BAYESIAN MULTIVARIATE INTERPOLATION WITH MISSING DATA AND ITS APPLICATIONS

by

WEIMIN SUN

B.Sc., University of Electric Science and Technology of China, 1983

M.Sc., University of Georgia, 1990

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

Department of Statistics

We accept this thesis as conforming

to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

December 1994

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

(Signature)

Department of Statistics

The University of British Columbia
Vancouver, Canada

Date Dec. 8, 1994

DE-6 (2/88)

Abstract

This thesis develops Bayesian multivariate interpolation theories when there are: (i) data missing-by-design; (ii) randomly missing data; (iii) monotone missing data patterns.

Case (i) is fully discussed both theoretically and empirically. A predictive distribution yields a Bayesian interpolator with associated standard deviation, a simultaneous interpolation region, and a hyperparameter estimation algorithm. These results are described in detail. The method is applied to interpolating data from Southern Ontario Pollution. An optimal redesign of a current network is proposed. A cross-validation study is conducted to judge the performance of our method. The method is compared with a Co-kriging approach to which the method is meant to be an alternate.

Case (ii) is briefly discussed. An approximation of a matrix T-distribution by a normal distribution is explored for obtaining an approximate predictive distribution. Based on the approximate distribution, an approximate Bayesian interpolator and an approach for estimating hyperparameters by the EM algorithm are described.

Case (iii) is only touched on. Only an iterative predictive distribution is derived. Further study is needed for finding ways of estimating hyperparameters.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank Prof. James Zidek for his excellent guidance, for his fundamental influence on my "statistical thinking" and for providing me with a opportunity to work on a practical problem. I would like to thank Dr. Nhu Le for the many fruitful discussions I had with him.

I am indebted to Dr. Rick Burnett for his comments and for providing me with the data used in my investigation. I thank Dr. Tim Hass; only with his generous help was the comparison of the method in this thesis with his own method feasible. I thank Mr. Hongbin Zhang, our systems analyst, for his support. I thank Mr. Rick White for providing part of the C codes I needed for my analysis. I thank the members of my Supervisory Committee for their help.

Very special thanks go to Ms. Chun Zhang, the McConnells and the Veecks for their substantial help.

Finally, I would like to thank the University and its Department of Statistics along with Health Canada for their financial support during my graduate studies.

Chapter 1

Introduction

A *spatial data* set consists of a collection of measurements or observations on one or more attributes taken at specified locations. At a fixed time, those data are observed over a restricted geographical or other region. Many models and theories have been established for spatial data analyses. The following several paragraphs give a brief overview of this newly established field.

Cressie (1991a) defined a general spatial model. Let $\mathbf{s} \in R^d$ be a generic sampling site in d-dimensional Euclidean space and $\mathbf{X}(\mathbf{s})$ the response vector at $\mathbf{s}$. As $\mathbf{s}$ varies over the index set $D \subset R^d$, $\{\mathbf{X}(\mathbf{s}) : \mathbf{s} \in D\}$ represents a multivariate random field (or alternatively, process). According to the form of D, spatial data analyses can be classified into four types. When D is a fixed, non-degenerate convex subset of R^d and $\mathbf{X}(\mathbf{s})$ is a random vector at site $\mathbf{s} \in D$, such analysis is called *geostatistical data analysis.* When D is a fixed collection of countably many points of R^d and $\mathbf{X}(\mathbf{s})$ is a random vector at site $\mathbf{s} \in D$, it is called *lattice data analysis.* When D is a random point process and $\mathbf{X}(\mathbf{s})$ is a random vector at site $\mathbf{s} \in D$, it is called *point pattern data analysis.* When D is a random point process and $\mathbf{X}(\mathbf{s})$ is a random set, it is called *object data analysis.* Cressie (1991a) gives a comprehensive survey of spatial data analyses.

An earlier form of spatial data occurred as a data map centuries ago (Halley 1686). Recently spatial data have been seen in many applications. For example, the grade of ore deposit is measured at various sites spread over a constrained geographical area. In terms of the general spatial model, D is the constrained geographical area, d is two, $\mathbf{s}$ is a two dimensional vector of the longitude and latitude of a site and $\mathbf{X}(\mathbf{s})$ is the measured ore deposit grade at location $\mathbf{s}$. Another example is the remotely sensed yield of wheat on earth at a fixed time. In the ore grade example, the ore grade at a site is likely similar to the grade at a nearby site. However, it will be less similar to that at a faraway site, provided these two sites are not on the same ore deposition ridge. It means an underlying relationship among the ore deposit grades does exist. This underlying relationship is called "spatial correlation". Spatial correlation plays an important role in spatial inference.

While the ore deposition grade does not change over time, observations like humidity, temperature and wind within a region are different from time to time. So they have to be measured not only across sites , but also over time. Such data are called *spatial-temporal* data. Another example of spatial-temporal data comes from air pollution monitoring. There, air pollution levels, e.g. sulphate or nitrate levels, are observed at different sites and regularly, say hourly, for a segment of time duration.

Among the four types of spatial analyses, geostatistics is most relevant to the topic of this dissertation. Geostatistics was established in the early 1980s as a hybrid discipline of mining engineering, geology, mathematics and statistics. Its study began with Matheron's early 1960's papers on *Kriging*, a name given by Matheron after D. G. Krige, a South African mining engineer who developed empirical methods to determine true ore grade distribution based on sampled ore grades. In Kriging, the model of a random process $\mathbf{X}(\mathbf{s})$ generally consists of two terms: a trend and an error. The trend

term catches large-scale variation of $\mathbf{X}(\mathbf{s})$ and is deterministic. The error term reflects small-scale variation of $\mathbf{X}(\mathbf{s})$ and is a random process. The Kriging approach gives a "best linear unbiased estimator" (BLUE) of the unknown ore-grade at sites in the prediction region using ore-grade samples from neighboring sites by exploring correlations among ore-grades at different sites. Later, in other applications, many different forms of Kriging are developed. Examples are ordinary Kriging; simple Kriging; universal Kriging; Bayesian Kriging; robust Kriging and Co-Kriging. These Kriging methods also give best linear predictors. Thus Kriging has become synonymous with *optimal spatial linear prediction.*

Before exploring the family of Kriging methods further, we define some basic concepts. In Kriging, the spatial correlation is expressed in terms of the *Variogram*, defined as $Var(\mathbf{X}(\mathbf{s_1}) - \mathbf{X}(\mathbf{s_2}))$, $\mathbf{s_1}, \mathbf{s_2} \in D$. When $E(\mathbf{X}(\mathbf{s_1}) - \mathbf{X}(\mathbf{s_2})) = 0$, $E(\mathbf{X}(\mathbf{s_1}) - \mathbf{X}(\mathbf{s_2}))^2 = Var(\mathbf{X}(\mathbf{s_1}) - \mathbf{X}(\mathbf{s_2}))$. The equation says that if the mean function of a random field is a constant, the variogram and expected mean squared difference are the same. Sometimes, the variogram has other names. For example, it is called *Dispersion* in Sampson and Guttorp (1992) (SG hereafter). Another important concept in Kriging is *intrinsic stationarity.* An intrinsically stationary process, $\mathbf{X}(\mathbf{s})$, has: (i)a constant mean for all $\mathbf{s} \in D$; (ii) $Var(\mathbf{X}(\mathbf{s_1}) - \mathbf{X}(\mathbf{s_2})) = 2r(\mathbf{s_1} - \mathbf{s_2})$, where $r(\cdot)$ is a real, non-negative function in R^d. In Cressie (1991a) $2r(\cdot)$ is called variogram and $r(\cdot)$ semi-variogram. The concept of variogram implicitly implies intrinsic stationarity. If (ii) is replaced by $Cov(\mathbf{X}(\mathbf{s_1}), \mathbf{X}(\mathbf{s_2})) = C(\mathbf{s_1} - \mathbf{s_2})$, $\mathbf{X}(\mathbf{s})$ is a second-order stationary process. The function $C(\cdot)$ is called a *covariogram.* If further, $2r(\mathbf{s_1} - \mathbf{s_2}) = 2r(\| \mathbf{s_1} - \mathbf{s_2} \|)$ ($C(\mathbf{s_1} - \mathbf{s_2}) = C \| \mathbf{s_1} - \mathbf{s_2} \|$), $2r(\cdot)$ ($C(\cdot)$) is isotropic. The cross-variogram and cross-covariogram between the random fields $\mathbf{X}_i(\mathbf{s})$ and $\mathbf{X}_j(\mathbf{s})$ have similar definitions.

Intrinsic stationarity is a strictly weaker assumption than second-order stationarity.

From $2r(\mathbf{h}) = 2(C(\mathbf{0}) - C(\mathbf{h}))$, we can prove that second-order stationarity implies intrinsic stationarity. However, the opposite is not true. For example, a Brownian motion is an intrinsically stationary process but not a second-order stationary process.

In many applications, the variogram (covariogram) is unknown. One has to estimate it. There are both parametric and nonparametric approaches for estimating variogram within a constrained region. Usually a parametric approach involves two steps. First, lags, $\mathbf{h_1}$, ..., $\mathbf{h_n}$, are chosen and sample variograms (covariograms) at these lags are estimated using observed data. Second, a proper, often an isotropic, variogram (covariogram) model is chosen and the model parameters are determined by fitting it to sample variograms (covariograms). Matheron (1962) proposes a natural way for the computation of a sample variogram (covariogram) by the method of moments. His estimator is unbiased but not robust. Cressie and Hawkins (1980) propose two robust sample variogram estimators. For choosing a proper variogram (covariogram) model, precautions are needed. For example, a variogram model should satisfy the *conditionally negative-definiteness* condition. That is, $\sum_{i=1}^{k}\sum_{j=1}^{k} a_i a_j \, 2\gamma(\mathbf{s_i} - \mathbf{s_j}) \leq 0$ for any real vector $\mathbf{a} \in R^k$ $\mathbf{s_1}, \ldots, \mathbf{s_k} \in D$ and any integer k. Similarly, a covariogram model must satisfy the *positive-definiteness* condition. That is, $\sum_{i=1}^{k}\sum_{j=1}^{k} a_i a_j \, C(\mathbf{s_i} - \mathbf{s_j}) \geq 0$ for any real vector $\mathbf{a} \in R^k$, $\mathbf{s_1}, \ldots, \mathbf{s_k} \in D$ and any integer k. There are other considerations too. For example, a variogram model should be able to reflect the sill, observed data may present a sill being defined as a variogram's non-zero limit at lag zero. Many isotropic variogram models have been proposed. Examples of such models are the rational quadratic model: $\gamma(\mathbf{h}) = a^2 \parallel \mathbf{h} \parallel^2 /(1+ \parallel \mathbf{h} \parallel^2)$,$\mathbf{h} \in R^d$ (Schoenberg 1938); (the Gaussian model) $\gamma(\mathbf{h}) = a^2\{1 - exp(- \parallel \mathbf{h} \parallel^2)\}$, $\mathbf{h} \in R^d$; the linear model; the exponential model and the spherical model (Journel and Huijbregts 1978). A covariogram model is chosen similarly. A natural model fitting criterion is "least squares". Other possible criteria include maximum likelihood, restricted maximum likelihood and

minimum norm quadratic. When observed data do not confirm a stationary or isotropic condition, measures are taken to make data stationary or isotropic. For example, Hass (1990a, 1990b, 1992, 1993) adopts a moving window approach so that all the observations inside a circular window are approximately isotropic.

SG describe a nonparametric approach to estimating a spatial dispersion matrix when the observed data are not stationary. Here a dispersion matrix has the same meaning as a variogram matrix except that no stationarity assumption is implicitly implied. Their method takes two steps. First, with the "nonmetric multidimensional scaling" (MDS) algorithm (see Mardia, Kent and Bibby 1979), a two-dimensional representation of sampling sites is sought. In this two dimensional Euclidean space, called the $D-plane$, a monotonic function, g, of the distance between any two points approximates the spatial dispersion between the two points. As a counterpart of the D-plane, the geographical coordinates plane of the sampling sites is called the $G-plane$. Second, thinplate splines, f, are found to provide smooth mappings of the G-plane representation into the D-plane representation. Then, the composition of f and g yields a nonparametric estimator of $var(Z(x_a) - Z(x_b))$ at any two geographic locations x_a and x_b. Other nonparametric estimation methods for a spatial covariance matrix are discussed , for example, in Loader and Switzer (1991), Leonard and Hsu (1992), Le and Zidek (1992) and Pilz (1991). While both adopting Bayesian approaches, Le and Zidek (1992) take conjugate priors, inverted Wishart, on the spatial covariance matrix while Leonard and Hsu (1992) propose a class of prior distributions, other than the inverted Wishart.

With the above preliminaries, we can now summarize ordinary Kriging as follows. Assume a model, $X(\mathbf{s}) = \mu + \delta(\mathbf{s})$, $\mathbf{s} \in D$, $\mu \in R$, where D is a convex subset in R^d, $\delta(\cdot)$ is a stochastic process and μ is an unknown, constant scalar. Kriging searches for an optimal linear predictor of any unobserved value $\mathbf{X}(\mathbf{s_0})$ within a family of linear func-

tions of $\mathbf{X}(\mathbf{s_i})$, $i = 1, \ldots, n$. Therefore the predictor takes the form $p(\mathbf{X};\ \mathbf{s_0}) = \sum_{i=1}^{n} \lambda_i$ $\mathbf{X}(\mathbf{s}_i)$ under the restriction, $\sum_{i=1}^{n} \lambda_i = 1$. The restriction makes $p(\mathbf{X};\ \mathbf{s_0})$ unbiased, since $E(p(\mathbf{X};\ \mathbf{s_0})) = \mu \sum_{i=1}^{n} \lambda_i = \mu = E(\mathbf{X}(\mathbf{s_0}))$.

Suppose $\mathbf{X}(\mathbf{s})$ is an intrinsically stationary process. By minimizing

$$\left(\mathbf{X}(\mathbf{s}_0) - \sum_{i=1}^{n} \lambda_i \mathbf{X}(\mathbf{s}_i)\right)^2 + 2m(1 - \sum_{i=1}^{n} \lambda_i) = 0, \tag{1.1}$$

optimal choices of $\lambda^t = (\lambda_1, \ldots, \lambda_n)$ and m have the forms,

$$\lambda^t = \left(\gamma + \mathbf{1}\frac{1 - \mathbf{1}^t\Gamma^{-1}\gamma}{\mathbf{1}^t\Gamma^{-1}\mathbf{1}}\right)^t \Gamma^{-1}$$

and

$$m = -(1 - \mathbf{1}^t\Gamma^{-1}\gamma)/(\mathbf{1}^t\Gamma^{-1}\gamma),$$

where $\gamma = (\gamma(\mathbf{s}_0 - \mathbf{s}_1), \ldots, \gamma(\mathbf{s}_0 - \mathbf{s}_n))^t$, an $n \times 1$ matrix and Γ is an $n \times n$ matrix whose $(i,j)^{th}$ element is $\gamma(\mathbf{s}_i - \mathbf{s}_j)$.

When $\mathbf{X}(\mathbf{s})$ is second-order stationary, the above optimal solution can be expressed in terms of covariograms. The solutions for λ and m are obtained by replacing λ with c, where $c = (C(\mathbf{s_0} - \mathbf{s_1}), \ldots, C(\mathbf{s_0} - \mathbf{s_n}))$, Γ with $\Sigma = (C(\mathbf{s_i} - \mathbf{s_j}))$ and m with -m. If the variogram or covariogram function is known, ordinary Kriging stops here. If unknown, it is estimated with either a parametric or nonparametric method.

When μ has a more complicated form, other Kriging approaches are developed. In simple Kriging (Matheron 1971), the model is $\mathbf{X}(\mathbf{s}) = \mu(\mathbf{s}) + \delta(\mathbf{s})$, where $\mu(\cdot)$ is known. The best linear predictor is sought within the family $p(\mathbf{X}; \mathbf{s}_0) = \sum_{i=1}^{n} l_i \mathbf{X}(\mathbf{s}_i) + k$ subject to an unbiasedness restriction. By minimizing $E(\mathbf{X}(\mathbf{s}_0) - p(\mathbf{X}; \mathbf{s}_0))^2$ over $l^t = \{l_1, \ldots, l_n\}$ and k, the optimal solution is, $k = \mu(\mathbf{s}_0) - \sum_{i=1}^{n} l_i \mu(\mathbf{s}_i)$ and $l^t = C^t\Sigma^{-1}$ where $C = (C(\mathbf{s}_0, \mathbf{s}_1), \ldots, C(\mathbf{s}_0, \mathbf{s}_n))^t$ and $\Sigma_{n\times n} = (C(\mathbf{s}_i, \mathbf{s}_j))$.

Universal Kriging is introduced when a more general form of μ is assumed. That general form is $\mathbf{X}(\mathbf{s}) = \sum_{i=1}^{p+1} f_{j-1}(\mathbf{s})\beta_{j-1} + \delta(\mathbf{s})$, where $\beta = (\beta_0, \ldots, \beta_p)^t \in R^{p+1}$ is an unknown vector of parameters, $f_{j-1}(\cdot), j = 1, \ldots, p+1$ are known functions and $\delta(\cdot)$ is a zero mean intrinsically stationary process with variogram $2\gamma(\cdot)$. The form of a best linear unbiased predictor is $p(\mathbf{X}; \mathbf{s}_0) = \sum_{i=1}^{n} \lambda_i \mathbf{X}(\mathbf{s}_i)$, subject to, $\lambda^t X = \mathbf{x}^t$, $\lambda = (\lambda_1, \ldots, \lambda_n)$, where X is an $n \times (p+1)$ matrix, its $(i,j)^{th}$ element being $f_{j-1}(\mathbf{s}_i)$, and $\mathbf{x} = (f_0(\mathbf{s}_0), \ldots, f_p(\mathbf{s}_0))^t$. If the variogram is known, solutions λ_i, $i = 1, \ldots, n$ are easily derived. If unknown, it needs to be estimated. One problem occurs when one estimates the variogram. Since in universal Kriging $\mathbf{X}(\mathbf{s})$ is not intrinsically stationary, the sample variograms computed with observed data, $\mathbf{X}(\mathbf{s_i})$, $i = 1 \ldots, n$ by the formulas of Matheron (1962), and Gressi and Hawkins (1980) are biased. To obtain unbiased variogram estimators, the residuals of $\mathbf{X}(\mathbf{s})$ must be known; but they are unknown, since β is generally unknown. To bypass this dilemma, Neuman and Jacobson (1984; see also Haas 1993) propose an iterative method starting with ordinary least square (o.l.s.). Note that the model for the vector $\mathbf{X}^t = (\mathbf{X}(\mathbf{s}_1), \ldots, \mathbf{X}(\mathbf{s}_n))$ can be rewritten in a matrix form, $\mathbf{X} = \mathbf{Z}\beta + \delta$, where $\mathbf{Z}$ is an $n \times (p+1)$ matrix with its $(i,j)^{th}$ element, $i = 1, \ldots, n$, $j = 0, 1, \ldots, p$ being $f_j(\mathbf{s}_i)$; β is regression coeficients and δ is a stochatic process vector. The iterative procedure is as follows. First, estimate β by $\hat{\beta}_{\mathbf{ols}} = (\mathbf{Z}^t\mathbf{Z})^{-1}\mathbf{Z}^{-1}\mathbf{X}$ and obtain $\hat{\Sigma}$ based on residuals, $\mathbf{X} - \mathbf{Z}\hat{\beta}_{\mathbf{ols}}$. Second, update β by $\hat{\beta}_{\mathbf{gls}} = (\mathbf{Z}^t\hat{\Sigma}^{-1}\mathbf{Z})^{-1}\mathbf{Z}^{-1}\hat{\Sigma}^{-1}\mathbf{X}$ and $\hat{\Sigma}$ with the updated residuals, $\mathbf{X} - \mathbf{Z}\hat{\beta}_{\mathbf{gls}}$. Repeat the above procedure until it converges.

In previous Kriging approaches, an estimate of $\mathbf{X}(\mathbf{s_0})$ is computed using information on the same process $\mathbf{X}(\mathbf{s_i})$ $i = 1, \ldots, n$. In some applications, additional information is available. In these cases, realizations of other correlated random processes are observed. CoKriging was developed to bring the additional information into the BLUE predictor. More specifically, if the observed data set is $\mathbf{X} = (\mathbf{X}(\mathbf{s}_1), \ldots, \mathbf{X}(\mathbf{s}_n))^t$, an $n \times k$ matrix with $(i,j)^{th}$ element being $\mathbf{X}_j(\mathbf{s}_i)$, $i = 1, \ldots, n$, $j = 1, \ldots, k$, one needs to predict

$\mathbf{X}_1(\mathbf{s}_0)$ using $\mathbf{X}$. Suppose $E(\mathbf{X}_j(\mathbf{s})) = \mu_,, j = 1, \ldots, k$, $\mathbf{s} \in D$ and $cov(\mathbf{X}(\mathbf{s}), \mathbf{X}(\mathbf{u})) = C(\mathbf{s}, \mathbf{u})$, $\mathbf{s}, \mathbf{u} \in D$, where $\mu = (\mu_1, \ldots, \mu_k)^t$ and $C(\mathbf{s}, \mathbf{u})$ is a $k \times k$ matrix. The best linear CoKriging predictor of $\mathbf{X}_1(\mathbf{s}_0)$ takes the form $p_1(Z; \mathbf{s}_0) = \sum_{i=1}^{n} \sum_{j=1}^{k} \lambda_{ji} \mathbf{X}_j(\mathbf{s}_i)$ with $\sum_{i=1}^{n} \lambda_{1i} = 1$, $\sum_{i=1}^{n} \lambda_{ji} = 0$, $j = 2, \ldots, k$. The remaining steps are the same as those for other Kriging methods.

Applications of the CoKriging method in environmetrics, soil science, and hydrology can be found in Haas (1993) , Yate and Warrick (1987), as well as Ahmed and de Marsily (1987). Haas (1993) proposes a *Moving Window Regression Residual CoKriging* (MWRRCK) method for predicting pollution carried in rainfall. There, the observed information includes: (i) wet deposition of pollutants monitored at 200 sampling sites in US; (ii) observations of precipitation at over 6000 sites of National Weather Service (NWS) network. We outline MWRRCK method below.

First, a moving circular window centered at a prediction site is selected to achieve local isotropy. A radius of the circular window is chosen so that the total number of monitoring sites inside the window will not be less than a predetermined value. This number is set to make sample variograms reasonably accurate. Second, the spatial trend surface (respectively spatial covariance matrix) in the window is removed (respectively estimated) through an iterative o.l.s-g.l.s-procedure. Third, a regular CoKriging method is applied to the residual process for estimating the residual at the prediction site. The final prediction is a sum of the predicted trend and the predicted residual.

In Hass's theory, care has been taken to make the covariance matrix within a window positive definite. However, the covariance matrix between windows need not be positive definite. The problem arises when two prediction sites are geographically close and one of the two sites falls inside the circular window of the other. Because different fitted

semivariogram and cross-semivariogram models are obtained for each window, it can happen that the covariance matrix between these two sites is negative.

Although Kriging is an appealingly simple method, it tends to understate uncertainties of predicted values, since model uncertainty is not included. One simple example given in Handcock and Stein (1993) shows how much effect the model uncertainty has on the confidence interval (CI) of a predicted value. There, the above authors showed that if the uncertainty of the unknown spatial covariance is not included, the 95% CI of a predicted elevation based on measured elevations is $(699, 707)$. When it is included by using a Bayesian approach, the 95% CI becomes $(694, 713)$. This example clearly supports the use of Bayesian approach to Kriging. Such Kriging are described as *Bayesian.*

Bayesian Kriging is relatively new. Not much work has been done in geostatistics. For a general model $\mathbf{X}(\mathbf{s}) = \mu(\mathbf{s}) + \delta(\mathbf{s})$, $\mathbf{s} \in D$, a Bayesian approach can be adopted by assuming $\mu(\mathbf{s})$ to be a random process that is independent of $\delta(\cdot)$. More specially, $\mu(\mathbf{s}) = \sum_{j=1}^{p+1} \beta_{j-1} f_{j-1}(\mathbf{s})$, $\mathbf{s} \in D$,where β_j, $j = 0, \ldots, p$ are random variables (see Nather 1985, Kitanidis 1986, and, Omre and Halvorsen 1989). For example, Omre and Halvarsen (1989) describe a version of Bayesian Kriging that puts prior on the mean function only. Their method appears to be a direct extension of traditional Bayesian linear model to case of spatial data. An empirical Bayesian predictor is obtained if the parameters of the prior are estimated from the data and substituted into the Bayesian predictor of $\mathbf{X}(\mathbf{s}_0)$. Similarly, one can assume the covariogram varies in the space of all positive-definite functions $P_D = \{C(\mathbf{s}, \mu) : \mathbf{s}, \mu \in D\}$ and put a (prior) distribution on P_D. Or one can assume a structural model on $C(\cdot)$, $C(\mathbf{s}, \mu; \theta)$ and put a prior on θ. For the latter case, the predictive density is

$$f(\mathbf{X}(\mathbf{s}_0) \mid \mathbf{X}) = \int_\Theta f(\mathbf{X}(\mathbf{s}_0) \mid \mathbf{X}, \theta) f(\theta \mid \mathbf{X}) d\theta.$$

One example of such an approach is proposed by Handcock and Stein (1993). In Hand-

cock and Wallis (1994), the method is applied to model meteorological fields in an attempt to assess the global warming problem from a statistical point of view.

One competitor to Kriging is the *smoothing spline* (Wahba 1990a, 1990b). The spline method can be briefly summarized as follows. For observed data, X_i, $i = 1, \ldots n$, the assumed model is $X_i = f(s_i) + e_i$, where s_i, $i = 1, \ldots, n$ are the locations of measurements, $f(\cdot)$ is a smooth function and e_i, $i = 1, \ldots, n$, are independently, identically distributed (i.i.d.) errors. For each value of γ, the smoothing function, $f(s)$, is estimated by minimizing

$$\sum_{j=1}^{n} \{X_i - f(s_i)\}^2 + \lambda \int [f''(x)]^2 dx$$

over all f with continuous first and squared-integerable second derivatives. The smooth parameter λ is chosen by "generalized cross-validation" (Wahba 1990b). The interpolated value at any location s_0 is taken as $\hat{f}(s_0)$. Oehlert (1993) shows one example which combines a smoothing spline technique with a Bayesian approach. There, Oehlert proposes a multiple time series model for data that have both temporal and spatial correlations. The smoothing spline is used when the mean and trend are extended from the rectangles where there are the monitoring sites to rectangles with no monitors. Many empirical comparisons have suggested that the interpolation performances of spline methods and Kriging methods are similar (see Laslett (1994) for references to these comparisons). Laslett (1994) demonstrates that in certain cases, Kriging surpasses splines by a big margin.

When the observed data have a spatial-temporal form, Kriging faces another rival, a "hierarchal Bayesian time series approach" developed by Le and Zidek (1992) (LZ hereafter). LZ assume independent responses over time, a common unknown spatial covariance structure at each fixed time point and conjugate priors on both trend parameters and the spatial covariance. As an alternative to Kriging, LZ's method has many

advantages. For example, it incorporates the model uncertainty into its interpolator. It uses incoming information to dynamically update estimation and gives a predictive distribution that enables construction of a simultaneous band for interpolated values. In its original form, the LZ method does not allow such data. In this thesis, we extend the LZ method to include missing data.

We develop and explain the extension in the context of interpolating air pollution. Let's first describe this context. Assume there are s sites scattering over a constrained region. At each site, k response values are measured. Examples of such measurements are air pollution levels, like sulphate or nitrate levels. Among the s sites, s_u sites are ungauged sites, where there are no observations but air pollution levels are needed. The other s_g sites are gauged; there are observations. In the case of missing data, some values are missing at the gauged sites. The missing data patterns discussed in the next chapter are called *randomly missing*, *missing-by-design* and *monotone missing*. Here, we use the term "randomly missing" to mean its probability of being missing is not related to its value (see Little and Rubin 1987). The term "monotone missing" is also from Little and Rubin (1987). The meanings of the three missing data patterns are explained in the next chapter.

Let X_t denote the complete, random response vector for all sites (both gauged and ungauged) at time t, where the first k elements represent pollution levels for k pollutants at site one, the second k elements for site two and so on. Thus, X_t is an sk-dimensional vector. The inferential objective treated in this thesis is to interpolate unobserved pollution levels at ungauged sites using incomplete observations at gauged sites. As in LZ, we assume a linear, Multivariate Gaussian model and conjugate priors on its parameters. Interpolation with missing data consists of two parts. First, by fixing hyperparameters, we find a predictive distribution and its posterior mean along with

a standard error. Second, we develop an estimation procedure for hyperparameters. Further, in two steps we estimate the hyper-covariance matrix of X_t. In step one, we adopt an EM algorithm to estimate the hypercovariance matrix at gauged sites. In step two, we apply the GS approach to extend this hyper-covariance matrix to all sites.

This thesis consists of six chapters. In Chapter 2, we describe three different interpolation theories depending on the missing data patterns. The patterns are: (i) missing-by-design; (ii) randomly missing; (iii) monotone. There, we fully develop the theory of interpolation with data missing-by-design; we briefly discuss the theory of interpolation with randomly missing data and only touch the theory of interpolation with monotone missing data. In Chapter 3, we apply the theory of interpolation with data missing-by-design to Southern Ontario pollution data; we implement it with S and C programs and carry out residual analysis. In Chapter 4, by combining the theory developed in Chapter 3 and the theory developed in Caselton, Kan and Zidek (1992), we show how to apply our results to an optimal network redesign problem. In Chapter 5, we compare the theory of "interpolation with data missing-by-design" with the general theory of LZ and also with Hass's CoKriging method. In Chapter 6, we drawn conclusions and list some future research topics. All figures referred in Chapter 3 and Chapter 5 are listed in Appendix C. We attach examples of S and C programs in Appendix A and B.

Chapter 2

Bayesian Multivariate Interpolation with Missing Data

2.1 Introduction

The problem of interpolating a spatial field has received a lot of attention. Kriging offers a well-known solution, but it has deficiencies. In particular, it overstates the precision of its interpolator because it fails to reflect the model uncertainty (LZ, Brown, Le and Zidek 1994a, hereafter BLZ). To avoid these deficiencies, LZ propose a hierarchical Bayesian alternative to Kriging. The LZ method takes a Bayesian time series approach with a Gaussian linear model and conjugate priors for the model parameters. Such a Bayesian approach incorporates model uncertainty and yields a heavier-than-normal tailed posterior distribution, the multivariate t. LZ's Bayesian alternative also has the advantage of dynamically updating the predictor as more observations come in. LZ developed their theory in a univariate case where at each of s sites, only one air pollutant is monitored.

BLZ extend the above LZ Bayesian interpolation theory to the multivariate case. At each site k air pollutants are monitored. BLZ adapt the original LZ Bayesian theory by stacking rows of the $s \times k$ response matrix into an $sk \times 1$ vector, where each row represents

k measurements at a gauged site. To reduce the total number of hyperparameters in the prior distributions, BLZ assume a Kronecker product structure on the hypercovariance matrix. Then BLZ describe an algorithm for hyperparameter estimation. But the BLZ theory has an important restriction. It does not permit missing values at gauged sites.

In this chapter, theories of Bayesian multivariate interpolation with different patterns of missing data are discussed. These patterns are: (i) missing-by-design ; (ii) randomly missing and (iii) monotone. In all three cases, we follow the hierarchical Bayesian approach of LZ. In the case of data missing-by-design, the proposed Bayesian interpolation method is an alternative to Co-Kriging. In a Bayesian analysis, a predictive distribution plays a key role and we need to derive that predictive distribution of the unknown pollution levels at ungauged sites given the observed at gauged sites.

Section 2 gives a brief review of the LZ theory. Section 3 serves as a technical preparation for the following sections. There, we define a matrix T-distribution, which plays a pivotal role in our inference, and explore its normal approximation. Section 4 spells out the theory of Bayesian multivariate interpolation with data missing-by-design. Section 5 gives a brief discussion to interpolation theory with randomly missing data. There, we describe an approximate predictive distribution and estimation of hyperparameters. Section 6 is about interpolation theory with monotone missing data patterns. There, only a recursive predictive distribution is described. We put the proofs of most theorems, lemmas and corollaries appearing in this Chapter, in the last section.

2.2 Bayesian Interpolation

In this section, the LZ Bayesian interpolation theory is briefly summarized. Let X_t be an s-dimensional random vector at time t, where the first s_u elements, denoted by X_t^1, correspond to the unobserved responses at s_u ungauged sites. The remaining

s_g elements, denoted by X_t^2, correspond to the observed responses at s_g gauged sites. Assume:

$$X_t \mid z_t, B, \Sigma \overset{\text{independent}}{\sim} N_s(Bz_t, \Sigma) \tag{2.1}$$

where z_t is an h-dimensional vector of known covariates at time t and B is an $s \times h$ matrix of regression coefficients,

$$B = \begin{pmatrix} \beta_{1,1} & \dots & \beta_{1,h} \\ & \vdots & \\ \beta_{s,1} & \dots & \beta_{s,h} \end{pmatrix}.$$

The priors of the unknown parameters B, Σ are taken as conjugates of the normal model,

$$B \mid B^0, \Sigma, F \sim N_{sh}(B^0, \Sigma \otimes F^{-1}), \tag{2.2}$$

$$\Sigma \mid \Phi, \delta^* \sim W_s^{-1}(\Phi, \delta^*). \tag{2.3}$$

Let A^t denote a matrix transpose of A. Since X_t is partitioned into X_t^1 and X_t^2, Σ and B are partitioned correspondingly as

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \quad \textit{and} \quad B = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix}.$$

Define

$$S = \sum_{j=1}^{n} (X_j^2 - \hat{B}_2 z_j)(X_j^2 - \hat{B}_2 z_j)^t,$$

$$\hat{B}_2 = CA^{-1},$$

$$C = \sum_{j=1}^{n} X_j^2 z_j^t,$$

$$A = \sum_{j=1}^{n} z_j z_j^t,$$

$$D = (X_1^2, \dots, X_n^2).$$

Note D is the set of all observed data.

With a straightforward calculation, the posterior distributions of B and Σ are easily found. Lemma 2.1 gives their forms.

Lemma 2.1 *The posterior distributions of B and Σ are:*

$$B \mid D, B^0, \Gamma_2 \sim N_{sh}(B^*, \Sigma^*),$$

$$\Sigma_{22} \mid D, \Phi_{22}, \delta^* \sim W_{s_g}^{-1}(\hat{\Phi}_{22}, \delta^* + n - s_u),$$

$$\Sigma_{1|2} \mid D, \Phi_{1|2}, \delta^* \sim W_{s_u}^{-1}(\Phi_{1|2}, \delta^*),$$

$$\tau_{12} \mid D, \eta_{12}, \Sigma_{1|2} \sim N_{s_u s_g}(\eta_{12}, \Sigma_{1|2} \otimes \Phi_{22}^{-1}),$$

where

$$\Gamma_2 = \{\Sigma_{22}, \tau_{12}, \Sigma_{1|2}\},$$

$$\tau_{12} = \Sigma_{12}\Sigma_{22}^{-1},$$

$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21},$$

$$B^* = B^0 + \begin{pmatrix} \tau_{12} \\ I_{s_g \times s_g} \end{pmatrix} (\hat{B}_2 - B_2^0)\hat{E}^t,$$

$$\Sigma^* = \Sigma \otimes F^{-1} - \left[\begin{pmatrix} \tau_{12} \\ I_{s_g} \end{pmatrix} (\Sigma_{22}\tau_{12}^t, \Sigma_{22}) \right] \otimes (\hat{E}F^{-1}),$$

$$\hat{E} = F^{-1}(A^{-1} + F^{-1})^{-1},$$

$$\hat{\Phi}_{22} = \Phi_{22} + S + (\hat{B}_2 - B_2^0)^t (A^{-1} + F^{-1})^{-1} (\hat{B}_2 - B_2^0),$$

$$\eta_{12} = \Phi_{12}\Phi_{22}^{-1}.$$

LZ show that the predictive distribution of a future unknown realization X_f consists of two multivariate t distributions. These distributions are given in Theorem 2.1. For completeness, the definition of a multivariate t-distribution is repeated here. A multivariate t-distribution, denoted as $t_r(\mu, H, \vartheta)$, is defined to have a density function,

$$f(x) \propto \mid H \mid^{-\frac{1}{2}} \left[\vartheta + (x - \mu)^t H^{-1} (x - \mu) \right]^{-\frac{1}{2}(\vartheta + r)}. \tag{2.4}$$

Theorem 2.1 *Let the posterior distribution of B and Σ be defined as in Lemma 2.1. The predictive distribution of $X_f^t = \left((X_f^1)^t, (X_f^2)^t\right)$, given covariate vector z_f and the hyperparameters B^0, Φ, δ^*, F, consists of*

$$X_f^2 \mid D \sim t_{s_g}\left(a_{(1)} + b, \frac{c-d}{l^*}\hat{\Phi}_{22}, l^*\right)$$

and

$$X_f^1 \mid X_f^2 = x_f^2, D \sim t_{s_u}\left(a_0 + \eta_{12}(x_f^2 - a_{(1)}), \frac{c + (a_{(1)} - x_f^2)^t \Phi_{22}^{-1}(a_{(1)} - x_f^2)}{q}\Phi_{1|2}, q\right),$$

where

$$l^* = \delta^* + n - s_u - s_g + 1,$$

$$q = \delta^* - s_u + 1,$$

$$a = \begin{pmatrix} a_0 \\ a_{(1)} \end{pmatrix} = B^0 z_f,$$

$$b = (\hat{B}_2 - B_2^0)\hat{E} z_f,$$

$$c = 1 + z_f^t F^{-1} z_f,$$

$$d = z_f^t \hat{E} F^{-1} z_f.$$

With the predictive distributions, the Bayesian predictor is simplely taken as the posterior mean.

Corollary 2.1 *Given the predictive distributions of Theorem 2.1 and z_f, the predictive means of X_f^1 and X_f^2 are*

$$\mu_1 = E(X_f^1) = a_1 + \eta_{12} b,$$

$$\mu_2 = E(X_f^2) = a_2 + b.$$

2.3 Matrix T-Distribution

A representation of normal distribution given as a lemma in Lindley (1972) is summarized here. For any given constant matrices A, B, assume

$$X \sim N(AY, \Sigma_1)$$

and

$$Y \sim N(B, \Sigma_2).$$

If two normal distributions are independent, Lindley's result asserts

$$X \sim AB + N(0, \Sigma_1 + A\Sigma_2 A^t).$$

The above fact will be repeatedly used in the sequel without explicit mention.

Besides the normal distribution, the matrix T-distribution plays a pivotal role in the theories developed in this chapter. Its definition and some properties are discussed next.

Definition 2.1 *A random matrix $T : p \times q$ is said to follow the matrix T-distribution, if its density is expressible as*

$$f(t) = \frac{k([m,q,p])^{-1}}{|\ P\ |^{\frac{1}{2}(m-q)} |\ Q\ |^{\frac{1}{2}p}} \cdot |\ P^{-1} + tQ^{-1}t^t\ |^{-\frac{1}{2}m}, \tag{2.5}$$

where $P_{p\times p} > 0$, $Q_{q\times q} > 0$, $m > p + q - 1$,

$$k[m,p,q] = \frac{\pi^{\frac{1}{2}pq}\Gamma_q\left(\frac{1}{2}(m-p)\right)}{\Gamma_q\left(\frac{1}{2}m\right)},$$

and $\Gamma_p(\lambda) = \pi^{\frac{p(p-1)}{4}}\Gamma(\lambda)\Gamma(\lambda - \frac{1}{2})\ldots\Gamma(\lambda - \frac{p}{2} + \frac{1}{2})$.

In the above theorem, the notation, "$P > 0$" means that P is positive-definite. An alternative form of $f(t)$ is

$$f(t) = \frac{|\ Q\ |^{\frac{1}{2}(m-p)} |\ P\ |^{\frac{1}{2}q}}{k\,[m,p,q]} \cdot |\ Q + t^tPt\ |^{-\frac{1}{2}m}\ . \tag{2.6}$$

Using the notation of Dickey (1967), Press(1982) , we express the matrix T-distribution by $T \sim T(P, Q, 0, m)$ or more generally $T + C \sim T(P, Q, C, m)$, where C is a constant matrix.

Lemma 2.2 *When $q = 1$ and Q is a scalar, $T(P, Q, 0, m)$ is equivalent to a multivariate $t_p(0, \frac{QP^{-1}}{m-p}, m - p)$.*

Proof: By Equation (2.6),

$$\begin{aligned} f(t) &\propto (Q + t^t[P^{-1}]^{-1}t)^{-\frac{1}{2}m} \\ &\propto [(m-p) + t^t \left(\frac{QP^{-1}}{m-p}\right)^{-1} t]^{-\frac{1}{2}((m-p)+p)}. \blacksquare \end{aligned}$$

By (2.5) and (2.6), it is easy to see that

$$T^t \sim T(Q^{-1}, P^{-1}, 0, m). \tag{2.7}$$

In Dickey (1967), a representation of a matrix T is given. The result is copied here.

Lemma 2.3 *Suppose that $U_{p\times p} \mid P, m \sim W(P, m-q)$, $X_{p\times q} \mid Q \sim N(0, I_p \otimes Q)$, $P > 0$, $Q > 0$,$m > p + q - 1$ and that X, U are independent. Let T be a random $p \times q$ matrix and $T = \left(U^{-\frac{1}{2}}\right)^t X$. Then T has the distribution given in (2.5).*

In the Lemma, $\otimes$ denotes the Kronecker product. A direct application of the above representation yields the mean of a matrix T.

Corollary 2.2 *If $T \sim T(P, Q, 0, m)$, the mean of $T + C$ is C, where C is any constant matrix.*

Proof: By Lemma 2.3, $E(T) = E(\left[U^{-\frac{1}{2}}\right]^t)E(X) = 0$. $\blacksquare$

Below, some properties of a matrix T distribution given in Press (1982) are listed without proof. Partition T as

$$T = \begin{pmatrix} T_1 \\ T_2 \end{pmatrix},$$

T_i being a $p_i \times q$ matrix, $i = 1, 2$, and conformably,

$$P = \begin{pmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{pmatrix}$$

for a matrix P_{11} of dimensions $p_1 \times p_1$. Let $P_{2|1} = P_{22} - P_{21}P_{11}^{-1}P_{12} > 0$ be positive definite.

Lemma 2.4 *Suppose* $T \sim T(P, Q, 0, m)$. *Then:*

1. *conditionally* $T_1 \mid T_2 = t_2 \sim T(P_{11}, Q + t_2^t P_{1|2} t_2, -P_{11}^{-1}P_{12}t_2, m)$;
2. *marginally* $T_2 \sim T(P_{2|1}, Q, 0, m - p_1)$;
3. $\Theta \sim T(P, C_1^t Q C_1, 0, m)$ *where* $\Theta_{p\times r} = TC_1$ *and* C_1 *is any* $q \times r$ *matrix.*

Dawid (1981) replaces the notation $T(P, Q, 0, m)$ with $T(\delta, P, Q)$. His notation differs from that of Dickey in the choice of the "degrees of freedom" parameter, that is, $m - p = \delta + q - 1$.

Dawid (1981) has another representation of a matrix T-distribution $T \sim T(I_p, I_q, 0, m)$ that can be defined as the marginal distribution of $T_{p\times q}$ with $T \mid \Sigma \sim N(0, I_p \otimes \Sigma)$ and $\Sigma \sim W_q^{-1}(I, m - p)$. In a general form, if $T_{p\times q} \mid \Sigma \sim N(0, P^{-1} \otimes \Sigma)$ and $\Sigma \mid Q \sim W_q^{-1}(Q, m - p)$ then

$$T \sim T(P, Q, 0, m), \tag{2.8}$$

where P, Q are invertible symmetric square matrices.

Proof:

$$\begin{aligned} f(t \mid P, Q) &= \int f(t \mid P, \Sigma) f(\Sigma \mid Q)\, d\Sigma \\ &\propto \int \mid \Sigma \mid^{-\frac{1}{2}(m+q+1)} e^{-\frac{1}{2}tr\left[\Sigma^{-1}(t^t P t + Q)\right]} d\Sigma \\ &\propto \mid Q + t^t P t \mid^{-\frac{1}{2}m} . \end{aligned}$$

The last step is true by integrating with respect to a partial Wishart density function.∎

Definition 2.2 *Let* $\mathrm{A_{n\times p}}$ *be a matrix with column vectors,* $a_1, \ldots, a_p$*, define:*

$$vec(\mathrm{A}) = \begin{pmatrix} a_1 \\ \vdots \\ a_p \end{pmatrix}.$$

Let $\mathrm{tr}(A)$ denote the trace of matrix A, which is the sum of diagonal elements of A. Here are some useful facts:

Lemma 2.5 *When the dimensions are proper:*

1. $vec(A+B) = vec(A) + vec(B)$;
2. $(B^t \otimes A)\, vec(Z) = vec(AZB)$;
3. $tr(AB) = tr(BA)$
4. $tr(AB) = (vec(A^t))^t vec(B)$;
5. $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$;
6. $(A \otimes B)^t = A^t \otimes B^t$;
7. $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$;
8. *if* X *is a random matrix,* $E(vec(X)) = vec(E(X))$.

The proof is straightforward and omitted.

Definition 2.3 *Assume* $Y = (Y_{ij})$ *is an* $m \times q$ *random matrix, elements of* Y *are independently, identically distributed as* $N(0,1)$*,* $M : n \times p$*,* $A : p \times q$*,* $B : n \times m$ *are constant matrices and*

$$X = M + BYA^t.$$

Then X *follows a matrix normal distribution, i.e.* $X \sim N_{n\times p}(M,\, (BB^t) \otimes (AA^t))$.

The density function of a matrix normal X is

$$f(x) = (2\pi)^{-\frac{1}{2}np} \mid W \mid^{-\frac{1}{2}p} \mid V \mid^{-\frac{1}{2}n} e^{-\frac{1}{2}tr\left[W^{-1}(X-M)V^{-1}(X-M)^t\right]}, \tag{2.9}$$

where $W = BB^t > 0$, $V = AA^t > 0$.

By Lemma 2.5,

$$\begin{aligned}
&[vec(X-M)^t]^t(W \otimes V)^{-1}vec(X-M)^t \\
&= [vec(X-M)^t]^t vec[V^{-1}(X-M)^t W^{-1}] \\
&= tr[(X-M)V^{-1}(X-M)^t W^{-1}] \\
&= tr[W^{-1}(X-M)V^{-1}(X-M)^t].
\end{aligned}$$

Thus, the following fact is proved.

Lemma 2.6 *When the dimensions are proper and W, V are invertible, symmetric matrices, the following is true,*

$$[vec(X-M)^t]^t(W \otimes V)^{-1}vec(X-M)^t = tr[W^{-1}(X-M)V^{-1}(X-M)^t].$$

By a direct application of Lemma 2.6 to Equation (2.9), an equivalent relation between multivariate normality and matrix normality is established.

Lemma 2.7 *Assume X be an $n \times p$ random matrix, then*

$$X \sim N_{np}(M, W \otimes V) \quad if \quad and \quad only \quad if \quad vec(X^t) \sim N_{np}(vec(M^t), W \otimes V).$$

As one can see, the covariance matrix of a matrix normal distribution is a Kronecker product of two matrices. Sometimes, for notational simplicity, a notation $X_{p\times q} \sim \mathcal{N}(W, V) + M$ due to Dawid (1981) replaces $X_{p\times q} \sim N(M_{p\times q}, W_{p\times q} \otimes V_{q\times q})$. With this new notation, some facts about the matrix normal distribution are given without proof:

Lemma 2.8 *If* $X_{p\times q} = \mathcal{N}(W, V) + M$,

1. $X^t \sim \mathcal{N}(V, W) + M^t$;
2. $CXD \sim \mathcal{N}(CWC^t,\ D^tVD)$, *where* C, D *are two nonrandom matrices with proper dimensions.*

Next, a normal approximation of a matrix T is derived. Suppose T follows a matrix T distribution $T_{p\times q} \sim T(P, Q, 0, \delta)$. Define a scaled matrix $T^* = \delta^{\frac{1}{2}}T$. By Lemma 2.3,

$$T^* = \delta^{\frac{1}{2}}T \stackrel{\mathrm{d}}{=} \left(\frac{U^t}{\delta}\right)^{-\frac{1}{2}} X, \tag{2.10}$$

where $=^{\mathrm{d}}$ means *equal in distribution.* T^* is a matrix analogue of the univariate t_δ.

Let $\vartheta = \delta - q$. By the definition of a Wishart distribution, there are ϑ, p dimensional random vectors, $Y_i \sim^{\text{i.i.d.}} N(0, P)$, such that

$$U^t \stackrel{\mathrm{d}}{=} U \stackrel{\mathrm{d}}{=} \sum_{i=1}^{\vartheta} Y_i Y_i^t. \tag{2.11}$$

By the multivariate strong law of large numbers (SLLN),

$$\frac{vec(U^t)}{\vartheta} \stackrel{\mathrm{d}}{=} \frac{1}{\vartheta}\sum_{i=1}^{\vartheta} vec(Y_iY_i^t) \longrightarrow E(vec(Y_1Y_1^t)) \quad a.s. \quad as \quad \vartheta \longrightarrow \infty,$$

where, *a.s.* represents *convergence almost surely* with respect to $f_{Y_1}(\cdot)$.

Note that $E(vec(Y_1Y_1^t)) = vec(E(Y_1Y_1^t)) = vec(P)$. Hence,

$$\frac{U^t}{\vartheta} \longrightarrow P \quad a.s. \quad as \quad \vartheta \longrightarrow \infty \quad or \quad \delta \longrightarrow \infty.$$

Since when p is fixed, $\vartheta \longrightarrow \infty$ is equivalent to $\delta \longrightarrow \infty$. Applying Slutsky's theorem, we have

$$\begin{aligned} T^* &= \left[\frac{U^t}{\delta}\right]^{-\frac{1}{2}} X \\ &= \left(\frac{\vartheta}{\delta}\right)^{-\frac{1}{2}} \left[\frac{U^t}{\vartheta}\right]^{-\frac{1}{2}} X \\ &\longrightarrow P^{-\frac{1}{2}}X \stackrel{\mathrm{d}}{=} \mathcal{N}(P^{-1}, Q) \quad in \quad distribution. \quad as \quad \vartheta \longrightarrow \infty. \end{aligned}$$

Thus the following theorem is proved.

Theorem 2.2 *When n, p are fixed, as $\delta \longrightarrow \infty$*

$$T^* \longrightarrow \mathcal{N}(P^{-1}, Q) \quad in \quad distribution.$$

Theorem 2.2 is an extension of a similar asymptotic result in the univariate case, that is $t_\delta \longrightarrow N(0,1)$ as $\delta \longrightarrow \infty$. The fact has been noted in Dickey (1967).

Lemma 2.9 *Let $U \sim W(P, \vartheta)$. Then*

$$\sqrt{\vartheta}\left(\frac{U^t}{\vartheta} - P\right) \xrightarrow{\mathrm{d}} \mathcal{N}(2P, P), \quad , as \quad \vartheta \longrightarrow \infty,$$

where in the lemma, $\longrightarrow^{\mathrm{d}}$ denotes *convergence in distribution.*

The above lemma is applied in the proof of the following theorem that gives an asymptotic distribution of a matrix T-distribution in higher order.

Theorem 2.3 *Let T^* be defined by (2.10), and $P > 0$, $Q > 0$. Then*

$$\sqrt{\vartheta}(T^* - P^{-\frac{1}{2}}X) \xrightarrow{\mathrm{d}} U * V, as \quad \vartheta \longrightarrow \infty,$$

where $U \sim \mathcal{N}(\frac{1}{2}P, P^{-1})$, $V \sim \mathcal{N}(P^{-1}, Q)$ and given the hyperparameters P, Q, U is independent of V.

Corollary 2.3 *When ϑ is big,*

$$T^* \simeq \left(1 + \frac{1}{\sqrt{\vartheta}}Y^*\right) X^*$$

where given P, Q, U, V are independent and

$$U \sim \mathcal{N}(P^{-1}, Q), \quad V \sim \mathcal{N}\left(\frac{1}{2}P, P^{-1}\right).$$

2.4 Interpolation with Data Missing-by-design

In this section, the theory of Bayesian spatial interpolation with data missing-by-design is established for the multivariate case in which there are s_u ungauged sites, s_g gauged sites and k monitored pollutants at each site.

Suppose, now, that observations at gauged sites are pooled from different air pollution monitoring network. Since by design each gauged site does not monitor the same subset of pollutants, some pollutants, concerned by us, may not be monitored at a site. Hence, these pollutants are missing because of design. Therefore, we call these "missing" values *missing-by-design*. More specifically, Let $X_t^t = ((X_t^0)^t, (X_t^{(1)})^t)$, where $(X_t^0)_{s_u k \times 1}$ represents the response vector at ungauged sites at time t and $(X_t^{(1)})_{s_g k \times 1}$ represents the response vector at gauged sites at time t. After a proper rearrangement of its elements, $X_t^{(1)}$ is further partitioned into $(X_t^1)_{l \times 1}$, and $(X_t^2)_{(s_g k - l) \times 1}$ that respectively correspond to the vector of missing-by-design pollutants and to the vector of observed pollutants at gauged sites. Since the same l pollutants are missing during the whole monitoring period and they comprise l columns in matrix X^t, they are sometimes referred as *missing columns* in the sequel.

Like LZ, a normal model for the conditional sampling distribution is adopted. In terms of a matrix normal distribution, the model can be written as,

$$X \mid Z, B, \Sigma \quad \sim N_{sk \times n}(BZ, \Sigma \otimes I_n), \tag{2.12}$$

where $X = (X_1, \ldots, X_n)_{sk \times n}$ is the response matrix; $Z = (Z_1, \ldots, Z_n)_{h \times n}$ is the matrix of covariates;

$$B = \begin{pmatrix} \beta_{1,1} & \cdots & \beta_{1,h} \\ & \vdots & \\ \beta_{sk,1} & \cdots & \beta_{sk,h} \end{pmatrix}_{sk \times h}$$

is the coefficient matrix; Σ is the unknown spatial covariance matrix of X_t and I_n is an

$n \times n$ identity matrix. The conjugate priors of Σ, B are,

$$B \mid B^o, \Sigma, F \sim N_{skh}(B^o, \Sigma \otimes F^{-1}) \tag{2.13}$$

and

$$\Sigma \mid \Phi, \delta^* \sim W_{sk}^{-1}(\Phi, \delta^*). \tag{2.14}$$

The random field interpolation theory developed in this section has two steps. In the first step, while the hyperparameters B^o, F, Φ and δ^* are assumed to be fixed, predictive distributions derived. In the second step, estimation procedure of hyperparameters is discussed.

2.4.1 Predictive Distributions and Interpolation

In this subsection, all hyperparameters B^o, F, Φ and δ^* fixed and are suppressed in the derivations. The estimation of these hyperparameters is left to the next subsection.

If indices of missing columns in $X_t^{(1)}$ are $i_1, \ldots, i_l$ and indices of observed columns are $i_{l+1}, \ldots, i_{s_g k}$, let $R_1 = (r_{i_1}, \ldots, r_{i_l})$ and $R_2 = (r_{i_{l+1}}, \ldots, r_{i_{s_g k}})$ where r_j, $j = 1, \ldots, s_g k$, is an $s_g k$-dimensional vector with j^{th} element being one and the remainder being zero. Thus R_1 and R_2 "mark" the position of missing columns. Now let $R = (R_1, R_2)$. Observe that $(X^1)^t = (X^{(1)})^t R_1$ consists of the missing columns. Similarly $(X^2)^t = (X^{(1)})^t R_2$ consists of the observed columns. Because vector X_t is partitioned into three parts, B, Σ, B^o and Φ are partitioned accordingly. For example, Σ is partitioned as

$$\Sigma = \begin{pmatrix} \Sigma_{00} & \Sigma_{0(1)} \\ \Sigma_{(1)0} & \Sigma_{(11)} \end{pmatrix},$$

where Σ_{00} and $\Sigma_{(11)}$ are $s_u k \times s_u k$, $s_g k \times s_g k$ matrices respectively; $R^t \Sigma_{(11)} R$ is further partitioned as

$$R^t \Sigma_{(11)} R = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} = \begin{pmatrix} R_1^t \Sigma_{(11)} R_1 & R_1^t \Sigma_{(11)} R_2 \\ R_2^t \Sigma_{(11)} R_1 & R_2^t \Sigma_{(11)} R_2 \end{pmatrix},$$

where Σ_{11} and Σ_{22} are $l \times l$, $(s_g k - l) \times (s_g k - l)$ matrices respectively and in general if $E^t = (E_0^t, E_1^t, E_2^t)^t$, $E_{(1)}$ means $(E_1^t, E_2^t)^t$. The partitions of $\Phi_{(11)}$, B, B^o have a meaning analogous to that of $\Sigma_{(11)}$. Further, denote

$$\Psi_{(11)} = R^t \Phi_{(11)} R = \begin{pmatrix} \Psi_{11} & \Psi_{12} \\ \Psi_{21} & \Psi_{22} \end{pmatrix} = \begin{pmatrix} R_1^t \Phi_{(11)} R_1 & R_1^t \Phi_{(11)} R_2 \\ R_2^t \Phi_{(11)} R_1 & R_2^t \Phi_{(11)} R_2 \end{pmatrix},$$

with Ψ_{11} being $l \times l$ and Ψ_{22} being $(s_g k - l) \times (s_g k - l)$. Let $\Psi_{1|2} = \Psi_{11} - \Psi_{12}\Psi_{22}^{-1}\Psi_{21}$ and $\gamma = \Psi_{12}\Psi_{22}^{-1}$ for use in the sequel. Let,

$$R^t B_{(1)}^o = \begin{pmatrix} R_1^t B_{(1)}^o \\ R_2^t B_{(1)}^o \end{pmatrix} = \begin{pmatrix} B_1^o \\ B_2^o \end{pmatrix}.$$

When a mean squared loss function is taken, the Bayesian interpolator is simply $E(X^0 \mid X^2)$. To find the posterior mean, one needs the predictive distribution function, $f(X^0 \mid X^2)$. That predictive distribution function plays a pivotal role in our Bayesian interpolation theory. As an indirect approach, since $E(X^0 \mid X^2) = E(E(X^0 \mid X^1, X^2) \mid X^1)$, one can instead find the predictive distributions of $f(X^0 \mid X^{(1)})$ and $f(X^1 \mid X^2)$. Those two predictive distributions are given in the following two theorems.

Theorem 2.4

$$X^0 \mid X^{(1)} = x^{(1)} \sim T(\Phi_{0|(1)}^{-1}, I_{n\times n} + Z^t F^{-1} Z + (x^{(1)} - B_{(1)}^o Z)^t \Phi_{(11)}^{-1}(x^{(1)} - B_{(1)}^o Z),$$

$$B_0^o Z + \Phi_{0(1)}\Phi_{(11)}^{-1}(x^{(1)} - B_{(1)}^o Z), \delta^* + n).$$

From Corollary 2.2 we obtain the next corollary.

Corollary 2.4

$$E(X^0 \mid X^{(1)} = x^{(1)}) = B_0^o Z + \Phi_{0(1)}\Phi_{(11)}^{-1}(x^{(1)} - B_{(1)}^o Z).$$

Theorem 2.5

$$X^1 \mid X^2 = x^2 \sim T(\Psi_{1|2}^{-1}, I_{n\times n} + Z^t F^{-1} Z + (x^2 - R_2^t B_{(1)}^o Z)^t \Psi_{22}^{-1}(x^2 - R_2^t B_{(1)}^o Z),$$

$$R_1^t B_{(1)}^o Z + \gamma(x^2 - R_2^t B_{(1)}^o Z), \delta^* - s_u k + n).$$

Corollary 2.5

$$E\left(X^1 \mid X^2 = x^2\right) = R_1^t B_{(1)}^o Z + \gamma(x^2 - R_2^t B_{(1)}^o Z).$$

With the above two corollaries, the Bayesian interpolator is found.

Corollary 2.6

$$E(X^0 \mid X^2 = x^2) = B_0^o Z + \Phi_{0(1)} R_2 \Psi_{22}^{-1}(x^2 - R_2^t B_{(1)}^o Z).$$

The above corollary shows that the interpolated value $E(X_t^0 \mid X^2)$ depends only on the observed vector, X_t^2, provided the hyperparameters are fixed. Later it will be seen that our estimator of the hyperparameters depends on all observed values X^2, therefore the interpolated values, $E(X_t^0 \mid X^2)$, do depend on all observed values.

Although it is difficult to find an analytical form of $Var(X^0 \mid X^2)$, it is possible to find a closed form for $Var(X_t^0 \mid X^2)$. The following theorem gives that closed form. Before the variance formula is presented, an useful fact, needed in the derivation of variance formula, is given.

Lemma 2.10 *Let S be an invertible matrix*

$$S_{n\times n} = \begin{pmatrix} a & A_{12} \\ A_{12}^t & A_{22} \end{pmatrix},$$

where a is a scalar and A_{22} is an $n-1$ by $n-1$ matrix. A^{-1} is denoted as

$$A^{-1} = \begin{pmatrix} b & B_{12} \\ B_{12}^t & B_{22} \end{pmatrix}.$$

Then $(b - B_{12}B_{22}^{-1}B_{12}^t)^{-1} = a$.

Theorem 2.6 *For $t = 1, \ldots, n$,*

$$Var(X_t^0 \mid X^2) = \frac{c_t}{m-2}\left[A_1\Psi_{1|2}A_1^t + \Phi_{0|(1)}\right]$$

where

$$(A_1, A_2) = \Phi_{0(1)}\Phi_{(11)}^{-1}R = (\Phi_{0(1)}\Phi_{(11)}^{-1}R_1, \Phi_{0(1)}\Phi_{(11)}^{-1}R_2);$$

$$c_t = 1 + Z_t^t F^{-1} Z_t + (X_t^2 - B_2^o Z_t)^t \Psi_{22}^{-1}(X_t^2 - B_2^o Z_t);$$

$$m = \delta^* - s_u k - l + 1.$$

Obviously when there are no missing columns, the variance is $\Phi_{0|(1)}c_t/(m-2)$. When there are missing columns, the variance is roughly increased by a factor $A_1\Psi_{1|2}A_1^t c_t/(m-2)$.

As indicated earlier, if the predictive distribution of $f(X^0 \mid X^2)$ is found, the Bayesian interpolator can easily be obtained. That approach is pursued here as a way of checking the formulas for the interpolator and its variance. Note that if hyperparameters are given, by Model (2.12)-(2.14) and Lemma 2.20 of Section 2.6:

$$\begin{pmatrix} X_t^0 \\ X_t^2 \end{pmatrix} \sim N\left(\begin{pmatrix} B_0 \\ B_2 \end{pmatrix} Z_t, \begin{pmatrix} \Sigma_{00} & \Sigma_{0(1)}R_2 \\ R_2^t\Sigma_{(1)0} & R_2^t\Sigma_{(11)}R_2 \end{pmatrix}\right);$$

$$\begin{pmatrix} B_0 \\ B_2 \end{pmatrix} \sim N\left(\begin{pmatrix} B_0^o \\ B_2^o \end{pmatrix}, \begin{pmatrix} \Sigma_{00} & \Sigma_{0(1)}R_2 \\ R_2^t\Sigma_{(1)0} & R_2^t\Sigma_{(11)}R_2 \end{pmatrix} \otimes F^{-1}\right);$$

$$\begin{pmatrix} \Sigma_{00} & \Sigma_{0(1)}R_2 \\ R_2^t\Sigma_{(1)0} & R_2^t\Sigma_{(11)}R_2 \end{pmatrix} \sim W^{-1}\left(\begin{pmatrix} \Phi_{00} & \Phi_{0(1)}R_2 \\ R_2^t\Phi_{(1)0} & \Psi_{22} \end{pmatrix}, \delta^* - l\right).$$

By following the same way as the proof of Theorem 2.4, the predictive distribution of $f(X_t^0 \mid X^2)$ is easily obtained.

Theorem 2.7

$$X_t^0 \mid X^2 \sim T\left(\Phi_{0|2}^{-1}, c_t, B_0^o Z_t + \Phi_{0(1)}R_2\Psi_{22}^{-1}(X_t^2 - B_2^0 Z_t), m + s_u k\right)$$

where c_t, m are defined in Theorem 2.6 and $\Phi_{0|2} = \Phi_{00} - \Phi_{0(1)}R_2\ \Psi_{22}^{-1}\ R_2^t\Phi_{(1)0}$.

From the above predictive distribution, the same Bayesian interpolator as that of Corollary 2.6 is obtained. By Lemma 2.2 and Press (1982), it is easy to see that the variance is

$$Var(X_t^0 \mid X_t^2) = \frac{c_t}{m-2}\Phi_{0|2}. \tag{2.15}$$

It is not difficult to prove that Equation (2.15) is equivalent to the variance formula of Theorem 2.6.

Lemma 2.11 *If A_1 is defined in Theorem 2.6 and $\Phi_{0|2}$ is defined in Theorem 2.7, the following equation is true.*

$$\Phi_{0|2} = A_1\Psi_{1|2}A_1^t + \Phi_{0|(1)}.$$

That verifies the variance formula of Theorem 2.6.

With the variance, only a pairwise confidence interval can be derived. With the posterior distribution of $X_t^0 \mid X^2$, we can derive a simultaneous region. The next lemma is Theorem 3.2.12 of Muirhead (1982 p96) and it is copied here.

Lemma 2.12 *If* $\mathbf{A}$ *is* $W_p(\Sigma, \delta^*)$. *So* δ^* *is a positive integer,* $\delta^* > p-1$, *and* $\mathbf{Y}$ *is any* $p \times 1$ *random vector distributed independently of* $\mathbf{A}$ *with* $P(\mathbf{A} = \mathbf{0}) = 0$ *then* $\mathbf{Y}^t\Sigma^{-1}\mathbf{Y}/\mathbf{Y}^t\mathbf{A}^{-1}\mathbf{Y}$ *is* $\chi^2_{\delta^*-p+1}$, *and is independent of* $\mathbf{Y}$.

By that lemma, the distribution of the quadratic form of matrix T is easily found.

Theorem 2.8 *If* $T_{p\times 1} \sim T(I_{p\times p}, 1, 0, \delta^*)$,

$$\frac{(\delta^* - p)T^tT}{p} \sim F_{p,\,\delta^*-p}.$$

Let $\hat{x}_t^0 = B_0^o Z_t + \Phi_{0(1)}R_2\Psi_{22}^{-1}(X_t^2 - B_2^o Z_t)$, where $\hat{x}_t^0$ is the Bayesian interpolator of X_t^0 when there are missing-by-design data. Then a simultaneous region of X_t^0 is given in Theorem 2.9.

Theorem 2.9 *The* $1-\alpha$, *where* $0 < \alpha < 1$, *simultaneously posterior region of the Bayesian interpolator* $\hat{x}_t^0$ *is a hyper-ellipsoid, defined by the set*

$$\{X_t^0 \ : (X_t^0 - \hat{x}_t^0)^t\Phi_{0|2}^{-1}(X_t^0 - \hat{x}_t^0) < b \mid X^2\},$$

where,

$$b = \frac{s_u k * c_t * F_{1-\alpha, s_u k, \delta^*-l-s_u k+1}}{\delta^* - l - s_u k + 1}$$

and c_t *is defined in Theorem 2.6.*

2.4.2 Estimation of Hyperparameters

In the previous subsection, the hyperparameters Φ, δ^*, F, B^o are assumed to be known. Often, they are unknown. To finish the interpolation procedure, they must be specified in some way. In a complete Bayesian hierarchical approach, another level of priors is usually laid on these hyperparameters. An alternative approach, suitable when the parameters are not sensitive to their prior specification, entails the use of empirical

Bayesian method. In an empirical method, the hyperparameters are estimated with the observed data.

In BLZ, Φ and δ^* are estimated through a maximum likelihood estimator (MLE) and B^o, F^{-1} are assumed to be zero. While it is difficult to estimate all of hyperparameters, Φ, δ^*, B^o, F^{-1} by MLE, in this subsection we propose two unbiased estimators for the estimation of B^o, F^{-1} and apply the MLE method to estimate Φ, δ^*. For the moment, the MLE of Φ, δ^* is discussed. The two unbiased estimators of B^o, F^{-1} are described at the end of this subsection.

Even with B^o, F^{-1} being fixed, the direct maximization of the marginal distribution $f(X^2 \mid \Phi, \delta^*)$ is difficult. With the help of EM algorithm proposed by Dempster, Laird and Rubin (1977), one can instead maximize the distribution of $f(X^2, \Sigma_{22}, B_2 \mid \Phi, \delta^*)$ (Chen 1979). Let the complete data set be x and assume only part of x is observed. That observed part is denoted as y. If the sampling density $f(x \mid \phi) = b(x)exp(\phi t(x)^t)/a(\phi)$ is from a regular exponential-family, the p^{th} iteration of EM algorithm consists of:

(E-step) estimating the complete data sufficient statistics, $t(x)$, by finding $t^{(p)} = E(t(x) \mid y, \phi^{(p)})$; *and*

(M-step) determining $\phi^{(p+1)}$ as the solution of the equation $E(t(x) \mid \phi) = t^{(p)}$.

It is proved, in C.F.J. Wu(1983), that the limit point ϕ^* does maximize the marginal likelihood function $l(\phi \mid l)$.

To apply the above EM algorithm, we need to find the sufficient statistics of Φ and δ^*. Following the approach of LZ, let

$$\hat{B}_2 = (X^2)^t Z^t (ZZ^t)^{-1}$$

and

$$S_2 = (X^2)^t(I - Z^t(ZZ^t)^{-1}Z)X^2.$$

Anderson (1984 p291) shows that given $B_{(1)}$ and $\Sigma_{(11)}$,

$$\hat{B}_2 \mid B_{(1)} \sim N(R_2^t B_{(1)}, \Sigma_{22} \otimes (ZZ^t)^{-1}), \tag{2.16}$$

$$S_2 \mid \Sigma_{(11)} \sim W_{s_gk-l}(\Sigma_{22}, n-h) \tag{2.17}$$

and that $\hat{B}_2$, S_2 are independent.

Since,

$$\begin{aligned}
&f(X^2, B_{(1)}, \Sigma_{(11)} \mid B^o_{(1)}, F^{-1}, \Phi_{(11)}, \delta^*) \\
&= f(X^2 \mid B_{(1)}, \Sigma_{(11)}) f(\Sigma_{(11)} \mid \Phi_{(11)}, \delta^*) f(B_{(1)} \mid B^o_{(1)}, F^{-1}, \Sigma_{(11)}) \\
&= f(S_2 \mid \Sigma_{(11)}) f(\Sigma_{(11)} \mid \Phi_{(11)}, \delta^*) f(\hat{B}_{(1)} \mid B_{(1)}, \Sigma_{(11)}) f(B_{(1)} \mid B^o_{(1)}, \Sigma_{(11)}, F^{-1}),
\end{aligned}$$

where the last step is true, because of Equations (2.16) and (2.17).

When B^o and F^{-1} are assumed to be known, the above likelihood function of $\Phi_{(11)}$ and δ^* is proportional to $f(\Sigma_{(11)} \mid \Phi_{(11)}, \delta^*)$. Because, by Lemma 2.20 of Section 2.6,

$$\Sigma_{(11)} \mid \Phi_{(11)}, \delta^* \sim W^{-1}_{s_gk}(\Phi_{(11)}, \delta^* - s_u k).$$

Therefore, the likelihood function is proportional to,

$$\begin{aligned}
f(X^2, B_{(1)}, \Sigma_{(11)} \mid \Phi_{(11)}, \delta^*) &\propto f(\Sigma_{(11)} \mid \Phi_{(11)}, \delta^*) \\
&\propto c_0^{-1} \mid \Phi_{(11)} \mid^{\frac{(\delta^* - s_u k)}{2}} \mid \Sigma_{(11)} \mid^{\frac{\delta^* + s_g k + 1 - s_u k}{2}} \\
&\quad \cdot exp\left[-\frac{1}{2} tr(\Phi_{(11)} \Sigma^{-1}_{(11)})\right]
\end{aligned} \tag{2.18}$$

where

$$c_0 = 2^{s_g k(\delta^* - s_u k)/2} \pi^{(s_g k - 1)s_g k/4} \prod_{i=1}^{s_g k} \Gamma\left(\frac{\delta^* - s_u k - i + 1}{2}\right).$$

With the likelihood Equation (2.18), the pair ($\Sigma_{(11)}$, $log \mid \Sigma_{(11)} \mid$) is readily identified as the sufficient statistics of Φ_{11}, δ^* (Chen 1979).

To reduce number of parameters that need to be estimated, BLZ adopt a Kronecker structure, $\Phi = \Lambda \otimes \Omega$, where Λ the between-sites-hypercovariance matrix and Ω between-pollutants-hypercovariance matrix. Since $\Phi_{(11)} = \Lambda_g \otimes \Omega$, where Λ_g is the hypercovariance matrix between the gauged sites, by the likelihood function (2.18), only Λ_g and Ω could be estimated. Hence, the estimation of Φ takes two steps. In Step one, Λ_g, Ω and δ^* are estimated by MLE through an EM algorithm. In Step two, the SG method is applied to extend Λ_g to Λ.

Expectation Step

At E-step, the posterior expectations of $\Sigma_{(11)}$ and $log \mid \Sigma_{(11)} \mid$ are of interest. The next lemma, due to Chen (1979, Theorem 2.1, p237), is used in the derivation of the expectations of these sufficient statistics. The lemma is copied without proof.

Lemma 2.13 *Given hyperparameters Φ, δ^*, B^o and F,*

$$E(\Sigma^{-1} \mid X^{(1)}) = \begin{pmatrix} \delta^* \Phi_{0|(1)}^{-1} & -\delta^* \Phi_{0|(1)}^{-1} \eta \\ -\delta^* \eta^t \Phi_{0|(1)}^{-1} & d \end{pmatrix},$$

where,

$$\eta = \Phi_{0(1)} \Phi_{(11)}^{-1},$$

$$A = ZZ^t,$$

$$C_{(1)} = X^{(1)} Z^t,$$

$$\hat{B}_{(1)} = C_{(1)} A^{-1},$$

$$S_{(1)} = X^{(1)}(I - Z^t(ZZ^t)^{-1}Z)(X^{(1)})^t,$$

$$\hat{\Phi}_{(11)} = \Phi_{(11)} + S_{(1)} + (\hat{B}_{(1)} - B^o_{(1)})(A^{-1} + F^{-1})^{-1}(\hat{B}_{(1)} - B^o_{(1)})^t,$$

$$d = (\delta^* + n - s_u k - h)\hat{\Phi}^{-1}_{(11)} + \delta^* \eta^t \Phi^{-1}_{0|(1)} \eta + s_u k \Phi^{-1}_{(11)}.$$

Note that $\Psi_{ij} = R_i^t \Phi_{(11)} R_j$.

Lemma 2.14 *If* $W \sim W_p(A^{-1}, n)$, *then given* A, n,

$$E(log \mid W \mid) = n * A^{-1}$$

and

$$E(log \mid W \mid) = p * log(2) + \sum_{i=1}^{p} \Psi(\frac{n - i + 1}{2}) - log \mid A \mid,$$

where $\Psi(\cdot)$ *represents a digamma function. A digamma function is defined to be the derivative of a log-gamma function.*

With the above lemma, the following lemma is true.

The expectation of $\Sigma_{(11)}$ is,

Theorem 2.10

$$E(\Sigma^{-1}_{(11)} \mid X^2) = R \begin{pmatrix} (\delta^* - s_u k)\Psi^{-1}_{1|2} & -(\delta^* - s_u k)\Psi^{-1}_{1|2}\eta^* \\ -(\delta^* - s_u k)(\eta^*)^t \Psi^{-1}_{1|2} & d_1 \end{pmatrix} R^t$$

where

$$G_1 = I - Z^t (ZZ^t)^{-1} Z,$$

$$G_2 = M_2^t Z^t (ZZ^t)^{-1} - B_2^o,$$

$$\eta^* = \Psi_{12}\Psi_{22}^{-1},$$

$$\tilde{S} = M_2^t G_1 M_2 + G_2((ZZ^t)^{-1} + F^{-1})^{-1} G_2^t,$$

$$\hat{\Psi}_{22} = \Psi_{22} + \tilde{S},$$

$$d_1 = (\delta^* - s_u k + n - l - h)\hat{\Psi}^{-1}_{22} + (\delta^* - s_u k)(\eta^*)^t \Psi^{-1}_{1|2}\eta^* + l\Psi^{-1}_{22}.$$

The expectation of $log \mid \Sigma_{(11)} \mid$ is,

Lemma 2.15

$$E(log \mid \Sigma_{(11)} \mid\mid X^2, \Phi_{(11)}, \delta^*) = -s_g k log(2) - \sum_{i=1}^{l} \Psi(\frac{\delta^* - s_u k - i + 1}{2}) -$$
$$\sum_{i=1}^{s_g k - l} \Psi(\frac{\delta^* + n - s_u k - l - h - i + 1}{2}) + log \mid \Psi_{1|2} \mid + log \mid \hat{\Psi}_{22} \mid .$$

Maximization Step

When the expectations of sufficient statistics are found, at M-step, the current values of the hyperparameters are updated with the values that maximize the "likelihood" function. Here, the "likelihood" function is obtained by plugging the expectations of the sufficient statistics into Equation (2.18). However maximizing the "likelihood" function over $\Phi_{(11)} = \Lambda_g \otimes \Omega$ is not easy. When there is more than one parameter involved in the maximization step, the following lemma leads to an iterated procedure. An advantage of the iterated procedure is that at each iterated step, only the maximization of a function over a single-parameter is required. That is generally easier than maximizing over several parameters simultaneously.

Lemma 2.16 *If a function $g(x, y)$ is upper bounded, the iteration procedure described below leads to the pair (x_0, y_0) that maximizes $g(\cdot, \cdot)$ locally. At the p^{th} iteration, update the current value $x^{(p)}$ with a value that maximizes the function $g(x, y^{(p)} \mid y^{(p)})$ and denote the updated x value as $x^{(p+1)}$; update the current value $y^{(p)}$ with a value that maximizes the function $g(x^{(p+1)}, y \mid x^{(p+1)})$ and denote the updated value y as $y^{(p+1)}$. (x_0, y_0) is the limit of $(x^{(p)}, y^{(p)})$.*

Proof: $g(x, y)$ is bounded up and

$$g(x^{(p)}, y^{(p)}) \leq g(x^{(p+1)}, y^{(p)}) \leq g(x^{(p+1)}, y^{(p+1)}).\blacksquare$$

Besides the above lemma, another lemma is used in the maximizing procedure. The lemma is Lemma 3.2.2 of Anderson (1984 p62). It is copied for convenience.

Lemma 2.17 *If $D_{p\times p} > 0$, the maximum of*

$$f(G) = -nlog \mid G \mid -tr(G^{-1}D),$$

with respect to positive definite matrices G exists, occurs at $G = (1/n)D$, and has the value

$$f[(1/n)D] = p * n * log(n) - n * log \mid D \mid -p * n.$$

With Lemma 2.16, an iterated procedure is adopted to find $\Phi_{(11)} = \Lambda_g \otimes \Omega$, δ^* that maximize Equation (2.18). Since $R^t\Phi_{(11)}R = \Psi_{(11)}$ and R is orthogonal, maximizing Equation (2.18) over $\Phi_{(11)}$ is equivalent to maximizing the same equation over $\Psi_{(11)}$.

When δ^* is fixed, the log-likelihood of Equation (2.18) is proportional to

$$\begin{aligned} L(\Lambda_g, \Omega) &= (\delta^* - s_u k)log \mid \Phi_{(11)} \mid -tr\left(\Phi_{(11)}\Sigma_{(11)}^{-1}\right) \\ &= -(\delta^* - s_u k)klog \mid \Lambda_9^{-1} \mid -(\delta^* - s_u k)s_g \mid \Omega^{-1} \mid -tr\left[(\Lambda_g \otimes \Omega)\Sigma_{(11)}^{-1}\right]. \end{aligned}$$

Again, an iterated procedure is applied to maximize $L(\Lambda_g, \Omega)$. When Ω is fixed, rewrite $tr\left(\Phi_{(11)}\Sigma_{(11)}^{-1}\right)$ as $tr(\Lambda_g Q)$. By Lemma 2.17, $\Lambda_g = k(\delta^* - s_u k)Q^{-1}$ maximizes $L(\Lambda_g, \Omega)$. Similarly, when Λ_g is fixed, $\Omega = s_g(\delta^* - s_u k)G^{-1}$, where $tr(\Phi_{(11)}\ \Sigma_{(11)}^{-1}) = tr(\Omega G)$, maximizes $L(\Lambda_g, \Omega)$.

When $\Phi_{(11)}$ is fixed, δ^* is updated by maximizing Equation (2.18) over δ^*. By taking the first derivative of Equation (2.18) with respect to δ^*, it becomes

$$-s_g klog(2) - log \mid \Sigma_{(11)} \mid +log \mid \Phi_{(11)} \mid -\sum_{i=1}^{s_g k} \Psi(\frac{\delta^* - s_u k - i + 1}{2}) = 0. \qquad (2.19)$$

When Λ_g, Ω are fixed, finding δ^* that maximizes Equation (2.18) is equivalent to finding a solution in Equation (2.19). For showing the existence of a solution of Equation (2.19), replace $log \mid \Sigma_{(11)} \mid$ with $E(log \mid \Sigma_{(11)} \mid\mid M_2^t, \Phi_{(11)}, \delta^*)$ in Lemma 2.15 to Equation (2.19) and use the relationship $log \mid \Phi_{(11)} \mid = log \mid R\Psi_{(11)}R^t \mid = log \mid \Psi_{1|2} \mid + log \mid \Psi_{22} \mid$, Equation (2.19) becomes,

$$\sum_{i=1}^{s_g k-l} \Psi(\frac{\delta^* + n - s_u k - l - h - i + 1}{2}) - \sum_{i=l+1}^{s_g k} \Psi(\frac{\delta^* - s_u k - i + 1}{2})$$

$$-log \mid \hat{\Psi}_{22} \mid + log \mid \Psi_{22} \mid = 0$$

or equivalently

$$\sum_{i=l+1}^{s_g k} \left[\Psi(\frac{\delta^* + n - s_u k - h - i + 1}{2}) - \Psi(\frac{\delta^* - s_u k - i + 1}{2})\right]$$

$$= log \mid \hat{\Phi}_{22} \mid - log \mid \Psi_{22} \mid .$$

Since the gamma function is convex, the digamma function is monotonically increasing. Thus, the left side of the above equation is always positive whenever $n - h > 0$ and it goes to zero when δ^* increases to infinity. The right side of the equation is positive. Therefore, a unique solution does exist.

2.4.3 EM Algorithm

By summarizing the above discussion, our EM algorithm becomes:

E-STEP. Given the current values of $\Psi_{(11)}$, δ^*,

$$E(\Sigma_{(11)}^{-1} \mid M_2^t) = R \begin{pmatrix} (\delta^* - s_u k)\Psi_{1|2}^{-1} & -(\delta^* - s_u k)\Psi_{1|2}^{-1}\eta^* \\ -(\delta^* - s_u k)(\eta^*)^t\Psi_{1|2}^{-1} & d_1 \end{pmatrix} R^t$$

where

$$d_1 = (\delta^* - s_u k + n - l - h)\hat{\Psi}_{22}^{-1} + (\delta^* - s_u k)(\eta^*)^t\Psi_{1|2}^{-1}\eta^* + l\Psi_{22}^{-1}$$

and

$$E(log \mid \Sigma_{(11)} \mid\mid M_2^t, \Phi_{(11)}, \delta^*) = -s_g k log(2) - \sum_{i=1}^{l} \Psi(\frac{\delta^* - s_u k - i + 1}{2}) -$$

$$\sum_{i=1}^{s_g k - l} \Psi(\frac{\delta^* + n - s_u k - l - h - i + 1}{2}) + log \mid \Psi_{1|2} \mid + log \mid \hat{\Psi}_{22} \mid;$$

M-STEP: Given the current values of $\Sigma_{(11)}^{-1}$ and $log \mid \Sigma_{(11)} \mid$, find the MLE of $\Phi_{(11)} = \Lambda_g \otimes \Omega$, δ^* by repeating the following steps until it converges.

Step 1:

Given the current $\Lambda_g^{(p)}$ and $\Omega^{(p)}$, represent $tr[(\Lambda_g^p \otimes \Omega^{(p)})\Sigma_{(11)}^{-1}]$ as $tr(\Omega^{(p)}G)$ and set $\Omega^{(p+1)} = s_g(\delta^* - s_u k)G^{-1}$;

Step 2.

Given current $\Lambda_g^{(p)}$ and $\Omega^{(p+1)}$, represent $tr[(\Lambda_g^p \otimes \Omega^{(p+1)})\Sigma_{(11)}^{-1}]$ as $tr(\Lambda_g^{(p)}Q)$ and set $\Lambda_g^{(p+1)} = k(\delta^* - s_u k)Q^{-1}$.

Step 3.

Given the current $\Psi_{(11)}$, estimate δ^* by solving the following equation:

$$\sum_{i=l+1}^{s_g k} \left[\Psi(\frac{\delta^* + n - s_u k - h - i + 1}{2}) - \Psi(\frac{\delta^* - s_u k - i + 1}{2})\right]$$

$$= log \mid \hat{\Phi}_{22} \mid - log \mid \Psi_{22} \mid . \tag{2.20}$$

Estimation of B^o and F^{-1}.

The estimation of the hyperparameters B^o and F^{-1} are based on two unbiased estimators. Suppose that the k air pollutants are labeled from 1 to k. The measurements of the same air pollutant at different sites are usually similar but the measurements among different pollutants are not similar. Based on this fact, we assume that B^o has an exchangeability structure, that is, the hypermean coefficients of i^{th} air pollutant at all gauged sites are the same and equal to u^i, $i = 1, \ldots, k$.

Since,

$$X^2 \mid \Sigma_{(11)} \stackrel{\mathrm{d}}{=} B_2^{**} Z + N(0, \Sigma_{22}^{**} \otimes I_n),$$

then the following lemma is true.

Lemma 2.18 *Given $B_{(1)}^o$ and $\Sigma_{(11)}$, $\hat{B}_2$ and S_2 are independent.*

Proof:

$$\begin{aligned} f(\hat{B}_2, S_2 \mid B_{(1)}^o, \Sigma_{(11)}) &= \int f(\hat{B}_2, S_2 \mid B_{(1)}, B_{(1)}^o, \Sigma_{(11)}) f(B_{(1)} \mid B_{(1)}^o, \Sigma_{(11)})\, dB_{(1)} \\ &= \int f(\hat{B}_2 \mid B_{(1)}, \Sigma_{(11)}) f(B_{(1)} \mid B_{(1)}^o, \Sigma_{(11)}) f(S_2 \mid \Sigma_{(11)})\, dB_{(1)} \\ &= f(\hat{B}_2 \mid B_{(1)}^o, \Sigma_{(11)}) f(S_2 \mid \Sigma_{(11)}). \quad \blacksquare \end{aligned}$$

The next theorem gives unbiased estimators of B_0 and F^{-1}. Let

$$\hat{B}_2 = \begin{pmatrix} \hat{\beta}_1^{j_1} \\ \vdots \\ \hat{\beta}_{s_g k-l}^{j_{s_g k-l}} \end{pmatrix}$$

where $j_v \in \{1, \ldots, k\}$, $v = 1, \ldots, s_g k - l$. Obviously $(\hat{\beta}_v^{j_v})^t = (ZZ^t)^{-1} Z^t X_v$ and j_v marks the pollutant type of X_v.

Let $\hat{u}^i$ be the means of all $\hat{\beta}_v^{j_v}$ with $j_v = i$, $i = 1, \ldots, k$. Therefore, under the exchangeability assumption, $\hat{u}^i$ is the estimator of u^i. Let,

$$\hat{F}^{-1} = \frac{n-h-2}{s_g k - l} \sum_{v=1}^{s_g k-l} \frac{(\hat{\beta}_v^{j_v} - \hat{u}^{j_v})^t (\hat{\beta}_v^{j_v} - \hat{u}^{j_v})}{a_{j_v}^t S_2 a_{j_v}} - (ZZ^t)^{-1}$$

where $\hat{\beta}_v^{j_v} - \hat{u}^{j_v} = a_{j_v}^t \hat{B}_2$.

Theorem 2.11 *Given X^2 and the exchangeability of B^o, $\hat{u}^i$, $i = 1, \ldots, k$ and $\hat{F}^{-1}$ are unbiased estimators of u^i, $i = 1, \ldots, k$ and F^{-1} respectively.*

2.5 Interpolation with Randomly Missing Data

We refer the term *randomly missing* to the same meaning as *missing at random* defined mathematically by Rubin (1976), Little and Rubin (1987). More precisely, let Y denote

the data that would occur in the absence of missing values. Further, we assume $Y = (Y_{obs}, Y_{mis})$, where Y_{obs} represents the observed values and Y_{mis}, missing values. Assume $f(Y \mid \theta)$ denotes the joint probability density of Y_{obs}, Y_{mis}. We define for each component of Y a missing data indicator, taking value 1 if the component is observed and 0 if missing. For example, if $Y = (Y_{ij})$, an $(n \times K)$ matrix of n observations measured for K variables, define the response indicator $R = (R_{ij})$, such that

$$R_{ij} = \begin{cases} 1 & \text{if } y_{ij} \text{ observed,} \\ 0 & \text{if } y_{ij} \text{ missing.} \end{cases}$$

When we treat R as a random variable and specify the joint distribution of R and Y, the conditional distribution R given Y indexed by an unknown parameter ψ is called the *distribution for the missing data mechanism.*

Rubin (1976) defines the missing data to be missing at random when the distribution of the missing data mechanism does not depend on the missing value Y_{mis}, that is, if

$$f(R \mid Y_{obs}, Y_{mis}, \psi) = f(R \mid Y_{obs}, \psi).$$

Little and Rubin (1987) point out that when data are missing at random, likelihood based inferences can ignore the missing data mechanism. In other words, likelihood based inferences can only base on the marginal distribution, $f(Y_{obs} \mid \theta)$.

Let M_1 represent the randomly missing subset of data from $X^{(1)}$ and M_2 the observed subset of data from $X^{(1)}$. Let l^* denote the total number of elements in M_1. Again, the predictive distribution $f(X^0 \mid M_2)$ leads to a Bayesian interpolator. Given the hyper-parameters, when there are data missing-by-design or no missing data, the predictive distribution follows a matrix T-distribution or a multivariate t-distribution. With randomly missing data, will the predictive distribution $f(X^0 \mid M_2)$ still follow some form of t-distribution? The following simple example gives a negative answer.

Consider a simple case, where there are two gauged sites and $n = 2$. Let T denote the random matrix at the gauged sites. That is,

$$T_{2\times 2} \mid y_1, y_2 = \begin{pmatrix} X_1 & y_1 \\ y_2 & X_2 \end{pmatrix},$$

where the upper case characters represent the missing values and lower case characters represent the observed values. Assume,

$$T \mid \Sigma \sim \mathcal{N}\left(I_{2\times 2}, \Sigma\right), \quad \Sigma \sim W_2^{-1}\left(I_{2\times 2}, \delta + 2\right).$$

Then,

$$\begin{aligned} \mid I + T^t T \mid &= \begin{vmatrix} X_1^2 + y_2^2 + 1 & y_1 X_1 + y_2 X_2 \\ y_1 X_1 + y_2 X_2 & X_2^2 + y_1^2 + 1 \end{vmatrix} \\ &= (X_1^2 + 1)X_2^2 + (-2y_1 y_2 X_1)X_2 + y_1^2 + y_2^2 + y_1^2 y_2^2 + 1 + X_1^2 \\ &= \left(a_1 X_2 + \frac{a_2}{a_1}\right)^2 + a_4; \end{aligned}$$

where $a_1 = (X_1^2 + 1)^{\frac{1}{2}}$, $a_2 = -y_1 y_2 X_1$, $a_3 = y_1^2 + y_2^2 + y_1^2 y_2^2 + 1 + X_1^2$ and $a_4 = a_3 - \frac{a_2^2}{a_1^2}$. Note,

$$\begin{aligned} a_4 &= y_1^2 + y_2^2 + y_1^2 y_2^2 + X_1^2 + 1 - \frac{y_1^2 y_2^2 X_1^2}{1 + X_1^2} \\ &= y_1^2 + y_2^2 + X_1^2 + 1 + \frac{y_1^2 y_2^2}{1 + X_1^2} \\ &> 0. \end{aligned}$$

Thus

$$\mid I + T^t T \mid^{-\frac{\delta+4}{2}} = \left[\left(a_1 X_2 + \frac{a_2}{a_1}\right)^2 + a_4\right]^{-\frac{\delta+4}{2}},$$

since

$$\begin{aligned} f\left(X_1 \mid Y_1 = y_1, Y_2 = y_2\right) &= \int_{-\infty}^{\infty} f\left(X_1, x_2 \mid y_1, y_2\right) dx_2 \\ &= \int_{-\infty}^{\infty} \frac{f(X_1, x_2, y_1, y_2)}{f(y_1, y_2)} dx_2 \\ &\propto \int_{-\infty}^{\infty} f(X_1, x_2, y_1, y_2)\, dx_2 \\ &\propto \int_{-\infty}^{\infty} f(T) dx_2. \end{aligned}$$

Use (2.5) to substitute for $f(T)$, and obtain

$$\begin{aligned} f(X_1 \mid Y_1 = y_1, Y_2 = y_2) &\propto \int_{-\infty}^{\infty} \left|I + T^t T\right|^{-\frac{\delta+4}{2}} dx_2 \\ &\propto a_4^{-\frac{\delta+4}{2}} \int_{-\infty}^{\infty} \left[1 + \frac{\left(a_1 x_2 + \frac{a_2}{a_1}\right)^2}{a_4}\right]^{-\frac{\delta+4}{2}} dx_2 \\ &\propto a_4^{-\frac{\delta+4}{2}} a_1^{-1} \int_{-\infty}^{\infty} \left[1 + \frac{t^2}{a_4}\right]^{-\frac{\delta+4}{2}} dt \\ &\propto a_4^{-\frac{\delta+3}{2}} a_1^{-1} \\ &\propto \left[(1+X_1)(y_1^2 + y_2^2 + X_1^2 + 1) + y_1^2 y_2^2\right]^{-\frac{\delta+3}{2}} \left[1 + X_1^2\right]^{\frac{\delta+2}{2}}. \end{aligned}$$

The last step is completed by integrating with respect to a partial t density.

Consider the special case, $y_1 = y_2 = 1$. Then

$$f(X_1 \mid y_1 = 1, y_2 = 1) \propto (X_1^2 + 2)^{-\delta+3}(1 + X_1^2)^{\frac{\delta+2}{2}},$$

$f(X_1 \mid 1, 1)$ is clearly not a t-distribution. Therefore $f(X_1, X_2 \mid y_1, y_2)$ does not follow a multivariate t-distribution. If it did, its marginal $f(X_1 \mid y_1, y_2)$ would follow a univariate t-distribution.

Since the search for an exact predictive distribution proves difficult, an alternative approach is to derive an approximate predictive distribution. Note that given hyperparameters, $f(X^0, M_1, M_2)$ follows a matrix T-distribution. By Theorem 2.2, $f(X^0, X^{(1)})$ is approximately normal and then by Lemma 2.7, $f(vec(X^0), vec(M_1), vec(M_2))$ is approximately multivariate normal. General normal theory implies that $f(vec(X^0) \mid vec(M_2))$ is approximately normal. Theorem 2.3 shows that the order of approximation is $O((\delta^*)^{-\frac{1}{2}})$.

The theorem below states an approximate predictive distribution of $f(vec(X^0) \mid vec(M_2))$. Let

$$R^* = \begin{pmatrix} I_{s_u k \times s_u k} & 0 \\ 0 & R \end{pmatrix},$$

where $R = (R_1, R_2)$ and $R_1^t vec((X^{(1)})^t) = vec(M_1)$, $R_2^t vec((X^{(1)})^t) = vec(M_2)$.

Theorem 2.12 *$f(vec((X^0)^t) \mid vec(M_2))$ is approximately normal with mean*

$$vec((B_0^o Z)^t) + \Sigma_{X^0 M_2}(\Sigma^*_{M_2 M_2})^{-1}(vec(M_2) - R_2^t vec((B^o_{(1)} Z)^t))$$

and covariance

$$\Sigma_{X^0 X^0} - \Sigma_{X^0 M_2}(\Sigma^*_{M_2 M_2})^{-1}\Sigma^t_{X^0 M_2},$$

where

$$\Sigma_{XX} = [(Z^t F^{-1} Z + I_n) \otimes \Phi]/(\delta^* + n);$$

$$\Sigma_{X^0 X^0} = (\, I_{s_u k \times s_u k} \quad 0\,)\,\Sigma_{XX} \begin{pmatrix} I \\ 0 \end{pmatrix};$$

$$\Sigma_{X^0 M_2} = (\, I \quad 0\,)\,\Sigma_{XX} \begin{pmatrix} 0_{(s_g k - l^*) \times (s_g k - l^*)} \\ R_2 \end{pmatrix};$$

$$\Sigma_{M_2 M_2} = (\, 0 \quad R_2^t\,)\,\Sigma_{XX} \begin{pmatrix} 0 \\ R_2 \end{pmatrix}.$$

If the loss function is "mean squared error", the approximately Bayesian interpolator is simply the mean of the above approximate predictive distribution.

Following the same approach of the previous section, an empirical Bayesian method is obtained. In other words, all the hyperparameters are estimated using the observations. Since no unbiased estimators of B^o, F^{-1} has yet been found, the estimation of these hyperparameters will not be discussed here. In applications, they will be specified as plausible values. Hence, in the following discussion, these two hyperparameters are assumed to be known. Again, the estimation of hyperparameters requires two steps. In step one, an EM algorithm is applied to estimate δ, $\Phi_{(11)} = \Lambda_g \otimes \Omega$. In step two, the SG method is applied to extend Λ_g so as to obtain Λ.

In the EM algorithm, the likelihood function is $f(X^{(1)} \mid \Phi_{(11)}, \delta)$. By the same argument as that of the proof of Theorem 2.12,

$$f(vec((X^{(1)})^t) \mid \Phi_{(11)}, \delta) \overset{\text{approx.}}{\sim} \mathcal{N}(\frac{Z^t F^{-1} Z + I}{\delta^* - s_u k + n}, \Phi_{(11)}) + vec((B^0_{(1)} Z)^t).$$

So by Equation (2.9), $f(X^{(1)} \mid \Phi_{(11)}, \delta)$ is approximately proportional to

$$\mid \Lambda_g \mid^{-\frac{hk}{2}} \mid \Omega \mid^{-\frac{hs_g}{2}} (\delta^* - s_u k + n)^{\frac{s_g kh}{2}} e^{-\frac{1}{2} tr[(\Lambda_g^{-1} \otimes \Omega^{-1})A]},$$

where

$$A = a^*(X^{(1)} - B^o_{(1)}Z)(Z^t F^{-1} Z + I)^{-1}(X^{(1)} - B^o_{(1)}Z)^t$$

and $a^* = \delta^* - s_u k + n$.

For the E-step, $E(M_1 \mid M_2)$ is given by the following lemma.

Lemma 2.19 *Given hyperparameters, $f(vec(M_1) \mid vec(M_2), \Phi_{(11)}, \delta)$ is approximately normal with mean*

$$R_1^t vec((B^o_{(1)}Z)^t) + \Sigma_{M_1M_2}\Sigma^{-1}_{M_2M_2}(vec(M_2) - R_2^t vec((B^o_{(1)}Z)^t)$$

and covariance

$$\Sigma_{M_1M_1} - \Sigma_{M_1M_2}\Sigma_{M_2M_2}\Sigma_{M_2M_1},$$

where

$$\Sigma_{X_{(1)}X_{(1)}} = [\Phi_{(11)} \otimes (Z^t F^{-1} Z + I)]/(\delta^* - s_u k + n);$$

$$\Sigma_{M_iM_j} = R_i^t \Sigma_{X_{(1)}X_{(1)}} R_j, i, j = 1, 2.$$

The M-step is similar to that in the previous section. Its discussion will not be repeated here.

The EM algorithm is summarized below.

E-STEP: Given the current values of $\Phi_{(11)}$, δ^*, $B^0_{(1)}$,

$$E(vec(M_1) \mid vec(M_2)) = R_1^t vec(B^o_{(1)}Z) + \Sigma_{M_1M_2}\Sigma^{-1}_{M_2M_2}(vec(M_2) - R_2^t vec(B^o_{(1)}Z))$$

M-STEP: Given the current values of $vec(M_1)$, update $\Phi_{(11)} = \Lambda_g \otimes \Omega$, δ^* by repeating the following steps until it converges.

Step 1:

Given the current $(\Lambda_g^{-1})^{(p)}$ and $(\Omega^{-1})^{(p)}$, represent $tr[((\Lambda_g^{-1})^p \otimes (\Omega^{-1})^{(p)})A]$ as $tr((\Omega^{-1})^{(p)}G)$ and set $(\Omega^{-1})^{(p+1)} = hs_gG^{-1}$;

Step 2:

Given the current $(\Lambda_g^{-1})^{(p)}$ and $(\Omega^{-1})^{(p+1)}$, represent $tr[((\Lambda_g^{-1})^p \otimes (\Omega^{-1})^{(p+1)})\ A]$ as $tr((\Lambda_g^{-1})^{(p)}Q)$ and set $(\Lambda_g^{-1})^{(p+1)} = hkQ^{-1}$.

Step 3:

Given the current $\Phi_{(11)} = \Lambda_g \otimes \Omega$, update a^* by

$$a^* = s_g kh/tr[\Phi_{(11)}^p(X^{(1)} - B_{(1)}Z)(Z^tF^{-1} + I)^{-1}(X^{(1)} - B_{(1)}Z)^t]$$

The estimated δ^* is $max\{sk, a^* - n - f\}$.

2.6 Interpolation with Monotone Missing Data Patterns

Let X_t^t be generally partitioned as $((X_t^0)^t, (X_t^1)^t, \ldots, (X_t^r)^t)$ and $(X_t^{(i)})^t = ((X_t^i)^t, \ldots, (X_t^r)^t)$, $i = 0, \ldots, r$. Σ is partitioned correspondingly as

$$\Sigma_{(ii)} = \begin{pmatrix} \Sigma_{ii} & \Sigma_{i(i+1)} \\ \Sigma_{(i+1)i} & \Sigma_{([i+1][i+1])} \end{pmatrix}, \quad i = 0, 1, \ldots, r-1.$$

In the above, Σ_{ii}, $\Sigma_{(ii)}$ are the covariance matrices of X_t^i and $X_t^{(i)}$, $i = 0, \ldots, r$ respectively. Note that $\Sigma_{(00)} = \Sigma$. A reparameterization is recursively taken as

$$\Sigma_{(ii)} \longrightarrow \{\Gamma_i, \Sigma_{([i+1][i+1])}\}, i = 0, \ldots, r-1,$$

where $\Gamma_i = \{\tau_{j(j+1)}, \Sigma_{j|(j+1)}\}_{j=0}^{i}$ and

$$\tau_{i(i+1)} = \Sigma_{i(i+1)}\Sigma^{-1}_{([i+1][i+1])}, \Sigma_{i|(i+1)} = \Sigma_{ii} - \tau_{i(i+1)}\Sigma^{-1}_{([i+1][i+1])}\tau^t_{i(i+1)}, \quad i = 0, \ldots, r-1.$$

B is partitioned as $B^t = (B_0^t, B_1^t, \ldots, B_r^t)$ accordingly.

The lemma below gives the intuition behind the reparameterization. In the lemma, Φ is partitioned in the same fashion as Σ.

Lemma 2.20 *If Σ has the prior distribution of (2.14), the following are true:*

1. $\Sigma_{([i+1][i+1])}$ *is independent of* $\Gamma_j, j = 0, 1, \ldots, i$.
2. $\Sigma_{([i+1][i+1])} \sim W^{-1}(\Phi_{([i+1][i+1])}, \delta^* - sk + l_{(i+1)})$, *where* $l_{(i+1)}$ *is the dimension of* $X_t^{(i+1)}$.
3. $\Sigma_{i|(i+1)} \sim W^{-1}(\Phi_{i|(i+1)}, \delta^* - sk + l_{(i)})$, *where* $l_{(i)}$ *is the dimension of* $X_t^{(i)}$ *and* $\Phi_{i|(i+1)} = \Phi_{ii} - \Phi_{i(i+1)}\Phi^{-1}_{([i+1][i+1])}\Phi_{(i+1)i}$.
4. $\tau_{i(i+1)} \mid \Sigma_{i|(i+1)} \sim N(\Phi_{i(i+1)}\Phi^{-1}_{([i+1][i+1])}, \Sigma_{i|(i+1)} \otimes \Phi^{-1}_{([i+1][i+1])})$

This result is called *Generalized Inverted Wishart Distribution* by Brown, Le and Zidek (1994b) (see Caselton, Kan and Zidek (1990), LZ for its earlier forms).

Suppose $(X_i^{(1)})^t = (U_i^t, G_i^t)$, $i = 1, \ldots, n$, where $U_i^t : 1 \times d_i$ is missing, $G_i^t : 1 \times (s_g k - d_i)$ is observed and $X_t^{(1)}$ is the matrix of response vectors at the gauged sites. When $d_1 \leq d_2 \leq \ldots \leq d_n$, the pattern of the missing data $\{U_i^t\}_{i=1}^n$ resembles an upside down staircase. An example of such missing data occurs when at gauged Site 1 the pollutants are measured up to time T_1, at gauged Site 2 up to T_2, ..., at gauged Site n up to T_n and when $T_1 < T_2 < \ldots < T_n$; the missing data of $[X^{(1)}]^t$ have the shape of staircase. Such a missing pattern is called "monotone"; in Little and Rubin (1987).

The following theorem gives a recursively predictive distribution.

Theorem 2.13 *Suppose X_t, $t = 1, \ldots, n$ follow the model of (2.12) and B, Σ follow the prior of (2.13), (2.14). Then, given all hyperparameters,*

$$U_j \mid \{G_i\}_{i=j}^n, D_j \sim T((\Phi^j_{1|(2)})^{-1}, a_2^j, a_1^j, \delta^* + j - s_u)k, \quad j = 1, \ldots, n,$$

where:

$$D_j = \{\{X_i^{(1)}\}_{i=1}^{j-1}\};$$

$$a_1^j = (B_1^j)^* Z_j - \hat{\Phi}^j_{1(2)}(\hat{\Phi}^j_{(22)})^{-1}(B^j_{(2)})^* Z_j + \hat{\Phi}^j_{1(2)}(\hat{\Phi}^j_{(22)})^{-1} G_j;$$

$$a_2^j = 1 + (G_j - (B^j_{(2)})^* Z_j)^t (\hat{\Phi}^j_{(22)})^{-1}(G_j - (B^j_{(2)})^* Z_j) + Z_j^t F_j^* Z_j;$$

$$(B^j_{(1)})^* = B^0_{(1)} + (\hat{B}^j_{(1)} - B^0_{(1)})\hat{E}_j^t;$$

$$(B^j_{(1)})^* = \begin{pmatrix} (B_1^j)^* \\ (B^j_{(2)})^* \end{pmatrix} \quad (B_1^j)^* : d_j \times h, \quad (B^j_{(2)})^* : (s_g k - d_j) \times h;$$

$$\hat{B}^j_{(1)} = C_j A_j^{-1};$$

$$S_j = \sum_{k=1}^{j-1} (X_k^{(1)} - \hat{B}^j_{(1)} Z_k)(X_k^{(1)} - \hat{B}^j_{(1)} Z_k)^t;$$

$$C_j = \sum_{k=1}^{j-1} X_k^{(1)} Z_k^t; \quad A_j^{-1} = \sum_{k=1}^{j-1} Z_k Z_k^t;$$

$$\hat{E}_j = F^{-1}(A_j^{-1} + F^{-1})^{-1}; \quad F_j^* = (I - \hat{E}_j)F^{-1};$$

$$\hat{\Phi}^j_{(11)} = \Phi_{(11)} + S_j + (\hat{B}^j_{(1)} - B^0_{(1)})^t (A_j^{-1} + F^{-1})^{-1}(\hat{B}^j_{(1)} - B^0_{(1)});$$

$$\hat{\Phi}^j_{(11)} = \begin{pmatrix} \hat{\Phi}^j_{11} & \hat{\Phi}^j_{1(2)} \\ \hat{\Phi}^j_{(2)1} & \hat{\Phi}^j_{(22)} \end{pmatrix}; \quad \hat{\Phi}^j_{(11)} : d_j \times d_j, \quad \hat{\Phi}^j_{(22)} : (s_g k - d_j) \times (s_g k - d_j).$$

For the case of a monotone missing data pattern, finding an EM algorithm estimation procedure for hyperparameters is difficult. As a make-shift measure, we can treat it as a randomly missing data problem; the EM algorithm proposed in the previous section may then be applied.

2.7 Proofs of Theorems, Corollaries and Lemmas

Proof of Lemma 2.9 :

Proof: By (2.11), $U = \sum_{i=1}^{\vartheta} Y_i Y_i^t$ and $Y_i Y_i^t \sim^{\text{i.i.d.}} N(0, P)$. Proposition 8.3 (ii) (cf. Eaton 1983 p305) implies $Cov(vec(Y_i Y_i^t)) = Cov(Y_i Y_i^t) = 2P \otimes P$. With the multivariate central limit theorem,

$$\sqrt{\vartheta}(\frac{vec(U^t)}{\vartheta} - vec(P)) \xrightarrow{\text{d}} \mathcal{N}(2P, P).$$

Since $vec(U^t)$ and U^t are only notationally different and a matrix normal distribution is distributionally equivalent to a multivariate normal distribution, by the definition of convergence in distribution (Billingsley 1968), it is easy to see that

$$\sqrt{\vartheta}(\frac{U^t}{\vartheta} - P) \xrightarrow{\text{d}} \mathcal{N}(2P, P).\blacksquare$$

Proof of Theorem 2.3 :

Since

$$\sqrt{\vartheta}(T^* - P^{-\frac{1}{2}}X) = \sqrt{\vartheta}\left(T^* - P^{-\frac{1}{2}}\left(\frac{\vartheta}{\delta}\right)^{-\frac{1}{2}}X\right) + \sqrt{v}\left(\left(\frac{\vartheta}{\delta}\right)^{-\frac{1}{2}} - 1\right)P^{-\frac{1}{2}}X$$

and

$$\begin{aligned}
&\sqrt{\vartheta}\left(T^* - P^{-\frac{1}{2}}\left(\frac{\vartheta}{\delta}\right)^{-\frac{1}{2}}X\right) \\
&= \sqrt{\vartheta}\left(\left(\frac{U^t}{\vartheta}\right)^{-\frac{1}{2}}P^{\frac{1}{2}} - I\right)P^{-\frac{1}{2}}\left(\frac{\vartheta}{\delta}\right)^{-\frac{1}{2}}X \\
&= \sqrt{\vartheta}\left(I - \left(\frac{U^t}{\vartheta}\right)^{\frac{1}{2}}P^{-\frac{1}{2}}\right)\left(\frac{U^t}{\vartheta}\right)^{-\frac{1}{2}}P^{\frac{1}{2}}P^{-\frac{1}{2}}\left(\frac{\vartheta}{\delta}\right)^{-\frac{1}{2}}X \\
&= \sqrt{\vartheta}\left(P - \frac{U^t}{\vartheta}\right)P^{-1}\left(I + \left(\frac{U^t}{\vartheta}\right)^{\frac{1}{2}}P^{-\frac{1}{2}}\right)^{-1}\left(\frac{U^t}{\vartheta}\right)^{-\frac{1}{2}}\left(\frac{\vartheta}{\delta}\right)^{-\frac{1}{2}}X.
\end{aligned}$$

By Lemma 2.9:

$$\sqrt{\vartheta}\left(P - \frac{U^t}{\vartheta}\right) \xrightarrow{\text{d}} \mathcal{N}(2P, P) \quad as \quad \vartheta \longrightarrow \infty.$$

By LLN and Slutsky Theorem, as $\vartheta \longrightarrow \infty$

$$\left(I + \left(\frac{U^t}{\vartheta}\right)^{\frac{1}{2}} P^{-\frac{1}{2}}\right)^{-1} \longrightarrow \frac{1}{2} I \quad a.s.;$$

$$\left(\frac{U^t}{\vartheta}\right)^{-\frac{1}{2}} \longrightarrow P^{-\frac{1}{2}} \quad a.s.;$$

$$\left(\frac{\vartheta}{\delta}\right)^{-\frac{1}{2}} \longrightarrow 1.$$

So by Slusky theorem

$$\sqrt{\vartheta}\left(T^* - P^{-\frac{1}{2}}\left(\frac{\vartheta}{\delta}\right)^{-\frac{1}{2}} X\right) \xrightarrow{\mathrm{d}} \frac{1}{2}\mathcal{N}(2P, P)P^{-1}P^{-\frac{1}{2}}\mathcal{N}(I, Q)$$

or

$$\sqrt{\vartheta}\left(T^* - P^{-\frac{1}{2}}\left(\frac{\vartheta}{\delta}\right)^{-\frac{1}{2}} X\right) \xrightarrow{\mathrm{d}} \mathcal{N}(\frac{1}{2}P, P^{-1})\mathcal{N}(P^{-1}, Q).$$

Given P, Q, the two normal distributions are independent, since U and X are independent.

Further, since

$$\begin{aligned}\sqrt{\vartheta}\left(\left(\frac{\vartheta}{\delta}\right)^{-\frac{1}{2}} - 1\right) &= \frac{\delta - \vartheta}{\sqrt{\vartheta}\left[\left(\frac{\delta}{\vartheta}\right)^{\frac{1}{2}} + 1\right]} \\ &= \frac{q}{\sqrt{\vartheta}\left[\left(\frac{\delta}{\vartheta}\right)^{\frac{1}{2}} + 1\right]} \\ &\longrightarrow 0 \qquad as \quad \vartheta \longrightarrow \infty.\end{aligned}$$

The above is true, since $\vartheta = \delta - q$ and q is fixed.

Reapply Slusky theorem. The conclusion is followed. ∎

Proof of Theorem 2.4 :

Given B^o, F, Φ, δ^*, (2.12), (2.13) imply

$$X \mid \Sigma, B \stackrel{\mathrm{d}}{=} BZ + N(0, \Sigma \otimes I_{n\times n})$$

and

$$B \mid \Sigma \stackrel{\mathrm{d}}{=} B^o + N(0, \Sigma \otimes F^{-1}).$$

In turn,

$$X \mid \Sigma \stackrel{\mathrm{d}}{=} B^o Z + N(0, \Sigma \otimes P)$$

where $P = Z^t F^{-1} Z + I_{n \times n}$. So

$$X^t \mid \Sigma \stackrel{\mathrm{d}}{=} Z^t (B^o)^t + N(0, P \otimes \Sigma).$$

Equation (2.14) and (2.8) imply,

$$X^t - Z^t (B^o)^t \sim T(P^{-1}, \Phi, 0, \delta^* + n).$$

By Bartlett's decomposition, $\Phi = \Delta \Lambda \Delta^t$, where

$$\Delta = \begin{pmatrix} I & \eta \\ 0 & I \end{pmatrix} \quad , \quad \Lambda = \begin{pmatrix} \Phi_{0|(1)} & 0 \\ 0 & \Phi_{(11)} \end{pmatrix}$$

and $\eta = \Phi_{0(1)} \Phi_{(11)}^{-1}$. Now apply Lemma 2.4 part 3 to get

$$(X^t - Z^t (B^o)^t)(\Delta^t)^{-1} \sim T(P^{-1}, \Lambda, 0, \delta^* + n).$$

Then by (2.7)

$$\begin{pmatrix} I & -\eta \\ 0 & I \end{pmatrix} X = \begin{pmatrix} t_1 - \eta t_2 \\ t_2 \end{pmatrix} \sim T(\begin{pmatrix} \Phi_{0|(1)}^{-1} & 0 \\ 0 & \Phi_{(11)}^{-1} \end{pmatrix}, P, \begin{pmatrix} B_0^o - \eta B_{(1)}^o \\ B_{(1)}^o \end{pmatrix} Z, \delta^* + n)$$

where $X^0 = t_1$ and $X^{(1)} = t_2$. By Lemma 2.4

$$t_1 - \eta t_2 - B_0^o Z + \eta B_{(1)}^o Z \mid t_2 \sim T(\Phi_{0|(1)}^{-1}, P + (t_2 - B_{(1)}^o Z)^t \Phi_{(11)}^{-1} (t_2 - B_{(1)}^o Z),$$

$$0, \delta^* + n).$$

The theorem is proved by reapplying the same lemma. ∎

Proof of Theorem 2.5 :

By (2.12)-(2.14) and Lemma 2.20

$$(X^{(1)})^t \mid \Sigma_{(11)} \stackrel{\mathrm{d}}{=} Z^t (B_{(1)}^o)^t + N(0, P \otimes \Sigma_{(11)}), \quad \Sigma_{(11)} \sim W_{s_g k}^{-1} \left(\Phi_{(11)}, \delta^* - s_u k \right)$$

where $P = Z^t F^{-1} Z + I_{n\times n}$. Thus

$$(X^{(1)})^t - (B^o_{(1)}Z)^t \sim T(P^{-1}, \Phi_{(11)}, 0, \delta^* - s_u k + n).$$

Apply Lemma 2.4 part 3:

$$(X^{(1)})^t R - (B^o_{(1)}Z)^t R \sim T(P^{-1}, R^t\Phi_{(11)}R, 0, \delta^* - s_u k + n).$$

The remaining is the same as in the proof of Theorem 2.4. ∎

Proof of Corollary 2.6 :

$$\begin{aligned}
E(X^0 \mid X^2) &= E(E(X^0 \mid X^{(1)}) \mid X^2) \\
&= B^o_0 Z + \Phi_{0(1)}\Phi^{-1}_{(11)}(E(X^{(1)} \mid X^2) - B^o_{(1)}Z) \\
&= B^o_0 Z + \Phi_{0(1)}\Phi^{-1}_{(11)}\left[R_1 E(X^1 \mid X^2) + R_2 X^2 - B^o_{(1)}Z\right].
\end{aligned}$$

The last equation is true since $I = RR^t = R_1R_1^t + R_2R_2^t$, $X^1 = R_1^t X^{(1)}$ and $X^2 = R_2^t X^{(1)}$. By Corollary 2.5

$$\begin{aligned}
E(X^1 \mid X^2) &= B^o_0 Z + \Phi_{0(1)}\Phi^{-1}_{(11)}\{R_1\left(R_1^t B^o_{(1)}Z + \gamma(X^2 - R_2^t B^o_{(1)}Z)\right) \\
&\quad + R_2 X^2 - B^o_{(1)}Z\} \\
&= B^o_0 Z + \Phi_{0(1)}\Phi^{-1}_{(11)}\{(R_1R_1^t - R_1\gamma R_2^t - I)B^o_{(1)}Z \\
&\quad + (R_2 + R_1\gamma)X^2\} \\
&= B^o_0 Z + \Phi_{0(1)}\Phi^{-1}_{(11)}\{(-R_1\gamma R_2^t - R_2R_2^t)B^o_{(1)}Z \\
&\quad + (R_2 + R_1\gamma)X^2\} \\
&= B^o_0 Z + \Phi_{0(1)}\Phi^{-1}_{(11)}(R_1\gamma + R_2)(X^2 - R_2^t B^o_{(1)}Z) \\
&= B^o_0 Z + \Phi_{0(1)}R_2\Psi^{-1}_{22}(X^2 - R_2^t B^o_{(1)}Z).
\end{aligned}$$

The third last step is true because $RR^t = R_1R_1^t + R_2R_2^t = I$ and the last step is true since

$$R^t\Phi_{(11)}R = \Psi_{(11)}$$

implies

$$R^t\Phi_{(11)}R_2 = \begin{pmatrix} \Psi_{12} \\ \Psi_{22} \end{pmatrix}.$$

So

$$\Phi_{(11)}R_2 = R\begin{pmatrix} \Psi_{12} \\ \Psi_{22} \end{pmatrix} = R_1\Psi_{12} + R_2\Psi_{22}.$$

That is

$$R_2\Psi_{22}^{-1} = \Phi_{(11)}^{-1}(R_1\Psi_{12}\Psi_{22}^{-1} + R_2) = \Phi_{(11)}^{-1}(R_1\gamma + R_2).\blacksquare$$

Proof of Lemma 2.10 :

By using Bartlett decomposition, it is easy to see that

$$b = (a - A_{12}A_{22}^{-1}A_{12}^t)^{-1};$$

$$B_{12} = -A_{12}A_{22}^{-1}b;$$

$$B_{22} = A_{22}^{-1} + A_{22}^{-1}A_{12}^t bA_{12}A_{22}^{-1}.$$

For notational simplicity, let $d = b^{-1} = a - A_{12}A_{22}^{-1}A_{12}^t$ which is a scalar. Then

$$\begin{aligned} b - B_{12}b_{22}^{-1}B_{12}^t &= d^{-1} - A_{12}A_{22}^{-1}d^{-1}(A_{22}^{-1} + A_{22}^{-1}A_{12}^t d^{-1}A_{12}A_{22}^{-1})^{-1}d^{-1}A_{22}^{-1}A_{12}^t \\ &= d^{-2}[d - A_{12}(A_{22} + A_{12}^t d^{-1}A_{12})^{-1}A_{12}^t] \\ &= d^{-1}[1 - A_{12}(A_{22}d + A_{12}^t A_{12})^{-1}A_{12}^t]. \end{aligned} \tag{2.21}$$

For showing the lemma, note that $A_{12}A_{22}^{-1}A_{12}^t$ is a scalar and

$$\begin{aligned} A_{12} &= A_{12} - a^{-1}A_{12}(A_{12}A_{22}^{-1}A_{12}^t) + a^{-1}A_{12}(A_{12}A_{22}^{-1}A_{12}^t) \\ &= A_{12}a^{-1}[a - A_{12}A_{22}^{-1}A_{12}^t] + a^{-1}A_{12}(A_{12}A_{22}^{-1}A_{12}^t) \\ &= A_{12}a^{-1}d + a^{-1}(A_{12}A_{22}^{-1}A_{12}^t)A_{12} \\ &= a^{-1}A_{12}A_{22}^{-1}(A_{22}d + A_{12}^t A_{12}). \end{aligned}$$

Equivalently,

$$A_{12}(A_{22}d + A_{12}^t A_{22})^{-1} = a^{-1}A_{12}A_{22}^{-1}.$$

Then,

$$1 - A_{12}(A_{22}d + A_{12}^t A_{12})^{-1} A_{12}^t = 1 - a^{-1} A_{12} A_{22}^{-1} A_{12}^t = a^{-1} d.$$

Combined with (2.21), the lemma is proved. ∎

Proof of Theorem 2.6 :

Given all the hyperparameters,

$$Var(X_t^0 \mid X^2) = Var(E(X_t^0 \mid X^{(1)}) \mid X^2)$$

$$+E[Var(X_t^0 \mid X^{(1)}) \mid X^2].$$

If $t = n$, by (2.7), Lemma 2.4, Theorem 2.4 and Lemma 2.10, it is obvious that

$$X_t^0 \mid X^{(1)} \sim T(\Phi_{0|(1)}^{-1}, g_t, B_0^o Z_t + \Phi_{0(1)}\Phi_{(11)}^{-1}(X_t^{(1)} - B_{(1)}^0 Z_t), \delta^* + 1), \tag{2.22}$$

where

$$g_t = 1 + Z_t^t F^{-1} Z_t + (X_t^{(1)} - B_{(1)}^o Z_t)^t \Phi_{(11)}^{-1}(X_t^{(1)} - B_{(1)}^o Z_t).$$

For $t \in \{1, \ldots, n-1\}$, let C be the orthogonal matrix such that

$$X^0 C = (X_1^0, \ldots, X_{t-1}^0, X_{t+1}^0, \ldots, X_n^0, X_t^0).$$

By Lemma 2.4 (III) and Theorem 2.4,

$$X^0 C \mid X^{(1)} \sim T(\Phi_{0|(1)}^{-1}, I_n + C^t Z^t F^{-1} Z C$$

$$+(X^{(1)}C - B_{(1)}^o ZC)^t \Phi_{(11)}^{-1}(X^{(1)}C - B_{(1)}^o ZC),$$

$$B_0^o ZC + \Phi_{0(1)}\Phi_{(11)}^{-1}(X^{(1)}C - B_{(1)}^o ZC), \delta^* + n).$$

Note that the last column of ZC is Z_t, the last diagonal element of $C^t Z^t F^{-1} Z C$ is $Z_t^t F^{-1} Z_t$ and the last column of $X^{(1)}C$ is $X_t^{(1)}$. So it shows that (2.22) is true for any t. By Lemma 2.2 and Press (1982 p128),

$$E(X_t^0 \mid X^{(1)}) = B_0^o Z_t + \Phi_{0(1)}\Phi_{(11)}^{-1}(X_t^{(1)} - B_{(1)}^o Z_t)$$

and

$$Var(X_t^0 \mid X_t^{(1)}) = \frac{g_t}{m+l-2}\Phi_{0|(1)}.$$

Further,

$$\begin{aligned}\Phi_{0(1)}\Phi_{(11)}^{-1}(X_t^{(1)} - B_{(1)}^o Z_t) &= \Phi_{0(1)}\Phi_{(11)}^{-1}RR^t(X_t^{(1)} - B_{(1)}^o Z_t) \\ &= (A_1, A_2)\begin{pmatrix} X_t^1 - B_1^o Z_t \\ X_t^2 - B_2^o Z_t \end{pmatrix} \\ &= A_1(X_t^1 - B_1^o Z_t) + A_2(X_t^2 - B_2^o Z_t).\end{aligned}$$

By Theorem 2.5 and a similar argument as the case of $X_t^0 \mid X^{(1)}$,

$$X_t^1 \mid X_t^2 \sim T(\Psi_{1|2}^{-1}, c_t, B_1^0 Z_t + \gamma(X_t^2 - B_2^0 Z_t), m+l). \tag{2.23}$$

So

$$\begin{aligned}Var(E(X_t^0 \mid X_t^{(1)}) \mid X^2) &= Var(\Phi_{0(1)}\Phi_{(11)}^{-1}(X_t^{(1)} - B_{(1)}^o Z_t) \mid X^2) \\ &= A_1 Var(X_t^1 \mid X^2)A_1^t \\ &= \frac{A_1 c_t \Psi_{1|2} A_1^t}{m-2}.\end{aligned}$$

The last equation is true, because of Lemma 2.2 and Press (1982 p128). To find $E[Var(X_t^0 \mid X_t^{(1)}) \mid X_t^2]$, let us start with

$$\begin{aligned}R^t\Phi_{(11)}^{-1}R &= \begin{pmatrix} \Psi_{11} & \Psi_{12} \\ \Psi_{21} & \Psi_{22} \end{pmatrix}^{-1} \\ &= \begin{pmatrix} I & -r \\ 0 & I \end{pmatrix}^t \begin{pmatrix} \Psi_{1|2}^{-1} & 0 \\ 0 & \Psi_{22}^{-1} \end{pmatrix} \begin{pmatrix} I & -r \\ 0 & I \end{pmatrix}.\end{aligned}$$

So

$$\begin{aligned}&E\left[(X_t^{(1)} - B_{(1)}^o Z_t)^t \Phi_{(11)}^{-1}(X_t^{(1)} - B_{(1)}^o Z_t) \mid X_t^2\right] \\ &= E\left[(R^t X_t^{(1)} - R^t B_{(1)}^o Z_t)^t R^t \Phi_{(11)}^{-1} R(R^t X_t^{(1)} - R^t B_{(1)}^o Z_t) \mid X_t^2\right] \\ &= E[\begin{pmatrix} (X_t^1 - B_1^o Z_t) - r(X_t^2 - B_2^o Z_t) \\ X_t^2 - B_2^o Z_t \end{pmatrix}^t \begin{pmatrix} \Psi_{1|2}^{-1} & 0 \\ 0 & \Psi_{22}^{-1} \end{pmatrix} \\ &\quad \begin{pmatrix} (X_t^1 - B_1^o Z_t) - r(X_t^2 - B_2^o Z_t) \\ X_t^2 - B_2^o Z_t \end{pmatrix} \mid X_t^2] \\ &= E\{[(X_t^1 - B_1^o Z_t) - r(X_t^2 - B_2^o Z_t)]^t \Psi_{1|2}^{-1}[(X_t^1 - B_1^o Z_t) \\ &\quad -r(X_t^2 - B_2^o Z_t)] \mid X_t^2\} + (X_t^2 - B_2^o Z_t)^t \Psi_{22}^{-1}(X_t^2 - B_2^o Z_t).\end{aligned}$$

Let $Y = \Psi_{1|2}^{-\frac{1}{2}}[(X_t^1 - B_1^o Z_t) - r(X_t^2 - B_2^o Z_t)]$. By (2.23) and Lemma 2.4,

$$Y \mid X_t^2 \sim T(I, c_t, 0, m+l).$$

Lemma 2.2 and Equation 2.4 imply

$$f(Y = y \mid X_t^2) = \frac{\Gamma\left(\frac{m+l}{2}\right)}{\pi^{\frac{l}{2}}\Gamma\left(\frac{m}{2}\right)} c_t^{\frac{m}{2}} (c_t + y^t y)^{-\frac{m+l}{2}}.$$

Then

$$\begin{aligned} E(Y^tY) &= E(Y^tY + c_t) - c_t \\ &= \frac{(m-2+l)c_t}{m-2} \int_0^\infty \frac{\Gamma\left(\frac{m-2+l}{2}\right)}{\pi^{\frac{l}{2}}\Gamma\left(\frac{m-2}{2}\right)} c_t^{\frac{m-2}{2}} (c_t + y^t y)^{-\frac{m-2+l}{2}} dy - c_t \\ &= \frac{m-2+l}{m-2} c_t - c_t. \end{aligned}$$

That implies

$$\begin{aligned} & E\{(X_t^{(1)} - B_{(1)}^o Z_t)^t \Phi_{(11)}^{-1} (X_t^{(1)} - B_{(1)}^o Z_t) \mid X_t^{(1)}\} \\ = & \frac{m-2+l}{m-2} c_t - c_t + (X_t^2 - B_2^o Z_t)^t \Psi_{22}^{-1} (X_t^2 - B_2^o Z_t). \end{aligned}$$

And

$$\begin{aligned} E[Var(X_t^0 \mid X_t^{(1)}) \mid X_t^2] &= \frac{1}{m+l-2}[c_t + \frac{m-2+l}{m-2} c_t - c_t] \Phi_{0|(1)} \\ &= \frac{c_t}{m-2} \Phi_{0|(1)}. \end{aligned}$$

Finally,

$$\begin{aligned} Var(X_t \mid X_t^2) &= \frac{A_1 c_t \Psi_{1|2} A_1^t}{m-2} + \frac{c_t \Phi_{0|(1)}}{m-2} \\ &= \frac{c_t}{m-2}[A_1 \Psi_{1|2} A_1^t + \Phi_{0|(1)}]. \blacksquare \end{aligned}$$

Proof of Lemma 2.11 :

By $R^t \Phi_{(11)} R = \Psi_{(11)}$, the following are true:

$$\Phi_{(11)} R_2 = R \begin{pmatrix} \Psi_{12} \\ \Psi_{22} \end{pmatrix} = R_1 \Psi_{12} + R_2 \Psi_{22}, \tag{2.24}$$

$$\Phi_{(11)} = R_1 \Psi_{11} R_1^t + R_2 \Psi_{21} R_1^t + R_1 \Psi_{12} R_2^t + R_2 \Psi_{22} R_2^t. \tag{2.25}$$

By (2.24) and (2.25)

$$\begin{aligned}\Phi_{(11)}R_2\Psi_{22}^{-1}R_2^t\Phi_{(11)} &= (R_1\Psi_{12}+R_2\Psi_{22})\Psi_{22}^{-1}(R_1\Psi_{12}+R_2\Psi_{22})^t \\ &= R_1\Psi_{12}\Psi_{22}^{-1}\Psi_{21}R_1^t+\Phi_{(11)}-R_1\Psi_{11}R_1^t \\ &= \Phi_{(11)}-R_1\Psi_{1|2}R_1^t.\end{aligned}$$

Hence,

$$R_1\Psi_{1|2}R_1^t=\Phi_{(11)}-\Phi_{(11)}R_2\Psi_{22}^{-1}R_2^t\Phi_{(11)}$$

which implies

$$\Phi_{(11)}^{-1}R_1\Psi_{1|2}R_1^t\Phi_{(11)}^{-1}=\Phi_{(11)}^{-1}-R_2\Psi_{22}^{-1}R_2^t.$$

So

$$\Phi_{0(1)}\Phi_{(11)}^{-1}R_1\Psi_{1|2}R_1^t\Phi_{(11)}^{-1}\Phi_{0(1)}^t=\Phi_{0(1)}\Phi_{(11)}^{-1}\Phi_{0(1)}^t-\Phi_{0(1)}R_2\Psi_{22}^{-1}R_2^t\Phi_{0(1)}^t$$

or

$$\begin{aligned}A_1\Psi_{1|2}A_1^t &= -[\Phi_{00}-\Phi_{0(1)}\Phi_{(11)}^{-1}\Phi_{0(1)}^t]+[\Phi_{00}-\Phi_{0(1)}R_2\Psi_{22}^{-1}R_2^t\Phi_{0(1)}^t] \\ &= -\Phi_{0|(1)}+\Phi_{0|2}.\end{aligned}$$

It proves that

$$A_1\Psi_{1|2}A_1^t+\Phi_{0|(1)}=\Phi_{0|2}.\blacksquare$$

Proof of Theorem 2.8:

By Lemma 2.3, there exists $U\sim W_p(I_{p\times p},\delta^*-1)$ and $Y_{p\times 1}\sim N(0,I_{p\times p}\otimes 1)$ such that $T=(U^{-\frac{1}{2}})^tY$. By Lemma 2.12, $\frac{Y^tY}{Y^tU^{-1}Y}\sim\chi^2_{\delta^*-p}$ is independent of $Y^tY\sim\chi^2_p$. Therefore,

$$\frac{(\delta^*-p)T^tT}{p}=\frac{Y^tY/p}{\frac{Y^tY}{Y^tU^{-1}Y}/(\delta^*-p)}\sim F_{p,\delta^*-p}.\blacksquare$$

Proof of Theorem 2.9:

By Equation (2.7), Lemma 2.4 and Theorem 2.7,

$$\frac{\Phi_{0|2}^{-\frac{1}{2}}(X_t^0-\hat{x}_t^0)}{c_t^{\frac{1}{2}}}\sim T(I,1,0,\delta^*-l+1).$$

Then by Theorem 2.8,

$$\frac{(\delta^* - l - s_u k + 1)(X_t^0 - \hat{x}_t^0)^t \Phi_{0|2}^{-1}(X_t^0 - \hat{x}_t^0)}{s_u k * c_t} \sim F_{s_u k,\, \delta^* - l - s_u k + 1}.$$

So,

$$1 - \alpha = P((X_t^0 - \hat{x}_t^0)^t \Phi_{0|2}^{-1}(X_t^0 - \hat{x}_t^0) < b).\blacksquare$$

Proof of Lemma 2.13 :

By Bartlett's decomposition, $\Sigma = T\Lambda T^t$, where

$$\Lambda = \begin{pmatrix} \Sigma_{0|(1)} & 0 \\ 0 & \Sigma_{(11)} \end{pmatrix} \quad , \quad T = \begin{pmatrix} I & \tau \\ 0 & I \end{pmatrix}$$

and $\tau = \Sigma_{0(1)}\Sigma_{(11)}^{-1}$. Thus

$$\begin{aligned} \Sigma^{-1} &= \left[T\Lambda T^t\right]^{-1} \\ &= \begin{pmatrix} I & 0 \\ -\tau^t & I \end{pmatrix} \begin{pmatrix} \Sigma_{0|(1)}^{-1} & 0 \\ 0 & \Sigma_{(11)}^{-1} \end{pmatrix} \begin{pmatrix} I & -\tau \\ 0 & I \end{pmatrix} \\ &= \begin{pmatrix} \Sigma_{0|(1)}^{-1} & -\Sigma_{0|(1)}^{-1}\tau \\ -\tau^t\Sigma_{0|(1)}^{-1} & \Sigma_{(11)}^{-1} + \tau^t\Sigma_{0|(1)}^{-1}\tau \end{pmatrix}. \end{aligned}$$

Given Φ, δ^*, Lemma 2.1 implies

$$\Sigma_{0|(1)}^{-1} \mid X^{(1)} \sim W_{s_u k}(\Phi_{0|(1)}^{-1}, \delta^*),$$

$$\Sigma_{(11)}^{-1} \mid X^{(1)} \sim W_{s_g k}(\hat{\Phi}_{(11)}^{-1}, \delta^* + n - s_u k - h)$$

and

$$\tau \mid X^{(1)}, \Sigma_{0|(1)}^{-1}, \eta \sim N_{(s_u k)(s_g k)}(\eta, \Sigma_{0|(1)} \otimes \Phi_{(11)}^{-1}).$$

So

$$\begin{aligned} E(\Sigma_{0|(1)}^{-1} \mid X^{(1)}) &= \delta^* \Phi_{0|(1)}^{-1}, \\ E(\Sigma_{0|(1)}^{-1}\tau \mid X^{(1)}) &= E\left(\Sigma_{0|(1)}^{-1} E(\tau \mid X^{(1)}, \Sigma_{0|(1)}) \mid X^{(1)}\right) \\ &= E(\Sigma_{0|(1)}^{-1} \mid X^{(1)})\eta \\ &= \delta^* \Phi_{0|(1)}^{-1}\eta \end{aligned}$$

and

$$E(\Sigma_{(11)}^{-1} + \tau^t \Sigma_{0|(1)}^{-1} \tau \mid X^{(1)}) =$$

$$(\delta^* + n - s_u k - h)\hat{\Phi}_{(11)}^{-1} + E(E(\tau^t \Sigma_{0|(1)}^{-1} \tau \mid X^{(1)}, \Sigma_{0|(1)}) \mid X^{(1)}).$$

Note that, $\tau^t \Sigma_{0|(1)}^{-1} \tau = (\Sigma_{0|(1)}^{-\frac{1}{2}} \tau)^t \Sigma_{0|(1)}^{-\frac{1}{2}} \tau$. Let $Y = \Sigma_{0|(1)}^{-\frac{1}{2}} \tau - \Sigma_{0|(1)}^{-\frac{1}{2}} \eta$. Obviously by definition of a Wishart distribution,

$$Y \mid \Sigma_{0|(1)}, X^{(1)} \sim N_{(s_u k)(s_g k)}(0, I_{s_u k \times s_u k} \otimes \Phi_{(11)}^{-1}).$$

Then,

$$Y^t Y \mid \Sigma_{0|(1)}, X^{(1)} \sim W_{s_g k}(\Phi_{(11)}^{-1}, s_u k).$$

Thus

$$\begin{aligned} E(\Sigma_{(11)} + \tau^t \Sigma_{0|(1)}^{-1} \tau \mid X^{(1)}) &= (\delta^* + n - s_u k - h)\hat{\Phi}_{(11)}^{-1} + \\ &\quad E(\eta^t \Sigma_{0|(1)}^{-1} \eta + s_u k \Phi_{(11)}^{-1} \mid X^{(1)}) \\ &= (\delta^* + n - s_u k - h)\hat{\Phi}_{(11)}^{-1} + \delta^* \eta^t \Phi_{0|(1)}^{-1} \eta + s_u k \Phi_{(11)}^{-1}. \blacksquare \end{aligned}$$

Proof of Theorem 2.10 :

Note that since,

$$R^t X^{(1)} \mid Z, B, \Sigma \sim R^t B_{(1)} Z + N(0, R^t \Sigma_{(11)} R \otimes I_n),$$

$$R^t B_{(1)} \mid B_{(1)}^o, \Sigma, F \sim R^t B_{(1)}^o + N(0, R^t \Sigma_{(11)} R \otimes F^{-1}),$$

$$R^t \Sigma_{(11)} R \mid \Phi_{(11)}, \delta^* \sim W^{-1}(R^t \Phi_{(11)} R, \delta^* - s_u k)$$

and

$$E(\Sigma_{(11)}^{-1} \mid X^2, \Phi_{(11)}, \delta^*) = R E((R^t \Sigma_{(11)} R)^{-1} \mid X^2, \Phi_{(11)}, \delta^*) R^t.$$

The theorem follows from Lemma 2.13. ∎

Proof of Lemma 2.15 :

Since

$$E(log \mid \Sigma_{(11)} \mid\mid X^2, \Phi_{(11)}, \delta^*) = E(log \mid R^t \Sigma_{(11)} R \mid\mid X^2, \Phi_{(11)}, \delta^*)$$
$$= E(log(\mid \Sigma_{1|2} \mid\mid \Sigma_{22} \mid) \mid X^2, \Phi_{(11)}, \delta^*).$$

Note that

$$\Sigma_{1|2} \mid X^2, \Phi_{(11)}, \delta^* \sim W_l^{-1}(\Psi_{1|2}, \delta^* - s_u k)$$

and that

$$\Sigma_{22} \mid X^2, \Phi_{(11)}, \delta^* \sim W_{s_g k - l}^{-1}(\hat{\Psi}_{22}, \delta^* + n - s_u k - l - h).$$

Now apply Lemma 2.14 to finish the proof. ∎

Proof of Theorem 2.11 :

In the following proof, we assume F^{-1} fixed. By (2.16),

$$E(\hat{\beta}_v^{j_v} \mid B_{(1)}^o) = \mu_o^{j_v}, \quad v = 1, \ldots, s_g k - l.$$

So

$$E(\hat{\mu}_o^i \mid B_{(1)}^o) = \mu_o^i \quad i = 1, \ldots, k.$$

To prove the other half of the Theorem, by (2.17) and Theorem 3.2.8 of Muirhead 1982 (p93),

$$\frac{a_{j_v}^t S_2 a_{j_v}}{a_{j_v}^t \Sigma_{22} a_{j_v}} \mid \Sigma_{22} \sim \chi_{n-h}^2$$

implying

$$E(\frac{a_{j_v}^t S_2 a_{j_v}}{a_{j_v}^t \Sigma_{22} a_{j_v}} \mid X^2 = x^2, \Sigma_{22})^{-1} = \frac{1}{n - h - 2} \tag{2.26}$$

and

$$E[(\hat{\beta}_v^{j_v} - \hat{\mu}_o^{j_v})^t (\hat{\beta}_v^{j_v} - \hat{\mu}_o^{j_v}) \mid X^2 = x^2, B_{(1)}^0, \Sigma_{22}] = Var(\hat{B}_2^t a_{j_v} \mid X^2 = x^2, B_{(1)}^0, \Sigma_{22}).$$

By (2.16) and the fact that $a_{j_v}^t \Sigma_{22} a_{j_v}$ is a scalar,

$$\hat{B}_2^t \mid X^2 = x^2, B_{(1)}^0, \Sigma_{22} \sim N([B_{(1)}^o]^t R_2, ((ZZ^t)^{-1} + F^{-1}) a_{j_v}^t \Sigma_{22} a_{j_v}).$$

Hence,

$$Var(\hat{B}_2^t a_{j_v} \mid X^2 = x^2, B_{(1)}^0, \Sigma_{22}) = (a_{j_v}^t \Sigma_{22} a_{j_v})(F^{-1} + (ZZ^t)^{-1}). \tag{2.27}$$

When $X^2 = x^2, B_{(1)}, \Sigma_{22}$ are fixed, S_2 and $\hat{B}_2$ are independent. If we replace $B_{(1)}$ with $B_{(1)}^o$, they are still independent (Lemma 2.18). Therefore, given $X^2 = x^2$, $B_{(1)}^o$ and Σ_{22}, by Lemma 2.18,

$$\begin{aligned} E(\hat{F}^{-1}) &= \frac{n-h-2}{s_g k - l} \sum_{v=1}^{s_g k - l} E\{\frac{a_{j_v}^t S_2 a_{j_v}}{a_{j_v}^t \Sigma_{22} a_{j_v}} (a_{j_v}^t \Sigma_{22} a_{j_v})\}^{-1} E[(\hat{\beta}_v^{j_v} - \hat{\mu}_o^{j_v})^t (\hat{\beta}_v^{j_v} - \hat{\mu}_o^{j_v})] \\ &\quad - (ZZ^t)^{-1} \\ &= \frac{1}{s_g k - l} \sum_{v=1}^{s_g k - l} (F^{-1} + (ZZ^t)^{-1}) - (ZZ^t)^{-1} \\ &= F^{-1}. \blacksquare \end{aligned}$$

Proof of Theorem 2.12 :

By (2.7), (2.8) and (2.12)-(2.14), given the hyperparameters Φ, δ^*, B^o, F,

$$X \sim T(\Phi^{-1}, Z^t F^{-1} Z + I_{n \times n}, B^o Z, \delta^* + n).$$

By Theorem 2.2,

$$X - B^o Z = a^{-\frac{1}{2}} (a^{\frac{1}{2}} (X - B^o Z)) \overset{\text{approx.}}{\sim} a^{-\frac{1}{2}} \mathcal{N}(\Phi, Z^t F^{-1} Z + I),$$

where $a = \delta^* + n$. Then by Lemma 2.7,

$$vec(X^t) \overset{\text{approx.}}{\sim} a^{-\frac{1}{2}} \mathcal{N}(Z^t F^{-1} Z + I, \Phi) + vec((B^o Z)^t).$$

So,

$$(R^*)^t vec(X^t) = \begin{pmatrix} vec[(X^0)^t] \\ vec(M_1) \\ vec(M_2) \end{pmatrix} \overset{\text{approx.}}{\sim}$$

$$\begin{pmatrix} vec[(B_0^o Z)^t] \\ R_1^t vec[(B_{(1)}^o Z)^t] \\ R_2^t vec[(B_{(1)}^o Z)^t] \end{pmatrix} + (R^*)^t a^{-\frac{1}{2}} \mathcal{N}(Z^t F^{-1} Z + I, \Phi).$$

By the general normal theory, the theorem is proved. ∎

Proof of Lemma 2.19 :

By the same argument of the proof of Theorem 2.12, one has

$$vec(X_{(1)}^t) \overset{\text{approx.}}{\sim} a_1^{-\frac{1}{2}} \mathcal{N}(Z^t F^{-1} Z + I, \Phi_{(11)}) + vec((B_{(1)}^o Z)^t),$$

where $a_1 = \delta^* - s_u k + n$. So,

$$\begin{pmatrix} vec(M_1) \\ vec(M_2) \end{pmatrix} \sim a_1^{-\frac{1}{2}} R^t \mathcal{N}(Z^t F^{-1} Z + I, \Phi_{(11)}) + R^t vec[(B_{(1)}^0 Z)^t].$$

By the general normal theory, the lemma is proved. ∎

Proof of Theorem 2.13 :

For notational simplicity, the superscript j is suppressed in the following derivation. Let X_j^t, $j = 1, \ldots, n$, be partitioned into $((X_j^0)_{1\times s_u k})$, $(U_j^t)_{1\times d_j}$, $((G_j^t)_{1\times(s_g k - d_j)})$. $B^t = (B_0^t, B_1^t, B_2^t)$ and Σ are partitioned confirmablely. Σ is reparameterized as $\{\Sigma_{0|(1)}, \tau_{0(1)}, \Gamma^*\}$, where $\Gamma^* = \{\Sigma_{1|(2)}, \tau_{1(2)}, \Sigma_{(22)}\}$. $B_{(1)}^t$ represents $\{B_1^t, B_2^t\}$. $B_{(1)}$ is also reparameterized as

$$b_1 = B_1 - \tau_{1(2)} B_{(2)}, \quad b_{(2)} = B_{(2)}.$$

Note,

$$\begin{aligned} f \quad & (U_j^t \mid \{G_i^t\}_{i=j}^n, D_j) \\ \propto \quad & \int f(U_j^t, \{G_i^t\}_{i=j}^n \mid D_j, b_1, b_{(2)}, \Gamma^*) f(b_1, b_{(2)}, \Gamma^* \mid D_j)\, db_1\, db_{(2)}\, d\Gamma^* \\ \propto \quad & \int f(U_j^t, G_j^t \mid D_j, b_1, b_{(2)}, \Gamma^*) f(\{G_i^t\}_{i=j+1}^n \mid D_j, b_1, b_{(2)}, \Gamma^*) \\ & \cdot f(b_1, b_{(2)} \mid \Gamma^*, D_j) f(\Gamma^* \mid D_j)\, db_1\, db_{(2)}\, d\Gamma^* \end{aligned} \tag{2.28}$$

To find the distribution of $f(b_1, b_{(2)} \mid D_j, \Gamma^*)$, note by lemma (2.1)

$$B_{(1)} \mid D_j, \Gamma^* \overset{\text{d}}{=} B_{(1)}^* + N(0, \Sigma_{(11)}^*),$$

where $B^*_{(1)} = B^0_{(1)} + (B_{(1)} - B^o_{(1)})\hat{E}^t$, $\Sigma^*_{(11)} = \Sigma_{(11)} \otimes F^*$, $F^* = (I - \hat{E})F^{-1}$. By Lemma 2.7,

$$vec(B^t_{(1)}) \mid D_j, \Gamma^* \stackrel{\mathrm{d}}{=} vec((B^*_{(1)})^t) + N(0, \Sigma_{(11)} \otimes F^*).$$

By a general normal theory,

$$\begin{aligned}
&vec(B^t_1) \mid D_j, vec(B^t_{(2)}), \Gamma^* \\
&\stackrel{\mathrm{d}}{=} vec((B^*_1)^t) + \Sigma_{vec(B^t_1)vec(B^t_{(2)})}\Sigma^{-1}_{vec(B^t_{(2)})vec(B^t_{(2)})}(vec(B^t_{(2)}) - vec((B^*_{(2)})^t)) \\
&\quad + N(0, \Sigma_{vec(B^t_1)\mid vec(B^t_{(2)})}) \\
&\stackrel{\mathrm{d}}{=} vec((B^*_1)^t) + (\tau_{1(2)} \otimes I)(vec(B^t_{(2)}) - vec((B^*_{(2)})^t)) + N(0, \Sigma_{1\mid(2)} \otimes F^*) \\
&\stackrel{\mathrm{d}}{=} vec((B^*_1 + \tau_{1(2)}(B_{(2)} - B^*_{(2)})^t) + N(0, \Sigma_{1\mid(2)} \otimes F^*).
\end{aligned}$$

In the above derivation, it is assumed that $(F^*)^{-1}$ exists and the last step is true because of Lemma 2.5. Again by Lemma 2.7,

$$B_1 \mid B_{(2)}, D_j, \Gamma^* \stackrel{\mathrm{d}}{=} B^*_1 + \tau_{1(2)}(B_{(2)} - B^*_{(2)}) + N(0, \Sigma_{1\mid(2)} \otimes F^*),$$

it implies

$$\begin{aligned}
b_1 \mid b_{(2)}, D_j, \Gamma^* &= (B_1 - \tau_{1(2)}B_{(2)}) \mid B_{(2)}, D_j, \Gamma^* \\
&=^{\mathrm{d}} b^*_1 + N(0, \Sigma_{1\mid(2)} \otimes F^*),
\end{aligned}$$

where $b^*_1 = B^*_1 - \tau_{1(2)}B^*_{(2)}$. Since the distribution of $b_1 \mid b_{(2)}$ does not depend on $b_{(2)}$, b_1 and $b_{(2)}$ are independent. So

$$\begin{aligned}
f(b_1, b_{(2)} \mid D_j, \Gamma^*) &= f(b_1 \mid D_j, \Gamma^*) f(b_{(2)} \mid D_j, \Gamma^*) \\
&= f(b_1 \mid, \Sigma_{1\mid(2)}, \tau_{1(2)}) f(b_{(2)} \mid \Sigma_{(22)}, D_j). \qquad (2.29)
\end{aligned}$$

Further since

$$f(U^t_j, G^t_j \mid b_1, b_{(2)}, \Gamma^*) \propto f(U^t_j \mid G^t_j, b_1, b_{(2)}, \Gamma^*) f(G^t_j \mid b_1, b_{(2)}, \Gamma^*)$$

and

$$U_j^t \mid G_j^t, b_1, b_{(2)}, \Gamma^* \stackrel{\mathrm{d}}{=} [B_1 Z_j + \tau_{1(2)}(G_j - B_{(2)} Z_j)]^t + N(0, \Sigma_{1|(2)})$$
$$= [b_1 Z_j + \tau_{1(2)} G_j]^t + N(0, \Sigma_{1|(2)}).$$

So

$$f(U_j^t, G_j^t \mid b_1, b_{(2)}, \Gamma^*) \propto f(U_j^t \mid G_j^t, b_1, \tau_{1(2)}, \Sigma_{1|(2)}) f(G_j^t \mid b_{(2)}, \Sigma_{(22)}). \tag{2.30}$$

Also, by Model (2.12)-(2.14),

$$f(\{G_i^t\}_{i=j+1}^n \mid D_j, b_1, b_{(2)}, \Gamma^*) = f(\{G_i^t\}_{i=j+1}^n \mid \Sigma_{(22)}, b_{(2)}). \tag{2.31}$$

By (2.28)-(2.31) and Lemma 2.20

$$\begin{aligned}
& f(U_j^t \mid \{G_i^t\}_{i=j}^n, D_j) \\
& \propto \int f(U_j^t \mid G_j^t, b_1, \tau_{1(2)}, \Sigma_{1|(2)}) f(b_1 \mid D_j, \tau_{1(2)}, \Sigma_{1|(2)}) \\
& \quad \cdot f(\tau_{1(2)}, \Sigma_{1|(2)} \mid D_j)\, db_1\, d\tau_{1(2)}\, d\Sigma_{1|(2)} \\
& \quad \cdot \int f(\{G_i^t\}_{i=j}^n \mid b_{(2)}, \Sigma_{(22)}) f(b_{(2)} \mid \Sigma_{(22)}, D_j) f(\Sigma_{(22)} \mid D_j)\, db_{(2)}\, d\Sigma_{(22)} \\
& \propto \int f(U_j^t \mid G_j^t, b_1, \tau_{1(2)}, \Sigma_{1|(2)}) f(b_1 \mid D_j, \tau_{1(2)}, \Sigma_{1|(2)}) \\
& \quad \cdot f(\tau_{1(2)}, \Sigma_{1|(2)} \mid D_j)\, db_1\, d\tau_{1(2)}\, d\Sigma_{1|(2)}.
\end{aligned}$$

For summary, we have,

$$U_j \mid G_j, b_1, \Gamma^* \stackrel{\mathrm{d}}{=} b_1 Z_j + \tau_{1(2)} G_j + N(0, \Sigma_{1|(2)}); \tag{2.32}$$

$$b_1 \mid D_j, \tau_{1(2)}, \Sigma_{1|(2)} \stackrel{\mathrm{d}}{=} b_1^* + N(0, \Sigma_{1|(2)} \otimes F^*), \tag{2.33}$$

where

$$\begin{aligned}
b_1^* &= B_1^* - \tau_{1(2)} B_{(2)}^* \\
&= B_1^o + (\hat{B}_1 - B_1^o)\hat{E}_t - \tau_{1(2)}[B_{(2)}^o + (\hat{B}_{(2)} - B_{(2)}^o)\hat{E}_t].
\end{aligned}$$

Since, by Lemma 2.1,

$$\Sigma_{(11)} \mid D_j \sim W_{s_g k}^{-1}(\hat{\Phi}_{(11)}, \delta^* + (j-1) - s_u k).$$

Then by Lemma 2.20,

$$\tau_{1(2)} \mid D_j, \Sigma_{1|(2)} \sim N(\hat{\Phi}_{1(2)} \hat{\Phi}_{(22)}^{-1}, \Sigma_{1|(2)} \otimes \hat{\Phi}_{(22)}^{-1}); \tag{2.34}$$

$$\Sigma_{1|(2)} \mid D_j \sim W^{-1}(\hat{\Phi}_{1|(2)}, \delta^* - s_u k + (j-1)). \tag{2.35}$$

By (2.32)-(2.34),

$$U_j \mid G_j, \Sigma_{1|(2)} \stackrel{\mathrm{d}}{=} a_1 + N(0, \Sigma_{1|(2)} a_2),$$

where

$$a_1 = B_1^* Z_j - \hat{\Phi}_{1(2)} \hat{\Phi}_{(22)}^{-1} B_{(2)}^* Z_j + \hat{\Phi}_{1(2)} \hat{\Phi}_{(22)}^{-1} G_j$$

and

$$a_2 = 1 + Z_j^t F^* Z_j + (G_j - B_{(2)}^* Z_j)^t \hat{\Phi}_{(22)}^{-1} (G_j - B_{(2)}^* Z_j).$$

Combined with Equation (2.8) and (2.35),

$$U_j \mid \{G_i\}_{i=j}^n, D_j \stackrel{\mathrm{d}}{=} U_j \mid G_j \sim T(\hat{\Phi}_{1|(2)}^{-1}, a_2, a_1, \delta^* - s_u k + j). \blacksquare$$

Chapter 3

Application to Air Pollution Data

In this chapter, the theory of interpolation for data missing-by-design developed in Chapter 2 is applied to obtain interpolated data.

The daily maximum hourly levels of nitrogen dioxide (NO_2), ozone (O_3) , sulphur dioxide (SO_2) and the daily mean levels of nitrate (NO_3), sulfate (SO_4) were recorded from January 1 of 1983 to December 31 of 1988 in Ontario and its surrounding areas. These data come from several monitoring networks in the province, including the Environment Air Quality Monitoring Network (OME), Air Pollution in Ontario Study (APIOS), the Canadian Acid and Participation Monitoring Network (CAPMON) (see Burnett et al (1992) for more description of the data set). Also available are some climatic data: daily maximum, minimum and mean temperature; daily maximum, minimum and mean humidity, and, mean pressure and mean wind speed, measured at other locations. In total, there are 37 different monitoring locations (sites), but not all sites monitor all of the five air pollutants. In the application below, we assume that the variation caused by networks to the observations at 37 sites is negligible. Therefore we can pool the observed pollution levels without worrying about that variation.

In general, there are two kinds of air pollutants: (i) a primary pollutant, which is

directly emitted by identifiable sources; (ii) a secondary pollutant, which is produced by chemical reactions within the atmosphere between pollutants and other constituents. SO_2 is a primary pollutant while NO_2, NO_3, O_3 and SO_4 are secondary pollutants. SO_2 is produced by burning fuels containing sulphur. Its level depends on local emission sources, like burning fuel oil or smelting.

The secondary pollutants studied here are all produced by oxidation of primary pollutants. This oxidation is driven by ultra-violet radiation from sunlight and comprises chemical reactions that are temperature dependent. Since the chemical reactions proceed while the polluted air is being adverted by winds, secondary pollutants are generally more widespread than primary pollutants. We thus refer to secondary pollutants as regional. Because of temperature dependence of the governing chemical reaction, NO_2, NO_3 and O_3 are high in early afternoon and midsummer, low overnight and in winter. The oxidation of SO_2 to SO_4 is dominated by photochemical processes in dry, warm atmospheres.

Monthly pollution data are interpolated down to the level of a Public Health Unit (PHU), the daily pollution data down to the level of a census subdivision (CSD). Both interpolation problems originate from environmental health studies not discussed in this thesis. In both cases, the relevance of the assumptions in Chapter 2 is investigated and the interpolated residuals are checked.

In Section 3.1, our interpolation theory is applied to monthly data and in Section 3.2, to daily data.

3.1 Application to Monthly Pollution Data

A monthly pollution level is simply computed as the mean of the observed daily levels in that month. Hence, a time series of observed monthly mean levels consists of 72 values. Since some of these time series contain excessively many randomly missing values, to control the quality of data all time series are screened and those with more than one-third of its values missing are deleted. Thus, the number of gauged sites is reduced from 37 to 31. Here, we assume that the probability of a time series having more than one-third of its values missing is not related to those values. Therefore, such a strategy for dropping time series from our study does not cause bias.

The locations of the remaining 31 sites, except two outlying sites, are plotted in Figure 3.1. The whole of the Province of Ontario divides into thirty-seven PHUs or districts (Duddek et al 1994). A PHU resembles a Census Division, the difference being marginal disagreements in boundaries. Some PHUs, for example, are aggregates of two Census Divisions. Figure 3.2 displays the locations of some approximate centroids of these PHU's in Southern Ontario. Our ungauged sites consist of the 37 approximate centroids. Hence, the total number of gauged sites, s_g, is 31 and that of ungauged sites, s_u, is 37.

For the monthly interpolation, four air pollutants are included. They are NO_2, SO_4, O_3 and SO_2. Table 3.1 lists the observed pollutants at each of gauged sites, where a, b, d and e represent NO_2, SO_4, O_3 and SO_2 respectively. From the table, we can see that at all gauged sites, there are 64 observed time series and 60 missing time series. In the sequel, the term "time series" is replaced with "column".

Among the 64 observed columns, about two percent are missing. The missing data include the randomly missing data and those censored from below. Here, *censored below* means that the actual air pollution level is below the detection limit of a monitor.

Table 3.1: Pollutants Measured at Each Gauged Site, Where a, b, d And e Represent NO_2, SO_4, O_3 And SO_2 Respectively.

Sites	1	2	3	4	5	6	7	8	9	10	11	12
Pollutants	b	b	b	b	abde	d	be	ade	d	d	ade	de
Sites	13	14	15	16	17	18	19	20	21	22	23	24
Pollutants	ade	ade	ade	b	abde	b	ade	ade	ade	d	e	ade
Sites	25	26	27	28	29	30	31					
Pollutants	de	ade	ade	de	b	e	de					

In statistics, the procedure for filling in missing data is called *imputation.* There are many existing imputation methods (c.f. Sarndel 1992, Little and Rubin 1987). Examples of these methods are: overall mean imputation; class mean imputation; hot-deck and cold-deck imputation; random overall imputation; random imputation within classes; sequential hot-deck imputation; distance function matching; regression imputation; EM algorithm based imputation and multiple imputation.

The imputation methods mentioned above, produce a single imputed value for each missing value except for the multiple imputation approach. Many authors used these methods. Afifi et al. (1961) suggested filling in the missing observations for each variable by that variable's mean. Buck (1960) and Stein et al. (1991) discussed imputing the missing values by regression models. Komungoma (1992) adopted a modified regression strategy. Johnson et al. (1994) combined a stepwise linear regression and a time series model to fill in missing values. Miller (1994) employed a nearest-neighbor hot-deck imputation procedure to fill in missing data. While the multiple imputation is appealing, it has a disadvantage of requiring more work for data handling and computation of estimates. Moreover, the multiple imputation method was designed for survey sampling. It is not suitable for a spatial application.

We may apply Theorem 2.12 to impute our missing data. However, that method has not been well tested yet. Further more, Theorem 2.12 bases on the normal approximation to a matrix T. As showed by the "empirical coverage percentage" at the end of this section, in the current scenario the matrix T cannot be replaced by its normal approximation.

For our monthly data, with such a small proportion of missing data, the choice of a procedure for filling in missing values is not critical. We filled in a value missing from the i^{th} column in the j^{th} month with the mean of the observed values of the same column in the j^{th} month of other years. In case, all six measurements of the same month of 1983 to 1988 are missing, the grand mean of the observed values of the column is used. However, with our data set, no such a case obtained. The above ad hoc filled-in method allows us to preserve the periodic property of the columns shown in Figure 3.3, where ozone is seen to have a strong periodic pattern. Again, because only a small percentage of the data are missing, our ad hoc fill-in method will cause negligible bias.

Our theory of Bayesian multivariate interpolation with data missing-by-design is developed under two important assumptions. One assumption is that the detrended residuals should spatially follow a multivariate normal distribution. The other is that each residual time series is a white noise process; in other words, the residuals are temporally independent. Checking multivariate normality is not easy since we lose hundreds of degrees of freedom to parameter estimation. In this chapter, only univariate marginal normality is checked. Temporal independence is checked with autocorrelation and partial autocorrelation plots.

The normal quantile-quantile plots of the detrended residuals of the original observed data do not show straight lines, suggesting the observed data be transformed. With a logarithmic transformation of the observed data, the residuals appear to be marginally

normal. A typical example of a normal q-q plot of the original and log-transformed data is shown in Figure 3.4. Autocorrelation and partial autocorrelation plots of the detrended residuals of log-transformed data are shown in Figure 3.5. The correlation plots do not show evidence of temporal correlation. By repeating the above initial data analysis for observations at all gauged sites, we are led to conclude that the log-transformed data do meet the assumptions of our interpolation theory. In the sequel, unless specify otherwise, we always refer observations on their log-scale.

For determining the linear and seasonal trends of our Gaussian model, time series plots are made. Based on the plots, Z_t is taken to be $1, t, cos(\frac{2\pi t}{12}), sin(\frac{2\pi t}{12})$,where $t = 1, \ldots 72$. Here $t = 1$ represents the January of 1983, $t = 2$ represents February of 1983 and so on, until $t = 72$ which represents December of 1988. The coefficients of the linear and seasonal trends are estimated with ordinary least squares. Figure 3.3 displays the time series plots and their least squares fitted curves for the four observed pollutants at Site 5. The fitted curve of $log(O_3)$ is far better than that of the other three because of its periodicity. The strong yearly pattern of ozone is partially explained by the fact that the creation of ozone is highly related to solar radiation.

For the environmental health study referred to earlier, the one which required interpolation of monthly means, the air pollution level in summer is of special interest. Therefore the pollution data of winter and summer are interpolated separately. In the following, only the interpolation for summer data is described. The procedure for winter data is similar. Here, the *summer* of a year is defined to be from May 1 to August 31 and "winter" the remainder of the year. Thus there are 24 values (24 months) in each summer column. The purpose of the analysis is to interpolate monthly NO_2, SO_4, O_3 and SO_2 levels in the summers of 1983 $\sim$ 1988 at 37 ungauged sites.

Table 3.2: The Estimated Between-pollutants-hypercovariance Matrix of the Log-transformed, Summer Monthly Data.

	NO2	SO4	O3	SO2
NO2	1.00000000	-0.2854130	0.03476434	0.1364513
SO4	-0.28541295	1.0000000	0.79402064	-0.3379370
O3	0.03476434	0.7940206	1.00000000	-0.1457342
SO2	0.13645127	-0.3379370	-0.14573424	1.0000000

The interpolation procedure follows the following order: first, the unbiased estimators of F^{-1} and $B^o_{(1)}$ are computed; second, the EM algorithm for the estimation of δ^* and Λ_g, Ω is invoked; third, the SG method is applied to extend Λ_g to Λ; then, with the exchangeable assumption on B^o, $B^o_{(1)}$ is extended to B^o; finally, with all the hyperparameters estimated, the interpolated values are computed by the Bayesian interpolator.

We use S macros to carry out the interpolation and call a C program to perform the EM algorithm. An example of the S macros and C program are attached in the Appendices. By running S macros and the C program, we get the estimated prior degrees of freedom to be 610; the estimated between-pollutants-hypercorrelation matrix, which is listed in Table 3.2 and the estimated hyper-covariances of NO_2, SO_4, O_3 and SO_2 are 0.6582724, 1.6263357, 0.2166192, 1.8503741, respectively. The biggest positive correlation among the four pollutants occurs between O_3 and SO_4. Since both O_3 and SO_4 are regional air pollutants and both are related to sunlight, a higher correlation between O_3 and SO_4 is expected. Meanwhile SO_2 is a primary pollutant, it has a lower correlation with the other pollutants.

The result of the SG step is summarized in Figures 3.6 $\sim$ 3.8. The right hand plot of Figure 3.6 is a twisted 30 by 30 checkerboard in the D-plane. The original 30 by 30 checkerboard is in the G-plane. The coordinates of its lower left corner are the

minimum latitude and longitude of gauged sites. The coordinates of its upper right corner are the maximum latitude and longitude of gauged sites. The left hand plot of Figure 3.6 shows an exponential fit between dispersions and the D-plane distances (refer to Chapter 1 for a brief summary of the SG method). The smoothness of the twisted checkerboard is controlled by parameter λ. A smoother checkerboard in D-plane is achieved by sacrificing the fit between the dispersions and D-plane distances. Figure 3.7 shows a smoother checkerboard in the D-plane but a rougher fit between the dispersions and D-plane distances when the smoothing parameter value is increased from 0 to 2500. A linear pattern in Figure 3.8 shows that the estimated covariance and the observed covariance are conformable.

By applying Corollary 2.6 and using the above estimated hyperparameter values, after SG step it is straightforward to compute the interpolated, summer monthly air pollution levels at ungauged sites over six years. As a way of checking the interpolated values, the overall average ozone levels the in summers of 1983 to 1988 at gauged sites are plotted in Figure 3.9. Those of interpolated ozone levels at ungauged sites are plotted in Figure 3.10. The two plots affirm our interpolation procedure. When a high mean O_3 level is observed at a gauged site, our interpolator gives a high O_3 values at nearby ungauged sites. Corresponding results obtain for lower observed O_3 levels.

Another way of checking the interpolation procedure is to look at the correlation between the observed and estimated data by cross-validation (CV hereafter). CV is a procedure which deletes observed datum one at a time and estimates these datum from the remaining data as if that datum were never observed. It is a popular diagnostic tool. In our CV study, we deleted one gauged site at a time and interpolated pollutant levels at that same site using observed levels at other sites. To avoid spuriously high computed correlations between the estimated and observed levels of pollutants, we first

removed the trends from both the estimated and observed columns, and then calculated the correlations among the residuals.

The correlations between the detrended, estimated and observed levels of each pollutant aggregating across sites and over time are:

	Summer	Winter
NO2	0.243	0.242
SO4	0.494	0.438
O3	0.534	0.429
SO2	0.238	0.200.

The correlations between the estimated and observed levels at each gauged site and for each observed pollutant are given in Table 3.3.

In Table 3.3, the correlations of SO_4 and O_3 in both summer and winter are generally higher than those of SO_2 with other pollutants. In other words, the predictions of SO_4 and O_3 are more accurate than those of SO_2. Figure 3.11 displays the plot residuals of log-transformed, monthly observed and estimated pollutant levels in both summer and winter. Figure 3.12 shows the scatteplots of log-transformed observed pollutants against estimated pollutant levels for each pollutant in summer and winter, respectively. In the plots a linear pattern means accurate interpolation. The plots confirm conclusions suggested by the tables; these results are consistent with the fact that O_3 and SO_4 are regional pollutants. It is easy to predict them with the observed data from other sites. In contrast SO_2 is a local pollutant and so more difficult to predict.

Can a simpler to use, normal distribution be substituted for the multivariate T predictive distribution? That might naively seem possible since the univariate normal approximates its longer tailed relative very well. However, our results suggest this

Table 3.3: Correlations Between Residuals of Log-Transformed, Observed and Estimated Pollution Levels at Gauged Sites.

Sites	summer				winter			
	NO2	SO4	O3	SO2	NO2	SO4	O3	SO2
1		0.96				0.81		
2		0.96				0.85		
3		0.87				0.74		
4		0.93				0.81		
5	0.39	0.82	0.75	0.57	0.09	0.67	0.70	0.40
6			0.92				0.74	
7		0.90		0.67		0.71		0.14
8	0.42		0.81	0.76	0.56		0.74	0.58
9			0.87				0.78	
10			0.85				0.41	
11	0.57		0.97	0.66	0.61		0.75	0.32
12			0.87	0.65			0.59	0.30
13	0.44		0.75	0.54	0.11		0.57	0.04
14	0.11		0.88	0.53	0.13		0.69	0.33
15	0.66		0.93	0.69	0.52		0.79	0.24
16		0.90				0.87		
17	0.66	0.91	0.80	0.78	0.25	0.68	0.55	0.18
18		0.85				0.75		
19	0.63		0.77	0.80	0.34		0.59	0.38
20	0.36		0.72	0.61	0.57		0.72	0.40
21	0.67		0.80	0.63	0.47		0.63	0.40
22			0.86				0.49	
23				0.78				0.25
24	0.48		0.74	0.68	0.56		0.65	0.56
25			0.80	0.52			0.44	0.26
26	0.71		0.82	0.81	0.24		0.47	0.23
27	0.55		0.77	0.66	0.44		0.59	0.50
28			0.87	0.83			0.40	0.49
29		0.97				0.49		
30				0.78				0.10
31			0.82	0.62			0.23	0.52

substitution cannot be recommended without additional study. Our initial impression comes from an evaluation we did of the empirical coverage percentage of three-standard-deviation confidence intervals (CI). If the predictive distribution were normal, all the three-standard-deviation CIs would include the true values about 100 percent of the time. As the percentages by pollutants presented below indicate, this high coverage probability is not achieved here. The heavier tailed predictive matrix T distribution seems to be required.

By pollutants the percentages are,

	Summer	Winter
NO2	100%	94.2%
SO4	100%	99.2%
O3	98.6%	98.8%
SO2	94.5%	100%

The unbiasedness of residuals is also checked. In the top plot of Figure 3.13 are the boxplots of the prediction errors for four pollutants, these being defined as the difference between the predicted and observed values. Except for SO_2, the mean prediction errors for the other three pollutants are almost zero. In other words, the predictor is unbiased. In the same figure, the other two plots indicate that observed values and the boxplots of the predicted values have similar patterns except that the predicted values have bigger variances.

3.2 Application to Daily Pollution Data

For one of the environmental health studies mentioned above we needed to interpolate daily air pollution levels down to the centroids of CSD's. In Southern Ontario, 733 such

centroids are chosen and all five pollutants NO_2, NO_3, O_3, SO_2 and SO_4, are included. The general interpolation procedure and many intermediate results are similar to those for the monthly data. Thus, in the following, only the results which differ from those in Section 1 are discussed. Again, only the interpolation of summer air pollution levels at the CSD centroids is discussed.

The observed data come from daily measurements at 37 sites. By removing the time series (columns) where there are excessively many missing data, the number of sites is reduced from 37 to 27 and the number of observed columns to 55. Then the number of missing columns is 80 ($= 27 * 5 - 55$). Figure 3.14 displays the locations of the 27 gauged sites, except for two outlying sites.

As in Section 3.1, an ad hoc method is used to fill-in the missing data. However, in this section, a different ad hoc method is used. Since in the daily observed columns, eleven percent are missing, a more delicate approach is needed. The new ad hoc method replys on multivariate normal theory. With the new method, missing data are filled in by

$$X_{ijt} = \hat{B}_{ij} Z_t + \hat{r}_{ij},$$

where X_{ijt} is the missing value for the i^{th} pollutant ($i = 1, \ldots, 5$) at j^{th} gauged site ($j = 1, \ldots, 27$) at time t; $\hat{B}_{ij}$ is estimated by the ordinary least squares method using observed values for the i^{th} pollutant at the j^{th} gauged site and $\hat{r}_{ij}$ is the estimated residual. The estimated residual is computed by using well known normal theory. That is,

$$E(X \mid Y = y) = E(X) + \Sigma_{XY}\Sigma_{YY}^{-1}(y - E(Y)),$$

where X and Y are jointly normal, Σ_{UV} is the covariance matrix between random variables U and V and $E(U)$ is the mean of U. By letting Y represent the set of all observed data at time t, replacing X with X_{ijt}, and taking both $E(X)$ and $E(Y)$ to be

zero, we apply the above formula to calculate $\hat{r}_{ij}$. However, the formula is not directly applicable, since the joint covariance matrix of X_t is unknown. A method of moments is used to estimate the covariance matrix in a pairwise manner. In other words, the covariance of X_{lmt} and X_{hkt} is estimated by $\frac{1}{f}\sum_t(X_{lmt} - \bar{X}_{lm})(X_{hkt} - \bar{X}_{hk})$, where f is the total number of observed pairs of X_{lmt} and X_{hkt}. The estimated covariance matrix has to be checked for positive definiteness. If the estimated matrix is not positive definite, it cannot be used to obtain $\hat{r}_{ij}$. If so, other imputation methods mentioned in the previous section would need to be investigated. However, in this application, the estimated matrix is indeed positive definite. Compared with the method used in Section 1, the advantage of new method is that it brings less autocorrelation into each time series when the missing values are fill-in. We assume such a filled-in procedure will not cause serious bias, because of the small percentage of missing data.

Figure 3.15 shows that a logarithmic transformation of daily data is also necessary. Since the time series plots of the daily summer data do not have obviously periodic patterns, the linear trend of the Model (2.12) is instead taken to be the grand mean effect; time t effect; weekday effect (from Monday to Thursday); monthly effect (May, June, July); yearly effect (from 1983 to 1987) and the mean daily temperature. The mean daily temperature is the mean of observed daily mean temperatures in Southern Ontario. Such an arrangement is due to the fact that Model (2.12) only allows covariates with spatially equal measurements. Other variables, for instances, the daily mean humidity, mean pressure etc. are investigated too. They are not included in the final model because their ordinary-least-squares fit coefficients are not significantly different from zero.

We checked the autocorrelation and partial autocorrelation plots of daily, detrended summer observations. Figure 3.16 shows an example of such a plot. We find that

Table 3.4: The Estimated Between-pollutants-hypercovariance Matrix of the Log-transformed Daily Summer Pollution Levels in Southern Ontario.

	NO2	SO4	NO3	O3	SO2
NO2	1.0000000	0.1224039	0.24357919	0.11314233	0.20852698
SO4	0.1224039	1.0000000	0.41186592	0.26680085	0.12408473
NO3	0.2435792	0.4118659	1.00000000	0.24745489	0.05080721
O3	0.1131423	0.2668008	0.24745489	1.00000000	0.09569043
SO2	0.2085270	0.1240847	0.05080721	0.09569043	1.00000000

they are lag-1 correlated. For removing the lag-1 correlation of the residuals, an AR(1) transformation of the observed values is applied, that is, $Y_{ijt} = X_{ijt} - \hat{\phi}_{ij} X_{ij(t-1)}$ for $t = 2, \ldots, 738$, where $\hat{\phi}_{ij}$ is estimated based on the observed values of i^{th} pollutant at j^{th} gauged site. Such a transformation removes the lag-1 correlation among the residuals. For comparison, the autocorrelation and partial autocorrelation plots of the AR(1)-transformed residuals of O_3 at Gauged Site 6 are shown in Figure 3.17.

When the normal and independence assumptions are satisfied, the EM algorithm can be applied to the residuals to estimate the hyperparameters. The between-pollutants hypercorrelation matrix is given in Table 3.4 and the hypervariances of the $NO2$, $SO4$, $NO3$, $O3$, $SO2$ are 0.8093218, 1.8325046, 1.3923836, 0.4061250, 1.9200310, respectively. The estimated number of degrees of freedom is 4365 for nearest integer. A big number reflects a lot of prior information about the covariance matrix Σ. That large number stems from the fact that there are only 27 gauged sites with 55 observed columns while 733 ungauged sites with 3665 columns missing-by-design need to be interpolated.
The SG and interpolation steps are similar to those of the monthly data application above. One remaining problem needs to be solved. Since the interpolation is based on the AR(1)-transformed data, the interpolated values are in the form of AR(1)-transformed residuals. To obtain the true interpolated residuals, the following fact

is must be used.

Lemma 3.1 *If* $U_k - \phi U_{k-1} = V_k$ *and* $\phi < 1$, *then* $U_k \simeq \sum_{k=2}^{n} \phi^{n-k} V_k$ *provided* n *is big enough.*

Proof: Since

$$U_2 - \phi U_1 = V_2$$

$$\vdots$$

$$U_{n-1} - \phi U_{n-2} = V_{n-1}$$

$$U_n - \phi U_{n-1} = V_n,$$

multiply both sides of the first equation by ϕ^{n-2}, the next equation by ϕ^{n-3}, etc and then add all equations. It becomes,

$$U_k - \phi^{n-1} U_1 = \sum_{k=2}^{n} \phi^{n-k} V_k.$$

When n is big enough, ϕ^{n-1} is almost zero. ∎

To apply the above fact, the ϕ's at ungauged sites are needed. These values are not available. Let us assume the ϕ's are the same for the same pollutant at all sites. Then in all, only 5 different coefficients are needed. The five coefficients can be estimated by ordinary linear regression subject to the above assumption. By checking the observed data, we know that the assumption is valid. Now with the new assumption, the interpolation procedure is repeated.

Table 3.5: Correlations Between the Residual of Log-transformed, Summer Daily Observed and Estimated pollutants at Gauged Sites.

Sites	Summer					AR(1) Summer				
	NO2	SO4	NO3	O3	SO2	NO2	SO4	NO3	O3	SO2
1		0.68	0.56				0.63	0.52		
2		0.79	0.65				0.76	0.63		
3		0.56	0.55				0.48	0.50		
4		0.72	0.64				0.70	0.63		
5	0.44	0.62	0.45	0.67	0.12	0.44	0.60	0.48	0.65	0.12
6				0.69					0.69	
7		0.65	0.49		0.10		0.64	0.50		0.10
8	0.49			0.87		0.51			0.88	
9				0.90					0.90	
10	0.55			0.82		0.55			0.83	
11				0.82					0.83	
12	0.18			0.78		0.17			0.78	
13				0.72					0.73	
14	0.41			0.75		0.38			0.73	
15		0.74	0.68				0.71	0.67		
16	0.69	0.50	0.58	0.86		0.68	0.50	0.59	0.85	
17		0.70	0.53				0.68	0.52		
18	0.63			0.78		0.63			0.79	
19	0.63			0.84		0.62			0.86	
20	0.71			0.88	0.02	0.69			0.88	0.03
21				0.82					0.82	
22	0.54			0.82		0.56			0.83	
23				0.79					0.78	
24	0.45			0.68		0.44			0.64	
25	0.34			0.65	0.04	0.32			0.61	0.05
26		0.60	0.58				0.52	0.54		
27					0.15					0.14

As before, a CV study of the residuals for both the observed and predicted values is carried out. The pollutant-wise correlations between the residuals of observed and predicted summer daily pollution levels at gauged sites are:

	Summer	AR(1) Summer
NO2	0.49128973	0.48457402
SO4	0.66564685	0.63188994
NO3	0.58376676	0.56927091
O3	0.78495848	0.78477716
SO2	0.08807141	0.08978648.

In the above table, the predicted values used for computing the correlations in Column one are obtained using the original observed values and the predicted values, and for those in Column two obtained using the AR(1)-transformed data. From the same table, it can be seen that the correlations of the same pollutant in both columns are close. We interpret this result positively, as saying that our interpolation procedure is robust to the assumption of temporal independence. In other words, in terms of the above correlations, by taking an AR(1) transformation, the prediction of observed values has not been improved. From the same table, the correlation for SO_2 is much lower than those of the other pollutants. This is explained as follows. First, from Table 3.4, we see that the hypercorrelations of SO_2 with other pollutants are very low. This fact indicates that by including other pollutants, the predictions on SO_2 levels are not improved much. Second, from Table 3.5 it is seen that SO_2 was only observed at only 5 gauged sites. So in the CV analysis, the prediction of SO_2 levels at one of its gauged sites are based only on the observations at the other four sites gauged for SO_2. For monthly data, the problems associated with SO_2 do not exist. Therefore, the computed correlations between the predicted and observed SO_2 values are higher.

The correlations of the observed and estimated residuals at each gauged site for each pollutant are given in Table 3.5. From Table 3.5, it can be seen that it is easier to predict O_3 and SO_4, because they are regional pollutants and more difficult to SO_2, because it is a local pollutant.

Chapter 4

Application to Environmental Monitoring Network Redesign

Another application of the "Bayesian interpolation with missing-by-design data" theory developed in Chapter 2 is to the problem of redesigning an environmental monitoring network. The term "redesigning" means adding or deleting sites from a current existing monitoring network. Guttorp, Le, Sampson and Zidek (1992), Caselton, Kan and Zidek (1992), Wu and Zidek (1992) have discussed the above redesign problem. These authors derived their optimally redesigning strategy based on following reasoning. Maintaining and collecting data from an environmental monitoring network is quite expensive, therefore the network is set up and maintained by a nation wide institute. Thus the collected data may be used by different users for different purposes. This fact implies that an optimal redesign of an environmental monitoring network should be based on certain common and fundamental purposes of the users of monitoring networks. They choose "reducing uncertainty about some aspect of the world, regardless of how that need may be expressed in any particular situation" (Guttorp, Le, Sampson and Zidek 1992) as the redesign goal. They developed a general network redesign theory by combing the entropy optimal criteria and the Bayesian paradigm.

In Section 1, the general theory of network redesign in univariate case, which is discussed in Guttorp, Le, Sampson and Zidek (1992), is summarized. In Section 2, it is demonstrated that how our theory can be applied to the redesign problem.

4.1 Theory of Network Redesign with Entropy

In this section, the theory of network redesign with the entropy is summarized in the univariate case, which means that there is one pollutant at each site. Although the theory deals only with an augmentation of a network, the reduction of a network is handled in a similar fashion and the conclusion is similar.

Suppose that currently there are g gauged sites in a environmental monitoring network and it is planned to add u_1 more sites to the network in future. These u_1 sites are chosen among u candidate sites. Call all the u candidate sites "ungauged sites". Let X_f represent a future realization at gauged and ungauged sites. Decompose $X_f^t into((X_f^g)^t, (X_f^u)^t)$, where X_f^g is a realization at gauged sites and X_f^u at ungauged. By properly rearranging the coordinates, $(X_f^u)^t$ is further decomposed into $((X_f^{add})^t, (X_f^{(rem)})^t)$. X_f^{add} is the response vector at ungauged sites that will be added into the network and $X_f^{(rem)}$ is the vector at sites that are not chosen. We assume the same Gaussian linear model of (2.12) $\sim$ (2.14).

As to a measurement of uncertainty, one natural choice is the *entropy*, which is defined as

$$H(X) = E\left[-log\frac{f(X)}{h(X)}\right]$$

where $h(X)$ is a reference density, $h(x)$ makes $H(X)$ be invariant to scale transformation of X. From the definition, it is easy to see that

$$H(X,Y) = H(Y) + H(X \mid Y), \quad provided \quad h(X,Y) = h_1(X)h_2(Y).$$

Let D be the set of all observed data at gauged sites in the past, $\theta = (\Sigma, B)$ be the set of parameters in our Guassian model. Then, given D, the total uncertainty of a future realization X_f and unknown parameter, θ, is $H(X_f, \theta \mid D)$. Since,

$$H(X_f, \theta \mid D) = H(U \mid G, \theta, D) + H(\theta \mid G, D) + H(G \mid D), \tag{4.1}$$

where $G = ((X_f^{add})^t, (X_f^{(g)})^t)$, $U = X_f^{(rem)}$ and

$$H(U \mid G, \theta, D) = E[-log(f(U \mid G, \theta, D)/h_{11}(U)) \mid D],$$

$$H(\theta \mid G, D) = E[-log(f(\theta \mid G, D)/h_2(\theta)) \mid D],$$

$$H(G \mid D) = E[-log(f(G \mid D)/h_{12}(G)) \mid D].$$

In the above, it is assumed that $h(X, \theta) = h_1(X)h_2(\theta)$ and $h_1(X) = h_{11}(U)h_{12}(G)$.

Note that adding or deleting any site to or from the current network will not change the total uncertainty $H(X_f, \theta \mid D)$. When in future, the response vector at gauged sites and added sites is observed and if the measurement errors are negligible, the uncertainty represented by $H(G \mid D)$ becomes known. By Equation (4.1), one can see that a fixed total present uncertainty is decomposed into two parts, one part will become absolutely known by taking observation in future and the other part is still unknown. Therefore minimizing the future uncertainty by taking additional sites is equivalent to minimizing the unknown future uncertainty, is in turn equivalent to maximizing the uncertainty of what will be known in future. So the problem is equivalent to adding sites to maximize $H(G \mid D)$.

For finding expression, $H(G \mid D)$, the entropy of a multivariate t-distribution needs to be computed, since $f(G \mid D)$ consists of two multivariate t-distributions.

Now assume, for a random vector Y,

$$Y \mid \Sigma \sim N_g(0, \Sigma),$$

$$\Sigma \mid \Phi, \delta^* \sim W_g^{-1}(\Phi, \delta^*),$$

by (2.8),

$$Y \mid \Phi, \delta^* \sim t(0, (\delta^* - g + 1)\Phi, \delta^* - g + 1).$$

Note that

$$H(Y, \Sigma \mid \Phi, \delta^*) = H(Y \mid \Sigma, \Phi, \delta^*) + H(\Sigma \mid \Phi, \delta^*) = H(\Sigma \mid Y, \Phi, \delta^*) + H(Y \mid \Phi, \delta^*).$$

To find the multivariate t-distribution's entropy, $H(Y \mid \Phi, \delta^*)$, we only need compute $H(Y \mid \Phi, \delta^*, \Sigma)$, $H(\Sigma \mid \Phi, \delta^*)$ and $H(\Sigma \mid Y, \Phi, \delta^*)$ respectively. Since $Y \mid \Sigma, \Phi, \delta^*$ is multivariate normal , $\Sigma \mid \Phi, \delta^*$ and $\Sigma \mid Y, \Phi, \delta^*$ are inverse Wishart. By using the result presented in Caselton, Kan and Zidek (1992), the entropies of the multivariate normal and the inverse Wishart. Thus, after a straightforward calculation, the entropy of a multivariate t-distribution is,

$$H(Y \mid \Phi, \delta^*) = \frac{1}{2} log \mid \Phi \mid + c(g, \delta^*), \tag{4.2}$$

where $c(g, \delta^*)$ is a function of g and δ^* only.

Since by Theorem 2.1, the distributions of $X_f^g \mid D$ and $X_f^{add} \mid X_f^g, D$ are multivariate t-distributions and

$$H(G \mid D) = H(X_f^g \mid D) + H(X_f^{add} \mid X_f^g, D).$$

By applying (4.2), it is easy to see that

$$H(G \mid D) = \frac{1}{2} log \mid \Phi_{add|g} \mid + c(\hat{\Phi}_{22}, \Phi_{22}, c, d, l, g).$$

Where $\Phi_{add|g}$ is the residual covariance matrix of X^{add} conditional on X^g and $\hat{\Phi}_{22}$ is defined in Lemma 2.1. Because only the term $log \mid \Phi_{add|g} \mid$ is related to a choice of the newly added sites, maximizing $H(G \mid D)$ is the same as maximizing $\mid \Phi_{add|g} \mid$.

4.2 Network Redesign with Data Missing-by-design

The result presented in the previous section can be easily extended to the multivariate case by assuming k pollutants at each site. The value of $\mid \Phi_{add|g} \mid$ comes from Φ that is estimated with the method described in BLZ. Similarly the result can be extended to the case when there is missing-by-design data. To justify this extension, one needs to notice that, given hyperparameters,

$$\begin{pmatrix} X_t^0 \\ X_t^2 \end{pmatrix} \mid B = R^* X_t \sim N(R^* B, (R^*)^t \Sigma R^*)$$

$$(R^*)^t \Sigma R^* \sim W^{-1}((R^*)^t \Phi R^*, \delta^* - l)$$

$$R^* B \mid \Sigma \sim N(R^* B^o, (R^*)^t \Sigma R^* \otimes F^{-1}),$$

where

$$R^* = \begin{pmatrix} I_{s_u k \times s_u k} & 0 \\ 0 & R_2 \end{pmatrix}$$

and R_2 is defined the same as that in Section 2.4.1 of Chapter 2. Then applying the above model and a similar argument as in the previous section, an optimal criteria for redesign of a current network is maximizing $log \mid \Psi_{add|2} \mid$, where $\Psi_{add|2}$ is the conditional hypercovariance matrix of the future realization at added pollutants and sites given the observed pollutants at gauged sites. The matrix $\Psi_{add|2}$ can be obtained from matrix Φ that is estimated in Chapter 2. In an application, the added sites need not monitor all air pollutants. This relaxation may be useful when the optimal network redesign is required for multiple networks that were set to monitor different pollutants. A hypothetical example is that there are three networks, labeled 1, 2, 3. Network 1 measures ozone only, network 2 nitrate only and network 3 both pollutants. The above discussed optimal redesign enables us, say, to optimally add one site to network 1 and another site to network 2, in terms of maximally reducing uncertainty of a future realization in these three networks.

As an example of implementing the above discussion, the monthly air pollution data in Chapter 3 is used. Thus, there are 31 gauged sites and their latitudes range from 42.246 to 49.800 and longitudes range from 74.736 to 94.400. Suppose in future two sites will be added to these monitoring networks in Southern Ontario. One added site will monitor ozone only and the other nitrate only. Further suppose the possible site locations for the two would-be added sites are constrained to the grids of 10 by 10 checkerboard with the latitudes and longitudes of the four corners being $(42.246, 74.736)$, $(42.246, 94.400)$, $(49.800, 74.736)$ and $(49.800, 94.400)$. By taking the estimated Φ from Section 3.1 and applying the above optimal redesign criteria, the following answer is reached.

We should add the measuring-ozone only site at latitude 43.92467 and longitude 85.66044 and add the site measuring-nitrate-only site at latitude 44.76400 and longitude 87.84533. However, further simulation study on the sensitivity of Φ to the locations of added sites need be done.

Chapter 5

Cross-validation Study

In this chapter, we describe three CV studies designed to judge performance of the newly developed theory of interpolation with data missing-by-design. The first study is to justify the necessity of developing a new theory, in situation where the LZ method could be applied to solve the same problem. In the second study, an artificial example is made for studying the trend of adjusted mean squared predicted error(AMSPE), where, starting from a complete data set, columns are deleted one-by-one and the AMSPE computed. The third is a comparison of our theory against Hass's CoKriging method, since both methods can handle data missing-by-design.

5.1 Simultaneous Interpolation Versus Univariate Interpolation

By interpolating one pollutant at a time, one can apply LZ theory to the Southern Ontario pollution interpolation problem. Why then is a new theory needed when an old theory is available? The answer lies in the difference of two methods. With the LZ method, only partial data is used for the interpolation, while the new method includes all available data in the procedure. Take the monthly pollution data in Southern Ontario, for example. When ozone levels are interpolated at ungauged sites by the LZ theory, only

the observed O_3 levels at gauged sites are included in the analysis. The new method uses all the observed values of NO_2, SO_4, O_3 and SO_2. For distinguishing between these two methods, that of LZ will be called *univariate interpolation* and that of the new method, *simultaneous interpolation.*

One way of showing the superiority of the simultaneous interpolation over the univariate interpolation is to prove that its interpolator leads to a smaller mean square error. That fact is shown below.

Theorem 5.1 *Let $\underline{X}_0$, $\underline{Y}_0$ be any two random vectors and X a random variable. Then*

$$E(X - E(X \mid \underline{X}_0, \underline{Y}_0))^2 \leq E(X - E(X \mid \underline{X}_0))^2. \tag{5.1}$$

Proof: Observe that

$$\begin{aligned} E(X - E(X \mid \underline{X}_0))^2 &= E(X^2 - 2XE(X \mid \underline{X}_0) + E^2(X \mid \underline{X}_0)) \\ &= E(X^2) - 2E(XE(X \mid \underline{X}_0)) + E(E^2(X \mid \underline{X}_0)) \\ &= E(X^2) - 2E(E(XE(X \mid \underline{X}_0) \mid \underline{X}_0)) + E(E^2(X \mid \underline{X}_0)) \\ &= E(X^2) - E(E^2(X \mid \underline{X}_0)). \end{aligned}$$

Similarly,

$$E(X - E(X \mid \underline{X}_0, \underline{Y}_0))^2 = E(X^2) - E(E^2(X \mid \underline{X}_0, \underline{Y}_0)).$$

Further, by Jensen's inequality, for any random variable Z,

$$E(Z^2) \geq (E(Z))^2.$$

Apply it,

$$\begin{aligned} E(E^2(X \mid \underline{X}_0, \underline{Y}_0)) &= E(E(E^2(X \mid \underline{X}_0, \underline{Y}_0) \mid \underline{X}_0)) \\ &\geq E([E(E(X \mid \underline{X}_0, \underline{Y}_0) \mid \underline{X}_0)]^2) \\ &= E(E^2(X \mid \underline{X}_0)). \end{aligned}$$

So equivalently,

$$E(X - E(X \mid \underline{X}_0, \underline{Y}_0))^2 \leq E(X - E(X \mid \underline{X}_0))^2 \blacksquare$$

Returning to the ozone example, we take $\underline{X}_0$ to be the observed levels of O_3 at gauged sites, $\underline{Y}_0$ the observed levels of the other pollutants and X, the unobserved pollution level at an ungauged site. Then the univariate Bayesian interpolator is $E(X \mid \underline{X}_0)$ and the simultaneous interpolator $E(X \mid \underline{X}_0, \underline{Y}_0)$. When the model is correctly specified and the hyperparameters are known, Theorem 5.1 implies that the simultaneous interpolator does no worse than the univariate interpolator.

The following CV study supports the above claim empirically. Again, monthly air pollution data in Southern Ontario is used. At each gauged site, the observed pollutants are deleted as if they were not observed. Then both univariate and simultaneous Bayesian interpolators are applied to obtain the predicted values of the "deleted" values based on the data at the other gauged sites. When the predicted values by both methods are computed for all 31 gauged sites, the mean squared predicted error (MSPE) is calculated for the univariate interpolator and the simultaneous interpolator respectively. The results for the monthly summer data and monthly winter data are listed below.

	Simultaneous		Univariate	
	summer	winter	summer	winter
NO2	0.18543849	0.14342632	0.2829322	0.1292677
SO4	0.13848447	0.21311221	1.274841	0.7310782
O3	0.04369236	0.05382523	0.12536	0.2367643
SO2	0.62173407	0.28098323	0.758635	0.4344104

The values confirm our theoretical result, except the case of NO_2 in winter, where the

MSPE of the univariate interpolator is smaller than that of the simultaneous interpolator. One interesting point is worth mentioning here. The above numbers show that the relative reduction in the MSPE's achieved by simultaneous interpolation over univariate interpolation, is much higher for SO_4 and O_3 than for SO_2. For SO_4 and O_3, the relative reduction is from 300% to 900%. For SO_2, the reduction is under 50%. This is because SO_4 , O_3 are regional pollutants and SO_2 is a local pollutant. Intuitively, a regional air pollutant intuitively has a higher correlation with the other pollutants than a local pollutant, as indicated by the estimated between-pollutants-hypercorrelation in Section 3.1. By including the other correlated pollutants in the analysis, we would expect to enhance the interpolation. For a local pollutant, since it has little or no correlation with other pollutants, the inclusion of additional pollutants in the analysis will not improve the interpolator as much. Therefore, we can conclude that the interpolator with data missing-by-design does better than that of LZ on regional pollutants. It does not do much better than LZ on local pollutants. The conclusion has theoretical support: if $\underline{X}_0$ are X are independent equality in Equation (5.1) obtains.

5.2 Trends in Adjusted Mean Squared Prediction Error (AMSPE)

We now check the performance of the Bayesian interpolator (2.6) in terms of the AMSPE. Here the adjusted prediction error is defined to be $(X_{pred} - X_{obs})/std(X_{pred})$, that is, the difference between the predicted value and the observed value divided by the standard deviation of the predicted value. AMSPE is the mean of all squared adjusted predicted errors. We prefer AMSPE over the mean squared error (MSE) because different pollutants have different units of measurement, hence different variability. Therefore, mean prediction errors of different pollutants must be normalized to make them comparable.

The data set for our study comes from the monthly Southern Ontario pollutant data used earlier. Among 31 gauged sites 13 sites where NO_2, O_3 and SO_2 are all observed, are chosen. So the original data set consists of thirty-nine columns (three observed pollutants at each of thirteen sites). In our study, a randomly chosen column at a randomly chosen site among the 13 sites is deleted. A CV study is performed on the remaining data set that has data missing-by-design. An AMSPE is computed. Next, in a similar way, a randomly chosen column from the remaining data is deleted and the CV process is repeated on the new data set, which has two columns less than the complete data set. So a new AMSPE is computed. By repeating this and plotting AMSPEs, a trend in AMSPE is obtained.

One example of AMSPE trends is shown in Figure 5.1. There are twenty-four AMSPE's in the plot. The first AMSPE is computed after the column of observed NO_2 at Site 2 is deleted from the original data set, the second AMSPE while the previously chosen column and the column of observed SO_2 at Site 3 are deleted. By repeating the same procedure, the 24^{th} AMSPE is computed after 24 randomly chosen columns are deleted. Below we show the order of the deleted columns.

```
Order: 1   2   3   4   5   6   7   8   9   10  11  12
Pollutant:  NO2 SO2 NO2 O3  O3  SO2 O3  O3  NO2 SO2 NO2 SO2
Site:  2   3   4   1   7   4   10  6   12  7   10  12

Order: 13  14  15  16  17  18  19  20  21  22  23  24
Pollutant:  NO2 O3  O3  SO2 SO2 SO2 NO2 SO2 NO2 SO2 O3  SO2
Site:  3   2   11  9   6   1   5   5   9   11  8   8
```

The AMSPE trend plot in Figure 5.1 shows a generally increasing pattern with some bumps along the way. The incremental changes in AMSPE are not very dramatic. This perhaps indicates good performance of the simultaneous interpolator. However, the result must be interpreted cautiously, since the AMSPE is not a robust index. When the standard deviation of a predicted value is relatively small, it could explode the corresponding adjusted prediction error so much that an AMSPE value will be dominated by a particular term. If a common term plays a dominant role among all the AMSPE values, the trend will be a flat line. Such a plot does not imply superiority of the simultaneous interpolator.

5.3 Comparison with CoKriging

Both Bayesian multivariate interpolation with data missing-by-design (referred to below as the new method) and Hass's CoKriging method can handle data missing-by-design.

Here are some direct contrasts of the two methods. The new method gives a predictive distribution, therefore a simultaneous interpolation region, while Hass's method only gives the interpolated value and its standard deviation. The new method easily handles any number of variables (pollutants), ignoring for the limitation of computer capacity, while Hass's method only allows two variables in order to retain mathematical tractability. The new method includes all available data in the interpolation process, while Hass's method only uses only partial data. However, when sk is big, Hass's method has a substantial computing advantage over the new method, since the convergence of the EM algorithm used in the new method is slow.

A detailed comparison follows. Generally speaking, our simulation study shows that Hass's method enjoys computational advantages over the new method when there are

many sites. When there are not enough sites, Hass's method encounters difficulty. Since in each moving window, Hass's method requires a minimal number of sites in the window for the purpose of variogram estimation. When gauged sites are widely spread out or there are not enough sites, that requirement cannot be met. This affects the accuracy of the estimated variogram. On the other hand, since the new method involves a lot of matrix operations, when there are too many sites, the dimensions of the matrix can be so big that computation becomes excessively slow. In the case of a small number of gauged sites, the new method works fine. Another difference between Hass's method and the new method is in the models they use. While Hass's method models only a "snapshot" of a spatial trend, the new method incorporates the temporal trend for each pollutant at each site. Therefore, the new method can handle spatial-temporal data, but may have problems with spatial data alone. In the case of spatial data alone, it is hard to model a temporal trend with one datum for each observed pollutant at each gauged site.

Because of the complexity of the cross-variogram formulas, Hass's method cannot take more than two variables (e.g., two air pollutants) in practice and it is hard to expand beyond that. The new method can handle any number of variables (pollutants). As shown in Chapter 2, a simultaneous prediction region is derivable when a predictive distribution is available. Hass's method only gives a pairwise confidence interval for the interpolated value.

A fundamental distinction between the two methods lies in their Bayesian against non-Bayesian distinction. We do not take up that issue here.

For an empirical comparison of the two methods, we use a real data set. The data come from selected sites in the National Acidic Deposition (NADP) Network and the

National Trends Network (NTN) in United States. Refer to Wu and Zidek (1992), Watson and Olsen (1984) for more details. Forty-five gauged sites, the latitudes and longitudes of which are in Table 5.1, are randomly chosen. Since both sulfate and nitrate are regional pollutants, they should be highly correlated. That guess is confirmed by an analysis of the data, where the estimated hypercorrelation between sulfate and nitrate is higher than 0.70. The high correlation enables both methods to fully demonstrate their strength and weakness, making their comparison more reliable. At the forty-five sites, both sulfate and nitrate are observed monthly from 1983 to 1986. In terms of Hass's method, one pollutant is treated as a response variable and the other as a covariate. The covariate is included in the analysis to improve the interpolation precision. We arbitrarily chose nitrate as the response variable and sulfate as the covariate. To agree with the form of a CV study provided by Dr. Tim Hass, both observed levels of nitrate and sulfate at the first 35 sites of Table 5.1 are used for our study and only the monthly sulfate observations at the remaining 10 sites are used for the study. The number, 35, is chosen because according to Hass's method, in each moving window there must be a minimum of 30 gauged sites.

For each month of the four years, a CV study of Hass's method is done and the MSPE is computed for the original and log-transformed data respectively. The same study is done for the new method. The MSPE, for both methods on the two cases are listed in Table 5.2. In terms of MSPE, with the original data, the new method beats Hass's method in 32 out of 48 months and in terms of the grand mean of MSPE, the new method also wins by 0.4437143 against 1.25524. Figure 5.2 displays the monthly MSPE's of both methods on the original data. In the figure, "Bayesian" means the new method and CoKriging means Hass's method. With the log-transformed data, the ranking is reversed. The new method loses to Hass's method in 39 out of 48 months and 0.3813128 against 0.2347863 of the mean MSPE. See Figure 5.3 for a graphical comparison of the MSPE's at each

month on the log-transformed data.

By checking the MSPE's of the new method used on the log-transformed data at each site, we found that among 35 MSPE's, the MSPE value at Site 35 (i.e, Site 075a) is much higher than those at other sites. At Site 35, the MSPE value is 7.48, while at other sites, all the MSPE values are below 0.493 (except that at Site 21). Observations at Site 35 are unnaturally compressed into a small interval near zero, making their logarithm extraordinarily large in magnitude. Thus in Table 5.2, we observe MSPE's for transformed data larger than those for the original because of this outlying site. As well, the poor performance at Site 35 makes the monthly MSPE values of the new method applied to the log-transformed data systematically higher than those of Hass's method. Figure 5.4 shows that the boxplot of log-transformed nitrate levels at Site 35 is well below other boxplots as noted above.

Recall, in Chapter 2, we assumed that the hyperparameter B^o has an exchangeable structure. From our simulation study, we know that if that assumption is violated, the proposed new method will not do well. For example, if B^o were taken to be identical over sites for each pollutant while the actual data shows that it is not true, the performance of our interpolator would be poor. So one explanation of the poor performance of the new method on the log-transformed data is that the exchangeability assumption on B^o is violated because of Site 35. That is, the actual hyperparameter B^o at that site is not the same as that of the same pollutant at other sites. To check this conjecture, Sites 21, 35, 36, 44, 46, where either observed sulfate levels or nitrate levels are unusual, are removed from the data set and five new sites are added in (the codes are 076a, 077a, 078a, 160a, 161a). CV studies of both Hass's method and the new method are then carried out on the new log-transformed data. In terms of MSPE, over 48 months the new method beats Hass's method in 35 out of 48 months and in terms of the mean

MSPE, the new method gains an edge by 0.2087363 against 0.3676921. These results support our conjecture.

However, the diagnosis of the exchangeability assumption is difficult. The following theory offers a way to make a rough check.

Note, by Model (2.12)-(2.14) and Equation (2.16),

$$\hat{B}_2^t \mid B_2^o \sim N((B_2^o)^t, ((ZZ^t)^{-1} + F^{-1}) \otimes \Sigma_{22}).$$

Equation (2.8) implies,

$$\hat{B}_2 \mid B_2^o \sim T(\Psi_{22}^{-1}, (ZZ^t)^{-1} + F^{-1}, B_2^o, \delta^* - s_u k - l + h).$$

By Lemma 2.4, the marginal distribution of $\hat{\beta}_r$ $(i = 1, \ldots, s_g k - l)$, the r^{th} row of $\hat{B}_2$, has distribution

$$T(d, (ZZ^t)^{-1} + F^{-1}, \beta_r^o, \delta^* - sk + h + 1),$$

where β_r^o is the r^{th} row of B_2^o. To define d, let

$$D = \begin{pmatrix} D_{11} & D_{12} \\ D_{21} & d_{22} \end{pmatrix} = C_r^t \Psi_{22}^{-1} C_r,$$

where C_r is an orthogonal matrix that exchanges the r^{th} column of Ψ_{22}^{-1} with its last column, D_{11} is $(s_g k - l - 1)$ by $(s_g k - l - 1)$ and d_{22} is 1 by 1. Then d is $d_{22} - D_{21} D_{11}^{-1} D_{12}$. By applying Theorem 2.8, the following theorem is proved.

Theorem 5.2

$$F = \frac{\delta^* - sk + 1}{h} (\hat{\beta}_r - \beta_r^o) \left[\frac{(ZZ^t)^{-1} + F^{-1}}{d} \right]^{-1} (\hat{\beta}_r - \beta_r^o)^t \sim F_{h,\, \delta^* - sk + 1}. \qquad (5.2)$$

Then by the above theorem, the p-value for $\hat{\beta}_{35}$ at Site 35 can be computed as

$$1 - [1 - P_F(F > F_{obs})]^{35},$$

because $\hat{\beta}_{35}$ is the extreme value among the 35 $\hat{\beta}_r$'s, $i = 1, \ldots, 35$. F_{obs} can be computed by plugging in the estimated hyperparameters and $\hat{\beta}_{35}$ into Equation (5.2). With the above p-value formula, unless $P_F(F > F_{obs})$ is extremely small, it is very unlikely we would reject the exchangeability assumption at Site 35. Therefore, the power of the above test is low. An application of the above test to our data shows that we failed to reject it. However, the boxplots of Figure 5.4 make us believe that the observed levels of nitrate at Site 35 are abnormal.

Theoretically, Hass's method is optimal only when the processes of pollutants are multivariate normal with a flat prior on the parameters. That explains why the MSPE values of Hass's method drop a lot when it is applied to the log-transformed data. (See Table 5.2.) A log-transformation made the observed levels of the pollutants follow a multivariate normal more closely, although it is not necessarily the most appropriate transformation. The new method is also sensitive to the normality assumption. From Table 5.2, it can be seen that the MSPE values were reduced in 31 out of 48 months when the new method was applied to the log-transformed data.

Table 5.1: Latitudes and Longitudes of the Gauged Sites.

Sites	Lat.	Long.
004a	36.1006	94.1733
010a	40.3644	105.56
011a	39.1011	105.092
012a	40.8064	104.754
017a	33.1778	84.4061
020a	40.0533	88.3719
021a	41.7011	87.9953
022a	37.71	89.2689
023a	37.4356	88.6719
024a	41.8414	88.8511
025a	41.6325	87.0878
028a	35.6644	83.5903
029a	37.1989	108.491
030a	45.4897	69.6644
031a	45.5611	84.6783
032a	42.4103	85.3928
033a	44.2244	85.8186
034a	47.5311	93.4686
035a	44.2372	95.3006
036a	32.3344	88.745
037a	48.5103	113.996
038a	41.1531	96.4928
039a	43.9431	71.7033
040a	42.7339	76.6597
041a	42.2994	79.3964
046a	43.5261	75.9472
047a	42.1061	77.5356
049a	36.1278	77.175
051a	35.6967	80.6228
052a	35.0239	78.2792
053a	35.7286	78.6811
055a	40.3553	83.0661
056a	39.7928	81.5311
058a	40.78	81.9253
059a	44.3869	123.623
061a	44.2231	122.242
063a	41.5978	78.7678
064a	40.6589	77.9361
065b	40.7883	77.9464
068a	36.0717	112.153
070a	29.3019	103.177
071a	28.8453	96.92
073a	37.335	80.5578
074a	47.86	123.933
075a	39.0897	79.6622

Table 5.2: MSPEs by Both Methods.

month	*original data*		*log-transformed data*	
	CoKriging	new method	CoKriging	new method
1983.1	3.28392	0.52072	0.341134	0.323611
2	5.652717	0.366484	0.549418	0.357877
3	0.950273	0.321812	0.165156	0.303732
4	0.617414	0.396189	0.115926	0.27441
5	0.232695	0.509551	0.083971	0.320196
6	0.914407	0.881811	0.183823	0.407721
7	0.668226	0.786001	0.290934	0.405173
8	1.347384	0.592036	0.263396	0.183026
9	0.300603	0.543329	0.092434	0.332301
10	0.521691	0.375799	0.166289	0.226903
11	0.238963	0.25169	0.079822	0.394561
12	0.320864	0.332461	0.095796	0.389354
1984.1	3.209043	0.402137	0.339717	0.392407
2	0.817165	0.539348	0.234923	0.453388
3	0.469469	0.18821	0.079713	0.394719
4	0.517674	0.288131	0.28535	0.781333
5	0.143388	0.334989	0.047676	0.253829
6	1.500481	0.787056	0.23976	0.400371
7	0.987726	0.775919	0.34463	0.340348
8	0.457944	0.60891	0.208633	0.606851
9	0.210432	0.354206	0.082369	0.290468
10	0.499223	0.502362	0.163754	0.427758
11	0.808873	0.234341	0.146061	0.405709
12	0.254086	0.356278	0.089412	0.501983
1985.1	1.876528	0.238328	0.51466	0.265733
2	1.821478	0.36958	0.181693	0.427747
3	0.398158	0.244937	0.13868	0.267305
4	2.397513	0.430129	0.326478	0.326975
5	0.570281	0.307239	0.13937	0.26641
6	1.89153	0.37827	0.405738	0.461921
7	0.509808	0.51783	0.153515	0.297702
8	0.758525	0.733556	0.215295	0.545632

month	*original data*		*log-transformed data*	
	CoKriging	new method	CoKriging	new method
1985.9	0.302321	0.470564	0.200713	0.58109
10	1.559379	0.484305	0.314374	0.291196
11	0.24711	0.165849	0.188169	0.265927
12	9.080794	0.585204	1.434542	0.395802
1986.1	6.937748	0.553377	0.685439	0.346247
2	1.717417	0.32262	0.29135	0.261732
3	1.376672	0.194057	0.217498	0.315148
4	0.410942	0.32283	0.116953	0.5226
5	0.240855	0.568083	0.069861	0.380813
6	0.270805	0.755552	0.09048	0.36987
7	0.377201	0.637667	0.08781	0.598339
8	0.664169	0.411857	0.226513	0.28267
9	0.492622	0.570382	0.14715	0.43719
10	0.363747	0.310728	0.118633	0.417164
11	0.199882	0.194151	0.109205	0.307664
12	0.861362	0.281423	0.205528	0.502112
Mean	1.25524	0.4437143	0.234786	0.381312

Chapter 6

Concluding Remarks and Future Studies

In this thesis, the theories of Bayesian multivariate interpolation with missing data are discussed. The proposed interpolation theories have two main characteristics, a hierarchical Bayesian method and a multivariate method.

Some obvious advantages of a Bayesian method: with a Bayesian paradigm, the prior information is easily incorporated into an interpolation procedure, if such prior information is indeed available; with a Bayesian approach, one can include the model uncertainty into the confidence interval of interpolated values, while traditional CoKriging fails to do so. Thus, the confidence intervals computed with traditional CoKriging methods are narrower they should be. With a hierarchical Bayesian approach, the specification of the spatial covariance structure can be pushed up to a higher level in the hierarchy. That is, the covariance structure can be specified at Stage 2 of the modeling. Under this setup, if there is any mistake in the specification of the covariance structure, future observations can modify the wrongly specified structure. Eventually if enough data are observed, the mistake will be corrected. Therefore, Bayesian modeling provides robustness. Another benefit of the Bayesian approach is that it provides predictive distribution, while

CoKriging only provides an interpolated value with its standard deviation. By knowing the predictive distribution, one can do more things, e.g., constructing a simultaneous confidence region, creating random numbers for a simulation study and so on.

Another characteristic of the theories in this thesis is their *multivariate* nature. The general theory of BLZ also yields a multivariate interpolator. However when there are missing data, as when data are missing-by-design, with the BLZ theory, interpolation can only be done pollutant-by-pollutant. That reduces the method to a univariate method. The advantage of a multivariate approach is that it allows the interpolation to be carried out by including all the available information. Theorem 5.1 of Chapter 3 concludes that in terms of mean squared error, when hyperparameters are known, a Bayesian interpolator based on all available information performs at least as well as a Bayesian interpolator based on partial information. With that theorem, the extension of the general BLZ theory to the new theory is intuitively justified. As an empirical check, a CV study in Chapter 5 shows that the quantitative gains of the new method over the LZ method are significant.

The theories developed in this thesis are by no means complete. There is much room for further study. For example, a general estimation method of the unknown hyperparameters is needed for the theory of interpolation with monotone missing data. Other future research topics are listed below.

In Chapter 5, while the theory of interpolation with data missing-by-design is compared with Hass's CoKriging theory, we pointed out that if the observed data is a spatial data set only, the theory in this thesis is not directly applicable. That is because the model of our Bayesian interpolation consists of a temporal trend, which can not be estimated by a single datum. Therefore, hierarchical Bayesian CoKriging is of interest. Here *hierarchal*

Bayesian CoKriging means putting a prior on the unknown coefficients of the trend and spatial covariance matrix. One possible approach is as follows. When there are many sites involved, the dimensions of the spatial covariance matrix can be very big. To reduce the number of random variables in the spatial covariance matrix, one may divide all related sites into m clusters, with n sites in each cluster. When the means are assumed to be zero, a model of the random process at the i^{th} cluster is set to be,

$$X_i = R_i \mathbf{1} + L_i,$$

where the random variable R_i reflects the correlations of Cluster i with other clusters, $\mathbf{1}$ is a $1 \times n_i$ vector of 1's and L_i reflects the local spatial correlation at cluster i. Let R represents a $m \times 1$ random vector that takes care of the between-clusters spatial correlation. Conjugate priors of inverted Wishart distribution are assumed on $Var(R)$ and $Var(L_i)$, $i = 1, \ldots, m$. Details of interpolation theory remain to be filled in.

A Gaussian model is assumed for all interpolation theories discussed in this thesis. In some applications, the observed data may not be Gaussian and they cannot be transformed to be Gaussian. An interpolation theory with non-Gaussian data is needed. With a Gaussian model, the interpolator with data missing-by-design has an analytical form. With a general distribution model, an interpolator may not have closed form. In that case, some Bayesian computing tools, like Gibbs' importance sampling may need to be brought in for a numerical answer.

The prior distribution of the unknown spatial covariance matrix, Σ, in this thesis, is taken to be an inverted Wishart, a conjugate distribution of the Gaussian model. Since many authors criticize the Wishart distribution (c.f. Press 1982, Brown, Le and Zidek 1993b), an extension of the inverted Wishart may be needed. On possible choice, which keeps the conjugate property of the priors, therefore, retains mathematical tractability, and at the same time, provides more parameters to offset one deficiency of the Wishart,

is proposed in Brown, Le and Zidek (1993b). The proposed prior is called a *Generalized Inverted Wishart Distribution.* However, the following is worth mentioning. Based on our application and simulation studies in Chapter 3 and 5, our method works well with the inverted Wishart prior.

In Chapter 3, we see that the daily air pollution levels in Southern Ontario show significant lag-1 correlation. There, an $AR(1)$ transformed is adopted to remove the correlation. A better method is to extend the interpolation theory of LZ to include a lag-1 correlated time series, or more generally, to remove the assumption of temporal independence. That extension will have immediate applications.

From the same Southern Ontario study, we see that among the observed data, some are censored below. In Chapter 3, the censored data are simply treated as missing and filled with an ad hoc method. Therefore, a new theory that can handle the censored data will be needed for applications.

A last remark goes to the optimal redesign of a current monitoring network. In Chapter 4, an example demonstrates how to add two sites to a current monitoring network. An interesting question may be raised. Instead of adding a new monitoring site, that measures ozone only, an alternative approach would put an ozone monitor at one of the currently gauged sites where ozone is not monitored. The question is which is optimal? While this can be done with the same entropy approach given in Chapter 4, there are some problems. Normally it is true that the cost of maintaining a separate monitoring site will cost more than putting a meter at a current operating site. The entropy approach does not take that into consideration. Therefore, to solve such a problem, we have to bring additional factors like cost into the objective function.

Bibliography

Afifi, A. A. and Elashoff, R. M. (1961). "Missing Observations in Multivariate Statistics." *J. Amer. Statist. Assoc.* Vol. 61, 595-604.

Ahmed, S. and de Marsily, G., (1987). "Comparison of Geostatistical Methods for Estimating Transmissivity Using Data on Transmissivity and Specific Capacity". *Water Resources Research*, 23, 1717-1737.

Anderson, T.W., (1984). "An Introduction to Multivariate Statistical Analysis". New York: Wiley.

Berndtsson, R., (1988). "Temporal Variability in Spatial Correlation of Daily Rainfall". *Water Resources Research,* Vol. 24, 1511-1517.

Billingsley, P., (1968). "Convergence of Probability Measures." New York: Wiley.

Brown, P.J., Le, N. D. and Zidek, J.V. (1994a). "Multivariate Spatial Interpolation and Exposure to Air Pollutants" . *Canadian Journal of Statistics.* To appear.

Brown P.J., Le, N. D. and Zidek, J.V. (1994b). "Inference for A Covariance Matrix". Aspects of Uncertainty: A Tribute to D.V. Lindley. Ed. A.F.M. Smith and P.R. Freeman. John Wiley & Sons.

Buck, S. F. (1960). "A Method of Estimation of Missing Values in Multivariate Data Suitable for Use with an Electronic Computer." *J. of the Royal Statistical Society*, Ser. B, 22, 302-307.

Burnett, R. T., Dales R. E., Rainenne M. D. and Krewski D. (1992). "The Relationship Between Hospital Admissions and Ambient Air Pollution in Ontario, Canada: A Preliminary Report". Unpublished Report.

Caselton, W.F., Kan, L. and Zidek, J. V. (1992) "Quality Data Networks that Minimize Entropy". Statistics in the Environmental and Earth Sciences. Eds. P. Guttorp and Walden. Griffin, London.

Chen, C.F., (1979). "Bayesian Inference for a Normal Dispersion Matrix and its Application to Stochastic Multiple Regression Analysis". *J. Roy. Statist. Soc.*, B, 41, 235-248.

Cressie, N. and Hawkins, D.M.(1980). "Robust Estimation of the Variogram, I." *J. of the International Association for Mathematical Geology*, 12, 115-125.

Cressie, N. (1986). "Kriging Nonstationary Data". *JASA,* 81, 625-634.

Cressie, N. (1989). "Geostatistics". *The American Statistician,* 43, 197-202.

Cressie, N., (1991a). "Statistics for Spatial Data". New York: Wiley.

Cressie, N., (1991b). "Modeling Growth with Random Sets." In *Spatial Statistics and Imaging* (Proceedings of the AMS-IMS-SIAM Joint Summer Research Conference), A. Possolo, ed. Institute of Mathematical Statistics, Hayward, CA.

Dawid, A.P., (1978). "Extendibility of Spherical Matrix Distribution". *J. Mult. Anal.* 8, 559-66.

Dawid, A.P., (1981). "Some Matrix–variate Distribution Theory". *Biometrika*, 68, 265-74.

Dempster, A.P., Laird, N.M. and Rubin, D.B., (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm (with discussion)". *J. Roy. Statist. Soc* B, 39, 1-38.

Delhomme, J.P. (1978). "Kriging in the Hydrosciences." *Advances in Water Resources,* 1, 251-266.

Dickey, J.M, (1967). "Matricvariate Generalizations of the Multivariate t Distribution and the Inverted Multivariate t–distribution". *Ann. Math. Statist.* 38, 511-8.

Dickey, J.M., Lindley, D.V. and Press, S.J. (1985). "Bayesian Estimation of the Dispersion Matrix of a Multivariate Normal Distribution ". *Communication in Statistics A,* 14:1019-1034.

Duddek, C., Le N. D., Sun W., White R., Wong H., Zidek J.V. (PI) (1994). "Assessing the Impact of Ambient Air Pollution on Hospital Admissions Using Interpolated Exposure Estimates in Both Space and Time: Final Report to Health Canada under DSS Contract h4078-3-C059/01-SS". Unpublished Report.

Dunnett, C. W. and Sobel, (1954) . "A Bivariate Generalization of Student's t-distribution with Tables for Certain Special Cases". *JASA,* 50, 1096-1121

Eaton, M.L., (1983). "Multivariate Statistics: A Vector Space Approach". Wiley, New York.

Guttorp, P., Le N. D., Sampson P.D. and Zidek, J.V., (1993) "Using Entropy in the Redesign of an Environmental Monitoring Network". Multivariate Environmental Statistics; edited by G.P. Patil and C.R. Rao. North Holland/Elsevier Science, NY.

Harrison, P. J. and Stevens, C. F.(1976) "Bayesian Forecasting" (with discussion) *J. of the Royal Statistical Society*, Ser. B, .38, 205-247.

Hass, T. C., (1993). "Cokriging Variables That Are First and Second Order Nonstationary ". Paper presented at the annual meeting of the Statistical Society of Canada, Wolfville, Nova Scotia, June, 1993.

Hass, T. C., (1992). "Redesigning Continental-Scale Monitoring Networks". *Atmospheric Environment* Vol.26A, No.18, 3323-3333.

Hass, T. C., (1990a). "Lognormal and Moving Window Methods of Estimating Acid Deposition". *J. Amer. Statist. Assoc.* 85(412), 950-963.

Hass, T. C., (1990b). "Kringing and Automated Variogram Modeling within a Moving Window". *Atmospheric Environment* Vol.24A, No.7, 1759-1769.

Halley, E. (1686). "An Historical Account of the Trade Winds, and Monsoons Observable in the Seas between and near the Tropics: With an Attempt to Assign the Physical Cause of Said Winds". *Philosophical Transactions* 183, 153-168.

Handcock, M. S. and Stein, M. L.(1993) "A Bayesian Analysis of Kriging" *Technometrics*, Vol.35, No.4, 403-410.

Handcock, M. S. and Wallis, J. R.(1994) "An Approach to Statistical Spatial-Temporal Modeling of Meteorological Fields" *JASA*, Vol.89, No.426, 368-390.

Huijibregts, C. J. and Matheron, G. (1971) "Universal Kriging (An Optimal Method for Estimating and Contouring in Trend Surface Analysis)." In *Proceedings of Ninth International Symposium on Techniques for Decision-Making and Metallurgy, Social Volume*, 12, 159-169.

Johnson, T., Capel, J., McCoy, M. and Warnasch, J. (1994) "Estimation of Carbon Monoxide Exposures and Associated Carboxyhemoglobin Levels Experienced by Residents of Toronto, Ontario Using a Probabilistic Version of NEM." Unpublished Report.

Journel, A. G. (1980) "The Lognormal Approach to Predicting Local Distributions of Selective Mining Unit Grades." *Journal of the International Association for Mathematical Geology*, Vol.12, 285-303.

Kannan, D. (1979). "An Introduction to Stochastic Processes ". New York: North Holland.

Kitanidis, P. K.(1986) "Parameter Uncertainty in Estimation of Spatial Functions: Bayesian Analysis." *Water Resources Research*,22,499-507.

Komungoma, S. (1992) "Assessment of the Quality for the NADP/NTN Data Based on Their Predictability." M.Sc. Thesis, Department of Statistics, University of British Columbia, Vancouver, Canada.

Laslett, G. M.(1994) "Kriging and Splines: An Empirical Comparison of Their Predictive Performance in Some Applications." *JASA*,Vol.89, No.426, 391-409.

Le, N. D. and Zidek, J.V. (1992). "Interpolation with Uncertain Spatial Covariance: A Bayesian Alternative to Kriging". *J. Mult. Anal*, 43, 351-74.

Leonard, T. and Hsu, S.J.H. (1992). "Bayesian Inference for a Covariance Matrix". *Annals of Statistics* , 20, 1669-1696.

Lindley, D. V. and Smith, A.F.M. (1972). "Bayes Estimates for the Linear Model". *J. Roy. Statist. Soc.* B, 34, 1-32.

Little, R. J. A. and Rubin, D.B. (1987). "Statistical Analysis with Missing Data". John Wiley & Sons.

Loader, C. and Switzer, P. (1992). "Spatial Covariance Estimation for Monitoring Data". In *Statistics in the Environmental and Earth Sciences,* eds. A. T. Walden and P. Guttorp, London: Edward Arnold, pp.52-70.

Mardia, K. V., Kent, J.T. and Bibby, J.M. (1979). "Multivariate Analysis". Academic Press, New York.

Mardia, K.V. and Marshall, R.J. (1984). "Maximum Likelihood Estimation of Models for Residual Covariance in Spatial Regression". *Biometrika* , 71, 135-146.

Matheron, G. (1962). "Traite de Geostatistique, Tome I." *Memoires du Bureau de Recherches Geologiques et Minieres*, No. 14. Editions Technip, Paris.

Matheron, G. (1969). "Le Kriegeage Universel." *Cahiers du Centre de Morphologie Mathematique* , No. 1. Fontainebleau, France.

Matheron, G. (1971). "The Theory of Regionalized Variables and Its Applications". *Cahiers du Centre de Morphologie Mathematique*, No. 5. Fontainebleau, France.

Miller, A. A. and Sager, T. W. (1994). "Site Redundancy in Urban Ozone Monitoring." *J. Air & Waste Manage. Assoc.*, 44, 1097-1102.

Muirhead, R. J. (1982). "Aspects of Multivariate Statistical Theory". John Wiley, New York.

Nather, W. (1985). "Effective Observations of Random Fields." *Teubner-Texte zur Mathematik,* Band 72. Teubner, Leigzig.

Neuman S. P. and Jacobson, E.A. (1984). "Analysis of Nonintrinsic Spatial Variability Be Residual Kriging with Application to Regional Groundwater Levels". *J. of the International Association for Mathematical Geology* 16, 499-521.

Olea, R.A. (1984). "Sampling Design Optimization for Spatial Functions ". *Journal of Geophysical Research,* 79, 695-702.

Oehlert, G. W. (1993). "Regional Trends in Sulfate Wet Deposition". *JASA,* 88, 390.3599.

Omre, H. (1987). "Bayesian Kriging–Merging Observations and Qualified Guesses in Kriging ". *Mathematical Geology,* 19, 25-39.

Omre, H. and Halvorsen, K.B. (1989a). "the Bayesian Bridge between Simple and Universal Kriging ". *Mathematical Geology,* 21, 767-786.

Omre, H. and Halvorsen, KB. (1989b). "A Bayesian Approach to Kriging ". *In Proceedings of the Third International Geostatistics Congress I,* ed. M. Armstrong, Dordrecht: Academic Publishers, pp. 49-68.

Pilz, J., (1991). "Bayesian Estimation and Experimental Design in Linear Regression Models." John Wiley, New York. . New York: Holt, Rinehart & Winston

Press, S. J., (1982). "Applied Multivariate Analysis-Using Bayesian and Frequentist Methods of Inference". Holt, Rinehart & Winston, New York.

Rubin, D. B. (1976). "Comparing Regressions When Some Predictor Variables Are Missing". *Technometrics* 18, 201-206.

Sampson, P. and Guttorp, P., (1992). "Nonparametric Estimation of Nonstationary Spatial Covariance Structure". *J. Amer. Statist. Assoc.* Vol.87 No. 417, 108-119.

Sarndel, C. E. and Swensson, B. and Wretman, J. (1992). "Model Assisted Survey Sampling". Springer-Verlag, New York, Inc.

Stein, A. and Corsten, L.C.A., (1991). "Universal Kriging and Cokriging as a Regres-

sion Procedure". *Biometrics* 47, 575-587.

Sten, M. L., Shen, X. and Styer, P.E. (1991). "Applications of a Simple Regression Model to Acid Rain Data." Technical Report No. 276. Department of Statistics, University of Chicago, Chicago, IL., USA.

Taylor, A. E. and Lay, D.C. (1980). "Introduction to Functional Analysis". Wiley, New York.

Tanner, M.A. and Wong, W.H., (1987). "The Calculation of Posterior Distribution by Data Augmentation". *JASA* 82, 528-550.

Wahba, G., (1990a). "Comment on Cressie," *The American Statistician*, 44, 255-256.

Wahba, G., (1990b). "Spline Models for Observational Data," Philadelphia: Society for Industrial and Applied Mathematics.

West, M. and Harrison, P.J., (1986). "Monitoring and Adaptation in Bayesian Forecasting Models". *JASA* 81, 741-750.

Woodbury, A.D. (1989). "Bayesian Updating Revisited,". *Mathematical Geology* 21, 285-308.

Wu, C.F.J., (1983). "On the Convergence Properties of the EM Algorithm," *Ann. Statist.* 11, 95-103.

Wu, S. and Zidek, J.V. (1992). "An Entropy Based Review of Selected NADP/NTN Network Sits for 1983-1986". *Atmospheric Environment.*

Yates, S. and Warrick, A.W. (1987). "Estimating Soil Water Content Using Cokriging." *Soil Science Society of America Journal*,51, 23-30.

Zellner, A. (1971). "An Introduction to Bayesian Inference in Econometrics." New York: John Wiley.

Watson, C.R. and Olsen A. R. (1984). "Acid Deposition System (ADS) for Statis-

tical Reporting–System Design and User's Code Manual". U.S. Environmental Protection Agency.

Appendix A:

In this Appendix, the S function, *iwm(m,cn,p,flag=1,s1,M2,misindx,Tint)*, and its S help file is listed. The iwm S function is designed to perform spatial interpolation and estimation of Λ_g, Ω, δ^* with an EM algorithm that is described at the end of Chapter 2. "iwm" calls a C program, the source file of which is listed in Appendix B.

iwm S funtion help file:

Returns a list of estimated hypercovariance matrices and prior degrees of freedom.

USAGE: $iwm(m, cn, p, flag = 1, s1, M2, misindx, Tint)$

ARGUMENTS:

m: number of measured or non-measured air pollutants at a site

cn: when flag=0, cn is the known degrees of freedom in the prior distribution, "inverse Wishart", of the unknown spatial covariance matrix. When flag=1, cn is unknown and the input is an initial value. That value must be greater than m times the total number of sites, p+s1.

p: number of gauged sites.

flag: a value indicates whether cn is known or not.

s1: number of ungauged sites.

M2: a response matrix. The rows are the observations at gauged sites over the time. The columns are "snapshot" air pollutant levels observed at gauged sites.

misindx: an index vector of missing columns and observed columns.

Tint: a p column matrix with each column being a mean of all columns at a gauged site. In iwm, Tint is used for getting an initial estimation of the between-sites-hypercovariance matrix T1

VALUE:

T1: the estimated between-gauged-sites-hypercovariance matrix.

B1: the estimated between-air-pollutants-hypercovariance matrix.

cn: if flag=0, cn is the same as the input value. If flag=1, cn is the estimated degrees of freedom.

C OBJECT FILE:

The C objective file "Phi11.o" is supposed to be under your S working directory that normally is at position 1 of your S search list. Use S function "search()" to check it. If that is not true, you may need to customize "iwm" function to your own version by replacing the following S commands in your copy of iwm function,

```
filename_search()[1]
filename_paste(c(filename,"/Phi11.o"),collapse="")
dyn.load(filename)
```

with

```
dyn.load("pathname/Phi11.o")
```

where "pathname" is the complete pathname under which "Phi11.o" locates.

DETAILS:

IWM, standing for "Interpolation With Data Missing-by-design", implements the theory of interpolation with data missing-by-design. The theory is developed under a normal distribution model when there are data missing-by-design. The other hyperparameters B°, F^{-1} are estimated by unbiased estimators. The output T1 of iwm is used later as an input for the Sampson and Guttorp nonparametric approach for extending T1 to including all sites. The structure of iwm is that: first, iwm does some initial data manipulation and second, it calls "Phi11.o". It is in the C program where the EM algorithm of the interpolation theory is implemented.

As a warning, iwm adopts an EM algorithm, which is notoriously slow. Sometime when the dimensions of the observed data matrix are big, it could take hours, even days to finish. So first of all, be patient! Secondly, either submit it as a batch job or lower the precision value at line 70 of file "Phi11.c". When you change the source code file, be

sure that you recompile the file with unix command "cc -c Phi11.c" and copy the new object file to your S working directory. The current value is set to be 0.0005. You may set it to a different number. Also you can try various transformations of the data matrix to make it close to normal. By doing those, the convergence is hopefully faster.

EXAMPLES:

Consider an artificial example for an explanation of how to use "iwm". Suppose there are four Sites 1, 2, 3 and 4. Site 1 is an ungauged site (s1=1), Sites 2, 3 and 4 are gauged sites (p=3). Three air pollutants: SO_4, O_3, NO_3 are measured at all sites. Only SO_4 is measured at Site 2, O_3 at Site 3, and, SO_4, O_3 and NO_3 at Site 4. Now label 1 to 3 for SO_4, O_3, NO_3 at Site 2; 4 to 6 for SO_4, O_3, NO_3 at Site 3 and 7 to 9 for SO_4, O_3, NO_3 at Site 4. Since the columns of air pollutants labeled with 2, 3, 4, 6 are missed, and the columns labeled with 1, 5, 7, 8, 9 are observed, misindx=(2,3,4,6,1,5,7,8,9). If further, the observed data matrix is:

	Site #2	Site #3	Site #4		
	SO_4	O_3	SO_4	O_3	NO_3
t=1	1.2	3.4	1.6	4.0	2.7
t=2	1.7	3.1	1.1	3.9	2.5

then

$$M2 = \begin{pmatrix} 1.2 & 3.4 & 1.6 & 4.0 & 2.7 \\ 1.7 & 3.1 & 1.1 & 3.9 & 2.5 \end{pmatrix}$$

and

$$Tint = \begin{pmatrix} 1,2 & 3.4 & 2.667 \\ 1.7 & 3.1 & 2.5 \end{pmatrix}$$

where the third column of Tint is the mean of the columns 4,5,6 of $M2$.

Macros of IWM:

```
function(m,cn,p,flag=1,s1,M2,misindx,Tint)
{
f_dim(M2)[1]
```

```
l_p*m-dim(M2)[2]
tm10_c(1:f)
Z_c(rep(1,f),tm10,cos(2*pi*tm10/12),sin(2*pi*tm10/12))
k_length(Z)/f
Z_matrix(Z,k,f,byrow=T)
Bhat_t(M2)%*%t(Z)%*%solve(Z%*%t(Z))
Mu0_apply(Bhat,2,mean)
S_t(M2)%*%M2-Bhat%*%Z%*%M2
tmpi_as.integer(p*m-l)
F1_matrix(0,k,k)
for (i in 1:tmpi) {
   a_rep(-1/tmpi,tmpi)
   a[i]_1+a[i]
   a_matrix(a, tmpi,1)
   F1_F1+t(Bhat)%*%a%*%t(a)%*%Bhat/c(t(a)%*%S%*%a)
   }
F1_(n+f-k-2)*F1/tmpi
B02_matrix(1,tmpi,1)%*%matrix(Mu0,1,k)
tmp_Bhat-B02
Ssim_S+tmp%*%solve(F1)%*%t(tmp)
B1_diag(rep(1,m))
Bhat2_t(Tint)%*%t(Z)%*%solve(Z%*%t(Z))
T1_t(Tint)%*%Tint-Bhat2%*%Z%*%Tint
n_0
filename_search()[1]
filename_paste(c(filename,"/Phi11.o"),collapse="")
dyn.load(filename)
rvalues_ .C("Phi11", as.double(T1), as.double((B1)),
  as.integer(misindx), as.double((Ssim)),as.integer(flag),
  as.integer(n), as.integer(f), as.integer(p), as.integer(m),
  as.integer(k), as.integer(l), as.double(cn), as.integer(s1))
```

```
list(T1=rvalues[[1]],B1=rvalues[[2]],degree=rvalues[[12]])
}
```

Appendix B:

Source file of C program Phi11.o:

```
#include <math.h>
#include <stdio.h>

void Phi11(T1, B1, misindx1, Ssim1,flag1,n1,f1,p1,m1,k1,l1,cn1,s11)
int *flag1, *n1,*f1,*p1,*m1,*k1,*l1, *misindx1, *s11;
double *T1, *B1, *Ssim1, *cn1;
{
   int flag, n,f,p,m,k,l,s1;
   double **T, **B, **Ssim,cn;
   double **,**matrix();
   double ma,t0,t1,t2,det(),sn(),t3,t4;
   int i,j,*ivector(), *misindx, tmpi2;
   void max1(), 1(),f_mat(), detlog();

   /* flag = 0 degrees of freedom is known
          (ie. m is constant in the paper)
        = 1 estimate the df as well
   n = number of observations in the past
   f = number of observations in the future
   l = number of missing columns
   p = number of (gauged) stations
   s1= number of (ungauged) stations
   m = number of ions
   k = number of coefficients in the linear model
       (i.e. 4  in the paper)
   cn= initial estimate of m;
```

```
        that is, if flag =0 then df =cn else if
    flag=1 then use the algorithm in the paper with
        cn as initial value
    Ssim=S^{~}
    misindx = the index matrix for the missing columns and
    the not missing cls.
*/
    /* T is Lambda, B is Omega */
    flag=*flag1;
    n=*n1;
    f=*f1;
    p=*p1;
    m=*m1;
    k=*k1;
    l=*l1;
    cn=*cn1;
    s1=*s11;
    T=matrix(p,p);
    B=matrix(m,m);
    misindx=ivector(m*p);
    =matrix(m*p,m*p);
    Ssim=matrix(p*m-l,p*m-l);
    /******************/
    for (j=0;j<m*p;++j)
       misindx[j]=*(misindx1+j);
    for (i=0;i<p*m-l;++i)
       for (j=0;j<p*m-l;++j)
          Ssim[i][j]=*(Ssim1+i*(p*m-l)+j);
    for (i=0;i<p;++i)
       for (j=0;j<p;++j)
          T[i][j]=*(T1+i*p+j);
```

```
for (i=0;i<m;++i)
   for (j=0;j<m;++j)
      B[i][j]=*(B1+i*m+j);
/******************/
printf("The program will finish when ma < 0.0005\n");
while (1) {
   /*    printf("step 1\n");  */
   1(sigma, misindx, T, B, cn, p,f,l, m,n,k,Ssim,s1);
   /*    printf("step 2\n");  */
   while (1) {
      t1=det(T,p);
      t2=det(B,m);
      if ( fabs(t1) < 1e-50) t1=1.0;
      if ( fabs(t2) < 1e-50) t2=1.0;
      /*    printf("step 3\n");  */
      max1(cn,T,B,,p,m,s1);
      /*    printf("step 4\n");  */
      t3=det(T,p);
      t4=det(B,m);
      if ( fabs(t3) < 1e-50) t3=1.0;
      if ( fabs(t4) < 1e-50) t4=1.0;
      if ((fabs(log(t1)-log(t3))<1e-4)&&
          (fabs(log(t2)-log(t4))<1e-4)) break;
   }
   ma=cn;
   tmpi2=p*m;
   /*    printf("step 5\n");  */
   detlog(T,B,&t2,&t3,p,m,l,misindx, Ssim);
   /*    printf("step 6\n");  */
   if (flag>0)  cn=sn(cn,n,p,m,s1,f,k,l,t2,t3);
   /*     printf("step 7\n");  */
```

```
      ma=fabs(ma-cn)/cn;
      printf("ma is %lf\n",ma);
      if (ma<0.0005) break;
   }
   for (i=0;i<p;++i)
      for (j=0;j<p;++j)
         *(T1+i*p+j)=T[i][j];
   for (i=0;i<m;++i)
      for (j=0;j<m;++j)
         *(B1+i*m+j)=B[i][j];
   *cn1=cn;
}
/*********************************************/
void max1(c,T,B,L,p,m,s1)
double **T,**B,**L,c;
int p,m,s1;
{
   int i,j,k,l;
   double x0,x1,det();
   void invert_matrix();

   for (k=0;k<p;++k)
      for (l=0;l<p;++l)
         for (i=0,T[k][l]=0.0;i<m;++i)
            for (j=0;j<m;++j)
               T[k][l]+=L[m*k+i][m*l+j]*B[i][j]/
                   (c-(double)(s1*m))/(double)m;
   invert_matrix(T,p);
   for (k=0;k<m;++k)
      for (l=0;l<m;++l)
         for (i=0,B[k][l]=0.0;i<p;++i)
```

```
            for (j=0;j<p;++j)
               B[k][l]+=L[m*i+k][m*j+l]*T[i][j]/
                   (c-(double)(s1*m))/(double)p;
   invert_matrix(B,m);
}
/*********************************************/
double sn(c1,n,p,m,s1,f,k,l,t2,t3)
double c1,t2,t3;
int n,p,m,s1,f,k,l;
{
   double k1,k2,c2,func1(),trigamma();
   int i, tmpi;

   while (1) {
      tmpi=p*m-l;
      k1=func1(c1,n,p,m,s1,f,k,l,t2,t3);
      c2=c1;
      for (i=l,k2=0.0;i<p*m;++i)
         k2=k2+trigamma((c1+(double)(n+f-i-k-s1*m))/2.0)
             -trigamma((c1-(double)(i+s1*m))/2.0);
      k2=k2/2.0;
      c1=c2-k1/k2;
      /* printf("k1,k2,c1,c2  is %lf %lf %lf %lf\n",k1,k2,c1,c2); */
      if (fabs(k1)<1.e-08) break;
      if (c1<(double)(m*(p+s1)))
      {
         if (c2/2>(double)(m*(p+s1)+1)) c1=c2/2;
         else
            c1 = (double)(m*(p+s1)+1);
      }
   }
```

```
   return c1;
}
/*********************************************/
double func1(c,n,p,m,s1,f,k,l,t2,t3)
double c,t2,t3;
int p,m,n,k,s1,f,l;
{
   double t1,digamma(),**matrix(),det();
   int i;
   void f_mat();

   for (i=1,t1=0.0;i<p*m;++i)
      t1=t1+digamma((c+(double)(n+f-i-k-s1*m))/2.0)
          -digamma((c-(double)(i+s1*m))/2.0);
   /*  printf("t1,t2,t3 %lf %lf %lf\n",t1,t2,t3);  */
   return t1-t2+t3;
}
/*********************************************/
void detlog(T,B,t2,t3,p,m,l,misindx,ssim)
double **T, **B, *t2,*t3,**ssim;
int p,m,l, *misindx;
{

   double **Psi22, **tm1, **tm2, **R, **matrix(),det() ;
   double **Phi11, **Psi;
   int    tmpi2,tmpi,i,j,i1,i2,i3,tpi;

   tmpi=p*m-l;
   tmpi2=p*m;
   R=matrix(p*m, p*m);
   Phi11=matrix(tmpi2, tmpi2);
```

```
for (i1=0;i1<tmpi2;++i1)
   for (i2=0;i2<tmpi2;++i2)
      R[i1][i2]=0.0;
for (i=0; i<p*m; i++) {
   tpi=misindx[i]-1;
   R[tpi][i]=1.0;
}
for (i=0;i<tmpi2;++i)
   for (j=0;j<tmpi2;++j)
      Phi11[i][j]=T[i/m][j/m]*B[i%m][j%m];
tm1=matrix(tmpi2, tmpi2);
tm2=matrix(tmpi2, tmpi2);
for (i1=0;i1<tmpi2;++i1)
   for (i2=0;i2<tmpi2;++i2)
      tm1[i1][i2]=R[i2][i1];
for (i1=0;i1<tmpi2;++i1)
   for (i2=0;i2<tmpi2;++i2)
      for (i3=0,tm2[i1][i2]=0.0;i3<tmpi2;++i3)
         tm2[i1][i2]=tm2[i1][i2]+tm1[i1][i3]*Phi11[i3][i2];
f_mat(Phi11,tmpi2);
f_mat(tm1,tmpi2);
Psi=matrix(tmpi2, tmpi2);
for (i1=0;i1<tmpi2;++i1)
   for (i2=0,Psi[i1][i2]=0.0;i2<tmpi2;++i2)
      for (i3=0;i3<tmpi2;++i3)
         Psi[i1][i2]=Psi[i1][i2]+tm2[i1][i3]*R[i3][i2];
f_mat(tm2,tmpi2);
f_mat(R, tmpi2);
Psi22=matrix(tmpi, tmpi);
for (i1=0; i1<tmpi; i1++)
   for (i2=0; i2<tmpi; i2++)  {
```

```
         Psi22[i1][i2]=Psi[i1+l][i2+l]+ssim[i1][i2];
      }
   *t2=log(det(Psi22,tmpi));
   for (i1=0; i1<tmpi; i1++)
      for (i2=0; i2<tmpi; i2++)  {
         Psi22[i1][i2]=Psi[i1+l][i2+l];
      }
   *t3=log(det(Psi22,tmpi));
   f_mat(Psi, tmpi2);
   f_mat(Psi22, tmpi);
}
/******i*********************************************************/

#include <stdio.h>
#include <math.h>
/*****************************************************************/
/*                                                               */
/* This module gives E(\_{(11)} | M_{2}, \Phi, \cn^{*})          */
/* and E(log |\_{(11)}| | m_{2}, \Phi, \cn^{*}) for the          */
/* systematic pattern with whole column missing                  */
/*  is \Sigma_{(11)}                                             */
/* log is log |\Sigma_{(11)} |                                   */
/* phi is the hyperparameter \Phi_{(11)}=                       */
/* \Lambda_{g} \otimes \Omega                                    */
/* cn is the degrees of freedom, another hyperparameter          */
/* s2 is number of gauged stations, s1 is number of ungauged     */
/*   stations                                                    */
/* l is number of missing columns, misindx is the index of       */
/*    missing columns                                            */
/* k is number of responses at each station                      */
/* n is number time spots in the past, f is that in the future   */
```

```
/* h is number of covariates                                          */
/*********************************************************************/
void 1(sigma, misindx, Lambda, Omega,
delta, s2,f,l, k,n,h,Ssim,s1)
double **,  **Lambda, **Omega, **Ssim,delta;
int  l, k,h,n, s2,f, *misindx,s1;
{
   double **Psi11star, **Psi11, **Psi12, **Psi22, **matrix(), **Psi;
   double **R, **Rt, **tm1, **tm2, **tm3, **tm4,**tm5, **1;
   double **Psi22in, **Phi22, **Psi11starin, **Phi11, cont,cont1;
   double **eta12, **eta12star;
   int    i1, i2, i3,tmpi,  tmpi2, tpi;
   void   mat_inverse(), f_mat();
   FILE *out;

   tmpi=s2*k -l ;
   R=matrix(s2*k, s2*k);
   for (i1=0;i1<s2*k;++i1)
      for (i2=0;i2<s2*k;++i2)
         R[i1][i2]=0.0;
   for (i1=0; i1<s2*k; i1++) {
      tpi=misindx[i1]-1;
      R[tpi][i1]=1.0;
   }

   tmpi2=s2*k;
   Phi11=matrix(tmpi2, tmpi2);
   Psi=matrix(tmpi2, tmpi2);
   for (i1=0;i1<tmpi2;++i1)
      for (i2=0;i2<tmpi2;++i2)
         Phi11[i1][i2]=Lambda[i1/k][i2/k]*Omega[i1%k][i2%k];
```

```
tm1=matrix(tmpi2, tmpi2);
tm2=matrix(tmpi2, tmpi2);
for (i1=0;i1<tmpi2;++i1)
   for (i2=0;i2<tmpi2;++i2)
      tm1[i1][i2]=R[i2][i1];
for (i1=0;i1<tmpi2;++i1)
   for (i2=0;i2<tmpi2;++i2)
      for (i3=0,tm2[i1][i2]=0.0;i3<tmpi2;++i3)
         tm2[i1][i2]=tm2[i1][i2]+tm1[i1][i3]*Phi11[i3][i2];
for (i1=0;i1<tmpi2;++i1)
   for (i2=0,Psi[i1][i2]=0.0;i2<tmpi2;++i2)
      for (i3=0;i3<tmpi2;++i3)
         Psi[i1][i2]=Psi[i1][i2]+tm2[i1][i3]*R[i3][i2];
f_mat(tm1, tmpi2);
f_mat(tm2, tmpi2);
Psi11=matrix(l,l);
for (i1=0; i1<l; i1++)
   for (i2=0; i2<l; i2++) {
      Psi11[i1][i2]=Psi[i1][i2];
   }
Psi12=matrix(l,tmpi);
for (i1=0; i1<l; i1++)
   for (i2=0; i2<tmpi; i2++) {
      Psi12[i1][i2]=Psi[i1][i2+l];
   }
Psi22=matrix(tmpi,tmpi);
for (i1=0; i1<tmpi; i1++)
   for (i2=0; i2<tmpi; i2++) {
      Psi22[i1][i2]=Psi[i1+l][i2+l];
   }
eta12=matrix(l,tmpi);
```

```
Psi22in=matrix(tmpi, tmpi);
mat_inverse(Psi22, tmpi, Psi22in);
f_mat(Psi22, tmpi);
Psi22=matrix(tmpi, tmpi);
for (i1=0; i1<tmpi; i1++)
   for (i2=0; i2<tmpi; i2++)  {
      Psi22[i1][i2]=Psi[i1+l][i2+l];
   }
f_mat(Psi,tmpi2);
for (i1=0;i1<l;++i1)
   for (i2=0;i2<tmpi;++i2)
      for (i3=0,eta12[i1][i2]=0.0;i3<tmpi;++i3)
         eta12[i1][i2]=eta12[i1][i2]+Psi12[i1][i3]*Psi22in[i3][i2];
tm1=matrix(tmpi, l);
for (i1=0;i1<tmpi;++i1)
   for (i2=0;i2<l;++i2)
      tm1[i1][i2]=Psi12[i2][i1];
tm2=matrix(l, l);
for (i1=0;i1<l;++i1)
   for (i2=0;i2<l;++i2)
      for (i3=0,tm2[i1][i2]=0.0;i3<tmpi;++i3)
         tm2[i1][i2]=tm2[i1][i2]+eta12[i1][i3]*tm1[i3][i2];
Psi11star=matrix(l,l);
for (i1=0; i1<l; i1++)
   for (i2=0; i2<l; i2++)
      Psi11star[i1][i2]=Psi11[i1][i2]-tm2[i1][i2];
Psi11starin=matrix(l,l);
/*out=fopen("junk","w"); */
/*for (i1=0;i1<l;i1++) {*/
/*  for (i2=0;i2<l;++i2) {*/
/*    fprintf(out,"%5.2lf ",Psi11star[i1][i2]);*/
```

```
/*    if (i1%10==9) fprintf(out,"\n"); }*/
/*fprintf("\n\n"); }*/
/*fclose(out);*/
/*exit(0);*/
if (l != 0)  mat_inverse(Psi11star, l, Psi11starin);
if (l !=0) {
   f_mat(Psi11star,l);
   f_mat(Psi11,l);
   f_mat(tm2, l);
   f_mat(Psi12,l);
}
f_mat(tm1, tmpi);
cont=delta-(double)(s1*k);
1=matrix(tmpi2,tmpi2);
for (i1=0; i1<l; i1++)
   for (i2=0; i2<l; i2++)
      1[i1][i2]=cont*Psi11starin[i1][i2];
tm1=matrix(l,tmpi);
for (i1=0;i1<l;++i1)
   for (i2=0;i2<tmpi;++i2)
      for (i3=0,tm1[i1][i2]=0.0;i3<l;++i3)
         tm1[i1][i2]=tm1[i1][i2]+Psi11starin[i1][i3]*eta12[i3][i2];
for (i1=0; i1<l; i1++)
   for (i2=0; i2<tmpi; i2++) {
      1[i1][i2+l]=-1.0*cont*tm1[i1][i2];
      1[i2+l][i1]=sigma1[i1][i2+l];
   }
cont1=delta-(double)(s1*k-n-f+l+h);
tm2=matrix(tmpi,l);
for (i1=0;i1<tmpi;++i1)
   for (i2=0;i2<l;++i2)
```

```
      tm2[i1][i2]=eta12[i2][i1];
tm3=matrix(tmpi,tmpi);
for (i1=0;i1<tmpi;++i1)
   for (i2=0;i2<tmpi;++i2)
      for (i3=0,tm3[i1][i2]=0.0;i3<l;++i3)
         tm3[i1][i2]=tm3[i1][i2]+tm2[i1][i3]*tm1[i3][i2];
if (l !=0) f_mat(tm1,l);
f_mat(tm2, tmpi);
tm4=matrix(tmpi,tmpi);
for (i1=0;i1<tmpi;++i1)
   for (i2=0;i2<tmpi;++i2)
      tm4[i1][i2]=Psi22[i1][i2]+Ssim[i1][i2];
tm5=matrix(tmpi,tmpi);
mat_inverse(tm4, tmpi, tm5);
tm1=matrix(tmpi,tmpi);
for (i1=0; i1<tmpi; i1++)
   for (i2=0; i2<tmpi; i2++) {
      1[i1+l][i2+l]=cont1*tm5[i1][i2]+cont*tm3[i1][i2]
          +(double)l*Psi22in[i1][i2];
   }
f_mat(tm1,tmpi);
f_mat(tm4, tmpi);
f_mat(Phi11,tmpi2);
f_mat(tm5, tmpi);
f_mat(Psi22, tmpi);
f_mat(tm3, tmpi);
f_mat(Psi22in, tmpi);
if (l !=0) {
   f_mat(eta12,l);
   f_mat(Psi11starin,l);
}
```

```
   tm1=matrix(tmpi2,tmpi2);
   Rt=matrix(tmpi2, tmpi2);
   for (i1=0;i1<tmpi2;++i1)
      for (i2=0;i2<tmpi2;++i2)
         Rt[i1][i2]=R[i2][i1];
   for (i1=0;i1<s2*k;++i1)
      for (i2=0;i2<s2*k;++i2) {
         tm1[i1][i2]=0.0;
         for (i3=0;i3<s2*k;++i3) tm1[i1][i2]=
         tm1[i1][i2]+1[i1][i3]*Rt[i3][i2];
      }
   for (i1=0;i1<s2*k;++i1)
      for (i2=0;i2<s2*k;++i2) {
         [i1][i2]=0.0;
         for (i3=0;i3<s2*k;++i3) [i1][i2]=
         [i1][i2]+R[i1][i3]*tm1[i3][i2];
      }
   f_mat(tm1, tmpi2);
   f_mat(Rt, tmpi2);
   f_mat(R,tmpi2);
   f_mat(1,tmpi2);
}
/*********************************************************/

#include <math.h>
#define SIGN(a,b) ((b)<0.0 ? -fabs(a) : fabs(a))
/*********************************************************/
void ord_mat(nvr,eig,mat)
int nvr;
double *eig,**mat;
{
```

```
   int i=0,j=0,k=0;
   double tmp=0.0;

   for (i=0;i<nvr;i++)
      for (j=i;j<nvr;j++)
         if (eig[j]<eig[i]) {
            tmp=eig[j];
            eig[j]=eig[i];
            eig[i]=tmp;
            for (k=0;k<nvr;k++) {
               tmp=mat[k][j];
               mat[k][j]=mat[k][i];
               mat[k][i]=tmp;
            }
         }
}
/*********************************************************/
void tred2(a,n,d,e)
double **a,d[],e[];
int n;
{
   int l=0,k=0,j=0,i=0;
   double scale=0.0,hh=0.0,h=0.0,g=0.0,f=0.0;

   for (i=n-1;i>=1;i--) {
      l=i-1;
      h=scale=0.0;
      if (l > 0) {
         for (k=0;k<=l;k++)
            scale += fabs(a[i][k]);
         if (scale == 0.0)
```

```
            e[i]=a[i][l];
        else {
            for (k=0;k<=l;k++) {
                a[i][k] /= scale;
                h +=a[i][k]*a[i][k];
            }
            f=a[i][l];
            g = f>0.0 ? -sqrt(h) : sqrt(h);
            e[i]=scale*g;
            h -= f*g;
            a[i][l]=f-g;
            f=0.0;
            for (j=0;j<=l;j++) {
                a[j][i]=a[i][j]/h;
                g=0.0;
                for (k=0;k<=j;k++)
                    g +=a[j][k]*a[i][k];
                for (k=j+1;k<=l;k++)
                    g +=a[k][j]*a[i][k];
                e[j]=g/h;
                f += e[j]*a[i][j];
            }
            hh=f/(h+h);
            for (j=0;j<=l;j++) {
                f=a[i][j];
                e[j]=g=e[j]-hh*f;
                for (k=0;k<=j;k++)
                    a[j][k] -= (f*e[k]+g*a[i][k]);
            }
        }
    } else
```

```
            e[i]=a[i][l];
         d[i]=h;
      }
      d[0]=0.0;
      e[0]=0.0;
      for (i=0;i<n;i++) {
         l=i-1;
         if (d[i]) {
            for (j=0;j<=l;j++) {
               g=0.0;
               for (k=0;k<=l;k++)
                  g +=a[i][k]*a[k][j];
               for (k=0;k<=l;k++)
                  a[k][j] -= g*a[k][i];
            }
         }
         d[i]=a[i][i];
         a[i][i]=1.0;
         for (j=0;j<=l;j++)
            a[i][j]=a[j][i]=0.0;
      }
   }
   /*************************************************************/
   void tqli(d,e,n,z)
   double d[],e[],**z;
   int n;
   {
      int m=0,l=0,iter=0,i=0,k=0;
      double s=0.0,r=0.0,p=0.0,g=0.0,f=0.0,dd=0.0,c=0.0,b=0.0;
      void nrerror();
```

```
for (i=1;i<n;i++)
   e[i-1]=e[i];
e[n-1]=0.0;
for (l=iter=0;l<n;l++)
   do {
      for (m=l;m<n-1;m++) {
         dd=fabs(d[m])+fabs(d[m+1]);
         if (fabs(e[m])+dd == dd) break;
      }
      if (m != l) {
         if ((iter++) == 10000)
            nrerror("Too many iterations in TQLI");
         g=(d[l+1]-d[l])/(2.0*e[l]);
         r=sqrt((g*g)+1.0);
         g=d[m]-d[l]+e[l]/(g+SIGN(r,g));
         s=c=1.0;
         p=0.0;
         for (i=m-1;i>=l;i--) {
            f=s*e[i];
            b=c*e[i];
            if (fabs(f) >= fabs(g)) {
               c=g/f;
               r=sqrt((c*c)+1.0);
               e[i+1]=f*r;
               c *= (s=1.0/r);
            } else {
               s=f/g;
               r=sqrt((s*s)+1.0);
               e[i+1]=g*r;
               s *= (c=1.0/r);
            }
```

```
                g=d[i+1]-p;
                r=(d[i]-g)*s+2.0*c*b;
                p=s*r;
                d[i+1]=g+p;
                g=c*r-b;
                for (k=0;k<n;k++) {
                   f=z[k][i+1];
                   z[k][i+1]=s*z[k][i]+c*f;
                   z[k][i]=c*z[k][i]-s*f;
                }
             }
             d[l]=d[l]-p;
             e[l]=g;
             e[m]=0.0;
          }
       } while (m != l);
    /* printf("Number iterations in tqli = %d\n",iter); */
}
/*************************************************************/
#include <malloc.h>
#include <stdio.h>
/*************************************************************/
void nrerror(error_text)
char error_text[];
{
    fprintf(stderr,"Numerical Recipes run-time error...\n");
    fprintf(stderr,"%s\n",error_text);
    fprintf(stderr,"...now exiting to system...\n");
    exit(1);
}
/*************************************************************/
```

```
double *vector(nh)
int nh;
{
   double *v;

   v=(double *)calloc(nh,sizeof(double));
   if (!v) nrerror("allocation failure in dvector()");
   return v;
}
/*************************************************************/
int *ivector(nh)
int nh;
{
   int *v;

   v=(int *)calloc(nh,sizeof(int));
   if (!v) nrerror("allocation failure in ivector()");
   return v;
}
/*************************************************************/
double **matrix(nrh,nch)
int nrh,nch;
{
   int i,j;
   double **m;

   m=(double **) calloc(nrh,sizeof(double*));
   if (!m) nrerror("allocation failure 1 in matrix()");
   for(i=0;i<nrh;i++) {
      m[i]=(double *)calloc(nch,sizeof(double));
      if (!m[i]) nrerror("allocation failure 2 in matrix()");
```

```
    }
    return m;
}
/*********************************************************/
void f_mat(m,nrh)
double **m;
int nrh;
{
    int i=0;

    for(i=nrh-1;i>=0;i--) free((char*) (m[i]));
    free((char*) (m));
}
/*********************************************************/
double **dmatrix(nrh,nch)
int nrh,nch;
{
    int i;
    double **m;

    m=(double **) calloc(nch,sizeof(double*));
    if (!m) nrerror("allocation failure 1 in dmatrix()");
    for(i=0;i<nch;i++) {
        m[i]=(double *)calloc(nrh,sizeof(double));
        if (!m[i]) nrerror("allocation failure 2 in dmatrix()");
    }
    return m;
}
/*********************************************************/
#include <math.h>
double det(h,m)
```

```
double **h;
int m;
{
   double *dv,*er,tmp,**temp;
   double *vector(),**matrix();
   void tqli(),tred2(),f_mat();
   int i,j;

   temp=matrix(m,m);
   for (i=0;i<m;++i)
      for (j=0;j<m;++j)
         temp[i][j]=h[i][j];
   er=vector(m);
   dv=vector(m);
   tred2(temp,m,dv,er);
   tqli(dv,er,m,temp);
   for (i=0,tmp=1.0;i<m;++i)
      tmp*=dv[i];
   free((char*)(dv));
   free((char*)(er));
   f_mat(temp,m);
   return tmp;
}
/*********************************************************/
void mat_inverse(h,m,invh)
double **h, **invh;
int m;
{
   double **hs,*dv,*er,**lam;
   double *vector(),**matrix();
   void tqli(),tred2(),f_mat(),m_mat(),nrerror();
```

```
   int i,j;

   er=vector(m);
   lam=matrix(m,m);
   hs=matrix(m,m);
   dv=vector(m);
   tred2(h,m,dv,er);
   tqli(dv,er,m,h);
   for (i=0;i<m;++i)
      if (dv[i]!=0.0)
         lam[i][i]=1.0/dv[i];
      else nrerror("trying to invert a singular matrix");
   m_mat(hs,h,lam,m,m,m);
   for (i=0;i<m;++i)
      for (j=0;j<m;++j)
         lam[i][j]=h[j][i];
   m_mat(invh,hs,lam,m,m,m);
   f_mat(hs,m);
   f_mat(lam,m);
   free((char*)(dv));
   free((char*)(er));
}
/*************************************************************/
void invert_matrix(h,m)
double **h;
int m;
{
   double **hs,*dv,*er,**lam;
   double *vector(),**matrix();
   void tqli(),tred2(),f_mat(),m_mat(),nrerror();
   int i,j;
```

```
   er=vector(m);
   lam=matrix(m,m);
   hs=matrix(m,m);
   dv=vector(m);
   tred2(h,m,dv,er);
   tqli(dv,er,m,h);
   for (i=0;i<m;++i)
      if (dv[i]!=0.0)
         lam[i][i]=1.0/dv[i];
      else nrerror("trying to invert a singular matrix");
   m_mat(hs,h,lam,m,m,m);
   for (i=0;i<m;++i)
      for (j=0;j<m;++j)
         lam[i][j]=h[j][i];
   m_mat(h,hs,lam,m,m,m);
   f_mat(hs,m);
   f_mat(lam,m);
   free((char*)(dv));
   free((char*)(er));
}
/*************************************************************/
void m_mat(m0,m1,m2,n,p,m)
double **m0,**m1,**m2;
int n,p,m;
{
   int i,j,k;

   for (i=0;i<n;++i)
      for (j=0;j<m;++j)
         for (k=0,m0[i][j]=0.0;k<p;++k)
```

```
            m0[i][j]+=m1[i][k]*m2[k][j];
}
/*************************************************************/
void s_mat(m0,m1,m2,n,m,k)
double **m0,**m1,**m2;
int n,m,k;
{
   int i,j;

   if (k==0)
      for (i=0;i<n;++i)
         for (j=0;j<m;++j)
            m0[i][j]=m1[i][j]-m2[i][j];
   else
      for (i=0;i<n;++i)
         for (j=0;j<m;++j)
            m0[i][j]=m1[i][j]+m2[i][j];
}
/*************************************************************/
#include <math.h>

double trigamma(z)
double z;
{
   double retval,r1,d,w,x,ww;
   int i1,i2,j;

   i2=(int)(z);
   if (i2>20) {
      r1=z;
      ww=r1*r1;
```

```
      x=1-(0.2-1/(ww*7))/ww;
      d=1/z+(x/(z*3)+1)/(ww*2);
   }
   else {
      i2=20-i2;
      w=z+i2;
      r1=w;
      ww=r1*r1;
      x=1-(0.2-1/(ww*7))/ww;
      d=1/w+(x/(w*3)+1)/(ww*2);
      i1=i2;
      for (j=0;j<i1;++j) {
         w-=1;
         r1=w;
         d+=1/(r1*r1);
      }
   }
   retval=d;
   return retval;
}
/*********************************************************/
double digamma(x)
double x;
{
   double y,retval,s3,s4,s5,d1,r,gamma();

   s3=1.0/12;
   s4=1.0/120;
   s5=0.003968253968;
   d1=-0.5772156649;
   retval=0.0;
```

```
   y=x;

   if (y<=0.0) return retval;
   if (y<=1e-5) return (d1-1.0/y);
   while (y<8.5) {
      retval=retval-1.0/y;
      y=y+1.0;
   }
   r=1.0/y;
   retval=retval+log(y)-r/2;
   r*=r;
   retval=retval-r*(s3-r*(s4-r*s5));
   return retval;
}
/*****************************************************/
double gamma(x)
double x;
{
   double gcs[24],pi,sq2pil,xmin,xmax,dxrel,y,sinpiy,gamm;
   double r9lgmc(),csevl(),gamlin();
   int ngcs,i,n,inits();
   gcs[1]=  0.008571195590989331e0;
   gcs[2]=  0.004415381324841007e0;
   gcs[3]=  0.05685043681599363e0;
   gcs[4]= -0.004219835396418561e0;
   gcs[5]=  0.001326808181212460e0;
   gcs[6]= -0.000189302452979888O0;
   gcs[7]=  0.000036069253274412 4e0;
   gcs[8]= -0.000006056761904460 8e0;
   gcs[9]=  0.000001055829546302 2e0;
   gcs[10]= -0.000000181196736554 2e0;
```

```
gcs[11]=  0.00000003117724964 71e0;
gcs[12]= -0.00000000535421963 90e0;
gcs[13]=  0.00000000091932755 19e0;
gcs[14]= -0.00000000015779412 80e0;
gcs[15]=  0.00000000002707980 62e0;
gcs[16]= -0.00000000000464681 86e0;
gcs[17]=  0.00000000000079733 50e0;
gcs[18]= -0.00000000000013680 78e0;
gcs[19]=  0.00000000000002347 31e0;
gcs[20]= -0.00000000000000402 74e0;
gcs[21]=  0.00000000000000069 10e0;
gcs[22]= -0.00000000000000011 85e0;
gcs[23]=  0.00000000000000002 03e0;
pi=3.14159265358979324e0;
sq2pil=0.91893853320467274e0;
ngcs = inits(gcs,23,5.9604645e-8);
gamlim(&xmin,&xmax);
dxrel = sqrt(1.1920929e-6);
y = fabs(x);
if (y>10.0) {
   gamm = exp((y-0.5)*log(y)-y+sq2pil+r9lgmc(y));
   if (x<=0.) {
      sinpiy = sin(pi*y);
      gamm = -pi/(y*sinpiy*gamm);
   }
}
else {
   n = x/1;
   if (x<0.)
      n = n-1;
   y = x-n/1.0;
```

```
      n = n-1;
      gamm = 0.9375+csevl(2.*y-1.,gcs,ngcs);
      if (n>0)
         for (i=1;i<=n;i++)
            gamm = (y+i/1.0)*gamm;
      else {
         n = -n;
         for (i=1;i<=n;i++)
            gamm = gamm/(x+(i-1)/1.0);
      };
   };
   return(gamm);
}
gamlim(xmin,xmax)
double *xmin,*xmax;
{
   double alnsml,alnbig,xln,xold,log(),fabs();
   int i;
   alnsml=log(2.9387359e-37);
   *xmin= -alnsml;
   for (i=1;i<=10;i++)
   {
      xold= *xmin;
      xln=log(*xmin);
      *xmin= *xmin- *xmin*((*xmin+0.5)*xln-
          *xmin-0.2258+alnsml)/(*xmin*xln+0.5);
      if (fabs(*xmin-xold)<0.005) break;
   };
   *xmin= - *xmin+0.01;
   alnbig=log(1.7014117e38);
   *xmax=alnbig;
```

```
   for (i=1;i<=10;i++)
   {
      xold= *xmax;
      xln=log(*xmax);
      *xmax= *xmax- *xmax*((*xmax-0.5)*xln-
          *xmax+0.9189-alnbig)/(*xmax*xln-0.5);
      if (fabs(*xmax-xold)<0.005) break;
   };
   *xmax= *xmax-0.01;
   *xmin=(*xmin>1.0- *xmax)? *xmin:1.0- *xmax;
}
double r9lgmc(x)
double x;
{
   double algmcs[7],xbig,xmax,y,z,csevl(),sqrt(),exp(),log();
   int nalgm,inits();
   algmcs[1]=  0.166638948045186;
   algmcs[2]= -0.0000138494817606;
   algmcs[3]=  0.0000000098108256;
   algmcs[4]= -0.0000000000180912;
   algmcs[5]=  0.0000000000000622;
   algmcs[6]= -0.0000000000000003;
   nalgm=inits(algmcs,6,5.9604645e-7);
   xbig=1.0/sqrt(5.9604645e-7);
   y=log(1.7014117e38/12.0);
   z= -log(12.0*2.9387359e-37);
   xmax=(y<z)? exp(y):exp(z);
   if (x<xbig) return(csevl(2.0*(10.0/x)*(10.0/x)-
       1.0,algmcs,nalgm)/x);
   else return(1.0/(12.0*x));
}
```

```
double csevl(x,cs,n)
int n;
double x,cs[201];
{
   double b0,b1,b2,twox;
   int i,ni;
   b1=0.0;
   b0=0.0;
   twox=2.0*x;
   for (i=1;i<=n;i++)
   {
      b2=b1;
      b1=b0;
      ni=n+1-i;
      b0=twox*b1-b2+cs[ni];
   };
   return(0.5*(b0-b2));
}
int inits(os,nos,eta)
int nos;
double os[201],eta;
{
   int i,ii;
   double err,fabs();
   err=0.0;
   ii=1;
   while (ii<=nos && err<eta)
   {
      i=nos+1-ii;
      err=err+fabs(os[i]);
      ii++;
```

```
    };
    return (i);
}
```

Appendix C: Figures

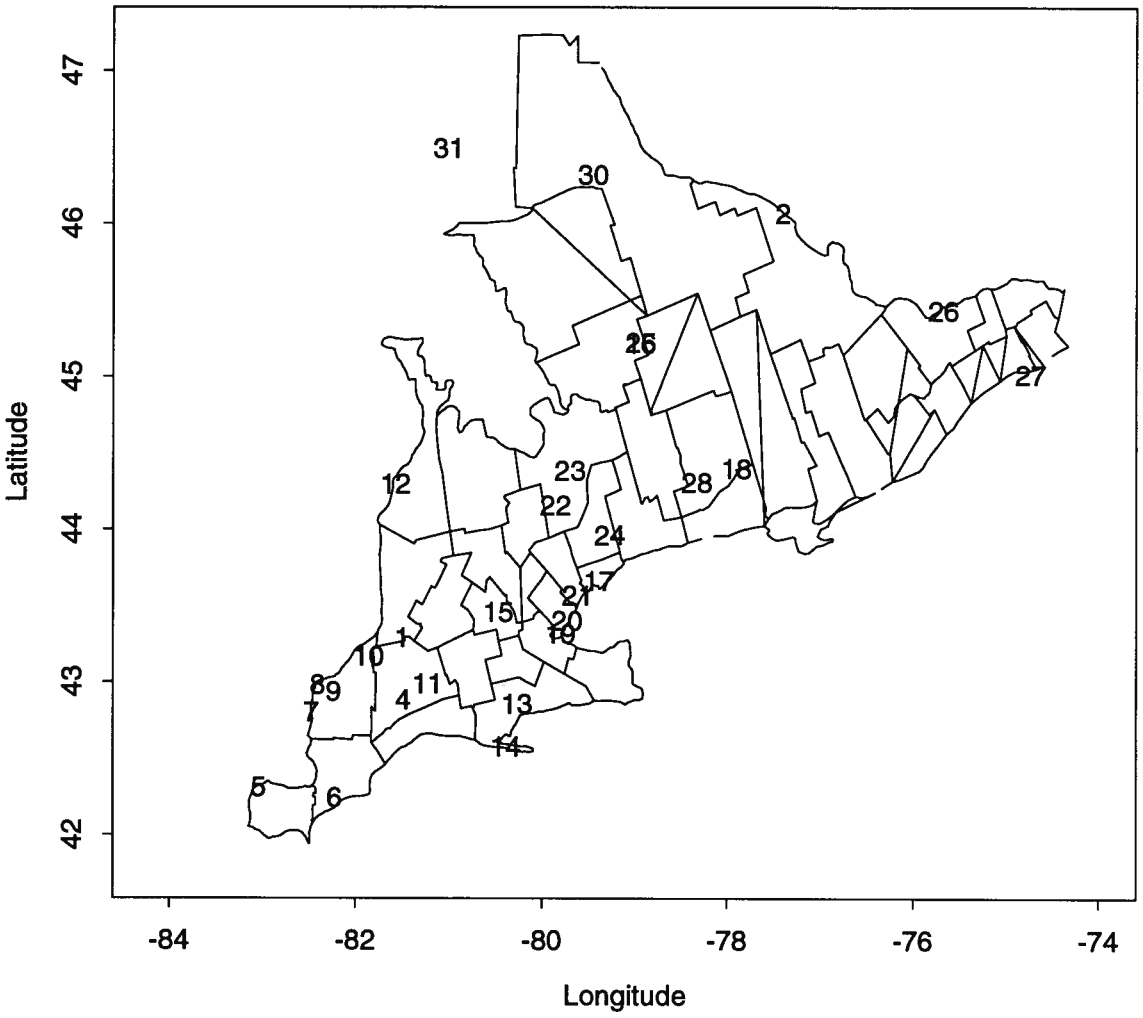


Figure 3.1: Locations of gauged sites in Southern Ontario plotted with Census Subdivision boundaries, where monthly pollution levels are observed and Sites 3, 29 (outliers) are not plotted.

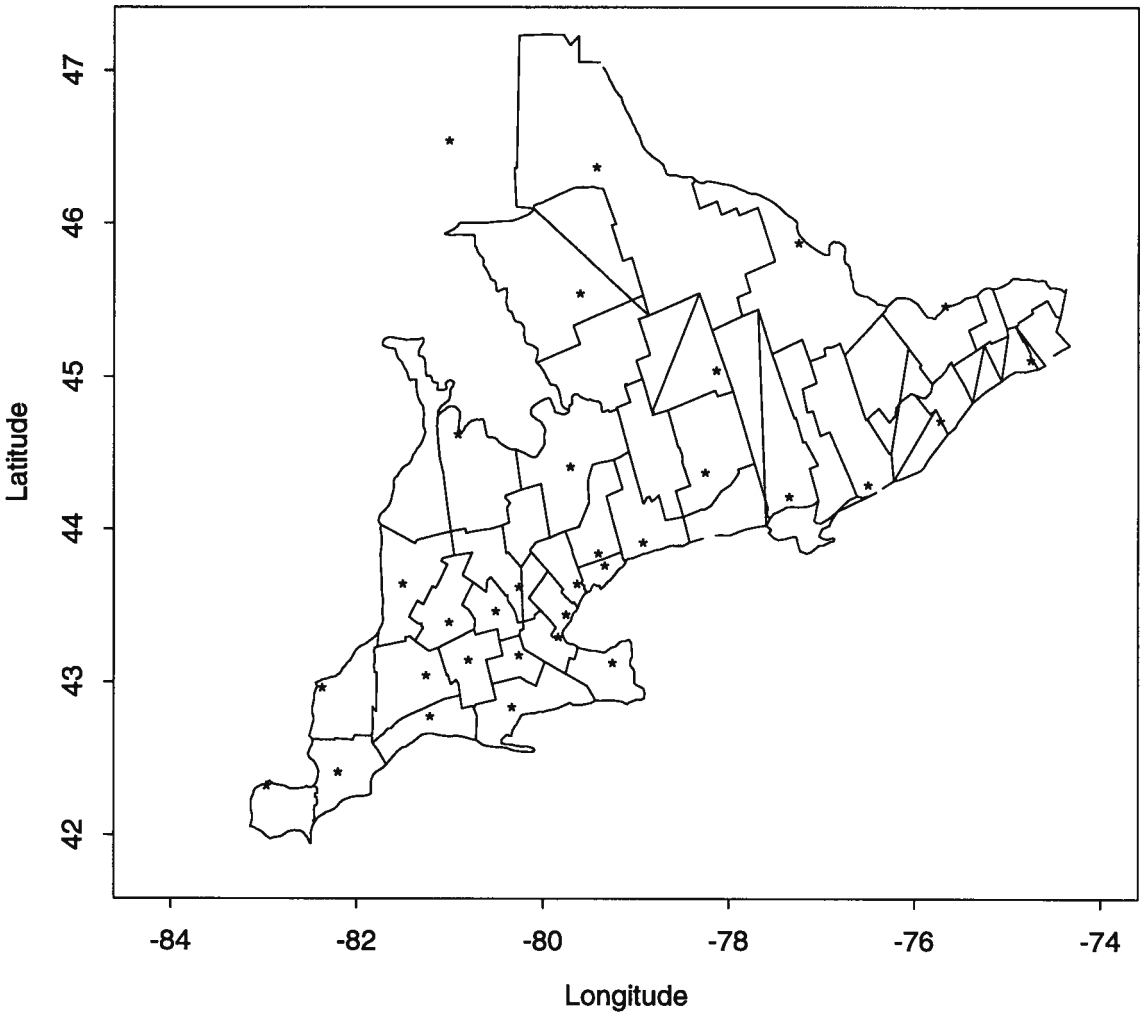


Figure 3.2: Locations of selected sites in Southern Ontario plotted with Census Subdivision boundaries, where monthly interpolated pollution levels are needed.

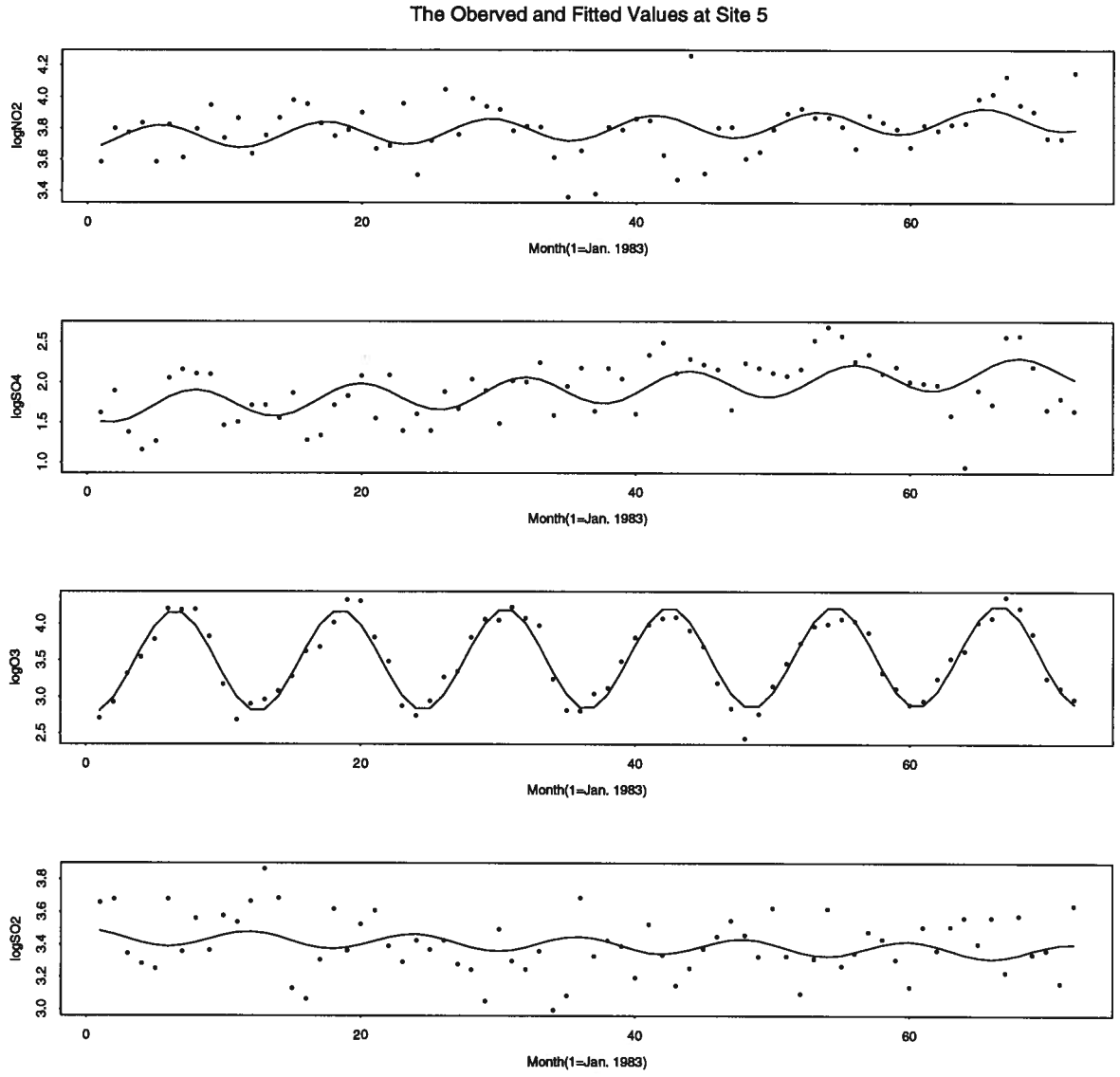


Figure 3.3: Plots for monthly observed and fitted, log-transformed levels of O_3 in *ppb*, SO_2, NO_2 and SO_4 in $\mu g/m^3$, at Gauged Site 5.

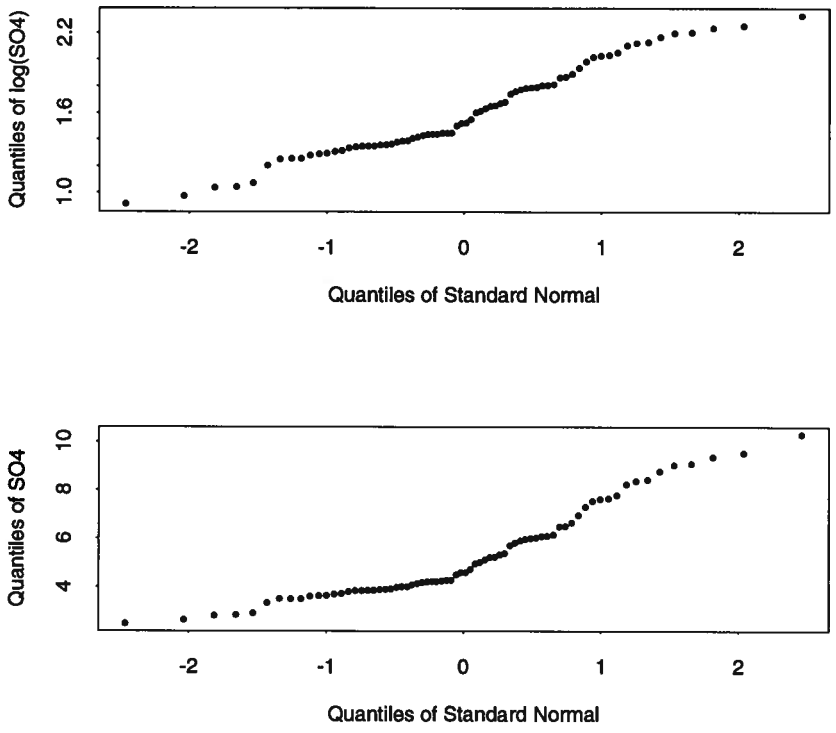


Figure 3.4: Normal quantile-quantile plots for original and log-transformed monthly levels of SO_4 in $\mu g/m^3$ at Gauged Site 4.

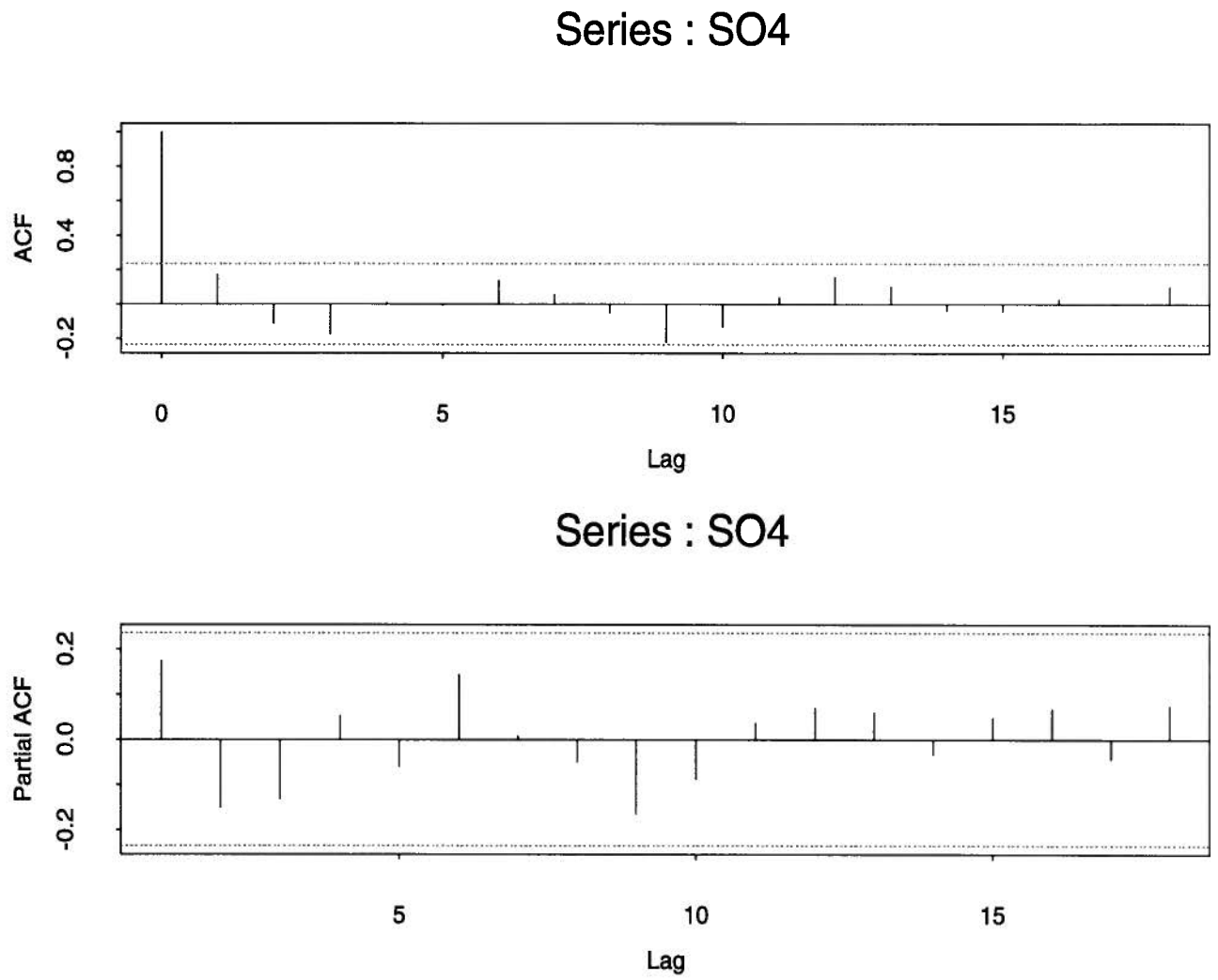


Figure 3.5: Plots for autocorrelation and partial autocorrelation of monthly, log-transformed levels of SO_4 in $\mu g/m^3$ at Gauged Site 4.

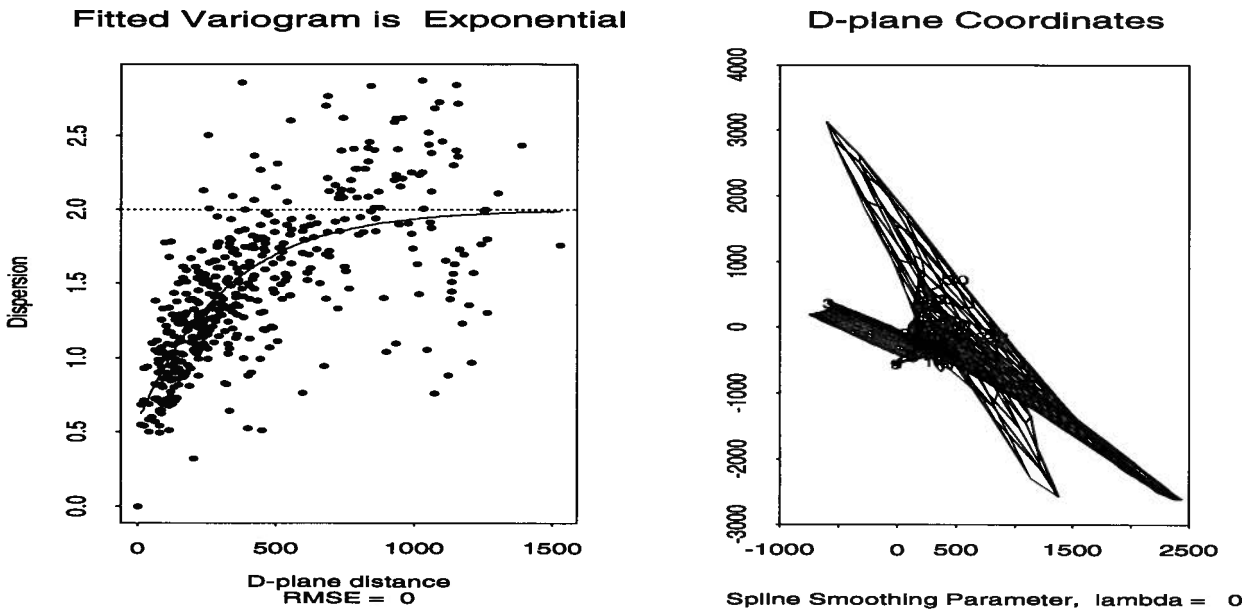
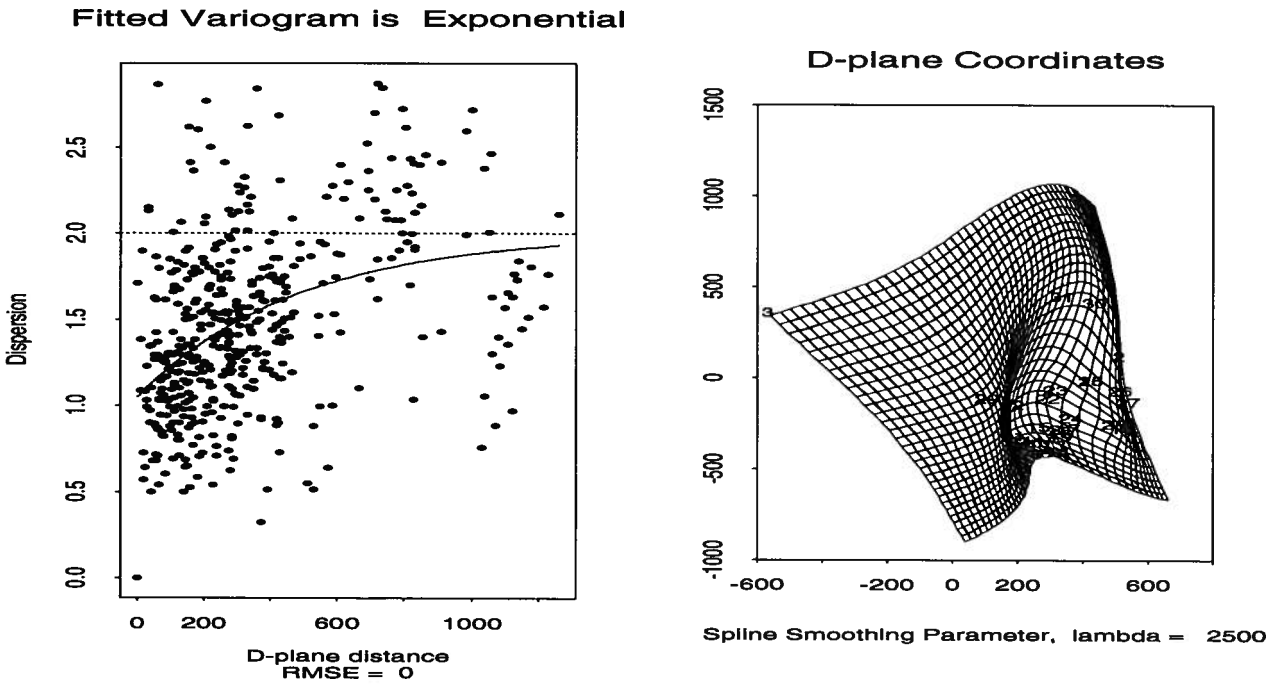


Figure 3.6: A rough checkerboard obtained in the SG step.

Figure 3.7: A smoother checkerboard obtained in the SG step.



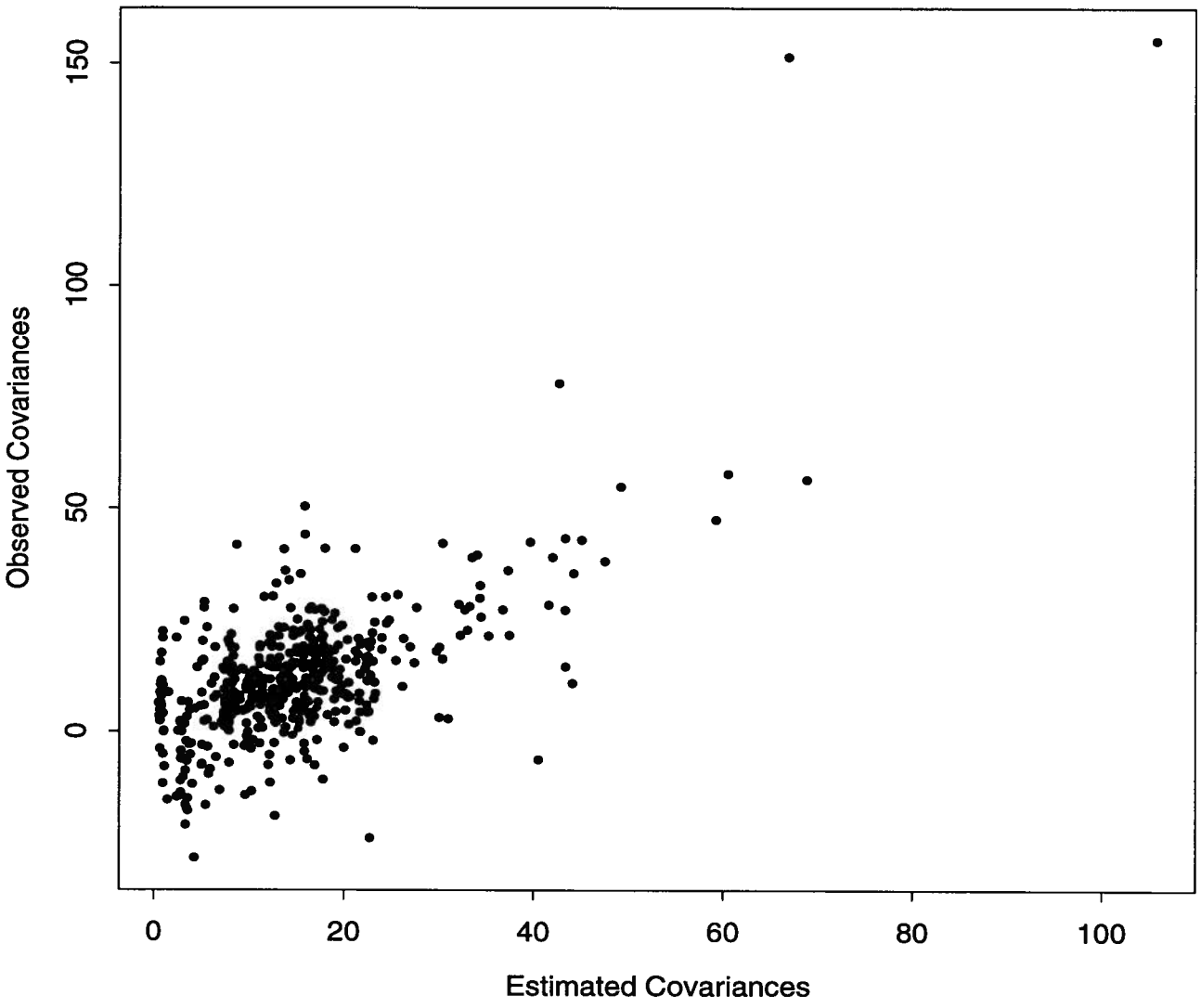


Figure 3.8: Scatter plot of observed covariances vs predicted covariances obtained by the GS approach.

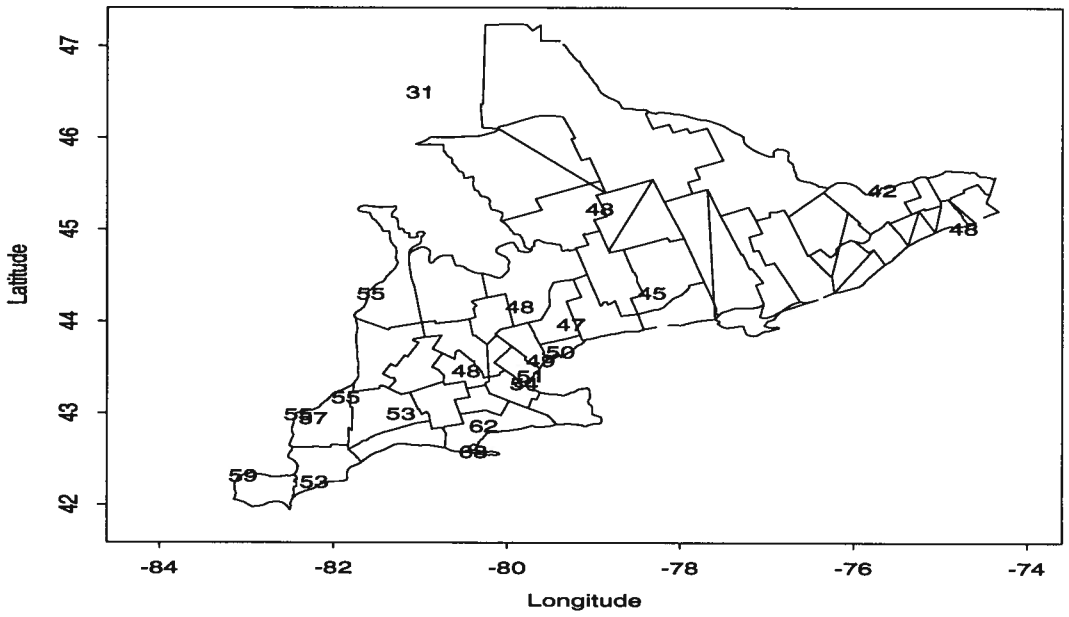
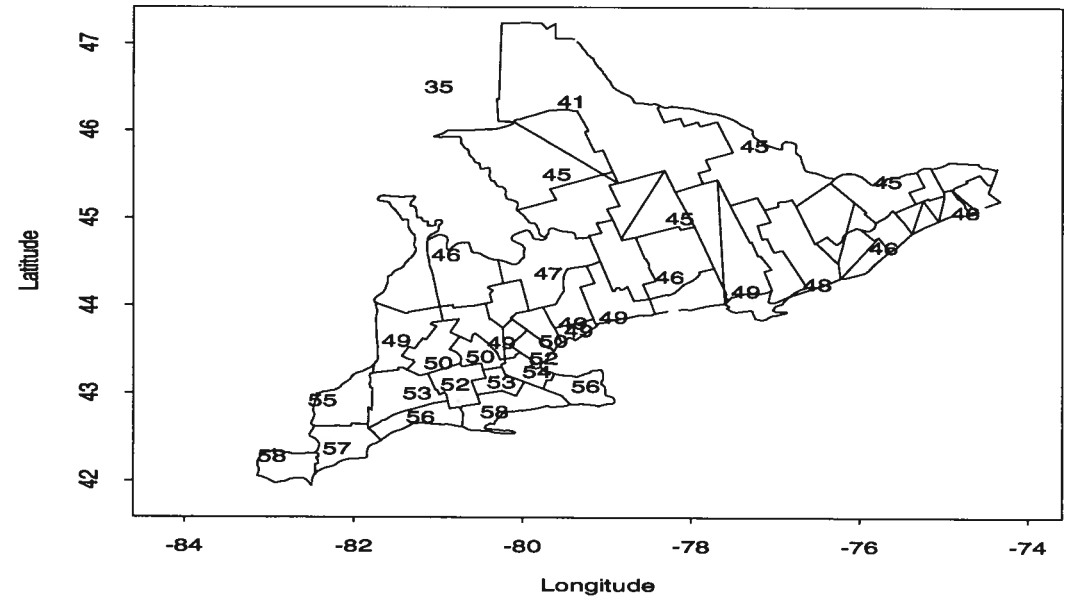
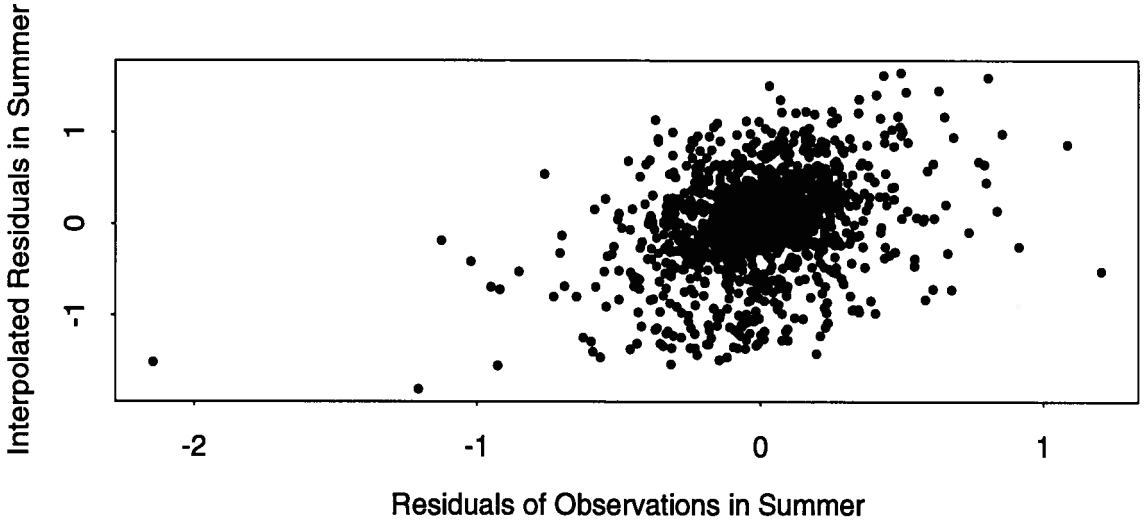


Figure 3.9: Means of monthly levels of O_3 in *ppb*, in summers of 1983 $\sim$ 1988 at gauged sites in Southern Ontario plotted with CSD boundaries.

Figure 3.10: Means of monthly levels of O_3 in *ppb*, in summers of 1983 $\sim$ 1988 at selected sites in Southern Ontario plotted with CSD boundaries.





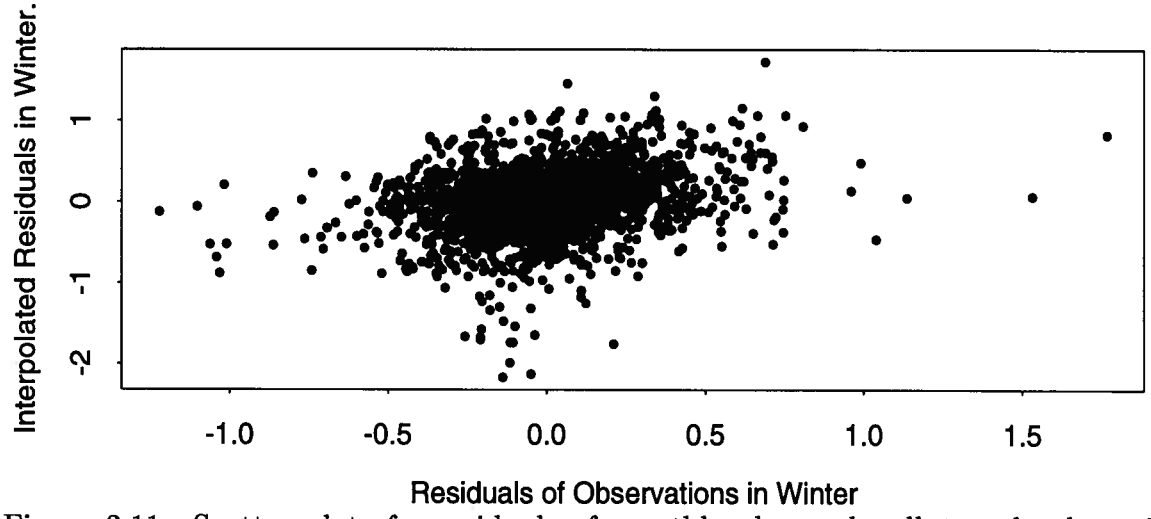


Figure 3.11: Scatter plots for residuals of monthly observed pollutant levels against residuals of interpolated levels at the log-scale in winter and summer respectively, where levels of O_3 are in ppb, SO_2, NO_2 and SO_4 in $\mu g/m^3$.

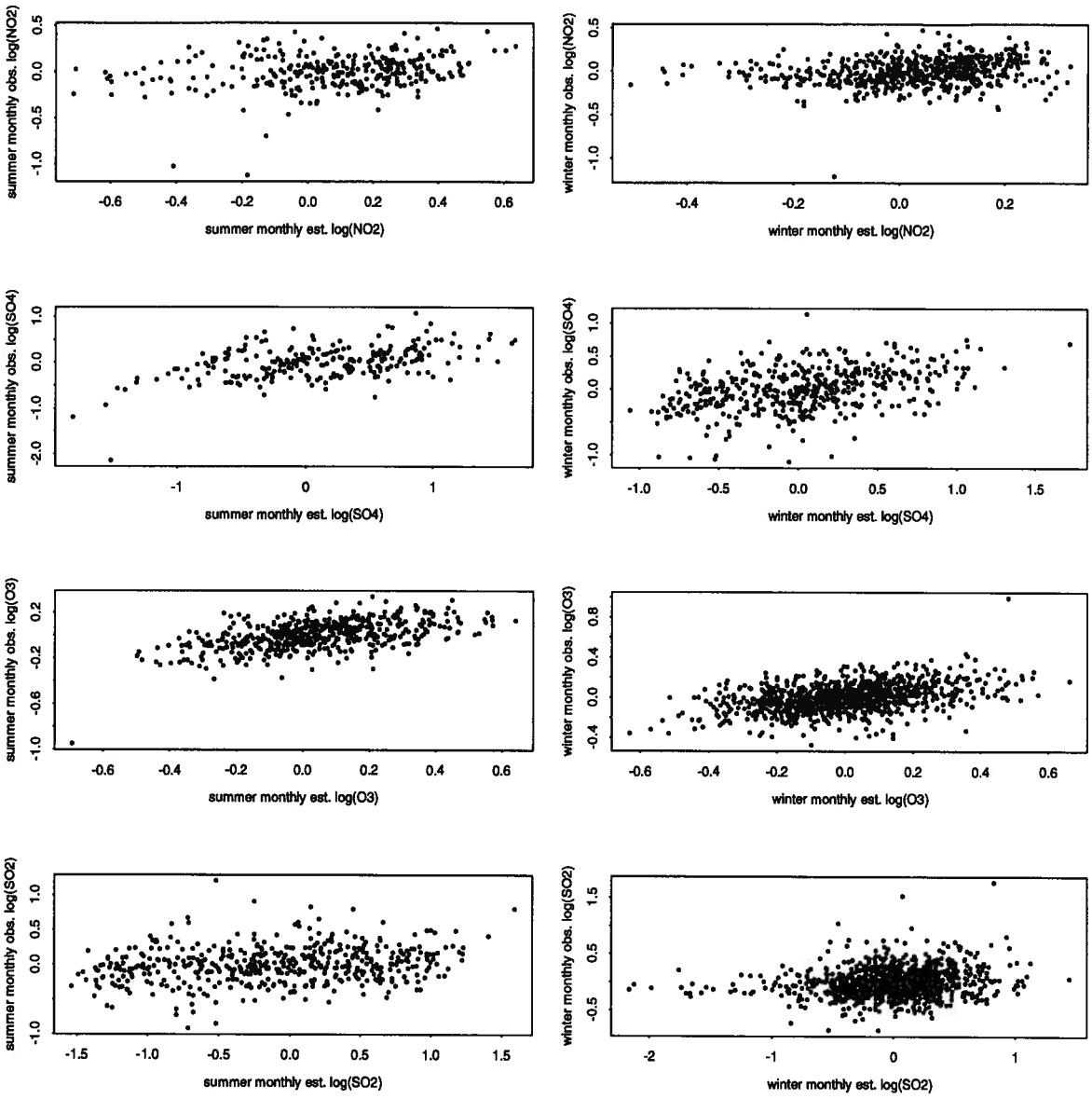


Figure 3.12: Pollutant-wise scatter plots for residuals of monthly observed pollutant levels against residuals of interpolated levels at the log-scale in winter and summer respectively, where levels of O_3 are in *ppb*; SO_2, NO_2 and SO_4 in $\mu g/m^3$.

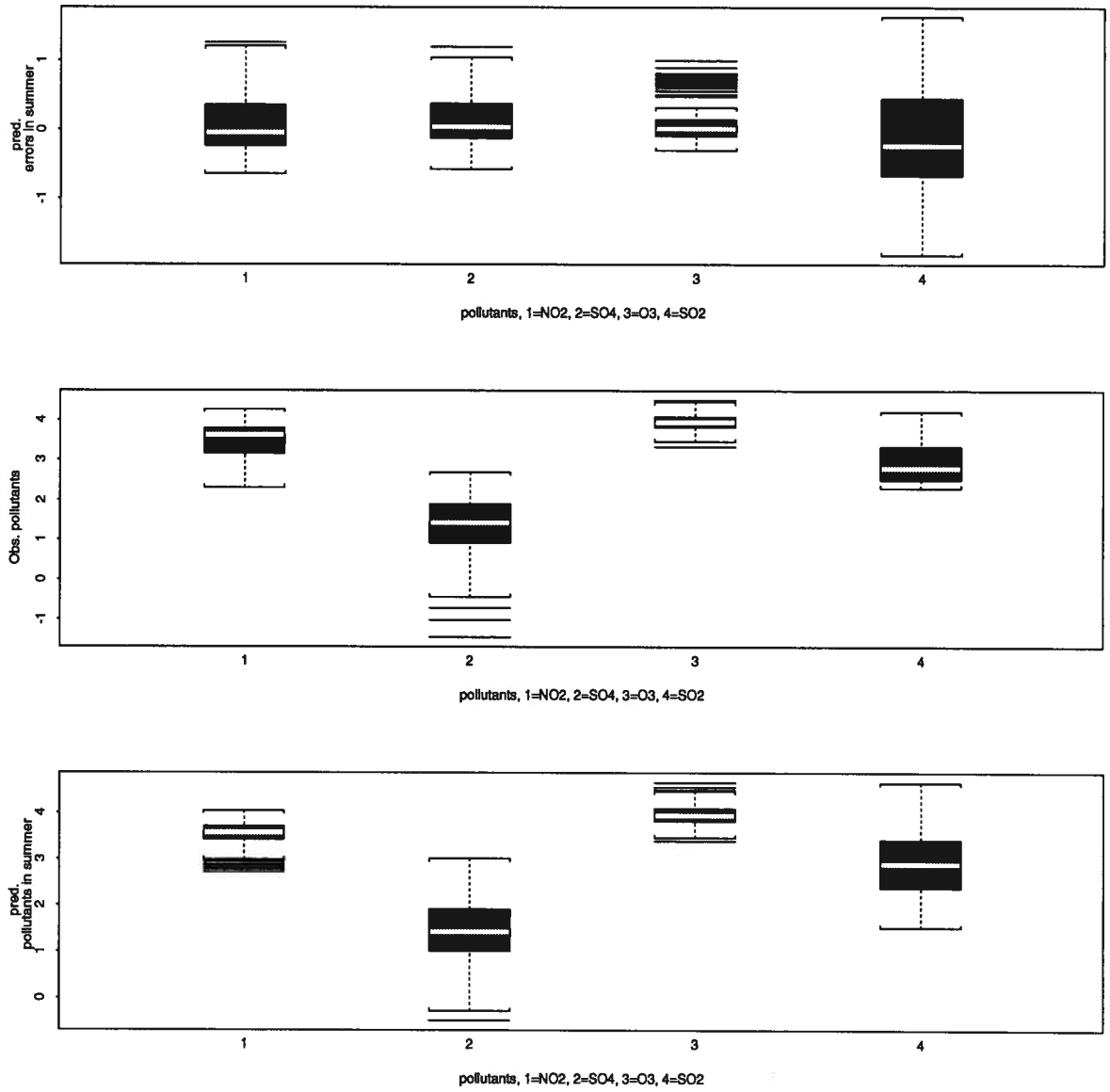


Figure 3.13: Boxplots for predicted, observed and residual levels of log-transformed, monthly concentrations of O_3 in *ppb*, SO_2, NO_2 and SO_4 in $\mu g/m^3$, respectively.

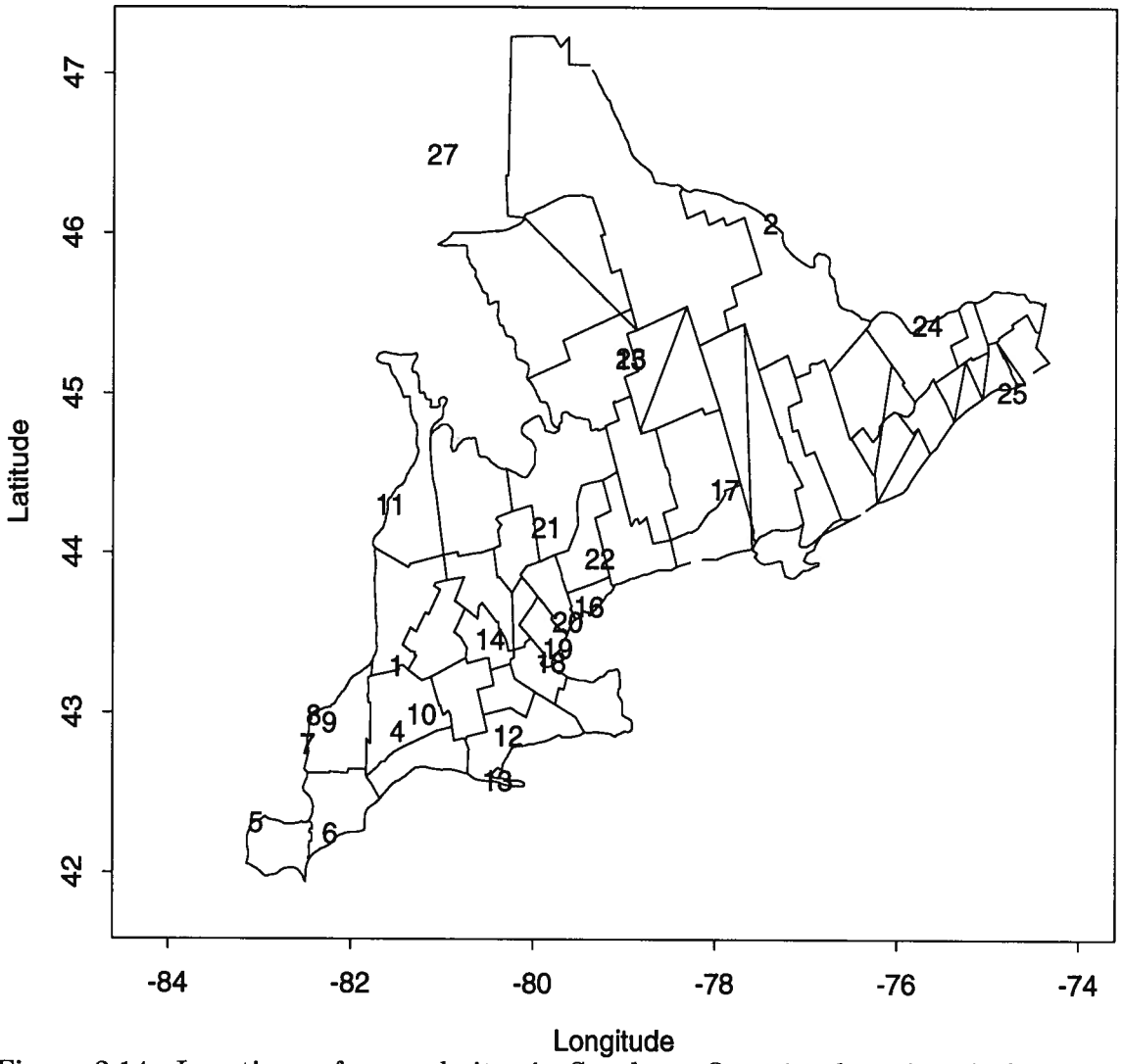


Figure 3.14: Locations of gauged sites in Southern Ontario plotted with Census Subdivision boundaries, where daily pollution levels are observed and Sites 3, 26 (outliers) are not plotted.

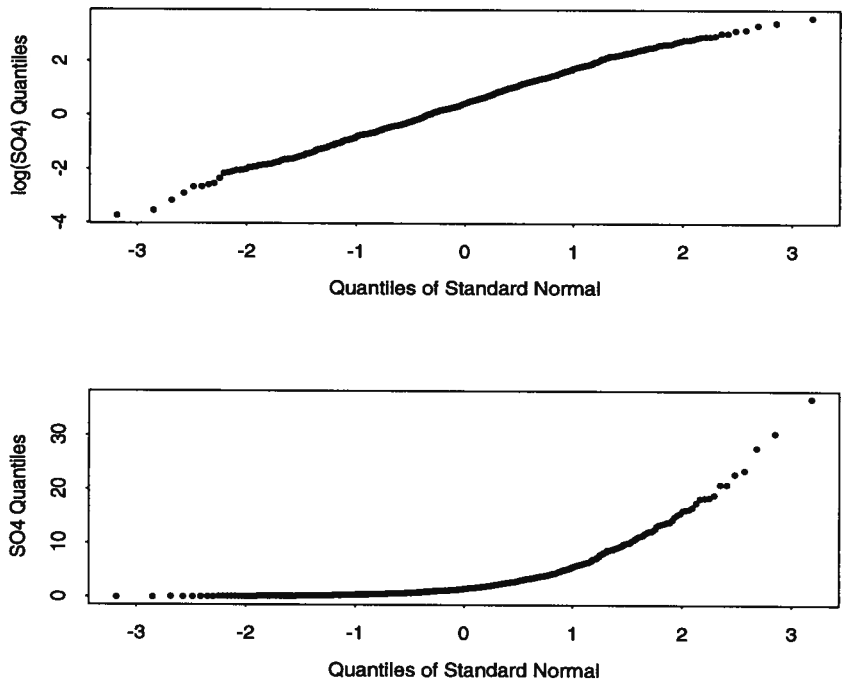


Figure 3.15: Normal quantile-quantile plots for original and log-transformed daily levels of SO_4 in $\mu g/m^3$ at Gauged Site 1.

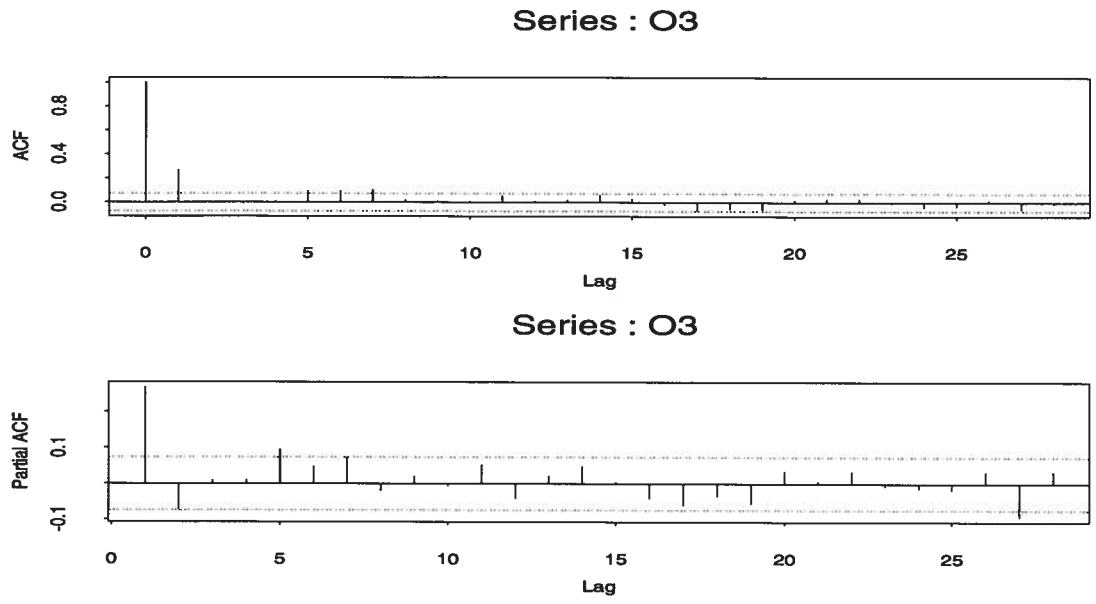


Figure 3.16: Plots for autocorrelation and partial autocorrelation of daily, log-transformed levels of O_3 in *ppb* at Gauged Site 6, before an AR(1) transformation is taken.

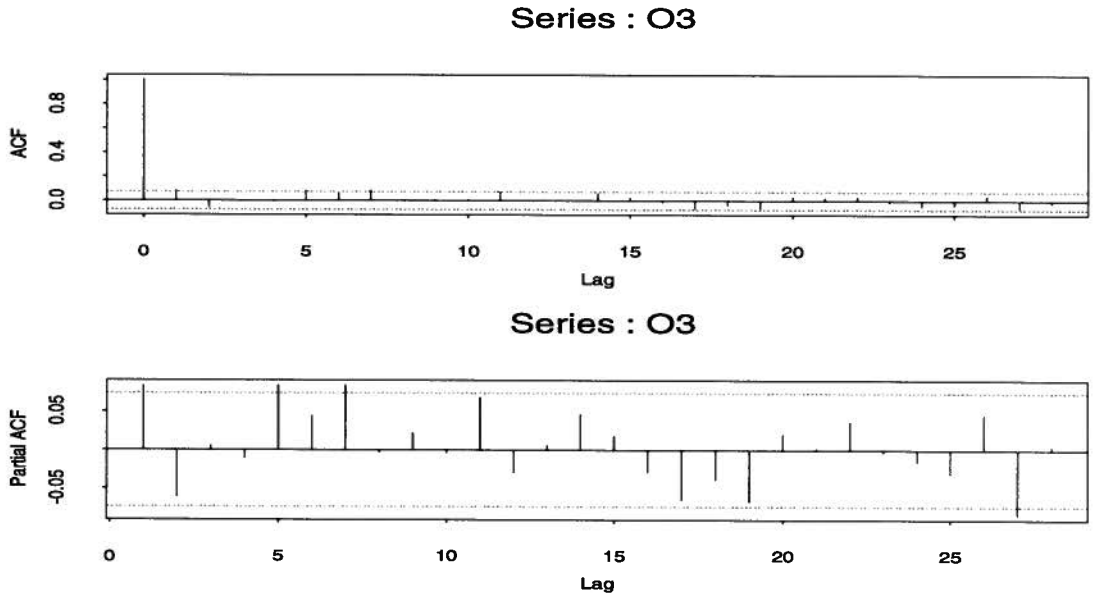


Figure 3.17: Plots for autocorrelation and partial autocorrelation of daily, log-transformed levels of O_3 in *ppb* at Gauged Site 6, after an AR(1) transformation is taken.

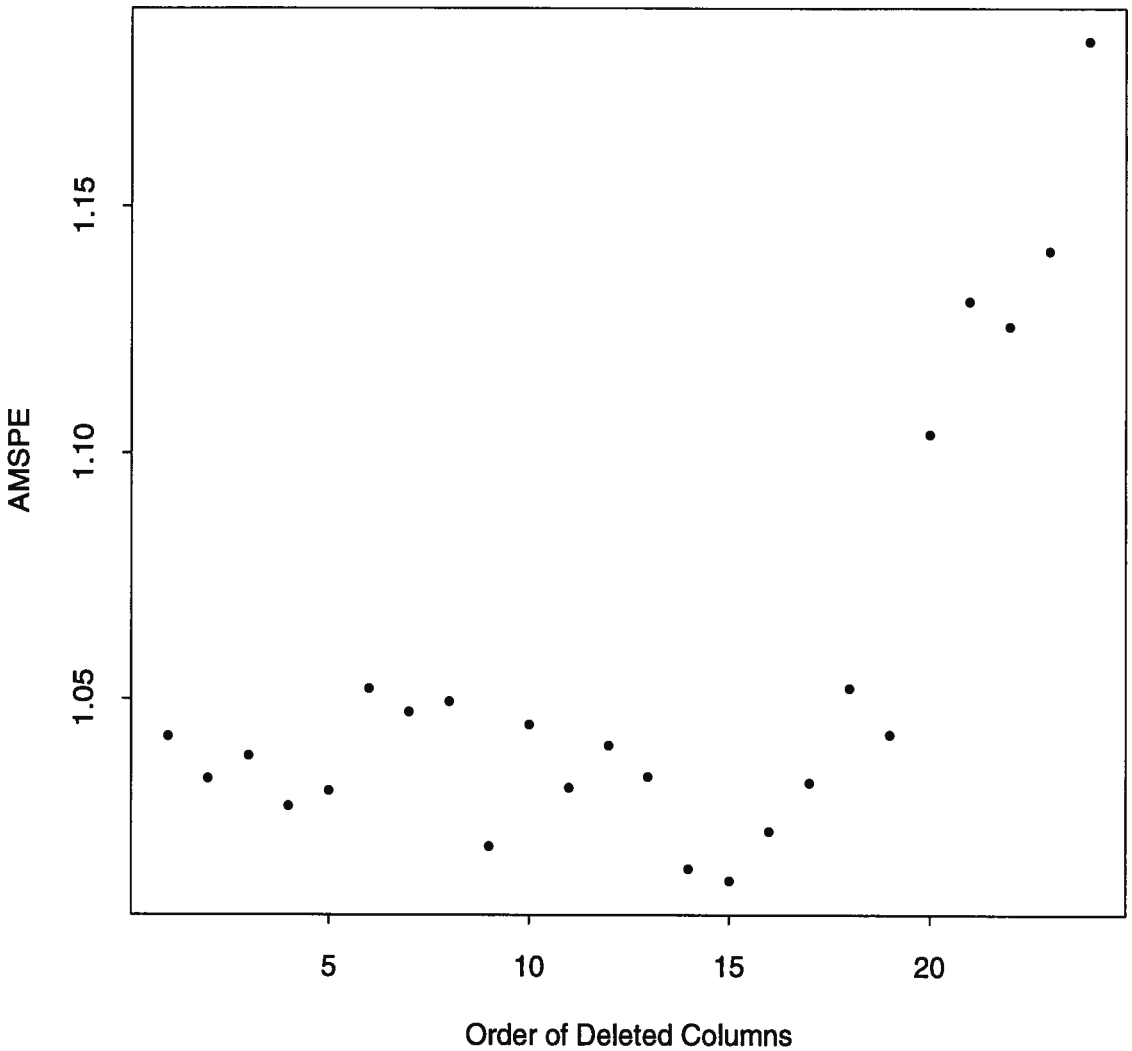


Figure 5.1: Plot for trends in AMSPE.

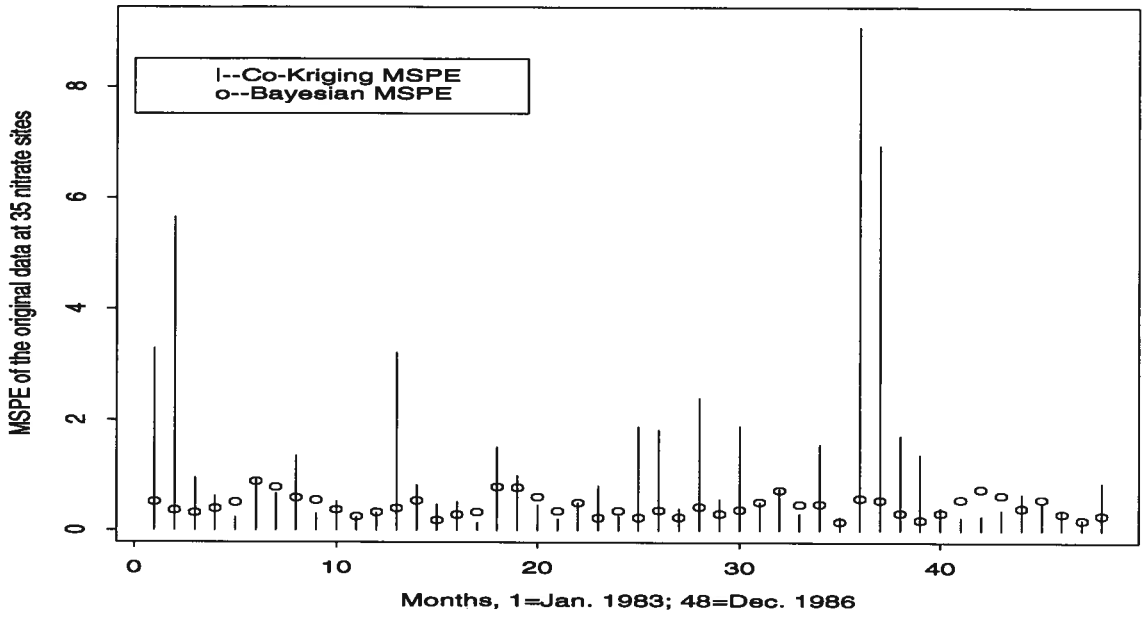
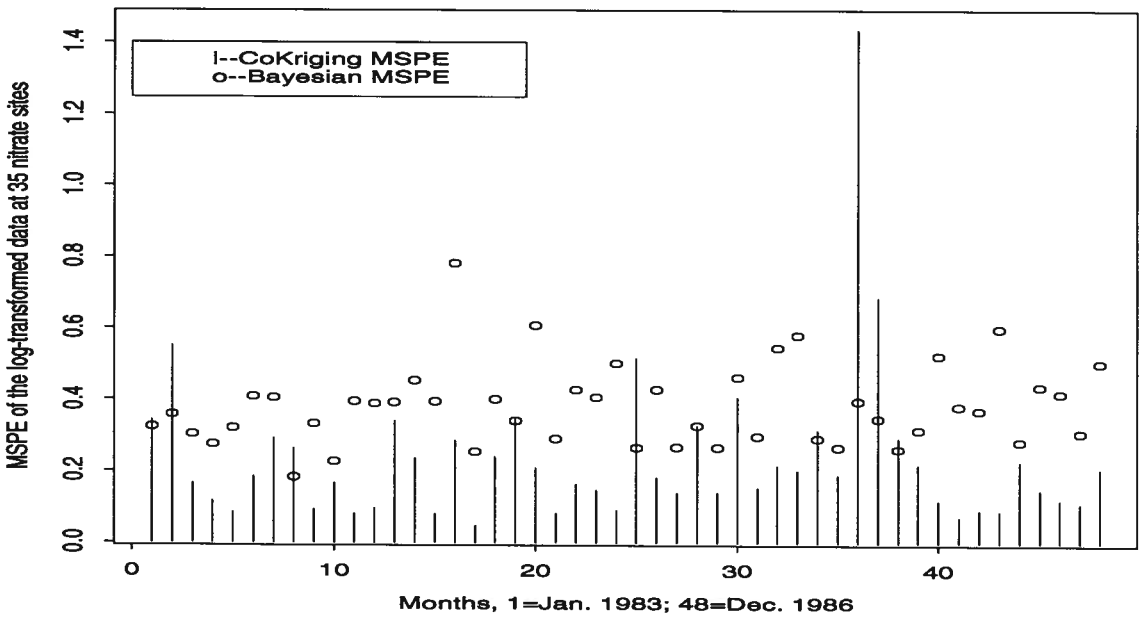


Figure 5.2: MSPEs obtained by Hass's interpolator and the Bayesian interpolator with original acid rain data.

Figure 5.3: MSPEs obtained by Hass's interpolator and the Bayesian interpolator with log-transformed acid rain data.



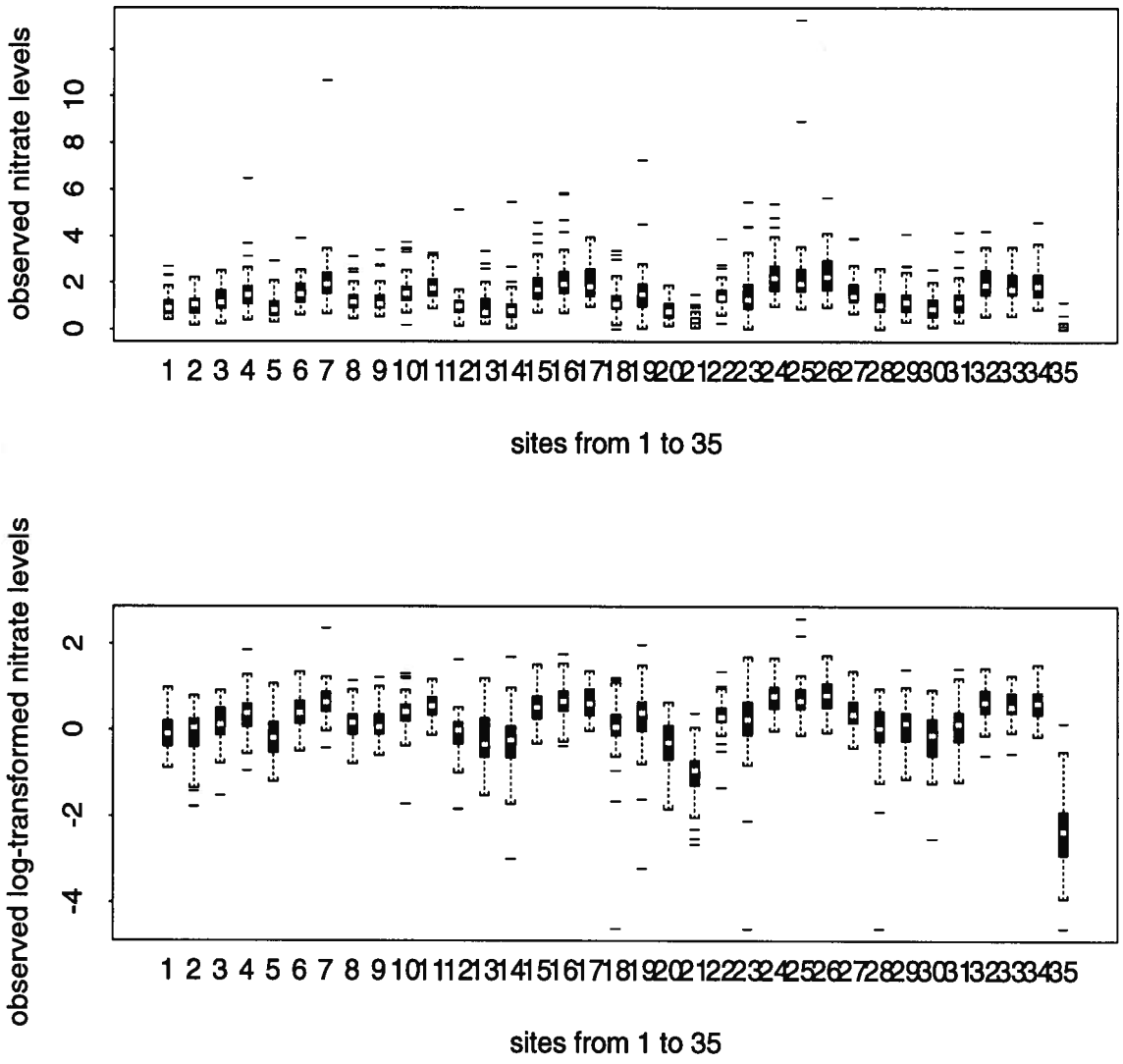


Figure 5.4: Boxplots of observed nitrate levels with/out log-transformation at 35 sites in US.