{\text{girl}} 1\{i \in \mathcal{G}\}\right) + F_i^T \gamma + \epsilon_i, \quad (2.1)$$

where $\mathcal{B}$ and $\mathcal{G}$ are the set of observation indices corresponding to subjects who are boys and girls, respectively. Now $R_{\text{true},i}$ is sampled from the random field of the log of the ratio of yearly rainfall to long term average at the outcome locations and F includes location (district) dummies, season dummies, time trend, and interactions. $R_{\text{true},i}$'s are not observed but $\mathbf{R}^*$, a vector of observations from the same random field albeit at different locations, is observed.

It is important to specify in which ways my two-step analysis follows that of Maccini and Yang (2009), and in which ways it differs from theirs. The objective of the analysis is the same, namely to estimate the effect of a shock in rainfall during infancy on adult socioeconomic outcomes. Furthermore, as detailed above, I used their choice of transformation for the rainfall data.

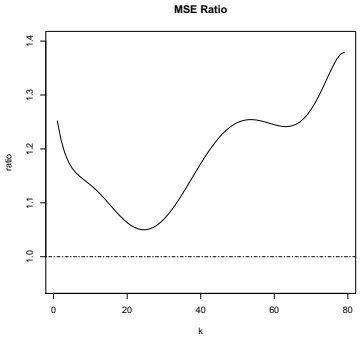

Figure 2.9: *Ratio of Mean Squared Errors*

Ratio of the mean squared errors of Kriging and k-means, for different k, in leave-one-out cross-validation.

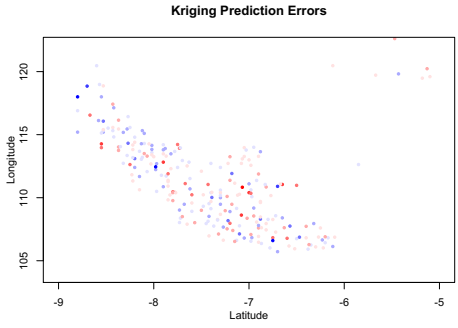

Figure 2.10: *Residuals*

Geographic plot of prediction errors from leave-one-out cross-validation. Negative residuals are in blue, positive residuals in red. Darker shades correspond to greater absolute values.

I use a different two-step method than Maccini and Yang (2009) did. Whilst I do the interpolation step with the best linear predictor, they use an instrumental variables approach. Their strategy is to estimate (1.22) by running a two-stage least-squares regression in which the nearest rainfall is instrumented with slightly more distant rainfall stations (the second to fifth closest stations). This approach is problematic for two reasons. Conceptually, there is no reason to believe that instrumenting the wrong rainfall measurements with other wrong rainfall measurements will "wash away" the error term which can be thought of as the difference between the rainfall at the outcome location and the rainfall at the nearest station (for instance, all stations could be near each other and have correlated "errors"). Practically, the data set at hand makes this strategy difficult to implement; the missing rainfall observations are many and sporadic such that, for any given outcome location, there will be few years for which the nearest rainfall station and all of the next four nearest stations will have observed measurements. Eliminating the incomplete observations reduces the data set substantially and imputing, say, medians impacts the outcome tangibly.

A natural alternative, for purposes of interpolation, is k-means. I compare the performance of k-means with with Kriging to justify the use of a more sophisticated method. Figure 9 displays the ratio of mean squared errors in a leave-one-out cross-validation exercise using the rainfall data. We see that there are benefits justifying the use of Kriging for interpolation; for all reasonable choices of number of neighbors k, Kriging dominates k-means in terms of mean squared error.

To further assess the quality of the Kriging estimates, I plot the geographically located interpolation errors from leave-one-out cross-validation in Figure 10. It appears there is not much spatial correlation left in the residuals.

The second stage for the Krig-and-Regress is done by ordinary least-squares. I found that using feasible generalized least-squares did not help in this case. In particular, fitting a covariance function to the regression errors is difficult, and the fit needs to be quite good to tangibly ameliorate the point estimates.

Implementation of the Krig-and-Regress approach yields results that are similar to those

of Maccini and Yang (2009).

To apply the QML approach to model (1), I need to accommodate an interaction term on the regression coefficient for rainfall. Observing, as in Subsection 1.1.1, that the mean is overparametrized, I can write the likelihood as

$$\begin{pmatrix} Y \\ R^* \end{pmatrix} \sim N_{N+M}\left(\begin{pmatrix} X\zeta \\ \mathbf{m}^* \end{pmatrix}, \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix}\right), \tag{2.2}$$

where X includes F as well as gender effects for boys and girls. The blocks of the covariance matrix are

$$(\Omega_{11})_{ij} = (\beta_{\text{boy}} 1\{i \in \mathcal{B}\} + \beta_{\text{girl}} 1\{i \in \mathcal{G}\})(\beta_{\text{boy}} 1\{j \in \mathcal{B}\} + \beta_{\text{girl}} 1\{j \in \mathcal{G}\}) K_{ij} + \Sigma_{ij},$$

$$(\Omega_{12})_{ij} = (\beta_{\text{boy}} 1\{i \in \mathcal{B}\} + \beta_{\text{girl}} 1\{i \in \mathcal{G}\}) \bar{\mathbf{K}}_{ij},$$

$$\Omega_{21} = \Omega_{12}^T \text{ and } \Omega_{22} = K^*,$$

where $\mathbf{1}_{\mathcal{B}} = (1\{i \in \mathcal{B}\})_i$ is the vector of indicator for boys, and $\mathbf{1}_{\mathcal{G}} = (1\{i \in \mathcal{G}\})_i$ is the vector of indicator for girls.

The coefficient of interest only figures in the covariance, and I obtain estimates using REML, which is tantamount to maximizing the likelihood for

$$\begin{pmatrix} QY \\ R^* \end{pmatrix} \sim N_{N-p+M}\left(\begin{pmatrix} 0 \\ \mathbf{m}^* \end{pmatrix}, \begin{pmatrix} Q\Omega_{11}Q^T & Q\Omega_{12} \\ \Omega_{21}Q^T & \Omega_{22} \end{pmatrix}\right), \tag{2.3}$$

where Q is a full rank $(n - \text{rank}(X)) \times n$ matrix satisfying $QX = 0$.

When maximizing the log-likelihood, I use the output of the Krig-and-Regress procedure as starting values for the optimization algorithm.

As did Maccini and Yang (2009), I find that the effects are more pronounced for girls than for boys. Table 4 gives the estimated coefficients for five outcome variables considered in the original analysis. I find, as they did, a significant effect of rainfall shocks on the indicator of self-reported very good health. However, although they found rainfall shocks in infancy to have a significant impact on adult height as well as on the indicator of self-reported poor

health, these findings are not maintained when using robust uncertainty assessments.

Point estimates appear to behave as suggested by the theory and simulations. As observed in the cross-validation exercise of Madsen et al. (2008), the standard errors of the (assumed to be well-specified) maximum likelihood output are much too narrow and may mislead the researcher into concluding statistical significance when unwarranted. In all instances the standard errors increased when using a covariance formula which is robust to covariance function misspecification.

As expected from the well-behaved histograms and QQ-plots of Figure 6, nonparametric estimation of the fourth-order moments did not alter the standard errors substantially, although they did increase in some instances.

The comparison of the output for the Krig-and-Regress method and the Bayesian bootstrap method is informative. In particular, the standard errors provided for the two-step Bayesian bootstrap do take into account the uncertainty in the imputation of the missing covariate, while those provided in the output of the Krig-and-Regress method do not. As ought to be expected, the standard errors for the two-step Bayesian bootstrap are larger than for Krig-and-Regress.

The maximum likelihood point estimate is about 0.1 higher than that of the two-step method. This means that the estimated impact of one standard deviation in yearly rainfall, away from the long term average, on the probability of declaring oneself very healthy increases from 3.5% to 6.3%, which is a difference of considerable economic magnitude.

All of Krig-and-Regress, robust estimator, and two-step Bayesian bootstrap provide point estimates which can be compared to that of maximum likelihood to help assess its robustness to the specification of a covariance function for the regression errors. One must exercise caution when carrying out such a robustness check; the change in the value of the regression coefficient estimate can be brought about both by the gain in information from moving to a more efficient method, and by misspecification of the regression error covariance. The fact that the one-step estimate is within the confidence interval of the robust estimate (for reasonable critical values) corroborates the former explanation.

Discussion of the specific sources of misspecification is warranted. Inspection of Table 4

		M&Y	2-step (K-R)	2-step (BB)	ML	QML (Wick)	QML	Robust
(a)	β_{girl}	0.101 (0.058)*	0.104 (0.085)	0.107 (0.094)	0.195 (0.034)***	0.195 (0.061)**	0.195 (0.063)**	0.116 (0.118)
	β_{boy}	−0.029 (0.072)	−0.079 (0.084)	−0.079 (0.094)	0.001 (0.037)	0.001 (0.062)	0.001 (0.067)	−0.149 (0.126)
(b)	β_{girl}	−0.192 (0.082)**	−0.074 (0.089)	−0.070 (0.101)	−0.088 (0.062)	−0.088 (0.103)	−0.088 (0.107)	−0.131 (0.195)
	β_{boy}	−0.100 (0.098)	0.013 (0.083)	0.015 (0.108)	0.001 (0.063)	0.001 (0.105)	0.001 (0.110)	−0.010 (0.200)
(c)	β_{girl}	2.832 (0.821)***	1.441 (1.402)	1.333 (1.637)	1.247 (0.484)***	1.247 (0.903)	1.247 (0.942)	1.720 (1.783)
	β_{boy}	0.998 (1.795)	1.210 (1.685)	1.010 (2.010)	1.160 (0.534)*	1.160 (0.935)	1.160 (0.971)	1.517 (1.930)
(d)	β_{girl}	−1.175 (0.831)	−1.169 (0.842)	−1.232 (0.972)	−0.868 (0.401)*	−0.868 (0.632)	−0.868 (0.657)	−1.225 (1.150)
	β_{boy}	0.515 (0.779)	−0.761 (1.020)	−0.735 (1.231)	−0.011 (0.443)	−0.011 (0.675)	−0.011 (0.683)	0.143 (1.347)
(e)	β_{girl}	1.086 (0.453)**	0.392 (1.035)	0.405 (1.161)	0.414 (0.472)	0.414 (0.599)	0.414 (0.625)	0.214 (1.174)
	β_{boy}	−0.474 (1.490)	−0.125 (1.044)	−0.140 (1.184)	0.007 (0.478)	0.007 (0.615)	0.007 (0.639)	−0.005 (1.201)

Table 2.4: *Effect of Birth Year Rainfall*

Effect of birth year rainfall on (a) indicator for very good self-reported health status; (b) indicator for poor or very poor self-reported health status; (c) adult height; (d) days absent due to illness (on four weeks preceding survey); (e) completed grades of schooling. The Maccini and Yang (M&Y) estimates are taken from the original paper; Krig-and-Regress (K-R) uses OLS in the second stage and the standard errors of the OLS output; the procedure for the two-step Bayesian bootstrap (BB) is as detailed in Section 4, with the difference that additional covariates are added in the second stage. The point estimate is the posterior mean. Maximum likelihood (ML) uses the inverse Hessian (the Hessian is computed in closed form) to obtain the asymptotic variance. Quasi-maximum likelihood with Wick's formula (QML (Wick)) uses the sandwich formula for the asymptotic variance but estimates the fourth-order moments using the Gaussian formula, and Quasi-maximum likelihood (QML) uses nonparametric estimates for fourth-order moments.

shows that standard errors changed more when allowing for covariance function misspecification than for distribution misspecification, leading one to suspect that the former misspecification issue is the most important one. In a sense, it is the less preferred one as I have shown that with misspecified distributional assumptions but well specified first and second order moments, QMLE yields consistent estimates. In fact, the large discrepancy between the standard errors obtained from the inverse Hessian (ML) and those obtained from the sandwich formula using the Gaussian approximation to fourth-order moments (QML (Wick)) suggests that the covariance function misspecification might be important. However, the standard errors displayed the same behavior in Table 3, in the cross-validation exercise, but the point estimate was very accurate. This is perhaps not so surprising. In the spatial statistics literature it has been shown that, under infill asymptotics, quite weak conditions suffice to guarantee that interpolation with the wrong covariance function will yield consistent and even efficient estimates of the interpolated variable (Stein, 1988; Yakowitz and Szidarovsky, 1985); Stein, 1999). Inquiring into how this translates into good properties of estimates of β is an upcoming item on my research agenda.

2.5 Discussion and Conclusion

The main purpose of this article was to modernize econometrics for spatial data, with a special focus on the case of misaligned data. As detailed in Chapter I, one-step maximum likelihood methods provide more efficient estimators than two-step methods. However, the theory and computational methodology were lacking for robust inference using the maximum likelihood estimator. I have provided asymptotic distribution theory for this estimator when the assumptions of Gaussianity and well-specified covariance functions need to be relaxed. I have provided two empirical examples using a quasi-maximum likelihood estimator for more efficient inference.

I also developed methods for the case in which the researcher does not want to specify the covariance structure of the regression errors, thus leaving the likelihood function for the full model unavailable. First, I proposed a one-step minimum distance estimator and developed

its limit distribution theory, thus producing correct standard errors for the estimator. Second, I suggested a modified two-step method which, by sampling the covariance coefficient from its posterior distribution, produces credible intervals that account for the variation due to the estimation of the imputed regressor. These are easy to compute and have a frequentist interpretation. This contribution is important because, although Krig-and-Regress is the commonly preferred method for two-step estimation in the misaligned data regression problem, to the best of my knowledge the literature did not offer standard errors which account for estimation in the first stage without requiring specification of the regression error covariance structure.

I would like to leave the reader with a clear sense of which method to use when. For regression with misaligned data, if the researcher can posit a covariance structure for the regression errors, he should use the maximum likelihood estimator. If the researcher is confident that both the covariance functions and the Gaussianity are well specified, he should use the inverse Hessian as the formula for the asymptotic variance. If the researcher believes only the covariance function specification is suspect, then he should use the sandwich covariance formula and estimate the fourth-order moments with Wick's formula. If Gaussianity is also suspect, then the researcher should estimate the fourth-order moments using the shrinkage estimator of Subsection 1.2.2. A good robustness check is to compare the point estimates to those of the minimum distance estimator and Krig-and-Regress. If the researcher concludes that specification of the regression error covariance is too restrictive, he should use either the minimum distance estimator or the Bayesian bootstrap two-step estimator. For both of these, I provided standard errors which account for estimation of the covariance structure of the misaligned variable. In economics, the effects estimated are often weak (e.g. impact of rainfall shocks during the first year of life on adult health) and the more efficient QML method ought to be preconized.

Estimation of fourth-order moments is an important topic for future research. I have suggested a method, inspired from the statistical finance literature, to directly estimate the fourth-order moments. More sophisticated (and difficult to implement) estimation methods

have been studied in the statistical geology literature (Goodfellow et al., 2012; Dimitrakopoulos et al., 2010). I believe, however, that further research ought to focus on developing robust estimators which do not require the estimation of fourth-order moments. Estimation in the spectral domain offers a promising line of inquiry. In recent work, Rao (2015) produces limit distribution theory for the spatial periodogram under distribution misspecification. The variance of the estimator then depends on fourth-order terms (through the tri-spectra). A reasonable conjecture is that by smoothing the spatial periodogram, one could decrease the fourth-order variance term enough to be comfortable ignoring it altogether, while introducing bias sufficiently small that the point estimates would remain reliable.

Another fascinating line of research is the study of properties of the regression coefficient when the covariance function of the misaligned covariate is misspecified. Under infill asymptotics and assuming only weak conditions, the interpolated variable is consistent and even efficient. This certainly guarantees consistency of the regression coefficient estimator, and perhaps efficiency. It will be important to assess the properties of the regression coefficient estimates under increasing domain asymptotics when the covariance function is misspecified, as this asymptotic framework seems to better describe the situations that economists typically face when dealing with misaligned data.

In conclusion, I have provided efficient and robust estimators for regression analysis with misaligned data, which is an important problem in spatial econometrics. I have developed general limit distribution theory where it was lacking, and I carried out a detailed reanalysis of a misaligned data set, where it appeared that the methods I recommend were indeed required. I believe this article will serve as a useful reference for applied economists carrying out regression analysis with misaligned data, and a good starting point for more theoretically inclined econometricians who need to develop further results for this problem.

Chapter 3

Imputation with k-Nearest Neighbors

3.1 Introduction

Large scale surveys such as the Current Population Survey (CPS) or the US Census are crucial sources of information for the study and conduct of public policy (Card, 1990; DiNardo et al., 1995). However, the quantity of missing or misrecorded data points has been increasing to such an extent that these must be attended to carefully in statistical studies using such survey data (Meyer et al., 2016). Furthermore, as is for instance the case with CPS supplementary surveys, some variables are only available in non-overlapping partial surveys, and in that sense are each missing from the sample in which the other is observed, thus making regression analysis nontrivial.

Survey data that is incomplete and missing information may bring about serious problems for the estimation of parameters of interest. Depending on what one is willing to assume about the mechanism for missing data, different methods and different levels of modeling may be required for the conduct of regression analysis. Mechanisms for missing data in which the absence of data informs variables of interest often need to be modeled explicitly in order to conduct correct inference (Little and Rubin, 2014). On the other hand, when for instance a covariate is missing at random or when we are only interested in analysis conditional on the covariate being missing, the modeling requirements are much less. In such an instance,

machine learning algorithms, some of which have low modeling requirements and are known to have good prediction accuracy (Friedman et al., 2001), are very well adapted to the task at hand −which is essentially a prediction problem.

The k-nearest neighbors (KNN) algorithm is a well-established and easily implemented supervised classification method (Ripley, 1994; Devroye et al., 1996). In this article, I investigate the use of KNN as an imputation method when some regressor is unobserved but additional variables, predictive of the missing regressor, are available as well as a KNN algorithm trained to predict the missing regressor from these additional variables.

One of the lessons of Chapter I was that the imputation of missing regressors can be tangibly improved by allowing the outcome variable to inform the imputation. This insight carries over in the survey imputation application of this chapter.

In order to allow the outcome variable to inform imputation, the first order of business is to obtain a probabilistic interpretation of the imputation method, in this case the k-nearest neighbors classifier. Holmes and Adams (2002, 2003) develop such an interpretation and give explicit formulas for the conditional distribution of missing category labels. However, Cucala et al. (2012) point out that there does not exist a joint distribution with full conditionals equal to those specified by Holmes and Adams (2002, 2003), and suggest instead a symmetrized version of the algorithm for which well-defined joint and conditional distributions immediately obtain. Cucala et al. (2012) present a formal Bayesian treatment of k-nearest neighbors, and investigate Bayesian computational strategies. In particular, sophisticated strategies are employed to deal with normalization constants.

The main objective of this chapter is to extend their work to the case in which the predicted categorical variable enters a regression function as a dependent variable. I lay out a full Bayesian model for the coefficients of interest as well as for the unobserved regressors, and provide a straightforward Gibbs algorithm for sampling the posterior distributions. I suggest alternative modeling strategies for speeding up the computation time. Furthermore, analogously to Chapters II, I propose a consistent 2-step estimator as a more robust and easily implemented alternative.

To offer comparison with the popular method of support vector machines (Vapnik, 1998; Smola & Schölkopf, 1998), I also consider using that method to impute the missing variables before evaluating the regression function, and then correcting for the induced attenuation biased. I do not embed the support vector machines method in a probabilistic framework (and thus do not let the outcome variables inform imputation) because, as explained below, such an embedding is much more challenging than with the k-nearest neighbors method.

The general set-up and notation for the problem are as follows. We are interested in estimating the regression function

$$E^* \left[Y \left| T, \tilde{X} \right. \right] = \tau_T + \phi \tilde{X},$$

where E^* is the best linear predictor operator and T is a categorical variable taking values in the set C, of cardinality $|C|$. We are particularly interested in inference for τ.

The analyst faces a missing data problem; the independent categorical variable T is not observed in the sample in which the outcome variable Y is observed. However, a variable X_{reg} is observed along with Y, and is believed to contain information about T. The exclusion restriction

$$E^* \left[Y \left| T, \tilde{X}, X_{\text{reg}} \right. \right] = E^* \left[Y \left| T, \tilde{X} \right. \right],$$

gives a sufficient condition for the identification of the parameter of interest τ.[1]

Explicitly, the analyst has access to two datasets, $\mathcal{D}_{\text{reg}} = \left\{ Y_i, \tilde{X}_i, X_{\text{reg},i} : i = 1, ..., N_{\text{reg}} \right\}$

[1]It is worthwhile to consider the tradeoff between assumptions on the form of the regression function and the exclusion restriction assumption. The best linear predictor

$$E^* \left[Y \left| T, \tilde{X} \right. \right] = \tau_T + \phi \tilde{X}$$

provides an interpretation of the estimands τ_i, $i = 1, ..., |C|$. We estimate a linear approximation, rather than assuming linearity of the conditional expectation. However, the interpretation of the exclusion restriction $E^* \left[Y \left| T, \tilde{X}, X_{\text{reg}} \right. \right] = E^* \left[Y \left| T, \tilde{X} \right. \right]$ is less straightforward; in claiming that the exclusion restriction is satisfied, one must argue that, even though the conditional expectation may not be linear, controlling linearly for T and $\tilde{X}$ suffices for X_{reg} to have a zero coefficient in the best linear predictor (i.e., for the linear projection on X_{reg} to vanish). This is a tricky argument to make because one must make the case that even though the conditional expectation may not be linear in T and $\tilde{X}$, introducing X_{reg} in the best linear predictor will not diminish the approximation error. A perhaps more transparent way of asserting this identifying assumption is to make the stronger assumption of linearity of the conditional expectation function and to state the exclusion

and $\mathcal{D}_{\text{train}} = \{T_{\text{train},i}, X_{\text{train},i} : i = 1, ..., N_{\text{train}}\}$. That is, the variable $T_{\text{reg},i},\ i = 1, ..., N_{\text{reg}}$, is missing from $\mathcal{D}_{\text{reg}}$ and must, implicitly or explicitly, be imputed in order to carry out regression analysis. The analyst believes that a machine learning method such as the k-nearest neighbors algorithm trained on $\mathcal{D}_{\text{train}}$ would provide accurate estimates of $T_{\text{reg},i}$ based on $X_{\text{reg},i},\ i = 1, ..., N_{\text{reg}}$.

I will use the shorthands $\mathbf{y} = \{Y_i : i = 1, ..., N_{\text{reg}}\}$, $\mathbf{T}_{\text{reg}} = \{T_{\text{reg},i} : i = 1, ..., N_{\text{reg}}\}$, $\mathbf{T}_{\text{train}} = \{T_i : i = 1, ..., N_{\text{train}}\}$, and $\mathbf{X} = \left\{\tilde{X}_i, X_{\text{reg},i}, X_{\text{train},j} : i = 1, ..., N_{\text{reg}}, j = 1, ..., N_{\text{train}}\right\}$. I will omit $\mathbf{X}$ in probabilistic expressions when no confusion arises, but all inference is conditional on it.

The standard Bayesian approach treats the unobserved $T_{\text{reg},i},\ i = 1, ..., N_{\text{reg}}$, as random coefficients whose posterior distribution is backed out from their prior and likelihood. A computationally economical alternative modeling strategy posits that the argument T of the regression function is not some realized yet unobserved value $T_{\text{reg},i}$, but is a predicted value $T_{\text{knn},i}$. This is investigated in details in Section 3.

It may be that the computational burden (e.g., approximating the normalizing constant) or robustness concerns will motivate the analyst to use a two-step method. Hence, I provide a consistent two-step method which requires neither computation of the normalizing constant or that of an attenuation bias correction term.

The rest of the paper is organized as follows. Section 2 gives the Bayesian treatment of k-nearest neighbors classification. Section 3 gives the full Bayesian treatment of the imputation of missing regressors with k-nearest neighbors. Section 4 provides a robust and easy-to-implement two-step method as an alternative approach. Section 5 applies the developed methods to CPS data, where the effect of work hour flexibility on voter turnout is investigated. Sections 6 and 7 discuss and conclude, respectively. A two-step approach consisting in first imputing with support vector machines, and then carrying out the regression estimation, is described in the

---

restriction assumption directly in terms of the conditional expectation, i.e.,

$$E\left[Y \middle| T, \tilde{X}, X_{\text{reg}}\right] = \tau T + \phi \tilde{X}.$$

Although a stronger assumption, it offers an interpretable sufficient condition for the exclusion restriction on the best linear predictor, as well as a richer interpretation of the regression coefficients.

appendix.

3.2 Bayesian Treatment of k-Nearest Neighbor Classification

In order to specify a posterior distribution for the coefficients of interest as well as for the unobserved variables, we must first have a correct Bayesian treatment of the k-nearest neighbor classifier –which we will later be able to model jointly with the regression component, for which the usual semi-conjugate normal regression model will be suiting.

One of the contributions of Cucala et al. (2012) is to present a symmetrized version of the count of a given point's neighbors such that the corresponding conditional distributions can be obtained from a well-defined full distribution.

Cucala et al. (2012) define the full distribution as

$$f(\mathbf{T}\,|\beta,k) = \frac{1}{Z(\beta,k)}\exp\left(\beta\sum_{i=1}^{n}\sum_{l\sim_k i}\delta_{T_i}(T_l)/k\right) \tag{3.1}$$

$$= \frac{1}{Z(\beta,k)}\exp\left(\beta S(\mathbf{T})/k\right) \tag{3.2}$$

where $\beta > 0$, $\delta_A(B) = 1$ if $A = B$ and zero otherwise, $Z(\beta,k)$ is the normalizing constant of the distribution, $S(\mathbf{T}) = \sum_{i=1}^{n}\sum_{l\sim_k i}\delta_{T_i}(T_l)$ and $l \sim_k i$ means that the summation is taken over the X_l 's that are the k-nearest neighbors of X_i (with respect to Euclidian distance).

The corresponding full conditional distributions[2] are

$$f\left(T_i \,|\mathbf{T}_{-i}, \beta, k\right) \propto \exp\left(\beta/k \left(\sum_{l\sim_k i} \delta_{T_i}(T_l) + \sum_{i\sim_k l} \delta_{T_l}(T_i)\right)\right) \tag{3.3}$$

$$= \exp\left(\beta S_i(T_i, \mathbf{T}_{-i})/k\right) \tag{3.4}$$

where $i \sim_k l$ means that the summation is taken over the observations X_l for which X_i is a k-nearest neighbor, $\mathbf{T}_{-i} = (T_1, ..., T_{i-1}, T_{i+1}, ..., T_n)$, and $S_i(T_i, \mathbf{T}_{-i}) = \sum_{l\sim_k i} \delta_{T_i}(T_l) + \sum_{i\sim_k l} \delta_{T_l}(T_i)$.

I thus take (1) as the model for the category labels $\mathbf{T}$. For β and k, I follow Cucala et al. (2012) and use a uniform prior distribution

$$\pi\left(\beta, k\right) \propto 1_{\{1,...,K\}\times[0,\beta_{\max}]},$$

where K may be taken to be the number of observations of the label with the least number of observations, and $\beta_{\max}$ as the "threshold" β at which (and for all β greater than $\beta_{\max}$) all labels are fitted to be of the same category; see Moller (2003) for an explanation of this phase transition phenomenon.[3]

The predictive distribution of a new unclassified observation T_{n+1} is

$$P\left(T_{n+1} = c \,|\mathbf{T}, X_{n+1}, \beta, k\right) \propto \exp\left\{\beta/k \left(\sum_{l\sim_k (n+1)} \delta_c(T_l) + \sum_{(n+1)\sim_k l} \delta_{T_l}(c)\right)\right\}, \tag{3.5}$$

---

[2]These are derived as follows. Note that

$$f\left(T_i \,|\mathbf{T}_{-i}, \beta, k\right) = \frac{f\left(\mathbf{T} \,|\beta, k\right)}{f\left(\mathbf{T}_{-i} \,|\beta, k\right)} = \frac{f\left(\mathbf{T} \,|\beta, k\right)}{\sum_{T_i\in C} f\left(\mathbf{T} \,|\beta, k\right)}$$

and

$$\sum_{T_i\in C} f\left(\mathbf{T} \,|\beta, k\right) = \frac{1}{Z(\beta, k)} \exp\left(\beta \sum_{j\neq i} \sum_{l\sim_k j \wedge l\neq i} \delta_{T_j}(T_l)/k\right) \sum_{T_i\in C} \exp\left(\beta/k \left(\sum_{l\sim_k i} \delta_{T_i}(T_l) + \sum_{i\sim_k l} \delta_{T_l}(T_i)\right)\right).$$

Hence

$$f\left(T_i \,|\mathbf{T}_{-i}, \beta, k\right) \propto \exp\left(\beta/k \left(\sum_{l\sim_k i} \delta_{T_i}(T_l) + \sum_{i\sim_k l} \delta_{T_l}(T_i)\right)\right).$$

[3]There is no analytical formula available for $\beta_{\max}$, but it is straightforward to find it approximately by exploration.

where

$$\sum_{l \sim_k (n+1)} \delta_c(T_l) \text{ and } \sum_{(n+1) \sim_k l} \delta_{T_l}(c)$$

are the numbers of observations with label c among the k nearest neighbors of X_{n+1} and among the observations for which X_{n+1} is a k-nearest neighbor, respectively.

The sempiternal difficulties brought about by the normalizing constant impede Bayesian inference for model (1). The problem is made all the more thorny by the demonstrated poor performance of the pseudo-likelihood approach (Cucala et al., 2013), which would have reduced the number of operations necessary for the calculation of the normalizing constants to the order of $n \cdot |C|$.

Cucala et al. (2012) survey and compare different strategies, one of which extends to the case with more than two categories. That strategy consists in artificially canceling the normalization constant in the Metropolis-Hastings acceptance probability by introducing an auxiliary variable.

To be sure, the computational issue is that the normalizing constant,

$$Z(\beta, k) = \sum_{\mathbf{T}} \exp\left(\beta \sum_{i=1}^{n} \sum_{l \sim_k i} \delta_{T_i}(T_l)/k \right),$$

is a summation over $|C|^n$ elements. For even a moderate dataset of $n = 100$ observations and $|C| = 3$ possible categories, the summation is of length on the order of 10^{47}, which is computationally intractable. This makes direct simulation as well as standard Metropolis-Hastings sampling algorithms impracticable. Indeed, the acceptance probability of a Metropolis-Hastings sampler for (β, k) would take the form

$$\frac{Z(\beta, k)}{Z(\beta', k')} \frac{\exp\left(\beta' S(\mathbf{T})/k'\right) \pi(\beta', k')}{\exp\left(\beta S(\mathbf{T})/k\right) \pi(\beta, k)} \times \frac{q_1\left(\beta, k \,|\beta', k'\right)}{q_1\left(\beta', k' \,|\beta, k\right)} \tag{3.6}$$

where q_1 is an arbitrary proposal distribution, and computation of the ratio requires computation of the intractable normalization constants.

The data augmentation strategy which Cucala et al. (2012) preconize for dealing with the normalization constant altogether circumvents its computation; by using a Metropolis-Hastings

sampler but adding an auxiliary variable with proposal distribution equal to the density of $\mathbf{T}$, the normalization constants of both densities cancel each other out in the Metropolis-Hastings acceptance probability formula. Explicitly, introduce the auxiliary variable $\mathbf{z} = \{z_1, ..., z_n\}$ where $z_i \in C$ for $i = 1, ..., n$, and consider the joint posterior

$$\begin{aligned} \pi(\beta, k, \mathbf{z}|\mathbf{T}) &\propto \pi(\beta, k, \mathbf{z}, \mathbf{T}) \\ &= g\left(\mathbf{z}\,|\beta, k, \mathbf{T}\right) \cdot f\left(\mathbf{T}\,|\beta, k\right) \cdot \pi(\beta, k). \end{aligned}$$

With a proposal

$$q_2(\beta', k', \mathbf{z}'|\beta, k, \mathbf{z}) = q_1(\beta', k'|\beta, k, \mathbf{z}) f(\mathbf{z}'|\beta', k'),$$

the acceptance probability for a Metropolis-Hastings algorithm becomes

$$\begin{aligned} &\frac{P\left(\beta', k', \mathbf{z}'\,|, \mathbf{T}\right)}{P\left(\beta, k, \mathbf{z}\,|\mathbf{T}\right)} \times \frac{q_2\left(\beta, k, \mathbf{z}\,|\beta', k', \mathbf{z}'\right)}{q_2\left(\beta', k', \mathbf{z}'\,|\beta, k, \mathbf{z}\right)} \\ = &\frac{Z(\beta, k)}{Z(\beta', k')} \frac{\exp\left(\beta' S(\mathbf{T})/k'\right) \pi(\beta', k')}{\exp\left(\beta S(\mathbf{T})/k\right) \pi(\beta, k)} \frac{g\left(\mathbf{z}'\,|\beta', k', \mathbf{T}\right)}{g\left(\mathbf{z}\,|\beta, k, \mathbf{T}\right)} \times \frac{q_1\left(\beta, k\,|\beta', k', \mathbf{z}'\right) \cdot \exp\left(\beta S(\mathbf{z})/k\right)}{q_1\left(\beta', k'\,|\beta, k, \mathbf{z}\right) \cdot \exp\left(\beta' S(\mathbf{z})/k'\right)} \frac{Z(\beta', k')}{Z(\beta, k)} \\ = &\frac{\exp\left(\beta' S(\mathbf{T})/k'\right) \pi(\beta', k')}{\exp\left(\beta S(\mathbf{T})/k\right) \pi(\beta, k)} \frac{g\left(\mathbf{z}'\,|\beta', k', \mathbf{T}\right)}{g\left(\mathbf{z}\,|\beta, k, \mathbf{T}\right)} \times \frac{q_1\left(\beta, k\,|\beta', k', \mathbf{z}'\right) \cdot \exp\left(\beta S(\mathbf{z})/k\right)}{q_1\left(\beta', k'\,|\beta, k, \mathbf{z}\right) \cdot \exp\left(\beta' S(\mathbf{z})/k'\right)}, \end{aligned} \tag{3.7}$$

where the cancelation of the normalization constant is explicit.

An astute choice of g is key for good mixing. Moller et al. (2006) point out that, upon comparison of (6) and (7), one notices that $q_1\left(\beta, k\,|\beta', k', \mathbf{z}'\right) / g\left(\mathbf{z}\,|\beta, k, \mathbf{T}\right)$ takes the place of $Z(\beta, k)$, and $q_1\left(\beta', k'\,|\beta, k, \mathbf{z}\right) / g\left(\mathbf{z}'\,|\beta', k', \mathbf{T}\right)$ takes the place of $Z(\beta', k')$. This suggests that one may obtain good mixing if one draws $\mathbf{z}$ such that

$$E_g\left[\frac{q_1\left(\beta, k\,|\beta', k', \mathbf{z}'\right)}{g\left(\mathbf{z}\,|\beta, k, \mathbf{T}\right)}\right] \approx Z(\beta, k) \text{ and } E_g\left[\frac{q_1\left(\beta', k'\,|\beta, k, \mathbf{z}\right)}{g\left(\mathbf{z}'\,|\beta', k', \mathbf{T}\right)}\right] \approx Z(\beta', k'), \tag{3.8}$$

where E_g denotes integration with respect to the variable $\mathbf{z}$ following the law g. The above display makes immediate the idea that the auxiliary variable technique can be understood as an indirect way of estimating the normalization constants (albeit individually, as opposed to their ratio) via importance sampling, which is a standard approach for tackling such problems (Gelman & Meng, 1998). Since $Z(\beta, k) = \exp\left(\beta S(\mathbf{T})/k\right) / f\left(\mathbf{T}\,|\beta, k\right)$, picking

$g(\mathbf{z}\,|\beta,k,\mathbf{T}) = f(\mathbf{T}\,|\beta,k)$ would make approximations (8) hold exactly; in fact the argument of the integral would be degenerate and (7) would equal (6) exactly. However, this would of course not be practical because we would be stuck, once again, having to evaluate the normalization constant.[4] Nevertheless, the importance sampling analogy suggests that

$$g(\mathbf{z}\,|\beta,k,\mathbf{T}) \approx f(\mathbf{T}\,|\beta,k)$$

may provide good mixing. Cucala et al. (2012) propose

$$g(\mathbf{z}\,|\beta,k,\mathbf{T}) = g(\mathbf{z}|\,\mathbf{T}) = g\left(\mathbf{z}\,\middle|\hat{\beta},\hat{k},\mathbf{T}\right),$$

with estimates $(\hat{\beta},\hat{k})$ obtained by iteratively updating using previous MCMC output.[5] The Gibbs sampler for the data-augmented model is the following. Pick S and define $\tilde{\mathbf{z}}_{-i}^{(s)} = \left(\tilde{\mathbf{z}}_{1:(i-1)}^{(s)}, \tilde{\mathbf{z}}_{(i+1):n}^{(s-1)}\right)$. For $j = 1, ..., J$,

1. propose (β', k') from $q_1\left(\beta, k\,\middle|\beta^{(j)}, k^{(j)}\right)$
2. propose $\mathbf{z}'$ from $f(\mathbf{z}\,|\beta', k')$:
   (a) set $\tilde{\mathbf{z}}^{(0)} = \mathbf{z}^{(j)}$; for $l = 1, ..., L$, and for each $i = 1, ..., n$,
      i. compute $Z_i(\beta', k') = \sum_{z\in C} \exp\left(\beta S_i(z, \tilde{\mathbf{z}}_{-i}^{(l)})/k\right)$
      ii. sample $\tilde{z}_i^{(l)}$ from $\frac{1}{Z_i(\beta',k')} \exp\left(\beta S_i(z, \tilde{\mathbf{z}}_{-i}^{(l)})/k\right)$
   (b) set $z_i' = \tilde{z}_i^{(L)}$
3. draw $u \sim \text{Unif}[0,1]$
4. if

$$u \leq \frac{\exp(\beta' S(\mathbf{T})/k')}{\exp\left(\beta^{(j)} S(\mathbf{T})/k^{(j)}\right)} \frac{g(\mathbf{z}'\,|\beta', k', \mathbf{T})}{g\left(\mathbf{z}^{(j)}\,\middle|\beta^{(j)}, k^{(j)}, \mathbf{T}\right)}$$

[4] This furthermore highlights how this procedure is directly trading off computational expenditure for quality of mixing.

[5] Cucala et al. (2012) lay out a perfect sampling approach for this data-augmented model. However, the question of how to implement a perfect sampler with $|C| > 2$ is still open, and even the perfect sampler laid out for the case of $|C| = 2$ is very computationally expensive.

$$\times \frac{q_1\left(\beta^{(j)}, k^{(j)}, \mathbf{z}^{(j)} \,|\beta', k', \mathbf{z}'\right) \cdot \exp\left(\beta^{(j)} S(\mathbf{z}^{(j)})/k^{(j)}\right)}{q_1\left(\beta', k', \mathbf{z}' \,\middle|\beta^{(j)}, k^{(j)}, \mathbf{z}^{(j)}\right) \cdot \exp\left(\beta' S(\mathbf{z}')/k'\right)},$$

let $\left(\beta^{(j+1)}, k^{(j+1)}, z^{(j+1)}\right) = (\beta', k', z')$; otherwise $\left(\beta^{(j+1)}, k^{(j+1)}, z^{(j+1)}\right) = \left(\beta^{(j)}, k^{(j)}, z^{(j)}\right)$.

To be sure, Step 2 is a Gibbs sampler of the proposal density for $\mathbf{z}$; we sample $\tilde{z}_i^{(s)}$ conditional on $\tilde{\mathbf{z}}_{-i}^{(s)}$ for each $i = 1, ..., n$ (inner loop), this is repeated S times (outer loop) and the final draw of the full vector $\tilde{\mathbf{z}}^{(L)} = \left\{\tilde{z}_1^{(L)}, ..., \tilde{z}_n^{(L)}\right\}$ is the proposed value $\mathbf{z}'$, to be accepted or rejected according to the acceptance probability in step 4. In that sense, step 2 is a Gibbs step within a Metropolis-Hastings algorithm. The tuning parameter S should be taken to be large enough that the Gibbs sampler in step 2 reaches its stationary distribution. Since, for proposals (β', k') close enough to $\left(\beta^{(j)}, k^{(j)}\right)$, the Gibbs sampler will begin with a "warm start" $\mathbf{z}^{(1)}$, a small S may suffice.

I assess the comparative performance of the Bayesian approach in a simulation exercise. I generate artificial data with three different labels, i.e., $T_i \in \{1, 2, 3\}$, $i = 1, ..., n$, with corresponding independent variables drawn from three differently centered normals, i.e., $X_i \sim N(\mu_{T_i}, \sigma^2 I_2)$, $i = 1, ..., n$. The simulation parameters are $\sigma^2 = 0.5$, $\mu_1 = (0, 1)$, $\mu_2 = (\cos(5/3 \cdot \pi)$, and $\sin(5/3 \cdot \pi))$, $\mu_3 = (\cos(7/6 \cdot \pi), \sin(7/6 \cdot \pi))$. I draw 10 observations for each label, totaling $n = 30$ observations. The draws are plotted in Figure 1.

Interestingly, the Bayesian approach outperforms the standard KNN approach[6] (where the tuning parameter k was selected by cross-validation) with a probability of correct out-of-sample classification of 77% versus 72%. This comparison is informed by the observation that if one classifies using the symmetrized classification function but without treating it probabilistically,[7] then one obtains a probability of correct classification of 76%. This suggests that, at least

[6]The standard KNN approach attributes to X_i the label

$$\underset{T}{\text{argmax}} \sum_{l \sim_k i} \delta_T(T_l).$$

[7]The symmetrized KNN approach attributes to X_i the label

$$\underset{T}{\text{argmax}} \sum_{l \sim_k i} \delta_T(T_l) + \sum_{i \sim_k l} \delta_{T_l}(T).$$

in this example, an important part of the improvement from using the Bayesian approach is in fact due to the symmetrization of the count of the neighbors. Likewise, the heatmaps displayed in Figure 1 reveal that the Bayesian approach makes for a smoother and more convex delineation of the fitted categories, but also indicate that this may be attributable to the symmetrization of the neighbor count, rather than to the probabilistic approach.

The additional gains from the probabilistic approach may be obtained, for instance, via model averaging over k. Figure 2 gives MCMC output for k. We can see that the posterior distribution of k has a clear mode at $k = 2$. It is important to note that this may however be uncharacteristic: Cucala et al. (2012) report very jagged posteriors for k which, along with jagged histograms of the cross-validation estimates of the misclassification probability with respect to different values of k, suggest strong potential gains from model averaging over k. Such gains from model averaging are investigated below. Figure 3 displays posterior MCMC output for β.

3.3 Bayesian Treatment of k-Nearest Neighbor Imputation

The observed outcome variable Y informs the value of the unobserved dependent variable T through the regression function $E^*[Y|T, \tilde{X}]$. Imputation methods not availing themselves of that information may be inefficient. The Bayesian modeling of the KNN imputation method allows for a natural extension in which the unobserved regressors are modeled jointly with the regression coefficients. Such gains in efficiency rely crucially on $E^*[Y|T, \tilde{X}]$ being a good approximation of $E[Y|T, \tilde{X}]$; robustness to misspecification is discussed in Sections 4 and 6.

Consider for simplicity the regression model in which T is the only regressor (adding $\tilde{X}$ back in is a trivial extension). Recall that all the analysis is conditional on $\mathbf{X}$.

Specifically, the model is $f\left(\mathbf{T}, \mathbf{y} \,|\tau, \beta, k\right)$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} \left\|\mathbf{y} - \tau_{\mathbf{T}}\right\|^2\right) \cdot \frac{1}{Z(\beta, k)} \exp\left(\beta S(\mathbf{T})/k\right) \tag{3.9}$$

$$= \phi(\mathbf{y} - \tau_{\mathbf{T}}, \sigma^2) \cdot \frac{1}{Z(\beta, k)} \exp\left(\beta S(\mathbf{T})/k\right)$$

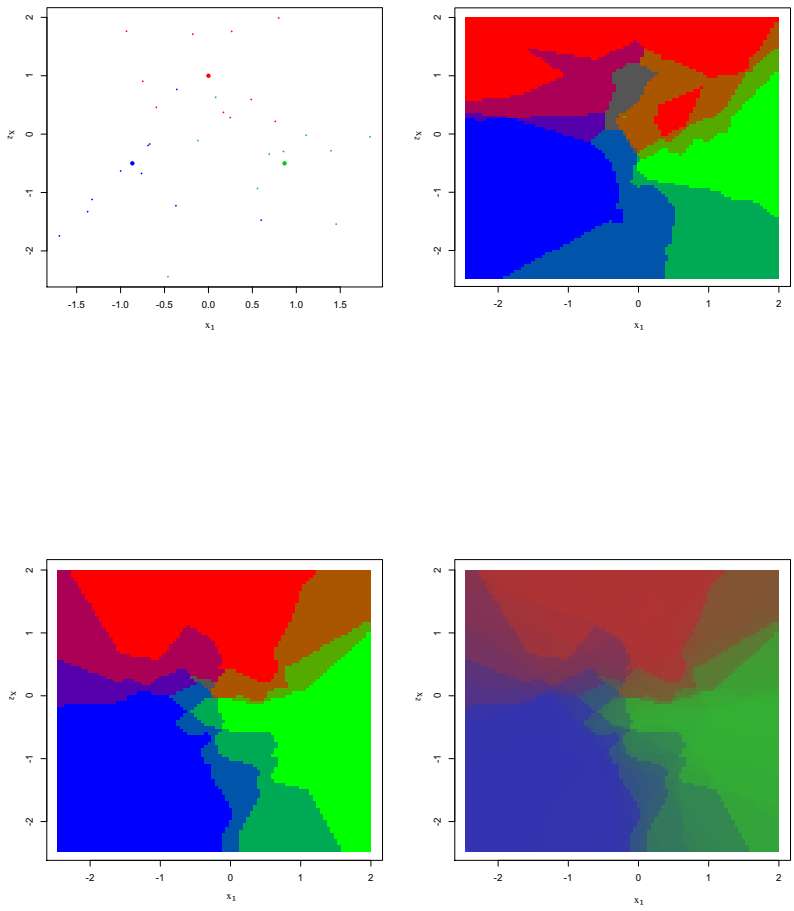

Figure 3.1: *Prediction of Categories*

(Top Left) The raw data. *(Top Right)* Fitted categories using standard (non-probabilistic) KNN. *(Bottom Left)* Fitted categories using symmetrized (non-probabilistic) KNN. *(Bottom Right)* Fitted categories using Bayesian KNN. For non-probabilistic methods, the count of labels amongst neighbors at each point on a fine grid is collected and normalized. For Bayesian KNN, the posterior predictive probabilities for each category at each point on the grid is collected. The color at any given point is obtained using the red-green-blue palette with red corresponding to label 1, green to label 2, and blue to label 3.

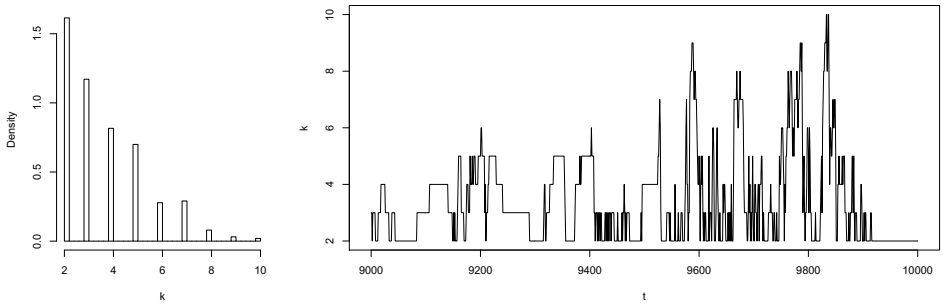

Figure 3.2: *Posterior Distribution of k*

MCMC output for the posterior distribution of k with 10,000 draws. *(Left-Hand Side)* Histogram of posterior draws of k. *(Right-Hand Side)* Last 1000 draws of the MCMC sampler.

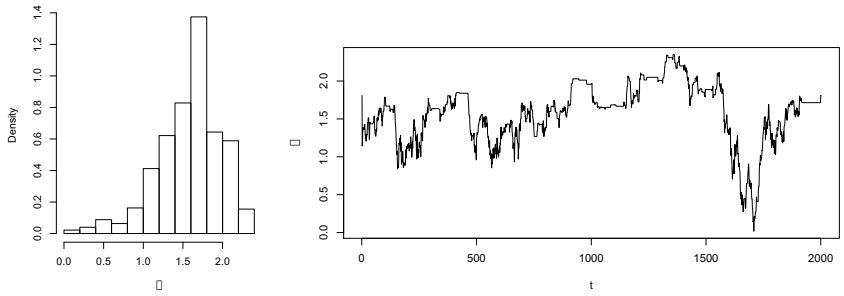

Figure 3.3: *Posterior Distribution of β*

MCMC output for the posterior distribution of β with 10,000 draws. *(Left-Hand Side)* Histogram of posterior draws of β. *(Right-Hand Side)* Last 2000 draws of the MCMC sampler.

where now $\mathbf{T} = \left(\mathbf{T}_{\text{reg}}^T, \mathbf{T}_{\text{train}}^T\right)^T$.

The regression parameters have the usual semi-conjugate priors. The pure effects have prior distribution $\pi(\tau) = N(\tau_0, \Sigma_0)$, and the precision $\gamma = 1/\sigma^2$ of the regression has prior $\pi(\gamma) = \text{Gamma}(\nu_0/2, \nu_0\sigma_0^2/2)$.

The Bayesian inference procedure for classification is readily adapted for the generated regressor problem. For the simplified case with no additional controls $\tilde{X}$, and where $\mathbf{W_T}$ is the design matrix corresponding to the classification vector $\mathbf{T}$, the Gibbs sampler for Bayesian imputation with k-nearest neighbors is the following.

For $j = 1, ..., J$,

1. sample $\left(\tau^{(j+1)}, \left(\sigma^2\right)^{(j+1)}\right)$ from $P\left(\tau, \sigma^2 \left| \mathbf{y}, \mathbf{T}, \beta^{(j)}, k^{(j)}, \mathbf{z}^{(j)} \right.\right)$

   (a) compute

   $$\mathbf{V} = Var\left[\tau \left| \mathbf{y}, \mathbf{T}, \left(\sigma^2\right)^{(j)} \right.\right] = \left(\Sigma_0^{-1} + \mathbf{W}_\mathbf{T}^T \mathbf{W}_\mathbf{T} / \sigma^{2(j)}\right)^{-1}$$

   and

   $$\mathbf{m} = E\left[\tau \left| \mathbf{y}, \mathbf{T}, \left(\sigma^2\right)^{(j)} \right.\right] = \mathbf{V}\left(\Sigma_0^{-1}\tau_0 + \mathbf{W}_\mathbf{T}^T \mathbf{y} / \left(\sigma^2\right)^{(j)}\right)$$

   (b) sample $\tau^{(j+1)} \sim N(\mathbf{m}, \mathbf{V})$

   (c) compute $\text{SSR}\left(\tau^{(j+1)}, \mathbf{T}\right)$

   (d) sample $\left(\sigma^2\right)^{(j+1)} \sim$Inv-Gamma$\left((\nu_0 + n)/2, \left(\nu_0\sigma_0^2 + \text{SSR}\left(\tau^{(j+1)}, \mathbf{T}\right)\right)/2\right)$

2. sample $\left(\beta^{(j+1)}, k^{(j+1)}, \mathbf{z}^{(j+1)}\right)$ from $P\left(\beta, k, \mathbf{z} \left| \mathbf{y}, \mathbf{T}^{(j)}, \tau^{(j+1)}, \left(\sigma^2\right)^{(j+1)} \right.\right)$

   (a) propose (β', k') from $q_1\left(\beta, k \left| \beta^{(j)}, k^{(j)} \right.\right)$

   (b) propose $\mathbf{z}'$ from $f\left(\mathbf{z} \left| \beta', k' \right.\right)$

   (c) draw $u \sim \text{Unif}[0, 1]$

   (d) if

   $$u \leq \alpha,$$

   let $\left(\beta^{(j+1)}, k^{(j+1)}, z^{(j+1)}\right) = (\beta', k', z')$, otherwise $\left(\beta^{(j+1)}, k^{(j+1)}, z^{(j+1)}\right) = \left(\beta^{(j)}, k^{(j)}, z^{(j)}\right)$.

3. sample $\mathbf{T}_{\text{reg}}^{(j+1)}$ from $P\left(\mathbf{T}_{\text{reg}} \middle| \mathbf{y}, \mathbf{T}_{\text{train}}, \tau^{(j+1)}, \left(\sigma^2\right)^{(j+1)}, \beta^{(j+1)}, k^{(j+1)}, z^{(j+1)}\right)$: for $i = 1, ..., N_{\text{reg}}$,

   (a) compute $\theta = \left\{P\left(T_{\text{reg},i} = t \middle| \mathbf{T}_{-i}^{(j)}, \mathbf{y}, \mathbf{T}_{\text{train}}, \beta^{(j)}, k^{(j)}\right) : t \in G\right\}$

   (b) draw $T_{\text{reg},i}^{(j+1)}$ from Multinomial(θ),

where $\mathbf{T}_{-i}^{(j)} = \left(\mathbf{T}_{\text{reg},1:(i-1)}^{(j)}, \mathbf{T}_{\text{reg},(i+1):N_{\text{reg}}}^{(j-1)}, \mathbf{T}_{\text{train}}\right)$ and $\text{SSR}\left(\tau, \mathbf{T}\right) = \|\mathbf{y} - \mathbf{W}_{\mathbf{T}}\tau\|_2^2$.

Remark that the proposal of the auxiliary variable $\mathbf{z}$ in step 2.b is is done in the same way as in Section 2, with $\mathbf{T}_{\text{train}}$ added to the conditioned upon variables.

The acceptance probability α in step 2.d is equal to

$$\begin{aligned}
\alpha_{\text{missing data}} &= \frac{P\left(\beta', k', \mathbf{z}' \middle| \tau, \sigma^2, \mathbf{T}\right)}{P\left(\beta, k, \mathbf{z} \middle| \tau', \sigma^2, \mathbf{T}\right)} \times \frac{q_2\left(\beta, k, \mathbf{z} \middle| \beta', k', \mathbf{z}'\right)}{q_2\left(\beta', k', \mathbf{z}' \middle| \beta, k, \mathbf{z}\right)} \\
&= \frac{Z(\beta, k)}{Z(\beta', k')} \frac{\phi(\mathbf{y} - \tau_{\mathbf{T}}, \sigma^2) \cdot \exp\left(\beta' S(\mathbf{T})/k'\right) \pi(\beta', k')}{\phi(\mathbf{y} - \tau_{\mathbf{T}}, \sigma^2) \cdot \exp\left(\beta S(\mathbf{T})/k\right) \pi(\beta, k)} \frac{g\left(\mathbf{z}' \middle| \mathbf{T}\right)}{g\left(\mathbf{z} \middle| \mathbf{T}\right)} \\
&\quad \times \frac{q_1\left(\beta, k, \mathbf{z} \middle| \beta', k', \mathbf{z}'\right) \cdot \exp\left(\beta S(\mathbf{z})/k\right)}{q_1\left(\beta', k', \mathbf{z}' \middle| \beta, k, \mathbf{z}\right) \cdot \exp\left(\beta' S(\mathbf{z})/k'\right)} \frac{Z(\beta', k')}{Z(\beta, k)} \\
&= \frac{\exp\left(\beta' S(\mathbf{T})/k'\right) \pi(\beta', k')}{\exp\left(\beta S(\mathbf{T})/k\right) \pi(\beta, k)} \frac{g\left(\mathbf{z}' \middle| \mathbf{T}\right)}{g\left(\mathbf{z} \middle| \mathbf{T}\right)} \times \frac{q_1\left(\beta, k, \mathbf{z} \middle| \beta', k', \mathbf{z}'\right) \cdot \exp\left(\beta S(\mathbf{z})/k\right)}{q_1\left(\beta', k', \mathbf{z}' \middle| \beta, k, \mathbf{z}\right) \cdot \exp\left(\beta' S(\mathbf{z})/k'\right)},
\end{aligned}$$

and the full conditional for $T_{\text{reg},i}$ is

$$P_{\text{missing data}}\left(T_{\text{reg},i} = t \middle| \mathbf{T}_{\text{reg},1:(i-1)}^{(j)}, \mathbf{T}_{\text{reg},(i+1):N_{\text{reg}}}^{(j-1)}, \mathbf{y}, \mathbf{T}_{\text{train}}, \beta, k^{(j)}\right)$$

$$= \frac{\phi(y_i - \tau_t) \cdot P\left(T_{\text{reg},i} = t \middle| \mathbf{T}_{-i}, \beta^{(j)}, k^{(j)}\right)}{\sum_t \phi(y_i - \tau_t) \cdot P\left(T_{\text{reg},i} = t \middle| \mathbf{T}_{-i}, \beta^{(j)}, k^{(j)}\right)}$$

$$= \frac{\phi(y_i - \tau_t) \cdot \exp\left(\beta^{(j)}/k^{(j)} \left(\sum_{l \sim_k (\text{reg},i)} \delta_t(\mathbf{T}_{-i,l}^{(j)}) + \sum_{(\text{reg},i) \sim_k l} \delta_{\mathbf{T}_{-i,l}^{(j)}}(t)\right)\right)}{\sum_t \phi(y_i - \tau_t) \cdot \exp\left(\beta^{(j)}/k^{(j)} \left(\sum_{l \sim_k (\text{reg},i)} \delta_t(\mathbf{T}_{-i,l}^{(j)}) + \sum_{(\text{reg},i) \sim_k l} \delta_{\mathbf{T}_{-i,l}^{(j)}}(t)\right)\right)}.$$

Imputed regressors as predicted variables

There is another reasonable modeling approach for this imputed regressor problem which makes for a computationally easier alternative to model (9). Instead of treating the imputed

values as missing data, one could treat them as predicted values. That is, one could assume that the data generating process is the following:

$$(\beta, k, \gamma, \phi) \sim \pi(\beta, k, \gamma, \phi), \tag{3.10}$$

$$\mathbf{T}_{\text{train}} \sim f\left(\mathbf{T} \,|\mathbf{X}_{\text{train}}, \beta, k\right), \tag{3.11}$$

$$T_{\text{knn},i} \sim P\left(T_{n+1} = c \,|\mathbf{T}_{\text{train}}, \mathbf{X}_{\text{train}}, X_{\text{reg},i}, \beta, k\right), \; i = 1, ..., N_{\text{reg}}, \tag{3.12}$$

$$Y_i \sim N(\tau_{T_{\text{knn},i}} + \phi\tilde{X}, 1/\gamma). \tag{3.13}$$

Conceptually, the key difference about this model is the way in which we think of the imputed regressors. Under this model, one thinks of each $\mathbf{T}_{\text{knn},i}$ as a variable drawn from the predictive distribution (5). This may feel conceptually close to the usual (non-probabilistic) implementation of the k-nearest neighbors algorithm, which predicts points out-of-sample using only points in sample. Under this approach, $\mathbf{T}_{\text{knn},i}$ is independent of $\mathbf{T}_{\text{knn},j}$ for all $i \neq j$, conditional on $\mathbf{T}_{\text{train}}$, β and k, because each $\mathbf{T}_{\text{knn},i}$ is drawn from the posterior distribution and the $Y_i|T_{\text{knn},i}$ are assumed to be independent across i's. This is in contrast with the earlier, missing data model, under which $\mathbf{T}_{\text{reg},i}$ informs the posterior for (β, k) and $\mathbf{T}_{\text{reg},-i}$.

Naturally, since only $\mathbf{T}_{\text{train}}$ is observed, the distinction is entirely superficial for the imputation problem taken in isolation. However, the distinction is relevant for the application at hand because we append a regression model to the missing data model. To be sure, the posterior distribution for $\mathbf{T}_{\text{reg},i}$ does carry information not only from $\mathbf{T}_{\text{train}}$, but also from Y_i; that is, the posterior distribution for $\mathbf{T}_{\text{reg},i}$ has information pertaining to the imputation of a neighbor $\mathbf{T}_{\text{reg},j}$ that is not contained in $\mathbf{T}_{\text{train}}$. Conversely, the posterior predictive distribution for $\mathbf{T}_{\text{knn},i}$ contains no such information.

Sampling from the posterior distribution under model (10)-(13) can be done by making direct modifications to the Gibbs sampler for the missing data model. This will facilitate

comparison of the computational requirements of the two samplers. One draws the predicted variable model by using the same Gibbs sampler as for the missing variable model, but using instead, as the acceptance probability,

$$\alpha_{\text{predicted}} = \frac{\exp\left(\beta' S(\mathbf{T}_{\text{train}})/k'\right)\pi(\beta',k')}{\exp\left(\beta S(\mathbf{T}_{\text{train}})/k\right)\pi(\beta,k)} \frac{g\left(\mathbf{z}'\,|\beta',k',\mathbf{T}_{\text{train}}\right)}{g\left(\mathbf{z}\,|\beta,k,\mathbf{T}_{\text{train}}\right)} \times \frac{q_1\left(\beta,k,\mathbf{z}\,|\beta',k',\mathbf{z}'\right)\cdot\exp\left(\beta S(\mathbf{z})/k\right)}{q_1\left(\beta',k',\mathbf{z}'\,|\beta,k,\mathbf{z}\right)\cdot\exp\left(\beta' S(\mathbf{z}')/k'\right)},$$

and, with $\mathbf{T}_{\text{knn}}$ playing the role of the imputed value (*en lieu* of $\mathbf{T}_{\text{reg}}$), using the full conditional

$$P_{\text{predicted}}\left(T_{\text{knn},i} = t \,\middle|\, \mathbf{T}^{(j)}_{\text{knn},1:(i-1)}, \mathbf{T}^{(j-1)}_{\text{knn},(i+1):N_{\text{knn}}}, \mathbf{y}, \mathbf{T}_{\text{train}}, \sigma^2, \tau, \beta^{(j)}, k^{(j)}\right)$$

$$= P\left(T_{\text{knn},i} = t \,\middle|\, y_i, \mathbf{T}_{\text{train}}, \sigma^2, \tau, \beta^{(j)}, k^{(j)}\right)$$

$$= \frac{\phi(y_i - \tau_t)\cdot P\left(T_{\text{knn},i} = t \,\middle|\, \mathbf{T}_{\text{train}}, \beta^{(j)}, k^{(j)}\right)}{\sum_t \phi(y_i - \tau_t)\cdot P\left(T_{\text{knn},i} = t \,\middle|\, \mathbf{T}_{\text{train}}, \beta^{(j)}, k^{(j)}\right)}$$

$$= \frac{\phi(y_i - \tau_t)\cdot \exp\left(\beta^{(j)}/k^{(j)}\left(\sum_{l\sim_k(\text{knn},i)}\delta_t(T^*_l) + \sum_{(\text{knn},i)\sim_k l}\delta_{T^*_l}(t)\right)\right)}{\sum_t \phi(y_i - \tau_t)\cdot \exp\left(\beta^{(j)}/k^{(j)}\left(\sum_{l\sim_k(\text{knn},i)}\delta_t(T^*_l) + \sum_{(\text{knn},i)\sim_k l}\delta_{T^*_l}(t)\right)\right)},$$

where the normalization terms $Z_{\text{knn},i}(\beta^{(j)}, k^{(j)})$ canceled out.

The simplifications in the Gibbs sampler are thus made salient. The realized value of $\mathbf{T}_{\text{knn}}$ does not inform the distribution of (β, k), and as such does not enter their full conditional distribution. Likewise, under this perspective, the full conditionals for (β, k) will not depend on τ or σ^2. Furthermore, the full conditional for $T_{\text{knn},i}$ will not depend on $\mathbf{T}_{\text{knn},-i}$. This simplifies the Gibbs sampler substantially. In particular, the Monte Carlo sampling of (β, k) *can be done separately* from that of the other parameters $\tau, \sigma^2, \mathbf{T}_{\text{knn}}, \mathbf{z}$.

This suggests an even simpler approach which I implement for comparison: one can simply treat (β, k) as parameters to estimate, and fix their value at some estimate $(\hat{\beta}, \hat{k})$, for instance by fixing them at their mode values from previous runs of the MCMC of Section 2. We will find in simulation that such increases in ease of computation come at a cost in accuracy.

I extend the simulation study from the previous section and append a regression function to the data generating process described therein. I pick $N_{\text{train}} = N_{\text{reg}} = 30$ and generate

Estimates	τ_1	τ_2	τ_3
OLS (true T_{reg})	1.38 (0.68,2.08)	1.59 (0.89,2.29)	3.33 (2.63,4.03)
KNN Emp. Bayes	1.74 (0.87,2.65)	1.49 (0.67,2.26)	2.62 (1.68,3.47)
KNN Predict	1.70 (0.66,2.85)	1.43 (0.55,2.36)	2.75 (1.61,3.78)
KNN MD	1.32 (0.33,2.39)	1.52 (0.84,2.25)	2.77 (1.98,3.60)
SVM	1.21 (−0.05,2.47)	1.48 (0.23,2.73)	2.88 (1.58,4.18)

Mean Squared Error	$P(T = T_{\text{true}})$
OLS (true T_{reg})	-
KNN Emp. Bayes	0.80
KNN Predict	0.83
KNN MD	0.87
SVM	0.86

Table 3.1: *Output Comparison*

OLS (true T_{reg}) gives the least-squares estimates when regressors are *not* missing. *KNN Emp. Bayes* gives the posterior mode from the Bayesian KNN when (β, k) is fixed at some estimated value $(\hat{\beta}, \hat{k})$. *KNN Predict* gives the posterior mode from the Bayesian KNN where $T = T_{\text{knn}}$ is treated as a predicted value. *KNN MD* gives the posterior mode from the Bayesian KNN where $T = T_{\text{reg}}$ is treated as a missing data. *SVM* gives output from the two-step method with imputation in the first step done with SVM, a bias correction is applied (see appendix), and standard errors account for its uncertainty.

balanced datasets for both the training (train) and regression (reg) data. The observed training data is $\mathcal{D}_{\text{train}} = \{T_{\text{train},i}, X_i : i = 1, ..., N_{\text{train}}\}$ and the observed regression data is $\mathcal{D}_{\text{reg}} = \{Y_i, X_i : i = 1, ..., N_{\text{reg}}\}$, which omits $T_{\text{reg},i}$, and where

$$Y_i = \tau_{T_{\text{reg},i}} + \varepsilon_i,$$

with $\tau_t = t$, $t = 1, 2, 3$, and $\varepsilon_i \sim N(0, 0.5)$.

From Table 5, we can tell that the model in which $\mathbf{T}_{\text{reg}}$ is treated as missing data −which is more informative− offers a performance superior to that of the alternatives. The regression coefficient estimates are closer to both the true population values and the sample OLS estimates when regressing on the true values of T.

Even more telling is the improvement observed in the probability of correct imputation, from treating T_{reg} as missing data entering the joint posterior, which allows the variation in $\mathbf{y}_{-i}$ to inform the imputation of $T_{\text{reg},i}$.

This is an important observation because it means that including a regression in the model is beneficial even if the ultimate objective is imputation of $\mathbf{T}_{\text{reg}}$. To be sure, Y cannot be used directly for training the KNN algorithm because it is not available for observations for which

we observe T. One could in fact use multiple regression functions which one knows T enters non-trivially in order to improve imputation.[8]

3.4 Two-Step Estimation of Regression Parameters

The joint modeling approach described in Section 3 makes efficient use of the information in order to estimate the regression coefficient, and a Gibbs sampler is detailed in order to draw from their joint posterior distribution. However, the algorithm laid out in Section 3 for the estimation of the joint model is not trivial to implement. As detailed above, convergence must be assessed by inspection, and tuning parameters must be chosen. Even more problematic is the need to approximate a normalizing constant.

Cucala et al. (2012) survey different approaches for dealing with the normalization constant, and suggest the auxiliary variable strategy detailed in Section 2. This approach remains tedious and computationally expensive, because the analyst still needs to obtain a good approximation to the normalizing constant. In a recent article, Yoon and Friel (2015) circumvent this issue by using a pseudolikeihood approach. However, as noted in Cucala et al. (2012), the pseudolikeihood approach may perform very poorly. It is thus worthwhile asking whether a two-step approach may provide useful estimates that are easy to compute.

Two-step estimation of the regression function may remain troublesome. On the one hand, consistent estimation using directly the KNN or symmetrized KNN algorithm, and thus imputing directly a prediction of the category, necessitates the computation of a bias correction term −which consists, as exemplified in the appendix, in an unenviable task. On the other hand, imputation of category probabilities (which, as detailed below, can provide a consistent regression coefficient estimate) using probabilistic KNN requires the estimation of the normalizing constant.

A first motivation for a two-step approach is computational simplicity. Consequently, one would like an estimator which requires neither the computation of a bias correction term or of

[8] At this point, it may be fruitful to rephrase the exercise as a smoothing problem.

an approximation to an intractable normalizing constant. In this section, I suggest such an estimator and investigate its performance.

A second motivation for the use of a two-step method is robustness. If one is concerned about model misspecification in the second stage, then it could be that estimation of the second-stage parameters will be more accurate if estimation is done in two stages, i.e. if imputation is not informed by the outcome variable through a misspecified regression function. It is thus important, if only for robustness checks, that the analyst has access to a reliable two-step method.

A further observation motivates our efforts to develop a two-step procedure similar in nature to that of Chapters II. There, motivated by a desire to circumvent the specification of the covariance of the regression errors, we proposed a model in which randomness arises from survey sampling, along with an inference procedure based on the bootstrap. A somewhat different concern will, in the case at hand, motivate the investigation of such an assumption and procedure. Under a standard, frequentist resampling scheme, computation of closed form standard errors for two-step estimators (see, for more standard cases, Murphy and Topel (1985)) requires knowledge of the asymptotic distribution of the parameters estimated in the first step. Asymptotic distribution of continuous maximum likelihood parameters is standard (Van der Vaart, 2000), and that of discrete parameters is largely elucidated (Choirat and Raffaello, 2012). However, the question of how to obtain the joint distribution of a discrete (i.e., k) and a continuous (i.e., β) maximum likelihood estimate is, to the best of my knowledge, still open.[9] The bootstrap approach and survey sampling assumption circumvent the need to estimate the covariance of the parameters estimated in the first stage, while still capturing the uncertainty due to their estimation.

As in Chapter II, the assumption of sampling with replacement is not exact, since the same potential respondent may not be surveyed twice (not even in separate supplementary surveys). However, the population of potential respondents is so large (and thus the chance of sampling the same respondent twice so small) that the assumption of drawing without replacement is

[9]For efforts in that direction, see references in Choirat and Raffaello (2012).

arguably innocuous.

Let $\mathscr{Y}$, $\mathscr{X}$, $\mathscr{\tilde{X}}$, and $\mathscr{T}$ be the vectors and matrices collecting the random values, for all individuals of the surveyed population, of $Y, X, \tilde{X}$, T, and let $\mathscr{S} = (\mathscr{Y}, \mathscr{X}, \mathscr{\tilde{X}}, \mathscr{T})$. Let $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{reg}}$ be the sampled training and regression datasets, respectively, both of which are subsets of $\mathscr{S}$ and have data missing as described in Section 1.

3.4.1 Two-Step Bootstrap

I give a two-step method which, analogously to that described in Chapter II, provides estimates of the coefficients of interest, as well as standard errors accounting for first-stage uncertainty. The method sacrifices efficiency for gains in ease of implementation and robustness (and, perhaps to some, interpretability). The two-step method for imputation with k-nearest neighbors may be described with the following pseudocode (for ease of exposition, I consider the case with binary category, i.e., $|C| = 2$).

For $r = 1, ..., R$,

1. Find $\hat{k}^{(r)}$ by cross-validation, using dataset $\mathcal{D}^{(0)}_{\text{train}}$ drawn with replacement from $\mathcal{D}_{\text{train}}$
2. For $l = 1, ..., L$,
   (a) Draw $\mathcal{D}^{(l)}_{\text{train}}$ from $\mathcal{D}_{\text{train}}$ with replacement
   (b) Fit $\hat{\mathbf{T}}^{(l)}_{\text{reg}} = \left\{\hat{T}^{(l)}_{\text{reg},i}\right\}_{i=1,...,N_{\text{reg}}}$ using KNN with $\hat{k}^{(r)}$ neighbors
3. Compute $\bar{\mathbf{T}}^{(r)}_{\text{reg}} = \left\{\frac{1}{L}\sum_{l=1}^{L} \hat{T}^{(l)}_{\text{reg},i}\right\}_{i=1,...,N_{\text{reg}}}$
4. Draw dataset $\mathcal{D}^{(r)}$ with replacement from $\left(\mathcal{D}_{\text{reg}}, \bar{\mathbf{T}}^{(r)}_{\text{reg}}\right)$
5. Compute the regression coefficients $(\tau^{(r)}, \phi^{(r)})$ from dataset $\mathcal{D}^{(r)}$.

Estimates may then be obtained by taking the average of the bootstrap draws $\tau^{(r)}, \phi^{(r)}$, $r = 1, ..., R$, and confidence intervals can be obtained by taking the appropriate quantiles. The key to imputing classification probabilities without computing the normalizing constant lays in step 2. The algorithm does not compute just a single vector of imputed categories $\hat{\mathbf{T}}_{\text{reg}}$

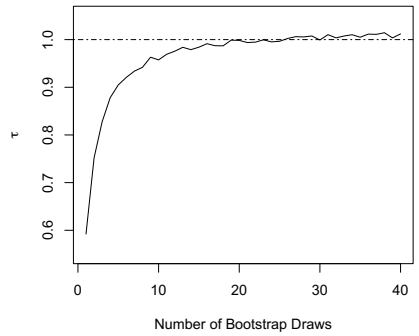

Figure 3.4: *Precision Gains from Bootstrapping the First Step*

The average (over the $R = 25$ draws from outer loop bootstrap) value of τ is plotted against the number of bootstrap draws L used in the inner loop.

(which, if used as imputed variables, would induce attenuation bias; see Lewbel (2007), and the appendix for the $|C| > 2$ case) but L of them, and imputes their average $\bar{\mathbf{T}}_{\text{reg}}$, where $\bar{\mathbf{T}}_{\text{reg},i}$ is an estimate of $E\left[\mathbf{1}\left\{T_i = 1\right\}\right] = P\left(\mathbf{1}\left\{T_i = 1\right\}\right)$. If the bootstrap sub-procedure of step 2 provides accurate estimates of these expectations, then a consistency argument (detailed in Subsection 4.2 below) suggests we may obtain accurate regression coefficient estimates.

It is worth assessing the gain in the estimation accuracy of the regression coefficients from bootstrapping the KNN imputation procedure. In order to illustrate the gain, I simulated data mimicking that of the application (Section 5), with the regression coefficient for the imputed variable set to 1. We can see from Figure 4 that in the case of standard imputation ($L = 1$), attenuation bias is an issue and we seriously underestimate the target parameter. However, with as few as $L = 20$ bootstrap draws in the subroutine, the attenuation bias can be greatly reduced. This is very fortunate, because it allows us to circumvent the tradeoff, previously encumbering the literature, between plugging a best guess which induced attenuation bias, and computing a probabilistic prediction which often required approximating the intractable normalizing constant.

3.4.2 Identification and Inference

Define individual dummy (random) variables for each category, i.e., $(D_1, ..., D_{|C|})$ satisfying $\sum_{i=1}^{|C|} D_i = 1$ where $D_i \in \{0, 1\}$ for $i = 1, ..., |C|$. Let the vectors of dummies for the surveyed population be denoted $\mathscr{D} = \left(\mathscr{D}_1, ..., \mathscr{D}_{|C|}\right)$. That is, I make explicit the linear dependence on T by defining a dummy variable for each of the values T may take on, and one can write $\mathscr{S} = (\mathscr{Y}, \mathscr{X}, \tilde{\mathscr{X}}, \mathscr{D})$. We are interested in estimating

$$E^*\left[\mathscr{Y}\left|\mathscr{D}, ..., \mathscr{D}_{|C|}, \tilde{\mathscr{X}}\right.\right] = \tau_1\mathscr{D}_1 + \cdots + \tau_{|C|}\mathscr{D}_{|C|} + \phi\tilde{\mathscr{X}}. \tag{3.14}$$

To be sure, a poor estimate of $E^*\left[\mathscr{D}_i\left|\tilde{\mathscr{X}}, \mathscr{X}\right.\right]$, $i = 1, ..., |C|$, would in turn make for a poor estimate of $E^*\left[\mathscr{Y}\left|\mathscr{D}, \tilde{\mathscr{X}}\right.\right]$. We may, however, consider the conditional expectations $\mathscr{Q}_j := E\left[\mathscr{D}_i\left|\tilde{\mathscr{X}}, \mathscr{X}\right.\right]$, $j = 1, ..., |C|$, and define the best linear predictor

$$E^*\left[\mathscr{Y}\left|\mathscr{Q}_1, ..., \mathscr{Q}_{|C|}, \tilde{\mathscr{X}}, \mathscr{X}\right.\right] = \tau\mathscr{Q}_1 + \cdots + \tau_{|C|}\mathscr{Q}_{|C|} + \phi\tilde{\mathscr{X}}. \tag{3.15}$$

The τ_i's defined in (14) are estimated by plugging in estimates of Q_i in the regression equation (1^{st} step), and then proceeding with the estimation of the linear predictor by least squares methods (2^{nd} step).

Even though inference is done conditional on $\mathscr{S}$, an ergodicity argument may establish the coefficients defined by the best linear predictor in population, $E^*\left[Y\left|D_1, ..., D_{|C|}, \tilde{X},\right.\right]$, as the target parameters (i.e., those which are consistently estimated). Our use of k-nearest neighbors as an imputation procedure relies on the belief that there is some dependence between nearby observations (with respect to Euclidian distance between different $(\tilde{X}, X_{\text{reg}})$ values, or whichever distance may be used to define "neighborliness"). However, if that dependence declines fast enough (with respect to distance), and the data set is "large" enough (in terms of the said distance), then it may be argued that the one sample may be broken down in many mutually independent subsamples which may be treated as independent replications of a survey, thus allowing for the consistent estimation of quantities such as $E^*\left[Y\left|D_1, ..., D_{|C|}, \tilde{X}\right.\right]$ (assuming the exclusion restriction) and $E\left[D_i\left|\tilde{X}_i, X_i\right.\right]$.

3.5 Application: Flexible Work Hours and Voter Turnout

There is a well-established literature in political science investigating the importance of voter turnout (Wolfinger and Rosenstone, 1980; Fowler, 2015) as well as its determinants (Enos and Fowler, 2016). The determinants of voter turnout (to the extent that we aim at increasing it) have immediate implications for public policy, and their accurate assessment is critical to the promulgation of optimal policies. Should we invest more financial and political capital in increasing the geographical proximity of voting booths? Should we rather concentrate our efforts on making sure that employers allow their employees to take time off to go vote? I tackle one specific question and ask: *controlling for other observables, is work hour flexibility correlated with higher voter turnout?* To be sure, I do not claim to identify a causal effect, but more modestly aim to estimate the coefficient on the indicator T for "flexible work hours" in a specific best linear predictor for Y, here defined as the indicator variable for having voted or not.

As the main application, I inquire into a specific use of CPS data in which some form of imputation is necessary. The survey has a large set of core questions which are asked to every surveyed individual. In addition, there are supplementary surveys which inquire about additional topics (such as work flexibility or voting habits). Supplementary surveys only cover part of the population covered by the standard survey, and in general distinct supplementary surveys cover distinct, non-overlapping sub-populations. Table 2 details the data availability for two supplementary surveys.

We will be interested in the impact of having flexible work hours on voting behavior. This will require imputing covariates because both variables are only available in (non-overlapping) supplementary surveys.

In order to implement the Bayesian k-nearest neighbors approach to missing data detailed in Section 3, some parts of the method need to be specified. The prior parameters are $\tau_0 = 0$, $\sigma_0^2 = 1000$, $\nu_0 = 1$. The chosen proposal distribution $q_1\left(\beta, k \left| \beta^{(j)}, k^{(j)} \right.\right)$ is defined as follows,

	Core variables	Vote data	Work schedule data
Voter supp.	YES	YES	NO
Work schedule supp.	YES	NO	YES

Table 3.2: *CPS Supplementary Survey*

The work schedule CPS supplementary survey ($\mathcal{D}_{\text{train}}$) and the voter turnout supplementary survey ($\mathcal{D}_{\text{reg}}$) both include all questions from the standard CPS questionnaire. However, they are non-overlapping; no surveyed individual answered both supplementary surveys.

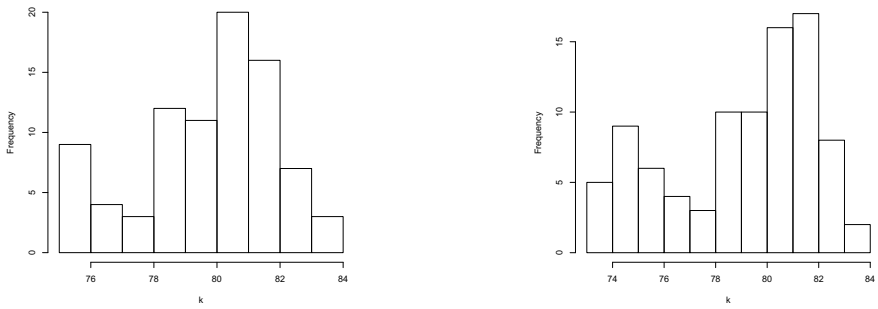

Figure 3.5: *Choice of k*

(Left) Histogram of out-of-sample mean-squared error in cross-validation as a function of k. *(Right)* Posterior distribution of k.

$$q(k'|k) = \begin{cases} \text{Unif}\left\{k-1, k+1\right\}, & k > 1 \text{ and } k < K \\ \text{Unif}\left\{1, 2\right\}, & k = 1 \\ \text{Unif}\left\{K-1, K\right\}, & k = K \end{cases},$$

and

$$q(\beta'|\beta) = \phi_{0,\beta_{\max}}(\beta' - \beta, \sigma^2_{\beta'}),$$

where $\phi_{0,\beta_{\max}}$ is the Gaussian density truncated below at 0 and above at $\beta_{\max}$, and $\sigma^2_{\beta'} = 0.05$. Convergence was assessed by starting the chain at various, far off starting points and witnessing it stabilizing at ostensibly equal distributions.

To capture the effect of flexible work hours as best possible, I include in $\tilde{X}$ the control variables for age, gender, race, geographic location, whether one is a citizen, as well as one's education, all of which are available in the CPS data. These control variables make up $\tilde{X}$ and

	τ	s.d.
Naive KNN	0.031	0.003
Naive SVM	0.035	0.002
Bias Corrected SVM	0.055	0.009
Bayesian KNN	0.049	0.005
Two-Step KNN	0.044	0.017

Table 3.3: *Coefficients and Uncertainty Estimates*

Naive KNN imputes with KNN and gives estimates conditional on the imputed regressors. *Naive SVM* imputes with SVM and gives estimates conditional on the imputed regressors. *Bias Corrected SVM* imputes a predicted $T_{\text{SVM}} \in C$ using SVM and corrects for the attenuation bias from misclassification; the standard errors account for uncertainty in the estimated attenuation bias coefficient but are conservative. *Bayesian KNN* implements the method laid out in Section 3. *Two-Step KNN* is an implementation of the bootstrap procedure described in Section 4.

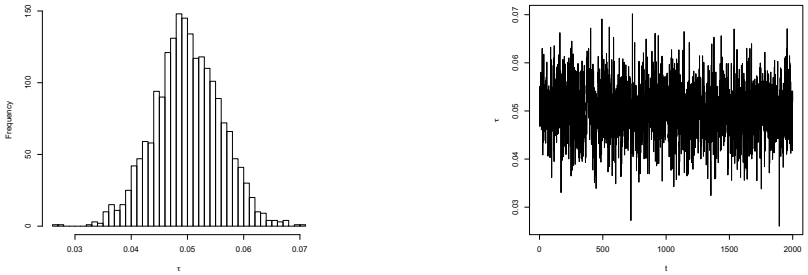

Figure 3.6: *Posterior Distribution of τ*

(Left) Posterior distribution of τ using 2000 draws (after burn-in). *(Right)* Plot of posterior draws of τ.

are described in Table B.1 in the appendix.

The data structure detailed in Table 2 makes the application at hand a natural candidate for the exclusion restriction $E^*\left[Y \middle| T, \tilde{X}, X_{\text{reg}}\right] = E^*\left[Y \middle| T, \tilde{X}\right]$. The analyst may choose the variables constituting X_{reg} from the core CPS data only, which is to say that all variables constituting X_{reg} are also available in the voting turnout supplementary survey to be used as controls $\tilde{X}$. If some variable X_c in the CPS core data is not added to the control variables $\tilde{X}$, it is therefore because the analyst assumed it was not explanatory, i.e., $E^*\left[Y \middle| T, \tilde{X}, X_c\right] = E^*\left[Y \middle| T, \tilde{X}\right]$.

In the original CPS data, the flexible work hours and voting categorical variables take on

more than two values. The possible values for flexible work hours are {"no", "yes", "refuse to answer", "don't remember", "missing data"}. I keep only the subset {"no", "yes"}. This reduces the size of the data set from 378, 892 to 165,318 observations. The performance of SVM can be impacted when the points to classify are highly imbalanced; in this case, with 104,907 "no" and 60,411 "yes", the balance turns out not to cause any problems. The possible values for voting are {"did not register", "registered, did not vote", "voted", "Don't know/refused/no response", "missing data"}. Only the subset {"did not register", "registered, did not vote", "voted"} is informative, and could be used directly. However, I encode both "did not register" and "registered, did not vote" as "did not vote" and carry out analysis with a binary response. This has the practical advantage that we will be able to get some traction with a simple linear regression model.

The jagged posterior for k and cross-validation estimates of mean-squared error with respect to k suggest there may be gains from model averaging over k. To be sure, gains from model averaging (over k) can be expected if minimal values of out-of-sample misclassification probabilities are attained with different k's yielding different out-of-sample predictions, which we may expect to be the case when witnessing a multimodal posterior or cross-validation histogram of mean-squared error against k, as is the case in this application.

Table 3 gives estimates of the effect of flexible work hours from different methods. Which one should be given the most credence? Although SVM appeared to provide superior estimates in the imputation problem taken in isolation, the information gain from using a "one-step"[10] approach apparently trumps such gains. Furthermore, the (admittedly conservative) standard errors obtained for the 2-step approach with SVM are quite large. The KNN two-step method gives a point estimate of 0.044, which is encouragingly close to the Bayesian KNN estimate of 0.049. However, its estimated standard errors are by far the largest.

Holding constant the aforementioned control variables, the Bayesian procedure indicates that we should expect the probability of voting to be about 5 percentage points higher (on

[10]The distinction between a "two-step" and a "one-step" approach is detailed in Chapter I. In a "two-step" approach, one first imputes the missing covariate ignoring the outcome variable, and then carries out regression on the imputed data set. In a "one-step" approach, all parameters are estimated jointly.

average) if one had flexible work hours. Overall in the dataset, 61.4% of the surveyed individuals did vote. If the analyst is confident that the data set is sufficiently large to invoke asymptotic results, the credible intervals may be interpreted as confidence intervals. The corresponding 5% size test would reject the nulls that $\tau = \hat{\tau}_{KNN,\text{ naive}}$ and $\tau = \hat{\tau}_{SVM,\text{ naive}}$, i.e., setting the null parameter at the value we obtained with the dataset at hand by carrying out first-stage imputation (without bias correction) with KNN and SVM, respectively. However, it would fail to reject the null that $\tau = \hat{\tau}_{SVM}$ or $\tau = \hat{\tau}_{2SKNN}$, the parameters obtained in the two-step method with imputation done using SVM with attenuation bias correction and the two-step KNN, respectively. This suggests that although the estimated standard errors for two-step SVM with bias correction and two-step KNN are unpractically large, the point estimates may be reliable.

We were of course limited in the selection of control variables, and the estimates of Table 3 could fall prey to, e.g., omitted variable bias. Endogeneity, more generally, could be an issue; people who value more greatly things outside of work, which could include civic duties such as voting, may seek work with flexible hours.

3.6 Discussion

In two-step methods for the imputation of missing regression covariates, if we can impute using the best linear predictor for T in the first step, then the second step will deliver a consistent estimator of τ. Thus, we can circumvent the issue of computing the bias correction term altogether. Above, we developed a distribution function for $\mathbf{T}_{\text{reg}}$ treated as a random variable, hence making such an approach feasible. That is an advantage of the two-step method with a probabilistic KNN as the first step over that with SVM as the first step, since probabilistic treatment with the latter is much more problematic (Williams & Rasmussen, 2006), and imputing directly the −best guess of the− categories generally requires correction of the attenuation bias.

As detailed in the appendix, inference with a correction term for the attenuation bias involves quite a bit of additional work and uncertainty (due to the estimation of the correction

term). This is further motivation for using an imputation method with a probabilistic implementation, even for use as the first step of a two-step method (see Chapter I for further discussion of the distinction between one-step and two-step methods).

That being said, imputation with probabilist KNN remains challenging because one must approximate an intractable normalizing constant, and closed form standard errors are not available. A bootstrap approach, along with the assumption of a survey sampling data generating process as described in Section 4, circumvents all those issues at once while delivering reasonable accuracy. It thus seems to be the most practical approach.

If the analyst is confident that the stated model is well-specified, and uses a probabilistic implementation of, say, KNN in the first step of a two-step estimator, then the analyst may gain much information at little cost in ease of implementation by using the equivalent one-step method.

However, the suggestion that if one uses a probabilistic implementation of the first-step imputation method, then one should use the more efficient one-step alternative method, cannot be offered as a general recommendation. For instance, if one is concerned about model misspecification in the second stage, then it could be that estimation of the second-stage parameters will be more accurate if estimation is done in two stages, i.e. if imputation is not informed by the outcome variable through a severely misspecified regression function.

We noted in Section 2 that gains in accuracy were obtained from using a modified, symmetrized count of a point's neighbors. It would be interesting to see if the gains from symmetrization obtain more generally and, were this to be the case, to provide a statistical explanation.

3.7 Conclusion

There is a host of problems in economics that are essentially prediction problems, and these are often well tackled with machine learning methods, themselves optimized for prediction tasks (Kleinberg et al., 2015). Missing data problems fall in this realm, as the task of imputing missing data is essentially a prediction problem.

However, it may be inefficient to simply use the machine learning predictions as plug-in values for the missing observations (i.e., two-step methods) because there are efficiency gains from letting the outcome variable inform imputation. Consequently, efficient estimation cannot be carried out using off-the-shelf machine learning methods. Machine learning methods, such as KNN, need to be extended and adapted in order to be embedded in a one-step estimation procedure, as was done in this article.

Furthermore, off-the-shelf machine learning methods do not always come with uncertainty assessments, hence requiring additional efforts to compute statistics such as standard errors for two-step estimators where the first step consists in imputation using machine learning methods.

In this article, I gave a full Bayesian treatment of the problem of regression analysis with an unobserved independent variable. In doing so, I extended the work of Cucala et al. (2012), who gave a correct Bayesian treatment of KNN. I furthermore suggested modifications to the model which yielded simplifications of the sampling algorithm. In order to compare the performance of the Bayesian KNN approach, I implemented a two-step estimator with the popular SVM algorithm as the imputation method used in the first step, as well as a two-step bootstrap method with KNN imputation in the first step.

Because it corresponded to the structure of the application's data, the method was laid out for data in which the missing covariate was observed for none of the observations for which the outcome variable was observed. It is trivial to extend the methods to the case in which the covariate is missing only for some of the observations for which the outcome variable is observed.

References

Abramowitz, Milton, and Irene A. Stegun. *Handbook of mathematical functions*. Vol. 1046. New York: Dover, 1965.

Acemoglu, Daron, Simon Johnson, and James A. Robinson. "The colonial origins of comparative development: An empirical investigation." *American Economic Review* 91, no. 5 (2001): 1369-1401.

Bandyopadhyay, Soutir, Soumendra N. Lahiri, and Daniel J. Nordman. "A frequency domain empirical likelihood method for irregularly spaced spatial data." *The Annals of Statistics* 43, no. 2 (2105): 1-28.

Billingsley, Patrick. *Probability and Measure*. John Wiley & Sons, 2008.

Card, David. "The impact of the Mariel boatlift on the Miami labor market." *Industrial & Labor Relations Review* 43, no. 2 (1990): 245-257.

Choirat, Christine, and Raffaello Seri. "Estimation in discrete parameter models." *Statistical Science* 27, no. 2 (2012): 278-293.

Chay, Kenneth Y., and Michael Greenstone. *The impact of air pollution on infant mortality: evidence from geographic variation in pollution shocks induced by a recession*. No. w7442. National Bureau of Economic Research, 1999.

Cressie, Noel, Soumendra Nath Lahiri, and Yoondong Lee. "On asymptotic distribution and asymptotic efficiency of least squares estimators of spatial variogram parameters." *Journal of Statistical Planning and Inference* 103, no. 1 (2002): 65-85.

Cressie, Noel. *Statistics for spatial data*. John Wiley & Sons, 2015.

Cressie, Noel, and Soumendra Nath Lahiri. "The asymptotic distribution of REML

estimators." *Journal of multivariate analysis* 45, no. 2 (1993): 217-233.

Cucala, Lionel, Jean-Michel Marin, Christian P. Robert, and D. Michael Titterington. "A Bayesian reassessment of nearest-neighbor classification." *Journal of the American Statistical Association* (2012).

Davidson, James. *Stochastic Limit Theory: An Introduction for Econometricians*. Oxford University Press, UK, 1994.

Dell, Melissa, Benjamin F. Jones, and Benjamin A. Olken. "Temperature shocks and economic growth: Evidence from the last half century." *American Economic Journal: Macroeconomics* (2012): 66-95.

Dell, Melissa, Benjamin F. Jones, and Benjamin A. Olken. *What do we learn from the weather? The new climate-economy litrature*. No. w19578. National Bureau of Economic Research, 2013.

Devroye, Luc, Lazlo Györfi, and Gabor Lugosi. *A probabilistic theory of pattern recognition*. Vol. 31. Springer Science & Business Media, 2013.

Dimitrakopoulos, Roussos, Hussein Mustapha, and Erwan Gloaguen. "High-order statistics of spatial random fields: exploring spatial cumulants for modeling complex non-Gaussian and non-linear phenomena." *Mathematical Geosciences 42*, no. 1 (2010): 65-99.

DiNardo, John, Nicole M. Fortin, and Thomas Lemieux. *Labor market institutions and the distribution of wages, 1973-1992: A semiparametric approach*. No. w5093. National Bureau of Economic Research, 1995.

Doukhan, Paul. *Mixing*. Springer New York, 1994.

Enos, Ryan D., and Anthony Fowler. "Aggregate Effects of Large-Scale Campaigns on Voter Turnout." 2016.

Ferguson, Thomas Shelburne. *A course in large sample theory*. Vol. 49. London: Chapman & Hall, 1996.

Fowler, A., 2015. Regular voters, marginal voters and the electoral effects of turnout. *Political Science Research and Methods*, 3(02), pp.205-219.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. *The elements of statistical*

learning. Springer, Berlin: Springer series in statistics, 2001.

Fuentes, Montserrat. "Approximate Likelihood for Large Irregularly Spaced Spatial Data." *Journal o*f the American Statistical Association 102, no. 477 (2007): 321-331.

Gelfand, Alan E., Peter Diggle, Peter Guttorp, and Montserrat Fuentes, eds. *Handbook of spatial statistics*. CRC press, 2010.

Gelman, Andrew, and Xiao-Li Meng. "Simulating normalizing constants: From importance sampling to bridge sampling to path sampling." *Statistical science* (1998): 163-185.

Giraitis, L., and D. Surgailis. "A central limit theorem for quadratic forms in strongly dependent linear variables and its application to asymptotical normality of Whittle's estimate." *Probability Theory and Related Fields* 86, no. 1 (1990): 87-104.

Goodfellow, Ryan, Hussein Mustapha, and Roussos Dimitrakopoulos. "Approximations of high-order spatial statistics through decomposition." *Geostatistics Oslo 2012*, pp. 91-102. Springer Netherlands, 2012.

Hall, Peter, and Christopher C. Heyde. *Martingale Limit Theory and its Application*. Academic press, 2014.

Harville, David A. "Bayesian inference for variance components using only error contrasts." *Biometrika* 61, no. 2 (1974): 383-385.

Herlihy, Alan T., John L. Stoddard, and Colleen Burch Johnson. "The relationship between stream chemistry and watershed land cover data in the mid-Atlantic region, US." *Biogeochemical Investigations at Watershed, Landscape, and Regional Scales*, pp. 377-386. Springer Netherlands, 1998.

Holmes, C. C., and N. M. Adams. "A probabilistic nearest neighbor method for statistical pattern recognition." *Journal of the Royal Statistical Society: Series B* (Statistical Methodology) 64, no. 2 (2002): 295-306.

Holmes, Christopher C., and Niall M. Adams. "Likelihood inference in nearest-neighbor classification models." *Biometrika* 90, no. 1 (2003): 99-112.

Ibragimov, Il´dar Abdulovich. *Independent and stationary sequences of random variables*. (1971).

Ibragimov, Ildar Abdulovich, and Yurii Antolevich Rozanov. *Gaussian random processes*. Vol. 9. Springer Science & Business Media, 2012.

Jiang, Jiming. "REML estimation: asymptotic behavior and related topics." *The Annals of Statistics* 24, no. 1 (1996): 255-286.

De Jong, Robert M. "Central limit theorems for dependent heterogeneous random variables." *Econometric Theory* 13, no. 03 (1997): 353-367.

de Jong, Peter. "A central limit theorem for generalized quadratic forms." *Probability Theory and Related Fields* 75, no. 2 (1987): 261-277.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. "Prediction policy problems." *The American economic review* 105, no. 5 (2015): 491-495.

Kolmogorov, A. N., and Yu A. Rozanov. "On strong mixing conditions for stationary Gaussian processes." *Theory of Probability & Its Applications* 5, no. 2 (1960): 204-208.

Kremer, M., Raissa Fabregas, Jon Robinson, and Frank Schilbach. *What Institutions are Appropriate for Generating and Disseminating Local Agricultural Information?* Working paper. 2015.

Lahiri, S. N. "Central limit theorems for weighted sums of a spatial process under a class of stochastic and fixed designs." *Sankhyā: The Indian Journal of Statistics* (2003): 356-388.

Lewbel, Arthur. "Estimation of average treatment effects with misclassification." *Econometrica* 75, no. 2 (2007): 537-551.

Lin, Fuming. "The asymptotic behavior of quadratic forms in φ-mixing random variables." *Journal of computational and applied mathematics* 233, no. 2 (2009): 437-448.

Little, Roderick JA, and Donald B. Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 2014.

Maccini, S., and D. Yang. 2009." Under the Weather: Health, Schooling, and Economic Consequences of Early-Life Rainfall." *American Economic Review* 99, no. 3: 1006-26.

Madsen, Lisa, David Ruppert, and N. S. Altman. "Regression with spatially misaligned data." *Environmetrics* 19.5 (2008): 453.

Mardia, Kanti V., and R. J. Marshall. "Maximum likelihood estimation of models for

residual covariance in spatial regression." *Biometrika* 71, no. 1 (1984): 135-146.

Matheron, Georges. *Traité de géostatistique appliquée.* Editions Technip, 1962.

McCullagh, Peter, and John A. Nelder. *Generalized linear models.* Vol. 37. CRC press, 1989.

Meyer, Bruce D., and Robert Goerge. "Errors in survey reporting and imputation and their effects on estimates of food stamp program participation." *US Census Bureau Center for Economic Studies Paper* No. CES-WP-11-14 (2011).

Miguel, Edward, Shanker Satyanath, and Ernest Sergenti. "Economic shocks and civil conflict: An instrumental variables approach." *Journal of political Economy* 112, no. 4 (2004): 725-753.

Mikosch, Thomas. "Functional limit theorems for random quadratic forms." *Stochastic Processes and their Applications* 37, no. 1 (1991): 81-98.

Moller, Jesper, and Rasmus P. Waagepetersen. "An introduction to simulation-based inference for spatial point processes." In *Spatial statistics and computational methods*, pp. 143-198. Springer New York, 2003.

Moller, J., Pettitt, A.N., Reeves, R. and Berthelsen, K.K., 2006. An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika*, 93(2), pp.451-458.

Murphy, K.M. and Topel, R.H., 2002. Estimation and inference in two-step econometric models. *Journal of Business & Economic Statistics*, 20(1), pp.88-97.

Nunn, Nathan, and Diego Puga. "Ruggedness: The blessing of bad geography in Africa." *Review of Economics and Statistics* 94, no. 1 (2012): 20-36.

Phillips, Donald L., Jayne Dolph, and Danny Marks. "A comparison of geostatistical procedures for spatial analysis of precipitation in mountainous terrain." *Agricultural and Forest Meteorology* 58, no. 1 (1992): 119-141.

Piterbarg, Vladimir I. *Asymptotic methods in the theory of Gaussian processes and fields.* Vol. 148. American Mathematical Soc., 2012.

Rao, Suhasini Subba. *Fourier Based Statistics for Irregular Spaced Spatial Data.* Working

document. 2015.

Rao, Calyampudi Radhakrishna, and Jorgen Kleffe. *Estimation of variance components and applications*. Amsterdam: North-Holland, 1988.

Ripley, Brian D. "Neural networks and related methods for classification." *Journal of the Royal Statistical Society. Series B* (Methodological) (1994): 409-456.

Shah, Manisha, and Bryce Millett Steinberg. *Drought of opportunities: contemporaneous and long term impacts of rainfall shocks on human capital*. No. w19140. National Bureau of Economic Research, 2013.

Smola, Alex J., and Bernhard Schölkopf. *Learning with kernels*. GMD-Forschungszentrum Informationstechnik, 1998.

Stein, Michael L. *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media, 2012.

Stein, Michael L. "Asymptotically efficient prediction of a random field with a misspecified covariance function." *The Annals of Statistics* 16, no. 1 (1988): 55-63.

Stein, Michael L., Zhiyi Chi, and Leah J. Welty. "Approximating likelihoods for large spatial data sets." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 66, no. 2 (2004): 275-296.

Sweeting, T. J. "Uniform asymptotic normality of the maximum likelihood estimator." *The Annals of Statistics* (1980): 1375-1381.

Tabios, Guillermo Q., and Jose D. Salas. "A comparative analysis of techniques for spatial interpolation of precipitation." *Water Resource Bulleting* 28, no. 3 (1985): 365-380.

Van der Vaart, Aad W. *Asymptotic statistics*. Vol. 3. Cambridge university press, 2000.

Vapnik, Vladimir Naumovich. *Statistical learning theory*. Vol. 1. New York: Wiley, 1998.

Vecchia, Aldo V. "Estimation and model identification for continuous spatial processes." *Journal of the Royal Statistical Society. Series B (Methodological)* (1988): 297-312.

Watkins, A. J., and F. H. M. Al-Boutiahi. "On maximum likelihood estimation of parameters in incorrectly specified models of covariance for spatial data." *Mathematical Geology* 22, no. 2 (1990): 151-173.

White, Halbert. "Maximum likelihood estimation of misspecified models." *Econometrica* (1982): 1-25.

Williams, C.K. and Rasmussen, C.E., 2006. *Gaussian processes for machine learning*. the MIT Press, 2(3), p.4.

Wolfinger, R.E. and Rosenstone, S.J., 1980. *Who votes?* (Vol. 22). Yale University Press.

Yakowitz, S. J., and F. Szidarovszky. "A comparison of kriging with nonparametric regression methods." *Journal of Multivariate Analysis* 16, no. 1 (1985): 21-53.

Yoon, Ji Won, and Nial Friel. "Efficient model selection for probabilistic K nearest neighbour classification." *Neurocomputing* 149 (2015): 1098-1108.

Appendix A

Supplement to Chapter 2

A.1 Mixing Assumptions and Distribution Theory

In this subsection, I detail and discuss the assumption on spatial dependence required for the limit distribution theory and give the postponed proofs for the limit distribution theory of the quasi-maximum likelihood estimator with dependent data.

A.1.1 Mixing Assumptions

The distribution theory relies on the assumption that random variables "far enough" from each other are approximately independent. Useful measures to quantify such concepts of limiting independence can be borrowed from the theory of mixing random variables (Doukhan, 1994; White, 1984; Davidson, 1994; Billingsley, 1995).

For our purposes, the natural assumption is in term of spatial dependence.

I introduce some terminology (Doukhan, 1994; Davidson, 1994). Given a probability space $(\Omega, \mathcal{F}, P)$, let $\mathcal{G}$ and $\mathcal{H}$ be σ-subfields of $\mathcal{F}$, then the strong mixing coefficient is defined

$$\alpha(\mathcal{G}, \mathcal{H}) = \sup_{G \in \mathcal{G}, H \in \mathcal{H}} |P(G \cap H) - P(G) P(H)| .$$

Given some distance[1] function $d : \mathscr{D} \to \mathbb{R}_+$, $\mathscr{D} \subset \mathbb{R}^2$, I define

[1] For our purposes, it will be a Euclidian or geodesic distance.

$$\alpha_Y(k;u,v) = \sup\left\{\alpha_Y\left(G,H\right);\ d(G,H)\geq k,\ |G|\leq u,\ |H|\leq v\right\},$$

for all $G, H \subset \mathscr{D}$ and $u, v > 0$, where $|\cdot|$ denotes the area, $d(A,B) = \min\{d(a,b) : a \in A, b \in B\}$, and

$$\alpha_Y\left(G,H\right) = \alpha\left(\mathcal{X}_G, \mathcal{X}_H\right),$$

where $\mathcal{X}_F$ is the σ-algebra generated by $Y_F = \{Y_x : x \in F\}$, $F \subset \mathscr{D}$. See Bradley (1989, 1993) for a discussion of the importance of the bound on G and H in the definition of α_Y.

We say that the random field Y is strongly mixing (or α-mixing) if

$$\lim_{k\to\infty} \alpha_Y\left(k;u,v\right) = 0,$$

for all integers $u, v \geq 0$. I also define mixing sequences. If the random variables are indexed by an ordered set T, then another notion of mixing is defined for the random sequence $X = \{X_t\}_{t\in T}$. Let $\alpha_{X,k;u,v} = \sup\{\alpha_X\left(G,H\right)\}$ where the supremum is over G and H with $|G| \leq u$, $|H| \leq v$, and $g < h - k$ for $g \in G$ and $h \in H$. The process X is said to be strongly mixing (α-mixing) if

$$\lim_{k\to\infty} \alpha_{X,k;u,v} = 0$$

for all $u, v > 0$.

In both cases, we say that the process is mixing of size $-\varphi$ if $\alpha_m = O\left(\alpha_m^{-\varphi-\varepsilon}\right)$ for some $\varepsilon > 0$.

Although the spatial nature of our motivating analysis makes more natural and verifiable mixing assumptions on the random field, the proof technique calls for mixing assumptions on sequences. Fortunately, for any relabeling of $Y_T = (Y(t_1), ..., Y(t_n))'$ as $Y_S = (Y_1, ..., Y_n)$, treated as a mixing sequence, we have that as long as the distance between any two sites is bounded below, it obtains that $\lim_{k\to\infty} \alpha_{Y_T}\left(k;u,v\right) = 0$ implies $\lim_{k\to\infty} \alpha_{Y_S,k;u,v} = 0$, for all $u, v > 0$. Hence, this is accommodated by the increasing domain asymptotic set-up, which specifies precisely such a lower bound.

I comment on the implication of the strong mixing assumption in the case in which the

data generating process is in fact Gaussian. Piterbarg (2012) briefly explores the question, whilst Ibragimov and Rozanov (2012) investigate the topic more thoroughly. An unfortunate fact is that one cannot conveniently specify a sufficient condition for strong mixing in terms of the covariance of the process; short of assuming its zeroing from some moment on, one can impose any conditions on the limiting rate of decrease of the covariance, and it will always be possible to find a Gaussian stationary process with this rate of decrease of its covariance, but which is not strongly mixing.

The problem is that the condition of strong mixing requires the appearance of almost all frequencies in the spectrum, so that the process "is similar to" white noise. See also Kolmogorov and Rozanov (1960) and Ibragimov (1971) for further discussion.

Piterbarg (2012) endeavors to rehabilitate the mixing condition for Gaussian processes by showing that by discretizing the Gaussian process at any level of precision, a fast enough decreasing of the correlation will imply strong mixing of the discretized process. For our purposes, this is innocuous, because we can comfortably treat weather shocks at very far distances (e.g. Boston vs Hong Kong) as independent.

A.1.2 Distribution Theory for QMLE

Let $X_{n,i} = a_{ii}\left(Z_i^2 - E\left[Z_i^2\right]\right) + 2\sum_{j<i} a_{ij}\left(Z_i Z_j - E\left[Z_i Z_j\right]\right)$ and $\mathcal{F}_{n,i} = \sigma\left(Z_{n1}, ..., Z_{n,i}\right), 1 \leq i \leq k_n$. Suppose Z_i is α-mixing of size $-\varphi$. Further suppose that $\left\|Z_i^2\right\|_4 < \infty$. I give the proof of Theorem 2.

Proof of Lemma 2

First, $E\left[X_{n,i}\middle|\mathcal{F}_{n,i+m}\right] = X_{n,i}$, and thus $\left\|X_{n,i} - E\left[X_{n,i}\middle|\mathcal{F}_{n,i+m}\right]\right\|_2 = 0$.

Second, observe that

$$\left\|E\left[X_{n,i}\middle|\mathcal{F}_{n,i-m}\right]\right\|_2 \leq a_{ii}\left\|E\left[Z_i^2\middle|\mathcal{F}_{n,i-m}\right] - E\left[Z_i^2\right]\right\|_2 + 2\sum_{j<i} a_{ij}\left\|E\left[Z_i Z_j\middle|\mathcal{F}_{n,i-m}\right] - E\left[Z_i Z_j\right]\right\|_2.$$

I can bound each summand using moment inequalities for mixing variables. Note that the Z_i^2's are likewise α-mixing of size $-\varphi$. From (Davidson, 1994, theorem 14.2), I have that

$$\left\| E\left[Z_i^2 \middle| \mathcal{F}_{n,i-m} \right] - E\left[Z_i^2 \right] \right\|_2 \leq 6\alpha_m^{1/2} \left\| Z_i^2 \right\|_2 .$$

A bound for $\| E[Z_i Z_j | \mathcal{F}_{n,i-m}] \|_2$ obtains either when $\mathcal{F}_{n,i-m}$ contains little information about both i and j, and when i and j are far apart. Let $m' = \lfloor m/2 \rfloor$. The first case is $j - i + m \geq m'$. That is, both i and j are greater than the indices of the random variables generating $\mathcal{F}_{n,i-m}$ by at least m'. Let $k = i - j$, then $Z_i Z_j = Z_i Z_{i-k}$ is likewise α-mixing of size $-\varphi$. Since $\mathcal{F}_{i-m} \supset \mathcal{F}_{i-m'}$, I obtain

$$\begin{aligned}
\| E[Z_i Z_j | \mathcal{F}_{n,i-m}] - E[Z_i Z_j] \|_2^2 &= \left\| E\left[E\left[Z_i Z_j | \mathcal{F}_{n,i-m'} \right] \middle| \mathcal{F}_{n,i-m} \right] - E[Z_i Z_j] \right\|_2^2 \\
&= \left\| E\left[E\left[Z_i Z_j | \mathcal{F}_{n,i-m'} \right] - E[Z_i Z_j] \middle| \mathcal{F}_{n,i-m} \right] \right\|_2^2 \\
&\leq E\left[E\left[\left(E\left[Z_i Z_j | \mathcal{F}_{n,i-m'} \right] - E[Z_i Z_j] \right)^2 \middle| \mathcal{F}_{n,i-m} \right] \right] \\
&= E\left[\left(E\left[Z_i Z_j | \mathcal{F}_{n,i-m'} \right] - E[Z_i Z_j] \right)^2 \right] \\
&\leq \left(6\alpha_{m'}^{1/2} \| Z_i Z_j \|_2 \right)^2 \leq \left(6\alpha_{m'}^{1/2} \| Z_i \|_4 \| Z_j \|_4 \right)^2 .
\end{aligned}$$

In the second case, when $j - i + m < m'$, or equivalently $i - j > m'$, it is the weak dependence of Z_i with Z_j which allows to control the norm . Indeed, observe that

$$\begin{aligned}
\| E[Z_i Z_j | \mathcal{F}_{n,i-m}] - E[Z_i Z_j] \|_2^2 &= \left\| E\left[E\left[Z_i Z_j | \mathcal{F}_{n,i-m'} \right] \middle| \mathcal{F}_{n,i-m} \right] - E[Z_i Z_j] \right\|_2^2 \\
&= \left\| E\left[E\left[Z_i | \mathcal{F}_{n,i-m'} \right] Z_j - E[Z_i Z_j] \middle| \mathcal{F}_{n,i-m} \right] \right\| \\
&\leq E\left[E\left[\left(E\left[Z_i | \mathcal{F}_{n,i-m'} \right] Z_j - E[Z_i Z_j] \right)^2 \middle| \mathcal{F}_{n,i-m} \right] \right] \\
&= E\left[\left(E\left[Z_i | \mathcal{F}_{n,i-m'} \right] Z_j - E[Z_i Z_j] \right)^2 \right] \\
&\leq \left(2(2^{3/4} + 1)\alpha_{m'}^{1/2} \| Z_i \|_4 \| Z_j \|_4 \right)^2 ,
\end{aligned}$$

using the covariance inequality for mixing sequences. Thus

$$\| E[Z_i Z_j | \mathcal{F}_{n,i-m}] - E[Z_i Z_j] \|_2 \leq 6\alpha_{m'}^{1/2} \| Z_i \|_4^2$$

always. Consequently,

$$\begin{aligned}\left\|E\left[X_{n,i}\middle|\mathcal{F}_{n,i-m}\right]\right\|_2 &\leq a_{ii}6\alpha_m^{1/2}\left\|Z_i^2\right\|_2+\sum_{j<i}a_{ij}6\alpha_{m'}^{1/2}\left\|Z_i\right\|_4^2\\ &\lesssim \alpha_m^{1/2}\left\|Z_i\right\|_4^2\sum_{j\leq i}a_{ij}\\ &\leq \alpha_m^{1/2}iM=O(m^{-1/2}),\end{aligned}$$

where M is an implicitly defined constant in m. In particular, $a_{n,i}=iM$ are the mixingale magnitude indices. □

I move onto the central limit theorem for quadratic forms in mixing random variables. I obtain a general result showing that, upon conditions on the sequences of matrices $\{A_n\}$, a CLT for the quadratic forms $Z_n^TA_nZ_n$ obtains.

For each n, let $X_{n1},...,X_{nk_n}$ be independent with mean 0 and normalized to have variance 1. The proof of the mixingale central limit theorem relies on a Bernstein blocks strategy. The idea is to divide the data into large and small blocks, with a small block separating every two consecutive large blocks, so to treat the sum of entries within each large block as (asymptotically) independent observations. Let a_n and b_n be the sizes of the big and small blocks, respectively. Define big blocks

$$\chi_{ns}=\sum_{i=(s-1)(b_n+l_n)+1}^{sb_n}X_{ni},\ s=1,...,r_n,$$

where $r_n=\left\lfloor\frac{n}{b_n+l_n}\right\rfloor$. I define the asymptotic framework such that $b_n/l_n\to\infty$, $l_n/n\to 0$, $b_n/n\to 0$, and that b_n^2/n is bounded in probability.

This corresponds to a pure increasing domain asymptotic framework for the underlying spatial data. In applications to merged data, the sample size of both data sets must go to infinity for the conclusions of this theory to carry through.

The mixingale central limit theorem requires conditions on extremum statistics. For

mixingale magnitude indices a_{ni}, define

$$M_{ns} = \max_{(s-1)(b_n+l_n)+1 \leq i \leq sb_n} a_{ni},$$

the maximum magnitude index for big block s.

Technical conditions are imposed in terms of expectations taken in the bulk and tails of the underlying data generating process. Define

$$\delta_{ni}^{(1)} = E(X_{ni}^2 - 1)^2 \mathbf{1}_{\{|X_{ni}| > L_{ni}\}},\ \delta_{nij}^{(2)} = E(X_{ni}^2 - 1)\left(X_{ni}X_{nj} - \mu_{nij}\right)\mathbf{1}_{\{|X_{ni}| > L_{ni}\}},$$

$$\delta_{nij}^{(3)} = E\left(X_{ni}^2 - 1\right)\left(X_{nj}^2 - 1\right)\mathbf{1}_{\{|X_{ni}| > L, |X_{nj}| > L_{ni}\}\}},$$

where $\mu_{nij} = E[X_i X_j]\ \forall\ i, j$, and

$$\delta_{nij} = \begin{cases} \delta_{ni}^{(1)}, & \text{if } i = j \\ \delta_{nij}^{(2)} + \delta_{nij}^{(3)}, & \text{if } i \neq j \end{cases}.$$

Define also

$$\gamma_{nij}^{(1)} = E(X_{ni}^2 - 1)^2 (X_{nj}^2 - 1)^2 \mathbf{1}\{|X_{nd}| \leq L_{nd}, d = i, j\},$$

$$\gamma_{nijs}^{(2)} = E(X_{ni}^2 - 1)(X_{nj}^2 - 1)(X_{ni}X_{nl} - \mu_{nil})(X_{nj}X_{nl} - \mu_{njl})\mathbf{1}\{|X_{nd}| \leq L_{nd}, d = i, j, l, s\},$$

$$\gamma_{nijsrt}^{(3)} = E(X_{ni}X_{nl} - \mu_{nil})(X_{ni}X_{nr} - \mu_{nir})(X_{nj}X_{ns} - \mu_{njs})(X_{nj}X_{nt} - \mu_{njt})\mathbf{1}\{|X_{nd}| > L_{nd}, d = i, l, j, s, r, \}.$$

Theorem 3 Suppose that $Z_{n1}, ..., Z_{nk_n}$ are α-mixing of size 1, and let $X_{n1}, ..., X_{nk_n}$ and $\mathcal{F}_{n1}, ..., \mathcal{F}_{nk_n}$ be defined as in (1.15) and (1.16), respectively. Assume that A_n, $n = 1, 2, ...$, is symmetric and that all fourth-order products $Z_{ni}Z_{nj}Z_{nk}Z_{nl}$ are uniformly integrable. Assume that there are numbers $\{L_{ni}\}_{1 \leq i \leq k_n, n \geq 1}$ such that

$$\frac{1}{\sigma_n^2} \sum_{i,j}^{k_n} a_{nij}^2 \delta_{nij} \to 0 \tag{A.1}$$

$$\frac{1}{\sigma_n^2}\sum_{i=1}^{k_n}\left(\sum_{j=1}^{k_n}a_{nij}^{(1)}\gamma_{nij}^{(1)}+\sum_{i'=1}^{k_n}\sum_{j\neq i,j'\neq i'}\left(a_{nijls}^{(2)}\gamma_{nijls}^{(2)}+\sum_{r\neq i,t\neq j}a_{niljsrt}^{(3)}\gamma_{niljsrt}^{(3)}\right)\right)\to 0, \tag{A.2}$$

where $a_{nij}^{(1)}=a_{nii}^2a_{njj}^2$, $a_{nijls}^{(2)}=a_{nii}a_{njj}a_{nil}a_{njs}$, $a_{niljsrt}^{(3)}=a_{nil}a_{njs}a_{njr}a_{njt}$ and $\sigma_n^2=Var(\mathscr{X}_n^TA_n\mathscr{X}_n)$ with $\mathscr{X}_n^T=(X_{n1},...,X_{nk_n})$. Further assume that

$$\frac{\lambda_{max}\left(\left(A_n^0\right)^2\right)}{\sigma_n^2}\to 0, \tag{A.3}$$

where A_n^0 is equal to A_n with the diagonal terms replaced by zeros. Then

$$\frac{\mathscr{X}_n^TA_n\mathscr{X}_n-E\mathscr{X}_n^TA_n\mathscr{X}_n}{\sigma_n}\xrightarrow{L_2}N(0,1). \tag{A.4}$$

PROOF

By lemma 2, I know that $\{X_{ni},\mathcal{F}_{ni}\}$ is an L_2 mixingale of size $-1/2$. Let a_{ni}, $i=1,...,k_n$ denote its mixingale magnitude indices. In order to establish (4), it suffices to show that $X_{n1},...,X_{nk_n}$ satisfies the conditions of De Jong (1997), Theorem 1.

The distance between any two points Z_{ni}, and thus their dependence, is bounded. The dependence vanishes as the distance between the location corresponding to two points Z_{ni} and Z_{nj} gets bigger (or, eventually, as the difference in their indices gets arbitrarily large). Thus I know that the mixingale magnitude indices are bounded in n and i. In particular, from the condition on the relative size of big blocks, $b_n/n\to 0$, I know that $n^{-1/2}/b^{-1/2}\to 0$, which gives $a_{ni}/b_n^{-1/2}\to 0$. Likewise, $\frac{r_nb_n^2}{n^2b_n^{-1}}=O(1)$ must be bounded, and thus $\sum_{s=1}^{r_n}M_{ns}^2=O(b_n^{-1})$. By uniform integrability of the fourth-order products of Z_{ni}'s, I know that X_{ni}^2/a_{ni}^2 is uniformly integrable.

Write

$$\frac{\mathscr{X}_n^TA_n\mathscr{X}_n-E\mathscr{X}_n^TA_n\mathscr{X}_n}{\sigma_n}=\sum_{i=1}^{k_n}\xi_{ni}+\sum_{i=1}^{k_n}\eta_{ni}$$

where

$$\xi_{ni}=\frac{1}{\sigma_n}\left(a_{nii}U_{ni}+2\left(\sum_{j<i}a_{nij}U_{nij}\right)\right),$$

$$\eta_{ni} = \frac{1}{\sigma_n}\left(a_{nii}V_{ni} + 2\left(\sum_{j<i} a_{nij}V_{nij}\right)\right),$$

with

$$U_{ni} = \left(X_{ni}^2 - 1\right)\mathbf{1}_{\{|X_{ni}|\leq L_{ni}\}} - E\left(X_{ni}^2 - 1\right)\mathbf{1}_{\{|X_{ni}|\leq L_{ni}\}},\ V_{ni} = X_{ni}^2 - 1 - U_{ni},$$

$$U_{nij} = \left(X_{ni}X_{nj} - \mu_{nij}\right)\mathbf{1}_{\{|X_{nd}|\leq L_{nd}, d=i,j\}} - E\left(X_{ni}X_{nj} - \mu_{nij}\right)\mathbf{1}_{\{|X_{nd}|\leq L_{nd}, d=i,j\}},\ V_{nij} = X_{ni}X_{nj} - \mu_{nij} - U_{nij}.$$

By (1), I have that $\sum_{i=1}^{k_n}\eta_i \xrightarrow{L_2} 0$. Note that the statements for the mixingale magnitude indices of the untruncated mixingale hold for those of the truncated mixingale. Hence, it suffices to show that the array of mixingales ξ_{ni} satisfy the mixingale central limit theorem. That is, it suffices to establish that condition (12) from De Jong (1997) Theorem 1, is satisfied. This requires showing that the sum of the squared sums over big Bernstein blocks is 1, i.e.

$$\sum_{s=1}^{r_n}\mathcal{X}_{ns}^2 \xrightarrow{P} 1.$$

Note that $\sum_{s=1}^{r_n}\left(\sum_{i=(s-1)b_n+l_n+1}^{sb_n}\xi_{ni}^2\right) = \left(\sum_{i=1}^{k_n}\xi_{ni}\right)^2 - \sum_{i\nsim j}\xi_{ni}\xi_{nj}$, where $i \nsim j$ means that the i^{th} and j^{th} observations are not in the same big Bernstein block. By the mixing condition, and because $b_n \to \infty$, I can establish that $\sum_{i\nsim j}\xi_{ni}\xi_{nj} = o_p(1)$. Consider

$$\sum_{i\nsim j}\xi_{ni}\xi_{nj} = \frac{1}{\sigma_n^2}\sum_{i\nsim j}\left(a_{nii}U_{ni} + 2\left(\sum_{l<i}a_{nil}U_{nil}\right)\right)\left(a_{njj}U_{nj} + 2\left(\sum_{s<j}a_{njs}U_{njs}\right)\right)$$

$$= \frac{1}{\sigma_n^2}\sum_{i\nsim j}\left(a_{nii}a_{njj}U_{ni}U_{nj} + 2\left(\sum_{l<i}a_{nil}U_{nil}\right)a_{njj}U_{nj} + 2\left(\sum_{s<j}a_{njs}U_{njs}\right)a_{nii}U_{ni} + 4\left(\sum_{l<i}a_{nil}U_{nil}\right)\left(\sum_{s<j}a_{njs}U_{njs}\right)\right)$$

$$= \frac{4}{\sigma_n^2}\sum_{i\nsim j}\left(\sum_{l<i}a_{nil}U_{nil}\right)\left(\sum_{s<j}a_{njs}U_{njs}\right) + o_p(1)$$

$$= \frac{4}{\sigma_n^2}\sum_{i\nsim j}\left(B_{(i)}^T U_{(i)}\right)\left(B_{(j)}^T U_{(j)}\right)$$

$$= \frac{4}{\sigma_n^2} \sum_{i \nsim j} U_{(i)}^T B_{(i)}^T B_{(j)} U_{(j)}$$

$$= \frac{4}{\sigma_n^2} U^T \tilde{A}_0 U$$

where $B_{(l)} = \left(a_{nil} \mathbf{1}_{\{l<i\}}\right)_i$, , $B = (B_{(1)}^T, ..., , B_{(k)}^T)^T$, $U_{(l)} = (U_{nil})_i$, $U = \left(U_{(1)}, ..., U_{(k)}\right)$. Note that

$$\tilde{A}_0 = \text{vec}(B)\text{vec}(B)^T - (I \otimes \text{vec}(B))(I \otimes \text{vec}(B))^T,$$

where the second term removes the block-diagonal corresponding to entries in the same big Bernstein block. Now note that

$$\frac{1}{\sigma_n^4} E \left(U^T \tilde{A}_0 U \right)^2$$

$$= \frac{1}{\sigma_n^4} E \left(\sum_{i,j=1}^{k_n^2} \tilde{A}_{0ij} U_i U_j \right)^2$$

$$\leq \frac{2c_1}{\sigma_n^4} \left(\sum_{i,j=1}^{k_n^2} \tilde{A}_{0ij}^2 E\left(U_i^2 U_j^2\right) \right) \tag{A.5}$$

$$\leq \frac{2c_2}{\sigma_n^2} \left\| \tilde{A}_0 \right\|_F^2 \leq \frac{\lambda_{\max}\left(\tilde{A}_0\right)}{\sigma_n^2}$$

$$\leq \frac{1}{\sigma_n^2} \max_{i,s} \left\{ |a_{nis}| \cdot \sum_{\substack{j \nsim i \\ l<j}} |a_{jl}| \right\} + o_p(1) \tag{A.6}$$

$$\leq \frac{1}{\sigma_n^2} \|B\|_F^2$$

$$\leq \frac{c_3}{\sigma_n^2} \lambda_{\max}\left(BB^T\right)$$

$$\leq \frac{c_4}{\sigma_n^2} \left(\text{tr}\left(\left(B^T B\right)^2 \right) \right)^{1/2}$$

$$\leq c_5 \left(\frac{\lambda_{\max}\left((A_0)^2\right)}{\sigma_n^2} \right)^{1/2} \frac{\sqrt{\operatorname{tr}\left(B^T B\right)}}{\sigma_n} \tag{A.7}$$

where (5) follows from Rosenthal's inequality, (6) from Gersgorin's circle theorem, and (7) from lemma 5.1 of Jiang (1996). Then, since $\frac{\lambda_{\max}\left((A_0)^2\right)}{\sigma_n^2} \to 0$, I have that $\sum_{i \sim j} \xi_{ni}\xi_{nj} \to 0$.

Therefore it suffices to show that

$$\left(\sum_{i=1}^{k_n} \xi_{ni} \right)^2 \xrightarrow{P} 1.$$

Write

$$\left(\sum_{i=1}^{k_n} \xi_{ni} \right)^2 = \sum_{i,j=1}^{k_n} \xi_{ni}\xi_{nj} = \sum_{i=i}^{3} U_i + \sum_{i=i}^{3} V_i$$

where

$$U_1 = \frac{1}{\sigma_n^2} \sum_{i,j=1}^{k_n} a_{nii}a_{njj} \left(U_{ni}U_{nj} - EU_{ni}U_{nj} \right),$$

$$U_2 = \frac{2}{\sigma_n^2} \sum_{i,j=1}^{k_n} a_{nii} \left(\sum_{s<j} a_{njs}U_{ni}U_{njs} - E\left(\sum_{s<j} a_{njs}U_{ni}U_{njs} \right) \right),$$

$$U_3 = \frac{4}{\sigma_n^2} \sum_{i,j=1}^{k_n} \left(\left(\sum_{l<i} a_{nil}U_{nil} \right)^2 - E\left(\sum_{s<j} a_{njs}U_{njs} \right)^2 \right),$$

$$V_1 = \frac{1}{\sigma_n^2} \sum_{i=1}^{k_n} a_{nii}a_{njj}EU_{ni}U_{nj},$$

$$V_2 = \frac{2}{\sigma_n^2} \sum_{i,j=1}^{k_n} a_{nii}E\left(\sum_{s<j} a_{njs}U_{ni}U_{njs} \right)$$

$$V_3 = \frac{4}{\sigma_n^2} \sum_{i,j=1}^{k_n} E\left(\sum_{s<j} a_{njs}U_{njs} \right)^2.$$

By (2) I have that $U_i \xrightarrow{L_2} 0$ for $i = 1, 2, 3$. By (3) I then have that

$$V_1 = \frac{1}{\sigma_n^2} \sum_{i,j=1}^{k_n} a_{nii} a_{njj} Cov(X_{ni}^2, X_{nj}^2) + o_p(1),$$

$$V_2 = \frac{2}{\sigma_n^2} \sum_{i,j=1}^{k_n} \sum_{s<j} a_{nii} a_{njs} Cov\left(X_{nj} X_{ns}, X_{ni}^2\right) + o_p(1),$$

$$V_3 = \frac{4}{\sigma_n^2} \sum_{i,j=1}^{k_n} \sum_{l<i} \sum_{s<j} a_{nij} a_{nij'} Cov(X_{ni} X_{nl}, X_{nj} X_{ns}) + o_p(1)$$

Thus I have that

$$\sum_{s=1}^{r_n} \mathcal{X}_{ns}^2 = \frac{\sum_{iji'j'} a_{nij} a_{ni'j'} Cov\left(X_{nj} X_{ni}, X_{nj'} X_{ni'}\right) + o_p(1)}{\sigma_n^2} \xrightarrow{P} 1.$$

□

Remark The conditions of theorem 3 reduce to the conditions of theorem 5.2 of Jiang (1996) in the case of independent random variables.

Remark Note that the argument of Jiang (1996), theorem 5.1 applies here, and conditions (1), (2) and (3) can be replaced with

$$\inf_n (\min_i Var(X_{ni})) \wedge (\min_{i:a_{ii}\neq 0} Var(X_{ni}^2)) > 0,$$

$$\sup_n (\max_{1\leq i\leq n} EX_{ni}^2 \mathbf{1}_{\{|X_{ni}|>x\}}) \vee (\max_{i:a_{ii}\neq 0} EX_{ni}^4 \mathbf{1}_{\{|X_{ni}|>x\}}) \to 0,$$

as $x \to 0$, and

$$\frac{\lambda_{\max}\left((A_0)^2\right)}{\sigma_n^2} \to 0,$$

for a proper choice of L_n.

Distribution Theory for Robust Estimation

To keep with the literature on asymptotics for semivariograms, I denote the mixing coefficients slightly differently in this subsection. Define

$$\alpha(a;b) = \sup\left\{\tilde{\alpha}(T_1, T_2) : d(T_1, T_2) \geq a,\ T_1, T_2 \in \mathcal{S}_3(b)\right\},$$

where $\mathcal{S}_3(b) = \left\{ \bigcup_{i=1}^3 D_i : \sum_{i=1}^3 |D_i| \leq b \right\}$ is the collection of all disjoint unions of three cubes D_1, D_2, D_3 in $\mathbb{R}^2$, $\tilde{\alpha}(T_1, T_2) = \sup \{ |P(A \cap B) - P(A)P(B)| : A \in \sigma \langle R(s) | s \in T_1 \rangle, B \in \sigma \langle R(s) | s \in T_2 \rangle \}$, and $d(T_1, T_2) = \min \{ \|x_1 - x_2\| : x_1 \in T_1, x_2 \in T_2 \}$.

To simplify the exposition, following Lahiri (2003), I further assume that there exist a non-increasing function $\alpha_1(\cdot)$ with $\lim_{a \to \infty} \alpha_1(a) = 0$ and a nondecreasing function $g(\cdot)$ such that the strong-mixing coefficients $\alpha(a, b)$ satisfy

$$\alpha(a, b) \leq \alpha_1(a) g(b), \ a > 0, b > 0.$$

In order to obtain a central limit theorem for the minimum distance estimator, I must first obtain a central limit theorem for the statistics from which I want to minimize distance. The following lemma is a useful preliminary result.

Lemma A.1 *Let $\{Y(s) : s \in \mathbb{R}^d\}$ be a stationary random field with $EY(0) = 0$ and auto-covariance function $\sigma(s) = EY(0)Y(s)$, $s \in \mathbb{R}^d$. Suppose $E|Y(0)|^{2+\delta} < \infty$ for some $\delta > 0$. Suppose f is strictly greater than 0 and continuous on R_0. Assume that $\alpha(a, b) \leq Ca^{-3} \log(a) g(b)$, $a, b > 1$, for a bounded g. Then for $n/\lambda_n^2 \to C_1$,*

$$n^{-1/2} \sum_{i=1}^n Y(s_i) \xrightarrow{d} N \left(0, \sigma(0) + Q \cdot C_1 \cdot \int_{\mathbb{R}^d} \sigma(x) dx \right),$$

almost surely, where $Q = \int f^2(x) dx$ and $\sigma(x) = EY(0)Y(x)$.

Proof of Lemma A.1

The claim follows directly from Proposition 3.1 and Theorem 3.2 of Lahiri (2003). □

I can now give the central limit theorem for the objective function of the robust estimation optimization problem.

A key observation is that even while maintaining the pure-increasing asymptotic framework for stochastic designs (as opposed to positing one in which the observations also get denser), the bins of the nonparametric semivariogram, $N_n(\cdot)$, can both shrink and contain more observations as we move along the sequence in n. The set of pairs of observations which will be used to approximate the semivariogram evaluated at a certain distance d, the "bin", is defined by the

lower and upper bounds on the distance, $\underline{b}$ and $\overline{b}$. The nonparametric variogram is

$$g_n(d) = \frac{1}{|N_n(d)|} \sum_{(i,j)\in N_n(d)} (\hat{R}(x_i) - \hat{R}(x_j))^2,$$

and the bins $N_n(d)$ used to estimate the squared difference of pairs distance d apart, with sample size n, are defined

$$N_n(d) = \left\{(s_i, s_j) : 1 \leq i, j \leq n,\ \underline{b}_n(d) \leq d(s_i - s_j) \leq \overline{b}_n(d)\right\}.$$

It is assumed that $\underline{b}_n(d), \overline{b}_n(d) \to d$, pointwise. Furthermore, it is assumed that $|N_n(h)| = O(n)$.

Remark I prove a more general result than stated in the main body of the article. Let the mean of the random field be $x(s)^T\rho$ instead of the constant m, and replace the assumption $\lambda_n^2 \|\hat{m} - m\|_2^4 = o_p(1)$ by $\sup\{\|x(s) - x(s+h)\| : s \in \mathbb{R}\} \leq C(h) < \infty$ and $\lambda_n^2 \|\hat{\rho} - \rho\|_2^4 = o_p(1)$.

Proof of Theorem 6

Using the Cramér-Wold device, it suffices to show that $n^{1/2}a^T g(\phi_0) \xrightarrow{d} N\left(0, a^T\Sigma_g(\phi_0)a\right)$ as $n \to \infty$ for any $a \in \mathbb{R}^K$. Let

$$g_{1n} = n^{1/2} \sum_{k=1}^{K} a_k \left(\frac{1}{|N_n(h_k)|} \sum_{N_n(h_k)} (R(s_i) - R(s_j))^2 - 2\gamma(h_k; \theta_0) \right),$$

and

$$g_{2n} = n^{1/2} \sum_{k=1}^{K} a_k \left(\frac{1}{n} \sum_{i=1}^{n} (R(s_i) - R(s_i + h_k))^2 - 2\gamma(h_k; \theta_0) \right).$$

The strategy is to show that g_n is close to g_{1n}, which is close to g_{2n}, which satisfies the conditions of the central limit theorem obtained in Lemma A.1. Note that

$$|g_{1n} - g_n| \leq n^{1/2} \sum_{k=1}^{K} \frac{|a_k|}{|N_n(h_k)|} \sum_{N_n(h_k)} \left| \left(\hat{R}(s_i) - \hat{R}(s_j)\right)^2 - (R(s_i) - R(s_j))^2 \right|$$

$$n^{1/2} \sum_{k=1}^{K} \frac{|a_k|}{|N_n(h_k)|} \sum_{N_n(h_k)} \left| \left((\hat{\rho} - \rho)(x(s_i) - x(x_j)) - (R(s_i) - R(s_j))\right)^2 - (R(s_i) - R(s_j))^2 \right|$$

$$\leq C_1 n^{1/2} \sum_{k=1}^{K} |a_k| \left(\|\rho - \hat{\rho}\|_2^2 \, C(\bar{b}_n(h_k))^2 + 2 \|\rho - \hat{\rho}\|_2 \|E_{k,n}\|_2 \right)$$

where $E_{kn} = \sum_{N_n(h_k)} \frac{1}{|N_n(h_k)|} (x(s_i) - x(s_j)) (R(s_i) - R(s_j)) = O(n^{-1})$. Then under increasing-domain asymptotics and since $E \|\hat{\rho} - \rho\|_2^2 = O(\lambda_n^{-1})$, I obtain that $|g_{1n} - g_n| = o_p(1)$.

Since $\lambda_n \to \infty$ and $f > a > 0$, I know that $|J_{kn}| = \int_{J_{kn}} f(x)dx \to 0$ where $J_{kn} = R_0 \backslash (1 - h_{k,n}/\lambda_n) R_0$, for $k = 1, ..., K$, and therefore $|N_n(h_k)| = n(1 + o(1))$. Consequently, by mean square continuity, it follows that

$$E|g_{1n} - g_{2n}| \leq C_2 E \left| n^{1/2} \sum_{k=1}^{K} a_k \left(\sum_{(i,j) \in N_n(h_k)} \frac{1}{|N_n(h_k)|} (R(s_i) - R(s_j))^2 - \frac{1}{n} \sum_{i=1}^{n} (R(s_i) - R(s_i + h_k))^2 \right) \right|$$

$$\leq C_2 E \left| n^{1/2} \sum_{k=1}^{K} a_k \left(\sum_{(i,j) \in N_n(h_k)} \frac{1}{|N_n(h_k)|} (R(s_i) - R(s_j))^2 - \frac{1}{n} \sum_{i:(i,j) \in N_n(h_k)} (R(s_i) - R(s_i + h_k))^2 \right) \right|$$

$$+ C_2 E \left| n^{1/2} \sum_{k=1}^{K} a_k \frac{1}{n} \sum_{i:(i,j) \notin N_n(h_k)} (R(s_i) - R(s_i + h_k))^2 \right|$$

$$\leq C_2 n^{1/2} \sum_{k=1}^{K} |a_k| \left(\left| \frac{1}{|N_n(h_k)|} - \frac{1}{n} \right| E \left(\sum_{(i,j) \in N_n(h_k)} (R(s_i) - R(s_i + h_k))^2 \right) \right)$$

$$+ C_2 n^{1/2} \sum_{k=1}^{K} |a_k| \frac{1}{n} E \left(\sum_{i:(i,j) \notin N_n(h_k)} (R(s_i) - R(s_i + h_k))^2 \right) + o_p(1)$$

by mean square continuity of $E\left((R(s) - R(s'))^2\right)$, and is

$$\leq C_3 n^{1/2} \sum_{k=1}^{K} |a_k| \left(||N_n(h_k)| - n| \, n^{-2} E \left(\sum_{i=1}^{n} (R(s_i) - R(s_i + h_k))^2 \right) \right)$$

$$+ C_3 n^{1/2} \sum_{k=1}^{K} |a_k| \frac{1}{n} \left(E \left(\sum_{i:(i,j) \notin N_n(h_k)} (R(s_i) - R(s_i + h_k))^2 \right)^2 \right)^{1/2}$$

$$\leq C_4 n^{1/2} \sum_{k=1}^{K} |a_k| \left(||N_n(h_k)| - n| \, n^{-2} \left(E \left(\sum_{i=1}^{n} \left(R(s_i) - R(s_i + h_k) \right)^2 \right) \right)^{2} \right)^{1/2}$$

$$+C_5 n^{1/2} \sum_{k=1}^{K} |a_k| \frac{1}{n} \left(\sum_{i:(i,j) \notin N_n(h_k)} Cov \left(\sum_{i=1}^{n} \left(R(s_i) - R(s_i + h_k) \right)^2, \left(R(0) - R(0 + h_k) \right)^2 \right) \right)^{1/2}$$

$$\leq C \sum_{k=1}^{K} |a_k| \left(n^{-1} \, ||N_n(h_k)| - n| + |J_{nk}|^{1/2} \, n^{-1/2} \right)$$

which goes to 0, as $n \to \infty$.

Then it suffices to show that $g_{2n} \to N(0, a^T \Sigma_g(\phi_0) a)$, which follows from an application of Lemma A.1. □

With the asymptotic distribution of the statistic g_n in hand, I can call on Theorem 3.2 of Cressie and Lahiri (2002) and obtain Theorem 7 as an immediate corollary. I record the additional necessary assumptions under Assumption 1.

Assumption *A*.1 Suppose that

- *For any* $\epsilon > 0$, *there exists* $\delta > 0$ *such that* $\inf \left\{ \sum_{i=1}^{K} \left(2\gamma(h_i; \phi_1) - 2\gamma(h_i; \phi_2) \right)^2 : \|\phi_1 - \phi_2\| \geq \epsilon \right\} > \delta$,
- $\sup \left\{ \gamma(h; \phi) : h \in \mathbb{R}^2, \phi \in \Phi \right\} < \infty$, $\gamma(h; \phi)$ is continuously differentiable with respect to each entry of ϕ.
- $W(\phi)$ is positive definite for all $\phi \in \Phi$ and $\sup \left\{ \|W(\phi)\| + \left\| W(\phi)^{-1} \right\| : \phi \in \Phi \right\} < \infty$, and $W(\phi)$ is continuously differentiable with respect to each entry of ϕ.

A.1.3 Proof of Theorem 5

I give the proof of Theorem 5.

Proof of theorem 5

First observe that, by simple differentiation, I get

$$\begin{aligned}
\frac{\partial^4}{\partial\lambda_1 ... \partial\lambda_4} E\left[\langle\lambda, Y\rangle^4\right] &= \frac{\partial^3}{\partial\lambda_2 ... \partial\lambda_4} E\left[\frac{\partial}{\partial\lambda_1} \sum_{1\le i,j,k,l\le 4} \lambda_i\lambda_j\lambda_k\lambda_l Y_i Y_j Y_k Y_l\right] \\
&= \frac{\partial^3}{\partial\lambda_2 ... \partial\lambda_4} E\left[4Y_1 \sum_{1\le j,k,l\le 4} \lambda_i\lambda_j\lambda_k\lambda_l Y_j Y_k Y_l\right] \\
&= \cdots \\
&= 4! E\left[Y_1 Y_2 Y_3 Y_4\right],
\end{aligned}$$

where $\lambda = (\lambda_1, ..., \lambda_4)$. I obtain the identity

$$E\left[Y_1 Y_2 Y_3 Y_4\right] = \frac{1}{4!} \frac{\partial^4}{\partial\lambda_1 ... \partial\lambda_4} E\left[\langle\lambda, Y\rangle^4\right]. \tag{A.8}$$

By Pinsker's inequality, I know that $D_{TV}(F, P) \le \sqrt{\frac{1}{2}b}$, where $b = D_{KL}(F||P)$. Note that total variation can be expressed as

$$D_{TV}(F, P) = \frac{1}{2} \sup_{|h|\le 1} \left| \int h dP - \int h dF \right|,$$

where $h : \mathbb{R}^n \to \mathbb{R}$.

In order to apply the bound on $E\left[\langle\lambda, Y\rangle^4\right]$, I separate the integrand into a bounded function and a remainder whose integral is negligible:

$$E\left[\langle\lambda, Y\rangle^4\right] = E\left[\langle\lambda, Y\rangle^4 \mathbf{1}\{\langle\lambda, Y\rangle^4 \le T\}\right] + E\left[\langle\lambda, Y\rangle^4 \mathbf{1}\{\langle\lambda, Y\rangle^4 > T\}\right]$$

Since X is Gaussian and Y is sub-Gaussian, $\langle\lambda, X\rangle^4$ are $\langle\lambda, Y\rangle^4$ are both ν-subgaussian for some ν. Then I have that

$$\begin{aligned}
\left| E\left[\langle\lambda, Y\rangle^4 \mathbf{1}\{\langle\lambda, Y\rangle^4 > T_\epsilon\}\right] - E\left[\langle\lambda, X\rangle^4 \mathbf{1}\{\langle\lambda, X\rangle^4 > T_\epsilon\}\right] \right| &\le 4\sqrt{2}\nu^4 \Gamma(2, T_\epsilon) \\
&= o(T_\epsilon^2 e^{-T_\epsilon}),
\end{aligned}$$

where $\Gamma(\cdot, \cdot)$ is the incomplete Gamma function, and can be made arbitrarily small for a big enough T_ϵ.

Using $h(x) = \frac{1}{T_\epsilon} \langle \lambda, x \rangle^4 \mathbf{1}\{\langle \lambda, x \rangle^4 \leq T_\epsilon\}$, I have

$$\left| E\left[\langle \lambda, Y \rangle^4 \mathbf{1}\{\langle \lambda, Y \rangle^4 \leq T_\epsilon\} \right] - E\left[\langle \lambda, X \rangle^4 \mathbf{1}\{\langle \lambda, X \rangle^4 \leq T_\epsilon\} \right] \right| \leq T_\epsilon 2\sqrt{\frac{1}{2}b}.$$

That is, as $b \to 0$, $\left| E\left[\langle \lambda, Y \rangle^4 \right] - E\left[\langle \lambda, Y \rangle^4 \right] \right| \to 0$. To see how this controls the quantity of interest, recall the identity (8), and note that by Cauchy's integral formula, for $f_1 : \mathbb{C} \to \mathbb{C}$ defined by $f_1 : \lambda_1 \mapsto f(\lambda)$ for any given $(\lambda_2, \lambda_3, \lambda_4)$, where $f : \lambda \mapsto E\left[\langle \lambda, Y \rangle^4 \right] - E\left[\langle \lambda, Y \rangle^4 \right]$, for R_1 small enough we have

$$\frac{\partial f_1}{\partial \lambda_1}(\lambda_1) = \int_{C(\lambda_1, R_1)} \frac{f_1(z)}{(\lambda_1 - z)^2} dz.$$

Consequently, $\frac{\partial f_1}{\partial \lambda_1}(\lambda_1)$ can be made arbitrarily small (for a proper choice of ϵ), for any choice of λ_1. Likewise, for $f_2 : \lambda_2 \mapsto \frac{\partial f}{\partial \lambda_1}(\lambda)$

$$\frac{\partial^2 f_2}{\partial \lambda_1 \partial \lambda_2}(\lambda_2) = \int_{C(\lambda_2, R_2)} \frac{f_2(z)}{(\lambda_2 - z)^2} dz,$$

can be made arbitrarily small. Preceding iteratively, I find that

$$\frac{\partial^4}{\partial \lambda_1 ... \partial \lambda_4} E\left[\langle \lambda, Y \rangle^4 \right] - \frac{\partial^4}{\partial \lambda_1 ... \partial \lambda_4} E\left[\langle \lambda, X \rangle^4 \right]$$

can be made arbitrary small, which proves the claim. □

A.2 Approximation of the Likelihood

Exact calculation of the the likelihood, a computational task of order $O(n^3)$, can be prohibitively expensive. In order to lighten the computational load, approximations to the likelihood have been developed.

Some successful strategies have employed spectral domain tools to approximate the likelihood. These generally rely on the estimating the spatial periodogram with the fast Fourier transform. This works well when the data is on a lattice. However, despite modern advances (Fuentes, 2007; Bandyopadhyay, 2015; Rao, 2015), Fourier inference with irregular space data

Figure A.1: *Approximating the Conditional Densities*

Vecchia recommends keeping only the nearest points from from the condition set. Stein recommends keeping some farther away points to capture lower frequencies of the covariance. For misaligned data, I suggest as a rule of thumb to apply Stein's rule (indiscriminately on dependent and independent variables) for the conditional density of independent variables, and to apply Vecchia's rule (only keeping independent variables) for the conditional density of dependent variables.

is still largely an open problem. In particular, I do not know of any fast Fourier algorithms for irregular space data.

Common strategies are based on the idea that a density $f(z_1, z_2, ..., z_n)$ can always be represented as a product of the densities of conditionally independent variables,

$$f(z_1)f(z_2|z_1)\cdots f(z_n|z_1, z_2, ..., z_{n-1}), \tag{A.9}$$

and with a proper ordering of the data (which will rely on the position of the points), $f(z_j|z_{j-1}, ..., z_1)$ terms can be approximated by $f(z_j|z_{j-1}, ..., z_{j-m})$ when $m < j$, for a chosen m. For a small m, this will considerably alleviate the computational burden. Furthermore, if any two random variables whose index are very far apart are also far apart geographically, we may be confident that

$$f(z_j|z_{j-1}, ..., z_1) \approx f(z_j|z_{j-1}, ..., z_{j-m}).$$

This approach was first suggested by Vecchia (1988), who proposed approximating $f(z_j|z_{j-1}, ..., z_1)$ by conditioning only on the nearest neighbors of z_j among $z_{j-1}, ..., z_1$. Stein

et al. (2004) later developed on this approach, allowing for conditional likelihoods of vector variables, giving an equivalent approximation for the residual likelihood, and importantly remarking that including farther away variables may be very beneficial.

As one simple rule for including farther away observations in the conditioned upon set, Stein et al. (2004) suggest the following. To approximate $f(z_j|z_{j-1},...,z_1)$, keep the $m' < m$ nearest neighbors, plus $m - m'$ points whose order starting from z_j is equal to $m + \lfloor l(j-m-1)/(m-m') \rfloor$ for $l = m - m'$. In the examples they consider, Stein et al. (2004) find that, when the correlation does not die out too rapidly with distance and the m is big enough (> 15), picking m' to be as low as $0.5 \cdot m$ can bring about tangible benefits, especially for parameter estimation. Both the Stein and Vecchia rule are displayed figuratively in figure 16.

The situation at hand has more structure because we deal with conditional densities for random variables which are conceptually different objects. The likelihood for the data set $(y_1,...,y_{n_1},r_1^*,...,r_{n_2}^*)$, after reindexing, decomposes into conditional densities of the form $f(y_j|\{r_i^*,y_i;i<j\})$, for which it is important that the conditioned upon variables allow a good prediction of the value of R at the location of y_j, and conditional densities of the form $f(r_j^*|\{r_i^*,y_i;i<j\})$, for which it is important that the conditioned upon variables allow good estimation of the covariance parameters.

For misaligned data, I suggest as a rule of thumb to apply Stein's rule indiscriminately on dependent and independent variables for the conditional density of independent variables, and to apply Vecchia's rule but only keeping independent variables (discarding all dependent variables) for the conditional density of dependent variables, see Figure A.1. The intuition for this rule of thumb is that in estimating the conditional likelihood for r_j^*, distant outcome observations y_i will not be very informative if its neighboring rainfall measurements are included. Likewise, the conditional density for y_j is informed very little by other outcome variables if their neighboring rainfall measurements are included. This strategy has an immediate analogue for REML (See Stein et al., 2004).

A.3 Censored Gaussian Random Field

In many applications with rainfall data sets, the data may not be transformed, or at least not in such a way that a potential point mass at zero is done away with.

In order to deal with such an issue, I preconize the importance sampling method laid out by Stein (1991) for its simplicity and effectiveness. Consider the model (1.5) and let ϕ be the vector of parameters. Further suppose that $R(x^*)$ is not observed, and I only observe the censored version $\bar{R}(x^*) = R(x^*)_+$.

I want to find the maximum likelihood estimate ϕ given the observed data[2] $Y(x), \bar{R}(x^*)$. Suppose, for ease of notation, that x^* is ordered such that $\bar{R}(x^*) = \left(\bar{R}(x_1^*), \bar{R}(x_2^*)\right)$ with $\bar{R}_1 = \bar{R}(x_1^*) \geq 0$ and $\bar{R}_2 = \bar{R}(x_2^*) = 0$. Then the likelihood of ϕ given the observed data is

$$L\left(\phi; Y(x), \bar{R}(x^*)\right) = p_\phi\left(Y(x), \bar{R}(x_1^*)\right) \cdot \int_{(-\infty,0]^{m_2}} p_\phi\left(s \middle| Y(x), \bar{R}(x_1^*)\right) ds,$$

where $m_2 = |x_2^*|$. The computational difficulty is the estimation of the integral. I approximate the maximum likelihood with the maximand of the ratio

$$\frac{p_\phi\left(Y(x), \bar{R}(x_1^*)\right)}{p_{\phi_0}\left(Y(x), \bar{R}(x_1^*)\right)} \frac{\sum_{j=1}^N p_\phi\left(s_j \middle| Y(x), \bar{R}(x_1^*)\right) / h_\theta(s)}{\sum_{j=1}^N p_{\phi_0}\left(s_j \middle| Y(x), \bar{R}(x_1^*)\right) / h_{\theta_0}(s)}, \tag{A.10}$$

with the function h_θ specified below.

Generate the s_j's as follows. Denote $s_j = (s_{j1}, ..., s_{jm_2})$ and $s_j^q = (s_{j1}, ..., s_{jq})$ for $1 \leq q \leq m_2$. Let μ_{jq} denote the mean and σ_q the variance of s_{jq} conditional on the data $Y(x), \bar{R}(x_1^*)$ and the realizations of s_j^{q-1}. Then for $R_{jq} \sim U[0,1]$, let

$$s_{jq} = \mu_{jq} + \sigma_q \Phi^{-1}\left(R_q \Phi\left(\frac{-\mu_{qj}}{\sigma_q}\right)\right).$$

It follows that $h(s_j) = \Pi_{q=1}^{m_2} p(s_j^q | s_j^{q-1}, Y(x), \bar{R}(x_1^*)) / \Pi_{q=1}^{m_2} \Phi\left(\frac{-\mu_{qj}}{\sigma_q}\right) = p(s_j | Y(x), \bar{R}(x_1^*)) / \Pi_{q=1}^{m_2} \Phi\left(\frac{-\mu_{qj}}{\sigma_q}\right)$, and thus

[2] I omit dependence on X, other aligned variables should be thought of as included in Y.

$$\frac{p(s_j|Y(x),\bar{R}(x_1^*))}{h(s_j)} = d_j(\theta),$$

where $d_j = \Pi_{q=1}^{m_2}\Phi\left(\frac{-\mu_{qj}}{\sigma_q}\right)$. Hence, (10) takes on the simple expression

$$\frac{p_\phi\left(Y(x),\bar{R}(x_1^*)\right)}{p_{\phi_0}\left(Y(x),\bar{R}(x_1^*)\right)}\frac{\sum_{j=1}^N d_j(\theta)}{\sum_{j=1}^N d_j(\theta_0)}. \tag{A.11}$$

The quantity $\sum_{j=1}^N d_j(\theta)$ tends to have high variance. But by using the ratio with $\sum_{j=1}^N d_j(\theta_0)$ and the same common random numbers (the R_{qj}'s) for all θ's and θ_0, I induce strong positive correlation between $\sum_{j=1}^N d_j(\theta)$ and $\sum_{j=1}^N d_j(\theta_0)$ for θ near the picked θ_0. Consequently, although $\sum_{j=1}^N d_j(\theta)$ and $\sum_{j=1}^N d_j(\theta_0)$ may not be good estimates of $\sum_{j=1}^N p_\phi\left(s_j|\,Y(x),\bar{R}(x_1^*)\right)/h_\theta(s)$ and $\sum_{j=1}^N p_{\phi_0}\left(s_j|\,Y(x),\bar{R}(x_1^*)\right)/h_{\theta_0}(s)$, respectively, (11) is a good estimate of (10).

A.4 Additional Tables and Figures

I give summary statistics tables.

Table A.1: *Summary Statistics for Women*

	Mean	Deviation	Minimum	Median	Maximum
Self-reported health status very good (indicator)	0.080	0.272	0	0	1
Self-reported health status poor/very poor (indicator)	0.116	0.320	0	0	1
Ln (lung capacity)	5.609	0.223	4.248	5.623	6.391
Height (centimeters)	150.649	5.375	110.600	150.600	169.800
Days absent due to illness (last 4 weeks)	1.155	3.096	0	0	28
Years of schooling	6.679	4.376	0	6	16
Ln (expenditures per capita in household)	12.640	0.952	9.208	12.562	17.338
Ln (total assets per capita in household)	14.644	1.664	0	14.748	19.612
Owns television (indicator)	0.647	0.478	0	1	1
Owns refrigerator (indicator)	0.309	0.462	0	0	1
Owns private toilet with septic tank (indicator)	0.478	0.500	0	0	1
Owns stove (indicator)	0.657	0.475	0	1	1

Table A.2: *Summary Statistics for Men*

	Mean	Deviation	Minimum	Median	Maximum
Self-reported health status very good (indicator)	0.089	0.285	0	0	1
Self-reported health status poor/very poor (indicator)	0.090	0.286	0	0	1
Ln (lung capacity)	6.000	0.238	4.700	6.024	6.697
Height (centimeters)	162.175	5.961	130.600	162.200	183.400
Days absent due to illness (last 4 weeks)	1.042	3.135	0	0	28
Years of schooling	8.036	4.383	0	8	16
Ln (expenditures per capita in household)	12.709	0.945	9.208	12.630	17.637
Ln (total assets per capita in household)	14.577	1.796	0	14.721	19.186
Owns television (indicator)	0.630	0.483	0	1	1
Owns refrigerator (indicator)	0.297	0.457	0	0	1
Owns private toilet with septic tank (indicator)	0.475	0.499	0	0	1
Owns stove (indicator)	0.668	0.471	0	1	1

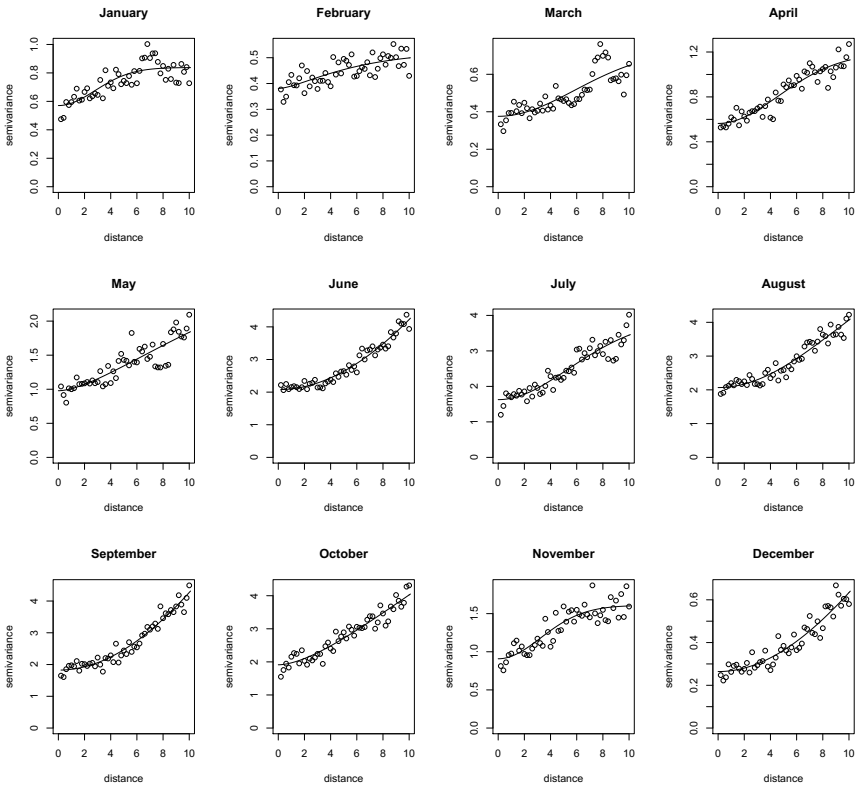

Figure A.2 1972 rainfall data, monthly sample (dots) and fitted (line) semivariogram. Note that the scales differ on the vertical axes.

Appendix B

Supplement to Chapter 3

B.1 Imputation with Support Vector Machines

The original motivation for employing the k-nearest neighbors methodology was to borrow from machine learning some scalable methods optimized for prediction. Support vector machines (SVM) may have seemed a natural choice for the imputation method and indeed make for a natural benchmark.

The linear support vector machines[1] decision rule is defined

$$T_{\text{SVM}} = \text{sign}\left(\hat{\omega}^T x\right),$$

and the coefficient $\hat{\omega}$ solves the problem

$$\min_{\omega} \frac{1}{n} \sum_{i=1}^{n} \left(1 - Y_i \left(\omega^T X_i\right)\right)_+ + \lambda \left\|\omega\right\|_2^2$$

where λ is a tuning parameter (usually selected via cross-validation), $(a)_+ = \max\{0, a\}$, and the outcome variable is encoded $Y_i \in \{-1, 1\}$ for all $i = 1, ..., n$.

SVM does not easily embed in a probabilistic framework, hence making it difficult to use that method while allowing the outcome variable to inform imputation.[2] One might however

[1]The linear version is used here for ease of exposition. However, due to the "kernel trick" and the representer theorem, this is done without loss of generality. See Smola & Schölkopf (1998).

[2]Williams & Rasmussen (2006) develop on this. The hinge loss function cannot be considered as a negative

wonder whether imputation with SVM on its own may make for superior imputation and estimation, subsequently, of the regression coefficient of interest. To investigate that question, I analyze the data by first imputing the missing covariate using SVM, and then estimating the regression function, correcting for the attenuation bias due to the uncertainty of the imputation.

Support vector machines provide us with a function SVM_λ:$\mathscr{D}_2 \mapsto T_{\text{SVM}}(\cdot)$. That is, it provides generated categorical regressors $T_{\text{SVM}}(x_i)$ that approximate the unobserved $T_{\text{reg},i}$'s.

A desirable feature of SVM is that, like KNN, it is Fisher consistent. If the class of functions over which we optimize includes the Bayes rule T_{Bayes}, then T_{SVM} will consistently estimate it (Lin, 2002). More general results are available for classifiers which are Fisher consistent, yet do not require the estimation of $P(T = 1|X = x)$ (Stone, 1977; Devroye et al., 1996).

Imputation with SVM is hoped to be good, but is nevertheless expected to be imperfect. Hence, attenuation bias due to misclassification ought to be attended to. The problem of attenuation bias has received attention in the econometrics literature, and correction terms are available that can be computed, as long as the misclassification probabilities can be properly quantified.

The form of the bias correction term is obtained as follows. Note that, using the over-parametrization restriction $\tau_0 = 0$,

$$\begin{aligned}\tau_{SVM,t} &= \mathbb{E}[Y|X, T_{SVM} = t] - \mathbb{E}[Y|X, T_{SVM} = 0] \\ &= \sum_{j=2}^{|C|} \tau_j \left(P\left(T = j \,|T_{\text{SVM}} = t\right) - P\left(T = j \,|T_{\text{SVM}} = 0\right)\right),\ t = 2, ..., |C|.\end{aligned}$$

If we let $\tau_{\text{SVM}} = (\tau_{\text{SVM},2}, ..., \tau_{\text{SVM},|C|})$ and $\tau = (\tau_2, ..., \tau_{|C|})$, the above can be rewritten

$$\tau_{\text{SVM}} = P\tau,$$

where $P = \left\{P\left(T = j \,|T_{\text{SVM}} = i\right) - P\left(T = j \,|T_{\text{SVM}} = 0\right)\right\}_{2 \le i,j \le |C|}$. In particular, if we can obtain an estimate of P, then we can provide an estimate of $\tau = P^{-1}\tau_{\text{SVM}}$.

---

loglikelihood. In particular, if we tried to define the density $V(f) = 1/Z \cdot e^{-C(1-f)_+}$, upon computing the normalization constant we would find that $Z = e^{-C(1-f)_+} + e^{-C(1+f)_+}$ depends on f.

Note that, by Bayes rule,

$$P\left(T = j \,|T_{\text{SVM}} = i\right) = \frac{b_{ij} r_j^*}{r_i},$$

where $b_{ij} = P\left(T_{\text{SVM}} = i \,|T = j\right)$, $r_j^* = P(T = j)$, and $r_i = P(T_{\text{SVM}} = i)$. Note that r_i has an immediate sample estimate $\hat{r}_i$.

The probability estimates b_{ij} can be obtained either using a holdout set or via cross-validation. I expand on this below, and suggest estimation via leave-one-out bootstrap.[3]

Finally, the r_j^*'s can be given in terms of r_i's and b_{ij}'s. Indeed, consider

$$\begin{aligned} r_i = P\left(T_{\text{SVM}} = i\right) &= \sum_{j=1}^{|C|} P\left(T_{\text{SVM}} = i \,|T = j\right) P\left(T = j\right) \\ &= \sum_{j=1}^{|C|} b_{ij} r_j^*. \end{aligned}$$

Letting $r = (r_1, ..., r_{|C|})$, $r^* = (r_1^*, ..., r_{|C|}^*)$ and $B = \{b_{ij}\}_{1 \leq i,j \leq |C|}$, we have that

$$r = Br^*,$$

where r, r^* and the columns of B satisfy a sum-to-one constraint, hence the inverse problem of solving for r^* is well-posed.

In the binary case, the above derivation produces the correction term given by Lewbel (2007), that is

---

[3] As in the binary case, the challenge is to estimate the misclassification probability of the estimated rule. Decomposing into sample uncertainty and asymptotic error suggests an estimation strategy. Consider

$$\begin{aligned} b_{ij} = P\left(T = i \,|T^* = j\right) &= \sum_{k=1}^{l} P\left(T = i \,|T^\infty = k, T^* = j\right) P\left(T^\infty = k \,|T^* = j\right) \\ &= \sum_{k=1}^{l} P\left(T = i \,|T^\infty = k\right) P\left(T^\infty = k \,|T^* = j\right), \end{aligned}$$

assuming $P\left(T = i \,|T^\infty = k\right) = P\left(T = i \,|T^\infty = k, T^* = j\right)$for all $i, k, j \in C$. For support vector machines, the piece $P\left(T = i \,|T^\infty = k\right)$ can be estimated using the central limit theorem from Pouliot (2015). Generally speaking, it can be evaluated using a CLT or the bootstrap.

For k-nearest neighbors, $P\left(T^\infty = k \,|T = j\right)$ can be estimated in sample using the asymptotic expansion for the asymptotic risk. For support vector machines, an important question is how to estimate $P\left(T^\infty = k \,|T^* = j\right)$ in sample and, short of doing so, how reliable is the naive approximation simply using in-sample misclassification? Can we guarantee it is conservative?

$$\tau_{\text{SVM}} = M \cdot \tau,$$

where

$$M = \frac{1}{1-2b}\left(1 - \frac{(1-b)b}{a} - \frac{(1-b)b}{1-a}\right),$$

$$\begin{aligned} a &= P\left(T_{\text{SVM}} = 1\right), \\ b &= P\left(T_{\text{SVM}} \neq T\right). \end{aligned}$$

B.1.1 Estimation of the Probability of Imputation Error

Applications at the intersection of big data and econometrics require a statistical assessment of risk which is usually not of immediate relevance in machine learning applications.

In typical machine learning applications, an estimate of risk is required in order to calibrate tuning parameters. There, a good but rough estimate of risk may suffice. However, in applied economics, and regression analysis with imputed categorical variables in particular, there is a need for accurate point estimates -as well as measures of uncertainty- of the risk of the estimated decision rule.

The estimation of the probability of imputation error involves the evaluation of out-of-sample performance, hence a natural approach and the one preconized here is to use k-fold cross-validation. This leaves us with the task of choosing the tuning parameter k. A large value of k, hence one for which estimators in the cross-validation exercise will be based on data sets the size of which is only one observation smaller than that of the estimator whose performance we are assessing, will be nearly unbiased, making leave-one-out cross-validation ($k = n$) a seemingly natural choice.

Since cross-validation may produce an estimate of risk with high variance, a smoothing strategy using the bootstrap and trading off some bias for a smaller variance will be suggested. This strategy has the further benefit of providing standard errors for the estimated risk (which will allow us to produce standard errors for regression coefficients that account, albeit conservatively, for the uncertainty in the estimated bias correction term).

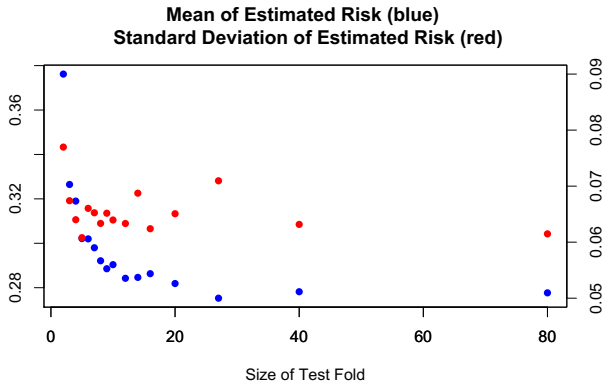

Figure B.1: *Mean Squared Errors*

The left axis corresponds to the blue points and gives the estimated MSE averaged over 40 simulations. The right-hand side axis corresponds to the red points and gives the standard deviation of the estimated MSE over the simulation draws.

The decrease in bias as the number of folds increases is salient in Figure 1, which presents output from a simulation calibrated to mimic the CPS data investigated in section 5. The intuition is straightforward: we wish to estimate R_n, the risk of the decision loss estimated with n data points, and we wish to separate the data into a test and sample set so to have some out-of-sample points (the test set) on which to evaluate the predictions and thus measure the error rate. In k-fold cross-validation, the error rate being estimated is $R_{n-\lfloor n/k \rfloor}$. This quantity should be expected to decrease in k since, for a greater k, it will correspond to the expected error rate of a more precise decision function (because estimated on more data).

The usual "no free lunch" intuition would suggest that this decrease in bias will come at the price of an increase in variance, i.e. that there is a bias-variance tradeoff as we increase the number of folds k. However, the intuition for the change in variance of the estimate of risk, as we increase k, is not entirely straightforward. There is, so to speak, a "variance-variance" trade-off: as we increase the number of folds k, we average over fewer and fewer loss functions, each of them being a function of more and more observations.[4] Importantly, the way this

[4]This is assuming that the estimated loss is the average over each of the k folds of each of the $\frac{n!}{(n/k)!^k k!}$ ways to divide the data into k folds. If we didn't average over choices of the folds, it would not be so surprising to see that the variance decreases as k increases; that it does, in some cases, even under this "complete averaging" is the interesting phenomenon.

tradeoff operates depends on the loss (and underlying method) that is being estimated,[5] and so the proper choice of k must be investigated on a case-by-case basis.

Another important thing to notice about Figure 1 is that the variance of the estimated risk is high. This invites the use of smoothing strategies for estimating the risk. Efron and Tibshirani (1997) suggest bootstrapping the training data as follows. We evaluate the risk of our estimated decision function

$$E_{(y_0,x_0)}\left[Q\left(y_0, T_{\text{SVM}}(x_0)\right)\right],$$

where $Q(y,T) = \mathbf{1}\{y \neq T\}$. Leave-one-out cross-validation uses the estimate

$$\frac{1}{n}\sum_{i=1}^{n} Q\left(T_i, T_{\text{SVM}}(x_i; \mathcal{D}_{(-i)})\right),$$

where $T_{\text{SVM}}(\cdot; \mathcal{D})$ is the SVM decision rule fitted with data $\mathcal{D}$, and

$$\mathcal{D}_{(-i)} = \{(x_1, y_1), ..., (x_{i-1}, y_{i-1}), (x_{i+1}, y_{i+1}), ..., (x_n, y_n)\}.$$

A leave-one-out bootstrap uses the estimate

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{(-i)} Q\left(T_i, T_{\text{SVM}}(x_i; \tilde{\mathcal{D}})\right),$$

where $\mathbb{E}_{(-i)}$ is the expectation corresponding to $\mathbb{P}_{(-i)}$, the empirical distribution itself corresponding $\mathcal{D}_{(-i)}$.

Efron and Tibshirani (1997) suggest a slightly modified, more computationally economical version of the leave-one-out bootstrap. Let $\mathcal{D}^{*1}, \mathcal{D}^{*2}, ..., \mathcal{D}^{*B}$ be bootstrap resamplings, let N_i^b be the number of times the i^{th} observation is drawn in the b^{th} bootstrap resampling, and let

$$I_i^b = \begin{cases} 1 & \text{if } N_i^b = 0 \\ 0 & \text{if } N_i^b > 0 \end{cases}.$$

[5] For least-squares regression and loss given by mean-squared error of the predicted dependent variable, for instance, the variance of the estimated risk can decrease noticeably as k increases, so in that case we are in fact given an allowance along with our lunch.

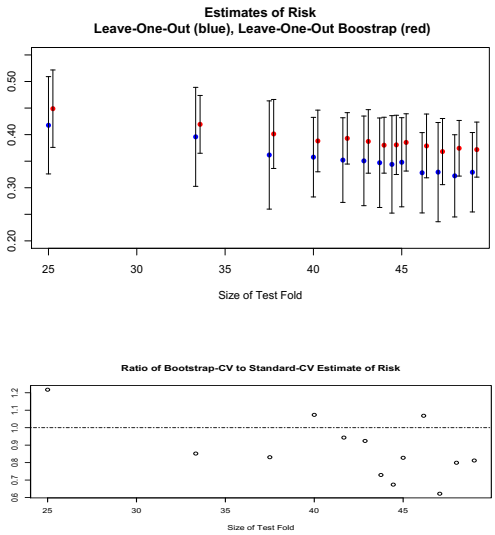

Figure B.2: *Smoothed Estimates*

(top) the average risk estimate from the leave-one-out cross-validation and from $\widehat{\text{Err}}^{(1)}$, the brackets give their standard deviations. *(bottom)* the ratio of the MSE of $\widehat{\text{Err}}^{(1)}$ to that of leave-one-out cross-validation.

Let $Q_i^b = Q\left(T_i, T_{\text{SVM}}(x_i; \mathcal{D}^{*b})\right)$, and define

$$\widehat{\text{Err}}^{(1)} = \frac{1}{n}\sum_{i=1}^{n} \hat{E}_i \text{ where } \hat{E}_i = \sum_b I_i^b Q_i^b / \sum_b I_i^b.$$

It is more economical because for every bootstrap resampling, all points not drawn in the bootstrap resampling are used in the test set.

The variance attenuation from smoothing comes at some cost in bias because of "underfitting"; Efron (1983) argues that by bootstrapping the training set, we "increase the distance" between the test points and the training points.[6] The rule of thumb (Efron, 1983) is that we should expect a bias on the order of that obtained when cross-validating with $k = n/2$. Hence, according to the (admittedly anecdotical) simulation data presented in Figure 6, even a mild reduction in variance would make for a favorable bias variance tradeoff.

Figure 7 plots the mean and standard deviation of the empirical risk for different choices of

[6]It would be convenient to use only the "closest" point not drawn in the bootstrap resampling as the test set, were there a reasonable and computationally affordable way to do this.

fold sizes. We can see both the increase in bias and reduction in variance across k's. To assess the tradeoff, we plot the ratio of mean-squared errors and observe that the mean-squared error of the smoothed estimate of risk tends to be lower.

B.1.2 Standard Errors for the Estimate of Risk

An attractive feature of $\widehat{\text{Err}}^{(1)}$ is that standard errors are available for it. Defining

$$\hat{D}_i = \frac{1}{n}\frac{\partial}{\partial \epsilon}\left.\frac{\partial \widehat{\text{Err}}^{(1)}(\hat{F}_{\epsilon,i})}{\partial \epsilon}\right|_{\epsilon=0},$$

where

$$\hat{F}_{\epsilon,i} = \begin{cases} \frac{1-\epsilon}{n} + \epsilon & \text{on } x_i \\ \frac{1-\epsilon}{n} & \text{on } x_j,\ j \neq i \end{cases},$$

Efron & Tibshirani (1997) suggest estimating standard errors using "delta method after bootstrap", that is

$$\widehat{SE}_{\text{del}} = \left(\sum_{i=1}^{n} \hat{D}_i^2\right)^{1/2}.$$

Efron and Tibshirani (1995) had worked out[7]

$$\hat{D}_i = \left(2 + \frac{1}{n-1}\right)\frac{\hat{E}_i - \widehat{\text{Err}}^{(1)}}{n} + e_n \hat{C}_i,$$

where $\hat{C}_i = \frac{1}{B}\sum_{b=1}^{B}(N_i^b - 1)q_i^b$ and $e_n = (1 - 1/n)^{-n}$.

This is important because it allows us to give standard errors, albeit conservative ones, that account for the sampling variation in $\hat{M}^{-1}$, which enters as a multiplicative constant in the estimated coefficient of interest $\tau = \hat{M}^{-1}\tau_{SVM}$.

By monotonicity, confidence intervals for the estimate of the risk of the sample decision rule immediately yield confidence intervals for $\hat{M}^{-1}$. Define the confidence interval for M as

$$M \in \left(\underline{\hat{M}}_\alpha, \overline{\hat{M}}_\alpha\right) \text{ with prob } 1 - \alpha.$$

Let $(\hat{\tau}_{SVM} \pm C_\alpha(M))$ be the size α confidence interval for τ (conditional on M). Then, using

[7]Presumably, their method could be used to get standard errors directly for $\hat{M}^{-1}$ or τ by working out the corresponding derivative.

the Bonferroni correction,

$$\tau \in \left(\overline{\hat{M}}_{\alpha/2}^{-1} \tau_{SVM} - C_{\alpha/2}(\underline{\hat{M}}_{\alpha/2}), \underline{\hat{M}}_{\alpha/2}^{-1} \tau_{SVM} - C_{\alpha/2}(\underline{\hat{M}}_{\alpha/2}) \right)$$

with probability greater than $1 - \alpha$.

B.2 Additional Tables and Figures

Table B.1: *Description of the Control Variables in the Second-Stage Regression*

Statistic	Mean	St. Dev.	Min	Max
Age	45.188	16.736	15	90
Sex	1.517	0.500	1	2
Black/Negro	0.089	0.285	0	1
American Indian/Aleut/Eskimo	0.012	0.107	0	1
Asian or Pacific Islander	0.013	0.113	0	1
Asian only	0.015	0.121	0	1
Hawaiian/Pacific Islander only	0.001	0.030	0	1
White-Black	0.001	0.023	0	1
White-American Indian	0.003	0.058	0	1
White-Asian	0.001	0.025	0	1
White-Hawaiian/Pacific Islander	0.0002	0.016	0	1
Black-American Indian	0.0003	0.017	0	1
Black-Asian	0.00003	0.006	0	1
Black-Hawaiian/Pacific Islander	0.00001	0.004	0	1
American Indian-Asian	0.00001	0.003	0	1
Asian-Hawaiian/Pacific Islander	0.0002	0.015	0	1
White-Black-American Indian	0.0001	0.011	0	1
White-Black-Asian	0.00000	0.002	0	1
White-American Indian-Asian	0.00001	0.003	0	1
White-Asian-Hawaiian/Pacific Islander	0.0002	0.013	0	1
White-Black-American Indian-Asian	0.00000	0.001	0	1
Two or three races, unspecified	0.0001	0.011	0	1
Four or five races, unspecified	0.00002	0.005	0	1
Born abroad of American parents	0.008	0.090	0	1
Naturalized citizen	0.047	0.212	0	1
Not a citizen	0.017	0.128	0	1
Both parents native-born	0.841	0.366	0	1
Father foreign, mother native	0.024	0.152	0	1
Mother foreign, father native	0.019	0.135	0	1
Both parents foreign	0.038	0.190	0	1
Foreign born	0.075	0.264	0	1
Mexican	0.014	0.117	0	1
Puerto Rican	0.009	0.094	0	1
Cuban	0.003	0.057	0	1
Dominican	0.007	0.082	0	1
Central/South American	0.009	0.094	0	1
Grades 1, 2, 3, or 4	0.005	0.068	0	1
Grades 5 or 6	0.009	0.095	0	1
Grades 7 or 8	0.025	0.157	0	1
Grade 9	0.018	0.133	0	1
Grade 10	0.028	0.164	0	1
Grade 11	0.033	0.179	0	1
12th grade, no diploma	0.012	0.110	0	1
High school diploma or equivalent	0.326	0.469	0	1
Some college but no degree	0.198	0.399	0	1
Associate's degree, occupational/vocational program	0.046	0.210	0	1
Associate's degree, academic program	0.039	0.193	0	1

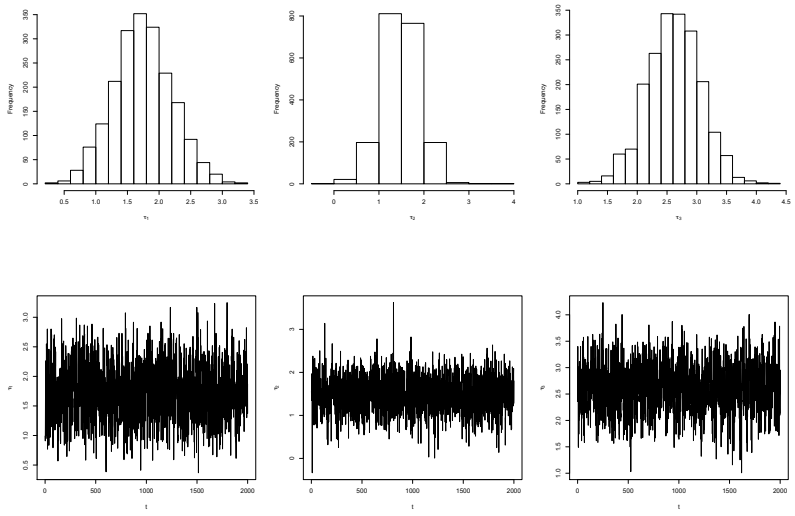

Figure B.3 MCMC output for empirical Bayesian KNN with fixed $(\hat{\beta}, \hat{k})$ obtained as MAP estimates of MCMC procedure of section 2. *(Top)* Posterior distribution of τ. *(Bottom)* Time series of draws from Gibbs sampler.

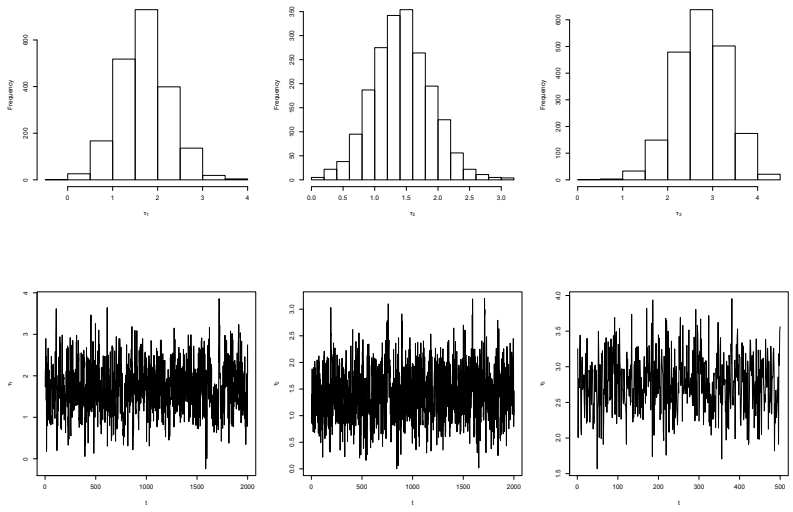

Figure B.4 MCMC output for empirical Bayesian KNN where $\mathbf{T}_{\text{knn}}$ is treated as a predicted value. *(Top)* Posterior distribution of τ. *(Bottom)* Time series of draws from Gibbs sampler.

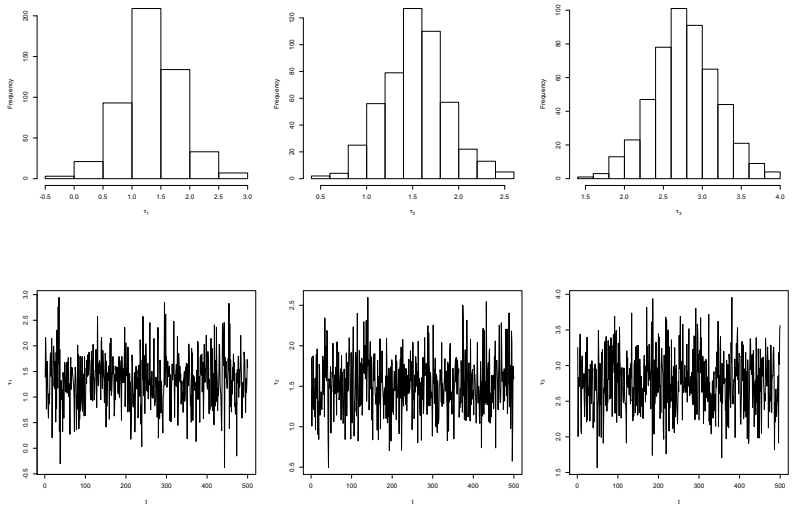

Figure B.5 MCMC output for empirical Bayesian KNN where $\mathbf{T}_{\text{reg}}$ is treated as missing data. *(Top)* Posterior distribution of τ. *(Bottom)* Time series of draws from Gibbs sampler.