Approximate Likelihoods for Spatial Processes

Petruţa C. Caragea

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.

Chapel Hill

2003

Approved by

Advisor: Dr. Richard L. Smith

Reader: Dr. Amarjit Budhiraja

Reader: Dr. Edward Carlstein

Reader: Dr. David Holland

Reader: Dr. Chuanshu Ji

ABSTRACT

Petruţa C. Caragea: Approximate Likelihoods for Spatial Processes

(Under the direction of Dr. Richard L. Smith)

Many applications of spatial statistics involve evaluating a likelihood over samples of several hundred data locations. If the underlying field is Gaussian with some spatial covariance structure, this evaluation involves calculating the inverse and determinant of the covariance matrix. Although this is feasible for up to about 100 observations, it is often troublesome for sample sizes larger than 100. To take advantage of the benefits of maximum likelihood estimates for large arrays of data, it is necessary to establish efficient approximations to the likelihood. We consider several such approximations based on grouping the observations into clusters and building an estimating function by accounting for variability both between and within groups. This way, the estimation becomes practical for considerably larger data sets. In this thesis we present the proposed alternatives to the likelihood function, and an analysis of the asymptotic efficiency of the estimators yielded by them. The theoretical method applies to any kind of spatial process, but an analogous time series model is used for illustration and explicit computation. In this context, since the standard Fisher information techniques of calculating the asymptotic variance of the alternative estimators would not lead to correct conjectures, we employ a method based on the "information sandwich" technique and a Corollary to the Martingale Central Limit Theorem (application to quadratic forms of independent normal random variables). Furthermore, we illustrate the asymptotic behavior of the alternative parameters in the spatial setting with results from a simulation study.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter I

Introduction

1.1 Motivation

Many applications of spatial statistics involve evaluating a likelihood function over a sample with an increasing number of data locations. For example, Holland *et. al.* (2002) analyzed the Clean Air Status and Trends Network (CASTNet) data set, which was developed by the U.S. Environmental Protection Agency (EPA) in conjunction with National Oceanic and Atmospheric Administration (NOAA) in order to monitor air quality and meteorological measurements across the United States. Established in 1987, CASTNet almost doubled the number of site locations from 38 in 1989 (35 in the Eastern part of the U.S.) to 70 sites across the U.S. in 2001. Holland *et. al.* focused on establishing a spatial map for trends in two pollutants: sulfur dioxide (SO_2), and sulfate particles (SO_4^{2-}); that is, the paper estimated the spatial parameters associated with these trends. To estimate these spatial parameters, the authors developed and used an algorithm which involved maximum likelihood estimation. In fact, the underlying field was assumed to be Gaussian with a spatial covariance function in some given family. The evaluation of the likelihood function involved calculating the inverse and determinant of the covariance matrix. Although this analysis is computationally feasible for CASTNet, it would not be so for a much larger network. Experience shows that by the time the number of locations increases to be in the hundreds, the impact of the high dimensionality on calculating the inverse and the determinant of the covariance matrix makes computing maximum likelihood estimates intractable. Moreover, data sets that encompass hundreds if not thousands of location sites are becoming more prevalent. For example, the Historical Climate Network (HCN) developed and maintained by NOAA now has near 6000 location sites. In order to be able to take advantage of the benefits of maximum likelihood estimates in the setting of such high dimensionality, it is necessary to establish efficient approximations to the likelihood.

In this thesis, we consider several approximate likelihoods based on grouping the observations into clusters and building an estimating function by accounting for variability both between and within clusters. Theoretical results derived for an analogous time series problem allow us to compare the three approximation schemes. These results are built around the general idea that calculations for the variance of the alternative estimator can be performed using the "information sandwich" principle, after we have expressed the derivatives of the pseudo-likelihood function as a quadratic sum of independent normal random variables. We conclude by illustrating the new method with simulations.

1.2 Outline

This thesis is made up of four distinct parts. Chapter II introduces some theoretical concepts necessary for evaluating the approximations. The application of the Central Limit Theorem for Martingales to quadratic forms of normal random variables plays a central role in the evaluations. Given its prominent function, we review the theorem and one of its corollaries.

Chapter III is the most mathematically intensive. We first propose an alternative calculation for the asymptotic variance of the maximum likelihood estimator in the classical case (i.e. the estimator obtained by maximizing the exact likelihood function). This method involves the use of the application of the Martingale Central Limit Theorem to quadratic forms of independent normal random variables and the "information sandwich" technique. We call this technique the *expansion method.* This method occupies a central place throughout this work and we apply its principles for all the alternative estimation methods we propose. Next we analyze in detail three approximation schemes — Big Blocks, Small Blocks, and Hybrid — for an AR(1) model. All three estimation methods are based on the concept of dividing the original time series in a number of blocks containing an equal number of sample points.

The first estimator, “Big Blocks”, is the most simplistic of all. We first compute the mean value for each block. Assuming the original covariance structure is known, it is routine algebra to compute the variance covariance matrix for the block means and hence the underlying likelihood function for the means process. This is the function we maximize in order to obtain the estimator. To compare its efficiency with the classical case, we calculate the asymptotic variance following the expansion method.

The second estimator is called the “Small Blocks”. We derive this estimator under the assumption that there is independence between blocks. We calculate the likelihood function for each block, which is readily available since the original covariance structure is known. The function to be minimized under the block independence assumption is obtained by multiplying the individual block-likelihoods. The last step is to compare this estimator with the classical MLE. In order to measure its efficiency we need to calculate its asymptotic variance. This is performed applying the principles outlined in the expansion method.

The last estimator is based on a combination of the two schemes mentioned above. Naturally, we expect the “Big Blocks” estimator to exhibit some loss in efficiency due to representing the whole block through its mean only, while the assumption of independence between blocks in the second case will also induce some efficiency reduction, although not as large as for the previous estimator. We construct the “Hybrid” estimation function as follows: first compute the block means, their covariance structure and their underlying likelihood function. Then, assume that given the block mean, the blocks are independent. Although this assumption cannot always be verified in practice, it is a reasonable working assumption. The next step is to compute the conditional likelihood of each block, given its mean. This step involves calculation of the conditional mean and variance-covariance matrix, and it is a direct application of general multivariate normal techniques. Based on the conditional block independence, we construct the pseudo-likelihood function to be the product of the block means likelihood and the individual conditional block likelihoods. This is not an exact likelihood, due to the conditional independence. As in the previous cases, we want to compare the efficiency of this estimator to the classical case, and thus we need to compute its asymptotic variance. This is performed using the expansion method.

As a check, we calculate the relative efficiencies of these three estimators on simulated data sets and compare the results with the ones obtained through theoretical derivations.

Chapter IV opens with a brief overview of spatial statistics with emphasis on the statistical concepts utilized in the development of the approximations. Next we present a theoretical description of the alternative estimation schemes adapted to the spatial setting. We conclude by illustrating the asymptotic behavior of the estimators through a simulation study.

Chapter V consists of an application of the proposed approximations to the likelihood to a large spatial data set. This concludes the presentation of this thesis, by illustrating the performance of the estimators when dealing with a real data set, presenting the impact of these methods on practical issues.

Chapter VI presents an extension for the time series problem to its spatial equivalent and describes a promising approach for the more general case.

Chapter II

Theoretical Background

2.1 Some Properties of the Multivariate Normal Distribution

Throughout this work we refer to properties of the multivariate normal Distributions. This section is intended as a very brief summary of these properties relevant to this work. For a more detailed description of these results see, for example, the Appendix A in Mardia, Kent and Bibby (1979).

The random vector $\mathbf{X}$ of length m is said to follow a multivariate normal distribution if $\mathbf{a}^T\mathbf{X}$ follows a univariate normal distribution for every $\mathbf{a} \in R^m$.

Every multivariate normal distribution has a well-defined mean vector and covariance matrix. Furthermore, if $\mathbf{X}$ is a multivariate normal random vector of length m with $E[\mathbf{X}] = \mu$ and covariance matrix $\mathrm{Cov}[\mathbf{X}, \mathbf{X}^T] = \mathbf{\Sigma}$ and if $\mathbf{\Sigma}$ is positive definite, then $\mathbf{X}$ has density

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{m/2} | \mathbf{\Sigma} |^{1/2}} \exp\left\{ -\frac{1}{2}(\mathbf{x} - \mu)^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \mu)) \right\}. \tag{2.1}$$

Suppose the multivariate normal random vector $\mathbf{X}$ is partitioned into two components: $\mathbf{X} = (\mathbf{X}_1^T, \mathbf{X}_2^T)^T$ where $\mathbf{X_1}$ has m_1 components and $\mathbf{X_2}$ has m_2 components. Then we can write the distribution of $(\mathbf{X}_1^T, \mathbf{X}_2^T)^T$ as

$$N\left(\begin{pmatrix} \mu_{\mathbf{1}} \\ \mu_{\mathbf{2}} \end{pmatrix}, \begin{pmatrix} \mathbf{\Sigma_{11}} & \mathbf{\Sigma_{12}} \\ \mathbf{\Sigma_{21}} & \mathbf{\Sigma_{22}} \end{pmatrix} \right) \tag{2.2}$$

where for $i = 1, 2$, $\mu_\mathbf{i}$ has length m_i, and for $i, j = 1, 2$, $\mathbf{\Sigma_{ij}}$ is a $m_i \times m_j$ matrix. Then, we have the following theorem:

Theorem 2.1 (Multivariate Normal Conditional Distribution). *The conditional distribution of* $\mathbf{X_1}$ *given* $\mathbf{X_2} = \mathbf{x_2}$ *is*

$$\mathcal{N}(\mu_\mathbf{1} + \mathbf{\Sigma_{12}}\, \mathbf{\Sigma_{22}^{-1}}\, (\mathbf{x_2} - \mu_\mathbf{2}),\ \mathbf{\Sigma_{11}} - \mathbf{\Sigma_{12}}\, \mathbf{\Sigma_{22}^{-1}}\, \mathbf{\Sigma_{21}}). \tag{2.3}$$

In a few cases we work with covariance matrices of a particular structure, and make use of the following basic linear algebra facts:

Let $\mathbf{A}$ be a $(n \times n)$ non-singular matrix and let $\mathbf{a}$ and $\mathbf{b}$ be n-dimensional vectors. Then the following identities are satisfied:

$$|A + a\ b^T| = |A|\ (1 + b^T A^{-1} a) \tag{2.4}$$

and

$$(A + a\ b^T)^{-1} = A^{-1} - \{(A^{-1}a)\ (b^T A^{-1})\ (1 + b^T A^{-1} a)^{-1} \tag{2.5}$$

2.2 "Information Sandwich" Formula

Throughout this thesis we refer to what has come to be known as the "information sandwich" approach. This method is a technique used when calculating the theoretical covariance matrix of estimators yielded by pseudo-likelihood functions. This method is illustrated, among others, by Liang and Zeger (1986) and White (1982). We present here a general discussion of this technique.

2.2.1 Asymptotic Normality

Suppose we have a statistical model indexed by a finite-dimensional parameter θ, and suppose an estimate $\tilde{\theta}_n$ is constructed by minimizing a criterion function $S_n(\theta)$. The parameter n is just an index which we shall let tend to ∞; in most cases, however, n will represent the sample size. The expression $S_n(\theta)$ will denote some "measure of fit" such as sum of squares, a likelihood or a pseudo-likelihood. We assume the true parameter value is θ_0 and that $\tilde{\theta}_n$ is a consistent estimator. We also assume that $S_n(\theta)$ is at least twice continuously differentiable in θ, and that its underlying distribution is sufficiently smooth so that the function $H(\theta)$, defined below, is also continuous in a neighborhood of θ_0. Let $\nabla f(\theta)$ for any function f denote the vector of first-order partial derivatives of f with respect to the components of θ, and $\nabla^2 f$ the matrix of second-order partial derivatives.

By a Taylor expansion, we have

$$0 = \nabla S_n(\tilde{\theta}_n) = \nabla S_n(\theta_0) + \nabla^2 S_n(\theta_n^*)(\tilde{\theta}_n - \theta_0)$$

where θ_n^* lies on the straight line joining $\tilde{\theta}_n$ to θ_0. Hence

$$\tilde{\theta}_n = \theta_0 - \{\nabla^2 S_n(\theta_n^*)\}^{-1} \nabla S_n(\theta_0) \ . \tag{2.6}$$

We assume

(SA1) $\frac{1}{n}\nabla^2 S_n(\theta) \xrightarrow{p} H(\theta)$ as $n \to \infty$ uniformly on some neighborhood of θ_0, where $H(\cdot)$ is a matrix-valued function, continuous near θ_0, with $H(\theta_0)$ invertible,

(SA2) $\frac{1}{\sqrt{n}}\nabla S_n(\theta_0) \xrightarrow{d} \mathcal{N}(0, V(\theta_0))$ for some covariance matrix $V(\theta_0)$.

Assumptions (SA1) and (SA2) are satisfied for regular maximum likelihood problems in which $S_n(\theta)$ is the negative log likelihood for the parameter θ, since this case reduces to dealing with a sum over n i.i.d. terms. However, these assumptions are also valid much more generally for a wide variety of estimation criteria.

Since $H(\theta)$ is continuous in θ and invertible at θ_0, it follows that $H(\theta)^{-1}$ is continuous near θ_0, and hence that

$$\left\{\frac{1}{n}\nabla^2 S_n(\theta_n^*)\right\}^{-1} \xrightarrow{p} H(\theta_0)^{-1} \ . \tag{2.7}$$

To reach the final conclusion, we refer to Slutsky's Theorem (a reference would be, for example, Casella and Berger (1990), Theorem 5.3.5):

Theorem 2.2. *If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, where c is a constant, then* *(i)* $Y_n \, X_n \xrightarrow{d} c \, X$

(ii) $X_n + Y_n \xrightarrow{d} X + c$

Therefore, combining (SA1), equation (2.7) and conclusion (i) in Theorem 2.2, we conclude that

$$\begin{aligned} \sqrt{n}(\tilde{\theta}_n - \theta_0) &= -\left\{ \frac{1}{n} \nabla^2 S_n(\theta_n^*) \right\}^{-1} \times \frac{1}{\sqrt{n}} \, \nabla S_n(\theta_0) \\ &\xrightarrow{d} \mathcal{N}(0, \; H(\theta_0)^{-1} \, V(\theta_0) \; H(\theta_0)^{-1} \;) \, . \end{aligned} \tag{2.8}$$

In standard maximum likelihood theory, V and H both define the Fisher information matrix so (2.8) is just a restatement of the well-known asymptotic normality of the maximum likelihood estimator. In general, however, V and H are not the same, and the phrase *information sandwich* has been coined to describe the matrix $H^{-1}VH^{-1}$.

2.2.2 Consistency Assumption

One of the assumptions of the argument in section 2.2.1 is the consistency of the estimators. As a general comment, criterion functions for the general processes that we describe in this work are complicated, there is no guarantee about their convexity, and it is unreasonable to try to prove that S_n attains its global maximum near θ_0. Therefore we concentrate on verifying consistency of a local maximum. Takeshi Amemiya (1985) states the general conditions under which we have local consistency of the estimators (Theorem 4.1.2):

Theorem 2.3. *Assume:*

(A) Θ is an open subset of the Euclidean p-space (the true value θ_0 is an interior point of Θ)

(B) The criterion function $S_n(\theta)$ is a measurable function for all $\theta \in \Theta$, and ∇S_n exists and is continuous in an open neighborhood of θ_0

(C) $\frac{1}{n}S_n(\theta)$ converges in probability uniformly to a nonstochastic function $S(\theta)$ in an open neighborhood of θ_0, and $S(\theta)$ attains a strict local maximum at θ_0

Then *there is a consistent root of the equation* $\nabla S_n = 0$
(i.e. for some sequence $\epsilon_n \to 0$*,*

$$P\{\exists\ \theta^* \text{s.t.} \mid \theta^* - \theta_0 \mid < \epsilon_n, \nabla S_n(\theta^*) = 0\} \to 1, \quad \text{as } n \to \infty\ .)$$

The conditions of Theorem 2.3 are not too difficult to check. For the most general cases, we illustrate the method here. Note first that assumptions (A) and (B) are immediately satisfied by the criterion functions. To verify the assumption (C), suppose that the first order derivatives of S_n are bounded on a neighborhood of θ_0, and that $\frac{1}{n}E|\nabla S_n(\phi)| \leq K$, on a neighborhood of θ_0. Using a first order Taylor's expansion, it is clear that for some θ_n^* and θ_n^{**} between θ_0 and θ, we have

$$\begin{aligned} \frac{1}{n}S_n(\theta) - \frac{1}{n}S_n(\theta_0) &= \frac{1}{n}\ \nabla S_n(\theta_n^*)(\theta - \theta_0) \qquad (2.9)\\ S(\theta) - S(\theta_0) &= \nabla S(\theta_n^{**})(\theta - \theta_0) \qquad (2.10)\end{aligned}$$

Therefore the difference between (2.9) and (2.10) is

$$\left\| \left(\frac{1}{n}S_n(\theta) - S(\theta)\right) - \left(\frac{1}{n}S_n(\theta_0) - S(\theta_0)\right) \right\| \leq \Gamma\ \|\theta - \theta_0\ \| \qquad (2.11)$$

where Γ has finite expectation. Note that the right-hand side of the equation (2.11) converges to 0 uniformly over a decreasing sequence of neighborhoods of the form $\|\ \theta - \theta_0\ \| < \epsilon_n$, for any sequence of ϵ_n tending to 0.
Also, note that

$$\frac{1}{n}S_n(\theta_0) - S(\theta_0) \xrightarrow{p} 0$$

by the Law of Large Numbers. Therefore,

$$\frac{1}{n}S_n(\theta) - S(\theta)$$

converges to 0 uniformly on a neighborhood of θ_0 such that $\|\ \theta - \theta_0\ \| < \epsilon_n$, which proves condition (C) in Theorem 2.3.

2.3 A Central Limit Theorem for Martingales

An important tool in computing the asymptotic variances for the proposed alternative estimators is based on an application of the Martingale Central Limit Theorem to quadratic forms. Here we give a brief description of this CLT and the application we utilize throughout this work.

Suppose we have a discrete-time stochastic process $\{X_n,\ n = 0, \pm 1, \pm 2, ..., \}$ and also an increasing sequence of σ-fields $\mathcal{F}_n$.

We say that X_n is a *martingale* with respect to $\{\mathcal{F}_n\}$ if

(i) X_n is measurable with respect to $\mathcal{F}_n$ for each n,

(ii) $\mathrm{E}\{X_n | \mathcal{F}_{n-1}\} = X_{n-1}$ for each n.

If X_n is a martingale then we say that the process $Y_n = X_n - X_{n-1}$ is a *martingale difference sequence*, or MDS for short.

Billingsley (1995, Theorem 35.12, page 476) gives the following discrete-time *Martingale Central Limit Theorem*:

Suppose $\{Y_{nk},\ k \geq 1\}$ is a MDS with respect to $\{\mathcal{F}_{nk}\}$ for each n, and $\sigma^2_{nk} = \mathrm{E}\{Y^2_{nk} | \mathcal{F}_{n,k-1}\}$. If

(1) $\sum_{k=1}^{\infty} \sigma^2_{nk} \xrightarrow{p} \sigma^2 > 0$ as $n \to \infty$,

(2) $\sum_{k=1}^{\infty} \mathrm{E}\{Y^2_{nk} I(|Y_{nk}| \geq \epsilon)\} \to 0$ as $n \to \infty$ for each $\epsilon > 0$,

then $\sum_{k=1}^{\infty} Y_{nk} \xrightarrow{d} \mathcal{N}[0, \sigma^2]$.

2.3.1 Application to Quadratic Forms in Normal Random Variables

Consider the sequence

$$S_n = \sum_{\{i,j:\ i \leq j\}} a_{n,i,j} \xi_i \xi_j, \tag{2.12}$$

where $\{\xi_i\}$ are independent $\mathcal{N}[0,1]$, coefficients $\{a_{n,i,j}\}$ are defined for each n. We are interested in limits as $n \to \infty$. In principle the sum in (2.12) extends across $1 \leq i \leq j < \infty$ though in practice the sum is often truncated, with n denoting the length of the sequence.

To calculate the mean,

$$m_n = \mathrm{E}\{S_n\} = \sum_i a_{n,i,i}\ . \tag{2.13}$$

For the variance,

$$v_n = \mathrm{Var}[S_n] = \sum_{\{i,j,k,\ell:\ i \leq j,\quad k \leq \ell\}} a_{n,i,j}\ a_{n,k,\ell}\ \mathrm{Cov}[\xi_i\,\xi_j,\ \xi_k\,\xi_\ell].$$

However, to compute $\mathrm{Cov}[\xi_i\ \xi_j, \xi_k\ \xi_\ell]$, we note four cases, where the covariance does not automatically equal 0:

$$\begin{aligned} &\text{(a)}\quad i = j = k = \ell, \\ &\text{(b)}\quad i = j,\ k = \ell \neq i \\ &\text{(c)}\quad i = k,\ j = \ell \neq i \\ &\text{(d)}\quad i = \ell,\ j = k \neq i. \end{aligned} \tag{2.14}$$

Under the additional condition that $i \leq j$ and $k \leq \ell$, case (d) is vacuous. In case (b), $\mathrm{Cov}[\xi_i\ \xi_j, \xi_k\ \xi_\ell]$ reduces to $\mathrm{Cov}[\xi_i^2, \xi_k^2]$, which is 0 since ξ_i and ξ_k are independent. For case (a), we note that $\mathrm{Var}[\xi_i^2] = 2$, while for case (c), $\mathrm{Cov}[\xi_i\xi_j, \xi_k\xi_\ell] = \mathrm{Var}[\xi_i\ \xi_j] = \mathrm{E}[\xi_i^2\ \xi_j^2] = 1$.

Therefore,

$$v_n = 2\sum_i a_{n,i,i}^2 + \sum_{\{i,j:\ i<j\}} a_{n,i,j}^2. \tag{2.15}$$

This implies the natural conjecture that with m_n and v_n defined by (2.13) and (2.15),

$$\frac{S_n - m_n}{\sqrt{v_n}} \xrightarrow{d} \mathcal{N}[0,1]. \tag{2.16}$$

Theorem 2.4. *Suppose*

$$(A1) \qquad \max_i a_{n,i,i}^2/v_n \to 0 \text{ as } n \to \infty$$

$$(A2) \qquad \max_k \left(\sum_{i:\ i<k} a_{n,i,k}^2 \right) /v_n \to 0 \text{ as } n \to \infty$$

Then (2.16) holds.

Proof. Define

$$Y_{nk} = \frac{1}{\sqrt{v_n}} \left(\sum_{\{i,j:\ i\le j\le k\}} a_{n,i,j}\xi_i\,\xi_j - \sum_{\{i,j:\ i\le j\le k-1\}} a_{n,i,j}\xi_i\,\xi_j - a_{n,k,k} \right)$$

and $\mathcal{F}_{n,k}$ the σ-field generated by all $\{\xi_i,\ i \le k\}$. We can rewrite this in the form

$$\begin{aligned} Y_{nk} &= \frac{1}{\sqrt{v_n}} \left\{ a_{n,k,k}(\xi_k^2 - 1) + \sum_{i:\ i<k} a_{n,i,k}\xi_i\,\xi_k \right\} \\ &= A_{nk}\xi_k^2 + B_{nk}\xi_k + C_{nk} \quad \text{say,} \end{aligned}$$

where A_{nk}, B_{nk} and C_{nk} are $\mathcal{F}_{n,k-1}$-measurable random variables with $A_{nk}+C_{nk} = 0$. However, we immediately have $\mathrm{E}\{Y_{nk}|\mathcal{F}_{n,k-1}\} = 0$, so $\{Y_{nk}\}$ is a martingale difference sequence, and

$$\sigma_{nk}^2 \;=\; \mathrm{E}\{Y_{nk}^2|\mathcal{F}_{n,k-1}\}$$

$$= \quad 3A_{nk}^2 + B_{nk}^2 + C_{nk}^2 + 2A_{nk}C_{nk}$$

$$= \quad \frac{1}{v_n}\left\{2a_{n,k,k}^2 + \left(\sum_{i:\ i<k} a_{n,i,k}\xi_i\right)^2\right\}.$$

Then

$$\sum_k \sigma_{nk}^2 = \frac{1}{v_n}\left(2\sum_k a_{n,k,k}^2 + \sum_{\{i,k:\ i<k\}} a_{n,i,k}^2\xi_i^2 + 2\sum_{\{i,j,k:\ i<j<k\}} a_{n,i,k}a_{n,j,k}\xi_i\,\xi_j\right). \quad (2.17)$$

From (2.17) and (2.15) it is clear that $\mathrm{E}\{\sum_k \sigma_{nk}^2\} = 1$.

The variance of $\sum_k \sigma_{nk}^2$ can be broken up into three terms:

$$\mathrm{Var}\left[\sum_k \sigma_{nk}^2\right] = \frac{1}{v_n^2}\mathrm{Cov}\left[\sum_{\{i,k:\ i<k\}} a_{n,i,k}^2\xi_i^2, \sum_{\{j,k:\ j<k\}} a_{n,j,k}^2\xi_j^2,\right] + \quad (2.18)$$

$$\frac{4}{v_n^2}\mathrm{Cov}\left[\sum_{\{i,j,k: i<j<k\}} a_{n,i,k}a_{n,j,k}\xi_i\,\xi_j, \sum_{\{r,s,k: r<s<k\}} a_{n,r,k}a_{n,s,k}\xi_r\,\xi_s\right] \quad (2.19)$$

$$+ \quad \frac{4}{v_n^2}\mathrm{Cov}\left[\sum_{\{i,k: i<k\}} a_{n,i,k}^2\xi_i^2, \sum_{\{r,s,k: r<s<k\}} a_{n,r,k}a_{n,s,k}\xi_r\,\xi_s\right]. \quad (2.20)$$

The expression in (2.20) is identically 0, but for (2.18) and (2.19) to tend to 0 as $n \to \infty$ (a sufficient condition for $\sum_k \sigma_{nk}^2 \xrightarrow{p} 1$), we require

$$\frac{1}{v_n^2}\sum_{\{i,k:\ i<k\}} a_{n,i,k}^4 \quad \to 0, \quad (2.21)$$

$$\frac{1}{v_n^2}\sum_{\{i,j,k:\ i<j<k\}} a_{n,i,k}^2 a_{n,j,k}^2 \quad \to 0. \quad (2.22)$$

Now (2.22) follows from (2.21) by the Cauchy-Schwartz inequality. Moreover, if (A2) holds, we have

$$\begin{aligned}
\frac{1}{v_n^2} \sum_{\{i,k:\ i<k\}} a_{n,i,k}^4 &= \sum_k \frac{1}{v_n^2} \sum_{i:\ i<k} a_{n,i,k}^4 \\
&\leq \sum_k \frac{1}{v_n^2} \left(\sum_{i:\ i<k} a_{n,i,k}^2 \right)^2 \\
&\leq \sum_k \frac{1}{v_n} \left(\sum_{i:\ i<k} a_{n,i,k}^2 \right) \cdot \max_k \frac{1}{v_n} \left(\sum_{i:\ i<k} a_{n,i,k}^2 \right) \\
&\leq \max_k \frac{1}{v_n} \left(\sum_{i:\ i<k} a_{n,i,k}^2 \right) \\
&\to 0.
\end{aligned}$$

Therefore, (A2) is sufficient for both (2.21) and (2.22) to be true.

This proves that condition (1) of the Martingale Central Limit Theorem is true, with $\sigma^2 = 1$. To prove condition (2), we note that

$$\mathrm{E}\left\{ Y_{nk}^2 I(|Y_{nk}| > \epsilon) \right\} \leq \mathrm{E}\left\{ \frac{Y_{nk}^4}{\epsilon^2} I(|Y_{nk}| > \epsilon) \right\} \leq \frac{1}{\epsilon^2} \mathrm{E}\{Y_{nk}^4\},$$

and therefore it suffices to prove

$$\sum_k \mathrm{E}\{Y_{nk}^4\} \to 0. \tag{2.23}$$

Now, $Y_{nk} = A_{nk}(\xi_k^2 - 1) + B_{nk}\xi_k$, using the inequality $(a+b)^4 \leq 8(a^4 + b^4)$ shows that

$$Y_{nk}^4 \leq 8A_{nk}^4(\xi_k^2 - 1)^4 + 8B_{nk}^4\xi_k^4,$$

Since $A_{nk} = a_{n,k,k}/\sqrt{v_n}$ is non-random, B_{nk} and ξ_k are independent and $\mathrm{E}\xi_k^8 < \infty$, it suffices for (2.23) to show

$$\sum_k A_{nk}^4 \quad \to 0, \tag{2.24}$$

$$\sum_k \mathrm{E}\{B_{nk}^4\} \quad \to 0. \tag{2.25}$$

For (2.24), we have

$$\begin{aligned}
\sum_k A_{nk}^4 &= \frac{1}{v_n^2} \sum_k a_{n,k,k}^4 \\
&\leq \frac{1}{v_n} \sum_k a_{n,k,k}^2 \times \frac{1}{v_n} \max_k a_{n,k,k}^2 \\
&\leq \frac{1}{v_n} \max_k a_{n,k,k}^2 \\
&\to 0
\end{aligned}$$

by (A1). For (2.25), since $B_{nk} \sim \mathcal{N}\,[0, (\sum_{i:\ i<k} a_{n,i,k}^2)/v_n]$, we have

$$\begin{aligned}
\sum_k \mathrm{E}\{B_{nk}^4\} &= \sum_k \frac{3}{v_n^2} \left(\sum_{i:\ i<k} a_{n,i,k}^2 \right)^2 \\
&\leq \sum_k \frac{3}{v_n} \left(\sum_{i:\ i<k} a_{n,i,k}^2 \right) \times \frac{1}{v_n} \max_k \left(\sum_{i:\ i<k} a_{n,i,k}^2 \right) \\
&\leq 3 \max_k \frac{1}{v_n} \left(\sum_{i:\ i<k} a_{n,i,k}^2 \right) \\
&\to 0 \quad ,
\end{aligned}$$

by (A2). Hence both conditions (1) and (2) of the Martingale Central Limit Theorem have been verified, and Theorem 2.4 follows.

Remark 1. The method does not critically rely on the Gaussian assumption and would extend to other distributions with suitable moment restrictions.

Remark 2. There are other methods of proof for the result stated before, for Gaussian random variables, but the martingale technique has been presented here because of its elegance and generality.

Remark 3. It is possible to restate the result the following way. Suppose $a_{n,i,j} = a_{n,j,i}$ for all n, i, j and define S_n by

$$S_n = \sum_i \sum_j a_{n,i,j} \xi_i \, \xi_j \ . \tag{2.26}$$

In other words, in contrast to (2.12), the sum is over all pairs (i, j) and not just $i \leq j$. Define m_n again by (2.13), and v_n by

$$v_n = 2 \sum_i \sum_j a^2_{n,i,j}. \tag{2.27}$$

Also, combine conditions (A1) and (A2) of Theorem 2.4 into a single condition:

(A3) $\max_k \sum_i a^2_{n,i,k}/v_n \to 0$.

Then Theorem 2.4 holds under these revised conditions.
This is just a restatement of Theorem 2.4, since we rewrite (2.26) as

$$S_n = \sum_i a_{n,i,j} \, \xi_i^2 \ + 2 \sum_{i<j} a_{n,i,j} \, \xi_i \, \xi_j$$

which is of the form (2.12). Therefore, (2.15) becomes

$$v_n = 2 \sum_i a^2_{n,i,i} + 4 \sum_{\{i,j:\ i<j\}} a^2_{n,i,j} = 2 \sum_{i,j} a^2_{n,i,j}$$

and conditions (A1) and (A2) follow from (A3); therefore, under the conditions of *Remark 3*, Theorem 2.4 applies to show $\frac{S_n - m_n}{\sqrt{v_n}} \xrightarrow{d} \mathcal{N}[0, 1]$ as before.

One attractive feature of this revised form of Theorem 2.4 is that it is no longer dependent on any ordering of the indices. In this setting there is no need for the indices to represent positive integers — they could be an arbitrary countable set. This is relevant to later work, because for the random field context, we are interested in cases where the index set is a subset of $\Re^d,\ d > 1$.

2.4 Autocovariance functions for ARMA(1,1) and AR(1) time series

Throughout the following discussion, we often refer to the structure and properties of AR(1) and ARMA(1,1) time series. Therefore, we give here a brief description of these two processes and their covariance structures. Detailed results and derivations can be found, for example, in Brockwell and Davis (1991).

The process $\{X_t,\ t = 0, \pm 1, \pm 2, \dots\}$ which can be represented as

$$X_t = \phi X_{t-1} + \epsilon_t$$

where the coefficient satisfies $|\phi| < 1$ and $\{\epsilon_t\}$ are independent random errors such that $\epsilon_t \sim \mathcal{N}[0, \sigma_\epsilon^2]$ is said to be the *autoregressive process* of order 1, AR(1). We call ϕ the AR(1) coefficient. From standard time series theory (see Wei (1989)), the autocovariance structure has the form:

$$\gamma_m = \begin{cases} \left[\frac{1}{1-\phi^2}\right] \sigma_\epsilon^2 & \text{if } m = 0 \\ \\ \phi^m \gamma_0 & \text{if } m \geq 1 \ . \end{cases}$$

Similarly, a process $\{X_t,\ t = 0, \pm 1, \pm 2, \dots\}$ which has the following property:

$$X_t = \theta_1 \, X_{t-1} + \epsilon_t - \theta_2 \, \epsilon_{t-1}$$

where both coefficients satisfy $|\theta_1| < 1,\ |\theta_2| < 1$ and $\{\epsilon_t\}$ are independent random errors such that $\epsilon_t \sim \mathcal{N}[0, \sigma_\epsilon^2]$ is said to be the *mixed autoregressive moving average process* of order 1, ARMA(1, 1). We call θ_1 the AR coefficient and θ_2 the MA coefficient.

Following routine time series calculations one obtains the autocovariance structure of an ARMA(1, 1) process (for a complete derivation see, for example, Wei (1989)):

$$\gamma_m = \begin{cases} \left[1 + \frac{(\theta_2+\theta_1)^2}{1-\theta_1^2}\right]\sigma_\epsilon^2 & \text{if } m = 0 \\ \left[(\theta_1 + \theta_2) + \frac{(\theta_2+\theta_1)^2\theta_1}{1-\theta_1^2}\right]\sigma_\epsilon^2 & \text{if } m = 1 \\ \theta_1^{m-1}\gamma_1 & \text{if } m \geq 2. \end{cases} \tag{2.28}$$

Chapter III

Expansion Technique for AR(1) Time Series

One of the practical challenges of the maximum likelihood estimation, in spite of its merits, is the computational consequences of the high dimensionality of many data sets. While maximizing the log-likelihood function, one needs to invert the covariance matrix, and calculate its determinant. This is often intractable for large samples, and this problem manifests itself rather rapidly in spatial settings. To avoid this problem, we propose here a series of approximations of the likelihood function, which we use as estimating functions. The idea that links these approximations is grouping the observations into blocks of approximately same size. All these methods significantly reduce the dimensionality of the original problem, but they do not always lead to reasonable estimators. As theoretical calculations become very involved in the general spatial setting, we begin by analyzing these block methods for time series problems. This will give us some insight into how these approximation methods would apply for spatial processes. As an example, we are going to examine in detail the special case of a first order autoregressive process.

Maximum likelihood methods of estimating the autoregressive parameter of an AR(1) time series achieve asymptotic efficiency under suitable regularity conditions when the size of the sample tends to infinity. Chapter 3 of Akahira and Takeuchi (1981) discusses in detail the asymptotic efficiency in an autoregressive process. We compare the asymptotic performance of the estimators of the AR(1) parameter yielded by the alternative methods with that of the maximum likelihood estimator. We consider as measure of fit the asymptotic relative efficiency, defined as the ratio between the asymptotic variance of the MLE and the asymptotic variance of the alternative estimators.

In order to calculate the asymptotic variance of the proposed estimators, we use both the properties specific to an AR(1) process (which enable us to use an application of the Martingale Central Theorem to compute the variance of a quadratic form in normal random normal variables), as well as the "information sandwich" formula, which combines the expected value of the second derivative and the variance of the first derivative of the pseudo-likelihood function. Since a critical step in the analysis is to expand the process in terms of independent random variables, we refer to this method as the *expansion technique.*

The three methods proposed here vary from making strong independence assumptions to incorporating both variability between and among groups in the pseudo-likelihood function. The efficiency of the estimates vary from very poor to very good.

We begin by illustrating how our alternative approach works in the "classical" case, using the exact maximum likelihood function. This will allow us to go through all the theoretical derivations in a widely studied case. Thus we have the advantage of comparing our conjecture with the ones based on Fisher information approximations, for example. The other major advantage we have in this case is that we know the exact forms of the determinant and inverse covariance matrix, hence we are able to find the MLE estimators even for very large data sets, and compare them with the ones obtained through one of the alternative methods.

We continue with the method expected to yield the least efficient estimators. The fundamental idea of this approach is summarizing each block through its mean, constructing the likelihood of the means time series and maximizing this alternative function. We call this method "Big Blocks". The loss of information will in some cases be rather significant, this method being expected to lose efficiency with any increase in the block size.

The next method makes the assumption that the blocks are independent (but using the correct dependence structure within blocks) and the pseudo-likelihood function is calculated by multiplying the individual likelihoods for each block. Although the independence assumption is a rather strong one, the loss of efficiency is usually smaller than in the previous case. It is interesting to note that this technique becomes more efficient with the increase of block size. We call this method "Small Blocks".

The last method, called the "Hybrid", is a combination of the above proposed alternatives. The basic principle for constructing the estimating function is incorporating both correlation between blocks and among blocks. The assumption made in this case, although not necessarily verifiable in practice, is a reasonable working assumption: the blocks are independent given the block means. Hence, in constructing the pseudo-likelihood here we multiply the conditional likelihood functions of each block given their means with the likelihood of the means. This method is expected to yield the most efficient estimator among the three approximations considered, comparable with the exact maximum likelihood estimator.

3.1 Classical Maximum Likelihood Estimator

Consider an AR(1) process, which is represented as $x_{i+1} = \phi x_i + \epsilon_{i+1}$, where $|\phi| < 1$ and $\epsilon_i \sim \mathcal{N}[0, \sigma_\epsilon^2]$ is independent of x_{i+1}. Defining $\sigma_x^2 = \sigma_\epsilon^2/(1-\phi^2)$, this also has the representation

$$x_i = \sigma_\epsilon \sum_{r=-\infty}^{i} \phi^{i-r}\xi_r = \sigma_x\sqrt{1-\phi^2} \sum_{r=-\infty}^{i} \phi^{i-r}\xi_r, \tag{3.1}$$

with ξ_r independent $\mathcal{N}[0,1]$.

Cochrane and Orcutt (1949) write U, the covariance matrix of $X = (x_1, ..., x_n)$ as:

$$U \; = \frac{\sigma_\epsilon^2}{1-\phi^2} \begin{pmatrix} 1 & \phi & \phi^2 & \dots & \dots & \phi^{n-1} \\ \phi & 1 & \phi & \dots & \dots & \phi^{n-2} \\ \phi^2 & \phi & 1 & \dots & \dots & \phi^{n-3} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \phi^{n-1} & \phi^{n-2} & \phi^{n-3} & \dots & \dots & 1 \end{pmatrix}, \tag{3.2}$$

and

$$U^{-1} \quad = \tfrac{1}{\sigma_\epsilon^2} \begin{pmatrix} 1 & -\phi & 0 & \dots & \dots & 0 \\ -\phi & 1+\phi^2 & -\phi & \dots & \dots & 0 \\ 0 & -\phi & 1+\phi^2 & \dots & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1+\phi^2 & -\phi \\ 0 & 0 & 0 & \dots & -\phi & 1 \end{pmatrix} . \tag{3.3}$$

It can also be shown that

$$|U| = \frac{\sigma_\epsilon^{2n}}{1-\phi^2} .$$

For simplicity of subsequent calculations, assume σ_ϵ^2 is known (this assumption will have little bearing on the final result). The negative log-likelihood function has therefore the following form:

$$\ell(\phi) = \frac{1}{2}\left(X^T U^{-1} X - \log|U^{-1}|\right)$$

and thus the MLE for ϕ minimizes this function. In the standard fashion, we compute the first derivative of the negative log likelihood function and the MLE is the solution for the equation $\partial_\phi \ell(\phi) = 0$ where, modulo fixed constants,

$$\begin{aligned} \partial_\phi \ell(\phi) &= \frac{\phi}{1-\phi^2} - \frac{1}{\sigma_\epsilon^2}\left[\sum_{i=1}^{n-1} x_i\, x_{i+1} - \phi \sum_{i=2}^{n-1} x_i^2\right] \\ &= \frac{\phi}{1-\phi^2} - \frac{1}{\sigma_\epsilon^2}\left[\sum_{i=1}^{n-1} x_i\,(x_{i+1} - \phi\, x_i) + \phi\, x_1^2\right] . \end{aligned}$$

Note immediately that $E\left[\partial_\phi \ell(\phi)\right] = 0$ and we are interested in calculating its variance, $\mathrm{Var}\left[\partial_\phi \ell(\phi)\right]$. For simplicity of notation, let us denote

$$S_n = \sum_{i=1}^{n-1} x_i\,(x_{i+1} - \phi\, x_i) + \phi\, x_1^2$$

and thus

$$\mathrm{Var}\left[\partial_\phi \ell(\phi)\right] = \frac{1}{\sigma_\epsilon^4} \mathrm{Var}[S_n] \ .$$

Using (3.1), we rewrite S_n as

$$S_n = \sigma_\epsilon^2 \left(\sum_{i=2}^{n} \sum_{j=-\infty}^{i-1} \phi^{i-r-1} \xi_i \xi_r + 2 \sum_{i=-\infty}^{1} \sum_{r=-\infty}^{i-1} \phi^{3-i-r} \xi_i \xi_r + \sum_{i=-\infty}^{1} \phi^{3-2i} \xi_i^2 \right) .$$

This is of the form (2.12), with

$$a_{n,i,r} = \begin{cases} \sigma_\epsilon^2 \phi^{3-2i} \ , & \text{if } r = i \leq 1 \\ 2\ \sigma_\epsilon^2 \phi^{3-i-r} \ , & \text{if } i \leq 1 \text{ and } r \leq i-1 \\ \sigma_\epsilon^2 \phi^{i-r-1} \ , & \text{if } 2 \leq i \leq n-2 \text{ and } r \leq i-1 \\ 0 \ , & \text{elsewhere.} \end{cases} \tag{3.4}$$

Then

$$m_n = E[S_n] = \frac{\phi}{1-\phi^2}\ \sigma_\epsilon^2$$

(which confirms that the expected value of the first derivative of the negative log-likelihood is 0). Furthermore,

$$\begin{aligned} v_n &= \mathrm{Var}[S_n] = 2 \sum_i a_{n,i,i}^2 + \sum_{i<j} a_{n,i,j}^2 \\ &= \sigma_\epsilon^4 \left[2 \sum_{i=-\infty}^{1} \phi^{2(3-2i)} + \sum_{i=2}^{n} \sum_{r=-\infty}^{i-1} \phi^{2(i-r-1)} + \sum_{i=-\infty}^{1} \sum_{r=-\infty}^{i-1} 4\ \phi^{2(3-i-r)} \right] \\ &= \frac{n-1-(n-3)\phi^2}{(1-\phi^2)^2}\ \sigma_\epsilon^4 \ . \end{aligned}$$

Hence

$$\mathrm{Var}\left[\partial_\phi \ell(\phi)\right] = \frac{n-1-(n-3)\phi^2}{(1-\phi^2)^2} \ .$$

Note here that for any fixed r, we have that $\sum_{i<r} a_{n,i,r}^2$ is bounded by a constant, therefore the condition (A2) in Theorem 2.4 is satisfied, and so is (A1). Therefore we can deduce that

$$\partial_\phi \ell(\phi) \to \mathcal{N}(0, v_n) \ .$$

To complete the calculations of the asymptotic variance of the ML estimator, we follow the "information sandwich" technique for which we need to calculate the mean of the second derivative of the negative log-likelihood function:

$$\partial_\phi^2 \ell(\phi) = \frac{1+\phi^2}{(1-\phi^2)^2} + \frac{1}{\sigma_\epsilon^2} \sum_{i=2}^{n-1} x_i^2 \ .$$

Note first that the second derivative is always positive, indicating that the solution of the equation $\partial_\phi \ell(\phi) = 0$ will indeed minimize the negative log-likelihood. Proceeding further, note that

$$E[\partial_\phi^2 \ell(\phi)] = \frac{1+\phi^2}{(1-\phi^2)^2} + \frac{1}{\sigma_\epsilon^2} \sum_{i=2}^{n-1} E[x_i^2] = \frac{(n-1)-(n-3)\phi^2}{(1-\phi^2)^2} \ .$$

This confirms that when using the likelihood function as the estimating criterion, the expected value of the second derivative equals the variance of the first derivative of the likelihood. Further we apply the "information sandwich" formula to obtain the asymptotic variance of the MLE estimator:

$$\mathrm{Var}[\hat{\phi}] \overset{a}{\cong} \frac{\mathrm{Var}[\partial_\phi \ell(\phi)]}{E^2[\partial_\phi^2 \ell(\phi)]} = \frac{(1-\phi^2)^2}{n-1-\phi^2(n-3)} \ , \tag{3.5}$$

which agrees with the asymptotic variance derived by Brockwell and Davis (1991), pages 258—259.

3.2 Analysis of Means Time Series: Big Blocks Method

Consider the first alternative method of estimation. Let us assume we can divide the entire time series in b blocks, of length k. The next step is to compute the mean of each block and let us denote by $\{X_m^*\}$ the time series consisting of these means.

In other words,

$$X_m^* = \frac{1}{k}\sum_{j=1}^{k} X_{(m-1)k+j} \quad \text{where } \ 1 \le m \le b. \tag{3.6}$$

Therefore, the covariance structure of this new time series is characterized by

$$\gamma_{m-1}^* = \text{Cov}[X_1^*, X_m^*] = \frac{1}{k^2}\sum_{i=1}^{k}\sum_{j=1}^{k}\gamma_{(m-1)k+j-i}. \tag{3.7}$$

Routine algebra manipulations show that

$$\gamma_0^* = \text{Cov}[X_1^*, X_1^*] = \left[\frac{2\phi^{k+1} - 2\phi - k\phi^2 + k}{k^2(1-\phi)^2(1-\phi^2)}\right]\sigma_\epsilon^2, \tag{3.8}$$

$$\tag{3.9}$$

$$\gamma_1^* = \text{Cov}[X_i^*, X_{i+1}^*] = \left[\frac{\phi(1-\phi^k)^2}{k^2\,(1-\phi)^2\,(1-\phi^2)}\right]\sigma_\epsilon^2 \tag{3.10}$$

and that

$$\gamma_m^* = \phi^k\gamma_{m-1}^* \quad \text{for } m \ge 2. \tag{3.11}$$

Thus, following (3.8), (3.10) and (3.11) we conclude that $\{X_m^*\}$ has the following covariance structure:

$$\gamma_m^* = \begin{cases} \left[\frac{2\phi^{k+1}-2\phi-k\phi^2+k}{k^2\,(1-\phi)^2\,(1-\phi^2)}\right]\sigma_\epsilon^2\ , & \text{if m=0} \\ \left[\frac{\phi(1-\phi^k)^2}{k^2\,(1-\phi)^2\,(1-\phi^2)}\right]\sigma_\epsilon^2\ , & \text{if m = 1} \\ (\phi^k)^{m-1}\gamma_1^*, & \text{if m} \ge 2. \end{cases}$$

According to equation (2.28), this covariance structure corresponds to an ARMA(1,1) process.

The next step is to compute the likelihood function. Since the goal here is to find the value of ϕ which maximizes the likelihood function, one could first identify the ARMA(1,1) coefficients as functions of the original AR(1) parameter ϕ and then use already established results (Brockwell and Davis, 1991, page 258) for the asymptotic covariance matrix for the ARMA(1,1) parameters to derive the variance of the Big Blocks estimator. However, the identification of the autoregressive coefficient is very involved. Another weakness of this approach is that since it is so specific, it would not be of any help in the development of Small Blocks and Hybrid methods. Therefore, we illustrate here how the expansion method is applied.

First we calculate the likelihood function for the means time series, using the covariance structure derived in equation (3.12). Since we have recognized the means time series to be an ARMA(1,1) process, we expect the Big Blocks estimator to be unbiased. We compute the variance of the estimator using the "information sandwich" technique.

Recall the definition

$$X^* = \{X_1^*, X_2^*, \ldots, X_b^*\} \ ,$$

where

$$X_j^* = \frac{x_{k(j-1)+1} + x_{k(j-1)+2} + \ldots + x_{k(j-1)+k}}{j} \ .$$

Therefore the likelihood function is given by

$$L_{means} = \frac{1}{(2\pi)^{b/2}} \frac{1}{|V_{means}|^{1/2}} \exp\left\{ -\frac{1}{2} X^{*T} V_{means}^{-1} X^* \right\} \ .$$

Define

$$V = \frac{\partial}{\partial \phi} V_{means}^{-1} = (v_{ij})_{1 \,\leq\, i,\, j\, \leq b}$$

and assume that σ_ϵ^2 is known (this assumption will have little bearing on the final result and it considerably simplifies the computations). Then the first derivative of the negative log-likelihood function, modulo fixed constants, is given by

$$\begin{aligned} \partial_\phi \ell(\phi) &= \frac{1}{2} \left[X^{*T} V \, X^* - \frac{\partial_\phi |V_{means}|}{|V_{means}|} \right] \\ &= \frac{1}{2} \left[\frac{1}{k^2} \sum_{i=1}^{b} \sum_{j=1}^{b} v_{ij} \sum_{\ell=1}^{k} \sum_{m=1}^{k} x_{(i-1)k+\ell} \, x_{(j-1)k+m} - \frac{\partial_\phi |V_{means}|}{|V_{means}|} \right] \ . \quad (3.12) \end{aligned}$$

Again, there is no apparent closed form solution for the equation $\partial_\phi \ell(\phi) = 0$. We compute the variance of the first derivative of the negative log-likelihood function and the expected value of the second derivative. As before, in order to compute the aforementioned quantities, we rewrite the expression in (3.12) using the representation of x_i as an AR(1) process. Thus

$$x_i = \sigma_\epsilon \sum_{r=-\infty}^{i} \phi^{i-r} \xi_i$$

and it follows that

$$x_{(i-1)k+\ell} x_{(j-1)k+m} = \sigma_\epsilon^2 \sum_{r=-\infty}^{(i-1)k+\ell} \sum_{s=-\infty}^{(j-1)k+m} \phi^{(i+j-2)k+\ell+m-r-s} \xi_r \, \xi_s \; .$$

Define

$$S_n = \frac{1}{k^2} \sum_{i=1}^{b} \sum_{j=1}^{b} v_{ij} \sum_{\ell=1}^{k} \sum_{m=1}^{k} x_{(i-1)k+\ell} \, x_{(j-1)k+m}$$

and thus rewrite it as

$$S_n = \frac{\sigma_\epsilon^2}{k^2} \sum_{i=1}^{b} \sum_{j=1}^{b} \sum_{\ell=1}^{k} \sum_{m=1}^{k} \sum_{r=-\infty}^{(i-1)k+\ell} \sum_{s=-\infty}^{(j-1)k+m} v_{ij} \phi^{(i+j-2)k+\ell+m-r-s} \xi_r \, \xi_s \; .$$

This is equivalent to

$$S_n = \sum_r a_{rr}^{(1)} \, \xi_r^2 + 2 \sum_{r<s} a_{rs}^{(1)} \, \xi_r \, \xi_s$$

where

$$a_{rs}^{(1)} = \begin{cases} \frac{\sigma_\epsilon^2}{k^2} \sum_{i=1}^{b} \sum_{j=1}^{b} \sum_{\ell=1}^{k} \sum_{m=1}^{k} v_{ij}\, \phi^{(i+j-2)k+\ell+m-2r} \ , & \text{if } s = r \leq 1, \\ \\ \frac{\sigma_\epsilon^2}{k^2} \sum_{i=r_1+1}^{b} \sum_{j=r_1+1}^{b} \sum_{\ell=1}^{k} \sum_{m=1}^{k} v_{ij}\, \phi^{(i+j-2)k+\ell+m-2r} + & \\ \frac{\sigma_\epsilon^2}{k^2} \sum_{\ell=r_2}^{k} \sum_{m=r_2}^{k} v_{r_1,r_1}\, \phi^{(2r_1-2)k+\ell+m-2r} + & \\ 2\,\frac{\sigma_\epsilon^2}{k^2} \sum_{i=r_1+1}^{b} \sum_{\ell=1}^{k} \sum_{m=r_2}^{k} \ v_{i\, r_1}\, \phi^{(i+r_1-2)k+\ell+m-2r} \ , & \text{if } 2 \leq r = s, \\ \\ \frac{\sigma_\epsilon^2}{k^2} \sum_{i=1}^{b} \sum_{j=1}^{b} \sum_{\ell=1}^{k} \sum_{m=1}^{k} v_{ij}\, \phi^{(i+j-2)k+\ell+m-r-s} \ , & \text{if } r < s \leq 1 \ , \\ \\ \frac{\sigma_\epsilon^2}{k^2} \sum_{i=1}^{b} \sum_{j=s_1+1}^{b} \sum_{\ell=1}^{k} \sum_{m=1}^{k} v_{ij}\, \phi^{(i+j-2)k+\ell+m-r-s} + & \\ \frac{\sigma_\epsilon^2}{k^2} \sum_{i=1}^{b} \sum_{\ell=1}^{k} \sum_{m=s_2}^{k} v_{i\, s_1}\, \phi^{(i+s_1-2)k+\ell+m-r-s} \ , & \text{if } r \leq 1 < s \ , \\ \\ \frac{\sigma_\epsilon^2}{k^2} \sum_{i=r_1+1}^{b} \sum_{j=s_1+1}^{b} \sum_{\ell=1}^{k} \sum_{m=1}^{k} v_{ij}\, \phi^{(i+j-2)k+\ell+m-r-s} + & \\ \frac{\sigma_\epsilon^2}{k^2} \sum_{\ell=r_2}^{k} \sum_{m=s_2}^{k} v_{r_1,s_1}\, \phi^{(r_1+s_1-2)k+\ell+m-r-s} + & \\ \frac{\sigma_\epsilon^2}{k^2} \sum_{i=r_1+1}^{b} \sum_{\ell=1}^{k} \sum_{m=s_2}^{k} v_{i\, s_1}\, \phi^{(i+s_1-2)k+\ell+m-r-s} + & \\ \frac{\sigma_\epsilon^2}{k^2} \sum_{j=s_1+1}^{b} \sum_{\ell=r_2}^{k} \sum_{m=1}^{k} v_{r_1\,,j}\, \phi^{(r_1+j-2)k+\ell+m-r-s} \ , & \text{if } 2 \leq r < s \ . \end{cases} \tag{3.13}$$

Notations: $r_1 = \lceil r/k \rceil$, $s_1 = \lceil s/k \rceil$ *and* $r_2 = r - (r_1 - 1)\ k$, $s_2 = s - (s_1 - 1)\ k$. [1]

Next we apply the Martingale Central Limit Theorem to S_n which is a quadratic form of independent normal random variables and we obtain

$$m_n = E[S_n] = \sum_r a_{rr}^{(1)} = 0$$

[1] *By* $\lceil x \rceil$ *we mean the smallest integer greater than or equal to* x, *and by* $\lfloor x \rfloor$ *we mean the largest integer smaller than or equal to* x.

and

$$v_n = \mathrm{Var}[S_n] = 2 \sum_r {a_{rr}^{(1)}}^2 + \sum_{r,s:\, r<s} 4\, {a^{(1)}}^2_{rs} \qquad (3.14)$$

and the variance for the first derivative of the log-likelihood function follows from (3.12).

To compute the expected value for the second derivative of the pseudo-likelihood function, note from (3.12) that

$$\begin{aligned} \partial^2_\phi\, \ell(\phi) &= \frac{1}{2} \left[\frac{1}{k^2} \sum_{i=1}^{b} \sum_{j=1}^{b} v'_{ij} \sum_{\ell=1}^{k} \sum_{m=1}^{k} x_{(i-1)k+\ell}\, x_{(j-1)k+m} \right. \\ &\quad - \left. \frac{(\partial^2_\phi\, |V_{means}|)\, |V_{means}| - (\partial_\phi |V_{means}|)^2}{|V_{means}|^2} \right] . \qquad (3.15) \end{aligned}$$

Hence to calculate the expected value of the above function, one has two alternatives: either expand it as a quadratic form of independent normal random variables, identify the coefficients and apply the Central Limit Theorem for quadratic forms, or take advantage of the underlying AR(1) correlation structure and compute it as:

$$\begin{aligned} E[\partial^2_\phi\, \ell(\phi)] &= \frac{1}{2} \left[\frac{1}{k^2} \sum_{i=1}^{b} \sum_{j=1}^{b} \sum_{\ell=1}^{k} \sum_{m=1}^{k} v'_{ij} \gamma_{|(i-j)\, k+\ell-m|} \right. \\ &\quad - \left. \frac{(\partial^2_\phi\, |V_{means}|)\, |V_{means}| - (\partial_\phi |V_{means}|)^2}{|V_{means}|^2} \right] \qquad (3.16) \end{aligned}$$

keeping in mind that

$$\gamma_i = \frac{\phi^i}{1-\phi^2}\, \sigma^2_\epsilon \ .$$

To conclude the calculations, we apply the “information sandwich” formula and obtain the asymptotic variance of the estimator $\hat{\phi}_1$ as

$$\mathrm{Var}[\phi_1] = \frac{v_n}{[E[\partial^2_\phi\, \ell(\phi)]]^2} \ ,$$

ϕ	b=5 k=100		b=10 k=50		b=50 k=10	
	Theory	Sim	Theory	Sim	Theory	Sim
-0.750	0.00214	0.002	0.00330	0.003	0.00549	0.005
-0.250	0.01166	0.013	0.02265	0.020	0.08982	0.080
-0.010	0.01925	0.018	0.03773	0.036	0.15929	0.158
0.010	0.02003	0.019	0.03929	0.039	0.16702	0.165
0.250	0.03280	0.032	0.06434	0.055	0.27280	0.269
0.750	0.13367	0.132	0.25465	0.255	0.73897	0.724

Table 3.1: Time Series: Big Blocks Asymptotic Relative Efficiency

where v_n is given by equation (3.14) and $E[\partial^2_\phi \ell(\phi)]$ by equation (3.16). As a measure of performance for the estimator $\hat{\phi}_1$, we compute the relative efficiency with respect to the classical maximum likelihood estimator for an AR(1) process,

$$e_1(\hat{\phi}, \hat{\phi}_1) = \frac{\text{Var}[\hat{\phi}]}{\text{Var}[\hat{\phi}_1]}$$

where $\text{Var}[\hat{\phi}]$ is the asymptotic variance of the MLE.

One issue here is to calculate the inverse covariance matrix. Due to the complex nature of this matrix, it would be natural to compute its inverse using numerical algorithms like the Cholesky decomposition. However, there is one property of this matrix that leads us to computing its inverse analytically, and this is the fact that it is a Toeplitz matrix (this is obvious, since it is the covariance matrix of an ARMA(1,1) process). Trench (1964) proposes an algorithm that enables us to compute analytically the inverse of an $n \times n$ Toeplitz matrix. His algorithm presents another appealing feature, in the sense that it requires an order $O(n^2)$ calculations, compared to $O(n^3)$ required by the Cholesky decomposition.

For a Toeplitz matrix of the form

$$T = \begin{pmatrix} \theta_0 & \theta_1 & \theta_2 & \dots & \theta_n \\ \theta_1 & \theta_0 & \theta_1 & \dots & \theta_{n-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \theta_n & \theta_{n-1} & \theta_{n-2} & \dots & \theta_0 \end{pmatrix} \tag{3.17}$$

denote by $B = T^{-1}$. Then we have:

- $\psi_{0,0} = \phi_1$; $\Delta_0 = 1$
- $\Delta_m = (1 - \psi^2_{m-1,m-1})\Delta_{m-1}$
- $\psi_{m,m} = \frac{\phi_{m+1} - \sum_{s=0}^{m-1} \psi_{s,m-1}\phi_{m-s}}{\Delta_m}$
- $\psi_{r,m} = \psi_{r,m-1} - \psi_{m,m}\psi_{m-r-1,m-1}$, for $0 \leq r \leq m-1$

The last three formulas are used for $1 \leq m \leq n-1$. To obtain B, compute as follows:

- $b_{00} = \frac{1}{\Delta_n}$
- $b_{r,0} = -\frac{\psi_{r-1,n-1}}{\Delta_n}$, for $1 \leq r \leq n$
- $b_{rs} = b_{r-1,s-1} + \frac{\psi_{r-1,n-1}\psi_{s-1,n-1} - \psi n-r,n-1\psi n-s,n-1}{\Delta_n}$

and further exploit the symmetry about the principal diagonal and secondary diagonal to complete the calculations for the inverse Toeplitz matrix. However, although we can obtain the exact formula for the inverse, we have to calculate the first and second derivatives of the inverse matrix using numerical procedures.

Given the intractable analytical structure of the relative efficiency, we analyze its values numerically for a few particular cases. We proceed in the following manner. We first compute matrix V numerically, then use its elements to evaluate each of the coefficients a_{rs}. Each coefficient consists of a finite sum, thus its evaluation is routine algebra. The next step is to calculate the sum over these coefficients. Note first that the indices r and s have b as an upper bound. Then, taking a closer look at the structure of the coefficients, we distinguish two cases. For all $r,\ s \geq 2$ we need to evaluate a finite sum, therefore this case comes down to a standard finite summation. In the other case, when at least r or s are less than or equal to 1, we take advantage of the fact that we can separate the sums containing r and s from the other sums, and simply compute these infinite sums (over r and s) as geometric series. We conclude by combining all the above sums to obtain the final result.

As a side comment, note that for the theoretical calculations the bias for both the MLE and Big Blocks is 0, therefore the ratio between their mean squared errors is identical to the ratio between their variances.

Table 3.1 presents values of $e_1(\hat{\phi}, \hat{\phi}_1)$ for different ϕ and k, the number of observations per block (the columns labeled as "Theory").

Parallel to the theoretical calculations we also perform a simulation study to analyze the asymptotic relative efficiency for the Big Blocks estimator. In order to do this, we simulate an AR(1) time series of length n, which we divide in b blocks of equal length, say k. We start by computing the likelihood function for the entire time series, maximize it and obtain the classical ML estimator. Next step is to calculate the mean value for each block, as well as the correlations between the block means. Thus we obtain the means likelihood, as a function of the autoregressive parameter ϕ. We maximize this pseudo-likelihood through an iterative numerical procedure, and obtain an estimator for ϕ. We replicate this process 1000 times, for each choice of b and ϕ and conclude by computing the means and variances for each of the two vectors of estimators, MLE and Big Blocks. The measure of efficiency we are interested in is the ratio between the mean squared errors of the two estimators. We report these results in Table 3.1 under the column labeled "Sim".

We note from this analysis that the Big Blocks estimator is not very efficient compared to the maximum likelihood estimator. It seems to be more efficient only for the cases where the number of blocks is large in which case the method is not attractive from the computational point of view. Table 3.1 also provides a verification of the validity of the theoretical results, by comparing theoretically obtained values with their analogous results obtained through simulations.

3.3 Small Blocks Method

As seen in the previous section, the "Big Blocks" estimator, although appealing for its simplicity and considerable dimension reduction, tends to be very inefficient for even moderate block sizes. This caveat makes it unfit for realistic problems. Therefore we need to alter the way we compute the minimizing criterion, and take into account more adequately the underlying correlation structure. We therefore define the second alternative estimator, which we call "Small Blocks". In this second approach, we ignore the correlation between blocks but take into account the true dependence structure within blocks. Therefore, the assumption under which we develop this scheme is that blocks are independent. We construct the pseudo-likelihood in this case as the product between the b individual block likelihoods, each block containing k sample points.

We shall proceed in two steps. The first stage is to compute the likelihood for each block. In the second stage we construct the maximization criterion by multiplying the individual block-likelihoods.

Consider the j^{th} block, which we denote by $X_k^j = (x_{(j-1)k+1}, \ldots, x_{jk})$. Since the complete process is an AR(1) time series, the covariance matrix for this block is given by:

$$U_k^j = \frac{\sigma_\epsilon^2}{1-\phi^2} \begin{pmatrix} 1 & \phi & \phi^2 & \ldots & \ldots & \phi^{k-1} \\ \phi & 1 & \phi & \ldots & \ldots & \phi^{k-2} \\ \phi^2 & \phi & 1 & \ldots & \ldots & \phi^{k-3} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \phi^{k-1} & \phi^{k-2} & \phi^{k-3} & \ldots & \ldots & 1 \end{pmatrix} \tag{3.18}$$

and hence

$$(U_k^j)^{-1} = \frac{1}{\sigma_\epsilon^2} \begin{pmatrix} 1 & -\phi & 0 & \ldots & \ldots & 0 \\ -\phi & 1+\phi^2 & -\phi & \ldots & \ldots & 0 \\ 0 & -\phi & 1+\phi^2 & \ldots & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 1+\phi^2 & -\phi \\ 0 & 0 & 0 & \ldots & -\phi & 1 \end{pmatrix} \tag{3.19}$$

and also

$$|U_k^j| = \frac{\sigma_\epsilon^{2k}}{1-\phi^2} \, .$$

We note immediately that the covariance matrix is the same for all blocks, hence we can drop the index j from now on. We are now in the position to calculate the likelihood function for each block, L_j :

$$L_j = \frac{1}{(2\pi)^{k/2} |U_k|^{1/2}} \exp\left(-\frac{1}{2} X_k^{j^T} {U_k}^{-1} X_k^j \right) \, . \tag{3.20}$$

If we define $\tilde{L} = \prod_{j=1}^{b} L_j$ then the pseudo-likelihood to be maximized with respect to ϕ is:

$$\tilde{L} = \frac{1}{(2\pi)^{kb/2} |U_k|^{b/2}} \exp\left(-\frac{1}{2} \sum_{j=1}^{b} X_k^{j^T} U_k{}^{-1} X_k^j \right) \ . \tag{3.21}$$

Assume here that σ_ϵ^2 is known, for simplicity of calculations. This assumption has little bearing on the final results. Then the MLE for ϕ minimizes

$$b \log \frac{1}{1-\phi^2} + \sum_{j=1}^{b} X_k^{j^T} U_k^{-1} X_k^{j^T} \ . \tag{3.22}$$

On the other hand, it easily follows from "block-calculations" that

$$X_k^{j^T} U_k^{-1} X_k^j = \left[(1+\phi^2) \sum_{i=2}^{k-1} x_{k(j-1)+i}^2 - 2\phi \sum_{i=1}^{k-1} x_{k(j-1)+i} x_{k(j-1)+i+1} \right] / \sigma_\epsilon^2$$

and hence the MLE for ϕ minimizes the following estimating function:

$$h(\phi) = -b \ \log\left(1-\phi^2\right) + \frac{1}{\sigma_\epsilon^2} \sum_{j=1}^{b} \left[(1+\phi^2) \sum_{i=2}^{k-1} x_{k(j-1)+i}^2 - 2\phi \sum_{i=1}^{k-1} x_{k(j-1)+i} x_{k(j-1)+i+1} \right] \ .$$

The derivative of the estimating function is of the form

$$h'(\phi) = b \frac{\phi}{1-\phi^2} + \frac{\phi S_1 - S_2}{\sigma_\epsilon^2}, \tag{3.23}$$

where

$$S_1 = \sum_{\substack{i=1 \\ i \notin \mathcal{M}_k, \mathcal{M}_k+1}}^{n} x_i^2, \quad S_2 = \sum_{\substack{i=1 \\ i \notin \mathcal{M}_k}}^{n-1} x_i x_{i+1} \tag{3.24}$$

(where $\mathcal{M}_k + j$ is the set of $m\,k + j$ for any integer m).

A local minimum of (3.22) must satisfy the equation

$$b \ \frac{\phi}{1-\phi^2} + \frac{\phi \ S_1 - S_2}{\sigma_\epsilon^2} = 0. \tag{3.25}$$

There is not an obvious closed form solution for(3.25), so we proceed asymptotically. Let us write ϕ_0 for the true value of ϕ. As $n \to \infty$, $b \to \infty$ we have

$$\frac{S_1}{b\ (k-2)\sigma_x^2} \xrightarrow{p} 1, \quad \frac{S_2}{b\ (k-1)\sigma_x^2} \xrightarrow{p} \phi_0.$$

Of course, both σ_x^2 and ϕ_0 are unknown at this point but the purpose of the calculation is to understand how the solution of (3.25) actually behaves in limiting cases. As a side calculation, we note from (3.23) that

$$h''(\phi) = \frac{b\ (1+\phi^2)}{(1-\phi^2)^2} + \frac{S_1}{\sigma_\epsilon^2},$$

and therefore in the limit,

$$\frac{1}{b}\ h''(\phi) \to \frac{\sigma_x^2}{\sigma_\epsilon^2}\ \{1+\phi_0^2+(k-2)\}\ = \frac{(\phi_0^2+k-1)}{1-\phi_0^2}$$

Note that the above expression is always positive, indicating that the solution of (3.25) is indeed a local minimum for the estimating function.
Returning to (3.25) and taking the expected value, we obtain

$$b\,\frac{\phi}{1-\phi^2} + b\,\sigma_x^2\frac{(k-2)\phi-(k-1)\phi_0}{\sigma_\epsilon^2} = 0\ .$$

Dividing throughout by $\frac{b}{\sigma_\epsilon^2}$, the equation to be solved becomes

$$\begin{aligned} 0 &= \frac{\phi}{1-\phi^2}(1-\phi_0^2) + (k-2)\,\phi - (k-1)\,\phi_0 \quad \text{or equivalently,} \\ 0 &= \frac{\phi_0-\phi}{1-\phi^2}\ \left[(k-2)\,\phi^2 - \phi\,\phi_0 - (k-1)\right] \qquad (3.26) \end{aligned}$$

If we denote by $h_k(\phi) = [(k-2)\,\phi^2 - \phi\,\phi_0 - (k-1)]$, it follows that

$$h_k(1) = -1-\phi_0 \quad \text{and} \quad h_k(-1) = \phi_0 - 1.$$

Therefore, since $h_k \to \infty$ as $\phi \to \pm\infty$, $h_k(1) < 0$, $h_k(-1) < 0$ and h_k is a quadratic function of ϕ, it follows that either both its roots are greater than 1 or both are smaller than -1.

To proceed, we assume

$$\frac{S_1}{b\sigma_x^2} = (k-2)(1+\epsilon_1), \quad \frac{S_2}{b\sigma_x^2} = (k-1)(\phi_0+\epsilon_2) \tag{3.27}$$

and look for a solution to the equation $g(\hat{\phi}_2, \epsilon) = 0$, where

$$g(\phi,\epsilon) = \frac{\phi}{(1-\phi^2)}\,(1-\phi_0^2) + \phi\,(k-2)\,(1+\epsilon_1) - (k-1)(\phi_0+\epsilon_2). \tag{3.28}$$

We know $g(\phi_0, 0) = 0$ and calculate

$$\frac{\partial g}{\partial \phi} = \frac{2\phi_0^2}{1-\phi^2} + k - 1, \quad \frac{\partial g}{\partial \epsilon_1} = \phi_0\,(k-2), \quad \frac{\partial g}{\partial \epsilon_2} = -(k-1) \tag{3.29}$$

where all the partial derivatives are evaluated at $\phi = \phi_0,\ \epsilon = 0$. Therefore, asymptotically, we have

$$\hat{\phi}_2 - \phi_0 \overset{p}{\sim} -\frac{\phi(k-2)\epsilon_1 - (k-1)\epsilon_2}{\frac{2\phi_0^2}{1-\phi^2} + k - 1}, \tag{3.30}$$

where $X \overset{p}{\sim} Y$ means that the ratio X/Y converges in probability to 1 under some suitable limiting operation (here, $n \to \infty$).

From now on, there is no need to distinguish between ϕ and ϕ_0, so we just write ϕ. From (3.30), we see that the asymptotic distribution of $\sqrt{b}(\hat{\phi}_2 - \phi)$ is the same as that of

$$-\frac{1}{\sqrt{b}\sigma_x^2\left(\frac{2\phi^2}{1-\phi^2} + k - 1\right)}\left\{\phi \sum_{\substack{i=1 \\ i\notin\mathcal{M}_k,\mathcal{M}_k+1}}^{n} x_i^2 - \sum_{\substack{i=1 \\ i\notin\mathcal{M}_k}}^{n-1} x_i x_{i+1} + \sigma_\epsilon^2\, b\, \frac{\phi}{1-\phi^2}\right\}. \tag{3.31}$$

Define

$$\begin{aligned} T(\phi) &= \sum_{\substack{i=1 \\ i\neq\mathcal{M}_k}}^{n} (\phi x_i^2 - x_i x_{i+1}) - \phi \sum_{\substack{i=1 \\ i=\mathcal{M}_k+1}}^{n} x_i^2 \\ &= \sum_{\substack{i=1 \\ i\neq\mathcal{M}_k}}^{n} (\phi x_i^2 - x_i x_{i+1}) - \phi \sum_{m=0}^{b-1} x_{mk+1}^2\ . \end{aligned}$$

Since the observations come from an AR(1) model, we can use the following representation:

$$x_i = \sigma_\epsilon \sum_{r=-\infty}^{i} \phi^{i-r} \xi_r$$

and hence

$$\phi \, x_i^2 = \sigma_\epsilon^2 \sum_{r=-\infty}^{i} \sum_{s=-\infty}^{i} \phi^{i-r+i-s+1} \xi_r \, \xi_s$$

and

$$x_i x_{i+1} = \sigma_\epsilon^2 \sum_{r=-\infty}^{i} \sum_{s=-\infty}^{i+1} \phi^{i-r+i-s+1} \xi_r \, \xi_s \, .$$

Also,

$$\phi \, x_i - x_{i+1} = -\sigma_\epsilon \, \xi_{i+1}$$

and therefore

$$\phi \, x_i^2 \quad - \quad x_i x_{i+1} = -\sigma_\epsilon^2 \sum_{r=-\infty}^{i} \phi^{i-r} \xi_r \, \xi_{i+1} \, .$$

Also, note that

$$\phi \, x_{mk+1}^2 = \sigma_\epsilon^2 \sum_{r=-\infty}^{mk+1} \sum_{s=-\infty}^{mk+1} \phi^{2mk+3-r-s} \xi_r \, \xi_s \, .$$

Thus

$$T(\phi) = -\sigma_\epsilon^2 \left(\sum_{\substack{i=1 \\ i \neq \mathcal{M}_k}}^{n-1} \sum_{r=-\infty}^{i} \phi^{i-r} \xi_r \, \xi_{i+1} + \sum_{m=0}^{b-1} \sum_{r=-\infty}^{mk+1} \sum_{s=-\infty}^{mk+1} \phi^{2mk+3-r-s} \xi_r \, \xi_s \right)$$

or

$$T(\phi) = -\sigma_\epsilon^2 \left(\sum_{\substack{s=2 \\ s \neq \mathcal{M}_k+1}}^{n} \sum_{r=-\infty}^{s-1} \phi^{s-r-1} \xi_r \, \xi_s + \sum_{m=0}^{b-1} \sum_{r=-\infty}^{mk+1} \sum_{s=-\infty}^{mk+1} \phi^{2mk+3-r-s} \xi_r \, \xi_s \right) . \qquad (3.32)$$

Our goal now is to express $T(\phi)$ as a quadratic form, in the sense that we rewrite it as

$$T(\phi) = \sum_{s=-\infty}^{n} \sum_{r=-\infty}^{s} a_{rs} \xi_r \, \xi_s$$

so that we can apply the quadratic forms Corollary of the Martingale Central Limit Theorem and compute its mean and variance.
Due to the specific form that $T(\phi)$ has in equation (3.32) above, we rewrite

$$T(\phi) = \sum_{s=-\infty}^{n} a''_{ss}\xi_s^2 + \sum_{s=-\infty}^{n}\sum_{r=-\infty}^{s-1}(a'_{rs} + 2a''_{rs})\xi_r\,\xi_s$$

where

$$a'_{rs} = \begin{cases} -\sigma_\epsilon^2\,\phi^{s-r-1}\ , & \text{if } r < s,\ 2 \le s \le n,\ s \neq pk+1,\ 1 \le p \le b-1, \\ 0\ , & \text{otherwise} \end{cases} \tag{3.33}$$

and

$$a''_{rs} = \begin{cases} -\sigma_\epsilon^2\,\phi^{3-2s}\frac{\phi^{2bk}-1}{\phi^{2k}-1}\ , & \text{if } s = r \le 1, \\ -\sigma_\epsilon^2\,\phi^{3-2s}\frac{\phi^{2bk}-\phi^{2mk}}{\phi^{2k}-1}\ , & \text{if } r = s,\ (m-1)\,k+2 \le r = s \le m\,k+1, \\ & \quad 1 \le m \le b-1, \\ -\sigma_\epsilon^2\,\phi^{3-r-s}\frac{\phi^{2bk}-1}{\phi^{2k}-1}\ , & \text{if } r < s \le 1, \\ -\sigma_\epsilon^2\,\phi^{3-r-s}\frac{\phi^{2bk}-\phi^{2mk}}{\phi^{2k}-1}\ , & \text{if } r \le 1,\ r < s, \\ & \quad (m-1)\,k+2 \le s \le m\,k+1,\ \ 1 \le m \le b-1, \\ -\sigma_\epsilon^2\,\phi^{3-r-s}\frac{\phi^{2bk}-\phi^{2mk}}{\phi^{2k}-1}\ , & \text{if } r < s,\ (p-1)\,k+2 \le r \le p\,k+1, \\ & \quad (m-1)\,k+2 \le s \le m\,k+1, \\ & \quad p \le m,\ 1 \le p \le b-1,\ 1 \le m \le b-1. \end{cases} \tag{3.34}$$

Hence the expected value can be computed as

$$m_n = \sum_{r=-\infty}^{n} a_{rr} = \sum_{r=-\infty}^{n} a''_{rr} = -\sigma_\epsilon^2\,\frac{b\phi}{1-\phi^2}$$

which confirms the result obtained earlier, and the variance as

$$\begin{aligned}
v_n &= 2\sum_{s=-\infty}^{n} a_{ss}^2 + \sum_{s=-\infty}^{n}\sum_{r=-\infty}^{s-1} a_{rs}^2 \\
&= 2\sum_{s=-\infty}^{n} {a''_{ss}}^2 + \sum_{s=-\infty}^{n}\sum_{r=-\infty}^{s-1} (a'_{rs} + 2a''_{rs})^2 \\
&= 2\sum_{s=-\infty}^{n} {a''_{ss}}^2 + \sum_{s=2}^{n}\sum_{r=-\infty}^{s-1} {a'_{rs}}^2 + 4\sum_{s=-\infty}^{n}\sum_{r=-\infty}^{s-1} {a''_{rs}}^2 + 4\sum_{s=-\infty}^{n}\sum_{r=-\infty}^{s-1} a'_{rs}a''_{rs} \\
&\equiv 2\,T_1(\phi) + T_2(\phi) + 4\,T_3(\phi) + 4\,T_4(\phi)
\end{aligned}$$

where

$$T_1(\phi) = \sum_{s=-\infty}^{n} {a''_{ss}}^2, \quad T_2(\phi) = \sum_{s=2}^{n}\sum_{r=-\infty}^{s-1} {a'_{rs}}^2,$$

$$T_3(\phi) = \sum_{s=-\infty}^{n}\sum_{r=-\infty}^{s-1} {a''_{rs}}^2, \; T_4(\phi) = \sum_{s=-\infty}^{n}\sum_{r=-\infty}^{s-1} a'_{rs}a''_{rs}.$$

Therefore, calculating each partial sum separately we obtain

$$\begin{aligned}
T_1(\phi) &= \sigma_\epsilon^4 \left[\frac{\phi^2(\phi^{2bk}-1)^2}{(\phi^{2k}-1)^2(1-\phi^4)}\right. \\
&- \left.\frac{\phi^{4k}-1}{(\phi^{2k}-1)^2(1-\phi^4)}\left(\frac{\phi^{4bk+2}-\phi^{4k+2}}{\phi^{4k}-1} - 2\frac{\phi^{2bk+2}-\phi^{2k+2}}{\phi^{2k}-1} + (b-1)\phi^2\right)\right],
\end{aligned}$$

$$T_2(\phi) = \sigma_\epsilon^4\,\frac{b\,(k-1)}{1-\phi^2}\,,$$

$$\begin{aligned}
T_3(\phi) &= \sigma_\epsilon^4 \left[\frac{\phi^4(\phi^{2bk}-1)^2}{(\phi^{2k}-1)^2(1-\phi^2)(1-\phi^4)}\right. \\
&- \left.\frac{\phi^4(\phi^{2k}+1)}{(\phi^{2k}-1)(1-\phi^2)(1-\phi^4)}\left(\frac{(\phi^{2bk}-\phi^{2k})(\phi^{2bk}-\phi^{2k}-2)}{\phi^{4k}-1} + b - 1\right)\right]
\end{aligned}$$

and

$$T_4(\phi) = \sigma_\epsilon^4\,\frac{\phi^2}{(1-\phi^2)(\phi^{2k}-1)}\left\{(k-1)(\phi^2-1)\frac{\phi^{2bk+2k}-\phi^{2k}}{\phi^{2k}-1}\right.$$

$$- \quad b(k-1)\phi^{2k}(\phi^2-1)\frac{1}{\phi^2-1}\left[\frac{(\phi^{2bk}-1)(\phi^{2k}-\phi^2)}{\phi^{2k}-1}+b(\phi^2-\phi^{2k})\right]\Bigg\} \; .$$

Hence

$$\begin{aligned} v_n \quad &= \quad \frac{\sigma_\epsilon^4}{(1-\phi^2)^2(1-\phi^{2\,k})^2}\big\{-4\phi^2(\phi^{2\,b\,k}-1)\big[-\phi^2+\phi^{2\,k}(1-\phi^2)^2(k-1)\big] \\ &+ \quad b(\phi^{2k}-1)\big[1-3\phi^2-4\phi^4-k+k\,\phi^2 \\ &+ \quad \phi^{2k}\big(-1+k+\phi^2\,(3\,k-1+4\,\phi^2(-2+\phi^2)(k-1))\big)\big]\big\}\,. \end{aligned}$$

Now recall from equation (3.31) that the asymptotic distribution of $\sqrt{b}(\hat{\phi}_2-\phi)$ is the same as that of

$$-\frac{1}{\sqrt{b}\sigma_x^2(\frac{2\phi^2}{1-\phi^2}+k-1)}\left\{T(\phi)+\sigma_\epsilon^2\frac{b\,\phi}{1-\phi^2}\right\}.$$

which according to Theorem 2.4 is asymptotically normal, with mean 0 and variance $\frac{v_n}{b\sigma_x^4\left(\frac{2\phi^2}{1-\phi^2}+k-1\right)^2}$. Thus the asymptotic variance of the Small Blocks estimator is given by

$$\mathrm{Var}[\hat{\phi}_2] = \frac{v_n\,(1-\phi^2)^2}{b^2\;\sigma_\epsilon^4\;\left(\frac{2\phi^2}{1-\phi^2}+k-1\right)^2}$$

(recall here that $n = b\,k$, in other words b is a function of n as well, so maybe it would be appropriate to think of it as b_n).

Therefore the asymptotic relative efficiency of this estimator is given by the ratio

$$e_2(\hat{\phi},\hat{\phi}_2) = \frac{\mathrm{Var}[\hat{\phi}]}{\mathrm{Var}[\hat{\phi}_2]},$$

where by $\mathrm{Var}[\hat{\phi}]$ we denote the asymptotic variance of the MLE.

Since the expression for the relative efficiency as defined above is not simple enough to allow us study its limiting behavior analytically, we compute it for a few particular cases. These results are summarized in Table 3.2, under the column

ϕ	b=5 k=100		b=10 k=50		b=50 k=10	
	Theory	Sim	Theory	Sim	Theory	Sim
-0.750	0.98998	0.999	0.97878	0.990	0.92595	0.934
-0.250	0.99292	0.991	0.98407	0.977	0.91329	0.898
-0.010	0.99199	0.990	0.98197	0.980	0.90182	0.891
0.010	0.99199	0.990	0.98197	0.989	0.90182	0.892
0.250	0.99292	0.992	0.98407	0.985	0.91329	0.912
0.750	0.98998	0.993	0.97878	0.992	0.92595	0.942

Table 3.2: Time Series: Small Blocks Asymptotic Relative Efficiency

labeled "Theory". Again, note here that the asymptotic bias for both the MLE and Small Blocks estimators is 0, therefore the ratio of asymptotic variances corresponds to the ratio of asymptotic mean squared errors.

One method to check the validity of the theoretical results is to complement the exact calculations with a simulation study. For one iteration, we simulate an AR(1) time series of length n, which we subsequently divide in b disjoint blocks of equal length k. As in Section 3.2, we first calculate the ML estimator for the autoregressive parameter. The pseudo-likelihood under the assumption of independence between blocks is simply the product between the block likelihoods, which are just the autoregressive likelihoods for each block. Through maximization of the pseudo-likelihood function we obtain the Small Blocks estimator. For each choice of ϕ and b, we repeat the process 1000 times, and store the values of the two estimators. We compute the mean squared errors for both of them, and report their ratio as the measure for asymptotic relative efficiency in Table 3.2.

We observe that the asymptotic performance of the Small Blocks estimator is very good, comparable to the classical maximum likelihood estimator. As expected, the method leads to a slightly less efficient estimator when the blocks are smaller. Therefore, we can conclude that the Small Block estimating technique is not only computationally very efficient, but performs very well according to statistical measures as well.

As a side note, we consider here the issue of consistency of the Small Blocks estimator. Although the general theory, as stated before, holds in this case, we take advantage of the simplicity of the criterion function in this case, and derive the consistency results for this particular case. We apply Lemma 5.10 of A.W. van der Vaart (1998). It should be noted that this result does not extend to the more general two-dimensional cases, since one of the hypotheses involves the monotonicity of the criterion function, which is, by its nature, one-dimensional.

Lemma: Let Θ be a subset of the real line and let ψ_n be random functions and ψ a fixed function of θ such that $\psi_n(\theta) \xrightarrow{p} \psi(\theta)$ for every θ. Assume that each map $\psi_n(\theta)$ is nondecreasing with $\psi_n(\hat{\theta}_n) = o_p(1)$ or is continuous and has exactly one zero, $\hat{\theta}_n$. Let θ_0 be a point such that $\psi(\theta_0 - \epsilon) < 0 < \psi(\theta_0 + \epsilon)$ for every $\epsilon > 0$. Then $\hat{\theta}_n \xrightarrow{p} \theta_0$.

In this case, both the interest parameter ϕ, and the true value of the parameter ϕ_0 lie in the parameter set $(-1, 1)$. The estimating equation in this case is given by:

$$\psi_n(\phi) = 2\,b\frac{\phi}{1-\phi^2} + \frac{1}{\sigma_\epsilon^2}\sum_{j=1}^{b}\left[2\phi\sum_{i=2}^{k-1} x_{k(j-1)+i}^2 - 2\sum_{i=1}^{k-1} x_{k(j-1)+i}x_{k(j-1)+i+1}\right]$$

and denote by $\hat{\phi}_n$ the approximate solution of the equation $\psi_n(\phi) = 0$.
Following the CLT, $\psi_n(\phi) \xrightarrow{p} \psi(\phi) = E[\psi_n(\phi)]$ for every ϕ, where

$$\begin{aligned}\psi(\phi) &= 2\,b\,\frac{\phi}{1-\phi^2} + \frac{1}{\sigma_\epsilon^2}\sum_{j=1}^{b}\left[2\phi\sum_{i=2}^{k-1}\frac{\sigma_\epsilon^2}{1-\phi_0^2} - 2\sum_{i=1}^{k-1}\frac{\sigma_\epsilon^2\,\phi_0}{1-\phi_0^2}\right] \\ &= 2\,b\left[\frac{\phi}{1-\phi^2} + \frac{(k-2)\phi - (k-1)\phi_0}{1-\phi_0^2}\right]\end{aligned}$$

Note that

$$\psi'(\phi) = 2\,b\left[\frac{1+\phi^2}{(1-\phi^2)^2} + \frac{k-2}{1-\phi_0^2}\right] \geq 0, \quad \text{for every } \phi.$$

We immediately note that $\psi(\phi_0) = 0$ and since ψ is a continuous nondecreasing function of ϕ, it follows that for every $\epsilon > 0$, $\psi(\phi_0 - \epsilon) < 0 < \psi(\phi_0 + \epsilon)$.

Also, computing

$$\psi_n'(\phi) = 2\,b\left[\frac{1+\phi^2}{(1-\phi^2)^2} + \frac{1}{\sigma_\epsilon^2}\sum_{j=1}^{b}\sum_{i=2}^{k-1} x_{k(j-1)+i}^2\right] \geq 0, \quad \text{for every } \phi$$

it follows that $\psi_n(\phi)$ is nondecreasing.
Since all the conditions of the Lemma are satisfied, consistency of $\hat{\phi}_n$ follows.

3.4 Hybrid Method

This section computes a hybrid estimator by relaxing the independence between blocks imposed in the previous section and including information about block means, as described in the Big Blocks section. The assumption here is that given the block means, the blocks are independent. In this context we first compute a pseudo-likelihood function as the product between the means likelihood and the conditional likelihood for each block. Then we find the parameter $\hat{\phi}_3$ which maximizes this function. The last step is to calculate the asymptotic variance of this estimator, using the expansion method.

The pseudo-likelihood function is of the form

$$L_2 \;=\; L_{means} \times \prod_{j=1}^{b} L_{cond_j}\ , \tag{3.35}$$

where L_{means} is the likelihood of the means process and L_{cond_j} is the conditional likelihood for block j given its mean.

To calculate the likelihood function for the means time series, we take advantage of the fact that in the Big Blocks stage of the analysis we showed that this time series is an ARMA(1,1) process whose coefficients we could derive as functions of the parameter of interest, ϕ. Recall the definition

$$X^* = \{X_1^*, X_2^*, \ldots, X_b^*\}$$

where

$$X_j^* = \frac{x_{k(j-1)+1} + x_{k(j-1)+2} + \ldots + x_{k(j-1)+k}}{j} \ ,$$

and therefore the likelihood function of the means is

$$L_{means} = \frac{1}{(2\pi)^{b/2}} \frac{1}{|V_{means}|^{1/2}} \exp\left\{ -\frac{1}{2} X^{*T} \; V_{means}^{-1} \; X^* \right\} \ ,$$

where the inverse covariance matrix is calculated following Trench's algorithm for inversion of Toeplitz matrices, as described in Section 3.2.

Block-Conditional Likelihoods: The next step is to calculate the block-likelihoods conditional on the block mean, and compute their product. We concentrate on the first block and then generalize the procedure to all blocks.

Consider the vector of observations for the first block, say $X_1^{k-1} = (x_1, x_2, \ldots, x_{k-1})$ and $\{X_1^*\} = \frac{1}{k}\sum_{i=1}^{k} x_i$. It follows that the joint distribution of $(x_1, x_2, \ldots, x_{k-1}, X_1^*)$ has mean 0 and covariance matrix:

$$V_{joint_1} = \begin{pmatrix} U_{k-1} & \tau \\ \tau^T & \eta \end{pmatrix}$$

where U_{k-1} is just the covariance matrix for an AR(1) model as described in (3.18), whose inverse is given in (3.19) and the determinant is

$$|U_{k-1}| = \frac{\sigma_\epsilon^{2\,(k-1)}}{1 - \phi^2} \ . \tag{3.36}$$

Also, we know $\eta = \mathrm{Var}[\mathrm{X}_1^*] = \gamma_0^* = \left[\frac{2\phi^{\mathrm{k}+1} - 2\phi - \mathrm{k}\phi^2 + \mathrm{k}}{\mathrm{k}^2\,(1-\phi)^2\,(1-\phi^2)}\right] \sigma_\epsilon^2$ as given in (3.12), and $\tau_i = \mathrm{Cov}[\mathrm{X_i}, \mathrm{X}_1^*]$. By standard algebraic manipulations, one obtains

$$\begin{aligned} \tau_i &= \frac{\gamma_{i-1} + \gamma_{i-2} + \cdots + \gamma_0 + \gamma_1 + \cdots + \gamma_{k-i}}{k} \\ &= \frac{1 + \phi - \phi^i - \phi^{k-i+1}}{k\ (\phi^2 - 1)\ (\phi - 1)}\ \sigma_\epsilon^2 \ . \end{aligned} \tag{3.37}$$

Hence following standard multivariate normal results, the conditional distribution of $(x_1, x_2, \ldots, x_{k-1})$ given X_1^* has mean

$$\mu_{cond_1} = \frac{\tau}{\eta} X_1^*$$

and covariance matrix

$$V_{cond_1} = U_{k-1} - \tau\, \eta^{-1} \tau^T \; . \tag{3.38}$$

To compute the likelihood function, a potential difficulty lies in calculating the inverse of V_{cond_1} and its determinant. From standard linear algebra results, as stated in Section 2.1, we calculate the determinant of V_{cond_1}. Thus, if in equation (2.4) we use U_{k-1} as the matrix A, and $b = \frac{\tau}{\sqrt{\eta}}$ it is immediate that

$$|V_{cond_1}| = |U_{k-1} - \tau\, \eta^{-1} \tau^T| = |U_{k-1}| \; \frac{\eta - \tau^T\, U_{k-1}^{-1} \tau}{\eta} \; .$$

After simplifications we obtain, for three or more observations per block,

$$|V_{cond_1}^{-1}| = \frac{2\phi^{k+1} - k\phi^2 - 2\phi + k}{\sigma_\epsilon^{2\,(k-1)} \; (1-\phi)^2} \tag{3.39}$$

Also, from equation (2.5), with the same notations as before, the inverse of V_{cond_1} can be calculated as:

$$V_{cond_1}^{-1} = (U_{k-1} - \tau\, \eta^{-1} \tau^T)^{-1} = U_{k-1}^{-1} + \frac{U_{k-1}^{-1} \tau \; \tau^T U_{k-1}^{-1}}{\eta - \tau^T\, U_{k-1}^{-1} \tau} \; .$$

After more algebra calculations and simplifications we obtain an exact form of the inverse conditional covariance matrix for the first block as a function of ϕ, the unknown parameter:

$$V_{cond_1}^{-1} = \frac{1}{\sigma_\epsilon^2} \begin{pmatrix} 2 & 1-\phi & 1 & \ldots & \ldots & 1 & 1 & 1+\phi \\ 1-\phi & 2+\phi^2 & 1-\phi & \ldots & \ldots & 1 & 1 & 1+\phi \\ 1 & 1-\phi & 2+\phi^2 & \ldots & \ldots & 1 & 1 & 1+\phi \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 1 & \ldots & \ldots & 2+\phi^2 & 1-\phi & 1+\phi \\ 1 & 1 & 1 & \ldots & \ldots & 1-\phi & 2+\phi^2 & 1 \\ 1+\phi & 1+\phi & 1+\phi & \ldots & \ldots & 1+\phi & 1 & \phi^2+2\phi+2 \end{pmatrix} \tag{3.40}$$

Note that the above calculations of the conditional covariance matrix are the same for all blocks, therefore for each block j, $V_{cond_1} = V_{cond_j}$. To simplify notation, we denote V_{cond_1} by V_{cond}.

We are now in the position to calculate the block-product part of the negative log-likelihood as

$$\begin{aligned} \sum_{j=1}^{b} -\log L_{cond_j} &= -b \log \left[\frac{1}{(2\pi)^{(k-1)/2}} \frac{1}{|V_{cond}|^{1/2}} \right] \qquad (3.41) \\ &+ \frac{1}{2} \sum_{j=1}^{b} \left(X_j^{k-1} - \mu_{cond_1} \right)^T V_{cond}^{-1} \left(X_j^{k-1} - \mu_{cond_1} \right) \\ &= \frac{b(k-1)}{2} \log(2\pi) - \frac{b}{2} \log |V_{cond}^{-1}| \\ &+ \frac{1}{2} \sum_{j=1}^{b} \left(X_j^{k-1} - \frac{\tau}{\eta} X_j^* \right)^T V_{cond}^{-1} \left(X_j^{k-1} - \frac{\tau}{\eta} X_j^* \right) . \end{aligned}$$

Note that from (3.40) it follows immediately that

$$\frac{\partial}{\partial \phi} V_{cond}^{-1} = \frac{1}{\sigma_\epsilon^2} \begin{pmatrix} 0 & -1 & 0 & 0 & \dots & \dots & 0 & 0 & 1 \\ -1 & 2\phi & -1 & 0 & \dots & \dots & 0 & 0 & 1 \\ 0 & -1 & 2\phi & -1 & \dots & \dots & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & \dots & 2\phi & -1 & 1 \\ 0 & 0 & 0 & 0 & \dots & \dots & -1 & 2\phi & 0 \\ 1 & 1 & 1 & 1 & \dots & \dots & 1 & 0 & 2\phi + 2 \end{pmatrix} \tag{3.42}$$

and that

$$\frac{\partial^2}{\partial \phi^2} V_{cond}^{-1} = \frac{1}{\sigma_\epsilon^2} \begin{pmatrix} 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 2 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 2 & 0 \\ 0 & 0 & 0 & \dots & 0 & 2 \end{pmatrix} . \tag{3.43}$$

We assume from here on that σ_ϵ^2 is known, since it leads to significant computational simplifications and has not much bearing on the final result. In other words, denoting the inverse covariance matrix for the Big Blocks process by $V(\phi)$, the inverse covariance matrix for each individual block, conditional on the corresponding mean by $W(\phi)$ (given by (3.40)) and by $\mu^j(\phi)$ the conditional mean for block j, we obtain, modulo some fixed constants, the following form for the pseudo-likelihood function which needs to be maximized with respect to ϕ:

$$\begin{aligned} p(\phi) &= -\log|V(\phi)| + X^{*T} V(\phi)\, X^* \\ &- b \log|W(\phi)| + \sum_{j=1}^{b} \left(X_j^{k-1} - \mu^j(\phi)\right)^T W(\phi) \left(X_j^{k-1} - \mu^j(\phi)\right) \ , \end{aligned} \tag{3.44}$$

Recall that

$$X^* = \{X_1^*, X_2^*, \dots, X_b^*\}$$

where

$$X_j^* = \frac{x_{k(j-1)+1} + x_{k(j-1)+2} + \ldots + x_{k(j-1)+k}}{k}$$

and hence we obtain

$$X^{*T} V(\phi)\, X^* = \frac{1}{k^2} \sum_{i=1}^{b} \sum_{j=1}^{b} v_{ij}(\phi) \sum_{\ell=1}^{k} \sum_{m=1}^{k} x_{(i-1)k+\ell}\, x_{(j-1)k+m} \;.$$

On the other hand,

$$X_j^{k-1} = (x_{(j-1)k+1}, x_{(j-1)k+2}, \ldots, x_{(j-1)k+k-1})$$

and therefore

$$\begin{aligned} & \sum_{j=1}^{b} \left(X_j^{k-1} - \mu^j(\phi)\right)^T W(\phi) \left(X_j^{k-1} - \mu^j(\phi)\right) \\ = & \sum_{j=1}^{b} \sum_{\ell=1}^{k-1} \sum_{m=1}^{k-1} w_{\ell m}(\phi) \left[\left(x_{(j-1)k+\ell} - \mu_{(j-1)k+\ell}(\phi)\right)\right. \\ \times & \left.\left(x_{(j-1)k+m} - \mu_{(j-1)k+m}(\phi)\right)\right] \end{aligned} \tag{3.45}$$

To simplify notation, we denote $V = V(\phi)$, $W = W(\phi)$ and $\mu_j = \mu^j(\phi)$. Thus the expression in (3.45) becomes

$$\begin{aligned} & \sum_{j=1}^{b} \sum_{\ell=1}^{k-1} \sum_{m=1}^{k-1} w_{\ell m} \left[\left(x_{(j-1)k+\ell} - \mu_{(j-1)k+\ell}\right) \left(x_{(j-1)k+m} - \mu_{(j-1)k+m}\right)\right] \\ = & \sum_{j=1}^{b} \sum_{\ell=1}^{k-1} \sum_{m=1}^{k-1} w_{\ell m}\, x_{(j-1)k+\ell}\, x_{(j-1)k+m} - 2 \sum_{j=1}^{b} \sum_{\ell=1}^{k-1} \sum_{m=1}^{k-1} w_{\ell m}\, x_{(j-1)k+\ell}\, \mu_{(j-1)k+m} \\ + & \sum_{j=1}^{b} \sum_{\ell=1}^{k-1} \sum_{m=1}^{k-1} w_{\ell m}\, \mu_{(j-1)k+\ell}\, \mu_{(j-1)k+m} \;. \end{aligned}$$

Therefore we rewrite (3.44) as

$$p(\phi) \;=\; -(\log|V| + b\log|W|)$$

$$
\begin{aligned}
&+ \frac{1}{k^2}\sum_{j=1}^{b}\sum_{i=1}^{b}\sum_{\ell=1}^{k}\sum_{m=1}^{k} v_{ij} x_{(i-1)k+\ell}\, x_{(j-1)k+m} \\
&+ \sum_{j=1}^{b}\sum_{\ell=1}^{k-1}\sum_{m=1}^{k-1} w_{\ell m}\, x_{(j-1)k+\ell}\, x_{(j-1)k+m} - 2\sum_{j=1}^{b}\sum_{\ell=1}^{k-1}\sum_{m=1}^{k-1} w_{\ell m}\, x_{(j-1)k+\ell}\, \mu_{(j-1)k+m} \\
&+ \sum_{j=1}^{b}\sum_{\ell=1}^{k-1}\sum_{m=1}^{k-1} w_{\ell m}\, \mu_{(j-1)k+\ell}\, \mu_{(j-1)k+m} \qquad (3.46)
\end{aligned}
$$

and further define

$$g(\phi) = -(\log|V| + b\log|W|)$$

where $\mid V \mid$ is the determinant of the covariance matrix given in (3.17) and $\mid W \mid$ is given by (3.39) (as functions of ϕ). Since the hybrid estimator, let us call it $\hat{\phi}_3$, maximizes the estimation function defined by (3.46), it is a solution of the equation

$$p'(\phi) = 0\ . \qquad (3.47)$$

There is not an apparent closed form solution for equation (3.47). In order to quantify the efficiency of the hybrid estimator we need to compute the variance of the hybrid estimator. To do so, we make use of the "information sandwich" method. The key elements for this technique are the expected value of the second derivative and the variance of the first derivative of the estimating function, $p(\phi)$.

Since we need to compute the expected value and variance of the first and second derivative of the pseudo-likelihood function, and since the derivation computations are more complicated, we derive them first.

The first derivative is given by:

$$
\begin{aligned}
p'(\phi) \;&=\; g'(\phi) + \frac{1}{k^2}\sum_{j=1}^{b}\sum_{i=1}^{b}\sum_{\ell=1}^{k}\sum_{m=1}^{k} v'_{ij} x_{(i-1)k+\ell}\, x_{(j-1)k+m} \\
&+ \sum_{j=1}^{b}\sum_{\ell=1}^{k-1}\sum_{m=1}^{k-1} w'_{\ell m}\, x_{(j-1)k+\ell}\, x_{(j-1)k+m} \\
&- 2\sum_{j=1}^{b}\sum_{\ell=1}^{k-1}\sum_{m=1}^{k-1} w'_{\ell m}\, x_{(j-1)k+\ell}\, \mu_{(j-1)k+m} - 2\sum_{j=1}^{b}\sum_{\ell=1}^{k-1}\sum_{m=1}^{k-1} w_{\ell m}\, x_{(j-1)k+\ell}\, \mu'_{(j-1)k+m}
\end{aligned}
$$

$$+ \quad 2\sum_{j=1}^{b}\sum_{\ell=1}^{k-1}\sum_{m=1}^{k-1} w_{\ell m}\, \mu'_{(j-1)k+\ell}\, \mu_{(j-1)k+m} + \sum_{j=1}^{b}\sum_{\ell=1}^{k-1}\sum_{m=1}^{k-1} w'_{\ell m}\, \mu_{(j-1)k+\ell}\, \mu_{(j-1)k+m}$$

while the second derivative is:

$$\begin{aligned}
p''(\phi) \;&=\; g''(\phi) + \frac{1}{k^2}\sum_{j=1}^{b}\sum_{i=1}^{b}\sum_{\ell=1}^{k}\sum_{m=1}^{k} v''_{ij} x_{(i-1)k+\ell}\, x_{(j-1)k+m} \\
&+\; \sum_{j=1}^{b}\sum_{\ell=1}^{k-1}\sum_{m=1}^{k-1} w''_{\ell m}\, x_{(j-1)k+\ell}\, x_{(j-1)k+m} \\
&-\; 2\sum_{j=1}^{b}\sum_{\ell=1}^{k-1}\sum_{m=1}^{k-1} w''_{\ell m}\, x_{(j-1)k+\ell}\, \mu_{(j-1)k+m} - 4\sum_{j=1}^{b}\sum_{\ell=1}^{k-1}\sum_{m=1}^{k-1} w'_{\ell m}\, x_{(j-1)k+\ell}\, \mu'_{(j-1)k+m} \\
&-\; 2\sum_{j=1}^{b}\sum_{\ell=1}^{k-1}\sum_{m=1}^{k-1} w_{\ell m}\, x_{(j-1)k+\ell}\, \mu''_{(j-1)k+m} + 4\sum_{j=1}^{b}\sum_{\ell=1}^{k-1}\sum_{m=1}^{k-1} w'_{\ell m}\, \mu'_{(j-1)k+\ell}\, \mu_{(j-1)k+m} \\
&+\; 2\sum_{j=1}^{b}\sum_{\ell=1}^{k-1}\sum_{m=1}^{k-1} w_{\ell m}\, \mu'_{(j-1)k+\ell}\, \mu'_{(j-1)k+m} + 2\sum_{j=1}^{b}\sum_{\ell=1}^{k-1}\sum_{m=1}^{k-1} w_{\ell m}\, \mu_{(j-1)k+\ell}\, \mu''_{(j-1)k+m} \\
&+\; \sum_{j=1}^{b}\sum_{\ell=1}^{k-1}\sum_{m=1}^{k-1} w''_{\ell m}\, \mu_{(j-1)k+\ell}\, \mu_{(j-1)k+m} \;.
\end{aligned}$$

Also, recall that by definition that

$$\mu_{(j-1)k+\ell} = \frac{\tau_\ell}{\eta}\frac{1}{k}\sum_{p=1}^{k} x_{(j-1)k+p}$$

and denote by

$$\tau^*_\ell = \frac{\tau_\ell}{k\eta}\ . \tag{3.48}$$

Thus, for example, one can express the last term of the sum above as:

$$\sum_{j=1}^{b}\sum_{\ell=1}^{k-1}\sum_{m=1}^{k-1} w''_{\ell m}\, \mu_{(j-1)k+\ell}\, \mu_{(j-1)k+m} = \sum_{j=1}^{b}\sum_{\ell=1}^{k-1}\sum_{m=1}^{k-1}\sum_{p=1}^{k}\sum_{q=1}^{k} w''_{\ell m}\tau^*_\ell\tau^*_m\, x_{(j-1)k+p}\, x_{(j-1)k+q}$$

and similarly all the other factors containing the conditional mean.

We are now in the position to calculate the expected value of the second derivative:

$$P_2(\phi) = E[p''(\phi)] \; .$$

Since $\{X_i\}$ is just the original AR(1) time series, we use its covariance structure to deduce that:

$$\begin{aligned}
E[p''(\phi)] &= g''(\phi) + \sum_{j=1}^{b}\sum_{i=1}^{b}\sum_{\ell=1}^{k}\sum_{m=1}^{k} v''_{ij}\,\gamma_{|(i-j)k+\ell-m|} \\
&+ b\sum_{\ell=1}^{k-1}\sum_{m=1}^{k-1} w''_{\ell m}\,\gamma_{|\ell-m|} \\
&- 2\sum_{j=1}^{b}\sum_{\ell=1}^{k-1}\sum_{m=1}^{k-1}\sum_{p=1}^{k} w''_{\ell m}\tau^*_m\,\gamma_{|\ell-p|} \\
&- 4\sum_{j=1}^{b}\sum_{\ell=1}^{k-1}\sum_{m=1}^{k-1}\sum_{p=1}^{k} w'_{\ell m}\tau^{*\prime}_m\,\gamma_{|\ell-p|} \\
&- 2\sum_{j=1}^{b}\sum_{\ell=1}^{k-1}\sum_{m=1}^{k-1}\sum_{p=1}^{k} w_{\ell m}\tau^{*\prime\prime}_m\,\gamma_{|\ell-p|} \\
&+ 4\sum_{j=1}^{b}\sum_{\ell=1}^{k-1}\sum_{m=1}^{k-1}\sum_{p=1}^{k}\sum_{q=1}^{k} w'_{\ell m}\tau^{*\prime}_m\tau^*_\ell\,\gamma_{|p-q|} \\
&+ 2\sum_{j=1}^{b}\sum_{\ell=1}^{k-1}\sum_{m=1}^{k-1}\sum_{p=1}^{k}\sum_{q=1}^{k} w_{\ell m}\tau^{*\prime}_m\tau^{*\prime}_\ell\,\gamma_{|p-q|} \\
&+ 2\sum_{j=1}^{b}\sum_{\ell=1}^{k-1}\sum_{m=1}^{k-1}\sum_{p=1}^{k}\sum_{q=1}^{k} w_{\ell m}\tau^*_m\tau^{*\prime\prime}_\ell\,\gamma_{|p-q|} \\
&+ \sum_{j=1}^{b}\sum_{\ell=1}^{k-1}\sum_{m=1}^{k-1}\sum_{p=1}^{k}\sum_{q=1}^{k} w''_{\ell m}\tau^*_\ell\tau^*_m\,\gamma_{|p-q|} \; ,
\end{aligned}$$

where τ_i is given by (3.37), τ^*_ℓ given by (3.48), by $\tau^{*\prime}_\ell$ and $\tau^{*\prime\prime}_\ell$ we mean the first and second derivatives of τ_ℓ with respect to ϕ. Also, recall that W denotes the matrix $V^{-1}_{cond_j}$ given by (3.40), with entries $w_{\ell m}$ and $w'_{\ell m}$, $w''_{\ell m}$ are the first and second derivatives of $w_{\ell m}$ with respect to ϕ. Also, by $g'(\phi)$ and $g''(\phi)$ we mean the first and second derivatives of $g(\phi)$ with respect to ϕ. The main feature of the AR(1) time series that enables us to complete computing $E[p''(\phi)]$, is that

$$\gamma_i = \frac{\phi^i}{1-\phi^2}\ \sigma_\epsilon^2\ .$$

Next, consider the first derivative, $p'(\phi)$, and define

$$P_1(\phi) = \mathrm{Var}[p'(\phi)] \tag{3.49}$$

To compute $P_1(\phi)$, we first rewrite the first derivative as:

$$\begin{aligned}
p'(\phi) &= g'(\phi) + \sum_{j=1}^{b}\sum_{i=1}^{b}\sum_{\ell=1}^{k}\sum_{m=1}^{k} v'_{ij} x_{(i-1)k+\ell}\, x_{(j-1)k+m} \\
&+ \sum_{j=1}^{b}\sum_{\ell=1}^{k-1}\sum_{m=1}^{k-1} w'_{\ell m}\, x_{(j-1)k+\ell}\, x_{(j-1)k+m} \\
&- 2\sum_{j=1}^{b}\sum_{\ell=1}^{k-1}\sum_{m=1}^{k-1}\sum_{p=1}^{k} w'_{\ell m}\, \tau^*_\ell\, x_{(j-1)k+\ell}\, x_{(j-1)k+p} \\
&- 2\sum_{j=1}^{b}\sum_{\ell=1}^{k-1}\sum_{m=1}^{k-1}\sum_{p=1}^{k} w_{\ell m}\, \tau^{*\prime}_\ell\, x_{(j-1)k+\ell}\, x_{(j-1)k+p} \\
&+ 2\sum_{j=1}^{b}\sum_{\ell=1}^{k-1}\sum_{m=1}^{k-1}\sum_{p=1}^{k}\sum_{q=1}^{k} w_{\ell m}\, \tau^{*\prime}_\ell\, \tau^*_m\, x_{(j-1)k+p}\, x_{(j-1)k+q} \\
&+ \sum_{j=1}^{b}\sum_{\ell=1}^{k-1}\sum_{m=1}^{k-1}\sum_{p=1}^{k}\sum_{q=1}^{k} w'_{\ell m}\, \tau^*_\ell\, \tau^*_m\, x_{(j-1)k+p}\, x_{(j-1)k+q}\ .
\end{aligned}$$

Recall now that the process $\{X_i\}$ is an AR(1) time series, hence we use one more time the representation:

$$x_t = \sigma_\epsilon \sum_{r=-\infty}^{t} \phi^{t-r}\xi_r$$

and thus

$$x_t x_{t'} = \sigma_\epsilon^2 \sum_{r=-\infty}^{t}\sum_{s=-\infty}^{t'} \phi^{t+t'-r-s}\xi_r\ \xi_s\ .$$

Hence the expression of the first derivative becomes:

$$
\begin{aligned}
p'(\phi) &= g'(\phi) + \sum_{j=1}^{b}\sum_{i=1}^{b}\sum_{\ell=1}^{k}\sum_{m=1}^{k}\sum_{r=-\infty}^{(i-1)k+\ell}\sum_{s=-\infty}^{(j-1)k+m} v'_{ij}\,\phi^{(i+j-2)k+\ell+m-r-s}\,\xi_r\,\xi_s \\
&+ \sum_{j=1}^{b}\sum_{\ell=1}^{k-1}\sum_{m=1}^{k-1}\sum_{r=-\infty}^{(j-1)k+\ell}\sum_{s=-\infty}^{(j-1)k+m} w'_{\ell m}\,\phi^{2(j-1)k+\ell+m-r-s}\,\xi_r\,\xi_s \\
&- 2\sum_{j=1}^{b}\sum_{\ell=1}^{k-1}\sum_{m=1}^{k-1}\sum_{p=1}^{k}\sum_{r=-\infty}^{(j-1)k+\ell}\sum_{s=-\infty}^{(j-1)k+p} w'_{\ell m}\,\tau_\ell^*\,\phi^{2(j-1)k+\ell+p-r-s}\,\xi_r\,\xi_s \\
&- 2\sum_{j=1}^{b}\sum_{\ell=1}^{k-1}\sum_{m=1}^{k-1}\sum_{p=1}^{k}\sum_{r=-\infty}^{(j-1)k+\ell}\sum_{s=-\infty}^{(j-1)k+p} w_{\ell m}\,\tau_\ell^{*\prime}\,\phi^{2(j-1)k+\ell+p-r-s}\,\xi_r\,\xi_s \\
&+ 2\sum_{j=1}^{b}\sum_{\ell=1}^{k-1}\sum_{m=1}^{k-1}\sum_{p=1}^{k}\sum_{q=1}^{k}\sum_{r=-\infty}^{(j-1)k+p}\sum_{s=-\infty}^{(j-1)k+q} w_{\ell m}\,\tau_\ell^{*\prime}\,\tau_m^*\,\phi^{2(j-1)k+p+q-r-s}\,\xi_r\,\xi_s \\
&+ \sum_{j=1}^{b}\sum_{\ell=1}^{k-1}\sum_{m=1}^{k-1}\sum_{p=1}^{k}\sum_{q=1}^{k}\sum_{r=-\infty}^{(j-1)k+p}\sum_{s=-\infty}^{(j-1)k+q} w'_{\ell m}\,\tau_\ell^*\,\tau_m^*\,\phi^{2(j-1)k+p+q-r-s}\,\xi_r\,\xi_s \;.
\end{aligned}
$$

Making the corresponding notations, rewrite

$$
\begin{aligned}
p'(\phi) &= g'(\phi) + \sum_{r,s:\ r\le s} a_{rs}^{(1)}\xi_r\,\xi_s + \sum_{r,s:\ r\le s} a_{rs}^{(2)}\xi_r\,\xi_s - 2\sum_{r,s:\ r\le s} a_{rs}^{(3)}\xi_r\xi_s \\
&- 2\sum_{r,s:\ r\le s} a_{rs}^{(4)}\xi_r\,\xi_s + 2\sum_{r,s:\ r\le s} a_{rs}^{(5)}\xi_r\,\xi_s + \sum_{r,s:\ r\le s} a_{rs}^{(6)}\xi_r\,\xi_s \;.
\end{aligned}
$$

Note here that $a_{rs}^{(1)}$ are the Big Blocks coefficients, as given by (3.13).

For the next 5 cases, the coefficients are computed as follows (denote by $r_1 = \lceil r/k \rceil$, $s_1 = \lceil s/k \rceil$ and $r_2 = r - (r_1 - 1)\,k$, $s_2 = s - (s_1 - 1)\,k$): [2]

[2] By $\lceil x \rceil$ we mean the smallest integer greater than or equal to x, and by $\lfloor x \rfloor$ we mean the largest integer smaller than or equal to x.

Case 2

$$a_{rs}^{(2)} = \begin{cases} \sum_{j=1}^{b} \sum_{\ell=1}^{k-1} \sum_{m=1}^{k-1} w'_{\ell m}\, \phi^{2(j-1)k+\ell+m-r-s} \ , & \text{if } r = s \leq 1, \\[2ex] 2 \sum_{j=1}^{b} \sum_{\ell=1}^{k-1} \sum_{m=1}^{k-1} w'_{\ell m}\, \phi^{2(j-1)k+\ell+m-r-s} \ , & \text{if } r < s \leq 1, \\[2ex] 2 \sum_{j=s_1+1}^{b} \sum_{\ell=1}^{k-1} \sum_{m=1}^{k-1} w'_{\ell m}\, \phi^{2(j-1)k+\ell+m-r-s} + & \\ 2 \sum_{\ell=1}^{k-1} \sum_{m=s_2}^{k-1} w'_{\ell m}\, \phi^{2(s_1-1)k+\ell+m-r-s} \ , & \text{if } r \leq 1 < s \ , \\[2ex] \sum_{j=s_1+1}^{b} \sum_{\ell=1}^{k-1} \sum_{m=1}^{k-1} w'_{\ell m}\, \phi^{2(j-1)k+\ell+m-r-s} + & \\ \sum_{\ell=r_2}^{k-1} \sum_{m=s_2}^{k-1} w'_{\ell m}\, \phi^{2(s_1-1)k+\ell+m-r-s} \ , & \text{if } 2 \leq r = s \ , \\[2ex] 2 \sum_{j=s_1+1}^{b} \sum_{\ell=1}^{k-1} \sum_{m=1}^{k-1} w'_{\ell m}\, \phi^{2(j-1)k+\ell+m-r-s} + & \\ 2 \sum_{\ell=r_2}^{k-1} \sum_{m=s_2}^{k-1} w'_{\ell m}\, \phi^{2(s_1-1)k+\ell+m-r-s} \ , & \text{if } 2 \leq r < s \text{ and } r_1 = s_1 \ , \\[2ex] 2 \sum_{j=s_1+1}^{b} \sum_{\ell=1}^{k-1} \sum_{m=1}^{k-1} w'_{\ell m}\, \phi^{2(j-1)k+\ell+m-r-s} + & \\ 2 \sum_{\ell=1}^{k-1} \sum_{m=s_2}^{k-1} w'_{\ell m}\, \phi^{2(s_1-1)k+\ell+m-r-s} \ , & \text{if } 2 \leq r < s \text{ and } r_1 < s_1 \ . \end{cases}$$

Case 3

$$a_{rs}^{(3)} = \begin{cases} \sum_{j=1}^{b} \sum_{\ell=1}^{k-1} \sum_{m=1}^{k-1} \sum_{p=1}^{k} w'_{\ell m}\, \tau_\ell^*\, \phi^{2(j-1)k+\ell+p-r-s} \,, & \text{if } r = s \leq 1, \\ 2\sum_{j=1}^{b} \sum_{\ell=1}^{k-1} \sum_{m=1}^{k-1} \sum_{p=1}^{k} w'_{\ell m}\, \tau_\ell^*\, \phi^{2(j-1)k+\ell+p-r-s} \,, & \text{if } r < s \leq 1, \\ 2\sum_{j=s_1+1}^{b} \sum_{\ell=1}^{k-1} \sum_{m=1}^{k-1} \sum_{p=1}^{k} w'_{\ell m}\, \tau_\ell^*\, \phi^{2(j-1)k+\ell+p-r-s} + & \\ \sum_{\ell=1}^{k-1} \sum_{m=1}^{k-1} \sum_{p=s_2}^{k} w'_{\ell m}\, \tau_\ell^*\, \phi^{2(s_1-1)k+\ell+p-r-s} + & \\ \sum_{\ell=s_2}^{k-1} \sum_{m=1}^{k-1} \sum_{p=1}^{k} w'_{\ell m}\, \tau_\ell^*\, \phi^{2(s_1-1)k+\ell+p-r-s} \,, & \text{if } r \leq 1 < s \text{ and} \\ & \quad s_2 < k \,, \\ 2\sum_{j=s_1+1}^{b} \sum_{\ell=1}^{k-1} \sum_{m=1}^{k-1} \sum_{p=1}^{k} w'_{\ell m}\, \tau_\ell^*\, \phi^{2(j-1)k+\ell+p-r-s} + & \\ \sum_{\ell=1}^{k-1} \sum_{m=1}^{k-1} w'_{\ell m}\, \tau_\ell^*\, \phi^{2(s_1-1)k+\ell+k-r-s} \,, & \text{if } r \leq 1 < s \text{ and} \\ & \quad s_2 = k \,, \\ \sum_{j=s_1+1}^{b} \sum_{\ell=1}^{k-1} \sum_{m=1}^{k-1} \sum_{p=1}^{k} w'_{\ell m}\, \tau_\ell^*\, \phi^{2(j-1)k+\ell+p-r-s} + & \\ \sum_{\ell=r_2}^{k-1} \sum_{m=1}^{k-1} \sum_{p=s_2}^{k} w'_{\ell m}\, \tau_\ell^*\, \phi^{2(s_1-1)k+\ell+p-r-s} \,, & \text{if } 2 \leq r = s \,, \\ 2\sum_{j=s_1+1}^{b} \sum_{\ell=1}^{k-1} \sum_{m=1}^{k-1} \sum_{p=1}^{k} w'_{\ell m}\, \tau_\ell^*\, \phi^{2(j-1)k+\ell+p-r-s} + & \\ \sum_{\ell=r_2}^{k-1} \sum_{m=1}^{k-1} \sum_{p=s_2}^{k} w'_{\ell m}\, \tau_\ell^*\, \phi^{2(s_1-1)k+\ell+p-r-s} + & \\ \sum_{\ell=s_2}^{k-1} \sum_{m=1}^{k-1} \sum_{p=r_2}^{k} w'_{\ell m}\, \tau_\ell^*\, \phi^{2(s_1-1)k+\ell+p-r-s} \,, & \text{if } 2 \leq r < s \text{ and} \\ & \quad r_1 = s_1 \text{ and } s_2 < k \,, \\ 2\sum_{j=s_1+1}^{b} \sum_{\ell=1}^{k-1} \sum_{m=1}^{k-1} \sum_{p=1}^{k} w'_{\ell m}\, \tau_\ell^*\, \phi^{2(j-1)k+\ell+p-r-s} + & \\ \sum_{\ell=r_2}^{k-1} \sum_{m=1}^{k-1} w'_{\ell m}\, \tau_\ell^*\, \phi^{2(s_1-1)k+\ell+k-r-s} \,, & \text{if } 2 \leq r < s \text{ and} \\ & \quad r_1 = s_1 \text{ and } s_2 = k \,, \\ 2\sum_{j=s_1+1}^{b} \sum_{\ell=1}^{k-1} \sum_{m=1}^{k-1} \sum_{p=1}^{k} w'_{\ell m}\, \tau_\ell^*\, \phi^{2(j-1)k+\ell+p-r-s} + & \\ \sum_{\ell=1}^{k-1} \sum_{m=1}^{k-1} \sum_{p=s_2}^{k} w'_{\ell m}\, \tau_\ell^*\, \phi^{2(s_1-1)k+\ell+p-r-s} + & \\ \sum_{\ell=s_2}^{k-1} \sum_{m=1}^{k-1} \sum_{p=1}^{k} w'_{\ell m}\, \tau_\ell^*\, \phi^{2(s_1-1)k+\ell+p-r-s} \,, & \text{if } 2 \leq r < s \text{ and} \\ & \quad r_1 < s_1 \text{ and } s_2 < k \,, \\ 2\sum_{j=s_1+1}^{b} \sum_{\ell=1}^{k-1} \sum_{m=1}^{k-1} \sum_{p=1}^{k} w'_{\ell m}\, \tau_\ell^*\, \phi^{2(j-1)k+\ell+p-r-s} + & \\ \sum_{\ell=1}^{k-1} \sum_{m=1}^{k-1} w'_{\ell m}\, \tau_\ell^*\, \phi^{2(s_1-1)k+\ell+k-r-s} \,, & \text{if } 2 \leq r < s \text{ and} \\ & \quad r_1 < s_1 \text{ and } s_2 = k \,. \end{cases}$$

Case 4 Similar to *Case 3*, except we replace $w'_{\ell m}\, \tau_\ell^*$ by $w_{\ell m}\, \tau_\ell^{*\prime}$.

Case 5

$$a_{rs}^{(5)} = \begin{cases} \sum_{j=1}^{b}\sum_{\ell=1}^{k-1}\sum_{m=1}^{k-1}\sum_{p=1}^{k}\sum_{q=1}^{k} w_{\ell m}\,\tau_\ell^{*\prime}\,\tau_m^{*}\,\phi^{2(j-1)k+p+q-r-s}\ , & \text{if } r=s\leq 1\ , \\ 2\sum_{j=1}^{b}\sum_{\ell=1}^{k-1}\sum_{m=1}^{k-1}\sum_{p=1}^{k}\sum_{q=1}^{k} w_{\ell m}\,\tau_\ell^{*\prime}\,\tau_m^{*}\,\phi^{2(j-1)k+p+q-r-s}\ , & \text{if } r<s\leq 1\ , \\ 2\sum_{j=s_1+1}^{b}\sum_{\ell=1}^{k-1}\sum_{m=1}^{k-1}\sum_{p=1}^{k}\sum_{q=1}^{k} w_{\ell m}\,\tau_\ell^{*\prime}\,\tau_m^{*}\,\phi^{2(j-1)k+p+q-r-s}+ & \\ 2\sum_{\ell=1}^{k-1}\sum_{m=1}^{k-1}\sum_{p=1}^{k}\sum_{q=s_2}^{k} w_{\ell m}\,\tau_\ell^{*\prime}\,\tau_m^{*}\,\phi^{2(s_1-1)k+p+q-r-s}\ , & \text{if } r\leq 1<s\ , \\ \sum_{j=s_1+1}^{b}\sum_{\ell=1}^{k-1}\sum_{m=1}^{k-1}\sum_{p=1}^{k}\sum_{q=1}^{k} w_{\ell m}\,\tau_\ell^{*\prime}\,\tau_m^{*}\,\phi^{2(j-1)k+p+q-r-s}+ & \\ \sum_{\ell=1}^{k-1}\sum_{m=1}^{k-1}\sum_{p=r_2}^{k}\sum_{q=s_2}^{k} w_{\ell m}\,\tau_\ell^{*\prime}\,\tau_m^{*}\,\phi^{2(s_1-1)k+p+q-r-s}\ , & \text{if } 2\leq r=s\ , \\ 2\sum_{j=s_1+1}^{b}\sum_{\ell=1}^{k-1}\sum_{m=1}^{k-1}\sum_{p=1}^{k}\sum_{q=1}^{k} w_{\ell m}\,\tau_\ell^{*\prime}\,\tau_m^{*}\,\phi^{2(j-1)k+p+q-r-s}+ & \\ 2\sum_{\ell=1}^{k-1}\sum_{m=1}^{k-1}\sum_{p=r_2}^{k}\sum_{q=s_2}^{k} w_{\ell m}\,\tau_\ell^{*\prime}\,\tau_m^{*}\,\phi^{2(s_1-1)k+p+q-r-s}\ , & \text{if } 2\leq r<s \\ & \text{and } r_1=s_1, \\ 2\sum_{j=s_1+1}^{b}\sum_{\ell=1}^{k-1}\sum_{m=1}^{k-1}\sum_{p=1}^{k}\sum_{q=1}^{k} w_{\ell m}\,\tau_\ell^{*\prime}\,\tau_m^{*}\,\phi^{2(j-1)k+p+q-r-s}+ & \\ 2\sum_{\ell=1}^{k-1}\sum_{m=1}^{k-1}\sum_{p=1}^{k}\sum_{q=s_2}^{k} w_{\ell m}\,\tau_\ell^{*\prime}\,\tau_m^{*}\,\phi^{2(s_1-1)k+p+q-r-s}\ , & \text{if } 2\leq r<s \\ & \text{and } r_1<s_1\ . \end{cases}$$

Case 6 Similar to *Case 5*, the only difference is that we replace $w_{\ell m}\,\tau_\ell^{*'}\,\tau_m^{*}$ by $w_{\ell m}'\,\tau_\ell^{*}\,\tau_m^{*}$.

Therefore we have expanded the first derivative of the pseudo-likelihood function as a quadratic form of independent normal random variables. Once the coefficients of the quadratic form are identified, we can proceed to compute the asymptotic variance, following the Martingale Central Limit Theorem (its application for quadratic forms, as described in Chapter II). We compute

$$v_n = 2\sum_{r} a_{r,r}^2 + \sum_{\{r,s:\ r<s\}} a_{r,s}^2$$

where

$$a_{r,s} = a_{r,s}^{(1)} + a_{r,s}^{(2)} + a_{r,s}^{(3)} + a_{r,s}^{(4)} + a_{r,s}^{(5)} + a_{r,s}^{(6)}$$

as described above. We calculate the coefficient $a_{r,s}$ for each of the 5 main different cases (e.g. $r = s \leq 1$, $2 < r = s$, $r < s \leq 1$, $r \leq 1 < s$ and $2 < r < s$). To compute the sums over r and s in the variance formula, we note there are two possibilities: finite sums when $2 < r,\ s < n$ or infinite sums when $r,\ s \leq 1$. To exemplify, I will explain in more detail one of each possibilities.

Finite sums: both $r > 2$ and $s > 2$: This case is basically a straightforward summation over a finite range.

Infinite sums: $r < s \leq 1$: In this case we use the fact that both r and s can be separated from the other power indices. Therefore, we compute the finite summations first and then sum over r and s. For example, in one of the cases when $r < s \leq 1$ one needs to calculate a sum of the following type:

$$\sum_{s=-\infty}^{1} \sum_{r=-\infty}^{s} \left[\sum_{j=1}^{b} \sum_{\ell=1}^{k} \sum_{m=1}^{k} w_{\ell m} \phi^{2(j-1)k+\ell-m-r-s} \right]^2 =$$

$$\sum_{s=-\infty}^{1} \sum_{r=-\infty}^{s} \phi^{-2r-2s} \left[\sum_{j=1}^{b} \sum_{\ell=1}^{k} \sum_{m=1}^{k} w_{\ell m} \phi^{2(j-1)k+\ell-m} \right]^2 .$$

Next evaluate the finite sums over j, ℓ and m, and denote the result by Ψ, which is independent of r and s. Continue the summation as:

$$\sum_{s=-\infty}^{1} \sum_{r=-\infty}^{s} \Psi \, \phi^{-2r-2s} = \Psi \sum_{s=-\infty}^{1} \phi^{-2s} \sum_{r=-\infty}^{s} \phi^{-2r} = \frac{\Psi}{\phi^4(1-\phi^2)(1-\phi^4)}$$

Combination – finite and infinite sums: $r \leq 1 < s$: Follow a similar argument as in the previous case, the only difference being that in the process of calculating the finite sums one also calculates the sum over s and only use the geometric series calculations for r.

We conclude by computing the variance of the "hybrid estimator" for which we follow the "sandwich information" technique, and obtain

$$\mathrm{Var}[\hat{\phi}_3] = P_2^{-1}(\phi)\, P_1(\phi) P_2^{-1}(\phi)\ . \tag{3.50}$$

An approximation for the Bias: As a side comment, note that we can alternatively calculate the expected value of the first derivative of the pseudo-likelihood function using the fact that $\{X_i\}$ is an AR(1) process. We exploit its underlying covariance structure and the form of the first derivative and proceed similarly as in the computations for the expected value of the second derivative. Since the analytical final form is too complicated, we compute the bias of the hybrid estimator numerically.

As a measure of efficiency, we use the Mean Squared Error (MSE) instead of the variance alone. To calculate the bias in this case, we use Taylor series expansion as given in the following derivation.

Since $\hat{\phi}_3$ maximizes the log-likelihood function, it follows that

$$p'(\hat{\phi}_3) = 0$$

and therefore

$$-p'(\phi) = p'(\hat{\phi}_3) - p'(\phi) \approx (\hat{\phi}_3 - \phi)^T\, p''(\phi)$$

where ϕ is the true value of the AR(1) parameter. Therefore, applying the Central Limit Theorem both side of the above equation we obtain the following formula for the bias of the estimator $\hat{\phi}_3$:

$$B_3[\phi] = E[\hat{\phi}_3 - \phi] \approx -[p''(\phi)]^{-1} E[p'(\phi)].$$

Thus we compute the MSE as

$$\mathrm{MSE}_3[\phi] = \mathrm{Var}[\hat{\phi}_3] + B_3^2[\phi] \tag{3.51}$$

and define the asymptotic relative efficiency as:

$$e_3(\hat{\phi}, \hat{\phi}_3) = \frac{\text{MSE}[\hat{\phi}]}{\text{MSE}_3[\phi]} \, .$$

Note here that for the classical ML estimator we can calculate the asymptotic bias which turns out to be 0. These calculations are feasible due to the much simpler form of the classical likelihood function. However, for the Hybrid estimator the analytical form of the asymptotic bias is intractable, therefore we analyze its asymptotic behavior on a set of particular cases. We present the theoretical results obtained for the asymptotic relative efficiency in Table 3.3 under the "Theory" column of each case.

For means of comparison, we perform a simulation study of the asymptotic relative efficiency for the Hybrid estimator. Just as described in Sections 3.2 and 3.3, for each iteration we simulate an AR(1) time series of length n, which is grouped in b disjoint groups, each of length k. We start by calculating the ML estimator for the autoregressive parameter. Next we calculate the likelihood of the block means, as described in Section 3.2. To calculate the conditional likelihood for each block we need to compute the conditional covariance structure, given the block mean, as described in this Section (see 3.38). The pseudo-likelihood is the product of the conditional block likelihoods and the likelihood of the block means. We maximize numerically this function, and obtain the Hybrid estimator for the autoregressive parameter. We repeat this process 1000 times, and compute the mean value and variance of the two vectors of estimators. In Table 3.3, under the column labeled "Sim", we present the various values of the ratio between asymptotic mean squared errors for the classical MLE and the Hybrid estimator.

We note that the Hybrid estimator is statistically very efficient, when comparing it to the maximum likelihood case. We note a slight decrease in efficiency when block sizes are small, a feature similar to the Small Blocks estimator.

It is evident from tables 3.1, 3.2 and 3.3 that the Big Blocks estimator is the least efficient one from the statistical point of view, which was to be expected. However, the more surprising element is that it is not clear, at least not in this context, that the Hybrid estimator is more efficient than the Small Blocks estimator. For example,

ϕ	b=5 k=100		b=10 k=50		b=50 k=10	
	Theory	Sim	Theory	Sim	Theory	Sim
-0.750	0.99953	0.999	0.99665	0.996	0.92267	0.943
-0.250	0.99802	0.998	0.97725	0.977	0.91373	0.921
-0.010	0.99495	0.995	0.97028	0.971	0.89989	0.898
0.010	0.99457	0.995	0.97033	0.970	0.89739	0.897
0.250	0.99203	0.992	0.97385	0.972	0.91409	0.903
0.750	0.99134	0.991	0.98952	0.990	0.91800	0.922

Table 3.3: Time Series: Hybrid Asymptotic Relative Efficiency

the Hybrid estimator seems to be more efficient for larger absolute values of the true autoregressive parameter, which is to be expected, since this means that long term correlation is strong. The Small Block estimator seems to be more efficient for values of ϕ closer to 0, but the difference between the asymptotic efficiency of the two estimators is very small.

Chapter IV

Extension to Spatial Processes

4.1 Introduction

Although scientists have been preoccupied with the spatial aspect of the empirical problems (such manifestations date as early as the seventeenth century), the origins of spatial statistics as we know it are embedded in the work initiated in the mining industry conducted in South Africa, refined and made rigorous in France, by Matheron at Ecole des Mines at Fontainebleau. This is why this area of statistics is sometimes referred to by the name of *Geostatistics.* While modern environmental applications have long surpassed its initial setting, most of the early terminology is still used. Matheron and Krige are among the first scientists to develop the basic equations for optimal linear interpolation in a spatially correlated field. Other statisticians showed a constant preoccupation in analyzing, understanding, and removing, if possible, spatial dependence in agricultural fields. Work of R.A. Fisher in the 1920's, Fairfield Smith in 1930's, Papadakis (1937), Bartlett (1938, 1976, 1978) and Whittle (1954) led to the development of a somewhat different branch of spatial statistics than the work in Geostatistics. Technological advances supplied statisticians with new problems, instances in which the old classical methods are less powerful. Many of the new challenges which need new tools in order to solve them lie in areas such as environmental science, medical imaging, ecology or health effects, and are spatial in nature. Examples of such environmental problems include trend analysis of SO_2 across Eastern U.S., deriving spatial maps of SO_4^{2-} or airborne nitrogen as a function of time and meteorological variables, and mapping spatial variability of ozone between rural and urban regions.

The extension from time series to spatial processes is a very natural one. Although the two display many similarities, there are also fundamental distinctions. Perhaps the biggest difference is the type of dependence structure that can be introduced. Sequential dependence in time series describes the relationship between variables over time. For instance, a time series variable depends only on past values. In spatial processes, it is natural to allow for spatial dependence which describes the relationship between variables across some region.

In this section we give a brief introduction to the basic terminology and methods used in the greater context of spatial statistics, tools that we rely on for the derivation of the following results. Most of the definitions can be found in Smith, R.L.(2001). In particular, our goal is to describe the maximum likelihood method of estimation in this context, outlining its strengths and weaknesses, and proposing an alternative methodology to solve some of the problems that the classical method cannot.

4.2 Spatial Theory: Background

4.2.1 Spatial Models

The basic object we consider is a stochastic process $\{Z(s), s \in D\}$ where D is a subset of $\Re^d$ (d-dimensional Euclidean space), usually though not necessarily $d = 2$. For example, $Z(s)$ may represent the daily quantity of sulfuric acid measured at a specific location s. Let

$$\mu(s) = E[Z(s)], \quad s \in D,$$

denote the mean value at location s. We also assume that the variance of $Z(s)$ exists for all $s \in D$.

The process Z is said to be *Gaussian* if, for any $k \geq 1$ and locations $s_1, s_2, \ldots, s_k$, the vector $(Z(s_1), Z(s_2), \ldots, Z(s_k))$ has a multivariate normal distribution.

The process Z is said to be *strictly stationary* if the joint distribution of $(Z(s_1), Z(s_2), \ldots, Z(s_k))$ is the same as that of $(Z(s_1 + h), Z(s_2 + h), \ldots, Z(s_k + h))$ for any k spatial points $s_1, s_2, \ldots, s_k$, and any $h \in \Re^d$.

The process Z is said to be *second-order stationary* (also called *weak stationary*) if $\mu(s) \equiv \mu$ and if we denote $C(s)$ by $\mathrm{Cov}[Z(s), Z(0)]$ we have:

$$\mathrm{Cov}[Z(s_1), Z(s_2)] = C(s_1 - s_2) \quad \text{for all } s_1 \in D,\ s_2 \in D\ .$$

Note that under the assumption of finite variance, strict stationarity implies second-order stationarity, but not conversely. However, if the underlying process is Gaussian, the two definitions are equivalent.

The next useful concept that we need to introduce is the *variogram*. Assume $\mu(s)$ is a constant, which we may take without loss of generality to be 0, and then define

$$\mathrm{Var}[Z(s_1) - Z(s_2)] = 2\gamma(s_1 - s_2).$$

This makes sense only if the expression in the left hand side depends on s_1 and s_2 only through their difference $s_1 - s_2$. Such a process is called *intrinsically stationary*. The function $2\gamma(\cdot)$ is called the *variogram* and $\gamma(\cdot)$ the *semivariogram*.
Intrinsic stationarity is weaker than second order stationarity. However, if the latter holds, we have

$$\gamma(h) = C(0) - C(h).$$

We usually assume second-order stationarity though many of the results hold under the weak intrinsic stationarity assumption.

A separate concept is *isotropy*. Suppose the process is intrinsically stationary with semivariogram $\gamma(h),\ h \in \Re^d$. If $\gamma(h) = \gamma_0(||h||)$ for some function γ_0, i.e. if the semivariogram depends on its vector argument h only through its length $||h||$, then the process is *isotropic*. Isotropic processes are convenient to deal with because there are a number of widely used parametric forms for $\gamma_0(\cdot)$. Here are several examples:

1. Exponential

$$\gamma_0(t) = \begin{cases} 0 \ , & \text{if } t = 0 \ , \\ c_0 + c_1(1 - e^{-t/R}) \ , & \text{if } t > 0 \ . \end{cases}$$

2. Gaussian

$$\gamma_0(t) = \begin{cases} 0 \ , & \text{if } t = 0 \ , \\ c_0 + c_1(1 - e^{-t^2/R^2}) \ , & \text{if } t > 0 \ . \end{cases}$$

3. Matérn

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

Here $\theta_1 > 0$ is the spatial scale parameter and $\theta_2 > 0$ is a shape parameter. The symbol $\Gamma(\cdot)$ denotes the usual gamma function, while $\mathcal{K}_{\theta_2}$ is the modified Bessel function of the third kind of order θ_2 (Abramovitz and Stegun 1964, Chapter 9). Special cases include $\theta_2 = \frac{1}{2}$ which corresponds to the exponential form of semivariogram, and the limiting case $\theta_2 \to \infty$ which results in a Gaussian form.

4. Linear

$$\gamma_0(t) = \begin{cases} 0 \ , & \text{if } t = 0 \ , \\ c_0 + c_1 \, t \ , & \text{if } t > 0 \ . \end{cases}$$

Note that c_0 and c_1 are positive constants; this function tends to ∞ as $t \to \infty$ and therefore it does not correspond to a stationarity process.

5. Spherical

$$\gamma_0(t) = \begin{cases} 0 \ , & \text{if } t = 0 \ , \\ c_0 + c_1 \left\{ \frac{3}{2}\frac{t}{R} - \frac{1}{2}\left(\frac{t}{R}\right)^3 \right\} \ , & \text{if } 0 < t \le R \ , \\ c_0 + c_1 \ , & \text{if } t \ge R \ . \end{cases}$$

This is valid if $d = 1, 2$ or 3, but for higher dimensions it fails the non-positive-definiteness condition described below. It is a convenient form because it increases from a positive value c_0 when t is small, leveling off at the constant $c_0 + c_1$ at $t = R$. This is of the "nugget/range/sill" form which is often considered a realistic and interpretable form for a semivariogram.

6. Exponential-power form

$$\gamma_0(t) = \begin{cases} 0 \ , & \text{if } t = 0 \ , \\ c_0 + c_1(1 - e^{-|t/R|^p}) \ , & \text{if } t > 0 \ , \end{cases}$$

where $0 \leq p \leq 2$.

7. Rational quadratic

$$\gamma_0(t) = \begin{cases} 0 \ , & \text{if } t = 0 \ , \\ c_0 + c_1 t^2/(1 + t^2/R) \ , & \text{if } t > 0 \ . \end{cases}$$

8. Wave

$$\gamma_0(t) = \begin{cases} 0 \ , & \text{if } t = 0 \ , \\ c_0 + c_1 \left\{1 - \frac{R}{t}\sin(\frac{t}{R})\right\} \ , & \text{if } t > 0 \ . \end{cases}$$

9. Power law

$$\gamma_0(t) = \begin{cases} 0 \ , & \text{if } t = 0 \ , \\ c_0 + c_1 t^\lambda \ , & \text{if } t > 0 \ . \end{cases}$$

Non-positive definiteness requires $0 \leq \lambda < 2$.

In Figure 4.1 (courtesy Professor R.L. Smith, 2001), we graph a few isotropic semivariograms. This illustrates the kind of shapes available: (a) Linear. (b) Spherical. (c) Exponential-power, $p = 0.5$. (d) Exponential. (e) Exponential-power, $p = 1.5$. (f) Gaussian. (g) Rational quadratic. (h) Wave. (i) Power law, $\lambda = 0.5$. (j) Power law, $\lambda = 1.5$. (k)-(o) Different forms of Matérn function with θ_2 respectively 0.1, 0.5, 1, 2, 10.

In a number of the above families, the general shape of the semivariogram is quite similar. We always have $\gamma_0(0) = 0$, but γ_0 increases from a non-negative value near $t = 0$ (*the nugget*) to a limiting value (*the sill*) which is either attained at a finite value $t = R$ (*the range*), or else approached asymptotically as $t \to \infty$. In the latter there is still a scale parameter which we may denote by R, and which may be defined precisely as the value of t at which $\gamma_0(t)$ comes within a specified distance of

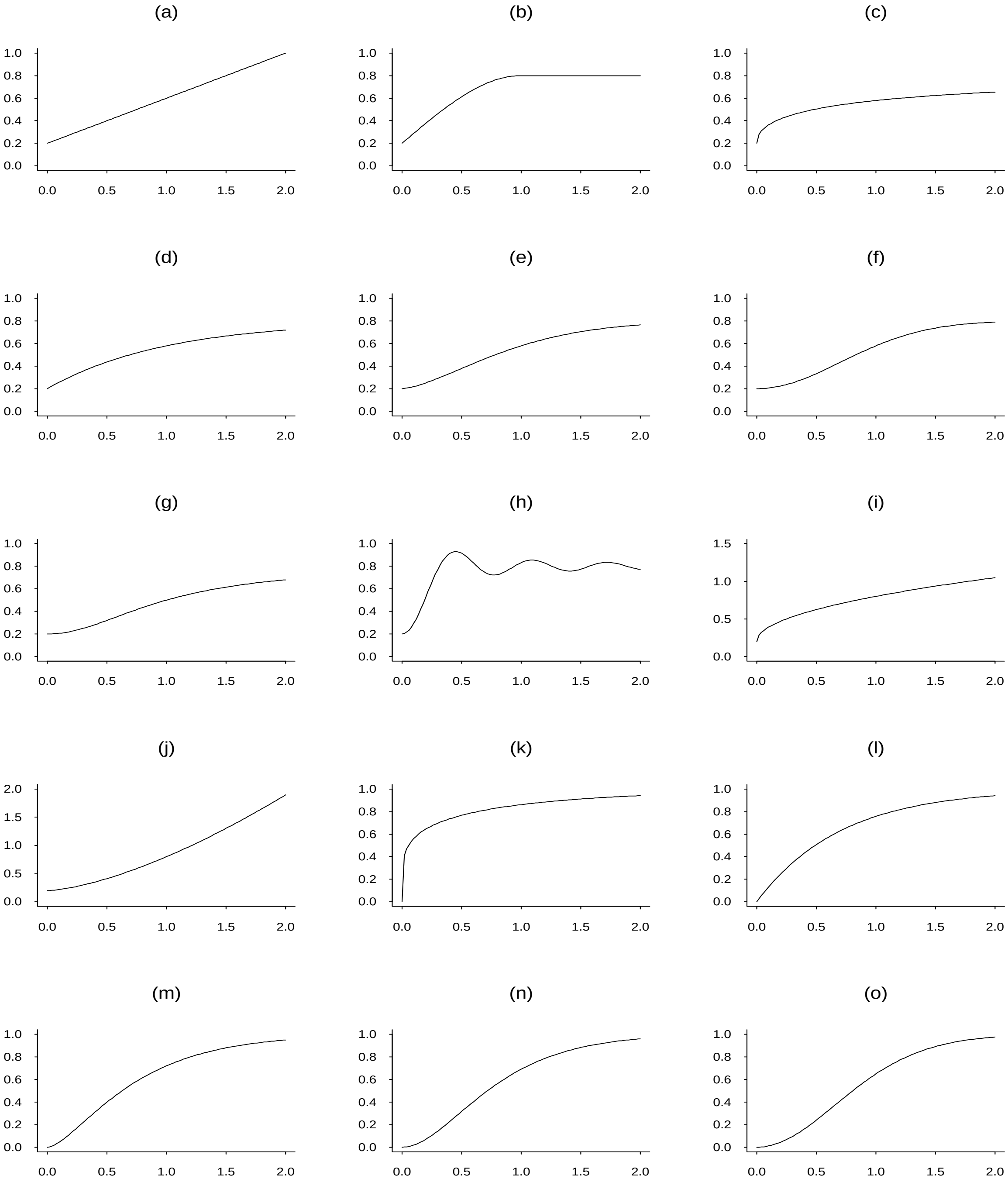

Figure 4.1: Examples of isotropic variogram functions.

its limiting value. The case where the nugget is strictly positive may appear paradoxical because it implies there is a discontinuity in the covariance function, but in fact this is a well-known feature of spatial data. There are various possible explanations, the simplest being that there is some residual white noise over and above any smooth spatial variation, or measurement error interpretation. Figure 4.2 is a plot

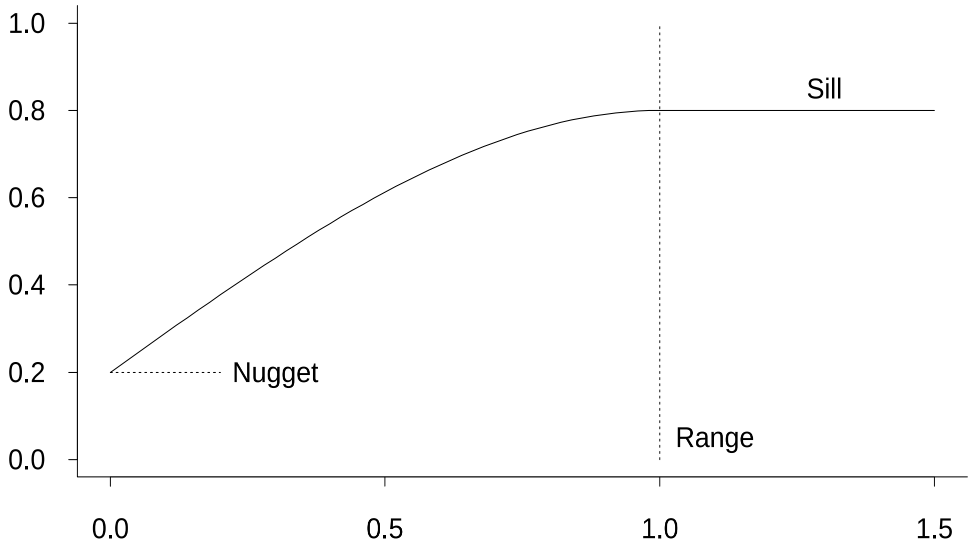

Figure 4.2: Idealized form of variogram function.

of the idealized form of the variogram function, illustrating the nugget, sill and range.

Positive definiteness

One cannot define a spatial covariance or semivariogram function in a totally arbitrary way. The key property which has to satisfy is *positive definiteness.* In the most general form where $\text{Cov}[Z(s_1), Z(s_2)] = C(s_1, s_2)$, which does not suppose any form of stationarity condition, positive definiteness means that the relation

$$\sum_i \sum_j a_i a_j C(s_i, s_j) \geq 0$$

holds for any finite set of points $s_1, \ldots, s_n$ and arbitrary real coefficients $a_1, \ldots, a_n$. There is a corresponding theory for the variogram. Suppose $\gamma(\cdot)$ is the semivariogram of a second-order stationary process; then, if $a_1, \ldots, a_n$ are constants with $\sum a_i = 0$, we have

$$\sum_i \sum_j a_i a_j \gamma(s_i - s_j) \leq 0 \ .$$

These conditions are necessary, and the fact that they are sufficient is a consequence of *Bochner's theorem.*

4.2.2 Outline of Various Estimation Techniques

So far, we have defined the main concepts of spatial covariances and variograms. We continue by considering the estimation problem. We treat this aspect under the following general setting: let $\{Z(s), s \in D\}$ be a process observed at a finite number of points $s_1, \dots, s_N$.

Sample Variogram: There are a few alternative ways of estimating the variogram. The simplest estimator is the *method of moments* (MoM) estimator, which can be defined both when the sampling points lie on a regular lattice or not. If the sampling points $s_1, \dots, s_N$ lie on a regular lattice, then the variogram is estimated by

$$2\hat{\gamma}(h) = \frac{\sum_{(s_i,s_j)\in N(h)}[Z(s_i) - Z(s_j)]^2}{|N(h)|}$$

where $N(h)$ denotes the pairs(s_i, s_j) for which $s_i - s_j = h$ and $|N(h)|$ denotes the cardinality of $N(h)$.
If the points do not lie on a regular lattice, we change the definition of $N(h)$ to

$$N(h) = \{(s_i, s_j) : \ s_i - s_j \in T(h)\} \ ,$$

$T(h)$ being some small neighborhood or tolerance region around h. Although the simplicity of this estimator is rather appealing, the main objection is that, like many methods based on sample averages, it is not robust against outlying values of Z. Another, more subtle objection, is related to the skewness of the sample distribution. Cressie and Hawkins (1980) take this into account and suggest an approximate unbiased estimator of $2\gamma(h)$:

$$2\bar{\gamma}(h) = \frac{1}{0.457 + 0.494/|N(h)|}\left\{\frac{\sum_{(s_i,s_j)\in N(h)}[Z(s_i) - Z(s_j)]^{1/2}}{|N(h)|}\right\}^4 \ .$$

Cressie(1993) goes on and suggests another alternative, a robust estimator. Meaningful information can be obtained by analyzing variogram graphs. One method is to compute these graphs as "variogram clouds": one point is plotted for each pair of stations s_i and s_j. The distance between them, say d_{ij} is plotted along the x axis, and an estimate of $\text{Var}[Z(s_i) - Z(s_j)]$ is plotted along the y axis. For the latter, we may use either the MoM or the robust method.

Parametric models: The properties of all three proposed semivariogram estimators have been extensively investigated, and it has been noted that they all lack the non-negative definiteness property. Hence the sample variogram is not acceptable as an estimator of the population variogram, since it is possible that spatial predictions derived from such estimators will appear to have negative variances. The most common way of avoiding this difficulty is to replace the empirical $\gamma(h)$ by some parametric form, such as one of the families listed in the previous section. Note that in general there is no need to restrict ourselves to isotropic models, though it is usually convenient to consider isotropic models first.

Four main methods are usually considered, least squares estimation, maximum likelihood (ML) and restricted maximum likelihood (REML), and Bayesian estimators. We describe in more detail only the ML method.

Least squares estimation

Suppose we have estimated the semivariogram $\hat{\gamma}(h)$ at a finite set of values of h and we wish to fit a model specified by the parametric function $\gamma(h;\theta)$ in terms of a finite parameter vector θ. There are three well-used versions of non-linear least squares estimators:

- *Ordinary least squares* or OLS: choose θ to minimize

$$\{\hat{\gamma} - \gamma(\theta)\}^T \{\hat{\gamma} - \gamma(\theta)\}.$$

- *Generalized least squares* or GLS: choose θ to minimize

$$\{\hat{\gamma} - \gamma(\theta)\}^T V(\theta)^{-1} \{\hat{\gamma} - \gamma(\theta)\}.$$

where $V(\theta)$ denotes the covariance matrix of $\hat{\gamma}$ which, since the problem is non-linear, depends on the unknown θ.

- *Weighted least squares* or WLS: choose θ to minimize

$$\{\hat{\gamma} - \gamma(\theta)\}^T W(\theta)^{-1} \{\hat{\gamma} - \gamma(\theta)\}.$$

where $W(\theta)$ is diagonal matrix whose diagonal entries are the variances of the entries of $\hat{\gamma}$. Thus WLS allows for the variances of $\hat{\gamma}$ but not the covariances, while GLS allows for both.

In general, we expect the three estimators OLS, WLS, GLS to be in increasing order of efficiency but in decreasing order of convenience to use. Note, in particular, that OLS can be immediately implemented by a nonlinear least squares procedure, whereas WLS and GLS require specification of the matrices $W(\theta)$ and $V(\theta)$. For example, for a Gaussian process we have the following expressions:

$$\mathrm{Var}[\{Z(s+h) - Z(s)\}^2] = 2\{2\gamma(h)\}^2$$

and

$$\begin{aligned} \mathrm{Corr} \quad & [\{Z(s_1+h_1) - Z(s_1)\}^2, \{Z(s_2+h_2) - Z(s_2)\}^2] \\ = \quad & \frac{\{\gamma(s_1 - s_2 + h_1) + \gamma(s_1 - s_2 - h_2) - \gamma(s_1 - s_2 + h_1 - h_2) - \gamma(s_1 - s_2)\}^2}{4\gamma(h_1)\gamma(h_2)} \end{aligned}$$

which may be used to evaluate the matrices $W(\theta)$ and $V(\theta)$. This GLS is possible in principle, but complicated to implement. For example, there is no guarantee that the resulting minimization problem has a unique solution.

As a compromise, Cressie (1985) proposed yet another alternative estimator, the following approximate WLS criterion: if $\hat{\gamma}$ is evaluated on a finite set $\{h_j\}$, choose θ to minimize

$$\sum_j |N(h_j)| \left\{ \frac{\hat{\gamma}(h_j)}{\gamma(h_j;\theta)} - 1 \right\}^2$$

where $N(h_j)$ is used to denote all pairs (s_i, s_j) for which $s_i - s_j = h_j$ and $|N(h_j)|$ denotes the cardinality of $N(h_j)$. This criterion is no more difficult to implement than OLS, and may be expected to be substantially more efficient, while avoiding the complications of GLS.

The idea behind the *restricted maximum likelihood* or REML estimation was originally proposed by Patterson and Thompson (1971) in connection with variance components in linear models. However, a number of authors have pointed out that this situation is essentially the same as arises with Gaussian models for spatial data: in both cases there is a linear model with correlated errors, whose covariance matrix depends on some additional parameters. Thus it is natural to try to separate the two parts of the estimation problem, the "linear model" part and the covariance structure part. Cressie (1993) is one author who has enthusiastically advocated this approach to spatial analysis.

4.3 Maximum Likelihood Estimation

4.3.1 Advantages and Disadvantages

As we note in the following technical description of this method, although the maximum likelihood estimation appears to be computationally feasible, there are still debates about its desirability when compared with some of the simpler methods mentioned above.

Mardia and Marshall (1984) considered its asymptotic properties. They concluded that the usual asymptotic consistency and normality are satisfied under a form of increasing domain asymptotics, by which we mean that the region of study is increased with the underlying density of sampling points being constant. Unfortunately, the conditions given by these authors in their paper are not necessarily easy to check, especially in the case of an irregular sampling lattice. Maybe even more problematic is that there is no indication of how large the samples need to be for asymptotic results to be reliable indicators of sampling properties.

Some authors, among which we mention Warnes and Ripley (1987) and Ripley (1988) drawn attention to the possible multimodality of the likelihood surface.

Another possible problem is sensitivity of this method to starting values. A general suggestion is to repeat the algorithm with different starting values and carefully analyze the results if any difficulties arise in the process.

The issue that we are most concerned with here is the dimensionality problem. Calculations of the likelihood based on n data points generally require $O(n^3)$ computations, which can be computationally demanding for sample sizes as large as 100 observations. We later expand on alternatives suggested by different authors to avoid this problem, together with our approach to solve it.

The theoretical advantage of maximum likelihood is that we can expect the estimates to be more efficient than the alternative methods in large samples. It is not clear at this point how big of a benefit this is. Zimmerman and Zimmerman (1991) presented a simulation study which compared a number of alternative estimators with MLE and they concluded that the MLE is only slightly superior to the approximate Weighted Least Squares(WLS) from this perspective. It has also been pointed out that the MLE procedure depends on the assumption of a Gaussian process and therefore it may perform poorly when the true distribution is non-Gaussian. This does not imply that the WLS procedure would be superior in this case. Their study does not address this issue since they restricted their analysis to Gaussian processes.

Our belief is that the potential computational complexity of maximum likelihood is outweighed by its advantages. It is a convenient, very widely applicable estimation technique, by which a variety of models can be estimated and compared using either likelihood ratio tests or automatic model selection criterion such as Akaike Information Criterion (AIC). Another advantage is that maximum likelihood methods naturally link up with Bayesian procedures.

In spite of the disadvantages mentioned at beginning of this section (which are caveats one should keep in mind when using the method), maximum likelihood is generally accepted as a valid estimation technique for spatial problems.

4.3.2 Technical Description

Assume that we are sampling from a Gaussian process. This enables us to write down the exact likelihood function and to maximize it numerically with respect to the unknown parameters.

We can incorporate deterministic linear regression terms with essentially no change in the methodology, so we consider the following model (also called the "universal kriging" model):

$$Z \sim \mathcal{N}(X\beta, \Sigma) \tag{4.1}$$

where Z an n-dimensional vector of observations, X an $n \times q$ matrix of known regressors ($q < n; X$ of full rank), β a q-vector of unknown regression parameters and Σ the covariance matrix of the observations. In many applications we may assume

$$\Sigma = \alpha V(\theta) \tag{4.2}$$

where α is an unknown parameter vector and $V(\theta)$ is a vector of standardized covariances determined by the unknown vector θ.

With Z defined by (4.1), its density is

$$(2\pi)^{-n/2}|\Sigma|^{-1/2}\exp\left\{-\frac{1}{2}(Z - X\beta)^T\Sigma^{-1}(Z - X\beta)\right\}. \tag{4.3}$$

Consequently, the negative log-likelihood is given by

$$\begin{aligned} \ell(\beta, \alpha, \theta) &= \frac{n}{2}\log(2\pi) + \frac{n}{2}\log\alpha + \frac{1}{2}\log|V(\theta)| \qquad (4.4) \\ &+ \frac{1}{2\alpha}(Z - X\beta)^T V(\theta)^{-1}(Z - X\beta). \end{aligned}$$

As a side calculation, if for a given V we define the GLS estimator of β based on the covariance matrix V as

$$\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} Z,$$

we have

$$(Z - X\hat{\beta})^T V^{-1} X = 0\ .$$

Therefore,

$$\begin{aligned}(Z - X\beta)^T V^{-1}(Z - X\beta) &= (Z - X\hat{\beta} + X\hat{\beta} - X\beta)^T V^{-1}(Z - X\hat{\beta} + X\hat{\beta} - X\beta) \\ &= (Z - X\hat{\beta})^T V^{-1}(Z - X\hat{\beta}) \\ &+ (\hat{\beta} - \beta)^T X^T V^{-1} X(\hat{\beta} - \beta) \ . \qquad (4.5)\end{aligned}$$

This confirms that this choice of β indeed minimizes the generalized sum of squares criterion (4.5) and leads to a sum of squares generalized residuals which we shall denote by

$$G^2 = (Z - X\hat{\beta})^T V^{-1}(Z - X\hat{\beta}) \ . \qquad (4.6)$$

Returning to (4.4), if we define $\hat{\beta}(\theta) = (X^T V(\theta)^{-1} X)^{-1} X^T V(\theta)^{-1} Z$ and the corresponding G^2 by $G^2(\theta)$ from (4.6) we have

$$\ell\ (\hat{\beta}(\theta), \alpha, \theta) = \frac{n}{2}\ \log(2\pi) + \frac{n}{2}\ \log\alpha + \ \frac{1}{2}\log|V(\theta)| + \frac{1}{2\alpha}\ G^2(\theta) \ . \qquad (4.7)$$

It is possible to minimize (4.7) numerically with respect to α and θ, or alternatively to minimize it analytically with respect to α defining

$$\hat{\alpha}(\theta) = \frac{G^2(\theta)}{n} \ .$$

In this case we have to minimize, with respect to θ, the function

$$\begin{aligned}\ell^*(\theta) &= \ell(\hat{\beta}(\theta), \hat{\alpha}(\theta), \theta) \\ &= \frac{n}{2}\ \log(2\pi) + \frac{n}{2}\ \log\frac{G^2(\theta)}{n} + \frac{1}{2}\ \log|V(\theta)| + \frac{n}{2} \ . \qquad (4.8)\end{aligned}$$

The quantity (4.7) or (4.8) is often called a *profile negative log-likelihood* to reflect the fact that it is computed from the negative log-likelihood (4.4) by minimizing analytically over some of the parameters. The method suggested here is essentially that first which was proposed by Kitanidis (1993) and by Mardia and Marshall (1984). To calculate (4.8), the key element is the Cholesky decomposition which allows us to write $V = L\ L^T$ where L is a lower triangular matrix. Note here that this decomposition requires a number of operations of the order $O(n^3)$. Next, write (4.1) and (4.2) in the form

$$Z = X\beta + \eta, \quad \eta \sim \mathcal{N}(0, \alpha V) \tag{4.9}$$

and define $Z^* = L^{-1}Z$, $X^* = L^{-1}X$, $\eta^* = L^{-1}\eta$, we have

$$Z^* = X^*\beta + \eta^*, \quad \eta^* \sim \mathcal{N}(0, \alpha \mathbf{I}) \tag{4.10}$$

so that the calculation of $\hat{\beta}$ reduces to an ordinary least squares problem for (Z^*, X^*). Also, the calculation of $|V|$ is straightforward because this is just $|L|^2$, and $|L|$ is just the product of diagonal entries.

This method is summarized in the following algorithm:

CLASSICAL ESTIMATION ALGORITHM

1. For the current value of θ, compute $V = V(\theta)$ and hence the Cholesky decomposition $V = L\ L^T$.
2. Calculate L^{-1} which is easy, given that L is lower triangular.
3. Calculate $|L|$, which is just the product of the diagonal entries of L. Hence $|V| = |L|^2$.
4. Compute $Z^* = L^{-1}Z$ and $X^* = L^{-1}X$.
5. Solve the ordinary least squares problem $Z^* = X^*\beta + \eta^*$ — the residual sum of squares is $G^2(\theta)$.
6. Define $l^*(\alpha, \theta)$ by (4.7) or $l^*(\theta)$ by (4.8) so that g is the function to minimize.
7. Repeat each of the steps $1 - 6$ for each θ (or each (α, θ) pair) for which g has to be evaluated. The minimum will eventually be achieved at a point $\hat{\theta}$ (or $(\hat{\alpha}, \hat{\theta})$) and this defines the maximum likelihood estimator.

8. Define H to be the Hessian matrix of second-order derivatives of g with respect to the unknown parameters, evaluated at the maximum likelihood estimators. This is also known as the observed information matrix, and in the case of a quasi-Newton algorithm such as DFPMIN routine of Press et al. 1986, may be obtained approximately from the algorithm itself. (The algorithm does not attempt to evaluate H directly, but maintains an approximation of it which is improved as the algorithm continues). In this case, in accordance with standard maximum likelihood theory, the inverse matrix H^{-1} is an approximation to the sampling covariance matrix of the parameter estimates. In particular, the square roots of the diagonal entries of H^{-1} are approximate *standard errors* of the parameter estimates. Finally, the minimized value of g may be used for *likelihood ratio tests* in comparing one model with another.

Multiple replications

The algorithm described above is designed solely for the single replication case. It can be easily extended to the multiple replication case. Suppose there are m replications denoted $Z_1, \ldots, Z_m$. The steps to be changed are the ones related to calculating the profile likelihood. First, for given θ, solve the GLS problem for the mean $\overline{Z}$, letting $G_0^2(\theta)$ be the generalized residual sum of squares. Then calculate

$$G^2(\theta) = G_0^2(\theta) + \frac{1}{m} \sum (Z_i - \overline{Z})^T \, V(\theta)^{-1} \, (Z_i - \overline{Z}).$$

Finally, substitute into the previous profile log-likelihood, multiplied by m.

4.4 Alternative Estimation Algorithm

4.4.1 Theoretical Considerations

The above example is just one of the many practical instances when computational issues combined with data sets' lack of homogeneity impede yielding efficient estimators in the exact maximum likelihood method framework. Since the number of computations to calculate the inverse and the determinant of an $n \times n$ covariance matrix is of the order n^3, we expect serious delays in getting the results for large data sets. With the growing interest in monitoring and analyzing the ozone and particulate matter over the U.S., scenarios in which data is collected at as many as 8-900 sites a few times daily, computational problems become more and more stringent.

As the exact maximum likelihood function becomes intractable in such instances, we shall consider again three alternatives to approximating the estimating function. They are the analogues of the one-dimensional methods described in Chapter III, and we refer to them as “Big Blocks”, “Small Blocks” and “Hybrid”. We shall analyze the behavior of all three estimators, and use the results obtained for one-dimensional time series as guides in helping us decide which of them is the most efficient.

All of these alternative methods are based on the idea of clustering the sampling sites in a given number of groups, say b, of approximately equal sizes, say k.

For the “Big Blocks” estimator, we first compute the cluster means and then consider their likelihood as the optimization criterion. Just as before, we expect that summarizing the entire cluster correlation in a single component, the cluster mean, to lead to a non-negligible loss in efficiency in some cases, especially for large cluster sizes.

For the “Small Block” estimator, we compute the pseudo-likelihood function as the product of individual cluster likelihoods. We assume the cluster correlation structure is known, belonging to some parametric family. The underlying assumption here is that the clusters are independent, which will induce some efficiency loss, although we expect it to be less serious than in the previous case.

To give a general idea of the computational efficiency of the “Hybrid” estimator, we describe not only the algorithm we shall follow, but also the approximate number of calculations one needs to perform in order to obtain it. This estimation technique accounts for both within and between cluster correlation, so we expect it to be superior to both abovementioned methods. We proceed as follows:

1. Calculate the cluster means and evaluate their joint likelihood. To do so, we need to compute the inverse of the $b \times b$ covariance matrix corresponding to the cluster means, each of which requires approximately k^2 steps, followed by the Cholesky decomposition of a $b \times b$ matrix, which requires $O(b^3)$ steps. If

we summarize, the number of evaluation steps required here is $O(b^2 \times k^2 + b^3)$. If $b = n^{2/3}$, this is of order $O(n^2)$, compared with $O(n^3)$ for the full likelihood calculations.

2. Conditionally on the mean of each cluster, compute the joint likelihood for each cluster. This is an $O(k^3)$ operation, which is repeated b times, hence we perform $O(b\,k^3)$ evaluations. This is of the same or smaller order than the first step if $b \geq O(n^{1/2})$.

3. Finally, compute the pseudo-likelihood function by multiplying all the above $b+1$ likelihood components. This is the function that needs to be maximized in the estimation process. This method is clearly an approximation, since we work under the assumption that clusters are independent given the block means. This assumption is not necessarily verifiable in practice, but it is nevertheless a reasonable working assumption.

Vecchia (1988) describes a general method for efficiently approximating the likelihood function. Although his approach is less refined than the technique we suggest, the two procedures share some common ideas. The central concept in his paper is to write $p(z_1, \ldots, z_n) = p(z_1)\ \prod_{j=2}^{n} p(z_j | z_1, \ldots, z_{j-1})$ where $p(z_1, \ldots, z_n)$ denotes the joint density of $(Z(s_1), \ldots, Z(s_n))$ evaluated at $(z_1, \ldots, z_n)$. Then he approximates $p(z_j | z_1, \ldots, z_{j-1})$ by the conditional density of $Z(s_j)$ given only the minimum between m and $j-1$ observations among $Z(s_1), \ldots, Z(s_{j-1})$ that are nearest to x_j (in the Euclidean distance sense), where m has to be much smaller than n. His conjecture is that the smaller the value of m, the more efficient the computations, but the worse the approximation to the true joint density. Vecchia's ordering of data points is arbitrary, and he found it to have some effect on the results. Another weakness of his method is the fact that, unlike our approach, his ignores long-range correlation (which we incorporate into the correlation of the cluster means). Therefore, we expect our methods to yield a better approximation to the likelihood than his approach.

The next section is mainly concerned with the practical aspects of this algorithm and concludes by describing a complete approach we suggest one should use in analyzing real data. Theoretical issues will be described and dealt with in Chapter V.

Practical aspects of the proposed alternative approach

The general setting for this problem is the same as in the previous case. We assume that we are sampling from a Gaussian process, and we continue to incorporate the deterministic linear regression terms. Thus, the model under consideration is:

$$Z \sim \mathcal{N}(X\beta, \Sigma)$$

with Z a n-dimensional vector of observations, X an $n \times q$ matrix of known regressors ($q < n; X$ of full rank), β a q-vector of unknown regression parameters and Σ the covariance matrix of the observations, $\Sigma = (\sigma_{ij})_{1 \leq i,\, j \leq n}$. As before, we assume

$$\Sigma = \alpha V(\theta) \tag{4.11}$$

where α is an unknown parameter vector and $V(\theta)$ is a vector of standardized covariances determined by the unknown vector θ.

As mentioned before, we suggest clustering the sample points into a given number of blocks, say b. We perform a classical clustering procedure (such as Ward's method which minimizes the within cluster sum of squares, to which we add a supplementary constraint so that we obtain clusters with similar number of sites). The clustering is performed according to the latitude and longitude of each sample location. Denote by b the number of clusters and by k the cluster size. For simplicity and ease of future reference, let us assume that once the clustering step is completed, we order spatial observations according to the cluster to which they belong to. In other words, the observation Z_l is the i-th observation in the j-th cluster [1], where $i = \lceil \frac{l}{b} \rceil$ and $j = l - k\ \lfloor \frac{l}{b} \rfloor$ (or, equivalently, the i-th cluster consists of the observations $\{Z_{(i-1)k+1}, \ldots, Z_{ik}\}$). This section is mainly concerned with the technical details arising in the implementation of the Hybrid estimator, as it is the most complex one. However, due to its nature, we have to completely describe the construction

[1] By $\lceil x \rceil$ we mean the smallest integer greater than or equal to x, and by $\lfloor x \rfloor$ we mean the largest integer smaller than or equal to x.

of the Big Blocks pseudo-likelihood as well, which would be maximized if one was interested in finding the corresponding estimators.

Big Blocks (or the block-means analysis): We start by evaluating the joint likelihood of the block means. To do so, we need to compute the mean and covariance matrix of the process.

Define Z^* as the vector of cluster means, i.e. $Z^* = \{Z_1^*, \ldots, Z_b^*\}^T$ where by Z_i^* denotes the mean of cluster i, i.e. $Z_i^* = \frac{1}{k}\sum_{j=1}^k Z_{(i-1)k+j}$. We assume that the process Z^* is Gaussian, with mean μ^* and covariance matrix Σ^*.

Thus the joint density of the cluster means is of the form:

$$(2\pi)^{-b/2}\,|\Sigma^*|^{-1/2}\,\exp\left\{-\frac{1}{2}\,(Z^*-\mu^*)^T\Sigma^{*-1}(Z^*-\mu^*)\right\}$$

and hence the negative log-likelihood for the cluster means becomes:

$$\begin{aligned}\ell_{means}(\beta,\theta) &= \frac{b}{2}\,\log(2\pi) + \frac{1}{2}\,\log|\Sigma^*(\theta)| \\ &+ \frac{1}{2}(Z^*-\mu^*)^T\,\Sigma^*(\theta)^{-1}\,(Z^*-\mu^*)\,.\end{aligned}$$

In order to be able to proceed to the maximization stage, we need to express μ^* and Σ^* as functions of the original quantities. Note that

$$\begin{aligned}\mu_i^* &= E[Z_i^*] = \frac{1}{k}\sum_{j=1}^k E[\,Z_{(i-1)k+j}\,] = \frac{1}{k}\sum_{j=1}^k\sum_{r=1}^q x_{(i-1)k+j,r}\,\beta_r \\ &= \frac{1}{k}\sum_{j=1}^k X_{(i-1)k+j}\,\beta = \beta\,\frac{\sum_{j=1}^k X_{(i-1)k+j}}{k} = X_i^*\,\beta, \quad \text{for } 1 \le i \le b \qquad (4.12)\end{aligned}$$

and let μ^* be the vector mean, $\mu^* = \{\mu_1^*, \ldots, \mu_b^*\}^T$.

Next we compute the covariance matrix of the cluster means process. Hence for any i and j, compute

$$\begin{aligned}\sigma^*_{ij} &= \mathrm{Cov}[Z^*_i, Z^*_j] = \sum_{l=1}^{k}\sum_{l'=1}^{k} \mathrm{Cov}[Z_{(i-1)k+l},\ Z_{(j-1)k+l'}] \\ &= \frac{1}{k^2}\sum_{l=1}^{k}\sum_{l'=1}^{k} \sigma_{(i-1)k+l,\,(j-1)k+l'}\end{aligned}$$

and define $\Sigma^* = (\sigma^*_{ij})_{1\le i,j\le b}$.

Block-conditional analysis: The second step in constructing the Hybrid pseudo-likelihood function is calculating the conditional block likelihoods given their mean.

To this end, for each cluster i consider the joint density of $Z_{(i-1)k+1}, \ldots, Z_{(i-1)k+k-1}$ and Z^*_i, $1 \le i \le b$. It follows that the vector $(Z_{(i-1)k+1}, \ldots, Z_{ik-1}, Z^*_i)^T$ is normally distributed with vector mean $(\mu_{i\_}, \mu^*_i)$ and covariance matrix $\begin{pmatrix} \Sigma_{i\_} & \tau_{i\_} \\ \tau^T_{i\_} & \sigma^*_{ii} \end{pmatrix}$ where

- $\mu^*_i = X^*_i\ \beta$ as given by (4.12)
- $\mu^T_{i\_} = \{\mu_{(i-1)k+1}, \ldots, \mu_{ik-1}\} = X_{i\_}\beta$,
- $\Sigma_{i\_} = (\sigma_{(i-1)k+j,\,(i-1)k+j'})_{1\le j,j'\le k-1}$ where $\sigma_{(i-1)k+j,\,(i-1)k+j'}$ are elements of the original variance-covariance matrix,
- $\sigma^*_{ii} = \mathrm{Var}[Z^*_i] = \frac{1}{k^2}\sum_{l=1}^{k}\sum_{l'=1}^{k}\sigma_{(i-1)k+l\ ,\,(i-1)k+l'}$ as defined for the Big Blocks case, and
- $\tau_{i\_} = \{\tau_{(i-1)k+1}, \ldots, \tau_{ik-1}\}$ where for all $1 \le j \le k-1$ we have

$$\tau_{(i-1)k+j} = \mathrm{Cov}[Z^*_i, Z_{(i-1)k+j}] = \frac{1}{k}\sum_{j'=1}^{k}\sigma_{(i-1)k+j',\,(i-1)k+j}$$

From the theory of multivariate normal distributions, we obtain the joint density of $Z_{(i-1)k+1}, \ldots, Z_{ik-1}$ given Z^*_i to be

$$\mathcal{N}\left(\mu_{i\_} + \frac{\tau^T_{i\_}}{\sigma^*_{ii}}(Z^*_i - \mu^*_i),\ \ \Sigma_{i\_} - \frac{\tau_{i\_}\tau^T_{i\_}}{\sigma^*_{ii}}\right)\quad .$$

For brevity's sake, we denote

$$\begin{aligned}\mu^{c_i} &\equiv \mu_{i\_} + \frac{\tau_{i\_}^T}{\sigma_{ii}^*}(Z_i^* - \mu_i^*) \ , \\ \Sigma^{c_i} &\equiv \Sigma_{i\_} - \frac{\tau_{i\_}\tau_{i\_}^T}{\sigma_{ii}^*} = (\sigma_{jj'}^{c_i})_{1\le j,\, j'\le k-1} \ , \\ \eta_i &\equiv \frac{\tau_{i\_}^T}{\sigma_{ii}^*} \ .\end{aligned}$$

It is immediate that for any $1 \le i \le b$ and any $1 \le j,\ j' \le k-1$ we have

$$\sigma_{j,j'}^{c_i} = \sigma_{(i-1)k+j,\,(i-1)k+j'} - \frac{\tau_{(i-1)k+j}\tau_{(i-1)k+j'}}{\sigma_{ii}^*} \quad \text{and}$$

$$\mu^{c_i} = X_{i\_}\beta + \eta_i(Z_i^* - X_i^*\beta) = \eta_i \ Z_i^* + \left(X_{i\_} - \eta_i \ X_i^*\right)\beta \ \ .$$

Therefore, we write

$$Z_{i\_} - \mu^{c_i} = (Z_{i\_} - \eta_i \ Z_i^*) - \left(X_{i\_} - \eta_i \ X_i^*\right)\beta$$

and define

$$\begin{aligned}Z^{c_i} &\equiv Z_{i\_} - \eta_i \ Z_i^*, \\ X^{c_i} &\equiv X_{i\_} - \eta_i \ X_i^* \ .\end{aligned} \tag{4.13}$$

Hence the conditional log-likelihood for each cluster, given the cluster mean, is of the form:

$$\begin{aligned}\ell_{c_i}(\beta,\theta) &= \frac{k-1}{2}\ \log(2\pi) + \frac{1}{2}\log|\ \Sigma^{c_i}(\theta)| \\ &+ \ \frac{1}{2}\,(Z^{c_i} - X^{c_i}\beta)^T\ {\Sigma^{c_i}}^{-1}(\theta)\ (Z^{c_i} - X^{c_i}\beta) \ .\end{aligned} \tag{4.14}$$

Finally, we multiply all the aforecomputed individual likelihoods, and obtain the approximate likelihood function (or, equivalently, we sum the $b{+}1$ individual negative log likelihoods). Thus, from (4.12) and (4.14) it follows that

$$\ell_{full}(\beta,\theta) = \ell_{means}(\beta,\theta) + \sum_{i=1}^{b} \ell_{c_i}(\beta,\theta)$$

or, more explicitly, we obtain

$$\begin{aligned} \ell_{full} &= \frac{1}{2}\Bigg[b\,k\ \log(2\pi) + \log|\Sigma^*(\theta)| + \sum_{i=1}^{b} \log|\Sigma^{c_i}(\theta)| \\ &+ (Z^* - X^*\beta)^T\ \Sigma^{*-1}(\theta)\ (Z^* - X^*\beta) \\ &+ \sum_{i=1}^{b} (Z^{c_i} - X^{c_i}\beta)^T\ \Sigma^{c_i-1}(\theta)\ (Z^{c_i} - X^{c_i}\beta) \Bigg]\ . \end{aligned} \tag{4.15}$$

In the original algorithm, we first estimate the regression parameters β through generalized least squares, using the covariance matrix Σ. Since this process involves computing the inverse and the determinant of the full matrix, which is prohibitive for large data sets, we propose here the use an approximation to the covariance matrix, call it $\tilde{\Sigma}$.

Following the conditional independence assumption, we consider $\tilde{\Sigma}$ to be given by:

$$\tilde{\Sigma} = \begin{pmatrix} \Sigma^* & 0 & \dots & 0 \\ 0 & \Sigma^{c_1} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \Sigma^{c_b} \end{pmatrix} \ . \tag{4.16}$$

If in a similar manner we denote by

$$\tilde{X} = \begin{pmatrix} X_1^* \\ \vdots \\ X_b^* \\ X_1^{c_1} \\ \vdots \\ X_{k-1}^{c_1} \\ X_1^{c_2} \\ \vdots \\ X_{k-1}^{c_2} \\ \vdots \\ \vdots \\ X_{k-1}^{c_b} \end{pmatrix} \quad \text{and} \quad \tilde{Z} = \begin{pmatrix} Z_1^* \\ \vdots \\ Z_b^* \\ Z_1^{c_1} \\ \vdots \\ Z_{k-1}^{c_1} \\ Z_1^{c_2} \\ \vdots \\ Z_{k-1}^{c_2} \\ \vdots \\ \vdots \\ Z_{k-1}^{c_b} \end{pmatrix}$$

then an estimator for β is given below:

$$\tilde{\beta} = (\tilde{X}^T \tilde{\Sigma}^{-1} \tilde{X})^{-1} \tilde{X}^T \tilde{\Sigma}^{-1} \tilde{Z} \tag{4.17}$$

The obvious advantage of this approach is that the inverse of the approximate covariance matrix, $\tilde{\Sigma}^{-1}$ is the inverse of a block diagonal matrix. Thus,

$$\tilde{\Sigma}^{-1} = \begin{pmatrix} \Sigma^{*^{-1}} & 0 & \dots & 0 \\ 0 & \Sigma^{c_1^{-1}} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \Sigma^{c_b^{-1}} \end{pmatrix} .$$

Going back to the form of the pseudo-likelihood function in (4.15) and substituting β by $\tilde{\beta}$, we obtain the function of α and θ to be minimized through an iterative numerical procedure.

Finally, the steps described in the original case should be followed with several changes, as described below:

MODIFIED ALGORITHM

1. Note first that we can assume, without any loss of generality, that $\Sigma = \alpha\, V(\theta)$ and hence $\Sigma^* = \alpha\, V^*(\theta)$. For the current value of θ, compute $V^* = V^*(\theta)$ and $V_{c_i} = V_{c_i}(\theta)$. Next, perform the Cholesky decomposition $V^* = L^* L^{*T}$ and $V_{c_i} = L_{c_i} L_{c_i}^T$ for all $i = \overline{1, b}$.

2. Calculate L^{*-1} and $L_{c_i}^{-1}$ for all $i = \overline{1, b}$ (which is straightforward to do, since they are all lower triangular matrices).

3. Calculate $|L^*|$ and $|L_{c_i}|$ which are simply the product of the diagonal entries of L^* and L_{c_i} respectively, for all $i = \overline{1, b}$.
   Thus $|V^*| = |L^*|^2$ and $|V_{c_i}| = |L_{c_i}|^2$, for all $i = \overline{1, b}$.

4. Compute $Z^{**} = L^{*-1} Z^*$ and $X^{**} = L^{*-1} X^*$. Also compute $Z^{c_i **} = L_{c_i}^{-1}\, Z^{c_i}$ and $X_i^{c**} = L_{c_i}^{-1}\, X^{c_i}$ where for all $i = \overline{1, b}$, Z^{c_i} and X^{c_i} are given by (4.13).

5. Solve the approximate ordinary least squares problem $\tilde{Z} = \tilde{X}\beta + \tilde{\lambda}$ leading to $\tilde{\beta}$ the estimator of β as described in (4.17).

6. Define the profile negative log-likelihood as $\ell_{full}(\theta)$, (the function to be minimized), given by (4.15)

7. Repeat each of the steps $1 - 6$ for each θ for which g has to be evaluated. The minimum will eventually be achieved at a point $\hat{\theta}$ which defines the Hybrid estimator.

Small Blocks: This is the simplest of the three methods to implement, and there are no technical details that are worth describing in detail. The general idea is that the criterion function here is the product of block likelihoods, as the underlying assumption is block independence:

$$\ell_{full} = \prod_{j=1}^{b} \ell_j = \frac{1}{(2\pi)^{kb/2} |\Sigma_j|^{b/2}} \exp\left(-\frac{1}{2} \sum_{j=1}^{b} (Z_j - X_j\ \beta)^T \Sigma_j^{-1} (Z_j - X_j\ \beta) \right) . \tag{4.18}$$

The problem of the least squares estimation is approximated in a similar fashion as for the hybrid estimator, with the exception that the approximate covariance matrix, $\tilde{\Sigma}$ is just the block-diagonal matrix:

$$\tilde{\Sigma} = \begin{pmatrix} \Sigma_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \Sigma_b \end{pmatrix} \tag{4.19}$$

where Σ_j is the covariance matrix for the j^{th} group. Also, in this case, $\tilde{X} = X$ and $\tilde{Z} = Z$.

4.4.2 Simulation Study

It is clear from the previous section that theoretical derivation of the asymptotic efficiency of the alternative estimators are very involved. Even in the simpler case of AR(1) time series these computations are rather complicated, and we have seen that only in one case, the Small Blocks estimator, we have been able to derive a closed form expression for the asymptotic efficiency. Therefore, we will not attempt to follow the "Expansion Method" here, but rather analyze the asymptotic performance of the alternative estimators through a simulation study.

Let the model be the one specified in (4.1),

$$Z \sim \mathcal{N}(X\beta, \Sigma)$$

where Z an n-dimensional vector of observations, X an $n \times q$ matrix of known regressors ($q < n; X$ of full rank), β a q-vector of unknown regression parameters and Σ the covariance matrix of the observations.

For simulation purposes, we first create the location matrix, and in for this exercise we consider a 20 by 20 grid, equally spaced, with the (Euclidean) distance between any two locations equal to 2 units.

The next step is to choose the structure of the true covariance matrix, and we assume that it belongs to the *exponential family.* In other words, the structure of the spatial covariance matrix is of the form: $\Sigma = \alpha \, V(R)$, where α is a scaling parameter and R is the *Range.* If we denote by d_{ij} the Euclidean distance between any two locations i and j, then the matrix V is of the form:

$$v_{ij} = \begin{cases} 1 \ , & \text{if } i = j \ , \\ \exp(-d_{ij}/R) \ , & \text{if } i \neq j \ . \end{cases} \tag{4.20}$$

In order to decide what true value of the parameter α we should choose, we note that, since it is a scaling parameter, its magnitude should be irrelevant to the estimation process. Thus, throughout this simulation study, we consider only the case when $\alpha = 3$.

As the *Range* parameter is driving the spatial structure of the model, we expect that our simulation study will lead to different results according to the magnitude of R. Hence we consider here two cases, $R = 1$ and $R = 3$.

As noted earlier, the estimation process is a two-step procedure, the first one being analytical estimation of the regression parameters. In the classical case, this is just the generalized least squares technique using the full covariance matrix, but for the alternative methods, we are using the approximate covariance matrix given by (4.16). There might be some loss of efficiency in the estimation of the spatial parameters due to this approximation. Therefore, we consider two different cases for the simulation study, one including no regression terms, and the other including the intercept, longitude and latitude in the design matrix.

Another issue to be studied through this simulation exercise is how the number of blocks influences the relative efficiency of the proposed estimators. Recall that the data set consists of 400 locations, and we consider the special cases of 4,8,10,25,40,50 and 100 blocks.

For each simulation, we keep the grid of locations fixed, (therefore the design matrix X is completely specified) and construct the covariance and design matrix, i.e. Σ. Our goal is to generate the vector of observations Z, and to accomplish this we follow the classical technique employed in such cases. First perform a Cholesky decomposition, i.e. $\Sigma = L\ L^T$, where L is a lower triangular matrix. Therefore, it follows that the inverse matrix $\Sigma^{-1} = L^{T^{-1}}\ L^{-1}$. Next, we generate $Y \sim \mathcal{N}(0, \mathbf{I})$, where by $\mathbf{I}$ we understand the n by n identity matrix. As a consequence, $Z = X\ \beta + L\ Y$ follows the normal distribution, as described by model (4.1).

Since Z, X and Σ are completely specified at this point, we evaluate the exact likelihood function, maximize it and find the maximum likelihood estimators for α and R, say α_{MLE} and R_{MLE}.

Next step is to calculate the alternative estimators for a given number of blocks, say b. Employ the same data set generated for the classical case, and cluster it in b blocks; evaluate the pseudo-likelihood functions for the Small Blocks and Hybrid, as described in the previous section and maximize the criterion functions which yield the alternative estimators, say α_{SB} and R_{SB} for the Small Blocks, and α_{HYB} and R_{HYB} for the Hybrid. Repeat this process a large number of times (in this study we performed 4000 simulations per scenario.) As a final step we compute the average value of the resulting arrays of estimators for each case (MLE, SB, HYB), the bias and the variance, leading to the mean squared error. We conclude by evaluating the ratio between the mean squared error of the classical ML estimator and the mean squared error of the alternative estimators. The results are presented in Tables 4.1 through 4.4. Note that, based on our conclusions for the time series problem, we expect the Big Blocks estimator not to be very efficient, therefore we do not include this method in the simulation study.

PAR	MET	b=4	b=8	b=10	b=25	b=40	b=50	b=100
$\alpha = 3$	SB	1.000	0.997	0.997	0.997	0.996	0.998	0.997
	HYB	1.000	0.997	0.997	0.997	0.995	0.997	0.996
$R = 1$	SB	0.949	0.887	0.896	0.797	0.691	0.658	0.524
	HYB	0.949	0.892	0.897	0.808	0.720	0.702	0.618

Table 4.1: Small Blocks and Hybrid Relative Efficiency, R=1, No Regression

PAR	MET	b=4	b=8	b=10	b=25	b=40	b=50	b=100
$\alpha = 3$	SB	0.969	0.953	0.945	0.940	0.948	0.954	0.961
	HYB	0.967	0.943	0.929	0.918	0.895	0.915	0.959
$R = 3$	SB	0.955	0.911	0.910	0.878	0.848	0.832	0.776
	HYB	0.952	0.896	0.885	0.839	0.787	0.779	0.796

Table 4.2: Small Blocks and Hybrid relative efficiency, R=3, No Regression

PAR	MET	b=4	b=8	b=10	b=25	b=40	b=50	b=100
$\alpha = 3$	SB	1.000	1.000	0.999	0.999	0.999	0.999	0.999
	HYB	1.000	1.000	1.000	1.000	0.999	1.000	0.999
$R = 1$	SB	0.945	0.869	0.883	0.774	0.721	0.663	0.538
	HYB	0.941	0.857	0.864	0.751	0.691	0.644	0.559

Table 4.3: Small Blocks and Hybrid relative efficiency, R=1, With Regression

PAR	MET	b=4	b=8	b=10	b=25	b=40	b=50	b=100
$\alpha = 3$	SB	0.980	0.944	0.928	0.919	0.910	0.914	0.906
	HYB	0.985	0.961	0.949	0.952	0.944	0.958	0.957
$R = 3$	SB	0.966	0.929	0.926	0.909	0.853	0.861	0.788
	HYB	0.962	0.912	0.901	0.872	0.815	0.813	0.778

Table 4.4: Small Blocks and Hybrid relative efficiency, R=3, With Regression

From Tables 4.1 through 4.4 we note that both methods lead to relatively efficient estimators for most cases. Note that the estimator for the scaling parameter α is more efficiently estimated than the one for the *Range*. Block size does not appear to influence the efficiency when estimating α. However, the efficiency for the estimators for the *Range* is more sensitive to the size of the groups, smaller cluster sizes leading to less efficient estimators for the *Range*. The *Range* parameter seems to be more efficiently estimated when the true value is equal to 3, rather than when it is equal to 1.

Comparing Tables 4.1 and 4.2 with Tables 4.3 and 4.4 we note that inclusion of location columns in the design matrix does not affect the efficiency, therefore the approximations we used to compute the estimator for the regression parameters did not induce much loss in efficiency in this cases.

For a better visualization of the simulation results, we present here two plots. Figure 4.3 displays the resulting estimators from 4000 simulations, for 10 blocks, $\alpha = 3$ and $R = 1$, for each of the three methods (MLE, Small Blocks and Hybrid) while Figure 4.4 displays the resulting estimators from 4000 simulations, for 100 blocks, $\alpha = 3$ and $R = 1$ (again, for each of the three methods). We note from these two plots that both methods perform well compared to the classical MLE, for each of the two parameters, somewhat closer to the MLE when estimating α.

4.4.3 Simulation Error

The results from this simulation study suggest that the difference in relative efficiency between the two methods is not very large. The natural question is how much of the difference is due to simulation error. We address this issue here.

Our strategy is to construct some confidence intervals for the asymptotic relative efficiencies for the two methods, in order to asses whether and when one method is more efficient than the other. To this end, we calculate the ratio between the asymptotic mean squared errors of the Small Blocks and Hybrid estimators as well. The methodology we employ here is a simple bootstrap procedure. As a result of the simulation study, we have three vectors of 4000 estimators (independent within

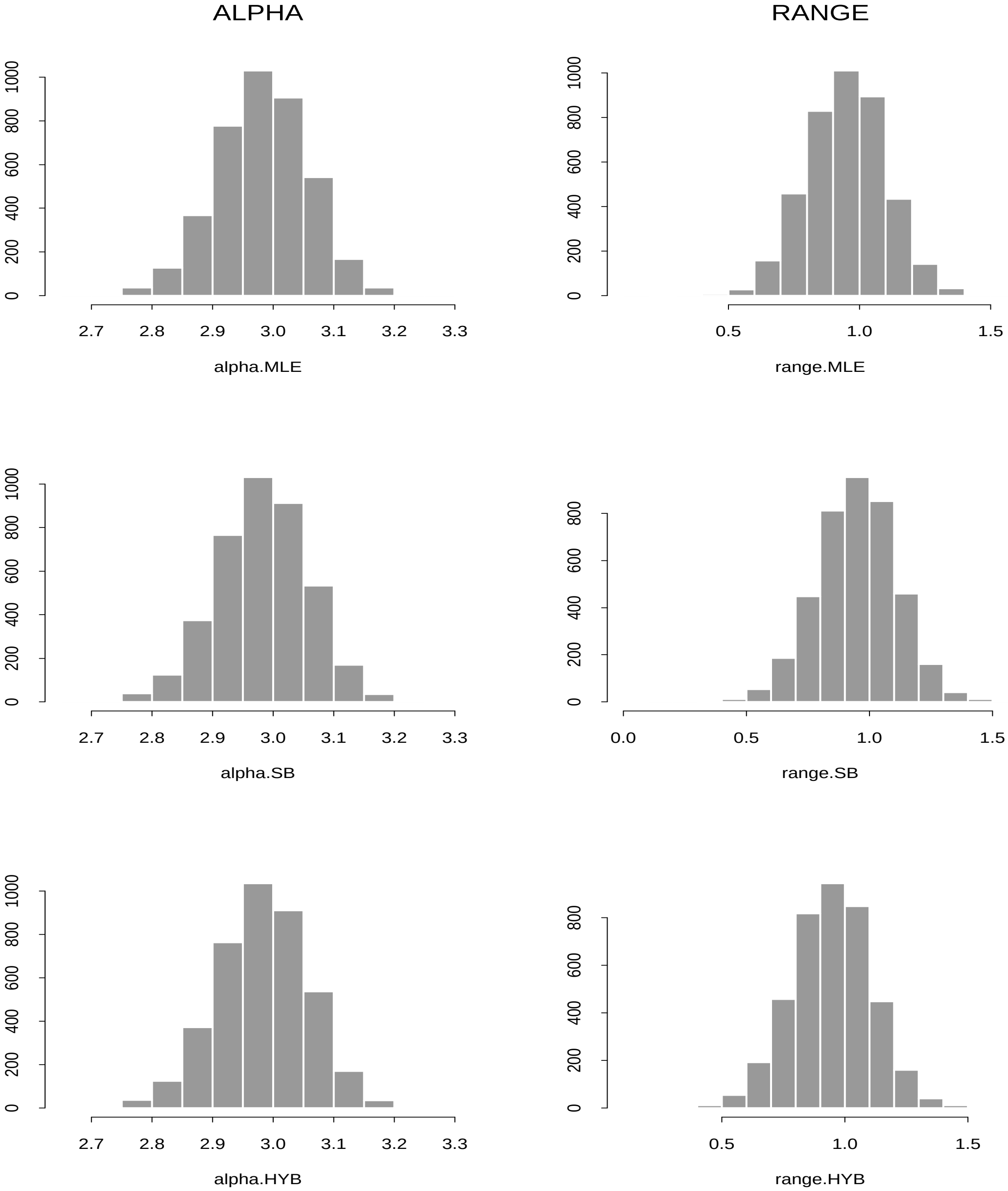

Figure 4.3: 4000 Estimators for 10 blocks, $R = 1$

the vector). From each vector, randomly select a sample of 4000 observations (with replacement), calculate the mean squared error of the resample, and compute the ratio between the mean squared errors for each vector. Repeat this process 10,000 times. The last step is to order these ratios for each scenario. For a 95% bootstrap confidence interval, using the percentile method, we find the values that cut off the

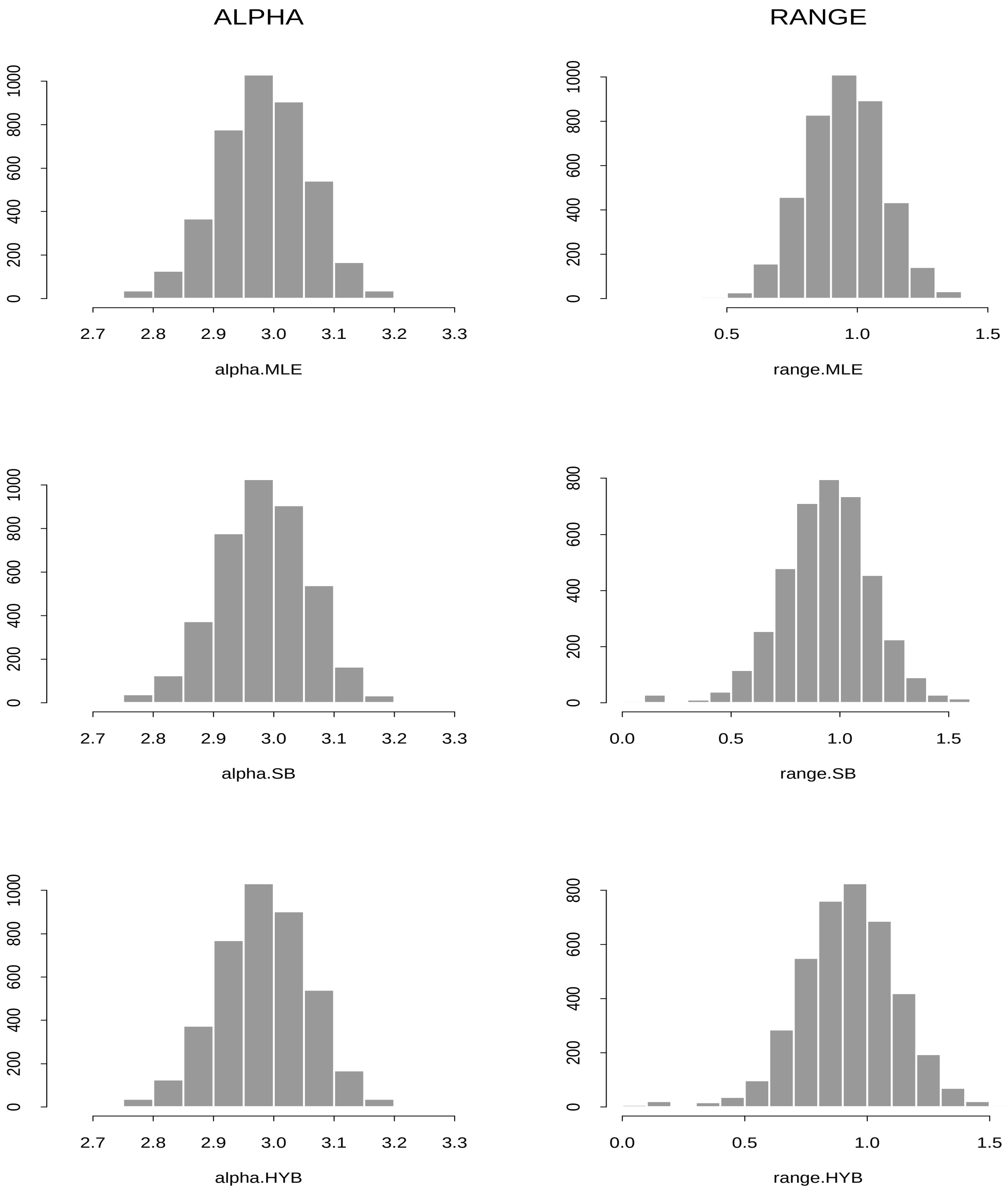

Figure 4.4: 4000 Estimators for 100 blocks, $R = 1$

lower 2.5% and the upper 2.5% of our data. Results for each of the cases are shown in Tables 4.5, 4.6, 4.7 and 4.8. For each of the cases, we present the asymptotic relative efficiency as well as the 95% bootstrap confidence intervals.

These results indicate, as we have already noted, that the difference in the efficiency between the two methods is very small. We can conclude that for these

	Method	$\alpha = 3$	$R = 1$
4 blocks	SB vs. MLE	1.000 (0.937 , 1.066)	0.949 (0.886 , 1.019)
	HYB vs. MLE	1.000 (0.936 , 1.065)	0.949 (0.887 , 1.020)
	HYB vs. SB	1.000 (0.939 , 1.069)	1.000 (0.932 , 1.069)
8 blocks	SB vs. MLE	0.997 (0.934 , 1.062)	0.887 (0.827 , 0.951)
	HYB vs. MLE	0.997 (0.937 , 1.061)	0.892 (0.832 , 0.955)
	HYB vs. SB	1.000 (0.940 , 1.066)	1.006 (0.936 , 1.080)
10 blocks	SB vs. MLE	0.997 (0.933 , 1.063)	0.896 (0.838 , 0.959)
	HYB vs. MLE	0.997 (0.936 , 1.064)	0.897 (0.838 , 0.960)
	HYB vs. SB	1.000 (0.938 , 1.068)	1.002 (0.937 , 1.072)
25 blocks	SB vs. MLE	0.997 (0.933 , 1.062)	0.797 (0.739 , 0.859)
	HYB vs. MLE	0.997 (0.936 , 1.063)	0.808 (0.751 , 0.867)
	HYB vs. SB	1.000 (0.940 , 1.067)	1.014 (0.940 , 1.095)
40 blocks	SB vs. MLE	0.996 (0.933 , 1.062)	0.691 (0.642 , 0.742)
	HYB vs. MLE	0.995 (0.933 , 1.058)	0.720 (0.670 , 0.774)
	HYB vs. SB	0.999 (0.938 , 1.063)	1.042 (0.963 , 1.125)
50 blocks	SB vs. MLE	0.998 (0.937 , 1.064)	0.658 (0.611 , 0.708)
	HYB vs. MLE	0.997 (0.935 , 1.061)	0.702 (0.654 , 0.751)
	HYB vs. SB	0.999 (0.937 , 1.064)	1.066 (0.991 , 1.154)
100 blocks	SB vs. MLE	0.997 (0.936 , 1.063)	0.524 (0.486 , 0.563)
	HYB vs. MLE	0.996 (0.933 , 1.061)	0.618 (0.574 , 0.664)
	HYB vs. SB	0.999 (0.938 , 1.064)	1.179 (1.087 , 1.276)

Table 4.5: Bootstrap Confidence Intervals for $\alpha = 3$, $R = 1$, Without Regression

examples the two methods yield to estimators that are comparable in efficiency. There are some instances, as in Table 4.5, 100 blocks, where there is a clear ordering of the methods, despite the simulation error. In this case, for example, it is clear that the Hybrid estimator for the *Range* is asymptotically more efficient than the Small Blocks estimator. However, the general conclusion is that it is not possible to decide which of the two methods leads to more efficient estimators, in part due to the error induced through simulation.

Another conclusion reinforced by the bootstrap analysis is that the scaling parameter is more efficiently estimated than the *Range*. The other previous conclusion that is confirmed is that the estimator for the *Range* seems to be more efficient when block sizes are large.

	Method	$\alpha = 3$	$R = 3$
4 blocks	SB vs. MLE	0.969 (0.909 , 1.035)	0.955 (0.889 , 1.026)
	HYB vs. MLE	0.967 (0.910 , 1.030)	0.952 (0.885 , 1.024)
	HYB vs. SB	0.998 (0.937 , 1.063)	0.997 (0.927 , 1.074)
8 blocks	SB vs. MLE	0.953 (0.896 , 1.014)	0.911 (0.847 , 0.980)
	HYB vs. MLE	0.943 (0.884 , 1.005)	0.896 (0.835 , 0.964)
	HYB vs. SB	0.990 (0.927 , 1.054)	0.984 (0.913 , 1.061)
10 blocks	SB vs. MLE	0.945 (0.888 , 1.008)	0.910 (0.846 , 0.978)
	HYB vs. MLE	0.929 (0.871 , 0.990)	0.885 (0.821 , 0.952)
	HYB vs. SB	0.984 (0.926 , 1.046)	0.973 (0.901 , 1.049)
25 blocks	SB vs. MLE	0.940 (0.883 , 1.001)	0.878 (0.816 , 0.946)
	HYB vs. MLE	0.918 (0.860 , 0.979)	0.839 (0.778 , 0.905)
	HYB vs. SB	0.976 (0.915 , 1.040)	0.956 (0.885 , 1.032)
40 blocks	SB vs. MLE	0.948 (0.891 , 1.011)	0.848 (0.791 , 0.912)
	HYB vs. MLE	0.895 (0.839 , 0.955)	0.787 (0.730 , 0.850)
	HYB vs. SB	0.945 (0.885 , 1.007)	0.928 (0.860 , 1.003)
50 blocks	SB vs. MLE	0.954 (0.896 , 1.017)	0.832 (0.771 , 0.894)
	HYB vs. MLE	0.915 (0.859 , 0.976)	0.779 (0.722 , 0.839)
	HYB vs. SB	0.960 (0.899 , 1.021)	0.936 (0.865 , 1.012)
100 blocks	SB vs. MLE	0.961 (0.901 , 1.020)	0.774 (0.719 , 0.834)
	HYB vs. MLE	0.959 (0.900 , 1.022)	0.796 (0.739 , 0.856)
	HYB vs. SB	0.998 (0.937 , 1.063)	1.029 (0.955 , 1.108)

Table 4.6: Bootstrap Confidence Intervals for $\alpha = 3$, $R = 3$, Without Regression

	Method	$\alpha = 3$	$R = 1$
4 blocks	SB vs. MLE	1.000 (0.936 , 1.069)	0.945 (0.875 , 1.018)
	HYB vs. MLE	1.000 (0.933 , 1.068)	0.941 (0.873 , 1.014)
	HYB vs. SB	1.000 (0.936 , 1.068)	0.996 (0.923 , 1.074)
8 blocks	SB vs. MLE	1.000 (0.936 , 1.067)	0.869 (0.803 , 0.942)
	HYB vs. MLE	1.000 (0.938 , 1.066)	0.857 (0.792 , 0.930)
	HYB vs. SB	1.000 (0.939 , 1.068)	0.986 (0.904 , 1.074)
10 blocks	SB vs. MLE	0.999 (0.935 , 1.066)	0.883 (0.820 , 0.949)
	HYB vs. MLE	1.000 (0.935 , 1.067)	0.864 (0.803 , 0.929)
	HYB vs. SB	1.000 (0.936 , 1.070)	0.978 (0.908 , 1.059)
25 blocks	SB vs. MLE	0.999 (0.936 , 1.065)	0.774 (0.715 , 0.840)
	HYB vs. MLE	1.000 (0.935 , 1.068)	0.751 (0.693 , 0.812)
	HYB vs. SB	1.000 (0.935 , 1.068)	0.970 (0.892 , 1.055)
40 blocks	SB vs. MLE	0.999 (0.937 , 1.066)	0.721 (0.666 , 0.780)
	HYB vs. MLE	0.999 (0.933 , 1.066)	0.691 (0.638 , 0.747)
	HYB vs. SB	1.000 (0.937 , 1.069)	0.958 (0.880 , 1.041)
50 blocks	SB vs. MLE	0.999 (0.935 , 1.067)	0.663 (0.613 , 0.718)
	HYB vs. MLE	1.000 (0.934 , 1.067)	0.644 (0.597 , 0.697)
	HYB vs. SB	1.000 (0.937 , 1.071)	0.970 (0.893 , 1.056)
100 blocks	SB vs. MLE	0.999 (0.937 , 1.065)	0.538 (0.496 , 0.580)
	HYB vs. MLE	0.999 (0.937 , 1.065)	0.559 (0.516 , 0.605)
	HYB vs. SB	1.000 (0.938 , 1.068)	1.039 (0.955 , 1.130)

Table 4.7: Bootstrap Confidence Intervals for $\alpha = 3$, $R = 1$, With Regression

	Method	$\alpha = 3$	$R = 1$
4 blocks	SB vs. MLE	0.982 (0.925 , 1.044)	0.972 (0.924 , 1.023)
	HYB vs. MLE	0.988 (0.931 , 1.049)	0.969 (0.919 , 1.022)
	HYB vs. SB	1.005 (0.948 , 1.064)	0.997 (0.946 , 1.050)
8 blocks	SB vs. MLE	0.953 (0.898 , 1.011)	0.940 (0.892 , 0.991)
	HYB vs. MLE	0.970 (0.915 , 1.031)	0.926 (0.877 , 0.976)
	HYB vs. SB	1.018 (0.959 , 1.081)	0.985 (0.932 , 1.038)
10 blocks	SB vs. MLE	0.936 (0.882 , 0.993)	0.927 (0.879 , 0.977)
	HYB vs. MLE	0.958 (0.902 , 1.015)	0.905 (0.856 , 0.957)
	HYB vs. SB	1.023 (0.964 , 1.085)	0.976 (0.922 , 1.032)
25 blocks	SB vs. MLE	0.926 (0.875 , 0.981)	0.907 (0.862 , 0.955)
	HYB vs. MLE	0.959 (0.906 , 1.017)	0.871 (0.824 , 0.920)
	HYB vs. SB	1.035 (0.978 , 1.097)	0.961 (0.909 , 1.013)
40 blocks	SB vs. MLE	0.926 (0.875 , 0.981)	0.907 (0.862 , 0.955)
	HYB vs. MLE	0.959 (0.906 , 1.017)	0.871 (0.824 , 0.920)
	HYB vs. SB	1.035 (0.978 , 1.097)	0.961 (0.909 , 1.013)
50 blocks	SB vs. MLE	0.920 (0.868 , 0.975)	0.855 (0.813 , 0.897)
	HYB vs. MLE	0.963 (0.909 , 1.018)	0.816 (0.773 , 0.860)
	HYB vs. SB	1.046 (0.987 , 1.107)	0.955 (0.905 , 1.008)
100 blocks	SB vs. MLE	0.912 (0.861 , 0.967)	0.794 (0.753 , 0.836)
	HYB vs. MLE	0.963 (0.909 , 1.020)	0.785 (0.747 , 0.828)
	HYB vs. SB	1.056 (0.997 , 1.119)	0.990 (0.940 , 1.042)

Table 4.8: Bootstrap Confidence Intervals for $\alpha = 3$, $R = 3$, With Regression

Chapter V

Application to the Precipitation Data Set

Large spatial data sets are not at all uncommon, and there is an increasing interest in inference based on the entire information available. One such example is the precipitation data set we present in this chapter, consisting of almost 6000 sites throughout the U.S. We begin by describing the general characteristics of the data. Our goal is to produce a spatial map for trends in monthly average precipitation. We perform a two-stage analysis: a site analysis to compute the monthly trends, followed by the spatial analysis. The estimation of spatial parameters is performed through the classical maximum likelihood technique (providing a standard for comparison), as well as through the alternative methods we proposed in the previous chapters, Small Blocks and Hybrid.

5.1 Data

The data set that we use to illustrate the impact of the alternative estimation techniques is the U.S. Daily Precipitation, compiled and made available by the National Climatic Data Center (NCDC). This data consists of a network of 5873 sites throughout the U.S. The coverage is given by: Southernmost Latitude—25N, Northernmost Latitude—50N, Westernmost Longitude—125W and Easternmost Longitude—65W. A map of all the sites is shown in Figure 5.1. At each site, precipitation is measured on a daily basis and the reporting unit is tenths of millimeters. The length of the individual site time series is 18,993 days, for the period

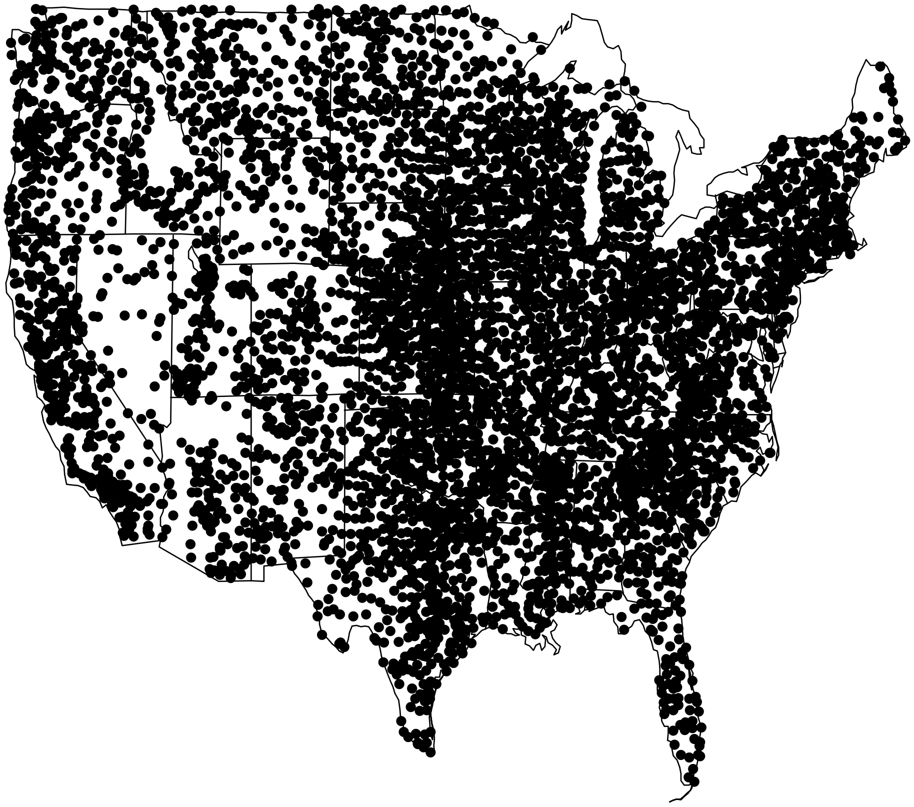

Figure 5.1: NCDC Daily Precipitation Data Set: 5873 Sites

from January 1, 1948 to December 31, 1999. We concentrate our attention to the time period between January 1, 1965 and December 31, 1999.

As a first step in understanding the data set, we draw a spatial map of average precipitation as measured by this network, for the period of interest. This is represented in Figure 5.2.

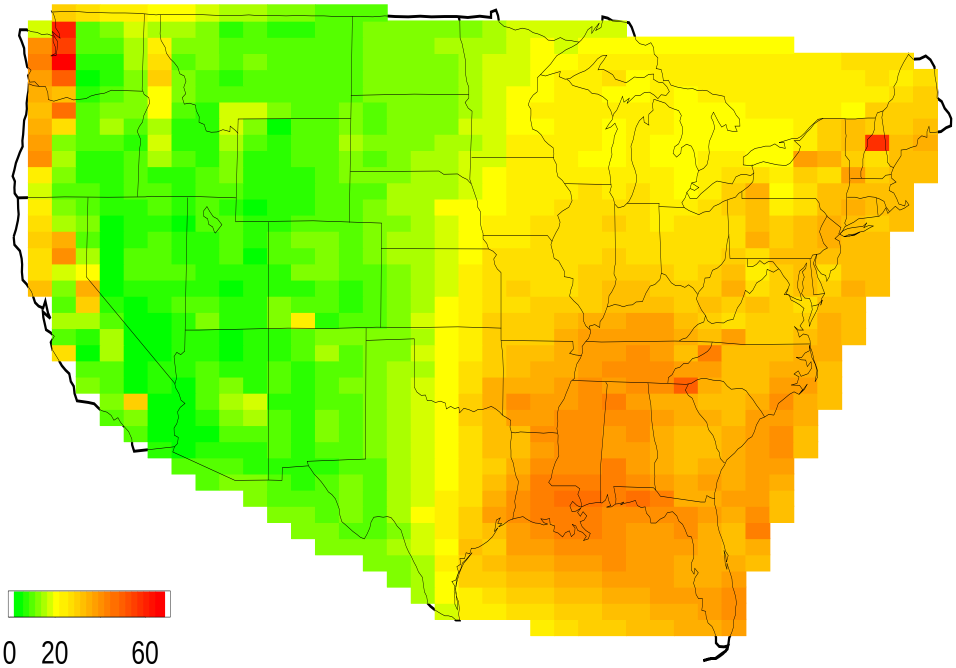

Figure 5.2: Interpolated Average Precipitation 1965-1985

5.2 Site Analysis

Throughout this section, we describe the aggregation of data performed at each site, in order to proceed to the spatial data analysis. The method employed is the same for all sites in the network, therefore, to simplify the notation, we shall drop the index corresponding to the site number.

To further simplify the spatial analysis, we first compute the trends in monthly average precipitation over the 20 years of interest. To do so, we consider the model:

$$Z_t = a\,\frac{t}{12} + b\,\mathbf{I} + \eta_t \tag{5.1}$$

where a is a unknown parameter, $t \in (1, .., 20 \times 12)$, b a unknown vector of 12 parameters, and $\mathbf{I}$ represents the 12 by 12 identity matrix (corresponding to the month variable).

We proceed by estimating the regression parameters through least squares, and record the values of $\hat{a}$ for future use in the spatial analysis.

5.3 Spatial Analysis

In the previous section, data at each site are aggregated into trends of monthly precipitation. This section is concerned with the evaluation of a spatial map based on these trends. A very crude interpolation technique is available in the Splus software. It is based on a triangulation scheme, where linear interpolation is used in the triangles bounded by data points. As a first step in understanding the spatial structure of the data, we employ this procedure for the aggregated site trends we computed previously, and the resulting map is shown in Figure 5.3.

Ideally, we would like to be able to estimate the parameters of the spatial covariance for the entire data set. Due to its large dimensions, it is not feasible to complete the estimation process employing the classical maximum likelihood techniques. Therefore, we start by confining our attention to a subset of the data, specifically 150 sites in Texas, as shown in Figure 5.4. As a first step in this analysis, we consider the covariance structure to belong to the exponential family, as given by equation (4.20). The design matrix X does not incorporate any functions of the site locations, therefore the only regression parameter that is to be estimated through least squares is the one corresponding to the intercept.

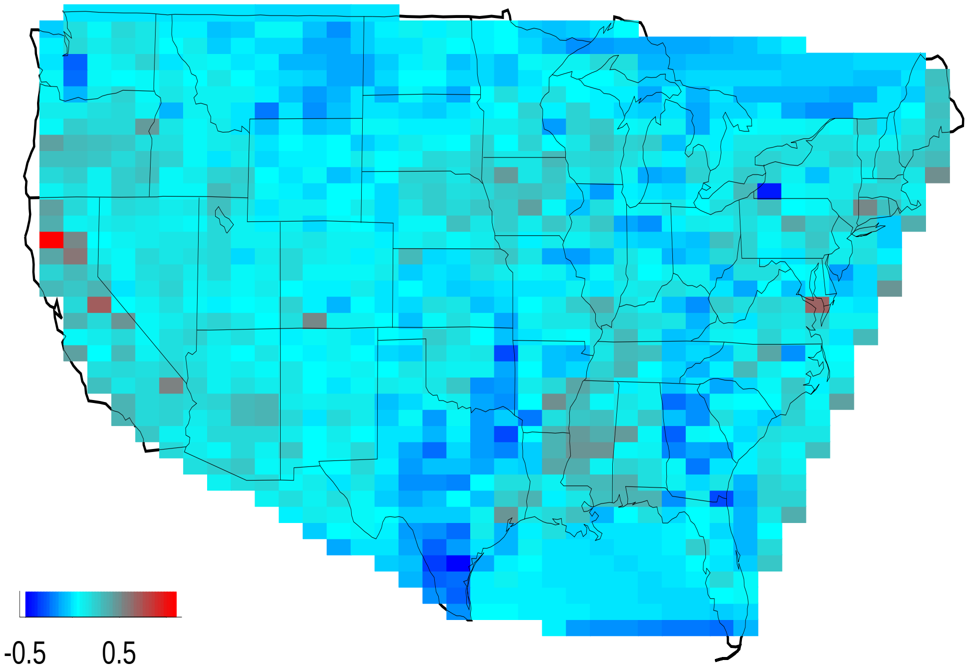

Figure 5.3: Interpolated Trends in Monthly Precipitation 1965-1985

We begin the analysis by computing the maximum likelihood estimators for the two spatial parameters driving this model, i.e. the scaling parameter α and the *Range*

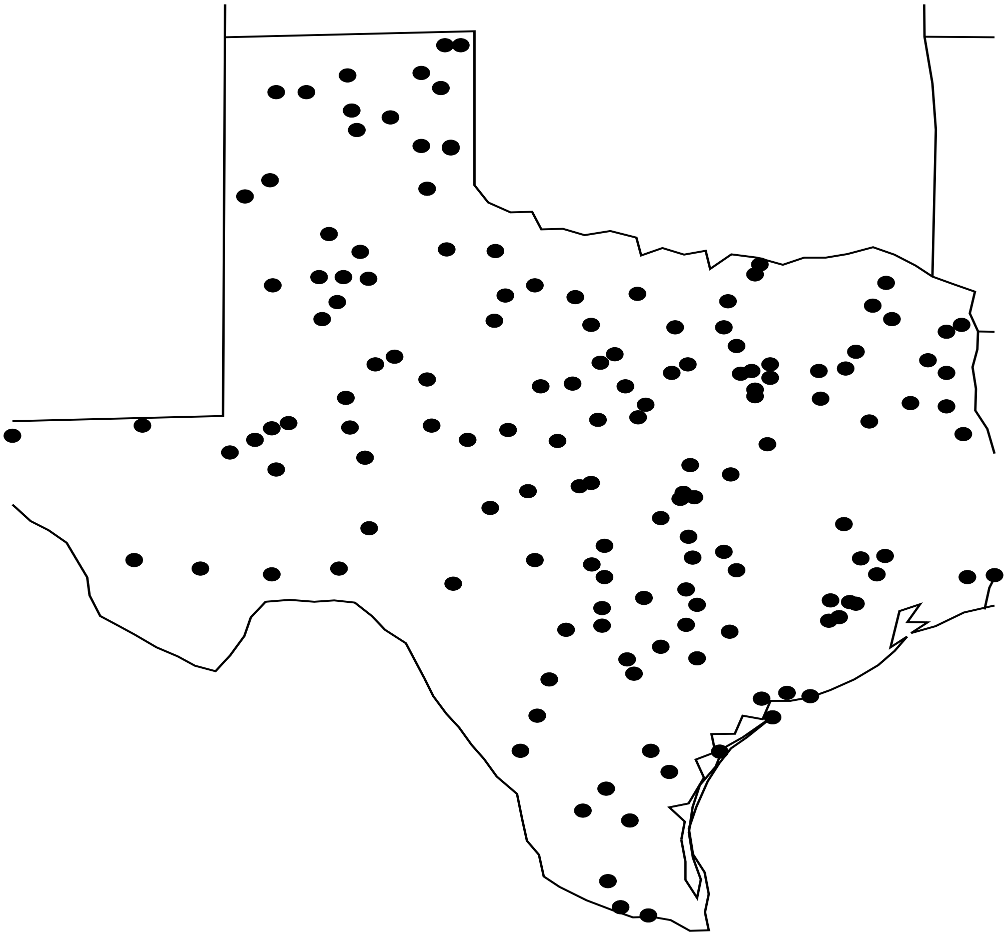

Figure 5.4: 150 Sites in Texas

parameter R. The results, which we denote by $\hat{\alpha}_{MLE}$ and $\hat{R}_{MLE}$ are:

$$\log(\hat{\alpha}_{MLE}) = -3.32855 \text{ and } \hat{R}_{MLE} = \ 0.54084\ . \tag{5.2}$$

To estimate the parameters through any of the three alternative techniques, we need to group the data into clusters. We consider first the case of 30 groups, each containing 5 sites. Applying the Small Blocks method we obtain the estimators $\hat{\alpha}_{SB}$ and $\hat{R}_{MLE}$, while employing the Hybrid method we obtain $\hat{\alpha}_{HYB}$ and $\hat{R}_{HYB}$. Results for such a clustering are:

$$\begin{aligned} \log(\hat{\alpha}_{SB}) &= -3.352 \text{ and } \hat{R}_{SB} = \ 0.509 \\ \log(\hat{\alpha}_{HYB}) &= -3.371 \text{ and } \hat{R}_{HYB} = \ 0.592 \ . \end{aligned}$$

We repeat this analysis a number of times, each time reclustering the data before proceeding with the estimation step. Also, we consider different cluster numbers, i.e. 15, 10 and 5 clusters of equal sizes. Results from all these analyses are displayed as a scatter plot in Figure 5.5. It is clear from this plot that the two alternative methods lead to estimators that are very close in magnitude to the maximum likelihood estimators, but it is not obvious which of the two produces better estimators.

Our objective is to evaluate a spatial map for the monthly precipitation trends over these 150 sites of interest. Therefore, it would be more sensible to compare the resulting maps for the three pairs of estimators, rather than their magnitudes. Such maps are displayed for two cases, the first one based on 30 groups of sites (Figure 5.6), and the second one based data clustered in 10 groups (Figure 5.7). Each of the two figures presents three maps, the first corresponding to the maximum likelihood case, the second to the Small Blocks method, and the third to the Hybrid technique. The maps are produced using kriging methodology on a 20 by 20 regular grid. Since the dimension of the problem still permits us to compute the inverse covariance matrix, we do a full kriging here, in order to compare the three methods. However, this would not be possible to do if the dimension of the problem were much larger, therefore we would need to rely on a conditional kriging approach.
It is again clear from Figures 5.6 and 5.7, that the alternative estimators lead to spatial maps that are almost identical to the map based on the maximum likelihood estimators.

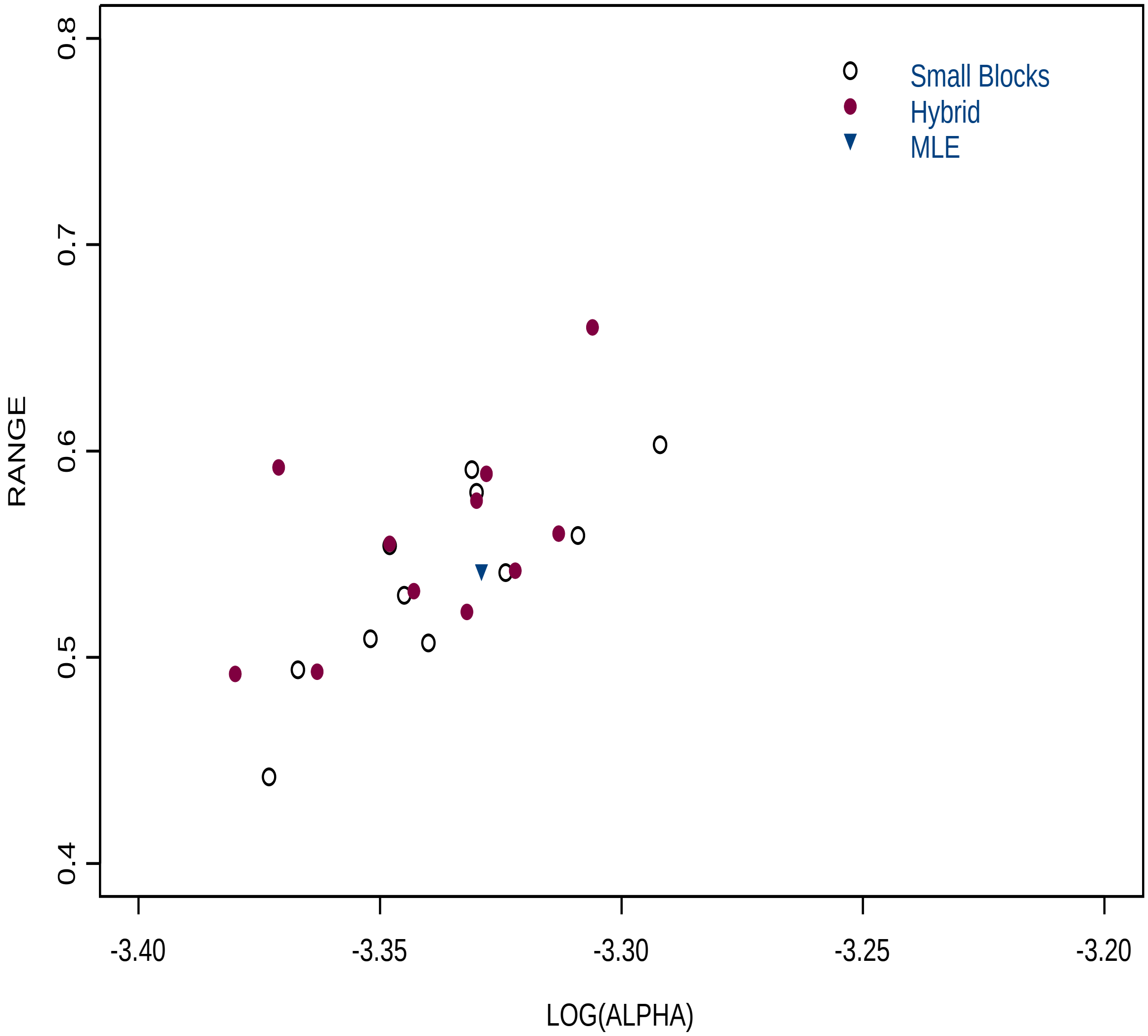

Figure 5.5: Scatterplot of Various Estimators for 150 Sites in Texas

In conclusion, according to the results obtained in this section, it follows that the two alternative methods, Small Blocks and Hybrid, do lead to results comparable to the classical case. Since they are computationally more efficient, it is therefore more advantageous to employ them for large data sets, where the maximum likelihood methods fail to produce the estimators.

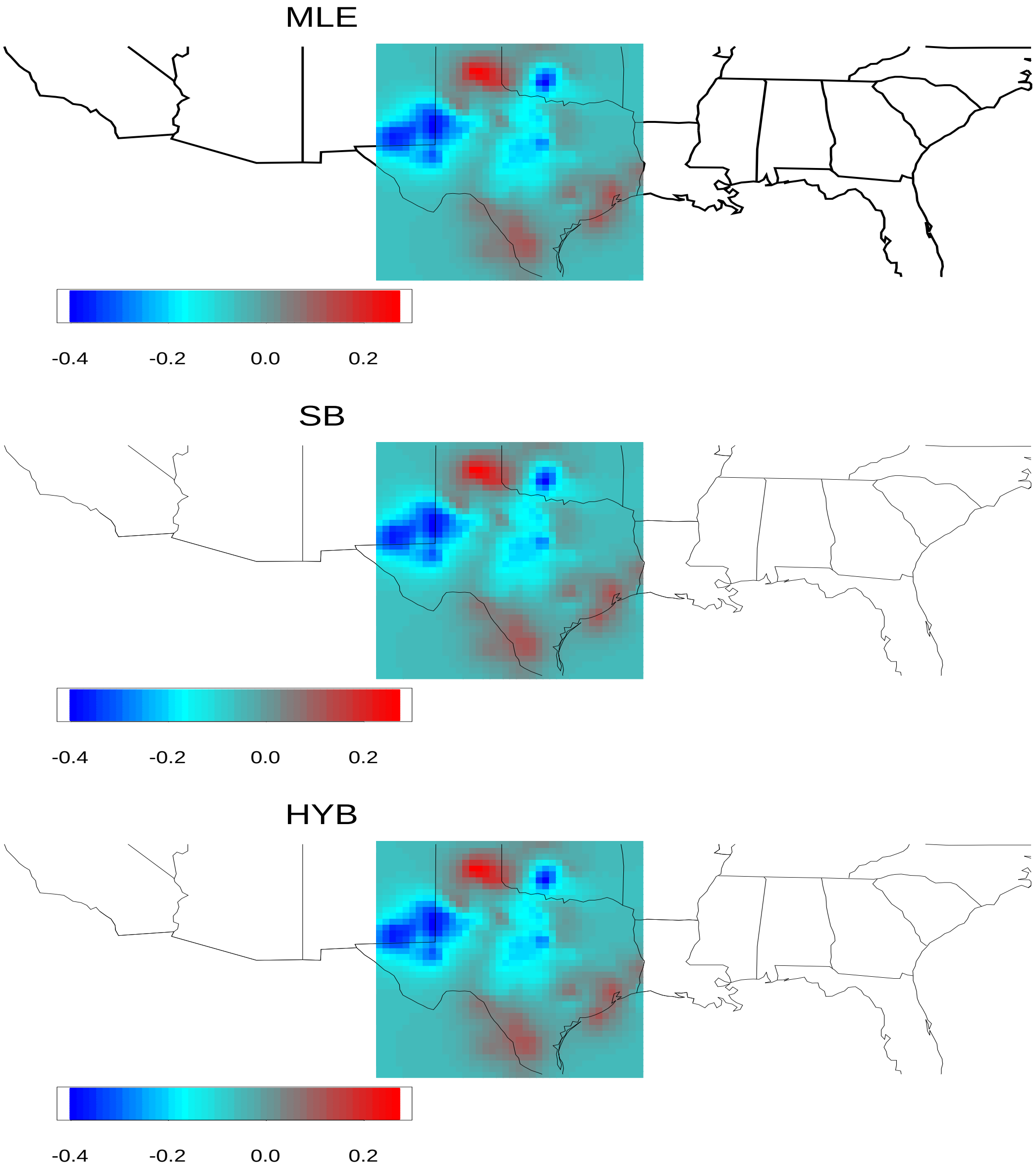

Figure 5.6: Interpolated Trends in Texas, 150 Sites, 30 blocks

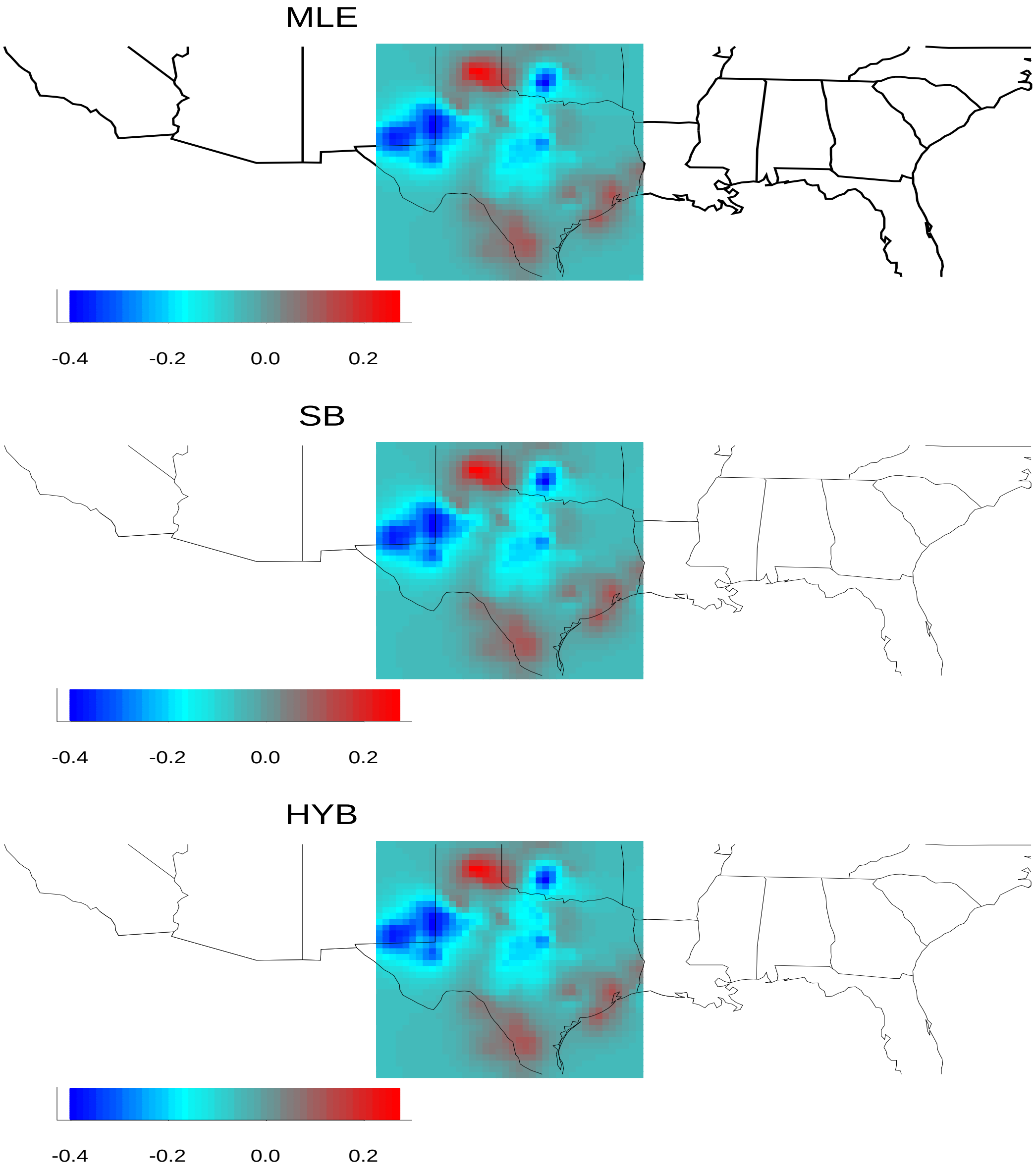

Figure 5.7: Interpolated Trends in Texas, 150 Sites, 10 blocks

Chapter VI

Forthcoming Work

The previous chapters describe the practical aspects related to the implementation of the alternative methods of estimation for the spatial parameters. Numerical studies illustrate performance of the estimators in the spatial setting, and an application to a real data set give an indication of their applicability. Theoretical derivations of the asymptotic efficiency could become extremely involved in more general cases. We illustrate here how would one extend the "Expansion technique" for the one dimensional time series problem to its analogous spatial process.

Extension to a Lattice Sampled Process

Consider a spatial process on an integer lattice, denoted x_{ij} where i and j are integers. Since we are going to model the process by its covariance structure, then one of the simplest forms to consider for the covariance structure is the Kronecker product form, i.e.

$$\mathrm{Cov}[x_{ij}, x_{t\ell}] = \gamma^{(1)}_{it}\gamma^{(2)}_{j\ell}, \tag{6.1}$$

where $\gamma^{(1)}$ and $\gamma^{(2)}$ are the covariances of one-dimensional time series in the horizontal and vertical directions. If we assume that these are both of AR(1) form, with the same autoregressive parameter, then we deduce

$$\mathrm{Cov}[x_{ij}, x_{t\ell}] = \sigma_x^2\phi^{|i-t|+|j-\ell|}, \tag{6.2}$$

where $|\phi| < 1$ for stationarity.

An equivalent definition, which represents (6.2) as a function of an array of independent $\mathcal{N}[0,1]$ random variables $\{\xi_{ij}\}$, is the formula

$$x_{ij} = \sigma_x(1-\phi^2)\sum_{r=0}^{\infty}\sum_{s=0}^{\infty}\phi^{r+s}\xi_{i+r,j+s}. \tag{6.3}$$

We may also represent the process equivalently by

$$x_{ij} - \phi(x_{i+1,j} + x_{i,j+1}) + \phi^2 x_{i+1,j+1} = \epsilon_{ij} \tag{6.4}$$

where $\epsilon_{ij} = \sigma_x(1-\phi^2)\xi_{ij}$ are independent $\mathcal{N}[0,\sigma_\epsilon^2]$, $\sigma_\epsilon^2 = \sigma_x^2(1-\phi^2)^2$. In the Kronecker product notation, the covariance function of the process is $U \otimes U$, and the inverse covariance function is $U^{-1} \otimes U^{-1}$, where U and U^{-1} are again given by (3.2) and (3.3). Note that the processes we have defined here lie within the general class of spatial processes on lattices first defined by Whittle (1954).

We now consider maximum likelihood estimation of ϕ. The model is that observations $\{x_{ij},\ 1 \le i \le m,\ 1 \le j \le n\}$ have a joint normal distribution with mean 0 and covariances given by (6.2). We also assume — because it simplifies the calculations and has little bearing on the final result — that σ_ϵ^2 is known. The negative log-likelihood for ϕ is then, modulo some fixed constants,

$$\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{t=1}^{m}\sum_{\ell=1}^{n} x_{ij}x_{t\ell}v_{ijt\ell} + \log|V|, \tag{6.5}$$

where V is the inverse covariance matrix and $v_{ijt\ell}$ is a component of the inverse covariance matrix evaluated at the $(i,j) \times (t,\ell)$ position. However the covariance matrix is $U_m \otimes U_n$, where U_n for any n is given by (3.2), and the inverse covariance matrix is therefore $U_m^{-1} \otimes U_n^{-1}$, where U_n^{-1} for any n is given by (3.3). Thus the analytical form of $v_{ijt\ell}$ is completely specified as follows:

$$V = \begin{pmatrix} A & B & 0 & 0 & \dots & 0 & 0 \\ B & C & B & 0 & \dots & 0 & 0 \\ 0 & B & C & B & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & B & C & B \\ 0 & 0 & 0 & \dots & 0 & B & A \end{pmatrix} \tag{6.6}$$

where A, B and C are $n \times n$ matrices given by

$$A = \begin{pmatrix} 1 & -\phi & \dots & 0 & 0 \\ -\phi & 1+\phi^2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 1+\phi^2 & -\phi \\ 0 & 0 & \dots & -\phi & 1 \end{pmatrix}, \tag{6.7}$$

$$B = \begin{pmatrix} -\phi & \phi^2 & 0 & \dots & 0 & 0 \\ \phi^2 & -\phi-\phi^3 & \phi^2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & -\phi-\phi^3 & \phi^2 \\ 0 & 0 & 0 & \dots & \phi^2 & -\phi \end{pmatrix} \text{ and} \tag{6.8}$$

$$C = \begin{pmatrix} 1+\phi^2 & -\phi-\phi^3 & 0 & \dots & 0 & 0 \\ -\phi-\phi^3 & (1+\phi^2)^2 & -\phi-\phi^3 & \dots & 0 & 0 \\ 0 & -\phi-\phi^3 & (1+\phi^2)^2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & (1+\phi^2)^2 & -\phi-\phi^3 \\ 0 & 0 & 0 & \dots & -\phi-\phi^3 & 1+\phi^2 \end{pmatrix}. \tag{6.9}$$

Observe that the estimating function in (6.5) can be rewritten as a quadratic form in independent normal random variables, taking advantage of the representation of

x_{ij} in (6.3) and (6.4). Therefore we can follow the expansion technique, much like we did in Chapter III, using the Martingale Central Limit Theorem and its Corollary for quadratic forms. For the classical MLE, we can also use the information matrix to calculate the asymptotic variance, and hence compare the results. Since the method involving Fisher information approximation would not lead to valid conjectures about the alternative estimators, i.e. Big Blocks, Small Blocks and Hybrid, we shall follow the expansion technique and obtain their asymptotic variances. Theoretical results will also be compared with their simulation derived analogues.

Small Blocks analysis To illustrate how the theoretical approach works for the lattice sampled process described in this section, we consider here the case that seemed to be the most manageable in the one-dimensional case. For simplicity, consider here that $m = n$, i.e. the sampling lattice is a square. Also, assume that we group the data in $b \times b$ disjoint groups, each consisting of $k \times k$ observations.

As for the one-dimensional process, in this case we assume that the blocks are independent. Therefore, the pseudo-likelihood in this case, is the product of the b^2 block likelihoods, i.e.

$$L(\phi) = \prod_{p=1}^{b} \prod_{q=1}^{b} L_{p,q}(\phi) \tag{6.10}$$

where by $L_{p,q}(\phi)$ we denote the likelihood corresponding to block $(p,\ q)$. On the other hand, the block likelihood is nothing but the classical likelihood reduced to a certain block. We denote by V^s the covariance matrix corresponding to any block (again, the covariance structure is identical for all the blocks). In other words, from equation (6.5) follows that the block negative log-likelihood is, modulo some fixed constants,

$$\ell_{p,q} = \sum_{i=1}^{k} \sum_{j=1}^{k} \sum_{t=1}^{k} \sum_{\ell=1}^{k} x_{(p-1)\,k+i,\ (q-1)\,k+j}\ x_{(p-1)\,k+t,\ (q-1)\,k+\ell}\ v^{s}_{(t-1)\,k+i,\ (\ell-1)\,k+j} + \log |V^s| \,. \tag{6.11}$$

We employed here a different notation for the indexes of the matrix V^s, which better illustrates the concept of block likelihood, as well as facilitates the actual calculations.

Also, in writing out the equation (6.11), we exploited the fact that the covariance matrix is going to be identical for each block.

From (6.5) and (6.11) it follows that the pseudo negative log-likelihood is given by

$$\sum_{p=1}^{b}\sum_{q=1}^{b}\sum_{i=1}^{k}\sum_{j=1}^{k}\sum_{t=1}^{k}\sum_{\ell=1}^{k} x_{(p-1)\,k+i,\,(q-1)\,k+j}\; x_{(p-1)\,k+t,\,(q-1)\,k+\ell}\; v^{s}_{(t-1)\;k+i,\;(\ell-1)\;k+j} + \log|V^{s}|\;. \tag{6.12}$$

The first comment we make about the function in equation (6.12) is that it is a degree four polynomial in the unknown parameter ϕ (this is just a result of the specific analytical form of the Kroneker product $U^{-1} \otimes U^{-1}$). Therefore, one can check that it satisfies the hypotheses in Theorem 2.3. In particular, assumptions (A) and (B) are straightforward. As we have shown in Chapter II, for condition (C) to be satisfied, all we need is that the expectations of the first order derivative are bounded on a neighborhood of the true value of the parameter. This condition is satisfied by the function in equation (6.12), therefore we can conclude that the two dimensional Small Blocks estimator is consistent.

Although the practical implementation of the Small Blocks method is very simple, computing the theoretical asymptotic efficiency is not. Our intent is to follow the principles outlined in the expansion method, as outlined in Chapter II. Using the alternative representation of the lattice process, as in equation (6.3), it is clear that we can expand the first derivative of the criterion function as a quadratic form of independent normal random variables. However, identification of the coefficients in the quadratic form is not a trivial step. For the moment, we proceed to the calculation of the expected value of the second derivative of the pseudo negative log-likelihood, as we need this quantity in the sandwich information technique. We can take advantage of the properties of the underlying AR(1) processes to compute the expected value of the second derivative avoiding coefficient identification. We proceed by computing the expected value of the second derivative for each block. Going back to equation (6.12), and denoting by $V^{s\prime\prime}$ the second derivative of the inverse covariance matrix, we have that

$$V^{s''} = \begin{pmatrix} A'' & B'' & 0 & 0 & \dots & 0 & 0 \\ B'' & C'' & B'' & 0 & \dots & 0 & 0 \\ 0 & B'' & C'' & B'' & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & B'' & C'' & B'' \\ 0 & 0 & 0 & \dots & 0 & B'' & A'' \end{pmatrix} \tag{6.13}$$

where A'', B'' and C'' are the second derivatives, with respect to ϕ, of the $k \times k$ matrices given by the expressions (6.7) through (6.9):

$$A'' = \begin{pmatrix} 0 & 0 & \dots & 0 & 0 \\ 0 & 2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 2 & 0 \\ 0 & 0 & \dots & 0 & 0 \end{pmatrix}, \quad B'' = \begin{pmatrix} 0 & 2 & 0 & \dots & 0 & 0 \\ 2 & -6\phi & 2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & -6\phi & 2 \\ 0 & 0 & 0 & \dots & 2 & 0 \end{pmatrix} \text{ and} \tag{6.14}$$

$$C'' = \begin{pmatrix} 2 & -6\,\phi & 0 & \dots & 0 & 0 \\ -6\,\phi & 4+12\,\phi^2 & -6\,\phi & \dots & 0 & 0 \\ 0 & -6\,\phi & 4+12\,\phi^2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 4+12\,\phi^2 & -6\,\phi \\ 0 & 0 & 0 & \dots & -6\,\phi & 2 \end{pmatrix}. \tag{6.15}$$

Therefore the analytical form of $V^{s''}$ is fully specified, and we can proceed to the actual calculation of the expected value for the second derivative of the criterion function. Thus, we have that

$$E[\ell''_{p,\,q}] \;=\; \sum_{i=1}^{k}\sum_{j=1}^{k}\sum_{t=1}^{k}\sum_{\ell=1}^{k} {v^{s''}}_{(t-1)\;k+i,\;(\ell-1)\;k+j} E[x_{(p-1)\,k+i,\;(q-1)\,k+j}\;x_{(p-1)\,k+t,\;(q-1)\,k+\ell}]$$

$$
\begin{aligned}
&+ \quad \frac{(\partial^2_\phi\,|V^s|)\,|V^s| - (\partial_\phi|V^s|)^2}{|V^s|^2} \\
&= \quad \sigma_x^2 \sum_{i=1}^{k}\sum_{j=1}^{k}\sum_{t=1}^{k}\sum_{\ell=1}^{k} v^{s\prime\prime}{}_{(t-1)\,k+i,\,(\ell-1)\,k+j}\ \phi^{|t-i|+|\ell-j|} \\
&+ \quad \frac{(\partial^2_\phi\,|V^s|)\,|V^s| - (\partial_\phi|V^s|)^2}{|V^s|^2} \qquad (6.16)
\end{aligned}
$$

which is straightforward to calculate.

Note that in equation (6.16) it was immaterial that we were referring to block (p, q). As a consequence, the expected value of the second derivative of the pseudo negative log-likelihood function is the same for all blocks.

As we have mentioned before, calculating the variance of the second derivative is a much more difficult procedure, and we shall develop this as part of our future research. We shall extend theoretical calculations for the other two alternative methods as well. Based on our experience with the one dimensional problem, we know we will not obtain a closed form solution for the inverse covariance matrix for the means process, and we might not even do so for the conditional covariance matrix for the Hybrid method. Therefore, we plan on completing all the calculations involved relying on numerical methods.

The applicability of the alternative methods described in this work goes well beyond the particular environmental data sets that we have employed to illustrate them. Large spatially correlated data sets are emerging from other sciences as well. One example would be biology, where there is need for estimating tools that can handle rich data sets. It is my intent to include such applications in my further research, as well as continuing the exploration of EPA and NOAA resources.

Another interesting extension to these methods is including the temporal aspect into the analysis, transitioning from a spatial analysis to a spatio-temporal one.

Also, the possible extension to nonstationary spatial processes is a very appealing feature of these methods, as one does not need to assume that the parametric model remains constant over the entire region, but rather assume stationarity on smaller subsets.

BIBLIOGRAPHY

Abramowitz,M. and Stegun, I.A. (1964), *Handbook of Mathematical Functions.* National Bureau of Standards, Washington D.C., reprinted by Dover, New York.

Akahira, M. and Takeuchi, K. (1981), *Asymptotic Efficiency of Statistical Estimators: Concepts and Higher Order Asymptotic Efficiency.* Lecture Notes in Statistics, Springer-Verlag, New York.

Amemiya, Takeshi (1985), *Advanced Econometrics.* Harvard University Press, Cambridge, Massachusetts.

Besag, J.E. (1974), Spatial interaction and the statistical analysis of lattice systems (with discussion). *J.R. Statist. Soc. B* **36**, 192–236.

Besag, J.E. (1975), Statistical analysis of non-lattice data. *The Statistician* **24**, 179–195.

Besag, J. (1989), A candidate's formula: A curious result in Bayesian prediction. *Biometrika* **76**, 183.

Billingsley, P. (1995), *Probability and Measure.* Third Edition, Wiley, New York.

Brockwell, P.J., Davis, R.A. (1991), *Time Series: Theory and Methods.* Second Edition, Springer-Verlag, New York.

Cassela, G. and Berger, R.L. (1990), *Statistical Inference*, Brooks/Cole Publishing Company

Cochrane, D. and Orcutt, G.H. (1949), Application of least squares regression to relationships containing autocorrelated error terms, *J. Amer. Statist. Assoc.* **44**, 32–61.

Cressie, N. (1985), Fitting variogram models by weighted least squares. *Mathematical Geology* **17**, 563–586.

Cressie, N. (1993), *Statistics for Spatial Data.* Second edition, John Wiley, New York.

Cressie, N. and Hawkins, D.M. (1980), Robust estimation of the variogram I. *Mathematical Geology* **12**, 115–125.

Handcock, M.S. and Wallis, J.R. (1994), An approach to statistical spatial-temporal modeling of meteorological fields (with discussion). *J. Amer. Statist. Assoc.* **89**, 368–390.

Hartigan, J.A. and Wong, M.A. (1979), A K-means clustering algorithm. *Applied Statistics* **28**, 101–108.

Holland, D.M., Caragea P.C. and Smith, R.L. (2002), Trends in Rural Sulfur Concentrations. *Submitted for publication*

Holland, D.M., De Oliveira, V., Cox, L.H. and Smith, R.L. (2000), Estimation of regional trends in sulfur dioxide over the eastern United States. *Environmetrics*, **11**, 373–393

Krige, D.G. (1951), A statistical approach to some basic mine valuation problems on the Witwatersrand. *Journal of the Chemical, Metallurgical and Mining Society of South Africa* **52**, 119-139

Mardia, K.V., Kent, J.T. and Bibby,J.M. (1979), *Multivariate Analysis.* New York: Academic Press

Liang, K.Y. and Zeger, S.L. (1986), Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

Mardia, K.V. and Marshall, R.J. (1984), Maximum Likelihood estimation of models for residual covariance in spatial regression. *Biometrika* **71**, 135–146.

Mardia, K.V. and Watkins, A.J. (1984), On multimodality of the likelihood in the spatial linear model. *Biometrika* **76**, 289–295.

Le, N.D. and Zidek, J.V. (1992), Interpolation with uncertain spatial covariances: A Bayesian alternative to kriging. *Journal of Multivariate Analysis* **43**, 351–374.

Patterson, H.D. and Thompson, R. (1971), Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**, 545–554.

Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T. (1986), *Numerical Recipes: The Art of Scientic Computing.* Cambridge University Press, Cambridge.

Ripley, B.D. (1988), *Statistical Inference for Spatial Processes.* Cambridge University Press, Cambridge, U.K.

Smith, R.L. (2001), CBMS Course in Environmental Statistics, University of Washington, June 25-29, 2001. http://www.stat.unc.edu/postscript/rs/envstat/env.html

Stein, M.L. (1999), *Interpolation of Spatial Data: Some Theory of Kriging.* Springer Verlag, New York.

Trench, W.F. (1964), An algorithm for the inversion of finite Toeplitz matrices. *J. Soc. Indust. Appl. Math.* **12**, 515–522.

Vecchia, A.V. (1988), Estimation and identification for continuous spatial processes. *J. Roy. Statist. B* **50** 297–312.

Van der Vaart, A.W. (1998), *Asymptotic Statistics.* Cambridge University Press

Warnes, J.J. and Ripley, B.D. (1987), Problems with likelihood estimation of covariance functions of spatial Gaussian processes. *Biometrika* **74**, 640–642.

Wei, William W.S. (1990), Time Series Analysis. Addison-Wesley.

White, H. (1982), Maximum Likelihood Estimation of Misspecified Models. *Econometrica* **50**, 1–26.

Whittle, P. (1954), On stationary processes in the plane. *Biometrika* **41**, 434–449.

Zimmerman, D.L. and Zimmerman, M.B. (1991), A comparison of spatial variogram estimators and corresponding ordinary kriging predictors. *Technometrics* **33**, 77–91.