A modification of the EM algorithm with applications to spatio-temporal modeling

by
Stanislav Kolenikov

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistics and Operations Research.

Chapel Hill
2005

Approved by:

Prof. Richard L. Smith, Advisor

Prof. Kenneth A. Bollen, Reader

Prof. Vidyadhar G. Kulkarni, Reader

Prof. Bahjat Qaqish, Reader

Prof. Zhengyuan Zhu, Reader

ABSTRACT
STANISLAV KOLENIKOV: A modification of the EM algorithm with applications to spatio-temporal modeling.
(Under the direction of Prof. Richard L. Smith.)

This dissertation outlines the use of the maximum likelihood procedures for estimation of the parameters of spatial and spatio-temporal processes when some observations are missing, and reviews the realizations of ML and EM procedures in the presence of missing data when the data are correlated. A version of the EM algorithm is suggested that has a promise of being computationally more efficient due to reduction of the number of matrix inversions, although at a price of the loss of the estimator's consistency and asymptotic efficiency. Corrections that restore unbiasedness of the estimating equations implied by the EM algorithm are proposed. Asymptotic properties (consistency and normality) of the resulting estimators are established. Applications of the new procedure are considered: an analytically tractable case of AR(1) process, and an application to real data on $PM_{2.5}$ measurements.

ACKNOWLEDGEMENTS

I am indebted to my advisor Richard L. Smith for the support he provided in this research. I am also grateful to Dave Holland who kindly provided the data for the analysis in Chapter 4, and to the committee members (Prof. Kenneth Bollen, Prof. Zhengyuan Zhu, Prof. Bahjat Qaqish, Prof. Vidyadhar Kulkarni) for helpful discussions that led the substantial improvements in the thesis. Paul Dudenhoffer provided useful editorial tips and suggestions. The work was partially supported by the EPA Coop Agreement CR-827737-01-0.

I am thankful to my family: my wife, Oksana Loginova, Ph.D., and my children, Timosha and Zlata, for their support and tolerance over all those years of graduate studies.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

This dissertation outlines the methods of dealing with the missing data in the context of spatially correlated environmental monitoring network data.

In their recent paper, Smith, Kolenikov & Cox (2003) analyzed a spatio-temporal data set that featured repeated measurement of spatially correlated data. For each week $t = 1, \ldots, T$, there were up to K measurements available at certain fixed locations of the monitors. The log likelihood of such model (assuming normality of the response variable) can be written down as

$$\ln L(\theta, \beta; \mathbf{y}, X) \sim -\frac{1}{2}\Big\{\sum_{t=1}^{T} \ln|\Sigma_t(\theta)| + \mathrm{tr}\big[(\mathbf{y}_t - X_t\beta_t)(\mathbf{y}_t - X_t\beta_t)^T\Sigma_t(\theta)^{-1}\big]\Big\} \qquad (1.1)$$

where X is the design matrix, the vector β represents the trends in time, space, and other covariates such as land use, and θ is the set of parameters for the geostatistical model of spatial covariance. The subindex t denotes a possible dependence of the dimensionality of the vectors and matrices on time t, as long as some data are missing. The independence over time is justified by the analysis of the residual correlation that shows no significant dependencies.

The likelihood (1.1) can be maximized explicitly with a nonlinear optimization routine, but it would possibly involve inverting T matrices of rather big size (in Smith et al. (2003), $K = 74$, but more realistic applications may have $K \sim 10^3 - 10^5$) and computing their determinants[1]. An appealing method to reduce those computational costs seems to be the EM algorithm (Dempster, Laird & Rubin 1977, McLachlan &

[1] Those two operations may be performed jointly thus reducing computational burden if appropriate matrix inversion methods relying on either spectral decomposition or Cholesky decomposition of a matrix (Demmel 1997), are used.

Krishnan 1997). This is an iterative procedure of Bayesian origin that increases the likelihood with each iteration by taking the conditional expectation of the missing data given the observed data and the current estimate of the parameters, and then maximizing the likelihood by the standard complete data methods. Various modification of the algorithm to simplify computations have been proposed. See Section 2.3 for more details.

The modification that is especially useful in our context involves splitting the maximization step of the EM algorithm into maximization over the trend parameter β subspace and the covariance parameter θ subspace (a version of the EM known as *expectation-conditional maximization*, ECM). Also, instead of maximizing the likelihood at each M step, the algorithm may aim at just increasing it in a single step (*generalized EM algorithm*, GEM). So the algorithm steps used in Smith et al. (2003) are as follows:

1. Initialize the trend parameters β by OLS over the available cases;

2. Initialize the covariance parameters θ by some reasonable guesses;

3. (E-step, h-th iteration) Compute $y_{it}^{(h)} = y_{it}$ if available, $\mathbf{x}_{it}^T \beta^{(h)}$ otherwise;

4. (E-step, h-th iteration) Compute the regression residuals $e_t^{(h)}$;

5. (E-step, h-th iteration) Compute the conditional expectation of the sufficient statistic
$$\mathbb{E} \sum_t \Big[e_t^{(h)} \big(e_t^{(h)}\big)^T | \theta^{(h)} \Big]; \tag{1.2}$$
or an approximation to it. Thus at the completion of the E-step, we have something a "current prediction" of the second term in (1.1): the cross-products in (1.1) are replaced by their (approximate) conditional expectations given by (1.2).

6. (M-step, h-th iteration, part 1) Maximize, by a nonlinear maximization routine, the log likelihood (1.1) with respect to the covariance parameters θ;

7. (M-step, h-th iteration, part 2) Run weighted least squares regression of $y^{(h)}$ with the weighting matrix $\Sigma(\theta^{(h)})$ to maximize the likelihood over the trend / regression parameter subspace. (If the weighting matrix were the true covariance matrix $\Sigma = \mathbb{C}\text{ov}[\mathbf{y}]$, then this step will become a GLS regression.)

8. Declare convergence according to a suitable criteria, or reiterate to step 3.

A computational difficulty remains in the above procedure at steps 4–5 to estimate the residuals and compute the sufficient statistic of the data given the observed values and the current parameter estimates. The exact implementation of the EM algorithm would require universal kriging (see Section 2.1.2) for each time point at the locations of the missing data at step 5 to predict the missing regression residuals, their variances and their covariances that are required for the likelihood. However, this provides no improvement in the computational speed over the classical MLE procedure based on straightforward maximization of (1.1) as this step would require the same matrix inversions separately for each t.

What this dissertation proposes is an alternative method based on the approximate expectations of the residuals at the E-step. An opportunity to save on computations may be to use $e_{it}e_{jt}$ in (i, j)-th position of the t-th term in (1.2) whenever both residuals were available, and use $\sigma_{ij}(\theta^{(h)})$ otherwise. In other words, the conditional expectations are replaced by unconditional, or marginal, expected values. A question must then be asked, by how much the above approximation to the conditional expectation is biased, can this bias be eliminated, what is the efficiency of the implied estimating equations, and what kind of other problems the procedure may lead to. This question is probably more pertinent for the covariance parameter estimates as long as the trend parameters estimates will be unbiased and consistent for any weighting matrix in the weighted least squares estimator, and asymptotically independent of the variance parameters.

The remainder of the thesis is organized as follows. Chapter 2 gives a basic introduction to the two prime themes of the proposal, the spatial and spatio-temporal models, and the missing data models. In particular, section 2.1 reviews the main models used in geostatistical research, while the sections 2.2 and 2.3 describe the basic approaches to the missing data analysis. Chapter 3 analyzes the properties of the proposed modification to the EM algorithm when the correlation structure is simple enough to lend itself to an analytical solution. Namely, we use AR(1) process to analyze the behavior of the estimates. Further, the application of the algorithm to the real data set (an abridged version of Smith et al. (2003)) is given in Chapter 4. The general treatment of the dissociated processes is given in Chapter 5. Finally, the possible directions for future research are suggested in Chapter 6. Certain technical results necessary for Chapter 5 are given in the Appendices.

Chapter 2

Literature review

This research attempts to find computationally efficient estimating equations for the spatio-temporal models, or repeated spatially correlated measurements. We shall firstly review the ways to model spatial correlations for a single time instance in Section 2.1, then proceed to the incorporation of the temporal component in Section 2.1.4. As long as the real data sets have missing observations, the methods that work nicely on the full data sets start running into problems when some portions of the data are missing, as discussed in Section 2.2. One of the most natural ways to take the missing data into account is to use the EM algorithm that was designed specifically for those purposes. A general introduction to the EM algorithm will be given in Section 2.3, with a focus on the repeated measurement and spatio-temporal models.

2.1 Geostatistical models

The models of covariance that assume some parametric relation between observations with given spatial coordinates are known as *geostatistical* models. A concise introduction to the topic is given in Smith (2003), and extended references are Cressie (1993) and Stein (1999).

2.1.1 Setup

Suppose we have a sample $Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n)$ of measurements taken at locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$ from a spatial process

$$Z(\mathbf{s}), \quad \mathbf{s} \in D \tag{2.1}$$

for some domain $D \subset \mathbb{R}^d$. In the discussion of models and applications, we have to deal primarily with $d = 2$, although there is nothing special to two dimensions, and analysis in higher dimensions is also possible.

To be able to estimate the mean and the variance of a linear functional of the process, such as the value in an unsampled location, or an areal mean, we have to assume that the underlying process has well defined means and variances, too:

$$\mu(\mathbf{s}) = \mathbb{E}\, Z(\mathbf{s}), \quad \mathbb{V}\, Z(\mathbf{s}) < \infty \quad \mathbf{s} \in D \tag{2.2}$$

The process is said to be *(strictly) stationary* if $\forall \mathbf{h} \in \mathbb{R}^d$ and $\forall k$, $\mathbf{s}_1, \ldots, \mathbf{s}_k \in D$ such that $\mathbf{s}_1 + \mathbf{h}, \ldots, \mathbf{s}_k + \mathbf{h} \in D$, the distributions of the original and shifted data are the same:

$$Z(\mathbf{s}_1, \ldots, \mathbf{s}_k) \stackrel{\mathcal{D}}{=} Z(\mathbf{s}_1 + \mathbf{h}, \ldots, \mathbf{s}_k + \mathbf{h}) \tag{2.3}$$

The process is said to be *second-order stationary* if $\mu(\mathbf{s}) = \mu\, \forall \mathbf{s} \in D$, and

$$\forall \mathbf{s}_1, \mathbf{s}_2 \in D\,, \mathbb{C}\mathrm{ov}\big[Z(\mathbf{s}_1), Z(\mathbf{s}_2)\big] = C(\mathbf{s}_1 - \mathbf{s}_2) \tag{2.4}$$

for some function $C(\cdot) : \mathbb{R}^d \to \mathbb{R}$.

The process said to be Gaussian if any finite sample from it has a multivariate normal distribution. For Gaussian processes, the two definitions of stationarity are equivalent.

2.1.2 Spatial prediction: kriging

Suppose one wants to obtain a point estimate and the standard error of that estimate at a new location $\mathbf{s}_0$ not available in the data set. Suppose we can model the spatial trend as

$$Z(\mathbf{s}_i) = X(\mathbf{s}_i)\beta + \eta(\mathbf{s}_i), \quad \eta \sim N(0, \Sigma),\, i = 1, \ldots, n \tag{2.5}$$

so that the mean, or the fixed effect, or the trend of the process is $\mu = X\beta$, η is the realization of a spatially correlated noise in given locations, and $\Sigma = \Sigma(\theta)$ is the known spatial covariance matrix of the observed elements of the spatial process parameterized by a low dimension vector θ. (In the examples below, θ has two to four components.) The new observation is supposed to also follow the model

$$Z(\mathbf{s}_0) = x_0\beta + \eta_0 \tag{2.6}$$

where η_0 comes from the same field, and its covariance with the observed data can be found as

$$\mathbb{E}\,\eta_0\eta = \tau \tag{2.7}$$

It can be shown (Cressie 1993, Smith 2003) that the best linear unbiased predictor (BLUP) for $Z(\mathbf{s}_0)$ can be found as

$$\hat{z}_0 = x_0^T\hat{\beta} + \tau^T\Sigma^{-1}(Z - X^T\hat{\beta}), \tag{2.8}$$

$$\hat{\beta} = \hat{\beta}_{GLS} = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Z \tag{2.9}$$

The first term of (2.8), $x_0^T\hat{\beta}$, is the prediction from the linear regression part, and the other term is capturing the spatial correlation of residuals η. 01 The mean squared prediction error (assuming $\Sigma(\theta)$ is known) is given by

$$\begin{aligned}\mathrm{MSPE}[\hat{z}_0] = \mathbb{V}[\hat{z}_0 - z_0] = \sigma_0^2 + \\ +(x_0 - X^T\Sigma^{-1}\tau)^T(X^T\Sigma^{-1}X)^{-1}(x_0 - X^T\Sigma^{-1}\tau) - \tau^T\Sigma^{-1}\tau\end{aligned} \tag{2.10}$$

where $\sigma_0^2 = \mathbb{V}[\eta_0]$. The first two terms represent the standard formulae for the prediction variance in a linear regression, and the terms involving τ show the reduction of the variance due to the information used from the spatially correlated measurements.

Those formulae are known as the *universal kriging*. A simpler version of the *ordinary kriging* is obtained when no regressors are present, so that the trend part of the model is simply

$$X\beta = \mu\mathbb{1} \tag{2.11}$$

where $\mathbb{1} = (1, \ldots, 1)^T$. The name of the method comes from the early works by Krige, a mining engineer, one of the founders of geostatistical models.

If the parameters θ of the spatial covariance process are to be estimated, the equation (2.10) is estimating only a part of the total variance that Cressie (1993) calls probabilistic prediction error, and that is related to the variability only in the new observation. The other part is the statistical prediction error, and it is related to the sampling variability of the estimates. The appropriate corrections are difficult to come by in the analytical frequentist framework, and Bayesian models have a greater promise in this respect.

A BLUP of an areal average can also be obtained in a similar way as a linear combination of the observed data along with the standard error. See Smith (2003).

2.1.3 Variograms and semivariograms

The dependence in a spatial field is usually characterized by a *variogram*. Assuming for transparency $\mu(\mathbf{s}) = 0$ (which we shall do later anyway as we shall be modelling the residual covariance in a regression model), denote

$$\mathbb{V}\big[Z(\mathbf{s}_1) - Z(\mathbf{s}_2)\big] = 2\gamma(\mathbf{s}_1 - \mathbf{s}_2) \tag{2.12}$$

The function $\gamma(\cdot)$ is referred to as *semivariogram*, and $2\gamma(\cdot)$, as *variogram*. If the process admits such representation, it is called *intrinsically stationary*, which is a weaker concept than stationarity. A 2D Brownian shield is not stationary, but intrinsically stationary with $\gamma(\mathbf{h}) \propto \|\mathbf{h}\|$. A process is called *isotropic* if the variogram only depends on the Euclidean distance between the two points:

$$\gamma(\mathbf{h}) = \gamma_0(\|\mathbf{h}\|), \quad \gamma : \mathbb{R}^d \mapsto \mathbb{R}, \gamma_0 : \mathbb{R} \mapsto \mathbb{R} \tag{2.13}$$

The typical shape of the variogram function of a stationary isotropic process has three main features shown on Fig. 2.1. The *sill* is the asymptotic value of $\gamma(\mathbf{h})$ as $\|\mathbf{h}\| \to \infty$, if such a value exists. The *nugget* models the jump of $\gamma(\cdot)$ in the vicinity of zero, and is most often attributed to the white noise measurement error, and the value of the nugget is the variance of this error. The *range* is the distance at which the

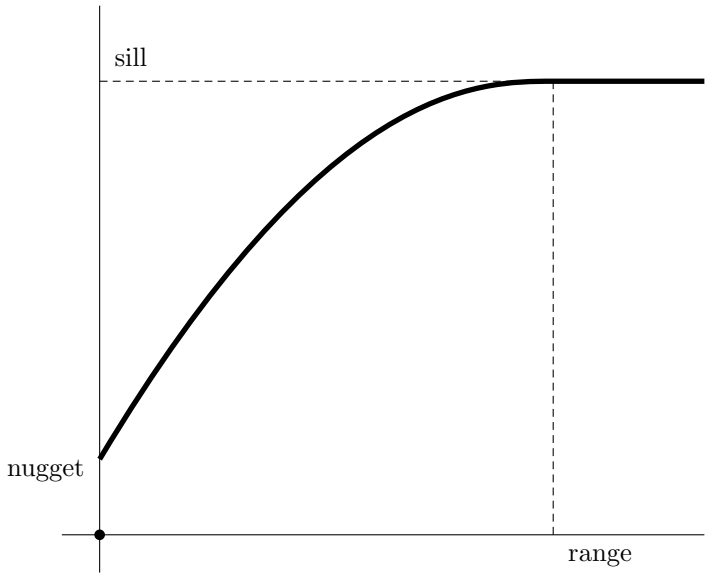


Figure 2.1: A stylized variogram function.

spatial correlation drops to zero, and the variogram reaches its maximum value (the sill). For some models, the sill is never achieved, but still one can talk about the range as the characteristic distance at which most of the action in the variogram occurs.

A special concern in variogram modelling is making sure the implied covariance function is *non-negative definite*:

$$\begin{gathered}\forall k,\, \forall \mathbf{s}_1, \ldots, \mathbf{s}_k \in D,\, \forall a_1, \ldots, a_k \in \mathbb{R}, \\ \mathbb{V}\big[\sum_i a_i Z(\mathbf{s}_i)\big] = \sum_i \sum_j a_i a_j\, \mathbb{C}\mathrm{ov}\big[Z(\mathbf{s}_i), Z(\mathbf{s}_j)\big] \geq 0 \end{gathered} \tag{2.14}$$

and the corresponding variograms are *non-positive definite*:

$$\forall k,\, \forall \mathbf{s}_1, \ldots, \mathbf{s}_k \in D,\, \forall a_1, \ldots, a_k \in \mathbb{R}, \sum_i a_i = 0 \Rightarrow \quad \sum_i \sum_j a_i a_j \gamma(\mathbf{s}_i - \mathbf{s}_j) \leq 0 \tag{2.15}$$

If this property is not satisfied, the variances of the spatial predictions (2.10) may become negative.

There are a number of analytical forms for variograms of stationary isotropic processes that guarantee sign definiteness. The most popular examples are given in Table 2.1.3 adopted from Smith (2003). The power law can be used to model a non-stationary, but intrinsically stationary, field, with a special case of $\lambda = 1$ corresponding to the linear form of the variogram (such as in Brownian motion). The spherical model is only applicable for $d \leq 3$. It fails positive definiteness in higher dimensions. The exponential power variogram has special cases $p = 1$, exponential form; and $p = 2$, Gaussian form. The Gaussian form is so called because of the similarity of the correlation function to a normal density, and does not imply that the process itself is Gaussian. The wave form is not monotonic and may be useful if the observations further apart in the space may have larger correlations than those closer in space.

A rather special class of specifications is Matérn class derived from the bivariate spectral density of the process:

$$f(\omega) = \frac{1}{(1 + \|\omega\|^2/\theta_1^2)^{-\theta_2 - 1}} \tag{2.16}$$

It is expressed in terms of covariances rather than a variogram:

$$C_0(t) = \frac{1}{2^{\theta_2 - 1}} \Gamma(\theta_2) \left(\frac{2\sqrt{\theta_2} t}{\theta_1} \right)^{\theta_2} \mathcal{K}_{\theta_2} \left(\frac{2\sqrt{\theta_2} t}{\theta_1} \right) \tag{2.17}$$

	$\gamma_0(t) =$	Notes
Power law	$c_0 + c_1 t^\lambda$	$c_1 > 0$, non-stationary, $0 < \lambda \leq 2$
Spherical ($d < 4$)	$c_0 + c_1 \left[\frac{3}{2}\frac{t}{R} - \frac{1}{2}\left(\frac{t}{R}\right)^3\right]$, $t < R$ $c_0 + c_1$, $t > R$	$c_1 > 0$ $c_0 + c_1$ is sill, R is range
Exponential-power	$c_0 + c_1(1 - e^{\lvert t/R\rvert^p})$	$c_1 > 0$; $c_0 + c_1$ is sill; $R > 0$ is range $0 < p \leq 2$ is shape parameter $p = 1$ — exponential, $p = 2$ — Gaussian
Rational quadratic	$c_0 + c_1 \frac{t^2}{1+t^2/R^2}$	$c_1 > 0$, $c_0 + c_1$ is sill $R > 0$ is range
Wave	$c_0 + c_1(1 - \frac{R}{t}\sin\frac{t}{R})$	$c_1 > 0$

Table 2.1: Parametric forms for variograms. For all specifications, $\gamma_0(0) = 0$, so $c_0 > 0$ is nugget.

where θ_1 is the scale parameter, $0 < \theta_2 < \infty$ is the shape parameter (the case $\theta \to \infty$ corresponds to the Gaussian variogram function), and $\mathcal{K}_\nu(z)$ is the modified Bessel function of the third kind of order ν (Abramovitz & Stegun 1964). It is a solution to the differential equation

$$z^2\frac{d^2w}{dz^2} + 2z\frac{dw}{dz} - [z^2 + \nu(\nu - 1)]w = 0 \tag{2.18}$$

The examples of all those functions are given in Fig. 2.2 (reproduced, with permission, from Smith (2003)).

The parameters of the variogram models can be estimated by (versions of) the least squares method fitting the parametric model to the empirical variogram, by the maximum likelihood, or by restricted maximum likelihood. In this paper, we shall concentrate on the likelihood-based methods as long as they allow explicit treatment of the missing data.

2.1.4 Spatio-temporal models

There are several approaches to incorporate the time dependence in spatio-temporal models.

The first big strand of literature starts off with the geostatistical models and allow the dependence on time. In the simplest form (which is what we concentrate further on in the proposal), such a model is a set of uncorrelated over time repeated measurements

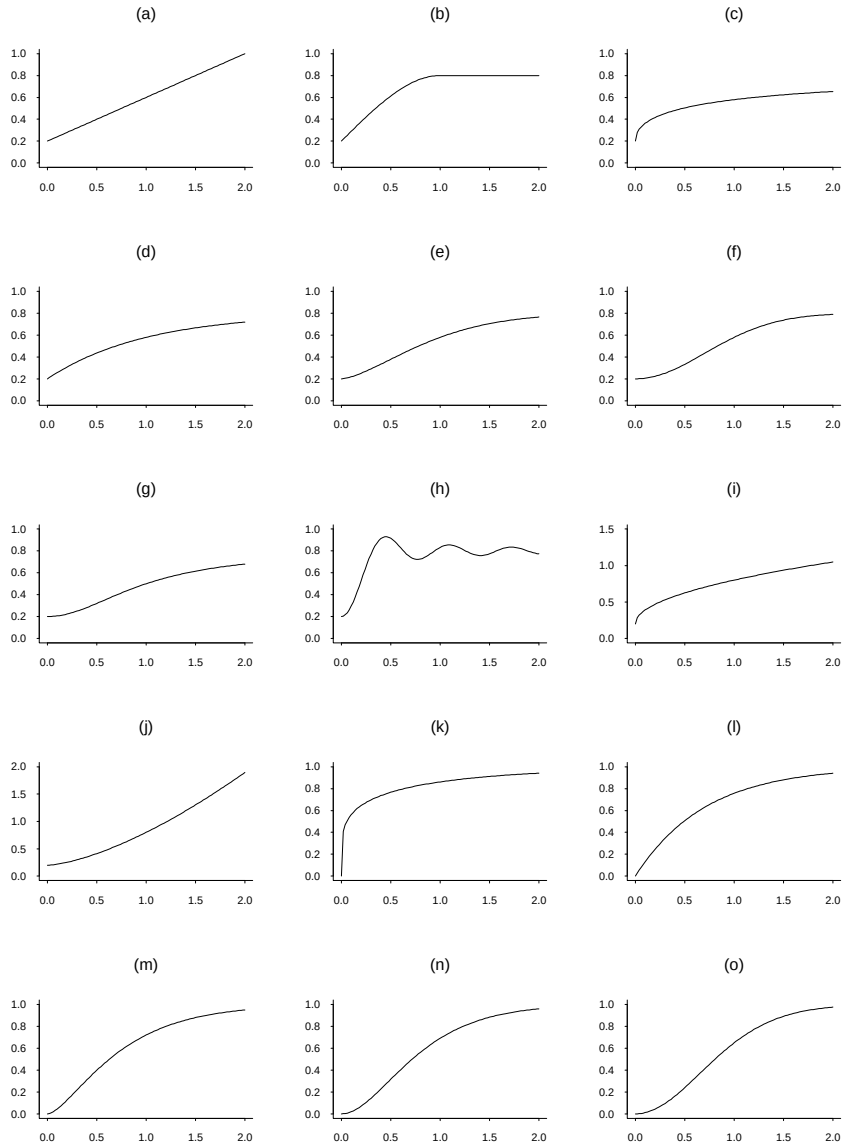


Figure 2.2: Examples of isotropic variogram functions: (a) linear; (b) spherical; (c) exponential-power, $p = 0.5$; (d) exponential (exponential-power with $p = 1$); (e) exponential-power, $p = 1.5$; (f) Gaussian (exponential-power with $p = 2$; Matérn with $\theta_2 = \infty$); (g) rational quadratic; (h) wave; (i) power law, $\lambda = 0.5$; (j) power law, $\lambda = 1.5$; (k)-(o) Matérn function with $\theta_2 = 0.1, 0.5, 1, 2, 10$. Courtesy of Smith (2003).

Z_{st} where the subindex s enumerates locations and t, time:

$$Z_t \sim \text{i.i.d. } N(\mu_t, \Sigma(\theta)) \tag{2.19}$$

where the trend μ_{st} may have both a spatial and temporal components. The model is treated in more detail in Section 2.4.

This model is plausible if correlations over time are small. If they are not, a model that accounts for such correlations needs to be built. Such model can be recast in the form of the spatio-temporal field that leads back to the formulation like (2.1), but the domain D will now be a subset of $\mathbb{R}^3$ to encompass both spatial and temporal dependence.

The simplest extension of (2.19) is to assume that despite the temporal correlation, the process is *separable*:

$$\mathbb{C}\text{ov}(Z(\mathbf{s}_1, t_1), Z(\mathbf{s}_2, t_2) = C_s(\mathbf{s}_1 - \mathbf{s}_2)C_t(t_1 - t_2) \tag{2.20}$$

In this case, the covariance matrix has a Kronecker structure that is reasonably easy to deal with (see Appendix B), and the log-likelihood can be straightforwardly maximized over the parameters of the spatial and temporal covariance functions $C_s(\cdot)$ and $C_t(\cdot)$, respectively.

In more interesting (and more realistic) cases, the covariances are non-separable. One approach to model such covariances is to use non-parametric or semi-parametric models (Haas 2002) that estimate the empirical spatio-temporal variogram based on certain concepts of the spatio-temporal neighborhoods, with some resemblance to the kernel density / regression estimates. An alternative approach is to find the appropriate global spatio-temporal variograms that possess all the required properties such as negative definiteness by inverting the appropriately behaved spectral densities (Cressie & Huang 1999, Gneiting 2002, Stein 2002).

Another big strand of literature has Bayesian origin. Those models (sometimes referred to as *physical-statistical models*, and also used in connection with *data assimilation models*) are aimed at combining the data observed from fixed monitors or satellites with the predictions obtained from the weather models based on complex systems of multidimensional partial differential equations describing atmosphere, ocean, and their interaction, and solved by grid methods. The scientifically motivated models provide the prior distributions (or the main parameters of such priors, such as means or modes), and the actual observations serve for Bayesian updating. Thus, regional

maps of the quantities of interest (such as winds, temperature, humidity and other characteristics of the atmosphere) can be obtained. The examples of this approach are Hoar, Milliff, Nychka, Wikle & Berliner (2003) and Wikle, Milliff, Nychka, & Berliner (2001) in atmospheric research, and Wikle (2003) also uses the same approach in an ecological problem.

2.2 Missing data challenges

One of the practical complications arising between the theory and the available data sets is the fact that some of the observations may be missing from the data.

The main principles of the analysis of the data with missing values are laid out in Little & Rubin (2002). Their framework is as follows.

The researcher's interest lies in the model of the form

$$Y_i \sim f(y; X, \theta) \tag{2.21}$$

where X represent explanatory exogenous variables, Y_i are observations on the dependent variable conditionally independent given x, and θ are parameters of interest, such as regression slopes and spatial covariance parameters in our application. Some of the observations on Y_i, however, may be missing, so along with Y_i's that may or may not be observed, the data set contains indicators of missing data

$$Z_i = \begin{cases} 1, & Y_i \text{ is missing} \\ 0, & Y_i \text{ is observed} \end{cases} \tag{2.22}$$

and the mechanism of the missing data is described by a model with parameters Ψ:

$$\mathbb{P}\mathrm{r}[Z_i = 1 | Y, X, \Psi] \tag{2.23}$$

The following typology is suggested. The data are said to be *missing completely at random* (MCAR), if

$$\mathbb{P}\mathrm{r}[Z_i = 1 | Y, X, \Psi] = p(\Psi) \tag{2.24}$$

that is, constant across all observations. The data are said to be *missing at random* (MAR) if the probability may depend on the observed variables:

$$\mathbb{P}\mathrm{r}[Z_i = 1 | Y, X, \Psi] = p(X, \Psi) \tag{2.25}$$

Finally, the data are not missing at random (NMAR), if the probability in (2.23) depends on the missing value Y itself. If a measurement on a pollutant concentration below or above a certain threshold cannot be obtained, then the mechanism is NMAR. If the probability of not getting the measurement depends on the day of the week and the altitude of the monitor, say, but not on Y, then the data is MAR. The missing data mechanism is ignorable (i.e. the straightforward maximum likelihood estimation, as if the data matrix were full, yields consistent and asymptotically efficient estimates for the given sample size, just as the estimation procedure that would model the missing data mechanism) if the mechanism is MAR, and the parameters of the primary model and the missing data mechanism are disjoint (i.e. the vectors θ and Ψ do not have any common or functionally dependent components).

The fact that some of the data are missing is formally not much of a trouble for maximum likelihood. If the random field is assumed to be normal and observations are independent over time, the likelihood function for each moment in time t can be written as

$$l(\theta|y_t) = (2\pi)^{-n_t/2}|\Sigma_t(\theta)|^{-1/2}\exp\big[-\frac{1}{2}(y_t - x_t\beta_t)\Sigma_t(\theta)^{-1}(y_t - x_t'\beta_t)\big] \tag{2.26}$$

where the subindex t indicates that observations from different sites are available at different points in time, so the dimensions of the measured $PM_{2.5}$ concentration y_t, the explanatory variables x_t, and the vector of the various trends coefficients β_t, are changing from one week to another, according to the number of available sites. However, computing many determinants and the inverse matrices is likely to be time consuming. For an arbitrary symmetric matrix, the fastest inversion algorithm is through Cholesky decomposition (Demmel 1997) that can also give the determinant as the product of eigenvalues of a matrix. For a matrix of size k, this decomposition requires $O(k^3)$ floating point operations. For larger matrices, a greater computational time arises due to different access times for different types of memory such as registers, processor cache, main memory, and disk memory, with access times differing by a factor of about $10^2 - 10^3$ from one type of memory to the next slowest.

One alternative to the straightforward MLE is the EM algorithm.

2.3 The EM algorithm

The expectation-maximization (EM) algorithm is a procedure to find the local extrema of the likelihood surface that works through incorporation of the missing data into the estimation procedure. The missing data in question may be an "authentic" missing data, i.e., the observation failed to be taken properly; or it may be an artificial construct, such as class labels in one of the important applications of the EM algorithm in k-means clustering / mixture decomposition.

The term was introduced by Dempster et al. (1977) where the history of similar methods was given, and the main convergence results were proved. The contemporary suggested monographs on the topic are Little & Rubin (2002) and McLachlan & Krishnan (1997).

The algorithm delivers the ML estimates for the case when the missing data mechanism is ignorable (see discussion in the previous section). It does so by alternating expectation (E) and maximization (M) steps.

At the expectation step, the conditional expected value of the log likelihood is computed given the observed data Y_{obs} and the current value of the parameter vector $\theta^{(h)}$ that combines both the model of interest (2.21) and the missing data model (2.23). That is, the expectation is taken over the distribution of the missing data Y_{miss}:

$$Q(\theta|\theta^{(h)}, Y_{obs}, X) = \int l(\theta|Y) g(Y_{miss}|Y_{obs}, X, \theta = \theta^{(h)})\, dY_{miss} \tag{2.27}$$

One can think of this step as of a sort of imputation step, although imputation usually refers to coming up with a number for a missing datum, while the E step of the EM algorithm works on other moments, cross-products, etc. Another expression often used is "to integrate out" the missing values. In fact, if there is a sufficient statistic for the model, then it is enough to compute the expected value of this statistic conditional on the observed values of the variables involved, and on the current parameter values.

At the maximization step, the full likelihood is maximized with respect to the parameters by using the "imputed" missing values or the expected values of the sufficient statistic:

$$\theta^{(h+1)} = \arg\max_{\theta} Q(\theta|\theta^{(h)}, Y_{obs}, X) \tag{2.28}$$

The procedure is iterated until convergence, which may be operationally defined that the successive parameter values do not change much, or the likelihood does not change much, or any sensible combination of the two. (As long as the EM algorithm does not

involve the gradients, the closeness of the gradient to zero, which is usually the best convergence criterion, cannot be used.)

It is shown (Dempster et al. 1977, Little & Rubin 2002) that the EM algorithm converges to a stationary point of the log likelihood functions under regularity conditions that seem to be quite general (smoothness of the likelihood function, interchangeability of expectation and differentiation operators, boundedness of the likelihood function from above). For some likelihoods and for some special starting values, however, the EM algorithm can converge to the saddle point of the likelihood, or even to a local minimum if that was a starting point.

A version of the algorithm (*generalized EM*, or GEM; see Section 3.3 of McLachlan & Krishnan (1997)) attempts to increase the likelihood at the M step rather than fully maximize it:

$$Q(\theta^{(h+1)}|\theta^{(h)}, Y_{obs}) \geq Q(\theta^{(h)}|\theta^{(h)}, Y_{obs}) \tag{2.29}$$

As the only requirement for the convergence of the EM algorithm is that the (true) likelihood is increasing (weakly) at each step, the GEM algorithm also converges, with the same qualifications on the likelihood surfaces and starting points that apply to the generic EM algorithm.

Another possible modification of the EM algorithm is to split the parameter space into subspaces $\Theta = (\Theta^{(1)}, \ldots, \Theta^{(R)})$ that fully span the original parameter space, so that the maximization at the M-step is performed separately over each of the subspaces. This version is referred to as *expectation–conditional maximization* (ECM) algorithm (McLachlan & Krishnan 1997, Section 5.2). In our application, the obvious choice is to maximize over the regression slopes subspace (where the maximization is simply a GLS regression), and the spatial covariance subspace. The latter may be further divided into overall scale, nugget, range and shape parameters. Thus a computationally expensive iterative nonlinear maximization is confined to a small number of parameters (two to four) and is likely to be much faster.

One of the weak points of the EM algorithm and its modifications is that they do not produce standard errors in the way Newton-Raphson likelihood maximization procedures do. It is still possible to obtain the standard errors by producing the empirical Jacobian $J(\hat{\theta})$ of the likelihood surface. The procedure is known as *supplemented EM algorithm* (Meng & Rubin 1991, McLachlan & Krishnan 1997, Section 4.5), and has some similarities with Louis (1982). In the presence of the missing data, the information

contained in the data can be described as

$$I(\hat{\theta}, y) = \mathbb{I}_c(\hat{\theta}, y)[I - J(\hat{\theta})], \quad I^{-1}(\hat{\theta}, y) = \mathbb{I}_c^{-1}(\hat{\theta}, y) + \Delta V,$$
$$\Delta V = [I - J(\hat{\theta})]^{-1} J(\hat{\theta}) \mathbb{I}_c^{-1}(\hat{\theta}, y) \tag{2.30}$$

where $\mathbb{I}_c(\cdot)$ is the conditional expected complete-data information matrix. After the convergence of the EM algorithm is achieved, so that $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_p)$ is the ML estimate of the parameters, the probing values $\tilde{\theta}_{(j)} = (\hat{\theta}_1, \ldots, \hat{\theta}_j + h_j, \ldots, \hat{\theta}_p)$ are formed, with a suitable step size h_j that does not lead too far away from the MLE, and one step of the EM algorithm is performed resulting in $\tilde{\theta}^d_{(j)}$. Then the approximate Jacobian is formed by combining the entries of the form

$$J_{ij} = \frac{\tilde{\theta}^d_{i,(j)} - \hat{\theta}_i}{h_j}, \tag{2.31}$$

and the corrected information matrix / asymptotic covariance matrix can be formed through (2.30).

2.4 Repeated measurement and dissociated models

The spatio-temporal model (2.19) has some similarities with the *repeated measures* (longitudinal, panel) models, in which each individual is observed a number of times, and also mixed models that have multiple random and fixed effects. The general form for such models can be given as follows:

$$\mathbf{y}_i = \mathbf{X_i}\alpha + \mathbf{Z}_i \mathbf{b}_i \epsilon_i \tag{2.32}$$

where $\mathbf{X}_i$ and $\mathbf{Z}_i$ are the covariates / design matrices corresponding to the fixed and random effects, respectively; α are the fixed effects coefficients; b_i are the random effects with mean zero and variance matrix $\mathbf{D}$ (so the assumption most often made is that $b_i \sim N(0, \mathbf{D})$); and ϵ_i are individual (measurement) errors independent of all other variables in the model (say $\epsilon_i \sim N(0, \sigma^2 I)$). The parameters α and the free elements of $\mathbf{D}$ are to be estimated.

In the repeated measurement model, the vector of y_i corresponds to all observations made on i-th unit, and the simplest model with a single random effect to the intercept

has a random (univariate) variable e_i that enters all observations on the unit through the unit design vector: $\mathbf{Z}_i = \mathbb{1}$. Laird, Lange & Stram (1987) derive the particulars of the EM algorithm for this model by working out the required sufficient statistics, where the missing data are the random effects $\mathbf{b}_i$. The matrix $\mathbf{D}$ is assumed to be free (besides being non-negative definite). Laird et al. (1987) also discuss a simplifying formulae for matrix inversion when not all of the components of the vector y are observed, so that the likelihood contribution involves (an inverse of) the matrix $\mathbf{D}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i$ (which is a minor of the matrix $\mathbf{D}$). Ibrahim, Chen & Lipsitz (2001) discuss the implementation of the Monte Carlo EM algorithm for the generalized linear mixed models with non-ignorable missing data mechanism. It uses the EM by the method of weights (Ibrahim 1990), and it samples from the conditional distributions of the missing data and unobserved random effects $\mathbf{b}_i$ via the Gibbs sampler. Each observation i with missing data is expanded into a set of completed data, and given a weight $1/m_i$ where m_i is the number of Monte Carlo samples taken for this observation. The complete data M-step can now be taken by the standard complete-data routines, and the process iterated until convergence. Ibrahim et al. (2001) also give the formulae to estimate the observed information matrix based on the derivative of $Q(\cdot|\cdot)$ and the completed likelihood scores (i.e., the information matrix reported by the complete-data routine). The procedure is unlikely to be of practical use in the current environmental applications, however, as in the typical data sets, one has $10^1 - 10^3$ missing observations at each time point, and Ibrahim et al. (2001) note that this can lead to inefficient and computationally unstable Gibbs samplers, as well as highly autocorrelated series of MCMC samples. Besides, the spatial correlation cannot be represented in form (2.32) unless the dimension of $\mathbf{b}_i$ is the same as number of sites, and sampling from the conditional distribution again requires inversion of many covariance matrices for each instance of time.

There are methodological differences between repeated measurement / mixed / panel models described above, and spatio-temporal applications we are interested in. In the latter, potentially all observations may be correlated through both spatial correlations across the units at the same time, and through the time series process for a single location. Sometimes, it turns out to be possible to model the temporal dependence through a trend, so the residual processes at each site are approximately uncorrelated temporally. The spatial correlation is more difficult to deal with, so one ends up with a model that is a "transposed" version of (2.32) where the correlated measurement are those taken at the same time. The index i is then running over time, and there is a spatial correlation built into matrix $\mathbf{D}$. As mentioned in section 2.1, this matrix is assumed to be dependent on a relatively few parameters such as the overall

variance, the range and the shape parameters, unlike the case of the model (2.32) where the matrix D is often assumed to be a free symmetric and non-negative definite matrix. The individual errors e_i can be thought of as the measurement errors contributing to the nugget effect if it is not modelled explicitly in the specification of the semivariogram. Such models can be called *dissociated models*, so as not to create confusion with the repeated measurement and panel models, where the incidental correlations are the ones over time within the same unit, while in our spatio-temporal models, the important correlations are those that occur between different units contemporaneously.

Another methodological aspect is the need for averaging that does not arise very often with the biostatistical longitudinal data. Some of the EPA standards, including those on particulate matter, require averaging over time, although they do not require the use of spatial averaging or computing areal averages. The averaging is understood as computing arithmetic averages of the available observations. The EPA does note however that "the use of averages from single or multiple community-oriented sites is more closely linked to the underlying health effects information, which relates area wide health statistics to averaged measurements of area wide air quality" (EPA 1997*a*). Averaging over space can be an important problem, too. Researchers studying trends in air pollution often use spatial averages as an indicator of the overall exposure in a population.

Chapter 3

Approximate EM algorithm for AR(1) process

The previous chapters have overviewed the main estimation problems in the geostatistical spatio-temporal model framework, including the general approach to model the spatial covariances with parameterized variograms, and the difficulties that arise because of missing data. The EM algorithm seems to be a promising tool in this problem, but a straightforward implementation of the algorithm seems to require a lot of computation. As was proposed in Chapter 1 and in Smith et al. (2003), some computational time savings are possible with an approximate EM algorithm (or its generalizations) that uses the unconditional expected values for the sufficient statistic in $Q(\cdot)$. The properties of such approximation are unknown at this point, and investigation into those is the prime goal of this proposal and further research, as outlined in Chapter 6.

This chapter takes a simple analytical model of an AR(1) time series process to analyze the performance of the proposed approximation to the EM algorithm. In this example, we are going to have a single instance of the correlated data, with an analogy of a single instance of the observed spatial field in the environmental applications, or an inseparable spatio-temporal field with correlations penetrating throughout the whole data set. Another specific feature of the time series is a distinct ordering of observations which does not have any analogue in the spatial or spatio-temporal context.

The problem of missing data in time series has been receiving substantial attention in the literature, as the missing data break otherwise nice structure of time series. Besides, due to autocovariance of the observations, a single missing data point may lead to a loss of many degrees of freedom ($p + q$ in a naïve implementation of an ARMA(p, q) model) as long as a single datum appears several times in the estimation

procedure. A number of approaches has been outlined in the time series literature. The likelihood of the time series with missing observations can be derived explicitly, as in Ljung (1982) or Penzer & Shea (1997). Alternatively, state-space models such as Kalman filter can be used to estimate the missing observations and parameters of the model (Kohn & Ansley 1986, Harvey & Pierse 1984). Regression-based methods of estimating the missing observations can be used, too (Beveridge 1992).

We, however, only use the AR(1) time series to provide a tractable analytic framework. We begin with the likelihood and estimating equations when the sample is complete, and no data are missing (Section 3.1). Then we shall introduce a gap of missing data and derive the estimating equations in the likelihood context (Section 3.2). The EM algorithm, as should have been expected from the general theory, gives the same estimating equations (Section 3.3). The proposed approximation, however, gives biased estimates (Section 3.4), and a correction to eliminate the bias is proposed. Finally, a more realistic scenario with many gaps of size 1 is considered in Sections 3.5 (ML approach) and 3.6 (approximate EM). The approximate EM gives biased estimates, but the bias terms in the estimating equations can be compensated for.

3.1 AR(1) process

Consider an AR(1) time-series process

$$y_t = a + \rho y_{t-1} + \epsilon_t, \quad t \in \mathcal{I} \subset \mathbb{Z},\ \epsilon_t \sim \text{i.i.d. } N(0, \sigma_\epsilon^2) \tag{3.1}$$

Assume $|\rho| < 1$, so the series is stationary. The mean of the process is

$$\mu = \mathbb{E}\, y_t = \frac{a}{1-\rho}, \tag{3.2}$$

so the process can also be written in deviations form as

$$y_t - \mu = \rho(y_{t-1} - \mu) + \epsilon_t \tag{3.3}$$

The variance of the process is

$$\sigma_y^2 = \mathbb{V}\, y_t = \frac{\sigma_\epsilon^2}{1-\rho^2} \tag{3.4}$$

If a sample of $y_t, t = 1, ..., T$ is observed, then the log likelihood of the data is

$$\ln L(\mu, \rho, \sigma_\epsilon^2; \mathbf{y}) = -\frac{T}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma| - \frac{1}{2}(\mathbf{y} - \mathbb{1}\mu)^T\Sigma^{-1}(\mathbf{y} - \mathbb{1}\mu) \tag{3.5}$$

where $\mathbb{1} = (1, \ldots, 1)^T$ is the column vector of ones. The covariance matrix, its inverse and its determinant are

$$\Sigma = \frac{\sigma_\epsilon^2}{1-\rho^2}\begin{pmatrix} 1 & \rho & \rho^2 & \ldots\ldots & \rho^{T-1} \\ \rho & 1 & \rho & \rho^2 & \ldots & \rho^{T-2} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \rho^{T-2} & \rho^{T-3} & \ldots & \rho & 1 & \rho \\ \rho^{T-1} & \rho^{T-2} & \ldots & \rho^2 & \rho & 1 \end{pmatrix}, \tag{3.6}$$

$$\Sigma^{-1} = \frac{1}{\sigma_\epsilon^2}\begin{pmatrix} 1 & -\rho & 0 & \ldots\ldots\ldots & 0 \\ -\rho & 1+\rho^2 & -\rho & 0 & \ldots & 0 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ 0 & \ldots\ldots\ldots & -\rho & 1+\rho^2 & -\rho \\ 0 & 0 & \ldots & 0 & -\rho & 1 \end{pmatrix}, \tag{3.7}$$

$$|\Sigma| = \frac{\sigma_\epsilon^{2T}}{1-\rho^2}, \tag{3.8}$$

so the likelihood simplifies to

$$\begin{aligned} \ln L(\mu, \rho, \sigma_\epsilon^2; \mathbf{y}) = &-\frac{T}{2}\ln 2\pi - \frac{1}{2}\big(T\ln\sigma_\epsilon^2 - \ln(1-\rho^2)\big) - \\ &- \frac{1}{2\sigma_\epsilon^2}\Big[(1+\rho^2)\sum_{t=1}^{T}(y_t-\mu)^2 - \rho^2\big((y_1-\mu)^2 + (y_T-\mu)^2\big)\Big) \\ &- 2\rho\sum_{t=2}^{T}(y_t-\mu)(y_{t-1}-\mu)\Big] \end{aligned} \tag{3.9}$$

From (3.9), the minimal sufficient statistic of the data is

$$\Big(\sum_{t=1}^{T} y_t, \sum_{t=1}^{T} y_t^2, \sum_{t=2}^{T} y_t y_{t-1}, y_1, y_T\Big) \tag{3.10}$$

As explained in Section 2.3, the sufficient statistic plays an important role in the EM algorithm. We shall use (3.10) in Section 3.3.

Note that the variance estimator is

$$\hat{\sigma}_\epsilon^2 = \frac{1}{T} Q(\mathbf{y}, \mu, \rho), \tag{3.11}$$

$$\begin{aligned} Q(\mathbf{y}, \mu, \rho) = (\mathbf{y} - \mathbb{1}\mu)^T \Sigma(\rho)^{-1} (\mathbf{y} - \mathbb{1}\mu) = \\ = (1+\rho^2) \sum_{t=1}^{T} (y_t - \mu)^2 - \rho^2 \big((y_1 - \mu)^2 + (y_T - \mu)^2 \big) \\ -2\rho \sum_{t=2}^{T} (y_t - \mu)(y_{t-1} - \mu) \end{aligned} \tag{3.12}$$

so the log likelihood concentrated with respect to σ_ϵ^2 is then

$$\begin{aligned} \ln L_c(\mu, \rho; \mathbf{y}) = \ln L(\mu, \rho, \hat{\sigma_\epsilon^2}; \mathbf{y}) = \\ = -\frac{T}{2} \ln 2\pi - \frac{1}{2} \big(T \ln \frac{Q(\mathbf{y}, \mu, \rho)}{T} - \ln(1-\rho^2) \big) - \frac{T}{2} \end{aligned} \tag{3.13}$$

Differentiating (3.13) with respect to μ and setting the derivative to zero gives

$$(1-\rho) \sum_{t=2}^{T-1} (y_t - \mu) + (y_1 - \mu) + (y_T - \mu) = 0 \tag{3.14}$$

Differentiating (3.13) with respect to ρ and setting the derivative to zero gives

$$\rho \sum_{t=2}^{T-1} (y_t - \mu)^2 - \sum_{t=2}^{T} (y_t - \mu)(y_{t-1} - \mu) + \alpha(\rho, T) Q(\mathbf{y}, \mu, \rho) = 0, \tag{3.15}$$

where

$$\alpha(\rho, T) = \frac{2\rho}{T(1-\rho^2)} = O(T^{-1}) \tag{3.16}$$

ML estimates of the parameters are

$$\hat{\mu}_{\text{c:ML}} = \frac{1}{T} \sum_{t=1}^{T} y_t - \alpha(\hat{\rho}_{\text{c:ML}}, T) \big(\frac{1}{T} \sum_{t=1}^{T} y_t - \frac{y_1 + y_T}{2} \big) + o_p(T^{-1}), , \tag{3.17}$$

$$\hat{\rho}_{\text{c:ML}} = \frac{\sum_{t=2}^{T-1} (y_t - \hat{\mu}_{\text{c:ML}})^2 + \alpha(\hat{\rho}_{\text{c:ML}}, T) Q(\mathbf{y}, \hat{\mu}_{\text{c:ML}}, \hat{\rho}_{\text{c:ML}})}{\sum_{t=2}^{T} (y_t - \hat{\mu}_{\text{c:ML}})(y_{t-1} - \hat{\mu}_{\text{c:ML}})} =$$

$$= \frac{\sum_{t=2}^{T-1}(y_t - \hat{\mu}_{\text{c:ML}})^2}{\sum_{t=2}^{T}(y_t - \hat{\mu}_{\text{c:ML}})(y_{t-1} - \hat{\mu}_{\text{c:ML}})} +$$
$$+\alpha(\hat{\rho}_{\text{c:ML}}, T)\frac{Q(\mathbf{y}, \hat{\mu}_{\text{c:ML}}, \hat{\rho}_{\text{c:ML}})}{\sum_{t=2}^{T}(y_t - \hat{\mu}_{\text{c:ML}})(y_{t-1} - \hat{\mu}_{\text{c:ML}})} + o_p(T^{-1}) \tag{3.18}$$

The subindex c:ML indicates that the estimates are obtained from the complete data set by maximum likelihood. The estimates can be computed by a numeric maximization procedure. Alternatively, one-step estimates that are asymptotically equivalent to the ML estimates can be obtained as follows. First, the OLS estimates

$$\hat{\mu}_{\text{OLS}} = \bar{y} = \frac{1}{T}\sum_{t=1}^{T} y_t \tag{3.19}$$

$$\hat{\rho}_{\text{OLS}} = \frac{\sum_{t=2}^{T-1}(y_t - \hat{\mu}_{\text{OLS}})^2}{\sum_{t=2}^{T}(y_t - \hat{\mu}_{\text{OLS}})(y_{t-1} - \hat{\mu}_{\text{OLS}})} \tag{3.20}$$

are obtained which are consistent estimates of the corresponding parameters, then $\alpha(\hat{\rho}_{\text{OLS}}, T)$ and $Q(\mathbf{y}, \hat{\mu}_{\text{OLS}}, \hat{\rho}_{\text{OLS}})$ are computed to improve the estimates up to the first order terms by (3.17) and (3.18).

3.2 AR(1) with missing data: ML approach

Now, suppose a portion of data that came from the AR(1) process is missing, so we first have n observed data points $(y_1, \ldots, y_n)$, then l missing observations $(y_{n+1}, \ldots, y_{n+l})$, and then again m observed points $(y_{n+l+1}, \ldots, y_T)$, $T = n + l + m$. The log likelihood is then

$$\ln L(\mu, \rho, \sigma_\epsilon^2; \mathbf{y}) = -\frac{m+n}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma^o| - \frac{1}{2}(\mathbf{y} - \mathbb{1}\mu)^T\Sigma^{o\,-1}(\mathbf{y} - \mathbb{1}\mu) \tag{3.21}$$

where Σ^o has a block structure:

$$\Sigma^o = \begin{pmatrix} T_n & R \\ R^T & T_m \end{pmatrix}, \tag{3.22}$$

$$T_s = \frac{\sigma_\epsilon^2}{1-\rho^2}\begin{pmatrix} 1 & \rho & \rho^2 & \ldots\ldots & \rho^{s-1} \\ \rho & 1 & \rho & \rho^2 & \ldots & \rho^{s-2} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \rho^{s-2} & \rho^{s-3} & \ldots & \rho & 1 & \rho \\ \rho^{s-1} & \rho^{s-2} & \ldots & \rho^2 & \rho & 1 \end{pmatrix}, \tag{3.23}$$

which is a $s \times s$ matrix, $s = n$, m, and its inverse is given by (3.7). Further,

$$R = \frac{\sigma_\epsilon^2}{1-\rho^2}\begin{pmatrix} \rho^{l+n} & \rho^{l+n+1} & \ldots\ldots\ldots & \rho^{l+m+n-1} \\ \rho^{l+n-1} & \rho^{l+n} & \rho^{l+n+1} & \ldots & \rho^{l+m+n-2} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \rho^{l+1} & \rho^{l+2} & \ldots\ldots\ldots & \rho^{l+m} \end{pmatrix} =$$

$$= \frac{\sigma_\epsilon^2 \rho^{l+1}}{1-\rho^2}\begin{pmatrix} \rho^{n-1} \\ \vdots \\ \rho \\ 1 \end{pmatrix}(1, \rho, \ldots, \rho^{m-1}) \tag{3.24}$$

is a rank 1 matrix of dimensions $n \times m$.

By using the formulae for block inverses (see e.g. Mardia, Kent & Bibby (1980)),

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} E & F \\ G & H \end{pmatrix},$$
$$E = \left(A - BD^{-1}C\right)^{-1}, \quad G = -D^{-1}CE,$$
$$H = \left(D - CA^{-1}B\right)^{-1}, \quad F = -A^{-1}BH \tag{3.25}$$

the component blocks of

$$\left(\Sigma^o\right)^{-1} = \begin{pmatrix} \Sigma^{o\,11} & \Sigma^{o\,12} \\ \Sigma^{o\,21} & \Sigma^{o\,22} \end{pmatrix} \tag{3.26}$$

can be obtained as follows.

$$RT_m^{-1}R^T = \frac{\sigma_\epsilon^2 \rho^{l+1}}{1-\rho^2}(\rho^{n-1}, \ldots, \rho, 1)^T(1, \rho, \ldots, \rho^{m-1}, \rho^m)\times$$

$$\times\frac{1}{\sigma_\epsilon^2}\begin{pmatrix} 1 & -\rho & 0 & \dots\dots\dots & 0 \\ -\rho & 1+\rho^2 & -\rho & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots\dots\dots & -\rho & 1+\rho^2 & -\rho \\ 0 & 0 & \dots & 0 & -\rho & 1 \end{pmatrix}\times$$

$$\times\frac{\sigma_\epsilon^2\rho^{l+1}}{1-\rho^2}(1,\rho,\dots,\rho^{m-1},\rho^m)^T(\rho^{n-1},\rho^{n-1},\dots,\rho,1) =$$

$$= \frac{\sigma_\epsilon^2\rho^{2(l+1)}}{(1-\rho^2)^2}(\rho^{n-1},\dots,\rho,1)^T(1-\rho^2,0,\dots,0)\times$$

$$\times(1,\rho,\dots,\rho^{m-1},\rho^m)^T(\rho^{n-1},\rho^{n-1},\dots,\rho,1) =$$

$$= \frac{\sigma_\epsilon^2\rho^{2(l+1)}}{1-\rho^2}(\rho^{n-1},\dots,\rho,1)^T(\rho^{n-1},\dots,\rho,1) =$$

$$= \frac{\sigma_\epsilon^2\rho^{2(l+1)}}{1-\rho^2}\begin{pmatrix} \rho^{2n-2} & \rho^{2n-3} & \dots & \rho^n & \rho^{n-1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \dots & \rho & 1 \end{pmatrix} \tag{3.27}$$

To proceed with $(T_n - RT_m^{-1}R^T)^{-1}$, the following matrix identity can be used:

$$(V+W)^{-1} = V^{-1}\big(I+WV^{-1}\big)^{-1} \tag{3.28}$$

In using this formula, we shall have $V = T_n$ with known inverse, and W is the negative of (3.27). Then

$$WV^{-1} = -\frac{\rho^{2(l+1)}}{1-\rho^2}\begin{pmatrix} \rho^{n-1} \\ \vdots \\ \rho \\ 1 \end{pmatrix}\begin{pmatrix} \rho^{n-1} \\ \vdots \\ \rho \\ 1 \end{pmatrix}^T\begin{pmatrix} 1 & -\rho & 0 & \dots\dots\dots & 0 \\ -\rho & 1+\rho^2 & -\rho & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots\dots\dots & -\rho & 1+\rho^2 & -\rho \\ 0 & 0 & \dots & 0 & -\rho & 1 \end{pmatrix} =$$

$$= -\frac{\rho^{2(l+1)}}{1-\rho^2}\begin{pmatrix} \rho^{n-1} \\ \vdots \\ \rho \\ 1 \end{pmatrix}(0,0,\dots,0,1-\rho^2) = \begin{pmatrix} 0 & 0 & \dots & 0 & -\rho^{2l+n+1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & -\rho^{2l+3} \\ 0 & 0 & \dots & 0 & -\rho^{2l+2} \end{pmatrix} \tag{3.29}$$

Further,

$$(I + WV^{-1}) = \begin{pmatrix} 1 & 0 & \dots & -\rho^{2l+n+1} \\ 0 & 1 & \dots & -\rho^{2l+n} \\ \dots & \dots & \dots & \dots \\ 0 & \dots & 1 & -\rho^{2l+3} \\ 0 & \dots & 0 & 1-\rho^{2l+2} \end{pmatrix},$$

$$(I + WV^{-1})^{-1} = \begin{pmatrix} 1 & 0 & \dots & \rho^{2l+n+1}/(1-\rho^{2l+2}) \\ 0 & 1 & \dots & \rho^{2l+n}/(1-\rho^{2l+2}) \\ \dots & \dots & \dots & \dots \\ 0 & \dots & 1 & \rho^{2l+3}/(1-\rho^{2l+2}) \\ 0 & \dots & 0 & 1/(1-\rho^{2l+2}) \end{pmatrix},$$

$$V^{-1}(I + WV^{-1})^{-1} =$$

$$= \frac{1}{\sigma_\epsilon^2} \begin{pmatrix} 1 & -\rho & \dots & 0 \\ -\rho & 1+\rho^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & \dots & -\rho & 1+\rho^2 & -\rho \\ 0 & \dots & 0 & -\rho & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & \dots & \rho^{2l+n+1}/(1-\rho^{2l+2}) \\ 0 & 1 & \dots & \rho^{2l+n}/(1-\rho^{2l+2}) \\ \dots & \dots & \dots & \dots \\ 0 & \dots & 1 & \rho^{2l+3}/(1-\rho^{2l+2}) \\ 0 & \dots & 0 & 1/(1-\rho^{2l+2}) \end{pmatrix}$$

$$= \frac{1}{\sigma_\epsilon^2} \begin{pmatrix} 1 & -\rho & 0 & \dots & 0 & 0 \\ -\rho & 1+\rho^2 & -\rho & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & -\rho & 1+\rho^2 & -\rho \\ 0 & \dots & \dots & 0 & -\rho & \phi(\rho,l) \end{pmatrix} = \Sigma^{o\,11}, \tag{3.30}$$

$$\phi(\rho, l) = \frac{1-\rho^{2l+4}}{1-\rho^{2l+2}} \tag{3.31}$$

which is the upper left block of $\left(\Sigma^o\right)^{-1}$.

The next block of $\Sigma^{o\,-1}$ is

$$\Sigma^{o\,21} = -T_m^{-1} R^T \Sigma^{o\,11} = \frac{1}{\sigma_\epsilon^2} \begin{pmatrix} 1 & -\rho & 0 & \dots\dots & 0 \\ -\rho & 1+\rho^2 & -\rho & 0 \quad \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots\dots & -\rho & 1+\rho^2 & -\rho \\ 0 & 0 \quad \dots & 0 & -\rho & 1 \end{pmatrix} \times$$

$$\times \frac{\sigma_\epsilon^2 \rho^{l+1}}{1-\rho^2} \begin{pmatrix} 1 \\ \rho \\ \vdots \\ \rho^{m-1} \end{pmatrix} (\rho^{n-1}, \dots, 1)\, \Sigma^{o\,11} =$$

$$= -\rho^{l+1} \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} (0, \dots, 0, \frac{1-\rho^2}{1-\rho^{2l+2}}) = \begin{pmatrix} 0 & \dots & 0 & -\psi(\rho,l) \\ 0 & \dots & 0 & 0 \\ \vdots & \ddots & \dots\dots & \\ 0 & \dots\dots & & 0 \end{pmatrix}, \tag{3.32}$$

$$\psi(\rho, l) = \rho^{l+1} \frac{1-\rho^2}{1-\rho^{2l+2}} \tag{3.33}$$

The computations identical to (3.27)–(3.30) yield

$$\Sigma^{o\,22} = \frac{1}{\sigma_\epsilon^2} \begin{pmatrix} \phi(\rho,l) & -\rho & 0 & \dots\dots & 0 & 0 \\ -\rho & 1+\rho^2 & -\rho & 0 \quad \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & -\rho & 1+\rho^2 & -\rho \\ 0 & \dots\dots & & 0 & -\rho & 1 \end{pmatrix}, \tag{3.34}$$

and finally $\Sigma^{o\,12} = \left(\Sigma^{o\,21}\right)^T$.

The changes due to the missing data are concentrated near the place where it is missing, as it should have been expected from the Markovian character of AR(1) process. Let us consider the two limiting cases.

If $l = 0$, no data is missing, and one can verify that $\phi(\rho, 0) = 1 + \rho^2$, $\psi(\rho, 0) = \rho$. Then $\left(\Sigma^o\right)^{-1}$ has the form of (3.7).

If $l \to \infty$, there are two independent samples of possibly different length from AR(1), and both the covariance matrix and its inverse are block matrices with off-

diagonal block equal to zero matrices.

The determinant of $\left(\Sigma^o\right)^{-1}$ can be shown to be

$$\left|\Sigma^{o-1}\right| = \left[\phi(\rho,l) - \rho^2\right]^2 - \psi^2(\rho,l), \tag{3.35}$$

which can be verified to give the appropriate limits $1-\rho^2$ and $(1-\rho^2)^2$ when $l=0$ and $l=\infty$, respectively.

Let us now derive the normal equations. The generalized sum of squares is

$$\begin{aligned} Q^o(\mathbf{y},\mu,\rho) &= (\mathbf{y}-\mathbb{1}\mu)^T\Sigma^o(\rho)^{-1}(\mathbf{y}-\mathbb{1}\mu) = \\ &= (y_1-\mu)^2 + (y_T-\mu)^2 + (1+\rho^2)\Big[\sum_{t=2}^{n}(y_t-\mu)^2 + \sum_{t=n+l+2}^{T}(y_t-\mu)^2\Big] + \\ &+\phi(\rho)\Big[(y_n-\mu)^2 + (y_{n+l+1}-\mu)^2\Big] - \\ &-2\rho\Big[\sum_{t=2}^{n}(y_t-\mu)(y_{t-1}-\mu) + \sum_{t=n+l+2}^{T}(y_t-\mu)(y_{t-1}-\mu)\Big] \\ &-2\psi(\rho)(y_n-\mu)(y_{n+l+1}-\mu) \end{aligned} \tag{3.36}$$

The likelihood can again be concentrated w.r.t. σ_ϵ^2:

$$\hat{\sigma^2_{\epsilon\,\text{i:ML}}} = \frac{Q^o(\mathbf{y},\mu,\rho)}{n+m}, \tag{3.37}$$

$$\begin{aligned} \ln L_c(\mu,\rho;\mathbf{y}) &= -\frac{n+m}{2}\ln 2\pi - \\ &-\frac{1}{2}\Big((n+m)\ln\frac{Q^o(\mathbf{y},\mu,\rho)}{n+m} - \ln\big\{\left[\phi(\rho,l)-\rho^2\right]^2 - \psi^2(\rho,l)\big\}\Big) - \frac{n+m}{2} \end{aligned} \tag{3.38}$$

The subindex i:ML denotes that this is the ML estimate for the incomplete data.

Differentiating w.r.t. μ gives

$$\begin{aligned} &(1-\rho)(y_1-\mu+y_T-\mu) + (1-\rho)^2\Big[\sum_{t=2}^{n-1}(y_t-\mu) + \sum_{t=n+l+2}^{T-1}(y_t-\mu)\Big] + \\ &+(\phi(\rho,l)-\psi(\rho,l)-\rho)\big[(y_n-\mu)+(y_{n+l+1}-\mu)\big] = 0, \end{aligned} \tag{3.39}$$

so

$$\hat{\mu}_{\text{i:ML}} =$$

$$= \frac{(1-\rho)(y_1+y_T)+(1-\rho)^2\Big[\sum_{t=2}^{n-1} y_t + \sum_{t=n+l+2}^{T-1} y_t\Big] + (\phi(\rho,l)-\psi(\rho,l)-\rho)(y_n+y_{n+l+1})}{2(1-\rho)+(n+m-4)(1-\rho)^2+2(\phi(\rho)-\psi(\rho)-\rho)} =$$

$$= \frac{1}{n+m}\Big[\sum_{t=1}^{n} y_t + \sum_{t=n+l+1}^{T} y_t\Big] + \alpha(\rho,n+m)\frac{y_1+y_T}{2} -$$

$$-\frac{1}{n+m}\Big[\sum_{t=1}^{n} y_t + \sum_{t=n+l+1}^{T} y_t\Big]\frac{2}{n+m}\frac{1-\rho-\phi(\rho,l)-\psi(\rho,l)}{1-\rho} +$$

$$+\frac{\phi(\rho,l)-\psi(\rho,l)-1}{(n+m)(1-\rho)}(y_n+y_{n+l+1}) + o_p\big((n+m)^{-1}\big) \tag{3.40}$$

where the first term is $O_p(1)$, and all others are $O_p\big((n+m)^{-1}\big)$. The second term is the correction due to the terminal points similar to the complete data case, and the other two are corrections for the missing data.

The derivatives with respect to ρ are as follows:

$$\frac{d\phi(\phi,l)}{d\rho} = \rho^{2l+1}\frac{(2l+2)(1-\rho^{2l+4})-(2l+4)\rho^2(1-\rho^{2l+2})}{(1-\rho^{2l+2})^2} \tag{3.41}$$

$$\frac{d\psi(\phi,l)}{d\rho} = \rho^{l}\frac{\big[(l+1)(1-\rho^2)-2\rho^2\big](1-\rho^{2l+2})+\rho^{2l+2}(1-\rho^2)(2l+2)}{(1-\rho^{2l+2})^2} \tag{3.42}$$

$$\frac{1}{2}\frac{\partial Q^o(\mathbf{y},\mu,\rho)}{\partial\rho} =$$

$$\rho\Big[\sum_{t=2}^{n}(y_t-\mu)^2 + \sum_{t=n+l+2}^{T}(y_t-\mu)^2\Big] + \phi'(\rho,l)\Big[(y_n-\mu)^2+(y_{n+l+1}-\mu)^2\Big] -$$

$$-\Big[\sum_{t=2}^{n}(y_t-\mu)(y_{t-1}-\mu) + \sum_{t=n+l+2}^{T}(y_t-\mu)(y_{t-1}-\mu) +$$

$$+\psi'(\rho,l)(y_n-\mu)(y_{n+l+1}-\mu)\Big] \tag{3.43}$$

$$\frac{1}{2}\frac{d\ln|\Sigma^{o-1}|}{d\rho} = \big(\phi(\rho,l)-\rho^2\big)\big(\phi'(\rho,l)-2\rho\big) - \psi'(\rho,l)\psi(\rho,l) \tag{3.44}$$

Then the normal equation for ρ is

$$\frac{1}{2}\frac{\partial Q^o(\mathbf{y},\mu,\rho)}{\partial\rho} =$$

$$\frac{Q^o(\mathbf{y},\mu,\rho)}{n+m}\,\frac{(\phi(\rho,l)-\rho^2)(\phi'(\rho,l)-2\rho)-\psi(\rho,l)\psi'(\rho,l)}{(\phi(\rho,l)-\rho^2)^2-\psi^2(\rho,l)} = R(\mathbf{y},\mu,\rho), \tag{3.45}$$

$$\hat{\rho}_{\text{i:ML}}\,\frac{\sum_{t=1}^{n}(y_t-\mu)^2+\sum_{t=n+l+1}^{T}(y_t-\mu)^2}{n+m} -$$

$$-\frac{\sum_{t=2}^{n}(y_t-\mu)(y_{t-1}-\mu)+\sum_{t=n+l+2}^{T}(y_t-\mu)(y_{t-1}-\mu)-(y_{n+l+1}-\mu)(y_n-\mu)}{n+m} =$$

$$= \frac{1}{n+m}R(\mathbf{y},\mu,\rho) - \frac{\phi'(\hat{\rho}_{\text{i:ML}},l)-\hat{\rho}_{\text{i:ML}}}{n+m}\big[(y_n-\mu)^2+(y_{n+l+1}-\mu)^2\big] -$$

$$-\frac{\psi'(\hat{\rho}_{\text{i:ML}},l)-1}{n+m}(y_n-\mu)(y_{n+l+1}-\mu) \tag{3.46}$$

where the terms in the LHS are of the order $O_p(1)$, and the terms in the RHS are of the order $O_p\big((n+m)^{-1}\big)$.

Just as in the case of the ML estimates for the complete data, the explicit analytic solution cannot be obtained. In practice, one would need to use two-step or iterative procedures. The starting values can be based on the sample mean and the sample correlation between y_t and y_{t-1}, as before. It should be noted however that the corrections due to the missing data are of order $O_p\big((n+m)^{-1}\big)$.

3.3 AR(1) with missing data: the EM algorithm

The EM algorithm is an iterative procedure of finding the critical points of the likelihood function. See Section 2.3 for a general overview and the main features of this estimating procedure. The particular version of the EM algorithm that we want to consider is the expectation-conditional maximization (ECM) variant of the EM algorithm in which the maximization can be performed separately over subspaces of the parameter space spanning the whole space. We would like to split the maximization into maximization over the mean parameter subspace (μ in our case, or regression coefficients if covariates are allowed for) and the variance subspace that may be split further into the overall constant σ_ϵ^2 that lends itself to a simple enough estimate such as (3.11), and the covariance structure parameter, which in this case is ρ.

As was shown above in (3.10), the sufficient statistic of the complete data set is

$$\Big(\sum_{t=1}^{T} y_t, \sum_{t=1}^{T} y_t^2, \sum_{t=2}^{T} y_t y_{t-1}, y_1, y_T\Big) \tag{3.47}$$

Thus, to get an implementation of the EM algorithm, we need to be able to predict the three sums given the parameter estimates and the available observations, which boils down to prediction of y_t, y_t^2 and $y_t y_{t-1}$ when at least one of y_t, y_{t-1} is missing.

To derive the aforementioned conditional expectation, consider four-variate normal distribution

$$\begin{pmatrix} y_n \\ y_{n+r-1} \\ y_{n+r} \\ y_{n+l+1} \end{pmatrix} \sim N\left(\mathbb{1}\mu, \frac{\sigma_\epsilon^2}{1-\rho^2} \begin{pmatrix} 1 & \rho^{r-1} & \rho^r & \rho^{l+1} \\ \rho^{r-1} & 1 & \rho & \rho^{l-r+2} \\ \rho^r & \rho & 1 & \rho^{l-r+1} \\ \rho^{l+1} & \rho^{l-r+2} & \rho^{l-r+1} & 1 \end{pmatrix} \right), \; 2 \le r \le l \tag{3.48}$$

By the standard multivariate normal theory, the distribution of the unobserved y_{n+r-1}, y_{n+r} conditional on y_n, y_{n+l+1} is bivariate normal with mean

$$\mathbb{E}\left[\begin{pmatrix} y_{n+r-1} \\ y_{n+r} \end{pmatrix} \middle| \begin{pmatrix} y_n \\ y_{n+l+1} \end{pmatrix} \right] =$$
$$\begin{pmatrix} \mu + \frac{1}{1-\rho^{2l+2}} \big[(\rho^{r-1} - \rho^{2l-r+3})(y_n - \mu) + (\rho^{l-r+2} - \rho^{l+r})(y_{n+l+1} - \mu) \big] \\ \mu + \frac{1}{1-\rho^{2l+2}} \big[(\rho^{r} - \rho^{2l-r+2})(y_n - \mu) + (\rho^{l-r+1} - \rho^{l+r+1})(y_{n+l+1} - \mu) \big] \end{pmatrix}, \tag{3.49}$$

and covariance matrix

$$\frac{1-\rho^2}{\sigma_\epsilon^2} \mathbb{V}\left[\begin{pmatrix} y_{n+r-1} \\ y_{n+r} \end{pmatrix} \middle| \begin{pmatrix} y_n \\ y_{n+l+1} \end{pmatrix} \right] = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} -$$
$$- \begin{pmatrix} \rho^{r-1} & \rho^{l-r+2} \\ \rho^r & \rho^{l-r+1} \end{pmatrix} \begin{pmatrix} 1 & \rho^{l+1} \\ \rho^{l+1} & 1 \end{pmatrix}^{-1} \begin{pmatrix} \rho^{r-1} & \rho^r \\ \rho^{l-r+2} & \rho^{l-r+1} \end{pmatrix} =$$

$$= \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} - \frac{1}{1-\rho^{2l+2}} \times$$
$$\times \begin{pmatrix} \rho^{2r-2} - 2\rho^{2l+2} + \rho^{2l-2r+4} & \rho^{2r-1} - \rho^{2l+1} + \rho^{2l-2r+3} - \rho^{2l+3} \\ \rho^{2r-1} - \rho^{2l+1} + \rho^{2l-2r+3} - \rho^{2l+3} & \rho^r - 2\rho^{2l+2} + \rho^{2l-2r+2} \end{pmatrix}, \tag{3.50}$$

This is also the kriging estimator in the univariate case.

Hence, the conditional expectations of the relevant sums are as follows:

$$\mathbb{E}\left[\sum_{t=n+1}^{n+l} y_t \,\middle|\, \begin{pmatrix} y_n \\ y_{n+l+1} \end{pmatrix}\right] = l\mu + \frac{1}{1-\rho^{2l+2}}\Big\{(y_n-\mu)\sum_{r=1}^{l}\big[\rho^r - \rho^{2l-r+2}\big] + (y_{n+l+1}-\mu)\sum_{r=1}^{l}\big[\rho^{l-r+1} - \rho^{l+r+1}\big]\Big\} =$$

$$= l\mu + \frac{\rho(1-\rho^l)(1-\rho^{l+1})}{(1-\rho)(1-\rho^{2l+2})}\big[(y_n-\mu)+(y_{n+l+1}-\mu)\big], \tag{3.51}$$

$$\mathbb{E}\left[\sum_{t=1}^{T} y_t \,\middle|\, \text{observed data}\right] = \sum_{t=1}^{n-1} y_t + \sum_{t=n+l+2}^{T} y_t + l\mu + \frac{(1-\rho^{l+1})(1-\rho^{l+2})}{(1-\rho)(1-\rho^{2l+2})}\big[(y_n-\mu)+(y_{n+l+1}-\mu)\big] \tag{3.52}$$

This can now be substituted into the likelihood and the normal equations. For instance, the normal equation for μ, equation (3.14), becomes

$$0 = (1-\rho)\,\mathbb{E}\left[\sum_{t=1}^{T} y_t \,\middle|\, \text{observed data}\right] - (1-\rho)T\mu + (y_1-\mu) + (y_T-\mu) =$$

$$= (1-\rho)\Big[\sum_{t=1}^{n-1} y_t + \sum_{t=n+l+2}^{T} y_t\Big] + (y_1-\mu) + (y_T-\mu)$$

$$+\frac{(1-\rho^{l+1})(1-\rho^{l+2})}{(1-\rho^{2l+2})}\big[(y_n-\mu)+(y_{n+l+1}-\mu)\big] - (1-\rho)(T-l)\mu \tag{3.53}$$

which after a number of simplifications coincides with (3.39). Other equations can be verified in the same way to demonstrate the equivalence of the EM algorithm and the maximum likelihood estimates.

3.4 AR(1) with missing data: the approximate EM algorithm

The proposed approximate version of the EM algorithm suggests that the predictions of the missing data statistics be simply the total, or unconditional, expected values.

Denoting the approximate conditional expectation as $\tilde{\mathbb{E}}$, we have

$$\tilde{\mathbb{E}}\big[y_t|\text{observed data, } \mu,\rho,\sigma_\epsilon^2\ \big]=\mu, \tag{3.54}$$

$$\tilde{\mathbb{E}}\big[y_t^2|\text{observed data, } \mu,\rho,\sigma_\epsilon^2\ \big]=\mu^2+\frac{\sigma_\epsilon^2,}{1-\rho^2}, \tag{3.55}$$

$$\tilde{\mathbb{E}}\big[y_t y_{t-1}|\text{observed data, } \mu,\rho,\sigma_\epsilon^2\ \big]=\mu^2+\rho\,\frac{\sigma_\epsilon^2,}{1-\rho^2} \tag{3.56}$$

which is obviously quite different from what was derived in the previous section, and does not account for the observed correlated data at all. Those are the values that can be substituted in place of the missing components of $(\mathbf{y}-\mathbb{1}\mu)(\mathbf{y}-\mathbb{1}\mu)^T$ of (1.1).

Let us see what the estimating equations are that this version of the EM algorithm implies. The approximate conditional expectation of the generalized sum of squares is

$$\begin{gathered}\tilde{\mathbb{E}}Q^l(\mathbf{y},\mu,\rho)=(y_1-\mu)^2-2\rho(y_1-\mu)(y_2-\mu)+\ldots-2\rho(y_n-\mu)(y_{n-1}-\mu)+\\ +(1+\rho^2)(y_n-\mu)^2-2\frac{\rho^2\sigma_\epsilon^2}{1-\rho^2}+\frac{\sigma_\epsilon^2(1+\rho^2)}{1-\rho^2}-\ldots+\frac{\sigma_\epsilon^2(1+\rho^2)}{1-\rho^2}-2\frac{\rho^2\sigma_\epsilon^2}{1-\rho^2}+\\ +(1+\rho^2)(y_{n+l+1}-\mu)^2-2\rho(y_{n+l+2}-\mu)(y_{n+l+1}-\mu)+\ldots-\\ -2\rho(y_T-\mu)(y_{T-1}-\mu)+(y_T-\mu)^2=\\ =(1+\rho^2)\big[\sum_{t=1}^{n}(y_t-\mu)^2+\sum_{t=n+l+1}^{T}(y_t-\mu)^2\big]-\rho^2\big[((y_1-\mu)^2+(y_T-\mu)^2\big]-\\ -2\rho\,\big[\sum_{t=2}^{n}(y_t-\mu)(y_{t-1}-\mu)+\sum_{t=n+l+2}^{T}(y_t-\mu)(y_{t-1}-\mu)\big]-\frac{2\sigma_\epsilon^2(1+\rho^2)}{1-\rho^2}+l\sigma_\epsilon^2\end{gathered} \tag{3.57}$$

and the approximate, or pseudo-likelihood, being maximized is

$$\begin{gathered}\ln\tilde{L}(\mu,\rho,\sigma_\epsilon^2;\mathbf{y})=-\frac{T}{2}\ln 2\pi-\frac{1}{2}\Big\{(T-l)\ln\sigma_\epsilon^2-\ln(1-\rho^2)+\\ +(1+\rho^2)\Big[\sum_{t=1}^{n}\frac{(y_t-\mu)^2}{\sigma_\epsilon^2}+\sum_{t=n+l+1}^{T}\frac{(y_t-\mu)^2}{\sigma_\epsilon^2}\Big]-\frac{\rho^2}{\sigma_\epsilon^2}\big[((y_1-\mu)^2+(y_T-\mu)^2\big]-\\ -2\rho\Big[\sum_{t=2}^{n}\frac{(y_t-\mu)(y_{t-1}-\mu)}{\sigma_\epsilon^2}+\sum_{t=n+l+2}^{T}\frac{(y_t-\mu)(y_{t-1}-\mu)}{\sigma_\epsilon^2}+\frac{1+\rho^2}{1-\rho^2}\Big]\Big\}\end{gathered} \tag{3.58}$$

Let us concentrate with respect to σ_ϵ^2:

$$\frac{\partial \ln \tilde{L}}{\partial \sigma_\epsilon^2} = -\frac{T-l}{2\sigma_\epsilon^2} + \frac{\tilde{Q}(\mathbf{y},\mu,\rho)}{2\sigma_\epsilon^4}; \tag{3.59}$$

$$\begin{aligned}\tilde{Q}(\mathbf{y},\mu,\rho) = (1+\rho^2)\Big[\sum_{t=1}^{n}(y_t-\mu)^2 + \sum_{t=n+l+1}^{T}(y_t-\mu)^2\Big] - \\ -\rho^2\big[(y_1-\mu)^2 + (y_T-\mu)^2\big] - \\ -2\rho\Big[\sum_{t=2}^{n}(y_t-\mu)(y_{t-1}-\mu) + \sum_{t=n+l+2}^{T}(y_t-\mu)(y_{t-1}-\mu)\Big]\end{aligned} \tag{3.60}$$

$$\hat{\sigma}^2_{\epsilon\ \mathrm{i:aEM}} = \frac{1}{T-l}\tilde{Q}(\mathbf{y},\mu,\rho) = \frac{1}{n+m}\tilde{Q}(\mathbf{y},\mu,\rho) \tag{3.61}$$

which is different by $O_p\big((n+m)^{-1}\big)$ from the ML estimate given by (3.2) since the terms $Q^o(\cdot)$ and $\tilde{Q}(\cdot)$ differ by terms of $O_p(1)$ related to the corrections for the missing data at the boundaries of the gap. The subindex i:aEM stands for incomplete data, estimation by approximate EM algorithm.

The concentrated pseudo-likelihood is now

$$\begin{aligned}\ln \tilde{L}_c(\mu,\rho;\mathbf{y}) = \\ = -\frac{T}{2}\ln 2\pi - \frac{1}{2}\Big\{(n+m)\ln\frac{\tilde{Q}(\mathbf{y},\mu,\rho)}{n+m} - \ln(1-\rho^2) + n + m + \frac{1+\rho^2}{1-\rho^2}\Big\}\end{aligned} \tag{3.62}$$

Differentiating with respect to μ gives

$$\begin{aligned}-\frac{1}{2}\frac{\partial\tilde{Q}(\mathbf{y},\mu,\rho)}{\partial\mu} = (1+\rho^2)\Big[\sum_{t=2}^{n}(y_t-\mu) + \sum_{t=n+l+1}^{T-1}(y_t-\mu)\Big] + \\ +(y_1-\mu) + (y_T-\mu) - 2\rho\Big[\sum_{t=2}^{n-1}(y_t-\mu) + \sum_{t=n+l+2}^{T-1}(y_t-\mu) + \\ +\frac{1}{2}\big\{(y_1-\mu) + (y_n-\mu) + (y_{n+l+1}-\mu) + (y_T-\mu)\big\}\Big] = \\ = (1-\rho)^2\Big[\sum_{t=2}^{n}(y_t-\mu) + \sum_{t=n+l+1}^{T-1}(y_t-\mu)\Big] + (1-\rho)\big[(y_1-\mu) + (y_T-\mu)\big] + \\ +\rho\big[(y_n-\mu) + (y_{n+l+1}-\mu)\big],\end{aligned} \tag{3.63}$$

$$\hat{\mu}_{\text{i:aEM}} = \frac{\sum_{t=2}^{n} y_t + \sum_{t=n+l+1}^{T-1} y_t + \frac{1}{1-\rho}(y_1 + y_T) + \frac{\rho}{(1-\rho)^2}(y_n + y_{n+l+1})}{(n+m-2) + 2/(1-\rho)^2} \tag{3.64}$$

This differs from the ML estimates (3.40) by $O_p\big((n+m)^{-1}\big)$. Note the same principal term $\sum_t y_t$ in the numerator; all other terms are of order $O_p\big((n+m)^{-1}\big)$.

The derivatives with respect to ρ are as follows.

$$\frac{1}{2}\frac{\partial \tilde{Q}}{\partial \rho} = \rho\Big[\sum_{t=2}^{n}(y_t-\mu)^2 + \sum_{t=n+l+1}^{T-1}(y_t-\mu)^2\Big] - \\ -\Big[\sum_{t=2}^{n}(y_t-\mu)(y_{t-1}-\mu) + \sum_{t=n+l+2}^{T}(y_t-\mu)(y_{t-1}-\mu)\Big], \tag{3.65}$$

$$\frac{\partial \ln L_c}{\partial \rho} = -\frac{n+m}{\tilde{Q}(\mathbf{y},\mu,\rho)}\Big\{\rho\Big[\sum_{t=2}^{n}(y_t-\mu)^2 + \sum_{t=n+l+1}^{T-1}(y_t-\mu)^2\Big] - \\ -\Big[\sum_{t=2}^{n}(y_t-\mu)(y_{t-1}-\mu) + \sum_{t=n+l+2}^{T}(y_t-\mu)(y_{t-1}-\mu)\Big]\Big\} - \frac{\rho(3-\rho^2)}{(1-\rho^2)^2} \tag{3.66}$$

$$\rho\,\frac{\sum_{t=2}^{n}(y_t-\mu)^2 + \sum_{t=n+l+1}^{T-1}(y_t-\mu)^2}{n+m} - \\ -\frac{\sum_{t=2}^{n}(y_t-\mu)(y_{t-1}-\mu) + \sum_{t=n+l+2}^{T}(y_t-\mu)(y_{t-1}-\mu)}{n+m} = \\ = -\frac{1}{n+m}\frac{\rho(3-\rho^2)}{(1-\rho^2)^2}\frac{\tilde{Q}(\mathbf{y},\mu,\rho)}{n+m} \tag{3.67}$$

The terms in the LHS are of the order $O_p(1)$, and the term in the RHS are of the order $O_p\big((n+m)^{-1}\big)$. Note that the principal term is the same as for the ML case except for the term $(y_{n+l+1}-\mu)(y_n-\mu)/(n+m)$ which is $O_p\big((n+m)^{-1}\big)$, so the two expressions differ by $O_p\big((n+m)^{-1}\big)$.

As noted in the end of Section 3.2, the corrections for the missing data are of order (sample size)$^{-1}$. The suggested procedure essentially affects the weights of the bordering observations y_n and y_{n+l+1} only, with differences in resulting estimating equations of order (sample size)$^{-1}$. It is then asymptotically equivalent to the maximum likelihood estimates since their deviations from the true values of the parameters are of the order (sample size)$^{-1/2}$, as given by the asymptotic normality of the estimates.

Thus, the approximate EM algorithm is asymptotically first-order equivalent to the exact EM algorithm and the MLE for the considered case of a single gap of missing data in the time series. This equivalence, however, only holds for a small amount of missing data, i.e., a small number of affected observations. In the AR(1) case, the data missing as a gap in the middle of the series, however long that gap is, only affects the bordering observations y_n, y_{n+l+1}, and those observations have a weight of (sample size)$^{-1}$ in the estimating equations.

The situation will change as we move towards more realistic missing data mechanisms.

3.5 AR(1) with many gaps

Let us now assume that there is not a single gap, but a number of small gaps. Within each gap, there is only one observation missing, and two gaps are separated by at least two non-missing observations[1]. If we denote the observed data by "$\circ$", and missing data by "$\times$", then a plausible series may look like

$$\circ\ \circ\ \circ\ \times\ \circ\ \circ\ \times\ \circ\ \circ\ \circ\ \times\ \circ\ \circ\ \circ\ \circ\ \circ\ \times\ \circ\ \circ\ \times\ \circ\circ$$

These assumptions look a bit artificial, but their sole purpose at this point is to make the analysis tractable. The contributions to the likelihood of the missing observations come only from the neighboring observations, and they are of the same structure that has been derived earlier. Also, each of the non-missing observations contributes to at most one missing observation.

If y_t is an observation from AR(1) process (3.1), we can derive the conditional distributions of the next two observations as follows:

$$y_{t+1}|y_t \sim N(\mu + \rho(y_t - \mu), \sigma_\epsilon^2), \tag{3.68}$$

$$y_{t+2}|y_t \sim N(\mu + \rho^2(y_t - \mu), \sigma_\epsilon^2(1 + \rho^2)), \tag{3.69}$$

The overall likelihood is then

$$\ln L(\theta) = \ln l(y_1|\theta) + \sum_{t\in\mathcal{I}} \ln l(y_{t+1}|y_t, \theta) + \sum_{t\in\mathcal{B}} \ln l(y_{t+2}|y_t, \theta)$$

[1] These assumptions are needed so that the likelihood function is simple enough to include only first and second lag correlations.

$$\theta = (\mu, \rho, \sigma_\epsilon^2)$$

$$\ln l(y_1|\theta) = -\frac{1}{2}\ln 2\pi - \frac{1}{2}\ln \sigma_\epsilon^2 + \frac{1}{2}\ln(1-\rho^2) - \frac{1-\rho^2}{2\sigma_\epsilon^2}(y_1-\mu)^2,$$

$$\ln l(y_{t+1}|y_t, \theta) = -\frac{1}{2}\ln 2\pi - \frac{1}{2}\ln \sigma_\epsilon^2 - \frac{1}{2\sigma_\epsilon^2}\big[(y_{t+1}-\mu) - \rho(y_t-\mu)\big]^2$$

$$\ln l(y_{t+2}|y_t, \theta) = -\frac{1}{2}\ln 2\pi - \frac{1}{2}\ln \sigma_\epsilon^2 - \frac{1}{2}\ln(1+\rho^2)$$
$$-\frac{1}{2\sigma_\epsilon^2(1+\rho^2)}\big[(y_{t+2}-\mu) - \rho^2(y_t-\mu)\big]^2 \tag{3.70}$$

where $t \in \mathcal{I}$ are all the time points for which the next point is observed (interior), and $t \in \mathcal{B}$ are the points for which the next one is missing (boundary). If there are $M = \nu T$ missing data points over the period $1, \ldots, T$ (ν is the fraction of the missing data), then there are M terms in the second (boundary) sum, and $T - 2M - 1 \approx T(1-2\nu)$ in the first (interior) sum. Combining the terms, we obtain:

$$\ln L(\theta) = -\frac{T(1-\nu)}{2}\ln 2\pi - \frac{T(1-\nu)}{2}\ln \sigma_\epsilon^2 + \frac{1}{2}\ln(1-\rho^2) - \frac{T\nu}{2}\ln(1+\rho^2) -$$
$$-\frac{1-\rho^2}{2\sigma_\epsilon^2}(y_1-\mu)^2 - \frac{1}{2\sigma_\epsilon^2}\sum_{t\in\mathcal{I}}\big[(y_{t+1}-\mu) - \rho(y_t-\mu)\big]^2 -$$
$$-\frac{1}{2\sigma_\epsilon^2(1+\rho^2)}\sum_{t\in\mathcal{B}}\big[(y_{t+2}-\mu) - \rho^2(y_t-\mu)\big]^2 \tag{3.71}$$

The score equations can be derived as follows:

$$\frac{\partial \ln L}{\partial \mu} = \frac{1}{\sigma_\epsilon^2}\Big\{(1-\rho^2)(y_1-\mu) + \sum_{t\in\mathcal{I}}(1-\rho)\big[(y_{t+1}-\mu) - \rho(y_t-\mu)\big]$$
$$+\frac{1}{1+\rho^2}\sum_{t\in\mathcal{B}}(1-\rho^2)\big[(y_{t+2}-\mu) - \rho^2(y_t-\mu)\big]\Big\} \tag{3.72}$$

with the sufficient statistic $(y_1, \sum y_t, \sum_{t\in\mathcal{B}} y_t, \sum_{t\in\mathcal{B}} y_{t+2})$. The next equation is

$$\frac{\partial \ln L}{\partial \sigma_\epsilon^2} = -\frac{T(1-\nu)}{2}\frac{1}{\sigma_\epsilon^2} + \frac{1}{2\sigma_\epsilon^4}\breve{Q},$$

$$\breve{Q}(\mathbf{y}, \mu, \rho) = (1-\rho^2)(y_1-\mu^2) + \sum_{t\in\mathcal{I}}\big[(y_{t+1}-\mu) - \rho(y_t-\mu)\big]^2 +$$
$$+\frac{1}{1+\rho^2}\sum_{t\in\mathcal{B}}\big[(y_{t+2}-\mu) - \rho^2(y_t-\mu)\big]^2 \tag{3.73}$$

Finally, the parameter of the greatest interest, the variance structure, is found from the following equation:

$$\frac{\partial \ln L}{\partial \rho} =$$
$$-\frac{\rho}{1-\rho^2} - \frac{\rho\nu T}{1+\rho^2} + \frac{\rho}{\sigma_\epsilon^2}(y_1-\mu)^2 + \frac{1}{\sigma_\epsilon^2}\sum_{t\in\mathcal{I}}\big[(y_{t+1}-\mu)-\rho(y_t-\mu)\big](y_t-\mu)$$
$$+\frac{1}{\sigma_\epsilon^2(1+\rho^2)^2}\Big\{\sum_{t\in\mathcal{B}} 2\rho(1+\rho^2)\big[(y_{t+2}-\mu)-\rho^2(y_t-\mu)\big](y_t-\mu)$$
$$+\sum_{t\in\mathcal{B}}\rho\big[(y_{t+2}-\mu)-\rho^2(y_t-\mu)\big]^2\Big\} \tag{3.74}$$

The direct verification, of course, gives that the expected value of each of the score equations is zero. The MLEs of the parameters can be obtained by solving the score equations (3.72)–(3.74), or by maximization of (3.71).

The implementation of the (exact) EM algorithm would imply computation of the conditional expectations of the relevant functions of the missing data. Suppose y_{t+1} is missing, and it is bordered by observed y_t, y_{t+2}:

$$\mathbb{E}[y_{t+1}|y_t, y_{t+2}, \theta] = \mu + \frac{\rho}{1+\rho^2}[(y_t-\mu)+(y_{t+2}-\mu)] \tag{3.75}$$

from (3.49),

$$\mathbb{E}[y_{t+1}^2|y_t, y_{t+2}, \theta] = \Big\{\mathbb{E}[y_{t+1}|y_t, y_{t+2}, \theta]\Big\}^2 + \mathbb{V}[y_{t+1}|y_t, y_{t+2}, \theta],$$
$$\mathbb{V}[y_{t+1}|y_t, y_{t+2}, \theta] = \frac{(1-\rho^2)^2\sigma_\epsilon^2}{1+\rho^2} \tag{3.76}$$

from (3.50), and

$$\mathbb{E}[y_t y_{t+1}|y_t, y_{t+2}, \theta] = y_t\,\mathbb{E}[y_{t+1}|y_t, y_{t+2}, \theta],$$
$$\mathbb{E}[y_{t+1} y_{t+2}|y_t, y_{t+2}, \theta] = y_{t+2}\,\mathbb{E}[y_{t+1}|y_t, y_{t+2}, \theta], \tag{3.77}$$

These expressions can now be used in the E-step in computing the expected values of the sufficient statistics in Section 3.1. For instance, equation (3.14) becomes

$$0 = (1-\rho)\,\mathbb{E}\big[\sum_{t=2}^{T-1}(y_t-\mu)|\text{observed data}\big] + (y_1-\mu) + (y_T-\mu) =$$
$$= (1-\rho)\big[\sum y_t + \frac{\rho}{1+\rho^2}\sum(y_{t-1}+y_{t+1}-2\mu) - (T-2)\mu\big] + (y_1-\mu) + (y_T-\mu) \tag{3.78}$$

where the first sum in the last line is over the available observations, and the second one, over the missing observations. This form of the estimating equation is much clearer than equivalent (3.72).

3.6 AR(1) with many gaps: approximate EM

Let us utilize our approach of using unconditional estimates for the sufficient statistics of the missing data. Then, denoting the approximate expectations by $\tilde{\mathbb{E}}$,

$$\begin{aligned}\tilde{\mathbb{E}}[y_{t+1}|\theta] &= \mu,\\ \tilde{\mathbb{E}}[(y_{t+1}-\mu)^2|\theta] &= \frac{\sigma_\epsilon^2}{1-\rho^2},\\ \tilde{\mathbb{E}}[(y_t-\mu)(y_{t+1}-\mu)|\theta] &= \frac{\rho\sigma_\epsilon^2}{1-\rho^2}\end{aligned} \tag{3.79}$$

Substituting those back to (3.9), we obtain

$$\begin{aligned}\ln\tilde{L}(\theta;\mathbf{y}) =& -\frac{1}{2}\big(T\ln 2\pi\sigma_\epsilon^2 + \ln(1-\rho^2)\big) + \frac{\rho^2}{2\sigma_\epsilon^2}\big((y_1-\mu)^2+(y_T-\mu)^2\big)\\ &- \frac{1}{2\sigma_\epsilon^2}\Big\{(1+\rho^2)\sum_{t\in I}(y_t-\mu)^2 + (1+\rho^2)\frac{\sigma_\epsilon^2\nu T}{1-\rho^2}\\ &- 2\rho\sum_{t\in\mathcal{I}^*}(y_t-\mu)(y_{t-1}-\mu) - 2\rho\frac{\rho\sigma_\epsilon^2 2\nu T}{1-\rho^2}\Big\} =\\ =& -\frac{1}{2}\big(T\ln 2\pi\sigma_\epsilon^2 + \ln(1-\rho^2)\big) - \frac{\nu T(1-3\rho^2)}{2(1-\rho^2)}\\ &- \frac{1}{2\sigma_\epsilon^2}\Big[(1+\rho^2)\sum_{t\in I}(y_t-\mu)^2 - \rho^2\big((y_1-\mu)^2+(y_T-\mu)^2\big)\big)\\ &- 2\rho\sum_{t\in\mathcal{I}^*}(y_t-\mu)(y_{t-1}-\mu)\Big]\end{aligned} \tag{3.80}$$

where $t \in \mathcal{I}^*$ are the points for which neither y_{t-1} nor y_{t+1} is missing. We made an assumption earlier that only one of the two can be missing. $|\mathcal{I}^*| = T - 2M = T(1-2\nu)$, so the last term in the first equality has a factor of $2\nu T$.

The derivatives of this objective function yield the following estimating equations.

$$\begin{aligned} 0 = \frac{\partial \ln \tilde{L}}{\partial \mu} = \frac{1}{\sigma_\epsilon^2}\Big[(1+\rho^2)\sum_{t\in I}(y_t-\mu) - \rho^2\big((y_1-\mu)+(y_T-\mu)\big) \\ - \rho\sum_{t\in\mathcal{I}^*}(y_t-\mu)+(y_{t-1}-\mu)\Big] = 0 \end{aligned} \tag{3.81}$$

This equation gives a consistent estimate of μ as its expected value is zero.

$$\begin{aligned} \frac{\partial \ln \tilde{L}}{\partial \rho} = \frac{\rho}{1-\rho^2} - \nu T\frac{2\rho}{(1-\rho^2)^2} - \frac{1}{\sigma_\epsilon^2}\Big[\rho\sum_{t\in I}(y_t-\mu)^2 \\ - \rho\big((y_1-\mu)^2+(y_T-\mu)^2\big)\big) - \sum_{t\in\mathcal{I}^*}(y_t-\mu)(y_{t-1}-\mu)\Big] = 0 \end{aligned} \tag{3.82}$$

Dividing through by T and taking probability limits, we obtain:

$$-\nu\frac{2\operatorname{plim}\hat{\rho}_{\text{im:aEM}}}{(1-\operatorname{plim}\hat{\rho}^2_{\text{im:aEM}})^2} - \frac{1}{\operatorname{plim}\hat{\sigma_\epsilon^2}}\Big[\operatorname{plim}\hat{\rho}_{\text{im:aEM}}\,(1-\nu)\sigma_\epsilon^2 - (1-2\nu)\rho\sigma_\epsilon^2\Big] = 0 \tag{3.83}$$

The subindex im:aEM shows that this is the case of *i*ncomplete data with *m*any gaps, estimation by the *a*pproximate *EM* algorithm. We would also need to derive the estimate of σ_ϵ^2 and analyze the two limits together. If we can get a consistent estimate of σ_ϵ^2, then (3.83) shows inconsistency of $\hat{\rho}$, as the equation (3.83) becomes

$$\nu\frac{2\operatorname{plim}\hat{\rho}}{(1-\operatorname{plim}\hat{\rho}^2)^2} + (1-\nu)\operatorname{plim}\hat{\rho} - (1-2\nu)\rho = 0 \tag{3.84}$$

Clearly, the $\operatorname{plim}\hat{\rho}_{\text{im:aEM}} \neq \rho$ unless $\nu = 0$. The expansion by ν near zero gives

$$\begin{aligned} \operatorname{plim}\hat{\rho}_{\text{im:aEM}} &= \rho + A\nu + B\nu^2 + o(\nu^2), \\ A &= -\frac{2+(1-\rho)^2}{(1-\rho)^2}\rho, \\ B &= \frac{A}{(1-\rho)^2}\big[(1-\rho)^2 + 2A(1-\rho) + 2\rho(1-\rho) - 2\big] \end{aligned} \tag{3.85}$$

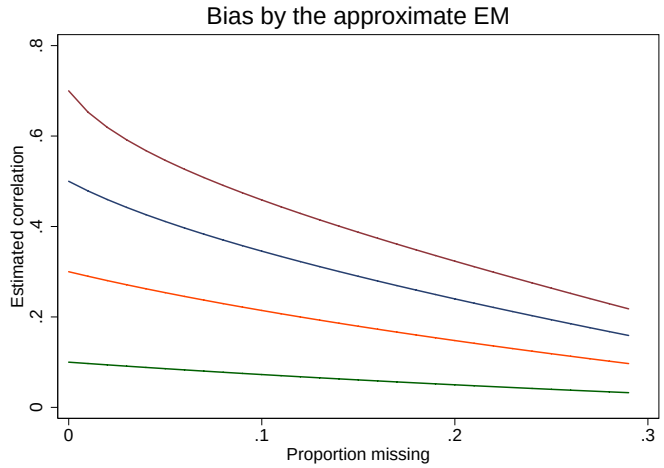
Fig. 3.1 characterizes the resulting bias of the estimator $\hat{\rho}_{\text{im:aEM}}$. The probability limits of (3.84) are understating the true correlation, and the ratio of the estimate to the true ρ primarily depends on ν rather than on ρ.

If we further take expansion over ρ near 0,

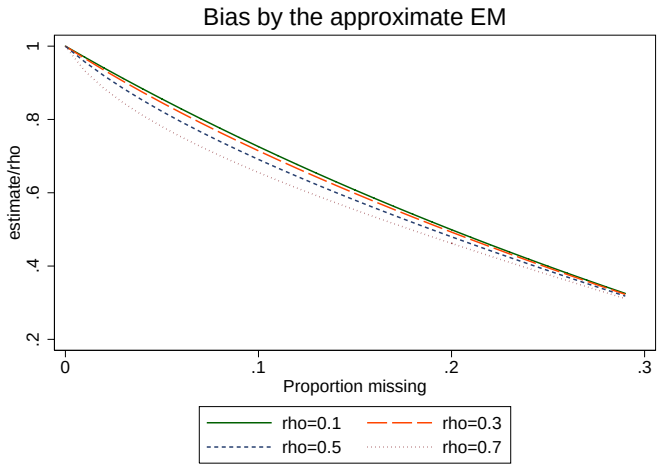
$$\operatorname{plim} \frac{\hat{\rho}}{\rho} = 1 - 3\nu + o(\nu) + o(\rho), \tag{3.86}$$

so the correction for the bias can be made as

$$\tilde{\rho} = \hat{\rho}(1 - 3\nu)^{-1} \tag{3.87}$$



(a) Absolute bias



(b) Relative bias

Figure 3.1: plim $\hat{\rho}_{\text{im:aEM}}$ from the approximate EM algorithm.

Finally, the estimating equation for σ_ϵ^2 is

$$0 = \frac{\partial \ln \tilde{L}}{\partial \sigma_\epsilon^2} = \frac{T}{2\sigma_\epsilon^2} - \frac{1}{2\sigma_\epsilon^4}\Big[(1+\rho^2)\sum_{t\in I}(y_t-\mu)^2 - \rho^2\big((y_1-\mu)^2+(y_T-\mu)^2\big)\Big) - 2\rho\sum_{t\in\mathcal{I}^*}(y_t-\mu)(y_{t-1}-\mu)\Big] \tag{3.88}$$

or

$$\hat{\sigma_\epsilon^2}_{\text{im:aEM}} = \frac{1}{T}\Big[(1+\hat{\rho}^2_{\text{im:aEM}})\sum_{t\in I}(y_t-\mu)^2 - \hat{\rho}^2_{\text{im:aEM}}\big((y_1-\mu)^2+(y_T-\mu)^2\big)\Big) - 2\hat{\rho}_{\text{im:aEM}}\sum_{t\in\mathcal{I}^*}(y_t-\mu)(y_{t-1}-\mu)\Big] \tag{3.89}$$

The probability limit of this estimate is

$$\operatorname{plim}\hat{\sigma_\epsilon^2} = (1+\operatorname{plim}\hat{\rho}^2_{\text{im:aEM}})(1-\nu)\frac{\sigma_\epsilon^2}{1-\rho^2} - 2\operatorname{plim}\hat{\rho}_{\text{im:aEM}}\,(1-2\nu)\frac{\rho\sigma_\epsilon^2}{1-\rho^2} \tag{3.90}$$

so even if ρ is estimated consistently,

$$\operatorname{plim}\hat{\sigma_\epsilon^2}_{\text{im:aEM}} = \sigma_\epsilon^2\Big[(1-\nu)+2\nu\frac{\rho^2}{1-\rho^2}\Big] \tag{3.91}$$

which gives the right answer only for $\rho^2 = 1/3$. The remedy may be found by noting that the term in the curly brackets of (3.80) divided by T gives a consistent estimate of σ_ϵ^2 provided that ρ can be consistently estimated by some $\tilde{\rho}$, too:

$$\begin{aligned}\tilde{\sigma_\epsilon^2} &= \frac{1}{T}\Big\{(1+\tilde{\rho}^2)\sum_{t\in I}(y_t-\hat{\mu})^2 + (1+\tilde{\rho}^2)\frac{\tilde{\sigma_\epsilon^2}\nu T}{1-\tilde{\rho}^2} \\ &\quad - 2\tilde{\rho}\sum_{t\in\mathcal{I}^*}(y_t-\hat{\mu})(y_{t-1}-\hat{\mu}) - 2\tilde{\rho}\frac{\tilde{\rho}\tilde{\sigma_\epsilon^2}2\nu T}{1-\tilde{\rho}^2}\Big\} = \\ &= \Big\{(1+\tilde{\rho}^2)\sum_{t\in I}(y_t-\hat{\mu})^2 - 2\tilde{\rho}\sum_{t\in\mathcal{I}^*}(y_t-\hat{\mu})(y_{t-1}-\hat{\mu})\Big\}\Big/ \\ &\quad T\Big\{1-\nu\frac{1-3\tilde{\rho}^2}{1-\tilde{\rho}^2}\Big\}\end{aligned} \tag{3.92}$$

So the estimate of σ_ϵ^2 is better obtained from the (generalized) residual sum of squares $\breve{Q}$ rather than from the maximization procedure itself.

That said, we need to come up with the way of correcting bias in the estimation of the correlation coefficient. Let us look back at (3.84) and rewrite it as

$$-\nu\frac{2\operatorname{plim}\hat{\rho}_{\text{im:aEM}}}{(1-\operatorname{plim}\hat{\rho}^2_{\text{im:aEM}})^2}-\nu\rho=(1-\nu)(\operatorname{plim}\hat{\rho}_{\text{im:aEM}}-\rho) \tag{3.93}$$

Integrating the left hand side, we get

$$-\nu\int[\frac{2\rho}{(1-\rho^2)^2}+\rho]\,d\rho=-\nu\frac{1}{1-\rho^2}+\nu\frac{1-\rho^2}{2} \tag{3.94}$$

Thus, if the "penalty" term

$$\mathcal{P}(\rho)=\nu T\Big[\frac{1}{1-\rho^2}-\frac{1-\rho^2}{2}\Big] \tag{3.95}$$

is added to the likelihood (3.80), then the following procedure will yield consistent estimates:

Penalized approximate ECM algorithm:

1. Initialize the estimates in some reasonable way (say $\mu^{(0)}=\bar{y}$, $\rho^{(0)}=0$, $\sigma_\epsilon^{2(0)}=s^2$)
2. Update $\mu^{(j)}$ by (3.81)
3. Update $\rho^{(j)}$ by maximizing $\ln\tilde{L}(\theta;\mathbf{y})+\mathcal{P}(\rho)$ from (3.80) and (3.95) conditional on $\mu^{(j)}$, $\sigma_\epsilon^{2(j-1)}$ by numerical maximization
4. Update $\sigma_\epsilon^{2(j)}$ by (3.92)
5. $j\leftarrow j+1$, iterate steps 2-4 until convergence

3.7 Conclusion

Let us summarize the main results of this chapter. We have derived the maximum likelihood estimates for the AR(1) process with missing data, and demonstrated that the (exact) EM algorithm produces the same estimating equations as those implied by the MLE.

The approximate EM algorithm where the unconditional expectations are taken at the E step faced problems. The estimation of the mean parameter μ does not pose any specific concerns, as even such simple estimator as the sample mean is going to be unbiased and consistent no matter what the correlation structure is. The estimation of the overall variance parameter σ_ϵ^2 can also be performed consistently from the generalized sum of squares rather than from the likelihood maximization procedure given that a consistent estimate of the correlation structure can be found.

The estimation of the correlation structure parameter ρ from the basic approximate EM is the most difficult part, as neglection of the correlation structure is the main loss of information incurred by using the approximation in the approximate EM algorithm. Generally, it leads to an inconsistent estimate in non-trivial cases when the correlation is not zero and the proportion of missing data is not zero. This inconsistency can be corrected in a number of ways. First, a penalty term can be added to the likelihood to compensate for the bias, and the resulting procedure gives consistent estimates of ρ. Second, a simple correction for the proportion of missing data can be performed that would make the estimate approximately consistent. Both of those ways require knowing the data generating process, and being able to derive the properties of the estimating equations for the approximate EM algorithm. Neither of the corrections, however, is intuitive enough to lend itself easily for the spatio-temporal process we are most interested in.

A number of other possible implications for the spatio-temporal processes may be put further. First, the (exact) EM algorithm that involves kriging is equivalent to the maximum likelihood, so depending on the availability and versatility of the general maximization routine in someone's favorite software and kriging procedures, either method can be chosen.

Second, the approximate EM algorithm, although easier to implement, needs attention, at least in the estimation of the parameters affecting the correlation structure. Most likely, it needs to be supplemented with correction terms whose structure needs to be derived analytically. The need for and availability of such corrections in the context of dissociated processes is the topic of Chaper 5.

Chapter 4

Application to the spatio-temporal modelling

This chapter applies the EM algorithm and its modifications discussed in Chapter 2 to the EPA data set on the particulate matter. It is based on Smith et al. (2003) and earlier drafts of it. The substantive output of the paper are the maps of the estimated and kriged $PM_{2.5}$ concentrations for the states of Georgia, North Carolina, and South Carolina that showed that all of them are at risk of violating the new EPA standard on $PM_{2.5}$.

Section 4.1 describes interest in the particulate matter, the main measurement issues and the EPA standards. Section 4.2 describes the data and poses the research questions. Section 4.3 presents the semiparametric model that accounts for trends in space and time, as well as for the residual spatial covariance. Then Section 4.4 briefly reviews the EM algorithm, and shows how it can be applied in our setting. Section 4.5 presents the estimation results and discusses kriging to obtain maps of the $PM_{2.5}$ concentrations, and Section 4.6 concludes

4.1 Particulate matter

Airborne particulate matter has become an important topic of epidemiological and environmental studies in the last decade when it was understood that particulate matter is an important determinant of deaths, especially in the elderly, even though the biological mechanisms of its effect are not quite clear yet. The United States Environmental Protection Agency regulates the admissible levels of PM_{10} and $PM_{2.5}$, the indicators of the concentration of the particulate matter of sizes 10 and 2.5 μm, re-

spectively[1]. The federal standard for $PM_{2.5}$, the particulate matter size studied in this paper, was introduced in 1997. The long term exposure part states that the 3-year average of annual arithmetic mean $PM_{2.5}$ concentrations from single or multiple community-oriented monitors should not exceed $15\mu g/m^3$. The extreme exposure part states that the 3-year average of the 98-th percentile of 24-hour $PM_{2.5}$ concentrations at each population-oriented monitor within an area should not exceed $65\mu g/m^3$ (EPA (1997*b*)). (EPA 1997*b*). The standard had a thorny path towards its implementation. It was immediately rebutted by the industrial lobby, and in May 1999, a panel of the U.S. Court of Appeals for the D.C. Circuit, in a split decision, held that the Clean Air Act was unconstitutional as an improper delegation of legislative authority to EPA. The EPA appealed the decision to the U.S. Supreme Court, and in February 2001, the latter upheld EPAs authority to set national air quality standards. In March 2002, following the Supreme Court decision on the constitutional issues, the Court of Appeals rejected all remaining challenges to the 1997 standards.

The EPA has also outlined a number of research topics related to the particulate matter, and one of the statistical questions raised is, "Can spatial interpolation methods provide more accurate estimates of individual exposures to particulate air pollution?" (Cox 2000).

A further step can be made to incorporate the temporal dimension of the data, especially as long as this is the natural way data comes from monitoring stations. We show that the data can be thought of as independent over time, and propose to refer to this type of data as *dissociated models*. Mathematically, they have a lot in common with repeated measurement / panel data, but in the latter, the important correlations are over time within the same unit, while in our application, the important correlations are those in space across units, while the time correlations are negligible.

4.2 The data

The data used in this research are a part of the EPA data set for 1999 on the monitors of particulate matter[2]. The total number of continental US monitors in the data set is 780. The measured variable is the concentration of the particles with aerodynamic diameter less than 2.5 microns ($PM_{2.5}$). The observation frequencies generally vary from site to

[1] The definition of the $PM_{2.5}$ is the particle size at which 50 per cent of the particles of this size (aerodynamic diameter) are collected by the monitoring device (Cox 2000).

[2] The data were provided by David Holland of EPA.

site. The majority of sites have observations recorded once in three days; there are some that have daily records, and there are some that only have a few observations for the whole period. The characteristics of the monitor itself include the geographic position (latitude and longitude), the area type as a combination of two categorical factors, urbanization (rural, urban, suburban) and land use (agricultural, industrial, commercial, residential, forest)[3], altitude of the monitor, the testing method, and some other technical information.

We only used a fraction of this rich data set related to North Carolina, South Carolina, and Georgia. There were 74 monitors across those states (23 in Georgia, 31 in North Carolina, 20 in South Carolina). The map of the monitors is given on Fig. 4.1. Heavily populated metropolitan areas of Atlanta and Charlotte have clusters of closely located observations sites. No data are available for Georgia in the fourth quarter of the year. The data were further aggregated into weekly averages to reduce temporal autocorrelation and reduce the fraction of the missing data. Some biases might have been introduced at this stage due to the day of the week effect[4].

We ended up with 2613 observations. The proportion of missing data is rather high at 27.9%: compare the above figure with $74 \times 49 = 3626$ observations that should be in the complete data set.

4.3 The spatio-temporal model

There are several components of the spatio-temporal model in Smith et al. (2003). Some of them are also similar to the approach in Holland, De Oliveira, Cox & Smith (2000).

The Box-Cox transformation (Box & Cox 1964) was used to combat skewness of the original data and stabilize variance. The chosen transform was the square root of the original data based on an exploratory analysis of the variance trends.

The mean of the spatio-temporal process at each station was modelled as a generalized additive model (Hastie & Tibshirani 1990) with components representing the time trend, the spatial trend, and the land use (individual characteristic of the monitor). The time trend was initially modelled by the B-splines (Green & Silverman 1994), which is

[3] Some cells are empty: there are no combinations of forest and urban or suburban, as well as agricultural and urban.

[4] The $PM_{2.5}$ concentrations are generally lower on the weekends when there is not as much industrial activity and traffic as during the business days, so if the weekends were under- or overrepresented in a given week, then the weekly average would be biased up- or downwards.

an expansion of a flexible spline function through piecewise cubic basis functions, with the coefficients of those functions that can be estimated by OLS or maximum likelihood in more general GAM context. The between week variability was found to be so high that the best fit was provided by the saturated model with weekly dummy variables. One can also think of this as a saturated model with 49 B-spline basis function, by the number of the weeks available in the data. The time trend picked the effect of hurricane Floyd that had a devastating effect on North Carolina in September 1999. It showed up as a sharp twofold increase of the $PM_{2.5}$ concentrations that subsided about six weeks later. This natural experiment gives a rough estimate of the relative scope of natural and anthropogenic particulate matter processes.

The space trend was also estimated in a semiparametric fashion with the bivariate splines that allow expansion by the *thin plate spline* basis (Green & Silverman 1994). In both of those expansions, the smoothness of the spline is controlled by the number of knots, or spline centers. The number and location of nodes for the spatial trend J required certain decision making. Holland et al. (2000) use k-means clustering to reduce the number of nodes and analyze the data by region. Smith et al. (2003) use clustering to reduce the number of nodes and thus control the smoothness of the

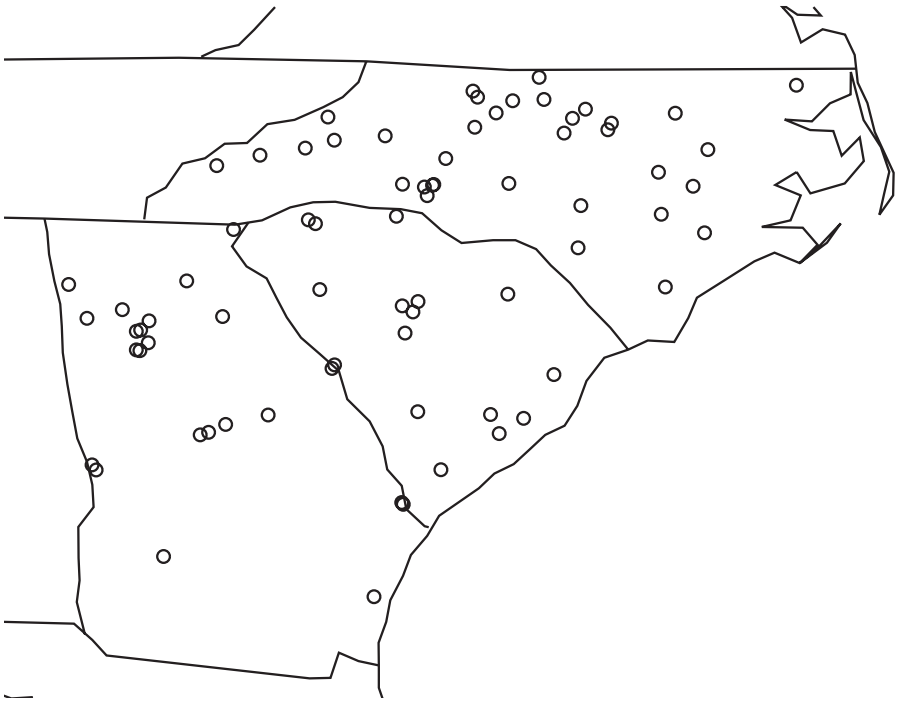


Figure 4.1: Monitor locations in the data set.

spatial trend. From the models with $J = 10, 20, 30, 40, 50$, the model with $J = 20$ was chosen according to the information criteria as compromise between AIC and BIC (Akaike 1973, Schwarz 1978). Finally, the last component of the generalized additive model was the set of indicators of the land use, the four dummies corresponding to the agricultural, commercial, forest, and industrial vicinity of the monitor, with the base category of the residential land use. Those variables are highly jointly significant.

Smith et al. (2003) then proceeded to the analysis of temporal and spatial correlations. Due to averaging of the observations over a week period, the temporal correlations were found to be insignificant.

The spatial distance between two sites was defined as the geodesic distance, or the length of great-circle arc (the shortest route between the two points on the sphere). The spatial correlations were found to be non-stationary as evidenced by the increasing

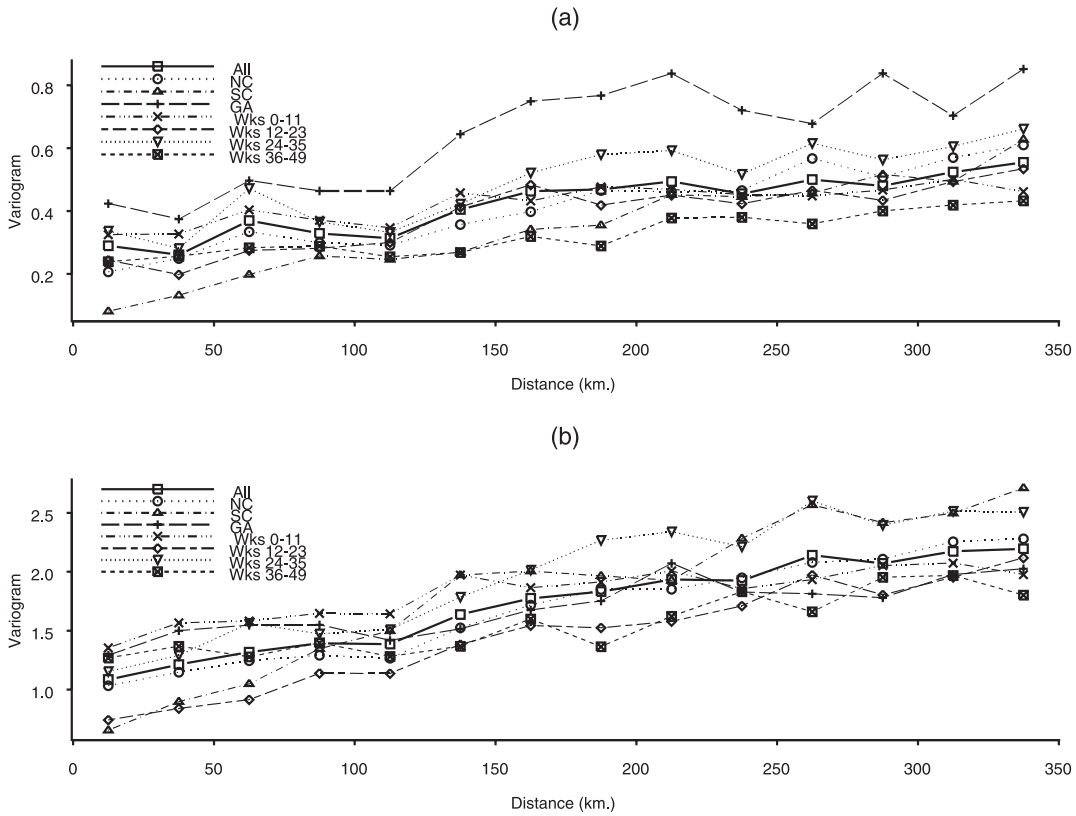


Figure 4.2: Empirical variograms for residuals after fitting the time trend, spatial trend, and type effects. All data are combined, and there are separate plots by state and by season: (a) without standardizing variances and (b) after standardizing the sample variance of residuals at each station to be 1. Fig. 5 of Smith et. al. (2003).

variogram that does not flat off at any finite sill (Fig. 4.2; c.f. Fig. 2.1). No substantial differences were found across the variograms for different states (NC, SC, GA) or across different time periods (length of 12 weeks). Thus in the maximum likelihood estimation, a non-stationary variogram for an intrinsically stationary process was chosen:

$$\gamma(h) = \begin{cases} 0, & h = 0 \\ \alpha\big(\theta_1 + h^{\theta_2}\big), & h > 0 \end{cases} \tag{4.1}$$

It shows the nugget effect θ_1, the overall variance scaling constant α and the shape parameter $0 < \theta_2 < 2$.

4.4 Estimation

Overall, the resulting generalized additive model looks like

$$y_{it} = g_t(t) + g_s(\mathbf{s}_i) + x_i^{\text{use}}\beta_{\text{use}} + \epsilon_{it} \equiv x_{it}\beta + \epsilon_{it} \tag{4.2}$$

where $g_t(\cdot)$ and $g_s(\cdot)$ are the temporal and spatial trends, that are in turn can be represented as a sum of basis function with some coefficients. The generalized additive model parameters can be stacked into β, and all of the design (splines and land use) variables can be stacked into x. If the additional assumption that the errors ϵ_{it} follow a suitable multivariate normal distribution is made, then the model can be estimated by maximum likelihood or equivalent procedures (exact EM algorithm).

An additional complication arises since the spatial process ϵ_{it} was found to be non-stationary. As was mentioned in Section 2.1.3, the linear combinations that have finite variances are the contrasts. Smith et al. (2003) subtracted the weekly average

$$\bar{y}_{\cdot t} = \frac{1}{n_t} \sum_{i \in \mathcal{I}_t} y_{it} \tag{4.3}$$

where $\mathcal{I}_t$ is the set of observations made at time t, $|\mathcal{I}_t| = n_t$. That way, the time trend was essentially taken out of the model, as the quantity subtracted corresponds to the individual week effect.

Even with the presence of missing data, the likelihood for the vector of observations y_t taken at time t can be written down as

$$l(\theta|y_t) = (2\pi)^{-n_t/2}|\Sigma_t(\theta)|\times$$
$$\exp\big[-\frac{1}{2}(y_t - x_t\beta_t)\Sigma_t(\theta)^{-1}(y_t - x_t'\beta_t)\big] \tag{4.4}$$

where the subindex t indicates that observations from different sites are available at different points in time, so the dimensions n_t of the measured $PM_{2.5}$ concentration y_t, the explanatory variables x_t, the vector of coefficients β_t, and the covariance matrix $\Sigma_t(\theta)$ are changing from one week to another, according to the number of available sites n_t. The overall likelihood can be obtained as a product of likelihoods of the form (4.4) once independence over time is assumed (and eventually tested).

The direct likelihood maximization thus involves identifying T matrices $\Sigma_t(\theta)$, extracting them from the "master" matrix $\Sigma(\theta)$, inverting them and computing their determinants. The latter two stages can be combined by the means of Cholesky decomposition. Computing many determinants and the inverse matrices is likely to be time consuming, as either of them is an $O(k^3)$ operation, where k is the dimension of the matrix, so other alternatives might be sought. One such alternative is the EM algorithm (see Section 2.3, as well as Dempster et al. (1977), Little & Rubin (2002) and McLachlan & Krishnan (1997)).

Apparently, only the response variable is missing, which is the measurement of the $PM_{2.5}$ concentrations, $\mu g/m^3$. All the design variables are observed perfectly. In using the EM algorithm, we implicitly assume that the data are missing at random. This assumption would be violated if an observation is not registered when the observed value is too high or too low, which may be the case if the measurements were outside the measurement range of a monitor.

For the version of the algorithm we used, the maximization step was split into two steps each maximizing the likelihood over a partition of the parameter space. This is known as the *expectation-conditional maximization*, or ECM, algorithm. It possesses the generic convergence properties of the EM-algorithm, too. At the first stage, the log likelihood was maximized over the covariance matrix parameters subspace (θ_1, θ_2, α of (4.1)) with fixed values of the additive model parameters β. Then at the second stage of the M step, a GLS regression model was estimated with the current covariance matrix estimate thus optimizing over the regression parameters subspace.

As in many implementations of the E-step where a sufficient statistic can be found, we computed the expectation of the $\sum_t e_t e_t'$ where e_t is the vector of the complete-

data residuals, conditional on the observed values of the variables involved, and on the current parameter values. In particular, at the h-th iteration, the E-step predicted the fitted values for the GLS regression, and calculated the current step EM predictions of the Y's and the residuals.

$$\tilde{Y}_{EM\,fit} = \begin{cases} Y_{obs}, & Y \text{ is non-missing} \\ x'\hat{\beta}^{(h)}_{GLS}, & Y \text{ is missing} \end{cases} \tag{4.5}$$

Then the residuals $\tilde{e}_{it} = \tilde{Y}_{EM\,fit} - x'\hat{\beta}^{(h)}_{GLS}$ were extracted, and their approximate conditional expectation was computed:

$$\tilde{\mathbb{E}}e_{it}e_{jt} = \begin{cases} \tilde{e}_{it}\tilde{e}_{jt}, & Y_{it}, Y_{jt} \text{ are both non-missing} \\ \sigma_{ij}(\theta), & \text{at least one of } Y_{it}, Y_{jt} \text{ is missing} \end{cases} \tag{4.6}$$

where $\sigma_{ij}(\theta)$ is the i,j-th entry of the spatial covariance matrix $\Sigma(\theta)$ evaluated at the current values of the parameters.

This is an approximate version of the EM algorithm. In computing the conditional expectations of the missing data, we only use the GAM parameters, and ignore the variance parameters. In computing the second moments of the missing data, we only use the variance parameter estimates, but not the available data. As discussed in Chapters 1 and 3, the exact implementation of the EM algorithm would require kriging to use all of the available information, and thus we would have to go back to inverting many matrices $\Sigma_t(\theta)$ losing all potential computational efficiency gains.

The process iterated until convergence: the approximate conditional expectation of the sufficient statistic ee' is calculated, where the current estimates of the covariances are used when the residuals (or Y's) are missing; the maximization over the variance parameter subspace is performed; GLS regression is run, and so on. The estimation procedure was coded in Stata software (Stata Corp. 2001, Kolenikov 2001) and in Fortran.

The starting values of the parameters for the algorithm are the available case OLS regression results for the regression part of the parameter vector, and some "reasonable" guesses for the covariance part.

4.5 Results

Smith et al. (2003) performed the estimation based on both the approximate EM al-

Table 4.1: Comparison of the approximate EM and ML estimates.

Method	θ_1	θ_2	α
MLE			
Point estimate	2.06	0.92	0.061
Standard error	0.35	0.097	0.0017
EM			
Point estimate	2.13	0.92	0.049
Standard error	0.29	0.083	0.0012
Corrected s.e.	0.35	0.098	0.0019

gorithm and full likelihood maximization. The comparison of the results is given in Table 4.1.The nugget and shape parameters are estimated quite well by the approximate EM algorithm, while the variance parameter is underestimated. This is in rough correspondence with the results in Chapter 3 that showed biases in parameter estimates by the (uncorrected) approximate EM algorithm.

The EM algorithm *per se* does not give the standard errors, but as far as the parameters of the trend and of the variance subspace are independent in the normal model, the estimates that were coming from maximization of $Q(\cdot|\cdot)$ in the M-step of the approximate EM algorithm should give some idea of the sampling variability. Somewhat better standard errors are obtained from the following argument. The information contained in the full data can be thought of as

$$I_{\text{complete}} = I_{\text{observed}} + I_{\text{missing}} \tag{4.7}$$

If the matrices are proportional to each other (which would be true if the missing data process is MCAR, and is used here only an assumption to derive a working approximation), then

$$(1 - \nu) I_{\text{complete}} = I_{\text{observed}} \tag{4.8}$$

Hence, the information is overestimated by a factor of $1 - \nu = 0.721$, and the standard errors should be multiplied by $1/\sqrt{0.721}$ to correct for the missing information. Those are reported in the last line of Table 4.1 and show surprisingly good correspondence to the errors obtained from the maximum likelihood.

After the parameters of the complete model (4.2) were estimated, Smith et al. (2003) proceed to kriging of the spatial field for different points in time, and for the average over the year.

As is readily seen, the universal kriging formula (2.9) is equivalent to the following:

$$\hat{y}_0 = x_0^T \hat{\beta} + \tau^T \Sigma^{-1} e \tag{4.9}$$

where $\hat{y}_0$ is the best linear prediction at the point characterized by the regressors x_0; $\hat{\beta}$ is the GLS estimate of the regression coefficient of the process $Y = X\beta + \nu$, $\operatorname{Cov} \nu = \Sigma$, so that $x_0^T \hat{\beta}$ is the linear fit, or the trend term, from the model; τ is the vector of cross-covariances between the observed values of the field Y and the unobserved y_0 given by the spatial model; and e is the residuals of the process from the fitted linear regression: $e = Y - X\hat{\beta}$.

The universal kriging prediction fully incorporates the available information on the parameters estimates. We only used the first term of (4.9) in the prediction of the missing data in the E-step because otherwise we would end up with different size matrix inversion operations that we tried to avoid by using the approximate version of the EM-algorithm.

To implement kriging following the ML (or EM) estimation, the estimates $\hat{\beta}, \hat{\Sigma}$ obtained at the last iteration can be used. It should be justly mentioned that the predicted variances in this case will understate the true variability as long as the parameters of the variogram are treated as fixed rather than as estimated.

As far as the only time-varying part of the model is the time trend, the estimated fields at different points in time differ by the overall shift, plus residual fluctuations. We added back the week effects (4.3) once the spatial prediction by kriging was obtained. Also, the choice of a single land use variable had to be made for the GAM part of (4.2), and Smith et al. (2003) used residential areas since this is where people tend to spend most of their time thus receiving the major part of their $PM_{2.5}$ exposure. It should be noted that only sites with commercial land use had higher average estimated levels of $PM_{2.5}$. The resulting maps are shown on Fig. 4.3

4.6 Conclusions

The paper Smith et al. (2003) has proposed and exemplified the use of likelihood based methods in the presence of missing data in the generalized additive model framework, with trends accounting for (most of the) variation in space and time, as well as across the sites in different area types. The estimation of the parameters is done through a version of the EM-algorithm to correct for missing data in the longitudinal data sets.

The procedure helped to use a large fraction of data even though the monitoring stations reported the data infrequently. The spatial field was found to be non-stationary. Kriging identified the problematic areas of Charlotte and Atlanta, and also demonstrated the effects of hurricane Floyd that swept through South-Eastern United States in September 1999.

The substantive results imply that the three analyzed states (Carolinas and Georgia) are in danger of violating the federal standard on $PM_{2.5}$, except for the coastal areas, and Appalachians. This suggests the anthropogenic origins of the $PM_{2.5}$. (This statement might be attenuated by the fact that the monitors might have been located in the "problematic" areas that are known to have polluted air, so that the design of the monitoring network is biased toward higher $PM_{2.5}$ concentrations.)

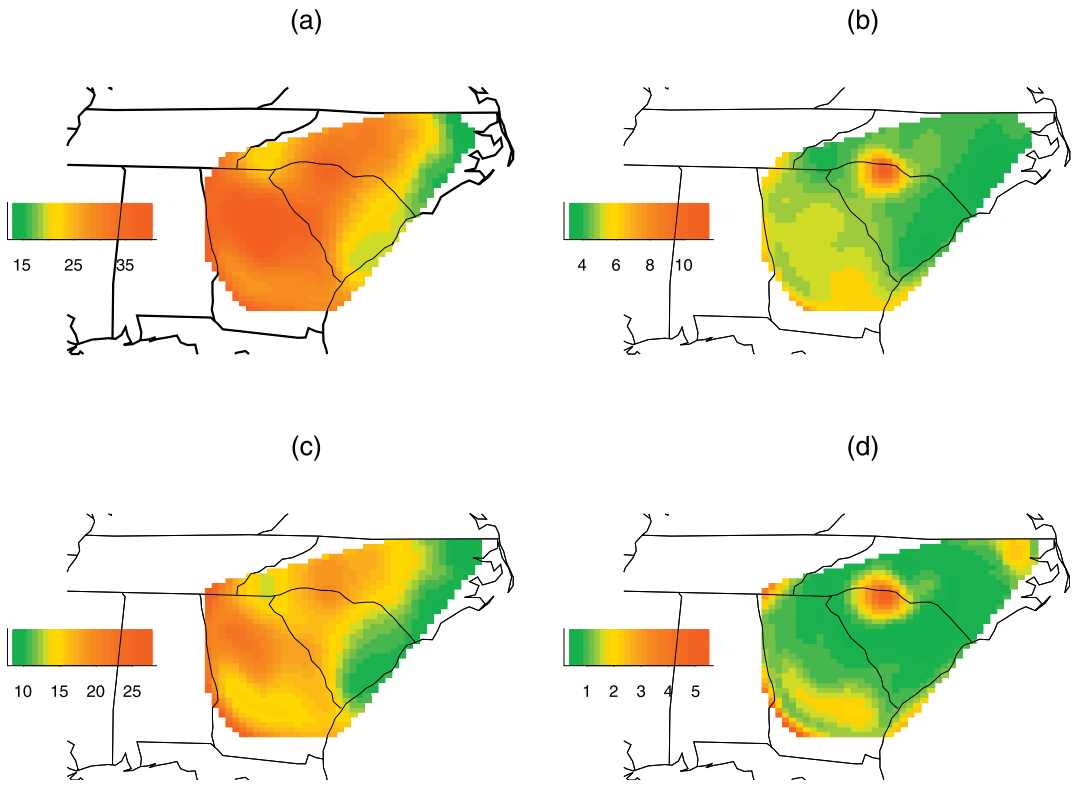


Figure 4.3: Plots of the predicted surface for $PM_{2.5}$ (Fig. 6 of Smith et. al. (2003). (a) Predicted surface for week 33. (b) Estimated prediction standard error for week 33. (c) Predicted surface for average of weeks 1-49. (d) Estimated prediction standard error for average of weeks 1-49.

Chapter 5

Dissociated processes

In this chapter, we shall consider *dissociated* processes where spatially correlated measurement are taken at multiple points in time, and observations are assumed to be independent over time. The derivations are carried out in the most general form using matrix formulations, and a number of matrix formalisms are necessary. Section 5.1 introduces incidence matrices that describe the patterns of observed and missing data. Section 5.2 deals with the MLE estimation. Section 5.3 introduces the approximate EM algorithm formulas and deals with the biases in the estimating equations. Corrections restoring unbiasedness of the estimating equations are proposed in Sections 5.3.5–5.3.7. The following two sections deal with the derivatives and variances of the estimating equations, and the next two sections use those as inputs to establish consistency (Section 5.6) and asymptotic normality of the estimates (Section 5.7). The derivatives and variances then become the components of the information sandwich estimator of the asymptotic covariance matrix of the estimates. Section 5.9 summarizes the results.

Several Appendices are also used in the derivations of this chapter. Appendix A introduces matrix calculus, defines matrix differentials and provides differentials of the most important matrix functions. These results are used in deriving the estimating equations of the approximate EM algorithm and in establishing their asymptotic properties. Appendix B gives main results on Kronecker products. Appendix C deals with the expected values of matrix functions in the presence of missing data, or, in other words, when some rows and corresponding columns of matrices are missing. Appendix D deals with the general results on convergence of the estimates implied by a set of estimating equations, and is used to demonstrate consistency and asymptotic normality of the estimates.

5.1 Incidence matrices

Suppose the complete data $Y_1^c, \ldots, Y_N^c$ are i.i.d. $N(\mu, \Sigma)$ where both μ and Σ may depend on some parameters and/or covariates. Two interpretations we might want to keep in mind are that (i) for each $i = 1, \ldots, N$, the Y_i^c is the set of measurements coming from environmental monitors, with Σ describing their spatial correlation, or that (ii) the Y_i^c's are the measurement of indicators in a social science data set for a specific individual linked through a latent variable or a factor model. The observed data are $Y_1^o, \ldots, Y_N^o$ where for each i, Y_i^o is a subvector of Y_i^c. Let us assume for simplicity that the data are MCAR (see Section 2.2):

$$\mathbb{P}\mathrm{r}[Y_{ik} \text{ is missing} | \mathbf{Y}, \mathbf{X}, \theta] = \nu \tag{5.1}$$

independently from other missing data. The MCAR assumption may or may not be justifiable in different settings. In the environmental statistics setting, this assumption will be violated if the monitors fail to record certain values of the pollutant concentrations (excessively high or excessively low, for instance), or if monitors in a certain area tend to report or not to report the data together. In the social science examples, the MCAR assumption is likely to be violated for sensitive questions. If a respondent is ashamed of their too low income, or is protective about their high income, then such a person may choose not to report their income.

Define the incidence matrices P_i and M_i (present in instance i and missing in that instance) as following. If there are $d_{i,o}$ observed sites and $d_{i,m}$ missing ones, so that $d = d_{i,m} + d_{i,o} = \dim Y$, the matrix P_i is of dimensions $d_{i,o} \times d$, and consists of rows $e_k = (0, \ldots, 0, 1, 0, \ldots, 0)$ of length d with 1 in k-th position, where k runs through the indices of available cases. That way,

$$Y_i^o = P_i Y_i^c \sim N(P_i\mu, P_i \Sigma P_i^T) \tag{5.2}$$

where the entries of Y_i^o are arranged in the same order as those of Y_i^c, with the missing values taken out.

Likewise, the matrix M_i is a $d_{i,m} \times d$ with unit rows corresponding to the missing cases, and

$$Y_i^m = M_i Y_i^c \sim N(M_i\mu, M_i \Sigma M_i^T) \tag{5.3}$$

The complete data vector can be reconstructed as

$$Y_i^c = P_i^T Y_i^o + M_i^T Y_i^m \tag{5.4}$$

It may be noted that $P_i P_i^T = I_{d_{i,o}}$ and $M_i M_i^T = I_{d_{i,m}}$. Matrices $D_i = P_i^T P_i$ and $E_i = M_i^T M_i$ are idempotent matrices of size d and ranks $d_{i,o}$ and $d_{i,m}$, respectively, with ones on the diagonal and zeroes off-diagonal. If no data are missing, $P_i = I_d$, and M_i is not defined.

The likelihood of the complete data for i-th observation is then

$$l_i^c = -\frac{d}{2}\ln 2\pi \; - \; \frac{1}{2}\ln|\Sigma(\theta)| - \frac{1}{2}\operatorname{tr}\Big[\Sigma^{-1}(Y_i^c - \mu)(Y_i^c - \mu)^T\Big] \tag{5.5}$$

and the likelihood of the observed portion is

$$l_i = -\frac{d_{i,o}}{2}\ln 2\pi \; - \; \frac{1}{2}\ln|P_i\Sigma(\theta)P_i^T| - \frac{1}{2}\operatorname{tr}\Big[(P_i\Sigma P_i^T)^{-1}(Y_i^o - P_i\mu)(Y_i^o - P_i\mu)^T\Big] \tag{5.6}$$

The likelihood of all observations, under the assumption of independence over i, is

$$\begin{aligned} l(\theta, Y^o) = &-\frac{1}{2}\ln 2\pi \sum_{i=1}^{N} d_{i,o} \; - \; \frac{1}{2}\sum_{i=1}^{N}\ln|P_i\Sigma(\theta)P_i^T| - \\ &- \frac{1}{2}\sum_{i=1}^{N}\operatorname{tr}\Big[(P_i\Sigma P_i^T)^{-1}(Y_i^o - P_i\mu)(Y_i^o - P_i\mu)^T\Big] \end{aligned} \tag{5.7}$$

Suppose we could rearrange the entries of Y_i^c so that the missing entries come first: $Y_i^c = (Y_i^m, Y_i^o)$, with dimensions $d = d_{i,m} + d_{i,o}$, respectively. If that were the missing data pattern, the matrices P_i and M_i would be blocks of the identity matrix:

$$I_d = \begin{pmatrix} M_i \\ P_i \end{pmatrix} \tag{5.8}$$

With the vector μ split into μ_i^m, μ_i^o and the covariance matrix and its inverse blocked accordingly,

$$\Sigma = \begin{pmatrix} \Sigma_{i,mm} & \Sigma_{i,mo} \\ \Sigma_{i,om} & \Sigma_{i,oo} \end{pmatrix}, \quad \Sigma^{-1} = \begin{pmatrix} \Sigma_i^{mm} & \Sigma_i^{mo} \\ \Sigma_i^{om} & \Sigma_i^{oo} \end{pmatrix}, \tag{5.9}$$

the contribution to the likelihood of the full data vector in i-th observation is

$$l_i^c = -\frac{d}{2}\ln 2\pi \; - \; \frac{1}{2}\ln|\Sigma(\theta)| - \frac{1}{2}\operatorname{tr}\big[\Sigma^{-1}(Y_i^c - \mu)(Y_i^c - \mu)^T\big] \tag{5.10}$$

and the likelihood of the observed part is

$$l_i = -\frac{d_o}{2}\ln 2\pi \; - \; \frac{1}{2}\ln|\Sigma_{i,oo}(\theta)| - \frac{1}{2}\operatorname{tr}\left[\Sigma_{i,oo}^{-1}(Y_i^o - \mu_i^o)(Y_i^o - \mu_i^o)^T\right] \tag{5.11}$$

For the exact EM algorithm (equivalent to the maximum likelihood, by the general EM theory), one would need to compute the expected values of the sufficient statistics of the missing data. In the normal case, sufficient statistics are the first two moments of the data. Computing the required conditional expectations of the missing data given the observed data and parameters is exactly the kriging problem: treating the model parameters as fixed, find means, variances and covariances at unobserved locations.

For the approximate EM algorithm that uses marginal expected values in place of conditional ones, if Y_{ik} is missing, then the conditional approximate expected values are

$$\tilde{\mathbb{E}}Y_{ik}^2 = \mu_k^2 + \sigma_{kk}, \quad \tilde{\mathbb{E}}Y_{ik}Y_{il} = \mu_k\mu_l + \sigma_{kl} \tag{5.12}$$

Then the approximate conditional expectation in the approximate likelihood is

$$\tilde{\mathbb{E}}\left[(Y_i^c - \mu)(Y_i^c - \mu)^T\right] = \begin{pmatrix} \Sigma_{i,mm} & \Sigma_{i,mo} \\ \Sigma_{i,om} & (Y_i^o - \mu_i^o)(Y_i^o - \mu_i^o)^T \end{pmatrix} \tag{5.13}$$

and further

$$\begin{aligned}
\Sigma^{-1}\tilde{\mathbb{E}}\left[(Y_i^c - \mu)(Y_i^c - \mu)^T\right] &= \begin{pmatrix} \Sigma_i^{mm} & \Sigma_i^{mo} \\ \Sigma_i^{om} & \Sigma_i^{oo} \end{pmatrix}\begin{pmatrix} \Sigma_{i,mm} & \Sigma_{i,mo} \\ \Sigma_{i,om} & (Y_i^o - \mu_i^o)(Y_i^o - \mu_i^o)^T \end{pmatrix} = \\
&= \begin{pmatrix} \Sigma_i^{mm} & \Sigma_i^{mo} \\ \Sigma_i^{om} & \Sigma_i^{oo} \end{pmatrix}\begin{pmatrix} \Sigma_{i,mm} & \Sigma_{i,mo} \\ \Sigma_{i,om} & (Y_i^o - P_i\mu)(Y_i^o - P_i\mu)^T \end{pmatrix} = \\
&= \begin{pmatrix} I_{d_{i,m}} & U_i \\ U_i^T & \Sigma_i^{om}\Sigma_{i,mo} + \Sigma_i^{oo}(Y_i^o - P_i\mu)(Y_i^o - P_i\mu)^T \end{pmatrix} = \\
&= \begin{pmatrix} I_{d_{i,m}} & U_i \\ U_i^T & I_{d_{i,o}} - \Sigma_i^{oo}\Sigma_{i,oo} + \Sigma_i^{oo}(Y_i^o - P_i\mu)(Y_i^o - P_i\mu)^T \end{pmatrix} = \\
&= \begin{pmatrix} I_{d_{i,m}} & U_i \\ U_i^T & I_{d_{i,o}} + \Sigma_i^{oo}R_i \end{pmatrix}
\end{aligned} \tag{5.14}$$

since

$$\Sigma_i^{om}\Sigma_{i,mo} + \Sigma_i^{oo}\Sigma_{i,oo} = I_{d_{i,o}} \tag{5.15}$$

by the block inverse formulae (3.25);

$$R_i = (Y_i^o - P_i\mu)(Y_i^o - P_i\mu)^T - \Sigma_{i,oo} \tag{5.16}$$

is the matrix residual;

$$U_i = \Sigma_i^{mm}\Sigma_{i,mo} + \Sigma_{i,mo}(Y_i^o - P_i\mu)(Y_i^o - P_i\mu)^T$$

is a matrix that can be safely disregarded, as long as the likelihood depends on the trace of (5.14) through

$$\operatorname{tr}\Big\{\Sigma^{-1}\tilde{\mathbb{E}}\big[(Y_i^c - \mu)(Y_i^c - \mu)^T\big]\Big\} = d + \operatorname{tr}\Big\{\Sigma_i^{oo}\big[(Y_i^o - P_i\mu)(Y_i^o - P_i\mu)^T - \Sigma_{i,oo}\big]\Big\} \tag{5.17}$$

Thus, the approximate likelihood for the i-th observation from the approximate EM-algorithm is

$$\tilde{l}_i = -\frac{d}{2}(\ln 2\pi + 1) - \frac{1}{2}\ln|\Sigma(\theta)| - \frac{1}{2}\operatorname{tr}\Big\{\Sigma_i^{oo}\big[(Y_i^o - P_i\mu)(Y_i^o - P_i\mu)^T - \Sigma_{i,oo}\big]\Big\} \tag{5.18}$$

The full matrix $\Sigma(\theta)$ need to be inverted only once, and if spectral methods (or Cholesky decomposition) are used, then the determinant can also be obtained as the product of eigenvalues (or diagonal entries of the Cholesky decomposition matrices).

Other computational considerations should be kept in mind for the matrices P_i and M_i. Storing the matrices "as is" is a very memory inefficient solution, as they are very sparse, and are of very well defined structure. A more efficient storage and handling solution might be to store vectors of indices k from the definition of those matrices on page 60, and have user-defined subroutines for multiplication operations. The multiplications of the form PA or AP^T for an arbitrary matrix A are submatrix extraction operations, and multiplications P^TA or AP are insertion of rows and columns of zeroes.

In a general case of an arbitrary pattern of missing data with no rearrangement of observations (and hence rows/columns of Σ), the approximate likelihood is given by a generalization of (5.18):

$$\tilde{l}_i = -\frac{d}{2}\ln 2\pi e - \frac{1}{2}\ln|\Sigma(\theta)| - \frac{1}{2}\operatorname{tr}\Big\{P_i\Sigma^{-1}P_i^T R_i\Big\} \tag{5.19}$$

where now the matrix residual is

$$R_i = (Y_i^o - P_i\mu)(Y_i^o - P_i\mu)^T - P_i\Sigma P_i^T \tag{5.20}$$

and for all observations, assuming independence over i,

$$\tilde{l}(\theta, Y^o) = -\frac{dN}{2}\ln 2\pi e - \frac{N}{2}\ln|\Sigma(\theta)| - \\ -\frac{1}{2}\sum_{i=1}^{N}\mathrm{tr}\Big\{P_i\Sigma^{-1}P_i^T\big[(Y_i^o - P_i\mu)(Y_i^o - P_i\mu)^T - P_i\Sigma P_i^T\big]\Big\} \tag{5.21}$$

5.2 Estimating equations: maximum likelihood

Let us rewrite the observed data likelihood (5.7), allowing μ to change from location to location according to a regression $\mu_i = X_i\beta$ where X_i is $d \times p$ matrix of design variables corresponding to the site i, and β are regression coefficients. The likelihood of the observed data, under the assumption of independence over i, is

$$l(\theta, Y^o) = -\frac{1}{2}\ln 2\pi\sum_{i=1}^{N} d_{i,o} - \frac{1}{2}\sum_{i=1}^{N}\ln|P_i\Sigma(\theta)P_i^T| - \\ -\frac{1}{2}\sum_{i=1}^{N}\mathrm{tr}\Big[(P_i\Sigma P_i^T)^{-1}(Y_i^o - P_iX_i\beta)(Y_i^o - P_iX_i\beta)^T\Big] \tag{5.22}$$

5.2.1 The differential of the log likelihood

Let us derive the estimating equations for the coefficient estimates by using matrix calculus. The differential notation (d) and the main matrix calculus results that are necessary in those derivations are introduced in Appendix A.

Lemma 5.1.

$$\mathsf{d}\, l(\theta, Y^o) = -\sum_{i=1}^{N}\mathrm{tr}\Big\{(P_i\Sigma P_i^T)^{-1}P_i\Big[\{\mathsf{d}\,\Sigma\}P_i^T(P_i\Sigma P_i^T)^{-1}R_i - 2X_i\{\mathrm{d}\beta\}(Y_i^o - P_iX_i\beta)^T\Big]\Big\} \tag{5.23}$$

Proof. We shall take the differentials of the terms in the likelihood sequentially.

$$\mathsf{d}\sum_{i=1}^{N}\ln|P_i\Sigma P_i^T| = \sum_{i=1}^{N}\mathsf{d}\ln|P_i\Sigma P_i^T| = \sum_{i=1}^{N}\mathrm{tr}\Big\{\big[P_i\Sigma P_i^T\big]^{-1}\mathsf{d}\big[P_i\Sigma P_i^T\big]\Big\}, \tag{5.24}$$

$$\mathsf{d}\sum_{i=1}^{N}\mathrm{tr}\Big[(P_i\Sigma P_i^T)^{-1}(Y_i^o - P_iX_i\beta)(Y_i^o - P_iX_i\beta)^T\Big] =$$

$$= \sum_{i=1}^{N}\mathrm{tr}\,\mathsf{d}\Big[(P_i\Sigma P_i^T)^{-1}(Y_i^o - P_iX_i\beta)(Y_i^o - P_iX_i\beta)^T\Big] =$$

$$= \sum_{i=1}^{N}\mathrm{tr}\Big[\mathsf{d}\big\{(P_i\Sigma P_i^T)^{-1}\big\}(Y_i^o - P_iX_i\beta)(Y_i^o - P_iX_i\beta)^T +$$

$$+(P_i\Sigma P_i^T)^{-1}\big\{\mathsf{d}(Y_i^o - P_iX_i\beta)\big\}(Y_i^o - P_iX_i\beta)^T +$$

$$+(P_i\Sigma P_i^T)^{-1}(Y_i^o - P_iX_i\beta)\big\{\mathsf{d}(Y_i^o - P_iX_i\beta)^T\big\}\Big] =$$

$$= \sum_{i=1}^{N}\mathrm{tr}\Big[-(P_i\Sigma P_i^T)^{-1}P_i\{\mathsf{d}\,\Sigma\}P_i^T(P_i\Sigma P_i^T)^{-1}(Y_i^o - P_iX_i\beta)(Y_i^o - P_iX_i\beta)^T -$$

$$-2(P_i\Sigma P_i^T)^{-1}P_iX_i\{\mathsf{d}\,\beta\}(Y_i^o - P_iX_i\beta)^T\Big] \tag{5.25}$$

Hence, combining (5.24) and (5.25),

$$-2\,\mathsf{d}\,l(\theta, Y^o) = \sum_{i=1}^{N}\mathrm{tr}\big[(P_i\Sigma P_i^T)^{-1}P_i\{\mathsf{d}\,\Sigma\}P_i^T\big] +$$

$$+\sum_{i=1}^{N}\mathrm{tr}\Big[-(P_i\Sigma P_i^T)^{-1}P_i\{\mathsf{d}\,\Sigma\}P_i^T(P_i\Sigma P_i^T)^{-1}(Y_i^o - P_iX_i\beta)(Y_i^o - P_iX_i\beta)^T -$$

$$-2(P_i\Sigma P_i^T)^{-1}P_iX_i\{\mathsf{d}\,\beta\}(Y_i^o - P_iX_i\beta)^T\Big] = \sum_{i=1}^{N}\mathrm{tr}\Big\{(P_i\Sigma P_i^T)^{-1}P_i\Big[\{\mathsf{d}\,\Sigma\}P_i^T\times$$

$$\times\big[I_{d_{i,o}} - (P_i\Sigma P_i^T)^{-1}(Y_i^o - P_iX_i\beta)(Y_i^o - P_iX_i\beta)^T\big] - 2X_i\{\mathsf{d}\,\beta\}(Y_i^o - P_iX_i\beta)^T\Big]\Big\} =$$

$$= -\sum_{i=1}^{N}\mathrm{tr}\Big\{(P_i\Sigma P_i^T)^{-1}P_i\Big[\{\mathsf{d}\,\Sigma\}P_i^T(P_i\Sigma P_i^T)^{-1}R_i - 2X_i\{\mathrm{d}\beta\}(Y_i^o - P_iX_i\beta)^T\Big]\Big\} \tag{5.26}$$

where we had to redefine the matrix residual R_i once again:

$$R_i = (Y_i^o - P_iX_i\beta)(Y_i^o - P_iX_i\beta)^T - P_i\Sigma P_i^T \tag{5.27}$$

□

5.2.2 $\mathsf{d}\,\Sigma$ for geostatistical models

In general, $\mathsf{d}\,\Sigma$ is a matrix that has a quite involved structure. Let the covariance structure of the spatial process be described by a typical sill and nugget structure (see Sec. 2.1.3):

$$\mathbb{C}\mathrm{ov}[Z_k, Z_l] = \begin{cases} \alpha(1+\kappa), & k = l \\ \alpha\rho(\varphi, k, l), & k \neq l \end{cases} \tag{5.28}$$

where α is the overall variance parameter, so that

$$\Sigma(\alpha, \kappa, \varphi) = \alpha C(\kappa, \varphi), \tag{5.29}$$

κ is the nugget effect, and (possibly a vector) φ describes the spatial correlation, then

$$\mathsf{d}\,\Sigma(\theta) = d\alpha\, C(\kappa, \varphi) \;+\; \alpha\, d\kappa\, I_d \;+\; \sum_j \alpha C_j(\varphi)\, d\varphi_j \tag{5.30}$$

where $C_j(\varphi)$ is the matrix with zero on diagonal, and k, l-th off-diagonal entry equal to $\frac{\partial\rho(\varphi,k,l)}{\partial\varphi_j}$. The matrix differential d reduces to combination of scalar differentials d multiplying fixed matrices.

Let us give some examples.

For the exponential-power variogram,

$$\gamma[Z(\mathbf{s}_k), Z(\mathbf{s}_l)] = \alpha(\kappa + 1 - e^{-(t_{kl}/R)^p}) \tag{5.31}$$

where

$$t_{kl} = \|Z(\mathbf{s}_k) - Z(\mathbf{s}_l)\| > 0 \tag{5.32}$$

is the distance between sites k and l. The two components of vector φ are the range R and the shape p parameters. The spatial correlation is then

$$\rho(\varphi, k, l) = e^{-|t/R|^p}, \quad t = \|Z(\mathbf{s}_k) - Z(\mathbf{s}_l)\| > 0 \tag{5.33}$$

and the derivatives are

$$\frac{\partial\rho}{\partial R} = \frac{p t_{kl}^p}{R^{p+1}} e^{-(t_{kl}/R)^p}, \tag{5.34}$$

$$\frac{\partial\rho}{\partial p} = -\Big(\frac{t_{kl}}{R}\Big)^p e^{-(t_{kl}/R)s^p} \ln\frac{t_{kl}}{R} \tag{5.35}$$

so that the two matrices C_1, C_2 consists of off-diagonal entries given by (5.34) and (5.35), respectively:

$$c_{1,kl} = \frac{pt_{kl}^p}{R^{p+1}} e^{-(t_{kl}/R)^p}, \quad k \neq l, \tag{5.36}$$

$$c_{2,kl} = \left(\frac{t_{kl}}{R}\right)^p e^{-(t_{kl}/R)^p} \ln \frac{t_{kl}}{R}, \quad k \neq l, \tag{5.37}$$

$$c_{1,kk} = c_{2,kk} = 0 \tag{5.38}$$

For the spherical variogram,

$$\gamma[Z(\mathbf{s}_k), Z(\mathbf{s}_l)] = \alpha \left(\kappa + \left[\frac{3}{2} \frac{t_{kl}}{R} - \frac{1}{2} \left(\frac{t_{kl}}{R} \right)^3 \right] \right), \quad 0 < t = \|Z(\mathbf{s}_k) - Z(\mathbf{s}_l)\| < R \tag{5.39}$$

the spatial correlation is

$$\rho(\varphi, k, l) = \left[1 - \frac{3}{2} \frac{t_{kl}}{R} + \frac{1}{2} \left(\frac{t_{kl}}{R} \right)^3 \right] \mathbb{1}\{0 < t_{kl} < R\} \tag{5.40}$$

and the derivative is given by

$$\frac{\mathrm{d}\rho}{\mathrm{d}R} = \left(\frac{3t_{kl}}{2R^2} - \frac{3t_{kl}^3}{2R^4} \right) \mathbb{1}\{0 < t_{kl} < R\}, \tag{5.41}$$

For non-stationary process, like those defined by the power law

$$\gamma[Z(\mathbf{s}_k), Z(\mathbf{s}_l)] = \alpha(\kappa + t^\lambda) \tag{5.42}$$

the function ρ cannot be interpreted as spatial correlation. Non-stationary variograms have to be analyzed using generalized covariances, i.e., covariances of the linear combinations such as contrasts, rather than the original observations; or through the use of restricted maximum likelihood, or REML (Zimmerman 1989).

5.2.3 Estimating equations

Using (5.23), we can derive the estimating equations for the MLE:

$$\frac{\partial l(\theta, Y^o)}{\partial \beta_j} = -\sum_{i=1}^{N} \mathrm{tr}\big[(P_i \Sigma P_i^T)^{-1}(Y_i^o - P_i X_i \beta) X_{ij}^T P_i^T\big], \tag{5.43}$$

$$\frac{\partial l(\theta, Y^o)}{\partial \alpha} = \frac{1}{2}\sum_{i=1}^{N} \mathrm{tr}\big[(P_i \Sigma P_i^T)^{-1} P_i C(\kappa, \varphi) P_i^T (P_i \Sigma P_i^T)^{-1} R_i\big], \tag{5.44}$$

$$\frac{\partial l(\theta, Y^o)}{\partial \kappa} = \frac{1}{2}\sum_{i=1}^{N} \mathrm{tr}\big[(P_i \Sigma P_i^T)^{-1} P_i \alpha P_i^T (P_i \Sigma P_i^T)^{-1} R_i\big], \tag{5.45}$$

$$\frac{\partial l(\theta, Y^o)}{\partial \varphi_j} = \frac{1}{2}\sum_{i=1}^{N} \mathrm{tr}\big[(P_i \Sigma P_i^T)^{-1} P_i \alpha C_j(\varphi) P_i^T (P_i \Sigma P_i^T)^{-1} R_i\big] \tag{5.46}$$

Equation (5.43) gives GLS estimators for β_j:

$$\begin{aligned}
0 &= \sum_{i=1}^{N} \mathrm{tr}\big[(P_i \Sigma P_i^T)^{-1}(Y_i^o - P_i X_i \beta) X_{ij}^T P_i^T\big] = \\
&= \sum_{i=1}^{N} \mathrm{tr}\big[X_{ij}^T P_i^T (P_i \Sigma P_i^T)^{-1}(Y_i^o - P_i X_i \beta)\big] = \\
&= \sum_{i=1}^{N} X_{ij}^T P_i^T (P_i \Sigma P_i^T)^{-1}(Y_i^o - P_i X_i \beta), \\
\mathbf{0} &= \sum_{i=1}^{N} X_i^T P_i^T (P_i \Sigma P_i^T)^{-1}(Y_i^o - P_i X_i \beta), \\
\hat{\beta} &= \Big[\sum_{i=1}^{N} X_i^T P_i^T (P_i \Sigma P_i^T)^{-1} P_i X_i\Big]^{-1} \sum_{i=1}^{N} X_i^T P_i^T (P_i \Sigma P_i^T)^{-1} Y_i^o
\end{aligned} \tag{5.47}$$

Note that the cross-derivatives of the likelihood with respect to trend parameters β and spatial covariance parameters α, κ, φ have zero expected values. E.g.,

$$\begin{aligned}
\frac{\partial^2 l(\theta, Y^o)}{\partial \beta_j \partial \alpha} &= \frac{\partial}{\partial \alpha}\frac{\partial l(\theta, Y^o)}{\partial \beta_j} = -\frac{\partial}{\partial \alpha}\sum_{i=1}^{N} \mathrm{tr}\big[(P_i \Sigma P_i^T)^{-1}(Y_i^o - P_i X_i \beta) X_{ij}^T P_i^T\big] = \\
&= -\sum_{i=1}^{N} \mathrm{tr}\big[(P_i \Sigma P_i^T)^{-1} P_i C(\kappa, \varphi) P_i^T (P_i \Sigma P_i^T)^{-1} P_i X_{ij} (Y_i^o - P_i X_i \beta)^T\big],
\end{aligned}$$

$$\mathbb{E}_Y\Big[\frac{\partial^2 l(\theta, Y^o)}{\partial\beta_j\partial\alpha}\Big] =$$

$$= -\sum_{i=1}^{N} \mathrm{tr}\big\{(P_i\Sigma P_i^T)^{-1}P_iC(\kappa,\varphi)P_i^T(P_i\Sigma P_i^T)^{-1}P_iX_{ij}\,\mathbb{E}_Y[Y_i^o - P_iX_i\beta]^T\big\} = 0 \quad (5.48)$$

since $\mathbb{E}_Y(Y_i^o - P_iX_i\beta) = 0$, and all other terms are constant matrices. ($\mathbb{E}_Y$ is a notation for the expectation over distribution of Y; the total expectation is to be combined with the expectation over the structure of the missing data which will be denoted by $\mathbb{E}_s$.) All other cross-derivatives have similar structure with the regression residual multiplying a term that only depends on the structure of the missing data. Hence, the information matrix entries between β and any of the spatial covariance parameters are zero. The information matrix is block-diagonal, and the parameter estimates of β are independent of the estimates of (jointly) α, κ and φ. This is a quite general result of independence of the mean and variance parameters of the normal distribution that has also been derived in the geostatistical context by Cressie (1993) and Smith (2003).

The asymptotic variances can be found by taking the expectations of the outer product of the estimating equations (5.43)–(5.44), or by taking the expected value of the Hessian matrix, i.e., derivatives of the above equations. By the general MLE theory, the two approaches are equivalent. Smith (2003, pp. 250–251) derives the asymptotic variance-covariance matrix for the case of complete data.

5.3 Estimating equations: approximate EM algorithm

5.3.1 The differential of the approximate likelihood

As was shown in (5.21), the approximate conditional expectations taken at the E-step of the EM algorithm imply a particular pseudo-likelihood to be maximized. Let us derive the differential and the estimating equations for that pseudo-likelihood in the way similar to one taken in the previous section.

Lemma 5.2.

$$\mathsf{d}\,\tilde{l}(\theta, Y^o) = \frac{1}{2}\sum_{i=1}^{N}\Big\{ -\operatorname{tr}\big[\Sigma^{-1}\{\mathsf{d}\,\Sigma\}\big] + \operatorname{tr}\Big(P_i\Sigma^{-1}\{\mathsf{d}\,\Sigma\}\Sigma^{-1}P_i^T R_i + P_i\Sigma^{-1}P_i^T\times \\ \times\big[2P_iX_i\{\mathsf{d}\,\beta\}(Y_i^o - P_iX_i\beta)^T + P_i\{\mathsf{d}\,\Sigma\}P_i^T\big]\Big)\Big\} \tag{5.49}$$

Proof.

$$\mathsf{d}\ln|\Sigma(\theta)| = \operatorname{tr}\big[\Sigma^{-1}\,\mathsf{d}\,\Sigma\big], \tag{5.50}$$

$$\mathsf{d}\operatorname{tr}\Big\{\sum_{i=1}^{N} P_i\Sigma^{-1}P_i^T\big[(Y_i^o - P_iX_i\beta)(Y_i^o - P_iX_i\beta)^T - P_i\Sigma P_i^T\big]\Big\} = \\ = \sum_{i=1}^{N}\operatorname{tr}\Big(P_i\{\mathsf{d}[\Sigma^{-1}]\}P_i^T R_i + P_i\Sigma^{-1}P_i^T\Big[-P_iX_i\{\mathsf{d}\,\beta\}(Y_i^o - P_iX_i\beta)^T + \\ +(Y_i^o - P_iX_i\beta)(-P_iX_i\,\mathsf{d}\,\beta)^T - P_i\{\mathsf{d}\,\Sigma\}P_i^T\Big]\Big) = \\ = \sum_{i=1}^{N}\operatorname{tr}\Big(-P_i\Sigma^{-1}\{\mathsf{d}\,\Sigma\}\Sigma^{-1}P_i^T R_i - P_i\Sigma^{-1}P_i^T\times \\ \times\Big[2P_iX_i\{\mathsf{d}\,\beta\}(Y_i^o - P_iX_i\beta)^T + P_i\{\mathsf{d}\,\Sigma\}P_i^T\Big]\Big) \tag{5.51}$$

Combining (5.50) and (5.51), the result follows. □

5.3.2 Regression parameter estimates for the approximate EM

The differential (5.49) implies an unbiased estimating equations for β:

$$\frac{\partial\tilde{l}(\theta, Y^o)}{\partial\beta_j} = \sum_{i=1}^{N}\operatorname{tr}\Big(X_{ij}^T P_i^T P_i\Sigma^{-1}P_i^T(Y_i^o - P_iX_i\beta)\Big) = \\ = \sum_{i=1}^{N} X_{ij}^T P_i^T P_i\Sigma^{-1}P_i^T(Y_i^o - P_iX_i\beta), \tag{5.52}$$

$$\tilde{\beta} = \Big(\sum_i X_i^T D_i\Sigma^{-1}D_iX_i\Big)^{-1}\Big(\sum_i X_i^T D_i\Sigma^{-1}P_i^T Y_i^o\Big) \tag{5.53}$$

where the matrix under the trace in the first line is a scalar, and $D_i = P_i^T P_i$ as defined on page 61. The resulting estimate is a version of weighted least squares estimate with a weighting matrix $P_i\Sigma^{-1}P_i^T$. The estimate is less efficient than the GLS estimate (5.47)

with the appropriate weighting matrix $\left(P_i \Sigma P_i^T\right)^{-1}$, but it is unbiased and consistent.

The argument in the end of the previous section about independence of the parameter estimates directly translates here, as well: the cross derivatives of the regression slopes and covariance parameters have zero expectations. But, unlike the MLE case, the asymptotic covariance matrix has the information sandwich form (see Appendix D), so to establish the zero cross-covariances, we would need to have other components of the sandwich estimator. They are derived later in the chapter.

The asymptotic variance for $\tilde{\beta}$ from (5.53) can be derived by either taking its variance explicitly, or by using the general sandwich formula (D.26). Computing the variance directly from (5.53), one obtains

$$\begin{aligned} \mathbb{V}[\tilde{\beta}] = \mathbb{V}\Big[\Big(\sum_i X_i^T D_i \Sigma^{-1} D_i X_i\Big)^{-1} \Big(\sum_i X_i^T D_i \Sigma^{-1} P_i^T Y_i^o\Big)\Big] = \\ = \Big(\sum_i X_i^T D_i \Sigma^{-1} D_i X_i\Big)^{-1} \Big(\sum_i X_i^T D_i \Sigma^{-1} D_i \Sigma D_i \Sigma^{-1} D_i X_i^T\Big) \Big(\sum_i X_i^T D_i \Sigma^{-1} D_i X_i\Big)^{-1} \end{aligned} \tag{5.54}$$

where the terms related to $\mathbb{V}[\tilde{\theta}]$ are of smaller order, and thus are ignored.

5.3.3 Estimating equations for spatial covariance parameters

Invoking (5.30), one can now obtain the estimating equations for the spatial covariance parameter subspace.

$$\begin{aligned} \frac{\partial \tilde{l}(\theta, Y^o)}{\partial \alpha} = \frac{1}{2} \sum_{i=1}^N \Big[- \operatorname{tr}\big[\Sigma^{-1} C(\kappa, \varphi)\big] + \\ + \operatorname{tr}\big(P_i \Sigma^{-1} C(\kappa, \varphi) \Sigma^{-1} P_i^T R_i + P_i \Sigma^{-1} P_i^T \, P_i C(\kappa, \varphi) P_i^T\big)\Big], \end{aligned} \tag{5.55}$$

$$\frac{\partial \tilde{l}(\theta, Y^o)}{\partial \kappa} = \frac{1}{2} \alpha \sum_{i=1}^N \Big[- \operatorname{tr} \Sigma^{-1} + \operatorname{tr}\big(P_i \Sigma^{-1} \Sigma^{-1} P_i^T R_i + P_i \Sigma^{-1} P_i^T\big)\Big], \tag{5.56}$$

$$\begin{aligned} \frac{\partial \tilde{l}(\theta, Y^o)}{\partial \varphi_j} = \frac{1}{2} \alpha \sum_{i=1}^N \Big[- \operatorname{tr}\big[\Sigma^{-1} C_j(\varphi)\big] + \\ + \operatorname{tr}\big(P_i \Sigma^{-1} C_j(\varphi) \Sigma^{-1} P_i^T R_i + P_i \Sigma^{-1} P_i^T \, P_i C_j(\varphi) P_i^T\big)\Big], \end{aligned} \tag{5.57}$$

All those equations are biased: even though $\mathbb{E}_Y \, R_{=} 0$, the expected value of the RHS

is not generally equal to zero.

5.3.4 Bias in the estimating equations

Equation (5.49) involves two residuals: the covariance space matrix residual R_i and the regression space residual $(Y_i^o - P_i X_i \beta)$ with zero expectations, but unlike the maximum likelihood estimating equation, it also has extra terms

$$B(\Sigma, \mathsf{d}\,\Sigma) = \sum_{i=1}^{N} \big[\mathrm{tr}(P_i \Sigma^{-1} P_i^T P_i \{\mathsf{d}\,\Sigma\} P_i^T) - \mathrm{tr}(\Sigma^{-1}\{\mathsf{d}\,\Sigma\})\big] \tag{5.58}$$

that generally would lead to bias in estimating equations.

If $B(\cdot)$ could be integrated out as a penalty term, the approximate EM algorithm would give consistent estimates of the parameters of interest. While the second term, by its origin, is $\mathsf{d} \ln |\Sigma|$, there does not seem to be a general expression for the first term.

Certain heuristic argument can be built for the contribution of $B(\cdot)$ for large N and data missing completely at random, so that the rows of P_i, $i = 1, \ldots, N$ represent (an ordered version of) a random sample without replacement of rows of I_d. By virtue of results in Appendix C,

$$\mathbb{E}_s \,\mathrm{tr}(P_i \Sigma^{-1} P_i^T P_i \{\mathsf{d}\,\Sigma\} P_i^T) = (1-\nu)\,\mathrm{tr}\big\{\Sigma^{-1}\big[(1-\nu)\,\mathsf{d}\,\Sigma + \nu\,\mathrm{diag}\,\mathsf{d}\,\Sigma\big]\big\} \tag{5.59}$$

where the expectation $\mathbb{E}_s$ is taken over the patterns of missing data, or observed samples s_i from the set of all monitors. If the missing data process is MCAR, then for $N \to \infty$, by the law of large numbers,

$$\frac{1}{N}\sum_{i=1}^{N} \mathrm{tr}(P_i \Sigma^{-1} P_i^T P_i \{\mathsf{d}\,\Sigma\} P_i^T) \xrightarrow{\;p\;} (1-\nu)\,\mathrm{tr}\big\{\Sigma^{-1}\big[(1-\nu)\,\mathsf{d}\,\Sigma + \nu\,\mathrm{diag}\,\mathsf{d}\,\Sigma\big]\big\} \tag{5.60}$$

where the probability limit is again taken over repeated sampling of sites, assuming samples are independent for different i's (which is a part of MCAR assumption).

5.3.5 Correction for κ

Let us first derive the correction for the nugget effect. By (C.4),

$$\frac{1}{N}\frac{\partial\tilde{l}(\theta,Y^o)}{\partial\kappa} = \frac{1}{N}\frac{\alpha}{2}\sum_{i=1}^{N}\Big[-\operatorname{tr}\Sigma^{-1} + \operatorname{tr}\big(P_i\Sigma^{-1}\Sigma^{-1}P_i^TR_i + P_i\Sigma^{-1}P_i^T\big)\Big] \xrightarrow{p} \\ \xrightarrow{p} \frac{\alpha}{2}\big[-\operatorname{tr}\Sigma^{-1} + (1-\nu)\operatorname{tr}\Sigma^{-1}\big] \tag{5.61}$$

where the limit involving the matrix residuals R_i is zero by the LLN under appropriate regularity conditions, and the limits of other terms are taken with respect to the MCAR missing data mechanism. The first term in the last expression needs to be attenuated by $1-\nu$ to make the whole expression equal to zero. As long as this is the differential of $\ln|\Sigma(\theta)|$, then the correction to the estimating procedure that needs to be made is the following:

> In the approximate ECM, the function that needs to be maximized with respect to κ should be
>
> $$\tilde{l}_\kappa(\theta,Y^o) = -\frac{N(1-\nu)}{2}\ln|\Sigma(\theta)| - \frac{1}{2}\sum_{i=1}^{N}\operatorname{tr}\Big\{P_i\Sigma^{-1}P_i^TR_i\Big\} \tag{5.62}$$

5.3.6 Correction for the spatial correlation parameters

As shown in (5.30), the contributions of parameters φ responsible for spatial correlations, such as the shape and the range of variogram, to $\mathsf{d}\,\Sigma$ have zero elements on the diagonal, and the relevant result from Appendix C is (C.6):

$$\frac{1}{N}\frac{\partial\tilde{l}(\theta,Y^o)}{\partial\varphi_j} = \frac{1}{N}\frac{1}{2}\alpha\sum_{i=1}^{N}\Big[-\operatorname{tr}\big[\Sigma^{-1}C_j(\varphi)\big] + \\ + \operatorname{tr}\big(P_i\Sigma^{-1}C_j(\varphi)\Sigma^{-1}P_i^TR_i + P_i\Sigma^{-1}P_i^T\,P_iC_j(\varphi)P_i^T\big)\Big] \xrightarrow{p} \\ \xrightarrow{p} \frac{\alpha}{2}\Big\{-\operatorname{tr}\big[\Sigma^{-1}C_j(\varphi)\big] + (1-\nu)^2\operatorname{tr}\big[\Sigma^{-1}C_j(\varphi)\big]\Big\} \tag{5.63}$$

where again the limit involving the matrix residuals R_i is zero by the LLN, and the limits of other terms are taken with respect to the MCAR missing data mechanism. The first term in the last expression needs to be attenuated by $(1-\nu)^2$ to make the whole expression equal to zero. As long as this is the differential of $\ln|\Sigma(\theta)|$, then the correction to the estimating procedure that needs to be made is the following:

In the approximate ECM, the function that needs to be maximized with respect to φ should be

$$\tilde{l}_{\varphi}(\theta, Y^o) = -\frac{N(1-\nu)^2}{2} \ln|\Sigma(\theta)| - \frac{1}{2}\sum_{i=1}^{N} \operatorname{tr}\Big\{P_i \Sigma^{-1} P_i^T R_i\Big\} \tag{5.64}$$

5.3.7 Correction for α

Unlike the previous two sets of parameters, the estimating equation for the overall scale parameter α involves both diagonal and off-diagonal terms as contributions to $\mathsf{d}\,\Sigma$. Hence,

$$\begin{aligned}
\frac{1}{N}\frac{\partial \tilde{l}(\theta, Y^o)}{\partial \alpha} &= \frac{1}{N}\frac{1}{2}\sum_{i=1}^{N}\Big[-\operatorname{tr}\big[\Sigma^{-1}C(\kappa,\varphi)\big] + \\
&+ \operatorname{tr}\big(P_i\Sigma^{-1}C(\kappa,\varphi)\Sigma^{-1}P_i^T R_i + P_i\Sigma^{-1}P_i^T\, P_i C(\kappa,\varphi)P_i^T\big)\Big] \xrightarrow{p} \\
\xrightarrow{p} \frac{1}{2}\Big[-\operatorname{tr}\big[\Sigma^{-1}C(\kappa,\varphi)\big] &+ (1-\nu)\operatorname{tr}\big\{\Sigma^{-1}\big[(1-\nu)C(\kappa,\varphi) + \nu \operatorname{diag} C(\kappa,\varphi)\big]\big\}\Big] = \\
&= \frac{1}{2}\Big[-\alpha^{-1}d + (1-\nu)^2\alpha^{-1}d + (1-\nu)\nu(1+\kappa)\operatorname{tr}\big\{\alpha^{-1}C(\kappa,\varphi)^{-1}\big\}\Big]
\end{aligned} \tag{5.65}$$

using the definition (5.29). The first term, which is the derivative of $\ln|\Sigma(\theta)|$, needs to be attenuated by $(1-\nu)^2$. The last term does not seem to fit anything in the quasi-likelihood, but can be integrated back over α to give

$$P(\alpha, \kappa, \varphi) = (1-\nu)\nu(1+\kappa)\operatorname{tr}\big[C(\kappa,\varphi)^{-1}\big]\, \ln\alpha \tag{5.66}$$

Thus the correction for α is given by:

In the approximate ECM, the function that needs to be maximized with respect to α should be

$$\tilde{l}_{\varphi}(\theta, Y^o) = -\frac{N(1-\nu)^2}{2} \ln|\Sigma(\theta)| - \frac{1}{2}P(\alpha,\kappa,\varphi) - \frac{1}{2}\sum_{i=1}^{N} \operatorname{tr}\Big\{P_i \Sigma^{-1} P_i^T R_i\Big\} \tag{5.67}$$

5.3.8 Summary of corrections

Equation (5.60) implies different corrections for different parameters. For the estimating equations of the nugget effect κ, the contribution to $\mathsf{d}\,\Sigma$ is $\alpha I\, d\kappa$, with zeroes off

diagonal, and the term $\operatorname{tr}\Sigma^{-1}\,\mathsf{d}\,\Sigma$ in (5.58) needs to be attenuated by $1-\nu$. For the estimating equations of the spatial correlations, the contribution to $\mathsf{d}\,\Sigma$ has zero diagonal, and the RHS becomes $(1-\nu)^2\operatorname{tr}\left[\Sigma^{-1}\alpha\sum_j C_j(\varphi)d\varphi_j\right]$, and thus the suggested correction would be to multiply $\operatorname{tr}\Sigma^{-1}\,\mathsf{d}\,\Sigma$ by $(1-\nu)^2$. For the estimating equations for α, the correction involves both multiplying by $(1-\nu)^2$, and adding an extra penalty term. Note again that an important assumption that we had to make to derive those corrections was the one of the data missing completely at random (MCAR), which also implies independence over i. This may be too strong in reality.

Thus, the overall structure of the approximate EM algorithm will be the following:

1. Start with some initial values (available OLS estimates for regression parameters; some reasonable guesses for the spatial covariance, e.g., the OLS regression mean squared error for α, 0.1 for κ, median distance between sites for the range parameter, etc.)

2. The E-step: compute $\tilde{\mathbb{E}}[(Y-X\beta)(Y-X\beta)|Y^o, X, \beta, \alpha, \kappa, \varphi]$ (involves the observed data and the marginal predictions $\sigma_{ij}(\alpha, \kappa, \varphi)$ for the missing data)

3. The conditional maximization step 1: update the estimate of κ maximizing (5.62) w.r.t. κ

4. The conditional maximization step 2: update the estimate of φ maximizing (5.64) w.r.t. φ

5. The conditional maximization step 3: update the estimate of α maximizing (5.67) w.r.t. α

6. The conditional maximization step 4: perform WLS regression (5.53)

7. Repeat steps 2–6 until convergence, properly defined

The above procedure results in the following set of estimating equations, rescaled by $1/N$ for convenience.

$$
\begin{aligned}
\psi_{\beta,N}(\mathbf{Y},\mathbf{X},\beta,\theta) =& \frac{1}{N}\sum_{i=1}^{N}\psi_{\beta}(Y_i,X_i,\beta,\theta),\\
\psi_{\beta}(Y_i,X_i,\beta,\theta) =& X_i^T P_i^T P_i \Sigma^{-1} P_i^T (Y_i^o - P_i X_i \beta), \qquad (5.68)\\
\psi_{\kappa,N}(\mathbf{Y},\mathbf{X},\beta,\theta) =& \frac{1}{N}\sum_{i=1}^{N}\psi_{\kappa}(Y_i,X_i,\beta,\theta),\\
\psi_{\kappa}(Y_i,X_i,\beta,\theta) =& \frac{\alpha}{2}\Big[-(1-\nu)\operatorname{tr}\Sigma^{-1} + \operatorname{tr}\big(P_i\Sigma^{-1}\Sigma^{-1}P_i^T R_i + P_i\Sigma^{-1}P_i^T\big)\Big], \qquad (5.69)\\
\psi_{\varphi_j,N}(\mathbf{Y},\mathbf{X},\beta,\theta) =& \frac{1}{N}\sum_{i=1}^{N}\psi_{\varphi_j}(Y_i,X_i,\beta,\theta),\\
\psi_{\varphi_j}(Y_i,X_i,\beta,\theta) =& \frac{\alpha}{2}\Big[-(1-\nu)^2\operatorname{tr}\big[\Sigma^{-1}C_j(\varphi)\big] +\\
&+ \operatorname{tr}\big(P_i\Sigma^{-1}C_j(\varphi)\Sigma^{-1}P_i^T R_i + P_i\Sigma^{-1}P_i^T\, P_i C_j(\varphi) P_i^T\big)\Big], \qquad (5.70)\\
\psi_{\alpha,N}(\mathbf{Y},\mathbf{X},\beta,\theta) =& \frac{1}{N}\sum_{i=1}^{N}\psi_{\alpha}(Y_i,X_i,\beta,\theta),\\
\psi_{\alpha}(Y_i,X_i,\beta,\theta) =& \frac{1}{2}\alpha^{-1}\Big[-(1-\nu)^2 d - (1-\nu)\nu(1+\kappa)\operatorname{tr}\big[C(\kappa,\varphi)^{-1}\big] +\\
&+ \operatorname{tr}\big(P_i\Sigma^{-1}P_i^T R_i + P_i\Sigma^{-1}P_i^T\, P_i\Sigma P_i^T\big)\Big] \qquad (5.71)
\end{aligned}
$$

5.4 Derivatives of the estimating equations

Let us establish the basic statistical properties of the estimating equations (5.68)–(5.71). For both consistency and asymptotic normality, the first order derivatives of those equations will be needed.

Derivatives of ψ_β

The derivatives of the estimating equations for β are as follows.

$$
\begin{aligned}
\mathsf{d}\,\psi_{\beta}(Y_i,X_i,\beta,\theta) = \mathsf{d}\big[X_i^T P_i^T P_i \Sigma^{-1} P_i^T (Y_i^o - P_i X_i\beta)\big] =\\
= X_i^T P_i^T P_i \Sigma^{-1}\{\mathsf{d}\,\Sigma\}\Sigma^{-1} P_i^T (Y_i^o - P_i X_i\beta) - X_i^T P_i^T P_i \Sigma^{-1} P_i^T P_i X_i \{d\beta\} \qquad (5.72)
\end{aligned}
$$

Hence,

$$\frac{\partial}{\partial\beta}\psi_{\beta,N}(\mathbf{Y},\mathbf{X},\beta,\theta) = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial\beta}\psi_{\beta}(Y_i,X_i,\beta,\theta) =$$

$$= -\frac{1}{N}\sum_{i=1}^{N} X_i^T P_i^T P_i \Sigma^{-1} P_i^T P_i X_i, \tag{5.73}$$

$$\frac{\partial}{\partial\kappa}\psi_{\beta,N}(\mathbf{Y},\mathbf{X},\beta,\theta) = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial\kappa}\psi_{\beta}(Y_i,X_i,\beta,\theta) =$$

$$= \frac{1}{N}\sum_{i=1}^{N} X_i^T P_i^T P_i \Sigma^{-1}\Sigma^{-1} P_i^T (Y_i^o - P_i X_i\beta), \tag{5.74}$$

$$\frac{\partial}{\partial\varphi_j}\psi_{\beta,N}(\mathbf{Y},\mathbf{X},\beta,\theta) = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial\varphi_j}\psi_{\beta}(Y_i,X_i,\beta,\theta) =$$

$$= \frac{1}{N}\sum_{i=1}^{N} X_i^T P_i^T P_i \Sigma^{-1} C_j(\kappa,\varphi)\Sigma^{-1} P_i^T (Y_i^o - P_i X_i\beta) \tag{5.75}$$

$$\frac{\partial}{\partial\beta}\psi_{\beta,N}(\mathbf{Y},\mathbf{X},\beta,\theta) = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial\beta}\psi_{\beta}(Y_i,X_i,\beta,\theta) =$$

$$= \frac{1}{N}\sum_{i=1}^{N} \alpha^{-1} X_i^T P_i^T P_i \Sigma^{-1} P_i^T (Y_i^o - P_i X_i\beta) \tag{5.76}$$

The expectations of the equations (5.74)–(5.76) with respect to the distribution Y are all 0 due to zero expectation of the residual terms $Y_i^o - P_i X_i\beta$ (under normality, or, more generally, under symmetry of Y around its mean):

$$\mathbb{E}_Y\Big[\frac{\partial}{\partial\kappa}\psi_{\beta,N}(\mathbf{Y},\mathbf{X},\beta,\theta)\Big] = \mathbb{E}_Y\Big[\frac{\partial}{\partial\varphi_j}\psi_{\beta,N}(\mathbf{Y},\mathbf{X},\beta,\theta)\Big] =$$

$$= \mathbb{E}_Y\Big[\frac{\partial}{\partial\alpha}\psi_{\beta,N}(\mathbf{Y},\mathbf{X},\beta,\theta)\Big] = 0 \tag{5.77}$$

Derivatives of ψ_κ

For the derivatives of the estimating equations for the covariance parameters, we would need the following differential:

$$\mathsf{d}\, R_i = \mathsf{d}\big[(Y_i^o - P_i X_i\beta)(Y_i^o - P_i X_i\beta)^T - P_i\Sigma P_i^T\big] =$$

$$= -2(Y_i^o - P_i X_i\beta)d\beta^T X_i^T P_i^T - P_i\{\mathsf{d}\,\Sigma\}P_i^T \tag{5.78}$$

Then the differential of the estimating equation for the nugget effect can be obtained as follows:

$$\begin{aligned}
\mathsf{d}\,\psi_\kappa(Y_i, X_i, \beta, \theta) &= \mathsf{d}\,\frac{\alpha}{2}\Big[-(1-\nu)\operatorname{tr}\Sigma^{-1} + \operatorname{tr}\big(P_i\Sigma^{-1}\Sigma^{-1}P_i^T R_i + P_i\Sigma^{-1}P_i^T\big)\Big] = \\
&= \frac{d\alpha}{2}\Big[-(1-\nu)\operatorname{tr}\Sigma^{-1} + \operatorname{tr}\big(P_i\Sigma^{-1}\Sigma^{-1}P_i^T R_i + P_i\Sigma^{-1}P_i^T\big)\Big] + \\
&+\frac{\alpha}{2}\Big[(1-\nu)\operatorname{tr}\big[\Sigma^{-1}\{\mathsf{d}\,\Sigma\}\Sigma^{-1}\big] + \operatorname{tr}\big(-P_i[\Sigma^{-1}\{\mathsf{d}\,\Sigma\}\Sigma^{-2} + \Sigma^{-2}\{\mathsf{d}\,\Sigma\}\Sigma^{-1}]P_i^T R_i - \\
&-P_i\Sigma^{-2}P_i^T[2(Y_i^o - P_iX_i\beta)d\beta^T X_i^T P_i^T + P_i\{\mathsf{d}\,\Sigma\}P_i^T] - P_i\Sigma^{-1}\{\mathsf{d}\,\Sigma\}\Sigma^{-1}P_i^T\big)\Big] \qquad (5.79)
\end{aligned}$$

Then,

$$\begin{aligned}
\frac{\partial}{\partial\beta_j}\psi_{\kappa,N}(\mathbf{Y}, \mathbf{X}, \beta, \theta) &= \frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial\beta_j}\psi_\kappa(Y_i, X_i, \beta, \theta) = \\
&= -\frac{1}{N}\sum_{i=1}^{N}\alpha\operatorname{tr}\big[X_{ij}^T P_i^T P_i\Sigma^{-2}P_i^T(Y_i^o - P_iX_i\beta)\big] \qquad (5.80)
\end{aligned}$$

$$\mathbb{E}_Y\Big[\frac{\partial}{\partial\beta^T}\psi_{\kappa,N}(\mathbf{Y}, \mathbf{X}, \beta, \theta)\Big] = -\frac{1}{N}\sum_{i=1}^{N}\alpha\operatorname{tr}\big\{X_{ij}^T P_i^T P_i\Sigma^{-2}P_i^T\,\mathbb{E}_Y[Y_i^o - P_iX_i\beta]\big\} = 0, \qquad (5.81)$$

$$\begin{aligned}
\frac{\partial}{\partial\kappa}\psi_{\kappa,N}(\mathbf{Y}, \mathbf{X}, \beta, \theta) &= \frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial\kappa}\psi_\kappa(Y_i, X_i, \beta, \theta) = \\
&= \frac{1}{N}\frac{\alpha}{2}\sum_{i=1}^{N}\Big[(1-\nu)\operatorname{tr}\Sigma^{-2} + \operatorname{tr}\big(-2P_i\Sigma^{-3}P_i^T R_i - 2P_i\Sigma^{-2}P_i^T\big)\Big], \qquad (5.82)
\end{aligned}$$

$$\mathbb{E}_Y\Big[\frac{\partial}{\partial\kappa}\psi_{\kappa,N}(\mathbf{Y}, \mathbf{X}, \beta, \theta)\Big] = \frac{1}{N}\frac{\alpha}{2}\sum_{i=1}^{N}\Big[(1-\nu)\operatorname{tr}\Sigma^{-2} - 2\operatorname{tr}P_i\Sigma^{-2}P_i^T\Big],$$

$$\begin{aligned}
\mathbb{E}_s\,\mathbb{E}_Y\Big[\frac{\partial}{\partial\kappa}\psi_{\kappa,N}(\mathbf{Y}, \mathbf{X}, \beta, \theta)\Big] &= \\
= \frac{\alpha}{2}\big[(1-\nu)\operatorname{tr}\Sigma^{-2} - 2(1-\nu)\operatorname{tr}\Sigma^{-2}\big] &= -\frac{\alpha(1-\nu)}{2}\operatorname{tr}\Sigma^{-2} \qquad (5.83)
\end{aligned}$$

where in the last expression, we first have taken the expectation ($\mathbb{E}_Y$) with respect to the distribution of Y, and then with respect to the missing data mechanism ($\mathbb{E}_s$) which is assumed to be independent of the Y's. The results from Appendix C were used, where the last trace was represented as $\operatorname{tr}\big[P_i\Sigma^{-2}P_i^T P_i I_d P_i^T\big]$.

The cross-derivatives with other spatial covariance parameters are:

$$\begin{aligned}
\frac{\partial}{\partial\varphi_j}\psi_{\kappa,N}(\mathbf{Y},\mathbf{X},\beta,\theta) &= \frac{\alpha}{2N}\sum_{i=1}^{N}\Big[(1-\nu)\operatorname{tr}\big[\Sigma^{-1}C_j(\varphi)\Sigma^{-1}\big]+\\
&+\operatorname{tr}\big(-P_i[\Sigma^{-1}C_j(\varphi)\Sigma^{-2}+\Sigma^{-2}C_j(\varphi)\Sigma^{-1}]P_i^TR_i-\\
&-P_i\Sigma^{-2}P_i^TP_iC_j(\varphi)P_i^T-P_i\Sigma^{-1}C_j(\varphi)\Sigma^{-1}P_i^T\big)\Big],\\
\mathbb{E}_Y\Big[\frac{\partial}{\partial\varphi_j}\psi_{\kappa,N}(\mathbf{Y},\mathbf{X},\beta,\theta)\Big] &= \frac{\alpha}{2N}\sum_{i=1}^{N}\Big[(1-\nu)\operatorname{tr}\big[\Sigma^{-1}C_j(\varphi)\Sigma^{-1}\big]-\\
&-\operatorname{tr}\big(P_i\Sigma^{-2}P_i^TP_iC_j(\varphi)P_i^T+P_i\Sigma^{-1}C_j(\varphi)\Sigma^{-1}P_i^T\big)\Big],\\
\mathbb{E}_s\,\mathbb{E}_Y\Big[\frac{\partial}{\partial\varphi_j}\psi_{\kappa,N}(\mathbf{Y},\mathbf{X},\beta,\theta)\Big] &= -\frac{\alpha}{2}(1-\nu)^2\operatorname{tr}\big(\Sigma^{-2}C_j(\varphi)\big),
\end{aligned}\tag{5.84}$$

$$\begin{aligned}
\frac{\partial}{\partial\alpha}\psi_{\kappa,N}(\mathbf{Y},\mathbf{X},\beta,\theta) &= -\frac{1}{2N}\sum_{i=1}^{N}\operatorname{tr}\big[P_i\Sigma^{-2}P_i^T(R_i+\alpha P_iC(\kappa,\varphi)P_i^T)\big],\\
\mathbb{E}_Y\Big[\frac{\partial}{\partial\alpha}\psi_{\kappa,N}(\mathbf{Y},\mathbf{X},\beta,\theta)\Big] &= -\frac{1}{2N}\sum_{i=1}^{N}\alpha\operatorname{tr}(P_i\Sigma^{-2}P_i^TP_iC(\kappa,\varphi)P_i^T),\\
\mathbb{E}_s\,\mathbb{E}_Y\Big[\frac{\partial}{\partial\alpha}\psi_{\kappa,N}(\mathbf{Y},\mathbf{X},\beta,\theta)\Big] &= -\frac{1-\nu}{2}\operatorname{tr}\big[(1-\nu)\Sigma^{-1}+\alpha\nu(1+\kappa)\Sigma^{-2}\big]
\end{aligned}\tag{5.85}$$

Derivatives of ψ_φ

Let us now work out the estimating equations (5.70) for the correlation parameters φ.

$$\begin{aligned}
\mathsf{d}\,\psi_{\varphi_j}(Y_i,X_i,\beta,\theta) &= \mathsf{d}\,\frac{\alpha}{2}\Big[-(1-\nu)^2\operatorname{tr}\big(\Sigma^{-1}C_j(\varphi)\big)+\\
&+\operatorname{tr}\big(P_i\Sigma^{-1}C_j(\varphi)\Sigma^{-1}P_i^TR_i+P_i\Sigma^{-1}P_i^T\,P_iC_j(\varphi)P_i^T\big)\Big]=\\
&=\frac{d\alpha}{2}\Big[-(1-\nu)^2\operatorname{tr}\big(\Sigma^{-1}C_j(\varphi)\big)+\operatorname{tr}\big(P_i\Sigma^{-1}C_j(\varphi)\Sigma^{-1}P_i^TR_i+P_i\Sigma^{-1}P_i^T\,P_iC_j(\varphi)P_i^T\big)\Big]+\\
&+\frac{\alpha}{2}\Big[(1-\nu)^2\operatorname{tr}\big(\Sigma^{-1}\{\mathsf{d}\,\Sigma\}\Sigma^{-1}C_j(\varphi)-\Sigma^{-1}\,\mathsf{d}\,C_j(\varphi)\big)+\\
&+\operatorname{tr}\big(-P_i\Sigma^{-1}\{\mathsf{d}\,\Sigma\}\Sigma^{-1}C_j(\varphi)\Sigma^{-1}P_i^TR_i+P_i\Sigma^{-1}\{\mathsf{d}\,C_j(\varphi)\}\Sigma^{-1}P_i^TR_i-\\
&-P_i\Sigma^{-1}C_j(\varphi)\Sigma^{-1}\{\mathsf{d}\,\Sigma\}\Sigma^{-1}P_i^TR_i+P_i\Sigma^{-1}C_j(\varphi)\Sigma^{-1}P_i^T\,\mathsf{d}\,R_i-\\
&-P_i\Sigma^{-1}\{\mathsf{d}\,\Sigma\}\Sigma^{-1}P_i^T\,P_iC_j(\varphi)P_i^T+P_i\Sigma^{-1}P_i^T\,P_i\{\mathsf{d}\,C_j(\varphi)\}P_i^T\big)\Big]=
\end{aligned}$$

$$= \frac{d\alpha}{2}\Big[-(1-\nu)^2\,\mathrm{tr}\big(\Sigma^{-1}C_j(\varphi)\big) + \mathrm{tr}\big(P_i\Sigma^{-1}C_j(\varphi)\Sigma^{-1}P_i^T R_i + P_i\Sigma^{-1}P_i^T\,P_iC_j(\varphi)P_i^T\big)\Big]+$$
$$+\frac{\alpha}{2}\Big[(1-\nu)^2\,\mathrm{tr}\big(\Sigma^{-1}\{\mathsf{d}\,\Sigma\}\Sigma^{-1}C_j(\varphi) - \Sigma^{-1}\,\mathsf{d}\,C_j(\varphi)\big)+$$
$$+\mathrm{tr}\big(-P_i\Sigma^{-1}\{\mathsf{d}\,\Sigma\}\Sigma^{-1}C_j(\varphi)\Sigma^{-1}P_i^T R_i + P_i\Sigma^{-1}\{\mathsf{d}\,C_j(\varphi)\}\Sigma^{-1}P_i^T R_i -$$
$$-P_i\Sigma^{-1}C_j(\varphi)\Sigma^{-1}\{\mathsf{d}\,\Sigma\}\Sigma^{-1}P_i^T R_i -$$
$$-P_i\Sigma^{-1}C_j(\varphi)\Sigma^{-1}P_i^T\big[2(Y_i^o - P_iX_i\beta)d\beta^T X_i^T P_i^T + P_i\{\mathsf{d}\,\Sigma\}P_i^T\big] -$$
$$-P_i\Sigma^{-1}\{\mathsf{d}\,\Sigma\}\Sigma^{-1}P_i^T\,P_iC_j(\varphi)P_i^T + P_i\Sigma^{-1}P_i^T\,P_i\{\mathsf{d}\,C_j(\varphi)\}P_i^T\big)\Big], \tag{5.86}$$

$$\frac{\partial}{\partial\beta_j}\psi_{\varphi_j,N}(\mathbf{Y},\mathbf{X},\beta,\theta) = -\frac{\alpha}{N}\sum_{i=1}^{N}\mathrm{tr}\big[X_{ij}^T P_i^T\Sigma^{-1}C_j(\varphi)\Sigma^{-1}P_i^T(Y_i^o - P_iX_i\beta)\big],$$
$$\mathbb{E}_Y\Big[\frac{\partial}{\partial\beta^T}\psi_{\varphi_j,N}(\mathbf{Y},\mathbf{X},\beta,\theta)\Big] = 0 \tag{5.87}$$

$$\frac{\partial}{\partial\kappa}\psi_{\varphi_j,N}(\mathbf{Y},\mathbf{X},\beta,\theta) = \frac{\alpha}{2N}\sum_{i=1}^{N}\Big[(1-\nu)^2\,\mathrm{tr}\big(\Sigma^{-2}C_j(\varphi)\big)+$$
$$+\mathrm{tr}\big(-P_i\Sigma^{-2}C_j(\varphi)\Sigma^{-1}P_i^T R_i - P_i\Sigma^{-1}C_j(\varphi)\Sigma^{-2}P_i^T R_i -$$
$$-P_i\Sigma^{-1}C_j(\varphi)\Sigma^{-1}P_i^T - P_i\Sigma^{-2}P_i^T\,P_iC_j(\varphi)P_i^T\big)\Big],$$
$$\mathbb{E}_Y\Big[\frac{\partial}{\partial\kappa}\psi_{\varphi_j,N}(\mathbf{Y},\mathbf{X},\beta,\theta)\Big] = \frac{\alpha}{2N}\sum_{i=1}^{N}\Big[(1-\nu)^2\,\mathrm{tr}\big(\Sigma^{-2}C_j(\varphi)\big) -$$
$$-\mathrm{tr}\big(P_i\Sigma^{-1}C_j(\varphi)\Sigma^{-1}P_i^T + P_i\Sigma^{-2}P_i^T\,P_iC_j(\varphi)P_i^T\big)\Big],$$
$$\mathbb{E}_s\,\mathbb{E}_Y\Big[\frac{\partial}{\partial\kappa}\psi_{\varphi_j,N}(\mathbf{Y},\mathbf{X},\beta,\theta)\Big] = -\frac{\alpha(1-\nu)}{2}\,\mathrm{tr}\big(\Sigma^{-1}C_j(\varphi)\Sigma^{-1}\big) \tag{5.88}$$

$$\frac{\partial}{\partial\varphi_k}\psi_{\varphi_j,N}(\mathbf{Y},\mathbf{X},\beta,\theta) = \frac{\alpha}{2N}\sum_{i=1}^{N}\Big[(1-\nu)^2\,\mathrm{tr}\big(\Sigma^{-1}C_k(\varphi)\Sigma^{-1}C_j(\varphi) - \Sigma^{-1}C_{jk}(\varphi)\big)+$$
$$+\mathrm{tr}\big(-P_i\Sigma^{-1}C_k(\varphi)\Sigma^{-1}C_j(\varphi)\Sigma^{-1}P_i^T R_i + P_i\Sigma^{-1}C_{jk}(\varphi)\Sigma^{-1}P_i^T R_i -$$
$$-P_i\Sigma^{-1}C_j(\varphi)\Sigma^{-1}C_k(\varphi)\Sigma^{-1}P_i^T R_i - P_i\Sigma^{-1}C_j(\varphi)\Sigma^{-1}P_i^T P_iC_k(\varphi)P_i^T -$$
$$-P_i\Sigma^{-1}C_k(\varphi)\Sigma^{-1}P_i^T\,P_iC_j(\varphi)P_i^T + P_i\Sigma^{-1}P_i^T\,P_iC_{jk}(\varphi)P_i^T\big)\Big],$$
$$\mathbb{E}_Y\Big[\frac{\partial}{\partial\varphi_k}\psi_{\varphi_j,N}(\mathbf{Y},\mathbf{X},\beta,\theta)\Big] = \frac{\alpha}{2N}\sum_{i=1}^{N}\Big[(1-\nu)^2\,\mathrm{tr}\big(\Sigma^{-1}C_k(\varphi)\Sigma^{-1}C_j(\varphi) -$$
$$-\Sigma^{-1}C_{jk}(\varphi)\big) + \mathrm{tr}\big(-P_i\Sigma^{-1}C_j(\varphi)\Sigma^{-1}P_i^T P_iC_k(\varphi)P_i^T -$$
$$-P_i\Sigma^{-1}C_k(\varphi)\Sigma^{-1}P_i^T\,P_iC_j(\varphi)P_i^T + P_i\Sigma^{-1}P_i^T\,P_iC_{jk}(\varphi)P_i^T\big)\Big], \tag{5.89}$$

$$\mathbb{E}_s\,\mathbb{E}_Y\Big[\frac{\partial}{\partial\varphi_k}\psi_{\varphi_j,N}(\mathbf{Y},\mathbf{X},\beta,\theta)\Big] = -\frac{\alpha(1-\nu)^2}{2}\,\mathrm{tr}\big(\Sigma^{-1}C_j(\varphi)\Sigma^{-1}C_k(\varphi)\big) \tag{5.90}$$

where $C_{jk}(\varphi)$ are components of $\mathsf{d}\, C_j(\cdot)$ related to $d\varphi_k$:

$$\mathsf{d}\, C_j(\varphi) = \sum_k C_{jk}(\varphi) d\varphi_k \tag{5.91}$$

Just like $C_j(\cdot)$ matrices, $C_{jk}(\cdot)$ matrices have zeroes on diagonal, so the result from (C.6) applies to them. The expectations of cross-derivatives are equal,

$$\mathbb{E}_s\, \mathbb{E}_Y\Big[\frac{\partial}{\partial\varphi_k}\psi_{\varphi_j,N}(\mathbf{Y},\mathbf{X},\beta,\theta)\Big] = \mathbb{E}_s\, \mathbb{E}_Y\Big[\frac{\partial}{\partial\varphi_j}\psi_{\varphi_k,N}(\mathbf{Y},\mathbf{X},\beta,\theta)\Big] \tag{5.92}$$

by the equality of the mixed partial derivatives. (Recall that the correlation parameters require the same style corrections, and thus the same function is being maximized over them.)

Finally,

$$\frac{\partial}{\partial\alpha}\psi_{\varphi_j,N}(\mathbf{Y},\mathbf{X},\beta,\theta) = -\frac{1}{2N}\sum_{i=1}^{N}\operatorname{tr}\big[P_i\Sigma^{-1}C_j(\varphi)\Sigma^{-1}P_i^T(R_i + P_i\Sigma P_i^T)\big], \tag{5.93}$$

$$\mathbb{E}_Y\Big[\frac{\partial}{\partial\alpha}\psi_{\varphi_j,N}(\mathbf{Y},\mathbf{X},\beta,\theta)\Big] = -\frac{1}{2N}\sum_{i=1}^{N}\operatorname{tr}\big[P_i\Sigma^{-1}C_j(\varphi)\Sigma^{-1}P_i^T P_i\Sigma P_i^T\big], \tag{5.94}$$

$$\begin{aligned}\mathbb{E}_s\, \mathbb{E}_Y\Big[\frac{\partial}{\partial\alpha}\psi_{\varphi_j,N}(\mathbf{Y},\mathbf{X},\beta,\theta)\Big] &= -\frac{1-\nu}{2}\operatorname{tr}\big[\Sigma^{-1}C_j(\varphi)\Sigma^{-1}((1-\nu)\Sigma + \nu(1+\kappa)I)\big] = \\ &= -\frac{1-\nu}{2}\operatorname{tr}\Big[\Sigma^{-1}C_j(\varphi)\big((1-\nu)I + \nu(1+\kappa)\Sigma^{-1}\big)\Big]\end{aligned} \tag{5.95}$$

Derivatives of ψ_α

The last set of derivatives are those of the estimating equation (5.71) for α.

$$\begin{aligned}\mathsf{d}\,\psi_\alpha(Y_i,X_i,\beta,\theta) = -\frac{1}{2}\alpha^{-2}\,d\alpha\Big[&-(1-\nu)^2 d - (1-\nu)\nu(1+\kappa)\operatorname{tr}\big(C(\kappa,\varphi)^{-1}\big)+ \\ &+\operatorname{tr}\big(P_i\Sigma^{-1}P_i^T R_i + P_i\Sigma^{-1}P_i^T\, P_i\Sigma P_i^T\big)\Big]+ \\ +\frac{1}{2}\Big[&-(1-\nu)\nu\operatorname{tr}\big(d\kappa\,\Sigma^{-1} - (1+\kappa)\alpha\Sigma^{-1}\{\mathsf{d}\, C(\kappa,\varphi)\}\Sigma^{-1}\big)+ \\ +\alpha^{-1}\operatorname{tr}\big(&-P_i\Sigma^{-1}\{\mathsf{d}\,\Sigma\}\Sigma^{-1}(R_i + P_i\Sigma P_i^T) - 2P_i\Sigma^{-1}P_i^T(Y_i^o - P_iX_i\beta)d\beta^T X_i^T P_i^T\big)\Big],\end{aligned} \tag{5.96}$$

$$\frac{\partial}{\partial\beta_j}\psi_{\alpha,N}(\mathbf{Y},\mathbf{X},\beta,\theta) = -\frac{1}{2\alpha N}\sum_{i=1}^{N}\operatorname{tr}\big[X_{ij}^T P_i^T P_i \Sigma^{-1} P_i^T (Y_i^o - P_i X_i \beta)\big], \tag{5.97}$$

$$\mathbb{E}_Y\Big[\frac{\partial}{\partial\beta^T}\psi_{\alpha,N}(\mathbf{Y},\mathbf{X},\beta,\theta)\Big] = 0 \tag{5.98}$$

$$\frac{\partial}{\partial\kappa}\psi_{\alpha,N}(\mathbf{Y},\mathbf{X},\beta,\theta) =$$

$$= \frac{1}{2N}\sum_{i=1}^{N}\Big[-(1-\nu)\nu\operatorname{tr}\big\{\Sigma^{-1}\big(I-(1+\kappa)\alpha\Sigma^{-1}\big)\big\} - \operatorname{tr}\big\{P_i\Sigma^{-2}P_i^T(R_i+P_i\Sigma P_i^T)\big\}\Big], \tag{5.99}$$

$$\mathbb{E}_Y\Big[\frac{\partial}{\partial\kappa}\psi_{\alpha,N}(\mathbf{Y},\mathbf{X},\beta,\theta)\Big] =$$

$$= \frac{1}{2N}\sum_{i=1}^{N}\Big[-(1-\nu)\nu\operatorname{tr}\big\{\Sigma^{-1}\big(I-(1+\kappa)\alpha\Sigma^{-1}\big)\big\} - \operatorname{tr}\big\{P_i\Sigma^{-2}P_i^T P_i\Sigma P_i^T\big\}\Big], \tag{5.100}$$

$$\mathbb{E}_s\,\mathbb{E}_Y\Big[\frac{\partial}{\partial\kappa}\psi_{\alpha,N}(\mathbf{Y},\mathbf{X},\beta,\theta)\Big] = -\frac{1}{2}(1-\nu)\operatorname{tr}\Sigma^{-1} \tag{5.101}$$

$$\frac{\partial}{\partial\varphi_j}\psi_{\alpha,N}(\mathbf{Y},\mathbf{X},\beta,\theta) = \frac{1}{2N}\sum_{i=1}^{N}\Big[(1-\nu)\nu(1+\kappa)\alpha\operatorname{tr}\big(\Sigma^{-1}C_j(\varphi)\Sigma^{-1}\big) -$$

$$-\operatorname{tr}\big(P_i\Sigma^{-1}C_j(\varphi)\Sigma^{-1}P_i^T R_i + P_i\Sigma^{-1}C_j(\varphi)\Sigma^{-1}P_i^T P_i\Sigma P_i^T\big)\Big], \tag{5.102}$$

$$\mathbb{E}_Y\Big[\frac{\partial}{\partial\varphi_j}\psi_{\alpha,N}(\mathbf{Y},\mathbf{X},\beta,\theta)\Big] =$$

$$= \frac{1}{2N}\sum_{i=1}^{N}\Big[(1-\nu)\nu(1+\kappa)\alpha\operatorname{tr}\big(\Sigma^{-1}C_j(\varphi)\Sigma^{-1}\big) -$$

$$-\operatorname{tr}\big(P_i\Sigma^{-1}C_j(\varphi)\Sigma^{-1}P_i^T\,P_i\Sigma P_i^T\big)\Big], \tag{5.103}$$

$$\mathbb{E}_s\,\mathbb{E}_Y\Big[\frac{\partial}{\partial\varphi_j}\psi_{\alpha,N}(\mathbf{Y},\mathbf{X},\beta,\theta)\Big] = -\frac{1}{2}(1-\nu)^2\operatorname{tr}\big[\Sigma^{-1}C_j(\varphi)\Sigma^{-1}\big] \tag{5.104}$$

$$\frac{\partial}{\partial\alpha}\psi_{\alpha,N}(\mathbf{Y},\mathbf{X},\beta,\theta) = \frac{1}{2N}\alpha^{-2}\sum_{i=1}^{N}\Big[-(1-\nu)^2 d - (1-\nu)\nu(1+\kappa)\operatorname{tr}\big(C(\kappa,\varphi)^{-1}\big)\Big], \tag{5.105}$$

$$\mathbb{E}_Y\,\mathbb{E}_s\Big[\frac{\partial}{\partial\alpha}\psi_{\alpha,N}(\mathbf{Y},\mathbf{X},\beta,\theta)\Big] = -\frac{1-\nu}{2\alpha^2}\big[(1-\nu)d + \nu(1+\kappa)\operatorname{tr}\big(C(\kappa,\varphi)^{-1}\big)\big] \tag{5.106}$$

The derivatives matrix A

The results in Section 5.4 jointly define matrix A of derivatives that will be useful in establishing consistency and asymptotic normality of the estimates, as discussed in Appendix D:

$$A = \begin{pmatrix} A_{\beta\beta} & 0 \\ 0 & A_{\theta\theta} \end{pmatrix}, \tag{5.107}$$

$$\begin{aligned} A_{\beta\beta} &= -\,\mathbb{E}_s \sum_{i=1}^{N} X_i^T P_i^T P_i \Sigma^{-1} P_i^T P_i X_i, \\ (A_{\beta\beta})_{jk} &= -\,\mathbb{E}_s \sum_{i=1}^{N} X_{ij}^T P_i^T P_i \Sigma^{-1} P_i^T P_i X_{ik} = \\ &= -\sum_{i=1}^{N} \mathbb{E}_s \operatorname{tr}\big[P_i X_{ik} X_{ij}^T P_i^T P_i \Sigma^{-1} P_i^T\big] = \\ &= -(1-\nu)\operatorname{tr}\Big\{\mathbf{X}_k \mathbf{X}_j^T \big[\big((1-\nu)\Sigma^{-1} + \nu \operatorname{diag}\Sigma^{-1}\big) \otimes I_N\big]\Big\} = \\ &= -(1-\nu)\mathbf{X}_j^T \big[\big((1-\nu)\Sigma^{-1} + \nu \operatorname{diag}\Sigma^{-1}\big) \otimes I_N\big]\mathbf{X}_k, \\ A_{\beta\beta} &= -(1-\nu)\mathbf{X}^T \big[\big((1-\nu)\Sigma^{-1} + \nu \operatorname{diag}\Sigma^{-1}\big) \otimes I_N\big]\mathbf{X} \end{aligned} \tag{5.108}$$

Generally, expectations with respect to the missing data mechanism involving the regressors X_i are going to be awkward, as long as the regressors can vary between occasions i, unless the design is nicely balanced. Alternatively, as long as the entries of the design matrix $\mathbf{X}$ are observed in practice, one can use observed matrices X_i and compute

$$\hat{A}_{\beta\beta} = -\frac{1}{N} \sum_{i=1}^{N} X_i^T P_i^T P_i \Sigma^{-1} P_i^T P_i X_i,$$

Further,

$$A_{\theta\theta} = -\frac{1-\nu}{2} \begin{pmatrix} 3\alpha \operatorname{tr}\Sigma^{-2} & \alpha(1-\nu)\operatorname{tr}\big(\Sigma^{-2} C_j(\varphi)\big) & \dots \\ \alpha \operatorname{tr}\big(\Sigma^{-1} C_1(\varphi) \Sigma^{-1}\big) & \alpha(1-\nu)\operatorname{tr}\big(\Sigma^{-1} C_1(\varphi) \Sigma^{-1} C_1(\varphi)\big) & \dots \\ \vdots & \vdots & \ddots \\ \alpha \operatorname{tr}\big(\Sigma^{-1} C_q(\varphi) \Sigma^{-1}\big) & \alpha(1-\nu)\operatorname{tr}\big(\Sigma^{-1} C_1(\varphi) \Sigma^{-1} C_q(\varphi)\big) & \dots \\ \operatorname{tr}\Sigma^{-1} & (1-\nu)\operatorname{tr}\big[\Sigma^{-1} C_1(\kappa, \varphi)\big] & \dots \end{pmatrix}$$

$$\begin{array}{lll}
\dots & \operatorname{tr}((1-\nu)\Sigma^{-1}+\alpha\nu(1+\kappa)\Sigma^{-2}) & \operatorname{tr}\Sigma^{-1} \\
\dots & 2\operatorname{tr}\big\{\Sigma^{-1}C_1(\varphi)\big[(1-\nu)I+\nu(1+\kappa)\Sigma^{-1}\big]\big\} & (1-\nu)\operatorname{tr}\big[\Sigma^{-1}C_1(\varphi)\big] \\
\ddots & \vdots & \vdots \\
\dots & 2\operatorname{tr}\big\{\Sigma^{-1}C_q(\varphi)\big[(1-\nu)I+\nu(1+\kappa)\Sigma^{-1}\big]\big\} & (1-\nu)\big\{\operatorname{tr}\Sigma^{-1}C_q(\varphi)\big\} \\
\dots & \frac{1}{\alpha}\big[(1-\nu)d+\nu(1+\kappa)\operatorname{tr}\big(C(\kappa,\varphi)^{-1}\big)\big] & \alpha^{-2}\big[(1-\nu)d+\nu(1+\kappa)\operatorname{tr}\big(C(\kappa,\varphi)^{-1}\big)\big]
\end{array} \tag{5.109}$$

where the entries are ordered corresponding to entries of β, κ, $\varphi=(\varphi_1,\dots,\varphi_q)$ and α.

5.5 The variances of estimating equations

The variance-covariance matrix of the estimating equations,

$$B=\mathbb{E}\,\Psi(\mathbf{Y},\mathbf{X},\beta,\theta)\Psi(\mathbf{Y},\mathbf{X},\beta,\theta)^T \tag{5.110}$$

will be useful in establishing that the estimating equations are asymptotically normal, which will then be used in showing asymptotic normality of the estimates. Another simple implication of normality will be that $\Psi_n(\mathbf{Y},\mathbf{X},\beta,\theta)\sim O_p(N^{-1/2})$ which will be used in establishing consistency of the estimates.

The upper left block of this matrix is comprised of the expectations of the outer product $\psi_\beta(Y,X,\beta,\theta)\psi_\beta(Y,X,\beta,\theta)^T$:

$$\begin{aligned}
&\mathbb{E}_Y\Big[\psi_\beta(\mathbf{Y},\mathbf{X},\beta,\theta)\psi_{\beta,N}(\mathbf{Y},\mathbf{X},\beta,\theta)^T\Big]= \\
&=\mathbb{E}_Y\,\frac{1}{N^2}\sum_{i=1}^N\psi_\beta(Y_i,X_i,\beta,\theta)\psi_\beta(Y_i,X_i,\beta,\theta)^T= \\
&=\frac{1}{N^2}\sum_{i=1}^N\mathbb{E}_Y\big[X_i^TP_i^TP_i\Sigma^{-1}P_i^T(Y_i^o-P_iX_i\beta)\big]\big[X_i^TP_i^TP_i\Sigma^{-1}P_i^T(Y_i^o-P_iX_i\beta)\big]^T= \\
&=\frac{1}{N^2}\sum_{i=1}^N\mathbb{E}_Y\,X_i^TP_i^TP_i\Sigma^{-1}P_i^T\,\mathbb{E}_Y\big[(Y_i^o-P_iX_i\beta)(Y_i^o-P_iX_i\beta)^T\big]P_i\Sigma^{-1}P_i^TP_iX_i= \\
&=\frac{1}{N^2}\sum_{i=1}^N X_i^TP_i^TP_i\Sigma^{-1}P_i^TP_i\Sigma P_i^TP_i\Sigma^{-1}P_i^TP_iX_i
\end{aligned} \tag{5.111}$$

$B_{\beta\beta} = \mathbb{E}_s \mathbb{E}_Y \Big[\psi_{\beta,N}(\mathbf{Y}, \mathbf{X}, \beta, \theta)\psi_{\beta,N}(\mathbf{Y}, \mathbf{X}, \beta, \theta)^T\Big]$ is going to be a fourth order polynomial in ν, as in computing the expectation with respect to the missing data, four probabilities of $1-\nu$ would have to be multiplied and added up with different elements of the matrices Σ, Σ^{-1} and $\mathbf{X}\mathbf{X}^T$.

The blocks $B_{\beta\kappa}$, $B_{\beta\varphi}$ and $B_{\beta\alpha}$ are based on the expected values of the products of the estimating equation for β, (5.68) with (5.69), (5.70) and (5.71). They will all contain terms $Y_i - P_i X_i \beta$ and $(Y_i - P_i X_i \beta)R_i$ which all have expectation of zero under the assumption of normality (or, more broadly, symmetry of the error distribution, including cross-symmetries of terms involving different sites: $\mathbb{E}_Y (Y - X\beta)(Y - X\beta)^T \otimes (Y - X\beta) = 0$).

The expectations of the outer products of the estimating equations for the covariance parameters will contain products of traces, and identity (B.7) will be handy. We would also need the fourth moments of the data. Define

$$R_i^c = (Y_i^c - X_i\beta)(Y_i^c - X_i\beta)^T - \Sigma, \tag{5.112}$$

so that

$$R_i = P_i R_I^c P_i^T \tag{5.113}$$

Then

$$\mathbb{E}_Y \, r^c_{i\,jk} r^c_{i\,lm} = \big[(y^c_{ij} - x_{ij}\beta)(y^c_{ik} - x_{ik}\beta) - \sigma_{jk}\big]\big[(y^c_{il} - x_{il}\beta)(y^c_{im} - x_{im}\beta) - \sigma_{lm}\big] =$$

$$= \mathbb{E}_Y \big[(y^c_{ij} - x_{ij}\beta)(y^c_{ik} - x_{ik}\beta) - \sigma_{jk}\big](y^c_{il} - x_{il}\beta)(y^c_{im} - x_{im}\beta) =$$
$$= \mathbb{E}_Y \big[(y^c_{ij} - x_{ij}\beta)(y^c_{ik} - x_{ik}\beta)(y^c_{il} - x_{il}\beta)(y^c_{im} - x_{im}\beta)\big] - \sigma_{jk}\sigma_{lm} \equiv \mu_{jklm} - \sigma_{jk}\sigma_{lm},$$
$$\mathbb{E}_Y \, R_i^c \otimes R_i^c \equiv K \tag{5.114}$$

where the subindices j, k, l, m correspond to individual sites, and μ_{jklm} is the fourth order central moment. This is a generic entry of K, the matrix of the fourth order central moments.

Now, let us derive the products/variances of estimating equations for the covariance space. They all will be finite, with the only condition that $|\Sigma| \neq 0$. This condition, however, is guaranteed to hold by the parametric choice of Σ; see Section 2.1.

Variance of $\psi_\kappa(\cdot)$

Lemma 5.3. *$\mathbb{V}\,\psi_\kappa$ is finite.*

Proof.

$$\psi_\kappa(\mathbf{y}_i, \mathbf{X}_i, \beta, \theta)^2 =$$
$$= \frac{\alpha}{2}\Big[-(1-\nu)\operatorname{tr}\Sigma^{-1} + \operatorname{tr}\big(P_i\Sigma^{-2}P_i^T R_i + P_i\Sigma^{-1}P_i^T\big)\Big]\times$$
$$\times\frac{\alpha}{2}\Big[-(1-\nu)\operatorname{tr}\Sigma^{-1} + \operatorname{tr}\big(P_i\Sigma^{-2}P_i^T R_i + P_i\Sigma^{-1}P_i^T\big)\Big] =$$
$$= \frac{\alpha^2}{4}\Big[(1-\nu)^2(\operatorname{tr}\Sigma^{-1})^2 - 2(1-\nu)\operatorname{tr}\Sigma^{-1}\cdot\operatorname{tr}P_i\Sigma^{-1}P_i^T +$$
$$+(\operatorname{tr}P_i\Sigma^{-1}P_i^T)^2 + (\operatorname{tr}[P_i\Sigma^{-2}P_i^T R_i])^2 + \text{ terms, involving } R_i \text{ only once}\Big] \tag{5.115}$$

Using Lemma C.5, the expectations with respect to the missing data mechanism are:

$$\mathbb{E}_s \operatorname{tr}P_i\Sigma^{-1}P_i^T = (1-\nu)\operatorname{tr}\Sigma^{-1},$$

$$\mathbb{E}_s(\operatorname{tr}P_i\Sigma^{-1}P_i^T)^2 = \big[(1-\nu)\operatorname{tr}\Sigma^{-1}\big]^2 + (1-\nu)^2\Delta(\nu;\Sigma^{-1},I,\Sigma^{-1},I)$$
$$\mathbb{E}_s\big[\operatorname{tr}(P_iR_i^cP_i^TP_i\Sigma^{-2}P_i^T)\big]^2 =$$
$$= (1-\nu)^2\Big\{\big[\operatorname{tr}(R_i^c[(1-\nu)\Sigma^{-2}+\nu\operatorname{diag}\Sigma^{-2}])\big]^2 + \Delta(\nu;R_i^c,\Sigma^{-2},R_i^c,\Sigma^{-2})\Big\}, \tag{5.116}$$
$$\mathbb{E}_s\,\psi_\kappa(\mathbf{y}_i,\mathbf{X}_i,\beta,\theta)^2 = (1-\nu)^2\Big\{\big[\operatorname{tr}(R_i^c[(1-\nu)\Sigma^{-2}+\nu\operatorname{diag}\Sigma^{-2}])\big]^2 + \Delta(\nu;\Sigma^{-1},I,\Sigma^{-1},I)+$$
$$+\Delta(\nu;R_i^c,\Sigma^{-2},R_i^c,\Sigma^{-2})\Big\} + \text{ terms, involving } R_i \text{ only once} \tag{5.117}$$

Now, when expectation with respect to the distribution of Y is taken, the latter terms disappear, as well as the terms involving cross-products of R_i^c and R_j^c. Hence, involving (B.7) for the product of traces,

$$\mathbb{E}_Y\,\mathbb{E}_s\,\psi_\kappa(\mathbf{y}_i,\mathbf{X}_i,\beta,\theta)^2 =$$
$$= (1-\nu)^2\Big[\operatorname{tr}\big[K\cdot\big\{[(1-\nu)\Sigma^{-2}+\nu\operatorname{diag}\Sigma^{-2}]\otimes[(1-\nu)\Sigma^{-2}+\nu\operatorname{diag}\Sigma^{-2}]\big\}\big]+$$
$$+\Delta(\nu;\Sigma^{-1},I,\Sigma^{-1},I) + \mathbb{E}_Y\,\Delta(\nu;R,\Sigma^{-2},R,\Sigma^{-2})\Big] \tag{5.118}$$

The last term depends on the (entries of the) fourth moment matrix K defined in (5.114), and is spelled out in (C.14). □

Variance of $\psi_\varphi(\cdot)$

Lemma 5.4. $\mathbb{V}\,\psi_\varphi(\cdot)$ *is finite.*

Proof.

$$
\begin{aligned}
&\psi_{\varphi_j}(Y_i, \mathbf{X}_i, \beta, \theta)\, \psi_{\varphi_k}(Y_i, \mathbf{X}_i, \beta, \theta) = \\
=&\frac{\alpha^2}{4}\Big[-(1-\nu)^2 \operatorname{tr}\big(\Sigma^{-1}C_j(\varphi)\big) + \operatorname{tr}\big(P_i\Sigma^{-1}C_j(\varphi)\Sigma^{-1}P_i^T R_i + P_i\Sigma^{-1}P_i^T\, P_iC_j(\varphi)P_i^T\big)\Big]\times \\
&\times\Big[-(1-\nu)^2 \operatorname{tr}\big(\Sigma^{-1}C_k(\varphi)\big) + \operatorname{tr}\big(P_i\Sigma^{-1}C_k(\varphi)\Sigma^{-1}P_i^T R_i + P_i\Sigma^{-1}P_i^T\, P_iC_k(\varphi)P_i^T\big)\Big] = \\
=&\frac{\alpha^2}{4}\Big[(1-\nu)^4 \operatorname{tr}\big(\Sigma^{-1}C_j(\varphi)\big)\cdot\operatorname{tr}\big(\Sigma^{-1}C_k(\varphi)\big) \\
&-(1-\nu)^2 \operatorname{tr}\big[\Sigma^{-1}C_j(\varphi)\big]\cdot\operatorname{tr}\big[P_i\Sigma^{-1}P_i^T\, P_iC_k(\varphi)P_i^T\big] \\
&-(1-\nu)^2 \operatorname{tr}\big[P_i\Sigma^{-1}P_i^T\, P_iC_j(\varphi)P_i^T\big]\cdot\operatorname{tr}\big[\Sigma^{-1}C_k(\varphi)\big] \\
&+\operatorname{tr}\big[P_i\Sigma^{-1}P_i^T\, P_iC_j(\varphi)P_i^T\big]\cdot\operatorname{tr}\big[P_i\Sigma^{-1}P_i^T\, P_iC_k(\varphi)P_i^T\big] \\
&+\operatorname{tr}\big[P_i\Sigma^{-1}C_j(\varphi)\Sigma^{-1}P_i^T R_i\big]\cdot\operatorname{tr}\big[P_i\Sigma^{-1}C_k(\varphi)\Sigma^{-1}P_i^T R_i\big] \\
&+\ \text{terms, involving } R_i \text{ only once}\Big]
\end{aligned}
\tag{5.119}
$$

The expectation of the first four terms with respect to the missing data mechanism gives together $(1-\nu)^2\Delta(\nu;\Sigma^{-1}, C_j(\varphi), \Sigma^{-1}, C_k(\varphi))$. After denoting

$$
F_j = (1-\nu)\big[(1-\nu)\Sigma^{-1}C_j(\varphi)\Sigma^{-1} + \nu\operatorname{diag}(\Sigma^{-1}C_j(\varphi)\Sigma^{-1})\big], \tag{5.120}
$$

the expectation of the fifth term is:

$$
\begin{aligned}
&\mathbb{E}_s\big[\operatorname{tr}(P_i\Sigma^{-1}C_j(\varphi)\Sigma^{-1}P_i^T R_i)\cdot\operatorname{tr}(P_i\Sigma^{-1}C_k(\varphi)\Sigma^{-1}P_i^T R_i)\big] = \\
&= (1-\nu)^2\operatorname{tr}\big[(F_j\otimes F_k)(R_i\otimes R_i)\big] + (1-\nu)^2\Delta(\nu;\Sigma^{-1}C_j(\varphi)\Sigma^{-1}, R_i^c, \Sigma^{-1}C_k(\varphi)\Sigma^{-1}, R_i^c)
\end{aligned}
\tag{5.121}
$$

The expectation of the last term of (5.119) w.r.t. distribution of Y is zero. Thus,

$$
\begin{aligned}
&\mathbb{E}_Y\,\mathbb{E}_s\big[\psi_{\varphi_j}(Y_i, \mathbf{X}_i, \beta, \theta)\, \psi_{\varphi_k}(Y_i, \mathbf{X}_i, \beta, \theta)^T\big] = \\
&= (1-\nu)^2\Delta(\nu;\Sigma^{-1}, C_j(\varphi), \Sigma^{-1}, C_k(\varphi)) + (1-\nu)^2\operatorname{tr}\big[(F_j\otimes F_k)K\big] + \\
&\quad +(1-\nu)^2\,\mathbb{E}_Y\,\Delta(\nu;\Sigma^{-1}C_j(\varphi)\Sigma^{-1}, R_i^c, \Sigma^{-1}C_k(\varphi)\Sigma^{-1}, R_i^c)
\end{aligned}
\tag{5.122}
$$

The generic form of the last term is given in (C.14), and the whole expression is going to be finite provided Σ is invertible. □

Variance of $\psi_\alpha(\cdot)$

The final piece of the variance matrix we are interested in is the variance of $\psi_\alpha(\cdot)$.

Lemma 5.5. $\mathbb{V}\,\psi_\alpha(\cdot)$ *is finite.*

Proof.

$$\begin{aligned}\psi_\alpha(Y_i, X_i, \beta, \theta)^2 = \frac{1}{4}\alpha^{-2}\Big[&-(1-\nu)^2 d - (1-\nu)\nu(1+\kappa)\operatorname{tr}\big(C(\kappa,\varphi)^{-1}\big)+\\ &+\operatorname{tr}\big(P_i\Sigma^{-1}P_i^T R_i + P_i\Sigma^{-1}P_i^T\, P_i\Sigma P_i^T\big)\Big]^2 =\\ &=\frac{1}{4\alpha^2}\Big\{(1-\nu)^4 d^2 + (1-\nu)^2\nu^2(1+\kappa)^2\big[\operatorname{tr}\big(C(\kappa,\varphi)^{-1}\big)\big]^2+\\ &+2(1-\nu)^3 d\nu(1+\kappa)\operatorname{tr}\big(C(\kappa,\varphi)^{-1}\big) + \big[\operatorname{tr}(P_i\Sigma^{-1}P_i^T P_i R_i^c P_i^T)\big]^2-\\ &-2\big[(1-\nu)^2 d + (1-\nu)\nu(1+\kappa)\operatorname{tr}\big(C(\kappa,\varphi)^{-1}\big)\big]\operatorname{tr}(P_i\Sigma^{-1}P_i^T\, P_i\Sigma P_i^T)+\\ &+\big[\operatorname{tr}(P_i\Sigma^{-1}P_i^T\, P_i\Sigma P_i^T)\big]^2 + \text{ terms, involving } R_i \text{ only once}\Big\} \quad (5.123)\end{aligned}$$

The expectations with respect to the missing data mechanism of the terms involving P_i are:

$$\begin{aligned}&\mathbb{E}_s\big[\operatorname{tr}(P_i\Sigma^{-1}P_i^T P_i R_i^c P_i^T)\big]^2 = \mathbb{E}_s \operatorname{tr}\big[(P_i\Sigma^{-1}P_i^T\otimes P_i\Sigma^{-1}P_i^T)(P_i R_i^c P_i^T\otimes P_i R_i^c P_i^T)\big] =\\ &=(1-\nu)^2\Big[\operatorname{tr}\Big\{\big[(1-\nu)\Sigma^{-1}+\nu\operatorname{diag}\Sigma^{-1}\big]\otimes\big[(1-\nu)\Sigma^{-1}+\nu\operatorname{diag}\Sigma^{-1}\big]\Big\}(R_i^c\otimes R_i^c)+\\ &\qquad+\Delta(\nu;\Sigma^{-1},R_i^c,\Sigma^{-1},R_i^c)\Big], \quad (5.124)\\ &\mathbb{E}_s\big[\operatorname{tr}(P_i\Sigma^{-1}P_i^T\, P_i\Sigma P_i^T)\big]^2 = \mathbb{E}_s^2\big[\operatorname{tr}(P_i\Sigma^{-1}P_i^T\, P_i\Sigma P_i^T)\big] + (1-\nu)^2\Delta(\nu;\Sigma^{-1},\Sigma,\Sigma^{-1},\Sigma) =\\ &=(1-\nu)^2\big\{\operatorname{tr}\big[\Sigma^{-1}\big((1-\nu)\Sigma+\nu(1+\kappa)I_d\big)\big]\big\}^2 + (1-\nu)^2\Delta(\nu;\Sigma^{-1},\Sigma,\Sigma^{-1},\Sigma), \quad (5.125)\\ &\mathbb{E}_s\operatorname{tr}(P_i\Sigma^{-1}P_i^T\, P_i\Sigma P_i^T) = (1-\nu)\big\{\big[\operatorname{tr}\Sigma^{-1}\big((1-\nu)\Sigma+\nu(1+\kappa)I_d\big)\big]\big\} \quad (5.126)\end{aligned}$$

The expected value with respect to the distribution of Y of the last term of (5.123) is zero, and of (5.124), is

$$\begin{aligned}&\mathbb{E}_Y\,\mathbb{E}_s\big[\operatorname{tr}(P_i\Sigma^{-1}P_i^T R_i)\big]^2 = (1-\nu)^2\Big[\operatorname{tr}\Big\{\big[(1-\nu)\Sigma^{-1}+\nu\operatorname{diag}\Sigma^{-1}\big]\otimes\\ &\otimes\big[(1-\nu)\Sigma^{-1}+\nu\operatorname{diag}\Sigma^{-1}\big]\Big\}K + \mathbb{E}_Y\,\Delta(\nu;\Sigma^{-1},R_i^c,\Sigma^{-1},R_i^c)\Big], \quad (5.127)\end{aligned}$$

where the last term is dealt with in (C.14). Thus the total result is

$$\begin{aligned}\mathbb{E}_Y\,\mathbb{E}_s\psi_\alpha(Y_i,X_i,\beta,\theta)^2 &= \\ &= \frac{(1-\nu)^2}{4\alpha^2}\Big[\mathrm{tr}\Big\{\big[(1-\nu)\Sigma^{-1}+\nu\,\mathrm{diag}\,\Sigma^{-1}\big]\otimes\big[(1-\nu)\Sigma^{-1}+\nu\,\mathrm{diag}\,\Sigma^{-1}\big]\Big\}K+ \\ &+\mathbb{E}_Y\,\Delta(\nu;\Sigma^{-1},R_i^c,\Sigma^{-1},R_i^c)+\Delta(\nu;\Sigma^{-1},\Sigma,\Sigma^{-1},\Sigma)\Big] \qquad (5.128)\end{aligned}$$

□

Covariances of estimating equations

The cross-products $\psi_\kappa(\cdot)\psi_\varphi(\cdot)$, $\psi_\kappa(\cdot)\psi_\alpha(\cdot)$, $\psi_\varphi(\cdot)\psi_\alpha(\cdot)$ and their expectations can also be tediously derived. By Cauchy-Schwarz inequality, they will be bounded by the product of the standard errors of individual components. For the matter of the theoretical proofs, it suffices that they are finite.

5.5.1 An empirical estimate

In the empirical applications of the method, it will be more natural and easier to compute individual contributions $\psi_j(y_i,\mathbf{x}_i,\beta,\theta)$ and compute the empirical covariance matrix

$$\hat{B}=\frac{1}{N}\sum_{i=1}^{N}\Psi(y_i,\mathbf{x}_i,\beta,\theta)\Psi(y_i,\mathbf{x}_i,\beta,\theta)^T \qquad (5.129)$$

to be used in the sandwich variance estimation (see Section D.3 of Appendix D). This procedure will also be more beneficial in that ensures robustness against violations of assumptions of normality used throughout in derivation of the variances of estimating equations.

5.6 Consistency of $\tilde{\theta}$

In the earlier sections, we have obtained the derivatives of the estimating equations (5.68)–(5.71). Let us see if they satisfy the regularity conditions outlined in Appendix D for consistency of the estimates, in particular, conditions of Huber (1967).

The estimating equations have the form

$$\frac{1}{N}\sum_{i=1}^{N}\psi(X_i,\theta)$$

and by construction they satisfy (D.20) with exact equality (sans the computational accuracy).

Both the estimating equations and their derivatives are continuous functions of θ provided $|\Sigma| \neq 0$ which is guaranteed by the choice of negative definite variograms (see Section 2.1.3). Hence Huber's (1967) condition (B-1) is satisfied.

The existence of continuous derivatives also ensures conditions (B-2) and (B-2′) (aka (D.21) and (D.22)): for close enough θ and θ',

$$\frac{1}{2}\lambda_{\min}[-A]\|\theta'-\theta\| \leq \|\Psi(x,\theta')-\Psi(x,\theta)\| \leq 2\lambda_{\max}[-A]\|\theta'-\theta\| \tag{5.130}$$

where A is the matrix of derivatives given in (5.107), and $\lambda_{\min}[-A]$ and $\lambda_{\max}[-A]$ are the smallest and the largest eigenvalues of the corresponding matrix. Here, A depends on the parameter values, and can be taken say at $\frac{1}{2}(\theta+\theta')$. It should be verified that the matrix A is negative definite. It is the case for the regression block of it, as is obvious from (5.108) or (5.109), but the case is not clear for $A_{\theta\theta}$ part in (5.109). The condition (B-3) is satisfied by the choice of corrections in Section 5.3.8 that ensure that the population parameters solve $\lambda(\theta)=0$. The condition (B-4) is satisfied with

$$b(\theta) = M_1\frac{1}{N}\sum_{i=1}^{N}\|X_i\Sigma^{-1}X_i\beta\| + M_2\alpha\,\mathrm{tr}\,\Sigma^{-1} + M_3\sum_j|\,\mathrm{tr}(\Sigma^{-1}C_j(\varphi))| + M_4\alpha^{-1} \tag{5.131}$$

for appropriately chosen M_1, M_2, M_3, M_4. Note that one also needs to have α separated from zero.

Alternatively, conditions (D.34)–(D.35) of Appendix D.4 are easily established: the first one holds by CLT (provided variances of the estimating equations are finite, which has already been demonstrated), and the second one, by the LLN, as explained in that

appendix. Both of those conditions follow from the fact that the estimating equations and their derivatives are sample averages, as mentioned earlier. Also, (D.37) holds by CLT, and thus the estimates can be shown to be asymptotically normal.

5.7 Asymptotic normality of $\tilde{\theta}$

If the missing data mechanism is treated as a random component, then estimating equations (5.68)–(5.71) represent the sums of independent random variables. The independence in the distribution of Y is assumed throughout by the "dissociatedness" assumption, and independence of the missing data process, by MCAR. As was shown in Section 5.5, the variances of individual terms are finite, and thus the standard CLT is applicable (see e.g. Theorem 27.1 of Billingsley (1995) or Theorem 8.2 of Borovkov, Borovkov & Borovkova (1999)).

Alternatively, the missing data patterns can be treated as fixed, and the triangular array versions of the CLT should be used to show asymptotic normality of the estimating equations. The estimating equations for β (5.68) are marginally normal, as they only involve y_i as random variables. The estimating equations for the covariance space parameters, (5.69), (5.70), (5.71), involve the matrix residuals R_i and their traces. The individual entries of those matrices will have marginal χ^2, or gamma, distributions, scaled appropriately, with all moments being finite. The traces of the matrices involving R_i will be distributed as a mixture of gammas, and thus will also have finite moments. Thus Lyapunov's condition (Theorem 27.3 of Billingsley (1995); Section 8.4 of Borovkov et al. (1999)) will be satisfied, and the estimating equations will be asymptotically normal.

Upon having established normality of the estimating equations, one can proceed to establishing normality of the estimates themselves. As established above in Section 5.6, the estimating equations do satisfy the regularity conditions outlined in Appendix D.4. Corollary 2 of that appendix establishes asymptotic normality of the estimates, and applies here as well.

Alternatively, one can check the regularity conditions in Huber (1967). Condition (N-1) follows from continuity of $\Psi(\cdot)$. Condition (N-2) is satisfied by the choice of the corrections that guarantee that the population parameters θ_0 solve $\mathbb{E}\,\psi(\theta_0) = 0$. Condition (N-4), finiteness of the variances, was established in Section 5.5. The condition (N-3) will also follow from the fact that $\Psi(\cdot)$ have continuous derivatives, although this would require a longer explanation.

Let $\lambda_{\min}(\theta_0)$ and $\lambda_{\max}(\theta_0)$ be the smallest and largest eigenvalues of the derivative matrix $-A(\theta_0) = -D\Psi(\theta_0)$, and let $\mu_{\max}$ be the largest eigenvalue of the variance matrix $B = \mathbb{V}[\Psi(x, \theta_0)]$ (see section 5.5). As discussed on p. 5.6, there does not seem to be an easy general way of showing it is positive definite. Assuming it is, the following distances characterize the bounds near θ_0:

$$\exists d_1 : \forall\, d \le d_1 : \|\,\mathbb{E}\,\Psi(x, \theta_0)\| \ge \frac{1}{2}\lambda_{\min}(\theta_0)\|\theta - \theta_0\| \tag{5.132}$$

$$\exists d_2 : \forall\, \theta, \tau \in B(\theta_0, d_2) \text{ (in the neighborhood of } \theta_0) \tag{5.133}$$

$$\mathbb{E}\big[\|\Psi(x, \theta) - \Psi(x, \tau)\|\big] \le 2\lambda_{\max}(\theta_0)\|\theta - \tau\| \tag{5.134}$$

$$\exists d_3 : \forall\, \theta, \tau \in B(\theta_0, d_3) \text{ (in the neighborhood of } \theta_0) \tag{5.135}$$

$$\mathbb{E}\big[\|\Psi(x, \theta) - \Psi(x, \tau)\|^2\big] \le \mu_{\max}\|\theta - \tau\| \tag{5.136}$$

Then choosing $d_0 = \min(d_1, d_2, d_3)$ guarantees (N-3).

Thus the estimating equations for the approximate EM algorithm given in Section 5.3 give asymptotically normal estimates:

$$\sqrt{N}(\tilde{\theta} - \theta_0) \xrightarrow{d} N(0, A^{-1}BA^{T^{-1}}) \tag{5.137}$$

with A derived in Section 5.4, and B, in Section 5.5.

5.8 Numerical illustration

This section is based on a small Monte Carlo study comparing the performance of the exact maximum likelihood and the approximate EM with corrections. An underlying process was a Gaussian spatial field on a unit square $[0, 1]^2$ with a linear spatial trend and an exponential variogram of the spatially correlated regression error:

$$z_{it} = \beta_x x_i + \beta_y y_i + \beta_0 + \varepsilon_{it}, \tag{5.138}$$

$$\mathbb{V}[\epsilon_{it} - \epsilon_{js}] = \delta_{ts}\alpha\big[\delta_{ij}(1 + \kappa) + (1 - \delta_{ij})\exp(-d_{ij}/R)\big] \tag{5.139}$$

where d_{ij} is the Euclidean distance between sites i and j, and the parameters of the process are given in Table 5.1. 40 locations were randomly selected on that unit square; see Fig. 5.1. 50 independent draws from the field were taken to represent the time dimension of the dissociated process. 10% of the generated responses z were set to missing. 100 Monte Carlo samples were drawn. The population values were used as

Table 5.1: The simulation results.

	β_x	β_y	β_0	κ	α	R
Population	2	-1	0	0.2	1	0.5
Maximum likelihood						
Mean	1.9900	-0.9886	0.0084	0.2035	0.9874	0.4916
S.d.	0.1253	0.1290	0.1171	0.0784	0.0878	0.0709
Approximate EM with corrections						
Mean	1.9885	-0.9905	0.0081	0.1979	0.9957	0.4930
S.d.	0.1298	0.1353	0.1203	0.0879	0.0955	0.0751
Rel. MSE	0.932	0.911	0.949	0.795	0.862	0.897

starting values of the iterative maximization.

The variogram of the process is shown on Fig. 5.2 for one of the sample realizations. For this particular sample, the sampling fluctuations suggest the estimate of the nugget to be above its theoretical value, and the range is also moved up to allow the MLE curve to pass through the data cloud. (As the simulation results showed, the estimates of the nugget and range are highly correlated, in general.)

The summary statistics are given in Table 5.1. An alternative graphic representation of the distributions is given in Fig. 5.3, with the solid lines representing the kernel density estimates of the distribution of MLEs, and dashed lines, those of the approximate EM with corrections. The vertical lines represent the population values of the corresponding parameters.

The distributions of the parameter estimates are concentrated around the population values, with the latter in the 95% confidence intervals for the means of the simulated distributions. The distributions of the regression parameter estimates are close to normal; the distributions of the covariance estimates, especially the range and the scale, are somewhat skewed. The mean squared errors of the ML estimates and those from the approximate EM with corrections are compared in the last row of Table 5.1. The efficiency losses do not exceed 9% for the regression subvector, and are between 10 and 20% for the spatial covariance parameter estimates.

Table 5.2 addresses correlations between parameter estimates. For each of the estimation methods, the first row shows the highest (in absolute value) correlations among and between subsets of parameter estimates, and the second row, the lowest correlations. Higher correlations found in part (a) of the Table tend to lead to deteriorated convergence; more iterations will be required, especially for the linear convergent EM.

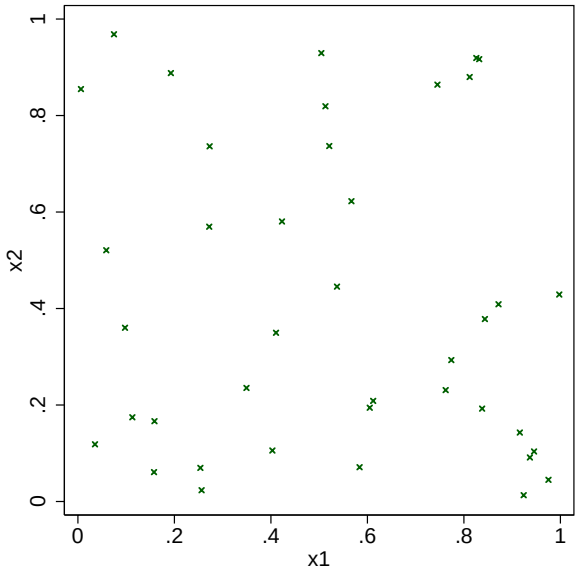


Figure 5.1: Locations of the simulated sites.

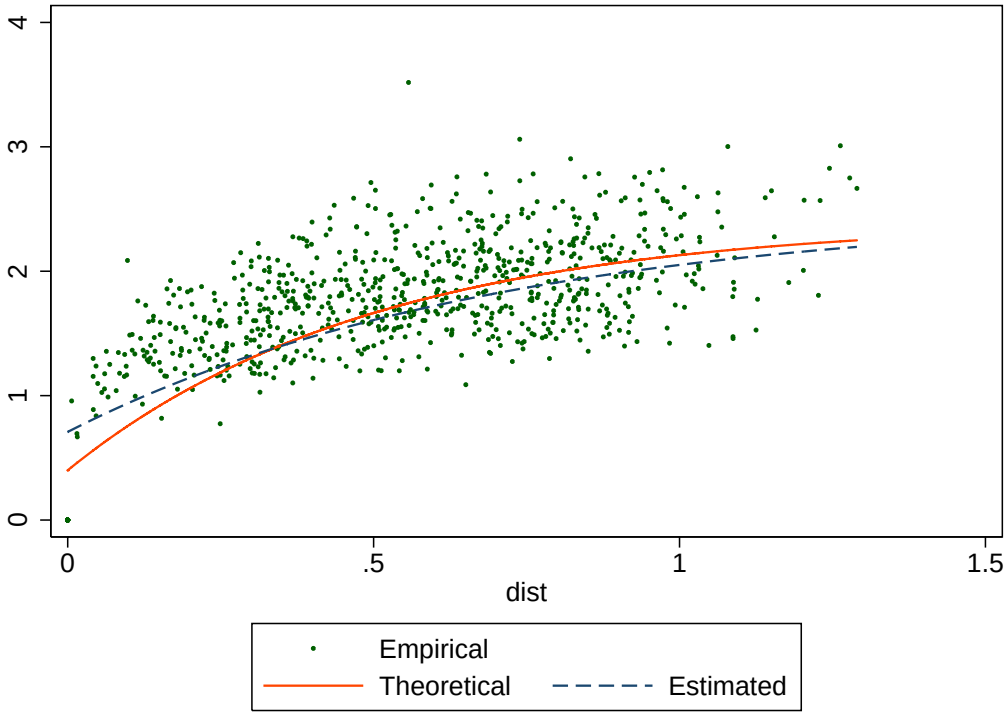


Figure 5.2: Variogram of the simulated process: pairwise variances, the theoretical variogram and the MLE.

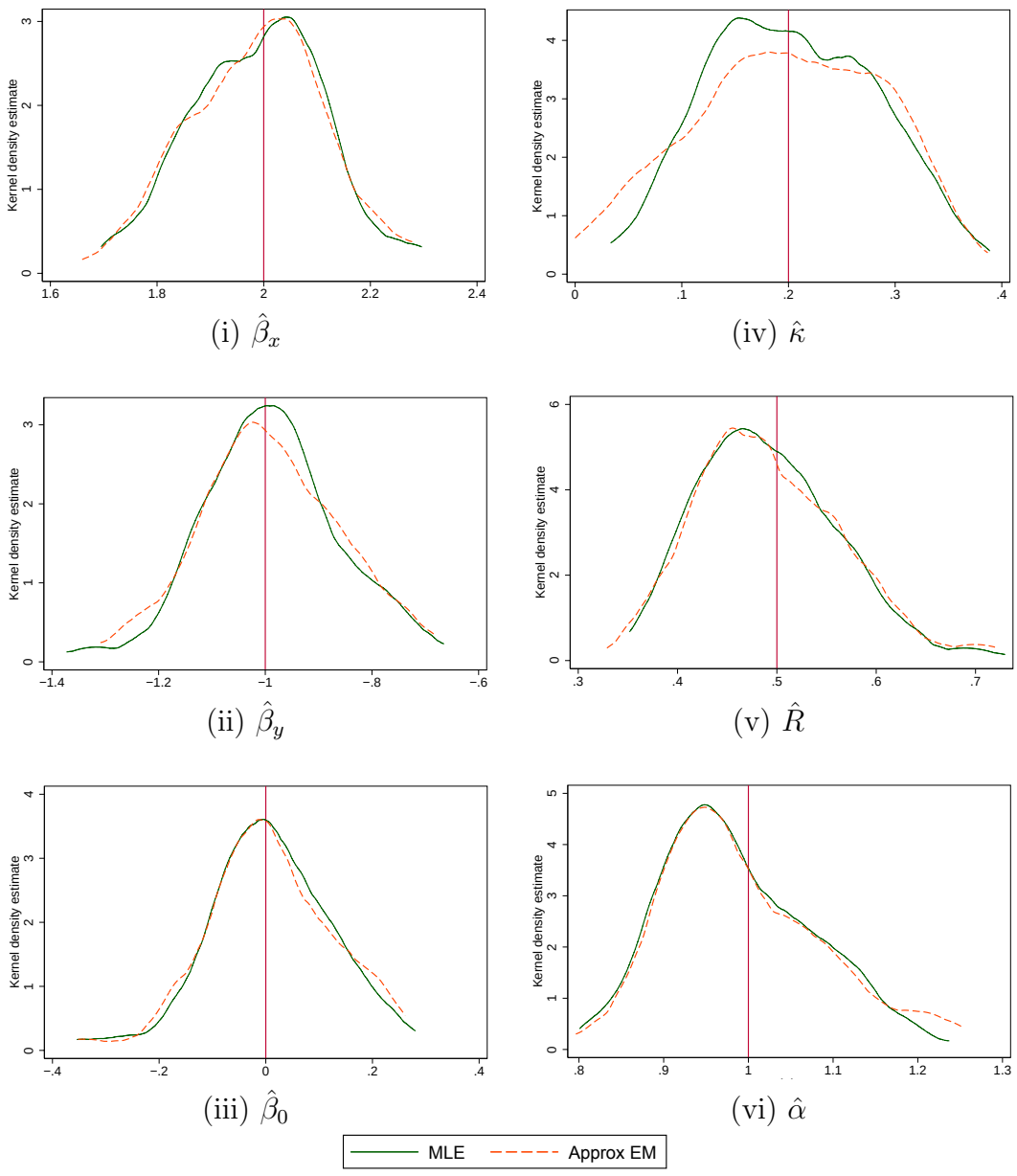


Figure 5.3: Simulated distributions of the estimates.

Table 5.2: Correlations of the parameter estimates.

(a) Between parameter estimates, within an estimation method.

	Across $\hat{\beta}$'s	Across $\hat{\theta}$'s	Between $\hat{\beta}$ and $\hat{\theta}$
Maximum likelihood (condition number = 76)			
high	$\mathbb{C}\text{orr}[\hat{\beta}_0, \hat{\beta}_x] = -0.63^{**}$	$\mathbb{C}\text{orr}[\hat{\kappa}, \hat{\alpha}] = -0.77^{**}$	$\mathbb{C}\text{orr}[\hat{\beta}_0, \hat{\kappa}] = -0.17^{*}$
low	$\mathbb{C}\text{orr}[\hat{\beta}_x, \hat{\beta}_y] = 0.13$	$\mathbb{C}\text{orr}[\hat{\alpha}, \hat{R}] = -0.16$	$\mathbb{C}\text{orr}[\hat{\beta}_x, \hat{\kappa}] = 0.01$
Approximate EM with corrections (condition number = 73)			
high	$\mathbb{C}\text{orr}[\hat{\beta}_0, \hat{\beta}_x] = -0.61^{**}$	$\mathbb{C}\text{orr}[\hat{\kappa}, \hat{\alpha}] = -0.80^{**}$	$\mathbb{C}\text{orr}[\hat{\beta}_0, \hat{\kappa}] = -0.18^{*}$
low	$\mathbb{C}\text{orr}[\hat{\beta}_x, \hat{\beta}_y] = 0.11$	$\mathbb{C}\text{orr}[\hat{\alpha}, \hat{R}] = -0.22^{*}$	$\mathbb{C}\text{orr}[\hat{\beta}_x, \hat{\kappa}] = -0.01$

**, significant at 1%; *, significant at 10%

(b) Between MLE and approximate EM with corrections estimates.

β_x	β_y	β_0	κ	α	R
0.974	0.952	0.960	0.858	0.902	0.934

The estimates and their standard errors are on about the same scale of a unity, so the higher condition numbers of the covariance matrices are due to correlations between parameter estimates.

The simulation results demonstrate that only one out of nine cross-correlations between the regression and covariance subspace parameter estimates have p-values below 10%; none of those cross-correlations is significant after Bonferroni correction (10%/9 = 1.11%). Thus the theoretical argument that the regression and covariance parameter estimates are asymptotically uncorrelated finds support in these simulation results.

Also, part (b) of Table 5.2 shows that the approximate EM with corrections estimates are following the MLEs quite closely. It might have been expected that with small proportions of missing data the estimates of the approximate EM with corrections would be close to MLEs, as long as they are based on similar estimating equations with corrections proportional to the missing data rate ν.

5.9 Discussion

This chapter has reviewed the proposed modification of the EM algorithm with application to the dissociated spatio-temporal models. The estimating equations, and thus estimates themselves, resulting from the approximate EM are biased and inconsistent,

except for the regression or trend coefficients. The corrections can be derived that restore consistency of the estimates. Those corrections have different forms for different (groups of) parameters. The resulting estimates are shown to be consistent and asymptotically normal. The asymptotic variances can be computed, either analytically (as was done here assuming normality), or empirically using the empirical estimate of the estimating equations covariance matrix.

A numerical demonstration with a small Monte Carlo study provides limited support of reasonable performance of the suggested estimator. The efficiency losses did not exceed 20%, and the estimates were found to be close to the MLEs.

The derivations in the chapter can also be used for other types of mixed models, as well as for structural equations with latent variables. In the latter applications, however, the corrections will be way more numerous, as each parameter would need to have its own correction. The estimation algorithms, and especially the secondary derivations such as variances and derivatives of the estimating equations, will become quite cumbersome.

The results of this chapter hinge on a number of assumptions. One of them is normality of the response, or measured quantity. From the perspective of M-estimation used throughout the chapter, this is just a starting point to derive the fitting function and estimating equations[1]. Other fitting functions may have been used as well, but the attractive feature of the normal likelihood (or quasi-likelihood) is that the spatial correlations can be built into the variogram specification and the observation variance-covariance matrix that may not have direct analogues with other forms such as least absolute deviations.

Another implication of the assumed normality is availability of sufficient statistic which makes the EM algorithm somewhat easier to implement. Other members of exponential family can be tried for similar approaches, although the required non-linear transformations may lead to additional terms in biases of the estimates that would have to be compensated for.

Another important assumption used in the derivations is the structural form of the covariance function. The author is not aware of any general tests of the "goodness of spatial fit" of general applicability. Comparison of different variogram specifications in empirical research is usually performed through information criteria, as different models are rarely nested. The derivations in this chapter are general enough to allow

[1] Normality can also be used to analytically derive the variance of the estimating equations, but, as mentioned above, an empirical estimate is likely to perform better under a wider variety of instances.

for quite arbitrary stationary variogram; on the other hand, as long as the model is misspecified, and there is no single likelihood evaluated and output as a result of the estimation procedure, the information criteria will not be applicable.

Finally, an important assumption being made is that of the data missing completely at random (MCAR). One of the components of this assumption is that the missing data occasions occur independently, and it is used, at least implicitly, in applying the asymptotic arguments (CLT and LLN) in deriving the corrections. If the adjacent monitoring stations are experiencing similar problems (such as understaffing in a certain part of the country, or inclement weather preventing from taking accurate measurements, etc.), then the missing data patterns will have some degree of spatial correlation.

Another component of MCAR assumption is that of constant probability of being missing that does not depend on any other covariates. This assumption is used to derive reasonably simple likelihoods, as well as corrections to the estimating equations through the lemmas of Appendix C. It will be even harder to justify given an observation that reporting rates, and thus the proportions of missing data, may differ by an order of magnitude across different monitoring sites. If the assumption of independence of the missing data at different locations can still be maintained, then the effect of varying missing data rates will be primarily on the analytical results of Appendix C, where each location will have to be assigned its own rate ν_s. The results of the Appendix will immediately become intractable, although approximations to them using some efficient missing data rates may still be possible.

Chapter 6

Future work

The research conducted within this dissertation, and in the last chapter, specifically, can be extended and further developed in a number of ways.

6.1 Separable processes

A spatio-temporal process is called *separable* if the correlations in time and in space can be separated from one another. If $\forall i, t\, Z_{it} = \mu$, then the separability property can be written down as

$$\mathbb{C}\mathrm{ov}[Z_{it}, Z_{js}] = \alpha C_s(i,j) C_t(t,s) \tag{6.1}$$

where $C_s(i,j)$ is the spatial correlation function between locations i and j, and $C_t(t,s)$ is the temporal correlation function between times t and s. Those correlation functions can include the nugget effect as $j \to i$ or $s \to t$. The covariance structure of such process has Kronecker structure, so inverses and differentials are easily available, as shown in Appendix B. In particular, computing an inverse of $NT \times NT$ covariance matrix only takes $O(N^3) + O(T^3)$ operations rather than $O(N^3T^3)$ operations for a same size matrix of an arbitrary structure; a gain of about six orders of magnitude for $N \approx 100$, $T \approx 100$, as in Chapter 4.

If some data are missing, however, the covariance matrix loses its convenient computational structure, and thus the computational costs will soar to $O(\nu^3 N^3 T^3)$ per matrix inversion (and each iteration of the numerical maximization algorithm may require several such iterations). It is then really worth studying alternative estimation procedures, and an extension of the approximate EM algorithm of Chapter 5 is a possible alternative getting us back to very efficient matrix algebra. Indeed, using an unconditional expectation $\mathbb{E}[(y_{it} - \mu_{it})(y_{js} - \mu_{js})] = \alpha C_s(i,j) C_t(t,s)$ as an approximation at the E-step

of the EM algorithm is a very simple rule. However, as was shown in Chapter 5, the estimates in general are going to be biased. Corrections to the estimating equations in the spirit of Chapter 5 will still be possible, but due to an increasing complexity of the problem, they will be more tedious.

In this setting, however, the data will no longer be independent over time, so in deriving the asymptotic properties of the estimates, stronger results on arrays of random variables would need to be used. In such derivations, a mixing property of the process must be satisfied for the estimates to be asymptotically normal[1].

By the formal symmetry of the separable process and corresponding Kronecker product structure with respect to the time and space interchange, similar mixing conditions would have to be satisfied for the spatial process for consistent estimation of the temporal correlations. This requirement seems to be ruling out non-stationary spatial processes, and only lends itself to the increasing domain asymptotics of geostatistics[2].

Also, additional results on sampling from Kronecker products of matrices extending the results of Appendix C will be necessary. In essence, Lemma C.4 aims at establishing the trace of a Kronecker product, and in that respect will serve as the basic tool, just as Lemma C.1 has been one for the results in Chapter 5.

If all such results can be established, then the corrections to restore unbiasedness of the estimating equations, and hence consistency of the estimates, can be derived. The asymptotic properties of the resulting estimates can then be established by using appropriate versions of CLT for random fields. It should be expected that the trend parameters can be estimated unbiasedly through a version of weighted least squares, while the estimates of the covariance space parameters, i.e., spatial and temporal correlations, will have to be obtained through a numeric minimization procedure, where the objective function will be a combination of the quasi-likelihood implied by the approximate EM algorithm, and the penalty terms restoring consistency.

---

[1] In the time series context, α-mixing is the following concept. Let

$$\alpha(k) \equiv \sup\{|\,\mathbb{P}\mathrm{r}\{A_t \cap B_{t+k}\} - \mathbb{P}\mathrm{r}\{A_t\}\,\mathbb{P}\mathrm{r}\{B_{t+k}\}| : A_t \in \mathcal{F}_{-\infty}^{t},\, B_{t+k} \in \mathcal{F}_{t+k}^{+\infty}\} \tag{6.2}$$

be the mixing coefficient of the time series Y_k, where $\mathcal{F}_a^b$ is the σ-algebra generated by a process Y_k, $a \leq k < b$. If $\alpha(k) \to 0$ as $k \to \infty$, the time series Y_k is said to be α-mixing. The mixing property means that correlations between "parts" of the process decay sufficiently quickly as one moves from one "part" of the space on which the process is defined to another. See Doukhan (1994).

[2] See discussion of increasing domain vs. infill asymptotics in Cressie (1993).

6.2 Unbiased estimating equations

Even in the simpler case of dissociated processes, because of the difficulties with deriving an appropriate correction to the likelihood required to estimate the spatial covariance parameters consistently, an entirely different approach can be taken to obtaining estimating equations that would be unbiased.

Consider the following function of the data that can be called *matrix weighted least squares* (MWLS):

$$Q(\theta, \beta; Y) = \frac{1}{2} \sum_{i=1}^{N} \operatorname{tr} W_i R_i W_i R_i^T \tag{6.3}$$

where the matrix residual R_i was defined in (5.27), and fixed symmetric weight matrix W_i is left undefined at the moment. Note that

$$\begin{aligned} \operatorname{tr} RR^T &= \sum_{j=1}^{d} \sum_{k=1}^{d} r_{jk}^2 = \|R\|_2^2, \\ \operatorname{tr} WRWR^T &= \|WR\|_2^2, \end{aligned} \tag{6.4}$$

due to the symmetry of the covariance residual matrix R.

The estimating equations can be derived from the differential of Q:

$$\begin{aligned} \mathsf{d}\, Q &= \frac{1}{2} \sum_{i=1}^{N} \operatorname{tr}(W_i \{\mathsf{d}\, R_i\} W_i R_i^T + W_i R_i W_i \{\mathsf{d}\, R_i\}^T) = \\ &= \sum_{i=1}^{N} \operatorname{tr}(W_i R_i W_i \{\mathsf{d}\, R_i\}^T) = \sum_{i=1}^{N} \operatorname{tr}\big(W_i \{\mathsf{d}\, R_i\} W_i R_i^T\big) = \\ &= \sum_{i=1}^{N} \operatorname{tr}\Big(W_i \big[2(Y_i^o - P_i X_i \beta)\{-\mathsf{d}\, \beta\}^T X^T P_i^T - P_i \{\mathsf{d}\, \Sigma\} P_i^T\big] W_i R_i\Big) \end{aligned} \tag{6.5}$$

The estimating equations for β are vector polynomials of third order:

$$0 = \sum_{i=1}^{N} X_i^T P_i^T W_i \big[(Y_i^o - P_i X_i \beta)(Y_i^o - P_i X_i \beta)^T - P_i \Sigma(\theta) P_i^T\big] W_i (Y_i^o - P_i X_i \beta) \tag{6.6}$$

Explicit solutions would be difficult to derive, but it is easy to establish that the equations are unbiased if the distribution of Y is symmetric (or at least has zero third

moment). Weighted least squares estimates, e.g., of the form

$$\hat{\beta} = \Big(\sum_i X_i^T W_i X_i\Big)^{-1} \Big(\sum_i X_i^T W_i Y_i\Big) \tag{6.7}$$

may be more advantageous, although they require a separate set of routines outside the framework of minimizing (6.3).

The estimating equations for θ are found from the second term of (6.5):

$$0 = \sum_{i=1}^{N} \mathrm{tr}\big[W_i P_i \{\mathsf{d}\,\Sigma\} P_i^T W_i R_i\big] \tag{6.8}$$

The comparison of (6.8) with the ML estimating equations (5.23) shows that efficient estimates asymptotically equivalent to the MLE are obtained by setting $W_i = (P_i \Sigma(\theta) P_i^T)^{-1}$. Those estimates are of course infeasible as they involve unknown θ. The practical alternatives might be:

1. use $W_i = (P_i \Sigma(\tilde{\theta}) P_i)^{-1}$ for some estimate $\tilde{\theta}$ obtained previously, either from an external source, or at the previous iteration of the optimization algorithm. This would imply higher computational costs, as an inversion of T potentially different matrices will be required. Note however that as long as the weight matrices are treated as fixed, they will only have to be inverted only once, and then the iterative algorithm can proceed by using the stored W_i's.

2. use $W_i = I$ without much concern for efficiency. This choice of the weighting matrix may be suitable to obtain the initial consistent estimates of the model parameters.

3. use $W_i = P_i \Sigma(\theta)^{-1} P_i^T$ in the hope that it will be "close" to the efficient choice. Indeed, if Σ is a diagonal matrix, then this choice is identical to the efficient choice, and it is quite possible that for a spatial field with quickly decaying correlations this choice will also be practical.

Some combinations of the weighting strategies may be in place, e.g. in a two-stage procedure: to start with an identity weighting matrix, obtain initial consistent estimates $\theta^{(1)}$, and then plug them in with a more refined weighting matrix of the form 1 or 3.

The estimators of this kind have been proposed in econometrics, where the estimation procedure is known as the *generalized method of moments* (GMM) (Hansen

1982, Mátyás 1999, Hall 2005), and quantitative sociology (structural equation models), where they are known as the *asymptotically distribution free* (ADF) method (Browne 1984). Generally, they give consistent estimates, and their asymptotic properties follow from the general theory of M-estimates outlined briefly in Appendix D. Their asymptotic distribution is normal, the variance is given by the information sandwich formula, and consistent standard errors are feasible. As mentioned above, the optimal choice of the weighting matrix is possible, and then the estimates are asymptotically efficient. If maximum likelihood estimates for the same model are available, then the optimal GMM/ADF estimates are asymptotically equivalent to the MLE estimates. An additional feature of the GMM models is that they allow to test for overidentifying restrictions, thus providing a goodness of fit test for the spatial covariance models.

Appendix A

Useful matrix calculus results

Matrix calculus is concerned with the infinitesimal properties of matrix functions. The basic book on the subject is Magnus & Neudecker (1999). They introduce the differential notation (d), show its properties, and argue quite convincingly (Chapter 9) why this notation is superior to more obvious notation such as $\partial U/\partial V$ for matrices U and V.

Let $F : S \to \mathbb{R}^{m\times p}$ be a matrix function defined on $S \subset \mathbb{R}^{n\times q}$. Let matrix $C \in \operatorname{int} S$, and let $U \in \mathbb{R}^{n\times q}$ be such that $\|U\| < r$ (i.e., $U \in B(0,r)$, an open ball of radius r centered at zero)[1], so that $C + U \in B(C,r) \in S$. If there exists a real matrix A of size $mp \times nq$ that depends on C, but not on U, such that

$$\operatorname{vec}[F(C+U)] = \operatorname{vec}[F(C)] + A(C)\operatorname{vec}[U] + \operatorname{vec}[R_C(U)] \;\; \forall U \in B(C,r) \tag{A.2}$$

and

$$\lim_{U\to 0} \frac{R_C(U)}{\|U\|} = 0 \tag{A.3}$$

then F is said to be *differentiable at* C, and $m \times p$ matrix $\mathsf{d}\,F(C;U)$ given by

$$\operatorname{vec}[\mathsf{d}\,F(C;U)] = A(C)\operatorname{vec}[U] \tag{A.4}$$

is the *first differential of* F *at* C *with an increment* U, and $mp \times nq$ matrix $A(C)$ is the *(first) derivative of* F *at* C.

[1]The norm of the matrix is taken to be the natural (spectral) norm

$$\|X\| = (\operatorname{tr} X^T X)^{1/2} \tag{A.1}$$

In this notation, the differentials of matrix functions U, V are:

$$\mathsf{d}(U+V) = \mathsf{d}\,U + \mathsf{d}\,V, \tag{A.5}$$

$$\mathsf{d}(\alpha U) = \alpha\,\mathsf{d}\,U, \tag{A.6}$$

$$\mathsf{d}\,U^T = (\mathsf{d}\,U)^T, \tag{A.7}$$

$$\mathsf{d}\,\mathrm{vec}[U] = \mathrm{vec}[\mathsf{d}\,U], \tag{A.8}$$

$$\mathsf{d}(UV) = (\mathsf{d}\,U)V + U(\mathsf{d}\,V), \tag{A.9}$$

$$\mathsf{d}(AUB) = A(\mathsf{d}\,U)B \text{ for constant matrices } A, B, \tag{A.10}$$

$$\mathsf{d}\,\mathrm{tr}\,U = \mathrm{tr}\,\mathsf{d}\,U, \tag{A.11}$$

$$\mathsf{d}(U \otimes V) = (\mathsf{d}\,U) \otimes V + U \otimes (\mathsf{d}\,V) \tag{A.12}$$

where $U \otimes V$ is Kronecker product (see Appendix B).

If additionally U is a non-degenerate square matrix, $|U| \neq 0$, then

$$\mathsf{d}\,|U| = |U|\,\mathrm{tr}[U^{-1}\,\mathsf{d}\,U], \tag{A.13}$$

$$\mathsf{d}\ln|U| = \mathrm{tr}[U^{-1}\,\mathsf{d}\,U], \tag{A.14}$$

$$\mathsf{d}\,U^{-1} = -U^{-1}(\mathsf{d}\,U)U^{-1}. \tag{A.15}$$

Mild regularity conditions (i.e., differentiability and invertibility in a neighborhood of U) that ensure existence of the differential are required for (A.13)–(A.15).

Appendix B

Kronecker products

Let A and B be two matrices of dimensions $m \times n$ and $p \times q$, respectively. The *Kronecker product* of two matrices is a matrix of dimension $mp \times nq$ defined as

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \dots & a_{1n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1}B & a_{n2}B & \dots & a_{mn}B \end{pmatrix} \tag{B.1}$$

The justification for the term "product" comes from the distributive property of Kronecker product:

$$A \otimes B \otimes C = (A \otimes B) \otimes C = A \otimes (B \otimes C), \tag{B.2}$$

$$(A + B) \otimes (C + D) = A \otimes C + A \otimes D + B \otimes C + B \otimes D, \tag{B.3}$$

$$(A \otimes B)(C \otimes D) = (AC) \otimes (BD) \tag{B.4}$$

whenever the terms in the above expressions are properly defined.

Other useful properties of Kronecker product:

$$(A \otimes B)^T = A^T \otimes B^T, \tag{B.5}$$

$$\operatorname{tr}(A \otimes B) = (\operatorname{tr} A)(\operatorname{tr} B), \tag{B.6}$$

$$\operatorname{tr}(AB)\operatorname{tr}(CD) = \operatorname{tr}(AB \otimes CD) = \operatorname{tr}\big[(A \otimes C)(B \otimes D)\big], \tag{B.7}$$

$$(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}, \tag{B.8}$$

$$A \text{ is } m \times m, B \text{ is } p \times p \Rightarrow |A \otimes B| = |A|^p |B|^m, \tag{B.9}$$

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \Rightarrow A \otimes B = A = \begin{pmatrix} A_{11} \otimes B & A_{12} \otimes B \\ A_{21} \otimes B & A_{22} \otimes B \end{pmatrix} \tag{B.10}$$

$$a^T \otimes b = ab^T = b \otimes a^T \text{ for vectors } a, b \tag{B.11}$$

whenever the appropriate matrices are defined properly.

Appendix C

Quadratic forms with missing data

Several results in Chapters 5 hinge on the expressions of the form $\operatorname{tr}[PAP^TPBP^T]$ for some matrices A and B of size d, and the incidence matrices P of dimensions $d^o \times d$ (see definition on page 60). Let us study the properties of this expression, in particular, its expectation over the missing data process. The matrices A and B would be related to the covariance matrices or the outer products of the observed data vectors. Thus the missing data mechanism works on those matrices by deleting matching rows and columns of the matrices. This is the setting of Hájek's binomial sampling scheme (Hájek 1960, Pathak 1988) where a finite population of size N is sampled at a predetermined rate p, with the number of individuals sampled being itself a random variable with distribution $\operatorname{Bin}(N, p)$. We shall denote the expectation over this sampling scheme as $\mathbb{E}_s$.

Lemma C.1. *If A and B are $d \times d$-matrices, and P is an incidence matrix consisting of (an ordered) binomial random sample without replacement of rows of I_d, then*

$$\mathbb{E}_s \operatorname{tr}(PAP^TPBP^T) = (1-\nu)\operatorname{tr}\big\{A[(1-\nu)B + \nu \operatorname{diag} B]\big\} \tag{C.1}$$

Proof. Note first that PAP^T is a random (with respect to the missing data process) minor of A, and PBP^T is a random minor of B. Then

$$\operatorname{tr}(PAP^TPBP^T) = \sum_{k \in s_i} \sum_{j \in s_i} a_{jk} b_{jk} \tag{C.2}$$

where s_i is the sample of indices (sites) available in i-th observation. The probability of sampling a given site is $1-\nu$, and each j, k combination is sampled/observed with probability $(1-\nu)^2$ by the independence assumption as a part of MCAR. The diagonal elements are sampled at the rate $1-\nu$, and the off-diagonal ones, at the rate $(1-\nu)^2$. Hence the expected value of (C.2) with respect to the missing data, or the process of

sampling the rows of I_d (random indices $j, k \in s_i$), is

$$\begin{aligned}
\mathbb{E}_S \operatorname{tr}(PAP^T PBP^T) &= \mathbb{E}_S \sum_{k\in s_i} \sum_{j\in s_i} a_{jk} b_{jk} = \\
&= (1-\nu) \sum_{k=1}^{d} a_{jj} b_{jj} + (1-\nu)^2 \sum_{1\le j\ne k\le d} a_{jk} b_{jk} = \\
&= (1-\nu)^2 \sum_{k=1}^{d} \sum_{j=1}^{d} a_{jk} b_{jk} + \nu(1-\nu) \sum_{k=1}^{d} a_{kk} b_{kk} = \\
&= (1-\nu)^2 \operatorname{tr}(AB) + \nu(1-\nu) \operatorname{tr}(A \odot B) = \\
&= (1-\nu) \operatorname{tr}\big\{A[(1-\nu)B + \nu \operatorname{diag} B]\big\} = (1-\nu) \operatorname{tr}\big\{A[B - \nu(B - \operatorname{diag} B)]\big\} \qquad \text{(C.3)}
\end{aligned}$$

where the expectation $\mathbb{E}_s$ is taken over all possible samples s_i, $\odot$ denotes Hadamard product (the entry-by-entry product of two matrices), and $\operatorname{diag} B = B \odot I$ is the diagonal part of the matrix (not a vector of the diagonal elements, but a matrix with the same diagonal as B, and zero off-diagonal elements). □

Corollary C.2. *If B is a diagonal matrix, then*

$$\mathbb{E}_s \operatorname{tr}(PAP^T PBP^T) = (1-\nu) \operatorname{tr} AB \qquad \text{(C.4)}$$

In particular, setting $B = I$, one obtains

$$\mathbb{E}_s \operatorname{tr}(PAP^T) = \mathbb{E}_s \operatorname{tr}(PAP^T PIP^T) = (1-\nu) \operatorname{tr} A \qquad \text{(C.5)}$$

Alternatively, by substituting $\operatorname{diag} B = 0$, one obtains

Corollary C.3. *If B has zeros on the diagonal, then*

$$\mathbb{E}_s \operatorname{tr}(PAP^T PBP^T) = (1-\nu)^2 \operatorname{tr} AB \qquad \text{(C.6)}$$

The derivations in Section 5.5 require computation of expectations of a cross-product involving different points in time (different incidence matrices P_i and P_j) $\mathbb{E}_s \operatorname{tr}(P_i A P_i^T P_i B P_i^T) \operatorname{tr}(P_j C P_j^T P_j D P_j^T)$ which will be dealt with in the following lemma.

Lemma C.4. *If A, B, C, D are $d \times d$ matrices, and P_i and P_j are random incidence matrices, then*

$$\mathbb{E}_s \operatorname{tr}(P_i A P_i^T P_i B P_i^T) \operatorname{tr}(P_j C P_j^T P_j D P_j^T) =$$
$$= \mathbb{E}_s \operatorname{tr}(P_i A P_i^T P_i B P_i^T)\, \mathbb{E}_s \operatorname{tr}(P_j C P_j^T P_j D P_j^T) + \delta_{ij}(1-\nu)^2 \Delta(\nu; A, B, C, D), \quad \text{(C.7)}$$
$$\Delta(\nu; A, B, C, D) =$$
$$= \frac{\nu}{1-\nu}^2 \sum_j a_{jj} b_{jj} c_{jj} d_{jj} + \nu(2-\nu)\Big(\sum_{j\neq k} a_{jk} b_{jk} c_{jk} d_{jk} + \sum_{j\neq k} a_{jk} b_{jk} c_{kj} d_{kj}\Big) +$$
$$+ (1-\nu)^2 \nu\Big(\sum_{j\neq m} a_{jj} b_{jj} c_{jm} d_{jm} + \sum_{j\neq l} a_{jj} b_{jj} c_{lj} d_{lj} + \sum_{j\neq k} a_{jk} b_{jk} c_{jj} d_{jj} +$$
$$+ \sum_{j\neq k} a_{jk} b_{jk} c_{kk} d_{kk}\Big) + (1-\nu)^3 \nu\Big(\sum_{j\neq k\neq m} a_{jk} b_{jk} c_{jm} d_{jm} + \sum_{j\neq k\neq l} a_{jk} b_{jk} c_{lj} d_{lj} +$$
$$+ \sum_{j\neq k\neq m} a_{jk} b_{jk} c_{km} d_{km} + \sum_{j\neq k\neq l} a_{jk} b_{jk} c_{lk} d_{lk}\Big) \quad \text{(C.8)}$$

where δ_{ij} is Kronecker's delta.

Proof. For different points in time, the locations sampled are independent, and so are stochastic matrices P_i and P_j. Thus the expectation in LHS of (C.7) is a product of two expectations of the form given by (C.3). For the same points in time, the product of traces becomes

$$\operatorname{tr}(P_i A P_i^T P_i B P_i^T) \operatorname{tr}(P_i C P_i^T P_i D P_i^T) = \sum_{k\in s_i} \sum_{j\in s_i} \sum_{l\in s_i} \sum_{m\in s_i} a_{jk} b_{jk} c_{lm} d_{lm} \quad \text{(C.9)}$$

The generic term of this expression is sampled at a rate $(1-\nu)^4$. However, some of the terms in this expression will be sampled at a rate higher than others. When two indices coincide, the rate becomes $(1-\nu)^3$; when three indices coincide, it is $(1-\nu)^2$; and when all four indices coincide (the diagonals of all four matrices), the rate is $1-\nu$. The breakdown of the rates and indices is given in Table C.1.

Let us compare (C.9) to a "regular" situation when the two traces are independent. There will be some extra terms in that expression due to the differences in sampling rates, as the assumption of independence understates how often a particular combination may appear in the expectation. See columns 2 and 4 of Table C.1. Then

Table C.1: Sampling probabilities for Lemma C.4.

Relation of indices	Correct Prob	Regular case	Prob under independence
$j=k=l=m$	$1-\nu$		$(1-\nu)^2$
$j=k\neq l=m$	$(1-\nu)^2$	*	$(1-\nu)^2$
$j=l\neq k=m$	$(1-\nu)^2$		$(1-\nu)^4$
$j=m\neq k=l$	$(1-\nu)^2$		$(1-\nu)^4$
$j=k=l\neq m$	$(1-\nu)^2$		$(1-\nu)^3$
$j=k=m\neq l$	$(1-\nu)^2$		$(1-\nu)^3$
$j=l=m\neq k$	$(1-\nu)^2$		$(1-\nu)^3$
$k=l=m\neq j$	$(1-\nu)^2$		$(1-\nu)^3$
$j=k\neq l,\neq m, l\neq m$	$(1-\nu)^3$	*	$(1-\nu)^3$
$j=l\neq k,\neq m, k\neq m$	$(1-\nu)^3$		$(1-\nu)^4$
$j=m\neq k,\neq l, k\neq l$	$(1-\nu)^3$		$(1-\nu)^4$
$k=l\neq j,\neq m, j\neq m$	$(1-\nu)^3$		$(1-\nu)^4$
$k=m\neq j,\neq l, j\neq l$	$(1-\nu)^3$		$(1-\nu)^4$
$l=m\neq j,\neq k, k\neq j$	$(1-\nu)^3$	*	$(1-\nu)^3$
$j\neq k\neq l\neq m$	$(1-\nu)^4$	*	$(1-\nu)^4$

$$
\begin{aligned}
&\mathbb{E}_s[\operatorname{tr}(P_iAP_i^TP_iBP_i^T)\operatorname{tr}(P_iCP_i^TP_iDP_i^T)]- \\
&\quad-\mathbb{E}_s[\operatorname{tr}(P_iAP_i^TP_iBP_i^T)]\,\mathbb{E}_s[\operatorname{tr}(P_iCP_i^TP_iDP_i^T)]= \\
&\quad=(1-\nu)\nu\sum_j a_{jj}b_{jj}c_{jj}d_{jj}+(1-\nu)^2\nu(2-\nu)\Big(\sum_{j\neq k}a_{jk}b_{jk}c_{jk}d_{jk}+\sum_{j\neq k}a_{jk}b_{jk}c_{kj}d_{kj}\Big)+ \\
&\quad+(1-\nu)^2\nu\Big(\sum_{j\neq m}a_{jj}b_{jj}c_{jm}d_{jm}+\sum_{j\neq l}a_{jj}b_{jj}c_{lj}d_{lj}+\sum_{j\neq k}a_{jk}b_{jk}c_{jj}d_{jj}+ \\
&\quad+\sum_{j\neq k}a_{jk}b_{jk}c_{kk}d_{kk}\Big)+(1-\nu)^3\nu\Big(\sum_{j\neq k\neq m}a_{jk}b_{jk}c_{jm}d_{jm}+\sum_{j\neq k\neq l}a_{jk}b_{jk}c_{lj}d_{lj}+ \\
&\quad+\sum_{j\neq k\neq m}a_{jk}b_{jk}c_{km}d_{km}+\sum_{j\neq k\neq l}a_{jk}b_{jk}c_{lk}d_{lk}\Big)\equiv(1-\nu)^2\Delta(\nu;A,B,C,D) \qquad \text{(C.10)}
\end{aligned}
$$

□

This excess part can be interpreted as extra variance due to random sampling of sites under MCAR process; it has the same stochastic structure as $\mathbb{V}[y]=\mathbb{E}[y^2]-(\mathbb{E}[y])^2$ for a numeric random variable y. The normalization $(1-\nu)^2$ is taken to make the results easier to conform with other terms in the expectations in Section 5.5.

From now on, matrices A, B, C and D will be assumed symmetric, as will be our case with the covariance matrices and their derivatives. Then equation (C.10) can be simplified:

$$\begin{aligned}\Delta(\nu; A, B, C, D) = \frac{\nu}{1-\nu}\sum_j a_{jj}b_{jj}c_{jj}d_{jj} + 2\nu\Big[(2-\nu)\sum_{j\neq k} a_{jk}b_{jk}c_{jk}d_{jk} + \\ + \sum_{j\neq m} a_{jj}b_{jj}c_{jm}d_{jm} + \sum_{j\neq k} a_{jk}b_{jk}c_{jj}d_{jj} + \\ + (1-\nu)\sum_{j\neq k\neq m} a_{jk}b_{jk}c_{jm}d_{jm} + (1-\nu)\sum_{j\neq k\neq l} a_{jk}b_{jk}c_{lk}d_{lk}\Big] \qquad \text{(C.11)}\end{aligned}$$

Appropriately, if there is no missing data ($\nu = 0$), then $\Delta(0; \cdot) = 0$, so there are no variance components associated with the missing data stochastic component.

Some special cases will be of interest.

- If say D is a diagonal matrix, then

$$\Delta(\nu; A, B, C, D) = \frac{\nu}{1-\nu}\Big[\sum_j a_{jj}b_{jj}c_{jj}d_{jj} + 2(1-\nu)\sum_{j\neq k} a_{jk}b_{jk}c_{jj}d_{jj}\Big] \qquad \text{(C.12)}$$

- If $A = C = A^T$ and $B = D = B^T$, then

$$\begin{aligned}\Delta(\nu; A, B, A, B) = \frac{\nu}{1-\nu}\sum_j a_{jj}^2 b_{jj}^2 + 2\nu\Big[(2-\nu)\sum_{j\neq k} a_{jk}^2 b_{jk}^2 + \\ + 2\sum_{j\neq m} a_{jj}b_{jj}a_{jm}b_{jm} + 2(1-\nu)\sum_{j\neq k\neq m} a_{jk}b_{jk}a_{jm}b_{jm}\Big] \qquad \text{(C.13)}\end{aligned}$$

In Section 5.5, we shall be dealing with the expressions of the form $\Delta(\nu; A, R_i^c, B, R_i^c)$ and their expectations, where A and B are symmetric matrices, and R_i^c is the matrix residual (5.112).

Lemma C.5. *If A and B are symmetric matrices, and c_{jk} is the (j,k)-th entry of the*

matrix of spatial correlations $C(\theta)$,

$$\mathbb{E}_Y\,\Delta(\nu; A, R_i^c, B, R_i^c) = \frac{\nu}{1-\nu}\sum_j 4a_{jj}b_{jj}\alpha^2(1+\kappa)^2 +$$
$$+2\nu\Big\{(2-\nu)\sum_{j\neq k} a_{jk}b_{jk}\alpha^2\big(c_{jk}^2+(1+\kappa)^2\big) + \sum_{j\neq k} 2(a_{jj}b_{jk}+a_{jk}b_{jj})c_{jk}\alpha^2(1+\kappa) +$$
$$+(1-\nu)\sum_{j\neq k} a_{jk}\sum_{j\neq l,k\neq l}(b_{jl}+b_{lk})\alpha^2\big[c_{jk}c_{jl}+(1+\kappa)c_{kl}\big]\Big\} \tag{C.14}$$

Proof. Note first that

$$\Delta(\nu; A, R_i^c, B, R_i^c) = \frac{\nu}{1-\nu}\sum_j a_{jj}r_{jj}^c b_{jj}r_{jj}^c + 2\nu\Big[(2-\nu)\sum_{j\neq k} a_{jk}r_{jk}^c b_{jk}r_{jk}^c +$$
$$+\sum_{j\neq m} a_{jj}r_{jj}^c b_{jm}r_{jm}^c + \sum_{j\neq k} a_{jk}r_{jk}^c b_{jj}r_{jj}^c +$$
$$+(1-\nu)\sum_{j\neq k\neq m} a_{jk}r_{jk}^c b_{jm}r_{jm}^c + (1-\nu)\sum_{j\neq k\neq l} a_{jk}r_{jk}^c b_{lk}r_{lk}^c\Big], \tag{C.15}$$
$$\mathbb{E}_Y\,\Delta(\nu; A, R_i^c, B, R_i^c) = \frac{\nu}{1-\nu}\sum_j 4a_{jj}b_{jj}\alpha^2(1+\kappa)^2 + 2\nu\Big[(2-\nu)\sum_{j\neq k} a_{jk}b_{jk}(\mu_{jkjk}-\alpha^2c_{jk}^2) +$$
$$+\sum_{j\neq m} a_{jj}b_{jm}(\mu_{jjjm}-\alpha^2(1+\kappa)c_{jm}) + \sum_{j\neq k} a_{jk}b_{jj}(\mu_{jkjj}-\alpha^2(1+\kappa)c_{jk}) +$$
$$+(1-\nu)\sum_{j\neq k\neq m} a_{jk}b_{jm}(\mu_{jkjm}-\alpha^2c_{jm}c_{jk}) + (1-\nu)\sum_{j\neq k\neq l} a_{jk}b_{lk}(\mu_{jklk}-\alpha^2c_{jk}c_{lk})\Big], \tag{C.16}$$

where c_{jk} is (j,k) entry of the matrix of spatial correlations $C(\varphi)$, see (5.29), and μ_{ijkl} is the (centered) fourth moment of data, see (5.114). Note also that $\alpha c_{jk} = \alpha(1+\kappa) - \gamma(\|s_j - s_k\|)$ where $\gamma(\cdot)$ is the semivariogram of the spatial field.

Consider further a decomposition of the random field at location $\mathbf{s}_k$ to a part associated with the value at location $\mathbf{s}_j$, and an idiosyncratic part:

$$y_k - \mu_k = \rho_{jk}(y_j - \mu_j) + u_{k|j}, \quad \mathbb{E}\,u_{k|j}y_j = 0, \tag{C.17}$$

where ρ_{jk} is the correlation between two observations,

$$\rho_{jk} = c_{jk}/(1+\kappa) \tag{C.18}$$

Then

$$\mu_{jjjk} = \mathbb{E}(y_j - \mu_j)^3(y_k - \mu_k) = \mathbb{E}\,\rho_{jk}(y_j - \mu_j)^4 = 3c_{jk}\alpha^2(1+\kappa), \tag{C.19}$$

$$\begin{aligned}\mu_{jjkk} &= \mathbb{E}(y_j - \mu_j)^2(y_k - \mu_k)^2 = \mathbb{E}\,\rho_{jk}^2(y_j - \mu_j)^4 + \mathbb{E}(y_j - \mu_j)^2 u_{k|j}^2 = \\ &= 3c_{jk}^2\alpha^2 + \alpha^2(1+\kappa)^2(1-\rho_{jk}^2) = \alpha^2\big(2c_{jk}^2 + (1+\kappa)^2\big)\end{aligned} \tag{C.20}$$

and any permutation of the subindices gives the same answer. Further, if the location $\mathbf{s}_m$ is next considered with

$$y_m - \mu_m = \rho_{jm}(y_j - \mu_j) + \gamma_{km|j}u_{k|j} + v_{m|kj} \tag{C.21}$$

where $v_{m|kj}$ is the part of y_m uncorrelated with either y_j and y_k, then

$$\gamma_{km|j} = \frac{\rho_{km} - \rho_{jk}\rho_{jm}}{1 - \rho_{jk}^2}, \tag{C.22}$$

$$\begin{aligned}\mu_{jjkm} &= \mathbb{E}(y_j - \mu_j)^2(y_k - \mu_k)(y_m - \mu_m) = \\ &= \mathbb{E}\,\rho_{kj}\rho_{mj}(y_j - \mu_j)^4 + \mathbb{E}\,\gamma_{km|j}(y_j - \mu_j)^2 u_{k|j}^2 = \\ &= \alpha^2(1+\kappa)^2\big[3\rho_{jk}\rho_{jm} + \rho_{km} - \rho_{jk}\rho_{jm}\big] = \alpha^2(1+\kappa)^2\big[2\rho_{jk}\rho_{jm} + \rho_{km}\big] = \\ &= \alpha^2\big[2c_{jk}c_{jm} + (1+\kappa)c_{km}\big]\end{aligned} \tag{C.23}$$

Grouping the results together, and using appropriate symmetry arguments (the value of μ_{jklm} is insensitive to the permutation of indices, $a_{jk} = a_{kj}$, $b_{jk} = b_{kj}$), one obtains (C.14). □

In Section 5.5, it will be necessary that $\Delta(\cdot)$ is finite. It is however clear that

$$|\Delta(\nu; A, B, C, D)| \leq \frac{\nu}{1-\nu} \sum_{i,j=1,d} |a_{ij}| \sum_{i,j=1,d} |b_{ij}| \sum_{i,j=1,d} |c_{ij}| \sum_{i,j=1,d} |d_{ij}| \tag{C.24}$$

Likewise, as each entry of the fourth order moment matrix K is bounded by $\mathbb{E}_Y[(y_i - \mu_i)^4] = \alpha^2(1+\kappa)^2$, $\mathbb{E}_Y\,\Delta(\nu; A, R_i^c, B, R_i^c)$ will also be finite.

Appendix D

Consistency and asymptotic normality of M-estimates

This appendix summarizes some results on asymptotic behavior of the estimates defined by a set of estimating equations.

D.1 Notation

The results in this Appendix apply to an i.i.d. sample:

$$X_1, \ldots, X_n \sim \text{i.i.d. } F(x, \eta), \quad \eta \in R^k, x \in R^m \tag{D.1}$$

and functionals of the distribution

$$\mathbb{E}\, \psi_j(x, \theta) = 0, \quad j = 1, \ldots, p \tag{D.2}$$

where the expectation may mean the expectation either with respect to the theoretical distribution

$$\mathbb{E}_\eta\, g(x) = \int g(x) dF(x) \tag{D.3}$$

or with respect to the sampling distribution

$$\mathbb{E}_n\, g(x) = \frac{1}{n} \sum_{i=1}^{n} g(X_i) \tag{D.4}$$

The equations define a p-dimensional parameter of a statistical model: $\theta \in \Theta \subset \mathbb{R}^p$.

An estimator $\hat{\theta}$ is obtained as a solution to

$$\mathbb{E}_n\, \psi_j(x, \theta_n) = 0 \tag{D.5}$$

(or an approximate solution if the exact solution is not feasible).

The population solution will be denoted θ_0:

$$\mathbb{E}_\eta \, \psi_j(x, \theta_0) = 0 \tag{D.6}$$

D.2 Consistency conditions

The consistency property can be viewed in global or local sense. Most consistency conditions are local, i.e., about the estimators that converge stochastically to a neighborhood of the "true parameter" θ_0. The global consistency would require the zero point of $\psi_j(\cdot)$ to be well separated, i.e., there are no other points θ_α (including infinity) s.t. $\lim_{\theta \to \theta_\alpha} \mathbb{E}_\eta \, \psi_j(\theta, x) = 0$.

van der Vaart (1998)

van der Vaart (1998) gives a description of dual estimating problem of either maximizing a criterion

$$M_n(x, \theta) = \frac{1}{n} \sum_i m(X_i, \theta) \quad vs. \quad M(\theta) = \mathbb{E}_\eta \, m(x, \theta) \tag{D.7}$$

or solving a system of estimating equations (D.2). He also introduces a concept of *near maximization*:

$$M_n(\hat{\theta}_n) \geq \sup_\theta M_n(\theta) - o_p(1) \tag{D.8}$$

and notes that solving for a zero of a set of estimating equations is related to maximizing the (GMM-type) criterion

$$-\| \, \mathbb{E}_n \, \psi_j(\theta, x) \| \to \max_\theta \tag{D.9}$$

Theorem 5.9: Let Ψ_n be random vector-valued functions, Ψ be a fixed vector-valued function of θ s.t. θ_0 is the solution of

$$\Psi(\theta_0) = 0 \tag{D.10}$$

If

$$\sup_{\theta \in \Theta} \| \Psi_n(\theta) - \Psi(\theta) \| \xrightarrow{p} 0, \tag{D.11}$$

and for every $\epsilon > 0$

$$\inf_{\theta : \| \theta - \theta_0 \| \geq \epsilon} \| \Psi(\theta) \| > 0 \tag{D.12}$$

Then

$$\forall \text{ sequence } \hat{\theta}_n : \Psi_n(\hat{\theta}_n) = o_p(1) \Rightarrow \hat{\theta}_n \xrightarrow{p} \theta_0 \tag{D.13}$$

van der Vaart (1998) discusses some sufficient conditions for Theorem 5.9.

$$\Theta \text{ is compact}, \Psi \in C(\Theta), \theta_0 \text{ is a unique max} \Rightarrow (D.12), \tag{D.14}$$

$$\begin{gathered} \forall\, x \in \mathbb{R}^m\, \psi_j(\theta, x) \in C(\Theta, x) \text{ and} \\ \exists \text{ integrable } g(x) : \forall\, \theta \in \Theta\, |\psi_j(\theta, x)| \leq g(x) \Rightarrow \\ \Rightarrow \mathbb{E}_n\, \psi_j(x, \theta) \text{ are Glivenko-Cantelli} \Rightarrow (D.11) \end{gathered} \tag{D.15}$$

His next result is to show how the conditions can be relaxed.

Lemma 5.10: Let $\Theta \subset \mathbb{R}$, Ψ_n be random functions, and Ψ, a fixed function, of θ.

$$\forall\, \theta\ \Psi_n(\theta) \xrightarrow{p} \Psi(\theta) \tag{D.16}$$

$$\forall\, \omega \in \Omega\ \Psi_n(\theta) \in C(\Theta), \tag{D.17}$$

$$\forall\, \omega \in \Omega\ \exists!\, \hat{\theta}_n : \Psi_n(\hat{\theta}_n) = 0 \text{ or } \Psi_n(\cdot) \text{ is non-decreasing with } \Psi_n(\hat{\theta}_n) = o_p(1), \tag{D.18}$$

$$\exists\, \theta_0 : \forall\, \epsilon > 0\ \Psi(\theta_0 - \epsilon) < 0 < \Psi(\theta_0 + \epsilon) \tag{D.19}$$

Then

$$\hat{\theta}_n \xrightarrow{p} \theta_0$$

van der Vaart (1998) finalizes discussion of consistency conditions with Wald's consistency proof that involves upper-semicontinuity and boundedness from above of the objective function $M(x, \theta)$, and finiteness of the objective function at the maximum θ_0.

Huber (1967)

Huber (1967) is a classic paper that establishes consistency and normality of the M-estimates, and introduces the sandwich formula for the asymptotic variance of M-estimates (see next section).

The following are the conditions under which Huber (1967) establishes consistency. The set of plausible parameter values Θ is assumed to be locally compact with a countable base, $\langle \mathcal{X}, \mathfrak{U}, P \rangle$ is the probability space. He considers the behavior of the

sequence of estimators $\theta_n : \mathcal{X}^n \to \Theta$ s.t.

$$\frac{1}{n}\sum_{i=1}^{n} \Psi(X_i, \theta_n) \to 0 \tag{D.20}$$

either a.s. or in probability.

(B-1) $\forall\, \theta$, $\Psi(x, \theta)$ is $\mathfrak{U}$-measurable, and $\Psi(x, \theta)$ is separable[1]

(B-2) The function Ψ is a.s. continuous in θ:

$$\lim_{\theta' \to \theta} \|\Psi(x, \theta') - \Psi(x, \theta)\| = 0 \quad \text{a.s.} \tag{D.21}$$

(B-2′) As the neighborhood $U(\theta)$ shrinks to $\{\theta\}$,

$$\mathbb{E}\big(\sup_{\theta' \in U} \|\Psi(x, \theta') - \Psi(x, \theta)\|\big) \to 0 \tag{D.22}$$

(B-3) $\lambda(\theta) = \mathbb{E}\,\Psi(x, \theta)$ exists $\forall\, \theta \in \Theta$; $\exists!\, \theta_0 : \lambda(\theta_0) = 0$

(B-4) $\exists\, b_0, b(\theta) \in C(\Theta) : b(\theta) \geq b_0 > 0$:

$$\sup_{\theta} \frac{\|\Psi(x, \theta)\|}{b(\theta)} \text{ is integrable}$$

$$\liminf_{\theta \to \infty} \frac{\|\lambda(\theta)\|}{b(\theta)} \geq 1,$$

$$\mathbb{E}\Big[\limsup_{\theta \to \infty} \frac{\|\Psi(x, \theta) - \lambda(\theta)\|}{b(\theta)}\Big] < 1$$

Lemma 2 of Huber (1967): (B-1), (B-4) and (D.20) $\Longrightarrow$ there is a compact set $C \subset \Theta$: any sequence θ_n a.s. ultimately stays in C (i.e., $\mathbb{P}\mathrm{r}\{\exists\, n_0 : \forall\, n > n_0, \theta_n \in C\} = 1$).

Theorem 2 of Huber (1967): (B-1), (B-2′), (B-3); θ_n satisfies (D.20) and conclusion of Lemma 2 of Huber (1967) $\Longrightarrow$ $\theta_n \to \theta_0$ a.s. and in probability.

[1] $\exists$ a null set $N \subset \mathcal{X} : P(N) = 0$, countable $\Theta' \subset \Theta$ s.t. for every open set $U \subset \Theta$ and for every closed interval A, the sets $\{x | \psi_j(x, \theta) \in A, \forall\, \theta \in U\}$ and $\{x | \psi_j(x, \theta) \in A, \forall\, \theta \in U \cap \Theta'\}$ differ by at most a subset of N.

D.3 Asymptotic normality

The question of asymptotic distribution of the estimates coming from a set of estimating equations is the next one to address once consistency of the estimators has been established. Versions of the central limit theorem are usually applicable that demonstrate asymptotic normality of the estimators. As with the proofs of consistency, the major part of showing the asymptotic normality in each particular situation is to verify the regularity conditions.

van der Vaart (1998)

van der Vaart (1998) gives the following heuristic argument which is a generic proof of the asymptotic normality based on a Taylor series expansion around θ_0:

$$\begin{aligned} 0 = \mathbb{E}_n\, \Psi(\hat{\theta}_n, X) = \mathbb{E}_n\, \Psi(\theta_0, X) + \mathbb{E}_n\, D\Psi(\theta_0, X)(\hat{\theta}_n - \theta_0) + \\ + \frac{1}{2}\big[I_p \otimes (\hat{\theta}_n - \theta_0)\big]^T \big[\mathbb{E}_n\, D^2\Psi(\tilde{\theta}_n, X)\big](\hat{\theta}_n - \theta_0) \end{aligned} \tag{D.23}$$

for some $\tilde{\theta}_n$ between θ_0 and $\hat{\theta}_n$, where $D\Psi(\cdot)$ is the matrix of derivatives of $\Psi(\cdot)$ with respect to θ, and $D^2\Psi(\cdot)$ is a stacked matrix of second order derivatives of $\Psi(\cdot)$:

$$D^2\Psi(\theta, X) = \begin{pmatrix} D^2\psi_1(\theta, X) \\ \vdots \\ D^2\psi_p(\theta, X) \end{pmatrix}, \quad D^2\psi_j(\theta, X) = \begin{pmatrix} \frac{\partial^2\psi_1}{\partial\theta_1\partial\theta_1} & \frac{\partial\psi_1}{\partial\theta_1\partial\theta_2} & \cdots & \frac{\partial\psi_1}{\partial\theta_1\partial\theta_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2\psi_1}{\partial\theta_p\partial\theta_1} & \frac{\partial\psi_1}{\partial\theta_p\partial\theta_2} & \cdots & \frac{\partial\psi_1}{\partial\theta_p\partial\theta_p} \end{pmatrix} \tag{D.24}$$

Then

$$\begin{aligned} \sqrt{n}(\hat{\theta}_n - \theta_0) = \\ = -\Big(\mathbb{E}_n\, D\Psi(\theta_0, X) + \frac{1}{2}\big[I_p \otimes (\hat{\theta}_n - \theta_0)^T\big]\big[\mathbb{E}_n\, D^2\Psi(\tilde{\theta}_n, X)\big]\Big)^{-1} \sqrt{n}\, \mathbb{E}_n\, \Psi(\theta_0, X) \end{aligned} \tag{D.25}$$

The last term is asymptotically normal, as it is a sum of i.i.d. random vectors, provided $\mathbb{E}_\eta\, \Psi(\theta_0, X)\Psi(\theta_0, X)^T$ is finite, so the CLT holds for this sum. The first term of the multiplying matrix is an average, and thus by the law of large numbers converges to $\mathbb{E}_\eta\, D\Psi(\theta_0, X)$. The second term in the multiplying matrix is a product of $o_P(1)$ and $O_P(1)$, and is negligible. The resulting matrix must be non-singular, or, in other words, the estimates must be functionally independent. If all those conditions are satisfied,

the whole expression is asymptotically normal:

$$\begin{aligned} \sqrt{n}(\hat{\theta}_n - \theta_0) &\xrightarrow{d} N\big(0, A^{-1}B(A^T)^{-1}\big), \\ A = \mathbb{E}_\eta\, D\Psi(\theta_0, X), &\quad B = \mathbb{E}_\eta\, \Psi(\theta_0, X)\Psi(\theta_0, X)^T \end{aligned} \tag{D.26}$$

For the correctly specified maximum likelihood estimates, $-A = B = \mathcal{I}$, the Fisher information matrix.

van der Vaart (1998) further presents several theorems that dwell on different regularity conditions needed to prove asymptotic normality. The next one is the best applicable to the general M-estimates.

Theorem 5.21. For each $\theta \in U \subset E$, a Euclidean space, let $x \mapsto \psi(x, \theta)$ be a measurable vector-valued function s.t. $\forall\, \theta_1, \theta_2$ in a neighborhood of θ_0 and a measurable function $\dot{\psi}$ with $\mathbb{E}\,\dot{\psi}^2 < \infty$,

$$\|\psi(x, \theta_1) - \psi(x, \theta_2)\| \leq \dot{\psi}(x)\|\theta_1 - \theta_2\|. \tag{D.27}$$

Assume that $\mathbb{E}\,\|\psi(x, \theta_0)\|^2 < \infty$ and that the map $\theta \mapsto \mathbb{E}\,\psi(x, \theta)$ is differentiable at θ_0 with non-singular derivative matrix V_{θ_0}. If $\mathbb{E}_n[\psi(x, \hat{\theta}_n)] = o_P(n^{-1/2})$, and $\hat{\theta}_n \xrightarrow{p} \theta_0$, then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -V_{\theta_0}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(X_i, \theta_0) + o_P(1) \tag{D.28}$$

In particular, the sequence $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically normal with mean zero and covariance matrix $V_{\theta_0}^{-1}\big(\mathbb{E}[\psi(x, \theta_0)\psi(x, \theta_0)^T]\big)V_{\theta_0}^{-1T}$.

For continuously differentiable functions $\psi(\cdot)$, a natural candidate for the dominating function $\dot{\psi}$ is $\sup_{\theta \in U(\theta_0)} \partial\psi/\partial\theta$ taken over a neighborhood of θ_0.

Huber (1967)

Upon discussing (ways to establish) consistency of an M-estimator, Huber (1967) continues to give the conditions under which the estimator will be asymptotically normal. Those are the assumptions he uses:

Set up:

$$\begin{aligned} \Theta \subset \mathbb{R}^m, &\quad \langle \mathcal{X}, \mathfrak{U}, P \rangle \text{ is probability space}, \quad \psi : \mathcal{X} \times \Theta \to \mathbb{R}^m \\ \lambda(\theta) = \mathbb{E}[\psi(x, \theta)], &\quad u(x, \theta, d) = \sup_{\|\tau - \theta\| \leq d} \|\psi(x, \tau) - \psi(x, \theta)\| \end{aligned} \tag{D.29}$$

Near zero of an estimator $T_n = T_n(x_1, \ldots, x_n)$:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(x_i, T_n) \xrightarrow{p} 0 \tag{D.30}$$

(N-1) $\forall\, \theta \in \Theta, \psi(x, \theta)$ is $\mathfrak{U}$-measurable, and $\psi(x, \theta)$ is separable (see Assumption (A-1) earlier)

(N-2) $\exists\, \theta_0 \in \Theta : \lambda(\theta_0) = 0$

(N-3) $\exists\, a > 0, b > 0, c > 0, d_0 > 0$:

1. $\forall \theta : \|\theta - \theta_0\| \leq d_0 \Rightarrow \|\lambda(\theta)\| \geq a\|\theta - \theta_0\|$
2. $\forall \theta : \|\theta - \theta_0\| + d \leq d_0, d \geq 0 \Rightarrow \mathbb{E}\, u(x, \theta, d) \leq bd$
3. $\forall \theta : \|\theta - \theta_0\| + d \leq d_0, d \geq 0 \Rightarrow \mathbb{E}[u(x, \theta, d)^2] \leq cd$

(N-4) $\mathbb{E}[\|\psi(x, \theta_0)\|^2] < \infty$

Here, $\|\cdot\|$ is any norm equivalent to Euclidean norm. He then shows the following results:

Lemma 3: Assumptions (N-1), (N-2) and (N-3) imply that

$$\sup_{|\tau - \theta| \leq d_0} \frac{\left\| \sum_{i=1}^{n} \left[\psi(x_i, \tau) - \psi(x_i, \theta) - \lambda(\tau) + \lambda(\theta) \right] \right\|}{\sqrt{n} + n\|\lambda(\tau)\|} \xrightarrow{p} 0 \text{ as } n \to \infty \tag{D.31}$$

Theorem 3: Assume that (N-1) to (N-4) hold and that T_n satisfies (D.30). If $\mathbb{P}\mathrm{r}\{|T_n - \theta_0| \leq d_0\} \to 1$, then

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(x, \theta_0) + \sqrt{n}\lambda(T_n) \xrightarrow{p} 0 \tag{D.32}$$

Corollary: Under the conditions of Theorem 3 of Huber (1967), assume that λ has a non-singular derivative A at θ_0, so that $\|\lambda(\theta) - \lambda(\theta_0) - A(\theta - \theta_0)\| = o(\|\theta - \theta_0\|)$. Then $\sqrt{n}(T_n - \theta_0)$ is asymptotically normal with mean 0 and covariance matrix $A^{-1}BA^{T^{-1}}$, where $B = \mathbb{C}\mathrm{ov}[\psi(x, \theta_0)]$.

D.4 A proof of consistency

This subsection is based on Smith (2005).

Suppose θ is a p-dimensional parameter and $\{\psi_{n,j},\ j = 1, ..., p\}$ are a family of p estimating functions for sample size n, such that the estimator $\hat{\theta}_n$ is defined as the solution of

$$\psi_{n,j}(\hat{\theta}_n) = 0,\ j = 1, ..., p. \tag{D.33}$$

We also write Ψ_n for the p-dimensional vector function whose individual coordinates are $\psi_{n,j},\ j = 1, ..., p$.

Let θ_0 be the true value of θ and suppose there is an open neighborhood of θ_0, denoted $\mathcal{N}$, and a constant $\beta > 0$ such that:

$$\psi_{n,j}(\theta_0) = o_p(n^{-\beta}),\ j = 1, ..., p,\ n \to \infty, \tag{D.34}$$

$$h_{n,j,k}(\theta) = \frac{\partial \psi_{n,j}(\theta)}{\partial \theta_k} \to h_{j,k}(\theta),\ j, k = 1, ..., p,\ n \to \infty, \tag{D.35}$$

where the convergence in (D.35) is in probability, uniformly over $\theta \in \mathcal{N}$.

We also assume that the matrix $H = \{h_{j,k}\}$ is negative definite in the sense that $u^T H u < 0$ whenever $u \neq 0$; note, however, that we do not assume H is symmetric. It's not clear whether this distinction is important, but the case of asymmetric H could arise in cases where the equation $\Psi_n(\theta) = 0$ does not arise from the minimization of some function of θ.

The canonical example is when $\psi_{n,1}, ..., \psi_{n,p}$ are $\frac{1}{n}$ times the first-order derivatives of the negative log likelihood; in that case, (D.34) holds for any $\beta < \frac{1}{2}$, and (D.35) with H the Fisher information matrix. However, for a wide class of estimating equations that are unbiased (in the sense that $E\{\psi_{n,j}(\theta_0)\} = 0$) we can expect (D.34) to hold for $\beta < \frac{1}{2}$ by the CLT, and (D.35) by the LLN.

Let $\alpha \in (0, \beta)$ and define $B_n = \{\theta :\ ||\theta - \theta_0|| \leq n^{-\alpha}\}$ where $||\cdot||$ is the L^2 norm. Then we claim:

Proposition 1. With probability tending to 1 as $n \to \infty$, there exists a solution $\hat{\theta}_n$ of the equations $\Psi_n(\hat{\theta}_n) = 0$, such that $\hat{\theta}_n \in B_n$.

Proof. The proof relies on *Brouwer's fixed point theorem:* if B is a ball in p-space and $f :\ B \to B$ is continuous, then f has a fixed point, i.e. there exists $x \in B$ such that $f(x) = x$.

The idea of the proof is to show there is a $t > 0$ such that, with probability tending to 1 as $n \to \infty$, the function $f_n(\theta) = \theta - t\Psi_n(\theta)$ maps B_n to itself. If this is true then,

by Brouwer's fixed point theorem, there must then exist a $\hat{\theta}_n \in B$ (with probability tending to 1) such that $\Psi_n(\hat{\theta}_n) = 0$, proving the desired result.

We use the approximation $\Psi_n(\theta) \approx \Psi_n(\theta) - \Psi_n(\theta_0) \approx H(\theta - \theta_0)$, and we first prove a preliminary result about the matrix H. Assume $\|u\| = 1$ and consider

$$\|(I + tH)u\| - \|u\| = 2tu^T Hu + t^2 u^T H^T Hu. \tag{D.36}$$

For any given u we have $u^T Hu < 0$, so there exists some positive t depending on u, call it $t(u)$, that minimizes the value of (D.36). But $t(u)$ is a continuous function of u, defined on the compact set $\|u\| = 1$, so the minimum of $t(u)$ is also positive. It then follows that for this choice of t, there exists $\delta < 1$ such that $\|(I - tH)u\| \leq \delta\|u\|$ for all u such that $\|u\| = 1$, and hence by scaling, for all u. Henceforth, this is the value of t we use to define the function f_n.

For a given θ_n such that $\|\theta_n - \theta_0\| < n^{-\alpha}$, we write $\Psi_n(\theta_n) = \Psi_n(\theta_0) + H_n(\theta_n^*)(\theta_n - \theta_0)$ where $H_n(\theta)$ is the matrix with entries $h_{n,j,k}(\theta)$ and θ_n^* is some value of θ between θ_0 and θ_n. Now write

$$\begin{aligned} f_n(\theta_n) - \theta_0 &= \theta_n - \theta_0 - t\Psi_n(\theta_n) \\ &= (I - tH)(\theta_n - \theta_0) - t(H_n(\theta_n^*) - H)(\theta_n - \theta_0) - t\Psi_n(\theta_0). \end{aligned}$$

Let E_n be the event $|t|\|H_n(\theta_n^*) - H\| \leq \frac{1-\delta}{3}$ and let F_n be the event $|t\Psi_n(\theta_0| \leq \frac{1-\delta}{3} n^{-\alpha}$. Both E_n and F_n have probability tending to 1 as $n \to \infty$ and on $E_n \cap F_n$,

$$\begin{aligned} |f_n(\theta_n) - \theta_0| &\leq \|(I - tH)(\theta_n - \theta_0)\| + |t|\|(H_n(\theta_n^*) - H)(\theta_n - \theta_0)\| + |t\Psi_n(\theta_0)| \\ &\leq \delta n^{-\alpha} + \frac{1-\delta}{3} n^{-\alpha} + \frac{1-\delta}{3} n^{-\alpha} \\ &= \frac{2+\delta}{3} n^{-\alpha} \\ &< n^{-\alpha}. \end{aligned}$$

Therefore, $f_n(\theta_n) \in B_n$, with probability tending to 1, which is what we were trying to prove.

Corollary 1. With probability tending to 1, the solution $\hat{\theta}_n$ is unique. (In other words, there is only one solution with the ball B_n. The result says nothing about the possibility of multiple solutions outside B_n.)

Proof. Let $\inf_{u:\ \|u\|=1}(-u^T H u) = \eta > 0$. Let G_n be the following event:

$$\inf_{u:\ \|u\|=1} \inf_{\theta \in B_n} (-u^T H_n(\theta) u) > \frac{\eta}{2}$$

Because of (D.35), $\Pr\{G_n\} \to 1$. On G_n, if $H_n(\theta)u = 0$ for any $\theta \in B_n$, we must have $u = 0$.

Suppose $\hat{\theta}_n \in B_n$ and $\tilde{\theta}_n \in B_n$ are two solutions of $\Psi_n(\theta) = 0$. Then $0 = \Psi_n(\hat{\theta}_n) - \Psi_n(\tilde{\theta}_n) = H_n(\theta_n^*)(\hat{\theta}_n - \tilde{\theta}_n)$ for some $\theta_n^* \in B_n$. On G_n, this implies $\hat{\theta}_n - \tilde{\theta}_n = 0$. The result follows.

Corollary 2. Suppose we strengthen (D.34) to

$$\sqrt{n}\Psi_n(\theta_0) \quad \xrightarrow{d} \quad N[0, J] \tag{D.37}$$

for some matrix J. Then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \quad \xrightarrow{d} \quad N[0, H^{-1} J H^{-T}]. \tag{D.38}$$

Proof. The previous approximations imply that $\hat{\theta}_n - \theta_0 = -H^{-1}\Psi_n(\theta_0)(1 + o_p(1))$. The result is immediate from this.

BIBLIOGRAPHY

Abramovitz, M. & Stegun, I. A. (1964), *Handbook of Mathematical Functions*, National Bureau of Standards, Washington, D.C.

Akaike, H. (1973), Information theory and an extension of the maximum likelihood principle, *in* B. N. Petrov & F. Csaki, eds, 'Second International Symposium on Information Theory', Akademiai Kiado, Budapest, Hungary, pp. 267–281.

Beveridge, S. (1992), 'Least-squares estimation of missing values in time-series', *Communications in Statistics—Theory and Methods* **21**(12), 3479–3496.

Billingsley, P. (1995), *Probability and Measure*, John Wiley and Sons, New York.

Borovkov, A. A., Borovkov, K. & Borovkova, O. (1999), *Probability Theory*, T&F STM.

Box, G. E. P. & Cox, D. R. (1964), 'An analysis of transformations', *Journal of the Royal Statistical Society* **26**, 211–252.

Browne, M. W. (1984), 'Asymptotically distribution-free methods for the analysis of the covariance structures', *British Journal of Mathematical and Statistical Psychology* **37**, 62–83.

Cox, L. H. (2000), 'Statistical issues in the study of air pollution involving airborne particulate matter', *Environmetrics* **11**, 611–626.

Cressie, N. (1993), *Statistics for Spatial Data*, 2nd edn, Wiley, New York.

Cressie, N. & Huang, H.-C. (1999), 'Classes of nonseparable, spatio-temporal stationary covariance functions', *The Journal of American Statistical Association* **94**, 1330–1340.

Demmel, J. W. (1997), *Applied Numerical Linear Algebra*, SIAM.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the EM algorithm (with discussion)', *Journal of the Royal Statistical Society B* **39**, 1–38.

Doukhan, P. (1994), *Mixing. Properties and Examples*, Vol. 85 of *Lecture Notes in Statistics*, Springer, New York.

EPA (1997*a*), 'EPA's revised particulate matter standards, fact sheet'. http://www.epa.gov/ttn/oarpg/naaqsfin/pmfact.html.

EPA (1997*b*), 'National ambient air quality standards for particulate matter', *Federal Register* **62**, 38651–38760. http://www.epa.gov/ttn/oarpg/naaqsfin/pmnaaqs.pdf.

Gneiting, T. (2002), 'Nonseparable, stationary covariance functions for space-time data', *The Journal of the American Statistical Association* **97**, 590–601.

Green, P. J. & Silverman, B. W. (1994), *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, Vol. 58 of *Monographs on Statistics and Applied Probability*, Chapman & Hall.

Haas, T. (2002), 'New systems for modeling, estimating, and predicting a multivariate spatio-temporal process', *Environmetrics* **13**, 311–332.

Hájek, J. (1960), 'Limiting distributions in simple random sampling from a finite population', *Publications of Mathematical Institute of Hungarian Academy of Sciences, Series A* **5**, 361–374.

Hall, A. R. (2005), *Generalized Method of Moments*, Oxford University Press.

Hansen, L. P. (1982), 'Large sample properties of generalized method of moments estimators', *Econometrica* **12**, 347–360.

Harvey, A. C. & Pierse, R. G. (1984), 'Estimating missing observations in economic time series', *Journal of the American Statistical Association* **79**(385), 125–131.

Hastie, T. & Tibshirani, R. (1990), *Generalized additive models*, Chapman & Hall/CRC.

Hoar, T. J., Milliff, R. E., Nychka, D., Wikle, C. K. & Berliner, L. M. (2003), 'Winds from a Bayesian hierarchical model: Computation for atmosphere-ocean research', *Journal of Computational and Graphical Statistics* **12**, 781–807.

Holland, D. M., De Oliveira, V., Cox, L. H. & Smith, R. L. (2000), 'Estimation of regional trends in sulfur dioxide over the eastern united states', *Environmetrics* **11**, 373–393.

Huber, P. (1967), The behavior of the maximum likelihood estimates under nonstandard conditions, *in* 'Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability', Vol. 1, University of California Press, Berkeley, pp. 221–233.

Ibrahim, J. G. (1990), 'Incomplete data in generalized linear models', *Journal of the American Statistical Association* **85**, 765–769.

Ibrahim, J. G., Chen, M.-H. & Lipsitz, S. R. (2001), 'Missing responses in generalised linear mixed models when the missing data mechanism is nonignorable', *Biometrika* **88**(2), 551–564.

Kohn, R. & Ansley, C. F. (1986), 'Estimation, prediction, and interpolation for ARIMA models with missing data', *Journal of the American Statistical Association* **81**(395), 751–761.

Kolenikov, S. (2001), 'Review of Stata 7', *Journal of Applied Econometrics* **16**, 637–646.

Laird, N., Lange, N. & Stram, D. (1987), 'Maximum likelihood computations with repeated measures: Application of the EM algorithm', *Journal of the American Statistical Association* **82**, 97–105.

Little, R. J. A. & Rubin, D. B. (2002), *Statistical Analysis with Missing Data*, Wiley Series in Probability and Statistics, 2nd edn, Wiley, New York.

Ljung, G. M. (1982), 'The likelihood function for a stationary gaussian autoregressive-moving average process with missing observations', *Biometrika* **69**(1), 265–268.

Louis, T. A. (1982), 'Finding the observed information matrix when using the EM algorithm', *Journal of the Royal Statistical Society* **44**, 226–233.

Magnus, J. R. & Neudecker, H. (1999), *Matrix differential calculus with applications in statistics and econometrics*, 2nd edn, John Wiley & Sons.

Mardia, K. V., Kent, J. T. & Bibby, J. M. (1980), *Multivariate Analysis*, Academic Press, London.

Mátyás, L., ed. (1999), *Generalized Method of Moments Estimation*, Cambridge University Press.

McLachlan, G. G. & Krishnan, T. (1997), *The EM Algorighm and Extensions*, John Wiley and Sons, New York.

Meng, X.-L. & Rubin, D. B. (1991), 'Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm', *Journal of the American Statistical Association* **86**(416), 899–909.

Pathak, P. K. (1988), Simple random sampling, *in* P. R. Krishnaiah & C. R. Rao, eds, 'Handbook of Statistics', Vol. 6, Elsevier Science Publishers, pp. 97–109.

Penzer, J. & Shea, B. (1997), 'The exact likelihood of an autoregressive-moving average model with incomplete data', *Biometrika* **84**(4), 919–928.

Schwarz, G. (1978), 'Estimating the dimension of a model', *Annals of Statistics* **6**, 461–464.

Smith, R., Kolenikov, S. & Cox, L. H. (2003), 'Spatio-temporal modeling of $PM_{2.5}$ data with missing values', *Journal of Geophysical Research – Atmospheres* **128**(D24). 9004, doi:10.1029/2002JD002914.

Smith, R. L. (2003), Environmental statistics. Unpublished manuscript under revision for publication as a book.
http://www.stat.unc.edu/postscript/rs/envnotes.ps.

Smith, R. L. (2005). Private communication.

Stata Corp. (2001), *Stata Statistical Software: Release 7*, College Station.

Stein, M. (2002), 'Models for spatial-temporal covariances'. Presentation at SAMSI/GSP Workshop on Spatio-Temporal Modeling. http://www.cgd.ucar.edu/stats/Workshop2003/stein.pdf.

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

van der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge University Press, Cambridge.

Wikle, C. K. (2003), 'Hierarchical Bayesian models for predicting the spread of ecological processes', *Ecology* **83**, 1382–1394.

Wikle, C. K., Milliff, R. E., Nychka, D., & Berliner, L. M. (2001), 'Spatiotemporal hierarchical Bayesian modeling: Tropical ocean surface winds', *Journal of the American Statistical Association* **96**, 382–398.

Zimmerman, D. (1989), 'Computationally efficient restricted maximum likelihood estimation of generalized covariance functions', *Mathematical Geology* **21**(655–672).