Spring 5-15-2018

Kriging and Non Parametric Regrssion

Chamila Ranaweera
ranawera@siu.edu

KRIGING AND NONPARAMETRIC REGRESSION

by

Chamila K. Ranaweera

B.S.,University of Peradeniya, 2006

M.S.,Southern Illinois University Carbondale, 2015

A Research Paper
Submitted in Partial Fulfillment of the Requirements for the
Master of Science

Department of Mathematics
in the Graduate School
Southern Illinois University Carbondale
May, 2018

RESEARCH PAPER APPROVAL

KRIGING AND NONPARAMETRIC REGRESSION

by

Chamila K. Ranaweera

A Research Paper Submitted in Partial

Fulfillment of the Requirements

for the Degree of

Master of Science

in the field of Mathematics

Approved by:

David J. Olive

Yaser Samadi

Kwangho Choiy

Graduate School
Southern Illinois University Carbondale
November 1 2017

AN ABSTRACT OF THE RESEARCH PAPER OF

CHAMILA K.RANAWEERA, for the Master of Science degree in MATHEMATICS, presented on NOVEMBER 1 2017, at Southern Illinois University Carbondale.

TITLE: KRIGING AND NONPARAMETRIC REGRESSION

MAJOR PROFESSOR: Dr. David J. Olive

Kriging and nonparametric regression are described and prediction intervals using GCV smoothing splines are developed for models of the form $Y = m(x) + e$.

ACKNOWLEDGMENTS

I would like to take this opportunity to thank my research advisor, Dr. David Olive for overseeing my Master's project and Dr. Yaser Samadi and Dr. Kwangho Choiy sitting on my committee. I would also like to thank the professors, staff members and my friends of Southern Illinois University for their instruction, support and care over the past two years. I want to thank my loving husband and three boys and my family for all of your support and encouragement. Last but not least I would like to thank all who supported, and encouraged me to succeed, I could not have done it without you!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

This chapter follows Christensen (1991, ch. 6) closely. Spatial data are modeled as a realization of a stochastic process: $Y(\boldsymbol{u})$ with $\boldsymbol{u} \in \mathbb{D} \subseteq \mathbb{R}^k$ where $\boldsymbol{u}$ is a location in space and k is usually 1, 2, or 3. Here $\mathbb{D}$ is a study region such as a tract of land divided into plots with locations $\boldsymbol{u}_i$. Then $Y(\boldsymbol{u}_i)$ might be the yield from plot $\boldsymbol{u}_i$ where the plots were given different treatments in an experimental design. Given the same treatment, responses from neighboring locations tend to be similar, and thus dependent rather than independent. Assume that for any $\boldsymbol{u} \in \mathbb{D}$, $E(Y(\boldsymbol{u})) = m(\boldsymbol{u})$ and $Var(Y(\boldsymbol{u}))$ exist. Then

$$Y(\boldsymbol{u}) = m(\boldsymbol{u}) + e(\boldsymbol{u}) \tag{1.1}$$

where $e(\boldsymbol{u})$ is a zero mean stochastic error process. The nonparametric regression model for spatial data has

$$Y_i = Y(\boldsymbol{u}_i) = m(\boldsymbol{u}_i) + e_i \tag{1.2}$$

for $i = 1, ..., n$ where often the e_i are assumed to be independent and identically distributed (iid).

The universal kriging model assumes a linear structure for $m(\boldsymbol{u})$: $x_i(\boldsymbol{u})$ are known functions of $\boldsymbol{u}$ for $i = 1, ..., p$ and

$$m(\boldsymbol{u}) = \sum_{i=1}^{p} \beta_i x_i(\boldsymbol{u}). \tag{1.3}$$

A special case of this model is the ordinary kriging model where $m(\boldsymbol{u}) = \mu = \beta_1$. For model (1.1), the covariance function is

$$C(\boldsymbol{u}, \boldsymbol{w}) = \sigma(\boldsymbol{u}, \boldsymbol{w}) = Cov(e(\boldsymbol{u}), e(\boldsymbol{w})) = Cov(Y(\boldsymbol{u}), Y(\boldsymbol{w})). \tag{1.4}$$

The nonparametric regression model (1.2) using cubic smoothing splines is often used as a competitor for the ordinary kriging model. See, for example, Laslett (1994) and Yakowitz and Szidarovszky (1985).

For the universal kriging model, observations are taken at $\boldsymbol{u}_1, ..., \boldsymbol{u}_n$. Let $Y_i = Y(\boldsymbol{u}_i)$, $x_{ij} = x_j(\boldsymbol{u}_i)$, $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip})^T$, $e_i = e(\boldsymbol{u}_i)$, $\boldsymbol{e} = (e_1, ..., e_n)^T$, the $n \times p$ matrix $\boldsymbol{X} = (x_{ij})$, $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$, and the $n \times n$ covariance matrix $Cov(\boldsymbol{e}) = \boldsymbol{\Sigma} = (\sigma_{ij})$ where $\sigma_{ij} = C(\boldsymbol{u}_i, \boldsymbol{u}_j) = \sigma(\boldsymbol{u}_i, \boldsymbol{u}_j) = Cov(e(\boldsymbol{u}_i), e(\boldsymbol{u}_j)) = Cov(Y(\boldsymbol{u}_i), Y(\boldsymbol{u}_j))$.

Then in matrix form, the universal kriging model is $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, which is a generalized least squares (GLS) model if $\boldsymbol{\Sigma}$ is known. This model is a feasible generalized least squares (FGLS) model if $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{V}(\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ needs to be estimated. See the definitions in chapter 2.

The covariance function is important. Let the mean function $\mu(\boldsymbol{u}_i) = E[Y(\boldsymbol{u}_i)]$ and the covariance function $C(\boldsymbol{u}_i, \boldsymbol{u}_j) = E[(Y(\boldsymbol{u}_i) - \mu(\boldsymbol{u}_i))(Y(\boldsymbol{u}_j) - \mu(\boldsymbol{u}_j))]$ for all $\boldsymbol{u}_i, \boldsymbol{u}_j \in \mathbb{D}$. Then the variance function is the variance of $Y(\boldsymbol{u}_i)$ given by $V(\boldsymbol{u}_i) = V(Y(\boldsymbol{u}_i)) = C(\boldsymbol{u}_i, \boldsymbol{u}_i)$. Then the variogram $= \gamma(\boldsymbol{u}_i - \boldsymbol{u}_j) = 0.5V[Y(\boldsymbol{u}_i) - Y(\boldsymbol{u}_j)]$. Then the random function or stochastic process $\{Y(\boldsymbol{u}), \boldsymbol{u} \in \mathbb{D}\}$ is weakly stationary or second order stationary if the mean function is constant $\mu(Y(\boldsymbol{u})) = E[Y(\boldsymbol{u})] \equiv \mu$, and the covariance function $C(Y(\boldsymbol{u}), Y(\boldsymbol{u} + \boldsymbol{h})) = C(\boldsymbol{h})$ only depends on $\boldsymbol{h}$ for $\boldsymbol{u}, \boldsymbol{u} + \boldsymbol{h} \in \mathbb{D}$. For a weakly stationary stochastic process, $V(\boldsymbol{u}) = \sigma^2 = C(\boldsymbol{0})$ and the variogram $\gamma(\boldsymbol{h}) = 0.5V[Y(\boldsymbol{u}+\boldsymbol{h}) - Y(\boldsymbol{u})] = C(\boldsymbol{0}) - C(\boldsymbol{h})$. Some authors call $2\gamma(\boldsymbol{h})$ the variogram. A stochastic process is intrinsically stationary if the differences $W_{\boldsymbol{u}}(\boldsymbol{h}) = Y(\boldsymbol{u} + \boldsymbol{h}) - Y(\boldsymbol{u})$ are second order stationary. Hence the drift $\mu(\boldsymbol{u}) = E[Y(\boldsymbol{u} + \boldsymbol{h}) - Y(\boldsymbol{u})]$ is constant with respect to $\boldsymbol{h}$ and $C(W_{\boldsymbol{u}}(\boldsymbol{h}), W_{\boldsymbol{u}}(\boldsymbol{h} + \boldsymbol{b})) = C(\boldsymbol{h}, \boldsymbol{h} + \boldsymbol{b}) = C(\boldsymbol{b})$ which is equivalent to $\gamma(\boldsymbol{h}) = 0.5V[Y(\boldsymbol{u}+\boldsymbol{h}) - Y(\boldsymbol{u})]$ only depending on $\boldsymbol{h}$. See Montero, Fernández-Avilés, and Mateau (2015, pp. 12-17).

Often the covariance function only depends on the distance $d = \|\boldsymbol{h}\|$. For kriging, the three common covariance functions are given below.

1) The spherical model

$$C(d) = m\left(1 - \left(\frac{3d}{2a} - \frac{d^3}{2a^3}\right)\right) I(0 \le d \le a).$$

Note that $m = C(0)$. As d increases from 0 to a, the covariance function drops almost linearly from m to 0.

2) The exponential model

$$C(d) = m \exp(-d/a)$$

for $a > 0$ decays exponentially with distance d and is close to $0.05m$ for $d = 3a$. Note that the correlation $Cor(Y(\boldsymbol{u}), Y(\boldsymbol{u} + \boldsymbol{h}))$ is near 0 for $d \geq 3a$.

3) The Gaussian model

$$C(d) = m \exp(-d^2/a^2)$$

for $a > 0$ decays exponentially with squared distance d^2, and is close to $0.05m$ for $d = a\sqrt{3}$. This model is unusual in practical implications.

Chapter 2 defines GLS and FGLS, chapter 3 defines the kriging estimator, Chapter 4 considers variogram estimation,chapter 5 discusses simulating correlated regression data in the statistical software package *R*. chapter 6 gives splire prediction intervals and chapter 7 presents simulation results.

CHAPTER 2

GLS AND FGLS

This chapter considers feasible generalized least squares and follows Olive (2017a, ch. 4) closely.

Definition 1. Suppose that the response variable and at least one of the predictor variables is quantitative. Then the generalized least squares (GLS) model is

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{2.1}$$

where $\boldsymbol{Y}$ is an $n \times 1$ vector of dependent variables, $\boldsymbol{X}$ is an $n \times p$ matrix of predictors, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients, and $\boldsymbol{e}$ is an $n \times 1$ vector of unknown errors. Also $E(\boldsymbol{e}) = \boldsymbol{0}$ and $\mathrm{Cov}(\boldsymbol{e}) = \sigma^2 \boldsymbol{V}$ where $\boldsymbol{V}$ is a known $n \times n$ positive definite matrix.

Definition 2. The GLS estimator

$$\hat{\boldsymbol{\beta}}_{GLS} = (\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{Y}. \tag{2.2}$$

The fitted values are $\hat{\boldsymbol{Y}}_{GLS} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{GLS}$.

Definition 3. Suppose that the response variable and at least one of the predictor variables is quantitative. Then the weighted least squares (WLS) model with weights $w_1, ..., w_n$ is the special case of the GLS model where $\boldsymbol{V}$ is diagonal: $\boldsymbol{V} = \mathrm{diag}(\mathrm{v}_1, ..., \mathrm{v}_\mathrm{n})$ and $w_i = 1/v_i$. Hence

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{2.3}$$

$E(\boldsymbol{e}) = \boldsymbol{0}$, and $\mathrm{Cov}(\boldsymbol{e}) = \sigma^2 \mathrm{diag}(\mathrm{v}_1, ..., \mathrm{v}_\mathrm{n}) = \sigma^2 \mathrm{diag}(1/\mathrm{w}_1, ..., 1/\mathrm{w}_\mathrm{n})$.

Definition 4. The WLS estimator

$$\hat{\boldsymbol{\beta}}_{WLS} = (\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{Y}. \tag{2.4}$$

The fitted values are $\hat{\boldsymbol{Y}}_{WLS} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{WLS}$.

Definition 5. The feasible generalized least squares (FGLS) model is the same as the GLS estimator except that $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\theta})$ is a function of an unknown $q \times 1$ vector of parameters $\boldsymbol{\theta}$. Let the estimator of $\boldsymbol{V}$ be $\hat{\boldsymbol{V}} = \boldsymbol{V}(\hat{\boldsymbol{\theta}})$. Then the FGLS estimator

$$\hat{\boldsymbol{\beta}}_{FGLS} = (\boldsymbol{X}^T \hat{\boldsymbol{V}}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \hat{\boldsymbol{V}}^{-1} \boldsymbol{Y}. \tag{2.5}$$

The fitted values are $\hat{\boldsymbol{Y}}_{FGLS} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{FGLS}$. The feasible weighted least squares (FWLS) estimator is the special case of the FGLS estimator where $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\theta})$ is diagonal. Hence the estimated weights $\hat{w}_i = 1/\hat{v}_i = 1/v_i(\hat{\boldsymbol{\theta}})$. The FWLS estimator and fitted values will be denoted by $\hat{\boldsymbol{\beta}}_{FWLS}$ and $\hat{\boldsymbol{Y}}_{FWLS}$, respectively.

Notice that the ordinary least squares (OLS) model is a special case of GLS with $\boldsymbol{V} = \boldsymbol{I}_n$, the $n \times n$ identity matrix. It can be shown that the GLS estimator minimizes the GLS criterion

$$Q_{GLS}(\boldsymbol{\eta}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\eta})^T \boldsymbol{V}^{-1} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\eta}).$$

Notice that the FGLS and FWLS estimators have $p + q + 1$ unknown parameters. These estimators can perform very poorly if $n < 10(p + q + 1)$.

The GLS and WLS estimators can be found from the OLS regression (without an intercept) of a transformed model. Typically there will be a constant in the model: the first column of $\boldsymbol{X}$ is a vector of ones. Following Seber and Lee (2003, pp. 66-68), there is a nonsingular $n \times n$ matrix $\boldsymbol{K}$ such that $\boldsymbol{V} = \boldsymbol{K}\boldsymbol{K}^T$. Let $\boldsymbol{Z} = \boldsymbol{K}^{-1}\boldsymbol{Y}$, $\boldsymbol{U} = \boldsymbol{K}^{-1}\boldsymbol{X}$, and $\boldsymbol{\epsilon} = \boldsymbol{K}^{-1}\boldsymbol{e}$. This method uses the fast, but rather unstable, Cholesky decomposition.

Proposition 1. a)

$$\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.6}$$

follows the OLS model since $E(\boldsymbol{\epsilon}) = \boldsymbol{0}$ and $\mathrm{Cov}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}_n$.

b) The GLS estimator $\hat{\boldsymbol{\beta}}_{GLS}$ can be obtained from the OLS regression (without an intercept) of $\boldsymbol{Z}$ on $\boldsymbol{U}$.

c) For WLS, $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$. The corresponding OLS model $\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ is equivalent to $Z_i = \boldsymbol{u}_i^T\boldsymbol{\beta} + \epsilon_i$ for $i = 1, ..., n$ where $\boldsymbol{u}_i^T$ is the ith row of $\boldsymbol{U}$. Then $Z_i = \sqrt{w_i}\, Y_i$

and $\boldsymbol{u}_i = \sqrt{w_i}\, \boldsymbol{x}_i$. Hence $\hat{\boldsymbol{\beta}}_{WLS}$ can be obtained from the OLS regression (without an intercept) of $Z_i = \sqrt{w_i}\, Y_i$ on $\boldsymbol{u}_i = \sqrt{w_i}\, \boldsymbol{x}_i$.

Proof. a) $E(\boldsymbol{\epsilon}) = \boldsymbol{K}^{-1}E(\boldsymbol{e}) = \boldsymbol{0}$ and

$$\mathrm{Cov}(\boldsymbol{\epsilon}) = \boldsymbol{K}^{-1}\mathrm{Cov}(\boldsymbol{e})(\boldsymbol{K}^{-1})^T = \sigma^2\boldsymbol{K}^{-1}\boldsymbol{V}(\boldsymbol{K}^{-1})^T$$

$$= \sigma^2\boldsymbol{K}^{-1}\boldsymbol{K}\boldsymbol{K}^T(\boldsymbol{K}^{-1})^T = \sigma^2\boldsymbol{I}_n.$$

Notice that OLS without an intercept needs to be used since $\boldsymbol{U}$ does not contain a vector of ones. The first column of $\boldsymbol{U}$ is $\boldsymbol{K}^{-1}\boldsymbol{1} \neq \boldsymbol{1}$.

b) Let $\hat{\boldsymbol{\beta}}_{ZU}$ denote the OLS estimator obtained by regressing $\boldsymbol{Z}$ on $\boldsymbol{U}$. Then

$$\hat{\boldsymbol{\beta}}_{ZU} = (\boldsymbol{U}^T\boldsymbol{U})^{-1}\boldsymbol{U}^T\boldsymbol{Z} = (\boldsymbol{X}^T(\boldsymbol{K}^{-1})^T\boldsymbol{K}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{K}^{-1})^T\boldsymbol{K}^{-1}\boldsymbol{Y}$$

and the result follows since $\boldsymbol{V}^{-1} = (\boldsymbol{K}\boldsymbol{K}^T)^{-1} = (\boldsymbol{K}^T)^{-1}\boldsymbol{K}^{-1} = (\boldsymbol{K}^{-1})^T\boldsymbol{K}^{-1}$.

c) The result follows from b) if $Z_i = \sqrt{w_i}\, Y_i$ and $\boldsymbol{u}_i = \sqrt{w_i}\, \boldsymbol{x}_i$. But for WLS, $\boldsymbol{V} = \mathrm{diag}(\mathrm{v}_1, ..., \mathrm{v}_\mathrm{n})$ and hence $\boldsymbol{K} = \boldsymbol{K}^T = \mathrm{diag}(\sqrt{\mathrm{v}_1}, ..., \sqrt{\mathrm{v}_\mathrm{n}})$. Hence

$$\boldsymbol{K}^{-1} = \mathrm{diag}(1/\sqrt{\mathrm{v}_1}, ..., 1/\sqrt{\mathrm{v}_\mathrm{n}}) = \mathrm{diag}(\sqrt{\mathrm{w}_1}, ..., \sqrt{\mathrm{w}_\mathrm{n}})$$

and $\boldsymbol{Z} = \boldsymbol{K}^{-1}\boldsymbol{Y}$ has ith element $Z_i = \sqrt{w_i}\, Y_i$. Similarly, $\boldsymbol{U} = \boldsymbol{K}^{-1}\boldsymbol{X}$ has ith row $\boldsymbol{u}_i^T = \sqrt{w_i}\, \boldsymbol{x}_i^T$. □

Following Johnson and Wichern (1988, p. 51) and Freedman (2005, p. 54), there is a symmetric, nonsingular $n \times n$ square root matrix $\boldsymbol{R} = \boldsymbol{V}^{1/2}$ such that $\boldsymbol{V} = \boldsymbol{R}\boldsymbol{R}$. Let $\boldsymbol{Z} = \boldsymbol{R}^{-1}\boldsymbol{Y}$, $\boldsymbol{U} = \boldsymbol{R}^{-1}\boldsymbol{X}$ and $\boldsymbol{\epsilon} = \boldsymbol{R}^{-1}\boldsymbol{e}$. This method uses the spectral theorem (singular value decomposition) and has better computational properties than transformation based on the Cholesky decomposition.

Proposition 2. a)

$$\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.7}$$

follows the OLS model since $E(\boldsymbol{\epsilon}) = \boldsymbol{0}$ and $\mathrm{Cov}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{I}_n$.

b) The GLS estimator $\hat{\boldsymbol{\beta}}_{GLS}$ can be obtained from the OLS regression (without an intercept) of $\boldsymbol{Z}$ on $\boldsymbol{U}$.

c) For WLS, $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$. The corresponding OLS model $\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ is equivalent to $Z_i = \boldsymbol{u}_i^T\boldsymbol{\beta} + \epsilon_i$ for $i = 1, ..., n$ where $\boldsymbol{u}_i^T$ is the ith row of $\boldsymbol{U}$. Then $Z_i = \sqrt{w_i}\, Y_i$ and $\boldsymbol{u}_i = \sqrt{w_i}\, \boldsymbol{x}_i$. Hence $\hat{\boldsymbol{\beta}}_{WLS}$ can be obtained from the OLS regression (without an intercept) of $Z_i = \sqrt{w_i}\, Y_i$ on $\boldsymbol{u}_i = \sqrt{w_i}\, \boldsymbol{x}_i$.

Proof. a) $E(\boldsymbol{\epsilon}) = \boldsymbol{R}^{-1}E(\boldsymbol{e}) = \boldsymbol{0}$ and

$$\mathrm{Cov}(\boldsymbol{\epsilon}) = \boldsymbol{R}^{-1}\mathrm{Cov}(\boldsymbol{e})(\boldsymbol{R}^{-1})^T = \sigma^2\boldsymbol{R}^{-1}\boldsymbol{V}(\boldsymbol{R}^{-1})^T$$

$$= \sigma^2\boldsymbol{R}^{-1}\boldsymbol{R}\boldsymbol{R}(\boldsymbol{R}^{-1}) = \sigma^2\boldsymbol{I}_n.$$

Notice that OLS without an intercept needs to be used since $\boldsymbol{U}$ does not contain a vector of ones. The first column of $\boldsymbol{U}$ is $\boldsymbol{R}^{-1}\boldsymbol{1} \neq \boldsymbol{1}$.

b) Let $\hat{\boldsymbol{\beta}}_{ZU}$ denote the OLS estimator obtained by regressing $\boldsymbol{Z}$ on $\boldsymbol{U}$. Then

$$\hat{\boldsymbol{\beta}}_{ZU} = (\boldsymbol{U}^T\boldsymbol{U})^{-1}\boldsymbol{U}^T\boldsymbol{Z} = (\boldsymbol{X}^T(\boldsymbol{R}^{-1})^T\boldsymbol{R}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{R}^{-1})^T\boldsymbol{R}^{-1}\boldsymbol{Y}$$

and the result follows since $\boldsymbol{V}^{-1} = (\boldsymbol{R}\boldsymbol{R})^{-1} = \boldsymbol{R}^{-1}\boldsymbol{R}^{-1} = (\boldsymbol{R}^{-1})^T\boldsymbol{R}^{-1}$.

c) The result follows from b) if $Z_i = \sqrt{w_i}\, Y_i$ and $\boldsymbol{u}_i = \sqrt{w_i}\, \boldsymbol{x}_i$. But for WLS, $\boldsymbol{V} = \mathrm{diag}(\mathrm{v}_1, ..., \mathrm{v}_n)$ and hence $\boldsymbol{R} = \mathrm{diag}(\sqrt{\mathrm{v}_1}, ..., \sqrt{\mathrm{v}_n})$. Hence

$$\boldsymbol{R}^{-1} = \mathrm{diag}(1/\sqrt{\mathrm{v}_1}, ..., 1/\sqrt{\mathrm{v}_n}) = \mathrm{diag}(\sqrt{\mathrm{w}_1}, ..., \sqrt{\mathrm{w}_n})$$

and $\boldsymbol{Z} = \boldsymbol{R}^{-1}\boldsymbol{Y}$ has ith element $Z_i = \sqrt{w_i}\, Y_i$. Similarly, $\boldsymbol{U} = \boldsymbol{R}^{-1}\boldsymbol{X}$ has ith row $\boldsymbol{u}_i^T = \sqrt{w_i}\, \boldsymbol{x}_i^T$. □

Remark 1. Standard software produces WLS output and the ANOVA F test and Wald t tests are performed using this output.

Remark 2. The FGLS estimator can also be found from the OLS regression (without an intercept) of $\boldsymbol{Z}$ on $\boldsymbol{U}$ where $\boldsymbol{V}(\hat{\boldsymbol{\theta}}) = \boldsymbol{R}\boldsymbol{R}$. Similarly the FWLS estimator can be

found from the OLS regression (without an intercept) of $Z_i = \sqrt{\hat{w}_i} Y_i$ on $\boldsymbol{u}_i = \sqrt{\hat{w}_i}\boldsymbol{x}_i$. But now $\boldsymbol{U}$ is a random matrix instead of a constant matrix. Hence these estimators are highly nonlinear. OLS output can be used for exploratory purposes, but the p–values for hypothesis testing are generally not correct. The Olive (2017ab) nonparametric bootstrap tests may be useful for FGLS and FWLS. The nonparametric bootstrap could also be applied to the OLS estimator.

Under regularity conditions, the OLS estimator $\hat{\boldsymbol{\beta}}_{OLS}$ is a consistent estimator of $\boldsymbol{\beta}$ when the GLS model holds, but $\hat{\boldsymbol{\beta}}_{GLS}$ should be used because it generally has higher efficiency.

Definition 8. Let $\hat{\boldsymbol{\beta}}_{ZU}$ be the OLS estimator from regressing $\boldsymbol{Z}$ on $\boldsymbol{U}$. The vector of fitted values is $\hat{\boldsymbol{Z}} = \boldsymbol{U}\hat{\boldsymbol{\beta}}_{ZU}$ and the vector of residuals is $\boldsymbol{r}_{ZU} = \boldsymbol{Z} - \hat{\boldsymbol{Z}}$. Then $\hat{\boldsymbol{\beta}}_{ZU} = \hat{\boldsymbol{\beta}}_{GLS}$ for GLS, $\hat{\boldsymbol{\beta}}_{ZU} = \hat{\boldsymbol{\beta}}_{FGLS}$ for FGLS, $\hat{\boldsymbol{\beta}}_{ZU} = \hat{\boldsymbol{\beta}}_{WLS}$ for WLS, and $\hat{\boldsymbol{\beta}}_{ZU} = \hat{\boldsymbol{\beta}}_{FWLS}$ for FWLS. For GLS, FGLS, WLS, and FWLS, a residual plot is a plot of $\hat{Z}_i$ versus $r_{ZU,i}$ and a *response plot* is a plot of $\hat{Z}_i$ versus Z_i.

Notice that the residual and response plots are based on the OLS output from the OLS regression without intercept of $\boldsymbol{Z}$ on $\boldsymbol{U}$. If the model is good, then the plotted points in the response plot should follow the identity line in an evenly populated band while the plotted points in the residual plot should follow the line $r_{ZU,i} = 0$ in an evenly populated band (at least if the distribution of ϵ is not highly skewed).

Plots based on $\hat{Y}_{GLS} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{ZU}$ and on $r_{i,GLS} = Y_i - \hat{Y}_{i,GLS}$ should be similar to those based on $\hat{\boldsymbol{\beta}}_{OLS}$. Although the plot of $\hat{Y}_{i,GLS}$ versus Y_i should be linear, the plotted points will not scatter about the identity line in an evenly populated band. Hence this plot can not be used to check whether the GLS model with $\boldsymbol{V}$ is a good approximation to the data. Moreover, the $r_{i,GLS}$ and $\hat{Y}_{i,GLS}$ may be correlated and usually do not scatter about the $r = 0$ line in an evenly populated band. The plots in Definition 8 are both a check on linearity and on whether the model using $\boldsymbol{V}$ (or $\hat{\boldsymbol{V}}$) gives a good approxima-

tion of the data, provided that $n > k(p+q+1)$ where $k \geq 5$ and preferably $k \geq 10$.

For GLS and WLS (and for exploratory purposes for FGLS and FWLS), plots and model building and variable selection should be based on $\boldsymbol{Z}$ and $\boldsymbol{U}$. Form $\boldsymbol{Z}$ and $\boldsymbol{U}$ and then use OLS software for model selection and variable selection. If the columns of $\boldsymbol{X}$ are $\boldsymbol{v}_1, ..., \boldsymbol{v}_p$, then the columns of $\boldsymbol{U}$ are $U_1, ..., U_p$ where $U_j = \boldsymbol{R}^{-1}\boldsymbol{v}_j$ corresponds to the jth predictor X_j. For example, the analog of the OLS residual plot of jth predictor versus the residuals is the plot of the jth predictor U_j versus r_{ZU}. The notation is confusing but the idea is simple: form $\boldsymbol{Z}$ and $\boldsymbol{U}$, then use OLS software and the OLS techniques to build the model.

CHAPTER 3

THE UNIVERSAL KRIGING ESTIMATOR

This chapter follows Christensen (1987: pp. 226-227, 1991: ch. 6) closely. A predictor $\hat{Y}_0$ of Y_0 is unbiased if $E[\hat{Y}_0] = Y_0$, and $b_0 + \boldsymbol{b}^T\boldsymbol{Y}$ is the best linear unbiased predictor of Y_0 if $b_0 + \boldsymbol{b}^T\boldsymbol{Y}$ is unbiased, and for any other unbiased linear predictor $a_0 + \boldsymbol{a}^T\boldsymbol{Y}$,

$$E[(Y_0 - b_0 - \boldsymbol{b}^T\boldsymbol{Y})^2] \leq E[(Y_0 - a_0 - \boldsymbol{a}^T\boldsymbol{Y})^2].$$

Then finding the best linear unbiased predictor for spatial data is called kriging.

Suppose we want to predict $Y_0 = Y(\boldsymbol{u}_0)$ given $\boldsymbol{x}_0$ and $Y_1, ..., Y_n$ where the Y_i satisfy (1.1) and (1.3). Assume $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{V}$ and $\sigma(\boldsymbol{u}, \boldsymbol{w})$ are known. Let

$$\boldsymbol{\Sigma}_{Y0} = \begin{bmatrix} \sigma(\boldsymbol{u}_1, \boldsymbol{u}_0) \\ \sigma(\boldsymbol{u}_2, \boldsymbol{u}_0) \\ \vdots \\ \sigma(\boldsymbol{u}_n, \boldsymbol{u}_0) \end{bmatrix}.$$

Under these extremely strong assumptions, Christensen (1991, p. 268) shows that the best unbiased linear predictor $\boldsymbol{b}^T\boldsymbol{Y}$ of Y_0 is

$$\hat{Y}_0 = \boldsymbol{x}_0^T\hat{\boldsymbol{\beta}}_{GLS} + \boldsymbol{\Sigma}_{Y0}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{GLS}) = \boldsymbol{b}^T\boldsymbol{Y} \tag{3.1}$$

where

$$\boldsymbol{b} = \boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_{Y0}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1})$$

with $Var(Y_0 - \hat{Y}_0) = \sigma(\boldsymbol{u}_0, \boldsymbol{u}_0) - 2\boldsymbol{b}^T\boldsymbol{\Sigma}_{Y0}^T + \boldsymbol{b}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{b}$.

Then $\boldsymbol{b}^T\boldsymbol{Y}$ is the universal kriging estimator of Y_0, and is used to create a 2 or 3 dimensional map of the variable $Y(\boldsymbol{u})$ over the region $\mathbb{D}$ by using a large number of locations $\boldsymbol{u}_{0,i}$.

Note: often Y_i is an observation on a neighborhood (block) B_i of $\boldsymbol{u}_i$. Let $|B_i|$ be the volume (area) of the ith bolck B_i. Then

$$Y_i = \frac{1}{|B_i|}\int_{B_i} Y(\boldsymbol{u})d\boldsymbol{u},$$

for $i = 0, 1, ..., n$, and

$$x_{ij} = \frac{1}{|B_i|}\int_{B_i} x_j(\boldsymbol{u})d\boldsymbol{u}.$$

Then $E(Y_i) = \sum_{k=1}^{p} \beta_k x_{ik} = \boldsymbol{x}_i^T\boldsymbol{\beta}$, but

$$Cov(Y_i, Y_j) = \sigma_{ij} = \frac{1}{|B_i|}\frac{1}{|B_j|}\int_{B_i}\int_{B_j} \sigma(\boldsymbol{u}, \boldsymbol{w})d\boldsymbol{u}d\boldsymbol{w}.$$

Then the predictor (3.1) is used with $\boldsymbol{\Sigma} = (\sigma_{ij})$ and $\sigma(\boldsymbol{u}_i, \boldsymbol{u}_0)$ replaced by $Cov(Y_i, Y_0)$.

CHAPTER 4

METHODS OF VARIOGRAM ESTIMATION

Consider ordinary kriging. (For the nonstationary kriging data such as the universal kriging model, detrend the data then use the variogram of the approximately stationary residuals.) Assume that the locations are from a uniform grid so that the $n_{\boldsymbol{h}}$ defined below are fairly large.

The classical empirical estimator of the variogram based on the data $\{Y(\boldsymbol{u}_i)\}_{i=1}^n$ is

$$2\hat{\gamma}(\boldsymbol{h}) = \frac{1}{n_{\boldsymbol{h}}} \sum_{N(\boldsymbol{h})} [Y(\boldsymbol{u}_i) - Y(\boldsymbol{u}_j)]^2 \tag{3.1}$$

where $N(\boldsymbol{h}) = \{(\boldsymbol{u}_i, \boldsymbol{u}_j) : \boldsymbol{u}_i - \boldsymbol{u}_j = \boldsymbol{h}\}$ and $n_{\boldsymbol{h}}$ is the number of distinct ordered pairs in $N(\boldsymbol{h})$.

Now let $2\gamma(\boldsymbol{h}|\boldsymbol{\theta})$ be a parametric variogram model that depends on unknown parameters $\boldsymbol{\theta}$. For the spherical, exponential, and Gaussian models given in chapter 1, $\boldsymbol{\theta} = (a, d, m)^T$. Consider discrete lags $\boldsymbol{h}_1, ..., \boldsymbol{h}_k$ where $\|\boldsymbol{h}_1\| < \cdots < \|\boldsymbol{h}_k\|$. The least squares method for estimating $\boldsymbol{\theta}$ chooses $\hat{\boldsymbol{\theta}}$ to minimize

$$\sum_{j=1}^{k} [2\hat{\gamma}(\boldsymbol{h}_j) - 2\gamma(\boldsymbol{h}_j|\boldsymbol{\theta})]^2$$

where $2\hat{\gamma}(\boldsymbol{h})$ is given by (4.1). There are also WLS and GLS methods for estimating $\boldsymbol{\theta}$.

CHAPTER 5

GENERATING DATA WITH A COVARIANCE MATRIX

There is a simple method for generating regression data with $Cov(\boldsymbol{e}) = \boldsymbol{\Sigma}$. First generate iid data $\boldsymbol{w}$ with $Cov(\boldsymbol{w}) = \sigma^2 \boldsymbol{I}_n$. Then $\boldsymbol{e} = \boldsymbol{A}\boldsymbol{w}$ has $Cov(\boldsymbol{e}) = \boldsymbol{\Sigma} = \sigma^2 \boldsymbol{A}\boldsymbol{A}^T$ where $\boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{A}^2$ if $\boldsymbol{A}$ is symmetric. The method is to specify the positive definite covariance matrix $\boldsymbol{\Sigma}$, take $\sigma^2 = 1$ and $\boldsymbol{A} = \boldsymbol{\Sigma}^{1/2}$, the symmetric square root matrix: $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{1/2}$. As an example of the this method, see the R code below. The `mvrnorm` function generates $n \times 1$ vectors $\boldsymbol{e}_i \sim N_n(\boldsymbol{0}, \boldsymbol{\Sigma})$ for $i = 1, ..., m$ and for a specified $n \times n$ covariance matrix $\boldsymbol{\Sigma}$.

```
library(MASS)
n <- 100    # number of errors e_1, ..., e_n
m <- 1 # number of n by 1 error vectors to generate, for regression 1
Sigma <- outer(1:n, 1:n, function(x,y) {.7^abs(x-y)})
e <- mvrnorm(m, rep(0,n), Sigma) #from MASS
#diagonals of the correlation matrix should be 0, 0.7, 0.7^2 0.7^3 etc
> n<-7 #take n small so it is easy to see the covariance matrix
> outer(1:n, 1:n, function(x,y) {.7^abs(x-y)}) #theta = 0.7
          [,1]    [,2]   [,3]  [,4]   [,5]    [,6]     [,7]
[1,] 1.000000 0.70000 0.4900 0.343 0.2401 0.16807 0.117649
[2,] 0.700000 1.00000 0.7000 0.490 0.3430 0.24010 0.168070
[3,] 0.490000 0.70000 1.0000 0.700 0.4900 0.34300 0.240100
[4,] 0.343000 0.49000 0.7000 1.000 0.7000 0.49000 0.343000
[5,] 0.240100 0.34300 0.4900 0.700 1.0000 0.70000 0.490000
[6,] 0.168070 0.24010 0.3430 0.490 0.7000 1.00000 0.700000
[7,] 0.117649 0.16807 0.2401 0.343 0.4900 0.70000 1.000000
> .7^6
```

```
[1] 0.117649
```

CHAPTER 6

SPLINE PREDICTION INTERVALS

This chapter describes the Olive (2013) asymptotically optimal prediction intervals for a regression model of the form $Y_i = m(\boldsymbol{x}_i) + e_i$ for $i = 1, ..., n$ where m is a function of $\boldsymbol{x}_i$ and the errors e_i are iid from a continuous unimodal distribution. If $\hat{m}$ is an estimator of m, then the ith residual is $r_i = Y_i - \hat{m}(\boldsymbol{x}_i) = Y_i - \hat{Y}_i$. The prediction intervals have coverage near or higher than the nominal coverage for many techniques even for moderate sample size n, say $n > 10$(model degrees of freedom). The prediction intervals are for a future response Y_f given a $p \times 1$ vector $\boldsymbol{x}_f$ of predictors.

A large sample $100(1 - \delta)\%$ prediction interval (PI) has the form $(\hat{L}_n, \hat{U}_n)$ where $P(\hat{L}_n < Y_f < \hat{U}_n) \xrightarrow{P} 1 - \delta$ as the sample size $n \to \infty$. Let ξ_δ be the δ percentile of the error e, i.e., $P(e \leq \xi_\delta) = \delta$. Let $\hat{\xi}_\delta$ be the sample δ percentile of the residuals. Consider predicting a future observation Y_f given a vector of predictors $\boldsymbol{x}_f$ where $(Y_f, \boldsymbol{x}_f)$ comes from the same population as the past data $(Y_i, \boldsymbol{x}_i)$ for $i = 1, ..., n$. Let $1 - \delta_2 - \delta_1 = 1 - \delta$ with $0 < \delta < 1$ and $\delta_1 < 1 - \delta_2$ where $0 < \delta_i < 1$. Then $P[Y_f \in (m(\boldsymbol{x}_f) + \xi_{\delta_1}, m(\boldsymbol{x}_f) + \xi_{1-\delta_2})] = 1 - \delta$.

Assume that $\hat{m}$ is consistent: $\hat{m}(\boldsymbol{x}) \xrightarrow{P} m(\boldsymbol{x})$ as $n \to \infty$. Then $r_i = Y_i - \hat{m}(\boldsymbol{x}_i) \xrightarrow{P} Y_i - m(\boldsymbol{x}_i) = e_i$ and, under "mild" regularity conditions, $\hat{\xi}_\delta \xrightarrow{P} \xi_\delta$. If $a_n \xrightarrow{P} 1$ and $b_n \xrightarrow{P} 1$, then

$$(\hat{L}_n, \hat{U}_n) = (\hat{m}(\boldsymbol{x}_f) + a_n \hat{\xi}_{\delta_1}, \hat{m}(\boldsymbol{x}_f) + b_n \hat{\xi}_{1-\delta_2}) \tag{6.1}$$

is a large sample $100(1 - \delta)\%$ PI for Y_f.

The shorth(c) estimator is useful for making prediction intervals. Let $Z_{(1)}, ..., Z_{(n)}$ be the order statistics of $Z_1, ..., Z_n$. Then let the shortest closed interval containing at least c of the Z_i be

$$\text{shorth(c)} = [\text{Z}_{(\text{s})}, \text{Z}_{(\text{s+c}-1)}]. \tag{6.2}$$

Let

$$k_n = \lceil n(1-\delta) \rceil. \tag{6.3}$$

Frey (2013) showed that for large $n\delta$ and iid data, the shorth(k_n) PI has maximum undercoverage $\approx 1.12\sqrt{\delta/n}$, and used the shorth($c$) estimator as the large sample $100(1-\delta)\%$ PI where

$$c = \min(n, \lceil n[1 - \delta + 1.12\sqrt{\delta/n}\ ]\ \rceil). \tag{6.4}$$

A problem with the prediction intervals that cover $\approx 100(1-\delta)\%$ of the training data cases Y_i (such as the shorth(k_n) PI), is that they have coverage lower than the nominal coverage of $1-\delta$ for moderate n. This result is not surprising since empirically statistical methods perform worse on test data. Increasing c will improve the coverage for moderate samples.

Example 1. (Example 5.3 from Olive (2017b).) Given below were votes for preseason 1A basketball poll from Nov. 22, 2011 WSIL News where the 778 was a typo: the actual value was 78. As shown below, finding shorth(3) from the ordered data is simple. If the outlier was corrected, shorth(3) = [76,78].

```
111   89   778   78   76

order data: 76 78 89 111 778

                  13 = 89 - 76

                      33 = 111 - 78

                         689 = 778 - 89

shorth(3) = [76,89]
```

Find the target population $100(1-\delta)\%$ covering interval. For small n, the coverage of the training data will be higher than that for the future case to be predicted. In simulations for a large group of models and distributions and $n = 20p$, the undercoverage could be as high as $\min(0.05, \delta/2)$. Let $q_n = \min(1-\delta+0.05, 1-\delta+p/n)$ for $\delta > 0.1$ and

$$q_n = \min(1-\delta/2, 1-\delta+10\delta p/n), \quad \text{otherwise.} \tag{6.5}$$

If $1-\delta < 0.999$ and $q_n < 1-\delta+0.001$, set $q_n = 1-\delta$. Then use the prediction interval or region that covers $100q_n\%$ of the training data. The coverage of the training data is $100q_n\%$ and converges to $100(1-\delta)\%$ as $n \to \infty$, even if the model assumptions fail to hold.

The technique used to produce asymptotically optimal PIs that perform well for moderate samples is simple. Find $\hat{Y}_f$ and the residuals from the regression model. Let

$$b_n = \left(1 + \frac{15}{n}\right)\sqrt{\frac{n+2p}{n-p}}. \tag{6.6}$$

Let $\delta_n = 1 - q_n$ where q_n is given by (6.5). Then

$$(\hat{L}_n, \hat{U}_n) = (\hat{m}(\boldsymbol{x}_f) + b_n\hat{\xi}_{\delta_n/2}, \hat{m}(\boldsymbol{x}_f) + b_n\hat{\xi}_{1-\delta_n/2}) \tag{6.7}$$

is a large sample $100(1-\delta)\%$ PI for Y_f. This semiparametric PI is only asymptotically optimal if the unimodal error distribution is symmetric about 0. The following PI does not need the symmetry assumption.

Let $c = \lceil nq_n \rceil$. Compute $r_{(c)} - r_{(1)}, r_{(c+1)} - r_{(2)}, ..., r_{(n)} - r_{(n-c+1)}$. Let $(r_{(d)}, r_{(d+c-1)}) = (\tilde{\xi}_{\delta_1}, \tilde{\xi}_{1-\delta_2})$ correspond to the interval with the smallest length. Then the asymptotically optimal 100 $(1-\delta)\%$ large sample PI for Y_f is

$$(\hat{m}(\boldsymbol{x}_f) + b_n\tilde{\xi}_{\delta_1}, \hat{m}(\boldsymbol{x}_f) + b_n\tilde{\xi}_{1-\delta_2}). \tag{6.8}$$

Olive (2013) shows that PI (6.8) is asymptotically optimal under mild regularity conditions if the sample percentiles of the residuals converge to the population percentiles of the iid unimodal errors: $\hat{\xi}_\delta \xrightarrow{P} \xi_\delta$. Even if these assumptions do not hold, the

PI covers $100q_n\%$ of the training data, and often the coverage of the future case will be close to $100(1-\delta)$ if the future case Y_f is similar to the training data.

For asymptotic optimality, we can not have extrapolation. Also, even if the coverage converges to the nominal coverage, the length of the PI need not be asymptotically shortest unless the highest $1-\delta$ density region of the probability density function of the iid errors is an interval. The highest density region is an interval for unimodal distributions, but need not be an interval for multimodal distributions for all δ.

Let d be an estimator of the model degrees of freedom. Since the smoothing spline d is typically much larger than $p = 1$, the Olive (2017c) and Pelawa Watagoda and Olive (2017) PI may be useful. This PI is similar to the Olive (2013) PI (6.8) with p replaced by d. Let $q_n = \min(1-\delta+0.05, 1-\delta+d/n)$ for $\delta > 0.1$ and

$$q_n = \min(1-\delta/2, 1-\delta+10\delta d/n), \quad \text{otherwise.} \tag{6.9}$$

If $1-\delta < 0.999$ and $q_n < 1-\delta+0.001$, set $q_n = 1-\delta$. Let

$$c = \lceil nq_n \rceil, \tag{6.10}$$

and let

$$b_n = \left(1+\frac{15}{n}\right)\sqrt{\frac{n+2d}{n-d}} \tag{6.11}$$

if $d \leq 8n/9$, and

$$b_n = 5\left(1+\frac{15}{n}\right),$$

otherwise. Compute the shorth(c) of the residuals $= [r_{(s)}, r_{(s+c-1)}] = [\tilde{\xi}_{\delta_1}, \tilde{\xi}_{1-\delta_2}]$. Then a 100 $(1-\delta)\%$ large sample PI for Y_f is

$$[\hat{m}(\boldsymbol{x}_f) + b_n\tilde{\xi}_{\delta_1}, \hat{m}(\boldsymbol{x}_f) + b_n\tilde{\xi}_{1-\delta_2}]. \tag{6.12}$$

For the R function `smooth.spline`, $p = 1$ and the function m is unknown but $\hat{m}$ is estimated using cubic splines and the GCV criterion. As described in Laslett (1994), Wahba and Cressie discussed the merits of kriging and smoothing splines with iid errors.

CHAPTER 7

EXAMPLES AND SIMULATIONS

Laslett (1994) gives a roller height data set that consists of 1150 heights measured at 1 micron intervals along the drum of a roller. The data set is divided into training data consisting of the odd numbered sites and test data consisting of the even number sites except cite 1150 to avoid extrapolation. For this data set ordinary kriging with the spherical covariance model yields better predictions than the GCV cubic spline. Figure 1 attempts to reproduce Figure 7.1 of Lastlett (1994), which used sites 900-1150, and shows the smoothing spline fit. The kriging fit, not shown, is saw toothed and looks like a time series except the unit of measurement is distance rather than time. See the *R* code below.

```
roller<-matrix(scan(),byrow=T,ncol=1) #copy and paste data
roller <- as.vector(roller)
x<-seq(1,1150)
y <- roller
a<-seq(1,1149,2)
b <- seq(2,1150,2)
xa <- x[a] #xa is the training data in Fig. 1
xb <- x[b] #xb is the test data
ya <- y[a]
yb <- y[b]
outa <- smooth.spline(xa,ya) #fit training data
pxb <- xb[-150]
pyb <- yb[-150]
fitatob <- predict(outa,pxb)
xashort <- xa[451:575]
```

```
yashort <- ya[451:575]

xpashort <- fitatob$x[450:573]

ypashort <- fitatob$y[450:573]

xbshort<-pxb[450:573]

ybshort<-pyb[450:573]

plot(xashort,yashort)

points(xbshort,ybshort,pch=4)#x = test data, 0 = training data

lines(xpashort,ypashort)  #about Laslett Fig. 1 fitted GCV cubic spline

 \caption{"Fitted GCV Cubic Spline for Roller Height Data"}
```

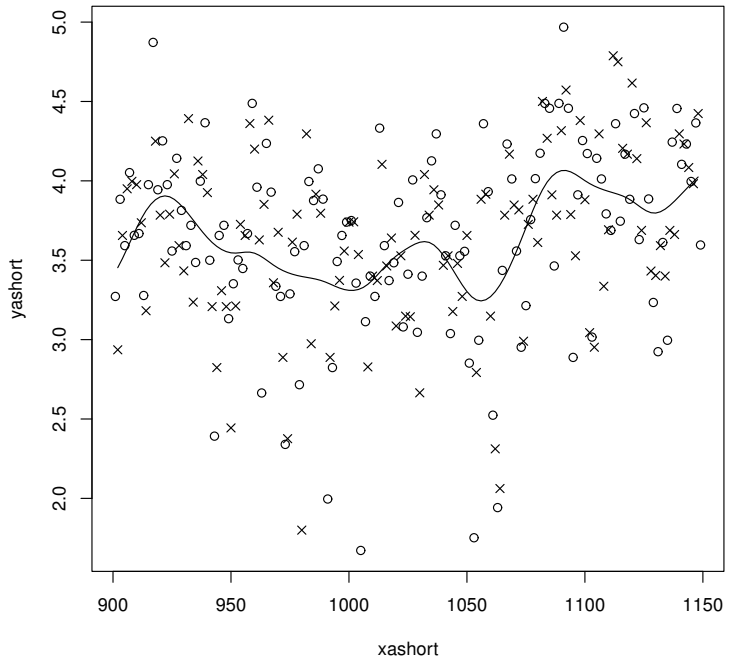

Figure 7.1. Fitted GCV Cubic Spline for Roller Height Data.

The smoothing spline simulation compares the PI lengths and coverages for sample sizes $n = 75, 100$ and 1000 for PIs (6.7), (6.8), and (6.12). Values for PI (6.7) were denoted by scov and slen, values for PI (6.8) were denoted by ocov and olen, and values for

PI (6.12) by dcov and dlen. The five error distributions in the simulation were 1) N(0,1), 2) t_3, 3) exponential(1) -1, 4) uniform$(-1, 1)$ and 5) $0.9N(0, 1)+0.1N(0, 100)$. The value $n = \infty$ gives the asymptotic coverages and lengths and does not depend on the model m if $\hat{m}$ is a consistent estimator of m. Three model types were used 1) $m(x) = x + x^2$, 2) $m(x) = \sin(x) + \cos(x) + \log(|x|)$, and 3) $m(x) = 3\sqrt{|x|}$. Tables 1, 2, and 3 show some results for these three models, and some R code is shown below.

```
pisimspline(n=100,nruns=5000,type=1,modt=1)
$adf
[1] 6.216255
$pimenlen
[1] 4.709470 4.694943 5.058490
$spicov
[1] 0.966
$opicov
[1] 0.9604
$dpicov
[1] 0.9736
```

Table 7.1. PIs for model type = 1

error type	n	95% PI slen	PI olen	95% PI dlen	PI scov	95% PI ocov	PI dcov	adf
1	75	4.7789	4.9436	5.4529	0.9612	0.9634	0.978	6.33
1	100	4.7095	4.6949	5.0585	0.9660	0.9604	0.9736	6.27
1	1000	4.0295	3.9867	4.0349	0.9508	0.9474	0.9502	9.03
1	∞	3.920	3.920	3.92	0.95	0.95	0.95	
2	75	8.4209	8.9654	10.0322	0.9464	0.9518	0.9696	7.45
2	100	8.3450	8.2741	8.9250	0.9567	0.9566	0.9668	6.69
2	1000	6.6341	6.5213	6.5898	0.9468	0.9458	0.9458	7.98
2	∞	6.365	6.365	6.365	0.95	0.95	0.95	
3	75	4.6873	4.7001	5.2655	0.9636	0.9638	0.9788	7.24
3	100	4.6234	4.3214	4.6873	0.97	0.9656	0.976	6.69
3	1000	3.8037	3.2392	3.2789	0.9578	0.9552	0.9572	9.08
3	∞	3.664	2.996	2.996	0.95	0.95	0.95	
4	75	2.3186	2.3525	2.6218	0.973	0.9742	0.9878	6.69
4	100	2.2395	2.2325	2.4330	0.9778	0.9754	0.9868	6.95
4	1000	1.9328	1.9246	1.9523	0.958	0.9534	0.9636	10.52
4	∞	1.900	1.900	1.900	0.95	0.95	0.95	
5	75	17.576	18.8094	21.1494	0.914	0.9158	0.9472	11.20
5	100	17.989	17.0503	18.4221	0.942	0.9368	0.9486	8.54
5	1000	14.248	13.4499	13.5612	0.949	0.9432	0.945	6.77
5	∞	13.490	13.490	13.490	0.95	0.95	0.95	

Table 7.2. PIs for model type = 2

error type	n	95% PI slen	PI olen	95% PI dlen	PI scov	95% PI ocov	PI dcov	adf
1	75	4.8071	4.9904	6.1904	0.94	0.9434	0.9766	12.88
1	100	4.7095	4.6949	5.0585	0.9660	0.9604	0.9736	6.27
1	1000	4.3765	4.3236	4.6204	0.963	0.959	0.9704	46.12
1	∞	3.920	3.920	3.92	0.95	0.95	0.95	
2	75	8.3747	8.2412	9.4625	0.9482	0.9456	0.9644	11.26
2	100	8.3039	8.7679	10.3497	0.9404	0.945	0.9728	10.89
2	1000	7.3369	7.1970	7.4934	0.959	0.959	0.9634	28.97
2	∞	6.365	6.365	6.365	0.95	0.95	0.95	
3	75	4.9677	5.2980	6.6525	0.9284	0.935	0.9672	14.13
3	100	4.9332	4.8722	5.9253	0.9446	0.9464	0.971	15.46
3	1000	4.3719	4.0909	4.3665	0.9612	0.9592	0.9692	46.26
3	∞	3.664	2.996	2.996	0.95	0.95	0.95	
4	75	2.4752	2.5300	3.4960	0.917	0.9216	0.976	18.62
4	100	2.3623	2.3270	3.1236	0.9196	0.919	0.977	22.15
4	1000	2.0690	2.0539	2.2859	0.9514	0.9464	0.9816	74.60
4	∞	1.900	1.900	1.900	0.95	0.95	0.95	
5	75	16.978	18.1773	20.5476	0.9112	0.9154	0.9432	11.37
5	100	18.065	17.8100	19.5232	0.9468	0.9464	0.9582	9.42
5	1000	15.706	14.9372	15.2679	0.953	0.9494	0.9512	15.97
5	∞	13.490	13.490	13.490	0.95	0.95	0.95	

Table 7.3. PIs for model type = 3

error type	n	95% PI slen	PI olen	95% PI dlen	PI scov	95% PI ocov	PI dcov	adf
1	75	4.7712	4.9373	5.6490	0.9582	0.963	0.9802	8.28
1	100	4.7095	4.6949	5.0585	0.9660	0.9604	0.9736	6.27
1	1000	4.1032	4.0609	4.1661	0.9644	0.961	0.966	18.26
1	∞	3.920	3.920	3.92	0.95	0.95	0.95	
2	75	8.3486	8.8706	10.0616	0.943	0.948	0.9658	8.57
2	100	8.3805	8.2387	9.0712	0.9586	0.96	0.9716	8.08
2	1000	6.7662	6.649	6.7732	0.9546	0.9526	0.957	13.49
2	∞	6.365	6.365	6.365	0.95	0.95	0.95	
3	75	4.7964	4.9050	5.7060	0.949	0.9528	0.9722	9.56
3	100	4.7997	4.5859	5.1645	0.9624	0.9592	0.9744	9.57
3	1000	3.9561	3.5351	3.6259	0.959	0.9566	0.96	18.27
3	∞	3.664	2.996	2.996	0.95	0.95	0.95	
4	75	2.4028	2.4515	2.9349	0.9572	0.9592	0.9868	10.66
4	100	2.3235	2.3068	2.6842	0.9594	0.9558	0.9872	11.70
4	1000	1.9703	1.9618	2.0340	0.9572	0.954	0.9748	25.26
4	∞	1.900	1.900	1.900	0.95	0.95	0.95	
5	75	17.321	18.3157	20.5934	0.9282	0.929	0.9538	11.06
5	100	18.0382	17.3865	18.8773	0.9432	0.9432	0.953	9.005
5	1000	14.6055	13.7890	13.9609	0.9558	0.9518	0.9516	9.19
5	∞	13.490	13.490	13.490	0.95	0.95	0.95	

CHAPTER 8

CONCLUSIONS

Simulations were done in *R*. See R Core Team (2016). The collection of *R* functions *slpack*, available from (http://lagrange.math.siu.edu/Olive/slpack.txt), has some useful functions for the inference. The function `pisimspline` was used to do the simulation.

The `spatial` package is due to Venables and Ripley (2002).

Some references for kriging, spatial statistics, and geostatistics include Atkinson, Riani, and Cerioli (2004, ch. 8), Banerjee, Carlin, and Gelfand (2015), Bivand, Pebesma, and Gómez-Rubio (2013), Brunsdon and Comber (2015), Chilès (2012), Christensen (1991: ch. 6, 2001), Chun and Griffith (2013), Crawley (2013, ch. 26), Cressie (1986, 1993), Cressie and Wilke (2011), Diggle (2014), Isaaks and Srivistava (1989), Kalkhan (2011), Montero, Fernández-Avilés, and Mateau (2015), Oliver and Webster (2015), Ripley (1981), Stein (1999), and Venables and Ripley (2003, ch. 15).

For model type 2 and 3, PIs (6.7) and (6.8) sometimes had undercoverage for n=75 and error type 5. PI (6.12) was sometimes rather long, especially for n=75.

REFERENCES

[1] Atkinson, A., Riani, R., and Cerioli, A. (2004), *Exploring Multivariate Data with the Forward Search*, Springer, New York, NY.

[2] Banerjee, S., Carlin, B.P., and Gelfand, A.E. (2015), *Hierarchical Modeling and Analysis for Spatial Data,* 2nd ed., Chapman & Hall/CRC, Boca Rotan, FL.

[3] Bivand, R.S., Pebesma, E., and Gómez-Rubio, V. (2013), *Applied Spatial Data Analysis with R*, 2nd ed., Springer, New York, NY.

[4] Brunsdon, C., and Comber, L. (2015), *An Introduction to R for Spatial Analysis and Mapping,* Sage Publications, Thousand Oaks, CA.

[5] Chilès, J.-P. (2012), *Geostatistics: Modeling Spatial Uncertainty*, 2nd ed., Wiley, Hoboken, NJ.

[6] Christensen, R. (1987), *Plane Answers to Complex Questions: the Theory of Linear Models,* Springer, New York, NY.

[7] Christensen, R. (1991), *Linear Models for Multivariate, Time Series, and Spatial Data*, Springer, New York, NY.

[8] Christensen, R. (2001), *Advanced Linear Modeling*, 2nd ed., Springer, New York, NY.

[9] Chun, Y., and Griffith, D.A. (2013), *Spatial Statistics & Geostatistics*, Sage Publications, Thousand Oaks, CA.

[10] Crawley, M.J. (2013), *The R Book,* 2nd ed., Wiley, Hoboken, NJ.

[11] Cressie, N. (1986), "Kriging Nonstationary Data," *Journal of the American Statistical Association*," 81, 625-634.

[12] Cressie, N. (1993), *Statistics for Spatial Data,* revised ed., Wiley, Hoboken, NJ.

[13] Cressie, N., and Wilke, C.K. (2011), *Statistics for Spatio-Temporal Data,* Wiley, Hoboken, NJ.

[14] Diggle, P.J. (2014), *Statistical Analysis of Spatial and Spatio-Temporal Point Pat-*

terns, 3rd ed. Chapman & Hall/CRC, Boca Raton, FL.

[15] Frey, J. (2013), "Data-Driven Nonparametric Prediction Intervals," *Journal of Statistical Planning and Inference,* 143, 1039-1048.

[16] Freedman, D.A. (2005), *Statistical Models Theory and Practice,* Cambridge University Press, New York, NY.

[17] Isaaks, E.H., and Srivistava, R.M. (1989), *An Introduction to Applied Geostatistics*, Oxford University Press, New York, NY.

[18] Johnson, R.A., and Wichern, D.W. (1988), *Applied Multivariate Statistical Analysis,* 2nd ed., Prentice Hall, Englewood Cliffs, NJ.

[19] Kalkhan, M.A. (2011), *Spatial Statistics: Geospatial Information Modeling and Thematic Mapping*, CRC Press, Boca Raton, FL.

[20] Laslett, G.M. (1994), "Kriging and Splines: an Empirical Comparison of Their Predictive Performance in Some Applications," *Journal of the American Statistical Association*, 89, 391-400.

[21] Montero, J.M., Fernández-Avilés, G., and Mateau, J. (2015), *Spatial and Spatio-temporal Geostatistical Modeling and Kriging*, Wiley, Chichester.

[22] Olive, D.J. (2013), "Asymptotically Optimal Regression Prediction Intervals and Prediction Regions for Multivariate Data," *International Journal of Statistics and Probability*, 2, 90-100.

[23] Olive, D.J. (2017a), *Linear Regression*, Springer, New York, NY.

[24] Olive, D.J. (2017b), *Robust Multivariate Analysis*, Springer, New York, NY, to appear.

[25] Olive, D.J. (2017c), *Prediction and Statistical Learning*, online course notes, see (http://lagrange.math.siu.edu/Olive/slearnbk.htm).

[26] Oliver, M.A., and Webster, R. (2015), *Basic Steps in Geostatistics: The Variogram and Kriging,* Springer, New York, NY.

[27] Pelawa Watagoda, L.C.R., and Olive, D.J. (2017), "Inference for Mul-

tiple Linear Regression After Model or Variable Selection," preprint at (http://lagrange.math.siu.edu/Olive/ppvsinf.pdf).

[28] R Core Team (2016), "R: a Language and Environment for Statistical Computing," R Foundation for Statistical Computing, Vienna, Austria, (www.R-project.org).

[29] Ripley, B.D. (1981), *Spatial Statistics*, Wiley, New York, NY.

[30] Seber, G.A.F., and Lee, A.J. (2003), *Linear Regression Analysis,* 2nd ed., Wiley, New York, NY.

[31] Stein, M.L. (1999), *Interpolation of Spatial Data: Some Theory for Kriging*, Springer, New York, NY.

[32] Venables, W.N., and Ripley, B.D. (2003), *Modern Applied Statistics with S,* 4th ed., Springer, New York, NY.

[33] Yakowitz, S.J., annd Szidarovszky, F. (1985), "A Comparison of Kriging with Nonparametric Regression Methods," *Journal of Multivariate Analysis*, 16, 21-53.

VITA

Graduate School
Southern Illinois University

Chamila K. Ranaweera

rcranaweera@yahoo.com

University of Peradeniys, Bachelor of Science, Geology, May 2005

Southern Illinois University Carbondale, Master of Science in Geology, May 2015

Research Paper Title: Kriging and nonparametric regression

Major Professor: Dr. David J. Olive