ABSTRACT

Title of dissertation: SMALL AREA ESTIMATION AND PREDICTION PROBLEMS: SPATIAL MODELS, BAYESIAN MULTIPLE COMPARISONS, AND ROBUST MSE ESTIMATION.

Nadarajasundaram Ganesh
Doctor of Philosophy, 2007

Dissertation directed by: Professor Partha Lahiri
Joint Program in Survey Methodology
and
Professor Eric V. Slud
Department of Mathematics

We study and partially solve three distinct problems in small area estimation. The problems are loosely connected by a common theme of prediction and (empirical) Bayesian models.

In the first part of the thesis we consider prediction in a survey small area context with spatially correlated errors. We introduce a novel asymptotic framework in which the spatially correlated small areas form clusters, the number of such clusters and the number of small areas in each cluster growing with sample size. Under such an asymptotic framework we show consistency and asymptotic normality of the parameter estimators. For empirical predictors based on model estimates, we show through simulation and a real data example, improved prediction over estimates ignoring spatial error-correlations.

The second part of the thesis involves using a hierarchical Bayes approach to

solve the problem of multiple comparison in small area estimation. In the context of multiple comparison, a new class of moment matching priors is introduced. This class includes the well-known *superharmonic* prior due to Stein. Through data analysis and simulation we illustrate the use of our class of priors.

In the third part of the thesis, for a special case of the nested error regression model, we derive a non-parametric second order unbiased estimator of the mean squared error of the empirical best linear unbiased predictor. For the balanced case, the Prasad-Rao estimator is shown to be second order unbiased when the small area effects are non-normal. Through simulation we show that the Prasad-Rao estimator is robust for departures from normality.

SMALL AREA ESTIMATION AND PREDICTION PROBLEMS: SPATIAL MODELS, BAYESIAN MULTIPLE COMPARISONS, AND ROBUST MSE ESTIMATION.

by

Nadarajasundaram Ganesh

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2007

Advisory Committee:
Dr. Partha Lahiri, Co-Chair
Dr. Eric V. Slud, Co-Chair
Dr. Benjamin Kedem
Dr. Marc Nerlove
Dr. Paul J. Smith

DEDICATION

To *Appa*.

In memory of my late father Kumaraswamy Nadarajasundaram

born May 13, 1941; died May 30, 1983.

ACKNOWLEDGEMENTS

Many people have contributed in helping me complete this dissertation. My advisors have been instrumental in training me as a researcher and I am greatly indebted to them for their excellent guidance and for being wonderful people to work with.

I thank my mother for everything she has done for me; I would also like to thank my aunt Rani and late uncle Thave for the many ways in which they have helped me; my brother for the many lunches and dinners we had during my last year of graduate school; my uncle Jega, aunt Navamani, Sanjay and Baba for their generosity; Anjali, Don and family for living in Washington, D.C.; Ravi for many years of teaching me mathematics; Umesh and Mara for two great vacations in San Diego; and all my other relatives who are simply too numerous to thank individually.

My sincere thanks to my friends Somantika Datta and Justin Brody for typing a portion of this thesis. I thank my friends in the Department of Mathematics, especially Justin Brody, Somantika Datta, Rob Delgado, Andy Kebo, Jessie Kim, Yabing Mai, Joanna Pressley, Suzanne Sindi and Pol Tangboondouangjit for sharing their time and enriching my graduate school experience.

Many thanks to Dr. Benjamin Kedem, Dr. Marc Nerlove and Dr. Paul Smith for serving on my committee. Thanks also to the graduate staff Linette Berry, Darcy Conant, Anita Dahms, Haydeé Hidalgo, Celeste Regalado and Patty Woodwell for *always* being helpful.

Finally, thank you to the Maryland Food Co-op for providing six years of "food for people, not for profit".

TABLE OF CONTENTS

LIST OF TABLES

,

LIST OF FIGURES

List of notations and abbreviations

I_m	$m \times m$ identity matrix
J_m	$m \times m$ matrix of 1's
O_m	$m \times m$ matrix of 0's
$\mathbf{1}_m$	$m \times 1$ vector of 1's
$\mathbf{0}_m$	$m \times 1$ vector of 0's
$\mathbf{x}$	column vector
$\mathbf{x}'$	transpose of vector $\mathbf{x}$
$\|\mathbf{x}\|$	Euclidean norm of vector $\mathbf{x}$
$\mathbf{f}_i$	i^{th} standard basis vector in $\mathbb{R}^M$
$\boldsymbol{\eta}_o$	true value of parameter $\boldsymbol{\eta}$
$\text{diag}(\psi_1, \ldots, \psi_m)$	diagonal matrix with entries $\psi_1, \ldots, \psi_m$
$\text{blockdiag}(A_1, \ldots, A_k)$	block diagonal matrix with blocks $A_1, \ldots, A_k$
A'	transpose of matrix A
$\|A\|$	determinant of matrix A
$\text{tr}(A)$	trace of matrix A
$A(\boldsymbol{\eta})$	entries of matrix A are functions of a parameter $\boldsymbol{\eta}$
$\frac{\partial A(\boldsymbol{\eta})}{\partial \eta_i}, \frac{\partial A}{\partial \eta_i}$	element wise partial derivative of matrix A
$A > B$	$A - B$ is a positive definite matrix
$A \to B$	for matrices A and B, entry by entry convergence
$A_o, A(\boldsymbol{\eta}_o)$	matrix A evaluated at true parameter value
A_{ij}	$(i, j)^{th}$ entry of matrix A

$\gamma_{\max}(A), \gamma_{\min}(A)$	largest, smallest eigenvalue of matrix A
$\xrightarrow{p}$	convergence in probability
$X_n \overset{p}{\approx} Y_n$	$X_n - Y_n \xrightarrow{p} 0$
$o_p(1)$	convergence in probability to zero
$O_p(1)$	bounded in probability
$\xrightarrow{d}$	convergence in distribution
$\nabla_{\boldsymbol{\eta}} g(\boldsymbol{\eta})$	gradient of g with respect to $\boldsymbol{\eta}$
$\nabla_{\boldsymbol{\eta}\boldsymbol{\eta}}\, g(\boldsymbol{\eta})$	hessian of g with respect to $\boldsymbol{\eta}$
$\nabla_{\boldsymbol{\eta}} g(\boldsymbol{\eta}_o; \mathbf{y})$	gradient of g evaluated at $\boldsymbol{\eta}_o$
$\frac{\partial g(\boldsymbol{\eta}_o)}{\partial \eta_i}$	partial derivative of g evaluated at $\boldsymbol{\eta}_o$
$I_{[\cdot]}$	indicator function
$\lvert S \rvert$	cardinality of set S
S^c	complement of set S
ind	independent
iid	independent and identically distributed
BP	best predictor
BLUP	best linear unbiased predictor
EBLUP	empirical best linear unbiased predictor
EB	empirical Bayes
MLE	maximum likelihood estimator
MSE	mean squared error
REML	restricted maximum likelihood estimator

List of conventions

Unless stated otherwise,

1. a lower case c with any subscript or superscript denotes a constant that may not necessarily be the same constant in different sections of this thesis.
2. $\mathrm{E}(\cdot)$, $\mathrm{var}(\cdot)$ and $\mathrm{cov}(\cdot,\cdot)$ are taken with respect to the true model.

Chapter 1

Introduction

1.1 Background

For effective planning of health, social and other services, and for apportioning government funds, there is a growing demand among various government agencies such as the U.S. Census Bureau, U.K. Central Statistical Office, and Statistics Canada to produce reliable estimates for smaller sub-populations, called small areas. For example, in both developed and developing countries, governmental policies increasingly demand income and poverty estimates for small areas. In fact, in the U.S.A., more than $130 billion of federal funds per year are allocated based on these estimates.

A sample survey designed for a large population may select a small number of elements - even no element - for the small area of interest. Other non-sampling errors such as non-response may further reduce the sample size for the small area. Thus, standard design-based methods that are solely based on the survey data, generally fail to provide small area estimates with the desired level of precision. Over

the last two decades, different model-based approaches that *borrow strength* from related data sources have been proposed in the literature. Such methods essentially use explicit models to combine information from the sample survey, various administrative/census records, and even previous surveys.

Depending on whether the data are available at the small area level or at the unit or respondent level, two popular small area models are used.

(a) Fay-Herriot model (area level model)

In order to estimate the per-capita income of small places (population less than 1000), Fay and Herriot [18] used the following two-level empirical Bayes model:

- Level 1 (sampling model): $y_i|\theta_i \overset{\text{ind}}{\sim} N(\theta_i, \psi_i),\ i = 1, \cdots, m;$
- Level 2 (linking model): $\theta_i \overset{\text{ind}}{\sim} N(\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2),\ i = 1, \cdots, m.$

In the above model, Level 1 is used to account for the sampling variability of the direct survey estimates y_i of the true small area means θ_i. Level 2 links the true small area means θ_i to a vector of q known auxiliary variables $\mathbf{x}_i$, often obtained from various administrative and census records. The parameters $\boldsymbol{\beta}$ and σ^2 of the linking model are unknown and are estimated from the available data. In order to estimate the sampling variability ψ_i, Fay and Herriot [18] employed the generalized variance function method (see Wolter [54]) that uses some external information from the survey. In the Fay-Herriot model, it is customary to assume that the ψ_i's are known without error, even though it is usually the case that some part of ψ_i is estimated.

The Fay-Herriot model can also be viewed as a mixed linear model:

$$y_i = \theta_i + e_i = \mathbf{x}_i'\boldsymbol{\beta} + v_i + e_i,\ i = 1, \cdots, m,$$

where v_i's and e_i's are independent with $v_i \overset{iid}{\sim} N(0, \sigma^2)$ and $e_i \overset{ind}{\sim} N(0, \psi_i)$. Fay and Herriot [18] used random effects (also referred to as small area effects) in order to capture the additional area-specific effects not explained by the area-specific auxiliary variables. In contrast, the corresponding regression model without random effects fails to capture this additional area-specific variability. Using the U.S. census data, Fay and Herriot [18] demonstrated that their empirical Bayes (EB) estimator [also an empirical best linear unbiased predictor (EBLUP)] performed better than the direct survey estimator and a synthetic estimator used earlier by the U.S. Census Bureau.

(b) Nested-error regression model (unit level model)

To estimate areas planted with corn and soybeans for twelve counties in North-Central Iowa, Battese et al. [3] used the following model:

$$y_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta} + v_i + e_{ij},\ i = 1, \ldots, m,\ j = 1, \ldots, n_i,$$

where y_{ij} is the j^{th} observation in the i^{th} small area, $\mathbf{x}_{ij}$ is a vector of covariates at the unit-level, v_i's and e_{ij}'s are independent with $v_i \overset{iid}{\sim} N(0, \sigma_v^2)$ and $e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2)$. Here, v_i's are area specific effects and e_{ij}'s are random effects associated with the j^{th} observation in the i^{th} small area. For the nested-error regression model, the usual parameter of interest is the small area mean $\theta_i = \overline{\mathbf{X}}_i'\boldsymbol{\beta} + v_i$, where $\overline{\mathbf{X}}_i$ is the known population mean of the covariates of the i^{th} small area.

1.2 Linear mixed models

In this section, we borrow ideas from semi-parametric regression, spatial statistics, geostatistics and disease mapping in order to discuss possible ways of generalizing the Fay-Herriot model. One such method is to assume that the mean function of the response variable is an unspecified smooth function. In such cases, splines could be used to approximate the smooth function. A second method is by modeling the random effects; we briefly discuss the conditional autoregressive model and the simultaneous autoregressive model. A third method is given by directly modeling the variance-covariance matrix of the random effects.

Consider a single covariate and assume

$$y_i = f(x_i) + \epsilon_i,\ i = 1, \ldots, m, \tag{1.1}$$

where $\epsilon_i \overset{iid}{\sim} \mathrm{N}(0, \sigma_\epsilon^2)$ and $f(x)$ is some unspecified smooth function. Kammann and Wand [23], Wahba [48] and Wand [50] suggest using splines to approximate the smooth function. For example, a cubic spline basis could be used to approximate the smooth function. However, due to the large number of parameters, usually a linear spline basis is used to approximate $f(x)$. We would fit

$$y_i = \beta_o + \beta_1 x_i + \sum_{j=1}^{r} u_j (x_i - \kappa_j)_+ + \epsilon_i,\ i = 1, \ldots, m,$$

via least squares, where

$$(x_i - \kappa_j)_+ = \begin{cases} 0 & \text{if } x_i \leq \kappa_j \\ x_i - \kappa_j & \text{if } x_i > \kappa_j, \end{cases}$$

and $\kappa_1, \ldots, \kappa_r$ are referred to as the 'knots'. Usually r, the number of knots, is chosen to be large, approximately one for every $3-4$ observations upto a maximum

of $20 - 40$ knots (Kammann and Wand [23]). However, the large number of knots will lead to a rough fit. By considering a penalty term on the coefficients of the knots (referred to as penalized splines), a much smoother fit can be achieved. That is, $(\boldsymbol{\beta}, \mathbf{u})$ are estimated by minimizing

$$\sum_{i=1}^{m} \left(y_i - \beta_o - \beta_1 x_i - \sum_{j=1}^{r} u_j (x_i - \kappa_j)_+ \right)^2 + \alpha \|\mathbf{u}\|^2 \tag{1.2}$$

where $\mathbf{u} = (u_1, \ldots, u_r)'$ and α is the penalty parameter. Note that in (1.2) instead of the Euclidean norm a number of other norms could be considered (Ruppert et al. [41]).

Minimizing (1.2) with respect to $(\boldsymbol{\beta}, \mathbf{u})$ is equivalent to treating $\mathbf{u}$ as a random effect in a linear mixed model (Wahba [49]). That is, let

$$X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_m \end{pmatrix}, \qquad Z = \begin{pmatrix} (x_1 - \kappa_1)_+ \ldots (x_1 - \kappa_r)_+ \\ \vdots \qquad\qquad \vdots \\ (x_m - \kappa_1)_+ \ldots (x_m - \kappa_r)_+ \end{pmatrix}.$$

Then penalized least squares is equivalent to the best linear unbiased prediction of $(\boldsymbol{\beta}, \mathbf{u})$ in the linear mixed model:

$$\mathbf{y} = X\boldsymbol{\beta} + Z\mathbf{u} + \boldsymbol{\epsilon} \tag{1.3}$$

where $\mathbf{u} \sim \mathrm{N}(\mathbf{0}_r, \sigma_u^2 I_r)$, $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}_m, \sigma_\epsilon^2 I_m)$ and $\mathbf{u}, \boldsymbol{\epsilon}$ are independent.

In the context of small area estimation, Opsomer et al. [34] used penalized splines to estimate the mean acid neutralizing capacity for 113 small areas. They fit their model using a bivariate spline on the geographical co-ordinates of the centroid of each small area. Even though we have only discussed splines for the univariate case, extension to the bivariate case can be done by considering radial or other specialized basis functions. We do not elaborate any further.

We also note several other contributors to the literature on spline smoothing. For example, Claeskens [8] derived a test statistic based on splines for testing a parametric mean model against a nonparametric alternative. When the number of knots and the smoothing parameter are selected as a function of m and the data, testing a parametric mean model against a nonparametric alternative is asymptotically equivalent to testing $\sigma_u^2 = 0$ in (1.3). Moreover, Ruppert and Carroll [42] considered spline fitting with a penalty parameter that varies spatially. That is, instead of assuming the penalty parameter α in (1.2) is a constant for all x, α is considered to be a function of x. The penalty parameter is allowed to vary spatially to adapt to possible spatial heterogeneity in the regression function.

For the linear mixed model

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i,\ i = 1, \cdots, m,$$

we next give an overview of different ways of modeling the mean zero random effect u_i. A popular model used in spatial statistics is the conditional autoregressive model (Besag [5] and Cressie [11]), where the u_i's are modeled as

$$u_i|\{u_j : j \neq i\} \sim \mathrm{N}\Big(\sum_{j\neq i,\, j=1}^{m} c_{ij}u_j,\ \sigma_u^2\Big), \tag{1.4}$$

and it follows from Besag [5],

$$\mathbf{u} \sim \mathrm{N}(\mathbf{0}_m, \sigma_u^2(I_m - C)^{-1})$$

where $\mathbf{u} = (u_1, \ldots, u_m)'$, C is a symmetric $m \times m$ matrix with $(i,j)^{th}$ element c_{ij}, and $c_{ii} = 0$. In the context of disease mapping, Clayton and Kaldor [8] considered a conditional autoregressive model with $c_{ij} = 1$ if i, j are neighboring districts

and $c_{ij} = 0$ otherwise. In order to estimate U.S. census undercount, Cressie [10] considered a more general conditional autoregressive model than the one given in (1.4).

A simultaneous autoregressive model (Cressie [11] and Ord [35]) treats the u_i's as

$$u_i = \sum_{j=1}^{m} c_{ij} u_j + \epsilon_i,$$

which can be expressed as

$$(I_m - C)\mathbf{u} = \boldsymbol{\epsilon}$$

$$\mathbf{u} = (I_m - C)^{-1}\boldsymbol{\epsilon},$$

where $\mathbf{u} = (u_1, \ldots, u_m)'$, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_m)'$, C is a $m \times m$ matrix with $(i,j)^{th}$ element c_{ij}. Moreover, assuming $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}_m, \sigma_\epsilon^2 I_m)$, we have

$$\mathbf{u} \sim \mathrm{N}(\mathbf{0}_m, \sigma_\epsilon^2 (I_m - C)^{-1}(I_m - C')^{-1}).$$

For example, Ord [35] took $C = \rho W$, where W is a known weighting matrix.

We also note that there are several other models for the random effects u_i that are mentioned in spatial statistics. For example, there is a spatial analog of the moving average model that is used in time series (Cliff and Ord [9]).

In contrast to the aforementioned models for the random effects u_i, the approach we take in Chapter 2 is directly modeling the variance-covariance matrix of the random effects (in our context, the random effects are referred to as small area effects). Rao [37] suggests using models from spatial statistics to model the

variance-covariance matrix of the u_i's. One such example is

$$\alpha_1 I_m + \alpha_2 W \tag{1.5}$$

where W_{ij}, the $(i,j)^{th}$ element of W, is given by

$$W_{ij} = \exp(-\alpha_3 \|\mathbf{h}_i - \mathbf{h}_j\|) \tag{1.6}$$

where $\|\mathbf{h}_i - \mathbf{h}_j\|$ is the distance between small areas i and j. The covariance models discussed in Chapter 2 are similar to (1.5)-(1.6). However, a major difference is the asymptotic framework we consider (see Section 2.2).

1.3 Overview of thesis

In Chapter 2, by introducing a scaling factor, we consider a hybrid asymptotic framework between infill asymptotics and increasing domain asymptotics (see Section 2.1-2.2). We assume that the small areas can be partitioned into clusters, the number of such clusters and the number of small areas in each cluster growing with sample size. Under such an asymptotic framework, we suggest a few models for the covariance matrix of the small area effects.

In Chapter 3, for the small area model we consider, we provide a method to estimate all parameters. Moreover, we show that our parameter estimators are consistent and asymptotically normal.

In Chapter 4, through a simulation study, we investigate the properties of the parameter estimators derived in Chapter 3. We compare the predictor obtained under our model and the predictor obtained under the misspecified Fay-Herriot

model. We also investigate the conjecture that under the asymptotic framework we consider, the maximum likelihood estimator is consistent and asymptotically normal.

In Chapter 5, we use the small area model that was proposed in Chapter 2 and the estimation methods developed in Chapter 3 to analyze a spatial data set.

In Chapter 6, we give a short summary of the small area estimation problems that are discussed and partially solved in Chapters 7 and 8. This chapter serves as a bridge between the problems with spatial covariates and correlated errors, discussed in earlier chapters, and the specialized non-spatial small area estimation problems treated in the rest of the thesis.

In Chapter 7, for the Fay-Herriot model, we use a hierarchical Bayes (HB) approach to develop a methodology to construct simultaneous $100(1-\alpha)\%$ credible intervals. We develop a new class of moment matching priors for the prior variance that has a desirable frequentist property.

In Chapter 8, for a special case of the nested error regression model, we derive a nonparametric second order unbiased estimator of the mean squared error (MSE) of the empirical best linear unbiased predictor (EBLUP). Through simulation, we show for various parameter combinations, the Prasad and Rao [36] estimator is quite robust for departures from normality.

In Chapter 9, we summarize our results and discuss future research problems.

Chapter 2

Modeling

In the Fay-Herriot [18] model, the small area effects are assumed to be independent - though in many data problems neighboring areas ought to be correlated - and by modeling the correlation, better predictors of the small area means could be achieved. For the models considered in Chapters 2-5, it is assumed that

$$\begin{aligned} y_{i\mathrm{M}} &= \theta_{i\mathrm{M}} + e_i,\ i \in S \\ \theta_{i\mathrm{M}} &= \mathbf{x}_i'\boldsymbol{\beta} + v_{i\mathrm{M}},\ i \in U \end{aligned} \tag{2.1}$$

where U is the set of all small areas, with $|U| = M$ elements, and $S \subset U$ is the set of sampled small areas, with $|S| = m$ elements. The small area means constitute a triangular array of the type $\{\theta_{i\mathrm{M}} : i = 1, 2, \ldots, M,\ M = 1, 2, \ldots\}$. We do not observe the small area means, but instead observe survey estimates $\{y_{i\mathrm{M}} : i \in S\}$. The $\mathbf{x}_i$'s are vector valued covariates for the i^{th} area and $\boldsymbol{\beta} \in \mathbb{R}^q$ is an unknown vector valued parameter. As in the Fay-Herriot model, the sampling errors $e_i \overset{\text{ind}}{\sim} N(0, \psi_i)$, the e_i's and the $v_{i\mathrm{M}}$'s are independent. The small area effect vector $\mathbf{v}_{\mathrm{U}} = (v_{1\mathrm{M}}, \ldots, v_{\mathrm{MM}})'$ is a mean zero normal random vector with covariance matrix $\Sigma_{\mathrm{U}}(\boldsymbol{\eta}) = \Sigma_{\mathrm{U}}$, where $\boldsymbol{\eta}$ is a vector valued parameter. See Section 2.2 for a discussion of variance-covariance

models for Σ_{U}.

For notational convenience, the set U is re-indexed so that the first m elements of U consist of the sampled small areas. Also, the subscript M in $y_{i\mathrm{M}}$, $\theta_{i\mathrm{M}}$, $v_{i\mathrm{M}}$ is dropped, with the understanding that the θ_i's constitute a triangular array. Given the set of sampled small areas, the vector of survey estimates $\mathbf{y} = (y_1, \ldots, y_m)'$ can be modeled as

$$\mathbf{y} = \boldsymbol{\theta} + \mathbf{e} = X\boldsymbol{\beta} + \mathbf{v} + \mathbf{e}, \tag{2.2}$$

where $X = (\mathbf{x}_1, \ldots, \mathbf{x}_m)'$, $\mathbf{v} = (v_1, \ldots, v_m)'$ and $\mathbf{e} = (e_1, \ldots, e_m)'$ are independent with $\mathbf{v} \sim \mathrm{N}(\mathbf{0}, \Sigma)$ and $\mathbf{e} \sim \mathrm{N}(\mathbf{0}, \Psi)$. Here $\Sigma = \Sigma(\boldsymbol{\eta})$ is the sub-matrix of Σ_{U} that corresponds to the sampled small areas, and $\Psi = \mathrm{diag}(\psi_1, \ldots, \psi_m)$. Also,

$$\mathrm{var}(\mathbf{y}) \equiv V = V(\boldsymbol{\eta}) = \Sigma + \Psi. \tag{2.3}$$

2.1 Overview of asymptotics for spatial data

For spatial data, two distinct asymptotic frameworks have been studied. Increasing domain asymptotics refers to more and more observations being sampled over an increasing domain $\mathcal{D} \subset \mathbb{R}^2$ such that the Lebesgue measure of $\mathcal{D}$ is unbounded. When referring to increasing domain asymptotics, it is assumed that the spatial locations of the observations do not become dense. That is, for some ϵ independent of M and $\epsilon > 0$, $\|\mathbf{h}_i - \mathbf{h}_j\| > \epsilon$, where $\mathbf{h}_i$, $\mathbf{h}_j$ are the spatial locations of the observations (Cressie [11] and Mardia and Marshall [26]). Under this asymptotic framework, Mardia and Marshall [26] showed that the maximum likelihood

estimator (MLE) of the covariance parameters of a Gaussian process is consistent and asymptotically normal.

Infill asymptotics refers to observations being increasingly sampled over a bounded domain. There are very few asymptotic results under infill asymptotics. For example, it is known that some covariance parameters of a zero mean Gaussian process cannot be consistently estimated, and for the remaining covariance parameters, the MLE is consistent and asymptotically normal. For such results, see Abt and Welch [1], Chen et al. [7], Stein [45], Ying [55], Zhang [56] and Zhang and Zimmerman [57].

One of the most popular covariance models for spatial data is given by

$$C(\mathbf{h}_i, \mathbf{h}_j) = \begin{cases} \sigma^2 + \delta & \text{if} \quad i = j, \\ \delta \exp(-\lambda \|\mathbf{h}_i - \mathbf{h}_j\|) & \text{if} \quad i \neq j, \end{cases} \tag{2.4}$$

where $\delta \geq 0$, $\lambda \geq 0$, $\sigma^2 > 0$. See Cressie [11], Stein [45] and Zimmermann and Harville [59]. The above model is referred to as the exponential covariance model with nugget effect. Under infill asymptotics, and assuming that the spatial process is Gaussian, when the covariance model is given by (2.4) and the spatial locations h_i are situated on a lattice in $[0, 1]$, Chen et al. [7] showed that the MLE for σ^2 is $m^{\frac{1}{2}}$-consistent. Moreover, δ and λ cannot be simultaneously consistently estimated, but the MLE for $\delta\lambda$ is $m^{\frac{1}{4}}$-consistent. Under infill asymptotics when either the spatial locations h_i are irregularly spaced on $[0, 1]$ or for any spatial pattern $\mathbf{h}_i \in [0, 1]^2$, there are no asymptotic results in the current literature for the MLE for $\boldsymbol{\eta} = (\delta, \lambda, \sigma^2)'$. On the other hand, under increasing domain asymptotics and assuming the spatial locations do not become dense, the MLE for $\boldsymbol{\eta} = (\delta, \lambda, \sigma^2)'$ is $m^{\frac{1}{2}}$-

consistent (Mardia and Marshall [26]).

2.2 Covariance models for the small area effects

Motivated by the results mentioned in Section 2.1, we assume the covariance model for the small areas effects is given by a model similar to (2.4), but we consider a hybrid asymptotic framework by introducing a scaling factor.

Since small area effects need not depend on geography alone, it would be reasonable to assume that a number of other covariates influence the correlation between two "neighboring" areas, and hence, $\mathbf{z}_i^\star$ is a s-dimensional vector of spatial locations and certain categorical and continuous variables which measure spatial similarity. The $\mathbf{z}_i^\star$'s are in a fixed, finite dimensional space whose dimension is independent of M. The vector of spatial locations and covariates $\mathbf{z}_i^\star$ of the small areas are thought to be in an increasing domain, but are scaled such that $\mathbf{z}_i$ are in a bounded domain.

The proposed covariance model for the small area effects is

$$\Sigma_{\mathrm{U}} = \sigma^2 I_M + \delta A_{\mathrm{U}} \tag{2.5}$$

where the $(i,j)^{th}$ entry of A_{U} is given by

$$A_{ij} = \exp(-\lambda M^p \|\mathbf{z}_i - \mathbf{z}_j\|), \tag{2.6}$$

where $\delta \geq 0$, $\lambda \geq 0$, $\sigma^2 > 0$, M^p is the scaling factor, $0 < p < 1/s$ is a user specified parameter (s is the dimension of $\mathbf{z}_i^\star$), and the scaling is such that $\|\mathbf{z}_i^\star - \mathbf{z}_j^\star\| = M^p\|\mathbf{z}_i - \mathbf{z}_j\|$. Note that when $\delta = 0$ we have the Fay-Herriot model. When $p = 1/s$,

and $\mathbf{z}_i^\star$ are only s-dimensional spatial locations, and $\|\mathbf{z}_i^\star - \mathbf{z}_j^\star\| > \epsilon > 0$, we have a special case of the usual increasing domain asymptotic framework.

Our main assumption in Chapters 2-5 is:

Assumption (C): *The set of small areas $|U|$ can be partitioned into k $(= k(M)$ increasing to ∞ with $M)$ clusters $C_1, \ldots, C_k$, with cluster sizes $N_1, \ldots, N_k$ such that $\sum_{l=1}^k N_l = M$. From each cluster C_l, n_l of the N_l small areas are sampled such that $\sum_{l=1}^k n_l = m$. The n_l's are assumed to be non-random. The asymptotic framework that is considered is $k \to \infty$ and for each l, $N_l \to \infty$, $n_l \to \infty$ such that* $0 < \lim_{n_l, N_l \to \infty} n_l/N_l < \infty$.

Moreover, for $l = 1, \ldots, k$,

$$\limsup_{M\to\infty} M^p \sup_{i,j\in C_l} \|\mathbf{z}_i - \mathbf{z}_j\| < \infty, \tag{2.7}$$

and for all $l_1 \neq l_2$,

$$\liminf_{M\to\infty} \frac{M^p}{\log M} \inf_{i\in C_{l_1}, j\in C_{l_2}} \|\mathbf{z}_i - \mathbf{z}_j\| = \infty. \tag{2.8}$$

The factor of $\log M$ in (2.8) is needed for technical reasons when deriving the asymptotic distribution of estimators for δ and λ (see Remark 7 in Section 3.4.2 and the proof of Theorem 3.3 in Section 3.5). Note the slightly unusual definition of what it means for two small areas to be in the same cluster. They are defined to be in the same cluster only if, asymptotically, their unscaled distance from one another is bounded.

Moreover, we do not want the clusters to shrink toward a point, that is, it is

assumed that for $l = 1, \ldots, k$, $\exists\ c_l > 0$ such that

$$\lim_{N_l \to \infty} \frac{1}{N_l^2} \sum_{i,j \in C_l} I_{[M^p \|\mathbf{z}_i - \mathbf{z}_j\| \geq c_l]} = \epsilon_l, \tag{2.9}$$

where $0 < \epsilon_l \leq 1$. As will be shown in Chapter 3, (2.9) is a sufficient condition for the parameter λ to be identified and consistently estimated.

In Chapter 3, estimation methods and asymptotic results are derived under Assumption (C), (2.1)-(2.3) and (2.5)-(2.9). In addition, several other possible covariance models for the small area effects are described below.

We also mention that an alternative method to ours would be to include the vector $\mathbf{z}_i^\star$ in the mean structure as an unspecified smooth function [similar to (1.1)]. We do not elaborate any further.

Based on how clusters were defined in (2.7)-(2.8), a cluster model for A_{U} is given by

$$A_{ij} = \begin{cases} \exp(-\lambda M^p \|\mathbf{z}_i - \mathbf{z}_j\|) & \text{if } \ i, j, \in C_l \text{ for some } l, \\ 0 & \text{otherwise.} \end{cases} \tag{2.10}$$

Note that for each cluster C_l and for all $i, j \in C_l$, $A_{ij} = \exp(-\lambda M^p \|\mathbf{z}_i - \mathbf{z}_j\|)$ is a valid covariance model. Since the clusters are uncorrelated and we assume the small area effects vector $\mathbf{v}_{\mathrm{U}}$ is a normal random vector, it follows that (2.10) is a valid covariance model. Moreover, we do not have to limit ourselves by defining distance as the Euclidean norm. We could use other norms in (2.6) and (2.10). However, we need to use a distance norm that gives a positive definite matrix A_{U}.

As is shown in Chapter 3, under Assumption (C), the asymptotic distribution of the parameter estimators when the true model is (2.1)-(2.3), (2.5), (2.7)-(2.10) is the same as when the true model is given by (2.1)-(2.3), (2.5)-(2.9).

Furthermore, (2.6) can be generalized to include a vector parameter $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)'$. For the i^{th} small area, let $\mathbf{h}_i^\star = (h_{i1}^\star, \ldots, h_{is}^\star)'$ and $\mathbf{d}_i^\star = (d_{i1}^\star, \ldots, d_{ia}^\star)'$ respectively denote the vector of spatial locations and the vector of certain categorical and continuous covariates. Since the scaling factors of each of the covariates and spatial locations could be different, an alternate covariance model for the small area effects is given by

$$A_{ij} = \exp\Big(-\lambda_1 M^{p_o}\|\mathbf{h}_i - \mathbf{h}_j\| - \lambda_2\Big(\sum_{r=1}^{a} M^{p_r}(d_{ir} - d_{jr})^2\Big)^{\frac{1}{2}}\Big) \tag{2.11}$$

where $\mathbf{h}_i = (h_{i1}, \ldots, h_{is})'$ and $\mathbf{d}_i = (d_{i1}, \ldots, d_{ia})'$ are respectively the scaled spatial locations and scaled covariates. For $r = 1, \ldots, a$, M^{p_r} is the scaling factor associated with covariate $d_{ir}^\star$, where $p_r \geq 0$ and $\sum_{r=1}^{a} p_r \leq 2$, and M^{p_o} is the scaling factor for the spatial locations, where $0 < p_o < 1/s$ (and s is the dimension of $\mathbf{h}_i^\star$). For any r, $p_r = 0$ means that $\{d_{ir}^\star : i \in U\}$ is always in a bounded interval. For example, the covariate, proportion of county residents with a college degree. Also, $p_r > 0$ means that $\{d_{ir}^\star : i \in U\}$ lies in an increasing interval. For example, with an increasing number of counties, it is possible that one may see a wider range for average county income. Also, (2.11) is a valid covariance model. See Cressie and Huang [13].

Similarly to (2.7)-(2.8), it is assumed that the set of small areas $|U|$ can be partitioned into clusters $C_1, \ldots, C_k$ such that for $l = 1, \ldots, k$,

$$\limsup_{M\to\infty} M^{p_o} \sup_{i,j\in C_l} \|\mathbf{h}_i - \mathbf{h}_j\| + \Big(\sum_{r=1}^{a} \limsup_{M\to\infty} M^{p_r} \sup_{i,j\in C_l} (d_{ir} - d_{jr})^2\Big)^{\frac{1}{2}} < \infty, \tag{2.12}$$

and for all $l_1 \neq l_2$,

$$\liminf_{M\to\infty} \frac{M^{p_o}}{\log M} \inf_{\substack{i\in C_{l_1}\\ j\in C_{l_2}}} \|\mathbf{h}_i - \mathbf{h}_j\| + \Big(\sum_{r=1}^{a} \liminf_{M\to\infty} \frac{M^{p_r}}{(\log M)^2} \inf_{\substack{i\in C_{l_1}\\ j\in C_{l_2}}} (d_{ir} - d_{jr})^2\Big)^{\frac{1}{2}} = \infty \tag{2.13}$$

Also, similarly to (2.9), we do not want the clusters to shrink toward a point. That is, we assume for $l = 1, \ldots, k$, $\exists\ c_{l,1}, c_{l,2} > 0$ such that

$$\lim_{N_l \to \infty} \frac{1}{N_l^2} \sum_{i,j \in C_l} I_{[M^{p_o}\|\mathbf{h}_i - \mathbf{h}_j\| \geq c_{l,1},\ (\sum_{r=1}^{a} M^{p_r}(d_{ir} - d_{jr})^2)^{\frac{1}{2}} \geq c_{l,2}]} = \epsilon_l, \tag{2.14}$$

where $0 < \epsilon_l \leq 1$.

A second cluster model for A_{U} is given by

$$A_{ij} = \begin{cases} \exp\Big\{ -\lambda_1 M^{p_o}\|\mathbf{h}_i - \mathbf{h}_j\| \\ \qquad -\lambda_2 \Big(\sum_{r=1}^{a} M^{p_r}(d_{ir} - d_{jr})^2 \Big)^{\frac{1}{2}} \Big\} & \text{if } i, j, \in C_l \text{ for some } l, \\ 0 & \text{otherwise.} \end{cases} \tag{2.15}$$

Finally, we consider the following covariance model for A_{U} in which small areas within clusters are equally correlated, and small area between clusters are uncorrelated:

$$A_{\mathrm{U}} = \mathrm{blockdiag}(J_{N_1}, \ldots, J_{N_k}) \tag{2.16}$$

The above model could be seen as a special case of (2.10) by defining $\|\mathbf{z}_i - \mathbf{z}_j\| = 0$ if i, j are in the same cluster.

2.3 Prediction

As mentioned previously, one of the objectives of modeling the small area effects is to obtain better predictors. Consider predicting a linear combination of fixed effects and small area effects; that is, for known $\mathbf{a} \in \mathbb{R}^q$, $\boldsymbol{\ell} \in \mathbb{R}^M$, we wish to

predict

$$t = \mathbf{a}'\boldsymbol{\beta} + \boldsymbol{\ell}'\mathbf{v}_{\mathrm{U}} \tag{2.17}$$

where $\mathbf{v}_{\mathrm{U}}$ is the (population) vector of small area effects. When using the Fay-Herriot model, the best linear unbiased predictor (BLUP) or the empirical BLUP (EBLUP) is used to predict the parameter of interest θ_i (Das et al. [14], Datta and Lahiri [15], Datta et al. [17], Lahiri and Rao [24] and Rao [37]). The same practice is adopted in this thesis as well.

Since it was assumed that the set of all small areas U was re-indexed such that the first m counties are observed, let $\mathbf{v}_{\mathrm{U}}$ and Σ_{U} be partitioned as follows:

$$\mathbf{v}_{\mathrm{U}} = \begin{pmatrix} \mathbf{v} \\ \mathbf{v}_\star \end{pmatrix}, \ \Sigma_{\mathrm{U}} = \begin{pmatrix} \Sigma & \Sigma_\star \\ \Sigma'_\star & \Sigma_{\star\star} \end{pmatrix}$$

where $\mathbf{v} \in \mathbb{R}^m$ is the vector of observed small area effects, $\mathbf{v}_\star \in \mathbb{R}^{M-m}$ is the vector of unobserved small area effects, $\Sigma = \mathrm{var}(\mathbf{v})$, $\Sigma_\star = \mathrm{cov}(\mathbf{v}, \mathbf{v}_\star)$ and $\Sigma_{\star\star} = \mathrm{var}(\mathbf{v}_\star)$.

For a general linear model, Rao [37] derived the BLUP of t. For the model given by Assumption (C), (2.1)-(2.3) and (2.5)-(2.9), the BLUP of t can be derived in a manner almost identical to the proof given in Rao [37], and hence, the proof is omitted. The BLUP $\hat{t}(\boldsymbol{\eta})$ of t is given by

$$\hat{t}(\boldsymbol{\eta}) = \mathbf{a}'\widetilde{\boldsymbol{\beta}}(\boldsymbol{\eta}) + \boldsymbol{\ell}'\Delta V^{-1}(\mathbf{y} - X\widetilde{\boldsymbol{\beta}}(\boldsymbol{\eta}))$$

where $V = \Sigma + \Psi = \sigma^2 I_m + \delta A + \Psi$, A is the sub-matrix of A_{U} that corresponds to the sampled small areas, $\Delta = \Delta(\boldsymbol{\eta}) = \begin{pmatrix} \Sigma \\ \Sigma'_\star \end{pmatrix}$ and $\widetilde{\boldsymbol{\beta}}(\boldsymbol{\eta})$ is the best linear unbiased estimator of $\boldsymbol{\beta}$. That is,

$$\widetilde{\boldsymbol{\beta}}(\boldsymbol{\eta}) = (X'V^{-1}X)^{-1}X'V^{-1}\mathbf{y}.$$

Moreover, the MSE of $\hat{t}(\boldsymbol{\eta})$ is given by

$$\text{MSE}[\hat{t}(\boldsymbol{\eta})] = g_1(\boldsymbol{\eta}) + g_2(\boldsymbol{\eta})$$

where

$$g_1(\boldsymbol{\eta}) = \boldsymbol{\ell}'(\Sigma_{\text{U}} - \Delta V^{-1}\Delta')\boldsymbol{\ell}$$

$$g_2(\boldsymbol{\eta}) = (\mathbf{a} - X'V^{-1}\Delta'\boldsymbol{\ell})'(X'V^{-1}X)^{-1}(\mathbf{a} - X'V^{-1}\Delta'\boldsymbol{\ell}).$$

Note that $g_1(\boldsymbol{\eta})$ is the MSE of the BLUP when $\boldsymbol{\beta}$ is known. Once again the proof is omitted as it is almost identical to that given in Rao [37].

Since the BLUP $\hat{t}(\boldsymbol{\eta})$ involves unknown parameters, an empirical version of the BLUP, referred to as the EBLUP, is given by

$$\hat{t}(\hat{\boldsymbol{\eta}}) = \mathbf{a}'\widetilde{\boldsymbol{\beta}}(\hat{\boldsymbol{\eta}}) + \boldsymbol{\ell}'\Delta(\hat{\boldsymbol{\eta}})[V(\hat{\boldsymbol{\eta}})]^{-1}(\mathbf{y} - X\widetilde{\boldsymbol{\beta}}(\hat{\boldsymbol{\eta}}))$$

where $\hat{\boldsymbol{\eta}}$ is a consistent estimator of $\boldsymbol{\eta}$.

We are interested in predicting $\theta_i = \mathbf{x}_i'\boldsymbol{\beta} + v_i$. For $i = 1, \ldots, M$, the EBLUP $\hat{\theta}_i(\hat{\boldsymbol{\eta}})$ of θ_i is given by

$$\hat{\theta}_i(\hat{\boldsymbol{\eta}}) = \mathbf{x}_i'\widetilde{\boldsymbol{\beta}}(\hat{\boldsymbol{\eta}}) + \mathbf{f}_i'\Delta(\hat{\boldsymbol{\eta}})[V(\hat{\boldsymbol{\eta}})]^{-1}(\mathbf{y} - X\widetilde{\boldsymbol{\beta}}(\hat{\boldsymbol{\eta}})) \tag{2.18}$$

where $\mathbf{f}_i$ is the i^{th} standard basis vector in $\mathbb{R}^M$.

In order to compare two predictors of t, we define the relative efficiency of two unbiased predictors $\hat{t}$ and $\tilde{t}$ by

$$R(\hat{t}, \tilde{t}) = \frac{\text{MSE}[\,\tilde{t}\,]}{\text{MSE}[\,\hat{t}\,]} \tag{2.19}$$

When all parameters are known, assuming that the true model is given by Assumption (C), (2.1)-(2.3) and (2.5)-(2.9), the following discussion seeks to calculate the relative efficiency of the BLUP $\hat{\theta}_i$ of θ_i obtained under the true model and the BLUP $\tilde{\theta}_i$ of θ_i obtained under the misspecified Fay-Herriot model. Knowing the relative efficiency of the two predictors gives an idea as to what parameter settings would require us to use the more complex model that is proposed as opposed to the simpler Fay-Herriot model. When $\boldsymbol{\beta}$ is known, $\tilde{\theta}_i$ is given by

$$\tilde{\theta}_i = \begin{cases} \mathbf{x}_i'\boldsymbol{\beta} + \frac{\sigma^2+\delta}{\sigma^2+\delta+\psi_i}(y_i - \mathbf{x}_i'\boldsymbol{\beta}) & \text{if } i \in S \\ \mathbf{x}_i'\boldsymbol{\beta} & \text{if } i \in S^c. \end{cases} \tag{2.20}$$

It is not obvious that in (2.20), $\sigma^2+\delta$ is the correct parameter choice. However, using the Kullback-Leibler Information Criterion (KLIC) in (3.4), we show that the aforementioned parameter choice minimizes the KLIC between the true model and the misspecified Fay-Herriot model, and is therefore the correct parameter choice.

Next, we compute the MSE of $\tilde{\theta}_i$, where $\tilde{\theta}_i$ is given by (2.20). For $i \in S$,

$$\begin{aligned} \mathrm{MSE}[\tilde{\theta}_i\,] &= \mathrm{E}\Big(\mathbf{x}_i'\boldsymbol{\beta} + \frac{\sigma^2+\delta}{\sigma^2+\delta+\psi_i}(y_i - \mathbf{x}_i'\boldsymbol{\beta}) - \mathbf{x}_i'\boldsymbol{\beta} - v_i\Big)^2 \\ &= \Big(\frac{\sigma^2+\delta}{\sigma^2+\delta+\psi_i}\Big)^2(\sigma^2+\delta+\psi_i) - 2\frac{\sigma^2+\delta}{\sigma^2+\delta+\psi_i}(\sigma^2+\delta) + (\sigma^2+\delta) \\ &= \frac{(\sigma^2+\delta)\psi_i}{\sigma^2+\delta+\psi_i} \end{aligned} \tag{2.21}$$

Also, for $i \in S^c$,

$$\mathrm{MSE}[\tilde{\theta}_i\,] = \mathrm{E}(v_i^2) = \sigma^2+\delta \tag{2.22}$$

When $\boldsymbol{\beta}$ is known, it follows from Rao [37] that the MSE of $\hat{\theta}_i$ is $g_1(\boldsymbol{\eta})$ where

$$\mathrm{MSE}[\hat{\theta}_i\,] = g_1(\boldsymbol{\eta}) = \mathbf{f}_i'(\Sigma_{\mathrm{U}} - \Delta V^{-1}\Delta')\mathbf{f}_i \tag{2.23}$$

As a simple example, using (2.21)-(2.23) and (2.19) we compute $R(\hat{\theta}_i, \tilde{\theta}_i)$ for $k = 1$, $n = m = 20$, $N = M = 40$ and $k = 1$, $n = m = 40$, $N = M = 80$. That is, we assume there is one cluster with 40 or 80 small areas of which we sample 20 or 40 small areas. We take $k = 1$ for the following reason: by Assumption (C), (2.7)-(2.8) and since all parameters are known, when predicting a small area only observed small areas from the same cluster are used. Furthermore, for simplicity we took $\psi_i = 0.5$ for all $i \in S$. The parameter λ was chosen so that the median within cluster values of the off diagonal entries of A_{U} was some number c. We chose values $c = 0.70, 0.35$ that respectively correspond to $\lambda = 0.12, 0.41$.

Table 2.1: $R(\hat{\theta}_i, \tilde{\theta}_i)$ for sampled and non-sampled small areas when $\psi_i = 0.5$. 'Obs.' refers to $R(\hat{\theta}_i, \tilde{\theta}_i)$ for an observed area, 'Unobs.' refers to $R(\hat{\theta}_i, \tilde{\theta}_i)$ for an unobserved area.

M	m	δ	λ	σ^2	Obs.	Unobs.
40	20	0.6	0.12	0.4	1.303	1.946
40	20	0.3	0.12	0.7	1.086	1.267
40	20	0.6	0.41	0.4	1.199	1.604
40	20	0.3	0.41	0.7	1.058	1.177
80	40	0.6	0.12	0.4	1.331	2.032
80	40	0.3	0.12	0.7	1.097	1.300
80	40	0.6	0.41	0.4	1.227	1.678
80	40	0.3	0.41	0.7	1.068	1.205

In Table 2.1 the column 'Obs' refers to $R(\hat{\theta}_i, \tilde{\theta}_i)$ for an observed small area. Similarly 'Unobs' refers to $R(\hat{\theta}_i, \tilde{\theta}_i)$ for an unobserved small area. As can be seen from Table 2.1 larger relative efficiency is obtained for the unobserved small areas. Moreover, large δ, m and small λ give larger relative efficiency. Note that since $k = 1$, m corresponds to the number of sampled small areas in a cluster. That is, relative efficiency depends on the number of sampled small areas in a cluster. We also refer to Tables 4.1 and 4.2 where $R(\hat{\theta}_i, \tilde{\theta}_i)$ is computed when all parameters are estimated. The tables are comparable in the sense that the parameter combinations of $(\delta, \lambda, \sigma^2)$ and n, N are the same. However, the ψ_i's were not all 0.5 in Tables 4.1 and 4.2 (see Section 4.2 for how the ψ_i's were generated to obtain Tables 4.1 and 4.2). We note that the relative efficiency in Tables 4.1 and 4.2 are similar to what we have obtained here.

Next we compute (2.19) for a much simpler model. Let $\hat{\theta}_i^{\star}$ be the BLUP of θ_i obtained under the model (2.1)-(2.3), (2.5),(2.16).

When the variance-covariance model for the small area effects is given by (2.5),(2.16), since the k small area clusters are independent and all parameters are known, for purposes of predicting θ_i, it may be assumed that $k = 1$ without loss of generality. This follows from noting that when predicting θ_i, since the clusters are independent, only observed small areas from the same cluster are used. For this one cluster, m of the M small areas are sampled. The set of all small areas U is once again re-indexed such that the first m elements of U consist of the sampled small areas S.

The BLUP $\hat{\theta}_i^{\star}$ of θ_i obtained under model (2.1)-(2.3), (2.5),(2.16) is

$$\hat{\theta}_i^{\star} = \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{f}_i' \begin{pmatrix} \sigma^2 I_m + \delta J_m \\ \delta J_{M-m} \end{pmatrix} [\sigma^2 I_m + \delta J_m + \Psi]^{-1}(\mathbf{y} - X\boldsymbol{\beta}),$$

and since $\boldsymbol{\beta}$ is known, the MSE of $\hat{\theta}_i^{\star}$ is given by the term g_1 (Rao [37]), that is

$$\begin{aligned} \mathrm{MSE}[\hat{\theta}_i^{\star}] = \mathbf{f}_i' \Big(& \sigma^2 I_M + \delta J_M \\ & - \begin{pmatrix} \sigma^2 I_m + \delta J_m \\ \delta J_{M-m} \end{pmatrix} [\sigma^2 I_m + \delta J_m + \Psi]^{-1} \big(\sigma^2 I_m + \delta J_m \mid \delta J_{M-m} \big) \Big) \mathbf{f}_i \end{aligned} \tag{2.24}$$

where $\big(\sigma^2 I_m + \delta J_m \mid \delta J_{M-m}\big)$ is an $m \times M$ matrix with the first m columns given by $\sigma^2 I_m + \delta J_m$ and the last $M - m$ columns given by δJ_{M-m}.

For any $i \in S$, using (2.24), we calculate the MSE of $\hat{\theta}_i^{\star}$. Let $E = \mathrm{diag}(\sigma^2 + \psi_1, \ldots, \sigma^2 + \psi_m)$. By multiplying both sides of (2.25) by $E + \delta J_m$, it can be checked that

$$[\sigma^2 I_m + \delta J_m + \Psi]^{-1} = E^{-1} - \frac{\delta}{1 + \delta \sum_{j=1}^{m} 1/(\sigma^2 + \psi_j)} E^{-1} J_m E^{-1} \tag{2.25}$$

For a sampled small area, that is, $i \in S$, note that

$$\begin{aligned} \mathbf{f}_i' \big(\sigma^2 I_M + \delta J_M \big) \mathbf{f}_i &= \sigma^2 + \delta, \\ \mathbf{f}_i' \begin{pmatrix} \sigma^2 I_m + \delta J_m \\ \delta J_{M-m} \end{pmatrix} &= \sigma^2 \mathbf{f}_i' + \delta \mathbf{1}_m', \\ (\delta \mathbf{1}_m + \sigma^2 \mathbf{f}_i)' E^{-1} (\delta \mathbf{1}_m + \sigma^2 \mathbf{f}_i) &= \delta^2 \sum_{j=1}^{m} \frac{1}{\sigma^2 + \psi_j} + \frac{2\delta\sigma^2}{\sigma^2 + \psi_i} + \frac{\sigma^4}{\sigma^2 + \psi_i}, \end{aligned}$$

$$\begin{aligned} (\delta \mathbf{1}_m + \sigma^2 \mathbf{f}_i)' E^{-1} J_m E^{-1} (\delta \mathbf{1}_m + \sigma^2 \mathbf{f}_i) &= (\delta \mathbf{1}_m + \sigma^2 \mathbf{f}_i)' E^{-1} \mathbf{1}_m \mathbf{1}_m' E^{-1} (\delta \mathbf{1}_m + \sigma^2 \mathbf{f}_i) \\ &= \Big((\delta \mathbf{1}_m + \sigma^2 \mathbf{f}_i)' E^{-1} \mathbf{1}_m \Big)^2 \\ &= \Big(\delta \sum_{j=1}^{m} \frac{1}{\sigma^2 + \psi_j} + \frac{\sigma^2}{\sigma^2 + \psi_i} \Big)^2. \end{aligned}$$

Hence, for $i \in S$, by (2.24)

$$\begin{aligned}
\mathrm{MSE}[\hat{\theta}_i^{\star}] &= \sigma^2 + \delta - \delta^2 \sum_{j=1}^{m} \frac{1}{\sigma^2+\psi_j} - \frac{2\delta\sigma^2}{\sigma^2+\psi_i} - \frac{\sigma^4}{\sigma^2+\psi_i} + \\
&\quad \frac{\delta}{1+\delta\sum_{j=1}^{m}\frac{1}{\sigma^2+\psi_j}}\Big(\delta^2\Big(\sum_{j=1}^{m}\frac{1}{\sigma^2+\psi_j}\Big)^2 + \frac{\sigma^4}{(\sigma^2+\psi_i)^2} + \frac{2\delta\sigma^2}{\sigma^2+\psi_i}\sum_{j=1}^{m}\frac{1}{\sigma^2+\psi_j}\Big) \\
&= \sigma^2 - \frac{\sigma^4}{\sigma^2+\psi_i} + \Big(\delta - \frac{\delta^2\sum_{j=1}^{m}1/(\sigma^2+\psi_j)}{1+\delta\sum_{j=1}^{m}1/(\sigma^2+\psi_j)}\Big) + \frac{\delta\sigma^4 - 2\delta\sigma^2(\sigma^2+\psi_i)}{\Big(1+\delta\sum_{j=1}^{m}1/(\sigma^2+\psi_j)\Big)(\sigma^2+\psi_i)^2} \\
&= \frac{\sigma^2\psi_i}{\sigma^2+\psi_i} + \Big(\frac{1}{1+\delta\sum_{j=1}^{m}1/(\sigma^2+\psi_j)}\Big)\frac{(\delta\sigma^4 - 2\delta\sigma^2(\sigma^2+\psi_i) + \delta(\sigma^2+\psi_i)^2)}{(\sigma^2+\psi_i)^2} \\
&= \frac{\sigma^2\psi_i}{\sigma^2+\psi_i} + \frac{\delta\psi_i^2}{\Big(1+\delta\sum_{j=1}^{m}1/(\sigma^2+\psi_j)\Big)(\sigma^2+\psi_i)^2} \qquad (2.26)
\end{aligned}$$

For a non-sampled small area, that is, $i \in S^c$, note that

$$\begin{aligned}
\mathbf{f}_i'(\sigma^2 I_M + \delta J_M)\mathbf{f}_i &= \sigma^2 + \delta, \\
\mathbf{f}_i'\begin{pmatrix}\sigma^2 I_m + \delta J_m \\ \delta J_{M-m}\end{pmatrix} &= \delta\mathbf{1}_m', \\
\delta\mathbf{1}_m' E^{-1} J_m E^{-1}\delta\mathbf{1}_m &= \delta^2\mathbf{1}_m' E^{-1}\mathbf{1}_m\mathbf{1}_m' E^{-1}\mathbf{1}_m \\
&= \delta^2\Big(\sum_{j=1}^{m}\frac{1}{\sigma^2+\psi_j}\Big)^2
\end{aligned}$$

Hence by (2.24) for $i \in S^c$, we get

$$\begin{aligned}
\mathrm{MSE}[\hat{\theta}_i^{\star}] &= \sigma^2 + \delta - \delta^2\Big[\sum_{j=1}^{m}\frac{1}{\sigma^2+\psi_j} - \frac{\delta}{1+\delta\sum_{j=1}^{m}1/(\sigma^2+\psi_j)}\Big(\sum_{j=1}^{m}\frac{1}{\sigma^2+\psi_j}\Big)^2\Big] \\
&= \sigma^2 + \delta - \frac{\delta^2\sum_{j=1}^{m}1/(\sigma^2+\psi_j)}{1+\delta\sum_{j=1}^{m}1/(\sigma^2+\psi_j)} \qquad (2.27)
\end{aligned}$$

From (2.26) and (2.21), for $i \in S$, and assuming $\psi_i > 0$ (note that if $\psi_i = 0$, then since it was assumed that $\boldsymbol{\beta}$ is known, v_i is also known), the relative efficiency

of the two predictors $\hat{\theta}_i^{\star}$ and $\tilde{\theta}_i$ is

$$R(\hat{\theta}_i^{\star}, \tilde{\theta}_i) = \frac{\Big(1 + \delta \sum_{j=1}^{m} 1/(\sigma^2 + \psi_j)\Big)(\sigma^2 + \psi_i)^2(\sigma^2 + \delta)/(\sigma^2 + \delta + \psi_i)}{\sigma^2(\sigma^2 + \psi_i)\Big(1 + \delta \sum_{j=1}^{m} 1/(\sigma^2 + \psi_j)\Big) + \delta\psi_i} \tag{2.28}$$

$$\rightarrow \Big(1 + \frac{\delta}{\sigma^2}\Big)\Big(\frac{\sigma^2 + \psi_i}{\sigma^2 + \delta + \psi_i}\Big) \text{ as } m \rightarrow \infty$$

Moreover, by (2.27) and (2.22), for $i \in S^c$, the relative efficiency of the two predictors $\hat{\theta}_i^{\star}$ and $\tilde{\theta}_i$ is

$$R(\hat{\theta}_i^{\star}, \tilde{\theta}_i) = \frac{(\sigma^2 + \delta)\Big(1 + \delta \sum_{j=1}^{m} 1/(\sigma^2 + \psi_j)\Big)}{(\sigma^2 + \delta)\Big(1 + \delta \sum_{j=1}^{m} 1/(\sigma^2 + \psi_j)\Big) - \delta^2 \sum_{j=1}^{m} 1/(\sigma^2 + \psi_j)} \tag{2.29}$$

$$\rightarrow 1 + \frac{\delta}{\sigma^2}$$

As can be seen from (2.28)-(2.29), the relative efficiency of $\hat{\theta}_i^{\star}$ and $\tilde{\theta}_i$ depends on $\delta m/\sigma^2$ (this was the case for Table 2.1 as well). Also, for non-sampled small areas, the predictor derived from the misspecified Fay-Herriot can perform poorly with respect to $\hat{\theta}_i^{\star}$ if the small areas are strongly correlated. For the sampled small areas, the loss in efficiency by using $\tilde{\theta}_i$ is not as large as the non-sampled small areas (see Table 2.2).

For various values of m, δ, σ^2 and ψ_i (which we took to be the same for all small areas) we calculate (2.28), (2.29) and its limits. In Table 2.1 'Obs.' refers to (2.28) and 'Obs.lim' refers to the limit of (2.28). Similarly we define 'Unobs', 'Unobs.lim'. Note that when deriving (2.28), (2.29) we took the number of clusters to be 1. Hence m refers to the number of sampled small areas in a cluster. It is interesting to note that increasing ψ_i from 0.5 to 1.5 increases $R(\hat{\theta}_i^{\star}, \tilde{\theta}_i)$ for the sampled small areas. As we previously mentioned, $R(\hat{\theta}_i^{\star}, \tilde{\theta}_i)$ depends on $\delta m/\sigma^2$.

Table 2.2: $R(\hat{\theta}_i^{\star}, \tilde{\theta}_i)$ and its limit for sampled and non-sampled small areas . 'Obs.' is $R(\hat{\theta_i}^{\star}, \tilde{\theta}_i)$ given by (2.28), 'Obs.lim' is the limit of (2.28), 'Unobs.' is $R(\hat{\theta_i}^{\star}, \tilde{\theta}_i)$ given by (2.29), 'Unobs.lim' is the limit of (2.29).

m	δ	σ^2	ψ_i	Obs.	Obs.lim	Unobs.	Unobs.lim
20	0.6	0.4	0.5	1.418	1.500	2.263	2.500
20	0.6	0.4	1.5	1.635	1.900	2.075	2.500
20	0.3	0.7	0.5	1.110	1.143	1.333	1.429
20	0.3	0.7	1.5	1.166	1.257	1.281	1.429
40	0.6	0.4	0.5	1.456	1.500	2.371	2.500
40	0.6	0.4	1.5	1.748	1.900	2.252	2.500
40	0.3	0.7	0.5	1.125	1.143	1.375	1.429
40	0.3	0.7	1.5	1.203	1.257	1.340	1.429

Based on Tables 2.1, 2.2 and (2.28)-(2.29), we conclude this chapter by remarking that for the true model given by Assumption (C), (2.1)-(2.3) and (2.5)-(2.9), in order to achieve large relative efficiency, we require:

1. There be a large number of observations in each cluster (from our computational experience we believe n should be at least 20).

2. The cluster radius be small and δ/σ^2 be large. Note that we could interpret the results in Table 2.2 as having come from the model given by Assumption (C), (2.1)-(2.3) and (2.5)-(2.9) with extremely small cluster radius.

However, as will be shown in Chapter 3, for good estimation of the parameters

we require:

1. A large number of clusters (at least 20) as opposed to a large number of observations in each cluster.

2. The cluster radius be bounded away from 0, as opposed to small cluster radius which is needed for better prediction.

Chapter 3

Parameter estimation

In this chapter, estimation methods and asymptotic theory of the estimators of the fixed effects parameter $\boldsymbol{\beta}$ and the covariance parameter $\boldsymbol{\eta} = (\delta, \lambda, \sigma^2)'$ are discussed.

In spatial statistics and geostatistics, numerous estimation methods have been developed for estimation of covariance models. We give a short review of these methods.

A classical method of estimation is by the empirical variogram. The variogram $\gamma(\mathbf{h})$ of a stationary increments process $Z(\mathbf{s})$ is defined as

$$\gamma(\mathbf{h}) = \frac{1}{2}\text{var}(Z(\mathbf{s}+\mathbf{h}) - Z(\mathbf{s})).$$

An empirical variogram is estimated using the data, and then by visual inspection, a parametric variogram is selected and fitted to the empirical variogram by least squares or generalized least squares (Cressie [11] and Zhu and Stein [58]). However, if only the residuals of a spatial process are stationary, the mean function is estimated by ordinary least squares, and then the residuals are used to estimate

the empirical variogram. The empirical variogram, while popular, has certain limitations. For one, the empirical variogram is correlated at different lags, making it difficult to visually inspect and choose a model for the variogram. Moreover, parameter estimators obtained by fitting a model to the empirical variogram may not be consistent (Zhu and Stein [58])

A second method is maximum likelihood. Due to the complex nature of the variance-covariance matrix in spatial models, Mardia and Marshall [26], Richardson et al. [39] and Zimmerman and Harville [59] suggest using gradient algorithms such as Newton-Raphson or scoring to maximize the likelihood function. Such methods require the inverse of the variance-covariance matrix to be evaluated at each iteration. Note that, in general, inverting the variance-covariance matrix has computational time $O(m^3)$ (Mardia and Marshall [26]). However, by taking into account the covariance structure, the computational time could be reduced.

There are also iterative methods that combine two different methods of estimation. For example, a spatial autoregressive model is defined as

$$\mathbf{y} = X\boldsymbol{\beta} + \mathbf{u}$$

$$\mathbf{u} = \rho W\mathbf{u} + \boldsymbol{\epsilon}$$

where $\boldsymbol{\epsilon} \sim N(\mathbf{0}_m, \sigma^2 I_m)$ and W is a known weighting matrix which is usually row normalized to equal 1. For this model, Ord [35] suggested the following iterative method to estimate all parameters: estimate $\boldsymbol{\beta}$ by ordinary least squares, then substitute the residuals into the likelihood function to estimate ρ, then obtain a generalized least squares estimator of $\boldsymbol{\beta}$, recompute ρ, and iterate until (numerical)

convergence. Another method suggested by Ord [35] to estimate the parameters $\boldsymbol{\beta}$ and ρ is to consider the profile log likelihood in terms of ρ alone. After obtaining the estimate for ρ by maximizing the profile log likelihood, an estimate for $\boldsymbol{\beta}$ is obtained by generalized least squares.

Finally, Haining [19] and Richardson et al. [39] suggest a non-parametric method to estimate the variance-covariance matrix by using the residuals of the ordinary least squares to estimate the covariance function at various lags. This method, like the empirical variogram method, has its limitations.

Let $\boldsymbol{\kappa} = (\boldsymbol{\beta}, \boldsymbol{\eta})$ denote the vector of all parameters in the model given by Assumption (C), (2.1)-(2.3) and (2.5)-(2.9). As mentioned, since determining the MLE of $\boldsymbol{\kappa}$ involves a large amount of computational time, and due to certain technical reasons discussed in Section 3.3, alternate methods of estimation are considered. To further simplify matters, we do not jointly estimate all parameters, but instead develop routines to estimate only a subset of the parameters at a time.

3.1 Estimation of $(\boldsymbol{\beta}, \tau^2)$

The parameter τ^2 in our model is defined as

$$\tau^2 = \delta + \sigma^2. \tag{3.1}$$

An estimator $(\widehat{\boldsymbol{\beta}}_{\text{FH}}, \hat{\tau}^2_{\text{FH}})$ for $(\boldsymbol{\beta}, \tau^2)$ is given by (subscript of FH for Fay-Herriot)

$$(\widehat{\boldsymbol{\beta}}_{\text{FH}}, \hat{\tau}^2_{\text{FH}}) = \underset{\boldsymbol{\beta} \in \mathbb{R}^q, \tau^2 > 0}{\operatorname{argmax}} \, g(\boldsymbol{\beta}, \tau^2; \mathbf{y}) \tag{3.2}$$

where

$$g(\boldsymbol{\beta}, \tau^2; \mathbf{y}) = -\frac{m}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{m}\log(\tau^2 + \psi_i) - \frac{1}{2}\sum_{i=1}^{m}\frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2}{\tau^2 + \psi_i}. \tag{3.3}$$

Note that (3.3) is the log likelihood when the direct survey estimates y_i are assumed to follow the Fay-Herriot model. That is, we estimate $(\boldsymbol{\beta}, \tau^2)$ by maximizing a misspecified log likelihood. White [53] showed that under certain regularity conditions the parameter vector that maximizes the log likelihood is a consistent estimator of the parameter vector that minimizes the Kullback-Leibler Information Criterion (KLIC). The KLIC is defined as

$$I(f_1, f_2) = E_{f_1}\Big(\log\frac{f_1}{f_2}\Big) \tag{3.4}$$

where f_1,f_2 are respectively the true and misspecified joint densities of the observations. In our case, the true model is given by Assumption (C), (2.1)-(2.3) and (2.5)-(2.9), and minimizing the KLIC is equivalent to maximizing

$$\mathrm{E}[g(\boldsymbol{\beta}, \tau^2; \mathbf{y})] = -\frac{m}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{m}\log(\tau^2 + \psi_i) - \frac{1}{2}\sum_{i=1}^{m}\frac{\tau_o^2 + \psi_i + (\mathbf{x}_i'(\boldsymbol{\beta}_o - \boldsymbol{\beta}))^2}{\tau^2 + \psi_i},$$

where $\mathrm{E}(\cdot)$ is taken with respect to the true model. Note that the correlations in the true model do not enter the KLIC. Assuming X has full rank, the parameter vector that minimizes the KLIC is $(\boldsymbol{\beta}, \tau^2) = (\boldsymbol{\beta}_o, \tau_o^2)$ [see (3.12) for a proof]. Hence, one would expect $(\widehat{\boldsymbol{\beta}}_{\mathrm{FH}}, \hat{\tau}^2_{\mathrm{FH}})$ to be consistent. However, White's [53] theory is not applicable here as White [53] assumed that the true model consisted of independent observations. In Theorems 3.1 and 3.2, sufficient conditions are given for $(\widehat{\boldsymbol{\beta}}_{\mathrm{FH}}, \hat{\tau}^2_{\mathrm{FH}})$ to be consistent and asymptotically normal.

Note: For each of the theorems that follow, see Section 3.4 for the assumptions and some remarks and Section 3.5 for the proofs.

Theorem 3.1. *Suppose the true model for* $\{y_i : i \in S\}$ *is given by (2.1)-(2.3), (2.5)-(2.6). Suppose* $\tau_o^2 > 0$, m, $M \to \infty$ *such that* $0 < \lim_{m,M\to\infty} \frac{m}{M} < \infty$ *and* $(\mathrm{A1}) - (\mathrm{A5})$ *in Section 3.4.1 are satisfied. Then* $(\widehat{\boldsymbol{\beta}}_{FH}, \hat{\tau}^2_{FH})$ *is (locally) consistent for* $(\boldsymbol{\beta}_o, \tau_o^2)$.

Theorem 3.2. *Suppose in addition to the assumptions of Theorem 3.1, (A6)-(A8) in Section 3.4.1 are also satisfied. Then*

$$\begin{pmatrix} (X'D_o^{-1}V_oD_o^{-1}X)^{-\frac{1}{2}}X'D_o^{-1}X & \mathbf{0}_q \\ \mathbf{0}_q' & \frac{\sum_{i=1}^m 1/(\tau_o^2+\psi_i)^2}{\sqrt{2\,\mathrm{tr}(D_o^{-2}V_oD_o^{-2}V_o)}} \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{\beta}}_{FH} - \boldsymbol{\beta}_o \\ \hat{\tau}^2_{FH} - \tau_o^2 \end{pmatrix} \xrightarrow{d} \mathrm{N}(\mathbf{0}_{q+1}, \mathbf{I}_{q+1})$$

where $D_o = \mathrm{diag}(\tau_o^2 + \psi_1, \ldots, \tau_o^2 + \psi_m)$, $V_o = \Sigma_o + \Psi$, $\Sigma_o = \Sigma(\boldsymbol{\eta}_o) = \sigma_o^2 I_m + \delta_o A_o$, $A_o = A(\lambda_o)$, *and where* A *is the sub-matrix of* A_U *that corresponds to the sampled small areas and* $\boldsymbol{\eta}_o = (\delta_o, \lambda_o, \sigma_o^2)'$.

3.2 Estimation of (δ, λ)

An estimator $(\hat{\delta}, \hat{\lambda})$ for (δ, λ) is given by

$$(\hat{\delta}, \hat{\lambda}) = \underset{\delta\geq 0,\, \lambda\geq 0}{\mathrm{argmax}}\, h(\delta, \lambda; \mathbf{y}) \tag{3.5}$$

where

$$h(\delta, \lambda; \mathbf{y}) = -\sum_{l=1}^{k} \sum_{\substack{i,j\in C_l \\ i\neq j}} \left(\hat{\epsilon}_i\hat{\epsilon}_j - \delta \exp(-\lambda M^p \|\mathbf{z}_i - \mathbf{z}_j\|)\right)^2 \tag{3.6}$$

where $\hat{\epsilon}_i = y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}_o$ that satisfies (B8), and for $l = 1, \ldots, k$, C_l is the l^{th} cluster [see Assumption (C) and (2.7)-(2.8)].

Theorem 3.3. *Assume there are k clusters, $C_1, \ldots, C_k$, with cluster sizes $N_1, \ldots, N_k$ such that $\sum_{l=1}^{k} N_l = M$. From each cluster C_l, n_l of the N_l small areas are sampled such that $\sum_{l=1}^{k} n_l = m$. Suppose $\delta_o > 0$, $\lambda_o > 0$, the true model for $\{y_i : i \in S\}$ is given by (2.1)-(2.3), (2.5)-(2.6), and (B1)-(B7) in Section 3.4.2 are satisfied, then $(\hat{\delta}, \hat{\lambda})$ is (locally) consistent for (δ_o, λ_o).*

Theorem 3.4. *Suppose in addition to the assumptions of Theorem 3.3, (B8) in Section 3.4.2 is also satisfied, then*

$$\frac{\sum_{l=1}^{k} n_l^2}{(\sum_{l=1}^{k} n_l^4)^{\frac{1}{2}}} K_o^{-\frac{1}{2}} L_o \begin{pmatrix} \hat{\delta} - \delta_o \\ \hat{\lambda} - \lambda_o \end{pmatrix} \xrightarrow{d} \mathrm{N}(\mathbf{0}_2, \mathbf{I}_2)$$

where

$$K_o = \frac{8}{\sum_{l=1}^{k} n_l^4} \begin{pmatrix} \mathrm{tr}[G_o V_o G_o V_o] & -\mathrm{tr}[G_o V_o H_o V_o] \\ -\mathrm{tr}[G_o V_o H_o V_o] & \mathrm{tr}[H_o V_o H_o V_o] \end{pmatrix}$$

$$L_o = \frac{2}{\sum_{l=1}^{k} n_l^2} \begin{pmatrix} \mathrm{tr}[G_o^2] & -\mathrm{tr}[G_o H_o] \\ -\mathrm{tr}[G_o H_o] & \mathrm{tr}[H_o^2] \end{pmatrix}$$

$$V_o = \Sigma(\boldsymbol{\eta}_o) + \Psi = \sigma_o^2 I_m + \delta_o A_o + \Psi$$

$$A_{o,ij} = \exp(-\lambda_o M^p \|\mathbf{z}_i - \mathbf{z}_j\|)$$

$$G_{o,ij} = \begin{cases} \exp(-\lambda_o M^p \|\mathbf{z}_i - \mathbf{z}_j\|) & \text{if } i \neq j,\ i, j \in C_l \text{ for some } l, \\ 0 & \text{otherwise} \end{cases}$$

$$H_{o,ij} = \begin{cases} \delta_o M^p \|\mathbf{z}_i - \mathbf{z}_j\| \exp(-\lambda_o M^p \|\mathbf{z}_i - \mathbf{z}_j\|) & \text{if } i \neq j,\ i, j, \in C_l \text{ for some } l, \\ 0 & \text{otherwise} \end{cases}$$

where $A_{o,ij}$, $G_{o,ij}$, $H_{o,ij}$ are respectively the $(i,j)^{th}$ entries of A_o, G_o, H_o.

3.3 Maximum likelihood estimator

As mentioned in Section 2.1, for spatial models under infill asymptotics there are very few results regarding consistency and asymptotic normality of the MLE. The difficulty of showing such results is because, for general patterns of $\mathbf{z}_i$, there is no explicit expression for the inverse of the variance-covariance matrix of the random vector $\mathbf{y}$. Moreover, yet another technical difficulty is that for certain spatial models, the rates of convergence of the parameter estimators are not necessarily identical. For example, under infill asymptotics, Chen et al. [7] showed that only σ^2 and $\delta\lambda$ are consistently estimable in model (2.4). Moreover, they showed that the MLE for σ^2 is $m^{\frac{1}{2}}$-consistent and the MLE for $\delta\lambda$ is $m^{\frac{1}{4}}$-consistent. This result was shown under the assumption that the z_i's are situated on a lattice in $[0, 1]$. Under such an assumption, it is possible to explicitly write the inverse of the variance-covariance matrix. Moreover, other results that show consistency and asymptotic normality of the MLE in spatial models, assume similar restrictive conditions on the spatial patterns to be able to write the inverse of the variance-covariance matrix in a manageable form (Loh and Lam [25] and Ying [55]). We encounter similar technical difficulties in trying to show the MLE is consistent and asymptotically normal. However, we conjecture that for the model given by Assumption (C), (2.1)-(2.3) and (2.5)-(2.9), the MLE $\widehat{\boldsymbol{\kappa}}_{\text{ML}}$ of $\boldsymbol{\kappa} = (\boldsymbol{\beta}, \boldsymbol{\eta})$ is consistent and

$$\Big(\mathcal{I}(\boldsymbol{\kappa})\Big)^{\frac{1}{2}}(\widehat{\boldsymbol{\kappa}} - \boldsymbol{\kappa}_o) \xrightarrow{d} \text{N}(\mathbf{0}_{q+3}, I_{q+3}) \tag{3.7}$$

where $\mathcal{I}(\boldsymbol{\kappa})$ is the information matrix, and using Lemma 3.7 (c), it is given by

$$\mathcal{I}(\boldsymbol{\kappa}) = \begin{pmatrix} X'V_oX & \mathbf{0}_q & \mathbf{0}_q & \mathbf{0}_q \\ \mathbf{0}_q' & \frac{1}{2}\mathrm{tr}[V_o^{-1}A_oV_o^{-1}A_o] & \frac{\delta_o}{2}\mathrm{tr}[V_o^{-1}A_oV_o^{-1}B_o] & \frac{1}{2}\mathrm{tr}[V_o^{-1}A_oV^{-1}] \\ \mathbf{0}_q' & \frac{\delta_o}{2}\mathrm{tr}[V_o^{-1}A_oV_o^{-1}B_o] & \frac{\delta_o^2}{2}\mathrm{tr}[V_o^{-1}B_oV_o^{-1}B_o] & \frac{\delta_o}{2}\mathrm{tr}[V_o^{-1}B_oV^{-1}] \\ \mathbf{0}_q' & \frac{1}{2}\mathrm{tr}[V_o^{-1}A_oV_o^{-1}] & \frac{\delta_o}{2}\mathrm{tr}[V_o^{-1}B_oV_o^{-1}] & \frac{1}{2}\mathrm{tr}[V_o^{-2}] \end{pmatrix} \tag{3.8}$$

where for $i, j = 1, \ldots, m$, $B_{o,ij} = -M^p\|\mathbf{z}_i - \mathbf{z}_j\| \exp(-M^p\lambda_o\|\mathbf{z}_i - \mathbf{z}_j\|)$.

The above conjecture is based on:

1. For the balanced one way random effects model, or equivalently, when the variance-covariance model for the small area effects is given by (2.5) and (2.16) with no sampling errors, $N_l = N$, $n_l = n$, and β is an intercept, consistency and asymptotic normality of the MLE follows as a special case of Miller [29]:

$$\begin{pmatrix} k^{\frac{1}{2}} & 0 & 0 \\ 0 & k^{\frac{1}{2}} & 0 \\ 0 & 0 & m^{\frac{1}{2}} \end{pmatrix} \begin{pmatrix} \hat{\beta}_{\mathrm{ML}} - \beta_o \\ \hat{\delta}_{\mathrm{ML}} - \delta_o \\ \hat{\sigma}^2_{\mathrm{ML}} - \sigma^2_o \end{pmatrix} \xrightarrow{d} \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \delta_o & 0 & 0 \\ 0 & 2\delta_o^2 & 0 \\ 0 & 0 & 2\sigma_o^4 \end{pmatrix} \right)$$

   Note that $\hat{\sigma}^2_{\mathrm{ML}}$ is $\sqrt{m}$-consistent, but $\hat{\beta}_{\mathrm{ML}}$ and $\hat{\delta}_{\mathrm{ML}}$ are only $\sqrt{k}$-consistent.

2. Under infill asymptotics, since some of the parameters are not consistently estimable, one cannot expect (3.7) to hold in general. However, this is not the case for our model. That is, we have already shown that all parameters are consistently estimable.

3. Simulation results (see Section 4.4) indicate that for large k, the empirical variance of the MLE of $\widehat{\boldsymbol{\eta}}$ matches the conjectured variance obtained by inverting the information matrix [(3.8)].

3.4 Assumptions and remarks

3.4.1 Theorems 3.1 and 3.2

(A1) X has full rank.

(A2) ψ_i's are bounded such that for all i, $0 \leq \psi_i \leq \psi_c$, for some $\psi_c < \infty$.

(A3) $0 < \lim_{m\to\infty}(1/m)\sum_{i=1}^{m} \|\mathbf{x}_i\|^2 < \infty$.

(A4) $\lim_{m\to\infty}(\delta_o/m^2)\sum_{i\neq j} \|\mathbf{x}_i\|\|\mathbf{x}_j\| \exp(-M^p\lambda_o\|\mathbf{z}_i - \mathbf{z}_j\|) = 0$.

(A5) $\lim_{m\to\infty}(\delta_o^2/m^2)\sum_{i\neq j} \exp(-2M^p\lambda_o\|\mathbf{z}_i - \mathbf{z}_j\|) = 0$.

(A6) $\lim_{m\to\infty} \gamma_{\max}(D_o^{-1}V_oD_o^{-1})/(\sum_{i=1}^{m}\gamma_i^2)^{\frac{1}{2}} = 0$, where the γ_i's are the eigenvalues of $D_o^{-1}V_oD_o^{-1}$.

(A7) $X'V_oX$ has entries of the order $O(\mathrm{tr}[V_o^2])$.

(A8) $(X'X)^{-1}$ has entries that are $O(1/m)$.

Remark 1:

Neither of the Theorems 3.1-3.2 require Assumption (C) and (2.7)-(2.9). However, these assumptions are needed when deriving the asymptotic theory of the estimators for δ and λ.

Remark 2:

Assumption (A3) is satisfied if for $i = 1, \ldots, m$, $\|\mathbf{x}_i\| < \infty$. To derive the asymptotic distribution of $(\widehat{\boldsymbol{\beta}}_{\mathrm{FH}}, \hat{\tau}^2_{\mathrm{FH}})$, we require that the y_i's are uniformly asymptotically negligible [(A6)]. (A7) and (A8) are needed to bound (3.27) in probability.

Remark 3:

Even though we have assumed that the covariance model for the small area effects is given by (2.5)-(2.6), we can relax this assumption. Instead of (A4)-(A5),

it suffices to assume that the off-diagonal entries σ_{ij} of the covariance matrix of the small area effects satisfy

(A4′) $\lim_{m\to\infty}(1/m^2)\sum_{i\neq j}\|\mathbf{x}_i\|\|\mathbf{x}_j\|\sigma_{ij}=0$

(A5′) $\lim_{m\to\infty}(1/m^2)\sum_{i\neq j}(\sigma_{ij})^2=0$

Remark 4:

If in addition to the assumptions of Theorems 3.1 and 3.2, also Assumption (C) and (2.7)-(2.8) hold, then (A3)-(A6) are satisfied if for $i=1,\ldots,m$, $\|\mathbf{x}_i\|<\infty$, and

$$\frac{\max_{1\leq l\leq k} n_l}{(\sum_{l=1}^{k} n_l^2)^{\frac{1}{2}}}\to 0.$$

From Theorem 3.2 the asymptotic variance of $\widehat{\boldsymbol{\beta}}_{\text{FH}}$ is

$$\text{var}(\widehat{\boldsymbol{\beta}}_{\text{FH}})=(X'D_o^{-1}X)^{-1}X'D_o^{-1}V_oD_o^{-1}X(X'D_o^{-1}X)^{-1}.$$

By (A7), $X'V_oX$ has entries that are $O(\text{tr}[V_o^2])$. Also,

$$\text{tr}[V_o^2]=\sum_{i,j=1}^{m}V_{o,ij}^2=\sum_{l=1}^{k}\sum_{i,j\in C_l}V_{o,ij}^2+\sum_{l_1\neq l_2}\sum_{\substack{i\in C_{l_1}\\ j\in C_{l_2}}}V_{o,ij}^2. \tag{3.9}$$

where $V_{o,ij}$ is the $(i,j)^{th}$ entry of $V_o=\sigma_o^2I_m+\delta_oA_o+\Psi$. But

$$\begin{aligned}\sum_{l_1\neq l_2}\sum_{\substack{i\in C_{l_1}\\ j\in C_{l_2}}}V_{o,ij}^2 &\leq \sum_{l_1\neq l_2}\sup_{\substack{i\in C_{l_1}\\ j\in C_{l_2}}}\exp(-2\lambda_oM^p\|\mathbf{z}_i-\mathbf{z}_j\|)\,n_{l_1}n_{l_2}\\ &\leq \sup_{l_1\neq l_2}\sup_{\substack{i\in C_{l_1}\\ j\in C_{l_2}}}\exp(-2\lambda_oM^p\|\mathbf{z}_i-\mathbf{z}_j\|)\,m^2\\ &= \sup_{l_1\neq l_2}\sup_{\substack{i\in C_{l_1}\\ j\in C_{l_2}}}M\exp(-2\lambda_oM^p\|\mathbf{z}_i-\mathbf{z}_j\|)\frac{m}{M}\,m=O(m)\ \text{[by (2.8)]}\end{aligned} \tag{3.10}$$

Hence, under Assumption (C) and (2.7)-(2.8), formulas (3.9)-(3.10) imply that $O(\mathrm{tr}[V_o^2]) = O(\sum_l^k n_l^2)$. Now by (A8), it follows that for $i = 1, \ldots, q$, $\mathrm{var}(\widehat{\beta}_i) = O(\sum_l^k n_l^2/m^2)$. If we further assume that all the n_i's grow at the same rate, that is there exists n such that for $l = 1, \ldots, k$, $0 < \lim_{n_l, n \to \infty} n_l/n < \infty$, then for $i = 1, \ldots, q$, $\mathrm{var}(\widehat{\beta}_i) = O(1/k)$.

Also, from Theorem 3.2 the asymptotic variance of $\hat{\tau}^2_{\mathrm{FH}}$ is

$$\mathrm{var}(\hat{\tau}^2_{\mathrm{FH}}) = \frac{2\mathrm{tr}(D_o^{-2} V_o D_o^{-2} V_o)}{\left(\sum_{i=1}^m \frac{1}{(\tau_o^2 + \psi_i)^2} \right)^2}$$

By (2.7)-(2.8), and the last several paragraphs, $\mathrm{var}(\hat{\tau}^2_{\mathrm{FH}}) = O(\sum_l^k n_l^2/m^2)$. Again, if we further assume that all the n_i's grow at the same rate, then for $i = 1, \ldots, q$, $\mathrm{var}(\widehat{\beta}_i) = O(1/k)$.

Remark 5: In practice, we could verify (A4),(A5) by checking that for all $c < \infty$,

$$\frac{1}{m^2} \sum_{i \neq j} I_{[M^p \|\mathbf{z}_i - \mathbf{z}_j\| \leq c]} \to 0. \tag{3.11}$$

Note that Assumption (C), (2.7)-(2.8) imply (3.11).

3.4.2 Theorems 3.3 and 3.4

(B1) As $M \to \infty$, also $m \to \infty$ such that $0 < \lim_{m, M \to \infty} m/M < \infty$ and $k \to \infty$, and for $l = 1, \ldots, k$, $n_l, N_l \to \infty$ such that $0 < \lim_{n_l, N_l \to \infty} n_l/N_l < \infty$.

(B2) For $l = 1, \ldots, k$,

$$\limsup_{M \to \infty} M^p \sup_{i,j \in C_l} \|\mathbf{z}_i - \mathbf{z}_j\| < \infty,$$

and for all $l_1 \neq l_2$,

$$\liminf_{M \to \infty} \frac{M^p}{\log M} \inf_{i \in C_{l_1}, j \in C_{l_2}} \|\mathbf{z}_i - \mathbf{z}_j\| = \infty.$$

(B3) For $l = 1, \ldots, k$, $\exists\ c_l$ such that

$$\lim_{N_l \to \infty} \frac{1}{N_l^2} \sum_{i,j \in C_l} I_{[M^p \|\mathbf{z}_i - \mathbf{z}_j\| \geq c_l]} = \epsilon_l > 0,$$

where $\mathbf{z}_i$ are in a finite dimensional space.

(B4) For $l = 1, \ldots, k$, $0 < \lim_{N_l \to \infty} (1/N_l) \sum_{i=1}^{N_l} \|\mathbf{x}_i\| < \infty$.

(B5) For $l = 1, \ldots, k$, $0 < \lim_{N_l \to \infty} (1/N_l) \sum_{i=1}^{N_l} \|\mathbf{x}_i\|^2 < \infty$.

(B6) Assume for $l = 1, \ldots, k, N_l \to \infty$ such that $\lim_{k \to \infty} \frac{\max_{1 \leq l \leq k} N_l^2}{(\sum_{l=1}^k N_l^4)^{\frac{1}{2}}} = 0$.

(B7) ψ_i's are bounded such that for all i, $0 \leq \psi_i \leq \psi_c$, for some $\psi_c < \infty$.

(B8) $\frac{\sum_{l=1}^k N_l^2}{(\sum_{l=1}^k N_l^4)^{\frac{1}{2}}} \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_o\|^2 \xrightarrow{p} 0$, for some consistent estimator $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}_o$.

Remark 6:

For the purely spatial analog of the model given by Assumption (C), (2.1)-(2.3) and (2.5)-(2.9) - that is, the model with no sampling errors - the asymptotic distribution of $(\widehat{\boldsymbol{\beta}}_{\text{FH}}, \hat{\tau}^2_{\text{FH}})$ and $(\hat{\delta}, \hat{\lambda})$ can be obtained from Theorems 3.2 and 3.4 by taking $\Psi = O_m$.

Remark 7:

The factor $\log M$ in (B2) is needed to disregard the contribution of the between cluster terms when showing consistency and asymptotic normality. The asymptotic properties of parameter estimators for (δ, λ) are derived when the true covariance model for the small area effects is given by (2.5)-(2.6). However, because of assumption (B2), even if the parameter estimators were derived under the assumption that the covariance model for the small area effects is given by the model (2.5) and (2.10), the estimators for (δ, λ) derived under this misspecification would have the

same asymptotic properties as the estimators given in Theorems 3.3 and 3.4.

Remark 8:

If the $\mathbf{z}_i$'s are equally spaced from one another, the data contains no information on the parameter λ. Hence, for $\mathbf{z}_i$ in a finite dimensional space, it is required for each cluster that the $\mathbf{z}_i$'s do not collapse to a point [(B3)]. We could generalize (B3) by not assuming the $\mathbf{z}_i$'s are in a finite dimensional space, but instead assume for $l = 1, \ldots, k$, the distribution of $\mathbf{z}_i$ in cluster C_l is such that

$$\lim_{N_l \to \infty} \frac{1}{N_l^2} \mathrm{var}\Big(\sum_{\substack{i,j \in C_l \\ i \neq j}} M^p \|\mathbf{z}_i - \mathbf{z}_j\| \Big) > 0$$

Remark 9:

Even though we have assumed (B8) for a consistent estimator $\widehat{\boldsymbol{\beta}}$, for $\widehat{\boldsymbol{\beta}}_{\mathrm{FH}}$ given in Theorems 3.1 and 3.2 this assumption is automatically satisfied. This assertion follows from Remark 4, (B1), (B6), Lemma 3.4 (c) and noting that $m = \sum_{l=1}^{k} n_l$, $M = \sum_{l=1}^{k} N_l$. Moreover, note that the asymptotic variance of $(\hat{\delta}, \hat{\lambda})$ given in Theorem 3.4 does not depend on the estimator $\widehat{\boldsymbol{\beta}}$. That is, for any estimator $\widehat{\boldsymbol{\beta}}$ that satisfies (B8), the asymptotic distribution of $(\hat{\delta}, \hat{\lambda})$ would be the same. In particular, we have shown that the asymptotic distribution of $(\hat{\delta}, \hat{\lambda})$ would be the same if $\boldsymbol{\beta}_o$ were known.

Remark 10:

The variance of $\hat{\delta}$ and $\hat{\lambda}$ is of the order $O\Big(\frac{\sum_{l=1}^{k} n_l^4}{(\sum_{l=1}^{k} n_l^2)^2}\Big)$. If all the n_i's grow at the same rate, that is, there exists n such that for $l = 1, \ldots, k$, $0 < \lim_{n_l, n \to \infty} \frac{n_l}{n} < \infty$,

then $\text{var}(\hat{\delta}) = O(1/k)$ and $\text{var}(\hat{\lambda}) = O(1/k)$.

Remark 11:

Since $\tau^2 = \delta + \sigma^2$, we can estimate σ^2 by $\hat{\sigma}^2 = \hat{\tau}^2_{\text{FH}} - \hat{\delta}$, which is a consistent estimator of σ^2_o. However, we do not have a formula for $\text{var}(\hat{\sigma}^2)$.

Remark 12:

(B6) is a uniform asymptotic negligibility condition similar to the one given in Remark 4 and (A6). However, (B6) is more restrictive than the condition in Remark 4, an assertion which follows from (B1) and Lemma 3.4 (a).

3.5 Proofs

The proofs for consistency of parameter estimators given in this section use two theorems given in Andersen and Gill [2] and van der Vaart [47], and are stated for convenience.

Theorem 3.5. *(Andersen and Gill [2]) Let E be an open convex set of $\mathbb{R}^s$, and let $F_1, F_2, \ldots,$ be a sequence of random strictly concave functions on E such that $\forall\, \mathbf{a} \in E$, $F_m(\mathbf{a}) \xrightarrow{p} f(\mathbf{a})$ as $m \to \infty$ where f is some real function on E. Then f is also concave and for all compact $H \subset E$,*

$$\sup_{\mathbf{a} \in H} |F_m(\mathbf{a}) - f(\mathbf{a})| \xrightarrow{p} 0.$$

Theorem 3.6. *(van der Vaart [47], p.45) Let F_m be a sequence of random functions and let f be some fixed function of $\mathbf{a}$ such that for every $\epsilon > 0$*

$$\sup_{\mathbf{a}\in H} |F_m(\mathbf{a}) - f(\mathbf{a})| \xrightarrow{p} 0$$

$$\sup_{\mathbf{a}:\|\mathbf{a}-\mathbf{a}_o\|\geq\epsilon} f(\mathbf{a}) < f(\mathbf{a}_o),$$

where H is as given in Theorem 3.5. Then any sequence of estimators $\hat{\mathbf{a}}_m$ with $F_m(\hat{\mathbf{a}}_m) \geq F_m(\mathbf{a}_o) - o_p(1)$ converges in probability to $\mathbf{a}_o$.

We state and prove a few lemmas that are used in the remarks and proofs of the main theorems in this chapter.

Lemma 3.1. *If* $\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathrm{N}\Big(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix} \Big)$, *then* $\mathrm{E}(X^2Y^2) = \sigma_x^2\sigma_y^2 + 2\sigma_{xy}^2$.

Proof of Lemma 3.1. Note that $X - \frac{\sigma_{xy}}{\sigma_y^2}Y \sim \mathrm{N}\Big(0, \sigma_x^2 - \frac{\sigma_{xy}^2}{\sigma_y^2}\Big)$ is independent of Y. Hence,

$$\begin{aligned}\mathrm{E}(X^2Y^2) &= \mathrm{E}\Big[\Big(X - \frac{\sigma_{xy}}{\sigma_y^2}Y + \frac{\sigma_{xy}}{\sigma_y^2}Y\Big)^2 Y^2\Big] = \mathrm{E}\Big[\Big(X - \frac{\sigma_{xy}}{\sigma_y^2}Y\Big)^2 Y^2\Big] + \mathrm{E}\Big[\frac{\sigma_{xy}^2}{\sigma_y^4}Y^4\Big] \\ &= \Big(\sigma_x^2 - \frac{\sigma_{xy}^2}{\sigma_y^2}\Big)\sigma_y^2 + \frac{\sigma_{xy}^2}{\sigma_y^4}3\sigma_y^4 = \sigma_x^2\sigma_y^2 + 2\sigma_{xy}^2\end{aligned}$$

□

Lemma 3.2. *Let $a_1, \ldots, a_n$ be a sequence of positive numbers such that*

$$\lim_{n\to\infty} \frac{\max_{1\leq i\leq n} a_i}{(\sum_{i=1}^n a_i^2)^{\frac{1}{2}}} = 0.$$

If c is any positive number such that $c > 2$, then

$$\lim_{n\to\infty} \frac{\sum_{i=1}^n a_i^c}{\Big(\sum_{i=1}^n a_i^2\Big)^{\frac{c}{2}}} = 0.$$

Proof of Lemma 3.2.

$$\frac{\sum_{i=1}^n a_i^c}{\Big(\sum_{i=1}^n a_i^2\Big)^{\frac{c}{2}}} = \frac{(\max_{1\le i\le n} a_i)^{c-2}}{\Big(\sum_{i=1}^n a_i^2\Big)^{\frac{c-2}{2}}} \frac{\sum_{i=1}^n a_i^c/(\max_{1\le i\le n} a_i)^{c-2}}{\Big(\sum_{i=1}^n a_i^2\Big)}$$
$$< \Big(\frac{\max_{1\le i\le n} a_i}{(\sum_{i=1}^n a_i^2)^{\frac{1}{2}}}\Big)^{c-2} \frac{\sum_{i=1}^n a_i^c/a_i^{c-2}}{\sum_{i=1}^n a_i^2} \quad \to 0 \text{ as } n\to\infty.$$

□

Lemma 3.3. *Let $a_1,\dots,a_n$ and $b_1,\dots,b_n$ be any real numbers, then*

$$\sum_{i=1}^n a_i \sum_{i=1}^n a_i b_i^2 - \Big(\sum_{i=1}^n a_i b_i\Big)^2 = \frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n a_i a_j (b_i-b_j)^2$$

Proof of Lemma 3.3.

$$\sum_{i=1}^n a_i \sum_{i=1}^n a_i b_i^2 - \Big(\sum_{i=1}^n a_i b_i\Big)^2 = \sum_{i,j=1}^n a_i a_j b_j^2 - \sum_{i,j=1}^n a_i a_j b_i b_j$$
$$= \sum_{i>j} a_i a_j (b_i-b_j)^2 = \frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n a_i a_j (b_i-b_j)^2$$

□

Lemma 3.4. *For k and $N_l,\dots,N_k$ given in (B1), assume for $l=1,\dots,k$, $N_l\to\infty$ such that*

$$\lim_{k\to\infty} \frac{\max_{1\le l\le k} N_l^2}{(\sum_{l=1}^k N_l^4)^{\frac{1}{2}}} = 0.$$

Then as $k\to\infty$, the following three ratios all tend to 0:

$$(a)\ \frac{\max_{1\le l\le k} N_l}{(\sum_{l=1}^k N_l^2)^{\frac{1}{2}}} \qquad (b)\ \frac{\sum_{l=1}^k N_l^4}{(\sum_{l=1}^k N_l^2)^2} \qquad (c)\ \frac{\sum_{l=1}^k N_l^2}{\sum_{l=1}^k N_l (\sum_{l=1}^k N_l^4)^{\frac{1}{4}}}.$$

Proof of Lemma 3.4.

$$(a) \qquad \frac{\max_{1\le l\le k} N_l}{(\sum_{l=1}^k N_l^2)^{\frac{1}{2}}} = \Big(\frac{\max_{1\le l\le k} N_l^2}{\sum_{l=1}^k N_l^2}\Big)^{\frac{1}{2}} < \Big(\frac{\max_{1\le l\le k} N_l^2}{(\sum_{l=1}^k N_l^4)^{\frac{1}{2}}}\Big)^{\frac{1}{2}} \to 0.$$

$$(b) \qquad \frac{\sum_{l=1}^k N_l^4}{(\sum_{l=1}^k N_l^2)^2} < \frac{\sum_{l=1}^k N_l^2 \max_{1\le l\le k} N_l^2}{(\sum_{l=1}^k N_l^2)^2} \to 0 \text{ [by } (a)].$$

$$(c) \qquad \frac{\sum_{l=1}^k N_l^2}{\sum_{l=1}^k N_l(\sum_{l=1}^k N_l^4)^{\frac{1}{4}}} < \frac{\sum_{l=1}^k N_l \max_{1\le l\le k} N_l}{\sum_{l=1}^k N_l(\sum_{l=1}^k N_l^4)^{\frac{1}{4}}} = \Big(\frac{\max_{1\le l\le k} N_l^2}{(\sum_{l=1}^k N_l^4)^{\frac{1}{2}}}\Big)^{\frac{1}{2}} \to 0.$$

□

The following results are from Rencher [38] and McCulloch and Searle [33].

Lemma 3.5. *(Rencher [38]) If G and H are any $m \times m$ matrices, then the eigenvalues of GH are the same as those of HG.*

Lemma 3.6. *(McCulloch and Searle [33]) Assume $\mathbf{y} \sim \mathrm{N}(\boldsymbol{\mu}, G)$, where $G = G(\boldsymbol{\varphi})$, $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_l)$. Let $L(\boldsymbol{\varphi}; \mathbf{y})$ denote the log likelihood, and H be a symmetric matrix, then*

$$(a) \qquad \mathrm{E}(\mathbf{y}'H\mathbf{y}) = \mathrm{tr}(HG) + \boldsymbol{\mu}'H\boldsymbol{\mu}$$

$$(b) \qquad \mathrm{var}(\mathbf{y}'H\mathbf{y}) = 2\,\mathrm{tr}[(HG)^2] + 4\boldsymbol{\mu}'HGH\boldsymbol{\mu}$$

$$(c) \qquad -\mathrm{E}\Big(\frac{\partial^2 L(\boldsymbol{\varphi};\mathbf{y})}{\partial\varphi_i\partial\varphi_j}\Big) = \frac{1}{2}\mathrm{tr}\Big(G^{-1}\frac{\partial G}{\partial\varphi_i}G^{-1}\frac{\partial G}{\partial\varphi_j}\Big)$$

$$(d) \qquad \frac{\partial \log|G|}{\partial\varphi_i} = \mathrm{tr}\Big(G^{-1}\frac{\partial G}{\partial\varphi_i}\Big)$$

$$(e) \qquad \frac{\partial G^{-1}}{\partial\varphi_i} = -G^{-1}\frac{\partial G}{\partial\varphi_i}G^{-1}$$

Proof of Theorem 3.1. Let $\boldsymbol{\zeta} = (\boldsymbol{\beta}, \tau^2)' = (\zeta_1, \ldots, \zeta_q, \zeta_{q+1})' \in \mathbb{R}^{q+1}$, and $\boldsymbol{\zeta}_o = (\boldsymbol{\beta}_o, \tau_o^2)' = (\zeta_{o1}, \ldots, \zeta_{oq}, \zeta_{o(q+1)})'$. Also, let

$$g(\boldsymbol{\zeta}; \mathbf{y}) = -\frac{m}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{m}\log(\tau^2 + \psi_i) - \frac{1}{2}\sum_{i=1}^{m}\frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2}{\tau^2 + \psi_i}$$

$$\begin{aligned} g(\boldsymbol{\zeta}) &= \mathrm{E}(g(\boldsymbol{\zeta}; \mathbf{y})) \\ &= -\frac{m}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{m}\log(\tau^2 + \psi_i) - \frac{1}{2}\sum_{i=1}^{m}\frac{\tau_o^2 + \psi_i + (\mathbf{x}_i'(\boldsymbol{\beta}_o - \boldsymbol{\beta}))^2}{\tau^2 + \psi_i} \end{aligned}$$

Consistency is established by showing that there exists a sequence $\widehat{\boldsymbol{\zeta}}$ $(= \widehat{\boldsymbol{\zeta}}_m)$ of local maxima of $\frac{1}{m}g(\boldsymbol{\zeta}; \mathbf{y})$ which is consistent for $\boldsymbol{\zeta}_o$. The proof involves showing (i) $\frac{1}{m}g(\boldsymbol{\zeta})$ has a unique maximum at $\boldsymbol{\zeta} = \boldsymbol{\zeta}_o$; (ii) for all $\boldsymbol{\zeta}$ in a sufficiently small non-shrinking neighborhood of $\boldsymbol{\zeta}_o$, $\frac{1}{m}(g(\boldsymbol{\zeta}; \mathbf{y}) - g(\boldsymbol{\zeta})) \xrightarrow{p} 0$ and (iii) the hessian of $\frac{1}{m}g(\boldsymbol{\zeta}; \mathbf{y})$, $\frac{1}{m}\nabla_{\zeta\zeta}g(\boldsymbol{\zeta}; \mathbf{y})$, is negative definite. Then, it follows from Theorems 3.5 and 3.6 that $\widehat{\boldsymbol{\zeta}} \xrightarrow{p} \boldsymbol{\zeta}_o$. For better readability, we indicate the steps given above in the proof.

Step (i): It is shown that $\frac{1}{m}g(\boldsymbol{\zeta})$ has a unique maximum at $\boldsymbol{\zeta} = \boldsymbol{\zeta}_o$. For fixed τ^2, since X has full rank, $\frac{1}{m}g(\boldsymbol{\zeta})$ is maximized at $\boldsymbol{\beta} = \boldsymbol{\beta}_o$. Consider

$$\frac{1}{m}g(\boldsymbol{\beta}_o, \tau^2) = -\frac{1}{2}\log 2\pi - \frac{1}{2m}\sum_{i=1}^{m}\log(\tau^2 + \psi_i) - \frac{1}{2m}\sum_{i=1}^{m}\frac{\tau_o^2 + \psi_i}{\tau^2 + \psi_i} \tag{3.12}$$

$$\frac{1}{m}\frac{\partial g(\boldsymbol{\beta}_o, \tau^2)}{\partial \tau^2} = -\frac{1}{2m}\sum_{i=1}^{m}\frac{1}{\tau^2 + \psi_i} + \frac{1}{2m}\sum_{i=1}^{m}\frac{\tau_o^2 + \psi_i}{(\tau^2 + \psi_i)^2}$$

For $\tau^2 < \tau_o^2$, $\frac{1}{m}\frac{\partial g(\boldsymbol{\beta}_o, \tau^2)}{\partial \tau^2} > 0$ and for $\tau^2 > \tau_o^2$, $\frac{1}{m}\frac{\partial g(\boldsymbol{\beta}_o, \tau^2)}{\partial \tau^2} < 0$, and since $\frac{1}{m}\frac{\partial g(\boldsymbol{\beta}_o, \tau_o^2)}{\partial \tau^2} = 0$, it follows that $\frac{1}{m}g(\boldsymbol{\beta}_o, \tau^2)$ is maximized at $\tau^2 = \tau_o^2$, and $\frac{1}{m}g(\boldsymbol{\beta}, \tau^2)$ has a unique maximum at $\boldsymbol{\zeta} = \boldsymbol{\zeta}_o$.

Step (ii): For any small $\epsilon > 0$, consider the following neighborhood of $\boldsymbol{\zeta}_o$ (by (A3)

the neighborhood is non-shrinking with m):

$$B_\epsilon = \{(\boldsymbol{\beta}, \tau^2) : \|\boldsymbol{\beta} - \boldsymbol{\beta}_o\| < \frac{\tau^4 \epsilon}{\frac{1}{m}\sum_{i=1}^m \|\mathbf{x}_i\|^2}, |\tau^2 - \tau_o^2| < \frac{9}{10}\tau_0^2\}$$

For all $\boldsymbol{\zeta} \in B_\epsilon$, consider

$$\frac{1}{m}(g(\boldsymbol{\zeta}; \mathbf{y}) - g(\boldsymbol{\zeta})) = -\frac{1}{2m}\sum_{i=1}^m \frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta}_o)^2 - (\tau_o^2 + \psi_i)}{\tau^2 + \psi_i} - \frac{1}{m}\sum_{i=1}^m \frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta}_o)\mathbf{x}_i'(\boldsymbol{\beta}_o - \boldsymbol{\beta})}{\tau^2 + \psi_i}.$$

Note that

$$\begin{aligned}
&\mathrm{var}\Big(\frac{1}{m}\sum_{i=1}^m \frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta}_o)^2 - (\tau_o^2 + \psi_i)}{\tau^2 + \psi_i}\Big) \\
&= \frac{1}{m^2}\Big(\sum_{i=1}^m \frac{\mathrm{var}[(y_i - \mathbf{x}_i'\boldsymbol{\beta}_o)^2]}{(\tau^2 + \psi_i)^2} + \sum_{i \neq j} \frac{\mathrm{cov}[(y_i - \mathbf{x}_i'\boldsymbol{\beta}_o)^2, (y_j - \mathbf{x}_j'\boldsymbol{\beta}_o)^2]}{(\tau^2 + \psi_i)(\tau^2 + \psi_j)}\Big) \\
&= \frac{1}{m^2}\Big(\sum_{i=1}^m \frac{2(\tau_o^2 + \psi_i)^2}{(\tau^2 + \psi_i)^2} + \sum_{i \neq j} \frac{2\delta_o^2 \exp(-2M^p\lambda_o\|\mathbf{z}_i - \mathbf{z}_j\|)}{(\tau^2 + \psi_i)(\tau^2 + \psi_j)}\Big) \qquad \text{(by Lemma 3.1)} \\
&\leq \frac{1}{\tau^4 m^2}\Big(2m(\tau_o^2 + \psi_c)^2 + 2\delta_o^2 \sum_{i \neq j} \exp(-2M^p\lambda_o\|\mathbf{z}_i - \mathbf{z}_j\|)\Big) \qquad \text{(by (A2))} \\
&\to 0 \ \text{ as } \ m \to \infty \qquad \text{(by (A5))}
\end{aligned}$$

$$\Rightarrow \frac{1}{m}\sum_{i=1}^m \frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta}_o)^2 - (\tau_o^2 + \psi_i)}{\tau^2 + \psi_i} \xrightarrow{p} 0 \tag{3.13}$$

$$\begin{aligned}
&\mathrm{var}\Big(\frac{1}{m}\sum_{i=1}^m \frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta}_o)\mathbf{x}_i'(\boldsymbol{\beta}_o - \boldsymbol{\beta})}{\tau^2 + \psi_i}\Big) \\
&= \frac{1}{m^2}\Big(\sum_{i=1}^m \frac{[\mathbf{x}_i'(\boldsymbol{\beta}_o - \boldsymbol{\beta})]^2(\tau_o^2 + \psi_i)}{(\tau^2 + \psi_i)^2} + \sum_{i \neq j} \frac{\mathbf{x}_i'(\boldsymbol{\beta}_o - \boldsymbol{\beta})\mathbf{x}_j'(\boldsymbol{\beta}_o - \boldsymbol{\beta})\delta_o \exp(-M^p\lambda_o\|\mathbf{z}_i - \mathbf{z}_j\|)}{(\tau^2 + \psi_i)(\tau^2 + \psi_j)}\Big) \\
&\leq \frac{1}{m^2}\Big(\sum_{i=1}^m \frac{\|\mathbf{x}_i\|^2\|\boldsymbol{\beta}_o - \boldsymbol{\beta}\|^2(\tau_o^2 + \psi_c)}{\tau^4} + \sum_{i \neq j} \frac{\|\mathbf{x}_i\|\|\mathbf{x}_j\|\|\boldsymbol{\beta}_o - \boldsymbol{\beta}\|^2\delta_o \exp(-M^p\lambda_o\|\mathbf{z}_i - \mathbf{z}_j\|)}{\tau^4}\Big) \\
&\to 0 \ \text{ as } \ m \to \infty \qquad \text{[by (A3), (A4)]}
\end{aligned}$$

$$\Rightarrow \frac{1}{m}\sum_{i=1}^m \frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta}_o)\mathbf{x}_i'(\boldsymbol{\beta}_o - \boldsymbol{\beta})}{\tau^2 + \psi_i} \xrightarrow{p} 0 \tag{3.14}$$

Hence, for all $\boldsymbol{\zeta} \in B_\epsilon$,

$$\frac{1}{m}(g(\boldsymbol{\zeta};\mathbf{y}) - g(\boldsymbol{\zeta})) \xrightarrow{p} 0 \tag{3.15}$$

Step (iii): Next it is shown that for sufficiently large m, and for all $\boldsymbol{\zeta} \in B_\epsilon$, $\frac{1}{m}\nabla_{\boldsymbol{\zeta}\boldsymbol{\zeta}} g(\boldsymbol{\zeta};\mathbf{y})$ is negative definite.

$$\frac{1}{m}\nabla_{\boldsymbol{\beta}\boldsymbol{\beta}} g(\boldsymbol{\zeta};\mathbf{y}) = -\frac{1}{m}\sum_{i=1}^{m}\frac{\mathbf{x}_i\mathbf{x}_i'}{\tau^2+\psi_i}$$

By (A1), $\frac{1}{m}\nabla_{\boldsymbol{\beta}\boldsymbol{\beta}} g(\boldsymbol{\zeta};\mathbf{y})$ is negative definite.

$$\frac{1}{m}\nabla_{\boldsymbol{\beta}\tau^2} g(\boldsymbol{\zeta};\mathbf{y}) = -\frac{1}{m}\sum_{i=1}^{m}\frac{y_i - \mathbf{x}_i'\boldsymbol{\beta}}{(\tau^2+\psi_i)^2}\mathbf{x}_i$$

Let $\mathbf{x}_i = (x_{i1},\ldots,x_{iq})'$. Consider the r^{th} element of $\frac{1}{m}\nabla_{\boldsymbol{\beta}\tau^2} g(\boldsymbol{\zeta};\mathbf{y})$:

$$-\frac{1}{m}\sum_{i=1}^{m}\frac{y_i - \mathbf{x}_i'\boldsymbol{\beta}}{(\tau^2+\psi_i)^2}x_{ir} = -\frac{1}{m}\sum_{i=1}^{m}\frac{y_i - \mathbf{x}_i'\boldsymbol{\beta}_o}{(\tau^2+\psi_i)^2}x_{ir} - \frac{1}{m}\sum_{i=1}^{m}\frac{\mathbf{x}_i'(\boldsymbol{\beta}_o - \boldsymbol{\beta})}{(\tau^2+\psi_i)^2}x_{ir}$$

Since $\|\boldsymbol{\beta} - \boldsymbol{\beta}_o\| < \frac{\tau^4\epsilon}{\frac{1}{m}\sum_{i=1}^{m}\|\mathbf{x}_i\|^2}$,

$$\Big| -\frac{1}{m}\sum_{i=1}^{m}\frac{\mathbf{x}_i'(\boldsymbol{\beta}_o - \boldsymbol{\beta})}{(\tau^2+\psi_i)^2}x_{ir}\Big| < \frac{1}{m}\sum_{i=1}^{m}\frac{\|\mathbf{x}_i\|^2\|\boldsymbol{\beta} - \boldsymbol{\beta}_o\|}{\tau^4} < \epsilon$$

Moreover, by (A3),(A4) and the Cauchy-Schwarz inequality (the proof is similar to (3.14)) ,

$$-\frac{1}{m}\sum_{i=1}^{m}\frac{y_i - \mathbf{x}_i'\boldsymbol{\beta}_o}{(\tau^2+\psi_i)^2}x_{ir} \xrightarrow{p} 0 \tag{3.16}$$

For $r = 1,\ldots,q$, let

$$B_r = \Big\{\mathbf{y} : \Big|\frac{1}{m}\sum_{i=1}^{m}\frac{y_i - \mathbf{x}_i'\boldsymbol{\beta}_o}{(\tau^2+\psi_i)^2}x_{ir}\Big| < \epsilon\Big\}$$

Then for $\mathbf{y} \in B_r$ and for all $\boldsymbol{\zeta} \in B_\epsilon$,

$$\Big|\frac{1}{m}\sum_{i=1}^{m}\frac{y_i - \mathbf{x}_i'\boldsymbol{\beta}}{(\tau^2+\psi_i)^2}x_{ir}\Big| < \epsilon + \epsilon = 2\epsilon$$

and by (3.16), $P(B_r) > 1 - \epsilon$.

Finally, consider

$$\begin{aligned}\frac{1}{m}\nabla_{\tau^2\tau^2} g(\boldsymbol{\zeta}; \mathbf{y}) &= \frac{1}{2m}\sum_{i=1}^{m}\frac{1}{(\tau^2+\psi_i)^2} - \frac{1}{m}\sum_{i=1}^{m}\frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2}{(\tau^2+\psi_i)^3} \\ &= \frac{1}{2m}\sum_{i=1}^{m}\frac{1}{(\tau^2+\psi_i)^2} - \frac{1}{m}\sum_{i=1}^{m}\frac{\tau_o^2+\psi_i}{(\tau^2+\psi_i)^3} - \frac{1}{m}\sum_{i=1}^{m}\frac{(\mathbf{x}_i'(\boldsymbol{\beta}_o - \boldsymbol{\beta}))^2}{(\tau^2+\psi_i)^3} \\ &\quad - \frac{1}{m}\sum_{i=1}^{m}\frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta}_o)^2 - (\tau_o^2+\psi_i)}{(\tau^2+\psi_i)^3} - \frac{2}{m}\sum_{i=1}^{m}\frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta}_o)\mathbf{x}_i'(\boldsymbol{\beta}_o - \boldsymbol{\beta})}{(\tau^2+\psi_i)^3}\end{aligned}$$

Since $|\tau^2 - \tau_o^2| < \frac{9}{10}\tau_0^2$,

$$\begin{aligned}\frac{1}{2m}\sum_{i=1}^{m}\frac{1}{(\tau^2+\psi_i)^2} - \frac{1}{m}\sum_{i=1}^{m}\frac{\tau_o^2+\psi_i}{(\tau^2+\psi_i)^3} &= -\frac{1}{2m}\sum_{i=1}^{m}\frac{2\tau_o^2+\psi_i-\tau^2}{(\tau^2+\psi_i)^3} \\ &< -\frac{1}{2m}\sum_{i=1}^{m}\frac{\frac{1}{10}\tau_o^2+\psi_i}{(\frac{19}{10}\tau_o^2+\psi_i)^3} \\ &< -\frac{1}{20}\frac{\tau_o^2}{(\frac{19}{10}\tau_o^2+\psi_c)^3}\end{aligned}$$

Moreover, by (A3)-(A5) and the Cauchy-Schwarz inequality (the proofs are similar to (3.13) and (3.14)),

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^{m}\frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta}_o)^2 - (\tau_o^2+\psi_i)}{(\tau^2+\psi_i)^3} &\xrightarrow{p} 0 \\ \frac{1}{m}\sum_{i=1}^{m}\frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta}_o)\mathbf{x}_i'(\boldsymbol{\beta}_o - \boldsymbol{\beta})}{(\tau^2+\psi_i)^3} &\xrightarrow{p} 0\end{aligned}$$

Let

$$\begin{aligned}B_{q+1} &= \Big\{\mathbf{y} : \Big|\frac{1}{m}\sum_{i=1}^{m}\frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta}_o)^2 - (\tau_o^2+\psi_i)}{(\tau^2+\psi_i)^3}\Big| < \epsilon\Big\} \\ B_{q+2} &= \Big\{\mathbf{y} : \Big|\frac{1}{m}\sum_{i=1}^{m}\frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta}_o)\mathbf{x}_i'(\boldsymbol{\beta}_o - \boldsymbol{\beta})}{(\tau^2+\psi_i)^3}\Big| < \epsilon\Big\}\end{aligned}$$

Hence, for $\mathbf{y} \in B_{q+1} \bigcap B_{q+2}$ and for all $\boldsymbol{\zeta} \in B_\epsilon$, it has been established that

$$\frac{1}{m}\nabla_{\tau^2\tau^2} g(\boldsymbol{\zeta}; \mathbf{y}) < -\frac{1}{20}\frac{\tau_o^2}{(\frac{19}{10}\tau_o^2+\psi_c)^3}$$

Let $B = B_1 \bigcap \ldots \bigcap B_{q+2}$, then for $\mathbf{y} \in B$ and for all $\boldsymbol{\zeta} \in B_\epsilon$, it has been shown that $\frac{1}{m}\nabla_{\boldsymbol{\zeta}\boldsymbol{\zeta}} g(\boldsymbol{\zeta}; \mathbf{y})$ is negative definite. So, for $\mathbf{y} \in B$ and for all $\boldsymbol{\zeta} \in B_\epsilon$, $\frac{1}{m} g(\boldsymbol{\zeta}; \mathbf{y})$ is strictly concave.

Also, $P(B^c) = P(B_1^c \bigcup \ldots \bigcup B_{q+2}^c) \leq P(B_1^c) + \ldots + P(B_{q+2}^c) = (q+2)\epsilon$. Hence, for sufficiently large m, $\mathbf{y} \in B$. By (3.15) and since for sufficiently large m, $\frac{1}{m} g(\boldsymbol{\zeta}; \mathbf{y})$ is strictly concave for all $\boldsymbol{\zeta} \in B_\epsilon$, an application of Theorem 3.5 gives

$$\sup_{\boldsymbol{\zeta} \in B_\epsilon^\star} \left| \frac{1}{m}(g(\boldsymbol{\zeta}; \mathbf{y}) - g(\boldsymbol{\zeta})) \right| \xrightarrow{p} 0 \tag{3.17}$$

where $B_\epsilon^\star$ is any compact set such that $B_\epsilon^\star \subset B_\epsilon$.

Then, by (3.17) and since $\frac{1}{m} g(\boldsymbol{\zeta})$ has a unique maximum at $\boldsymbol{\zeta} = \boldsymbol{\zeta}_o$, an application of Theorem 3.6 gives the desired result: there exists a sequence of local maxima $\widehat{\boldsymbol{\zeta}}_m$ which is consistent for $\boldsymbol{\zeta}_o$. □

Proof of Theorem 3.2. Since it has been shown that $\widehat{\boldsymbol{\zeta}}$ is consistent for $\boldsymbol{\zeta}_o$, we expand $\nabla_\zeta g(\widehat{\boldsymbol{\zeta}}; \mathbf{y})$ around $\boldsymbol{\zeta}_o$ to obtain

$$\begin{aligned} 0 \equiv \nabla_\zeta g(\widehat{\boldsymbol{\zeta}}; \mathbf{y}) = \nabla_\zeta g(\boldsymbol{\zeta}_o; \mathbf{y}) + \nabla_{\zeta\zeta} g(\boldsymbol{\zeta}_o; \mathbf{y})(\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_o) + \\ \frac{1}{2}\sum_{j=1}^{q+1} \frac{\partial}{\partial \zeta_j}\{\nabla_{\zeta\zeta} g(\boldsymbol{\zeta}_\star; \mathbf{y})\}(\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_o)(\hat{\zeta}_j - \zeta_{oj}) \end{aligned} \tag{3.18}$$

where $\boldsymbol{\zeta}_\star$ lies between $\widehat{\boldsymbol{\zeta}}$ and $\boldsymbol{\zeta}_o$, and $\widehat{\boldsymbol{\zeta}} = (\hat{\zeta}_1, \ldots, \hat{\zeta}_{q+1})'$. For better readability, we indicate the steps in the proof.

Step (i)*:* We first seek to derive the asymptotic distribution of a properly normalized $\nabla_\zeta g(\boldsymbol{\zeta}_o; \mathbf{y})$, where

$$\nabla_\zeta g(\boldsymbol{\zeta}_o; \mathbf{y}) = \begin{pmatrix} \sum_{i=1}^m \frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta}_o)\mathbf{x}_i}{\tau_o^2 + \psi_i} \\ \frac{1}{2}\sum_{i=1}^m \frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta}_o)^2}{(\tau_o^2 + \psi_i)^2} - \frac{1}{2}\sum_{i=1}^m \frac{1}{\tau_o^2 + \psi_i} \end{pmatrix}.$$

Note that

$$\sum_{i=1}^{m} \frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta}_o)\mathbf{x}_i}{\tau_o^2 + \psi_i} = X'D_o^{-1}(\mathbf{y} - X\boldsymbol{\beta}_o) \tag{3.19}$$

where $D_o = \text{diag}(\tau_o^2 + \psi_1, \ldots, \tau^2 + \psi_m)$, and so,

$$\text{var}\Big(\sum_{i=1}^{m} \frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta}_o)\mathbf{x}_i}{\tau_o^2 + \psi_i}\Big) = X'D_o^{-1}V_oD_o^{-1}X$$

Also,

$$\sum_{i=1}^{m} \frac{1}{\tau_o^2 + \psi_i} = \text{tr}[D_o^{-1}V_oD_o^{-1}] = \sum_{i=1}^{m} \gamma_i \quad \text{[by (A6)]}$$

$$\frac{1}{2}\sum_{i=1}^{m} \frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta}_o)^2}{(\tau_o^2 + \psi_i)^2} - \frac{1}{2}\sum_{i=1}^{m} \frac{1}{\tau_o^2 + \psi_i} = \frac{1}{2}(\mathbf{y} - X\boldsymbol{\beta}_o)'D_o^{-2}(\mathbf{y} - X\boldsymbol{\beta}_o) - \frac{1}{2}\sum_{i=1}^{m} \gamma_i \tag{3.20}$$

and so,

$$\begin{aligned}\text{var}\Big(\frac{1}{2}\sum_{i=1}^{m} \frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta}_o)^2}{(\tau_o^2 + \psi_i)^2} - \frac{1}{2}\sum_{i=1}^{m} \gamma_i\Big) &= \frac{1}{4}2\,\text{tr}[D_o^{-2}V_oD_o^{-2}V_o] \qquad \text{[Lemma 3.6 (b)]} \\ &= \frac{1}{2}\sum_{i=1}^{m} \gamma_i^2 \qquad \text{[by (A6) and Lemma 3.5]}\end{aligned}$$

Consider the normalized linear and quadratic forms given in (3.19) and (3.20). Let

$$W = \begin{pmatrix} \mathbf{w}_1' \\ \vdots \\ \mathbf{w}_q' \end{pmatrix} = (X'D_o^{-1}V_oD_o^{-1}X)^{-\frac{1}{2}}X'D_o^{-1}, \tag{3.21}$$

and for $i = 1, \ldots, q$,

$$\begin{aligned} R_i &= \mathbf{w}_i'(\mathbf{y} - X\boldsymbol{\beta}_o) \\ R_{q+1} &= \frac{1/2}{(\frac{1}{2}\sum_{i=1}^{m}\gamma_i^2)^{\frac{1}{2}}}\Big((\mathbf{y} - X\boldsymbol{\beta}_o)'D_o^{-2}(\mathbf{y} - X\boldsymbol{\beta}_o) - \sum_{i=1}^{m}\gamma_i\Big)\end{aligned}$$

From Theorem 3.2c.1 in Mathai and Provost [27], the joint moment generating function of $(R_1, \ldots, R_{q+1})$ is given by

$$M_{R_1,\ldots,R_{q+1}}(t_1,\ldots,t_{q+1}) = \Big|I_m - \frac{t_{q+1}D_o^{-2}V_o}{(\frac{1}{2}\sum_{i=1}^m \gamma_i^2)^{\frac{1}{2}}}\Big|^{-\frac{1}{2}} \exp\Big(-\frac{\sum_{i=1}^m \gamma_i t_{q+1}}{2(\frac{1}{2}\sum_{i=1}^m \gamma_i^2)^{\frac{1}{2}}}\Big)$$
$$\cdot \exp\Big(\frac{1}{2}\Big(\sum_{j=1}^q t_j \mathbf{w}_j'\Big)V_o^{\frac{1}{2}}\Big[I_m - \frac{t_{q+1}V_o^{\frac{1}{2}}D_o^{-2}V_o^{\frac{1}{2}}}{(\frac{1}{2}\sum_{i=1}^m \gamma_i^2)^{\frac{1}{2}}}\Big]^{-1}V_o^{\frac{1}{2}}\Big(\sum_{j=1}^q t_j \mathbf{w}_j\Big)\Big)$$

Next it is shown that

$$\lim_{m\to\infty} M_{R_1,\ldots,R_{q+1}}(t_1,\ldots,t_{q+1}) = \exp\Big(\frac{1}{2}\sum_{j=1}^{q+1} t_j^2\Big). \tag{3.22}$$

Note that

$$I_m - \frac{t_{q+1}V_o^{\frac{1}{2}}D_o^{-2}V_o^{\frac{1}{2}}}{(\frac{1}{2}\sum_{i=1}^m \gamma_i^2)^{\frac{1}{2}}} > I_m - \frac{t_{q+1}\gamma_{\max}(D_o^{-1}V_oD_o^{-1})I_m}{(\frac{1}{2}\sum_{i=1}^m \gamma_i^2)^{\frac{1}{2}}} \quad \text{[by Lemma 3.5]}$$

$$\Rightarrow \Big(\sum_{j=1}^q t_j \mathbf{w}_j'\Big)V_o^{\frac{1}{2}}\Big(I_m - \frac{t_{q+1}V_o^{\frac{1}{2}}D_o^{-2}V_o^{\frac{1}{2}}}{(\frac{1}{2}\sum_{i=1}^m \gamma_i^2)^{\frac{1}{2}}}\Big)^{-1}V_o^{\frac{1}{2}}\Big(\sum_{j=1}^q t_j \mathbf{w}_j\Big)$$
$$< \Big(\sum_{j=1}^q t_j \mathbf{w}_j'\Big)V_o\Big(\sum_{j=1}^q t_j \mathbf{w}_j\Big)\Big(1 - \frac{t_{q+1}\gamma_{\max}(D_o^{-1}V_oD_o^{-1})}{(\frac{1}{2}\sum_{i=1}^m \gamma_i^2)^{\frac{1}{2}}}\Big)^{-1}$$
$$< \sum_{j=1}^q t_j^2\Big(1 - \frac{t_{q+1}\gamma_{\max}(D_o^{-1}V_oD_o^{-1})}{(\frac{1}{2}\sum_{i=1}^m \gamma_i^2)^{\frac{1}{2}}}\Big)^{-1} \to \sum_{j=1}^q t_j^2 \text{ as } m\to\infty$$

[Since $WV_oW' = I_q$ and (A6)].

Similarly,

$$\Big(\sum_{j=1}^q t_j \mathbf{w}_j'\Big)V_o^{\frac{1}{2}}\Big(I_m - \frac{t_{q+1}V_o^{\frac{1}{2}}D_o^{-2}V_o^{\frac{1}{2}}}{(\frac{1}{2}\sum_{i=1}^m \gamma_i^2)^{\frac{1}{2}}}\Big)^{-1}V_o^{\frac{1}{2}}\Big(\sum_{j=1}^q t_j \mathbf{w}_j\Big)$$
$$> \Big(\sum_{j=1}^q t_j \mathbf{w}_j'\Big)V_o\Big(\sum_{j=1}^q t_j \mathbf{w}_j\Big)\Big(1 - \frac{t_{q+1}\gamma_{\min}(D_o^{-1}V_oD_o^{-1})}{(\frac{1}{2}\sum_{i=1}^m \gamma_i^2)^{\frac{1}{2}}}\Big)^{-1} \to \sum_{j=1}^q t_j^2 \text{ as } m\to\infty.$$

Hence we obtain the same limit as an asymptotic upper and lower bound. That is

$$\Big(\sum_{j=1}^{q} t_j \mathbf{w}_j'\Big) V_o^{\frac{1}{2}} \Big(I_m - \frac{t_{q+1} V_o^{\frac{1}{2}} D_o^{-2} V_o^{\frac{1}{2}}}{(\frac{1}{2}\sum_{i=1}^{m}\gamma_i^2)^{\frac{1}{2}}}\Big)^{-1} V_o^{\frac{1}{2}} \Big(\sum_{j=1}^{q} t_j \mathbf{w}_j\Big) \rightarrow \sum_{j=1}^{q} t_j^2. \tag{3.23}$$

Also, note that $D_o^{-1} V_o D_o^{-1}$ and $D_o^{-2} V_o$ have the same eigenvalues (Lemma 3.5). Hence, the eigenvalues of $I_m - c D_o^{-2} V_o$ for any scalar c are given by $1 - c\gamma_1, \ldots, 1 - c\gamma_m$, where the γ_i's are as given in (A6).

$$\Rightarrow \Big| I_m - \frac{t_{q+1} D_o^{-2} V_o}{(\frac{1}{2}\sum_{i=1}^{m}\gamma_i^2)^{\frac{1}{2}}} \Big|^{-\frac{1}{2}} = \prod_{i=1}^{m} \Big(1 - \frac{t_{q+1}\gamma_i}{(\frac{1}{2}\sum_{i=1}^{m}\gamma_i^2)^{\frac{1}{2}}}\Big)^{-\frac{1}{2}}$$

Hence,

$$\begin{aligned}
&\log\Big[\prod_{i=1}^{m} \Big(1 - \frac{t_{q+1}\gamma_i}{(\frac{1}{2}\sum_{i=1}^{m}\gamma_i^2)^{\frac{1}{2}}}\Big)^{-\frac{1}{2}} \exp\Big(-\frac{\sum_{i=1}^{m}\gamma_i t_{q+1}}{2(\frac{1}{2}\sum_{i=1}^{m}\gamma_i^2)^{\frac{1}{2}}}\Big)\Big] \\
&= -\frac{1}{2}\sum_{i=1}^{m} \log\Big[1 - \frac{t_{q+1}\gamma_i}{(\frac{1}{2}\sum_{i=1}^{m}\gamma_i^2)^{\frac{1}{2}}}\Big] - \frac{\sum_{i=1}^{m}\gamma_i t_{q+1}}{(2\sum_{i=1}^{m}\gamma_i^2)^{\frac{1}{2}}} \\
&= -\frac{1}{2}\Big[\sum_{i=1}^{m}\Big(-\frac{t_{q+1}\gamma_i}{(\frac{1}{2}\sum_{i=1}^{m}\gamma_i^2)^{\frac{1}{2}}} - \frac{1}{2}\frac{t_{q+1}^2\gamma_i^2}{\frac{1}{2}\sum_{i=1}^{m}\gamma_i^2}\Big)\Big] + o(t_{q+1}) - \frac{\sum_{i=1}^{m}\gamma_i t_{q+1}}{(2\sum_{i=1}^{m}\gamma_i^2)^{\frac{1}{2}}} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\text{[by (A6) and Lemma 3.2]} \\
&= \frac{1}{2}t_{q+1}^2 + o(t_{q+1}).
\end{aligned} \tag{3.24}$$

Now (3.22) follows from (3.23) and (3.24). That is,

$$C_o^{-\frac{1}{2}} \nabla_{\zeta} g(\boldsymbol{\zeta}_o; \mathbf{y}) \xrightarrow{d} N(\mathbf{0}_{q+1}, I_{q+1}) \tag{3.25}$$

where

$$C_o = \begin{pmatrix} X' D_o^{-1} V_o D_o^{-1} X & \mathbf{0}_q \\ \mathbf{0}_q' & \frac{1}{2}\mathrm{tr}[D_o^{-2} V_o D_o^{-2} V_o] \end{pmatrix}.$$

Step (ii): Next we show that

$$-C_o^{-\frac{1}{2}} \nabla_{\zeta\zeta} g(\boldsymbol{\zeta}_o; \mathbf{y}) C_o^{-\frac{1}{2}} F_o \xrightarrow{p} I_{q+1} \tag{3.26}$$

where

$$F_o = C_o^{\frac{1}{2}} \begin{pmatrix} X'D_o^{-1}X & \mathbf{0}_q \\ \mathbf{0}_q' & \frac{1}{2}\sum_{i=1}^m \frac{1}{(\tau_o^2+\psi_i)^2} \end{pmatrix}^{-1} C_o^{\frac{1}{2}}.$$

Since,

$$-\nabla_{\varsigma\varsigma} g(\boldsymbol{\zeta}_o; \mathbf{y}) = \begin{pmatrix} X'D_o^{-1}X & \sum_{i=1}^m \frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta}_o)\mathbf{x}_i}{(\tau_o^2+\psi_i)^2} \\ \sum_{i=1}^m \frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta}_o)\mathbf{x}_i'}{(\tau_o^2+\psi_i)^2} & \sum_{i=1}^m \frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta}_o)^2}{(\tau_o^2+\psi_i)^3} - \frac{1}{2}\sum_{i=1}^m \frac{1}{(\tau_o^2+\psi_i)^2} \end{pmatrix},$$

$$-C_o^{-\frac{1}{2}} \nabla_{\varsigma\varsigma} g(\boldsymbol{\zeta}_o; \mathbf{y}) C_o^{-\frac{1}{2}} F_o = \begin{pmatrix} I_q & \mathbf{z}_2 \\ \mathbf{z}_1' & \left(\frac{1}{2}\sum_{i=1}^m \frac{1}{(\tau_o^2+\psi_i)^2}\right)^{-1} \sum_{i=1}^m \frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta}_o)^2}{(\tau_o^2+\psi_i)^3} - 1 \end{pmatrix},$$

where

$$\mathbf{z}_1 = \left(\frac{1}{2}\mathrm{tr}[D_o^{-2}V_oD_o^{-2}V_o]\right)^{-\frac{1}{2}} (X'D_o^{-1}V_oD_o^{-1}X)^{\frac{1}{2}} (X'D_o^{-1}X)^{-1}X'D_o^{-2}(\mathbf{y} - X\boldsymbol{\beta}_o)$$

and

$$\mathbf{z}_2 = \frac{\left(\frac{1}{2}\mathrm{tr}[D_o^{-2}V_oD_o^{-2}V_o]\right)^{\frac{1}{2}}}{\frac{1}{2}\sum_{i=1}^m \frac{1}{(\tau_o^2+\psi_i)^2}} (X'D_o^{-1}V_oD_o^{-1}X)^{-\frac{1}{2}}X'D_o^{-2}(\mathbf{y} - X\boldsymbol{\beta}_o).$$

$$\begin{aligned} \mathrm{var}(\mathbf{z}_1) = &\left(\frac{1}{2}\mathrm{tr}[D_o^{-2}V_oD_o^{-2}V_o]\right)^{-1} (X'D_o^{-1}V_oD_o^{-1}X)^{\frac{1}{2}} \\ &\cdot (X'D_o^{-1}X)^{-1}X'D_o^{-2}V_oD_o^{-2}X(X'D_o^{-1}X)^{-1}(X'D_o^{-1}V_oD_o^{-1}X)^{\frac{1}{2}}. \end{aligned}$$

Note that

1. $X'D_o^{-2}V_oD_o^{-2}X < \gamma_{\max}(D_o^{-\frac{3}{2}}V_oD_o^{-\frac{3}{2}})\, X'D_o^{-1}X$
2. $\gamma_{\max}(D_o^{-\frac{3}{2}}V_oD_o^{-\frac{3}{2}})$, $\gamma_{\max}(D_o^{-\frac{1}{2}}V_oD_o^{-\frac{1}{2}})$ and $\gamma_{\max}(D_o^{-1}V_oD_o^{-1})$ are of the same order
3. $\mathrm{tr}[D_o^{-2}V_oD_o^{-2}V_o] = \sum_{i=1}^m \gamma_i^2$

and hence

$$\text{var}(\mathbf{z}_1) < \frac{\gamma_{\max}(D_o^{-\frac{1}{2}} V_o D_o^{-\frac{1}{2}})\gamma_{\max}(D_o^{-\frac{3}{2}} V_o D_o^{-\frac{3}{2}})}{\frac{1}{2}\sum_{i=1}^m \gamma_i^2} I_q \to O_{q\times q} \qquad \text{[by (A6)]}$$

Since $\text{E}(\mathbf{z}_1) = \mathbf{0}_q$, we have established that $\mathbf{z}_1 \xrightarrow{p} \mathbf{0}_q$. Similarly,

$$\text{var}(\mathbf{z}_2) = \frac{\frac{1}{2}\text{tr}[D_o^{-2}V_oD_o^{-2}V_o]}{\left(\frac{1}{2}\sum_{i=1}^m \frac{1}{(\tau_o^2+\psi_i)^2}\right)^2}(X'D_o^{-1}V_oD_o^{-1}X)^{-\frac{1}{2}}X'D_o^{-2}V_oD_o^{-2}X(X'D_o^{-1}V_oD_o^{-1}X)^{-\frac{1}{2}}.$$

Since

1. $\left(\sum_{i=1}^m \frac{1}{(\tau_o^2+\psi_i)^2}\right)^2 = O(m^2)$

2. $\text{tr}(D_o^{-2}V_oD_o^{-2}V_o)$ is the same order as $\text{tr}(V_o^2) = \sum_{i\neq j}\delta_o^2 \exp\left(-2M^p\lambda_o\|\mathbf{z}_i - \mathbf{z}_j\|\right) + \sum_{i=1}^m(\tau_o^2+\psi_i)^2$,

by (A5) we get

$$\text{var}(\mathbf{z}_2) \to O_{q\times q}.$$

Since $\text{E}(\mathbf{z}_2) = \mathbf{0}_q$, we have established that $\mathbf{z}_2 \xrightarrow{p} \mathbf{0}_q$. Also,

$$\begin{aligned}
&\text{var}\Big[\Big(\frac{1}{2}\sum_{i=1}^m \frac{1}{(\tau_o^2+\psi_i)^2}\Big)^{-1}\sum_{i=1}^m \frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta}_o)^2}{(\tau_o^2+\psi_i)^3} - 1\Big] \\
&= \Big(\frac{1}{2}\sum_{i=1}^m \frac{1}{(\tau_o^2+\psi_i)^2}\Big)^{-2}\text{var}\Big((\mathbf{y}-X\boldsymbol{\beta}_o)'D_o^{-3}(\mathbf{y}-X\boldsymbol{\beta}_o)\Big) \\
&= \Big(\frac{1}{2}\sum_{i=1}^m \frac{1}{(\tau_o^2+\psi_i)^2}\Big)^{-2} 2\text{tr}(D_o^{-3}V_oD_o^{-3}V_o).
\end{aligned}$$

Once again by (A5),

$$\text{var}\Big[\Big(\frac{1}{2}\sum_{i=1}^m \frac{1}{(\tau_o^2+\psi_i)^2}\Big)^{-1}\sum_{i=1}^m \frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta}_o)^2}{(\tau_o^2+\psi_i)^3} - 1\Big] \to 0 \text{ as } m \to \infty.$$

Hence, it follows that

$$-C_o^{-\frac{1}{2}}\nabla_{\varsigma\varsigma}g(\boldsymbol{\zeta}_o;\mathbf{y})C_o^{-\frac{1}{2}}F_o \xrightarrow{p} E\Big(-C_o^{-\frac{1}{2}}\nabla_{\varsigma\varsigma}g(\boldsymbol{\zeta}_o;\mathbf{y})C_o^{-\frac{1}{2}}F_o\Big) = I_{q+1}$$

Step (iii)*:* Finally it has to be shown that for $j = 1, \ldots, q+1$,

$$\frac{1}{2} C_o^{-\frac{1}{2}} \frac{\partial}{\partial \zeta_j} \{\nabla_{\varsigma\varsigma} g(\boldsymbol{\zeta}_\star; \mathbf{y})\} C_o^{-\frac{1}{2}} F_o \tag{3.27}$$

is bounded in probability. The proof for $j = 1, \ldots, q$ is similar to $j = q+1$, hence, we give the proof for (3.27) only for the case $j = q+1$ (Note: $\zeta_{q+1} = \tau^2$).

$$C_o^{-\frac{1}{2}} \frac{\partial}{\partial \zeta_{q+1}} \{\nabla_{\varsigma\varsigma} g(\boldsymbol{\zeta}_\star; \mathbf{y})\} C_o^{-\frac{1}{2}} F_o = \tag{3.28}$$

$$\begin{pmatrix} (X'D_o^{-1}V_oD_o^{-1}X)^{-\frac{1}{2}} X'D_\star^{-2}X(X'D_o^{-1}X)^{-1}(X'D_o^{-1}V_oD_o^{-1}X)^{\frac{1}{2}} & \mathbf{z}_{2\star} \\ \mathbf{z}'_{1\star} & \frac{3\sum_{i=1}^m \frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta}_\star)^2}{(\tau_\star^2+\psi_i)^4} - \sum_{i=1}^m \frac{1}{(\tau_\star^2+\psi_i)^3}}{\frac{1}{2}\sum_{i=1}^m \frac{1}{(\tau_o^2+\psi_i)^2}} \end{pmatrix}$$

where $\boldsymbol{\zeta}_\star$ lies between $\widehat{\boldsymbol{\zeta}}$ and $\boldsymbol{\zeta}_o$, $\boldsymbol{\zeta}_\star = (\boldsymbol{\beta}_\star, \tau_\star^2)$, $D_\star = \text{diag}(\tau_\star^2 + \psi_1, \ldots, \tau_\star^2 + \psi_m)$, and

$$\mathbf{z}_{1\star} = \frac{1}{\left(\frac{1}{2}\text{tr}[D_o^{-2}V_oD_o^{-2}V_o]\right)^{\frac{1}{2}}} (X'D_o^{-1}V_oD_o^{-1}X)^{\frac{1}{2}} (X'D_o^{-1}X)^{-1} X'D_\star^{-3}(\mathbf{y} - X\boldsymbol{\beta}_\star)$$

$$\mathbf{z}_{2\star} = \left(\frac{1}{2}\text{tr}[D_o^{-2}V_oD_o^{-2}V_o]\right)^{\frac{1}{2}} (X'D_o^{-1}V_oD_o^{-1}X)^{-\frac{1}{2}} \left(\frac{X'D_\star^{-3}(\mathbf{y} - X\boldsymbol{\beta}_\star)}{\frac{1}{2}\sum_{i=1}^m \frac{1}{(\tau_o^2+\psi_i)^2}}\right)$$

It is not difficult to see that apart from $z_{1\star}$ and $z_{2\star}$, the remaining terms in (3.28) are bounded in probability. By (A8), $(X'X)^{-1}$ has entries that are $O(1/m)$. Hence, by computing the variance of $(X'D_o^{-1}X)^{-1}X'D_\star^{-3}(\mathbf{y} - X\boldsymbol{\beta}_\star)$ it follows that $(X'D_o^{-1}X)^{-1}X'D_\star^{-3}(\mathbf{y} - X\boldsymbol{\beta}_\star)$ is bounded in probability. By (A7), the matrix $\left(\frac{1}{2}\text{tr}[D_o^{-2}V_oD_o^{-2}V_o]\right)^{\frac{1}{2}} (X'D_o^{-1}V_oD_o^{-1}X)^{-\frac{1}{2}}$ has entries that are $O(1)$. Hence, $z_{1\star}$ and $z_{2\star}$ are also bounded in probability.

Step (iv)*:* Putting (3.25)-(3.27) together we derive the asymptotic distribution of $\widehat{\boldsymbol{\zeta}}$. Left multiplying (3.18) by $C_o^{-\frac{1}{2}}$,

$$0 = C_o^{-\frac{1}{2}} \nabla_\varsigma g(\boldsymbol{\zeta}_o; \mathbf{y}) + C_o^{-\frac{1}{2}} \nabla_{\varsigma\varsigma} g(\boldsymbol{\zeta}_o; \mathbf{y}) C_o^{-\frac{1}{2}} F_o [F_o^{-1} C_o^{\frac{1}{2}} (\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_o)] +$$

$$\frac{C_o^{-\frac{1}{2}}}{2} \sum_{j=1}^{q+1} \frac{\partial}{\partial \zeta_j} \{\nabla_{\varsigma\varsigma} g(\boldsymbol{\zeta}_\star; \mathbf{y})\} C_o^{-\frac{1}{2}} F_o [F_o^{-1} C_o^{\frac{1}{2}} (\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_o)] (\hat{\zeta}_j - \zeta_{oj})$$

$$\Rightarrow\ F_o^{-1}C_o^{\frac{1}{2}}(\widehat{\boldsymbol{\zeta}}-\boldsymbol{\zeta}_o)=$$

$$\Big(-C_o^{-\frac{1}{2}}\nabla_{\varsigma\varsigma}g(\boldsymbol{\zeta}_o;\mathbf{y})C_o^{-\frac{1}{2}}F_o-\frac{C_o^{-\frac{1}{2}}}{2}\sum_{j=1}^{q+1}\frac{\partial}{\partial\zeta_j}\{\nabla_{\varsigma\varsigma}g(\boldsymbol{\zeta}_\star;\mathbf{y})\}C_o^{-\frac{1}{2}}F_o(\hat{\zeta}_j-\zeta_{oj})\Big)^{-1}C_o^{-\frac{1}{2}}\nabla_{\varsigma}g(\boldsymbol{\zeta}_o;\mathbf{y})$$

By (3.25)-(3.27) and since for $j=1,\ldots,q+1$, $\hat{\zeta}_j\xrightarrow{p}\zeta_{oj}$,

$$F_o^{-1}C_o^{\frac{1}{2}}(\widehat{\boldsymbol{\zeta}}-\boldsymbol{\zeta}_o)\xrightarrow{d}\mathrm{N}(\mathbf{0}_{q+1},I_{q+1})$$

where

$$F_o^{-1}C_o^{\frac{1}{2}}=\begin{pmatrix}(X'D_o^{-1}V_oD_o^{-1}X)^{-\frac{1}{2}}X'D_o^{-1}X & \mathbf{0}_q\\ \mathbf{0}_q' & \frac{\frac{1}{\sqrt{2}}\sum_{i=1}^m\frac{1}{(\tau_o^2+\psi_i)^2}}{\sqrt{\mathrm{tr}(D_o^{-2}V_oD_o^{-2}V_o)}}\end{pmatrix}$$

□

Proof of Theorem 3.3. Let $\boldsymbol{\nu}=(\delta,\lambda)'=(\nu_1,\nu_2)'$, and $\boldsymbol{\nu}_o=(\delta_o,\lambda_o)'=(\nu_{o1},\nu_{o2})'$. Also, let

$$h(\boldsymbol{\nu};\mathbf{y})=-\sum_{l=1}^k\sum_{\substack{i,j\in C_l\\ i\neq j}}\Big(\hat{\epsilon}_i\hat{\epsilon}_j-\delta\exp(-\lambda M^p\|\mathbf{z}_i-\mathbf{z}_j\|)\Big)^2$$

$$h(\boldsymbol{\nu})=-\sum_{l=1}^k\sum_{\substack{i,j\in C_l\\ i\neq j}}\Big(\delta_o\exp(-\lambda_o M^p\|\mathbf{z}_i-\mathbf{z}_j\|)-\delta\exp(-\lambda M^p\|\mathbf{z}_i-\mathbf{z}_j\|)\Big)^2$$

where $\hat{\epsilon}_i=y_i-\mathbf{x}_i'\widehat{\boldsymbol{\beta}}$. Our method of proof is identical to that of Theorem 3.1.

Step (i): We first show that

$$\frac{1}{\sum_{l=1}^k n_l^2}\Big(h(\boldsymbol{\nu};\mathbf{y})-h(\boldsymbol{\nu})\Big)\xrightarrow{p}0 \tag{3.29}$$

Note that

$$\frac{1}{\sum_{l=1}^k n_l^2}h(\boldsymbol{\nu};\mathbf{y})=-\frac{1}{\sum_{l=1}^k n_l^2}\sum_{l=1}^k\sum_{\substack{i,j\in C_l\\ i\neq j}}\Big((y_i-\mathbf{x}_i'\boldsymbol{\beta}_o)(y_j-\mathbf{x}_j'\boldsymbol{\beta}_o)-(y_i-\mathbf{x}_i'\boldsymbol{\beta}_o)\mathbf{x}_j'\widehat{\mathbf{b}}$$

$$-(y_j-\mathbf{x}_j'\boldsymbol{\beta}_o)\mathbf{x}_i'\widehat{\mathbf{b}}+\mathbf{x}_i'\widehat{\mathbf{b}}\,\mathbf{x}_j'\widehat{\mathbf{b}}-\delta\exp(-\lambda M^p\|\mathbf{z}_i-\mathbf{z}_j\|)\Big)^2,$$

where $\widehat{\mathbf{b}} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_o$. We argue that any term that involves $\widehat{\mathbf{b}}$ in the expression for $\frac{1}{\sum_{l=1}^k n_l^2} h(\boldsymbol{\nu}; \mathbf{y})$ will converge to zero in probability. For example,

$$\frac{1}{\sum_{l=1}^k n_l^2} \sum_{l=1}^k \sum_{\substack{i,j \in C_l \\ i \neq j}} (\mathbf{x}_i' \widehat{\mathbf{b}})^2 \, (\mathbf{x}_j' \widehat{\mathbf{b}})^2 \leq \frac{1}{\sum_{l=1}^k n_l^2} \sum_{l=1}^k n_l^2 \sum_{\substack{i,j \in C_l \\ i \neq j}} \frac{\|\mathbf{x}_i\|^2}{n_l} \frac{\|\mathbf{x}_j\|^2}{n_l} \|\widehat{\mathrm{b}}\|^4$$

$$\xrightarrow{p} 0 \qquad \text{[by (B1), (B5) and since } \|\widehat{\mathrm{b}}\| \xrightarrow{p} 0]$$

Consider the term $\frac{1}{\sum_{l=1}^k n_l^2} \sum_{l=1}^k \sum_{\substack{i,j \in C_l \\ i \neq j}} (y_i - \mathbf{x}_i' \boldsymbol{\beta}_o)^2 (\mathbf{x}_j' \widehat{\mathbf{b}})^2$.

$$\frac{1}{\sum_{l=1}^k n_l^2} \sum_{l=1}^k \sum_{\substack{i,j \in C_l \\ i \neq j}} (y_i - \mathbf{x}_i' \boldsymbol{\beta}_o)^2 (\mathbf{x}_j' \widehat{\mathbf{b}})^2 \leq \frac{1}{\sum_{l=1}^k n_l^2} \sum_{l=1}^k \sum_{\substack{i,j \in C_l \\ i \neq j}} (y_i - \mathbf{x}_i' \boldsymbol{\beta}_o)^2 \|\mathbf{x}_j\|^2 \|\widehat{\mathrm{b}}\|^2$$

$$\mathrm{E}\Big| \frac{1}{\sum_{l=1}^k n_l^2} \sum_{l=1}^k \sum_{\substack{i,j \in C_l \\ i \neq j}} (y_i - \mathbf{x}_i' \boldsymbol{\beta}_o)^2 \|\mathbf{x}_j\|^2 \Big| = \frac{1}{\sum_{l=1}^k n_l^2} \sum_{l=1}^k n_l^2 \sum_{\substack{i,j \in C_l \\ i \neq j}} \frac{E(y_i - \mathbf{x}_i' \boldsymbol{\beta}_o)^2}{n_l} \frac{\|\mathbf{x}_j\|^2}{n_l}$$

$$\leq c < \infty \qquad \text{[by (B1), (B5)]}$$

and since, $\|\widehat{\mathrm{b}}\| \xrightarrow{p} 0$,

$$\frac{1}{\sum_{l=1}^k n_l^2} \sum_{l=1}^k \sum_{\substack{i,j \in C_l \\ i \neq j}} (y_i - \mathbf{x}_i' \boldsymbol{\beta}_o)^2 (\mathbf{x}_j' \widehat{\mathbf{b}})^2 \xrightarrow{p} 0.$$

Also,

$$\frac{1}{\sum_{l=1}^k n_l^2} \Big| \sum_{l=1}^k \sum_{\substack{i,j \in C_l \\ i \neq j}} (y_i - \mathbf{x}_i' \boldsymbol{\beta}_o)^2 (y_j - \mathbf{x}_j' \boldsymbol{\beta}_o) \mathbf{x}_j' \widehat{\mathbf{b}} \Big|$$

$$\leq \frac{\|\widehat{\mathbf{b}}\|}{\sum_{l=1}^k n_l^2} \sum_{l=1}^k \sum_{\substack{i,j \in C_l \\ i \neq j}} (y_i - \mathbf{x}_i' \boldsymbol{\beta}_o)^2 |y_j - \mathbf{x}_j' \boldsymbol{\beta}_o| \|\mathbf{x}_j\|,$$

Next, $\|\widehat{\mathbf{b}}\| \xrightarrow{p} 0$ and

$$\frac{1}{\sum_{l=1}^{k} n_l^2} \sum_{l=1}^{k} n_l^2 \sum_{\substack{i,j \in C_l \\ i \neq j}} \frac{\mathrm{E}\Big((y_i - \mathbf{x}_i'\boldsymbol{\beta}_o)^2 |y_j - \mathbf{x}_j'\boldsymbol{\beta}_o|\Big)\|\mathbf{x}_j\|}{n_l^2}$$

$$\leq \frac{1}{\sum_{l=1}^{k} n_l^2} \sum_{l=1}^{k} n_l^2 \sum_{\substack{i,j \in C_l \\ i \neq j}} \frac{(\mathrm{E}(y_i - \mathbf{x}_i'\boldsymbol{\beta}_o)^4)^{\frac{1}{2}}}{n_l} \frac{(\mathrm{E}(y_j - \mathbf{x}_j'\boldsymbol{\beta}_o)^2)^{\frac{1}{2}}\|\mathbf{x}_j\|}{n_l}$$

$$\leq c < \infty \qquad \text{[by (B1), (B4)]},$$

hence,

$$\frac{1}{\sum_{l=1}^{k} n_l^2} \sum_{l=1}^{k} \sum_{\substack{i,j \in C_l \\ i \neq j}} (y_i - \mathbf{x}_i'\boldsymbol{\beta}_o)^2 (y_j - \mathbf{x}_j'\boldsymbol{\beta}_o) \mathbf{x}_j'\widehat{\mathbf{b}} \xrightarrow{p} 0$$

Also,

$$\frac{1}{\sum_{l=1}^{k} n_l^2} \Big| \sum_{l=1}^{k} \sum_{\substack{i,j \in C_l \\ i \neq j}} \mathbf{x}_i'\widehat{\mathbf{b}}\, \mathbf{x}_j'\widehat{\mathbf{b}}\, \delta \exp(-\lambda M^p \|\mathbf{z}_i - \mathbf{z}_j\|) \Big| \leq \frac{\delta \|\widehat{\mathbf{b}}\|^2}{\sum_{l=1}^{k} n_l^2} \sum_{l=1}^{k} n_l^2 \sum_{\substack{i,j \in C_l \\ i \neq j}} \frac{\|\mathbf{x}_i\|}{n_l} \frac{\|\mathbf{x}_j\|}{n_l}$$

$$\xrightarrow{p} 0 \qquad \text{[by (B1), (B4)]}$$

Similarly, the remaining terms that involve $\widehat{\mathbf{b}}$ in the expression for $\frac{1}{\sum_{l=1}^{k} n_l^2} h(\boldsymbol{\nu}; \mathbf{y})$ converge to zero in probability. The sum of the terms that do not involve $\widehat{\mathbf{b}}$ in the expression for $\frac{1}{\sum_{l=1}^{k} n_l^2} h(\boldsymbol{\nu}; \mathbf{y})$ are shown to converge in probability to its expectation. We expand variance terms within and between clusters.

$$\mathrm{var}\Big(\sum_{l=1}^{k} \sum_{\substack{i,j \in C_l \\ i \neq j}} (y_i - \mathbf{x}_i'\boldsymbol{\beta}_o)(y_j - \mathbf{x}_j'\boldsymbol{\beta}_o) \delta \exp(-\lambda M^p \|\mathbf{z}_i - \mathbf{z}_j\|) \Big)$$

$$\leq \delta^2 \sum_{l=1}^{k} \mathrm{var}\Big(\sum_{\substack{i,j \in C_l \\ i \neq j}} (y_i - \mathbf{x}_i'\boldsymbol{\beta}_o)(y_j - \mathbf{x}_j'\boldsymbol{\beta}_o) \Big) + \tag{3.30}$$

$$\delta^2 \sum_{l_1 \neq l_2} \mathrm{var}\Big(\sum_{\substack{i \in C_{l_1} \\ j \in C_{l_2}}} (y_i - \mathbf{x}_i'\boldsymbol{\beta}_o)(y_j - \mathbf{x}_j'\boldsymbol{\beta}_o) \exp(-\lambda M^p \|\mathbf{z}_i - \mathbf{z}_j\|) \Big) \tag{3.31}$$

For the term in (3.30), the sum within cluster involves n_l^2 terms, and the variance of this sum is $O(n_l^4)$. For the term in (3.31), the sum between cluster, involves $n_{l_1} n_{l_2}$ terms, and the variance of this sum is:

$$O(n_{l_1}^2 n_{l_2}^2) \sup_{\substack{i \in C_{l_1} \\ j \in C_{l_2}}} \exp(-\lambda M^p \|\mathbf{z}_i - \mathbf{z}_j\|).$$

Hence,

$$\begin{aligned}
&\operatorname{var}\Big(\frac{1}{\sum_{l=1}^k n_l^2} \sum_{l=1}^k \sum_{\substack{i,j \in C_l \\ i \neq j}} (y_i - \mathbf{x}_i'\boldsymbol{\beta}_o)(y_j - \mathbf{x}_j'\boldsymbol{\beta}_o)\delta \exp(-\lambda M^p \|\mathbf{z}_i - \mathbf{z}_j\|)\Big) \\
&= \frac{1}{(\sum_{l=1}^k n_l^2)^2} \sum_{l=1}^k O(n_l^4) + \frac{1}{(\sum_{l=1}^k n_l^2)^2} \sum_{l_1 \neq l_2} O(n_{l_1}^2 n_{l_2}^2) \sup_{\substack{i \in C_{l_1} \\ j \in C_{l_2}}} \exp(-\lambda M^p \|\mathbf{z}_i - \mathbf{z}_j\|) \\
&= \frac{O\Big(\sum_{l=1}^k n_l^4\Big)}{(\sum_{l=1}^k n_l^2)^2} + \sup_{l_1 \neq l_2} \sup_{\substack{i \in C_{l_1} \\ j \in C_{l_2}}} \exp(-\lambda M^p \|\mathbf{z}_i - \mathbf{z}_j\|) \\
&\to 0 \qquad \text{[by (B1), (B2), (B6) and Lemma 3.4(b)]}
\end{aligned}$$

Hence, the term given by (3.32) is non-zero and is one of the terms in the expression for $\frac{1}{\sum_{l=1}^k n_l^2} h(\boldsymbol{\nu})$.

$$\begin{aligned}
&\frac{1}{\sum_{l=1}^k n_l^2} \sum_{l=1}^k \sum_{\substack{i,j \in C_l \\ i \neq j}} (y_i - \mathbf{x}_i'\boldsymbol{\beta}_o)(y_j - \mathbf{x}_j'\boldsymbol{\beta}_o)\delta \exp(-\lambda M^p \|\mathbf{z}_i - \mathbf{z}_j\|) \\
&\qquad \overset{p}{\approx} E\Big(\frac{1}{\sum_{l=1}^k n_l^2} \sum_{l=1}^k \sum_{\substack{i,j \in C_l \\ i \neq j}} (y_i - \mathbf{x}_i'\boldsymbol{\beta}_o)(y_j - \mathbf{x}_j'\boldsymbol{\beta}_o)\delta \exp(-\lambda M^p \|\mathbf{z}_i - \mathbf{z}_j\|)\Big) \\
&\qquad = \frac{1}{\sum_{l=1}^k n_l^2} \sum_{l=1}^k \sum_{\substack{i,j \in C_l \\ i \neq j}} \delta_o \exp(-\lambda_o M^p \|\mathbf{z}_i - \mathbf{z}_j\|)\delta \exp(-\lambda M^p \|\mathbf{z}_i - \mathbf{z}_j\|) \quad (3.32)
\end{aligned}$$

Similar to the above proof,

$$\mathrm{var}\Big(\frac{1}{\sum_{l=1}^{k} n_l^2}\sum_{l=1}^{k}\sum_{\substack{i,j\in C_l\\ i\neq j}}(y_i-\mathbf{x}_i'\boldsymbol{\beta}_o)^2(y_j-\mathbf{x}_j'\boldsymbol{\beta}_o)^2\Big)\to 0,$$

and hence,

$$\begin{aligned}
&\frac{1}{\sum_{l=1}^{k} n_l^2}\sum_{l=1}^{k}\sum_{\substack{i,j\in C_l\\ i\neq j}}(y_i-\mathbf{x}_i'\boldsymbol{\beta}_o)^2(y_j-\mathbf{x}_j'\boldsymbol{\beta}_o)^2\\
&\qquad\overset{p}{\approx}\frac{1}{\sum_{l=1}^{k} n_l^2}\sum_{l=1}^{k}\sum_{\substack{i,j\in C_l\\ i\neq j}}\mathrm{E}((y_i-\mathbf{x}_i'\boldsymbol{\beta}_o)^2(y_j-\mathbf{x}_j'\boldsymbol{\beta}_o)^2)\\
&\qquad=\frac{1}{\sum_{l=1}^{k} n_l^2}\sum_{l=1}^{k}\sum_{\substack{i,j\in C_l\\ i\neq j}}\delta_o^2\exp(-2\lambda_o M^p\|\mathbf{z}_i-\mathbf{z}_j\|)\quad\text{[by Lemma 3.1]}\qquad(3.33)
\end{aligned}$$

Note that the term given by (3.33) is also one of the terms in the expression for $\frac{1}{\sum_{l=1}^{k} n_l^2}h(\boldsymbol{\nu})$. Putting all the above arguments together, we get (3.29).

Step (ii)*:* We next argue that $h(\boldsymbol{\nu})$ has a unique maximum at $\delta=\delta_o$ and $\lambda=\lambda_o$. First note that $h(\delta_o,\lambda_o)=0$ and it is clear that the maximum value of $h(\boldsymbol{\nu})$ is 0. For any $\epsilon>0$ and for any $\boldsymbol{\nu}=(\delta,\lambda)$ such that $\|\boldsymbol{\nu}-\boldsymbol{\nu}_o\|\geq\epsilon$, by (B3) we have that $h(\boldsymbol{\nu})$ is strictly less than 0. Hence we have shown $h(\boldsymbol{\nu})$ has a unique maximum at $\delta=\delta_o$ and $\lambda=\lambda_o$.

Step (iii)*:* Next it is shown that for large m, $\frac{1}{\sum_{l=1}^{k} n_l^2}\nabla_{\boldsymbol{\nu}\boldsymbol{\nu}}h(\boldsymbol{\nu};\mathbf{y})$ is negative definite in a non-shrinking neighborhood of $\boldsymbol{\nu}_o$.

Uniformly for large k, $n_1,\ldots,n_k$,

$$\frac{1}{\sum_{l=1}^{k} n_l^2}\frac{\partial^2}{\partial\delta^2}h(\boldsymbol{\nu};\mathbf{y})=\frac{-2}{\sum_{l=1}^{k} n_l^2}\sum_{l=1}^{k}\sum_{\substack{i,j\in C_l\\ i\neq j}}\exp(-2\lambda M^p\|\mathbf{z}_i-\mathbf{z}_j\|)<0.$$

$$\frac{1}{\sum_{l=1}^{k} n_l^2}\frac{\partial^2}{\partial\lambda^2}h(\boldsymbol{\nu};\mathbf{y}) = \frac{2\delta}{\sum_{l=1}^{k} n_l^2}\sum_{l=1}^{k}\sum_{\substack{i,j\in C_l\\ i\neq j}}\Big(\hat{\epsilon}_i\hat{\epsilon}_j(M^p\|\mathbf{z}_i-\mathbf{z}_j\|)^2\exp(-\lambda M^p\|\mathbf{z}_i-\mathbf{z}_j\|)$$

$$-2\delta(M^p\|\mathbf{z}_i-\mathbf{z}_j\|)^2\exp(-2\lambda M^p\|\mathbf{z}_i-\mathbf{z}_j\|)\Big)$$

Similar to the proof shown previously that any term that involves $\widehat{\mathbf{b}}$ in the expression for $\frac{1}{\sum_{l=1}^{k} n_l^2}h(\boldsymbol{\nu};\mathbf{y})$ converges to zero in probability, we get

$$\frac{1}{\sum_{l=1}^{k} n_l^2}\frac{\partial^2}{\partial\lambda^2}h(\boldsymbol{\nu};\mathbf{y}) = \frac{2\delta}{\sum_{l=1}^{k} n_l^2}\sum_{l=1}^{k}\sum_{\substack{i,j\in C_l\\ i\neq j}}\Big(-2\delta(M^p\|\mathbf{z}_i-\mathbf{z}_j\|)^2\exp(-2\lambda M^p\|\mathbf{z}_i-\mathbf{z}_j\|)$$

$$+(y_i-\mathbf{x}_i'\boldsymbol{\beta}_o)(y_j-\mathbf{x}_j'\boldsymbol{\beta}_o)(M^p\|\mathbf{z}_i-\mathbf{z}_j\|)^2\exp(-\lambda M^p\|\mathbf{z}_i-\mathbf{z}_j\|)\Big)+o_p(1)$$

Note (B2) is needed in the above proof for two reasons: within a cluster we need $M^p\|\mathbf{z}_i-\mathbf{z}_j\|<\infty$, and to make it possible to disregard between cluster terms.

Moreover,

$$\mathrm{var}\Big(\frac{1}{\sum_{l=1}^{k} n_l^2}\sum_{l=1}^{k}\sum_{\substack{i,j\in C_l\\ i\neq j}}(y_i-\mathbf{x}_i'\boldsymbol{\beta}_o)(y_j-\mathbf{x}_j'\boldsymbol{\beta}_o)(M^p\|\mathbf{z}_i-\mathbf{z}_j\|)^2\exp(-\lambda M^p\|\mathbf{z}_i-\mathbf{z}_j\|)\Big)\to 0$$

Hence,

$$\frac{1}{\sum_{l=1}^{k} n_l^2}\frac{\partial^2}{\partial\lambda^2}h(\boldsymbol{\nu};\mathbf{y}) \overset{p}{\approx} \frac{2\delta}{\sum_{l=1}^{k} n_l^2}\sum_{l=1}^{k}\sum_{\substack{i,j\in C_l\\ i\neq j}}(M^p\|\mathbf{z}_i-\mathbf{z}_j\|)^2\exp(-\lambda M^p\|\mathbf{z}_i-\mathbf{z}_j\|)$$

$$\cdot\Big(\delta_o\exp(-\lambda_o M^p\|\mathbf{z}_i-\mathbf{z}_j\|)-2\delta\exp(-\lambda M^p\|\mathbf{z}_i-\mathbf{z}_j\|)\Big)$$

Since for large m we wish to show that $\frac{1}{\sum_{l=1}^{k} n_l^2}\frac{\partial^2}{\partial\lambda^2}h(\boldsymbol{\nu};\mathbf{y})<0$ in a neighborhood of $\boldsymbol{\nu}_o$, it is sufficient to show that $\forall l,\forall i,j\in C_l$, $2\delta\exp(-\lambda M^p\|\mathbf{z}_i-\mathbf{z}_j\|)-\delta_o\exp(-\lambda_o M^p\|\mathbf{z}_i-\mathbf{z}_j\|)>0$ in a neighborhood of $\boldsymbol{\nu}_o$.

Let $\delta \in (0.6\delta_o, 1.4\delta_o)$, then

$$2\delta \exp(-\lambda M^p \|\mathbf{z}_i - \mathbf{z}_j\|) - \delta_o \exp(-\lambda_o M^p \|\mathbf{z}_i - \mathbf{z}_j\|) >$$

$$\delta_o[1.2 \exp(-\lambda M^p \|\mathbf{z}_i - \mathbf{z}_j\|) - \exp(-\lambda_o M^p \|\mathbf{z}_i - \mathbf{z}_j\|)]$$

Let

$$\lambda \in \lambda_o \pm \frac{\log 1.2}{\sup_l \limsup_{M\to\infty} \sup_{i,j\in C_l} \|\mathbf{z}_i - \mathbf{z}_j\|}$$

For the neighborhood of $\boldsymbol{\nu}_o$ given above, we have uniformly, for large k, $n_1, \ldots, n_k$, $\frac{1}{\sum_{l=1}^k n_l^2} \frac{\partial^2}{\partial \lambda^2} h(\boldsymbol{\nu}; \mathbf{y}) < 0$. Note that we need $\delta_o, \lambda_o > 0$, and (B2) is needed in showing the existence of a neighborhood at λ_o such that $\frac{1}{\sum_{l=1}^k n_l^2} \frac{\partial^2}{\partial \lambda^2} h(\boldsymbol{\nu}; \mathbf{y}) < 0$.

Finally, it needs to be shown that uniformly, for large k, $n_1, \ldots, n_k$, the determinant of $\frac{1}{\sum_{l=1}^k n_l^2} \nabla_{\nu\nu} h(\boldsymbol{\nu}; \mathbf{y})$ is positive in a neighborhood of $\boldsymbol{\nu}_o$.

$$\frac{1}{\sum_{l=1}^k n_l^2} \frac{\partial^2}{\partial \delta \partial \lambda} h(\boldsymbol{\nu}; \mathbf{y}) = \frac{-2}{\sum_{l=1}^k n_l^2} \sum_{l=1}^k \sum_{\substack{i,j\in C_l \\ i\neq j}} \Big(\hat{\epsilon}_i \hat{\epsilon}_j M^p \|\mathbf{z}_i - \mathbf{z}_j\| \exp(-\lambda M^p \|\mathbf{z}_i - \mathbf{z}_j\|)$$

$$- 2\delta M^p \|\mathbf{z}_i - \mathbf{z}_j\| \exp(-2\lambda M^p \|\mathbf{z}_i - \mathbf{z}_j\|) \Big)$$

Using a similar idea to the proof of (3.29), we get

$$\frac{1}{\sum_{l=1}^k n_l^2} \frac{\partial^2}{\partial \delta \partial \lambda} h(\boldsymbol{\nu}; \mathbf{y}) \overset{p}{\approx} \frac{-2}{\sum_{l=1}^k n_l^2} \sum_{l=1}^k \sum_{\substack{i,j\in C_l \\ i\neq j}} M^p \|\mathbf{z}_i - \mathbf{z}_j\| \exp(-\lambda M^p \|\mathbf{z}_i - \mathbf{z}_j\|)$$

$$\cdot \Big(\delta_o \exp(-\lambda_o M^p \|\mathbf{z}_i - \mathbf{z}_j\|) - 2\delta \exp(-\lambda M^p \|\mathbf{z}_i - \mathbf{z}_j\|) \Big)$$

The determinant of $\frac{1}{\sum_{l=1}^k n_l^2}\nabla_{\boldsymbol{\nu}\boldsymbol{\nu}} h(\boldsymbol{\nu};\mathbf{y})$ is

$$\Big|\frac{1}{\sum_{l=1}^k n_l^2}\nabla_{\boldsymbol{\nu}\boldsymbol{\nu}} h(\boldsymbol{\nu};\mathbf{y})\Big| =$$

$$\frac{4}{(\sum_{l=1}^k n_l^2)^2}\Big(\delta\sum_{l=1}^k\sum_{\substack{i,j\in C_l\\ i\neq j}}\exp(-2\lambda M^p\|\mathbf{z}_i-\mathbf{z}_j\|)\sum_{l=1}^k\sum_{\substack{i,j\in C_l\\ i\neq j}}(M^p\|\mathbf{z}_i-\mathbf{z}_j\|)^2\exp(-\lambda M^p\|\mathbf{z}_i-\mathbf{z}_j\|)$$

$$\cdot\Big\{2\delta\exp(-\lambda M^p\|\mathbf{z}_i-\mathbf{z}_j\|)-\delta_o\exp(-\lambda_o M^p\|\mathbf{z}_i-\mathbf{z}_j\|)\Big\}-\Big\{\sum_{l=1}^k\sum_{\substack{i,j\in C_l\\ i\neq j}}M^p\|\mathbf{z}_i-\mathbf{z}_j\|$$

$$\cdot\exp(-\lambda M^p\|\mathbf{z}_i-\mathbf{z}_j\|)[\delta_o\exp(-\lambda_o M^p\|\mathbf{z}_i-\mathbf{z}_j\|)-2\delta\exp(-\lambda M^p\|\mathbf{z}_i-\mathbf{z}_j\|)]\Big\}^2\Big)+o_p(1)$$

$$=\Pi(\delta,\lambda)+o_p(1)$$

It needs to be shown that there exists a non shrinking neighborhood around $\boldsymbol{\nu}_o$ such that on that neighborhood $\Pi(\delta,\lambda)>0$. Let $\delta=\delta_o+\delta_\star$. By choosing $\delta_\star$ to be small, we can bound any term involving $\delta_\star$ in $\Pi(\delta+\delta_\star,\lambda)$ by ϵ_1. Now let $\lambda=\lambda_o+\lambda_\star$, then by choosing $\lambda_\star$ to be small, we can bound any term involving $\exp(-\lambda_o M^p\|\mathbf{z}_i-\mathbf{z}_j\|)-\exp(-(\lambda_o+\lambda_\star)M^p\|\mathbf{z}_i-\mathbf{z}_j\|)$ by ϵ_2. That is

$$\Big|\frac{1}{\sum_{l=1}^k n_l^2}\nabla_{\boldsymbol{\nu}\boldsymbol{\nu}} h(\boldsymbol{\nu};\mathbf{y})\Big| = \frac{4\delta_o^2}{(\sum_{l=1}^k n_l^2)^2}\Big(\sum_{l=1}^k\sum_{\substack{i,j\in C_l\\ i\neq j}}\exp(-2(\lambda_o+\lambda_\star)M^p\|\mathbf{z}_i-\mathbf{z}_j\|)$$

$$\cdot\sum_{l=1}^k\sum_{\substack{i,j\in C_l\\ i\neq j}}\exp(-2(\lambda_o+\lambda_\star)M^p\|\mathbf{z}_i-\mathbf{z}_j\|)(M^p\|\mathbf{z}_i-\mathbf{z}_j\|)^2$$

$$-\Big\{\sum_{l=1}^k\sum_{\substack{i,j\in C_l\\ i\neq j}}M^p\|\mathbf{z}_i-\mathbf{z}_j\|\exp(-2(\lambda_o+\lambda_\star)M^p\|\mathbf{z}_i-\mathbf{z}_j\|)\Big\}^2\Big)+\epsilon_1+\epsilon_2+o_p(1)$$

$$=\frac{4\delta_o^2}{(\sum_{l=1}^k n_l^2)^2}\frac{1}{2}\sum_{l_1=1}^k\sum_{\substack{i_1,j_1\in C_{l_1}\\ i_1\neq j_1}}\sum_{l_2=1}^k\sum_{\substack{i_2,j_2\in C_{l_2}\\ i_2\neq j_2}}\exp(-2(\lambda_o+\lambda_\star)M^p\{\|\mathbf{z}_{i_1}-\mathbf{z}_{j_1}\|+\|\mathbf{z}_{i_2}-\mathbf{z}_{j_2}\|\})$$

$$\cdot\Big(M^p\|\mathbf{z}_{i_1}-\mathbf{z}_{j_1}\|-M^p\|\mathbf{z}_{i_2}-\mathbf{z}_{j_2}\|\Big)^2+\epsilon_1+\epsilon_2+o_p(1)\quad\text{[by Lemma 3.3]}$$

By (B3) and since $\mathbf{z}_i$ are in a finite dimensional space, the leading term on the right hand side of $\Big|\frac{1}{\sum_{l=1}^k n_l^2}\nabla_{\boldsymbol{\nu}\boldsymbol{\nu}} h(\boldsymbol{\nu};\mathbf{y})\Big|$ is strictly positive. Hence, we have shown

that there exists a non-shrinking neighborhood around $\boldsymbol{\nu}_o$ such that uniformly for large k, $n_1, \ldots, n_k$, $\left| \frac{1}{\sum_{l=1}^{k} n_l^2} \nabla_{\nu\nu} h(\boldsymbol{\nu}; \mathbf{y}) \right|$ is strictly positive.

Now combining all the above arguments, for large m, with probability one, $\frac{1}{\sum_{l=1}^{k} n_l^2} \nabla_{\nu\nu} h(\boldsymbol{\nu}; \mathbf{y})$ is negative definite and hence, strictly concave in a neighborhood around $\boldsymbol{\nu}_o$. Let that neighborhood be denoted by B_ϵ. Let $B_\epsilon^\star \subset B_\epsilon$ such that $B_\epsilon^\star$ is compact. Then an application of Theorem 3.5 gives

$$\sup_{\boldsymbol{\nu} \in B_\epsilon^\star} \left| \frac{1}{\sum_{l=1}^{k} n_l^2} \Big(h(\boldsymbol{\nu}; \mathbf{y}) - h(\boldsymbol{\nu}) \Big) \right| \xrightarrow{p} 0 \tag{3.34}$$

Then by (3.34) and since $\frac{1}{\sum_{l=1}^{k} n_l^2} h(\boldsymbol{\nu})$ has a unique maximum at $\boldsymbol{\nu} = \boldsymbol{\nu}_o$, an application of Theorem 3.6 gives the desired result: there exists a sequence of local maxima which is consistent for $\boldsymbol{\nu}_o$. □

Proof of Theorem 3.4. Since it has been shown that $\widehat{\boldsymbol{\nu}}$ is consistent for $\boldsymbol{\nu}_o$, we expand $\nabla_\nu h(\widehat{\boldsymbol{\nu}}; \mathbf{y})$ around $\boldsymbol{\nu}_o$ to obtain

$$\begin{aligned} 0 \equiv \nabla_\nu h(\widehat{\boldsymbol{\nu}}; \mathbf{y}) = \nabla_\nu h(\boldsymbol{\nu}_o; \mathbf{y}) + \nabla_{\nu\nu} h(\boldsymbol{\nu}_o; \mathbf{y})(\widehat{\boldsymbol{\nu}} - \boldsymbol{\nu}_o) + \\ \frac{1}{2} \sum_{j=1}^{2} \frac{\partial}{\partial \nu_j} \{ \nabla_{\nu\nu} h(\boldsymbol{\nu}_\star; \mathbf{y}) \} (\widehat{\boldsymbol{\nu}} - \boldsymbol{\nu}_o)(\hat{\nu}_j - \nu_{oj}) \end{aligned} \tag{3.35}$$

where $\boldsymbol{\nu}_\star$ lies between $\widehat{\boldsymbol{\nu}}$ and $\boldsymbol{\nu}_o$, and $\widehat{\boldsymbol{\nu}} = (\hat{\nu}_1, \hat{\nu}_2)'$.

Step (i): We first seek to derive the asymptotic distribution of a properly normalized $\nabla_\nu h(\boldsymbol{\nu}_o; \mathbf{y})$, where

$$\nabla_\nu h(\boldsymbol{\nu}_o; \mathbf{y}) = \begin{pmatrix} \frac{\partial}{\partial \delta} h(\boldsymbol{\nu}_o; \mathbf{y}) \\ \frac{\partial}{\partial \lambda} h(\boldsymbol{\nu}_o; \mathbf{y}) \end{pmatrix}$$

$$\frac{\partial}{\partial\delta}h(\boldsymbol{\nu}_o;\mathbf{y}) = 2\sum_{l=1}^{k}\sum_{\substack{i,j\in C_l\\ i\neq j}}\Big(\hat{\epsilon}_i\hat{\epsilon}_j - \delta_o\exp(-\lambda_o M^p\|\mathbf{z}_i-\mathbf{z}_j\|)\Big)\exp(-\lambda_o M^p\|\mathbf{z}_i-\mathbf{z}_j\|)$$

$$\frac{\partial}{\partial\lambda}h(\boldsymbol{\nu}_o;\mathbf{y}) = -2\delta_o\sum_{l=1}^{k}\sum_{\substack{i,j\in C_l\\ i\neq j}}\Big(\hat{\epsilon}_i\hat{\epsilon}_j - \delta_o\exp(-\lambda_o M^p\|\mathbf{z}_i-\mathbf{z}_j\|)\Big)M^p\|\mathbf{z}_i-\mathbf{z}_j\|$$

$$\cdot\exp(-\lambda_o M^p\|\mathbf{z}_i-\mathbf{z}_j\|)$$

In the expression for $\frac{\partial}{\partial\delta}h(\boldsymbol{\nu}_o;\mathbf{y})$, we write $\hat{\epsilon}_i$, $\hat{\epsilon}_j$ as $(y_i-\mathbf{x}_i'\boldsymbol{\beta}_o-\mathbf{x}_i'\widehat{\mathbf{b}})$, $(y_j-\mathbf{x}_j'\boldsymbol{\beta}_o-\mathbf{x}_j'\widehat{\mathbf{b}})$ and argue that all terms in $\frac{1}{(\sum_{l=1}^k n_l^4)^{\frac{1}{2}}}\frac{\partial}{\partial\delta}h(\boldsymbol{\nu}_o;\mathbf{y})$ that involve $\widehat{\mathbf{b}}$ are $o_p(1)$.

Consider the term $\frac{1}{(\sum_{l=1}^k n_l^4)^{\frac{1}{2}}}\sum_{l=1}^k\sum_{\substack{i,j\in C_l\\ i\neq j}}(y_i-\mathbf{x}_i'\boldsymbol{\beta}_o)\mathbf{x}_j'\widehat{\mathbf{b}}\exp(-\lambda_o M^p\|\mathbf{z}_i-\mathbf{z}_j\|)$.

$$\frac{1}{(\sum_{l=1}^k n_l^4)^{\frac{1}{2}}}\mathrm{E}\Big|\sum_{l=1}^{k}\sum_{\substack{i,j\in C_l\\ i\neq j}}(y_i-\mathbf{x}_i'\boldsymbol{\beta}_o)\|\mathbf{x}_j\|\exp(-\lambda_o M^p\|\mathbf{z}_i-\mathbf{z}_j\|)\Big|$$

$$\leq\Big(\frac{1}{\sum_{l=1}^k n_l^4}\mathrm{E}\Big(\sum_{l=1}^{k}\sum_{\substack{i,j\in C_l\\ i\neq j}}(y_i-\mathbf{x}_i'\boldsymbol{\beta}_o)\|\mathbf{x}_j\|\exp(-\lambda_o M^p\|\mathbf{z}_i-\mathbf{z}_j\|)\Big)^2\Big)^{\frac{1}{2}}$$

Now we expand the above expression as within and between cluster terms, that is,

$$\mathrm{E}\Big(\sum_{l=1}^{k}\sum_{\substack{i,j\in C_l\\ i\neq j}}(y_i-\mathbf{x}_i'\boldsymbol{\beta}_o)\|\mathbf{x}_j\|\exp(-\lambda_o M^p\|\mathbf{z}_i-\mathbf{z}_j\|)\Big)^2 \leq \sum_{l=1}^{k}\mathrm{E}\Big(\sum_{\substack{i,j\in C_l\\ i\neq j}}(y_i-\mathbf{x}_i'\boldsymbol{\beta}_o)\|\mathbf{x}_j\|\Big)^2$$

$$+\sup_{l_1\neq l_2}\sup_{\substack{i\in C_{l_1}\\ j\in C_{l_2}}}\exp(-\lambda_o M^p\|\mathbf{z}_i-\mathbf{z}_j\|)\sum_{l_1\neq l_2}\mathrm{E}\Big(\sum_{\substack{i\in C_{l_1}\\ j\in C_{l_2}}}(y_i-\mathbf{x}_i'\boldsymbol{\beta}_o)\|\mathbf{x}_i\|(y_j-\mathbf{x}_i'\boldsymbol{\beta}_o)\|\mathbf{x}_j\|\Big)^2$$

The sum within a cluster involves n_l^2 terms, and hence, the expectation of this sum is $O(n_l^4)$. The sum between cluster involves $n_{l_1}n_{l_2}$ terms, so that the expectation

of this latter sum is $O(n_{l_1}^2 n_{l_2}^2)$. Hence,

$$\begin{aligned}
&\frac{1}{\sum_{l=1}^k n_l^4}\mathrm{E}\Big(\sum_{l=1}^k \sum_{\substack{i,j\in C_l\\ i\neq j}} (y_i - \mathbf{x}_i'\boldsymbol{\beta}_o)\|\mathbf{x}_j\| \exp(-\lambda_o M^p \|\mathbf{z}_i - \mathbf{z}_j\|)\Big)^2 \\
&= \frac{1}{\sum_{l=1}^k n_l^4}\sum_{l=1}^k O(n_l^4) + \frac{1}{\sum_{l=1}^k n_l^4}\sum_{l_1\neq l_2} O(n_{l_1}^2 n_{l_2}^2) \sup_{\substack{i\in C_{l_1}\\ j\in C_{l_2}}} \exp(-\lambda_o M^p \|\mathbf{z}_i - \mathbf{z}_j\|) \\
&= \frac{O\Big(\sum_{l=1}^k n_l^4\Big)}{\sum_{l=1}^k n_l^4} + \sup_{l_1\neq l_2}\sup_{\substack{i\in C_{l_1}\\ j\in C_{l_2}}} M\exp(-\lambda_o M^p \|\mathbf{z}_i - \mathbf{z}_j\|)\,\frac{O\Big(\Big(\sum_{l=1}^k n_l^2\Big)^2\Big)}{M\sum_{l=1}^k n_l^4} \\
&\leq c \qquad \Big[\text{by (B1), (B2) and } \frac{O\Big(\Big(\sum_{l=1}^k n_l^2\Big)^2\Big)}{M\sum_{l=1}^k n_l^4} \to 0\Big]
\end{aligned}$$

Hence,

$$\frac{1}{(\sum_{l=1}^k n_l^4)^{\frac{1}{2}}}\mathrm{E}\Big|\sum_{l=1}^k \sum_{\substack{i,j\in C_l\\ i\neq j}} (y_i - \mathbf{x}_i'\boldsymbol{\beta}_o)\|\mathbf{x}_j\| \exp(-\lambda_o M^p \|\mathbf{z}_i - \mathbf{z}_j\|)\Big| \leq c \tag{3.36}$$

$$\frac{1}{(\sum_{l=1}^k n_l^4)^{\frac{1}{2}}}\sum_{l=1}^k \sum_{\substack{i,j\in C_l\\ i\neq j}} (y_i - \mathbf{x}_i'\boldsymbol{\beta}_o)\mathbf{x}_j'\widehat{\mathbf{b}} \exp(-\lambda_o M^p \|\mathbf{z}_i - \mathbf{z}_j\|) \xrightarrow{p} 0 \ \text{ [by (3.36) and } \widehat{\mathbf{b}} \xrightarrow{p} 0]$$

Also,

$$\begin{aligned}
&\frac{1}{(\sum_{l=1}^k n_l^4)^{\frac{1}{2}}}\sum_{l=1}^k \sum_{\substack{i,j\in C_l\\ i\neq j}} \mathbf{x}_i'\widehat{\mathbf{b}}\mathbf{x}_j'\widehat{\mathbf{b}} \exp(-\lambda_o M^p \|\mathbf{z}_i - \mathbf{z}_j\|)\Big| \leq \frac{\|\widehat{\mathbf{b}}\|^2}{(\sum_{l=1}^k n_l^4)^{\frac{1}{2}}}\sum_{l=1}^k n_l^2 \sum_{\substack{i,j\in C_l\\ i\neq j}} \frac{\|\mathbf{x}_i\|\|\mathbf{x}_j\|}{n_l^2} \\
&\qquad \xrightarrow{p} 0 \qquad [\text{by (B1), (B4), (B8)}]
\end{aligned}$$

Using the matrices defined in Theorem 3.4, we have shown

$$\begin{aligned}
\frac{1}{(\sum_{l=1}^k n_l^4)^{\frac{1}{2}}}\frac{\partial}{\partial\delta}h(\boldsymbol{\nu}_o;\mathbf{y}) &= \frac{2}{(\sum_{l=1}^k n_l^4)^{\frac{1}{2}}}\Big((\mathbf{y} - X\boldsymbol{\beta}_o)'G_o(\mathbf{y} - X\boldsymbol{\beta}_o) - \delta_o \mathrm{tr}[G_o^2]\Big) + o_p(1) \\
&= R_1 + o_p(1)
\end{aligned}$$

Similarly,

$$\frac{1}{(\sum_{l=1}^{k} n_l^4)^{\frac{1}{2}}}\frac{\partial}{\partial\lambda}h(\boldsymbol{\nu}_o;\mathbf{y}) = \frac{-2}{(\sum_{l=1}^{k} n_l^4)^{\frac{1}{2}}}\Big((\mathbf{y}-X\boldsymbol{\beta}_o)'H_o(\mathbf{y}-X\boldsymbol{\beta}_o) - \delta_o \text{tr}[G_oH_o]\Big) + o_p(1)$$

$$= R_2 + o_p(1)$$

Next we derive the asymptotic distribution of $(R_1, R_2)'$ and use Slutsky's Lemma to claim that the asymptotic distribution of $(R_1, R_2)'$ is the same as the asymptotic distribution of $\frac{1}{(\sum_{l=1}^{k} n_l^4)^{\frac{1}{2}}}\nabla_{\boldsymbol{\nu}} h(\boldsymbol{\nu}_o;\mathbf{y})$.

From Theorem 3.2c.2 of Mathai and Provost [27], the moment generating function of $(R_1, R_2)'$ is given by

$$M_{R_1,R_2}(t_1,t_2) = \Big|I_m - \frac{4t_1G_oV_o}{(\sum_{l=1}^{k} n_l^4)^{\frac{1}{2}}} + \frac{4t_2H_oV_o}{(\sum_{l=1}^{k} n_l^4)^{\frac{1}{2}}}\Big|^{-\frac{1}{2}}\exp\Big(\frac{-2\delta_o t_1 \text{tr}[G_o^2] + 2\delta_o t_2 \text{tr}[G_oH_o]}{(\sum_{l=1}^{k} n_l^4)^{\frac{1}{2}}}\Big)$$

Let $s(t_1,t_2) = \log\Big|I_m - \frac{4t_1G_oV_o}{(\sum_{l=1}^{k} n_l^4)^{\frac{1}{2}}} + \frac{4t_2H_oV_o}{(\sum_{l=1}^{k} n_l^4)^{\frac{1}{2}}}\Big|$. Using Lemma 3.6 (d),(e), we get

$$\frac{\partial s(0,0)}{\partial t_1} = -\frac{4\text{tr}[G_oV_o]}{(\sum_{l=1}^{k} n_l^4)^{\frac{1}{2}}}$$

$$\frac{\partial s(0,0)}{\partial t_2} = \frac{4\text{tr}[H_oV_o]}{(\sum_{l=1}^{k} n_l^4)^{\frac{1}{2}}}$$

$$\frac{\partial^2 s(0,0)}{\partial t_1^2} = -\frac{16\text{tr}[G_oV_oG_oV_o]}{\sum_{l=1}^{k} n_l^4}$$

$$\frac{\partial^2 s(0,0)}{\partial t_2^2} = -\frac{16\text{tr}[H_oV_oH_oV_o]}{\sum_{l=1}^{k} n_l^4}$$

$$\frac{\partial^2 s(0,0)}{\partial t_1\partial t_2} = \frac{16\text{tr}[G_oV_oH_oV_o]}{\sum_{l=1}^{k} n_l^4}$$

Note that $\frac{\partial^{(i+j)}s(0,0)}{\partial t_1^{(i)}\partial t_2^{(j)}} = o(1)$, if $i+j \geq 3$. This follows from (B1),(B2),(B6) and

Lemma 3.2. For example,

$$\begin{aligned}\frac{\partial^3 s(0,0)}{\partial t_1^3} &= c \cdot \frac{\mathrm{tr}[(G_o V_o)^3]}{(\sum_{l=1}^k n_l^4)^{\frac{3}{2}}} \\ &= \frac{O(\sum_{l=1}^k n_l^6)}{(\sum_{l=1}^k n_l^4)^{\frac{3}{2}}} \qquad \text{[by (B1), (B2)]} \\ &\to 0 \qquad \text{[by (B2), (B6) and Lemma 3.2]}\end{aligned}$$

Next we expand $\log M_{R_1,R_2}(t_1,t_2)$ around $(0,0)$,

$$\begin{aligned}\log M_{R_1,R_2}(t_1,t_2) = -\frac{1}{2}\Big(&\frac{-4\mathrm{tr}[G_o V_o]}{(\sum_{l=1}^k n_l^4)^{\frac{1}{2}}} t_1 + \frac{4\mathrm{tr}[H_o V_o]}{(\sum_{l=1}^k n_l^4)^{\frac{1}{2}}} t_2 - \frac{16\mathrm{tr}[G_o V_o G_o V_o]}{\sum_{l=1}^k n_l^4}\frac{t_1^2}{2} \\ &- \frac{16\mathrm{tr}[H_o V_o H_o V_o]}{\sum_{l=1}^k n_l^4}\frac{t_2^2}{2} + \frac{16\mathrm{tr}[G_o V_o H_o V_o]}{\sum_{l=1}^k n_l^4} t_1 t_2 + o(1)\Big) - \frac{2\delta_o(\mathrm{tr}[G_o^2]t_1 - \mathrm{tr}[G_o H_o]t_2)}{(\sum_{l=1}^k n_l^4)^{\frac{1}{2}}}\end{aligned}$$

But $\mathrm{tr}[G_o V_o] = \delta_o \mathrm{tr}[G_o^2]$ and $\mathrm{tr}[H_o V_o] = \delta_o \mathrm{tr}[G_o H_o]$, hence

$$\log M_{R_1,R_2}(t_1,t_2) = \frac{4\mathrm{tr}[G_o V_o G_o V_o]}{\sum_{l=1}^k n_l^4} t_1^2 + \frac{4\mathrm{tr}[H_o V_o H_o V_o]}{\sum_{l=1}^k n_l^4} t_2^2 - \frac{8\mathrm{tr}[G_o V_o H_o V_o]}{\sum_{l=1}^k n_l^4} t_1 t_2 + o(1)$$

$$\Rightarrow \quad K_o^{-\frac{1}{2}} \begin{pmatrix} R_1 \\ R_2 \end{pmatrix} \xrightarrow{d} \mathrm{N}(\mathbf{0}_2, I_2)$$

where K_o is defined in Theorem 3.4. By Slutsky's Lemma,

$$\frac{K_o^{-\frac{1}{2}}}{(\sum_{l=1}^k n_l^4)^{\frac{1}{2}}} \nabla_{\boldsymbol{\nu}} h(\boldsymbol{\nu}_o; \mathbf{y}) \xrightarrow{d} \mathrm{N}(\mathbf{0}_2, I_2) \tag{3.37}$$

Step (ii): Next we show

$$-\frac{1}{\sum_{l=1}^k n_l^2} \nabla_{\boldsymbol{\nu}\boldsymbol{\nu}} h(\boldsymbol{\nu}_o; \mathbf{y}) L_o^{-1} \xrightarrow{p} I_2 \tag{3.38}$$

In Theorem 3.3, it was established that each term of $\frac{1}{\sum_{l=1}^k n_l^2} \nabla_{\boldsymbol{\nu}\boldsymbol{\nu}} h(\boldsymbol{\nu}; \mathbf{y})$ con-

verges in probability. Evaluating those expressions at $\boldsymbol{\nu}_o$ we get

$$\frac{1}{\sum_{l=1}^k n_l^2}\frac{\partial^2}{\partial\delta^2}h(\boldsymbol{\nu}_o;\mathbf{y}) = \frac{-2}{\sum_{l=1}^k n_l^2}\sum_{l=1}^k\sum_{\substack{i,j\in C_l\\ i\neq j}}\exp(-2\lambda_o M^p\|\mathbf{z}_i-\mathbf{z}_j\|)$$

$$\frac{1}{\sum_{l=1}^k n_l^2}\frac{\partial^2}{\partial\lambda^2}h(\boldsymbol{\nu}_o;\mathbf{y}) \overset{p}{\approx} \frac{-2\delta_o^2}{\sum_{l=1}^k n_l^2}\sum_{l=1}^k\sum_{\substack{i,j\in C_l\\ i\neq j}}(M^p\|\mathbf{z}_i-\mathbf{z}_j\|)^2\exp(-2\lambda_o M^p\|\mathbf{z}_i-\mathbf{z}_j\|)$$

$$\frac{1}{\sum_{l=1}^k n_l^2}\frac{\partial^2}{\partial\delta\partial\lambda}h(\boldsymbol{\nu}_o;\mathbf{y}) \overset{p}{\approx} \frac{2\delta_o}{\sum_{l=1}^k n_l^2}\sum_{l=1}^k\sum_{\substack{i,j\in C_l\\ i\neq j}}M^p\|\mathbf{z}_i-\mathbf{z}_j\|\exp(-2\lambda_o M^p\|\mathbf{z}_i-\mathbf{z}_j\|)$$

By expressing the above terms by the matrices defined in Theorem 3.4, (3.38) follows immediately.

Step (iii)*:* Finally it has to be shown that $\frac{1}{\sum_{l=1}^k n_l^2}\sum_{j=1}^2\frac{\partial}{\partial\nu_j}\{\nabla_{\nu\nu}h(\boldsymbol{\nu}_\star;\mathbf{y})\}$ is bounded in probability. We show this for one of the terms; the remainder is similar and hence the proof is omitted. For example, consider

$$\frac{1}{\sum_{l=1}^k n_l^2}\frac{\partial^3}{\partial\lambda^3}h(\boldsymbol{\nu};\mathbf{y}) = \frac{-2\delta}{\sum_{l=1}^k n_l^2}\sum_{l=1}^k\sum_{\substack{i,j\in C_l\\ i\neq j}}\Big(\hat{\epsilon}_i\hat{\epsilon}_j(M^p\|\mathbf{z}_i-\mathbf{z}_j\|)^3\exp(-\lambda M^p\|\mathbf{z}_i-\mathbf{z}_j\|)$$
$$-4\delta(M^p\|\mathbf{z}_i-\mathbf{z}_j\|)^3\exp(-2\lambda M^p\|\mathbf{z}_i-\mathbf{z}_j\|)\Big).$$

Then,

$$\Big|\frac{1}{\sum_{l=1}^k n_l^2}\frac{\partial^3}{\partial\lambda^3}h(\boldsymbol{\nu}_\star;\mathbf{y})\Big| \leq \frac{2\,c^3\,\delta_\star}{\sum_{l=1}^k n_l^2}\sum_{l=1}^k\sum_{\substack{i,j\in C_l\\ i\neq j}}\Big(|(y_i-\mathbf{x}_i'\boldsymbol{\beta}_o-\mathbf{x}_i'\widehat{\mathbf{b}})(y_j-\mathbf{x}_j'\boldsymbol{\beta}_o-\mathbf{x}_j'\widehat{\mathbf{b}})|+4\delta_\star\Big)$$

where $\boldsymbol{\nu}_\star=(\delta_\star,\lambda_\star)'$ and by (B2),

$$c=\sup_{1\leq l\leq k}\limsup_{M\to\infty}M^p\sup_{i,j\in C_l}\|\mathbf{z}_i-\mathbf{z}_j\|<\infty$$

Similar to the proofs shown earlier, terms that involve $\widehat{\mathbf{b}}$ are $o_p(1)$, that is,

$$\Big|\frac{1}{\sum_{l=1}^k n_l^2}\frac{\partial^3}{\partial\lambda^3}h(\boldsymbol{\nu}_\star;\mathbf{y})\Big| \leq \frac{2c^3(|\delta_\star-\delta_o|+\delta_o)}{\sum_{l=1}^k n_l^2}\sum_{l=1}^k\sum_{\substack{i,j\in C_l\\ i\neq j}}\Big(|(y_i-\mathbf{x}_i'\boldsymbol{\beta}_o)(y_j-\mathbf{x}_j'\boldsymbol{\beta}_o)| + 4(|\delta_\star-\delta_o|+\delta_o)\Big)+o_p(1)$$

And now we have the right hand side above converging in probability to a constant and hence, $\Big|\frac{1}{\sum_{l=1}^k n_l^2}\frac{\partial^3}{\partial\lambda^3}h(\boldsymbol{\nu}_\star;\mathbf{y})\Big|$ is bounded in probability.

Step (iv): Rearranging terms in (3.35) and normalizing by $\frac{\sum_{l=1}^k n_l^2}{(\sum_{l=1}^k n_l^4)^{\frac{1}{2}}}$, we get

$$\frac{\sum_{l=1}^k n_l^2}{(\sum_{l=1}^k n_l^4)^{\frac{1}{2}}}K_o^{-\frac{1}{2}}L_o(\widehat{\boldsymbol{\nu}}-\boldsymbol{\nu}_o) = K_o^{-\frac{1}{2}}L_o\Big(-\frac{\nabla_{\nu\nu}h(\boldsymbol{\nu}_o;\mathbf{y})}{\sum_{l=1}^k n_l^2} - \frac{1}{2\sum_{l=1}^k n_l^2}\sum_{j=1}^2\frac{\partial}{\partial\nu_j}\{\nabla_{\nu\nu}h(\boldsymbol{\nu}_\star;\mathbf{y})\}(\hat{\nu}_j-\nu_{oj})\Big)^{-1}K^{\frac{1}{2}}K^{-\frac{1}{2}}\frac{\nabla_\nu h(\boldsymbol{\nu}_o;\mathbf{y})}{(\sum_{l=1}^k n_l^4)^{\frac{1}{2}}}$$

By (3.37)-(3.38), and for $j=1,2$, $\hat{\nu}_j \xrightarrow{p} \nu_{oj}$, and since $\Big|\frac{1}{\sum_{l=1}^k n_l^2}\frac{\partial^3}{\partial\lambda^3}h(\boldsymbol{\nu}_\star;\mathbf{y})\Big|$ is bounded in probability, it follows that

$$\frac{\sum_{l=1}^k n_l^2}{(\sum_{l=1}^k n_l^4)^{\frac{1}{2}}}K_o^{-\frac{1}{2}}L_o(\widehat{\boldsymbol{\nu}}-\boldsymbol{\nu}_o) \xrightarrow{d} \mathrm{N}(\mathbf{0}_2,\mathbf{I}_2)$$

□

Chapter 4

Simulation study

In this chapter we are interested in analyzing the following through simulation:

1. Find the relative efficiency of the predictor obtained under the true model and the predictor obtained under the misspecified Fay-Herriot model.

2. Check if our estimation methods are adversely affected by small, but non-negligible correlation between neighboring clusters.

3. Analyze the properties of the estimators derived in Sections 3.1-3.2, and compare them with the MLE.

4. Check if increasing the dimension of $\mathbf{z}_i$, the vector of spatial locations and covariates, decreases the standard error of the estimators.

4.1 Simulation setup

The $\mathbf{z}_i$'s for the simulation study are spatial co-ordinates, and except for item 4 above, the domain for $\mathbf{z}_i$ is $[0, 10]^2$, while for item 4, $\mathbf{z}_i \in [0, 10]^4$. For a given

number of clusters k, the cluster centers were chosen to be equally spaced on a grid in $[0, 10]^2$. For each cluster, the $\mathbf{z}_i$'s were generated using an independent bivariate normal distribution such that all $\mathbf{z}_i$ in a cluster were within a radius of r of the cluster center (the standard deviation of the normal distribution was taken to be $3r/10$). For simplicity, for $l = 1, \ldots, k$, we took $N_l = N$. Note that for any specific parameter combination of $(M, N, k, \delta, \lambda, \sigma^2)$, the $\mathbf{z}_i$'s were generated only once and then fixed for all simulation runs for that parameter combination.

For the simulation, to mimic (2.7)-(2.8), when M is increased from M_1 to M_2, we need to decrease the radius of the clusters and the standard deviation of the bivariate normal distribution that generates the $\mathbf{z}_i$'s. This is done as follows:

$$r_2 = r_1 \frac{M_1^p}{M_2^p}$$

where r_1 is the original radius and r_2 is the new radius of the clusters. The standard deviation of the bivariate normal distribution that generates the $\mathbf{z}_i$'s is then decreased to $3r_2/10$.

Except for item 1 above, we generated data without sampling errors, we took $M = m$, $N = n$ (that is, a purely spatial model with no sampling) and we included a covariate (fraction of adult population with bachelor's degree) from the data set analyzed in Chapter 5. The fixed effect parameter $\boldsymbol{\beta} = (\beta_1, \beta_2)' = (1, 1)'$ was fixed for the entire simulation study. For item 1, the data were generated with sampling errors (see Section 4.2): 50% of the population was sampled and we included an intercept $\beta = 1$. Moreover, except when generating data for item 2 above, the data were generated so that the clusters were independent. Note that generating data so

that the clusters are independent is equivalent to assuming that the clusters are well separated [Assumption (C) and (2.7)-(2.8)] . In item 2 above, we generated data so that the clusters were dependent (see Section 4.3).

For the entire simulation study the following parameters were fixed: $p = 0.25$, $\tau^2 = \delta + \sigma^2 = 1$. To be able to interpret λ, we choose it as follows: λ is chosen so that the median value of the within cluster off-diagonal entries of A_{U} is some number c. Hence, the approximate median covariance among all pairs of small areas in a cluster is δc.

Next, we define some terminology that is used in this chapter. The estimators $(\widehat{\boldsymbol{\beta}}_{\text{FH}}, \hat{\tau}^2_{\text{FH}}, \hat{\delta}, \hat{\lambda})$ derived in Sections 3.1-3.2 are referred to as the "least squares estimator" (LSE). To differentiate the MLE from the LSE, we use a subscript of 'ML' to denote the MLE.

The *empirical mean squared error* of an estimator $\hat{\alpha}$ is defined as

$$mse(\hat{\alpha}) = \frac{1}{R}\sum_{r=1}^{R}(\hat{\alpha}_r - \alpha_o)^2 \tag{4.1}$$

where R is the number of simulation runs for a specific choice of parameter values, $\hat{\alpha}_r$ is the estimate of α in the r^{th} simulation run and α_o is the true value of the parameter. We refer to $\sqrt{mse(\hat{\alpha})}$ as the *empirical root mean squared error* (abbreviated as e.s.e.). Moreover, the theoretical standard error, abbreviated as s.e., is obtained for LSE from Theorems 3.2 and 3.4, while for MLE it is obtained from the information matrix (3.8).

The *relative efficiency* of two estimators $\hat{\alpha}_{\text{ML}}$ and $\tilde{\alpha}$ of a parameter α is defined

to be

$$\mathrm{RE}(\hat{\alpha}_{\mathrm{ML}}, \tilde{\alpha}) = \frac{var(\tilde{\alpha})}{var(\hat{\alpha}_{\mathrm{ML}})}. \tag{4.2}$$

where $var(\hat{\alpha}_{\mathrm{ML}})$ and $var(\tilde{\alpha})$ are the empirical variances of $\hat{\alpha}_{\mathrm{ML}}$ and $\tilde{\alpha}$.

For each set of parameter values, when estimating by LSE, we ran 500 simulation runs. However, when estimating by MLE, we ran 100 simulation runs. In particular, when computing the MLE, we used the first 100 batches of simulated data that was used to compute the LSE. The difference in simulation runs is due to time constraints. Using the R command unix.time, for $k = 40, N = 20$, it was estimated that the MLE takes approximately 25 times the running time of LSE. Because of the small number of simulations runs we do not claim the simulation is definitive, but we do claim that certain patterns emerge which support our conclusions.

Finally, note that in Sections 3.1-3.2, we derived estimators for $(\boldsymbol{\beta}, \tau^2, \delta, \lambda)$. Since $\tau^2 = \delta + \sigma^2$, we can estimate σ^2 by $\hat{\sigma}^2 = \hat{\tau}^2_{\mathrm{FH}} - \hat{\delta}$, which is a consistent estimator of σ_o^2. However, we do not have a formula for $\mathrm{var}(\hat{\sigma}^2)$. We defer deriving a least squares estimator for σ^2 to future research.

4.2 Comparison of predictors

Here we are interested in computing the relative efficiency of the EBLUP $\hat{\theta}_i(\widehat{\boldsymbol{\eta}})$ [(2.18)] obtained under the true model, and the EBLUP $\tilde{\theta}_i(\hat{\tau}^2_{\mathrm{FH}})$ obtained under the Fay-Herriot model, where $\tilde{\theta}_i(\hat{\tau}^2_{\mathrm{FH}})$ is given by

$$\tilde{\theta}_i(\hat{\tau}^2_{\mathrm{FH}}) = \begin{cases} \mathbf{x}_i'\widehat{\boldsymbol{\beta}}_{\mathrm{FH}} + \Big(\hat{\tau}^2_{\mathrm{FH}}/(\hat{\tau}^2_{\mathrm{FH}} + \psi_i)\Big)(y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}}_{\mathrm{FH}}) & \text{if } i \in S \\ \mathbf{x}_i'\widehat{\boldsymbol{\beta}}_{\mathrm{FH}} & \text{if } i \in S^c. \end{cases}$$

where $(\widehat{\boldsymbol{\beta}}_{\text{FH}}, \hat{\tau}^2_{\text{FH}})$ is given by (3.2). As mentioned in Chapter 2, the relative efficiency of two predictors $\hat{\theta}_i(\widehat{\boldsymbol{\eta}})$ and $\tilde{\theta}_i(\hat{\tau}^2_{\text{FH}})$ is defined as

$$R(\hat{\theta}_i(\widehat{\boldsymbol{\eta}}), \tilde{\theta}_i(\hat{\tau}^2_{\text{FH}})) = \frac{\text{MSE}[\,\tilde{\theta}_i(\hat{\tau}^2_{\text{FH}})\,]}{\text{MSE}[\,\hat{\theta}_i(\widehat{\boldsymbol{\eta}})\,]} \tag{4.3}$$

When computing $\hat{\theta}_i(\widehat{\boldsymbol{\eta}})$, since we need to estimate $\boldsymbol{\eta} = (\delta, \lambda, \sigma^2)'$, we do so by two different methods: LSE and MLE.

In our study, we fixed $M = 1200$, $m = 600$, $\beta = 1$ (the intercept) and considered two different patterns for (k, n, N): $k = 30$, $n = 20$, $N = 40$ and $k = 15$, $n = 40$, $N = 80$, and 4 different combinations for $(\delta, \lambda, \sigma^2)$. Due to time constraints we do not consider an elaborate experimental design for different parameter combinations, however in a future study we plan on doing so. Moreover, in a future study we will try to provide a more detailed summary as to what combinations of (k, n, N) and $(\delta, \lambda, \sigma^2)$ result in large relative efficiency. The simulation setup is as given in Section 4.1, and the sampling errors ψ_i were generated from a mixture of normal distributions; that is, $\psi_i \sim (1/2)\,\mathbf{N}(0.2, 0.03) + (1/2)\,\mathbf{N}(0.4, 0.07)$. We generate ψ_i's from a bivariate normal distribution so as to mimic a sample survey in which a fraction of the sampled small areas have a much smaller sampling error compared to the remaining sampled small areas.

Since the relative efficiency was significantly different for sampled and non-sampled areas, we summarize our results by these categories. For each of the two estimation methods, the column 'Obs' (sampled small areas) is obtained as follows: we compute $\text{MSE}[\,\tilde{\theta}_i(\hat{\tau}^2_{\text{FH}})\,]$ and $\text{MSE}[\,\hat{\theta}_i(\widehat{\boldsymbol{\eta}})\,]$ by empirically averaging over all sampled small areas and all simulation runs, and then compute the ratio to obtain

$R(\hat{\theta}_i(\widehat{\boldsymbol{\eta}}), \tilde{\theta}_i(\hat{\tau}^2_{\mathrm{FH}}))$. Similarly, we compute the column 'Unobs' (non-sampled small areas).

As can be seen from Table 4.1, the relative efficiency is almost identical for both methods of estimation. Moreover, much larger relative efficiencies are obtained for the non-sampled areas compared to the sampled areas, and large values of δ and small values of λ correspond to large relative efficiencies. Note that $\lambda = 0.12$ corresponds to $c = 0.70$ (that is, the median value of the within cluster off-diagonal entries of A_{U} is 0.70) and $\lambda = 0.41$ corresponds to $c = 0.35$.

Table 4.1: Relative efficiency of EBLUP obtained under the true model and the Fay-Herriot model: $k = 30$, $n = 20$, $N = 40$, $m = 600$, $M = 1200$, $\beta = 1$. 'Obs.', 'Unobs.' respectively refer to the relative efficiency for sampled small areas and the relative efficiency for non-sampled small areas.

			LSE		MLE	
δ	λ	σ^2	Obs.	Unobs.	Obs.	Unobs
0.60	0.12	0.40	1.231	1.720	1.238	1.729
0.30	0.12	0.70	1.072	1.211	1.073	1.212
0.60	0.41	0.40	1.095	1.331	1.096	1.331
0.30	0.41	0.70	1.033	1.099	1.034	1.099

When $k = 15$, $n = 40$, $N = 80$, in Table 4.2, we provide only the relative efficiency when LSE is used to estimate the variance components in the EBLUP. As can be seen, the results are similar to the corresponding parameter combinations in

Table 4.2: Relative efficiency of EBLUP obtained under the true model and the Fay-Herriot model: $k = 15$, $n = 40$, $N = 80$, $m = 600$, $M = 1200$, $\beta = 1$.

δ	λ	σ^2	Obs.	Unobs.
0.60	0.12	0.40	1.280	1.861
0.30	0.12	0.70	1.033	1.261

Table 4.1. We note that even though the relative efficiency of the LSE and the MLE can be small for certain parameters, it does not affect prediction. For example, from Table 4.3, by dividing column 9 by column 5 and then squaring it, we get the relative efficiency: the RE($\hat{\delta}$, $\hat{\delta}_{\text{ML}}$) and RE($\hat{\sigma}^2$, $\hat{\sigma}^2_{\text{ML}}$) is 0.70. However, the RE($\hat{\tau}^2_{\text{FH}}$, $\hat{\tau}^2_{\text{ML}}$) is 0.98. Hence one possible reason for the difference in relative efficiency between LSE and MLE not affecting prediction could be explained by $\hat{\theta}_i(\boldsymbol{\eta})$ depending on τ^2, but not individually on δ and σ^2.

For the corresponding parameter combinations, the relative efficiency in Tables 4.1 and 4.2 are quite similar to the relative efficiency in Table 2.1. Recall that, in Table 2.1, when all parameters are known, we computed the relative efficiency of the BLUP obtained under the true model and the BLUP obtained under the Fay-Herriot model. Also, for the calculation in Table 2.1, we took $\psi_i = 0.5$ for all $i \in S$. However, the relative efficiency in Table 2.1 is slightly larger than the corresponding relative efficiency in Tables 4.1 and 4.2. This is probably explained by the difference in the number of parameters that need to be estimated to obtain the EBLUP under the true model as opposed to the EBLUP under the Fay-Herriot model. That is, for

the EBLUP obtained under the true model two additional parameters (δ, λ) have to be estimated.

Table 4.3: Summary of LSE [3] and MLE for model with sampling errors, $k = 30$, $n = 20$, $N = 40$, $m = 600$, $M = 1200$.

		LSE				MLE			
par.	tr. val.[4]	mean [5]	med. [6]	e.s.e. [7]	s.e. [8]	mean	med.	e.s.e.	s.e.
β	1.00	1.006	0.999	0.093	0.094	0.998	0.999	0.091	0.093
δ	0.30	0.320	0.316	0.110	0.112	0.304	0.287	0.092	0.089
λ	0.12	0.133	0.116	0.102	0.102	0.145	0.123	0.099	0.077
σ^2	0.70	0.672	0.675	0.099	-[9]	0.685	0.678	0.083	0.077
τ^2	1.00	0.992	0.984	0.093	0.092	0.989	0.975	0.092	0.091

[3] LSE refers to $(\widehat{\boldsymbol{\beta}}_{\mathrm{FH}}, \hat{\tau}^2_{\mathrm{FH}})'$ and $(\hat{\delta}, \hat{\lambda})'$ given by (3.2) and (3.5).

[4] true value of the parameter.

[5] mean parameter estimate over all simulation runs.

[6] median parameter estimate over all simulation runs.

[7] empirical root mean squared error of parameter estimate, given by the square root of (4.1).

[8] theoretical standard error evaluated at true parameter value.

[9] σ^2 can be consistently estimated by $\hat{\tau}^2_{\mathrm{FH}} - \hat{\delta}$. However, we do not have a formula for its variance.

4.3 Correlated clusters

For this section, the point pattern was generated as described in Section 4.1, with the cluster radius $r = 1.25$ and the cluster centers $\{(i, j) : i, j = 1, 3, 5, 7, 9\}$. The data were generated using the covariance model (2.5)-(2.6), where $k = 25$, $N = 20$, $M = 500$. Note that the choice of r allows cluster boundaries of neighboring clusters to intersect and small areas from two neighboring clusters to have non-negligible correlation. $\lambda = 0.3$ was chosen so that the median within cluster value of the A_{ij}'s was 0.40. In this case, the median A_{ij} value for any two neighboring clusters is approximately 0.05.

In order to compare our estimation method of (δ, λ) with an estimation method that takes advantage of the between cluster correlation, we also consider estimating (δ, λ) as follows:

$$(\tilde{\delta}, \tilde{\lambda}) = \underset{\delta \geq 0,\, \lambda \geq 0}{\operatorname{argmax}} \tilde{h}(\delta, \lambda; \mathbf{y}) \tag{4.4}$$

where

$$\tilde{h}(\delta, \lambda; \mathbf{y}) = -\sum_{\substack{i,j=1 \\ i \neq j}}^{m} \left(\hat{\epsilon}_i \hat{\epsilon}_j - \delta \exp(-\lambda M^p \|\mathbf{z}_i - \mathbf{z}_j\|) \right)^2 \tag{4.5}$$

where $\hat{\epsilon}_i = y_i - \mathbf{x}_i' \widehat{\boldsymbol{\beta}}_{\text{FH}}$. As can be seen from Tables 4.4 and 4.5, there is little to choose between the two estimation methods. Moreover, we note that for the LSE the s.e.'s match the e.s.e.'s.

NOTE (1): In addition to the above mentioned estimator $(\tilde{\delta}, \tilde{\lambda})$, it is possible to derive an approximately unbiased estimator for (δ, λ) as follows.

Table 4.4: Summary of LSE and LSE-C [the estimator given by (4.4)-(4.5)] when clusters are correlated and no sampling errors, $k = 25$, $N = 20$, $M = 500$.

		LSE				LSE-C		
par.	tr. val.	mean	med.	e.s.e.	s.e.	mean	med.	e.s.e.
δ	0.3	0.305	0.294	0.105	0.112	0.311	0.305	0.107
λ	0.3	0.319	0.308	0.178	0.175	0.339	0.319	0.180

Table 4.5: Summary of LSE and LSE-C when clusters are correlated and no sampling errors, $k = 25$, $N = 20$, $M = 500$.

		LSE				LSE-C		
par.	tr. val.	mean	med.	e.s.e.	s.e.	mean	med.	e.s.e.
δ	0.6	0.618	0.603	0.175	0.171	0.619	0.613	0.160
λ	0.3	0.327	0.318	0.129	0.111	0.336	0.338	0.138

Let $D = \text{diag}(\tau^2 + \psi_1, \ldots, \tau^2 + \psi_m)$ and $\widetilde{\mathbf{y}} = [I_m - (X'D^{-1}X)^{-1}X'D^{-1}]\mathbf{y}$; that is, $\widetilde{\mathbf{y}}$ is the vector of residuals after estimating $\boldsymbol{\beta}$ by $\widehat{\boldsymbol{\beta}}_{\text{FH}}$ when τ^2 is known. The approximately bias corrected estimator is derived as:

$$\begin{aligned}
\text{E}(\widetilde{\mathbf{y}}\widetilde{\mathbf{y}}') &= \text{E}[(I_m - (X'D^{-1}X)^{-1}X'D^{-1})\mathbf{y}\mathbf{y}'(I_m - D^{-1}X(X'D^{-1}X)^{-1}X')] \\
&= (I_m - (X'D^{-1}X)^{-1}X'D^{-1})(\sigma^2 I_m + \delta A + \Psi)(I_m - D^{-1}X(X'D^{-1}X)^{-1}X') \\
&= (I_m - (X'D^{-1}X)^{-1}X'D^{-1})(D + \delta(A - I_m))(I_m - D^{-1}X(X'D^{-1}X)^{-1}X') \\
&= \delta(I_m - X(X'D^{-1}X)^{-1}X'D^{-1})(A - I_m)(I_m - D^{-1}X(X'D^{-1}X)^{-1}X') \\
&\qquad + (D - X(X'D^{-1}X)^{-1}X') \\
&= \delta Q^{(1)}(\lambda, \tau^2) + Q^{(2)}(\tau^2)
\end{aligned}$$

where

$$\begin{aligned}
Q^{(1)}(\lambda, \tau^2) &= (I_m - X(X'D^{-1}X)^{-1}X'D^{-1})(A - I_m)(I_m - D^{-1}X(X'D^{-1}X)^{-1}X') \\
Q^{(2)}(\tau^2) &= (D - X(X'D^{-1}X)^{-1}X')
\end{aligned}$$

An approximately unbiased estimator $(\breve{\delta}, \breve{\lambda})$ for (δ, λ) is given by

$$(\breve{\delta}, \breve{\lambda}) = \underset{\delta \geq 0,\, \lambda \geq 0}{\text{argmax}}\, h^\star(\delta, \lambda; \mathbf{y}) \tag{4.6}$$

where

$$h^\star(\delta, \lambda; \mathbf{y}) = -\sum_{l=1}^{k} \sum_{\substack{i,j \in C_l \\ i \neq j}} \left(\hat{\epsilon}_i \hat{\epsilon}_j - \delta Q_{ij}^{(1)}(\lambda, \hat{\tau}^2_{\text{FH}}) - Q_{ij}^{(2)}(\hat{\tau}^2_{\text{FH}}) \right)^2$$

where $\hat{\epsilon}_i = y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}}_{\text{FH}}$, $\widehat{\boldsymbol{\beta}}_{\text{FH}}$, $\hat{\tau}^2_{\text{FH}}$ are given in (3.2) and $Q_{ij}^{(1)}(\lambda, \hat{\tau}^2_{\text{FH}})$, $Q_{ij}^{(2)}(\hat{\tau}^2_{\text{FH}})$ are the $(i,j)^{th}$ element of $Q^{(1)}$, $Q^{(2)}$ evaluated at $\hat{\tau}^2_{\text{FH}}$. Note that since $h^\star(\delta, \lambda; \mathbf{y})$ does not include diagonal elements of $Q^{(1)}(\lambda, \hat{\tau}^2_{\text{FH}})$ and $Q^{(2)}(\hat{\tau}^2_{\text{FH}})$, the bias correction terms

are of lower order than the top order terms. Hence, $(\breve{\delta}, \breve{\lambda})$ obtained by maximizing $h^{\star}(\delta, \lambda; \mathbf{y})$ is asymptotically equivalent to $(\hat{\delta}, \hat{\lambda})$.

As we did not perform many simulation runs, we do not report summary statistics for the estimator given by (4.6). However, for the parameter combinations we considered, there was negligible difference between the estimators given by (4.6) and the least squares estimators we considered in Chapter 3.

NOTE (2): We can express the variance-covariance matrix of $\mathbf{y}$ as

$$\text{var}(\mathbf{y}) = V = \sigma^2 I_m + \delta A + \Psi = \tau^2 I_m + \delta(A - I_m) + \Psi \tag{4.7}$$

Using (4.7) we can estimate (δ, λ) as follows: Use the log likelihood function but substitute estimators $(\widehat{\boldsymbol{\beta}}_{\text{FH}}, \hat{\tau}^2_{\text{FH}})$ for $(\boldsymbol{\beta}, \tau^2)$ and maximize over (δ, λ). That is we define the estimator $(\hat{\delta}_{\text{H}}, \hat{\lambda}_{\text{H}})$ of (δ, λ) as

$$(\hat{\delta}_{\text{H}}, \hat{\lambda}_{\text{H}}) = \underset{\delta \geq 0,\, \lambda \geq 0}{\text{argmax}}\, L(\delta, \lambda; \mathbf{y}) \tag{4.8}$$

where

$$L(\delta, \lambda; \mathbf{y}) = -\frac{1}{2} \log |V(\delta, \lambda, \hat{\tau}^2_{\text{FH}})| - \frac{1}{2} (\mathbf{y} - X\widehat{\boldsymbol{\beta}}_{\text{FH}})' \Big[V(\delta, \lambda, \hat{\tau}^2_{\text{FH}}) \Big]^{-1} (\mathbf{y} - X\widehat{\boldsymbol{\beta}}_{\text{FH}}) \tag{4.9}$$

where

$$V(\delta, \lambda, \hat{\tau}^2_{\text{FH}}) = \hat{\tau}^2_{\text{FH}} I_m + \delta(A - I_m) + \Psi.$$

In a limited simulation study (not reported here), of the estimator given by (4.8)-(4.9), we found that it was more efficient than the least squares estimator for (δ, λ) given by (3.5). Moreover, the estimator given by (4.8)-(4.9) has a running time approximately one third the running time of the MLE when $k = 40$, $N = 20$.

The shorter running time is due to the reduced number of matrix multiplications that need to be computed for each iteration when (δ, λ) is estimated by (4.8)-(4.9) as opposed to the MLE. We defer showing large sample properties of the estimator given by (4.8)-(4.9) to the future.

4.4 Comparison of LSE and MLE

In this section, we compare the LSE and the MLE for various parameter combinations. Since it is not possible to be exhaustive and analyze all parameter combinations, we fix $\lambda = 0.54$, which corresponds to the median within cluster value of A_{ij} being approximately 0.32. The simulation setup is exactly as described in Section 4.1. We first summarize our results and then give some details for specific simulation runs.

1. The relative efficiency of LSE and MLE depends on δ. Large values of δ correspond to small values of relative efficiency of LSE and MLE (that is, for large δ, MLE is much more efficient).

2. When δ and k are small, estimating λ is problematic. Depending on δ and k, as many as 5% (and this could be much larger if k or δ is taken to be smaller than values considered in this simulation study) of the simulation runs result in the estimate for λ either being 0 or being extremely large. Here, by extremely large, we mean the estimate is greater than 4 s.e.'s from the true value. However, the frequency of such cases decreases as k increases. For all parameters except λ, in most cases, the s.e.'s of the LSE and the MLE match

the e.s.e.'s. Moreover, we do not have boundary problems with $(\beta_1, \beta_2, \tau^2)$. However, in very few cases the estimate of δ is 0. We note that whenever an estimate was at the boundary or very large (in the case of λ), in nearly all cases, randomly restarting the maximization routine did not help. If this occurred when estimating by LSE, in certain cases, the contour plots were checked and they indicated that the maximization routine had indeed converged correctly.

3. In certain cases, when estimating λ, the $\mathrm{RE}(\hat{\lambda}, \hat{\lambda}_{\mathrm{ML}}) > 1$. This is probably due to k being too small. In such cases, whenever we increase k, the $\mathrm{RE}(\hat{\lambda}, \hat{\lambda}_{\mathrm{ML}})$ decreases to a number smaller than 1. Moreover, for small k, the histogram of λ_{ML} has a fatter tail on the right when compared to the histogram of $\hat{\lambda}$. Also, we note that since we ran only 100 simulation runs for the MLE, the MLE of λ is unduly influenced by large estimated values for λ.

4. Large values of δ and N imply smaller values of k are needed so that the s.e. of the LSE and the e.s.e. of the LSE are close to one another. The same is true for the MLE.

5. Based on previous results mentioned in Sections 2.1 and 3.3, we conjectured that the MLE for σ^2 is $\sqrt{m}$-consistent. However, the e.s.e. of the MLE for σ^2 is approximately the same as the e.s.e. of the MLE for δ. This seems to indicate that the MLE for σ^2 is not $\sqrt{m}$-consistent but possibly only $\sqrt{k}$-consistent.

From Table 4.6 for the LSE and the MLE, we see that, except when estimating λ by LSE, for all the other cases the e.s.e.'s are close to the s.e.'s. Moreover, two

Table 4.6: Summary of LSE and MLE for model without sampling errors, $k = 20$, $N = 20$, $M = 400$.

		LSE				MLE			
par.	tr. val.	mean	med.	e.s.e.	s.e.	mean	med.	e.s.e.	s.e.
β_1	1.00	1.012	1.007	0.141	0.141	1.008	1.002	0.145	0.125
β_2	1.00	0.940	0.931	0.925	0.893	0.952	0.915	0.852	0.746
δ	0.30	0.314	0.301	0.131	0.135	0.301	0.309	0.108	0.103
λ	0.54	0.570	0.526	0.327	0.303	0.579	0.575	0.267	0.260
σ^2	0.70	0.673	0.681	0.121	-	0.697	0.690	0.100	0.095
τ^2	1.00	0.987	0.983	0.083	0.080	0.998	0.977	0.088	0.079

of the estimated values of $\hat{\lambda}$ are large, and without these two estimates, the e.s.e. of LSE reduces to 0.304, which is nearly identical to the s.e. Moreover, we note that $\mathrm{RE}(\hat{\beta}_{1,\mathrm{FH}}, \hat{\beta}_{1,\mathrm{ML}})$ is marginally greater than 1. However, when k is increased to 40 (Table 4.7), we have $\mathrm{RE}(\hat{\beta}_{1,\mathrm{FH}}, \hat{\beta}_{1,\mathrm{ML}})$ is much smaller than 1. Regarding estimation of τ^2, we note that for every parameter combination we took, the $\mathrm{RE}(\hat{\tau}^2, \hat{\tau}^2_{\mathrm{ML}}) \approx 1$. When k is increased to 40 (Table 4.7) with $(\delta, \lambda, \sigma^2)$ being the same, we calculate the factor by which the e.s.e.'s decrease. For β_1, β_2, δ, λ, σ^2, τ^2, the factors by which the e.s.e.'s decrease are 1.37, 1.28, 1.41, 1.54, 1.44, 1.48. Theory would suggest that they should reduce by approximately $\sqrt{2}$. However, we point out the need for a larger simulation study to be definitive about this claim.

Table 4.7: Summary of LSE and MLE for model without sampling errors, $k = 40$, $N = 20$, $M = 800$.

		LSE				MLE			
par.	tr. val.	mean	med.	e.s.e.	s.e.	mean	med.	e.s.e.	s.e.
β_1	1.00	1.003	1.002	0.103	0.103	0.996	0.987	0.077	0.092
β_2	1.00	0.935	0.944	0.722	0.729	0.983	1.009	0.522	0.620
δ	0.30	0.305	0.302	0.093	0.093	0.312	0.303	0.075	0.074
λ	0.54	0.539	0.534	0.212	0.212	0.589	0.555	0.223	0.187
σ^2	0.70	0.688	0.690	0.084	-	0.688	0.692	0.069	0.068
τ^2	1.00	0.993	0.989	0.056	0.056	1.000	1.000	0.056	0.056

Note that in Table 4.7, the RE($\hat{\lambda}$, $\hat{\lambda}_{\text{ML}}$) > 1. When estimating by maximum likelihood, there were two extremely large estimated values of λ which unduly influenced the e.s.e. After eliminating the large estimated values for both LSE and MLE of λ, the RE($\hat{\lambda}$, $\hat{\lambda}_{\text{ML}}$) decreases to 0.96.

When instead of increasing k, we increased $N = 40$ (Table 4.8) with $(\delta, \lambda, \sigma^2)$ being the same, we noticed that the e.s.e.'s decrease quite significantly, for example, when compared to Table 4.6, the e.s.e. of the LSE of δ, λ, σ^2, τ^2 reduce by a factor of 1.30, 1.28, 1.39, 1.32. While this may seem puzzling, as increasing N should result in an inconsistent estimator, we probably have to increase N significantly to be able to detect this. This is beyond the computing resources used for this thesis. We point out that for both methods of estimation, the s.e. for the estimator of λ is

Table 4.8: Summary of LSE and MLE for model without sampling errors, $k = 20$, $N = 40$, $M = 800$.

		LSE				MLE			
par.	tr. val.	mean	med.	e.s.e.	s.e.	mean	med.	e.s.e.	s.e.
β_1	1.00	0.990	0.983	0.122	0.124	0.986	0.990	0.102	0.102
β_2	1.00	1.099	1.134	0.811	0.829	1.035	1.038	0.625	0.614
δ	0.30	0.315	0.307	0.101	0.092	0.301	0.301	0.065	0.068
λ	0.54	0.569	0.541	0.256	0.219	0.589	0.555	0.201	0.179
σ^2	0.70	0.676	0.683	0.087	-	0.690	0.692	0.055	0.058
τ^2	1.00	0.991	0.985	0.063	0.062	0.991	0.984	0.059	0.061

much smaller than the e.s.e. But once again we point out that after eliminating the large estimated values for λ, the e.s.e.'s are almost identical to the s.e.'s

Comparing Tables 4.7 and 4.8, it is striking that when $N = 40, k = 20$, the e.s.e. of $\hat{\delta}_{\text{ML}}$, $\hat{\lambda}_{\text{ML}}$ and $\hat{\sigma}^2_{\text{ML}}$ is smaller than when $N = 20, k = 40$. We mention that for large k, N such a phenomenon should not occur. However, we have no way of verifying this claim using the computing resources available for this thesis. As expected, the LSE and MLE for $(\beta_1, \beta_2, \tau^2)$ have larger e.s.e. when $N = 40$, $k = 20$ compared to $N = 20$, $k = 40$. This should be the case as larger N implies more correlated observations which would make estimating the fixed effect parameter and the variance parameter more difficult.

Table 4.9: Summary of LSE and MLE for model without sampling errors, $k = 20$, $N = 20$, $M = 400$.

		LSE				MLE			
par.	tr. val.	mean	med.	e.s.e.	s.e.	mean	med.	e.s.e.	s.e.
β_1	1.00	0.998	1.010	0.184	0.173	0.994	1.005	0.133	0.134
β_2	1.00	1.046	0.970	1.148	1.050	1.015	0.998	0.730	0.696
δ	0.60	0.619	0.598	0.192	0.181	0.600	0.594	0.120	0.116
λ	0.54	0.538	0.519	0.207	0.214	0.599	0.565	0.184	0.159
σ^2	0.40	0.365	0.377	0.146	-	0.392	0.386	0.082	0.074
τ^2	1.00	0.984	0.972	0.106	0.103	0.992	0.991	0.097	0.096

In Tables 4.9 and 4.10, δ is increased to 0.6 with $k = 20, N = 20$ (Table 4.9) and $k = 40$, $N = 20$ (Table 4.10). We see that except when λ is estimated by MLE for $k = 20$, $N = 20$, for all other cases, the e.s.e. matches the s.e. Moreover, unlike in Table 4.7, when $\delta = 0.6$, $k = 40$ is large enough for both methods of estimation to not have estimated values of λ to be 0 or extremely large. Moreover, a comparison of the relative efficiency of LSE and MLE in Table 4.10 for parameters β_1, β_2, δ, λ, σ^2, τ^2, gives 0.49, 0.42, 0.38, 0.63, 0.25, 1.03, which we compare to the similar numbers in Table 4.7. which are 0.56, 0.52, 0.65, 1.11, 0.67, 1.00. It appears likely that the relative efficiency depends on δ for fixed λ, k, N. Moreover, we point out that in Table 4.9 for all parameters, the e.s.e. of LSE matches the s.e. which is not the case in Table 4.6.

Table 4.10: Summary of LSE and MLE for model without sampling errors, $k = 40$, $N = 20$, $M = 800$.

		LSE				MLE			
par.	tr. val.	mean	med.	e.s.e.	s.e.	mean	med.	e.s.e.	s.e.
β_1	1.00	0.994	0.997	0.123	0.124	0.989	0.995	0.086	0.097
β_2	1.00	1.052	0.961	0.913	0.866	1.093	1.153	0.594	0.580
δ	0.60	0.614	0.613	0.136	0.137	0.609	0.611	0.084	0.082
λ	0.54	0.541	0.525	0.148	0.150	0.559	0.549	0.117	0.112
σ^2	0.40	0.380	0.393	0.107	-	0.392	0.386	0.054	0.053
τ^2	1.00	0.994	0.989	0.071	0.071	1.001	0.996	0.072	0.068

Next, we consider $\delta = 0.2$, $k = 40$, $N = 15$ (Table 4.11). As mentioned previously, small δ results in difficulty in estimating λ. The LSE and the MLE for λ both show large bias, and the RE($\hat{\lambda}$, $\hat{\lambda}_{\text{ML}}$) > 1. Increasing to $k = 80$ (Table 4.12) with $(\delta, \lambda, \sigma^2)$ and N the same, still results in the e.s.e. not matching the s.e. for λ. However, from Table 4.12 we have RE($\hat{\lambda}$, $\hat{\lambda}_{\text{ML}}$) < 1. Using Table 4.12, we compute the relative efficiency of the LSE and the MLE. For the parameters β_1, β_2, δ, λ, σ^2, τ^2, the relative efficiency of the LSE and the MLE is 0.90, 0.87, 0.79, 0.82, 0.70, 0.77 which are much larger compared to the similar calculation we did for the relative efficiency of LSE and MLE using Tables 4.7 and 4.10. In an effort to attain the s.e. for the estimator for λ, we increase to $k = 160$. However, for this case it is not possible to run a simulation for the MLE using available computer resources.

Hence, we only consider the LSE. As Table 4.13 indicates, we have finally managed to match the s.e. for $\hat{\lambda}$ with the e.s.e. for $\hat{\lambda}$.

Table 4.11: Summary of LSE and MLE for model without sampling errors, $k = 40$, $N = 15$, $M = 600$.

		LSE				MLE			
par.	tr. val.	mean	med.	e.s.e.	s.e.	mean	med.	e.s.e.	s.e.
β_1	1.00	1.003	0.996	0.101	0.101	0.999	0.999	0.102	0.096
β_2	1.00	0.991	0.999	0.702	0.706	0.943	0.976	0.677	0.662
δ	0.20	0.220	0.204	0.112	0.099	0.201	0.198	0.090	0.087
λ	0.54	0.600	0.528	0.420	0.335	0.693	0.574	0.495	0.313
σ^2	0.80	0.777	0.788	0.113	-	0.784	0.779	0.101	0.090
τ^2	1.00	0.997	0.997	0.060	0.060	0.985	0.985	0.061	0.060

4.5 Change in dimension

Here, we consider $\mathbf{z}_i$ in a higher dimensional space - we take $\mathbf{z}_i \in [0, 10]^4$. Except for this, the simulation setup is exactly as described in Section 4.1. We consider the same set of parameter values for $(\delta, \lambda, \sigma^2)$ and k, N, M as Table 4.7. We notice that increasing the dimension of $\mathbf{z}_i$ (see Tables 4.14 and 4.7) reduces the e.s.e. of $(\beta_1, \beta_2, \tau^2)$. This is because in a higher dimensional space for $\mathbf{z}_i$, the observations within a cluster are not as highly correlated, which in turn results in better estimation of $(\beta_1, \beta_2, \tau^2)$. However, as can be seen from these two tables, the

Table 4.12: Summary of LSE and MLE for model without sampling errors, $k = 80$, $N = 15$, $M = 1200$.

		LSE				MLE			
par.	tr. val.	mean	med.	e.s.e.	s.e.	mean	med.	e.s.e.	s.e.
β_1	1.00	0.991	0.991	0.077	0.075	0.981	0.983	0.073	0.071
β_2	1.00	1.059	1.058	0.567	0.559	1.129	1.149	0.496	0.522
δ	0.20	0.208	0.204	0.073	0.069	0.203	0.194	0.065	0.061
λ	0.54	0.551	0.529	0.261	0.239	0.555	0.525	0.237	0.224
σ^2	0.80	0.789	0.793	0.074	-	0.795	0.796	0.062	0.063
τ^2	1.00	0.997	0.995	0.041	0.043	0.998	0.998	0.036	0.043

Table 4.13: Summary of LSE for model without sampling, $k = 160$, $N = 15$, $M = 2400$.

par.	tr. val.	mean	med.	e.s.e.	s.e.
β_1	1.00	1.007	1.008	0.057	0.056
β_2	1.00	0.933	0.920	0.432	0.424
δ	0.20	0.202	0.200	0.047	0.048
λ	0.54	0.534	0.514	0.170	0.168
σ^2	0.80	0.793	0.795	0.050	-
τ^2	1.00	0.995	0.994	0.030	0.030

e.s.e. of $\hat{\delta}$ and $\hat{\lambda}$ both increase, though they match the s.e. of LSE.

Table 4.14: Summary of LSE for model without sampling, $\mathbf{z}_i \in \mathbb{R}^4$, $k = 40$, $N = 20$, $M = 800$.

par.	tr. val.	mean	med.	e.s.e.	s.e.
β_1	1.00	0.997	0.998	0.094	0.092
β_2	1.00	0.992	0.963	0.686	0.677
δ	0.30	0.335	0.307	0.182	0.173
λ	0.54	0.539	0.525	0.261	0.258
σ^2	0.70	0.661	0.688	0.180	-
τ^2	1.00	0.996	0.996	0.054	0.052

4.6 Concluding remarks on simulation study

We conclude this chapter by remarking that while our simulation study is by no means exhaustive, it is supportive of our theoretical results. Moreover for large k, the simulation study provides evidence that the e.s.e. of the MLE matches the s.e. of the MLE. However, we need to run a much larger simulation study, which we will do in the future. We also suggest a few areas where it would be of interest to carry out further simulation studies:

1. Compare the EBLUP under the true model and the EBLUP under the misspecified Fay-Herriot model for more exhaustive combinations of $(n, N, m, M, \delta, \lambda, \sigma^2, \psi_i)$. Include covariates in addition to the intercept term. We believe by

doing so larger relative efficiency between the EBLUP obtained under the true model and the EBLUP obtained under the Fay-Herriot model can be achieved.

2. Consider different point patterns that generate the $\mathbf{z}_i$'s in each cluster.

3. Consider decreasing the cluster radius and check whether estimation is adversely affected, especially estimation of λ.

4. Increase N and see if this results in inconsistent estimators for LSE. As we mentioned in Sections 2.1 and 3.3, for certain spatial models under infill asymptotics $\hat{\sigma}^2_{\text{ML}}$ is $\sqrt{m}$-consistent. However, these results are for very special point patterns. Based on our simulation results when estimating by MLE, we have no evidence of different rates of convergence. We need to conduct a more exhaustive simulation study to check this claim.

5. Consider certain types of mispecifications. For example, when clusters are not well separated, check if the estimation method is adversely affected if a fraction of points are assigned to the incorrect cluster.

Chapter 5

Data analysis

We analyze a U.S. county level data set that was previously analyzed by Wheeler [51], [52] for a different purpose. The data set consists of civilian employment growth rates for all U.S. counties between 1980 and 1990 and includes 14 county level covariates. We are interested in clustering the counties, and by doing so we hope to obtain better predictors of the observed and unobserved counties when compared to the predictors obtained without clustering. We assume the clusters are Census Bureau regions (there are 9 such regions) and assume that the variance-covariance matrix of the county level random effects (small area effects) is given by (2.5), (2.10).

The set of covariates included in the data set were (the year is given in parenthesis): log employment (1980), log population (1980), employment density (1980), population density (1980), log land area (1980), fraction of adult population with bachelor's degree (1980), fraction of employment in manufacturing (1980), unemployment rate (1980), per capita income (1979), urban/rural indicator (1990), share of local government spending on education (1982), share of local government spend-

ing on police (1982), share of local government spending on highways (1982) and fraction of population that is not white (1980).

Among the 3106 U.S. counties, we deleted 4 counties with missing covariates. Unfortunately the deleted counties were all large counties. The deleted counties were Bronx, New York, Queens and Richmond with employment growth rates of 0.0926, 0.0972, 0.0992 and 0.1976. Among the deleted counties, the first 3 counties have approximately the median employment growth rate among all U.S. counties while the last county has a growth rate in the 75^{th} percentile. The missing covariates for the counties were local government spending on education, police and highways, and we were not able to impute these missing covariates.

In order to choose the best set of covariates, we considered the stepwise AIC criterion. Following the discussion in Wheeler [51] where he mentions a non-linear inverted 'U' relationship between the employment growth rate and log employment, we plotted the data to check for such a relationship. Having seen one, we fitted a model with a second degree polynomial in log employment in addition to including other covariates. The other covariates were urban/rural indicator, fraction of adult population with bachelor's degree, fraction of employment in manufacturing, log population, log land area, share of local government spending on police, share of local government spending on highways and fraction of population that is not white. Having chosen our main effects by stepwise AIC criterion, next we considered interaction terms. We considered interaction among all main effects, and the final model selected by the AIC criterion was: log employment, urban/rural indicator, fraction of adult population with bachelor's degree, fraction of employment in

manufacturing, log population, share of local government spending on police, share of local government spending on highways, fraction of population that is not white, and interaction between the following pairs of covariates: urban/rural indicator and fraction of employment in manufacturing, urban/rural indicator and log population, share of local government spending on police and fraction of employment in manufacturing, share of local government spending on police and fraction of adult population with bachelor's degree. We note that when we considered interaction terms, the quadratic term in log employment was not significant at the 0.05 level. In the final model, among all coefficients for the fixed effects, the largest p-value was 2.5×10^{-8}.

A histogram of the employment growth rates shows a slightly fatter tail to the right but is otherwise symmetric and unimodal. Having fitted the model with the above mentioned covariates, we checked a plot of the residuals against covariates and against the fitted values, searching for any pattern. In particular, we checked a plot of the residuals against log employment. We did not detect any pattern in any of the plots. Moreover, for the residuals we did a normal Q-Q plot to check our normality assumption. In the middle and the left tail, the plot looked fine, however beyond the second standard deviation the plot deviated significantly from the normal quantiles. This is associated with the slightly fatter tail to the right of the histogram of the observations.

We draw a simple random sample of 800 from the 3102 U.S. counties and pretend that only the sampled counties were observed. We refer to the remaining 2302 counties as the unobserved counties. Moreover, since the sampling errors ψ_i

were not included in the data set, we added noise to the data in order to mimic a typical setting in which the Fay-Herriot model is used. We pretend that the employment growth rates given in the data set are the true employment growth rates. We also analyze the data set without sampling errors. In this case, we do cross-validation to compute the relative efficiency of the two different predictors. To make the discussion easier, regardless of whether we add or do not add noise to the employment growth rates, we refer to the competing model as the Fay-Herriot model. Moreover, the above mentioned covariates were selected using the entire data set so as to not have the set of covariates be dependent on the observed sample. Also, in order not to have a specific sample unduly influence the outcome, we reanalyzed the data set by randomly selecting another 2 samples.

In order to choose the sampling errors we did the following. Usually the ψ_i's are chosen so that for all i, $\psi_i \propto \frac{1}{n_i}$, where n_i is the sample size of the direct survey estimate of the i^{th} small area. Moreover, in Fay-Herriot applications, the n_i's are taken to be roughly proportional to population size for the i^{th} small area. We chose the constant of proportionality by taking the smallest county with the number of civilians employed of at least $200,000$, and for such a county we made sure its direct estimate would not differ from its EBLUP under the Fay-Herriot model by more than 1%. Once we chose the constant of proportionality (which was 51.03), we kept it fixed for all 3 samples.

We chose the number $200,000$ by using the same practice employed by the Current Population Survey (CPS) where for a sampled county the sampling fraction is approximately 1 in 2000 . Hence we do not want a county with a sample of at

Table 5.1: Estimates of $(\delta, \lambda, \tau^2)$ for 3 random samples from employment growth rate data set.

	Sample 1		Sample 2		Sample 3	
Parameter	spatial	survey	spatial	survey	spatial	survey
δ	0.00958	0.01214	0.00834	0.00930	0.00768	0.00921
λ	0.00135	0.00150	0.00200	0.00212	0.00166	0.00189
τ^2	0.02104	0.02316	0.01822	0.01879	0.01744	0.01766

Table 5.2: Relative efficiency of EBLUP under true and Fay-Herriot for 3 randomly selected samples using employment growth rate data.

	Sample 1	Sample 2	Sample 3
sampled area	1.074	1.042	1.085
non-sampled area (survey)	1.395	1.341	1.368
non-sampled area (spatial)	1.483	1.420	1.447

least 100 to differ from its EBLUP under the Fay-Herriot model by more than 1%.

In Table 5.1, we give the set of parameter estimates $(\delta, \lambda, \tau^2)$ for each of the 3 samples we generated. Moreover, we considered estimating the parameters with and without adding sampling error which in the Table 5.1 is referred to as 'survey' and 'spatial'. As can be seen from Table 5.1, there seems to be some variation by sample in the estimates of the covariance parameters. Moreover, for any specific sample, there is a noticeable difference in the estimates of the covariance parameters when we add error compared to when we do not.

From Table 5.2 we note that the relative efficiency of the EBLUP under the true model and the Fay-Herriot is approximately the same for all 3 samples, even though as mentioned previously there is some variability in the estimates by sample. The row 'sampled area' refers to when we add sampling error to the county level employment growth rates and we compute the relative efficiency of the two predictors. We compute the relative efficiency by computing the ratio of the squared error averaged over all observed small areas for each of the 2 different types of predictors. Also, 'non-sampled area (survey)' refers to a similar computation as above when we add sampling error and are interested in the relative efficiency of the predictors for the non-sampled areas. Finally, 'non-sampled area (spatial)' refers to cross validation when we do not add sampling errors. As expected when we limit ourselves to non-sampled areas, the relative efficiency is larger when there is no sampling error as opposed to when we add sampling error. Also, as can be seen, the relative efficiency is quite small for the sampled areas. We compare these numbers to the limit of (2.28). For example, for sample 1, we substitute the estimated values of δ, σ^2 and the median value of ψ_i in the limit of (2.28) and we get 1.08. Assuming the estimated values for δ, σ^2 in sample 1 are close to the true values, we have achieved as large a relative efficiency as we can hope for.

In Figure 5.1, for one of the samples for the case when we did not add sampling error, we give a plot of the squared error of the EBLUP under our model (x co-ordinate) against the squared error of the EBLUP under the Fay-Herriot model (y co-ordinate) for each of the non-sampled areas. The plotted line is $y = x$. As can be seen, for most non-sampled small areas, the squared error of the EBLUP under the

Fay-Herriot model is larger than the squared error of the EBLUP under our model.

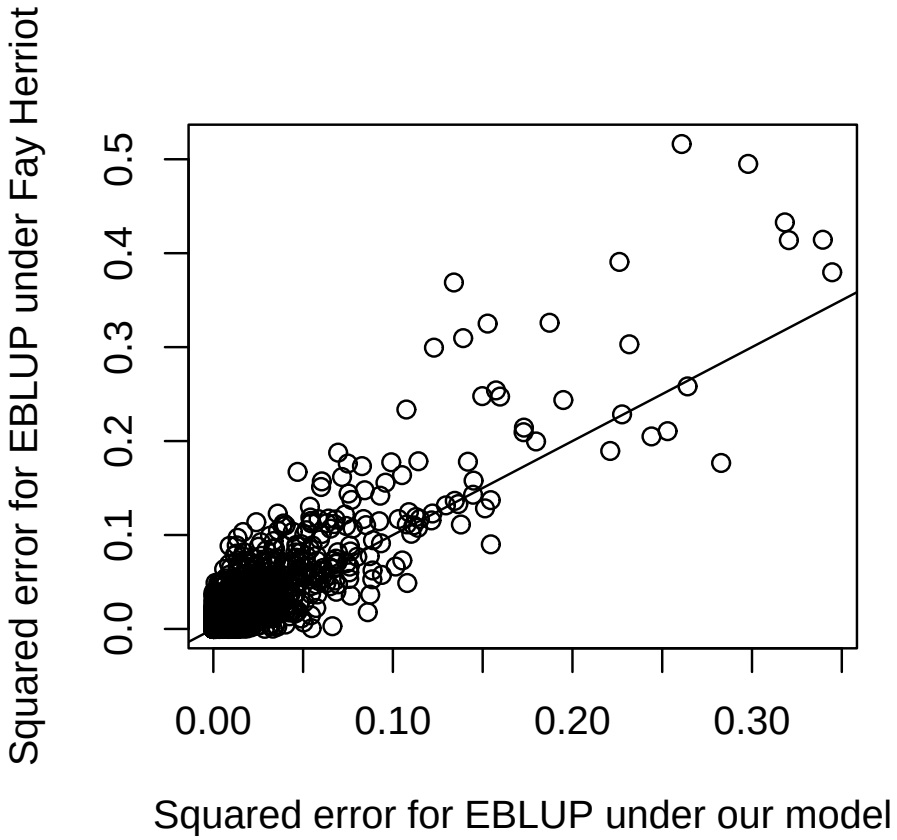

Figure 5.1: Plot of the squared error for EBLUP under our model vs. Fay Herriot model.

As a further validation of fitting our model we computed the deviance, that is 2 times the difference in the log likelihood under our model and the log likelihood under the Fay-Herriot model. The deviance when we added sampling error was 123.29 (in this case, the log likelihood under the Fay-Herriot model was 6463.59). The deviance when we did not add sampling error was 183.19 (in this case, the log likelihood under the Fay-Herriot model was 6311.36). We computed the deviance

for only one of the samples. We also note that usually deviance is computed when the parameters are estimated by MLE, however we point out that the large deviance even when we estimate the parameters by our method is indicative of a better fit of our model compared to the Fay-Herriot model.

We conclude this chapter by mentioning that in a future study of this data set we plan on clustering the counties in a more scientific manner by using spatial locations and various covariates. We did not do so in this thesis due to time constraints. However, as we have shown, even when we use Census Bureau regions as our clusters, we obtain improved prediction, especially for the non-sampled areas.

Chapter 6

Small area estimation problems

The second and third parts of this thesis consist of two non-spatial small area estimation problems. In Chapter 7, in a small area context, we consider the problem of simultaneous credible intervals. In the frequentist framework there are several multiple comparison procedures that have been used in linear models. For example, Scheffé's, Tukey's and Bonferroni's multiple comparison procedures (Hochberg & Tamhane [21], Miller [29] and Scheffé [43]). We adapt these procedures to the Bayesian framework in order to construct simultaneous credible intervals for small areas.

In the Bayesian framework the choice of prior for the hyperparameter(s) is important. Usually a flat prior is chosen for the regression coefficient and the prior variance is assumed to be independent of the regression coefficients and uniformly distributed over the positive part of the real line (Berger [4] Morris & Christiansen [30] and Morris [32]). However, there is significant of amount literature on choosing priors so that they have approximate frequentist validity. For example, probability matching priors refers to priors that give Bayesian credible sets that have approx-

imately correct frequentist coverage (Datta & Mukerjee [16] and Tibshirani [46]). The approximate agreement of the Bayesian and frequentist coverage probabilities of the associated credible sets gives these priors an external validation (Datta & Mukerjee [16]).

Datta et al. [17] refer to a prior as having dual justification if the posterior variance of the hierarchical Bayes estimator is second order unbiased for the mean squared error of the EBLUP (see 7.10). In Chapter 7, we follow a similar approach in deriving an objective prior for the prior variance.

In Chapter 8, for a special case of the nested error regression model we consider the problem of deriving a robust mean squared error (MSE) estimator for the empirical best linear unbiased predictor (EBLUP). The small area mean θ_i is predicted using the BLUP, but usually there will be unknown variance components that need to be estimated. The BLUP with estimated variance components is referred to as the EBLUP. The MSE of the BLUP cannot be used as an approximation of the MSE of the EBLUP as it does not take into account the variability of estimating the variance components.

For the Fay-Herriot model and the nested-error regression model, Prasad & Rao [36] used a moment estimator to estimate the variance components, and under normality derived a second order approximation of the MSE of the EBLUP, and a second order unbiased estimator of the MSE of the EBLUP. Here, second order approximation of the MSE means that the difference between the true MSE and the approximation of the true MSE is $o(1/m)$, where m is the number of small areas. Also, second order unbiased means that the difference between the expectation of

the estimator of the MSE and the true MSE is $o(1/m)$.

For the Fay-Herriot model, Lahiri & Rao [24] showed that the normality of the random effects could be relaxed by a certain moment condition on the random effects so that the Prasad-Rao MSE estimator (with the variance components estimated by a moment estimator) would still be second order unbiased. In a more general model, under normality, Datta & Lahiri [15] gave an approximation and an estimator of the MSE of the EBLUP when the variance components were estimated by either the MLE or REML.

Without assuming normality Das, Jiang & Rao [14] derived an approximation for the MSE of the EBLUP under a more general model that included the Fay-Herriot model and nested error regression model as special cases. However when deriving an estimator of the MSE of the EBLUP they assume normality.

For nested error regression model, Hall & Maiti [20] derived a non-parametric estimator of the MSE of the EBLUP. They do so by considering a double bootstrap method. In Chapter 8, unlike in Hall & Maiti, we derive a closed form expression for the estimator of the MSE of the EBLUP.

Chapter 7

Simultaneous credible intervals

A researcher in public health may report an estimate of the mean body mass index and the associated 95% individual confidence interval for each domain formed by different demographic groups (e.g., for different race×gender× age-group combinations), and then use these individual confidence intervals to find significant difference among pairs of domains. The problem with the above approach, often referred to as *data snooping*, is that even if a table of estimates of the domain mean differences and their associated 95% (individual) confidence intervals are reported for all possible pairs, the confidence level refers to a single comparison and not to a series of comparisons. In fact, the overall confidence level, that is, the probability that all confidence intervals cover their respective true values, could be much lower than the nominal 95% level. The problem of finding spurious significance results due to data snooping is referred to as the problem of *multiple comparison.*

Exploratory data analysis is a useful part of any scientific investigation, but any claim suggested by such analysis should be validated by an appropriate statistical procedure. Multiple comparison is the most common data snooping problem

encountered in small area research. The literature on multiple comparison for linear models is vast, for example, see Hochberg and Tamhane [21] and Miller [29].

Using the celebrated Fay-Herriot model, we demonstrate how the Bayesian method can be adapted to address the multiple comparison problem. The Bayesian method is conceptually straightforward. Once the posterior distribution of the parameter(s) of interest is found, this is used for all inferential purposes.

As mentioned in Chapter 1, the Fay-Herriot model is given by

- Level 1 (sampling model): $y_i|\theta_i \overset{\text{ind}}{\sim} N(\theta_i, \psi_i),\ i = 1, \cdots, m;$
- Level 2 (linking model): $\theta_i \overset{\text{ind}}{\sim} N(\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2),\ i = 1, \cdots, m.$

Suppose we are interested in finding a $100(1-\alpha)\%$ credible interval for a specific $\boldsymbol{\ell}'\boldsymbol{\theta}$, where $\boldsymbol{\ell}$ is a known $m \times 1$ column vector. We simply find the posterior distribution of $\boldsymbol{\ell}'\boldsymbol{\theta}$ and use this to find the desired credible interval. To illustrate the method, first assume σ^2 is known, but $\boldsymbol{\beta}$ unknown. We put a flat (improper) prior on $\boldsymbol{\beta}$, that is, $\pi(\boldsymbol{\beta}) \propto 1$. As we show in Section 6.5,

$$\boldsymbol{\theta}\,|\,\mathbf{y} \sim N(\Lambda\boldsymbol{\omega}, \Lambda) \tag{7.1}$$

where $\mathbf{y} = (y_1, \ldots, y_m)'$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)'$, $\boldsymbol{\omega} = (\frac{y_1}{\psi_1}, \ldots, \frac{y_m}{\psi_m})'$, $X = (\mathbf{x}_1, \ldots, \mathbf{x}_m)'$, and $\Lambda^{-1} = \text{diag}(\frac{1}{\psi_1} + \frac{1}{\sigma^2}, \ldots, \frac{1}{\psi_m} + \frac{1}{\sigma^2}) - \frac{X(X'X)^{-1}X'}{\sigma^2}$.

A $100(1-\alpha)\%$ credible interval for $\boldsymbol{\ell}'\boldsymbol{\theta}$ is given by

$$\boldsymbol{\ell}'\Lambda\boldsymbol{\omega} \pm \left(\boldsymbol{\ell}'\Lambda\boldsymbol{\ell}\chi^2_{(\alpha,1)}\right)^{\frac{1}{2}}, \tag{7.2}$$

where $\chi^2_{(\alpha,1)}$ is the upper α percentage point of the chi-squared distribution with one degree of freedom.

When σ^2 is unknown, we need to put priors on both $\boldsymbol{\beta}$ and σ^2. We assume that

$$\pi(\boldsymbol{\beta}, \sigma^2) = \pi(\boldsymbol{\beta})\pi(\sigma^2) \propto \pi(\sigma^2).$$

In this case, a closed-form density for

$$T^{(1)} = \frac{\Big(\boldsymbol{\ell}'(\boldsymbol{\theta} - E(\boldsymbol{\theta} \mid \mathbf{y}))\Big)^2}{\boldsymbol{\ell}' \mathrm{var}(\boldsymbol{\theta} \mid \mathbf{y})\boldsymbol{\ell}} \mid \mathbf{y}$$

cannot be obtained. Hence, a Monte Carlo method is used to construct a credible interval for $\boldsymbol{\ell}'\boldsymbol{\theta}$. The method is as follows: For large R, independently simulate $(\boldsymbol{\theta}_{(1)}, \boldsymbol{\beta}_{(1)}, \sigma^2_{(1)}), \ldots, (\boldsymbol{\theta}_{(R)}, \boldsymbol{\beta}_{(R)}, \sigma^2_{(R)}) \sim f(\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma^2 \,|\, \mathbf{y})$. Then $E(\boldsymbol{\theta} \,|\, \mathbf{y})$ and $\mathrm{var}(\boldsymbol{\theta} \,|\, \mathbf{y})$ are approximated by

$$\begin{aligned} E(\boldsymbol{\theta} \mid \mathbf{y}) &= \overline{\boldsymbol{\theta}}_{(.)} = \frac{1}{R}\sum_{i=1}^{R} \boldsymbol{\theta}_{(i)}, \\ \mathrm{var}(\boldsymbol{\theta} \mid \mathbf{y}) &= \frac{1}{(R-1)}\sum_{i=1}^{R} (\boldsymbol{\theta}_{(i)} - \overline{\boldsymbol{\theta}}_{(.)})(\boldsymbol{\theta}_{(i)} - \overline{\boldsymbol{\theta}}_{(.)})'. \end{aligned}$$

Also, $T^{(1)}_{\alpha}$, the upper α percentage point of the distribution of $T^{(1)}$, is given by the upper α percentage point of the ordered values $T^{(1)}_{(i)}$ $(i = 1, \ldots, R)$, where

$$T^{(1)}_{(i)} = \frac{\Big(\boldsymbol{\ell}'(\boldsymbol{\theta}_{(i)} - E(\boldsymbol{\theta} \mid \mathbf{y}))\Big)^2}{\boldsymbol{\ell}' \mathrm{var}(\boldsymbol{\theta} \mid \mathbf{y})\boldsymbol{\ell}}.$$

When σ^2 is unknown, a $100(1-\alpha)\%$ credible interval for $\boldsymbol{\ell}'\boldsymbol{\theta}$ is given by

$$\boldsymbol{\ell}' E(\boldsymbol{\theta} \mid \mathbf{y}) \pm \Big(\boldsymbol{\ell}' \mathrm{var}(\boldsymbol{\theta} \mid \mathbf{y})\boldsymbol{\ell}\, T^{(1)}_{\alpha}\Big)^{\frac{1}{2}}. \tag{7.3}$$

One important step in the Bayesian approach is the choice of the prior distribution for the hyperparameter(s). Morris and Christiansen [30] used a flat (Lebesgue

measure) prior distribution for the regression coefficients, and assumed the prior variance to be independent of the regression coefficients and uniformly distributed over the positive part of the real line. These prior distributions for the hyperparameters are simple to interpret to a nonstatistician and are often recommended. See Berger [4] and Morris [32]. The uniform prior for the variance, often referred to as the Stein's *superharmonic* prior, is noninformative and is known to provide minimax procedures. Unless more information on the hyperparameters is available, these simple prior distributions for the hyperparameters give good frequentist properties to the resulting rules (Morris and Christiansen [30]).

7.1 Multiple Comparison

We are interested in constructing simultaneous $100(1-\alpha)\%$ credible intervals, say I_ℓ, for $\boldsymbol{\ell}'\boldsymbol{\theta}$ for all $\boldsymbol{\ell} \in L$, where $L \subset \mathbb{R}^m$, the $m-$dimensional Euclidean space. That is, we want

$$P[\boldsymbol{\ell}'\boldsymbol{\theta} \in I_\ell \text{ for all } \boldsymbol{\ell} \in L|\mathbf{y}] = 1-\alpha,$$

where the probability is with respect to the posterior distribution of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)'$ given $\mathbf{y} = (y_1, \ldots, y_m)'$.

If one were to use (7.2) [when σ^2 is known] or (7.3) [when σ^2 is unknown] for multiple comparison, then the overall coverage probability will be much lower than the nominal $100(1-\alpha)\%$. Hence the need for our method. In the following three subsections, we discuss multiple comparison procedures for three useful classes L.

7.1.1 Pairwise comparison

Here we are only interested in constructing simultaneous credible intervals for all pairwise comparisons. We will restrict attention to the case where σ^2 is unknown. A Bayesian version of Tukey's simultaneous confidence intervals can be used. Define

$$\max_k \Big((\theta_k - E(\theta_k \mid \mathbf{y})) \mid \mathbf{y} \Big) - \min_k \Big((\theta_k - E(\theta_k \mid \mathbf{y})) \mid \mathbf{y} \Big) \equiv T^{(2)}.$$

Note that $\forall i, j$,

$$\begin{aligned} & \Big| \Big((\theta_i - E(\theta_i \mid \mathbf{y})) \mid \mathbf{y} \Big) - \Big((\theta_j - E(\theta_j \mid \mathbf{y})) \mid \mathbf{y} \Big) \Big| \leq T^{(2)} \\ \Rightarrow \quad & P\Big(\forall i, j, \, |(\theta_i - E(\theta_i \mid \mathbf{y})) - (\theta_j - E(\theta_j \mid \mathbf{y}))| \leq T^{(2)}_\alpha \mid \mathbf{y} \Big) \geq 1 - \alpha, \end{aligned}$$

where $T^{(2)}_\alpha$ is the upper α percentage point of the distribution of $T^{(2)}$. Simultaneous $100(1-\alpha)\%$ credible intervals for all pairwise comparisons, $\theta_i - \theta_j$, are given by

$$E(\theta_i \mid \mathbf{y}) - E(\theta_j \mid \mathbf{y}) \pm T^{(2)}_\alpha,$$

where, as before, Monte Carlo is used to compute $E(\theta_i \mid \mathbf{y})$, $E(\theta_j \mid \mathbf{y})$, $T^{(2)}_\alpha$.

7.1.2 Multiple comparison for all contrasts

Here we concentrate on all possible contrasts in $\boldsymbol{\theta}$ (that is, $\boldsymbol{\ell}'\boldsymbol{\theta}$ such that $\sum_{i=1}^m \ell_i = 0$). Define

$$(\boldsymbol{\theta} - E(\boldsymbol{\theta} \mid \mathbf{y}))' \Big\{ \Big(\mathrm{var}(\boldsymbol{\theta} \mid \mathbf{y}) \Big)^{-1} - \frac{\Big(\mathrm{var}(\boldsymbol{\theta} \mid \mathbf{y}) \Big)^{-1} J_m \Big(\mathrm{var}(\boldsymbol{\theta} \mid \mathbf{y}) \Big)^{-1}}{\mathbf{1}_m' \Big(\mathrm{var}(\boldsymbol{\theta} \mid \mathbf{y}) \Big)^{-1} \mathbf{1}_m} \Big\} (\boldsymbol{\theta} - E(\boldsymbol{\theta} \mid \mathbf{y})) \, \Big| \, \mathbf{y} \equiv T^{(3)}.$$

Note (see Section 6.5) that subject to the constraint $\sum_{i=1}^m \ell_i = 0$,

$$\max_{\boldsymbol{\ell}} \frac{\Big(\boldsymbol{\ell}'(\boldsymbol{\theta} - E(\boldsymbol{\theta} \mid \mathbf{y})) \Big)^2}{\boldsymbol{\ell}' \mathrm{var}(\boldsymbol{\theta} \mid \mathbf{y}) \boldsymbol{\ell}} \, \Big| \, \mathbf{y} = T^{(3)}. \tag{7.4}$$

When σ^2 is known, in Section 6.5, it is shown that

$$T^{(3)} \sim \chi^2_{(m-1)}. \tag{7.5}$$

Thus simultaneous $100(1-\alpha)\%$ credible intervals for all $\boldsymbol{\ell}'\boldsymbol{\theta}$ such that $\sum_{i=1}^m \ell_i = 0$ are given by

$$\boldsymbol{\ell}'\Lambda\boldsymbol{\omega} \pm \left(\boldsymbol{\ell}'\Lambda\boldsymbol{\ell}\chi^2_{(\alpha,m-1)}\right)^{\frac{1}{2}}.$$

When σ^2 is unknown, Monte Carlo is used to compute $E(\boldsymbol{\theta} \mid \mathbf{y})$, $\text{var}(\boldsymbol{\theta} \mid \mathbf{y})$, $T^{(3)}_\alpha$, and in this case simultaneous $100(1-\alpha)\%$ credible intervals for all $\boldsymbol{\ell}'\boldsymbol{\theta}$ such that $\sum_{i=1}^m \ell_i = 0$ are given by

$$\boldsymbol{\ell}'E(\boldsymbol{\theta} \mid \mathbf{y}) \pm \left(\boldsymbol{\ell}'\text{var}(\boldsymbol{\theta} \mid \mathbf{y})\,\boldsymbol{\ell}\,T^{(3)}_\alpha\right)^{\frac{1}{2}}.$$

7.1.3 Multiple comparison for all $\boldsymbol{\ell}'\boldsymbol{\theta}$

Note that (proof is similar to (7.4))

$$\max_{\boldsymbol{\ell}} \frac{\left(\boldsymbol{\ell}'(\boldsymbol{\theta} - E(\boldsymbol{\theta} \mid \mathbf{y}))\right)^2}{\boldsymbol{\ell}'\text{var}(\boldsymbol{\theta} \mid \mathbf{y})\boldsymbol{\ell}} \,\Big|\, \mathbf{y} = (\boldsymbol{\theta} - E(\boldsymbol{\theta} \mid \mathbf{y}))'\{\text{var}(\boldsymbol{\theta} \mid \mathbf{y})\}^{-1}(\boldsymbol{\theta} - E(\boldsymbol{\theta} \mid \mathbf{y}))\,|\,\mathbf{y} \equiv T^{(4)}.$$

When σ^2 is known, $T^{(4)} \sim \chi^2_{(m)}$. Thus simultaneous $100(1-\alpha)\%$ credible intervals for $\ell'\theta$ for all $\ell \in R^m$ are given by

$$\boldsymbol{\ell}'\Lambda\boldsymbol{\omega} \pm \left(\boldsymbol{\ell}'\Lambda\,\boldsymbol{\ell}\,\chi^2_{(\alpha,m)}\right)^{\frac{1}{2}}.$$

When σ^2 is unknown, Monte Carlo is used to compute $E(\theta \mid \mathbf{y})$, $\text{var}(\theta \mid \mathbf{y})$, $T^{(4)}_\alpha$, and in this case simultaneous $100(1-\alpha)\%$ credible intervals for all $\boldsymbol{\ell}'\boldsymbol{\theta}$ for all

$\boldsymbol{\ell} \in R^m$ are given by

$$\boldsymbol{\ell}' E(\boldsymbol{\theta} \mid \mathbf{y}) \pm \Big(\boldsymbol{\ell}' \operatorname{var}(\theta \mid \mathbf{y}) \, \boldsymbol{\ell} \, T_{\alpha}^{(4)} \Big)^{\frac{1}{2}}.$$

7.2 Prior Selection

There are several ways one can choose the prior distribution for σ^2. A popular choice is Stein's superharmonic prior distribution given by

$$\pi(\sigma^2) \propto I_{[\sigma^2 > 0]}.$$

The above choice of prior is non-informative and is known to provide an admissible procedure in the context of point estimation (Morris and Christiansen [30]). The superharmonic prior was also used by Morris [31] in obtaining a suitable measure of uncertainty of his empirical Bayes estimator. In what follows, we consider another approach to choosing a prior for σ^2.

Given $\{w_i \geq 0, \ i = 1, \cdots, m, \ \text{such that} \ \sum_{i=1}^m w_i = 1\}$, we seek a prior $\pi(\sigma^2)$ satisfying the following condition:

$$\sum_{i=1}^{m} w_i E\Big(\operatorname{var}(\theta_i \mid \mathbf{y}) - MSE[\hat{\theta}_i(\hat{\sigma}^2)] \Big) \ = \ o(1/m), \tag{7.6}$$

where $\operatorname{var}(.|y)$ is the variance under the prior $\pi(\sigma^2)$, $E(\cdot)$ and $MSE(\cdot)$ are taken with respect to the Fay-Herriot model; $\hat{\theta}_i(\hat{\sigma}^2)$ is the EBLUP of θ_i, that is

$$\hat{\theta}_i(\hat{\sigma}^2) = x_i' \tilde{\boldsymbol{\beta}}(\hat{\sigma}^2) + \frac{\hat{\sigma}^2}{(\hat{\sigma}^2 + \psi_i)} (y_i - x_i' \tilde{\boldsymbol{\beta}}(\hat{\sigma}^2))$$

$$\tilde{\boldsymbol{\beta}}(\sigma^2) = (X' \Omega^{-1} X)^{-1} X' \Omega^{-1} \mathbf{y}$$

where $\Omega = \text{diag}(\psi_1 + \sigma^2, \ldots, \psi_m + \sigma^2)$, and $\hat{\sigma}^2$ is the REML estimator of σ^2. The choice of REML estimator for σ^2 is for convenience.

Assuming a general prior $\pi(\psi)$, Datta et al. [17] proved that the hierarchical Bayes estimator of θ_i has frequentist validation in the sense that, the EBLUP of θ_i and the hierarchical Bayes estimator of θ_i differ by terms of the order $O_p(1/m)$. Moreover, Datta et al. [17] chose a prior for which the posterior variance, a Bayesian measure of variability, has a certain frequentist property (see 7.10). Similarly, we seek a prior that satisfies (7.6). Such a prior has the following property: the weighted average of the posterior variances over all small areas is second order unbiased for the corresponding weighted average of the MSE of the EBLUP.

In order to satisfy (7.6), as shown in Section 6.5, it is necessary and sufficient for $\pi(\sigma^2)$ to satisfy the following differential equation

$$\frac{d\pi(\sigma^2)}{d\sigma^2}\frac{1}{\pi(\sigma^2)} - 2\frac{\sum_{i=1}^m w_i\psi_i^2/(\psi_i+\sigma^2)^3}{\sum_{i=1}^m w_i\{\psi_i/(\psi_i+\sigma^2)\}^2} + 2\frac{\sum_{i=1}^m 1/(\psi_i+\sigma^2)^3}{\sum_{i=1}^m 1/(\psi_i+\sigma^2)^2} = 0. \tag{7.7}$$

It can be checked that the solution to (7.7) is given by

$$\pi(\sigma^2) \propto \frac{\sum_{i=1}^m 1/(\psi_i+\sigma^2)^2}{\sum_{i=1}^m w_i\{\psi_i/(\psi_i+\sigma^2)\}^2}. \tag{7.8}$$

When the prior is given by (7.8), it can be checked that for $m+2 > \text{rank}(X)$ the posterior distribution of θ is proper (Datta et al. [17]). It is interesting to note that Stein's super-harmonic prior is a special case of (7.8): simply take $w_i = \frac{1/\psi_i^2}{\sum_{j=1}^m 1/\psi_j^2}$. The superharmonic prior could be interpreted as a prior under which the weighted average of the posterior variance of θ_i is a second-order unbiased estimator of the corresponding weighted average of the MSE of the EBLUP, the average being taken

over all small areas and the weight for a given area being proportional to the inverse of the squared sampling variance.

By taking $w_i = 1/m$ (for $i = 1, \ldots, m$), we get the following prior which we refer to as the "average moment matching prior":

$$\pi(\sigma^2) \propto \frac{\sum_{i=1}^{m} 1/(\psi_i + \sigma^2)^2}{\sum_{i=1}^{m} \{\psi_i/(\psi_i + \sigma^2)\}^2}. \tag{7.9}$$

The prior given by (7.9) has the property that the average posterior variance of θ_i is second-order unbiased for the average MSE of the EBLUP of θ_i. Also, by taking $w_j = 1$ for $j = i$, and $w_j = 0$ for $j \neq i$, we get a prior obtained by Datta et al. [17]. Their main motivation was to choose a prior distribution for σ^2 such that the posterior variance of θ_i is second-order unbiased for the mean squared error of the EBLUP of θ_i, that is,

$$E\Big(\text{var}(\theta_i \mid \mathbf{y})\Big) = MSE\Big(\hat{\theta}_i(\hat{\sigma}^2)\Big) + o(1/m). \tag{7.10}$$

Datta et al. [17] showed that the prior which satisfies (7.10) is given by

$$\pi(\sigma^2) \propto \ (\psi_i + \sigma^2)^2 \sum_{j=1}^{m} \frac{1}{(\psi_j + \sigma^2)^2}. \tag{7.11}$$

Note that, unless $\psi_i = \psi$ for $i = 1, \ldots, m$, the prior for σ^2 is area specific and hence it is not possible to select a prior which satisfies (7.11) simultaneously for $i = 1, \ldots, m$.

7.3 Implementation by Monte Carlo

It is straightforward to show that (see Section 6.5)

$$f_{\sigma^2|y}(\sigma^2 \mid \mathbf{y}) \propto \pi(\sigma^2) \prod_{i=1}^{m} \frac{\exp(-\frac{1}{2}\mathbf{y}'(\Omega^{-1} - \Omega^{-1}X(X'\Omega^{-1}X)^{-1}X'\Omega^{-1})\mathbf{y})}{(\psi_i + \sigma^2)^{\frac{1}{2}}|X'\Omega^{-1}X|^{\frac{1}{2}}} \tag{7.12}$$

$$\boldsymbol{\beta} \mid \sigma^2, \mathbf{y} \sim N((X'\Omega^{-1}X)^{-1}X'\Omega^{-1}\mathbf{y}, (X'\Omega^{-1}X)^{-1}) \tag{7.13}$$

$$\theta \mid \boldsymbol{\beta}, \sigma^2, \mathbf{y} \sim N(\Gamma\boldsymbol{\zeta}, \Gamma) \tag{7.14}$$

where $\Omega = \text{diag}(\psi_1 + \sigma^2, \ldots, \psi_m + \sigma^2)$, $\Gamma = \text{diag}(\frac{\psi_1\sigma^2}{\psi_1+\sigma^2}, \ldots \frac{\psi_m\sigma^2}{\psi_m+\sigma^2})$, , $|X'\Omega^{-1}X|$ is the determinant of $X'\Omega^{-1}X$, $\boldsymbol{\zeta} = \frac{X\boldsymbol{\beta}}{\sigma^2} + \text{diag}(\frac{1}{\psi_1}, \ldots, \frac{1}{\psi_m})\mathbf{y}$.

We need to generate $(\boldsymbol{\theta}_*, \boldsymbol{\beta}_*, \sigma_*^2)$ from $f(\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma^2 \mid \mathbf{y})$. To this end, note that

$$f(\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma^2 \mid \mathbf{y}) \propto f_{\sigma^2|\mathbf{y}}(\sigma^2 \mid \mathbf{y}) f(\boldsymbol{\beta} \mid \sigma^2, \mathbf{y}) f(\theta \mid \boldsymbol{\beta}, \sigma^2, y).$$

Hence $(\boldsymbol{\theta}_*, \boldsymbol{\beta}_*, \sigma_*^2)$ will be generated as follows: $\sigma_*^2 \sim f_{\sigma^2|\mathbf{y}}(\sigma^2 \mid \mathbf{y})$, $\boldsymbol{\beta}_* \sim f(\boldsymbol{\beta} \mid \sigma_*^2, \mathbf{y})$, $\boldsymbol{\theta}_* \sim f(\boldsymbol{\theta} \mid \boldsymbol{\beta}_*, \sigma_*^2, \mathbf{y})$. Simulating $\boldsymbol{\beta}_* \sim f(\boldsymbol{\beta} \mid \sigma_*^2, \mathbf{y})$ and $\boldsymbol{\theta}_* \sim f(\boldsymbol{\theta} \mid \boldsymbol{\beta}_*, \sigma_*^2, \mathbf{y})$ is straightforward. To simulate $\sigma_*^2 \sim f_{\sigma^2|\mathbf{y}}(\sigma^2 \mid \mathbf{y})$, use the following accept-reject method (for a discussion of the accept-reject method, see Robert and Casella [40]):

1. Simulate $z \sim \chi^2_{(m-q-2)}$ [where $q = \text{rank}(X)$].

2. Compute $u = \frac{\mathbf{y}'(I - X(X'X)^{-1}X')\mathbf{y}}{z} - \varphi$. If $u \geq 0$, then $u \sim f_U(u)$, where

$$f_U(u) \propto \frac{\exp(-\frac{1}{2(\varphi+u)}\mathbf{y}'(I - X(X'X)^{-1}X')\mathbf{y})}{(\varphi + u)^{(m-q)/2}} I_{[u \geq 0]}.$$

   φ is chosen such that the acceptance rate in the accept-reject method is maximized or we could simply choose φ to be the median of the ψ_i's.

3. Generate $w \sim \text{Unif}[0, 1]$.

4. Check if $\frac{1}{K}\,\frac{f_{\sigma^2|\mathbf{y}}(u|\mathbf{y})}{f_U(u)} \geq w$, where $K = \max_t \frac{f_{\sigma^2|\mathbf{y}}(t|\mathbf{y})}{f_U(t)}$. If true, then $u \sim f_{\sigma^2|\mathbf{y}}(\sigma^2\,|\,\mathbf{y})$.

7.4 Data analysis and simulation

In this section, we use a well-known data set to illustrate to what extent the theoretically valid methods for multiple comparison differ from the naive comparison based on individual credible intervals. In our study, we include both pairwise comparisons and comparisons of general contrasts. Also, a simulation study is performed to compare the average moment matching prior (7.9) with that of Stein's superharmonic prior.

In our data analysis, we use the baseball run scoring data given in Morris and Christiansen [30]. The baseball data set (Table 7.1) gives the average runs scored per game and sample standard deviation of 14 baseball teams in the American League for the year 1993. Each of the teams played 162 games, and y_i denotes the average runs scored over those 162 games. A good approximation given in Morris and Christiansen [30] for the variance of runs scored for a single game is $\text{var}(\mu) = (1.375\mu)^{1.2}$, where μ is the mean runs scored for a single game. For the 162 games played, the variance ψ_i for the i^{th} team is then approximated by $\psi_i = \text{var}(y_i)/162 = (1.375y_i)^{1.2}/162$, and is assumed to be known without error. The normality assumption for y_i is justified by the central limit theorem. The estimates of the true runs per game θ_i and its standard error given in Table 7.1 were computed using $20,000$ independent samples for each of the two priors.

Table 7.1: Estimates of the true runs/game and its s.e., using the superharmonic prior (columns 5 and 6) and average moment matching prior (columns 7 and 8).

Obs	Team	y_i	$\sqrt{\psi_i}$	θ_i	s_i	θ_i^*	s_i^*
1	Det	5.549	0.266	5.287	0.250	5.290	0.250
2	Tor	5.228	0.257	5.070	0.227	5.073	0.230
3	Tex	5.154	0.254	5.022	0.225	5.021	0.226
4	NY	5.068	0.252	4.962	0.221	4.961	0.221
5	Cle	4.877	0.246	4.827	0.214	4.829	0.214
6	Bal	4.852	0.245	4.808	0.212	4.809	0.211
7	Chi	4.790	0.243	4.765	0.210	4.764	0.211
8	Sea	4.531	0.235	4.570	0.205	4.573	0.205
9	Mil	4.525	0.235	4.569	0.206	4.567	0.206
10	Oak	4.414	0.232	4.483	0.207	4.486	0.205
11	Min	4.278	0.227	4.379	0.205	4.381	0.205
12	Bos	4.235	0.226	4.346	0.205	4.348	0.205
13	Cal	4.222	0.226	4.336	0.204	4.337	0.205
14	KC	4.167	0.224	4.293	0.208	4.294	0.205

Table 7.2: Credible intervals for selected contrasts using the superharmonic prior

Contrast	All contrasts	Pairwise	Individual
$\theta_1 - \theta_{14}$	(-0.691,2.680)	(-0.026,2.015)	(0.341,1.682)
$\theta_2 - \theta_{14}$	(-0.785,2.339)	(-0.244,1.797)	(0.173,1.417)
$\theta_4 - \theta_{12}$	(-0.887,2.120)	(-0.404,1.637)	(0.034,1.228)
$\theta_5 - \theta_{13}$	(-0.965,1.947)	(-0.529,1.511)	(-0.070,1.084)
$\frac{1}{2}(\theta_2 + \theta_3) - \theta_{13}$	(-0.627,2.045)	not pairwise	(0.189,1.251)
$\frac{1}{3}(\theta_1 + \theta_2 + \theta_3 - \theta_{12} - \theta_{13} - \theta_{14})$	(-0.250,1.852)	not pairwise	(0.382,1.219)

It is interesting to note that the average moment matching prior gives very similar results to the ones obtained when the superharmonic prior is used. This is possibly because, for the baseball data set, there is little variability in the sampling errors. Hence, the weights $w_i = \frac{1/\psi_i^2}{\sum_{j=1}^{m} 1/\psi_j^2}$ that generate the superharmonic prior are more or less uniform across areas.

For the baseball data set, Tables 7.2-7.3 give 95% credible intervals for a few contrasts of interest. Note that when an appropriate multiple comparison method is used, the coverage probability holds simultaneously for all contrasts or pairwise comparisons. If instead, before looking at the data, a practitioner decides that a specific $\boldsymbol{\ell}'\boldsymbol{\theta}$ is the only contrast of interest, then a much shorter interval can be obtained by using (7.3). As can be seen from Tables 7.2-7.3, in a number of instances, after looking at the data, if a practitioner were to naively use (7.3), he/she would incorrectly reject the null hypothesis $H_o : \boldsymbol{\ell}'\boldsymbol{\theta} = 0$ when it should be accepted.

In an attempt to further investigate our class of priors, in a simulation study we

Table 7.3: Credible intervals for selected contrasts using the moment matching prior

Contrast	All contrasts	Pairwise	Individual
$\theta_1 - \theta_{14}$	(-0.666,2.657)	(-0,027,2.018)	(0.348,1.667)
$\theta_2 - \theta_{14}$	(-0.785,2.343)	(-0.244,1.801)	(0.176,1.419)
$\theta_4 - \theta_{12}$	(-0.900,2.125)	(-0.410,1.636)	(0.034,1.230)
$\theta_5 - \theta_{13}$	(-0.965,1.949)	(-0.531,1.515)	(-0.077,1.088)
$\frac{1}{2}(\theta_2 + \theta_3) - \theta_{13}$	(-0.636,2.055)	not pairwise	(0.187,1.253)
$\frac{1}{3}(\theta_1 + \theta_2 + \theta_3 - \theta_{12} - \theta_{13} - \theta_{14})$	(-0.252,1.854)	not pairwise	(0.379,1.217)

consider two different patterns for the sampling errors ψ_i and compare the average moment matching prior with the uniform prior. The simulation setup we consider is similar to the one given in Datta, Rao and Smith [17]. In the first pattern, the ψ_i's are more or less equal across areas. In the second pattern, there is considerable variation in ψ_i's, so that the weights that generate the superharmonic prior are also quite variable.

Similar to Datta et al. [17], our simulation setup is as follows: $m = 15$, $\sigma^2 = 1$, five groups G_1, G_2, G_3, G_4, G_5, with three small areas having the same ψ_i value of 0.7, 0.6, 0.5, 0.4, 0.3 [pattern (a)] and 4.0, 0.6, 0.5, 0.4, 0.1 [pattern (b)]. Note that our ψ_i patterns (a) and (b) are same as the Type I and Type III patterns of Datta, Rao and Smith [17] respectively. For the entire simulation, $\boldsymbol{\beta} = (1, 1)'$ was fixed, and the scalar covariate x_i was generated uniformly on $[0, 1]$, and then fixed for the entire simulation run. The above simulation was run 100 times, and for each simulation run, the posterior distributions of $\boldsymbol{\theta}$, $\boldsymbol{\beta}$, ψ were approximated by 10000

runs of the monte carlo method discussed in Section 7.3.

Tables 7.4-7.7 summarize for each of the groups G_1-G_5, the average coverage of a nominal 95% equal-tailed credible interval for θ_i, average length of the aforementioned credible interval, and the average integrated Bayes risk (same as the MSE) of the θ_i's. In computing the coverage and integrated Bayes risk, the joint distribution of $\mathbf{y}$ and $\boldsymbol{\theta}$ is used. We take the average over all small areas in the same group and over all simulation runs. The last column in Tables 7.4-7.7 gives similar summary statistics for the prior variance $\sigma^2 = 1$, although, unlike the θ_i's, the average is only taken over all simulation runs.

Table 7.4: Summary of simulation results for θ_i in each group and for σ^2 using the superharmonic prior when $m = 15$, $\sigma^2 = 1$ and pattern (a) for ψ_i's.

	G_1	G_2	G_3	G_4	G_5	σ^2
Coverage	0.930	0.970	0.967	0.950	0.940	0.930
Length	2.755	2.612	2.435	2.229	1.971	4.335
Risk	0.471	0.419	0.363	0.354	0.261	1.220

As can be seen from Tables 7.4-7.5, there is little to choose between the two priors for ψ_i pattern (a). When the ψ_i's have pattern (b), for G_1 there is a reduction of 10% in the average length of the credible interval for the same coverage, and a 6.75% reduction in average risk by using the moment matching prior as opposed to the superharmonic prior. For G_2 the gains are smaller, and for the remaining groups there is little or no difference between the two priors. In terms of estimation

Table 7.5: Summary of simulation results for θ_i in each group and for σ^2 using the moment matching prior when $m = 15$, $\sigma^2 = 1$ and pattern (a) for ψ_i's.

	G_1	G_2	G_3	G_4	G_5	σ^2
Coverage	0.930	0.970	0.960	0.950	0.943	0.940
Length	2.744	2.605	2.428	2.223	1.964	4.280
Risk	0.472	0.420	0.362	0.354	0.261	1.167

Table 7.6: Summary of simulation results for θ_i in each group and for σ^2 using the superharmonic prior when $m = 15$, $\sigma^2 = 1$ and pattern (b) for ψ_i's.

	G_1	G_2	G_3	G_4	G_5	σ^2
Coverage	1.000	0.933	0.967	0.933	0.967	0.900
Length	4.436	2.562	2.406	2.235	1.202	4.719
Risk	0.652	0.483	0.276	0.309	0.103	1.753

of σ^2, significant gains can be achieved by using the moment matching prior. For example, the average length of the credible interval is 26% shorter and the average risk is reduced by 45%.

In conclusion, we remark that our simulation study suggests that one may consider using the average moment matching prior over the superharmonic prior when there is substantial variation in the sampling errors.

Table 7.7: Summary of simulation results for θ_i in each group and for σ^2 using the moment matching prior when $m = 15$, $\sigma^2 = 1$ and pattern (b) for ψ_i's.

	G_1	G_2	G_3	G_4	G_5	σ^2
Coverage	1.000	0.933	0.933	0.933	0.967	0.900
Length	3.993	2.435	2.310	2.159	1.189	3.482
Risk	0.608	0.477	0.284	0.305	0.102	0.960

7.5 Appendix

Derivation of (7.1). For known σ^2 and $\psi_1, \ldots, \psi_m$,

$$
\begin{aligned}
f(\boldsymbol{\theta}|\mathbf{y}) &\propto \int_{\boldsymbol{\beta}} f(\mathbf{y}|\boldsymbol{\theta}) f(\boldsymbol{\theta}|\boldsymbol{\beta}) \pi(\boldsymbol{\beta}) d\boldsymbol{\beta} \\
&\propto \int_{\boldsymbol{\beta}} \exp\Big(- \sum_{i=1}^{m} \frac{(y_i - \theta_i)^2}{2\psi_i} - \sum_{i=1}^{m} \frac{(\theta_i - \mathbf{x}_i'\boldsymbol{\beta})^2}{2\sigma^2} \Big) d\boldsymbol{\beta} \\
&\propto \int_{\boldsymbol{\beta}} \exp\Big(- \sum_{i=1}^{m} \frac{(y_i - \theta_i)^2}{2\psi_i} - \frac{1}{2\sigma^2} \Big(\boldsymbol{\beta}' X' X \boldsymbol{\beta} - 2\boldsymbol{\theta}' X \boldsymbol{\beta} + \boldsymbol{\theta}'\boldsymbol{\theta} \Big) \Big) d\boldsymbol{\beta} \\
&\propto \exp\Big(- \sum_{i=1}^{m} \Big(\frac{\theta_i^2}{2\psi_i} - \frac{y_i \theta_i}{\psi_i} \Big) - \frac{1}{2\sigma^2} \boldsymbol{\theta}' \Big(I_m - X(X'X)^{-1} X' \Big) \boldsymbol{\theta} \Big) \\
&\propto \exp\Big(- \frac{1}{2} \Big(\boldsymbol{\theta}' \Lambda^{-1} \boldsymbol{\theta} - 2\boldsymbol{\omega}' \boldsymbol{\theta} \Big) \Big) \\
&\propto \exp\Big(- \frac{1}{2} (\boldsymbol{\theta} - \Lambda\boldsymbol{\omega})' \Lambda^{-1} (\boldsymbol{\theta} - \Lambda\boldsymbol{\omega}) \Big)
\end{aligned}
$$

where $\boldsymbol{\omega} = (\frac{y_1}{\psi_1}, \ldots, \frac{y_m}{\psi_m})'$ and $\Lambda^{-1} = \mathrm{diag}(\frac{1}{\psi_1} + \frac{1}{\sigma^2}, \ldots, \frac{1}{\psi_m} + \frac{1}{\sigma^2}) - \frac{X(X'X)^{-1}X'}{\sigma^2}$, and the result follows. □

Derivation of (6.4). For notational convenience, let $\mathbf{Z}$ be a random vector such that $\mathrm{E}(\mathbf{Z}) = \boldsymbol{\mu}$ and $\mathrm{var}(\mathbf{Z}) = \Upsilon$. We shall show that subject to the constraint $\sum_{i=1}^{m} \ell_i = 0$

$$
\max_{\boldsymbol{\ell}} \frac{(\boldsymbol{\ell}'(\mathbf{Z} - \boldsymbol{\mu}))^2}{\boldsymbol{\ell}' \Upsilon \boldsymbol{\ell}} = (\mathbf{Z} - \boldsymbol{\mu})' \Big\{ \Upsilon^{-1} - \frac{\Upsilon^{-1} J_m \Upsilon^{-1}}{\mathbf{1}_m \Upsilon^{-1} \mathbf{1}_m} \Big\} (\mathbf{Z} - \boldsymbol{\mu}).
$$

Let

$$f = \frac{(\boldsymbol{\ell}'(\mathbf{Z}-\boldsymbol{\mu}))^2}{\boldsymbol{\ell}'\Upsilon\,\boldsymbol{\ell}} + \lambda\boldsymbol{\ell}'\mathbf{1}_m.$$

Setting $\frac{\partial f}{\partial \ell} = 0$ and $\frac{\partial f}{\partial \lambda} = 0$, we have

$$\begin{aligned} \frac{2\{\boldsymbol{\ell}^{'}(\mathbf{Z}-\boldsymbol{\mu})\}}{\boldsymbol{\ell}^{'}\Upsilon\,\boldsymbol{\ell}}(\mathbf{Z}-\boldsymbol{\mu}) - \frac{2\{\boldsymbol{\ell}^{'}(\mathbf{Z}-\boldsymbol{\mu})\}^2}{\{\boldsymbol{\ell}^{'}\Upsilon\,\boldsymbol{\ell}\}^2}\Upsilon\,\boldsymbol{\ell} + \lambda\mathbf{1}_m &= 0 \qquad (7.15)\\ \boldsymbol{\ell}^{'}\mathbf{1}_m &= 0. \end{aligned}$$

Multiplying (7.15) by $\mathbf{1}'_m\Upsilon^{-1}$ and solving for λ

$$\lambda = -\frac{2\{\boldsymbol{\ell}'(\mathbf{Z}-\boldsymbol{\mu})\}\{\mathbf{1}'_m\Upsilon^{-1}(\mathbf{Z}-\boldsymbol{\mu})\}}{\{\mathbf{1}'_m\Upsilon^{-1}\mathbf{1}_m\}\{\boldsymbol{\ell}'\Upsilon\,\boldsymbol{\ell}\}}.$$

Substituting λ in (7.15), we get

$$\boldsymbol{\ell} = \frac{\boldsymbol{\ell}'\Upsilon\,\boldsymbol{\ell}}{\boldsymbol{\ell}'(\mathbf{Z}-\boldsymbol{\mu})}\Upsilon^{-1}(\mathbf{Z}-\boldsymbol{\mu}) - \frac{\{\boldsymbol{\ell}'\Upsilon\,\boldsymbol{\ell}\}\{\mathbf{1}'_m\Upsilon^{-1}(\mathbf{Z}-\boldsymbol{\mu})\}}{\{\mathbf{1}'_m\Upsilon^{-1}\mathbf{1}_m\}\{\boldsymbol{\ell}'(\mathbf{Z}-\boldsymbol{\mu})\}}\Upsilon^{-1}\mathbf{1}_m. \qquad (7.16)$$

Using (7.16), we obtain

$$\begin{aligned} \max_{\ell} f &= \max_{\ell} \frac{(\boldsymbol{\ell}'(\mathbf{Z}-\boldsymbol{\mu}))^2}{\boldsymbol{\ell}'\Upsilon\,\boldsymbol{\ell}} + \lambda\boldsymbol{\ell}'\mathbf{1}_m \\ &= \frac{\Big\{(\mathbf{Z}-\boldsymbol{\mu})^{'}\Upsilon^{-1}(\mathbf{Z}-\boldsymbol{\mu}) - \frac{\{\mathbf{1}'_m\Upsilon^{-1}(\mathbf{Z}-\boldsymbol{\mu})\}^2}{\mathbf{1}'_m\Upsilon^{-1}\mathbf{1}_m}\Big\}^2}{\Big\{(\mathbf{Z}-\boldsymbol{\mu})^{'}\Upsilon^{-1} - \frac{\mathbf{1}'_m\Upsilon^{-1}(\mathbf{Z}-\boldsymbol{\mu})}{\mathbf{1}'_m\Upsilon^{-1}\mathbf{1}_m}\mathbf{1}'_m\Upsilon^{-1}\Big\}\Upsilon\Big\{\Upsilon^{-1}(\mathbf{Z}-\boldsymbol{\mu}) - \frac{\mathbf{1}'_m\Upsilon^{-1}(\mathbf{Z}-\boldsymbol{\mu})}{\mathbf{1}'_m\Upsilon^{-1}\mathbf{1}_m}\Upsilon^{-1}\mathbf{1}_m\Big\}} \\ &= (\mathbf{Z}-\boldsymbol{\mu})^{'}\Upsilon^{-1}(\mathbf{Z}-\boldsymbol{\mu}) - \frac{\{\mathbf{1}'_m\Upsilon^{-1}(\mathbf{Z}-\boldsymbol{\mu})\}^2}{\mathbf{1}'_m\Upsilon^{-1}\mathbf{1}_m} \\ &= (\mathbf{Z}-\boldsymbol{\mu})'\Big\{\Upsilon^{-1} - \frac{\Upsilon^{-1}J_m\Upsilon^{-1}}{\mathbf{1}'_m\Upsilon^{-1}\mathbf{1}_m}\Big\}(\mathbf{Z}-\boldsymbol{\mu}). \qquad (7.17) \end{aligned}$$

□

Derivation of (6.5). Moreover, for $\mathbf{Z} \sim N(\boldsymbol{\mu}, \Upsilon)$, consider the quadratic form given in (7.17). By Searle [44] [Theorem 2, p.57], since

$$\Big(\Upsilon^{-1} - \frac{\Upsilon^{-1}J_m\Upsilon^{-1}}{\mathbf{1}'_m\Upsilon^{-1}\mathbf{1}_m}\Big)\Upsilon = I_m - \frac{\Upsilon^{-1}J_m}{\mathbf{1}'_m\Upsilon^{-1}\mathbf{1}_m}$$

is idempotent, and

$$\text{rank}\Big(I_m - \frac{\Upsilon^{-1}J_m}{\mathbf{1}_m'\Upsilon^{-1}\mathbf{1}_m}\Big) = \text{tr}\Big(I_m - \frac{\Upsilon^{-1}J_m}{\mathbf{1}_m'\Upsilon^{-1}\mathbf{1}_m}\Big) = m-1,$$

it follows that

$$(\mathbf{Z}-\boldsymbol{\mu})'\Big\{\Upsilon^{-1} - \frac{\Upsilon^{-1}J_m\Upsilon^{-1}}{\mathbf{1}_m'\Upsilon^{-1}\mathbf{1}_m}\Big\}(\mathbf{Z}-\boldsymbol{\mu}) \sim \chi^2_{(m-1)}.$$

□

Derivation of (6.7). Using the approximations for $E\Big(\text{var}(\theta_i \mid \mathbf{y})\Big)$ and $MSE\Big(\hat{\theta}_i(\hat{\sigma}^2)\Big)$ given in Datta et al. [17], we have

$$E\Big(\text{var}(\theta_i \mid \mathbf{y})\Big) = g_{1i}(\sigma^2) + g^{\star}_{1\pi i}(\sigma^2) + g_{2i}(\sigma^2) + o(1/m) \tag{7.18}$$

$$MSE\Big(\hat{\theta}_i(\hat{\sigma}^2)\Big) = g_{1i}(\sigma^2) + g_{2i}(\sigma^2) + g_{3i}(\sigma^2) + o(1/m) \tag{7.19}$$

where

$$\begin{aligned}
g_{1i}(\sigma^2) &= \frac{\psi_i\sigma^2}{\psi_i+\sigma^2}\\
g_{2i}(\sigma^2) &= \frac{\psi_i^2}{(\psi_i+\sigma^2)^2}\mathbf{x}_i^{'}\Big(\sum_{j=1}^{m}\frac{\mathbf{x}_j\mathbf{x}_j^{'}}{\psi_j+\sigma^2}\Big)^{-1}\mathbf{x}_i\\
g^{\star}_{1\pi i}(\sigma^2) &= \frac{2\psi_i^2}{(\psi_i+\sigma^2)^2}\frac{1}{\sum_{j=1}^{m}(\psi_j+\sigma^2)^{-2}}\Big(\frac{d\pi(\sigma^2)}{d\sigma^2}\frac{1}{\pi(\sigma^2)} - \frac{1}{\psi_i^2+\sigma^2} + 2\frac{\sum_{j=1}^{m}(\psi_j+\sigma^2)^{-3}}{\sum_{j=1}^{m}(\psi_j+\sigma^2)^{-2}}\Big)\\
g_{3i}(\sigma^2) &= \frac{2\psi_i^2}{(\psi_i+\sigma^2)^3}\frac{1}{\sum_{j=1}^{m}(\psi_j+\sigma^2)^{-2}}.
\end{aligned}$$

Using (7.6), (7.18) and (7.19), it follows that

$$\begin{aligned}
\frac{d\pi(\sigma^2)}{d\sigma^2}\frac{1}{\pi(\sigma^2)}\sum_{i=1}^{m}w_i\frac{\psi_i^2}{(\psi_i+\sigma^2)^2}\frac{1}{\sum_{i=1}^{m}(\psi_i+\sigma^2)^{-2}} - \sum_{i=1}^{m}w_i\frac{\psi_i^2}{(\psi_i+\sigma^2)^3}\frac{1}{\sum_{i=1}^{m}(\psi_i+\sigma^2)^{-2}} +\\
2\sum_{i=1}^{m}w_i\frac{\psi_i^2}{(\psi_i+\sigma^2)^2}\frac{\sum_{i=1}^{m}(\psi_i+\sigma^2)^{-3}}{\Big(\sum_{i=1}^{m}(\psi_i+\sigma^2)^{-2}\Big)^2} = \sum_{i=1}^{m}w_i\frac{\psi_i^2}{(\psi_i+\sigma^2)^3}\frac{1}{\sum_{i=1}^{m}(\psi_i+\sigma^2)^{-2}}
\end{aligned}$$

and by rearranging terms we get (7.7). □

Derivation of (6.12)-(6.14). For (6.12), see Datta et al. [17]. For (6.13):

$$
\begin{aligned}
f(\boldsymbol{\beta}|\sigma^2, \mathbf{y}) &\propto \int_{\boldsymbol{\theta}} f(\mathbf{y}|\boldsymbol{\theta}) f(\boldsymbol{\theta}|\boldsymbol{\beta}) \pi(\boldsymbol{\beta}, \sigma^2) d\boldsymbol{\theta} \\
&\propto \int_{\boldsymbol{\theta}} \exp\Big(- \sum_{i=1}^{m} \frac{(y_i - \theta_i)^2}{2\psi_i} - \frac{1}{2\sigma^2} (\boldsymbol{\theta} - X\boldsymbol{\beta})'(\boldsymbol{\theta} - X\boldsymbol{\beta}) \Big) d\boldsymbol{\theta} \\
&\propto \exp\Big(- \frac{1}{2\sigma^2} \boldsymbol{\beta}' X' X \boldsymbol{\beta} \Big) \int_{\boldsymbol{\theta}} \exp\Big(- \frac{1}{2} (\boldsymbol{\theta}' \Gamma^{-1} \boldsymbol{\theta} - 2\boldsymbol{\zeta}' \boldsymbol{\theta}) \Big) d\boldsymbol{\theta} \\
&\propto \exp\Big(- \frac{1}{2\sigma^2} \boldsymbol{\beta}' X' X \boldsymbol{\beta} + \frac{1}{2} \boldsymbol{\zeta}' \Gamma \boldsymbol{\zeta} \Big) \\
&\propto \exp\Big(- \frac{1}{2} (\boldsymbol{\beta}' X' \Omega^{-1} X \boldsymbol{\beta} - 2(X' \Omega^{-1} \mathbf{y})' \boldsymbol{\beta}) \Big) \\
&\propto \exp\Big(- \frac{1}{2} (\boldsymbol{\beta} - (X' \Omega^{-1} X)^{-1} X' \Omega^{-1} \mathbf{y})' X' \Omega^{-1} X (\boldsymbol{\beta} - (X' \Omega^{-1} X)^{-1} X' \Omega^{-1} \mathbf{y}) \Big)
\end{aligned}
$$

where $\Omega = \mathrm{diag}(\psi_1 + \sigma^2, \ldots, \psi_m + \sigma^2)$, $\Gamma = \mathrm{diag}(\frac{\psi_1 \sigma^2}{\psi_1 + \sigma^2}, \ldots, \frac{\psi_m \sigma^2}{\psi_m + \sigma^2})$, $\boldsymbol{\zeta} = \frac{X\boldsymbol{\beta}}{\sigma^2} + \mathrm{diag}(\frac{1}{\psi_1}, \ldots, \frac{1}{\psi_m})\mathbf{y}$, and the result follows.

For (6.14):

$$
\begin{aligned}
f(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\beta}, \sigma^2) &\propto f(\mathbf{y}|\boldsymbol{\theta}) f(\boldsymbol{\theta}|\boldsymbol{\beta}, \sigma^2) \\
&\propto \exp\Big(- \sum_{i=1}^{m} \frac{(y_i - \theta_i)^2}{2\psi_i} - \frac{1}{2\sigma^2} (\boldsymbol{\theta} - X\boldsymbol{\beta})'(\boldsymbol{\theta} - X\boldsymbol{\beta}) \Big) \\
&\propto \exp\Big(- \frac{1}{2} (\boldsymbol{\theta}' \Gamma^{-1} \boldsymbol{\theta} - 2\boldsymbol{\zeta}' \boldsymbol{\theta}) \Big) \\
&\propto \exp\Big(- \frac{1}{2} (\boldsymbol{\theta} - \Gamma\boldsymbol{\zeta})' \Gamma^{-1} (\boldsymbol{\theta} - \Gamma\boldsymbol{\zeta}) \Big)
\end{aligned}
$$

and the result follows. □

Chapter 8

Robust mean squared error estimator

As mentioned in Chapter 1, the nested error regression model is given by

$$y_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta} + v_i + e_{ij},\ i = 1, \ldots, m,\ j = 1, \ldots, n_i, \tag{8.1}$$

where y_{ij} is the j^{th} observation in the i^{th} small area, $\mathbf{x}_{ij}$ is a vector of known covariates at the unit-level, v_i's and e_{ij}'s are independent with $v_i \overset{iid}{\sim} N(0, \sigma_v^2)$ and $e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2)$.

The model given by (8.1) along with the Fay-Herriot model are the two most popular models in small area estimation. To estimate areas planted with corn and soybeans for twelve counties in North-Central Iowa, Battese et al. [3] used (8.1). The parameter of interest is the small area mean $\theta_i = \overline{\mathbf{X}}_i'\boldsymbol{\beta} + v_i$, where $\overline{\mathbf{X}}_i$ is the known population mean of the covariates of the i^{th} small area. Usually θ_i is predicted by the best linear unbiased predictor $\hat{\theta}_i(\boldsymbol{\sigma}^2)$ (Battese et al. [3], Prasad and Rao [36] and Rao [37]). Here, $\boldsymbol{\sigma}^2 = (\sigma_v^2, \sigma_e^2)'$ is the vector of variance components. Since $\hat{\theta}_i(\boldsymbol{\sigma}^2)$ contains unknown variance components, an empirical BLUP (EBLUP) is given by $\hat{\theta}_i(\hat{\boldsymbol{\sigma}}^2)$, where $\hat{\boldsymbol{\sigma}}^2 = (\hat{\sigma}_v^2, \hat{\sigma}_e^2)'$ is a consistent estimator of $\boldsymbol{\sigma}^2$.

An important problem in small area estimation has to do with estimating the mean squared error (MSE) of the EBLUP. For the model given by (8.1), under the assumption of normality of the v_i's and e_{ij}'s, Prasad and Rao [36] derived an estimator of the MSE of the EBLUP which was second order unbiased. An estimator of the MSE of $\hat{\theta}_i(\hat{\boldsymbol{\sigma}}^2)$ is said to be second order unbiased if

$$E\Big(\widehat{\mathrm{MSE}}[\,\hat{\theta}_i(\hat{\boldsymbol{\sigma}}^2)\,]\Big) = \mathrm{MSE}[\,\hat{\theta}_i(\hat{\boldsymbol{\sigma}}^2)\,] + o(m^{-1}) \tag{8.2}$$

where $\mathrm{MSE}[\,\hat{\theta}_i(\hat{\boldsymbol{\sigma}}^2)\,]$ is the mean squared error of $\hat{\theta}_i(\hat{\boldsymbol{\sigma}}^2)$, $\widehat{\mathrm{MSE}}[\,\hat{\theta}_i(\hat{\boldsymbol{\sigma}}^2)\,]$ is an estimator of $\mathrm{MSE}[\,\hat{\theta}_i(\hat{\boldsymbol{\sigma}}^2)\,]$, and m is the number of sampled small areas.

Moreover, recently there has been interest in relaxing the normality assumption. For example, Hall and Maiti [20] derived a non-parametric second order unbiased estimator of the MSE of the EBLUP using a double bootstrap method. In this chapter, for a special case of the nested error regression model, without assuming any distributional assumptions, we derive an estimator of the MSE of the EBLUP that satisfies (8.2). Unlike the estimator given by Hall and Maiti [20], our estimator is closed form. Moreover, for the balanced case (that is for all i, $n_i = k$), and when the e_{ij}'s are normally distributed, we show that the Prasad Rao MSE estimator is second order unbiased. Through simulation, we show that the Prasad Rao MSE estimator is robust for departures from normality.

The model we consider is $\mathbf{x}_{ij}'\boldsymbol{\beta} = \mu$, that is, a common means model

$$y_{ij} = \mu + v_i + e_{ij},\ i = 1, \ldots, m,\ j = 1, \ldots, n_i. \tag{8.3}$$

Moreover, we relax the normality assumption of the v_i's and e_{ij}'s, and we only assume v_i's are uncorrelated with $\mathrm{E}(v_i) = 0$, $\mathrm{var}(v_i) = \sigma_v^2$, e_{ij}'s are uncorrelated

with $\mathrm{E}(e_{ij}) = 0$, $\mathrm{var}(e_{ij}) = \sigma_e^2$, and v_i's and e_{ij}'s are uncorrelated. In addition to the aforementioned assumptions, for technical reasons, we require for some $c > 0$, $\mathrm{E}(e_{ij}^{8+c}) < \infty$ and $\mathrm{E}(v_i^{8+c}) < \infty$.

For the above model, the BLUP of the i^{th} small area mean $\theta_i = \mu + v_i$ is

$$\hat{\theta}_i(\boldsymbol{\sigma}^2) = \tilde{\mu} + \gamma_i(\bar{y}_i - \tilde{\mu}) \tag{8.4}$$

where $\boldsymbol{\sigma}^2 = (\sigma_v^2, \sigma_e^2)'$, $\gamma_i = \sigma_v^2/(\sigma_v^2 + \sigma_e^2/n_i)$, $\tilde{\mu} = \sum_{i=1}^m \gamma_i \bar{y}_i / \sum_{i=1}^m \gamma_i$ and $\bar{y}_i = (1/n_i)\sum_{j=1}^{n_i} y_{ij}$.

Since the BLUP contains unknown variance components, the EBLUP of θ_i is obtained by plugging in estimators for the unknown variance components in (8.4). That is, the EBLUP of θ_i is given by

$$\hat{\theta}_i(\hat{\boldsymbol{\sigma}}^2) = \hat{\mu} + \hat{\gamma}_i(\bar{y}_i - \hat{\mu}) \tag{8.5}$$

where $\hat{\boldsymbol{\sigma}}^2 = (\hat{\sigma}_v^2, \hat{\sigma}_e^2)'$, $\hat{\gamma}_i$ and $\hat{\mu}$ are the same as γ_i and $\tilde{\mu}$ except that unknown variance components are estimated by the analysis of variance estimators (Searle [44]):

$$\hat{\sigma}_e^2 = \frac{1}{n-m}\sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 = \frac{\mathrm{SSW}}{n-m} = \mathrm{MSW} \tag{8.6}$$

$$\hat{\sigma}_v^2 = \frac{(m-1)}{g}(\mathrm{MSB} - \mathrm{MSW}) \tag{8.7}$$

where $\bar{y}_i = (1/n_i)\sum_{j=1}^{n_i} y_{ij}$, $\bar{y} = (1/n)\sum_{i=1}^m n_i \bar{y}_i$, $n = \sum_{i=1}^m n_i$, $g = n - \sum_{i=1}^m (n_i^2/n)$ and

$$\mathrm{MSB} = \frac{\mathrm{SSB}}{m-1} = \frac{1}{m-1}\sum_{i=1}^m n_i(\bar{y}_i - \bar{y})^2.$$

8.1 Robust MSE approximation

In order to derive an estimator that satisfies (8.2), we need to first approximate the MSE of $\hat{\theta}_i(\hat{\boldsymbol{\sigma}}^2)$. To this end, note that

$$\begin{aligned}\text{MSE}[\hat{\theta}_i(\hat{\boldsymbol{\sigma}}^2)] &= \text{E}[\hat{\theta}_i(\hat{\boldsymbol{\sigma}}^2) - \theta_i]^2 = \text{E}[\hat{\theta}_i(\hat{\boldsymbol{\sigma}}^2) - \hat{\theta}_i(\boldsymbol{\sigma}^2) + \hat{\theta}_i(\boldsymbol{\sigma}^2) - \theta_i]^2 \\ &= \text{MSE}[\hat{\theta}_i(\boldsymbol{\sigma}^2)] + \text{E}[\hat{\theta}_i(\hat{\boldsymbol{\sigma}}^2) - \hat{\theta}_i(\boldsymbol{\sigma}^2)]^2 \\ &\quad + 2\text{E}[\hat{\theta}_i(\hat{\boldsymbol{\sigma}}^2) - \hat{\theta}_i(\boldsymbol{\sigma}^2)][\hat{\theta}_i(\boldsymbol{\sigma}^2) - \mu - v_i] \qquad (8.8)\end{aligned}$$

where

$$\begin{aligned}\text{MSE}[\hat{\theta}_i(\boldsymbol{\sigma}^2)] &= g_{1i}(\boldsymbol{\sigma}^2) + g_{2i}(\boldsymbol{\sigma}^2) \\ &= (1-\gamma_i)\sigma_v^2 + \frac{(1-\gamma_i)^2\sigma_v^2}{\sum_{i=1}^m \gamma_i}. \qquad (8.9)\end{aligned}$$

A second order approximation of the last two terms given in (8.8) is given below. Due to time constraints, in Section 8.6, we only give a sketch of the proof[1], which we note is similar to how Prasad and Rao [36] derived their MSE approximation.

$$\begin{aligned}&E[\hat{\theta}_i(\hat{\boldsymbol{\sigma}}^2) - \hat{\theta}_i(\boldsymbol{\sigma}^2)]^2 \\ &\quad= \frac{1/n_i^2}{(\sigma_e^2/n_i + \sigma_v^2)^3}\text{var}(\sigma_v^2\hat{\sigma}_e^2 - \sigma_e^2\hat{\sigma}_v^2) + o(m^{-1}) \\ &\quad= \frac{1/n_i^2}{(\sigma_e^2/n_i + \sigma_v^2)^3}\Big(\sigma_v^4\text{var}(\hat{\sigma}_e^2) + \sigma_e^4\text{var}(\hat{\sigma}_v^2) - 2\sigma_e^2\sigma_v^2\text{cov}(\hat{\sigma}_v^2, \hat{\sigma}_e^2)\Big) + o(m^{-1}) \qquad (8.10)\end{aligned}$$

[1]In a personal communication by Dr. Lahiri, he mentioned the decomposition of the MSE given in (8.8) and the approximations given in (8.10), (8.11) and that he had a proof for the approximation given in (8.13). The result was independently re-derived by the author.

$$
\begin{aligned}
&2E[\hat{\theta}_i(\hat{\boldsymbol{\sigma}}^2) - \hat{\theta}_i(\boldsymbol{\sigma}^2)][\hat{\theta}_i(\boldsymbol{\sigma}^2) - \mu - v_i] \\
&\quad = \frac{2}{n}\frac{\sigma_e^2/n_i}{(\sigma_e^2/n_i + \sigma_v^2)^3}\Big(\frac{\sigma_v^2}{n_i^2}(\delta_e - 3\sigma_e^4) - \sigma_e^2(\delta_v - 3\sigma_v^4)\Big) \\
&\qquad - \Big(\frac{n_i - 1}{n_i^3}\Big)\frac{1}{(n-m)}\frac{\sigma_v^4}{(\sigma_e^2/n_i + \sigma_v^2)^3}(\delta_e - 3\sigma_e^4) + o(m^{-1}) \\
&\quad = g_{4i}(\boldsymbol{\sigma}^2, \boldsymbol{\delta}) + o(m^{-1}) \qquad (8.11)
\end{aligned}
$$

where $\boldsymbol{\delta} = (\delta_v, \delta_e)'$ are the fourth moments of e, v. In Section 8.6, we give the derivations for $\mathrm{var}(\hat{\sigma}_e^2)$, $\mathrm{var}(\hat{\sigma}_v^2)$ and $\mathrm{cov}(\hat{\sigma}_v^2, \hat{\sigma}_e^2)$. From (8.55)-(8.57), we have

$$
\begin{aligned}
\mathrm{var}(\hat{\sigma}_e^2) &= \frac{1}{(n-m)^2}\Big((n-2m)\delta_e - (n-4m)\sigma_e^4 + \sum_{i=1}^{m}\frac{1}{n_i}(\delta_e - 3\sigma_e^4)\Big) \\
\mathrm{var}(\hat{\sigma}_v^2) &= \frac{1}{n^2}\Big(\sum_{i=1}^{m} n_i^2(\delta_v - \sigma_v^4) + 4n\sigma_v^2\sigma_e^2 + 2m\sigma_e^4 + \sum_{i=1}^{m}\frac{1}{n_i}(\delta_e - 3\sigma_e^4)\Big) \\
&\qquad + \frac{m^2}{n^2(n-m)^2}\Big((n-2m)\delta_e - (n-4m)\sigma_e^4 + \sum_{i=1}^{m}\frac{1}{n_i}(\delta_e - 3\sigma_e^4)\Big) \\
&\qquad - \frac{2m}{n^2(n-m)}\Big(m - \sum_{i=1}^{m}\frac{1}{n_i}\Big)(\delta_e - 3\sigma_e^4) + O(m^{-2}) \\
\mathrm{cov}(\hat{\sigma}_v^2, \hat{\sigma}_e^2) &= \frac{1}{n(n-m)^2}\Big(m^2(\delta_e - \sigma_e^4) - 2mn\sigma_e^4 - n\sum_{i=1}^{m}\frac{1}{n_i}(\delta_e - 3\sigma_e^4)\Big) + O(m^{-2})
\end{aligned}
$$

where δ_e, δ_v are the fourth moments of e, v and $n = \sum_{i=1}^{m} n_i$.

From (8.10) and the above formulas for $\mathrm{var}(\hat{\sigma}_e^2)$, $\mathrm{var}(\hat{\sigma}_v^2)$, $\mathrm{cov}(\hat{\sigma}_v^2, \hat{\sigma}_e^2)$, we can derive an approximation for $E[\hat{\theta}_i(\hat{\boldsymbol{\sigma}}^2) - \hat{\theta}_i(\boldsymbol{\sigma}^2)]^2$ correct upto order $o(m^{-1})$. Denote this approximation by $g_{3i}(\boldsymbol{\sigma}^2, \boldsymbol{\delta})$. That is,

$$
E[\hat{\theta}_i(\hat{\boldsymbol{\sigma}}^2) - \hat{\theta}_i(\boldsymbol{\sigma}^2)]^2 = g_{3i}(\boldsymbol{\sigma}^2, \boldsymbol{\delta}) + o(m^{-1}) \qquad (8.12)
$$

where $g_{3i}(\boldsymbol{\sigma}^2, \boldsymbol{\delta})$ is given by the first term on the right hand side of (8.10) and (8.55)-(8.57).

From (8.8)-(8.12), we obtain the following second order approximation of the MSE of $\hat{\theta}_i(\hat{\sigma}^2)$:

$$MSE[\hat{\theta}_i(\hat{\sigma}^2)] = g_{1i}(\boldsymbol{\sigma}^2) + g_{2i}(\boldsymbol{\sigma}^2) + g_{3i}(\boldsymbol{\sigma}^2, \boldsymbol{\delta}) + g_{4i}(\boldsymbol{\sigma}^2, \boldsymbol{\delta}) + o(m^{-1}). \tag{8.13}$$

Also, under normality of e and v, since $\delta_v = 3\sigma_v^4$, $\delta_e = 3\sigma_e^4$, from (8.11), we have $g_{4i}(\boldsymbol{\sigma}^2, \boldsymbol{\delta}) = 0$. Furthermore, substituting $\delta_v = 3\sigma_v^4$, $\delta_e = 3\sigma_e^4$ in $g_{3i}(\boldsymbol{\sigma}^2, \boldsymbol{\delta})$ and simplifying we get

$$MSE[\hat{\theta}_i(\hat{\sigma}^2)] = g_{1i}(\boldsymbol{\sigma}^2) + g_{2i}(\boldsymbol{\sigma}^2) + g_{3i}^{PR}(\boldsymbol{\sigma}^2) + o(m^{-1}) \tag{8.14}$$

where

$$g_{3i}^{PR}(\boldsymbol{\sigma}^2) = \frac{1/n_i^2}{(\sigma_e^2/n_i + \sigma_v^2)^3}\Big(\frac{2}{n-m}\sigma_v^4\sigma_e^4 + \frac{2}{n^2}\Big\{\frac{nm}{n-m}\sigma_e^4 + 2n\sigma_v^2\sigma_e^2 + \sum_{i=1}^{m} n_i^2\sigma_v^4\Big\}\sigma_e^4 + \frac{4m}{n(n-m)}\sigma_v^2\sigma_e^6\Big). \tag{8.15}$$

Note that $g_{3i}^{PR}(\boldsymbol{\sigma}^2)$ agrees upto terms $O(m^{-1})$ with the similar term derived in Prasad and Rao [36]. Also, under the assumption of normality, Kackar and Harville [22] showed

$$E[\hat{\theta}_i(\hat{\boldsymbol{\sigma}}^2) - \hat{\theta}_i(\boldsymbol{\sigma}^2)][\hat{\theta}_i(\boldsymbol{\sigma}^2) - \mu - v_i] = 0. \tag{8.16}$$

Hence, under normality of e, v, the above term need not be approximated.

8.2 MSE estimators

In this section, we give closed form expressions for the robust MSE estimator, the naive MSE estimator and the Prasad Rao MSE estimator.

8.2.1 Naive MSE estimator

The naive MSE estimator refers to estimating the MSE of the EBLUP by the MSE of the BLUP with estimated variance components $(\hat{\sigma}_v^2, \hat{\sigma}_e^2)$ substituted for (σ_v^2, σ_e^2). That is,

$$\widehat{MSE}_{i,N} \;=\; g_{1i}(\hat{\boldsymbol{\sigma}}^2) + g_{2i}(\hat{\boldsymbol{\sigma}}^2) \tag{8.17}$$

8.2.2 Robust MSE estimator

Since $g_{2i}(\boldsymbol{\sigma}^2)$, $g_{3i}(\boldsymbol{\sigma}^2, \boldsymbol{\delta})$ and $g_{4i}(\boldsymbol{\sigma}^2, \boldsymbol{\delta})$ are of order $O(m^{-1})$, second order unbiased estimators of these terms are given by the plug-in estimators. However, $g_{1i}(\boldsymbol{\sigma}^2)$ is of order $O(1)$, and since

$$E(g_{1i}(\hat{\boldsymbol{\sigma}}^2)) = g_{1i}(\boldsymbol{\sigma}^2) - g_{3i}(\boldsymbol{\sigma}^2, \boldsymbol{\delta}) + o(1/m),$$

we obtain the following robust second order unbiased MSE estimator:

$$\widehat{MSE}_{i,prop} = g_{1i}(\hat{\boldsymbol{\sigma}}^2) + g_{2i}(\hat{\boldsymbol{\sigma}}^2) + 2g_{3i}(\hat{\boldsymbol{\sigma}}^2, \hat{\boldsymbol{\delta}}) + g_{4i}(\hat{\boldsymbol{\sigma}}^2, \hat{\boldsymbol{\delta}}) \tag{8.18}$$

where $\hat{\boldsymbol{\delta}} = (\hat{\delta}_v, \hat{\delta}_e)'$ is the fourth moment estimator of v and e. In Section 8.6, we give a brief sketch justifying the second order unbiasedness of the robust MSE estimator given by (8.18). Following Hall and Maiti [20], we use the following moment based estimators for the fourth moments of e and v:

$$\hat{\delta}_e = \max\Big(\frac{1}{2} \frac{1}{\sum_{i=1}^m n_i(n_i-1)} \sum_{i=1}^m \sum_{\substack{j_1,j_2=1 \\ j_1 \neq j_2}}^{n_i} (y_{ij_1} - y_{ij_2})^4 - 3\hat{\sigma}_e^4,\ \hat{\sigma}_e^4 \Big)$$

$$\hat{\delta}_v = \max\Big(\frac{1}{n} \sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \hat{\mu})^4 - 6\hat{\sigma}_e^2\hat{\sigma}_v^2 - \hat{\delta}_e,\ \hat{\sigma}_v^4 \Big)$$

where $\hat{\mu} = \sum_{i=1}^m \hat{\gamma}_i \bar{y}_i / \sum_{i=1}^m \hat{\gamma}_i$ and $\hat{\gamma}_i = \hat{\sigma}_v^2/(\hat{\sigma}_v^2 + \hat{\sigma}_e^2/n_i)$.

8.2.3 Prasad Rao MSE estimator

The Prasad Rao MSE estimator can be derived from (8.18) by taking $\hat{\delta}_e = 3\hat{\sigma}_e^4$ and $\hat{\delta}_v = 3\hat{\sigma}_v^4$, and is given by

$$\widehat{MSE}_{i,PR} = g_{1i}(\hat{\boldsymbol{\sigma}}^2) + g_{2i}(\hat{\boldsymbol{\sigma}}^2) + 2g_{3i}^{PR}(\hat{\boldsymbol{\sigma}}^2) \tag{8.19}$$

where $g_{3i}^{PR}(\boldsymbol{\sigma}^2)$ is given by (8.15).

8.3 Simulation results for unbalanced case

For the simulation study both e, v were generated from either shifted exponential or double exponential, $m = 30$, with 5 areas each having $n_i = 2$, 3, 4, 5, 6, $\sigma_v^2 = 1$, $\sigma_e^2 = 0.5$, 4. To compare the previously mentioned MSE estimators, for each set of distributions for e,v and for each value of σ_e^2, 10000 independent samples were generated and the percent relative bias of each MSE estimator was evaluated. The percent relative bias of each MSE estimator was defined to be the average over all areas with the same n_i of

$$RB_i = 100 \cdot \frac{E(\widehat{MSE}_i) - MSE_i}{MSE_i}$$

where the expectation of the MSE estimator for the i^{th} area, $E(\widehat{MSE}_i)$, and the true MSE of the EBLUP for the i^{th} area, MSE_i, were estimated empirically.

In the tables that are provided, 'NAIVE' denotes the naive MSE estimator given by (8.17), 'prop' denotes the robust MSE estimator given by (8.18) and 'PR' denotes the Prasad Rao MSE estimator given by (8.19). Simulations indicate that when $\sigma_e^2/\sigma_v^2 = 0.5$, the Prasad Rao and the robust MSE estimators perform quite

Table 8.1: Percent relative bias of MSE estimators, $m = 30$, $\sigma_v^2 = 1$, $\sigma_e^2 = 0.5$.

	$e, v \sim$ Double Exponential					$e, v \sim$ Shifted Exponential				
n_i	2	3	4	5	6	2	3	4	5	6
NAIVE	-6.54	-3.83	-2.98	-1.77	-1.22	-7.36	-3.99	-3.45	-1.62	-1.42
prop	-1.21	0.26	0.26	0.89	1.00	-0.81	0.90	0.30	1.39	0.99
PR	-2.42	-0.38	-0.05	0.78	1.04	-3.10	-0.31	-0.28	1.21	1.11

Table 8.2: Percent relative bias of MSE approximations, $m = 30$, $\sigma_v^2 = 1$, $\sigma_e^2 = 0.5$.

	$e, v \sim$ Double Exponential					$e, v \sim$ Shifted Exponential				
n_i	2	3	4	5	6	2	3	4	5	6
prop	-0.25	0.38	-0.08	0.25	0.16	0.97	1.08	-0.43	0.07	-0.71
PR	-0.67	0.87	0.90	1.54	1.67	0.18	2.11	1.58	2.73	2.36

well. For example, from Table 8.1, for both sets of distributions and for all values of n_i, the robust MSE estimator has relative bias less that 1.5% while the Prasad Rao MSE estimator has relative bias less than 3%. In this case, even the naive MSE estimator for large n_i (=5, 6) has relative bias of under 2%. Moreover, in this case, from Table 8.2 we have that the robust MSE approximation has relative bias less than 1%. However, even the Prasad Rao MSE approximation given by (8.14)-(8.15) has small relative bias - less than 3% - this is due to a "canceling off" effect. What we mean, is in several settings that we considered (not reported here), the term $E[\hat{\theta}_i(\hat{\boldsymbol{\sigma}}^2) - \hat{\theta}_i(\boldsymbol{\sigma}^2)][\hat{\theta}_i(\boldsymbol{\sigma}^2) - \mu - v_i]$ in (8.8) was always negative and hence, $g_{4i}(\boldsymbol{\sigma}^2, \boldsymbol{\delta})$ was also negative. Moreover, $g_{3i}^{PR}(\boldsymbol{\sigma}^2)$ for various non-normal settings was smaller than the term $\mathrm{E}[\hat{\theta}_i(\hat{\boldsymbol{\sigma}}^2) - \hat{\theta}_i(\boldsymbol{\sigma}^2)]^2$. That is, the Prasad Rao MSE approximation does well in certain settings because $g_{3i}^{PR}(\boldsymbol{\sigma}^2)$ is smaller than $\mathrm{E}[\hat{\theta}_i(\hat{\boldsymbol{\sigma}}^2) - \hat{\theta}_i(\boldsymbol{\sigma}^2)]^2$ and the Prasad Rao MSE approximation assumes $E[\hat{\theta}_i(\hat{\boldsymbol{\sigma}}^2) - \hat{\theta}_i(\boldsymbol{\sigma}^2)][\hat{\theta}_i(\boldsymbol{\sigma}^2) - \mu - v_i] = 0$, which is negative in most settings.

When $\sigma_e^2/\sigma_v^2 = 4$, the relative bias of the Prasad Rao MSE estimator increases significantly with n_i, but for the robust MSE estimator it decreases with n_i (Tables 8.3 and 8.4). This result can be partially explained by looking at the relative bias of the robust and the Prasad Rao MSE approximations (Table 8.5). From Table 8.5, when $\sigma_e^2/\sigma_v^2 = 4$, $e, v \sim$ double exponential and n_i increases from 2 to 6, the relative bias of the robust MSE approximation decreases from 3.92% to 0.23%. When $\sigma_e^2/\sigma_v^2 = 4$, $e, v \sim$ shifted exponential and n_i increases from 2 to 6, the relative bias of the robust MSE approximation decreases from 4.88% to -1.03%. In contrast, when $e, v \sim$ double exponential, the relative bias of the Prasad Rao MSE approximation

Table 8.3: Percent relative bias of MSE estimators, $m = 30$, $\sigma_v^2 = 1$, $\sigma_e^2 = 4$, $e, v \sim$ Double Exponential.

n_i	2	3	4	5	6
NAIVE	-15.66	-17.71	-16.37	-16.49	-15.66
prop	11.89	9.28	8.70	4.64	0.61
PR	-1.30	-0.91	2.83	4.46	7.13

increases from 1.24% to 3.91%, and when $e, v \sim$ shifted exponential, the relative bias of the Prasad Rao MSE approximation increases from -0.18% to 6.79%.

It is difficult to give a general statement as to when the robust MSE approximation will do better than the Prasad Rao MSE approximation. In a future study, we will consider a more exhaustive simulation design for different parameter combinations. However, from different combinations of σ_e^2/σ_v^2 that we have tried (not all reported here), we draw the following conclusions: when σ_e^2/σ_v^2 is small (less than 2), the Prasad Rao and robust MSE approximations perform well. However, when σ_e^2/σ_v^2 is large, the robust MSE approximation does poorly for small n_i, but does exceedingly well for large n_i. The Prasad Rao MSE approximation does poorly for large n_i but well for small n_i. This needs to be investigated further.

Moreover, we note that compared to the Prasad Rao MSE estimator, the robust MSE estimator has a much larger mean squared error (not reported here). The much larger variability in the robust MSE estimator is due to estimation of fourth moments, in particular the estimation of the fourth moment of v.

Table 8.4: Percent relative bias of MSE estimators, $m = 30$, $\sigma_v^2 = 1$, $\sigma_e^2 = 4$, $e, v \sim$ Shifted Exponential.

n_i	2	3	4	5	6
NAIVE	-18.54	-17.71	-16.68	-16.12	-15.43
prop	14.26	14.40	12.02	7.18	1.16
PR	-4.50	-0.48	3.24	6.12	9.07

Table 8.5: Percent relative bias of MSE approximations, $m = 30$, $\sigma_v^2 = 1$, $\sigma_e^2 = 4$.

	$e, v \sim$ Double Exponential					$e, v \sim$ Shifted Exponential				
n_i	2	3	4	5	6	2	3	4	5	6
prop	3.92	1.61	2.32	0.78	0.23	4.88	4.52	3.26	1.04	-1.03
PR	1.24	0.81	3.19	3.13	3.91	-0.18	3.13	5.30	6.14	6.79

8.4 Balanced case

For the model given by (8.3), the balanced case refers to when we have $n_i = k$ for all i. In Section 8.6, for the balanced case, we derive the corresponding terms for $g_{3i}(\boldsymbol{\sigma}^2, \boldsymbol{\delta})$ and $g_{4i}(\boldsymbol{\sigma}^2, \boldsymbol{\delta})$. We drop the subscript i as in the balanced case $g_{3i}(\boldsymbol{\sigma}^2, \boldsymbol{\delta})$ and $g_{4i}(\boldsymbol{\sigma}^2, \boldsymbol{\delta})$ do not depend on i.

$$g_3(\boldsymbol{\sigma}^2, \boldsymbol{\delta}) = \frac{1}{m} \frac{1}{(\sigma_e^2 + k\sigma_v^2)^3} \Big(\delta_e \sigma_v^4 + k\delta_v \sigma_e^4 + \frac{2\sigma_e^8}{k-1} - \Big(\frac{k^2-3}{k-1} \Big) \sigma_e^4 \sigma_v^4 + \frac{4k\sigma_e^6 \sigma_v^2}{k-1} \Big) \quad (8.20)$$

$$g_4(\boldsymbol{\sigma}^2, \boldsymbol{\delta}) = \frac{2}{m} \frac{1}{(\sigma_e^2 + k\sigma_v^2)^3} \Big(\frac{1}{k} \sigma_e^2 \sigma_v^2 (\delta_e - 3\sigma_e^4) - \sigma_v^4 (\delta_e - 3\sigma_e^4) - k\sigma_e^4 (\delta_v - 3\sigma_v^4) \Big) \quad (8.21)$$

From (8.18), for the balanced case we have the following robust MSE estimator:

$$\widehat{MSE}_{prop} \;=\; g_1(\hat{\boldsymbol{\sigma}}^2) + g_2(\hat{\boldsymbol{\sigma}}^2) + 2g_3(\hat{\boldsymbol{\sigma}}^2, \hat{\boldsymbol{\delta}}) + g_4(\hat{\boldsymbol{\sigma}}^2, \hat{\boldsymbol{\delta}}). \tag{8.22}$$

However, note that by the derivations we have for $g_3(\boldsymbol{\sigma}^2, \boldsymbol{\delta})$ and $g_4(\boldsymbol{\sigma}^2, \boldsymbol{\delta})$ given in (8.20)-(8.21), it follows that

$$\begin{aligned} 2g_3(\hat{\boldsymbol{\sigma}}^2, \hat{\boldsymbol{\delta}}) + g_4(\hat{\boldsymbol{\sigma}}^2, \hat{\boldsymbol{\delta}}) &= \frac{2}{m}\frac{1}{(\hat{\sigma}_e^2 + k\hat{\sigma}_v^2)^3}\Big(\hat{\delta}_e\hat{\sigma}_v^4 + k\hat{\delta}_v\hat{\sigma}_e^4 + \frac{2\hat{\sigma}_e^8}{k-1} - \frac{(k^2-3)\hat{\sigma}_e^4\hat{\sigma}_v^4}{k-1} \\ &\quad \frac{4k\hat{\sigma}_e^6\hat{\sigma}_v^2}{k-1} + \frac{1}{k}\hat{\sigma}_e^2\hat{\sigma}_v^2(\hat{\delta}_e - 3\hat{\sigma}_e^4) - \hat{\sigma}_v^4(\hat{\delta}_e - 3\hat{\sigma}_e^4) - k\hat{\sigma}_e^4(\hat{\delta}_v - 3\hat{\sigma}_v^4)\Big) \\ &= \frac{2}{m}\frac{1}{(\hat{\sigma}_e^2 + k\hat{\sigma}_v^2)^3}\Big(\frac{2\hat{\sigma}_e^8}{k-1} + \frac{2k^2}{k-1}\hat{\sigma}_v^4\hat{\sigma}_e^4 + \Big(\frac{4k}{k-1} - \frac{3}{k}\Big)\hat{\sigma}_e^6\hat{\sigma}_v^2 + \frac{\hat{\sigma}_e^2\hat{\sigma}_v^2}{k}\hat{\delta}_e\Big). \end{aligned} \tag{8.23}$$

1. By (8.23), we have shown for the balanced case, the robust MSE estimator given by (8.22) does not depend on the estimated fourth moment of v_i.

2. Note that since the robust MSE estimator does not involve the estimated fourth moment of v_i, it follows that, under the assumption e_{ij} are normally distributed, the Prasad Rao MSE estimator is second order unbiased even if v_i are not normally distributed.

   To derive the Prasad Rao MSE estimator, in (8.22) we simply take $\hat{\delta}_e = 3\hat{\sigma}_e^4$.

$$\widehat{MSE}_{PR} \;=\; g_1(\hat{\boldsymbol{\sigma}}^2) + g_2(\hat{\boldsymbol{\sigma}}^2) + 2g_3^{PR}(\hat{\boldsymbol{\sigma}}^2) \tag{8.24}$$

where $2g_3^{PR}(\hat{\boldsymbol{\sigma}}^2)$ is given by substituting $\hat{\delta}_e = 3\hat{\sigma}_e^4$ in (8.23). That is,

$$2g_3^{PR}(\hat{\boldsymbol{\sigma}}^2) = \frac{2}{m}\frac{1}{(\hat{\sigma}_e^2 + k\hat{\sigma}_v^2)^3}\frac{1}{(k-1)}\Big(2\hat{\sigma}_e^8 + 2k^2\hat{\sigma}_v^4\hat{\sigma}_e^4 + 4k\hat{\sigma}_e^6\hat{\sigma}_v^2\Big).$$

From (8.23), if $\hat{\delta}_e > 3$, then $\widehat{MSE}_{prop} > \widehat{MSE}_{PR}$. Hence, if $\widehat{MSE}_{PR}$ is overestimating the true MSE, so will $\widehat{MSE}_{prop}$ (see Tables 8.6 and 8.7).

8.5 Simulation results for balanced case

To compare the above MSE estimators, 10000 independent samples were generated and the percent relative bias of each MSE estimator was evaluated. The percent relative bias of each MSE estimator was defined to be the average over all small areas of

$$RB_i = 100 \cdot \frac{E(\widehat{MSE_i}) - MSE_i}{MSE_i} \tag{8.25}$$

where once again the expectation of the MSE estimator for the i^{th} area, $E(\widehat{MSE_i})$, and the true MSE of the EBLUP for the i^{th} area, MSE_i, were estimated empirically. In the simulation, both e, v were generated from either shifted exponential or double exponential, k was either 3 or 6, $m = 30$, $\sigma_v^2 = 1$, $\sigma_e^2 = 0.5$, 1, 2, 4.

From Tables 8.6 and 8.7, for all simulated values of σ_e^2, increasing k from 3 to 6 reduces the relative bias of all MSE estimators. This is due to Prasad Rao and robust MSE approximation performing better in terms of relative bias when k is increased. For example, from Table 8.8, when $k = 3$, $e, v \sim$ shifted exponential and σ_e^2 varies from 0.5 to 4, the Prasad Rao MSE approximation and the robust MSE approximation have relative bias that varies from 2.71% to 7.62% and -0.34% to 5.36%. When k is increased to 6 (Table 8.9), the relative bias of the Prasad Rao MSE approximation varies from 1.42% to 4.01%, and the robust MSE approximation has negligible bias.

Moreover, the robust MSE approximation performs slightly better than the Prasad Rao MSE approximation for small values σ_e^2/σ_v^2, but performs as badly when σ_e^2/σ_v^2 is large and k is small (that is, $\sigma_e^2/\sigma_v^2 = 4$ and $k = 3$). As mentioned

Table 8.6: Percent relative bias of MSE estimators, $m = 30$, $k = 3$, $\sigma_v^2 = 1$.

	$e, v \sim$ Double Exponential				$e, v \sim$ Shifted Exponential			
σ_e^2	0.5	1	2	4	0.5	1	2	4
NAIVE	-3.87	-7.19	-13.17	-19.57	-3.77	-7.44	-13.37	-19.16
prop	0.32	0.94	2.19	7.35	1.00	1.67	3.67	10.21
PR	0.24	0.68	1.72	6.87	0.70	0.96	2.41	8.69

Table 8.7: Percent relative bias of MSE estimators, $m = 30$, $k = 6$, $\sigma_v^2 = 1$.

	$e, v \sim$ Double Exponential				$e, v \sim$ Shifted Exponential			
σ_e^2	0.5	1	2	4	0.5	1	2	4
NAIVE	-1.60	-3.51	-6.08	-10.44	-1.71	-3.22	-6.12	-10.88
prop	0.03	-0.33	0.12	1.47	0.08	0.35	0.72	1.90
PR	0.04	-0.34	0.05	1.33	0.06	0.24	0.42	1.33

earlier, in every case we considered whenever the Prasad Rao estimator has positive relative bias, the robust MSE estimator has larger relative bias than the Prasad Rao MSE estimator.

Table 8.8: Percent relative bias of MSE approximations, $m = 30$, $k = 3$, $\sigma_v^2 = 1$.

	$e, v \sim$ Double Exponential				$e, v \sim$ Shifted Exponential			
σ_e^2	0.5	1	2	4	0.5	1	2	4
prop	-0.53	0.28	0.41	2.54	-0.34	-0.29	0.64	5.36
PR	0.99	2.39	2.56	3.77	2.71	3.58	4.89	7.62

Table 8.9: Percent relative bias of MSE approximations, $m = 30$, $k = 6$, $\sigma_v^2 = 1$.

	$e, v \sim$ Double Exponential				$e, v \sim$ Shifted Exponential			
σ_e^2	0.5	1	2	4	0.5	1	2	4
prop	-0.20	-0.87	-0.93	-0.02	-0.22	-0.33	-0.77	-0.51
PR	0.63	0.51	1.07	2.28	1.42	2.44	3.25	4.01

8.6 Appendix

Notation:

$$\bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij} \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{m} n_i \bar{y}_i$$

$$\bar{e}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} e_{ij} \qquad \bar{e} = \frac{1}{n}\sum_{i=1}^{m} n_i \bar{e}_i$$

$$n = \sum_{i=1}^{m} n_i \qquad g = n - \sum_{i=1}^{m} \frac{n_i^2}{n}$$

$$\gamma_i = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_e^2/n_i} \qquad \tilde{\mu} = \frac{\sum_{i=1}^m \gamma_i \bar{y}_i}{\sum_{i=1}^m \gamma_i}$$

$$\hat{\sigma}_e^2 = \frac{1}{n-m}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 = \frac{\text{SSW}}{n-m} = \text{MSW}$$

$$\hat{\sigma}_v^2 = \frac{(m-1)}{g}(\text{MSB} - \text{MSW})$$

$$\text{MSB} = \frac{\text{SSB}}{m-1} = \frac{1}{m-1}\sum_{i=1}^{m} n_i(\bar{y}_i - \bar{y})^2.$$

Brief sketch on approximating $\text{E}[\hat{\theta}_i(\hat{\boldsymbol{\sigma}}^2) - \hat{\theta}_i(\boldsymbol{\sigma}^2)][\hat{\theta}_i(\boldsymbol{\sigma}^2) - \mu - v_i]$ and $\text{E}[\hat{\theta}_i(\hat{\boldsymbol{\sigma}}^2) - \hat{\theta}_i(\boldsymbol{\sigma}^2)]^2$. We first expand $\hat{\theta}_i(\hat{\boldsymbol{\sigma}}^2)$ around $\boldsymbol{\sigma}^2$.

$$\hat{\theta}_i(\hat{\boldsymbol{\sigma}}^2) = \hat{\theta}_i(\boldsymbol{\sigma}^2) + \frac{\partial\hat{\theta}_i(\boldsymbol{\sigma}^2)}{\partial\sigma_e^2}(\hat{\sigma}_e^2 - \sigma_e^2) + \frac{\partial\hat{\theta}_i(\boldsymbol{\sigma}^2)}{\partial\sigma_v^2}(\hat{\sigma}_v^2 - \sigma_v^2) + R_m$$

where R_m are the remainder terms. Since $\hat{\theta}_i(\boldsymbol{\sigma}^2) = \tilde{\mu} + \gamma_i(\bar{y}_i - \tilde{\mu}) = \sum_{j=1}^m b_j \bar{y}_j$, we have $\hat{\theta}_i(\boldsymbol{\sigma}^2)$ is linear in $\bar{y}_j$. In particular, the weights b_j attached to $\bar{y}_j$ sum to one. Moreover, we have

$$\frac{\partial\hat{\theta}_i(\boldsymbol{\sigma}^2)}{\partial\sigma_e^2} = \sum_{j=1}^{m} a_j \bar{y}_j \tag{8.26}$$

where $a_j = \frac{\partial b_j}{\partial\sigma_e^2}$, $\sum_{j=1}^m a_j = 0$. So, for example, in order to approximate

$\mathrm{E}[\hat{\theta}_i(\hat{\boldsymbol{\sigma}}^2) - \hat{\theta}_i(\boldsymbol{\sigma}^2)][\hat{\theta}_i(\boldsymbol{\sigma}^2) - \mu - v_i]$, we express $[\hat{\theta}_i(\hat{\boldsymbol{\sigma}}^2) - \hat{\theta}_i(\boldsymbol{\sigma}^2)][\hat{\theta}_i(\boldsymbol{\sigma}^2) - \mu - v_i]$ as

$$
\begin{aligned}
&[\hat{\theta}_i(\hat{\boldsymbol{\sigma}}^2) - \hat{\theta}_i(\boldsymbol{\sigma}^2)][\hat{\theta}_i(\boldsymbol{\sigma}^2) - \mu - v_i] \\
&= \Big(\frac{\partial \hat{\theta}_i(\boldsymbol{\sigma}^2)}{\partial \sigma_e^2}(\hat{\sigma}_e^2 - \sigma_e^2) + \frac{\partial \hat{\theta}_i(\boldsymbol{\sigma}^2)}{\partial \sigma_v^2}(\hat{\sigma}_v^2 - \sigma_v^2) + R_m\Big)(\hat{\theta}_i(\boldsymbol{\sigma}^2) - \mu - v_i) \qquad (8.27)
\end{aligned}
$$

By (8.26) we can express (8.27) as

$$
\begin{aligned}
&[\hat{\theta}_i(\hat{\boldsymbol{\sigma}}^2) - \hat{\theta}_i(\boldsymbol{\sigma}^2)][\hat{\theta}_i(\boldsymbol{\sigma}^2) - \mu - v_i] \\
&= \Big(\sum_{j=1}^{m} a_j \bar{y}_j(\hat{\sigma}_e^2 - \sigma_e^2) + \sum_{j=1}^{m} \tilde{a}_j \bar{y}_j(\hat{\sigma}_v^2 - \sigma_v^2) + R_m\Big)(\sum_{j=1}^{m} b_j \bar{y}_j - \mu - v_i)
\end{aligned}
$$

Now it has to be shown that the expectation for terms involving R_m is of the order $o(m^{-1})$. And for the terms that do not involve R_m by noting that both $\hat{\sigma}_e^2$ and $\hat{\sigma}_v^2$ are quadratic in y_{ij} and by expanding terms and taking expectations we get (8.11). The argument is similar but a lot lengthier when it comes to approximating $\mathrm{E}[\hat{\theta}_i(\hat{\boldsymbol{\sigma}}^2) - \hat{\theta}_i(\boldsymbol{\sigma}^2)]^2$.

In order to derive $\mathrm{var}(\hat{\sigma}_e^2)$, $\mathrm{var}(\hat{\sigma}_v^2)$ and $\mathrm{cov}(\hat{\sigma}_e^2, \hat{\sigma}_v^2)$, we first derive var(SSW), var(SSB) and cov(SSW,SSB).

Derivation of var(SSW).

$$
\begin{aligned}
var[SSW] &= var\Big[\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2\Big] = var\Big[\sum_{i=1}^{m}\sum_{j=1}^{n_i}(e_{ij} - \bar{e}_i)^2\Big] \\
&= \sum_{i=1}^{m} var\Big[\sum_{j=1}^{n_i}(e_{ij} - \bar{e}_i)^2\Big] \\
&= \sum_{i=1}^{m}\Big(E\Big[\sum_{j=1}^{n_i}(e_{ij} - \bar{e}_i)^2\Big]^2 - \Big[E\Big(\sum_{j=1}^{n_i}(e_{ij} - \bar{e}_i)^2\Big)\Big]^2\Big). \qquad (8.28)
\end{aligned}
$$

First we compute $E(e_{ij} - \bar{e}_i)^2$.

$$
\begin{aligned}
E(e_{ij} - \bar{e}_i)^2 &= E[e_{ij}^2 - 2e_{ij}\bar{e}_i + {\bar{e}_i}^2] \\
&= \sigma_e^2 - \frac{2}{n_i}\sigma_e^2 + \frac{1}{{n_i}^2} n_i \sigma_e^2 = \Big(1 - \frac{1}{n_i}\Big)\sigma_e^2. \qquad (8.29)
\end{aligned}
$$

Next we expand the first term in (8.28), and compute $E(e_{ij} - \bar{e_i})^4$.

$$\begin{aligned} E(e_{ij} - \bar{e_i})^4 &= E[e_{ij}^4 - 4e_{ij}^3\bar{e_i} + 6e_{ij}^2\bar{e_i}^2 - 4e_{ij}\bar{e_i}^3 + \bar{e_i}^4] \\ &= \delta_e - \frac{4}{n_i}\delta_e + \frac{6}{n_i^2}[(n_i-1)\sigma_e^4 + \delta_e] - \frac{4}{n_i^3}[\delta_e + 3(n_i-1)\sigma_e^4] + \\ &\quad + \frac{1}{n_i^4}[n_i\delta_e + 3n_i(n_i-1)\sigma_e^4]. \end{aligned} \tag{8.30}$$

Next we compute the cross term when expanding the first term in (8.28).

$$\begin{aligned} E(e_{ij} - \bar{e_i})^2(e_{ik} - \bar{e_i})^2 &= E[e_{ij}^2 - 2e_{ij}\bar{e_i} + \bar{e_i}^2][e_{ik}^2 - 2e_{ik}\bar{e_i} + \bar{e_i}^2] \\ &= \sigma_e^4 - \frac{2}{n_i}\sigma_e^4 + \frac{1}{n_i^2}[\delta_e + (n_i-1)\sigma_e^4] - \frac{2}{n_i}\sigma_e^4 + \frac{8}{n_i^2}\sigma_e^4 \\ &\quad - \frac{2}{n_i^3}[\delta_e + 3(n_i-1)\sigma_e^4] + \frac{1}{n_i^2}[\delta_e + (n_i-1)\sigma_e^4] \\ &\quad - \frac{2}{n_i^3}[\delta_e + 3(n_i-1)\sigma_e^4] + \frac{1}{n_i^4}[n_i\delta_e + 3n_i(n_i-1)\sigma_e^4] \end{aligned} \tag{8.31}$$

From (8.30) and (8.31) we get the first term in (8.28).

$$\begin{aligned} E\Big[\sum_{j=1}^{n_i}(e_{ij} - \bar{e_i})^2\Big]^2 &= \Big(n_i - 4 + \frac{6}{n_i} - \frac{3}{n_i^2}\Big)\delta_e + \Big[\frac{6(n_i-1)}{n_i} - \frac{12(n_i-1)}{n_i^2} + \frac{3(n_i-1)}{n_i^2}\Big]\sigma_e^4 \\ &\quad + n_i(n_i-1)\Big[\Big(\frac{2}{n_i^2} - \frac{4}{n_i^3} + \frac{1}{n_i^3}\Big)\delta_e + \Big(1 - \frac{4}{n_i} + \frac{n_i-1}{n_i^2} + \frac{8}{n_i^2} - \frac{6(n_i-1)}{n_i^3} + (\frac{n_i-1}{n_i^2}\Big) \\ &\quad - \frac{6(n_i-1)}{n_i^3} + \frac{3(n_i-1)}{n_i^3}\Big)\sigma_e^4\Big] \\ &= \Big(n_i - 2 + \frac{1}{n_i}\Big)\delta_e + \Big(7 - \frac{14}{n_i} + \frac{12}{n_i^3}\Big)\sigma_e^4. \end{aligned} \tag{8.32}$$

From (8.29) and (8.32) we get

$$\begin{aligned} var\Big[\sum_{j=1}^{n_i}(e_{ij} - \bar{e_i})^2\Big] &= \Big(n_i - 2 + \frac{1}{n_i}\Big)\delta_e + \Big(n_i^2 - 3n_i + 5 - \frac{3}{n_i}\Big)\sigma_e^4 - n_i^2\Big(1 - \frac{1}{n_i}\Big)^2\sigma_e^4 \\ &= \Big(n_i - 2 + \frac{1}{n_i}\Big)\delta_e + \Big(-n_i + 4 - \frac{3}{n_i}\Big)\sigma_e^4. \end{aligned} \tag{8.33}$$

Finally, from (8.33), we get (8.28). That is

$$
\begin{aligned}
var[SSW] &= \sum_{i=1}^{m}\Big(n_i - 2 + \frac{1}{n_i}\Big)\delta_e + \sum_{i=1}^{m}\Big(-n_i + 4 - \frac{3}{n_i}\Big)\sigma_e^4 \\
&= \Big(n - 2m + \sum_{i=1}^{m}\frac{1}{n_i}\Big)\delta_e + \Big(-n + 4m - \sum_{i=1}^{m}\frac{3}{n_i}\Big)\sigma_e^4 \qquad (8.34)
\end{aligned}
$$

Derivation of var(SSB)

$$
\begin{aligned}
var[SSB] &= var\Big[\sum_{i=1}^{m} n_i(\bar{y}_i - \bar{y})^2\Big] \\
&= var\Big[\sum_{i=1}^{m} n_i\Big(v_i - \frac{\sum n_j v_j}{n} + \bar{e}_i - \bar{e}\Big)^2\Big] \\
&= E\big[\sum_{i=1}^{m} n_i(v_i - \frac{\sum n_j v_j}{n} + \bar{e}_i - \bar{e})^2\,\big]^2 - \big[E[SSB]\big]^2. \qquad (8.35)
\end{aligned}
$$

We first compute $E[SSB]$.

$$
\begin{aligned}
E[SSB] &= E\Big[\sum_{i=1}^{m} n_i[v_i - \frac{\sum_j n_j v_j}{n} + \bar{e}_i - \bar{e}]^2\Big] \\
&= \sum_{i=1}^{m} n_i E\Big[\Big(v_i - \frac{\sum_j n_j v_j}{n}\Big)^2 + (\bar{e}_i - \bar{e})^2\Big] \\
&= \sum_{i=1}^{m} n_i\big[\sigma_v^2 + \frac{\sum n_i^2}{n^2}\sigma_v^2 - 2\frac{n_i}{n}\sigma_v^2 + \frac{1}{n_i}\sigma_e^2 + \frac{1}{n}\sigma_e^2 - \frac{2}{n_i n}n_i\sigma_e^2\,\big] \\
&= \Big(n - \frac{\sum_i n_i^2}{n}\Big)\sigma_v^2 + (m-1)\sigma_e^2 \qquad (8.36)
\end{aligned}
$$

We expand the first term of (8.35).

$$
\begin{aligned}
&E\big[n_i^2[v_i - \frac{\sum_j n_j v_j}{n} + \bar{e_j} - \bar{e}]^4\big] \\
&\quad = n_i^2\big[\, E\Big[\Big(v_i - \frac{\sum_j n_j v_j}{n}\Big)^4 + 6\Big(v_i - \frac{\sum_j n_j v_j}{n}\Big)^2(\bar{e_i} - \bar{e})^2 + (\bar{e_i} - \bar{e})^4\Big]\,\big] \\
&\quad = n_i^2\big[\, E\Big[v_i^4 - 4v_i^3\frac{\sum_j n_j v_j}{n} + 6v_i^2\frac{(\sum_j n_j v_j)^2}{n^2} - 4\frac{v_i(\sum_j n_j v_j)^3}{n^3} + \frac{(\sum_j n_j v_j)^4}{n^4}\Big] \\
&\qquad + 6\Big(\sigma_v^2 + \frac{\sum_j n_j^2\sigma_v^2}{n^2} - \frac{2n_i}{n}\sigma_v^2\Big)\Big(\frac{1}{n_i}\sigma_e^2 - \frac{1}{n}\sigma_e^2\Big) \\
&\qquad + E[\bar{e_i}^4 - 4\bar{e_i}^3\bar{e} + 6\bar{e_i}^2\bar{e}^2 - 4\bar{e_i}\bar{e}^3 + \bar{e}^4]\big] \\
&\quad = n_i^2\big[\delta_v - 4\frac{n_i}{n}\delta_v + 6\frac{n_i^2}{n^2}\delta_v + \frac{6}{n^2}\sum_{j\neq i} n_j^2\sigma_v^4 - \frac{4n_i^3}{n^3}\delta_v - \frac{12n_i}{n^3}\sum_{j\neq i} n_j\sigma_v^4 \\
&\quad + \frac{\sum_j n_j^4}{n^4}\delta_v + 3\frac{\sum_{j\neq k} n_j^2 n_k^2}{n^4}\sigma_v^4\big] + 6(\sigma_v^2 + \sigma_v^2\frac{\sum_j n_j^2}{n^2} - \frac{2n_i}{n}\sigma_v^2)(\frac{1}{n_i} - \frac{1}{n})\sigma_e^2 \\
&\quad + \frac{1}{n_i^4}[n_i\delta_e + 3n_i(n_i-1)\sigma_e^4] - \frac{4}{n_i^3 n}[n_i\delta_e + 3n_i(n_i-1)\sigma_e^4] \\
&\quad + \frac{6}{n_i^2 n^2}[n_i\delta_e + n_i(n-1)\sigma_e^4 + n_i(n_i-1)\sigma_e^4] - \frac{4}{n_i n^3}[n_i\delta_e + 3n_i(n-1)\sigma_e^4] \\
&\quad + \frac{1}{n^4}[n\delta_e + 3n(n-1)\sigma_e^4]. \qquad (8.37)
\end{aligned}
$$

From (8.37) we get

$$
\begin{aligned}
&E\Big[\sum_{i=1}^m n_i^2[v_i - \frac{\sum_j n_j v_j}{n} + \bar{e_j} - \bar{e}]^4\Big] \\
&\qquad = \sum_{i=1}^m n_i^2\delta_v + 6n\sigma_e^2\sigma_v^2 + \sum_{i=1}^m \frac{1}{n_i}\delta_e + \sum_{i=1}^m \frac{3(n_i-1)}{n_i}\sigma_e^4 + O(1). \qquad (8.38)
\end{aligned}
$$

When expanding the first term of (8.35), the cross term is

$$
\begin{aligned}
&E\Big[n_i n_j\big[v_i - \frac{\sum_k n_k v_k}{n} + \bar{e_i} - \bar{e}\big]^2\big[v_j - \frac{\sum_k n_k v_k}{n} + \bar{e_j} - \bar{e}\big]^2\Big] \\
&\quad = n_i n_j E\Big[\,\Big[\Big(v_i - \frac{\sum_k n_k v_k}{n}\Big)^2 + (\bar{e_i} - \bar{e})^2 + 2\Big(v_i - \frac{\sum_k n_k v_k}{n}\Big)(\bar{e_i} - \bar{e})\Big] \\
&\qquad \Big[\Big(v_j - \frac{\sum_k n_k v_k}{n}\Big)^2 + (\bar{e_j} - \bar{e})^2 + 2\Big(v_j - \frac{\sum_k n_k v_k}{n}\Big)(\bar{e_j} - \bar{e})\Big]\Big] \qquad (8.39)
\end{aligned}
$$

To derive (8.39), we compute (8.40)-(8.43)

$$
\begin{aligned}
&E\Big(\Big[v_i - \frac{\sum_k n_k v_k}{n}\Big][\bar{e_i} - \bar{e}]\Big[v_j - \frac{\sum_k n_k v_k}{n}\Big][\bar{e_j} - \bar{e}]\Big) \\
&\quad = E[v_i - \frac{\sum_k n_k v_k}{n}][v_j - \frac{\sum_k n_k v_k}{n}]E[\bar{e_i} - \bar{e}][\bar{e_j} - \bar{e}] \\
&\quad = \big(- \frac{n_i}{n}\sigma_v^2 - \frac{n_j}{n}\sigma_v^2 + \sum_k \frac{n_k^2}{n^2}\sigma_v^2\big)\big(- \frac{1}{n_i n}n_i\sigma_e^2 - \frac{1}{n}\sigma_e^2 + \frac{1}{n}\sigma_e^2\big) \\
&\quad = \Big(\frac{\sum_k n_k^2}{n^2} - \frac{n_i}{n} - \frac{n_j}{n}\Big)\Big(\frac{-1}{n}\Big)\sigma_e^2\sigma_v^2 = O(m^{-2}).
\end{aligned} \tag{8.40}
$$

$$
\begin{aligned}
&E\Big[v_i - \frac{\sum_k n_k v_k}{n}\Big]^2\Big[v_j - \frac{\sum_k n_k v_k}{n}\Big]^2 \\
&\quad = E[v_i^2 - \frac{2v_i}{n}\sum_k n_k v_k + \frac{1}{n^2}(\sum_k n_k v_k)^2][v_j^2 - \frac{2v_j}{n}\sum_k n_k v_k + \frac{1}{n^2}(\sum_k n_k v_k)^2] \\
&\quad = \sigma_v^4 - \frac{2n_j}{n}\sigma_v^4 + \frac{1}{n^2}\sum_{k\neq i} n_k^2\sigma_v^4 + \frac{n_i^2}{n^2}\delta_v - \frac{2n_i}{n}\sigma_v^4 + \frac{8n_i n_j}{n^2}\sigma_v^4 + \frac{1}{n^2}\sum_{k\neq j} n_k^2\sigma_v^4 \\
&\quad + \frac{n_j^2}{n^2}\delta_v + \frac{1}{n^4}[\sum_k n_k^4\delta_v + 3\sum_{k\neq l} n_k n_l\sigma_v^4] + O(m^{-2}) \\
&\quad = \sigma_v^4 - \frac{2n_i}{n}\sigma_v^4 + \frac{1}{n^2}\sum_{k\neq i} n_k^2\sigma_v^4 - \frac{2n_j}{n}\sigma_v^4 + \frac{1}{n^2}\sum_{k\neq j} n_k^2\sigma_v^4 + O(m^{-2}).
\end{aligned} \tag{8.41}
$$

$$
\begin{aligned}
E[v_i - \frac{\sum_k n_k v_k}{n}]^2 E[\bar{e}_j - \bar{e}]^2 &= \big(\sigma_v^4 - \frac{2n_i}{n}\sigma_v^2 + \frac{1}{n^2}\sum_k n_k^2\sigma_v^2\big)\big(\frac{1}{n_j^2}n_j\sigma_e^2 + \frac{1}{n}\sigma_e^2 - \frac{2}{n}\sigma_e^2\big) \\
&= \big(\frac{1}{n_j} - \frac{1}{n}\big)\sigma_v^2\sigma_e^2 - \frac{2n_i}{n_j n}\sigma_v^2\sigma_e^2 + \frac{\sum_k n_k^2}{n_j n^2}\sigma_v^2\sigma_e^2 + O(m^{-2}).
\end{aligned} \tag{8.42}
$$

$$
\begin{aligned}
&E(\bar{e_i} - \bar{e})^2(\bar{e_j} - \bar{e})^2 = E[\bar{e_i}^2 - 2\bar{e_i}\bar{e} + \bar{e}^2][\bar{e_j}^2 - 2\bar{e_j}\bar{e} + \bar{e}^2] \\
&\quad = \frac{1}{n_i}\frac{1}{n_j}\sigma_e^4 - \frac{2}{n_i^2 n_j n}n_i n_j\sigma_e^4 + \frac{1}{n_i^2 n^2}[n_i(n-1)\sigma_e^4] - \frac{2}{n_j^2 n_i n}\sigma_e^4 + \frac{8}{n_i n_j n^2}n_i n_j\sigma_e^4 \\
&\quad + \frac{1}{n_j^2 n^2}[n_j(n-1)\sigma_e^4] + \frac{1}{n^4}[n\delta_e + 3n(n-1)\sigma_e^4] + O(m^{-2}) \\
&\quad = \frac{1}{n_i n_j}\sigma_e^4 - \frac{2}{n_i n}\sigma_e^4 + \frac{1}{n_i n}\sigma_e^4 - \frac{2}{n_j n}\sigma_e^4 + \frac{1}{n_j n}\sigma_e^4 + O(m^{-2})
\end{aligned} \tag{8.43}
$$

$$
\begin{aligned}
&\Rightarrow E\Big[n_i n_j\big[v_i - \frac{\sum_k n_k v_k}{n} + \bar{e_i} - \bar{e}\big]^2\big[v_j - \frac{\sum_k n_k v_k}{n} + \bar{e_j} - \bar{e}\big]^2\Big] \\
&\quad = n_i n_j\Big[\Big(\frac{1}{n_j} - \frac{1}{n}\Big) - \frac{2n_i}{n_j n} + \frac{\sum_k n_k^2}{n_j n^2}\Big]\sigma_v^2\sigma_e^2 \\
&\quad + \Big[1 - \frac{2n_j}{n} + \frac{1}{n^2}\sum_{k\neq j} n_k^2 - \frac{2n_i}{n} + \frac{1}{n^2}\sum_{k\neq j} n_k^2\Big]\sigma_v^4 + \Big[\Big(\frac{1}{n_i} - \frac{1}{n}\Big) - \frac{2n_j}{n_i n} + \frac{\sum_k n_k^2}{n_i n^2}\Big]\sigma_v^2\sigma_e^2 \\
&\quad + \Big[\frac{1}{n_i n_j} - \frac{1}{n_i n} - \frac{1}{n_i n}\Big]\sigma_e^4 + O(m^{-2}) \qquad (8.44)
\end{aligned}
$$

Now we get the first term in (8.35), from (8.38) and (8.44).

$$
\begin{aligned}
&\Rightarrow E\Big[\sum_{i=1}^{m} n_i[v_i - \frac{\sum_k n_k v_k}{n} + \bar{e_i} - \bar{e}]^2\Big]^2 \\
&\quad = \sum_{i=1}^{t} n_i^2\delta_v + 6n\sigma_e^2\sigma_v^2 + \sum_{i=1}^{m}\frac{1}{n_i}\delta_e + 3m\sigma_e^4 - 3\sum_{i=1}^{m}\frac{1}{n_i}\sigma_e^4 \\
&\quad + [2n(m-1) - 2n + 2\frac{\sum_i n_i^2}{n} - 4\frac{\sum_i n_i^2}{n}(m-1) + 2\frac{\sum_k n_k^2(m-1)}{n}]\sigma_v^2\sigma_e^2 \\
&\quad + \Big[n^2 - \sum_i n_i^2 - \frac{4}{n}(n\sum_k n_k^2 - \sum_k n_k^3) + \frac{2}{n^2}\sum_k n_k^2 n^2 - \frac{2}{n^2}(\sum_k n_k^2)^2\Big]\sigma_v^4 \\
&\quad + [m(m-1) - \frac{2}{n}n(m-1)]\sigma_e^4 + O(1) \\
&\quad = \sum_{i=1}^{m} n_i^2\delta_v + 6n\sigma_e^2\sigma_v^2 + \sum_{i=1}^{m}\frac{1}{n_i}\delta_e - 3\sum_{i=1}^{m}\frac{1}{n_i}\sigma_e^4 + 3m\sigma_e^4 \\
&\quad + \Big[2nm - 4n - 2\sum_{i=1}^{m} n_i^2\frac{m}{n}\Big]\sigma_v^2\sigma_e^2 + [n^2 - \sum_i n_i^2 - 2\sum_k n_k^2]\sigma_v^4 + \\
&\quad + (m-1)(m-2)\sigma_e^4 + O(1). \qquad (8.45)
\end{aligned}
$$

Hence, from (8.35),(8.36) and (8.45), we get

$$
\begin{aligned}
var[SSB] &= \sum_{i=1}^{m} n_k^2\delta_v + \Big[2nm + 2n - 2\sum_k n_k^2\frac{m}{n}\Big]\sigma_v^2\sigma_e^2 + [n^2 - 3\sum_k n_k^2]\sigma_v^4 \\
&+ [3m + (m-1)(m-2)]\sigma_e^4 + \sum_{k=1}^{m}\frac{1}{n_k}(\delta_e - 3\sigma_e^4) - [n^2 - 2\sum_k n_k^2]\sigma_v^4 \\
&- (m-1)^2\sigma_e^4 - 2(m-1)(n - \frac{\sum_k n_k^2}{n})\sigma_v^2\sigma_e^2 \\
&= \sum_{k=1}^{m} n_k^2\delta_v - \sum_{k=1}^{m} n_k^2\sigma_v^4 + 4n\sigma_v^2\sigma_e^2 + 2m\sigma_e^4 + \sum_{k=1}^{m}\frac{1}{n_k}(\delta_e - 3\sigma_e^4) + O(1). \qquad (8.46)
\end{aligned}
$$

Derivation of cov(SSB,SSW)

$$
cov(SSB, SSW) = \mathrm{E}\Big[\sum_{i=1}^{m} n_i(\bar{y}_i - \bar{y})^2 \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2\Big] - \mathrm{E}(SSB)\mathrm{E}(SSW) \qquad (8.47)
$$

where

$$
\mathrm{E}(SSB) = \Big(n - \frac{\sum_{k=1}^{m} n_k^2}{n}\Big)\sigma_v^2 + (m-1)\sigma_e^2 \qquad (8.48)
$$

$$
\mathrm{E}(SSW) = (n-m)\sigma_e^2 \qquad (8.49)
$$

$$
\begin{aligned}
\mathrm{E}[SSBSSW] &= \mathrm{E}\Big[\sum_{i=1}^{m} n_i(v_i - \frac{\sum n_k v_k}{n} + \bar{e}_i - \bar{e}]^2 \sum_{i=1}^{m}\sum_{j=1}^{n_i}(e_{ij} - \bar{e}_i)^2\Big] \\
&= \sum_{i=1}^{m} n_i[\sigma_v^2 + \frac{\sum n_k^2}{n^2}\sigma_v^2 - \frac{2n_i}{n}\sigma_v^2](n-m)\sigma_e^2 \\
&\quad + \mathrm{E}\Big[\sum_{i=1}^{m} n_i(\bar{e}_i - \bar{e})^2 \sum_{i=1}^{m}\sum_{j=1}^{n_i}(e_{ij} - \bar{e}_i)^2\Big] \qquad (8.50)
\end{aligned}
$$

In (8.50) consider the first term.

$$
\begin{aligned}
\sum_{i=1}^{m} n_i\Big[\sigma_v^2 + \frac{\sum n_k^2\sigma_v^2}{n^2} - \frac{2n_i}{n}\sigma_v^2\Big](n-m)\sigma_e^2 &= \sigma_v^2\sigma_e^2(n-m)\Big[n + \frac{\sum n_k^2}{n} - \frac{2\sum n_k^2}{n}\Big] \\
&= (n-m)\sigma_v^2\sigma_e^2\Big[n - \sum\frac{n_k^2}{n}\Big] \qquad (8.51)
\end{aligned}
$$

In (8.50) consider the term $\mathrm{E}\Big[\sum_{k=1}^{m} n_k(\bar{e}_k-\bar{e})^2\sum_{i=1}^{m}\sum_{j=1}^{n_i}(e_{ij}-\bar{e}_i)^2\Big]$. For $i=k$,

$$\begin{aligned}
&\mathrm{E}\Big[\sum_{k=1}^{m}\sum_{j=1}^{n_i} n_k(\bar{e}_k-\bar{e})^2(e_{kj}-\bar{e}_k)^2\Big]\\
&\quad=\sum_{k=1}^{m}\sum_{j=1}^{n_i} n_k\mathrm{E}[e_{kj}^2+\bar{e}_k^2-2\bar{e}_k e_{kj}][\bar{e}_k^2-2\bar{e}_k\bar{e}+\bar{e}^2]\\
&\quad=\sum_{k=1}^{m}\sum_{j=1}^{n_i} n_k\Big[\frac{1}{n_k^2}(\delta_e+(n_k-1)\sigma_e^4)+\frac{1}{n_k^4}(n_k\delta_e+3n_k(n_k-1)\sigma_e^4)\\
&\qquad\qquad-\frac{2}{n_k^3}(\delta_e+3(n_k-1)\sigma_e^4)\Big]+O(1)\\
&\quad=\sum_{k=1}^{m}\sum_{j=1}^{n_i}\Big[\frac{1}{n_k}\delta_e+\Big(1-\frac{1}{n_k}\Big)\sigma_e^4-\frac{1}{n_k^2}\delta_e-3\Big(\frac{1}{n_k}-\frac{1}{n_k^2}\Big)\sigma_e^4\Big]+O(1)\\
&\quad=\Big(m-\sum_{k=1}^{m}\frac{1}{n_k}\Big)(\delta_e-3\sigma_e^4)+(n-m)\sigma_e^4+O(1)
\end{aligned}\tag{8.52}$$

For $i\neq k$,

$$\begin{aligned}
&\mathrm{E}\Big[\sum_{i\neq k}\sum_{j=1}^{n_i}(e_{ij}-\bar{e}_k)^2 n_k(\bar{e}_k-\bar{e})^2\Big]\\
&=\sum_{i\neq k}\sum_{j=1}^{n_i} n_k\mathrm{E}[(e_{ij}^2-2\bar{e}_j e_{ij}+\bar{e}_i^2)(\bar{e}_k^2-2\bar{e}_k\bar{e}+\bar{e}^2)]\\
&=\sum_{i\neq k}\sum_{j=1}^{n_i} n_k\Big[\frac{1}{n_k^2}n_k\sigma_e^4-\frac{2}{n_k n}n_k\sigma_e^4+\frac{1}{n^2}((n-1)\sigma_e^4+\delta_e)\Big]\\
&\qquad-2\sum_{i\neq k}\sum_{j=1}^{n_i} n_k\Big[\frac{1}{n_i n_k^2}n_k\sigma_e^4-\frac{2}{n_i n_k n}(n_k\sigma_e^4)+\frac{1}{n_i n^2}(\delta_e+(n-1)\sigma_e^4)\Big]\\
&\qquad+\sum_{i\neq k}\sum_{j=1}^{n_i} n_k\Big[\frac{1}{n_i n_k}\sigma_e^4-\frac{2}{n_i^2 n_k n}n_i n_k\sigma_e^4+\frac{1}{n_i n^2}(n_i(n-1)\sigma_e^4)\Big]+O(1)\\
&=\sigma_e^4\sum_{i\neq k}\sum_{j=1}^{n_i}\Big[1-\frac{2n_k}{n}+\Big(\frac{n-1}{n^2}\Big)n_k\Big]+\sigma_e^4\sum_{i\neq k}\sum_{j=1}^{n_i}\Big[-\frac{2}{n_i}+\frac{4n_k}{n_i n}-\frac{2n_k(n-1)}{n_i n^2}\Big]\\
&\qquad+\sigma_e^4\sum_{i\neq k}\sum_{j=1}^{n_i}\Big[\frac{1}{n_i}-\frac{2n_k}{n_i n}+\frac{n_k(n-1)}{n_i n^2}\Big]+O(1)
\end{aligned}$$

$$= \sigma_e^4 \sum_{i \neq k} \sum_{j=1}^{n_i} \Big[1 - \frac{n_k}{n} - \frac{1}{n_i} + \frac{2n_k}{n_i n} - \frac{n_k(n-1)}{n_i n^2} \Big] + O(1)$$

$$= [n(m-2) - (m-1)^2]\sigma_e^4 + O(1) \tag{8.53}$$

Putting (8.47)-(8.53) together we get

$$\begin{aligned} \mathrm{cov}(SSB, SSW) &= (n-m)\sigma_e^4 + (\delta_e - 3\sigma_e^4)(m - \sum_{k=1}^{m} \frac{1}{n_k}) \\ &\quad + [n(m-2) - (m-1)^2)]\sigma_e^4 - (n-m)(m-1)\sigma_e^4 + O(1) \\ &= (\delta_e - 2\sigma_e^4)(m - \sum_{k=1}^{m} \frac{1}{n_k}) + O(1) \end{aligned} \tag{8.54}$$

Derivation of $\mathrm{var}(\hat{\sigma}_e^2)$. Using (8.34) we get

$$\begin{aligned} \mathrm{var}(\hat{\sigma}_e^2) &= \mathrm{var}\Big[\frac{SSW}{n-m}\Big] \\ &= \frac{1}{(n-m)^2}[(n - 2m + \sum_{k=1}^{m} \frac{1}{n_k})\delta_e + (-n + 4m - 3\sum_{k=1}^{m} \frac{1}{n_k})\sigma_e^4] \\ &= \frac{1}{(n-m)^2}\Big((n-2m)\delta_e - (n-4m)\sigma_e^4 + \sum_{i=1}^{m} \frac{1}{n_i}(\delta_e - 3\sigma_e^4)\Big) \end{aligned} \tag{8.55}$$

Derivation of $\mathrm{var}(\hat{\sigma}_v^2)$. Using (8.34), (8.46) and (8.54) we get

$$\begin{aligned} &\mathrm{var}(\hat{\sigma}_v^2) = \mathrm{var}\Big[\Big(\frac{SSB}{m-1} - \frac{SSW}{n-m}\Big)\frac{(m-1)}{g}\Big] \\ &= \frac{(m-1)^2}{g^2}\Big[\frac{\mathrm{var}(SSB)}{(m-1)^2} + \frac{\mathrm{var}(SSW)}{(n-m)^2} - \frac{2}{(m-1)(n-m)}\mathrm{cov}(SSB, SSW)\Big] \\ &= \frac{(m-1)^2}{g^2}\Big[\frac{1}{(m-1)^2}\Big(\sum_{k=1}^{m} n_k^2\delta_v - \sum_{k=1}^{m} n_k^2\sigma_v^4 + 4n\sigma_v^2\sigma_e^2 + 2m\sigma_e^4 + \sum_{k=1}^{m} \frac{1}{n_k}(\delta_e - 3\sigma_e^4)\Big) \\ &\quad + \frac{1}{(n-m)^2}\Big((n - 2m + \sum_{k=1}^{m} \frac{1}{n_k})\delta_e + (-n + 4m - 3\sum_{k=1}^{m} \frac{1}{n_k})\sigma_e^4\Big) \\ &\quad - \frac{2}{(m-1)(n-m)}\Big((\delta_e - 3\sigma_e^4)(m - \sum_{k=1}^{m} \frac{1}{n_k})\Big)\Big] + O(m^{-2}) \end{aligned}$$

By noting that $\frac{1}{g^2} = \frac{1}{n^2} + O(m^{-3})$ we get

$$\begin{aligned}\operatorname{var}(\hat{\sigma}_v^2) = \frac{1}{n^2}\Big(\sum_{k=1}^{m} n_k^2(\delta_v - \sigma_v^4) + 4n\sigma_v^2\sigma_e^2 + 2m\sigma_e^4 + \sum_{k=1}^{m}\frac{1}{n_k}(\delta_e - 3\sigma_e^4)\Big) \\ + \frac{m^2}{n^2(n-m)^2}\Big((n-2m)\delta_e - (n-4m)\sigma_e^4 + \sum_{k=1}^{m}\frac{1}{n_k}(\delta_e - 3\sigma_e^4)\Big) \\ - \frac{2m}{n^2(n-m)}\Big(m - \sum_{k=1}^{m}\frac{1}{n_k}\Big)(\delta_e - 3\sigma_e^4) + O(m^{-2}) \qquad (8.56)\end{aligned}$$

Derivation of $\operatorname{cov}(\hat{\sigma}_e^2, \hat{\sigma}_v^2)$. By (8.34) and (8.50)-(8.53) we get

$$\begin{aligned}&\operatorname{cov}(\hat{\sigma}_e^2, \hat{\sigma}_v^2) = \mathrm{E}\Big[\frac{SSW}{n-m}\Big(SSB - \frac{(m-1)SSW}{n-m}\Big)g^{-1}\Big] - \sigma_e^2\sigma_v^2 \\ &= \frac{1}{g}\Big[\frac{1}{(n-m)}\mathrm{E}(SSW \cdot SSB) - \frac{(m-1)}{(n-m)^2}\mathrm{E}(SSW^2)\Big] - \sigma_e^2\sigma_v^2 \\ &= \frac{g^{-1}}{(n-m)}\Big[(\delta_e - 3\sigma_e^4)\Big(m - \sum_{k=1}^{m}\frac{1}{n_k}\Big) + (n-m)\Big(n - \frac{\sum_{k=1}^{m} n_k^2}{n}\Big)\sigma_e^2\sigma_v^2 \\ &\qquad + (n-m)(m-1)\sigma_e^4\Big] - \frac{(m-1)}{(n-m)^2 g}\Big[\Big(n - 2m + \sum_{k=1}^{m}\frac{1}{n_k}\Big)\delta_e \\ &\qquad - \Big(n - 4m + 3\sum_{k=1}^{m}\frac{1}{n_k}\Big)\sigma_e^4 + (n-m)^2\sigma_e^4\Big] - \sigma_e^2\sigma_v^2 + O(m^{-2}) \\ &= \frac{(\delta_e - 3\sigma_e^4)(m - \sum_{k=1}^{m}\frac{1}{n_k})}{g(n-m)} - \frac{(m-1)}{(n-m)^2 g}\Big[\Big(n - 2m + \sum_{k=1}^{m}\frac{1}{n_k}\Big)\delta_e \\ &\qquad - \Big(n - 4m + 3\sum_{k=1}^{m}\frac{1}{n_k}\Big)\sigma_e^4\Big] + O(m^{-2}) \\ &= \frac{1}{n(n-m)^2}\Big(m^2(\delta_e - \sigma_e^4) - 2mn\sigma_e^4 - n\sum_{i=1}^{m}\frac{1}{n_i}(\delta_e - 3\sigma_e^4)\Big) + O(m^{-2}) \qquad (8.57)\end{aligned}$$

Derivation of the robust second order unbiased MSE estimator. We give a brief argument outlining the proof. First it needs to be shown that $\mathrm{E}[g_{1i}(\hat{\boldsymbol{\sigma}}^2)] = g_{1i}(\boldsymbol{\sigma}^2) - g_{3i}(\boldsymbol{\sigma}^2)$. Note that

$$\frac{\partial g_{1i}(\boldsymbol{\sigma}^2)}{\partial \sigma_e^2} = \frac{n_i\sigma_v^4}{(\sigma_e^2 + n_i\sigma_v^2)^2}, \quad \frac{\partial g_{1i}(\boldsymbol{\sigma}^2)}{\partial \sigma_v^2} = \frac{\sigma_e^4}{(\sigma_e^2 + n_i\sigma_v^2)^2}, \quad \frac{\partial^2 g_{1i}(\boldsymbol{\sigma}^2)}{\partial (\sigma_e^2)^2} = -\frac{2n_i\sigma_v^4}{(\sigma_e^2 + n_i\sigma_v^2)^3},$$
$$\frac{\partial^2 g_{1i}(\boldsymbol{\sigma}^2)}{\partial (\sigma_v^2)^2} = -\frac{2n_i\sigma_e^4}{(\sigma_e^2 + n_i\sigma_v^2)^3}, \qquad \frac{\partial^2 g_{1i}(\boldsymbol{\sigma}^2)}{\partial \sigma_e^2\sigma_v^2} = \frac{2n_i\sigma_e^2\sigma_v^2}{(\sigma_e^2 + n_i\sigma_v^2)^3}.$$

Now we expand $g_{1i}(\hat{\boldsymbol{\sigma}}^2)$ around $\boldsymbol{\sigma}^2 = (\sigma_v^2, \sigma_e^2)$.

$$g_{1i}(\hat{\boldsymbol{\sigma}}^2) = g_{1i}(\boldsymbol{\sigma}^2) + (\hat{\sigma}_e^2 - \sigma_e^2)\frac{\partial g_{1i}(\boldsymbol{\sigma}^2)}{\partial \sigma_e^2} + (\hat{\sigma}_v^2 - \sigma_v^2)\frac{\partial g_{1i}(\boldsymbol{\sigma}^2)}{\partial \sigma_v^2} + \frac{1}{2}(\hat{\sigma}_e^2 - \sigma_e^2)^2\frac{\partial^2 g_{1i}(\boldsymbol{\sigma}^2)}{\partial (\sigma_e^2)^2}$$
$$+ \frac{1}{2}(\hat{\sigma}_v^2 - \sigma_v^2)^2\frac{\partial^2 g_{1i}(\boldsymbol{\sigma}^2)}{\partial (\sigma_v^2)^2} + (\hat{\sigma}_e^2 - \sigma_e^2)(\hat{\sigma}_v^2 - \sigma_v^2)\frac{\partial^2 g_{1i}(\boldsymbol{\sigma}^2)}{\partial \sigma_e^2 \sigma_v^2} + R_m$$

where R_m are the remainder terms. Computing the expectation of $g_{1i}(\hat{\boldsymbol{\sigma}}^2)$ we get

$$\begin{aligned}\mathrm{E}[g_{1i}(\hat{\boldsymbol{\sigma}}^2)] &= g_{1i}(\boldsymbol{\sigma}^2) - \frac{1}{2}var(\hat{\sigma}_e^2)\frac{2n_i\sigma_v^4}{(\sigma_e^2 + n_i\sigma_v^2)^3} - \frac{1}{2}var(\hat{\sigma}_v^2)\frac{2n_i\sigma_e^4}{(\sigma_e^2 + n_i\sigma_v^2)^3} \\ &\quad + cov(\hat{\sigma}_v^2, \hat{\sigma}_v^2)\frac{2n_i\sigma_e^2\sigma_v^2}{(\sigma_e^2 + n_i\sigma_v^2)^3} + \mathrm{E}(R_m) \\ &= g_{1i}(\boldsymbol{\sigma}^2) - g_{3i}(\boldsymbol{\sigma}^2) + \mathrm{E}(R_m)\end{aligned}$$

It needs to be shown that $\mathrm{E}(R_m) = o(m^{-1})$. Consider a typical remainder term:

$$\frac{1}{6}(\hat{\sigma}_e^2 - \sigma_e^2)^3\frac{\partial^3 g_{1i}(\boldsymbol{\sigma}_*^2)}{\partial (\sigma_e^2)^3}$$

where $\boldsymbol{\sigma}_*^2$ lies between $\hat{\boldsymbol{\sigma}}^2$ and $\boldsymbol{\sigma}^2$. Note that

$$\frac{\partial^3 g_{1i}(\boldsymbol{\sigma}_*^2)}{\partial (\sigma_e^2)^3} < 1$$

and since $\hat{\sigma}_e^2$ is $\sqrt{m}$-consistent we have $|\hat{\sigma}_e^2 - \sigma_e^2|^3 = o_p(m^{-\frac{3}{2}})$. It has to be shown $\mathrm{E}|\hat{\sigma}_e^2 - \sigma_e^2|^3 = o(m^{-1})$. Similarly it has to be shown the other remainder terms are also $o(m^{-1})$. Hence we have

$$\mathrm{E}[g_{1i}(\hat{\boldsymbol{\sigma}}^2)] = g_{1i}(\boldsymbol{\sigma}^2) - g_{3i}(\boldsymbol{\sigma}^2) + o(m^{-1}) \tag{8.58}$$

Next we argue that $\mathrm{E}[g_{2i}(\hat{\boldsymbol{\sigma}}^2)] = g_{2i}(\boldsymbol{\sigma}^2) + o(m^{-1})$. As we did previously expand $g_{2i}(\hat{\boldsymbol{\sigma}}^2)$ around $\boldsymbol{\sigma}^2 = (\sigma_v^2, \sigma_e^2)$.

$$g_{2i}(\hat{\boldsymbol{\sigma}}^2) = g_{2i}(\boldsymbol{\sigma}^2) + (\hat{\sigma}_e^2 - \sigma_e^2)\frac{\partial g_{2i}(\boldsymbol{\sigma}_*^2)}{\partial \sigma_e^2} + (\hat{\sigma}_v^2 - \sigma_v^2)\frac{\partial g_{2i}(\boldsymbol{\sigma}_*^2)}{\partial \sigma_v^2}$$

where $\boldsymbol{\sigma}_*^2$ lies between $\hat{\boldsymbol{\sigma}}^2$ and $\boldsymbol{\sigma}^2$. It needs to be shown that

$$\mathrm{E}\left|(\hat{\sigma}_e^2 - \sigma_e^2)\frac{\partial g_{2i}(\boldsymbol{\sigma}_*^2)}{\partial \sigma_e^2}\right| = o(m^{-1})$$

Note that $\frac{\partial g_{2i}(\boldsymbol{\sigma}_*^2)}{\partial \sigma_e^2} = O_p(m^{-1})$ and since $\hat{\sigma}_e^2$ is $\sqrt{m}$-consistent we have $\mathrm{E}\left|(\hat{\sigma}_e^2 - \sigma_e^2)\frac{\partial g_{2i}(\boldsymbol{\sigma}_*^2)}{\partial \sigma_e^2}\right| = o(m^{-1})$. So we have

$$\mathrm{E}[g_{2i}(\hat{\boldsymbol{\sigma}}^2)] = g_{2i}(\boldsymbol{\sigma}^2) + o(m^{-1}) \tag{8.59}$$

Similarly it needs to be shown that

$$\mathrm{E}[g_{3i}(\hat{\boldsymbol{\sigma}}^2, \hat{\boldsymbol{\delta}})] = g_{3i}(\boldsymbol{\sigma}^2, \boldsymbol{\delta}) + o(m^{-1}) \tag{8.60}$$

$$\mathrm{E}[g_{4i}(\hat{\boldsymbol{\sigma}}^2, \hat{\boldsymbol{\delta}})] = g_{4i}(\boldsymbol{\sigma}^2, \boldsymbol{\delta}) + o(m^{-1}) \tag{8.61}$$

Putting (8.58)-(8.61) together we have

$$\begin{aligned}\mathrm{E}[g_{1i}(\hat{\boldsymbol{\sigma}}^2) + g_{2i}(\hat{\boldsymbol{\sigma}}^2) + 2g_{3i}(\hat{\boldsymbol{\sigma}}^2, \hat{\boldsymbol{\delta}}) + g_{4i}(\hat{\boldsymbol{\sigma}}^2, \hat{\boldsymbol{\delta}})] &= g_{1i}(\boldsymbol{\sigma}^2) + g_{2i}(\boldsymbol{\sigma}^2) + g_{3i}(\boldsymbol{\sigma}^2, \boldsymbol{\delta}) \\ &\quad + g_{4i}(\boldsymbol{\sigma}^2, \boldsymbol{\delta}) + o(m^{-1}) \\ &= MSE[\hat{\theta}_i(\hat{\sigma}^2)] + o(m^{-1}) \quad \text{[by (8.13)]}\end{aligned}$$

which shows that the robust MSE estimator given by (8.18) is second order unbiased.

Derivation of $g_3(\boldsymbol{\sigma}^2, \boldsymbol{\delta})$ for the balanced case. In the balanced case we denote for all i, $n_i = k$. For balanced case after some algebra and disregarding terms of the order $O(m^{-2})$ it follows that (8.55)-(8.57) are given by

$$\mathrm{var}(\hat{\sigma}_e^2) = \frac{1}{km}\delta_e - \frac{(k-3)}{k(k-1)m}\sigma_e^4 \tag{8.62}$$

$$\mathrm{var}(\hat{\sigma}_v^2) = \frac{\delta_v - \sigma_v^4}{m} + \frac{4\sigma_v^2\sigma_e^2}{km} + \frac{2\sigma_e^4}{k^2m} - \frac{(k-3)\sigma_e^4}{k^3(k-1)m} + \frac{3\sigma_e^4}{k^3m} + O(m^{-2}) \tag{8.63}$$

$$\mathrm{cov}(\hat{\sigma}_e^2, \hat{\sigma}_v^2) = \frac{\delta_e - 3\sigma_e^4}{k^2m} - \frac{\delta_e}{k^2m} + \frac{(k-3)\sigma_e^4}{k^2(k-1)m} + O(m^{-2}) \tag{8.64}$$

By (8.10) and (8.62)-(8.64) we get

$$
\begin{aligned}
& g_3(\boldsymbol{\sigma}^2, \boldsymbol{\delta}) \\
&= \frac{1/k^2}{(\sigma_e^2/k + \sigma_v^2)^3}\Big(\sigma_v^4 \mathrm{var}(\hat{\sigma}_e^2) + \sigma_e^4 \mathrm{var}(\hat{\sigma}_v^2) - 2\sigma_e^2\sigma_v^2 \mathrm{cov}(\hat{\sigma}_v^2, \hat{\sigma}_e^2)\Big) \\
&= \frac{1/k^2}{(\sigma_e^2/k + \sigma_v^2)^3}\Big(\frac{\delta_e}{km}\sigma_v^4 + \frac{\mu_{4v}\sigma_e^4}{m} - \frac{\sigma_e^4\sigma_v^4}{m}\Big[1 + \frac{k-3}{k(k-1)}\Big] + \frac{2\sigma_e^8}{k(k-1)m} \\
&\qquad\qquad + \frac{\sigma_e^6\sigma_v^2}{km}\Big[4 + \frac{6}{k} - \frac{2(k-3)}{k(k-1)}\Big]\Big) \\
&= \frac{1/k^2}{(\sigma_e^2/k + \sigma_v^2)^3}\Big(\frac{\delta_e}{km}\sigma_v^4 + \frac{\mu_{4v}\sigma_e^4}{m} - \frac{\sigma_e^4\sigma_v^4}{k(k-1)m}(k^2-3) + \frac{2\sigma_e^8}{k(k-1)m} + \frac{4\sigma_e^6\sigma_v^2}{(k-1)m}\Big) \\
&= \frac{k}{m(\sigma_e^2 + k\sigma_v^2)^3}\Big[\frac{\delta_e}{k}\sigma_v^4 + \mu_{4v}\sigma_e^4 + \frac{2\sigma_e^8}{k(k-1)} - \frac{\sigma_e^4\sigma_v^4(k^2-3)}{k(k-1)} + \frac{4\sigma_e^6\sigma_v^2}{k-1}\Big] \\
&= \frac{1}{m(\sigma_e^2 + k\sigma_v^2)^3}\Big[\delta_e\sigma_v^4 + k\delta_v\sigma_e^4 + \frac{2\sigma_e^8}{k-1} - \frac{\sigma_e^4\sigma_v^4(k^2-3)}{k-1} + \frac{4k\sigma_e^6\sigma_v^2}{k-1}\Big] \qquad (8.65)
\end{aligned}
$$

Derivation of $g_4(\boldsymbol{\sigma}^2, \boldsymbol{\delta})$ for the balanced case. From (8.11) we have

$$
\begin{aligned}
g_4(\boldsymbol{\sigma}^2, \boldsymbol{\delta}) &= \frac{2}{n}\frac{\sigma_e^2/k}{(\sigma_e^2/k + \sigma_v^2)^3}\Big(\frac{\sigma_v^2}{k^2}(\delta_e - 3\sigma_e^4) - \sigma_e^2(\delta_v - 3\sigma_v^4)\Big) \\
&\quad - \Big(\frac{k-1}{k^3}\Big)\frac{1}{(n-m)}\frac{\sigma_v^4}{(\sigma_e^2/k + \sigma_v^2)^3}(\delta_e - 3\sigma_e^4) \\
&= \frac{2}{m}\frac{1}{(\sigma_e^2 + k\sigma_v^2)^3}\Big(\frac{1}{k}\sigma_e^2\sigma_v^2(\delta_e - 3\sigma_e^4) - \sigma_v^4(\delta_e - 3\sigma_e^4) - k\sigma_e^4(\delta_v - 3\sigma_v^4)\Big) \qquad (8.66)
\end{aligned}
$$

Chapter 9

Summary of results and future research problems

A summary of the results of this thesis and a list of research problems that have arisen from this thesis are listed below.

1. By considering spatial and non-spatial covariates to cluster the small areas, I have introduced a hybrid asymptotic framework between infill asymptotics and increasing domain asymptotics. By building on the popular exponential covariance model with nugget effect, I have introduced some variance-covariance models for the random effects.

2. Under my asymptotic framework, I have derived parameter estimators that are consistent and asymptotically normal. Moreover, I have provided some simulation evidence to show that the MLE exhibits its "usual" large sample behavior. However, I have not shown that the MLE is consistent and asymptotically normal. Proving consistency and asymptotic normality of the MLE for general patterns $\mathbf{z}_i$ will be very technical. However, for certain specialized spatial patterns I will attempt to do so.

3. I have shown that the estimators $(\boldsymbol{\beta}, \tau^2)$ derived under the Fay-Herriot model are somewhat robust for certain types of model misspecification. However, the relative efficiency of $\widehat{\boldsymbol{\beta}}_{\text{FH}}$ and $\widehat{\boldsymbol{\beta}}_{\text{ML}}$ can be small especially if the random effects are strongly correlated.

4. Through simulation and a real data example I have shown improved prediction over predictors that ignore small area correlations. Simulations indicate for purposes of prediction, the method of parameter estimation (my method and the MLE) does not seem to matter. I hope to consider a more comprehensive simulation study as suggested in Section 4.6.

5. I have not considered estimation of the parameters for the covariance models that include a vector parameter $\boldsymbol{\lambda}$. I should be able to generalize the estimation methods developed in Chapter 3 to derive the large sample properties of the estimators of this more general model. In addition, as mentioned in Chapter 4, I hope to derive a least squares estimator of σ^2. I also plan on showing consistency and asymptotic normality of the estimator for (δ, λ) given by (4.8)-(4.9).

6. Due to time constraints, the data analysis in Chapter 5 was done only using spatial locations to cluster the small areas (U.S. counties). However, in a future study of the data set analyzed in Chapter 5, I plan on using non-spatial covariates in addition to spatial locations to cluster the small areas.

7. By borrowing frequentist methods for multiple comparisons, I have shown

how they could be applied in a Bayesian setting for the problem of multiple comparisons of small areas. In the context of multiple comparisons, I have introduced a new class of moment matching priors.

8. For a special case of the nested error regression model, a robust MSE estimator of the EBLUP was derived. For the balanced case, the Prasad-Rao MSE estimator was shown to be second order unbiased when the errors e_{ij} are normally distributed. Moreover, my simulation study indicates that the Prasad-Rao MSE estimator is robust for departures from normality. I will be generalizing the robust MSE estimator to the regression case.

BIBLIOGRAPHY

[1] Abt, M. & Welch, W.J. (1998), Fisher information and maximum likelihood estimation of covariance parameters in Gaussian stochastic processes, *Canadian Journal of Statistics*, **26**, 127-137.

[2] Andersen, P.K. & Gill, R.D. (1982), Cox's regression model for counting processes: A large sample study, *Annals of Statistics* **10**, 1100-1120.

[3] Battese, G.E, Harter, R.M. & Fuller, W.A. (1988), An error-components model for prediction of county crop areas using survey and satellite data, *Journal of the American Statistical Association*, **80**, 28-36.

[4] Berger, J. (1985), *Statistical Decision Theory and Bayesian Analysis,* 2nd ed. New York: Springer-Verlag.

[5] Besag, J. (1974), Spatial interaction and the statistical analysis of lattice systems (with discussion), *Journal of the Royal Statistical Society*, Ser. B, **35**, 192-236.

[6] Clayton, D. & Kaldor J. (1987), Empirical Bayes estimates of age-standardized relative risks for use in disease mapping, *Biometrics*, **43**, 671-681.

[7] Chen, H., Simpson, D.G. & Ying, Z. (2000), Infill asymptotics for a stochastic process model with measurement error, *Statistical Sinica*, **10**, 141-156.

[8] Claeskens, G. (2004), Restricted likelihood ratio lack-of-fit tests using mixed spline models, *Journal of the Royal Statistical Society*, Ser. B, **66**, 909-926.

[9] Cliff, A.D. & Ord, J.K. (1980), *Spatial processes: models and applications*, London: Pion.

[10] Cressie, N. (1991), Small area prediction of undercount using the general linear model, *Proceedings of statistics symposium 90: measurement and improvement of data quality*, Ottawa: Statistics Canada, 93-105.

[11] Cressie, N. (1993), *Statistics for Spatial Data*, New York: Wiley.

[12] Cressie, N. & Chan, N.H. (1989), Spatial modeling of regional variables, *Journal of the American Statistical Association* **84**, 393-401.

[13] Cressie, N. & Huang, H. (1999), Classes of nonseparable, spatio-temporal stationary covariance functions, *Journal of the American Statistical Association* **94**, 1330-1340.

[14] Das, K., Jiming, J. & Rao, J.N.K. (2004), Mean square error of empirical predictor, *Annals of Statistics* **32**, 818-840.

[15] Datta, G.S. & Lahiri, P. (2000), A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems, *Statistical Sinica* **10**, 613-627.

[16] Datta, G.S. & Mukerjee, R. (2004), *Probability Matching Priors: Higher Order Asymptotics* , New York: Springer.

[17] Datta, G.S., Rao, J.N.K. & Smith, D.D. (2005), On measuring the variability of small area estimators under a basic area level model, *Biometrika*, **92**, 183-196.

[18] Fay, R.E. & Herriot, R. A. (1979), Estimates of income for small places: An application of the James-Stein procedures to census data, *Journal of the American Statistical Association* **74**, 269-277.

[19] Haining, R. (1987), Trend-surface models with regional and local scales of variation with an application to aerial survey data, *Technometrics* **29**, 461-469.

[20] Hall, P. & Maiti, T. (2006), Nonparametric estimation of mean-squared prediction error in nested-error regression models, *Annals of Statistics* **34**, 1733-1750.

[21] Hochberg, Y. & Tamhane, A.C. (1987), *Multiple Comparison Procedures*, New York: John Wiley & Sons.

[22] Kacker, R.N. & Harville, D.A. (1984), Approximations for standard errors of estimators of fixed and random effects in mixed linear models, *Journal of the American Statistical Association* **79**, 853-862.

[23] Kammann, E.E. & Wand, M.P. (2002), Geoadditive models, *Applied Statistics*, **52**, 1-18.

[24] Lahiri, P. & Rao, J.N.K. (1995), Robust estimation of mean square error of small area estimators, *Journal of the American Statistical Association* **90**, 758-766.

[25] Loh, W. & Lam, T. (2000), Estimating structured correlation matrices in Gaussian random field models, *Annals of Statistics*, **28**, 880-904.

[26] Mardia, K.V. & Marshall, R.J. (1984), Maximum likelihood estimation of models for residual covariance in spatial regression, *Biometrika*, **71**, 135-146.

[27] Mathai, A.M. & Provost, S.B. (1992), *Quadratic Forms in Random Variables*, New York: Marcel Dekker, Inc.

[28] Miller, J.J. (1977), Asymptotic properties of maximum likelihood estimates in the mixed model of the analysis of variance, *Annals of Statistics* **5**, 746-762..

[29] Miller, R.G., Jr. (1991), *Simultaneous Statistical Inference*, New York: Springer-Verlag.

[30] Morris, C.N., & Christiansen, C.L. (1995), Hierarchical models for ranking and for identifying extremes with applications, *Bayes Statistics 5*, Oxford: Oxford University Press.

[31] Morris, C.N. (1983), Parametric empirical Bayes inference: theory and applications, *Journal of the American Statistical Association*, **78**, 47-59.

[32] Morris, C.N. (1983), Parametric empirical Bayes confidence intervals, In *Proceeding on the Conference on Scientific Inference, Data Analysis and Robustness*, New York: Academic Press, 25-50.

[33] McCulloch, C.E. & Searle, S.R. (2001), *Generalized, Linear, and Mixed Models*, New York: Wiley.

[34] Opsomer, J.D., Claeskens, G., Ranalli, M.G., Kauermann, G. & Breidt, F.J. (2007), Nonparametric small area estimation using penalized spline regression, working paper.

[35] Ord, K. (1975), Estimation methods for models of spatial interaction, *Journal of the American Statistical Association* **70**, 120-126.

[36] Prasad, N.G.N. & Rao, J.N.K. (1990), The estimation of mean square errors of small area estimators, *Journal of the American Statistical Association* **85**, 163-171.

[37] Rao, J.N.K. (2003), *Small Area Estimation*, New York: Wiley.

[38] Rencher, A.C. (2000), *Linear Models in Statistics*, New York: Wiley.

[39] Richardson, S., Guihenneuc, C. & Lasserre, V. (1992), Spatial linear models with autocorrelated error structure, *The Statistician* **41**, 539-557.

[40] Robert, C.P. & Casella, G. (1999), *Monte Carlo Statistical Methods*, New York: Springer.

[41] Ruppert, D., Wand, M.P. & Carroll, R.J. (2003), *Semiparametric Regression*, New York: Cambridge.

[42] Ruppert, D. & Carroll, R.J. (2000), Spatially-adaptive penalties for spline fitting, *Australian & New Zealand Journal of Statistics*, **42**, 205-223.

[43] Scheffé, H. (1959), *The Analysis of Variance*. New York: Wiley.

[44] Searle, S.R. (1971), *Linear Models*, New York: Wiley.

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

[46] Tibshirani, R.J. (1989), Noninformative priors for one parameter of many, *Biometrika*, **76**, 604-608.

[47] van der Vaart, A.W. (2005), *Asymptotic Statistics*. New York: Cambridge.

[48] Wahba, G. (1975), Smoothing noisy data with spline functions, Numerical Mathematics, **24**, 383-393.

[49] Wahba, G. (1990), *Spline Models for Observational Data* , Philadelphia: SIAM.

[50] Wand, M. (2003), Smoothing and mixed models, *Computational Statistics*, **18**, 223-249.

[51] Wheeler, C.H. (2003), Evidence on agglomeration economies, diseconomies, and growth, *Journal of Applied Econometric*, **18**, 79-104.

[52] Wheeler, C.H. (2003), U.S. Counties 1998. http://qed.econ.queensu.ca/jae/2003-v18.1/wheeler.

[53] White, H. (1982), Maximum likelihood estimation of misspecified models, *Econometrica*, **50**, 1-25.

[54] Wolter, K.M. (1985), *Introduction to Variance Estimation*. New York: Springer-Verlag.

[55] Ying, Z. (1993), Maximum likelihood estimation of parameters under a spatial sampling scheme, *Annals of Statistics*, **21**, 1567-1590.

[56] Zhang, H. (2004), Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics, *Journal of the American Statistical Association*, **99**, 250-261.

[57] Zhang, H. & Zimmerman, D.L. (2005), Towards reconciling two asymptotic frameworks in spatial statistics, *Biometrika*, **92**, 921-936.

[58] Zhu, Z. & Stein, .L. (2005), Spatial sampling design for parameter estimation of the covariance function, *Journal of Statistical Planning and Inference*, **134**, 583-603.

[59] Zimmerman, D.L. & Harville, D.A. (1991), A random field approach to the analysis of field-plot experiments and other spatial experiments. *Biometrics*, **47**, 223-239.