Some Contributions to Statistical Inference on Small Sample Size Data: Small Area Estimation and High Dimension Low Sample Size Data Analysis

by

Hee Cheol Chung

(Under the Direction of Gauri Sankar Datta and Jeongyoun Ahn)

Abstract

This dissertation consists of four chapters regarding statistical inference on small sample size data. The first three chapters include small area estimation studies, and the last chapter addresses the high-dimensional outlier detection problem.

In the first two chapters, small area estimation models that account for heteroscedastic random effects are explored under the Bayesian framework. In particular, in Chapter 1, we study a hierarchical Bayes random regression coefficients model which accounts for the heteroscedasticity as an appropriate quadratic function of covariates. Chapter 2 considers spatially correlated random effects. Hierarchical Bayes spatial models based on four different autocorrelation structures are introduced to capture the extra variabilities caused by spatial dependence.

The third chapter of the dissertation studies measurement error models in small area estimation. In many cases, area-level models benefit from auxiliary variables that are observed with random errors such as covariates that are estimates drawn from another survey. The uncertainty in such covariates can be accounted for by fitting measurement error models. We examine and contrast two types of measurement error models with the alternative of simply ignoring the sampling errors in the covariates.

The last topic relates to high dimensional outlier detection problem. Specifically, we consider the case when the number of variables is much larger than the sample size. A randomization test called the subspace rotation test is proposed to conduct hypothesis tests for potential outliers. We justify the subspace rotation test by showing the unbiasedness of the distribution estimator and finite sample exactness. In the context of outlier detection, we also show that the power of the subspace rotation test converges to one as the dimension increases.

INDEX WORDS: High dimension low sample size data analysis, Outlier detection, Measurement error, Random regression coefficients, Randomization test, Small area estimation, Spatial modeling, Subspace rotation

Some Contributions to Statistical Inference on Small Sample Size Data: Small Area Estimation and High Dimension Low Sample Size Data Analysis

by

Hee Cheol Chung

B.A., Yonsei University, Republic of Korea, 2009

B.S., Yonsei University, Republic of Korea, 2009

M.A., Yonsei University, Republic of Korea, 2014

A Dissertation Submitted to the Graduate Faculty

of The University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2020

Some Contributions to Statistical Inference on Small Sample Size Data: Small Area Estimation and High Dimension Low Sample Size Data Analysis

by

Hee Cheol Chung

Approved:

Major Professors:	Gauri Sankar Datta
	Jeongyoun Ahn
Committee:	Daniel Hall
	Abhyuday Mandal
	Paul Schliekelman

Electronic Version Approved:

Ron Walcott
Interim Dean of the Graduate School
The University of Georgia
May 2020

Dedication

This dissertation is dedicated to my father and mother, Key Young Chung and Kae Ja Kim.

Acknowledgments

This dissertation has been made possible through the support of many people. First and foremost, I would like to express my deepest gratitude to my doctoral advisors, Dr. Gauri Sankar Datta and Dr. Jeongyoun Ahn, for their guidance. Without their endless encouragement and support, this dissertation could never have been completed. I am confident that the experiences I had with them will be invaluable assets for my academic career.

I would like to thank Dr. Daniel Hall, Dr. Abhyuday Mandal and Dr. Paul Schliekelman for serving on my dissertation committee. Their insightful comments have helped to improve my research. I would like to extend my thanks to Dr. T.N. Sriram, who always encouraged and believed in me. In addition, I would express my sincerest appreciation to Dr. Cheolwoo Park for giving me the opportunity to join this department.

I was fortunate enough to have great collaborators. I would like to thank Dr. William Bell, Dr. Carolina Franco, Dr. Jerry Maples and Dr. Yongho Jeon for helping me broaden my research horizons. Also, I am grateful to Scott Markley and Taylor Hafley for inviting me to be a co-author.

Last but certainly not least, I wish to thank my wife Bowon Yoon and lovely daughter Ina Chung. They obviously made a significant contribution to this dissertation.

This dissertation research is indirectly supported through an IPA from the Centers for Disease Control and Prevention, Center for Global Health Division of Global HIV and TB, and partially supported through the grant 2019-67023-29672 from the National Institute of Food and Agriculture.

Table of Contents

List of Figures

List of Tables

Chapter 1

Area-Level Random Regression Coefficients Model for Small Area Estimation

1.1 Introduction

Small area estimation has become an important topic in survey sampling. Both the public and private sectors during the last forty years felt an acute need to produce reliable estimates for many important economic, health or social characteristics for various segments of a population. For example, the U.S. Department of Education is required to produce reliable estimates of the poverty rates of school-age children (5-17 years old) for all the counties. The U.S. Department of Health and Human Services is in need to accurately estimate health insurance access for different population subgroups defined by age, race, sex and geography.

Sample surveys conducted to produce estimates of economic characteristics are usually targeted to achieve the desired accuracy, such as a margin of error for the estimate, of the estimates for the entire population or domain at the national level. The data available from a survey for sub-national domains or sub-domains which are subsets of the surveyed population can be small, or even non-existent for some domains. Consequently, sub-national estimates that are calculated based on the corresponding sub-national sample only, if there is any, are subject to large sampling variability. These estimates, referred to as the "direct" estimates in the small area estimation literature, often possess such large variability that they cannot be reliably used to make official policy decisions.

Populations for which estimates of characteristics are needed, but do not have adequate samples to produce reliable direct estimates, are called small areas, or small domains. Small areas are often formed by demographic and/or geographic division of the population. More

precise estimates for a common characteristic of many small areas can be constructed by "borrowing strength" from the other direct estimates and related auxiliary variables. Borrowing of information from the direct estimates of other small areas and related auxiliary variables to produce reliable "indirect" small area estimates is pursued through shrinkage estimation. A shrinkage estimator is obtained by shrinking a direct estimate toward another estimate of that characteristic, formed by using data on auxiliary variable and direct estimates from the other small areas.

Stein (1956) in a seminal paper introduced shrinkage estimation of a multivariate normal mean vector, where each component is measuring similar characteristic for one of the groups. Stein (1956) and James and Stein (1961) explicitly showed under sum of squared error loss that suitable shrinkage estimator of the mean vector can be obtained which is more precise than the direct or the standard estimator under appropriate balanced conditions.

Shrinkage estimators became immensely popular when in a series of articles Efron and Morris put forward an interesting empirical Bayes (EB) interpretation (see, for example, Efron and Morris 1973). Their EB interpretation of the shrinkage estimators provided a transparent and heuristic justification of the domination result established by James and Stein (1961). Moreover, since the EB method can be used to develop shrinkage estimators even when the balanced conditions needed for the explicit domination are not realized, it increases the utility and popularity of shrinkage estimators in applications.

In multi-level models, such as small area estimation models, as an alternative to EB method, a hierarchical Bayes (HB) method is well-suited to produce estimates of multiple sub-population means, which measure similar population characteristics across areas. An HB method is really a Bayesian method, where the prior distribution is expressed in multiple steps. To be specific, the prior distribution of the related parameters (for example, the sub-population means) are expressed in multiple or hierarchical steps. In the first step of the prior distribution hierarchy, conditional on some common parameters, a joint distribution of the true population means is specified, and in the second step, a prior distribution on

the common parameters is specified. This hierarchy can continue beyond two steps. If all these prior distributions are available, we can pursue an HB method to estimate the related parameters or any function of them.

Often in many applications, the main difficulty is to specify the prior distribution on the common parameters appearing in the last stage of an HB model. While Bayesians may use a vague prior at this stage, if necessary, other statisticians estimate these parameters from the data and use them to construct empirical Bayes (EB) estimators. While the shrinkage estimators based on the EB method may have explicit expressions, the HB estimators usually lack such expressions and are computed numerically. However, both the HB and EB estimators benefit from shrinking the direct estimators in producing more accurate estimators for all the parameters, assessed via frequentist risk (based on the sampling distribution of the direct estimators as given by the sampling part of the Fay-Herriot model below) resulting from the sum of squared error loss.

In a pioneering article, Fay and Herriot (1979) suggested an important extension of Stein's shrinkage estimation using Efron and Morris's EB proposal for estimation of small area characteristics. Many authors subsequently developed the fully Bayesian (that is, HB) version of the Fay-Herriot model (see, for example, Datta et al. 1991, Datta et al. 1996, Datta et al. 2005). In this work, we consider an area-level random regression coefficients model which includes both fixed and random regression coefficients. While the original Fay-Herriot model is based on the assumption of constant variability of the small area population means, this general model is more appropriate if the small area means have unequal variances. We pursue an HB approach by specifying a family of noninformative priors and establish posterior propriety. We note that this model includes the Fay-Herriot model as a special case.

In Section 1.3, we compare these two models by applying these to estimate four-person family median incomes for all fifty American states and the District of Columbia. We conduct a small simulation study in Section 1.4 to further our comparison of these two models. Section

1.5 provides a few concluding remarks. In Section 1.6, we provide a proof of the propriety of the posterior density for the random regression coefficients model that resulted from a noninformative improper prior density for the model parameters.

1.2 Fay-Herriot model and random regression coefficients model

Suppose there are m small areas, and, for a given characteristic of interest, we want to estimate θ_i, which we interpret as the mean of the characteristic for the ith small area. Suppose data from a survey have been summarized, and $Y_1, \ldots, Y_m$ are the summary statistics that we refer to as direct estimates of θ_i, $i = 1, \ldots, m$. Many direct estimates do not meet the accuracy requirement due to their small sample sizes. Often in small area estimation, in addition to the direct estimates, data are also available for auxiliary variables that are related to the area characteristic of interest. To develop more reliable estimates of the population small area means $\theta_1, \ldots, \theta_m$, Fay and Herriot (1979) proposed the following hierarchical model for EB prediction of θ_i, $i = 1, \ldots, m$.

(I) **Sampling model:** $Y_i|\theta_1, \ldots, \theta_m, \beta, \sigma_v^2 \overset{ind}{\sim} N(\theta_i, D_i)$, $i = 1, \ldots, m$;

(II) **Linking model:** $\theta_i|\beta, \sigma_v^2 \overset{ind}{\sim} N(x_i^{\mathrm{T}}\beta, \sigma_v^2)$, $i = 1, \ldots, m$,

where the D_i, $i = 1, \ldots, m$, are known sampling variances of the direct estimators Y_i, $i = 1, \ldots, m$, and the $x_i = (1, x_{i2}, \ldots, x_{ip})^T$ is the $p \times 1$ vector of covariates associated with Y_i. The regression parameter $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T$ and the model error variance σ_v^2 are the model parameters. Let $Y = (Y_1, \ldots, Y_m)^{\mathrm{T}}$, and $X = [x_1, \ldots, x_m]^{\mathrm{T}}$. This model can be expressed as a linear mixed effect model by writing

$$Y = X\beta + v + e,$$

where $e = (e_1, \ldots, e_m)^{\mathrm{T}} \sim N_m(0, D)$, $D = \mathrm{diag}\{D_i\}_{i=1}^m$ and $v = (v_1, \ldots, v_m)^{\mathrm{T}} \sim N_m(0, \sigma_v^2 I_m)$.

The above hierarchical model can also be written as

$$\begin{aligned} Y_i &= \theta_i + e_i, \quad \theta_i = \beta_{i1} + \sum_{j=2}^{p} \beta_j x_{ij}, \quad i = 1, \ldots, m, \\ \Rightarrow Y_i &= \beta_{i1} + \sum_{j=2}^{p} \beta_j x_{ij} + e_i, \; i = 1, \ldots, m, \end{aligned} \tag{1.1}$$

where the $\theta_1, \ldots, \theta_m$ are expressed as a sum of a *random* intercept β_{i1} and a linear function of the covariates $x_{i2}, \ldots, x_{ip}$ with *non-random* regression coefficients $\beta_2, \ldots, \beta_p$. The random intercept accounts for error in explaining the θ_i by the linear regression on the covariates. The Fay-Herriot (FH) model assumes that $\beta_{11}, \ldots, \beta_{m1}$ are independently and identically distributed as $N(\beta_1, \sigma_v^2)$.

In their application of shrinkage estimation of per capita income for small places, Fay and Herriot (1979) used the EB approach. Assuming the model parameters β and σ_v^2 were known, based on the hierarchical model in (I) and (II) above, they first obtained the Bayes estimators of $\theta_1, \ldots, \theta_m$. The Bayes "estimator" of θ_i, given by

$$\theta_i^B = Y_i - \frac{D_i}{D_i + \sigma_v^2}(Y_i - x_i^{\mathrm{T}}\beta),$$

shrinks the direct estimator Y_i to the regression function $x_i^{\mathrm{T}}\beta$. However, since the model parameters, β and σ_v^2, are unknown, these Bayes estimators of $\theta_1, \ldots, \theta_m$ could not be used. To obtain usable version from these estimators, they estimated the model parameters from the marginal distribution of $Y_1, \ldots, Y_m$. They used these estimates in place of β and σ_v^2 in the Bayes "estimators" of $\theta_1, \ldots, \theta_m$, resulting in the EB estimators of $\theta_1, \ldots, \theta_m$.

As an alternative to EB estimation, some practitioners use HB estimation. To deal with the unknown model parameters, a hierarchical Bayesian assigns a prior distribution to them. The resulting posterior distribution of the unknown model parameters is used to integrate them out from the Bayes estimators, described above. The HB estimation method explicitly accounts for the estimation error of the model parameters. However, the plug-in EB method suffers from underestimation of associated measures of uncertainty of the EB estimators. This is due to the fact that the method does not automatically account for estimation error of the

model parameters. Extensions to EB have attempted to incorporate the extra uncertainty due to estimating the parameters (Morris, 1983). It is well documented in the literature that EB predictors are also the empirical best linear unbiased predictors (EBLUP) of the small area means. Accurate approximation of the mean squared error (MSE) of the EBLUPs and estimation of the MSE have been extensively discussed in Rao and Molina (2015, Ch. 5); see also Prasad and Rao (1990), Datta and Lahiri (2000) and Datta et al. (2005).

While the HB method facilitates the use of prior information for the model parameters, in many applications, no specific proper prior distribution may be available. Even in such cases the method is found to be beneficial by using vague or diffuse or noninformative priors. Due to a rapid growth in computing capacity in the past few decades, implementation of HB methods has been relatively easily achieved. As a result, the HB methods, even with diffuse priors, are found to be viable for many complex scientific studies. In a parallel development to model-based estimation in survey sampling, the HB methods have also gained popularity in small area estimation.

We now introduce the HB version of the FH model based on a class of noninformative priors that has been extensively used for the model parameters.

(I) **Sampling model:** $Y_i|\theta_1,\ldots,\theta_m,\beta,\sigma_v^2 \overset{ind}{\sim} N(\theta_i, D_i)$, $i=1,\ldots,m$;

(II) **Linking model:** $\theta_i|\beta,\sigma_v^2 \overset{ind}{\sim} N(x_i^{\mathrm{T}}\beta,\sigma_v^2)$, $i=1,\ldots,m$,

(III) **A popular noninformative improper prior density function for β and σ_v^2:**

$$\pi(\beta,\sigma_v^2) = (\sigma_v^2)^{-\alpha} \quad \text{for } \beta\in\mathbb{R}^p, 0<\sigma_v^2,$$

where $1-(m-p)/2<\alpha<1$.

The above noninformative prior density function does not integrate to 1, unlike any regular probability density function (pdf). Thus, it is termed as an improper prior, and it is required to verify that the resulting posterior pdf will be a proper pdf, that is, it will integrate to 1 after being appropriately normalized. Adapting arguments of Theorem 2 in Datta and Smith

(2003), the resulting posterior pdf of β, σ_v^2 is a proper pdf provided $1-(m-p)/2 < \alpha < 1$. The prior corresponding to $\alpha = 0$, which is a uniform prior for the model parameters, is a popular choice. We use this prior in our illustrative application.

In the FH model in (1.1), the random intercept term is intended to account for variability among the $\theta_1, \ldots, \theta_m$ that is left unexplained by the covariates. It allows for a constant variability of the θ_i around its mean $x_i^{\mathrm{T}}\beta = \beta_1 + \sum_{j=2}^{p} \beta_j x_{ij}$. In that spirit, it is conceivable that the variability of the θ_i's around the regression line may not be constant, but it may also be dependent on covariates. This leads to the random regression coefficients (RRC) model. For unit-level small area estimation models Prasad and Rao (1990) suggested the RRC model (see also Datta and Ghosh, 1991). These random coefficients can be viewed as similar to interaction effects between the covariates and the small areas. For the area-level data, we propose below a RRC model, which can be viewed as a generalization of the FH model. Suppose $q-1$ $(q \leq p)$ covariates have random regression coefficients in the mean function of θ_i. Without loss of generality, let $x_{i2}, \ldots, x_{iq}$ be these $q-1$ covariates and denote $z_i = (1, x_{i2}, \ldots, x_{iq})^{\mathrm{T}}$. Then, for $q = p$, this model can be written as the following linear mixed effect model:

$$\begin{aligned} Y_i &= \theta_i + e_i, \quad \theta_i = x_i^{\mathrm{T}}\beta + z_i^{\mathrm{T}}u_i, \quad i = 1, \ldots, m, \\ \Rightarrow Y_i &= (\beta_1 + u_{i1}) + \sum_{j=2}^{p} (\beta_j + u_{ij}) x_{ij} + e_i, \; i = 1, \ldots, m, \end{aligned} \tag{1.2}$$

where $u_i \overset{iid}{\sim} N_q(0, \Sigma_v)$, Σ_v is a $q \times q$ positive definite matrix and $(\beta_j + u_{ij})$, $j = 2, \ldots, p$, are the random regression coefficients for the ith area. We introduce the HB version of the RRC model below.

(I) **Sampling model:** $Y_i | \theta_1, \ldots, \theta_m, \beta, \Sigma_v \overset{ind}{\sim} N(\theta_i, D_i)$, $i = 1, \ldots, m$,

(II) **Linking model:** $\theta_i | \beta, \Sigma_v \overset{ind}{\sim} N(x_i^{\mathrm{T}}\beta, z_i^{\mathrm{T}}\Sigma_v z_i)$, where $z_i = (1, x_{i2}, \ldots, x_{iq})^{\mathrm{T}}$, $i = 1, \ldots, m$ and Σ_v is a $q \times q$ positive definite matrix,

(III) **A noninformative prior for the model parameters:**

$$\pi(\beta, \Sigma_v) = g(\Sigma_v), \tag{1.3}$$

where g is a suitable function defined below. We introduce a noninformative prior on the model parameters below. Let $\Psi = \Sigma_v^{-1}$ be the precision matrix and define $\tau_{jk} = \psi_{jk}/(\psi_{jj}\psi_{kk})^{1/2}$, where $\Psi = \{\psi_{jk}\}_{1\le j,k\le q}$. Then, $\tau_{11}, \ldots, \tau_{qq}$ are all 1. Let $T = \{\tau_{jk}\}_{1\le j,k\le q}$. We propose a noninformative prior given by

$$\pi(\beta, \psi_{11}, \ldots, \psi_{qq}, \tau_{12}, \ldots, \tau_{1q}, \ldots, \tau_{q-1,q}) = \prod_{j=1}^{q} \psi_{jj}^{-a_j/2}\ I(T \text{ is p.d.}), \tag{1.4}$$

where the $a_1, \ldots, a_q$ (> 2) are suitably chosen, and $I(\cdot)$ is the indicator function. Here, $\beta \in \mathbb{R}^p$, and p.d. stands for positive definite. We note that the precision matrix is parameterized in terms of its diagonal elements and τ_{jk}, $1 \le j < k \le q$. We prove the following theorem in Section 1.6.

Theorem 1. *For $a_1 > 2, \ldots, a_q > 2$ and under some upper bound conditions on $a_1, \ldots, a_q$ (noted explicitly in the proof), the posterior pdf corresponding to the noninformative prior in (1.4) is a proper pdf.*

Remark 1. *We note that the HB FH model is a special case with $q = 1$ and $a_1 = 4 - 2\alpha$.*

In this study, we pursued an HB approach for the RRC model. However, a frequentist approach to this problem via EB approach or a mixed model formulation can be pursued. Frequentist estimation of the variance parameters may not be easy, as it is not obvious how to define $q(q+1)/2$ estimating equations for the components of Σ_v. Also, the maximum likelihood (ML) approach may present a hurdle since the likelihood (residual or profile likelihood) function may be quite flat, which would make convergence of the ML computing rather slow.

	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\sigma}_{v,11}$	$\hat{\sigma}_{v,22}$	$\hat{\sigma}_{v,33}$	$\hat{\rho}_{12}$	$\hat{\rho}_{13}$	$\hat{\rho}_{23}$
FH_0	-0.037	0.030	0.791	0.184	–	–	–	–	–
$\mathrm{FH}_{0.75}$	-0.042	0.032	0.789	0.160	–	–	–	–	–
$\mathrm{RRC}_{2.1}$	-0.041	0.056	0.774	0.142	0.009	0.015	0.053	0.007	-0.014
$\mathrm{RRC}_{2.5}$	-0.039	0.085	0.758	0.133	0.035	0.044	0.130	-0.021	-0.074

Table 1.1: Posterior means of the model parameters

1.3 Four-person family median income estimation

In this section, we consider an application of estimating four-person median incomes by states, i.e., the median income of four-person households for fifty states and Washington, D.C. The U.S. Department of Health and Human Services needs accurate estimates of medians for the state distributions of income of four-person families to implement a cash welfare program to provide energy assistance benefits to low-income American families. For over twenty years through the mid-1990s, the U.S. Census Bureau continued to calculate these state-level estimates annually using the Current Population Survey (CPS) data. The statewide samples available from the CPS to compute direct estimates of state characteristics are not sufficiently accurate for some states due to the smallness of available samples.

Let Y_i be the direct estimate of 1989 four-person median income of the ith state. We consider two auxiliary variables x_{i2} and x_{i3}, where x_{i2} is the median income for 1979 collected from the 1980 census and x_{i3} is the adjusted census median income. Adjusted census median incomes are obtained by utilizing 1979 and 1989 per capita incomes (PCI) from the Bureau of Economic Analysis of the U.S. Department of Commerce such that $x_{i3} = (\mathrm{PCI}_{i,1989}/\mathrm{PCI}_{i,1979})x_{i2}$, $i = 1, \ldots, 51$. Thus we have $x_i = (1, x_{i2}, x_{i3})^{\mathrm{T}}$, $i = 1, \ldots, 51$. Here $p = 3$, and we take $q = p$. Then the model parameters are given by β (a 3×1 vector) and Σ_v (a 3×3 p.d. matrix). Since the direct estimates and covariates are large values, we

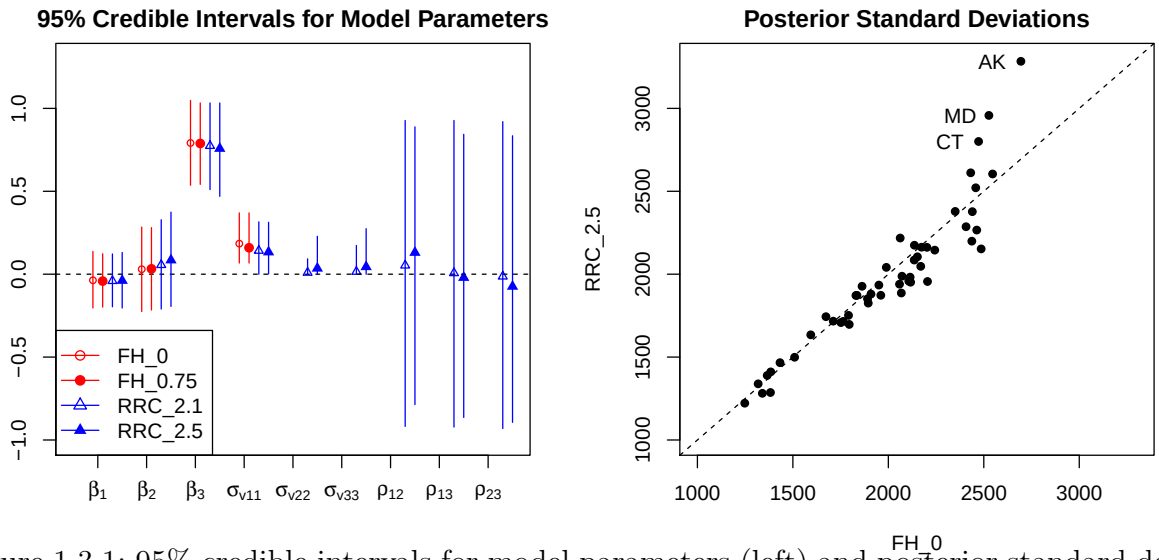


Figure 1.3.1: 95% credible intervals for model parameters (left) and posterior standard deviations of θ_i's (right) under the FH_0 and $\text{RRC}_{2.5}$ models.

standardize the direct estimates and scale auxiliary variables to prevent numerical overflow (underflow). We fit the HB FH model and the HB RRC model to the data using rstan (Stan Development Team, 2018). We consider two prior pdfs for the FH model, $\alpha = 0, 0.75$, and for the RRC model we consider two prior pdfs with $a_j = 2.1, 2.5$, $j = 1, 2, 3$ (close to their lower bound for propriety). We denote results for these models as FH_0, $\text{FH}_{0.75}$, $\text{RRC}_{2.1}$ and $\text{RRC}_{2.5}$, respectively. Posterior means of the model parameters are given in Table 1.1. As we can see from the left panel of Figure 1.3.1, the four models result in very similar posterior distributions in that a large portion of their 95% credible intervals overlap for the common parameters. Also, we can see that the posterior means of $\sigma_{v,22}$ and $\sigma_{v,33}$ are close to 0, and their lower bounds of the intervals get almost 0. This indicates that regression coefficients do not vary much across the states.

Figure 1.3.2 illustrates 95% credible intervals of the predictions under the FH_0 and $\text{RRC}_{2.5}$. Red (blue) lines and circles (triangles) represent 95% credible intervals and posterior means of the four-person family median incomes under the FH_0 ($\text{RRC}_{2.5}$) model. Black

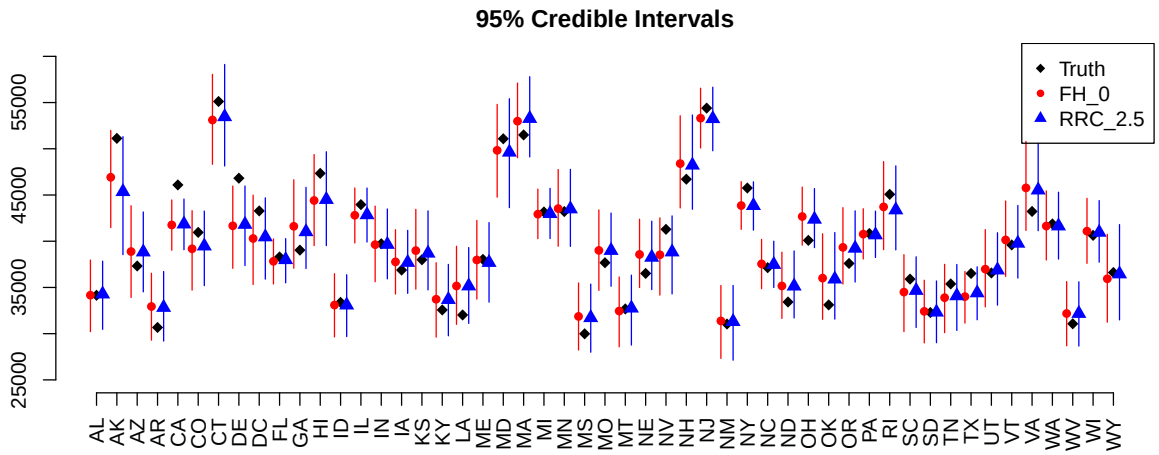


Figure 1.3.2: 95% credible intervals under the FH_0 and $RRC_{2.5}$ models.

Estimate	AAD	ASD	AARD	ASRD	AL
Direct	2928.824	13811122	0.074	0.008	11424.650
FH_0	1507.303	3682226	0.038	0.002	7856.558
$FH_{0.75}$	1438.580	3386631	0.036	0.002	7614.605
$RRC_{2.1}$	1405.772	3310202	0.035	0.002	7621.411
$RRC_{2.5}$	1435.107	3614879	0.035	0.002	7815.770

Table 1.2: Model comparison criteria.

diamonds are the true median incomes for 1989 collected from the 1990 census. Both models underestimate the median incomes of California and Delaware, and the credible intervals miss (fall below) the true values.

We compare both the models according to several criteria. We first compare posterior standard deviations which are illustrated in the right panel of Figure 1.3.1. The $RRC_{2.5}$ model shows smaller posterior standard deviations that the FH_0 model in 31 states. However, for the states of Alaska, Connecticut and Maryland, the FH_0 model has much smaller

posterior standard deviations. On average, $\text{RRC}_{2.5}$ showed 1.3% smaller posterior standard deviations than the FH_0 model. We also compare the average length (AL) of 95% credible intervals and several deviation criteria. For the true median income θ_i for the ith state, let $\hat{\theta}_i$ be any estimate of θ_i, $i = 1, \ldots, 51$. We compute average absolute deviation (AAD) $51^{-1} \sum_{i=1}^{m} |\hat{\theta}_i - \theta_i|$, average squared deviation (ASD) $51^{-1} \sum_{i=1}^{51} |\hat{\theta}_i - \theta_i|^2$, average absolute relative deviation (AARD) $51^{-1} \sum_{i=1}^{51} |(\hat{\theta}_i - \theta_i)/\theta_i|$ and average squared relative deviation (ASRD) $51^{-1} \sum_{i=1}^{51} \{(\hat{\theta}_i - \theta_i)/\theta_i\}^2$. These deviation criteria are listed in Table 1.2. We see a clear improvement for the model-based estimators over the direct estimators. However, there is only a slight improvement for the RRC models over the FH models. Also, we have seen from Table 1.1 that estimated random effect variances are very small except for the random intercept, $\hat{\sigma}_{v,11}$. This indicates that regression parameters do not vary much across the states. Thus, in this example, it is reasonable to use the FH model since the gain from using the more complicated model is rather limited.

1.4 A simulation study

To investigate the effectiveness of the RRC model further, we conducted a simulation study. In our simulation, we used the data for the covariates from the median income application in Section 1.3. To reduce computational burden, in our simulation we used $q = 2$ (instead of $q = 3$ that was in the data analysis) and $p = 3$. That is, we used only one random regression coefficient, corresponding to the most significant covariate found in the analysis. Specifically, we set $z_i = (1, x_{i3})^{\text{T}}$. We considered two scenarios; data are generated from the FH model and the RRC model, respectively.

In order to simulate from a model that will closely mimic the reality, we need some knowledge about model parameters for each model. To suitably choose these model parameter values to reflect the reality, we seek guidance from our data analysis in Section 1.3. Our data analysis showed a marginal improvement of the $\text{RRC}_{2.5}$ model over the FH_0 model. This is due to very small variances of the random regression coefficients. In order to have non-negligible

	β_1	β_2	β_3	$\sigma_{v,11}$	$\sigma_{v,22}$	ρ_{12}
FH	-0.037	0.027	0.793	0.185	–	–
RRC	-0.044	0.070	0.755	0.062	14.706	0.033

Table 1.3: Model parameters used for simulation.

contribution of the random regression coefficient in our simulation study, we considered a large value for the variance of the random regression coefficient compared to that of the random intercept term. We first fit the models to the data and obtained the posterior means of the model parameters. We used these values as guidance to our choice of model parameters in the simulation study. In our simulations from the $\text{RRC}_{2.5}$ model, we set $\sigma_{v,11} = \hat{\sigma}_{v,22}$, $\sigma_{v,12} = 10\hat{\sigma}_{v,12}$ and $\sigma_{v,22} = 100\hat{\sigma}_{v,11}$, where $\hat{\sigma}_{v,11}$, $\hat{\sigma}_{v,12}$ and $\hat{\sigma}_{v,22}$ are obtained by fitting the $\text{RRC}_{2.5}$ model to the four-person median income data with $z_i = (1, x_{i3})^{\text{T}}$. This allows the random regression coefficient to dominate the model error variance. It also preserves the correlation between the random intercept and random regression coefficient. Finally, we also multiplied the sampling variances by 100 in order to approximately preserve the ratio between model and sampling error variances. The parameter values are given in Table 1.3. To compare the performance of the FH and RRC models we considered 30 replicated data sets and computed the deviation criteria (introduced in the last section) and overall coverage probabilities of 95% credible intervals by averaging both over the areas and the data sets. Table 1.4 lists these values.

When data are generated from the FH model, the FH_0 model shows slightly better performance than the $\text{RRC}_{2.5}$ model but the differences are not significant. Under the true RRC model, the $\text{RRC}_{2.5}$ model outperforms the FH_0 model in all the deviation criteria. Also, it has larger coverage probabilities while the posterior standard deviations are approximately 13% smaller on average, which are not shown here. In conjunction with our finding in the

True Model	Fitted Model	AAD	ASD	AARD	ASRD	Coverage
FH	FH_0	2.072	7.040	1.577	7.236	0.952
	$RRC_{2.5}$	2.077	7.184	1.557	8.825	0.944
RRC	FH_0	2.232	8.993	4.003	92.246	0.936
	$RRC_{2.5}$	2.079	8.391	2.698	27.137	0.959

Table 1.4: Averages of various deviation criteria and coverage probabilities.

real data analysis given in the previous section, the RRC model is as much good as the FH model when all the states share common regression parameters. This is an expected result since the FH is a sub-model of the RRC model. The RRC model can fit any data generated under the FH model, although it requires an extra computing effort. However, when the regression coefficients vary across the states, the FH model will be a misspecified model. As a result, the reported posterior standard deviations for the small area means for some areas will be underestimates and for other areas, they will be overestimates. Similarly, the credible intervals of the small area means for some areas will be overly short and for some overly long.

1.5 Conclusions

In this chapter, we extended the FH model by replacing the constant model error variance assumption by a heteroscedastic variance assumption. The heteroscedasticity that we modeled results from the randomness of the regression coefficients in modeling the small area population means. The heteroscedasticity induced by this assumption models the variance by an appropriate quadratic function of the covariates. We developed a noninformative HB procedure for the RRC model. We applied the FH model and the RRC model to predict four-person family median incomes for the U.S. states. This application and limited simu-

lation study demonstrate the usefulness of the more general model in better prediction of small area means. The RRC model may also be a good check on the standard FH model to see if all areas follow the same regression function (regression plane). Sometimes there are subsets of areas that may have a different regression relationship with the covariates, such as due to a missing interaction effect (area and covariate). The RRC model can pick up these departures and provide a more robust modeling framework.

1.6 Technical details

Proof of Theorem 1. In our RRC model, $Y_i = \theta_i + e_i$, $\theta_i = x_i^{\mathrm{T}}\beta + z_i^{\mathrm{T}}v_i$, $i = 1, \ldots, m$, where $e_1, \ldots, e_m$, $v_1, \ldots, v_m$ are all independently distributed with $e_i \sim N(0, D_i)$, $i = 1, \ldots, m$, $x_i = (1, x_{i2}, \ldots, x_{ip})^{\mathrm{T}}$, $p \geq 2$, $z_i = (1, x_{i2}, \ldots, x_{iq})^{\mathrm{T}}$, $q \geq 2$, $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$, $v_i \overset{iid}{\sim} N(0, \Psi^{-1})$ and Ψ is a $q \times q$ positive definite precision matrix. We assume that $q \leq p$ and, if necessary, $Y = (Y_1, \ldots, Y_m)^{\mathrm{T}}$, $\tilde{x}_l = (x_{1l}, \ldots, x_{ml})^{\mathrm{T}}$, $l = 2, \ldots, p$, will be standardized with suitable locations and scales. We note that $\Psi = \{\psi_{jk}\}_{1 \leq j,k \leq q}$ and $T = \{\tau_{jk}\}_{1 \leq j,k \leq q}$ are defined as in the main text. We consider the noninformative prior given in (1.4).

Without loss of generality, we arrange the data set so that the first p observations lead to a nonsingular design matrix X_* given by $X_* = [1_p, \tilde{x}_{*2}, \ldots, \tilde{x}_{*p}]$, where 1_p is a $p \times 1$ vector of ones, $\tilde{x}_{*l} = (x_{1l}, \ldots, x_{pl})^{\mathrm{T}}$, $l = 2, \ldots, p$. We use these p observations to estimate β by $b = X_*^{-1}Y_*$, where $Y_* = (Y_1, \ldots, Y_p)^{\mathrm{T}}$. Let $Y_0 = (Y_{p+1}, \ldots, Y_m)^{\mathrm{T}}$. Note that $b|\beta, \Psi \sim N_p(\beta, \Phi(\Psi))$, where Φ is a $p \times p$ positive definite matrix depending on Ψ. Under this model and the prior, the joint posterior pdf of β and $\psi_{11}, \ldots, \psi_{qq}, \tau_{jk}$, $1 \leq j < k \leq q$ is given by

$$\begin{aligned}
\pi(\beta, \psi_{11}, \ldots, &\psi_{qq}, \tau_{jk}, 1 \leq j < k \leq q|Y) = cN_p(b|\beta, \Phi(\Psi)) \prod_{i=p+1}^{m} (D_i + z_i^{\mathrm{T}}\Psi^{-1}z_i)^{-1/2} \\
&\times \exp\left\{-\frac{(Y_i - x_i^{\mathrm{T}}\beta)^2}{2(D_i + z_i^{\mathrm{T}}\Psi^{-1}z_i)}\right\} \prod_{j=1}^{q} \psi_{jj}^{-a_j/2}\ I(T \text{ is p.d.}) \\
&\leq cN_p(\beta|b, \Phi(\Psi)) \prod_{i=p+1}^{m} (D_i + z_i^{\mathrm{T}}\Psi^{-1}z_i)^{-1/2} \prod_{j=1}^{q} \psi_{jj}^{-a_j/2}\ I(T \text{ is p.d.}), \qquad (1.5)
\end{aligned}$$

where $Y = (Y_1, \ldots, Y_m)^{\mathrm{T}}$. Integrating both sides of (1.5) with respect to β, we get

$$\pi(\psi_{11}, \ldots, \psi_{qq}, \tau_{jk}, 1 \le j < k \le q | Y) \\ \le c \prod_{i=p+1}^{m} (D_i + z_i^{\mathrm{T}} \Psi^{-1} z_i)^{-1/2} \prod_{j=1}^{q} \psi_{jj}^{-a_j/2} \; I(T \text{ is p.d.}) \tag{1.6}$$

Note that $z_i^{\mathrm{T}} \Psi^{-1} z_i \ge z_{ij}^2 / \psi_{jj}$, $j = 1, \ldots, q$. From the $m - p$ small areas, select n_j small areas for which $z_{ij}^2 > 0$. We need suitable conditions on $n_1, \ldots, n_q$, where $\sum_{j=1}^{q} n_j = m - p$. By these observations we have that

$$\prod_{i=p+1}^{m} (1 + z_i^{\mathrm{T}} \Psi^{-1} z_i)^{1/2} \le C (1 + \psi_{11}^{-1})^{-n_1/2} \ldots (1 + \psi_{qq}^{-1})^{-n_q/2}. \tag{1.7}$$

We now use (1.7) in (1.6) and transform ψ_{jj} to u_j by $u_j = \psi_{jj}^{-1}$, $j = 1, \ldots, q$, and keep τ_{jk}, $1 \le j < k \le q$, fixed. Then integration of both sides of (1.7) will lead to

$$\int \pi(\psi_{11}, \ldots, \psi_{qq}, \tau_{jk}, 1 \le j < k \le q | Y) d\psi_{11}, \ldots, d\psi_{qq} \\ \le C \prod_{j=1}^{q} \int_0^{\infty} u_j^{-2+a_j/2} (1 + u_j)^{-n_j/2} du_j \\ = C^* < \infty \tag{1.8}$$

provided $-2 + a_j/2 > -1$, $-2 + a_j/2 - n_j/2 < -1$, $j = 1, \ldots, q$, i.e., $a_j > 2$, $n_j > a_j - 2$, $j = 1, \ldots, q$.

Note that C^* is finite and free from τ_{jk}, $1 \le j < k \le q$. Thus, C^* will also be integrable with respect to τ_{jk}, $1 \le j < k \le q$. Consequently, subject to the conditions $a_j > 2, n_j > a_j - 2$, $j = 1, \ldots, q$, the posterior pdf for our RRC model will be proper.

□

Bibliography

Datta, G. S., Fay, R. E., and Ghosh, M. (1991), "Hierarchical and empirical multivariate Bayes analysis in small area estimation," in *Proceedings of the Seventh Annual Research Conference of the Bureau of the Census*, U.S. Department of Commerce, Bureau of the Census, vol. 7, pp. 63–79.

Datta, G. S. and Ghosh, M. (1991), "Bayesian prediction in linear models: Applications to small area estimation," *The Annals of Statistics*, 19, 1748–1770.

Datta, G. S., Ghosh, M., Nangia, N., and Natarajan, K. (1996), "Estimation of median income of four-person families: a Bayesian approach," in *Bayesian Analysis in Statistics and Econometrics*, eds. Berry, D. A., Chaloner, K. M., and Geweke, J. K., John Wiley & Sons, chap. 11, pp. 129–140.

Datta, G. S. and Lahiri, P. (2000), "A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems," *Statistica Sinica*, 10, 613–627.

Datta, G. S., Rao, J. N. K., and Smith, D. D. (2005), "On measuring the variability of small area estimators under a basic area level model," *Biometrika*, 92, 183–196.

Datta, G. S. and Smith, D. D. (2003), "On propriety of posterior distributions of variance components in small area estimation," *Journal of Statistical Planning and Inference*, 112, 175–183.

Efron, B. and Morris, C. N. (1973), "Stein's estimation rule and its competitors – an empirical Bayes approach," *Journal of the American Statistical Association*, 68, 117–130.

Fay, R. E. and Herriot, R. A. (1979), "Estimates of income for small places: an application of James-Stein procedures to census data," *Journal of the American Statistical Association*, 74, 269–277.

James, W. and Stein, C. (1961), "Estimation with quadratic loss," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, vol. 1, pp. 361–379.

Morris, C. N. (1983), "Parametric empirical Bayes inference: theory and applications," *Journal of the American Statistical Association*, 78, 47–55.

Prasad, N. G. N. and Rao, J. N. K. (1990), "The estimation of the mean squared error of small-area estimators," *Journal of the American Statistical Association*, 85, 163–171.

Rao, J. N. K. and Molina, I. (2015), *Small area estimation*, John Wiley & Sons.

Stan Development Team (2018), "RStan: the R interface to Stan," R package version 2.17.3.

Stein, C. (1956), "Inadmissibility of the usual estimator for the mean of a multivariate normal distribution," in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, vol. 1, pp. 197–206.

Chapter 2

Hierarchical Bayesian Spatial Models for Small Area Estimation

2.1 Introduction

Sample surveys have been useful tools in estimating numerous characteristics of a population of interest. However, when it comes to estimating a characteristic of a sub-population, a design-based direct estimate is usually inaccurate or sometimes nonexistent as the accessible sample size is small or no sample is available. Small areas refer to these sub-populations which lack enough sample size to produce reliable estimates. When a small area has no sample, we call it an un-sampled area. Model-based approach has been widely used in enhancing the accuracy of a small area direct estimate by borrowing information from direct estimates of other domains and auxiliary information available from other sources. For most cases, auxiliary information is obtained from other surveys and administrative data. These are called auxiliary variables or covariates. Model-based estimates are produced by suitably shrinking the direct estimates to the synthetic regression estimates based on auxiliary variables. Hence the accuracy improvement greatly depends on how much they are related to the sub-population characteristic.

Fay and Herriot (1979) proposed a very useful model to develop composite small area estimates based on direct survey estimates and synthetic estimates of a small area characteristic. Their model, which is essentially a mixed linear model, is widely known as the Fay-Herriot (FH) model in small area estimation literature. For $i = 1, \ldots, m$, let Y_i be the direct estimate of the small area characteristic θ_i obtained from a survey. Also let x_i and β be the p-component vector of covariates and $p \times 1$ regression coefficient vector, respectively.

Then the FH model can be written by

$$Y_i = \theta_i + e_i, \quad \theta_i = x_i^{\mathrm{T}}\beta + u_i, \quad i = 1, \ldots, m, \tag{2.1}$$

where sampling errors $e_1, \ldots, e_m$ are independently distributed as $e_i \overset{ind}{\sim} N(0, D_i)$, and are independent of random effects, small area effects, $u_i \overset{iid}{\sim} N(0, \sigma_v^2)$, $i = 1, \ldots, m$. This model assumes that the sampling variances D_i, $i = 1, \ldots, m$ are known but the regression parameter β and model error variance σ_v^2 are unknown quantities.

There has been extensive research on FH model and its variations. Taking the frequentist approach, Prasad and Rao (1990), Datta and Lahiri (2000) and Datta et al. (2005) derived the second-order approximation to mean squared error (MSE) of empirical best linear unbiased predictor (EBLUP) of θ_i and provided MSE estimators which are asymptotically unbiased. In the Bayesian framework, FH model in (2.1) can be expressed as the following hierarchical Bayes model:

$$Y_i|\theta_1, \ldots, \theta_m, \beta, \sigma_v^2 \overset{ind}{\sim} N(\theta_i, D_i), \quad i = 1, \ldots, m, \tag{2.2}$$

$$\theta_i|\beta, \sigma_v^2 \overset{ind}{\sim} N(x_i^{\mathrm{T}}\beta, \sigma_v^2), \quad i = 1, \ldots, m, \tag{2.3}$$

$$\pi(\beta, \sigma_v^2) \propto g(\beta, \sigma_v^2), \tag{2.4}$$

where $g(\cdot)$ is a suitably chosen function of β and σ_v^2, which expresses a prior probability density function (pdf) for these parameters. Without specifying a prior pdf as in (2.4), empirical Bayes predictor for θ_i was originally developed by Fay and Herriot (1979). Datta and Smith (2003) provided propriety of posterior distribution under the noninformative prior $\pi(\beta, \sigma_v^2) \propto 1$. The hierarchical Bayesian approach facilitates uncertainty quantification due to estimation of unknown parameters, β and σ_v^2. The uncertainty is fully captured by simulating the parameters from their posterior distributions by using Monte-Carlo methods.

In model-based small area estimation, small area effects are of great importance for capturing the remaining variabilities that are not explained by the regression model. In real applications, small areas generally involve features such as population size, age-group and education level, which might affect the variability of small area effects. Furthermore,

when an epidemic prevalence rates are of interest, it is reasonable to assume that small area effects of adjacent small areas are correlated in a certain way. In such a case, the FH model given in (2.1) or (2.2)–(2.3), which we refer to as the independent FH random-effects model, oversimplifies small area effects model by assuming independent and identical distributions over small areas or different geographical regions. Although benefit from model-based approach is substantial, it has been known that it can perform poorly under model misspecification as a domain sample size increases (Rao and Molina, 2015).

In this work, we explore small area estimation models which effectively account for heteroscedasticity and spatial dependence of the small area effects. The small area effects are modeled by four widely used spatial autocorrelation structures. These include conditional autoregressive (CAR), simultaneous autoregressive (SAR), intrinsic auto-regressive (IAR) and the spatial model suggested by Rao and Molina (2015) (SRM). We further extend the spatial models in estimating area characteristics of un-sampled small areas. By specifying a hierarchical Bayes model, which includes the five models, including the independent FH model, as special cases, we provide sufficient conditions for posterior propriety under a general family of prior distributions. With a noninformative prior on model parameters, we apply the spatial models to simulated and the Current Population Survey data. The numerical studies indicate that proposed spatial models greatly improve prediction accuracy with lesser uncertainty.

2.2 Some spatial alternatives to the independent FH model

Let $Y = (Y_1, \ldots, Y_m)^{\mathrm{T}}$ be the vector with the direct estimates of m small areas, $D = \mathrm{diag}\{D_i\}_{i=1}^m$ be the $m \times m$ diagonal matrix with the sampling variances of the direct estimates. We denote $\theta = (\theta_1, \ldots, \theta_m)^{\mathrm{T}}$ by the m-component vector with small area characteristics. Also, let $X = [x_1, \cdots, x_m]^{\mathrm{T}}$ be the $m \times p$ matrix of auxiliary variables, where $x_i \in \mathbb{R}^p$ is the $p \times 1$ vector of auxiliary variables for the ith small area. A special case of the hierarchical Bayes

model given in (2.2)–(2.4) can be expressed as

$$Y|\theta, \beta, \sigma_v^2 \sim N_m(\theta, D), \tag{2.5}$$

$$\theta|\beta, \sigma_v^2 \sim N_m(X\beta, \sigma_v^2 I_m), \tag{2.6}$$

$$\pi(\beta, \sigma_v^2) \propto 1, \tag{2.7}$$

where β is the $p \times 1$ regression coefficient vector and σ_v^2 is the model error variance. We call β and σ_v^2 model parameters. The uniform prior (2.7) on the model parameters is a popularly used non-informative prior. The resulting posterior density is proper provided that $m > p+2$ (Datta and Smith, 2003).

From (2.6), it can be seen that θ_i, $i = 1, \ldots, m$ are independently distributed with common random effects variance σ_v^2 over the small areas. However, in many cases, the area characteristic of interest is closely related to geographical factors such as population size, ethnicity, age-group and education level. In such case, due to misspecified random effects distribution (independence, in this case), inference based on the hierarchical model (2.5)–(2.7) may yield erroneous result.

Let $W = \{w_{ij}\}$, $1 \leq i, j \leq m$, be the adjacency matrix which plays an important role in capturing spatial dependency. In particular, for the $m \times m$ adjacency matrix W, $w_{ij} = 1$ if the ith and jth small areas are connected via a geographical boundary or through other consideration such as air traffic, and $w_{ij} = 0$, otherwise. Also, $w_{ii} = 0$ for $i = 1, \ldots, m$. Let $w_{i.} = \sum_{j=1}^{m} w_{ij}$ be the sum of the ith row of W and define $L = \text{diag}\{w_{i.}\}_{i=1}^{m}$. We assume that diagonal elements of L are positive and we define $\tilde{W} = L^{-1}W$. We denote by λ_i the ith largest eigenvalue of the adjacency matrix W such that $\lambda_1 \geq \cdots \geq \lambda_m$. Since W is a non-null matrix and $w_{ii} = 0$ for all the diagonals, we get as a result that $0 \in (\lambda_m, \lambda_1)$. Let I_m be the identity matrix of order m. We consider four different types of spatial dependencies: specifically, conditional autoregressive (CAR), simultaneous autoregressive (SAR), intrinsic auto-regressive (IAR), and the spatial model suggested by Rao and Molina (2015) (SRM).

These spatial dependencies can be represented by the following "precision" matrices:

$$\text{(CAR) } \Omega_2(\rho) = (I_m - \rho W), \quad \rho \in (\lambda_m^{-1}, \lambda_1^{-1}), \tag{2.8}$$

$$\text{(SAR) } \Omega_3(\rho) = (I_m - \rho \tilde{W})^{\mathrm{T}}(I_m - \rho \tilde{W}), \quad \rho \in (-1, 1), \tag{2.9}$$

$$\text{(IAR) } \Omega_4(\rho) = L - \rho W, \quad \rho \in (-1, 1), \tag{2.10}$$

$$\text{(SRM) } \Omega_5(\rho) = \rho R + (1 - \rho) I_m, \quad \rho \in (0, 1), \tag{2.11}$$

where ρ is the model parameter that captures the strength of spatial dependence. We call it the spatial autocorrelation parameter. Here the matrix R is given by $R = L - W$. The ith diagonal element of R is the number of neighborhoods of the ith small area, and the (i, j)th element is -1 if the ith small area is connected to the jth small areas and 0 otherwise. We note that the precision matrices are defined without the model error variance parameter σ_v^2. We argued below that each of these precision matrices is a positive definite matrix.

Lemma 1. *Let $\Omega_k(\rho)$ be the precision matrix of the kth model and let l_k and u_k be corresponding the lower and upper bound of ρ as in (2.8)–(2.11). Then $\Omega_k(\rho)$ is positive definite for $\rho \in (l_k, u_k)$, $k = 1, \ldots, 5$.*

The SAR adjacency matrix $\tilde{W}$ is row-normalized so that ρ is ranging from -1 to 1 while having the positive definite precision matrix. In doing so, one can also use other types of adjacency measures such as "length" of the shared boundary (or traffic connecting the places). The spatial autocorrelation parameter ρ represents a strength of spatial association. Figure 2.2.1 graphically illustrates the strength of spatial autocorrelation with the state of Georgia's $m = 159$ counties. Data are generated from $N_m(0_m, \sigma_v^2\{\Omega_3(\rho)\}^{-1})$, where $\sigma_v^2 = 1$ and $\rho = 0, 0.75, 0.85, 0.95$. Interpretation of this parameter varies from model to model and its feasible values change across the models. Feasible values of ρ for each model are listed next to the model precision matrices $\Omega_k(\rho)$, $k = 2, \ldots, 5$.

Suppose that the ith small area is only connected to jth small area. The CAR model assumes that θ_i depends only on θ_j. In other words, θ_i is correlated with other areas only through θ_j. Similar interpretations hold for the IAR and SRM models. On the other hand,

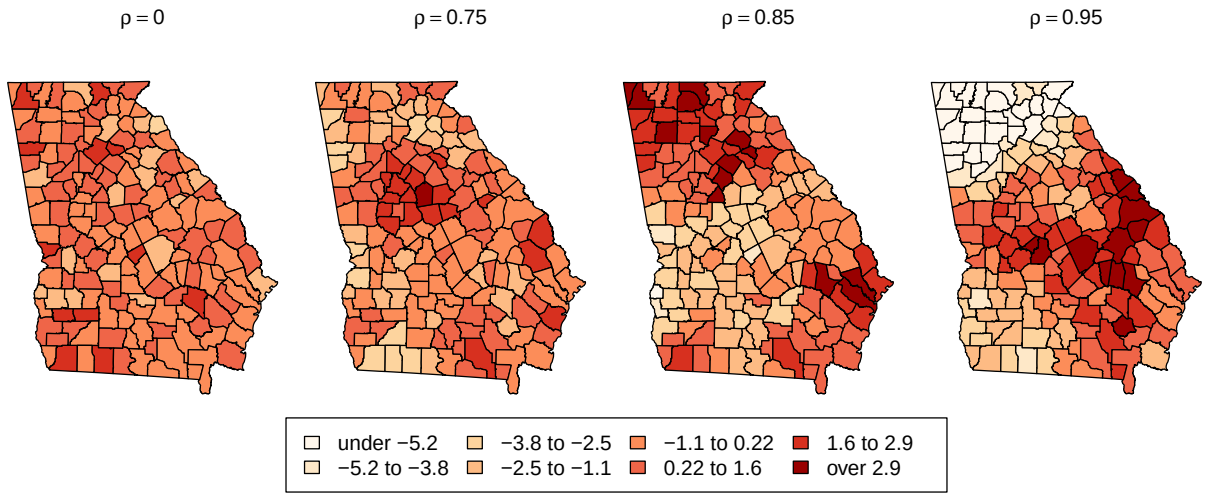


Figure 2.2.1: Geographical illustrations of spatial dependencies with various values of ρ. Each geographical region corresponds to one of the state of Georgia's 159 counties. Data are generated from the SAR model.

the SAR assumes that θ_i depends simultaneously on other θ_j, $j \neq i$, but has larger (weaker) correlations for connected (distant) areas. The independent FH model can be viewed as a special case of the above class of models with $\rho = 0$. For notational convenience, the precision matrix for the independent FH model is denoted $\Omega_1(\rho) = I_m$, which does not depend on ρ.

Let us denote the lower and upper bounds for ρ corresponding to $\Omega_k(\rho)$ by l_k and u_k, $k = 1, \ldots, 5$. We consider the following hierarchical Bayes spatial models:

$$Y|\theta \sim N_m(\theta, D), \tag{2.12}$$

$$\theta|\beta, \sigma_v^2, \rho \sim N_m(X\beta, \sigma_v^2\{\Omega_k(\rho)\}^{-1}), \quad k = 1, \ldots, 5, \tag{2.13}$$

$$\pi(\beta, \sigma_v^2, \rho) \propto g(\sigma_v^2)h(\rho), \quad \beta \in \mathbb{R}^p,\ \sigma_v^2 > 0,\ l_k < \rho < u_k, \tag{2.14}$$

where $g(\sigma_v^2)$ and $h(\rho)$ are suitable functions of σ_v^2 and ρ. Let $I(\cdot)$ be the indicator function taking the value 1 when its argument is true and 0 otherwise. Then the (joint) posterior distribution of model parameters is proper under the following conditions.

Theorem 2. *Under the hierarchical Bayes models given in* (2.12)–(2.14)*, the posterior distribution is proper if the following conditions hold for some positive constant* $c > 0$*:*

(a) $\int_0^\infty g(\sigma_v^2) I(\sigma_v^2 \leq c) d\sigma_v^2 < \infty$.

(b) $\int_0^\infty (\sigma_v^2)^{-(m-p)/2} g(\sigma_v^2) I(\sigma_v^2 > c) d\sigma_v^2 < \infty$.

(c) $\int_{l_k}^{u_k} h(\rho) d\rho < \infty$.

Any bounded function of ρ satisfies (c) in Theorem 2 as the domains are all bounded. Consider the following family of noninformative priors:

$$\pi(\beta, \sigma_v^2, \rho) \propto (\sigma_v^2)^{-\alpha} I(l_k < \rho < u_k), \quad \beta \in \mathbb{R}^p, \ \sigma_v^2 > 0. \tag{2.15}$$

Under (2.15), we provide the conditions for the posterior propriety in the corollary below.

Corollary 2.1. *For any of the hierarchical Bayes model given in* (2.12)–(2.13) *with the prior* (2.15)*, the posterior distribution is proper if and only if* $1 - (m-p)/2 < \alpha < 1$.

We use the prior corresponding to $\alpha = 0$, which is a uniform prior for the model parameters, in our numerical studies.

2.2.1 Estimation of un-sampled small area characteristics

In this section, we consider the case when there are small areas with no available sample, which we call un-sampled small areas. In such case, the direct estimates of the un-sampled areas are missing and thus predictions of the area characteristics of un-sampled small areas are solely based on synthetic estimators. Here, we propose to exploit spatial dependencies in predicting area characteristics of un-sampled small areas. Without loss of generality, let $Y_1, \ldots, Y_{m_1}$ be the direct estimates of un-sampled small areas (unobserved) and $Y_{m_1+1}, \ldots, Y_m$ be the direct estimates sampled small areas. Also, let $m_2 = m - m_1$. Although $Y_1, \ldots, Y_{m_1}$ are unobserved, we assume that the sampling variances of $Y_1, \ldots, Y_{m_1}$ would be finite if we had selected a sample from such a small area. So we assume that $D_i < \infty$, $i = 1, \ldots, m$. Let

$Y_{(1)} = (Y_1, \ldots, Y_{m_1})^{\mathrm{T}}$ and $Y_{(2)} = (Y_{m_1+1}, \ldots, Y_m)^{\mathrm{T}}$, and consider the following hierarchical Bayes models:

$$Y_{(2)}|\theta \sim N_{m_2}(\theta_{(2)}, D_{(2)}), \tag{2.16}$$

$$\theta|\beta, \sigma_v^2, \rho \sim N_m(X\beta, \sigma_v^2\{\Omega_k(\rho)\}^{-1}), \quad k = 1, \ldots, 5, \tag{2.17}$$

$$\pi(\beta, \sigma_v^2, \rho) \propto (\sigma_v^2)^{-\alpha} I(l_k < \rho < u_k) \quad \text{for } \beta \in \mathbb{R}^p, 0 < \sigma_v^2. \tag{2.18}$$

where $\theta_{(2)} = (\theta_{m_1+1}, \ldots, \theta_m)^{\mathrm{T}}$ and $D_{(2)} = \mathrm{diag}\{D_i\}_{i=m_1+1}^m$ are the vector of area characteristics and the diagonal matrix of sampling variances corresponding $Y_{(2)}$. Under this model, the posterior pdf is proper under the following condition.

Theorem 3. *Under the hierarchical Bayes models given in (2.16)–(2.18), the posterior pdf is proper if* $1 - (m - m_1 - p)/2 < \alpha < 1$.

We can directly see that equivalent condition for the posterior propriety is $\alpha < 1$ and $m - p - 2 + 2\alpha > m_1$. Thus, using the flat prior, $\alpha = 0$, the posterior distribution is proper as long as the number of un-sampled small areas is smaller than $m - p - 2$.

2.3 A simulation study

In this section, we evaluate the performance of independent FH model and the four spatial models by examining their prediction accuracy in the absence of "good" covariates. We consider various settings of sampling and model error variances. The geographical region of interest is the state of Georgia with $m = 159$ counties. Sampling variances D_i, $i = 1, \ldots, m$, are independently generated from the gamma distribution with shape 1.1 and scale 3. We consider two spatially correlated covariates x_1 and x_2 such that

$$x_j \overset{iid}{\sim} N_m(0_m, \{\Omega_3(\rho)\}^{-1}), \quad j = 1, 2.$$

We consider four different degrees of spatial autocorrelations, $\rho = 0, 0.75, 0.85, 0.95$. The strength of the spatial autocorrelations are graphically illustrated in Figure 2.2.1. For the

model error variance parameter σ_v^2, we also consider four values which are proportional to the average of the sampling variances $\bar{D}$, where $\bar{D} = m^{-1}\sum_{i=1}^{m} D_i$. Specifically, we use $\sigma_v^2 = \bar{D}/8, \bar{D}/2, \bar{D}, 2\bar{D}$.

For given values of ρ, σ_v^2 and D_i, $i = 1, \ldots, m$, we consider $S = 50$ replicated data sets generated as follows. For $i = 1, \ldots, m$, we generate

$$\theta_i \overset{ind}{\sim} N(\mu_i, \sigma_v^2),$$
$$Y_i|\theta_1, \ldots, \theta_m \overset{ind}{\sim} N(\theta_i, D_i),$$

where $\mu_i = \beta_1 x_{i1} + \beta_2 x_{i2}$, $i = 1 \ldots, m$, and $(\beta_1, \beta_2)^{\mathrm{T}} = (2, 0.5)^{\mathrm{T}}$. This makes x_1 more significant than x_2.

To examine how the spatial models can capture spatial dependence when one (or both) of the covariates are un-available for model fitting, we consider $L = 4$ different combinations of covariates. In particular, we consider four design matrices given by $X_1 = 1_m$, $X_2 = [1_m, x_1]$, $X_3 = [1_m, x_2]$ and $X_4 = [1_m, x_1, x_2]$, where 1_m is the $m \times 1$ vector of ones. For each $k = 1, \ldots, 5$, we fit the following hierarchical models:

$$Y|\theta \sim N_m(\theta, D), \tag{2.19}$$
$$\theta|\beta_l, \sigma_v^2, \rho \sim N_m(X_l\beta_l, \sigma_v^2\{\Omega_k(\rho)\}^{-1}), \quad l = 1, \ldots, 4, \tag{2.20}$$
$$\pi(\beta_l, \sigma_v^2, \rho) \propto I(l_k < \rho < u_k), \tag{2.21}$$

where the dimensions of β_l agree with the corresponding X_l.

To evaluate predictors, $\hat{\theta}_k = (\hat{\theta}_{k1} \ldots, \hat{\theta}_{km})^{\mathrm{T}}$, $k = 1, \ldots, 5$, in terms of prediction accuracy, we calculate average squared deviation, $\mathrm{ASD}_k = m^{-1}\sum_{i=1}^{m}(\hat{\theta}_{ki} - \theta_i)^2$, after fitting the models above for each replicated data set. To make overall comparisons, we again average ASD_k over the replicated data sets for each $k = 1, \ldots, 5$, which are denoted by AASD_k. By setting the independent FH model's AASD_1 as a baseline, spatial models are assessed by the amount of average improvement defined as

$$\frac{\mathrm{AASD}_k}{\mathrm{AASD}_1}, \quad \text{where } \mathrm{AASD}_k = \frac{1}{mS}\sum_{s=1}^{S}\sum_{i=1}^{m}(\hat{\theta}_{ki}^{(s)} - \theta_i^{(s)})^2.$$

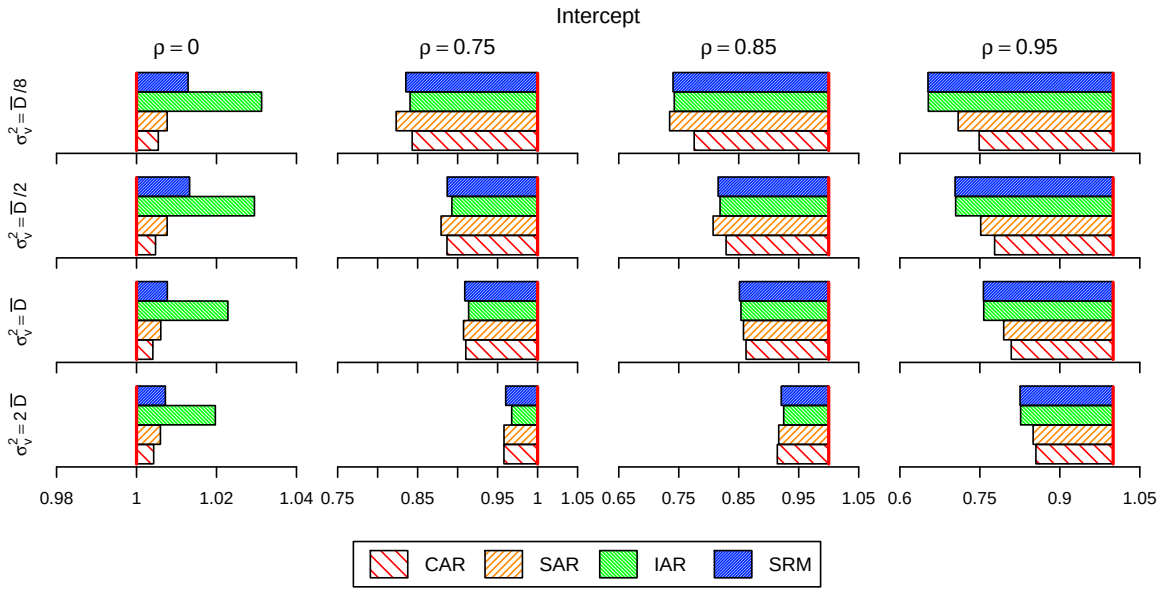


Figure 2.3.1: Ratios of average ASD_k to average ASD_1 based on 50 replications when there is no available covariate.

Here, $\theta_i^{(s)}$ is the true value of the ith small area mean based on sth replicated data set; and $\hat{\theta}_{ki}^{(s)}$ is the predicted value of $\theta_i^{(s)}$ of the kth model. Simulation results for the four different combinations of covariates are summarized in Figures 2.3.1 – 2.3.4, respectively. Reference lines are drawn vertically at one. Each bar located to the left (right) hand side of the reference line indicates the corresponding spatial model has a smaller (larger) AASD compared with the independent FH model's AASD. The length of a bar represents the percentage decrease (increase) in AASD_k, $k = 2, \ldots, 5$, compared with AASD_1.

Figures 2.3.1 and 2.3.2 are illustrating the simulation results with no covariate (intercept only) and the least significant covariate x_1, respectively. When $\rho = 0$, the true model is given by the independent FH model. In this case, spatial models have approximately 0.5% to 4% larger AASD and we interpret this as a cost of introducing an unnecessary parameter ρ. On the contrary, when $\rho > 0$, spatial models outperform the independent FH model by having much smaller AASDs. Figure 2.3.1 shows that the amount of reduction in AASD

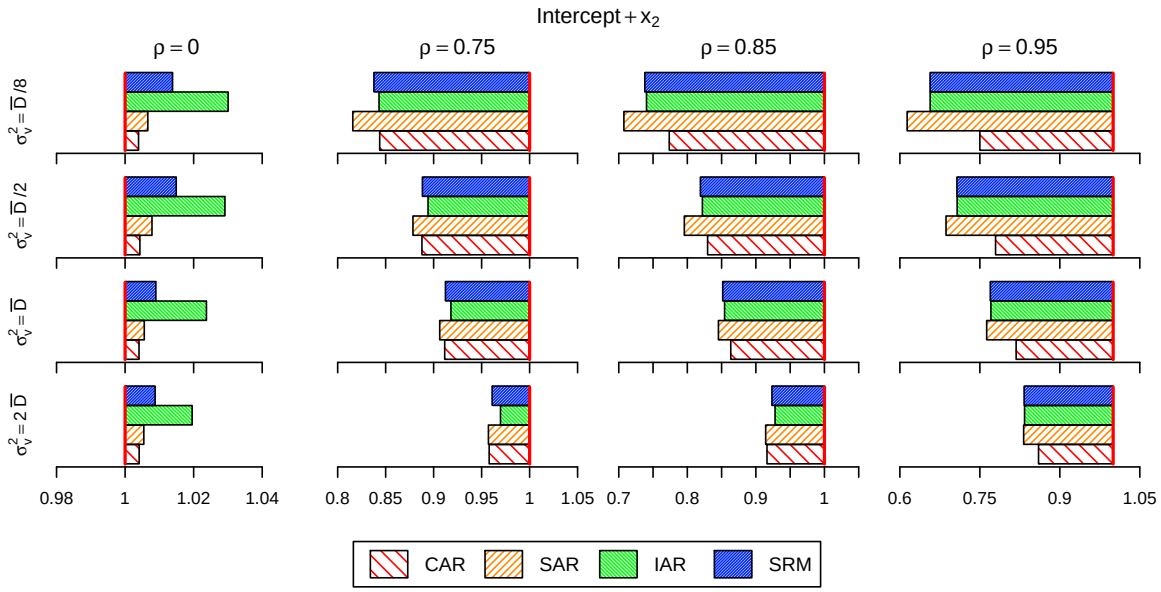


Figure 2.3.2: Ratios of average ASD_k to average ASD_1 based on 50 replications when x_1 is unavailable.

increases as the model error variance σ_v^2 decreases and the spatial autocorrelation increases. Specifically, spatial models have approximately 5% to 40% smaller AASD compared with the independent FH model. Even when σ_v^2 is two times larger than the average sampling variance $\bar{D}$, we can expect 5% to 10% improvement depending on the magnitude of ρ. This indicates that spatial model predictors provide better predictions when no good covariate to explain spatial dependence in $\theta_1, \ldots, \theta_m$ is available.

Figure 2.3.2 displays the simulation results when the good covariate, x_1, is unavailable. It can be seen that the results are similar to those with no covariate, given in Figure 2.3.1. On average, spatial models have 5% to 40% smaller AASD than the independent FH model. Although general patterns are analogous to Figure 2.3.1, it can be observed that the SAR model performs better in this setting. This is possibly due to the fact that x_1 is generated under the SAR model.

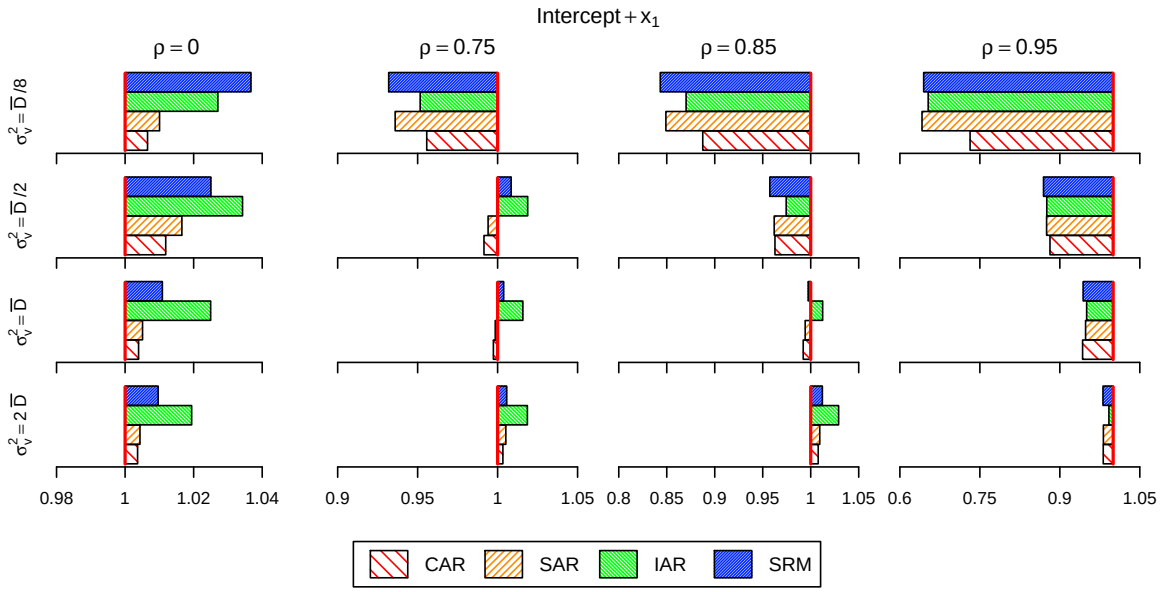


Figure 2.3.3: Ratios of average ASD_k to average ASD_1 based on 50 replications when x_2 is unavailable.

The simulation results with x_1 are given in Figures 2.3.3 and 2.3.4. Figure 2.3.3 summarizes the results when the less important covariate x_2 is unavailable. The results indicate that the meaningful AASD improvements are made only when σ_v^2 is small. Thus, given that small areas are spatially correlated, the spatial models allow more accurate predictions if noise (random effect variance) is not too large.

When all the covariates are available, the independent FH model demonstrates the best performance. Spatial models have larger AASD's; however, the differences are only up to about 5%. It is interesting to see that, as σ_v^2 increases, the performance of the SRM model also increases while other spatial models show similar AASDs.

From these simulation studies, we find that if there is any spatial pattern remains after fitting a model, then the spatial models effectively capture the extra variabilities and yield better predictions. Although we generate spatial patterns from the SAR model, the SRM model also shows competitive performances. On the other hand, when there is no spatial

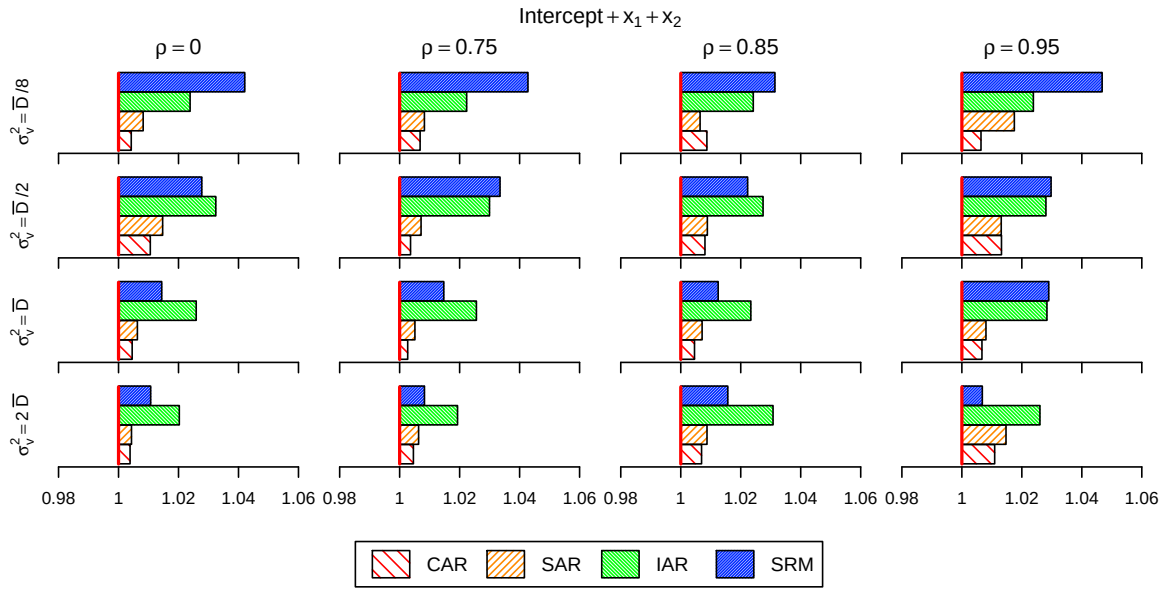


Figure 2.3.4: Ratios of average ASD_k to average ASD_1 based on 50 replications when all the covariates are included in the fitted model.

pattern in the residuals, we find that the SRM model is relatively sensitive to the magnitude of σ_v^2 compared with the other spatial models.

2.4 Application to the Current Population Survey data

In this section, we evaluate the spatial models in terms of their prediction accuracies for median incomes. The U.S. Department of Health and Human Services has a cash welfare program for low-income families in the U.S. as a means of providing energy assistance benefits. In order to make accurate policy decisions, the U.S. Census Bureau utilizes the Current Population Survey (CPS) data to obtain the direct estimates, Y_i, $i = 1, \ldots, m$, of state-level median incomes. However, the CPS is targeted to achieve a desired level of accuracy of the estimate at the national level. Thus, when it comes to estimating state-level median incomes, available sample sizes for individual states are typically small. We apply our proposed spatial

	ASD	ASD-PI	APSD	APSD-PI
FH	2.899	0%	3.879	0%
CAR	2.989	-3.12%	3.924	-0.40%
SAR	2.583	10.92%	3.887	0.15%
IAR	2.642	8.89%	3.792	1.94%
SRM	2.507	13.51%	3.575	4.41%

Table 2.1: Average squared deviations and average posterior standard deviations of independent FH model and four spatial models. The second and fourth columns represent the percentage improvements (PI) over the independent FH model.

models to estimate the 1990 state-level four-person family median incomes for all 49 contiguous states including the Washington, D.C. Prediction performances are measured with data from all 49 states, and, by removing data from some states.

2.4.1 Four-person family median income estimation

Let θ_i be the 1990 four-person family median income (median income) of the ith state, where $i = 1, \ldots, m$, where $m = 49$ is the number of contiguous states including the District of Columbia. The states of Alaska and Hawaii are excluded as they are not geographically connected. Let Y_i be the direct estimate of θ_i based on the 1989 CPS. Covariates of interest are 1980 census median income x_{i1} and an adjusted 1980 census median income x_{i2}. The adjusted census median income x_{i2} is defined as $(\text{PCI}_{i,1989}/\text{PCI}_{i,1979})x_{i1}$, $i = 1, \ldots, m$, where $\text{PCI}_{i,1979}$ and $\text{PCI}_{i,1989}$ are the 1979 and 1989 per capita income of the ith state from the Bureau of Economic Analysis of the U.S. Department of Commerce. We consider the state-level median incomes obtained from the 1990 census as the true values of θ_i, $i = 1, \ldots, m$. It has been known that the adjusted census median income is a good covariate which accounts for the variability of the 1990 four-person median incomes.

Let $X = [1_m, x_1, x_2]$, where $x_j = (x_{1j}, \ldots, x_{mj})^{\mathrm{T}}$, $j = 1, 2$. For each $k = 1, \ldots, 5$, we fit models (2.12)–(2.14). The prior distribution for the model parameters is the noninformative prior (2.15) with $\alpha = 0$. After fitting the models, we calculate the posterior means $\hat{\theta}_{ki}$, $i = 1, \ldots, m$, and obtain the average squared deviation ASD $= m^{-1} \sum_{i=1}^{m} (\hat{\theta}_{ki} - \theta_i)^2$ for $k = 1, \ldots, 5$. We also calculate posterior standard deviations of θ_{ki} and calculate average posterior standard deviations (APSD) for each $k = 1, \ldots, 5$. Table 2.1 lists ASD, APSD and respective percentage improvements. On average, the SRM model has approximately 13% smaller ASD and 4% smaller posterior standard deviations than the independent FH model. In terms of ASD, the second best performing model is the SAR having approximately 10% smaller ASD. In terms of APSD, the IAR model is the second best model with 2% smaller APSD.

2.4.2 Estimation of four-person family median income of un-sampled small areas

In this section, we evaluate spatial models in terms of un-sampled small area prediction accuracy. At each instance, we randomly exclude direct estimates of multiple states and make predictions for the median incomes of the excluded states. As there are 49 small areas, we have 12 data sets with $m_1 = 4$ or $m_1 = 5$ missing states, where m_1 is the number of un-sampled small areas given in Section 2.2.1. Excluded states corresponding to each data set are listed in Table 2.2. For notational convenience, we denote the median income and direct estimate of the jth un-sampled area by θ_j and Y_j, $j = 1, \ldots, m_1$.

For each data set, we fit the independent FH model and four spatial models as specified in (2.16)–(2.18) with $\alpha = 0$ for (2.18). Then the squared deviations of un-sampled areas are obtained for each model. Specifically, for $j = 1, \ldots, m_1$, we calculate $r_{kj}^2 = (\hat{\theta}_{kj} - \theta_j)^2$, where $\hat{\theta}_{kj}$ is the posterior mean of θ_j under the kth model with the missing direct estimates $Y_1, \ldots, Y_{m_1}$. Each panel displayed in Figure 2.4.1 illustrates the following quantities:

$$\eta_{ki} = 1 - \frac{r_{ki}^2}{r_{1i}^2} \quad i = 1, \ldots, m, \tag{2.22}$$

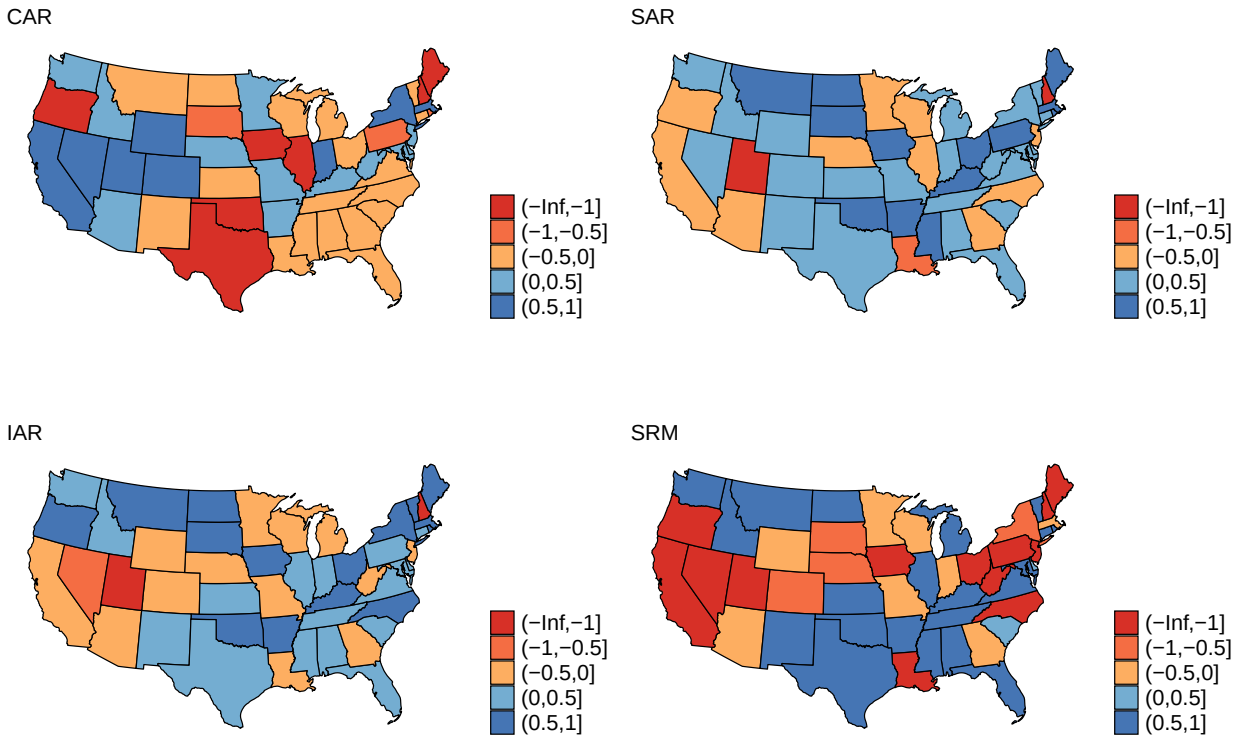


Figure 2.4.1: Relative squared deviations. Blue and red color schemes represent positive and negative values of the performance measure η_{ki}.

where $k = 2, \ldots, 5$. A negative value of η_{ki} is in favor of the independent FH model and depicted as value red color schemes. On the other hand, a positive value of η_{ki} indicates that the kth model has a smaller squared deviation on the ith small area. The darker the color is the larger the larger the absolute value of η_{ki} is.

The top left panel indicates that squared deviations of the CAR model are smaller than that of the independent FH model at 20 states. There are 8 states which have more than 50% smaller squared deviations. However, there are 11 states which have more than 50% larger squared deviations. Hence, the independent FH model does better than the CAR model on this analysis. The results with the SAR model are illustrated at the top right panel. It has

Set 1				Set 2				Set 3				
MS	OK	AR	SD	DE	AZ	CO	TN	WV	MD	NV	MI	
Set 4				Set 5				Set 6				
NC	MT	NY	NE	ID	ND	DC	GA	MO	VT	WY	AL	
Set 7				Set 8				Set 9				
LA	UT	FL	WA	MN	MA	TX	SC	KY	RI	VA	WI	
Set 10				Set 11				Set 12				
IN	IL	PA	NH	CA	ME	OH	NJ	OR	KS	NM	IA	CT

Table 2.2: Excluded small areas for each data set.

smaller squared deviations on 35 states and 13 of them are more than 50% smaller. There are only 3 states with more than 50% larger squared deviations and these are the states of Utah, Louisiana and New Hampshire. The IAR model outperforms the independent FH model in 33 states and more than 50% smaller squared deviations in 16 states. On the contrary, in the states of Nevada, Utah and New Hampshire, the independent FH model has more than 50% smaller squared deviations. Overall, it performs similarly to the SAR model while the results are more volatile than the SAR model. The bottom right panel demonstrates that the SRM model is more volatile than the IAR model. In 23 states, it has smaller squared deviations and 21 of them have more than 50% smaller squared deviations. On the other hand, there are 18 states with more than 50% larger squared deviations.

2.5 Conclusions

In this chapter, we proposed alternatives to the independent Fay-Herriot model for the estimation of spatially correlated small area means. In particular, four spatial models with different autocorrelation structures were considered under the hierarchical Bayes framework. We further extended the spatial models to allow multiple missing un-sampled small areas

in predicting small area characteristics. With a family of noninformative priors, posterior proprieties of the proposed models were established.

The simulation study presented in Section 2.3 suggested that prediction accuracy can be greatly improved by considering spatial models under the absence of good covariates. It has been noted by Datta et al. (2011) that the prediction accuracy of small area estimation models largely depends on the availability of good covariates. In other words, when good covariates are unavailable, we may not gain much from model-based approaches. The simulation results indicated that, in that event, the spatial models significantly improve the prediction accuracy by exploiting information from adjacent areas.

In Section 2.4, we applied the spatial models to four-person family median income estimation. Even when a good covariate exists, the spatial models exhibited noticeable performance in terms of squared deviation and posterior standard deviation. Furthermore, the SAR and IAR models provided more precise estimates in estimation of un-sampled small area means.

2.6 Technical details

Proof of Lemma 1 . Recall that $\lambda_1 \geq \cdots \geq \lambda_m$ are the eigenvalues of W. Thus, for $\rho \in (\lambda_m^{-1}, \lambda_1^{-1})$, we have $1 - \rho\lambda_i > 0$ for all $i = 1, \ldots, m$. This implies that $\Omega_2(\rho) = I - \rho W$ is a positive definite matrix for $\rho \in (\lambda_m^{-1}, \lambda_1^{-1})$.

Let $\tilde{\lambda}_1 \geq \cdots \geq \tilde{\lambda}_m$ be the eigenvalues of $\tilde{W}$. Since $\tilde{W}$ is a stochastic matrix, i.e., $\tilde{W}1_m = 1_m$ and $w_{ij} \geq 0$, all its eigenvalues will satisfy $|\tilde{\lambda}_i| \leq 1$ and at least one of them is one. Note that, for $\rho \in (-1, 1)$, we get $|\rho\tilde{\lambda}_i| < 1$, implying $0 < |1 - \rho\tilde{\lambda}_1| < 2$. Thus the eigenvalues of $I - \rho\tilde{W}$ are all non-zero, not exceeding 2 in absolute value. This results in the positive definiteness of $\Omega_3(\rho)$ for all $\rho \in (-1, 1)$.

Now for $\Omega_4(\rho)$, we note that for $t = (t_1, \ldots, t_m)^{\mathrm{T}}$,

$$t^{\mathrm{T}}\Omega_4(\rho)t = (1-\rho)\sum_{i=1}^{m} w_{i.}t_i^2 + \rho\sum_{i=1}^{m}\sum_{i<j}(t_i - t_j)^2 w_{ij} \tag{2.23}$$

$$= (1+\rho)\sum_{i=1}^{m} w_{i.}t_i^2 - \rho\sum_{i=1}^{m}\sum_{i<j}(t_i + t_j)^2 w_{ij}. \tag{2.24}$$

The right hand side of (2.23) is non-negative for $0 \leq \rho \leq 1$ for all $t \in \mathbb{R}^m$, and (2.24) is non-negative for $-1 \leq \rho \leq 0$ for all $t \in \mathbb{R}^m$. Thus the quadratic form $t^{\mathrm{T}}\Omega_4(\rho)t$ is non-negative definite. Furthermore, for $-1 < \rho < 1$, $t^{\mathrm{T}}\Omega_4(\rho)t = 0$ implies $t = 0_m$. Hence $\Omega_4(\rho)$ is positive definite for $-1 < \rho < 1$.

We noticed that $\Omega_4(1)$ is a non-negative definite matrix, where $\Omega_4(1) = R$. Thus $\Omega_5(\rho) = \rho\Omega_4(1) + (1-\rho)I$ is a positive definite matrix for $0 < \rho < 1$.

□

Lemma 2. *Let $\Omega_k(\rho)$, $k = 1, \ldots, 5$, be the precision matrices given in (2.8)–(2.11) and let $\tau_{ki}(\rho)$ be the ith eigenvalue of $\Omega_k(\rho)$ such that $\tau_{k1}(\rho) \geq \ldots \geq \tau_{km}(\rho) > 0$. Then, for each $k = 1, \ldots, 5$,*

$$\sup_{1\leq i\leq m}\sup_{l_k<\rho<u_k}\tau_{ki}(\rho) < \infty,$$

where l_k and u_k are the lower and upper bounds of ρ corresponding to $\Omega_k(\rho)$.

Proof of Lemma 2. Recall that for $\lambda_m^{-1} < \rho < \lambda_1^{-1}$, $\Omega_2(\rho) = I_m - \rho W$ is positive definite. Since $\mathrm{tr}(W) = 0$, we have $\mathrm{tr}\{\Omega_2(\rho)\} = m$. This implies $\sum_{i=1}^{m}\tau_{2i}(\rho) = m < \infty$, and therefore

$$\sup_{1\leq i\leq m}\sup_{\lambda_m^{-1}<\rho<\lambda_1^{-1}}\tau_{2i}(\rho) < \infty.$$

For the SAR model, $\Omega_3(\rho) = (I_m - \rho\tilde{W})^{\mathrm{T}}(I_m - \rho\tilde{W})$ and it is positive definite for $-1 < \rho < 1$. Let $\tilde{w}_{ij}$ be the (i,j)th element of $\tilde{W}$. As $|\tilde{w}_{ij}| \leq 1$, we have

$$\mathrm{tr}\{\Omega_3(\rho)\} = m + \rho^2\,\mathrm{tr}(\tilde{W}^{\mathrm{T}}\tilde{W}) \leq m + \sum_{i=1}^{m}\sum_{j=1}^{m}\tilde{w}_{ij}^2 < 2m.$$

This implies that $\tau_{3i}(\rho) < 2m$ for $1 \le i \le m$, $-1 < \rho < 1$, and thus,

$$\sup_{1\le i\le m} \sup_{-1<\rho<1} \tau_{3i}(\rho) < \infty.$$

Since $\Omega_4(\rho) = L - \rho W$ for the IAR model, $\mathrm{tr}\{\Omega_4(\rho)\} = \mathrm{tr}(L)$. We note that $\mathrm{tr}(L)$ is free from ρ and $\mathrm{tr}(L) = \sum_{i=1}^m \sum_{j=1}^m w_{ij} \le m^2$. Hence $\tau_{4i}(\rho) < m^2$ for $1 \le i \le m$, $-1 < \rho < 1$, and

$$\sup_{1\le i\le m} \sup_{-1<\rho<1} \tau_{4i}(\rho) < \infty.$$

Lastly, consider $\Omega_5(\rho) = \rho(L - W) + (1-\rho)I_m$ for the SRM model. For $0 < \rho < 1$, $\Omega_5(\rho)$ is positive definite and

$$\begin{aligned}\mathrm{tr}\{\Omega_5(\rho)\} &= \rho\,\mathrm{tr}\,(L) + (1-\rho)m = \rho \sum_{i=1}^m \sum_{j=1}^m w_{ij} + m(1-\rho) \\ &< \rho m^2 + (1-\rho)m \le m^2.\end{aligned}$$

Thus $\tau_i(\rho) < m^2$ for $1 \le i \le m$, $-1 < \rho < 1$, and

$$\sup_{1\le i\le m} \sup_{0<\rho<1} \tau_{5i}(\rho) < \infty.$$

□

Proof of Theorem 2. For notational convenience, we denote $\Omega_k(\rho)$ and $\tau_{ki}(\rho)$ by Ω and τ_i, respectively, and $\delta = \max_i D_i < \infty$. Let P^{T} be an orthogonal matrix of eigenvectors of Ω^{-1} such that $\Omega^{-1} = PT^{-1}P^{\mathrm{T}}$, where $T = \mathrm{diag}\{\tau_i\}_{i=1}^m$, and let $y^* = P^{\mathrm{T}}y$, $X^* = P^{\mathrm{T}}X$. Then, for a generic constant K,

$$f(y,\theta,\beta,\sigma_v^2,\rho) \le K g(\sigma_v^2) h(\rho) \exp\left\{-\frac{1}{2\delta}\sum_{i=1}^m (y_i - \theta_i)^2\right\} \tag{2.25}$$

$$\times |\sigma_v^2 \Omega^{-1}|^{-1/2} \exp\left\{-\frac{1}{2}(\theta - X\beta)^{\mathrm{T}}(\sigma_v^2\Omega^{-1})^{-1}(\theta - X\beta)\right\}. \tag{2.26}$$

Integrating both sides with respect to θ, we get

$$\begin{aligned}
&\int f(y,\theta,\beta,\sigma_v^2,\rho)d\theta \\
&\quad \leq Kg(\sigma_v^2)h(\rho)|\delta I_m + \sigma_v^2\Omega^{-1}|^{-1/2}\exp\left\{-\frac{1}{2}(y-X\beta)^{\mathrm{T}}(\delta I_m+\sigma_v^2\Omega^{-1})^{-1}(y-X\beta)\right\} \\
&\quad = Kg(\sigma_v^2)h(\rho)\prod_{i=1}^{m}\{\delta+\sigma_v^2\tau_i^{-1}\}^{-1/2}\exp\left\{-\frac{1}{2}(y^*-X^*\beta)^{\mathrm{T}}(\delta I_m+\sigma_v^2 T^{-1})^{-1}(y^*-X^*\beta)\right\}.
\end{aligned}$$

Suppose the rows of X^* corresponding to distinct p indices $\{i_1,\ldots,i_p\} \subseteq \{1,\ldots,m\}$ are linearly independent. We denote these rows by $x_{i_k}^{\mathrm{T}*}$, $k=1,\ldots,p$. Then we get from the last displayed expression

$$\begin{aligned}
&\int f(y,\theta,\beta,\sigma_v^2,\rho)d\theta \\
&\quad \leq Kg(\sigma_v^2)h(\rho)\prod_{i=1}^{m}\{\delta+\sigma_v^2\tau_i^{-1}\}^{-1/2}\exp\left\{-\frac{1}{2}\sum_{k=1}^{p}\frac{(y_{i_k}^*-x_{i_k}^{\mathrm{T}*}\beta)^2}{\delta+\sigma_v^2\tau_{i_k}^{-1}}\right\}.
\end{aligned}$$

Integrating both sides with respect to β, we get that

$$\int f(y,\theta,\beta,\sigma_v^2,\rho)d\theta d\beta \leq Kg(\sigma_v^2)h(\rho)\prod_{i\notin\{i_1,\ldots,i_p\}}\{\delta+\sigma_v^2\tau_i^{-1}\}^{-1/2}. \tag{2.27}$$

We note that for any i and some positive number N,

$$\{\delta+\sigma_v^2\tau_i^{-1}\}^{-1/2} \leq \delta^{-1/2}I(\sigma_v^2\leq N)+\nu^{1/2}(\sigma_v^2)^{-1/2}I(\sigma_v^2>N),$$

where ν is a constant satisfying $\sup_{1\leq i\leq m}\sup_{-1<\rho<1}\tau_i(\rho)=\nu$, $k=1,\ldots,5$. We note that $\nu<\infty$, $k=1,\ldots,5$, by Lemma 2. Hence,

$$\prod_{i\notin\{i_1,\ldots,i_p\}}\{\delta+\sigma_v^2\tau_i^{-1}\}^{-1/2} \leq KI(\sigma_v^2\leq N)+K(\sigma_v^2)^{-(m-p)/2}I(\sigma_v^2>N)$$

Using (2.27), we get

$$\int f(y,\theta,\beta,\sigma_v^2,\rho)d\theta d\beta \leq Kg(\sigma_v^2)h(\rho)\left\{I(\sigma_v^2\leq N)+(\sigma_v^2)^{-(m-p)/2}I(\sigma_v^2>N)\right\}.$$

This indicates that, if

$$\int_{l_k}^{u_k}h(\rho)d\rho<\infty, \quad \int_0^\infty g(\sigma_v^2)I(\sigma_v^2\leq N)d\sigma_v^2<\infty$$
$$\text{and} \quad \int_0^\infty(\sigma_v^2)^{-(m-p)/2}g(\sigma_v^2)I(\sigma_v^2>N)d\sigma_v^2<\infty,$$

then the posterior pdf is proper. □

Proof of Corollary 2.1 . The result is directly obtained by Theorem 2 as follows. Since $h(\rho) = I(l_k < \rho < u_k)$ is integrable, it suffices to show the integrability of $g(\sigma_v^2) = (\sigma_v^2)^{-\alpha}$. We know that

$$\int_0^N (\sigma_v^2)^{-\alpha} d\sigma_v^2 < \infty \quad \text{and} \quad \int_N^\infty (\sigma_v^2)^{-(2\alpha+m-p)/2} d\sigma_v^2 < \infty,$$

if $\alpha < 1$ and $(2\alpha + m - p)/2 > 1$, respectively. Thus the posterior pdf will be proper if $1 - (m-p)/2 < \alpha < 1$. □

Proof of Theorem 3. Let $Y_{(1)}$ and $Y_{(2)}$ be the $m_1 \times 1$ and $m_2 \times 1$ vectors with direct estimates corresponding to the un-sampled and sampled small areas, respectively. Without loss of generality, we assume that $Y_1, \ldots, Y_m$ and $\theta_1, \ldots, \theta_m$ are arranged so that $Y = (Y_{(1)}^{\mathrm{T}}, Y_{(2)}^{\mathrm{T}})^{\mathrm{T}}$ and $\theta = (\theta_{(1)}^{\mathrm{T}}, \theta_{(2)}^{\mathrm{T}})^{\mathrm{T}}$. Let $D_{(2)} = \{D_i\}_{i=m_1+1}^m$ be the diagonal matrix with sampling variances corresponding to the components of $Y_{(2)}$ and $\delta = \max_{m_1 < i \le m} D_i < \infty$. For notational convenience, we denote $\Omega_k(\rho)$ and $\tau_{ki}(\rho)$ by Ω and τ_i, and a generic constant will be denoted by K.

The joint pdf of $Y_{(2)}, \theta, \beta, \sigma_v^2$ and ρ is given by

$$f(y_{(2)}, \theta, \beta, \sigma_v^2, \rho) = N_{m_2}(y_{(2)}|\theta_{(2)}, D_{(2)}) N_m(\theta|X\beta, \sigma_v^2\Omega^{-1}) g(\sigma_v^2) h(\rho), \tag{2.28}$$

where $N(y_{(2)}|\theta_{(2)}, D_{(2)})$ is the normal pdf with the mean $\theta_{(2)}$ and covariance matrix $D_{(2)}$. Since

$$N_{m_2}(y_{(2)}|\theta_{(2)}, D_{(2)}) \le K \exp\left\{-\frac{1}{2\delta}(y_{(2)} - \theta_{(2)})^{\mathrm{T}}(y_{(2)} - \theta_{(2)})\right\},$$

we have

$$\begin{aligned}\pi(\theta, \beta, \sigma_v^2, \rho|y_{(2)}) &\le K \exp\left\{-\frac{1}{2\delta}(y_{(2)} - \theta_{(2)})^{\mathrm{T}}(y_{(2)} - \theta_{(2)})\right\} N_m(\theta|X\beta, \sigma_v^2\Omega^{-1}) g(\sigma_v^2) h(\rho) \\ &= K \int \exp\left\{-\frac{1}{2\delta}(y - \theta)^{\mathrm{T}}(y - \theta)\right\} dy_{(1)} N_m(\theta|X\beta, \sigma_v^2\Omega^{-1}) g(\sigma_v^2) h(\rho). \end{aligned} \tag{2.29}$$

By integrating both sides of (2.29) with respect to θ, we get

$$\pi(\beta, \sigma_v^2, \rho|y_{(2)}) \le K g(\sigma_v^2) h(\rho) \int \exp\left\{-\frac{1}{2}(y - X\beta)^{\mathrm{T}}(\delta I + \sigma_v^2\Omega^{-1})^{-1}(y - X\beta)\right\} dy_{(1)}. \tag{2.30}$$

Partition X as $X = [X_1, X_2]^{\mathrm{T}}$, where X_1^{T} is $m_1 \times p$ and X_2^{T} is $m_2 \times p$. We assume that $\mathrm{rank}(X_2) = p$. Let $d = (0_{m_1,1}^{\mathrm{T}}, y_{(2)}^{\mathrm{T}})^{\mathrm{T}}$, $\phi = (y_{(1)}^{\mathrm{T}}, \beta^{\mathrm{T}})^{\mathrm{T}}$ and

$$G = \begin{bmatrix} -I_{m_1} & X_1^{\mathrm{T}} \\ 0_{m_2,m_1} & X_2^{\mathrm{T}} \end{bmatrix}.$$

Then, we can write

$$y - X\beta = d - G\phi,$$

where G is $m \times (m_1 + p)$, ϕ is $(m_1 + p) \times 1$. Hence, (2.30) can be written as

$$\pi(\beta, \sigma_v^2, \rho | y_{(2)}) \leq Kg(\sigma_v^2)h(\rho) \int N_m(d | G\phi, \delta I_m + \sigma_v^2 \Omega^{-1}) dy_{(1)}. \tag{2.31}$$

By integrating both sides of (2.31) with respect to β, we get

$$\pi(\sigma_v^2, \rho | y_{(2)}) \leq Kg(\sigma_v^2)h(\rho) \int N_m(d | G\phi, \delta I_m + \sigma_v^2 \Omega^{-1}) d\phi. \tag{2.32}$$

Let P^{T} be an orthogonal matrix of eigenvectors of Ω^{-1} such that $\Omega^{-1} = PT^{-1}P^{\mathrm{T}}$, where $T = \mathrm{diag}\{\tau_i\}_{i=1}^m$. Also, let $d^* = P^{\mathrm{T}}d$ and $G^* = P^{\mathrm{T}}G$. Then, we can write (2.32) as

$$\begin{aligned} \pi(\sigma_v^2, \rho | y_{(2)}) \leq & Kg(\sigma_v^2)h(\rho) \prod_{i=1}^{m} (\delta + \sigma_v^2 \tau_i^{-1})^{-1/2} \\ & \times \int \exp\left\{ -\frac{1}{2} (d^* - G^*\phi)^{\mathrm{T}} (\delta I_m + \sigma_v^2 T^{-1})^{-1} (d^* - G^*\phi) \right\} d\phi. \end{aligned} \tag{2.33}$$

Since $\mathrm{rank}(X_2) = p$, we know that $\mathrm{rank}(G) = m_1 + p$. We choose $b = m_1 + p$ indices $\{i_1, \ldots, i_b\} \subset \{1, \ldots, m\}$ such that corresponding rows of G^* are linearly independent. Suppose these rows are $g_{i_1}^{\mathrm{T}*}, \ldots, g_{i_b}^{\mathrm{T}*}$. Then

$$\begin{aligned} \int \exp\left\{ -\frac{1}{2} (d^* - G^*\phi)^{\mathrm{T}} (\delta I_m + \sigma_v^2 T^{-1})^{-1} (d^* - G^*\phi) \right\} d\phi &\leq \int \exp\left\{ -\frac{1}{2} \sum_{k=1}^{b} \frac{(d_{i_k}^* - g_{i_k}^{\mathrm{T}*}\phi)^2}{\delta + \sigma_v^2 \tau_{i_k}^{-1}} \right\} d\phi \\ &= K \prod_{k=1}^{b} (\delta + \sigma_v^2 \tau_{i_k}^{-1})^{1/2}. \end{aligned} \tag{2.34}$$

Using (2.34) in (2.33), we get

$$\pi(\sigma_v^2, \rho | y_{(2)}) \leq Kg(\sigma_v^2)h(\rho) \prod_{i \notin \{i_1, \ldots, i_b\}} \{\delta + \sigma_v^2 \tau_i^{-1}\}^{-1/2}.$$

Now, the result follows by Theorem 2 and Corollary 2.1.

□

Bibliography

Datta, G. S., Hall, P., and Mandal, A. (2011), "Model selection by testing for the presence of small-area effects, and application to area-level data," *Journal of the American Statistical Association*, 106, 362–374.

Datta, G. S. and Lahiri, P. (2000), "A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems," *Statistica Sinica*, 10, 613–627.

Datta, G. S., Rao, J. N. K., and Smith, D. D. (2005), "On measuring the variability of small area estimators under a basic area level model," *Biometrika*, 92, 183–196.

Datta, G. S. and Smith, D. D. (2003), "On propriety of posterior distributions of variance components in small area estimation," *Journal of Statistical Planning and Inference*, 112, 175–183.

Fay, R. E. and Herriot, R. A. (1979), "Estimates of income for small places: an application of James-Stein procedures to census data," *Journal of the American Statistical Association*, 74, 269–277.

Prasad, N. G. N. and Rao, J. N. K. (1990), "The estimation of the mean squared error of small-area estimators," *Journal of the American Statistical Association*, 85, 163–171.

Rao, J. N. K. and Molina, I. (2015), *Small area estimation*, John Wiley & Sons.

Chapter 3

Measurement Error in Small Area Estimation: Functional Versus Structural Versus Naive Models

3.1 Introduction

Small area estimation methodology benefited immensely from Stein's shrinkage estimation to develop reliable small area statistics by borrowing strength from the direct estimates of the other areas and appropriate auxiliary variables available for all the areas. A part of the variability of the population small area means is explained through some regression model based on the auxiliary variables. The regression model connects the direct estimates from the survey data with the auxiliary variables to construct indirect "synthetic regression estimates" of the small area means used in shrinkage estimation.

In many cases, auxiliary variables are often subject to sampling or measurement errors as they obtained from another surveys. Thus blind using these auxiliary variables as covariates ignoring measurement error may produce misleading results. In this chapter, we compare two small area models which account for measurement errors in covariates and Fay-Herriot (FH) model focusing on their predictive performance for small area characteristics. We make comparisons by deriving analytic expressions of the first order mean squared errors.

Let z_i's and δ be covariates without measurement error and respective regression parameter vector so that the FH model can be written as

$$Y_i = \theta_i + e_i, \quad \theta_i = z_i^{\mathrm{T}}\delta + u_i, \quad i = 1, \ldots, m, \tag{3.1}$$

where $e_i \overset{ind}{\sim} N(0, D_i)$ $u_i \overset{iid}{\sim} N(0, \sigma_u^2)$. Note that D_i, $i = 1, \ldots, m$, are assumed to be known. Let X_i be a covariate which is itself taken from another survey. By simply ignoring the

sampling error in X_i, treating it like the covariates in z_i, (3.1) can be written as

$$Y_i = \theta_i + e_i, \quad \theta_i = \beta_N X_i + z_i^{\mathrm{T}} \delta_N + u_{i,N}, \tag{3.2}$$

which we call naive FH model. To distinguish the naive FH model from the measurement error models to come, the subscript N is added to the random effect $u_{i,N}$ and regression parameter vectors β_N and δ_N. The model incorrectly assumes common random effect variance σ_u^2 over small areas although with heteroscedastic sampling error in X_i. Hence, (3.2) is misspecified.

For $i = 1, \ldots, m$, let x_i denotes the population characteristic being estimated by X_i with sampling error η_i, where the $\eta_i \overset{ind}{\sim} N(0, C_i)$ with known measurement error variance C_i. A generalization of the model (3.1) to include the covariate X_i while accounting for its sampling error is

$$Y_i = \theta_i + e_i, \quad \theta_i = \beta X_i + z_i^{\mathrm{T}} \delta + u_i, \tag{3.3}$$

$$X_i = x_i + \eta_i. \tag{3.4}$$

If the x_i are assumed to be fixed unknown quantities, then the model defined by (3.3)–(3.4) is known as the functional measurement error (FME) model. This model is discussed by Fuller (2009) and has been studied for small area estimation by Ybarra and Lohr (2008), Arima et al. (2015, 2016) and Arima et al. (2017). Analogous unit-level measurement error models for small area estimation have been studied by Ghosh and Sinha (2007), Datta et al. (2010), and Arima et al. (2012).

Another measurement error model of interest specifies a model for x_i in (3.4) which implies bivariate models for $(\theta_i, x_i)^{\mathrm{T}}$ and $(Y_i, X_i)^{\mathrm{T}}$. This is known as a structural measurement error (SME) model. If x_i follows the regression model $x_i = z_{xi}^{\mathrm{T}} \delta_x + v_i$, with covariate z_{xi} and error term $v_i \overset{iid}{\sim} N(0, \sigma_v^2)$ independent of u_i, then the resulting model for $(Y_i, X_i)^{\mathrm{T}}$ can be written

as

$$\begin{bmatrix} Y_i \\ X_i \end{bmatrix} = \begin{bmatrix} \theta_i \\ x_i \end{bmatrix} + \begin{bmatrix} e_i \\ \eta_i \end{bmatrix} \tag{3.5}$$

$$= \begin{bmatrix} z_i^{\mathrm{T}} & \beta z_{xi}^{\mathrm{T}} \\ 0 & z_{xi}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \delta \\ \delta_x \end{bmatrix} + \begin{bmatrix} u_i + \beta v_i \\ v_i \end{bmatrix} + \begin{bmatrix} e_i \\ \eta_i \end{bmatrix}, \tag{3.6}$$

where

$$\begin{bmatrix} e_i \\ \eta_i \end{bmatrix} \overset{iid}{\sim} N_2(0, \Omega), \quad \Omega = \begin{bmatrix} D_i & 0 \\ 0 & C_i \end{bmatrix} \tag{3.7}$$

$$\begin{bmatrix} u_i + \beta v_i \\ v_i \end{bmatrix} \overset{iid}{\sim} N_2(0, \Sigma), \quad \Sigma = \begin{bmatrix} \sigma_u^2 + \beta^2 \sigma_v^2 & \beta \sigma_v^2 \\ \beta \sigma_v^2 & \sigma_v^2 \end{bmatrix}. \tag{3.8}$$

This model differs from a standard bivariate FH model in that the parameter β affects both the regression mean function for Y_i and the random effect covariance matrix Σ. However, if the covariates z_{xi} are linear functions of the covariates z_i, then the fixed effects regression part of (3.6) can be re-parameterized to unrestricted linear regression effect $[z_i^{\mathrm{T}}\delta_y, z_{xi}^{\mathrm{T}}\delta_x]$ with regression covariates z_i for the first equation and z_{xi} for the second. With this reparameterization, β no longer affects the regression fixed effects, so the matrix Σ can then be re-parameterized in the general form $\Sigma = \{\sigma_{jk}\}_{jk}$, or by σ_{11}, σ_{22}, and $\rho = \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}}$, as there is now a one-to-one correspondence between $(\sigma_u^2, \sigma_v^2, \beta)$ and $(\sigma_{11}, \sigma_{22}, \sigma_{12})$ or $(\sigma_{11}, \sigma_{22}, \rho)$. Two instances where this condition on z_{xi} holds are (i) if the regression covariates are the same in both equations, $z_{xi} = z_i$, or (ii) if z_{xi} is just an intercept term, $z_{xi} = 1$ and z_i also includes an intercept.

Datta et al. (2018) study the area-level SME model, while Huang and Bell (2012) present a study examining use of general bivariate models for small area estimation. Analogous unit-level models have been studied by Ghosh et al. (2006) and Torabi et al. (2009). Fuller (2009) and Buonaccorsi (2010) discuss additional measurement error models including nonlinear models and the Berkson model.

In this chapter, we compare the FME and SME models with the naive FH model by examining their first order prediction mean squared errors (MSE). Note that the FME and SME models design the relation between the true unobserved quantities θ_i and x_i, whereas the naive FH model models the relation between θ_i and the observed X_i. The $X_1, \ldots, X_m$ contain noise in the form of generally heteroscedastic sampling error, and this heteroscedasticity produces the naive FH model's misspecification noted earlier. An interesting finding of this study is that, when the true model is given as the FME model, naive FH predictors can have lower MSE than the FME predictors, although it is a misspecified model.

The rest of this chapter is organized as follows. Section 3.2 summarizes theoretical findings for the three models. These include finding the probability limits of three models' parameter estimators and derivations of asymptotic prediction MSEs under the FME and SME models. Detailed derivations of these results are provided in Section 3.6. In Section 3.3, we compare theoretical MSEs of three models' predictors with various parameter values assuming a true SME model. In Section 3.4, MSE comparisons are made using the American Community Survey data. We summarize our findings in Section 3.5.

3.2 Theoretical results

To facilitate interpretation of the results, we present the simplest possible versions of the models outlined in the introduction. Specifically, the vector of non-measurement error covariates reduces to just an intercept term, i.e., $z_i = 1$. We note that all the derivations given in Section 3.6 are based on the general models.

To revert to standard notation, we use α for the intercept coefficient instead of δ, so the simplified model for θ_i in the FME and SME models given in (3.3) becomes

$$\theta_i = \alpha + \beta x_i + u_i. \tag{3.9}$$

For the SME model, we assume that $x_i \overset{iid}{\sim} N(\mu, \sigma_x^2)$ so there are no regression terms other than the mean μ in the model for x_i.

For the naive FH model (3.2), the simplified model for θ_i becomes

$$\theta_i = \alpha_N + \beta_N X_i + u_{i,N} \tag{3.10}$$

where, as before, we use the "N" subscript to distinguish the coefficients and random effects in the naive model (3.10) for θ_i, since this model differs from (3.9) by substituting X_i in place of x_i.

We find that the estimators for α, β, and σ_u^2 are the same for the FME and SME models even though they are obtained from different estimating equations. Please see equations (3.17) and (3.18) versus equations (3.34)–(3.36) in Section 3.6 for detailed derivations. These estimators are given by

$$\begin{aligned}
\hat{\beta} &= \frac{\frac{1}{m}\sum_{i=1}^{m}(X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{m}\sum_{i=1}^{m}(X_i - \bar{X})^2 - \bar{C}}, \\
\hat{\alpha} &= \bar{Y} - \hat{\beta}\bar{X}, \\
\hat{\sigma}_u^2 &= \frac{1}{m}\sum_{i=1}^{m}(Y_i - \hat{\alpha} - \hat{\beta}X_i)^2 - \bar{D} - \hat{\beta}^2\bar{C},
\end{aligned}$$

where $\bar{X} = m^{-1}\sum_{i=1}^{m} X_i$, with analogous definitions of $\bar{Y}$, $\bar{C}$, and $\bar{D}$. For fitting the SME model we also have $\hat{\mu} = \bar{X}$ and $\hat{\sigma}_x^2 = m^{-1}\sum_{i=1}^{m}(X_i - \bar{X})^2 - \bar{C}$.

Theorem 4. *For fitting the FME and SME models given by (3.3)–(3.4) and by (3.5)–(3.8), respectively, but simplifying $z_i^T\delta$ to just the intercept α, we have the following:*

$$\hat{\beta} \xrightarrow{p} \beta, \quad \hat{\alpha} \xrightarrow{p} \alpha, \quad \hat{\sigma}_u^2 \xrightarrow{p} \sigma_u^2$$

where $\xrightarrow{p}$ denotes convergence in probability under the true model (whether FME or SME).

Theorem 4 indicates that they are consistent for the true model parameters. The consistency thus holds for these estimators regardless of whether the true model is the FME or the SME. These statements also hold for the more general version of the models considered in Section 3.6.

The parameter estimator for fitting the naive FH model with the simplified model for θ_i as in (3.2) are given by

$$\begin{aligned}\hat{\beta}_N &= \left(\frac{\frac{1}{m}\sum_{i=1}^m (X_i - \bar{X})^2 - \bar{C}}{\frac{1}{m}\sum_{i=1}^m (X_i - \bar{X})^2}\right)\hat{\beta}, \\ \hat{\alpha}_N &= \bar{Y} - \hat{\beta}_N \bar{X}, \\ \hat{\sigma}^2_{u,N} &= \frac{1}{m}\sum_{i=1}^m (Y_i - \hat{\alpha}_N - \hat{\beta}_N X_i)^2 - \bar{D}.\end{aligned}$$

The probability limits of these estimators are given in the following Theorem.

Theorem 5. *For fitting the naive FH model in (3.2), but simplifying $z_i^T\delta$ to just the intercept α, and when the corresponding FME (SME) model is true, we have the following:*

$$\hat{\beta}_N \xrightarrow{p} a\beta, \quad \hat{\alpha}_N \xrightarrow{p} \alpha + (1-a)\beta\bar{x}, \quad \hat{\sigma}^2_{u,N} \xrightarrow{p} \sigma_u^2 + a\beta^2\bar{C}, \text{ under the FME model}$$

$$\hat{\beta}_N \xrightarrow{p} a_*\beta, \quad \hat{\alpha}_N \xrightarrow{p} \alpha + (1-a_*)\beta\bar{x}, \quad \hat{\sigma}^2_{u,N} \xrightarrow{p} \sigma_u^2 + a_*\beta^2\bar{C}, \text{ under the SME model}$$

where $\bar{x} = m^{-1}\sum_{i=1}^m x_i$ and the "attenuation factors" a and a_* are given by

$$a = \frac{s_x^2}{s_x^2 + \bar{C}}, \quad a_* = \frac{\sigma_x^2}{\sigma_x^2 + \bar{C}} \tag{3.11}$$

with $s_x^2 = \lim_{m\to\infty} m^{-1}\sum_{i=1}^m (x_i - \bar{x})^2$.

The result $\hat{\beta}_N \xrightarrow{p} a\beta$ is the well-known attenuation of the estimate of the regression parameter when measurement error is ignored. Please refer Theil (1971) for the FME case and Fuller (2009) for the SME case. The limit for $\hat{\sigma}^2_{u,N}$ shows that the naive FH model inflates the estimate of the model error variance σ_u^2 for the true FME (SME) model by the amount $a\beta^2\bar{C}$ ($a_*\beta^2\bar{C}$) due to the failure to account for the measurement error.

3.2.1 Small area predictors and their mean squared errors

The predictors of θ_i for the three models are given below. We note that any of these predictors will be used whenever the corresponding model is assumed regardless of the true model.

$$\text{FME predictor} \quad : \quad \hat{\theta}_{i,F} = Y_i - \frac{D_i \left\{ Y_i - \hat{\alpha} - \hat{\beta} X_i \right\}}{D_i + \hat{\sigma}_u^2 + \hat{\beta}^2 C_i}$$

$$\text{SME predictor} \quad : \quad \hat{\theta}_{i,S} = Y_i - \frac{D_i \left\{ Y_i - \hat{\alpha} - \hat{\beta} X_i + \hat{\beta}(C_i/(\hat{\sigma}_x^2 + C_i))(X_i - \bar{X}) \right\}}{D_i + \hat{\sigma}_u^2 + \hat{\beta}^2 C_i \hat{\sigma}_x^2/(\hat{\sigma}_x^2 + C_i)}$$

$$\text{naive FH predictor} \quad : \quad \hat{\theta}_{i,N} = Y_i - \frac{D_i \left\{ Y_i - \hat{\alpha}_N - \hat{\beta}_N X_i \right\}}{D_i + \hat{\sigma}_{u,N}^2}.$$

These are the empirical versions of the optimal predictors (best linear unbiased predictors) under their respective assumed models. The FME predictor $\hat{\theta}_{i,F}$ is given by Theorem 1 of Ybarra and Lohr (2008), and $\hat{\theta}_{i,N}$ is the predictor of Fay and Herriot (1979) for the naive FH model.

Several special cases are worth noting from above predictors. First, as $D_i \to 0$ all the predictors converge to the direct survey estimate Y_i, and since its sampling variance is going to 0, all the predictors achieve design consistency. Second, if $C_i = \bar{C}$, it can be seen that the SME and naive FH predictors agree whereas the FME predictor generally remains different. These statements are also true for the more general model considered in the Section 3.6. Third, it can be seen that as $\hat{\sigma}_x^2 \to \infty$, the SME predictor converges to the FME predictor. The same holds as $C_i \to 0$, which implies that x_i is known in the limit. We can put these together and say that the SME and FME predictors behave similarly when C_i/σ_x^2 is small.

It can be shown that $\hat{\theta}_{i,S}$ can be obtained by taking $\hat{\theta}_{i,F}$ and replacing X_i in the numerator of the fraction by $\mathsf{E}(x_i|X_i) = X_i - \{C_i/(\hat{\sigma}_x^2 + C_i)\}(X_i - \bar{X})$, and $C_i = \mathsf{Var}(X_i - x_i)$ in the denominator of the fraction by $\mathsf{Var}(x_i|X_i) = C_i\hat{\sigma}_x^2/(\hat{\sigma}_x^2 + C_i)$, where these are the conditional mean and variance of x_i given X_i under the estimated model.

Table 3.1 summarizes the first order biases and prediction error variances of the three predictors assuming the FME model is true. This gives a realistic approximation for the case when m, the number of small areas, is large. The FME and SME model predictors then use

Prediction model	**Bias**	**Prediction error variance**
FME	0	$\frac{(\sigma_u^2+\beta^2 C_i)D_i}{\sigma_u^2+\beta^2 C_i+D_i}$
SME	$\frac{-\beta D_i C_i}{F_i}(x_i-\bar{x})$	$D_i - \frac{D_i^2[(s_x^2+C_i)+\beta^2 s_x^2 C_i^2/F_i]}{F_i}$
naive FH	$\frac{-\beta D_i(1-a)}{\sigma_u^2+a\beta^2\bar{C}+D_i}(x_i-\bar{x})$	$\frac{(\sigma_u^2+a\beta^2\bar{C})D_i}{\sigma_u^2+a\beta^2\bar{C}+D_i} + \left(\frac{\beta D_i}{\sigma_u^2+a\beta^2\bar{C}+D_i}\right)^2 a(aC_i-\bar{C})$

Table 3.1: Biases and prediction error variances when the FME model is true.

Prediction model	**Prediction error variance = MSE**
FME	$\frac{(\sigma_u^2+\beta^2 C_i)D_i}{\sigma_u^2+\beta^2 C_i+D_i}$
SME	$D_i - \frac{D_i^2(\sigma_x^2+C_i)}{F_i^*}$
naive FH	$\frac{(\sigma_u^2+a_*\beta^2\bar{C})D_i}{\sigma_u^2+a_*\beta^2\bar{C}+D_i} + \left(\frac{\beta D_i}{\sigma_u^2+a_*\beta^2\bar{C}+D_i}\right)^2 a_*^2(C_i-\bar{C})$

Table 3.2: Prediction error variances when the SME model is true

the true values of α, β, and σ_u^2, while for the naive FH model predictors, the limiting values of $\hat{\alpha}_N$, $\hat{\beta}_N$, and $\hat{\sigma}_{u,N}^2$ as shown in Theorem 5 are used. Table 3.1 entries for the naive FH model use the quantity

$$F_i = (\sigma_u^2 + D_i)(s_x^2 + C_i) + \beta^2 s_x^2 C_i. \tag{3.12}$$

Table 3.2 gives the results for the case when the SME model is true. In this case all the predictors are unbiased in the sense that $\mathsf{E}(\hat{\theta}_i - \theta_i) = 0$. Hence, the table just gives the prediction error variances, which are also the MSEs. For F_i^* in Table 3.2, we substitute σ_x^2 for s_x^2 in the expression (3.12), analogous to the definition of a_* in (3.11). Several points are worth noting about the results of Table 3.1 and Table 3.2.

(I) The results for the FME predictor are the same in both cases, i.e., whether the FME or SME model is true. To achieve unbiasedness under the assumption that the x_i are fixed but unknown quantities, the FME predictor eliminates them from the prediction error. Hence, its prediction error results are not affected by whether the x_i actually are fixed and unknown or are random variables following some distribution, as the SME model assumes.

(II) When the FME model is true, the biases of the SME and naive FH predictors are proportional to $(x_i - \bar{x})$, which is unconstrained, and so can be arbitrarily large in magnitude. Hence, for areas where $|x_i - \bar{x}|$ is large, the squared bias can dominate the prediction MSE. Since the x_i are unobserved, it will typically be difficult to estimate the squared bias (unless the C_i are small so the X_i are very good predictors of the x_i, in which case the motivation to use a measurement error model diminishes).

(III) The MSEs in Table 3.2 can be obtained by taking the expressions for squared bias plus prediction error variance from Table 3.1 and substituting σ_x^2 for s_x^2 and for $(x_i - \bar{x})^2$. This is the difference between assuming the x_i fixed and unknown versus assuming $x_i \overset{iid}{\sim} N(\mu, \sigma_x^2)$. Notice that doing this for the SME predictor results in cancellation of a term and the simpler expression for the MSE when the SME model is true. It also results, after a little algebra, in a simplification of the formula for the MSE of the naive FH predictor.

(IV) As noted earlier, if an area has $C_i = \bar{C}$ then the SME and naive FH predictors agree. Hence, when $C_i = \bar{C}$, the biases and prediction error variances of the SME and naive FH predictors are the same.

(V) When the SME model is true and when $C_i = \bar{C}$, the "reported MSE" for the naive FH model will agree with the true MSE. The reported MSE is the MSE one would compute assuming the naive FH model to be true, and is given by the first term in the naive FH MSE expression in Table 3.2. The second term is obviously zero when $C_i = \bar{C}$. We can

thus see in this case that when $C_i > \bar{C}$ the term in brackets is positive so the reported MSE understates the true MSE, while when $C_i < \bar{C}$ the term in brackets is negative so the reported MSE overstates the true MSE. The misspecification of the naive FH model when the SME model is true (and the C_i are not constant) can thus lead to substantial misstatement of the MSEs except for areas for which C_i is close to $\bar{C}$.

An implication of points (IV) and (V), and the analogous result stated earlier for the point predictors, is that if $C_i = \bar{C}$ for all $i = 1, \ldots, m$, then the prediction results for the SME and naive FH models are the same. This provides some basis for the statement sometimes made that measurement error in covariates doesn't affect model prediction. Put another way, this statement is true only if the C_i are constant for all areas, and only when comparing prediction results for the naive FH model to those for the SME model. Prediction results for the FME model will be different.

3.3 MSE comparisons of alternative predictors

We now compare the performance of the three alternative model predictors using simulated data from a true SME model. Comparisons are made by considering a range of values for the model parameters and D_i and C_i values. We use the MSE results of Table 3.2 and examine percentage differences in MSEs

$$100\left(\frac{MSE_F}{MSE_S} - 1\right) \tag{3.13}$$

for comparing MSEs of the FME and SME predictors. We similarly define the analogs to (3.13) for comparing MSEs for the naive FH and FME predictors, and for the naive FH and SME predictors, as well as for comparing the reported and actual MSEs of the naive FH predictor. We denote the reported MSE for the naive FH predictor by $\widehat{MSE}_N$. We assume that the SME model is true model as this facilitates the comparisons. Assuming true FME model leads to the complication that the MSEs for the SME and naive FH models depend

on x_i which has unrestricted variation over areas. This is examined in Section 3.4 using the American Community Survey data.

For making relative comparisons as in (3.13), the scale of the data does not matter, so rescaling Y_i and X_i to Y_i/σ_u and X_i/σ_x will not affect these comparisons. These rescalings reduce the number of varying parameters to consider by two, which lets us express σ_u^{-2} times MSE_F, MSE_S, MSE_N and $\widehat{MSE}_N$, all computed assuming the SME model is true, in terms of the following four scale invariant quantities:

$$r_D = \frac{D_i}{\sigma_u^2}, \quad r_C = \frac{C_i}{\sigma_x^2}, \quad \rho = \text{corr}(\theta_i, x_i), \quad \bar{r}_C = \frac{\bar{C}}{\sigma_x^2}. \tag{3.14}$$

The Section 3.6.3 illustrates such re-expression for the calculation of MSE_S/σ_u^2. To simplify the notation, we omit the subscript i from r_D and r_C. To compare MSEs, we examine contour plots over (r_D, r_C) for each of the MSE percentage differences defined as in (3.13), viewing the MSE percentage difference as a function of (r_D, r_C). We examine such plots for fixed values of ρ and $\bar{r}_C$, which do not vary over i. Figure 3.3.1 gives contour plots of (3.13) for $\rho = 0.3$ and $\rho = 0.7$. The x- and y-axes of the plots, representing the values of r_D and r_C, range from 0.1 to 10, and are shown with $\log_{10}$ scaling. We need not set $\bar{r}_C$ for these comparisons because MSE_F and MSE_S do not depend on $\bar{C}$. The percentage differences are all positive, favoring the SME model which here is assumed to be true, and the differences increase with both r_D and r_C so that the more sampling or measurement error is present, the larger is the advantage to use of the SME predictor. When either r_D or r_C is small, say generally below 1, the MSE percentage differences are small as no contours show up plotted in this area, and choice of model has little effect on prediction accuracy. In fact, when both r_D and r_C are sufficiently small the FME and SME predictors are both close to the direct estimator Y_i, leading to small MSE differences, a pattern repeated in subsequent graphs. Towards the upper right corner, the MSE percentage differences become substantial in both graphs, and the difference is larger for $\rho = 0.7$. Analysis for MSE_F/MSE_S reveals that, for given values of r_C and r_D, the MSE percent differences increase with $\rho > 0$ to the point, where $\rho = \{1 + r_C/\sqrt{(1 + r_C)(1 + r_D)}\}^{-1/2}$, and then they decline to 0 as ρ increases to

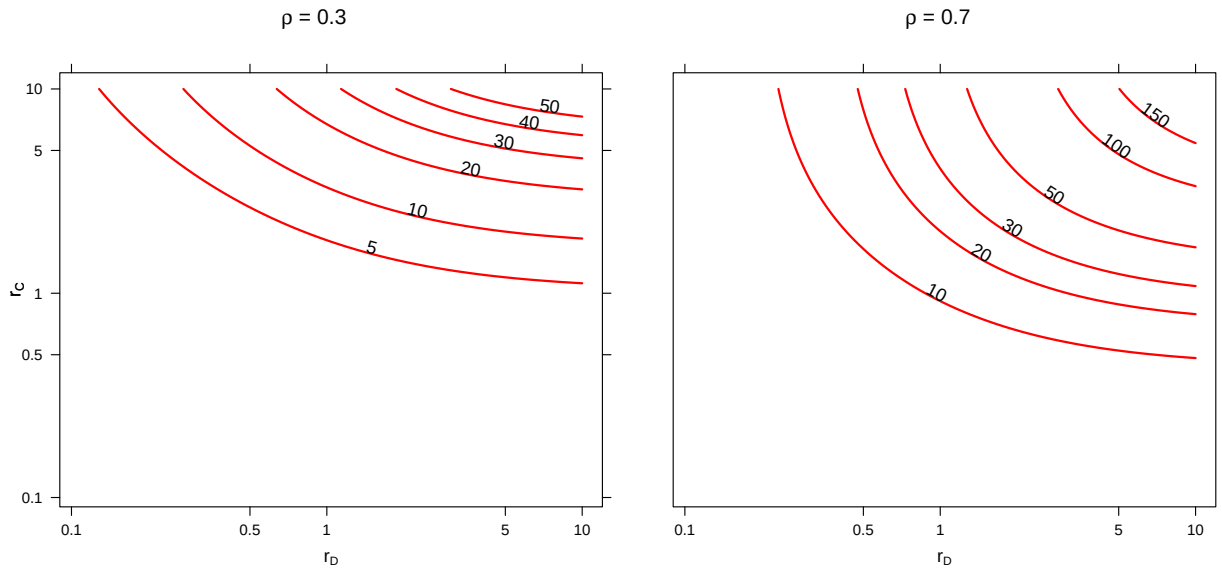


Figure 3.3.1: Contours of $100(MSE_F/MSE_S - 1)$ for two values of ρ when the SME model is true.

1. Over the range of values $[1, 10]$ for r_C and r_D, this maximum point varies from about $\rho = 0.57$ to $\rho = 0.91$. Note that the results for ρ and for $-\rho$ would be the same since the MSEs actually depend on ρ^2.

For the other MSE comparisons, which are shown in Figures 3.3.2–3.3.4, the percentage differences depend on all four quantities in (3.14). To get a general idea of how the comparisons vary, we take $\rho = 0.7$ as a representative value, and examine contour plots for $\bar{r}_C = 0.1, 1$ and 10. Figure 3.3.2 shows contour plots of (3.13) for the naive FH and FME predictors. In these plots, we see both positive and negative contours, indicating regions where the FME predictor does better, and other parts where the naive FH predictor does better. The patterns in these plots can be understood by keeping in mind that: (1) for small values of r_C, the FME predictor acts like the SME predictor, which here is optimal, so the FME predictor performs well and (2) for r_C close to $\bar{r}_C$ the naive FH predictor acts like the SME predictor and so performs well. Thus, in the plot for $\bar{r}_C = 0.1$, both the FME and

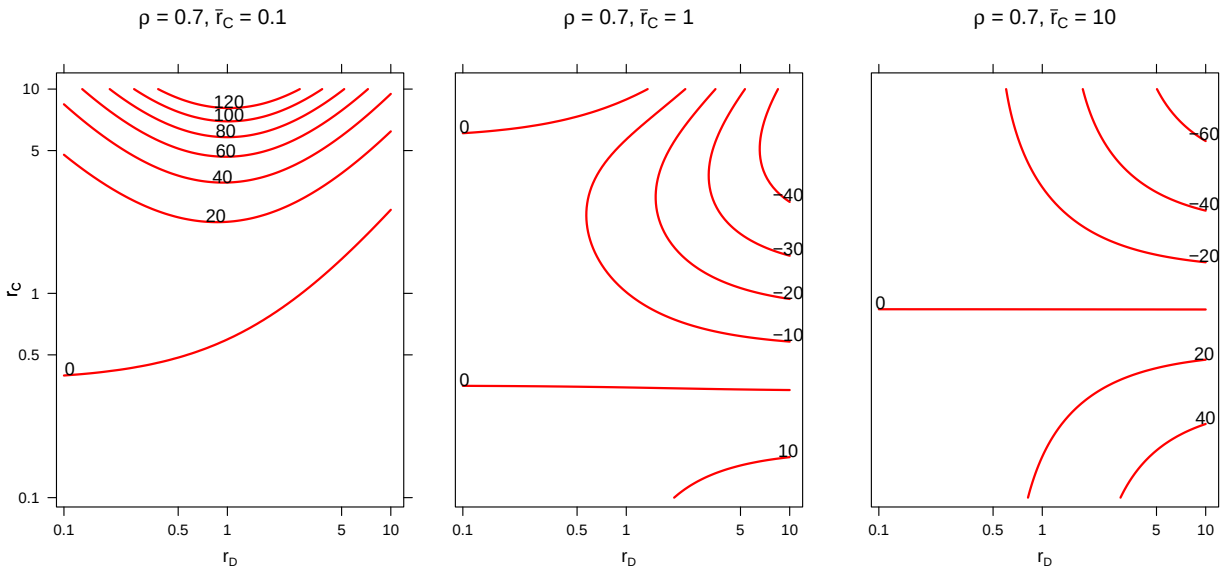


Figure 3.3.2: Contours of $100(MSE_N/MSE_F - 1)$ for $\rho = 0.7$ and $\bar{r}_C = 0.1, 1$ and 10 when the SME model is true.

naive FH predictors perform similarly to the optimal SME predictor for small values of r_C, so there is little difference in their MSEs. Apart from this case where both perform well, the naive FH predictor performs better when r_C is sufficiently close to $\bar{r}_C$, where the meaning of sufficiently close depends on the values of $\bar{r}_C$ and r_D.

The results in Figure 3.3.2 showing that for certain regions the naive FH predictor has lower prediction MSE than the FME predictor. As the true model is the SME model, the naive FH model is misspecified since it ignores the measurement error in X_i. In contrast, the FME model accounts for the measurement error in X_i. Since it makes no assumptions about the x_i, it is not inconsistent with the true SME model. In fact, as we move towards larger amounts of measurement error overall (larger values of $\bar{r}_C$), the MSE advantages of the naive FH predictor become more substantial and cover larger ranges of the r_C and r_D values. The general explanation for this is that, when measurement error is substantial, the FME models

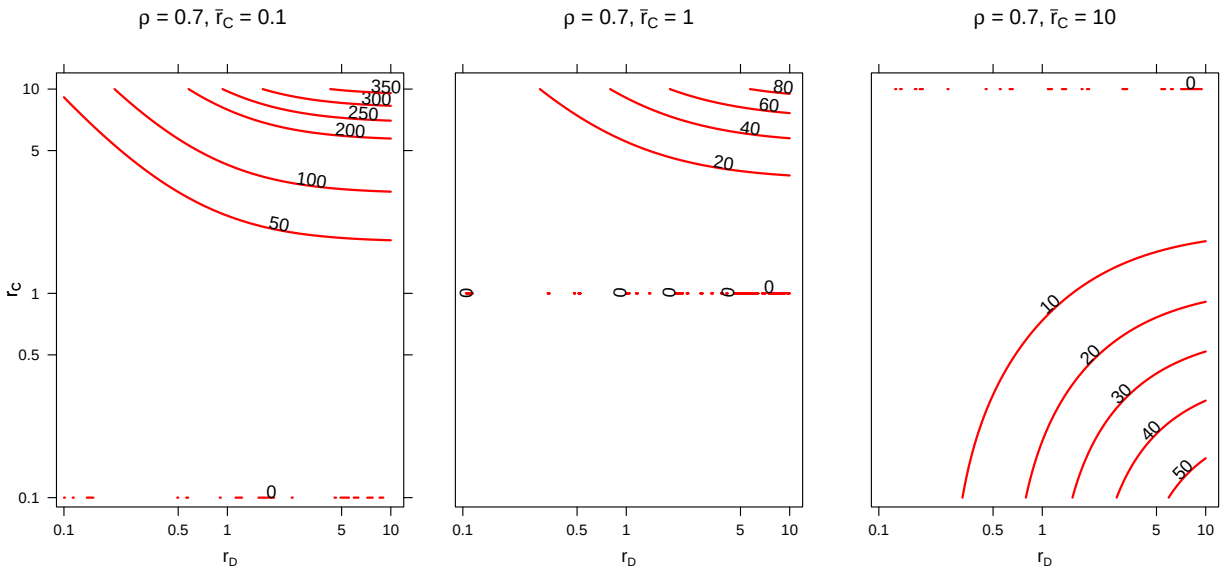


Figure 3.3.3: Contours of $100(MSE_N/MSE_S - 1)$ for $\rho = 0.7$ and $\bar{r}_C = 0.1, 1$ and 10 when the SME model is true.

avoidance of any modeling assumptions about the x_i can lead to rather inefficient use of the data X_i while the naive FH predictor makes suboptimal but better use of the X_i unless r_C is very different from $\bar{r}_C$.

Figure 3.3.3 gives contour plots of $100(MSE_N/MSE_S - 1)$, comparing MSEs for the naive FH and SME predictors. Since the SME model is assumed be the true model, all the contours shown are positive indicating naive FH has uniformly larger MSE. Each plot shows zero contour at $r_C = 0.1, 1$ and 10, and these zero contours occur as horizontal lines when $r_C = \bar{r}_C$ which is when the naive FH and SME predictors agree.

Apart from this, the plots for $\bar{r}_C = 0.1$ and 1 show substantial positive contours for large values of r_C that also increase with r_D, while for $\bar{r}_C = 10$ the substantial positive contours occur for small r_C as r_D grows large.

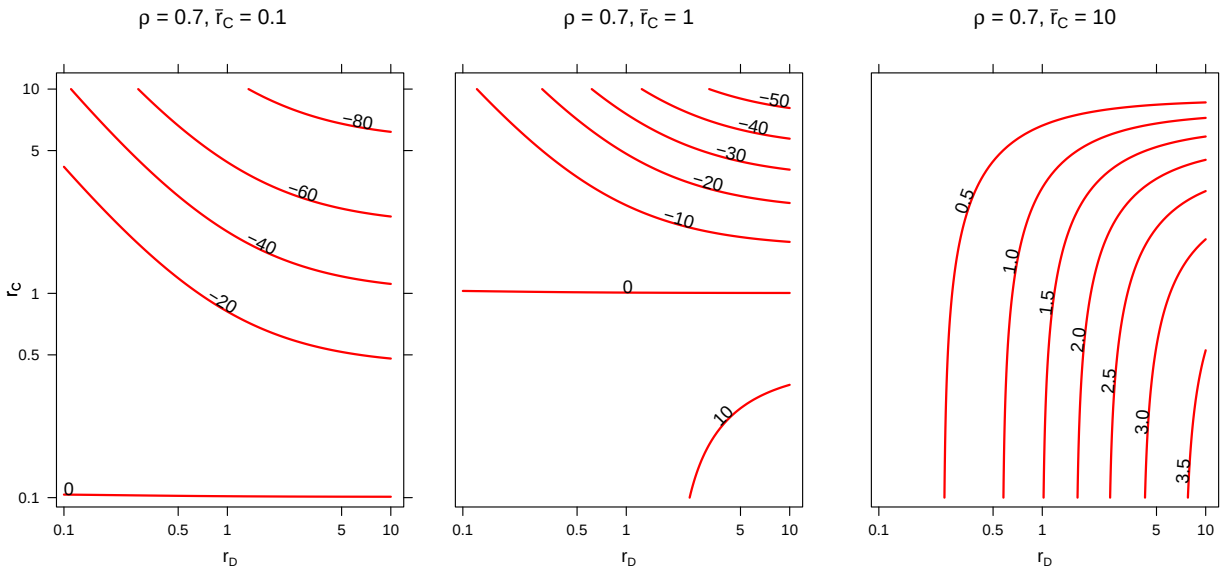


Figure 3.3.4: Contours of $100(\widehat{MSE}_N/MSE_N - 1)$ for $\rho = 0.7$ and $\bar{r}_C = 0.1, 1$ and 10 when the SME model is true.

Figure 3.3.4 gives contour plots of the MSE percentage difference of the reported and actual MSEs for the naive FH predictor. As with Figure 3.3.3, the three plots should show zero contours at the values $\bar{r}_C = 0.1, 1$, and 10, respectively. In these plots the regions above the zero contours have negative values that reflect understatement of the true MSE by the reported MSE, while the regions below the zero contours have positive values that reflect overstatement of the true MSE. The first two plots show regions for $r_C > \bar{r}_C$ with significant understatement of the true MSE, while the second two reflect at most very minor overstatement of the true MSE when $r_C < \bar{r}_C$. This pattern remains when the axis ranges are expanded to include larger values of r_C and r_C. While further extrapolation of these results to more general cases than those considered here is questionable, they nonetheless suggest that understatement of MSE by the naive FH model may be a potentially more serious problem than overstatement.

3.4 Application to county-level poverty rate data

The previous section compared MSEs of the SME, FME and naive FH predictors across a range of D_i, C_i and model parameter values. In this section, we fit the models to American Community Survey data to examine realistic behaviors of the models. In particular, we consider modeling county poverty rates of school-age children for the U.S. counties.

We fit the SME and FME models to the data, and then use the theoretical expressions given in Table 3.1 and Table 3.2 to obtain small area prediction MSEs for the three alternative model predictors. For each of which, we consider a fitted model as the truth. For the FME model, we generate the true covariate values x_i from the $N(0, \hat{\sigma}_x^2)$, where $\hat{\sigma}_x^2$ is obtained from the fitted SME model. We note that the purpose of using the poverty rate data is to illustrate the results obtained in Section 3.2 rather than producing poverty estimates.

The American Community Survey (ACS) is the largest U.S. household survey. the ACS produces annual estimates based on one year or five years of data collection, and we use the 2007-2011 aggregated county-level data with $m = 3129$ counties after removing missing values. We consider the estimates of county-level school-aged children poverty rates as the primary response variable Y_i and the total county poverty rates as a covariate X_i which is subject to measurement error. For D_i, C_i, we use estimates of the sampling variances of the school-aged children poverty and total poverty rates, respectively.

We first assume the SME model is true, and obtain maximum likelihood estimates $\hat{\sigma}_u^2$, $\hat{\sigma}_x^2$, $\hat{\alpha}$, and $\hat{\beta}$. Assuming these estimates are true value of the parameters, we compute the first order approximations to the MSEs of the naive, FME, and SME model predictors when the SME model is true. We note that the parameter estimates are fairly accurate as the number of small areas is large.

Figure 3.4.1 (a)–(c) displays ratios comparing first order approximations of three model predictors' MSEs plotted against C_i on the $\log_{10}$ scale, with a vertical line at $\log_{10} \bar{C}$ for reference. Panel (a) shows the ratios of MSEs for the SME and naive predictors. We note that, due to their optimality, the SME model predictors always have smaller prediction MSEs

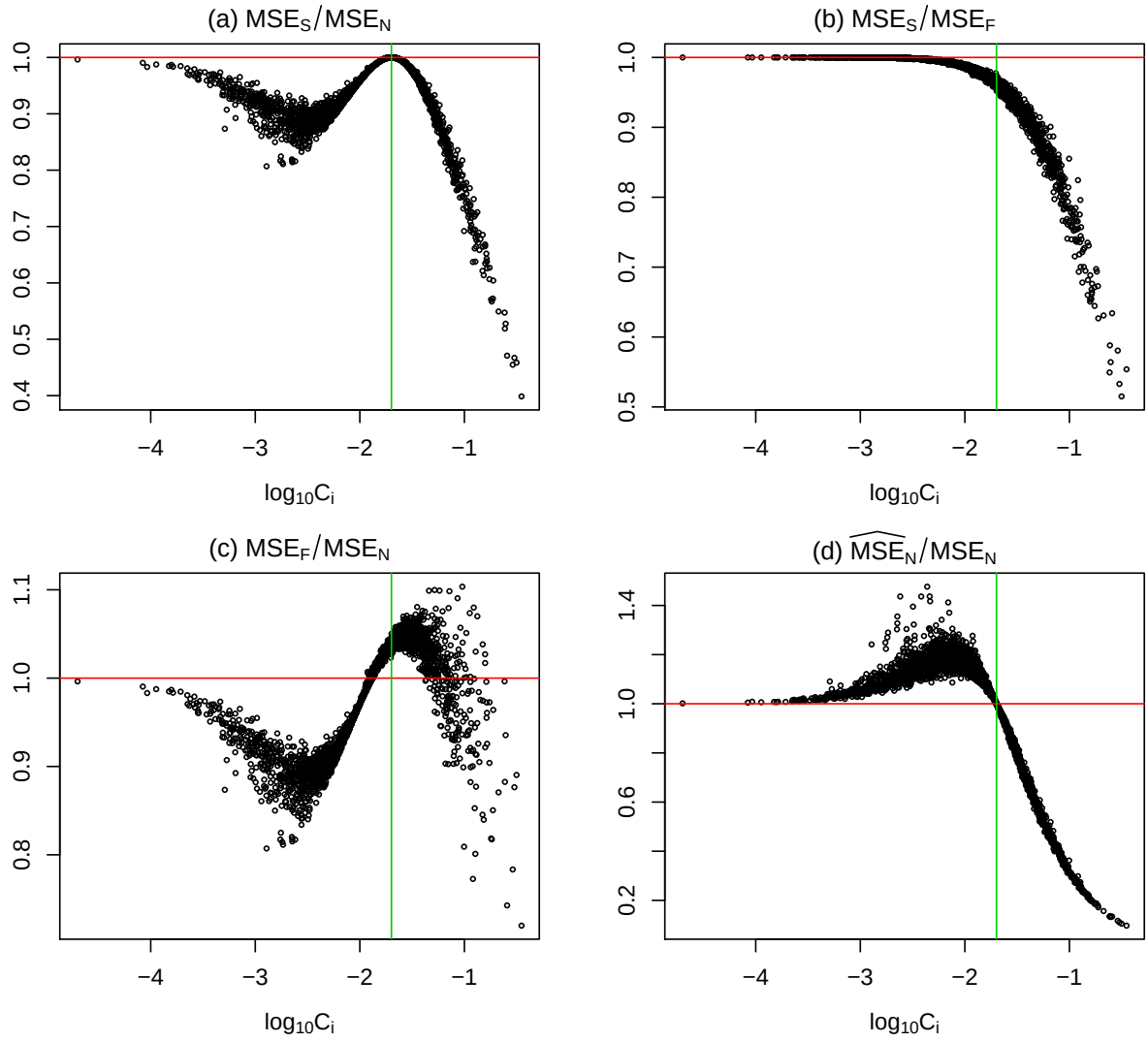


Figure 3.4.1: First order approximations of MSE ratios against $\log_{10} C_i$ when structural measurement error is true.

than the naive model predictors. Because the C_i's are strongly correlated with the D_i's (with a correlation of about 0.95), for small C_i's all three model predictors are approximately equal to the direct estimators, and hence the MSEs of the naive and SME predictors are similar. We will see this trend in all four panels of Figure 3.4.1. In panel (a), the ratio reaches its maximum of approximately one when $C_i \approx \bar{C}$. This agrees with the finding in Section 3.2.1

that naive and SME predictors are identical when $C_i = \bar{C}$. When the measurement error variance is larger than $\bar{C}$, $C_i > \bar{C}$, the ratios decline rapidly to values that reflect up to 60% larger MSEs for the naive FH model compared to the SME model.

Panel (b) shows the ratios of the SME and FME predictors' MSEs. As expected, the SME predictor performs uniformly better than the FME predictor. The differences become prominent as the measurement error variance C_i increases, with the FME predictor MSEs up to 50% higher than those for the SME predictor.

Panel (c) shows the corresponding ratios for the FME and naive FH predictors. The naive FH predictor has slightly higher MSEs than the FME predictor for small C_i and lower MSEs for large C_i's, a pattern expected from the results in panels (a) and (b). The two predictors' MSEs are approximately equal at some point which is lower than $\bar{C}$. The FME predictor's MSEs are up to 20% larger than the naive predictor's MSEs. This may seem surprising given that the naive model is misspecified while the FME model is not in the sense that the FME modeling assumptions agree with those of the SME model, while making no assumptions about the x_i.

Note that the MSE that is obtained for the FME predictor when the SME model is actually true is still correct to the first order, even though the FME predictor is not optimal. However, the MSE obtained assuming the naive model is true, what we call the "reported" MSE, differs from the naive model predictor's true MSE. Panel (d) plots the ratios of the first order approximations of the reported and true MSEs of the naive model predictor when the SME model is true. As noted in Section 3.2.1, the naive model overstates the MSEs for small C_i's and understates them for large C_i's, while correctly estimating the MSE at $C_i = \bar{C}$. The overstatement for $C_i < \bar{C}$ is relatively small, less than 40%, while the understatement becomes large, increasing with increasing C_i to more than 80%.

To investigate the performance of each of the predictors when the FME model holds, we generate $x_i \overset{iid}{\sim} N(0, \hat{\sigma}_x^2)$ as we centered X_i so that $\hat{\mu}_x = \bar{X} = 0$. We used the parameter

values obtained from fitting the SME model as the FME and SME parameter estimators agree. Please refer Section 3.6.1.3 for details.

Figure 3.4.2 panels (a)-(d) are analogous to Figure 3.4.1, but assume for the first order approximations that the FME model is true. Panel (a) plots the ratios of the SME and FME predictor MSEs. Despite having generated the data under the assumptions which make the FME predictor "optimal," it frequently performs worse than the SME predictor with respect to the MSE. However, the FME predictor's optimality is in the class of unbiased predictors, and both the SME and the naive predictors are biased, so there is no contradiction. The difference in MSEs can be up to about 70% in either direction, with two points where the SME MSEs are about 70% higher than the FME MSE. However, there are relatively few points for which the SME MSE is more than 25% higher than the FME MSE, while there are many points where the SME MSE is more than 25% lower than the FME MSE. Examining the bias and variance terms of the MSE separately reveals that the points where the SME predictor performs worse than the FME predictor in panel (a) are due to the bias of the former.

Panel (b) in Figure 3.4.2 displays the ratios of MSEs of the SME and naive FH predictors against C_i. It shows that when the FME model is true, the SME predictor sometimes performs better and sometimes performs worse than the naive predictor in terms of MSE. The same statement can be made about the functional and naive predictors based on panel (c), which shows the ratios of the FME and naive FH MSEs plotted against C_i.

Panel (d), which plots the ratio of reported to true MSE of the naive predictor, reminds us that the naive model will misstate the mean squared error, sometimes overstating it and sometimes understating it. The overstatement is relatively small, up to about 40%, but the understatement is more considerable, up to more than 80%.

Since both the SME and naive predictors are biased under the FME model, we can analyze the relationship between the respective biases. Panel (e) of Figure 3.4.2 shows the ratio of the bias squared of the SME predictors and the naive predictors. It shows that the

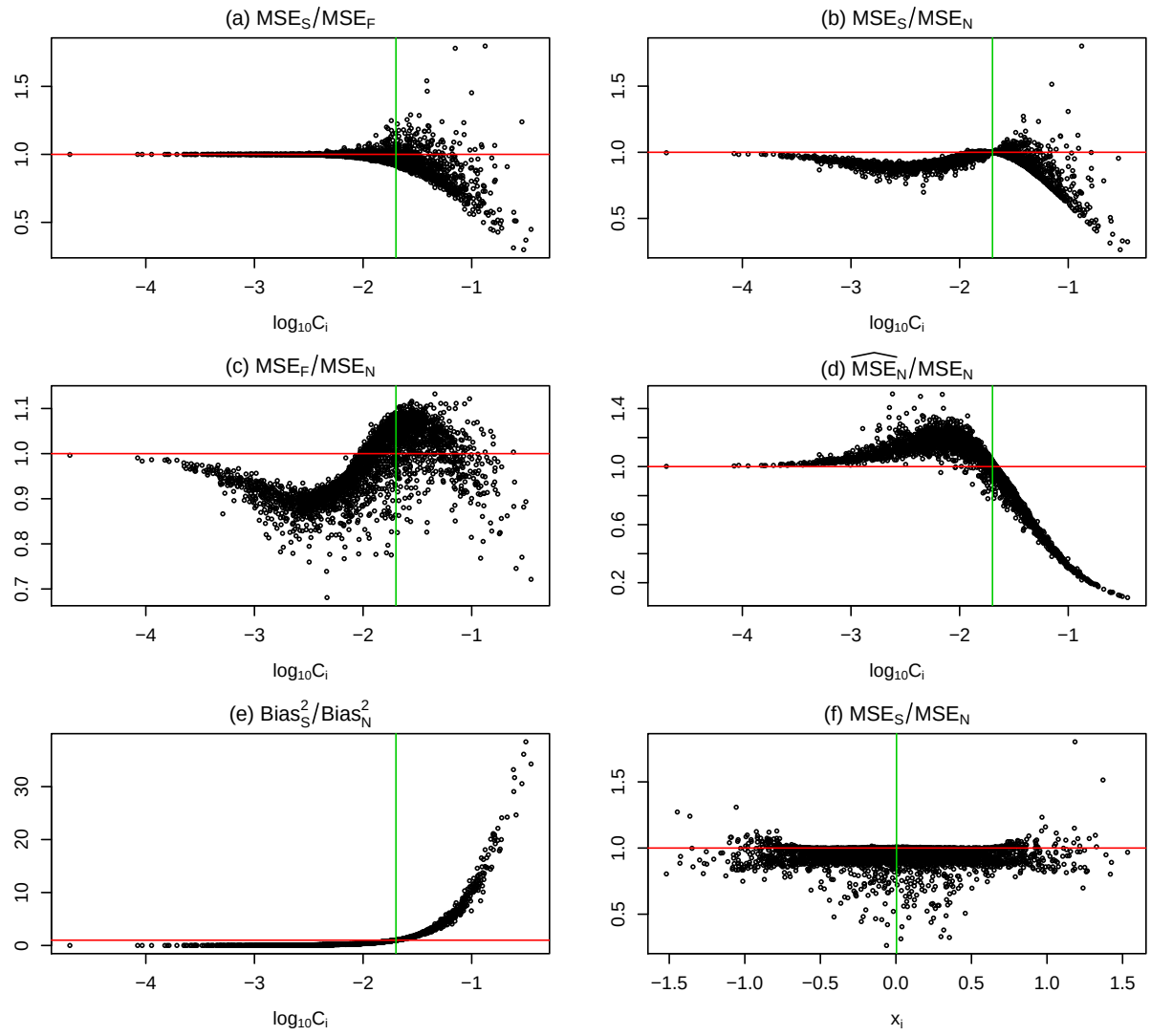


Figure 3.4.2: First order approximations of MSE ratios against $\log_{10} C_i$ when functional measurement error is true.

SME predictor has lower absolute bias for $C_i < \bar{C}$, and higher absolute bias for $C_i > \bar{C}$. Their absolute biases are the same when $C_i = \bar{C}$ as the two estimators are equivalent. This indicates that the extreme points in the top right quadrant in panel (b) are due to the bias. Panel (f) displays the ratio of the SME and naive FH true MSEs against x_i. The vertical line represents the mean of x_i, which is 0 as X_i's are centered. We note that the extreme

points in the top quadrants have high deviations of x_i from its mean. On the other hand, the most extreme points in the bottom quadrants correspond to values where x_i is close to $\bar{x}$. This indicates large deviations of x_i's from $\bar{x}$ will have more impact on the true MSEs of the SME predictors than on those of the naive FH predictors. For the majority of points, however, the SME model's MSEs are lower than those of the naive FH model based on our first order approximations.

3.5 Conclusions

In this chapter, we compared three small area estimation models, FME, SME and naive FH models, when one or more covariates are measured with error. Section 3.2 established relevant theoretical results for these models. These included asymptotic studies of parameter estimation, their predictions, and their corresponding MSEs. This led to several observations relating the models. First, the naive FH and SME model predictions and MSEs asymptotically agree for areas with $C_i = \bar{C}$. Secondly, SME predictor converges to FME predictor as $\sigma_x^2 \to \infty$ or $C_i \to 0$. Thus, when C_i/σ_x^2 is small, two predictors behave similarly. Lastly, in the presence of measurement error, the naive FH model over or under estimate the prediction MSE except for areas with $C_i = \bar{C}$ as it is misspecified model.

In Section 3.3, prediction MSEs of the three models are compared by considering various ranges of the parameter values for the case assuming SME model is the true model. Section 3.4 made analogous comparisons using the data on poverty rates of school-age children for U.S. counties. We fitted a particular SME model which resembles the models used by the Census Bureaus SAIPE program, so its use provided results for a realistic case of a true SME model. MSE comparisons were also obtained for an analogous FME model by simulating values of the unobserved true covariate values x_i. The MSE comparisons of Sections 3.3 and 3.4 tended to favor the SME model overall. Comparisons to the naive FH model showed that the naive FH predictor performs poorly for C_i not near $\bar{C}$, with substantial MSE increases compared to the SME predictor. Regarding the naive FH models additional problem of

misstatement of MSE for C_i apart from $\bar{C}$, understatement of MSE when $C_i > \bar{C}$ appeared more serious than overstatement of MSE for $C_i < \bar{C}$.

From the comparisons of the SME and FME models, it is noted that when the SME model is true, the FME predictor can have substantially higher prediction MSE when sampling and measurement errors are large (D_i and C_i are large). While the FME predictor can be best when the FME model is the underlying model, it was also not unusual in this case for the SME and naive FH predictors to actually have lower MSEs than the optimal FME predictor. We note that the optimality of the FME predictor under the FME model is among the class of unbiased predictors given fixed x_i. The SME and naive FH predictors, being biased, fall outside this class and so can have lower MSE. It appears that while the avoidance of modeling assumptions for the x_i gives the FME model some potential for robustness, this can come at a high cost in terms of larger prediction error variances for some areas.

3.6 Technical Details

3.6.1 Asymptotic evaluations under FME model

3.6.1.1 FME estimators and predictors

We consider the following functional measurement error (FME) model:

$$Y_i = \theta_i + e_i,\ \theta_i = x_i^{\mathrm{T}}\beta + z_i^{\mathrm{T}}\delta + u_i,\ X_i = x_i + \eta_i, \quad i = 1, \ldots, m, \tag{3.15}$$

where X_i and z_i are $p \times 1$ and $q \times 1$ vectors, and e_i's, u_i's and η_i's are independently distributed, each with zero mean, and respective variance (covariance matrix) given by D_i, σ_u^2 and C_i. We define $Y = (Y_1, \ldots, Y_m)^{\mathrm{T}}$, $C_i^* = \mathrm{diag}(C_i, 0_{qq})$, and

$$A = \begin{bmatrix} X_1^{\mathrm{T}} & z_1^{\mathrm{T}} \\ \vdots & \vdots \\ X_m^{\mathrm{T}} & z_m^{\mathrm{T}} \end{bmatrix}, \tag{3.16}$$

which is the observed design matrix of covariates. We use $\mathsf{E}_F(\cdot)$ to denote the expectation under the FME model. Let $\gamma = (\beta^{\mathrm{T}}, \delta^{\mathrm{T}})^{\mathrm{T}}$. Under the model in (3.15)

$$\mathsf{E}_F\left\{(A^{\mathrm{T}}A - \sum_{i=1}^{m} C_i^*)\gamma\right\} = \mathsf{E}_F(A^{\mathrm{T}}Y),$$

which leads to an unbiased estimating equation,

$$(A^{\mathrm{T}}A - \sum_{i=1}^{m} C_i^*)\gamma = A^{\mathrm{T}}Y, \tag{3.17}$$

for γ. Also, for $i = 1, \ldots, m$, $\mathsf{E}_F\left[(Y_i - X_i^{\mathrm{T}}\beta - z_i^{\mathrm{T}}\delta)^2\right] = D_i + \sigma_u^2 + \beta^{\mathrm{T}}C_i\beta$ leads to another unbiased estimating equation,

$$\frac{1}{m}\sum_{i=1}^{m}(Y_i - X_i^{\mathrm{T}}\beta - z_i^{\mathrm{T}}\delta)^2 - \bar{D} - \beta^{\mathrm{T}}\bar{C}\beta - \sigma_u^2 = 0. \tag{3.18}$$

Assuming $A^{\mathrm{T}}A - \sum_{i=1}^{m} C_i^*$ is nonsingular, an estimator of γ follows from (3.17). We denote this estimator by $\hat{\gamma}_F$ (the subscript F is used to denote the FME model fitting), which is

$$\hat{\gamma}_F = (A^{\mathrm{T}}A - \sum_{i=1}^{m} C_i^*)^{-1}A^{\mathrm{T}}Y. \tag{3.19}$$

Using this in (3.18), we have

$$\hat{\sigma}_{u,F}^2 = \frac{1}{m}\sum_{i=1}^{m}(Y_i - X_i^{\mathrm{T}}\hat{\beta}_F - z_i^{\mathrm{T}}\hat{\delta}_F)^2 - \bar{D} - \hat{\beta}_F^{\mathrm{T}}\bar{C}\hat{\beta}_F. \tag{3.20}$$

We assume that the following limits exist, where

$$\lim_{m\to\infty}\frac{1}{m}\sum_{i=1}^{m}\begin{bmatrix} x_i \\ z_i \end{bmatrix}\begin{bmatrix} x_i \\ z_i \end{bmatrix}^{\mathrm{T}} = K, \tag{3.21}$$

$$\lim_{m\to\infty}\frac{1}{m}\sum_{i=1}^{m} C_i^* = \mathrm{diag}(\bar{C}, 0_{qq}), \tag{3.22}$$

and K and $\bar{C}$ are assumed positive definite. Let $\xrightarrow{p_F}$ denote weak convergence under the FME model. Then under the model in (3.15), with the assumptions in (3.21) and (3.22), and by weak laws of large numbers, we get

$$\frac{1}{m}(A^{\mathrm{T}}A - \sum_{i=1}^{m} C_i^*) \xrightarrow{p_F} K, \quad \frac{1}{m}A^{\mathrm{T}}Y \xrightarrow{p_F} K\gamma \text{ and } \hat{\gamma}_F \xrightarrow{p_F} K^{-1}K\gamma = \gamma.$$

Using the consistency of $\hat{\gamma}_F$, it follows that

$$\hat{\sigma}^2_{u,F} - \left\{ \frac{1}{m} \sum_{i=1}^{m} (Y_i - X_i^{\mathrm{T}}\beta - z_i^{\mathrm{T}}\delta)^2 - \bar{D} - \beta^{\mathrm{T}}\bar{C}\beta \right\} \xrightarrow{p_F} 0.$$

Finally, using this and the weak law of large numbers we get

$$\frac{1}{m} \sum_{i=1}^{m} (Y_i - X_i^{\mathrm{T}}\beta - z_i^{\mathrm{T}}\delta)^2 - \bar{D} - \beta^{\mathrm{T}}\bar{C}\beta \xrightarrow{p_F} \sigma_u^2$$

and, consequently,

$$\hat{\sigma}^2_{u,F} \xrightarrow{p_F} \sigma_u^2.$$

This establishes consistency of the estimators $\hat{\gamma}_F$ and $\hat{\sigma}^2_{u,F}$ derived from the unbiased estimating equation approach. We know from Ybarra and Lohr (2008) that, under the FME model with known parameters γ and σ_u^2, the best linear unbiased predictor of θ_i is

$$\tilde{\theta}_{i,F} = Y_i - \frac{D_i}{D_i + \sigma_u^2 + \beta^{\mathrm{T}} C_i \beta} (Y_i - X_i^{\mathrm{T}}\beta - z_i^{\mathrm{T}}\delta).$$

Replacing the unknown model parameters in $\tilde{\theta}_{i,F}$ with the parameter estimators derived above, the EBLUP of θ_i is obtained as

$$\hat{\theta}^{EB}_{i,F} = Y_i - \frac{D_i}{D_i + \hat{\sigma}^2_{u,F} + \hat{\beta}_F^{\mathrm{T}} C_i \hat{\beta}_F} (Y_i - X_i^{\mathrm{T}}\hat{\beta}_F - z_i^{\mathrm{T}}\hat{\delta}_F). \tag{3.23}$$

Since $\hat{\gamma}_F$ is consistent of γ, we have

$$\begin{aligned}
\hat{\theta}^{EB}_{i,F} - \theta_i &= Y_i - \theta_i - \frac{D_i}{D_i + \hat{\sigma}^2_{u,F} + \hat{\beta}_F^{\mathrm{T}} C_i \hat{\beta}_F} (Y_i - X_i^{\mathrm{T}}\hat{\beta}_F - z_i^{\mathrm{T}}\hat{\delta}_F) \\
&= e_i - \frac{D_i}{D_i + \sigma_u^2 + \beta^{\mathrm{T}} C_i \beta} (Y_i - X_i^{\mathrm{T}}\beta - z_i^{\mathrm{T}}\delta) + O_p(m^{-1/2}) \\
&= e_i - \frac{D_i (e_i + u_i - \eta_i^{\mathrm{T}}\beta)}{D_i + \sigma_u^2 + \beta^{\mathrm{T}} C_i \beta} + O_p(m^{-1/2}).
\end{aligned}$$

This leads to $\mathsf{E}_F(\hat{\theta}^{EB}_{i,F} - \theta_i) = O(m^{-1})$, implying asymptotic unbiasedness of the EBLUP predictor under the FME model. Also,

$$\begin{aligned}
\mathsf{E}_F(\hat{\theta}^{EB}_{i,F} - \theta_i)^2 &= \mathsf{E}_F \left(e_i - \frac{D_i (e_i + u_i - \eta_i^{\mathrm{T}}\beta)}{D_i + \sigma_u^2 + \beta^{\mathrm{T}} C_i \beta} \right)^2 + O(m^{-1}) \\
&= \mathsf{E}_F \left(\frac{e_i (\sigma_u^2 + \beta^{\mathrm{T}} C_i \beta)}{D_i + \sigma_u^2 + \beta^{\mathrm{T}} C_i \beta} - \frac{D_i (u_i - \eta_i^{\mathrm{T}}\beta)}{D_i + \sigma_u^2 + \beta^{\mathrm{T}} C_i \beta} \right)^2 + O(m^{-1}) \\
&= \frac{D_i (\sigma_u^2 + \beta^{\mathrm{T}} C_i \beta)}{(D_i + \sigma_u^2 + \beta^{\mathrm{T}} C_i \beta)} + O(m^{-1}).
\end{aligned}$$

3.6.1.2 Naive FH estimators and predictors

If we ignore the measurement error in X_i, then we fit the following naive FH model

$$Y_i = \theta_i + e_i,\ \theta_i = X_i^{\mathrm{T}}\beta_N + z_i^{\mathrm{T}}\delta_N + u_{i,N},\quad i = 1, \ldots, m, \tag{3.24}$$

where, as usual, we assume that e_i's and $u_{i,N}$'s are independently distributed with means zero and variances D_i and $\sigma^2_{u,N}$, respectively. To avoid confusion, we use the subscript N for the parameters in the naive FH model. Let $\gamma_N = (\beta_N^{\mathrm{T}}, \delta_N^{\mathrm{T}})^{\mathrm{T}}$. We estimate γ_N and $\sigma^2_{u,N}$ by

$$\hat{\gamma}_N = (A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}Y, \tag{3.25}$$

$$\hat{\sigma}^2_{u,N} = \frac{1}{m}\sum_{i=1}^{m}(Y_i - X_i^{\mathrm{T}}\hat{\beta}_N - z_i^{\mathrm{T}}\hat{\delta}_N)^2 - \bar{D}. \tag{3.26}$$

Using these estimators, the naive FH EBLUP of θ_i can be obtained by

$$\hat{\theta}^{EB}_{i,N} = Y_i - \frac{D_i}{D_i + \hat{\sigma}^2_{u,N}}(Y_i - X_i^{\mathrm{T}}\hat{\beta}_N - z_i^{\mathrm{T}}\hat{\delta}_N). \tag{3.27}$$

With respect to the true model in (3.15), the naive FH model given by (3.24) is misspecified. It follows that

$$\frac{1}{m}A^{\mathrm{T}}A \xrightarrow{p_F} K + \mathrm{diag}(\bar{C}, 0) \quad \text{and} \quad \frac{1}{m}A^{\mathrm{T}}Y \xrightarrow{p_F} K\gamma.$$

Let $K_* = K + \mathrm{diag}(\bar{C}, 0)$ and partition the matrices K and K_* as

$$K = \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix},\ K_* = \begin{bmatrix} K_{*11} & K_{*12} \\ K_{*21} & K_{*22} \end{bmatrix}.$$

Note that $K_{ij} = K_{*ij}$, for $(i,j) \neq (1,1)$ and $K_{*11} = K_{11} + \bar{C}$. Denoting $K_{11.2}$ by $K_{11} - K_{12}K_{22}^{-1}K_{21}$, we have

$$K_*^{-1} = \begin{bmatrix} (K_{11.2} + \bar{C})^{-1} & -(K_{11.2} + \bar{C})^{-1}K_{12}K_{22}^{-1} \\ -K_{22}^{-1}K_{21}(K_{11.2} + \bar{C})^{-1} & K_{22}^{-1} + K_{22}^{-1}K_{21}(K_{11.2} + \bar{C})^{-1}K_{12}K_{22}^{-1} \end{bmatrix},$$

and, after writing $\tilde{A} = (K_{11.2} + \bar{C})^{-1}K_{11.2}$, we find that

$$K_*^{-1}K = \begin{bmatrix} \tilde{A} & 0 \\ K_{22}^{-1}K_{21}(I - \tilde{A}) & I \end{bmatrix}.$$

Putting all these together, it follows that

$$\hat{\gamma}_N \xrightarrow{p_F} K_*^{-1} K\gamma = \begin{bmatrix} \tilde{A}\beta \\ \delta + K_{22}^{-1} K_{21}(I - \tilde{A})\beta \end{bmatrix} \equiv \bar{\gamma}. \tag{3.28}$$

Let $\bar{\gamma} = (\bar{\beta}^{\mathrm{T}}, \bar{\delta}^{\mathrm{T}})^{\mathrm{T}}$. Then $\hat{\sigma}^2_{u,N}$ can be written as

$$\hat{\sigma}^2_{u,N} = \frac{1}{m}\sum_{i=1}^{m}(e_i + u_i - \eta_i^{\mathrm{T}}\bar{\beta})^2 + \frac{1}{m}\sum_{i=1}^{m}\left\{x_i^{\mathrm{T}}(\beta - \bar{\beta}) + z_i^{\mathrm{T}}(\delta - \bar{\delta})\right\}^2 - \bar{D} + O_p(m^{-1/2})$$

and it converges to

$$\hat{\sigma}^2_{u,N} \xrightarrow{p_F} \sigma_u^2 + \bar{\beta}^{\mathrm{T}}\bar{C}\bar{\beta} + (\gamma - \bar{\gamma})^{\mathrm{T}} K(\gamma - \bar{\gamma}). \tag{3.29}$$

We note that

$$\begin{aligned} K(\gamma - \bar{\gamma}) &= K(I - K_*^{-1}K)\gamma \\ &= \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix} \begin{bmatrix} (I - \tilde{A})\beta \\ -K_{22}^{-1}K_{21}(I - \tilde{A})\beta \end{bmatrix} \\ &= \begin{bmatrix} K_{11.2}(I - \tilde{A})\beta \\ 0 \end{bmatrix}. \end{aligned}$$

Thus, the right hand size of (3.29) can be simplified as

$$\hat{\sigma}^2_{u,N} \xrightarrow{p_F} \sigma_u^2 + \beta^{\mathrm{T}}\bar{C}\tilde{A}\beta \equiv \sigma_u^2 + \tau. \tag{3.30}$$

Note that $\tau = \beta^{\mathrm{T}}(\bar{C}^{-1} + K_{22.1}^{-1})^{-1}\beta > 0$ unless $\beta = 0$ since $\bar{C}$ is assumed to be positive definite. This indicates that $\hat{\sigma}^2_{u,N}$ overestimates σ_u^2. Using (3.28) and (3.30) in (3.27), we obtain an asymptotic expression for the bias of the naive EBLUP $\hat{\theta}^{EB}_{i,N}$ as

$$\mathsf{E}_F(\hat{\theta}^{EB}_{i,N} - \theta_i) = -\frac{D_i(x_i - K_{12}K_{22}^{-1}z_i)^{\mathrm{T}}(I - \tilde{A})\beta}{D_i + \sigma_u^2 + \tau} + O(m^{-1}). \tag{3.31}$$

After further simplification, we also get that

$$\mathsf{Var}_F(\hat{\theta}^{EB}_{i,N} - \theta_i) = \frac{D_i(\sigma_u^2 + \tau)}{D_i + \sigma_u^2 + \tau} + \frac{D_i^2\{\beta^{\mathrm{T}}(\tilde{A}^{\mathrm{T}}C_i - \bar{C})\tilde{A}\beta\}}{(D_i + \sigma_u^2 + \tau)^2} + O(m^{-1}). \tag{3.32}$$

Combining (3.31) and (3.32) gives the limiting expression for $MSE_F(\hat{\theta}^{EB}_{i,N})$

3.6.1.3 SME ESTIMATORS AND PREDICTORS

We now investigate the impact of model misspecification when, instead of the true FME model in (3.15), we fit the following structural measurement error (SME) model:

$$\begin{aligned} Y_i &= \theta_i + e_i, \qquad \theta_i = x_i^{\mathrm{T}}\beta + z_i^{\mathrm{T}}\delta + u_i, \\ X_i &= x_i + \eta_i, \qquad x_i = \mu + w_i, \quad i = 1, \ldots, m, \end{aligned} \tag{3.33}$$

where e_i's, u_i's, η_i's and w_i's are independently distributed with zero means, and variances D_i, σ_u^2, C_i and Φ, respectively. We assume that Φ is a positive definite matrix. In our SME formulation in (3.33) we assume that $z_{i1} = 1$ for an intercept, and for simplicity of calculations that $z_{i2}, \ldots, z_{iq}$ are centered about their respective means. Under the model in (3.33), we can create the following unbiased estimating equations:

$$\sum_{i=1}^{m}(Y_i - X_i^{\mathrm{T}}\beta - z_i^{\mathrm{T}}\delta)z_i = 0_{q,1}, \tag{3.34}$$

$$\sum_{i=1}^{m}(X_i - \mu) = 0_{p,1}, \tag{3.35}$$

$$\sum_{i=1}^{m}\left\{\begin{bmatrix} Y_i - z_i^{\mathrm{T}}\delta \\ X_i - \mu \end{bmatrix}\begin{bmatrix} Y_i - z_i^{\mathrm{T}}\delta \\ X_i - \mu \end{bmatrix}^{\mathrm{T}} - \begin{bmatrix} D_i + \sigma_u^2 + \beta^{\mathrm{T}}(\mu\mu^{\mathrm{T}} + \Phi)\beta & \beta^{\mathrm{T}}\Phi \\ \Phi\beta & \Phi + C_i \end{bmatrix}\right\} = 0_{p+1,p+1}. \tag{3.36}$$

Let $(\hat{\beta}_S^{\mathrm{T}}, \hat{\delta}_S^{\mathrm{T}})^{\mathrm{T}}$, $\hat{\mu}_S$, $\hat{\Phi}_S$ and $\hat{\sigma}_{u,S}^2$ denote a solution to (3.34)–(3.36). The subscript S is used to indicate the SME model is being fitted. Note that $\hat{\mu}_S = \bar{X}$, and

$$\hat{\Phi}_S = \frac{1}{m}\sum_{i=1}^{m}(X_i - \bar{X})(X_i - \bar{X})^{\mathrm{T}} - \bar{C}, \tag{3.37}$$

$$\hat{\Phi}_S\hat{\beta}_S = \frac{1}{m}\sum_{i=1}^{m}(Y_i - z_i^{\mathrm{T}}\hat{\delta}_S)(X_i - \bar{X}), \tag{3.38}$$

$$\hat{\sigma}_{u,S}^2 = \frac{1}{m}\sum_{i=1}^{m}(Y_i - z_i^{\mathrm{T}}\hat{\delta}_S)^2 - \bar{D} - \hat{\beta}_S^{\mathrm{T}}(\bar{X}\bar{X}^{\mathrm{T}} + \hat{\Phi}_S)\hat{\beta}_S, \tag{3.39}$$

$$\sum_{i=1}^{m}(Y_i - X_i^{\mathrm{T}}\hat{\beta}_S - z_i^{\mathrm{T}}\hat{\delta}_S)z_i = 0_{q,1}. \tag{3.40}$$

We assume $\hat{\Phi}_S$ is positive definite. We show below that the estimators $\hat{\delta}_S$, $\hat{\beta}_S$ and $\hat{\sigma}_{u,S}^2$ obtained by solving (3.37)–(3.40) also solve the estimating equations (3.17) and (3.18), which

correspond to the FME model fitting. Since $z_{i1} = 1$, from (3.34) it follows that $\bar{Y} - \bar{X}^{\mathrm{T}}\hat{\beta}_S - \bar{z}^{\mathrm{T}}\hat{\delta}_S = 0$. Using this, it follows from (3.37) and (3.38), after simplification, that

$$\sum_{i=1}^{m} (X_i X_i^{\mathrm{T}} - C_i)\,\hat{\beta}_S + \left(\sum_{i=1}^{m} X_i z_i^{\mathrm{T}}\right)\hat{\delta}_S = \sum_{i=1}^{m} Y_i X_i. \tag{3.41}$$

From (3.40), we get

$$\left(\sum_{i=1}^{m} z_i X_i^{\mathrm{T}}\right)\hat{\beta}_S + \left(\sum_{i=1}^{m} z_i z_i^{\mathrm{T}}\right)\hat{\delta}_S = \sum_{i=1}^{m} Y_i z_i. \tag{3.42}$$

Equations (3.41) and (3.42) can also be written as

$$\left(A^{\mathrm{T}}A - \sum_{i=1}^{m} C_i^*\right)\begin{bmatrix}\hat{\beta}_S \\ \hat{\delta}_S\end{bmatrix} = A^{\mathrm{T}}Y.$$

From this equation, and the assumption that $A^{\mathrm{T}}A - \sum_{i=1}^{m} C_i^*$ is nonsingular, it follows that $(\hat{\beta}_S^{\mathrm{T}}, \hat{\delta}_S^{\mathrm{T}})^{\mathrm{T}}$ also satisfies (3.17), so $\hat{\gamma}_S = \hat{\gamma}_F$. Using $\hat{\gamma}_S = \hat{\gamma}_F$ in (3.39) and equation (3.37), it can be shown after a lot of simplification that

$$\begin{aligned}
\hat{\sigma}^2_{u,S} &= \frac{1}{m}\sum_{i=1}^{m}\left(Y_i - z_i^{\mathrm{T}}\hat{\delta}_F\right)^2 - \bar{D} - \hat{\beta}_F^{\mathrm{T}}\left\{\frac{1}{m}\sum_{i=1}^{m}(X_i X_i^{\mathrm{T}} - C_i)\right\}\hat{\beta}_F \\
&= \frac{1}{m}\sum_{i=1}^{m}\left(Y_i - X_i^{\mathrm{T}}\hat{\beta}_F - z_i^{\mathrm{T}}\hat{\delta}_F\right)^2 + \frac{1}{m}\hat{\beta}_F^{\mathrm{T}}\left(\sum_{i=1}^{m} X_i X_i^{\mathrm{T}}\right)\hat{\beta}_F \\
&\quad + \frac{2}{m}\hat{\beta}_F^{\mathrm{T}}\left\{\sum_{i=1}^{m} X_i\left(Y_i - X_i^{\mathrm{T}}\hat{\beta}_F - z_i^{\mathrm{T}}\hat{\delta}_F\right)\right\} - \bar{D} - \hat{\beta}_F^{\mathrm{T}}\left(\frac{1}{m}\sum_{i=1}^{m} X_i X_i^{\mathrm{T}} - \bar{C}\right)\hat{\beta}_F.
\end{aligned} \tag{3.43}$$

Using $\sum_{i=1}^{m} X_i(Y_i - X_i^{\mathrm{T}}\hat{\beta}_S - z_i^{\mathrm{T}}\hat{\delta}_S) = -\sum_{i=1}^{m} C_i\hat{\beta}_S$ (from (3.41)), we get from (3.20) and (3.43) that

$$\hat{\sigma}^2_{u,S} = \frac{1}{m}\sum_{i=1}^{m}(Y_i - X_i^{\mathrm{T}}\hat{\beta}_F - z_i^{\mathrm{T}}\hat{\delta}_F)^2 - \bar{D} - \hat{\beta}_F^{\mathrm{T}}\bar{C}\hat{\beta}_F = \hat{\sigma}^2_{u,F}.$$

These calculations show that fitting the SME model using the estimating equations given by (3.34)–(3.36) will lead to estimators of δ, β and σ_u^2 that are identical to the FME estimators given in (3.19) and (3.20).

From our derivation that $\hat{\gamma}_S = \hat{\gamma}_F$ and $\hat{\sigma}^2_{u,S} = \hat{\sigma}^2_{u,F}$, and from the convergence under the FME model of the estimators $\hat{\gamma}_F$ and $\hat{\sigma}^2_{u,F}$ to γ and σ_u^2, we immediately conclude that the

parameters estimated by fitting SME model converge to their true values under the FME model. If we write $(Y_i, X_i^{\mathrm{T}}) = Y_i^{*\mathrm{T}}$, $(\theta_i, x_i^{\mathrm{T}}) = \theta_i^{*\mathrm{T}}$, $D_i^* = \mathrm{diag}(D_i, C_i)$ and

$$\Sigma = \begin{bmatrix} \sigma_u^2 + \beta^{\mathrm{T}}\Phi\beta & \Phi\beta \\ \beta\Phi & \Phi \end{bmatrix},$$

from the multivariate FH model, the EBLUP of θ_i^* is given by

$$\hat{\theta}_{i,MFH}^{*EB} = Y_i^* - D_i^* \left(D_i^* + \hat{\Sigma}\right)^{-1} \begin{bmatrix} Y_i - \hat{\mu}^{\mathrm{T}}\hat{\beta}_S - z_i^{\mathrm{T}}\hat{\delta}_S \\ X_i - \hat{\mu} \end{bmatrix}. \tag{3.44}$$

If $\lambda = (1, 0_{p,1}^{\mathrm{T}})^{\mathrm{T}}$ denotes a unit vector with the first component 1 and all others zero, then the EBLUP of θ_i, denoted by $\hat{\theta}_{i,S}^{EB}$, is the first component of $\hat{\theta}_{i,MFH}^{*EB}$, which is, $\lambda^{\mathrm{T}}\hat{\theta}_{i,MFH}^{*EB}$. Using the block diagonal structure of D_i^* and using the inverse formula for partitioned matrices, it follows after substantial simplifications that

$$\lambda^{\mathrm{T}} D_i^* \left(D_i^* + \hat{\Sigma}\right)^{-1} \begin{bmatrix} Y_i - \bar{X}^{\mathrm{T}}\hat{\beta}_S - z_i^{\mathrm{T}}\hat{\delta}_S \\ X_i - \bar{X} \end{bmatrix} = \frac{D_i \left\{ Y_i - X_i^{\mathrm{T}}\hat{\beta}_S - z_i^{\mathrm{T}}\hat{\delta}_S + \hat{\beta}_S^{\mathrm{T}} C_i (\hat{\Phi}_S + C_i)^{-1}(X_i - \bar{X}) \right\}}{D_i + \hat{\sigma}_{u,S}^2 + \hat{\beta}_S^{\mathrm{T}} C_i (\hat{\Phi}_S + C_i)^{-1} \hat{\Phi}_S \hat{\beta}_S},$$

and finally,

$$\hat{\theta}_{i,S}^{EB} = Y_i - \frac{D_i \left\{ Y_i - X_i^{\mathrm{T}}\hat{\beta}_S - z_i^{\mathrm{T}}\hat{\delta}_S + \hat{\beta}_S^{\mathrm{T}} C_i (\hat{\Phi}_S + C_i)^{-1}(X_i - \bar{X}) \right\}}{D_i + \hat{\sigma}_{u,S}^2 + \hat{\beta}_S^{\mathrm{T}} C_i (\hat{\Phi}_S + C_i)^{-1} \hat{\Phi}_S \hat{\beta}_S}. \tag{3.45}$$

We note that

$$\hat{\Phi}_S \xrightarrow{p_F} \frac{1}{m} \sum_{i=1}^{m} (x_i - \bar{x})(x_i - \bar{x})^{\mathrm{T}} \equiv S_x. \tag{3.46}$$

By the convergence properties of $\hat{\gamma}_F$ and $\hat{\sigma}_{u,F}^2$, and the result in (3.46), it follows from (3.45) that the asymptotic bias of the predictor $\hat{\theta}_{i,S}^{EB}$ under the FME model is given by

$$\mathsf{E}_F(\hat{\theta}_{i,S}^{EB} - \theta_i) = -\frac{D_i \beta^{\mathrm{T}} C_i (S_x + C_i)^{-1}(x_i - \bar{x})}{D_i + \sigma_u^2 + \beta^{\mathrm{T}} C_i (S_x + C_i)^{-1} S_x \beta} + O(m^{-1}). \tag{3.47}$$

After considerable simplifications, it can be shown that

$$\begin{aligned} \mathsf{Var}_F(\hat{\theta}_{i,S}^{EB} - \theta_i) &= \frac{D_i(\sigma_u^2 + \beta^{\mathrm{T}}(C_i^{-1} + S_x^{-1})^{-1}\beta)}{D_i + \sigma_u^2 + \beta^{\mathrm{T}}(C_i^{-1} + S_x^{-1})^{-1}\beta} \\ &- \left(\frac{D_i}{D_i + \sigma_u^2 + \beta^{\mathrm{T}}(C_i^{-1} + S_x^{-1})^{-1}\beta}\right)^2 \beta^{\mathrm{T}} \left\{ (C_i^{-1} + S_x^{-1}) S_x (S_x^{-1} + C_i^{-1}) \right\}^{-1} \beta + O(m^{-1}). \end{aligned} \tag{3.48}$$

Combining (3.47) and (3.48), we obtain an asymptotic expression of $MSE_F(\hat{\theta}_{i,S}^{EB})$.

3.6.2 Asymptotic evaluations of estimators and predictors under SME model

3.6.2.1 FME estimators and predictors

Let $\xrightarrow{p_S}$ denote convergence in probability under the SME model, and let $\mathsf{E}_S(\cdot)$ be the corresponding expectation. We obtained FME estimators of γ and σ_u^2 in Section 3.6.1.1. Recall that $z_{i1} = 1$ and $z_{i2}, \ldots, z_{iq}$ are centered so that $\bar{z} = (1, 0, \ldots, 0)^{\mathrm{T}}$. We denote $m^{-1}\sum_{i=1}^{m} z_i z_i^{\mathrm{T}}$ by $\mathrm{diag}(1, G_{22})$, and assume that this matrix converges to a positive definite matrix $\mathrm{diag}(1, \Gamma_{22})$, where $\mathrm{diag}(A, B)$ denotes the block diagonal matrix with square matrices A and B. Then, under SME model,

$$\frac{1}{m}A^{\mathrm{T}}A \xrightarrow{p_S} \begin{bmatrix} \mu\mu^{\mathrm{T}} + \Phi + \bar{C} & \mu\bar{z}^{\mathrm{T}} \\ \bar{z}\mu^{\mathrm{T}} & \mathrm{diag}(1, \Gamma_{22}) \end{bmatrix}.$$

and

$$\frac{1}{m}A^{\mathrm{T}}Y \xrightarrow{p_S} \begin{bmatrix} \mu\mu^{\mathrm{T}} + \Phi & \mu\bar{z}^{\mathrm{T}} \\ \bar{z}\mu^{\mathrm{T}} & \mathrm{diag}(1, \Gamma_{22}) \end{bmatrix} \begin{bmatrix} \beta \\ \delta \end{bmatrix}.$$

These give us that $\hat{\gamma}_F \xrightarrow{p_S} \gamma$ and $\hat{\sigma}_{u,F}^2 \xrightarrow{p_S} \sigma_u^2$ as similarly done by Section 3.6.1.1. Hence, FME estimators are still consistent under SME model. Using these results, it follows from (3.23) that

$$\mathsf{E}_S(\hat{\theta}_{i,F}^{EB} - \theta_i) = O(m^{-1}). \tag{3.49}$$

Moreover,

$$\mathsf{Var}_S(\hat{\theta}_{i,F}^{EB} - \theta_i) = \frac{D_i(\sigma_u^2 + \beta^{\mathrm{T}}C_i\beta)}{D_i + \sigma_u^2 + \beta^{\mathrm{T}}C_i\beta} + O(m^{-1}). \tag{3.50}$$

3.6.2.2 SME estimators and predictors

We have shown in Section 3.6.1.3 that estimators of γ and σ_u^2 under SME model are identical to those under FME model. As similarly done by (3.49), we have

$$\mathsf{E}_S(\hat{\theta}_{i,S}^{EB} - \theta_i) = O(m^{-1}) \tag{3.51}$$

and, by similar simplification that was used to obtain equation (3.45), we can derive that

$$\mathsf{Var}_S(\hat{\theta}_{i,S}^{EB} - \theta_i) = \frac{D_i(\sigma_u^2 + \beta^{\mathrm{T}} C_i(\Phi + C_i)^{-1}\Phi\beta)}{D_i + \sigma_u^2 + \beta^{\mathrm{T}} C_i(\Phi + C_i)^{-1}\Phi\beta} + O(m^{-1}). \tag{3.52}$$

3.6.2.3 Naive FH estimators and predictors

We now turn to estimators and predictors derived under the naive FH model. We note that

$$\mathsf{E}_S(\frac{1}{m}A^{\mathrm{T}}Y) = \begin{bmatrix} \mu\mu^{\mathrm{T}} + \Phi & \mu\bar{z}^{\mathrm{T}} \\ \bar{z}\mu^{\mathrm{T}} & \frac{1}{m}\sum_{i=1}^m z_i z_i^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \beta \\ \delta \end{bmatrix} \tag{3.53}$$

and

$$\mathsf{E}_S(\frac{1}{m}A^{\mathrm{T}}A) = \begin{bmatrix} \mu\mu^{\mathrm{T}} + \Phi + \bar{C} & \mu\bar{z}^{\mathrm{T}} \\ \bar{z}\mu^{\mathrm{T}} & \frac{1}{m}\sum_{i=1}^m z_i z_i^{\mathrm{T}} \end{bmatrix}. \tag{3.54}$$

Let R denote the matrix on the right hand side of (3.54) and let

$$R^{-1} = \begin{bmatrix} R^{11} & R^{12} \\ R^{21} & R^{22} \end{bmatrix}.$$

Then $(R^{11})^{-1} = (\mu\mu^{\mathrm{T}} + \Phi + \bar{C}) - \mu\bar{z}^{\mathrm{T}}(\sum_{i=1}^m z_i z_i^{\mathrm{T}}/m)\bar{z}\mu^{\mathrm{T}}$. Since, $\bar{z}^{\mathrm{T}}(\sum_{i=1}^m z_i z_i^{\mathrm{T}}/m)\bar{z} = 1$, we get

$$R^{11} = (\Phi + \bar{C})^{-1}, \tag{3.55}$$

and

$$R^{21} = -R_{22}^{-1}R_{21}R^{11} = -\operatorname{diag}(1, G_{22}^{-1})\bar{z}\mu^{\mathrm{T}}R^{11} = -\bar{z}\mu^{\mathrm{T}}(\Phi + \bar{C})^{-1}.$$

Let S be the square matrix on the right hand size of (3.53). Then $S = R - \operatorname{diag}(\bar{C}, 0_{qq})$. Consequently,

$$R^{-1}S = \begin{bmatrix} I - R^{11}\bar{C} & 0 \\ -R^{21}\bar{C} & I \end{bmatrix} \quad \text{and} \quad R^{-1}S\gamma = \begin{bmatrix} (I - R^{11}\bar{C})\beta \\ \delta - R^{21}\bar{C}\beta \end{bmatrix}.$$

Denote $(\Phi + \bar{C})^{-1}\Phi$ by $\check{A}$. After simplification, we get

$$R^{-1}S\gamma = \begin{bmatrix} \check{A}\beta \\ \delta_1 + \mu^{\mathrm{T}}(I - \check{A})\beta \\ \delta_2 \\ \vdots \\ \delta_q \end{bmatrix} \equiv \bar{\gamma}. \tag{3.56}$$

By (3.56), it follows that $\hat{\gamma}_N \xrightarrow{p_S} \bar{\gamma}$. This shows that under the naive model while the parameters $\delta_2, \ldots, \delta_q$ are consistently estimated, estimates of the parameter vector β are shrunk to the null vector. We now find the probability limit of $\hat{\sigma}^2_{u,N}$. Recall that

$$\hat{\sigma}^2_{u,N} = m^{-1}\sum_{i=1}^{m}(Y_i - X_i^{\mathrm{T}}\hat{\beta}_N - z_i^{\mathrm{T}}\hat{\delta}_N)^2 - \bar{D}.$$

Since

$$\begin{aligned} &\mathsf{E}_S(Y_i - X_i^{\mathrm{T}}\hat{\beta}_N - z_i^{\mathrm{T}}\hat{\delta}_N)^2 = \mathsf{E}_S(Y_i - X_i^{\mathrm{T}}\bar{\beta} - z_i^{\mathrm{T}}\bar{\delta})^2 + O(m^{-1}) \\ &= \sigma_u^2 + D_i + (\beta - \bar{\beta})^{\mathrm{T}}\Phi(\beta - \bar{\beta}) + \bar{\beta}^{\mathrm{T}}C_i\bar{\beta} + (\gamma - \bar{\gamma})^{\mathrm{T}} \begin{bmatrix} \mu \\ z_i \end{bmatrix} \begin{bmatrix} \mu \\ z_i \end{bmatrix}^{\mathrm{T}} (\gamma - \bar{\gamma}) + O(m^{-1}), \end{aligned}$$

as similarly done by in Section 3.6.1.2,

$$\hat{\sigma}^2_{u,N} \xrightarrow{p_S} \sigma_u^2 + \beta^{\mathrm{T}}\bar{C}\check{A}\beta. \tag{3.57}$$

As naive EBLUP of θ_i is given by

$$\hat{\theta}^{EB}_{i,N} = Y_i - \frac{D_i}{D_i + \hat{\sigma}^2_{u,N}}(Y_i - X_i^{\mathrm{T}}\hat{\beta}_N - z_i^{\mathrm{T}}\hat{\delta}_N),$$

we have

$$\hat{\theta}^{EB}_{i,N} - \theta_i = e_i + \frac{D_i}{D_i + \sigma_u^2 + \beta^{\mathrm{T}}\bar{C}\check{A}\beta}(Y_i - X_i^{\mathrm{T}}\bar{\beta} - z_i^{\mathrm{T}}\bar{\delta}) + O_p(m^{-1/2}),$$

and thus $\mathsf{E}_S(\hat{\theta}^{EB}_{i,N} - \theta_i) = O(m^{-1})$. Furthermore, after some simplification, we obtain

$$\mathsf{Var}_S(\hat{\theta}^{EB}_{i,N} - \theta_i) = \frac{D_i(\sigma_u^2 + \beta^{\mathrm{T}}\bar{C}\check{A}\beta)}{D_i + \sigma_u^2 + \beta^{\mathrm{T}}\bar{C}\check{A}\beta} + \left(\frac{D_i}{D_i + \sigma_u^2 + \beta^{\mathrm{T}}\bar{C}\check{A}\beta}\right)^2 \bar{\beta}^{\mathrm{T}}(C_i - \bar{C})\bar{\beta} + O(m^{-1}).$$

3.6.3 Re-expression of MSEs

In this section, we provide simplified MSE expressions in terms of r_D, r_D and ρ given in (3.14).With these quantities, we first have

$$\begin{aligned} F_i^* &= (\sigma_u^2 + D_i)(\sigma_x^2 + C_i) + \beta^2\sigma_x^2 \\ &= \sigma_u^2\sigma_x^2\{(1+r_D)(1+r_C) + \beta^2\sigma_x^2 r_C/\sigma_u^2\}. \end{aligned}$$

Further, noting that $\sigma_v^2 = \sigma_x^2$, we get

$$r_\rho = \frac{\rho^2}{1-\rho^2} = \frac{\beta^2\sigma_x^2}{\sigma_u^2},$$

where

$$\rho^2 = \text{corr}(\theta_i, x_i)^2 = \frac{\beta^2\sigma_x^2}{\sigma_u^2 + \beta^2\sigma_x^2}.$$

This implies that $F_i^* = \sigma_u^2\sigma_x^2\{(1+r_D)(1+r_C) + r_\rho r_C\}$. Thus, we have the expression

$$\begin{aligned} MSE_S &= \sigma_u^2\left\{\frac{D_i}{\sigma_u^2} - \frac{\sigma_u^{-2}D_i^2(\sigma_x^2+C_i)}{\sigma_u^2\sigma_x^2\{(1+r_D)(1+r_C)+r_\rho r_C\}}\right\} \\ &= \sigma_u^2\left\{r_D - \frac{r_D^2(1+r_C)}{(1+r_D)(1+r_C)+r_\rho r_C}\right\}. \end{aligned}$$

3.6.4 Equality of SME and Naive FH predictors when $C_i = \bar{C}$

In this section, we show the equality of SME and naive FH predictors when $C_i = \bar{C}$. Without loss of generality, we assume that the first element of z_i is 1 for an intercept and $C_i = C$, $i = 1, \ldots, m$. Let $\delta = (\delta_1, \delta_2^{\mathrm{T}})^{\mathrm{T}}$, then, for $i = 1 \ldots, m$, SME model θ_i is given by

$$\theta_i = x_i^{\mathrm{T}}\beta + \delta_1 + z_{2,i}^{\mathrm{T}}\delta_2 + u_i \quad \text{and} \quad X_i = x_i + \eta_i,$$

whereas naive FH model θ_i is

$$\theta_i = X_i^{\mathrm{T}}\beta_N + \delta_{1,N} + z_{2,i}^{\mathrm{T}}\delta_{2,N} + u_{i,N},$$

where X_i's are treated as fixed and $u_{i,N} \sim N(0, \sigma^2_{u,N})$. Under SME model, we can write $x_i = \mathsf{E}_S(x_i|X_i) + \{x_i - \mathsf{E}_S(x_i|X_i)\}$, where

$$\mathsf{E}_S(x_i|X_i) = \mu + \Phi(\Phi + C)^{-1}(X_i - \mu) = \mu + \check{A}(X_i - \mu)$$
$$\mathsf{Var}(x_i|X_i) = (\Phi^{-1} + C^{-1})^{-1}.$$

Thus, SME model θ_i is

$$\begin{aligned}\theta_i &= x_i^{\mathrm{T}}\beta + \delta_1 + z_{2,i}^{\mathrm{T}}\delta_2 + u_i \\ &= X_i^{\mathrm{T}}\check{A}\beta + \delta_1 + \mu^{\mathrm{T}}(\Phi + C)^{-1}C\beta + z_{2,i}^{\mathrm{T}}\delta_2 + u_i + \beta^{\mathrm{T}}\{x_i - \mathsf{E}_S(x_i|X_i)\}.\end{aligned}$$

By letting,

$$\begin{aligned}u_{i,N} &= u_i + \beta^{\mathrm{T}}\{x_i - \mathsf{E}_S(x_i|X_i)\}, \\ \beta_N &= \check{A}\beta, \\ \delta_{1,N} &= \delta_1 + \mu^{\mathrm{T}}(\Phi + C)^{-1}C\beta,\end{aligned}$$

we get $\sigma^2_{u,N} = \sigma^2_u + \beta^{\mathrm{T}}(\Phi^{-1} + C^{-1})^{-1}\beta$. The naive model parameters β_N, $\delta_{1,N}$, $\delta_{2,N}$ and $\sigma^2_{u,N}$ are estimated by estimating equations and these estimators are shown to be consistent. Furthermore, $\hat{\beta}_S$, $\hat{\delta}_{1,S}$, $\hat{\delta}_{2,S}$, $\hat{\mu}_S$ and $\hat{\Phi}$ are also consistent estimators. Recall that $\check{A} = (\Phi + \bar{C})^{-1}\Phi$ which is estimated by $\hat{A} = (\hat{\Phi} + C)^{-1}\hat{\Phi}$. By the consistency of two sets of estimators, we have the following results:

$$\hat{\beta}_N - \hat{A}^{\mathrm{T}}\hat{\beta}_S \xrightarrow{p_S} 0, \tag{3.58}$$
$$\hat{\sigma}^2_{u,N} - \hat{\sigma}^2_{u,S} - \hat{\beta}_S(\hat{\Phi}^{-1} + C^{-1})^{-1}\hat{\beta}_S \xrightarrow{p_S} 0, \tag{3.59}$$
$$\hat{\delta}_{2,N} - \hat{\delta}_{2,S} \xrightarrow{p_S} 0, \tag{3.60}$$
$$\hat{\delta}_{1,N} - \hat{\delta}_{1,S} - \hat{\mu}_S^{\mathrm{T}}(\hat{\Phi} + C)^{-1}C\hat{\beta}_S \xrightarrow{p_S} 0. \tag{3.61}$$

Using (3.58)–(3.61), we obtain

$$X_i^{\mathrm{T}}\hat{\beta}_N + z_i^{\mathrm{T}}\hat{\delta}_N - X_i^{\mathrm{T}}\hat{\beta}_S - z_i^{\mathrm{T}}\hat{\delta}_S + (X_i - \hat{\mu}_S)^{\mathrm{T}}(\hat{\Phi} + C)^{-1}C\hat{\beta}_S \xrightarrow{p_S} 0,$$
$$(D_i + \hat{\sigma}^2_{u,N}) - \{D_i + \hat{\sigma}^2_{u,S} + \hat{\beta}_S(\hat{\Phi}^{-1} + C^{-1})^{-1}\hat{\beta}_S\} \xrightarrow{p_S} 0.$$

As $\hat{\theta}_{i,N}^{EB}$ and $\hat{\theta}_{i,S}^{EB}$ given in (3.23) and (3.45) involve above expressions, we directly have

$$\hat{\theta}_{i,N}^{EB} - \hat{\theta}_{i,S}^{EB} \xrightarrow{p_S} 0.$$

We note that although $\delta_{2,N} = \delta$ and $\hat{\delta}_{2,N} - \hat{\delta} \xrightarrow{p_S} 0$, two estimators $\hat{\delta}_{2,N}$ and $\hat{\delta}$ are not identical in general.

Bibliography

Arima, S., Bell, W. R., Datta, G. S., Franco, C., and Liseo, B. (2017), "Multivariate Fay–Herriot Bayesian estimation of small area means under functional measurement error," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180, 1191–1209.

Arima, S., Datta, G. S., and Liseo, B. (2015), "Bayesian estimators for small area models when auxiliary information is measured with error," *Scandinavian Journal of Statistics*, 42, 518–529.

— (2016), "Models in Small Area Estimation when Covariates are Measured with Error," in *Analysis of Poverty Data by Small Area Estimation*, ed. Pratesi, M., John Wiley & Sons, chap. 8, pp. 149–170.

Arima, S., Datta, G. S., Liseo, B., et al. (2012), "Objective Bayesian analysis of a measurement error small area model," *Bayesian Analysis*, 7, 363–384.

Buonaccorsi, J. P. (2010), *Measurement error: models, methods, and applications*, CRC Press.

Datta, G. S., Delaigle, A., Hall, P., and Wang, L. (2018), "Semi-parametric prediction intervals in small areas when auxiliary data are measured with error," *Statistica Sinica*, 28, 2309–2335.

Datta, G. S., Rao, J. N. K., and Torabi, M. (2010), "Pseudo-empirical Bayes estimation of small area means under a nested error linear regression model with functional measurement errors," *Journal of Statistical Planning and Inference*, 140, 2952–2962.

Fay, R. E. and Herriot, R. A. (1979), "Estimates of income for small places: an application of James-Stein procedures to census data," *Journal of the American Statistical Association*, 74, 269–277.

Fuller, W. A. (2009), *Measurement error models*, vol. 305, John Wiley & Sons.

Ghosh, M. and Sinha, K. (2007), "Empirical Bayes estimation in finite population sampling under functional measurement error models," *Journal of Statistical Planning and Inference*, 137, 2759–2773.

Ghosh, M., Sinha, K., and Kim, D. (2006), "Empirical and hierarchical Bayesian estimation in finite population sampling under structural measurement error models," *Scandinavian Journal of Statistics*, 33, 591–608.

Huang, E. T. and Bell, W. R. (2012), "An empirical study on using previous American Community Survey data versus Census 2000 data in SAIPE models for poverty estimates," Research Report RRS2012-4. Center for Statistical Research and Methodology, U.S Census Bureau, Washington D.C. Available at https://www.census.gov/srd/papers/pdf/rrs2012-04.pdf.

Theil, H. (1971), *Principles of Econometrics*, Wiley/Hamilton publication, John Wiley & Sons.

Torabi, M., Datta, G. S., and Rao, J. N. K. (2009), "Empirical Bayes estimation of small area means under a nested error linear regression model with measurement errors in the covariates," *Scandinavian Journal of Statistics*, 36, 355–369.

Ybarra, L. M. and Lohr, S. L. (2008), "Small area estimation when auxiliary information is measured with error," *Biometrika*, 95, 919–931.

Chapter 4

Subspace Rotations for High-dimensional Outlier Detection

4.1 Introduction

High-throughput data are usually a product of long and complex experiments in laboratories or fields. Due to the multi-step process when generating data, a concern for possible contamination in high-dimensional data is naturally more severe than low-dimensional counterparts. However, even when it is suspected that the data contain some abnormal observations, it is difficult to identify outliers when the dimension is much higher than the sample size. The difficulty arises from the fact that there are insufficient observations to characterize the "regular" behavior of uncontaminated data. Nevertheless, it is crucial in any data analysis to identify outliers since a blind application of a statistical method without being aware of the existence of outliers will likely yield misleading scientific conclusions. For instance, in the area of cancer classification, mislabeled observations can disguise the contrast between classes, and thus result in an unreliable classifier. Furthermore, the existence of clustered outliers in a given class may indicate the presence of a sub-type of cancer. Accordingly, outlier detection is of great importance and a main objective of analysis in many situations.

In this work, we address the problem of detecting outliers among the observations in $\mathbb{R}^d$, when d is much larger than the sample size N, i.e., in the High Dimension, Low Sample Size (HDLSS) setting. Although there is abundant literature on multivariate outlier detection for the case of $d \ll N$, only a few of the existing methods can be applied to HDLSS setting. Sajesh and Srinivasan (2012) introduced a Comedian distance for high-dimensional outlier detection, which is a variation of Mahalanobis distance based on robust estimates of the mean and covariance matrix. When the dimension is large, the Comedian method

is computationally expensive since it requires the inversion of a $d \times d$ matrix. Ro et al. (2015) suggested a method based on a minimum covariance determinant estimator. Due to the strong assumption that the population covariance matrix is diagonal, their method may not perform well in real world situations with correlated variables. Filzmoser et al. (2008) proposed to use the kurtosis of the principal component (PC) scores as a measure for outlyingness. However, this method may not be preferred in HDLSS as it is known that sample PC directions are not generally consistent in HDLSS (Jung and Marron, 2009). Ahn et al. (2012) considered a distance-based HDLSS outlier detection method motivated by the high-dimensional geometric representation (Hall et al., 2005; Ahn et al., 2007). Our numerical study has found that their method is susceptible to the masking effect, possibly due to their one-by-one elimination process.

Difficulties of HDLSS outlier detection are two-fold. Firstly, the high-dimensionality causes all data points to be located on the boundary of a point cloud. Thus it is difficult to distinguish the true outliers from non-outliers. Secondly, it is difficult to construct a hypothesis test on the potential outliers due to the small sample size. We have found that possibly because of these issues, existing methods tend to suffer from large false positives. That is, the methods are able to identify the true outliers correctly; but, they are likely to misclassify non-outliers as outliers. False detection of outliers in HDLSS is a much more serious issue than in low-dimensional cases due to the high cost of data collection, e.g., high-throughput -omics data and medical images.

Our contributions in this research are the following. First, we propose a novel outlier detection method for data with much larger dimensionality than the sample size. Numerical studies suggest that our method is much less likely to falsely declare outliers than existing methods. Second, we extend a random rotation test (Langsrud, 2005; Perry and Owen, 2010) to a more general distributional family, with which we conduct hypothesis tests on the outlyingness of data vectors. Third, we provide theoretical properties of the rotation

tests such as unbiasedness and test size. Lastly, we show that as the dimension increases, the proposed method would detect multiple outliers correctly with overwhelming probability.

The proposed outlier detection procedure consists of two stages: screening and testing. Let $X \in \mathbb{R}^{N\times d}$ be an observed data matrix. In the screening step, we partition the data into row-wise sub-matrices, say X_0 and X_1, respectively, consisting of non-outliers and potential outliers which we call candidate outliers. Then each observation in X_1 is tested based on how far it is from X_0 in the testing step. Abnormality of a candidate outlier is quantified by the orthogonal distance to the affine hyperplane generated by the non-outliers in X_0 (Ahn and Marron, 2010). This distance measure is computationally efficient and has been found to be effective in the context of HDLSS classification (Ahn and Marron, 2010), clustering (Ahn et al., 2012), and outlier detection (Ahn et al., 2019).

The hypothesis test in the testing stage is conducted by a randomization test that generates data that possess the same scatters but free of outliers, which we call the subspace rotation test. The idea of testing via random rotations was suggested by Wedderburn in his unpublished manuscript in 1975. It has been considered in the context of multivariate regression with normal errors in Langsrud (2005) and Perry and Owen (2010), and with left-spherical errors in Solari et al. (2014). However, theoretical justification on the rotation procedure has been limited to the normal population. Moreover, the properties of the hypothesis test have not been fully established. In this work, we investigate the left-spherical distribution family to construct the rotation test and establish its finite-sample exactness formally. We also explore the asymptotic power of the test when the dimension tends to infinity in the context of outlier detection.

4.2 Background and preliminaries

We use, unless otherwise stated, a calligraphic uppercase alphabet to denote a set. For a set $\mathcal{A}$, we denote $\mathcal{A}^c$ and $|\mathcal{A}|$ by the complementary set and cardinality of $\mathcal{A}$, respectively. We also denote $\underline{x}$ and X by a realization of a random vector $\mathbf{x}$ and a random matrix $\mathbf{X}$. The

matrix of differentials of X is denoted by dX. The $a \times a$ identity matrix and $a \times b$ matrix with ones are denoted by I_a and $J_{a,b}$. The column space of a matrix X and its orthogonal complement space will be denoted by $\mathrm{Span}(X)$ and $\mathrm{Span}(X)^{\perp}$, respectively. We also assume that the dimension d is larger than sample size N.

4.2.1 The exterior product and differential forms

Let $f_{\mathbf{x}}(\cdot)$ be the density function of an $n \times 1$ random vector $\mathbf{x}$ and $\mathcal{E} \subset \mathbb{R}^n$. Then the multiple integral

$$\Delta = \int_{\mathcal{E}} f(x_1, \ldots, x_n) dx_1 \ldots dx_n \tag{4.1}$$

represents the probability $\Pr(\mathbf{x} \in \mathcal{E})$. Consider the following change of variables

$$\begin{aligned} x_1 &= x_1(y_1, \ldots, y_n) \\ &\vdots \\ x_n &= x_n(y_1, \ldots, y_n). \end{aligned}$$

Then, with respect to $y_1, \ldots, y_n$, (4.1) can be written as

$$\Delta = \int_{\mathcal{E}'} f\{\underline{x}(\underline{y})\} \det(J) dy_1 \ldots dy_n, \tag{4.2}$$

where $\underline{x} = (x_1, \ldots, x_n)^{\mathrm{T}}$, $\underline{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$, $\mathcal{E}'$ is the image of $\mathcal{E}$, $J = \{\frac{\partial x_i}{\partial y_j}\}_{i,j}$ is the Jacobian matrix and $\det(J)$ denotes the determinant of J.

We illustrate an alternative way to express (4.2). For each $i = 1, \ldots, n$, the differential of the transformation $x_i = x_i(y_1, \ldots, y_n)$ is defined as

$$dx_i = \frac{\partial x_i}{\partial y_1} dy_1 + \cdots + \frac{\partial x_i}{\partial y_n} dy_n. \tag{4.3}$$

By substituting (4.3) for dx_i in (4.1), we have

$$\Delta = \int_{\mathcal{E}'} f(\underline{x}(\underline{y})) \left(\sum_{j=1}^{n} \frac{\partial x_1}{\partial y_j} dy_j \right) \cdots \left(\sum_{j=1}^{n} \frac{\partial x_n}{\partial y_j} dy_j \right). \tag{4.4}$$

Now, we consider the exterior product of the differentials, which is an associative, distributive and anti-commutative binary operation. That is, for all i, j, k, it satisfies

$$(dx_i \bigwedge dx_j) \bigwedge dx_k = dx_i \bigwedge (dx_j \bigwedge dx_k),$$
$$(dx_i + dx_j) \bigwedge dx_k = dx_i \bigwedge dx_k + dx_j \bigwedge dx_k,$$
$$dx_i \bigwedge dx_j = -dx_j \bigwedge dx_i.$$

Note that the anti-commutative law makes $dx_i \wedge dx_j = 0$ if $i = j$. Then the resulting exterior product of the linear differential forms in (4.4) is equivalent to the determinant of the Jacobian matrix as the following theorem shows.

Theorem 6 (Theorem 2.1.1 of Muirhead (2009)). *If $d\underline{y}$ is an $n \times 1$ vector of differentials and if $d\underline{x} = Bd\underline{y}$, where B is an $n \times n$ nonsingular matrix, then*

$$\bigwedge_{i=1}^{n} dx_i = \det(B) \bigwedge_{i=1}^{n} dy_i.$$

Now, let us write (4.1) as

$$\Delta = \int_{\mathcal{E}} f(\underline{x}) dx_1 \bigwedge \ldots \bigwedge dx_n.$$

Since

$$dx_i = \sum_{j=1}^{n} \frac{\partial x_i}{\partial y_j} dy_j, \quad i = 1, \ldots, n,$$

we have

$$d\underline{x} = \begin{bmatrix} \frac{\partial x_1}{\partial y_1} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \vdots & & \vdots \\ \frac{\partial x_n}{\partial y_1} & \cdots & \frac{\partial x_n}{\partial y_n} \end{bmatrix} d\underline{y}.$$

Thus, by Theorem 6, we have

$$\bigwedge_{i=1}^{n} dx_i = \det(J) \bigwedge_{i=1}^{n} dy_i.$$

For an $n \times m$ matrix X, we denote (dX) by the exterior product of the elements of dX such that

$$(dX) \equiv \bigwedge_{j=1}^{m} \bigwedge_{i=1}^{n} dx_{ij}. \tag{4.5}$$

If X is a symmetric $n \times n$ matrix, then (dX) represents the exterior product of the $n(n+1)/2$ upper-triangular elements of dX, where

$$(dX) \equiv \bigwedge_{i \leq j} dx_{ij}.$$

Similarly, if X is a skew-symmetric matrix ($X = -X^{\mathrm{T}}$), then (dX) is defined as the exterior product of $n(n-1)/2$ super-diagonal of dX as

$$(dX) \equiv \bigwedge_{i < j} dx_{ij}. \tag{4.6}$$

Theorem 7 (Theorem 2.1.5 of Muirhead (2009)). *If $X = BYC$ where X and Y are $n \times m$ matrices and B and C are fixed nonsingular $n \times n$ and $m \times m$ matrices, then*

$$(dX) = \det(B)^m \det(C)^n (dY).$$

An exterior differential form of degree r in $\mathbb{R}^n$ is an expression of the type

$$\sum_{i_1 < \cdots < i_r} p_{i_1 \cdots i_r}(\underline{x}) \bigwedge_{l=1}^{r} dx_{i_l}, \tag{4.7}$$

where $\{i_1, \ldots, i_r\} \subset \{1, \ldots, n\}$ and $p_{i_1 \cdots i_r}(\underline{x})$ are analytic functions of $\underline{x}$. The expression (4.7) can be considered as the integrand of an r-dimensional surface integral in $\mathbb{R}^n$. The maximum degree the exterior differential form can have is n since at least one of the dx_i is repeated if $r > n$.

4.2.2 The invariant measure on the Stiefel manifold

Let us denote $\mathcal{V}_{m,n}$ by the Stiefel manifold whose elements are m-frames in $\mathbb{R}^n$ such that $\mathcal{V}_{m,n} = \{V \in \mathbb{R}^{n \times m} | V^{\mathrm{T}} V = I_m\}$, where the m-frame indicates an orthonormal set of m

vectors in $\mathbb{R}^n$. For a $n \times m$ matrix A, it can be shown by the Gram-Schmidt process that $m(m+1)/2$ functionally independent constraints are needed for A to be in $\mathcal{V}_{m,n}$. Thus, the dimension of $\mathcal{V}_{m,n}$ is $nm - m(m+1)/2$. When $m = n$, $\mathcal{V}_{m,n}$ is the orthogonal group, i.e., the set of m-dimensional orthogonal transformations, which is denote by $\mathcal{O}_m$.

For $O \in \mathcal{O}_n$, the differential of the $O^{\mathrm{T}}O$ is given by

$$dO^{\mathrm{T}}O + O^{\mathrm{T}}dO = 0_{n,n}, \tag{4.8}$$

where $0_{n,n}$ is the $n \times n$ null matrix. The equality (4.8) indicates that the $O^{\mathrm{T}}dO$ is skew-symmetric. Thus, by (4.6) and Theorem 7, we have

$$(dO) = \bigwedge_{i<j}^{n} \underline{o}_i^{\mathrm{T}} d\underline{o}_j, \tag{4.9}$$

where $\underline{o}_1, \ldots, \underline{o}_m$ are the columns of O, and it defines the invariant measure on $\mathcal{O}_n$. For $V \in \mathcal{V}_{m,n}$, we similarly have

$$dV^{\mathrm{T}}V + V^{\mathrm{T}}dV = 0_{m,m}. \tag{4.10}$$

Let $V_1 = [\underline{v}_{m+1}, \ldots, \underline{v}_n]$ be an $(n-m)$-frame such that $[V, V_1] \in \mathcal{O}_n$. By pre-multiplying $[V, V_1]^{\mathrm{T}}$ to dV, we get

$$\begin{bmatrix} V^{\mathrm{T}} \\ V_1^{\mathrm{T}} \end{bmatrix} dV = \begin{bmatrix} V^{\mathrm{T}}dV \\ V_1^{\mathrm{T}}dV \end{bmatrix}. \tag{4.11}$$

We first consider matrices $V^{\mathrm{T}}dV$ and $V_1^{\mathrm{T}}dV$ in the right side of (4.11). Knowing that $V^{\mathrm{T}}dV$ is skew-symmetric from (4.10), we have corresponding differential forms

$$(V^{\mathrm{T}}dV) = \bigwedge_{i<j}^{m} \underline{v}_i^{\mathrm{T}} d\underline{v}_j, \tag{4.12}$$

$$(V_1^{\mathrm{T}}dV) = \bigwedge_{i=1}^{m} \bigwedge_{j=1}^{n-m} \underline{v}_{m+j}^{\mathrm{T}} d\underline{v}_i. \tag{4.13}$$

By Theorem 7, the exterior product of the elements of the left side of (4.11) is equivalent to (dV). Also, exterior product of (4.12) and (4.13) is a $(2nm - m^2 - m)/2$ degree differential

form which is of maximum degree. Thus, we have

$$(dV) = \bigwedge_{i=1}^{m} \bigwedge_{j=1}^{n-m} \underline{v}_{m+j}^{\mathrm{T}} d\underline{v}_i \bigwedge_{i<j}^{m} \underline{v}_i^{\mathrm{T}} d\underline{v}_j, \tag{4.14}$$

and the differential form (4.14) defines invariant measure on Stiefel manifold $\mathcal{V}_{m,n}$.

The volume, or the area of the surface, of the Stiefel manifold $\mathcal{V}_{m,n}$ is

$$\mathrm{Vol}(\mathcal{V}_{m,n}) = \int_{\mathcal{V}_{m,n}} (dV) = \frac{2^m \pi^{mn/2}}{\Gamma_m(n/2)},$$

where $\Gamma_m(\cdot)$ is the multivariate gamma function (Muirhead, 2009). The uniform probability measure on $\mathcal{V}_{m,n}$ is obtained by normalizing the invariant measure (dV) by its volume, which will be denoted by $[dV] = (dV)/\mathrm{Vol}(\mathcal{V}_{m,n})$. We denote the uniform distribution on $\mathcal{V}_{m,n}$ by $\mathrm{Unif}(\mathcal{V}_{m,n})$.

4.2.3 The left-spherical family

Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^{\mathrm{T}}$ be a sample taking values in $\mathcal{X} \subseteq \mathbb{R}^{N \times d}$. Then the left-spherical distributional family is defined as follows.

Definition 1. *(Gupta and Nagar, 1999) Let $\mathbf{X}$ be an $N \times d$ random matrix according to a probability distribution $\mathbb{P}$. If $O\mathbf{X}$ is identically distributed as $\mathbf{X}$, for all $O \in \mathcal{O}_N$, then $\mathbb{P}$ is called a left-spherical distribution, denoted by $\mathbb{P} \in LS_{N,d}$.*

Necessary conditions for left-sphericity are zero mean and identity (up to a constant factor) row-wise covariance matrix. In other words, observations need not be independent but they should be uncorrelated. Throughout this chapter, we assume that the column-wise covariance matrix is nonsingular. The left-spherical family contains matrix normal, matrix t and scale-mixture of left-spherical distributions. In particular, under independence, left-sphericity implies matrix normal (James, 1954).

It is known that $\mathbf{S} = \mathbf{X}^{\mathrm{T}}\mathbf{X}$ is a sufficient statistic for the left-spherical family. To see this, we first consider the following lemma.

Lemma 3. *Suppose that a random matrix* $\mathbf{X}$ *is from a left-spherical distribution* $\mathbb{P} \in LS_{N,d}$. *Let* $\mathbf{X} = \mathbf{O}\mathbf{\Lambda}\mathbf{\Gamma}^T$ *be the nonsingular part of the singular value decomposition. Then,* $\mathbf{O} \sim \text{Unif}(\mathcal{O}_N)$ *and independent of* $\mathbf{\Lambda}\mathbf{\Gamma}^T$.

Lemma 3 directly show that, conditioning on $\mathbf{\Lambda}\mathbf{\Gamma}^{\mathrm{T}}$, $\mathbf{X}$ is uniformly distributed on the set $\{X \in \mathcal{X} \mid X^{\mathrm{T}}X = S\}$ regardless of $\mathbb{P}$, and thus $\mathbf{S}$ is a sufficient statistic for the left-spherical family.

The following fact will be useful for developing the testing procedure in Sections 4.3 and 4.4. For $W \in \mathbb{R}^{N\times d}$, let $\Psi_{\mathbf{X}}(W)$ be the characteristic function of $\mathbf{X}$. Since $\mathbf{O}\mathbf{X}$ and $\mathbf{X}$ are identically distributed, $\Psi_{\mathbf{X}}(W) = \Psi_{\mathbf{X}}(\mathbf{O}^{\mathrm{T}}W)$. Since $W^{\mathrm{T}}W$ is maximally invariant (Lehmann and Romano, 2006) under $\mathcal{O}_N$, $\Psi_{\mathbf{X}}(W)$ can be expressed as

$$\Psi_{\mathbf{X}}(W) = \psi(W^{\mathrm{T}}W), \tag{4.15}$$

for some function ψ. Based on this, we have the following theorem, which implies that any row-wise partition of a left-spherical $\mathbf{X}$ is also left-spherical.

Theorem 8. *(Gupta and Nagar, 1999) Let* $\mathbf{X} \sim \mathbb{P} \in LS_{N,d}$ *be partitioned as* $\mathbf{X} = [\mathbf{X}_0^T, \mathbf{X}_1^T]^T$. *Then,* $\mathbf{X}_i \sim \mathbb{P}_{n_i} \in LS_{n_i,d}$, *where* $\mathbf{X}_i \in \mathbb{R}^{n_i\times d}$, $i = 0, 1$.

4.3 Randomization via subspace rotations

Let X be the observed data and assume that the population distribution is left-spherical. The random rotation (Langsrud, 2005) perturbs X by pre-multiplying an independent $\mathbf{O} \sim \text{Unif}(\mathcal{O}_N)$ such that $\mathbf{Y} = \mathbf{O}X$ and considers $\mathbf{Y}$ as a new sample from the population. Noting that $\mathbf{Y}^{\mathrm{T}}\mathbf{Y} = X^{\mathrm{T}}X$, the rotation randomizes the data conditioning on $\mathbf{S}$. This implies that the perturbed data provide as much information about the population distribution as the original data.

As we noted in the previous section, the left-spherical assumption imposes the zero mean condition to the data, which is rarely satisfied in real applications. In this section, we assume a family of distributions which we called location-shifted left-spherical family and study a

group of transformations under which the probability distributions are invariant. Specifically, we consider a random matrix $\mathbf{X}$ whose population centered version, $\mathbf{X}-\mathsf{E}(\mathbf{X})$, is left-spherical and call the distribution of $\mathbf{X}$ a location-shifted left-spherical distribution.

We note that neither $\mathbf{OX}$ nor $\mathbf{OX}_c$ is identically distributed as $\mathbf{X}$, where $\mathbf{X}_c$ is the sample centered version of $\mathbf{X}$. Let us assume that $\mathsf{E}(\mathbf{X})$ is known up to its column space. In particular, we assume that $\mathsf{E}(\mathbf{X})$ can be written as $\mathsf{E}(\mathbf{X}) = M_0 B^{\mathrm{T}}$, where M_0 is a known $N \times m_0$ matrix with m_0 basis vectors, B is the $d \times m_0$ unobservable coefficient matrix, and $m_0 < N$. We denote by $LS_{N,d}(M_0)$ the set of location-shifted left-spherical distributions with mean matrix whose column space is spanned by M_0. For simplicity, we assume in this section that the columns of M_0 are orthonormal.

Let M_1 be a $N \times m_1$ matrix whose columns constitute an orthonormal basis of $\mathrm{Span}(M_0)^{\perp}$ and consider the following subgroup of $\mathcal{O}_N$:

$$\mathcal{R}(M_0) = \{R \mid R = M_0 M_0^{\mathrm{T}} + M_1 O M_1^{\mathrm{T}}, \quad O \in \mathcal{O}_{m_1}\}. \tag{4.16}$$

Perry and Owen (2010) provided the invariance of the normal family in $LS_{N,d}(M_0)$ under (4.16) by showing that transformations in $\mathcal{R}(M_0)$ preserve the first and second moments of the normal family. In what follows we extend the distributional invariance to the location-shifted left-spherical family $LS_{N,d}(M_0)$.

By slightly abusing the notations, let $\mathbf{V}\mathbf{\Lambda}\mathbf{\Gamma}^{\mathrm{T}}$ be the nonsingular part of the singular value decomposition of $M_1 M_1^{\mathrm{T}}\mathbf{X}$. It is known by Fang and Li (1999) that $(M_0 M_0^{\mathrm{T}}\mathbf{X}, \mathbf{\Lambda}\mathbf{\Gamma}^{\mathrm{T}})$ is a complete sufficient statistic for $LS_{N,d}(M_0)$. Hence, $\mathbf{V}$ determines the conditional distribution of $\mathbf{X}$ given the sufficient statistic as $\mathbf{X} = M_0 M_0^{\mathrm{T}}\mathbf{X} + \mathbf{V}\mathbf{\Lambda}\mathbf{\Gamma}^{\mathrm{T}}$. We find the distribution of $\mathbf{V}$ and the subspace in which the probability measure is defined as follows.

Theorem 9. *Let $M_1 M_1^{T}\mathbf{X} = \mathbf{V}\mathbf{\Lambda}\mathbf{\Gamma}^{T}$, where $\mathbf{X} \sim \mathbb{P} \in LS_{N,d}(M_0)$. Then $\mathbf{V}$ is independent of $\mathbf{\Lambda}\mathbf{\Gamma}^{T}$, and $\mathbf{V}$ is uniformly distributed on $\mathcal{V}^1_{m_1,N} = \{V \mid \mathrm{Span}(V) = \mathrm{Span}(M_1),\ V \in \mathcal{V}_{m_1,N}\}$.*

Corollary 9.1. *The volume of $\mathcal{V}^1_{m_1,N}$ is the same as the volume of $\mathcal{O}_{m_1}$.*

The probability measure of $\mathbf{V}$ is singular in that, for any $\underline{a} \in \text{Span}(M_0)$, $\underline{a}^{\text{T}}\mathbf{V} = 0$ with probability one. An interesting observation is that $\mathbf{V}$ is identically distributed as the conditional distribution of $\mathbf{W} \sim \text{Unif}(\mathcal{V}_{m_1,N})$ conditioning on $\mathbf{W}\mathbf{W}^{\text{T}} = M_1 M_1^{\text{T}}$. To illustrate, let $\mathcal{P}_{m_1,N}$ denote the set of $N \times N$ orthogonal projections of rank m_1. Also, let $[dP]$ and $[dO]$ be the differential forms for the uniform measures on $\mathcal{P}_{m_1,N}$ and $\mathcal{O}_{m_1}$, respectively. Then the differential form $[dW]$ defining the uniform measure on $\mathcal{V}_{m_1,N}$ can be decomposed into two independent uniform measures such that $[dW] = [dP][dO]$ (Chikuse, 2012). This implies that, conditioning on $\mathbf{P} = M_1 M_1^{\text{T}}$, $\mathbf{W}$ is in $\mathcal{V}^1_{m_1,N}$ with conditional probability measure $[dO]$, where $\mathbf{P} \sim \text{Unif}(\mathcal{P}_{m_1,N})$. Corollary 9.1 also confirms the above interpretation.

Theorem 10. *The volume of $\mathcal{R}(M_0)$ is the same as the volume of $\mathcal{O}_{m_1}$.*

From Theorems 9 and 10, we find that, although $\mathcal{V}^1_{m_1,N}$ and $\mathcal{R}(M_0)$ contain different sizes of matrices, they have the same volume in the common subspace. We denote the differential form for the normalized invariant measure on $\mathcal{R}(M_0)$ by $[dR]$. For notational convenience, we denote $\mathcal{R}(M_0)$ by $\mathcal{R}$ as we always consider the uniform distribution on $\mathcal{R}(M_0)$ together with $LS_{N,d}(M_0)$. Let $\text{Unif}(\mathcal{R})$ denote the uniform distribution on $\mathcal{R}$.

It can bee seen that, for any $R \in \mathcal{R}$, $M_0 M_0^{\text{T}}\mathbf{X}$ is invariant in that $RM_0 M_0^{\text{T}}\mathbf{X} = M_0 M_0^{\text{T}}\mathbf{X}$. Furthermore, since $\mathbf{V}$ and $\mathbf{\Lambda\Gamma}^{\text{T}}$ are independent by Theorem 9, it suffices to show that $\text{Unif}(\mathcal{V}^1_{m_1,N})$ is invariant under $\mathcal{R}$ to establish the distributional invariance of general location-shifted left-spherical distributions.

Lemma 4. *Let $\mathbf{R} \sim \text{Unif}(\mathcal{R})$ and $\mathbf{V} \sim \text{Unif}(\mathcal{V}^1_{m_1,N})$. Then, for $W_0 \in \mathbb{R}^{N\times N}$ and $W_1 \in \mathbb{R}^{N\times m_1}$, characteristic functions of $\mathbf{R}$ and $\mathbf{V}$ is given by*

$$\Psi_{\mathbf{R}}(W_0) = \text{etr}(\iota W_0^T M_0 M_0^T)\,{}_0F_1\left(\frac{m_1}{2}; -\frac{1}{4}M_1^T W_0^T M_1 M_1^T W_0^T M_1\right),$$
$$\Psi_{\mathbf{V}}(W_1) = {}_0F_1\left(\frac{m_1}{2}; -\frac{1}{4}W_1^T M_1 M_1^T W_1\right),$$

where $\text{etr}(A) = \exp\{\text{tr}(A)\}$, ι is the imaginary number, ${}_0F_1(b; A)$ is the hypergeometric function with a matrix argument A and a parameter b (Muirhead, 2009).

Using the characteristic functions of $\mathbf{R}$ and $\mathbf{V}$, we can readily show that $\mathbf{RV}$ and $\mathbf{V}$ are identically distributed. This provides the following theorem.

Theorem 11. *Let* $\mathbf{X} \sim \mathbb{P} \in LS_{N,d}(M_0)$ *and* $\mathbf{R} \sim \mathrm{Unif}(\mathcal{R})$*, where* $\mathbf{X}$ *and* $\mathbf{R}$ *are independent. Then,* $\mathbf{RX}$ *and* $\mathbf{X}$ *are identically distributed, and* $\mathbf{RX}$ *and* $\mathbf{R}$ *are independent.*

Based on Theorem 11, we consider $\mathbf{RX}$ as new data from the population distribution of $\mathbf{X}$. As R only transforms $\mathbf{V}$ keeping $\mathrm{Span}(\mathbf{V})$ and the sufficient statistic $(M_0 M_0^{\mathrm{T}}\mathbf{X}, \mathbf{\Lambda\Gamma}^{\mathrm{T}})$ fixed, we call the randomization carried by (4.16) "subspace rotation" (SR).

4.3.1 Subspace rotation tests for non-trivial mean

In this section, we construct the subspace rotation (SR) tests and explore their non-asymptotic properties. Using the subspace rotations, a SR test can be constructed as follows. Suppose that we have the following null hypothesis:

$$H_0 : \mathbb{P} \in LS_{N,d}(M_0), \tag{4.17}$$

and that $t(\cdot)$ is a chosen test statistic such that the rejection region has the form of $\{t(X) > c\}$ for some constant $c \in \mathbb{R}$. By Theorem 11, for independently distributed $\mathbf{R} \sim \mathrm{Unif}(\mathcal{R}), t(\mathbf{RX})$ is identically distributed as $t(\mathbf{X})$ under H_0. By conditioning on $\mathbf{X}$, the distribution function of $t(\mathbf{RX})$ is given by the surface integral

$$F_{t|\mathbf{X}}(z|X) = \int_{\mathcal{R}} 1\{t(RX) \le z\}[dR], \tag{4.18}$$

where $1(\cdot)$ denotes an indicator function. We call (4.18) as the SR distribution of the test statistic which unbiasedly estimates the true distribution function of $t(\mathbf{X})$ as follows.

Theorem 12. *The conditional distribution in* (4.18) *is an unbiased estimator of the true distribution function* $F_t(z) = \Pr(t(\mathbf{X}) \le z)$ *in the sense that* $\mathsf{E}_{\mathbf{X}}\{F_{t|\mathbf{X}}(z|\mathbf{X})\} = F_t(z)$.

For a given significance level α, the distribution function in (4.18) is used to find a critical value c_α such that

$$c_\alpha(X) = \inf\{z | F_{t|\mathbf{X}}(z|X) \ge 1 - \alpha\}. \tag{4.19}$$

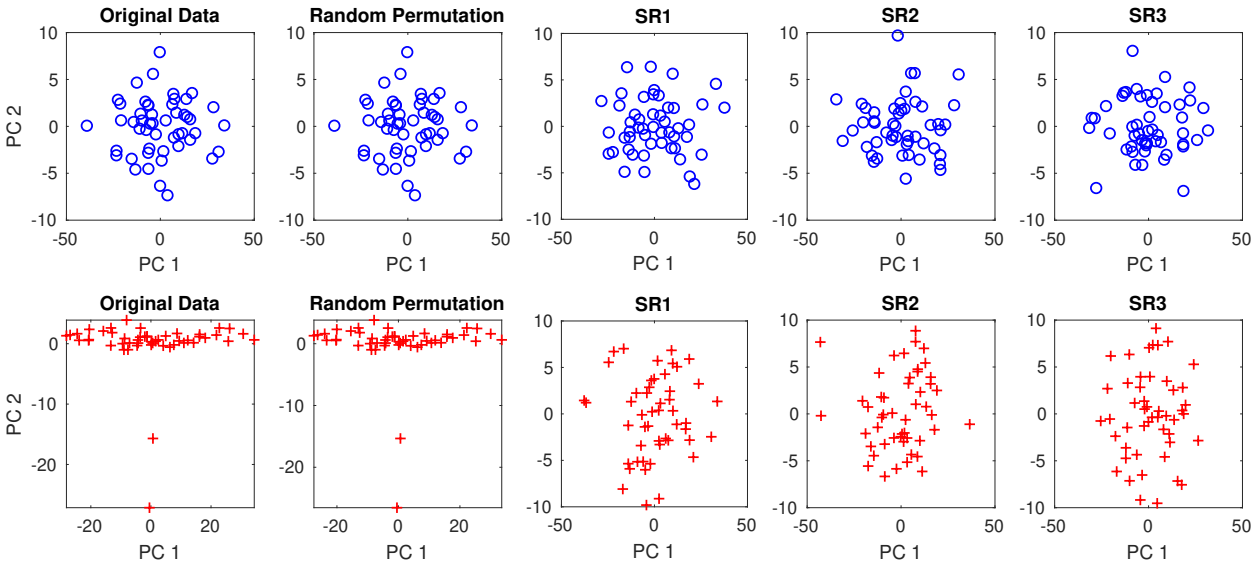


Figure 4.3.1: Graphical illustrations of random permutation and subspace rotation using two toy data sets with no outlier (top) and two outliers (bottom), where $d = 500$ and $N = 50$. The original, randomly permuted and three subspace rotated data are projected onto the first two principal component directions.

With the critical value, the SR test function is defined by

$$\phi(X) = 1\{t(X) \geq c_\alpha(X)\}. \tag{4.20}$$

The SR test can be considered as a continuous extension of the permutation tests (Lehmann and Romano, 2006) with the symmetric group being replaced by $\mathcal{R}$, a subgroup of the orthogonal group. We note that, in the context of outlier detection, the permutation distribution (Lehmann and Romano, 2006) of a suitable test statistic degenerates into a point as permutations do not change the geometric configuration of the data. On the contrary, the SR allows approximating the null distribution of the test statistic by perturbing all the data points in $\mathbb{R}^d$ while preserving its sample mean and covariance matrix. We compare SR with random permutation using two toy data sets in Figure 4.3.1. We set $d = 500$ and generate two data sets with $N = 50$ from zero mean normal distribution with covariance matrix $\Sigma = \{\sigma_{ij}\}$, where $\sigma_{ii} = 1$, $\sigma_{ij} = 0.5$, $i \neq j$. We add two outliers to the one of the

data sets by shifting the mean by $\sqrt{d}\theta$, where $\theta \sim \text{Unif}(\mathcal{V}_{1,d})$. The data set with no outlier is illustrated at the top row and the data set with two outliers is illustrated at the bottom. For each data set, one randomly permuted and three subspace rotated data sets are displayed in each column. It can be seen that the subspace rotated contaminated data sets have similar shapes as the ones from the data with no outlier. Hence, SR provides a way of generating "null" data sets which inherit the covariance structure of the original data but are free of specific features such as clusters or outliers.

Now we provide the level of the SR test (4.20). The following theorem shows that it controls the type-I error rate below α and achieves finite sample exactness when a test statistic is continuous.

Theorem 13. *Under H_0 given in (4.17) and a given significance level α, the SR test given in (4.20) is a level-α test. If the test statistic $t(\cdot)$ is continuous, then the p-value of the SR test is uniformly distributed over $(0, 1)$, i.e. the SR test is a size-α test.*

In addition to the properties under H_0, it is also important to investigate the power of the SR test, where the power function of the SR test can be written as

$$\beta(\mathbb{P}) = \mathsf{E}_{\mathbf{X}}[1\{t(\mathbf{X}) \geq c_\alpha(\mathbf{X})\}]$$

for $\mathbb{P} \notin LS_{N,d}(M_0)$. However, as Hoeffding (1952) stated, evaluating the power function is difficult as the critical value $c_\alpha(\mathbf{X})$ is also a random quantity. To explore the power of the SR test, we consider the p-value of the SR test given below.

$$\vartheta_t(\mathbf{X}) = \mathsf{E}_{\mathbf{R}}[1\{t(\mathbf{X}) \leq t(\mathbf{R}\mathbf{X})\}|\mathbf{X}]. \tag{4.21}$$

In Section 4.5.2, we study the high-dimensional asymptotic power of the SR test by investigating the behavior of $\vartheta_t(\mathbf{X})$.

4.4 Application to high-dimensional outlier detection

In this section, we propose a high-dimensional outlier detection methodology which consists of two stages: screening and testing potential outliers. In Section 4.4.1, we first introduce

a measure of abnormality. Then, in Section 4.4.2, we provide a screening algorithm for potential outliers that we call the candidate outliers. In Section 4.4.3, we provide a sequential hypothesis testing procedure for the candidate outliers using the SR tests.

4.4.1 Measure of abnormality

Let $\mathbf{x}_i$ be an observation in $\mathbf{X}$ and $\mathbf{X}_0$ be a row-wise sub-matrix of $\mathbf{X}$ that does not contain $\mathbf{x}_i$. Measuring how distant $\mathbf{x}_i$ is from $\mathbf{X}_0$ when $d > N$ is an important consideration and we propose to use the so-called Distance to Hyperplane (DH). The DH of $\mathbf{x}_i$ to $\mathbf{X}_0$ is the closest L_2 distance between the affine hyperplane generated by the rows of $\mathbf{X}_0$ and $\mathbf{x}_i$ (Ahn and Marron, 2010). The DH has been found effective in HDLSS setting in the context of clustering (Ahn et al., 2012) and multiple-outlier detection (Ahn et al., 2019). Note that DH is well-defined with probability one when the underlying distribution is continuous. Let $\mathbf{P}_0$ be the projection matrix onto the row space of $\mathbf{X}_0$ and $\bar{\mathbf{x}}_0$ be the average of the rows of $\mathbf{X}_0$. Then the DH of $\mathbf{x}_i$ to $\mathbf{X}_0$ is given by

$$DH(\mathbf{x}_i|\mathbf{X}_0) = \|(I_d - \mathbf{P}_0)(\mathbf{x}_i - \bar{\mathbf{x}}_0)\|_2, \tag{4.22}$$

where $\|\cdot\|_2$ refers to the L_2 norm.

4.4.2 Stage I: screening of candidate outliers

In this section, we introduce our outlier screening algorithm. Let n_0 and n_1 be the (unknown) numbers of non-outliers and outliers in the data such that $N = n_0 + n_1$. We assume that the number of non-outliers is greater than the number of outliers so that $n_0 > n_1 + 1$.

Suppose we know from prior knowledge that n_1 is bounded by $\tilde{n}_1$, a safe upper bound for the number of outliers. We propose to screen $\tilde{n}_1$ candidate outliers through the following steps. First, we identify $S < N - \tilde{n}_1$ observations, which are located at the center of the data cloud. In the second step, we find $\tilde{n}_1$ observations that are the most distant from the affine hyperplane generated by the S observations. We consider these $\tilde{n}_1$ observations as candidate outliers and let $\mathcal{I}$ denote the index set of the candidate outliers.

The screening algorithm is as follows. First, for each observation, we calculate L_2 distances to each of $N-1$ rest of observations and obtain their median. Once we obtain N median values, we take the S observations with the smallest median values. We consider these observations as surely non-outliers, as these observations are located at the center of the majority of observations. We call the index set of the S observations as the initial reference set and denote the initial reference set by $\mathcal{I}_0$. Then we measure how distant each of the $N-S$ observations not in $\mathcal{I}_0$ is to $\mathcal{I}_0$ based on DH. The index set of candidate outliers $\mathcal{I}$ is obtained by selecting $\tilde{n}_1$ observations with the largest DH values. We set $\mathbf{X}_1 = [\mathbf{x}_1^*, \ldots, \mathbf{x}_{\tilde{n}_1}^*]^{\mathrm{T}}$ with the observations in $\mathcal{I}$ and $\mathbf{X}_0$ with the remaining $\tilde{n}_0 = N - \tilde{n}_1$ observations. Finally, we calculate the DH measures between each candidate outlier to the affine hyperplane of $\mathbf{X}_0$ as

$$t_j := DH(\mathbf{x}_j^* | \mathbf{X}_0), \quad j = 1, \ldots, \tilde{n}_1, \tag{4.23}$$

which will be used as test statistics for the hypothesis testing in the second stage.

In practice, the upper bound $\tilde{n}_1$ can be chosen by experts based on prior knowledge of the data. Under low-dimensional setting, where d is fixed and N increases, it is suggested that a reasonable upper bound $\tilde{n}_1$ is of the form $\tilde{n}_1 = O(N^{\xi})$, where $0 < \xi < 1$ (Barnett and Lewis, 1974; Simar, 2003). This condition assumes that the proportion of outliers decreases as the sample size increases. However, this assumption seems unrealistic under the HDLSS setting, where d grows with fixed N. In our asymptotic and numerical studies, we assume that $\tilde{n}_1$ increases at the same rate as the sample size, which corresponds to $\tilde{n}_1 = O(N)$.

For the number of the initial reference observations S, we find in Section 4.5.1 that, as long as $S < n_0$, the proposed screening procedure successfully screens out all the outliers with overwhelming probability for large d. However, since n_0 is unknown in real situations, we recommend using S that is strictly less than half of the sample size. In particular, our numerical studies suggest that $S = \tilde{n}_1 = \lfloor 0.3N \rceil$ is reasonable choices for S and $\tilde{n}_1$, where $\lfloor a \rceil$ is the nearest integer of $a \in \mathbb{R}$.

4.4.3 Stage II: sequential SR tests on candidate outliers

The sampling distribution of (4.23) under spherical normal population is known to be a scaled χ^2 (Ahn et al., 2012), but, in general, it is not readily available. We propose to utilize the SR distribution in estimating the sampling distribution of (4.23). We note that any row-wise partition of a left-spherical random matrix is also left-spherical as in Theorem 8. Thus, for each $\mathbf{x}_j^*$ in $\mathbf{X}_1$, the null hypothesis that $\mathbf{x}_j^*$ is not an outlier can be written as

$$\mathbf{Y}_j \sim \mathbb{P} \in LS_{n,d}(J_{n,1}), \quad j = 1, \ldots, \tilde{n}_1,$$

where $\mathbf{Y}_j = [\mathbf{X}_0^{\mathrm{T}}, \mathbf{x}_j^*]^{\mathrm{T}}$ is the non-outliers augmented by a candidate outlier, $n = \tilde{n}_0 + 1$. By Theorem 11, the distribution of $\mathbf{Y}_j$ is invariant under transformations in $\mathcal{R}$, where $\mathcal{R}$ is analogously defined as (4.16) with $M_0 = n^{-1/2} J_{n,1}$.

We assume without loss of generality that the candidate outliers $\mathbf{x}_1^*, \ldots, \mathbf{x}_{\tilde{n}_1}^*$ are rearranged so that that $t_1 \geq t_2 \geq \cdots \geq t_{\tilde{n}_1}$. Starting with $\mathbf{x}_1^*$, we propose to carry out sequential SR tests as the following. In order to test $H_{0,1} : \mathbf{Y}_1 \sim \mathbb{P} \in LS_{n,d}(J_{n,1})$, we generate $\mathbf{R}_k \overset{iid}{\sim} \mathrm{Unif}(\mathcal{R})$, $k = 1, \ldots, K$, and obtain the subspace rotated data sets $\mathbf{Y}_{1,k} = \mathbf{R}_k \mathbf{Y}_1$. Note that under $H_{0,1}$, each $\mathbf{Y}_{1,k}$ will be a conditionally independent sample from the location-shifted left-spherical population. Then we find the observation which is the most far from the rest of the observations in $\mathbf{Y}_{1,k}$, and store the corresponding DH value. Specifically, for each $\mathbf{Y}_{1,k}$, we calculate n leave-one-out (LOO) DH such that

$$DH(\mathbf{y}_i | \mathbf{Y}_{1,k}^{-i}) = \|(I_d - \mathbf{P}_{-i})(\mathbf{y}_i - \bar{\mathbf{y}}_{-i})\|_2, \quad i = 1, \ldots, n,$$

where $\mathbf{Y}_{1,k}^{-i}$ is the subspace rotated data without the ith row $\mathbf{y}_i$, $\mathbf{P}_{-i}$ and $\bar{\mathbf{y}}_{-i}$ are analogously defined with $\mathbf{Y}_{1,k}^{-i}$ as (4.22). Then we find the maximum of the n LOO DH values, which is denoted as $t_{1,k}$, $k = 1, \ldots, K$. Using these K reference values, we estimate the critical value $c_\alpha(\mathbf{Y}_1)$ as

$$\hat{c}_{\alpha,1} = \min\{z | \hat{F}_{t|\mathbf{Y}_1}(z|\mathbf{Y}_1) \geq 1 - \alpha\}, \text{ where } \hat{F}_{t|\mathbf{Y}_1}(z|\mathbf{Y}_1) = K^{-1} \sum_{k=1}^{K} 1(t_{1,k} \leq z).$$

Finally we declare $\mathbf{x}_1^*$ as an outlier if $t_1 > \hat{c}_{\alpha,1}$. If $\mathbf{x}_1^*$ is found to be an outlier, we move on to $\mathbf{x}_2^*$ and repeat the above process to test $H_{0,2} : \mathbf{Y}_2 \sim \mathbb{P} \in LS_{n,d}(J_{n,1})$. We continue this until we fail to reject $H_{0,j}$, $j \leq \tilde{n}_1$. At this point, we finally conclude $\mathbf{x}_1^*, \ldots, \mathbf{x}_{j-1}^*$ as identified outliers. Algorithm 1 in Section 4.10 summarizes the proposed outlier detection algorithm.

4.5 High-dimensional asymptotics of outliers

In this section we justify the proposed outlier detection procedure using high-dimensional asymptotic arguments. In particular, we employ the HDLSS asymptotics (Hall et al., 2005; Ahn et al., 2007), assuming that the dimension grows while the sample size is fixed. With a slight abuse of notation, we let $\mathbf{x} = (x_1, \cdots, x_d)^{\mathrm{T}}$ represent a random vector of non-outliers and $\mathbf{x}^o = (x_1^o, \cdots, x_d^o)^{\mathrm{T}}$ represent a random vector of outliers. The following conditions on moments are commonly assumed in the HDLSS literature (Hall et al., 2005; Jung and Marron, 2009).

Condition 1. *Assume that the population structure of the data satisfies the following as d increases:*

(a) The fourth moments of the entries of the data vectors are uniformly bounded.

(b) $\sum_{l=1}^{d} \{\mathsf{E}(x_l) - \mathsf{E}(x_l^o)\}^2/d \longrightarrow \mu^2$.

(c) $\sum_{l=1}^{d} \mathsf{Var}(x_l)/d \longrightarrow \sigma^2$.

(d) $\sum_{l=1}^{d} \mathsf{Var}(x_l^o)/d \longrightarrow \tau^2$.

(e) For both $\mathbf{x}$ and $\mathbf{x}^o$, there exists a permutation of entries such that the sequence of the variables are ρ-mixing for functions that are dominated by quadratics.

The ρ-mixing in (e) is a mild condition to achieve the law of large numbers for sequence of correlated random variables (Kolmogorov and Rozanov, 1960). For $-\infty \leq i \leq j \leq \infty$, consider the σ-field $\mathcal{F}_i^j$ generated by the random variables $\{z_l | i \leq l \leq j\}$ and the following

maximal correlation coefficient (Bradley et al., 2005):

$$\rho(m) = \sup |\operatorname{corr}(x, y)|, \quad x \in L_2(\mathcal{F}_{-\infty}^{i}),\ y \in L_2(\mathcal{F}_{i+m}^{\infty}),$$

where the supremum is taken over all x, y and $1 \leq i < \infty$, and $L_2(\mathcal{F})$ is the set of square integrable and $\mathcal{F}$–measurable random variables. Then a sequence of random variables $\{z_l\}$ is called ρ-mixing if $\rho(m) \to 0$ as m tends to infinity.

4.5.1 Sure screening of outliers

Let $\mathcal{J}$ be the index set of outliers, $|\mathcal{J}| = n_1$. Under Condition 1, the HDLSS geometric representation implies that, for any $i \neq i'$ in $\mathcal{J}^c$ and $j \neq j'$ in $\mathcal{J}$,

$$d^{-1/2}\|\mathbf{x}_i - \mathbf{x}_{i'}\|_2 \xrightarrow{p} \sqrt{2\sigma^2} \tag{4.24}$$

$$d^{-1/2}\|\mathbf{x}_j^o - \mathbf{x}_{j'}^o\|_2 \xrightarrow{p} \sqrt{2\tau^2}$$

$$d^{-1/2}\|\mathbf{x}_j^o - \mathbf{x}_i\|_2 \xrightarrow{p} \sqrt{\mu^2 + \sigma^2 + \tau^2} \tag{4.25}$$

as $d \to \infty$, where $\xrightarrow{p}$ indicates convergence in probability. Thus, for each $\mathbf{x}_i$ and $\mathbf{x}_j^o$, the medians of L_2 distances to each of $N - 1$ rest of observations converge to $\sqrt{2\sigma^2}$ and $\sqrt{\mu^2 + \sigma^2 + \tau^2}$ after scaled by $\sqrt{d}$, provided that $n_0 > n_1 + 1$.

Recall that $\mathcal{I}_0$ and $\mathcal{I}$ are the index sets of surely non-outliers and candidate outliers based on the screening procedure described in Section 4.4.2. A necessary condition for the successful screening, $\mathcal{J} \subset \mathcal{I}$, is that $\mathcal{I}_0$ does not include any of the outlier indices in $\mathcal{J}$.

Lemma 5. *Under Condition 1, if $n_0 > n_1 + 1$ and $\mu^2 + \tau^2 > \sigma^2$, then*

$$\lim_{d \to \infty} \Pr(\mathcal{I}_0 \cap \mathcal{J} = \emptyset) = 1.$$

The condition $\mu^2 + \tau^2 > \sigma^2$ characterizes the high-dimensional asymptotic behavior of an outlier. In particular, a location and scale outliers can be characterized by $\mu^2 > 0$ and $\tau^2 > \sigma^2$, respectively. Clustered location outliers can be represented as $\sigma^2 > \tau^2$ and $\mu^2 > 0$. In this case, the location of outliers has to be at least a certain distance apart from the

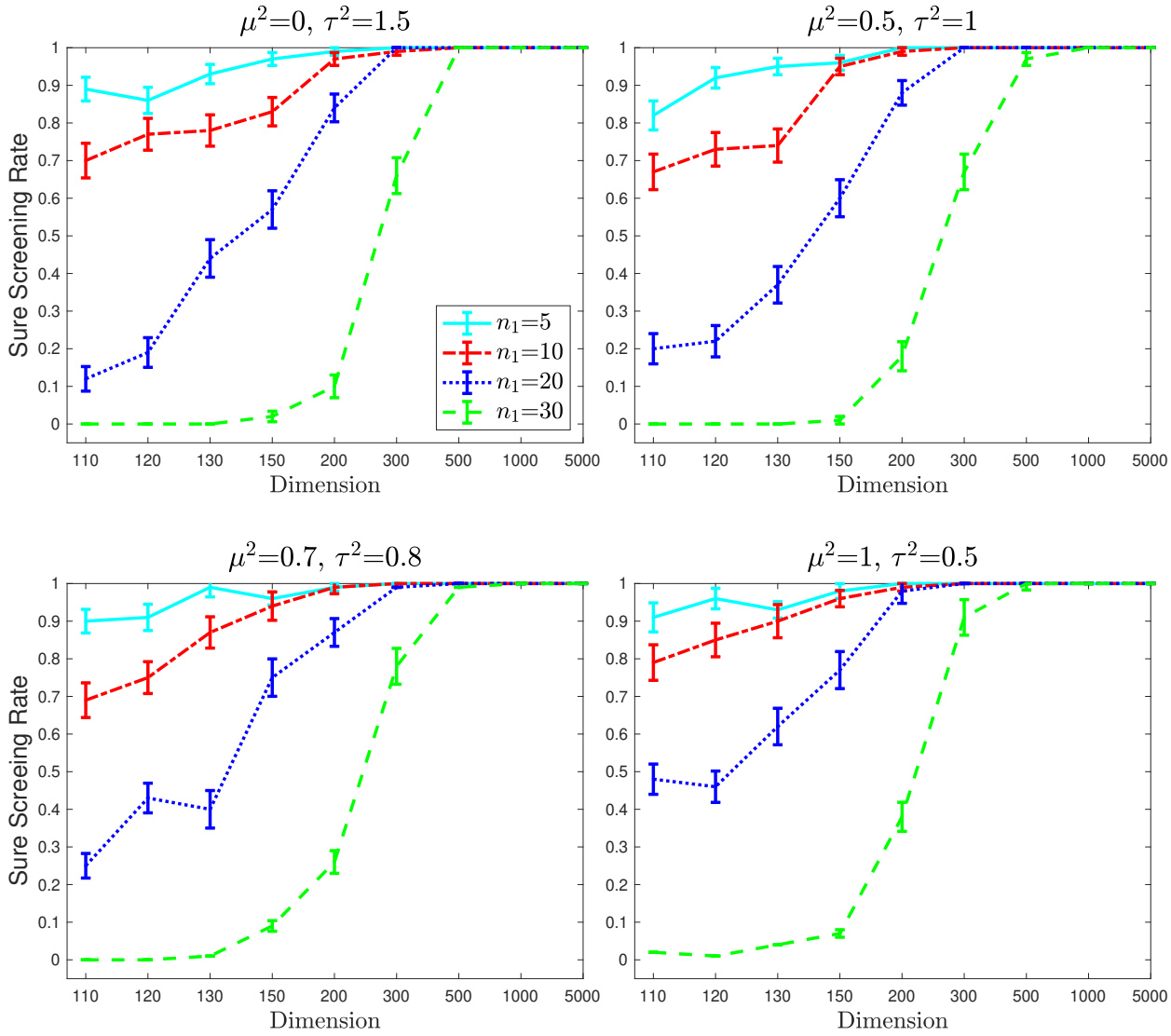


Figure 4.5.1: Average sure screening rates with standard error bars based on 100 replications. Data are independently generated from the standard multivariate normal with $N = 100$, i.e., $\sigma^2 = 1$. We consider $d = 110, \ldots, 5000$ and $n_1 = 5, 10, 20, 30$ outliers by shifting and scaling the standard normal vectors by μ and τ. For each setting, we fix $S = \tilde{n}_1 = 30$.

non-outliers so that $\mu^2 + \tau^2 > \sigma^2$. We assume that $\mu^2 + \tau^2 > \sigma^2$ and the number of outliers is strictly less than the number of non-outliers such that $n_0 > n_1 + 1$.

After obtaining $\mathcal{I}_0$, we set $\mathcal{I}$ by selecting $\tilde{n}_1$ observations that have the largest DH values to the affine hyperplane generated by $\mathcal{I}_0$. Thus, a sufficient condition for successful screening

is that outliers have larger DH values than non-outliers. As $d \to \infty$, the sufficient condition is satisfied with probability converging to one, and thus, combining with Lemma 5, we have the following theorem.

Theorem 14. *Under Condition 1, if* $n_0 > n_1 + 1$ *and* $\mu^2 + \tau^2 > \sigma^2$, *then*

$$\lim_{d\to\infty} \Pr\left(\mathcal{J} \subset \mathcal{I}\right) = 1.$$

We call the event $\mathcal{J} \subset \mathcal{I}$ sure screening and demonstrate the sure screening scenarios under various settings in Figure 4.5.1. Data are independently generated from the standard multivariate normal with $N = 100$, and outliers are obtained by shifting and scaling standard normal vectors by μ and τ, respectively. For each setting, we fix $\tilde{n}_1 = S = \lfloor 0.3N \rceil$ and consider $n_1 = 5, 10, 20, 30$. The y-axis represents the proportion of the sure screenings out of 100 replications. The top panels correspond to the cases when scale and location outliers are present, respectively. The top left panel represents the case of the *scale* outliers in that the location parameter is the same as non-outliers, while they have the larger variability. The top right panel illustrates the case where outliers have the same variability as the non-outliers but different location. It can be seen that even under the high contamination, $n_1 = 30$, the perfect sure screening is achieved under a moderately large dimension, $d > 500$. The two bottom panels display the cases when location outliers are clustered by having less variability than non-outliers. These settings are intended to mimic the situations where the masking effects are likely to occur. The bottom two panels in Figure 4.5.1 support the claim that the proposed screening method is robust to the possible masking effects.

4.5.2 Asymptotic power of SR tests for outlier detection

In this section, we investigate the high-dimensional asymptotic power of the SR test. Recall that in Section 4.4.3 we denote $[\mathbf{X}_0^{\mathrm{T}}, \mathbf{x}_j^*]^{\mathrm{T}}$ by $\mathbf{Y}_j$, and the test statistic is given by $t(\mathbf{Y}_j) = DH(\mathbf{x}_j^*|\mathbf{X}_0)$. For notational convenience, we drop the candidate outlier index j. The null hypothesis $H_0 : \mathbb{P} \in LS_{n,d}(J_{n,1})$, i.e., $\mathbf{x}^*$ is not an outlier so that $\mathbf{Y}$ is from $LS_{n,d}(J_{n,1})$, implies $\mu^2 = 0$ and $\tau^2 = \sigma^2$ in Condition 1.

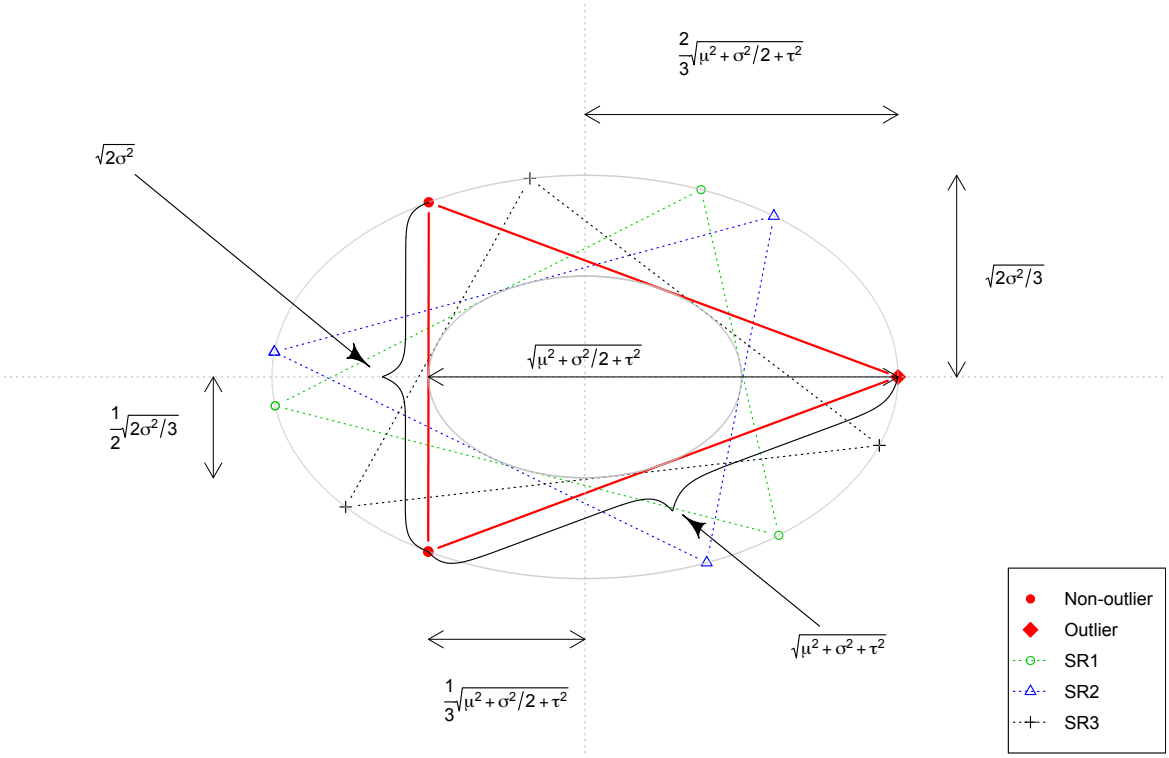


Figure 4.5.2: Illustration of the asymptotic geometry of HDLSS data and three randomly rotated data sets when $n_0 = 2$ and $n_1 = 1$. Original (rotated) data points are connected with solid (dotted) lines.

Using the HDLSS asymptotic argument, we can show that after proper scaling, the test statistic converges to a deterministic value under respective hypothesis, as the following lemma shows.

Lemma 6. *Under Condition 1, as* $d \to \infty$,

$$d^{-1/2}t(\mathbf{Y}) \xrightarrow{p} \begin{cases} \sqrt{(\tilde{n}_0+1)\sigma^2/\tilde{n}_0} & \text{under } H_0, \\ \sqrt{\mu^2+\sigma^2/\tilde{n}_0+\tau^2} & \text{under } H_a. \end{cases}$$

In the next theorem, we show that the power of the proposed SR test converges to one as the dimension tends to infinity.

Theorem 15. *Let $\vartheta_t(\mathbf{Y})$ be the p-value defined in (4.21), with N and $\mathbf{X}$ replaced by $n = \tilde{n}_0+1$ and $\mathbf{Y}$, respectively. Under Condition 1, if $n_0 > n_1 + 1$ and $\mu^2 + \tau^2 > \sigma^2$, then, as $d \to \infty$,*

$$\vartheta_t(\mathbf{Y}) \xrightarrow{p} \begin{cases} 1 & \text{under } H_0, \\ 0 & \text{under } H_a. \end{cases}$$

Figure 4.5.2 displays the geometry of three ultra-high-dimensional ($d = 50,000$) data points, one of which is an outlier. The vertical red line segment represents the simplex of $\mathbf{X}_0$. The three random vectors approximately form an isosceles triangle, with the edge lengths given in (4.24) and (4.25) . Dotted triangles represent three randomly selected, rotated data sets that are generated for the SR test. We can see that the rotated data and the original data have a common set of Steiner inellipse and circumellipse. All triangles in the figure have the same area, which is generalized by the statement that SR preserves the volume of the data, or yields the same sample second moments as the original data. The lengths of axes of the circumellipse are $\sqrt{n_0/N}$ times the square roots of eigenvalues of the sample covariance matrix of the data, where the inellipse has axes of a half length of the circumellipse. In the context of this figure, the proposed SR test would compare the maximum height of the red, solid triangle with the maximum heights of the subspace rotated, dotted triangles. Theorem 15 states that when d increases, the red triangle is likely to have the longest height among all triangles.

We note that HDLSS asymptotic scenario assumes that d is diverging with fixed N. The scenario where both d and N are diverging can be considered by progressively taking the limits, and this resembles the asymptotic scenario $d/n \to \infty$ (Jung et al., 2018; Lee et al., 2014). Under this scenario, after scaled by $\sqrt{d}$, the test statistic given in Lemma 6 converges to $\sqrt{\sigma^2}$ and $\sqrt{\mu^2 + \tau^2}$ in probability under H_0 and H_a, respectively. Thus the result of Theorem 7 will still hold.

We evaluate the performance of the proposed procedure with various data settings by comparing with existing methods. These include the Comedian (COM) by Sajesh and Srinivasan (2012), the distance-based method (DSO) by Ahn et al. (2019), the PCout (PCO) by Filzmoser et al. (2008) and MDP by Ro et al. (2015). We set $\tilde{n}_1 = S = \lfloor 0.3N \rceil$, $K = 500$ and $\alpha = 0.05$ for SR tests. For the competing methods, we use the default settings suggested by the authors.

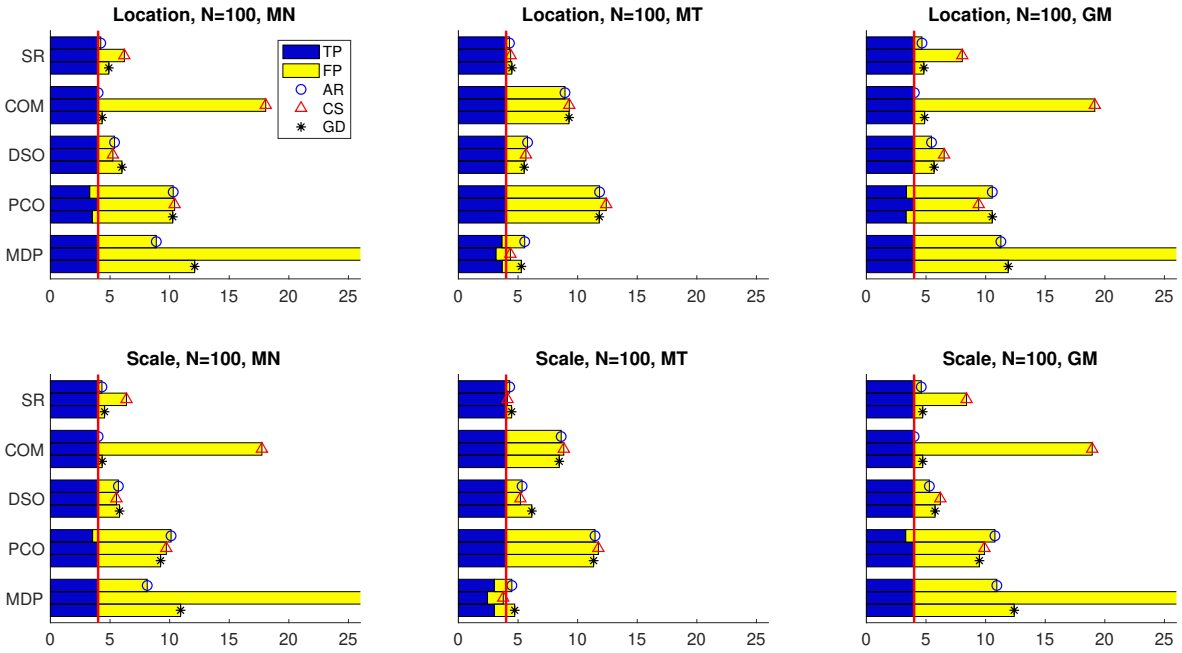
We generated $d = 1,000$ dimensional data with sample size $N = 50$ or $N = 100$ from the following distributions: matrix normal (denoted as MN), matrix t with four degrees of freedom (MT), and a generalized gamma distribution with skewness one and excess kurtosis 1.5 (GM). Note that the GM distributional setting violates the left-spherical assumption. Each distribution has mean zero and the following covariance structures:

- Auto-Regressive (AR): $\Sigma = \{0.8^{|l-l'|}\}_{l,l'}$, where $1 \leq l, l' \leq d$.
- Compound Symmetry (CS): $\Sigma = 0.7I_d + 0.3J_{d,d}$.
- Geometric Decay (GD): $\Sigma = \Gamma\Delta\Gamma^{\mathrm{T}}$, where Γ is generated from $\mathrm{Unif}(\mathcal{O}_d)$ and Δ is a diagonal matrix with geometrically decaying eigenvalues. Specifically,

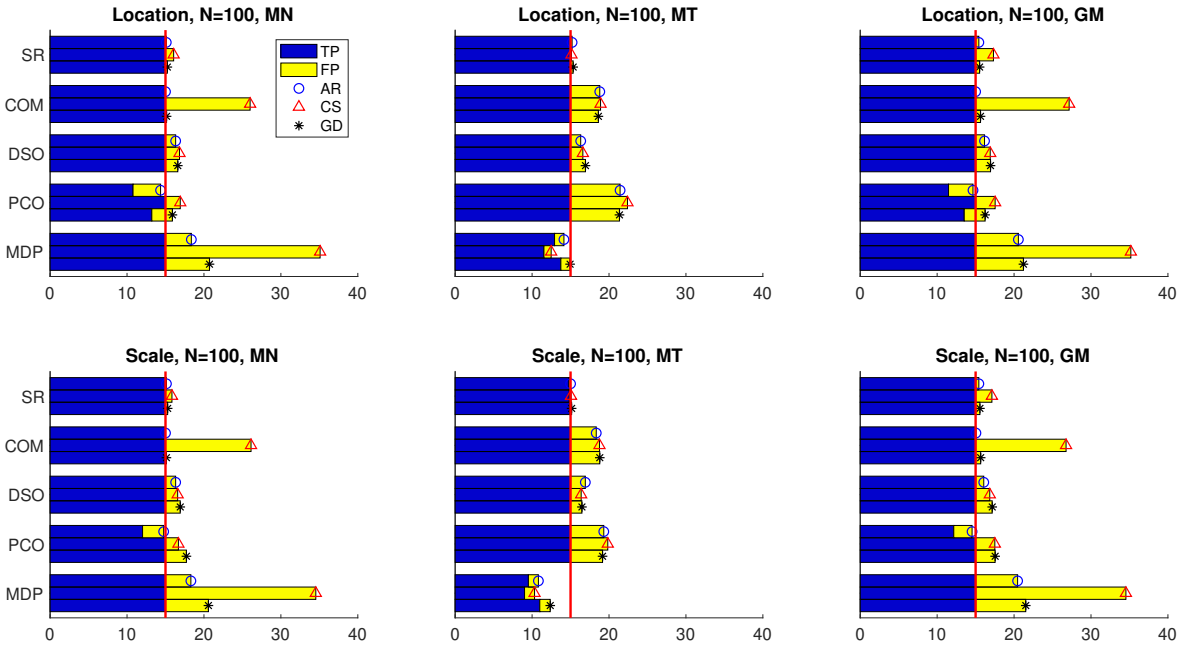
$$\delta_l = \frac{d(.9^{l-1} - .9^l)}{1 - .9^d}, \; l = 1, \ldots, d.$$

The variables in the AR setting are strongly correlated with only a couple of variables, while nearly uncorrelated with the rest. The CS setting assumes that variables are moderately positively correlated with each other. The GD setting is intended to mimic a covariance structure that has been empirically observed in real high-throughput data.

We consider two scenarios regarding the source of abnormality, location or scale, and also two levels of contamination, low or high. Location outliers are generated from $\mathcal{N}(\sqrt{d}\Theta, I_{n_1}, \Sigma)$, where $\Theta = [\theta_1, \cdots, \theta_{n_1}]^{\mathrm{T}}$ and θ_j are independently distributed as $\mathrm{Unif}(\mathcal{V}_{1,d})$,



(a) Low contamination ($n_1 = 4$)



(b) High contamination ($n_1 = 15$)

Figure 4.6.1: The average numbers of detected outliers with simulated data based on 50 replications. The dimension $d = 1000$, the sample size $N = 100$ and the number of outliers, $n_1 = 4, 15$. The matrix normal, matrix t_4 and generalized gamma distributions are denoted by MN, MT and GM. The blue (yellow) bar indicates the average number of true (false) positives, and circle, triangle and asterisk at the end of each bar corresponds to the auto-regressive (AR), compound-symmetry (CS) and geometric-decaying (GD) covariance settings.

$j = 1, \ldots, n_1$. Scale outliers are from the same distribution family as non-outliers, but with $\mathsf{Var}(\mathbf{x}_j^o) = 2I_d$, $j = 1, \ldots, n_1$.

Simulation results averaged over 50 replications for $N = 100$ are shown in Figure 4.6.1. The results with $N = 50$ can be found in Figure 4.11.1 in Section 4.11. Each horizontal bar represents the average number of declared outliers by each method. The blue (dark color) part of a bar is for true positives (TP), true outliers detected, and the yellow (light color) is for false positives (FP), falsely declared outliers. The vertical reference line is drawn at the true number of outliers n_1. It is immediately clear that the proposed method performs better than the compared methods in most of the settings. The SR method achieves near perfect TP, while having the minimum level of FP in all settings, except for a few cases with the CS covariance, under which Condition 1-(e) for the asymptotic study in the previous section is violated. The COM and MDP methods particularly suffer under this covariance setting. It is also noticeable that the effect of the CS covariance is relatively non-existent under the t distribution.

The right panels of the figure showing the result under the gamma distribution indicate that the proposed method is robust to the left-spherical assumption. The DSO method shows consistently good performance, achieving perfect TP as well in all cases. However it is overall worse than the proposed SR method in terms of FP. The PCO method also successfully detects all outliers except for a few cases, but its FP is mostly higher than other methods especially in the low contamination case. The MDP method fails to detect the true outliers under the t_4 distribution, and is the worst in overall performance.

To examine how different choices of $\tilde{n}_1$ change the performance of the proposed method, we re-run the simulations with various choices of $\tilde{n}_1$ under the same settings. Specifically, we consider $\eta = 0.1, 0.2, 0.3, 0.45$, where $\tilde{n}_1 = \lfloor \eta N \rceil$. We find that the SR method is robust to the choice of $\tilde{n}_1$ in terms of TP and FP. In term of TP, the proposed SR method achieves near perfect TP for all $\tilde{n}_1$. On the other hand, FP increases as $\tilde{n}_1$ increases but our method

		N=50			N=100		
		AR	CS	GD	AR	CS	GD
MN	SR	**0.44**	2.48	**0.50**	**0.54**	2.74	**0.84**
		(0.10)	(0.27)	**(0.10)**	**(0.12)**	(0.30)	**(0.15)**
	COM	1.00	7.62	1.26	1.00	14.52	1.12
		(0.00)	(0.38)	(0.07)	(0.00)	(0.49)	(0.05)
	DSO	1.48	**1.72**	1.66	1.62	**1.70**	2.40
		(0.15)	**(0.25)**	(0.14)	(0.16)	**(0.19)**	(0.31)
	PCO	4.84	4.66	4.80	7.94	7.40	8.14
		(0.42)	(0.35)	(0.37)	(0.53)	(0.47)	(0.47)
	MDP	3.30	18.96	5.40	5.76	34.06	7.78
		(0.25)	(0.43)	(0.39)	(0.39)	(0.78)	(0.42)
MT	SR	**0.54**	**0.44**	**0.54**	**0.44**	**0.52**	**0.44**
		(0.11)	**(0.11)**	**(0.12)**	**(0.08)**	**(0.12)**	**(0.11)**
	COM	6.06	6.80	6.40	5.56	5.98	5.54
		(0.31)	(0.39)	(0.33)	(0.37)	(0.32)	(0.31)
	DSO	1.70	1.64	1.60	2.20	2.36	2.58
		(0.21)	(0.17)	(0.14)	(0.24)	(0.29)	(0.34)
	PCO	5.36	5.90	5.70	10.58	9.70	10.02
		(0.35)	(0.42)	(0.41)	(0.61)	(0.57)	(0.50)
	MDP	1.68	1.32	1.72	2.18	1.68	2.08
		(0.28)	(0.22)	(0.28)	(0.27)	(0.24)	(0.30)
GM	SR	**0.68**	3.36	**0.78**	**0.88**	5.22	**1.30**
		(0.12)	(0.30)	**(0.13)**	**(0.14)**	(0.38)	**(0.17)**
	COM	1.00	8.58	1.60	1.00	16.62	1.60
		(0.00)	(0.33)	(0.11)	(0.00)	(0.47)	(0.12)
	DSO	1.40	**1.96**	1.88	1.88	**2.70**	2.14
		(0.16)	**(0.22)**	(0.22)	(0.23)	**(0.41)**	(0.22)
	PCO	4.98	5.12	5.62	7.34	7.94	8.06
		(0.38)	(0.38)	(0.39)	(0.42)	(0.44)	(0.46)
	MDP	5.26	20.30	6.64	8.78	36.04	9.98
		(0.39)	(0.33)	(0.48)	(0.47)	(0.86)	(0.54)

Table 4.1: Average numbers of false identifications and standard errors based on 50 replications under no-outlier settings. The matrix normal, matrix t_4 and generalized gamma distributions are denoted by MN, MT and GM, for each of which one of auto-regressive (AR), compound-symmetry (CS) and geometric-decaying (GD) covariance setting is applied. The numbers in boldface represent the smallest average number of false identifications under each setting.

outperforms existing methods even in the worst case $\eta = 0.45$. The results are summarized in Figure 4.11.2 and Figure 4.11.3 in Section 4.11.

One of the findings in the results in Figure 4.6.1 is that most methods are able to detect most of the true outliers, and an important criterion for an effective outlier detection method is the ability to not false declare non-outliers as an outlier. Hence we conduct a simulation under the null settings, i.e., the same distributional settings as above but with no contamination. Table 4.1 lists the average numbers of the observations that are falsely declared as outliers based on 50 replications. The proposed SR method shows the lowest level of false detection in most settings, which implies that the type-I error of the SR test is well controlled below the nominal level $\alpha = 0.05$. Under the CS covariance setting, DSO has the smallest false detection when the population is MN or GM. A similar interpretation to Figure 4.6.1 on the overall comparative performances of the methods can be applied in this experiment too.

4.7 Outliers in human face image data

We applied the compared methods to the Olivetti Research Laboratory (ORL) face image data set (Samaria and Harter, 1994) to see how they perform on a real data problem. The ORL data consist of a total of 400 face images from 40 individuals so that each person has ten images showing various facial expressions and brightness. Each image contains 112×92 pixels in 0–255 grayscale that makes the dimension $d = 10,304$. For each individual, we add $n_1 = 3$ outliers by adding images of three randomly chosen different people in the data set. The proposed SR tests were conducted with $\tilde{n}_1 = S = \lfloor 0.3N \rceil$, $K = 1,000$ and $\alpha = 0.05$. MDP could not be implemented due to the high-dimensionality.

Figures 4.11.4–4.11.7 in Section 4.11 display all cases with outlier detection results from the compared methods. Each column shows ten images belonging to one individual, followed by three contaminating images at the bottom. The red text next to an image indicates a method that declares that image as an outlier. A summary of all 40 individuals is shown in

	SR	COM	DSO	PCO
TPR	0.917	0.850	0.308	0.700
FPR	0.017	0.138	0.008	0.153

Table 4.2: Average true positive rates (TPR) and false positive rates (FPR) for the ORL data containing 400 face images from 40 individuals (10 each). For each set of 10 images belonging to one person, three randomly selected images are added to contaminate the data ($N = 13, n_1 = 3$).

Table 4.2, in which average true positive rate (TPR) and false positive rate (FPR) are shown. The proposed method outperforms all the competing methods in that it has the largest true positive rate and the lowest false positive rate. Unlike the simulation results, DSO fails to detect outliers in many examples. The PCO and COM methods show reasonable TPR but have a relatively large FPR compared with the proposed method.

4.8 Conclusions

In this chapter, we proposed a new high-dimensional outlier detection procedure based on a randomization test. Numerical studies suggested that it has a number of advantages over existing approaches. Most importantly, the proposed SR method showed smaller false positives than the compared methods. It seems for HDLSS data; it is much more challenging not to falsely accuse a non-outlier than finding true outliers. This is a crucially practical advantage as it is usually costly to collect high-throughput data; thus, it is important to keep as many observations as possible for analysis.

Although in the present chapter, we have focused on outlier detection, the SR tests can be applied to various hypothesis testing problems in multivariate analysis. It would be a natural alternative of the permutation test in a variety of situations. There are a few future research directions to pursue. For instance, we can verify the existence of clusters in unsupervised

learning using the SR test, as random rotations can create the null data without clusters while inheriting the overall volume of the data.

4.9 Technical details

Proof of Theorem 9. Let $M_1 M_1^{\mathrm{T}} X = X_{M_1}$. Since, for any $V \in \mathcal{V}^1_{m_1,N}$, there exists an $O \in \mathcal{O}_{m_1}$ such that $V = M_1 O$, the differential of X_{M_1} is given by

$$dX_{M_1} = M_1 dO\Lambda\Gamma^{\mathrm{T}} + M_1 O d\Lambda\Gamma^{\mathrm{T}} + M_1 O\Lambda d\Gamma^{\mathrm{T}}.$$

Let $V_2 = [M_1 O, M_0] \in \mathcal{O}_N$ and $\Gamma_2 = [\Gamma, \Gamma_1] \in \mathcal{O}_d$. By pre- and post-multiplying V_2^{T} and Γ_2 to dX_{M_1}, we have

$$V_2^{\mathrm{T}} dX_{M_1} \Gamma_2 = \begin{bmatrix} O^{\mathrm{T}} dO\Lambda + d\Lambda + \Lambda d\Gamma^{\mathrm{T}}\Gamma & \Lambda d\Gamma^{\mathrm{T}}\Gamma_1 \\ 0_{d-m_1,m_1} & 0_{d-m_1,d-m_1} \end{bmatrix},$$

where $0_{a,b}$ is the $a \times b$ matrix of zeros. As similarly done in Theorem 5 of Uhlig et al. (1994),

$$(O^{\mathrm{T}} dO\Lambda + d\Lambda + \Lambda d\Gamma^{\mathrm{T}}\Gamma) = \prod_{i<j}^{m_1}(\lambda_i^2 - \lambda_j^2) \bigwedge_{i=1}^{m_1} d\lambda_i \bigwedge_{i=1}^{m_1}\bigwedge_{j=i+1}^{m_1} \underline{o}_j^{\mathrm{T}} d\underline{o}_i \bigwedge_{i=1}^{m_1}\bigwedge_{j=i+1}^{m_1} \underline{\gamma}_j^{\mathrm{T}} d\underline{\gamma}_i \tag{4.26}$$

and

$$(\Lambda d\Gamma^{\mathrm{T}}\Gamma_\perp) = \det(\Lambda)^{d-m_1} \bigwedge_{i=1}^{m_1}\bigwedge_{j=1}^{d-m_1} \underline{\gamma}_{m_1+j}^{\mathrm{T}} d\underline{\gamma}_i, \tag{4.27}$$

where $\Lambda = \mathrm{Diag}\{\lambda_i\}_{i=1}^{m_1}$, $\underline{o}_i$ and $\underline{\gamma}_i$ are the ith columns of O and Γ_2, respectively. Since $(V_2^{\mathrm{T}} dX_{M_1}\Gamma_2) = (dX_{M_1})$ by Theorem 2.1.5 (Muirhead, 2009), the exterior product of (4.26) and (4.27) gives

$$\begin{aligned}(dX_{M_1}) &= \det(\Lambda)^{d-m_1} \prod_{i<j}^{m_1}(\lambda_i^2 - \lambda_j^2) \bigwedge_{i=1}^{m_1} d\lambda_i \bigwedge_{i=1}^{m_1}\bigwedge_{j=i+1}^{m_1} \underline{o}_j^{\mathrm{T}} d\underline{o}_i \bigwedge_{i=1}^{m_1}\bigwedge_{j=1}^{d-m_1} \underline{\gamma}_{m_1+j}^{\mathrm{T}} d\underline{\gamma}_i \bigwedge_{i=1}^{m_1}\bigwedge_{j=i+1}^{m_1} \underline{\gamma}_j^{\mathrm{T}} d\underline{\gamma}_i \\ &= \det(\Lambda)^{d-m_1} \prod_{i<j}^{m_1}(\lambda_i^2 - \lambda_j^2)(dO) \bigwedge (d\Lambda) \bigwedge (d\Gamma),\end{aligned} \tag{4.28}$$

where (dO) is the differential form for the invariant measure on $\mathcal{O}_{m_1}$.

Independence is obtained as follows. By Theorem 6.1.4 (Muirhead, 2009), the density of $\mathbf{X}_{M_1}$ is proportional to $g(\Gamma^{\mathrm{T}}\Lambda^2\Gamma)$ for some g. Since $g(\Gamma^{\mathrm{T}}\Lambda^2\Gamma)$ is free of V (or O) and (dO) can be factored out from (4.28), the independence is achieved. □

Proof of Theorem 10. Recall that $R = M_0 M_0^{\mathrm{T}} + M_1 O M_1^{\mathrm{T}}$ for some $O \in \mathcal{O}_{m_1}$. Since $R^{\mathrm{T}} R = I_N$, $R^{\mathrm{T}} dR$ is a skew-symmetric matrix, where $R^{\mathrm{T}} dR = M_1 O^{\mathrm{T}} dO M_1^{\mathrm{T}}$. Let $V_2 = [M_1, M_0] \in \mathcal{O}_N$. By pre- and post-multiplying V_2^{T} and V_2 to $R^{\mathrm{T}} dR$, we have

$$V_2^{\mathrm{T}} R^{\mathrm{T}} dR V_2 = \begin{bmatrix} O^{\mathrm{T}} dO & 0_{m_1, m_0} \\ 0_{m_0, m_1} & 0_{m_0, m_0} \end{bmatrix},$$

where $m_0 + m_1 = N$. Since $(V_2^{\mathrm{T}} R^{\mathrm{T}} dR V_2)$ is identical to $(R^{\mathrm{T}} dR)$ by Theorem 2.1.5 (Muirhead, 2009), $(R^{\mathrm{T}} dR) = (O^{\mathrm{T}} dO)$. □

Proof of Lemma 4. Since $\mathbf{R} = M_0 M_0^{\mathrm{T}} + M_1 \mathbf{O} M_1^{\mathrm{T}}$ for $\mathbf{O} \sim \mathrm{Unif}(O_{m_1})$, we have

$$\begin{aligned}
\Psi_{\mathbf{R}}(W_0) &= \mathsf{E}\{\mathrm{etr}(\iota W_0^{\mathrm{T}} \mathbf{R})\} \\
&= \int_{\mathcal{O}_{m_1}} \mathrm{etr}\{\iota W_0^{\mathrm{T}} (M_0 M_0^{\mathrm{T}} + M_1 O M_1^{\mathrm{T}})\} [dO] \\
&= \int_{\mathcal{O}_{m_1}} \mathrm{etr}(\iota W_0^{\mathrm{T}} M_0 M_0^{\mathrm{T}}) \, \mathrm{etr}(\iota W_0^{\mathrm{T}} M_1 O M_1^{\mathrm{T}})\} [dO] \\
&= \mathrm{etr}(\iota W_0^{\mathrm{T}} M_0 M_0^{\mathrm{T}}) \int_{\mathcal{O}_{m_1}} \mathrm{etr}(\iota M_1^{\mathrm{T}} W_0^{\mathrm{T}} M_1 O)\} [dO] \\
&= \mathrm{etr}(\iota W_0^{\mathrm{T}} M_0 M_0^{\mathrm{T}}) \, {}_0F_1 \left(\frac{m_1}{2}; -\frac{1}{4} M_1^{\mathrm{T}} W_0^{\mathrm{T}} M_1 M_1^{\mathrm{T}} W_0^{\mathrm{T}} M_1 \right).
\end{aligned}$$

Also, since we can write $\mathbf{V} = M_1 \mathbf{O}$, we have

$$\begin{aligned}
\Psi_{\mathbf{V}}(W_1) &= \mathsf{E}\{\mathrm{etr}(\iota W_1^{\mathrm{T}} \mathbf{V})\} \\
&= \int_{\mathcal{O}_{m_1}} \mathrm{etr}(\iota W_1^{\mathrm{T}} M_1 O) [dO] \\
&= {}_0F_1 \left(\frac{m_1}{2}; -\frac{1}{4} W_1^{\mathrm{T}} M_1 M_1^{\mathrm{T}} W_1 \right).
\end{aligned}$$

□

Proof of Theorem 11. The result follows if we show that, for independently distributed $\mathbf{V} \sim \mathrm{Unif}(\mathcal{V}^1_{m_1, N})$, $\mathbf{RV}$ and $\mathbf{R}$ are identically distributed. We write $\mathbf{R} = M_0 M_0^{\mathrm{T}} + M_1 \mathbf{O}_0 M_1^{\mathrm{T}}$ and $\mathbf{V} = M_1 \mathbf{O}_1$, where $\mathbf{O}_0$ and $\mathbf{O}_1$ are independent and identically distributed as $\mathrm{Unif}(\mathcal{O}_{m_1})$.

For $W = [W_0^{\mathrm{T}}, W_1^{\mathrm{T}}]^{\mathrm{T}}$, the joint characteristic function of $\mathbf{R}$ and $\mathbf{Y} = \mathbf{RV}$ is given by

$$\begin{aligned}
\Psi_{\mathbf{R},\mathbf{Y}}(W) &= \mathsf{E}[\mathrm{etr}\{\iota(W_0^{\mathrm{T}}\mathbf{R} + W_1^{\mathrm{T}}\mathbf{Y})\}] \\
&= \int_{\mathcal{O}_{m_1}} \int_{\mathcal{O}_{m_1}} \mathrm{etr}(\iota W_0^{\mathrm{T}} R)\, \mathrm{etr}(\iota W_1^{\mathrm{T}} M_1 O_0 O_1)[dO_1][dO_0] \\
&= \int_{\mathcal{O}_{m_1}} {}_0F_1\left(\frac{m_1}{2}; -\frac{1}{4} W_1^{\mathrm{T}} M_1 O_0 O_0^{\mathrm{T}} M_1^{\mathrm{T}} W_1\right) \mathrm{etr}(\iota W_0^{\mathrm{T}} R)[dO_0] \\
&= {}_0F_1\left(\frac{m_1}{2}; -\frac{1}{4} W_1^{\mathrm{T}} M_1 M_1^{\mathrm{T}} W_1\right) \int_{\mathcal{O}_{m_1}} \mathrm{etr}(\iota W_0^{\mathrm{T}} R)[dO_0] \\
&= {}_0F_1\left(\frac{m_1}{2}; -\frac{1}{4} W_1^{\mathrm{T}} M_1 M_1^{\mathrm{T}} W_1\right) \mathrm{etr}(\iota W_0^{\mathrm{T}} M_0 M_0^{\mathrm{T}})\, {}_0F_1\left(\frac{m_1}{2}; -\frac{1}{4} M_1^{\mathrm{T}} W_0^{\mathrm{T}} M_1 M_1^{\mathrm{T}} W_0^{\mathrm{T}} M_1\right).
\end{aligned}$$

Combining this and Lemma 4 indicates that $\mathbf{RV}$ and $\mathbf{V}$ are identically distributed, and $\mathbf{RV}$ and $\mathbf{R}$ are independent. □

Proof of Theorem 12. Let $f_{\mathbf{X}}(X)$ be the density function of $\mathbf{X}$ and (dX) be the Lebesgue measure for $X \in \mathbb{R}_{N,d}$ as in (4.6). We denote $f_{\mathbf{X}}(X)(dX)$ by $[dX]$. Then the unbiasedness of the estimator in (4.18) can be seen as

$$\begin{aligned}
\mathsf{E}_{\mathbf{X}}\{F_{t|\mathbf{X}}(z|\mathbf{X})\} &= \int_{\mathcal{X}} \int_{\mathcal{R}(M_0)} 1\{t(RX) \le z\}[dR][dX] \\
&= \int_{\mathcal{R}(M_0)} \int_{\mathcal{X}} 1\{t(RX) \le z\}[dX][dR] \\
&= \int_{\mathcal{R}(M_0)} \int_{\mathcal{X}} 1\{t(X) \le z\}[dX][dR], \quad \text{since } \mathbf{X} \overset{d}{=} \mathbf{RX}, \\
&= \int_{\mathcal{R}(M_0)} F_t(z)[dR] = F_t(z).
\end{aligned}$$

□

Proof of Theorem 13. Consider the expectation of $\phi(\mathbf{RX})$ given $\mathbf{X} = X$. By (4.18) and (4.20), we have

$$\mathsf{E}_{\mathbf{R}}\{\phi(\mathbf{R}X)|X\} = \int_{\mathcal{R}(M_0)} \phi(RX)[dR] = \int_{\mathcal{R}(M_0)} 1\{t(RX) > c_\alpha(X)\}[dR] \le \alpha.$$

Note that the last equality holds when $t(\cdot)$ is continuous. We also have

$$\mathsf{E}_{\mathbf{X}}\{\phi(\mathbf{X})\} = \mathsf{E}_{\mathbf{X}}\{\phi(\mathbf{RX})\} = \mathsf{E}_{\mathbf{R}}[\mathsf{E}_{\mathbf{X}}\{\phi(\mathbf{RX})\}] = \mathsf{E}_{\mathbf{X}}[\mathsf{E}_{\mathbf{R}}\{\phi(\mathbf{RX})\}].$$

The first and second equalities hold by the invariance and independence shown in Theorem 11. By the definitions of the SR distribution (4.18) and $c_\alpha(X)$ (4.19), we have

$$c_\alpha(X) = c_\alpha(RX)$$

for all $R \in \mathcal{R}_{M_0}$. Since $\mathbf{R}$ and $\mathbf{X}$ are independent, $\mathsf{E}_{\mathbf{X}}[\mathsf{E}_{\mathbf{R}}\{\phi(\mathbf{RX})\}] = \mathsf{E}_{\mathbf{X}}[\mathsf{E}_{\mathbf{R}}\{\phi(\mathbf{RX})|\mathbf{X}\}] \leq \alpha$. The equality holds, provided that $t(\cdot)$ is continuous.

□

Proof of Lemma 5. Let med_i and med_j^o be the medians of the pairwise distances of $\mathbf{x}_i$ and $\mathbf{x}_j^o$, respectively. Under the given assumption $n_0 > n_1 + 1$, the asymptotic geometric representation given in Section 4.5.1 implies that

$$\text{med}_j^2 - \text{med}_i^{o2} = d(\mu^2 + \tau^2 - \sigma^2) + o_p(d).$$

The condition $\mu^2 + \tau^2 \geq \sigma^2$ directly gives us that

$$\lim_{d\to\infty} \Pr(\text{med}_j^o < \text{med}_i) = 0.$$

Since $\Pr(\mathcal{I}_0 \cap \mathcal{J} \neq \emptyset) \leq \Pr(\text{med}_j^o < \text{med}_i)$, we have the desired result. □

The following lemma will be needed to prove Theorem 14 and Lemma 6.

Lemma 7. *Let $(I_n - n^{-1}J_{n,n})\mathbf{Y} = \mathbf{V}\mathbf{\Lambda}\mathbf{\Gamma}^T$ be the nonsingular part of the singular value decomposition. Under Condition 1, as $d \to \infty$,*

$$d^{-1}\mathbf{V}\mathbf{\Lambda}^2\mathbf{V}^T \xrightarrow{p} UD^2U^T,$$

where $U = [\underline{u}_1, \ldots, \underline{u}_{\tilde{n}_0}]$, $D^2 = \text{diag}\{\delta_i^2\}_{i=1}^{\tilde{n}_0}$,

$$\underline{u}_1 = \left(\frac{\tilde{n}_0}{n}\right)^{1/2} [\tilde{n}_0^{-1} J_{\tilde{n}_0,1}^T, -1]^T \quad \textit{and}$$

$$\delta_1^2 = \sqrt{\frac{\tilde{n}_0}{n}\left(\mu^2 + \frac{\sigma^2}{\tilde{n}_0} + \tau^2\right)}.$$

Proof of Lemma 7. Condition 1 gives us that, by the weak law of large numbers, as $d \to \infty$,

$$d^{-1}\mathbf{V}\mathbf{\Lambda}^2\mathbf{V}^{\mathrm{T}} \xrightarrow{p} \begin{bmatrix} \sigma^2 Q_{\tilde{n}_0} + (\frac{\tilde{n}_0}{n})^2\mu_0^2 J_{\tilde{n}_0,\tilde{n}_0} & -\frac{\tilde{n}_0}{n^2}\mu_0^2 J_{\tilde{n}_0,1} \\ -\frac{\tilde{n}_0}{n^2}\mu_0^2 J_{1,\tilde{n}_0} & (\frac{\tilde{n}_0}{n})^2\mu_0^2 \end{bmatrix} \equiv \Omega,$$

where

$$Q_{\tilde{n}_0} = I_{\tilde{n}_0} - \frac{1}{\tilde{n}_0} J_{\tilde{n}_0,\tilde{n}_0}, \quad \text{and} \quad \mu_0^2 = \mu^2 + \frac{\sigma^2}{\tilde{n}_0} + \tau^2.$$

Let $\underline{u}_1 = \sqrt{\tilde{n}_0/n}[\tilde{n}_0^{-1} J_{\tilde{n}_0,1}^{\mathrm{T}}, -1]^{\mathrm{T}}$. Then,

$$\Omega\underline{u}_1 = \frac{\tilde{n}_0}{n}\mu_0^2\ \underline{u}_1 \equiv \delta_1^2 \underline{u}_1.$$

Thus, $\underline{u}_1$ is an eigenvector corresponding to δ_1^2. Without loss of generality, let $\underline{e}_i$ be the ith eigenvector of $Q_{\tilde{n}_0}$. Then $\underline{u}_i = [\underline{e}_{i-1}^{\mathrm{T}}, 0]^{\mathrm{T}}$, $i = 2, \ldots, \tilde{n}_0$, are eigenvectors of Ω corresponding to the common eigenvalue σ^2.

We now show that $\underline{u}_1$ is the eigenvector corresponding to the largest eigenvalue, under the condition $\mu^2 + \tau^2 > \sigma^2$. This can be seen as

$$\begin{aligned} \frac{\tilde{n}_0}{n}\mu^2 + \frac{\tilde{n}_0}{n}\tau^2 &> \frac{\tilde{n}_0}{n}\sigma^2 \\ \frac{\tilde{n}_0}{n}\mu^2 + \frac{\tilde{n}_0}{n}\tau^2 + \frac{1}{n}\sigma^2 &> \sigma^2 \\ \frac{\tilde{n}_0}{n}\left(\mu^2 + \frac{\sigma^2}{\tilde{n}_0} + \tau^2\right) &> \sigma^2. \end{aligned}$$

Thus, the δ_1^2 is the distinct largest eigenvalue and $\underline{u}_1$ is unique up to constant. □

Proof of Lemma 6. Let $\underline{\ell} = [J_{\tilde{n}_0,1}^{\mathrm{T}}, -1]^{\mathrm{T}}$ be an n-dimensional label vector for $\mathbf{x}^*$. By Ahn and Marron (2010), we can express $d^{-1}t(\mathbf{Y})^2$ by

$$d^{-1}t(\mathbf{Y})^2 = \frac{4}{d\underline{\ell}^{\mathrm{T}}\mathbf{V}\mathbf{\Lambda}^{-2}\mathbf{V}^{\mathrm{T}}\underline{\ell}}.$$

By Lemma 7, we have

$$d^{-1}t(\mathbf{Y})^2 \xrightarrow{p} \frac{4}{\underline{\ell}^{\mathrm{T}}\Omega^{+}\underline{\ell}} \quad \text{as} \quad d \to \infty, \tag{4.29}$$

where A^+ is the Moore-Penrose generalized inverse of matrix A. Note that the label vector $\underline{\ell}$ can be expressed as a linear combination of $\underline{u}_1$ and $\underline{u}_n$ such that $\underline{\ell} = a_1\underline{u}_1 + a_n\underline{u}_n$, where

$$\underline{u}_n = n^{-1/2}J_{n,1}, \quad a_1 = 2\sqrt{\frac{\tilde{n}_0}{n}}, \quad \text{and} \quad a_n = \frac{\tilde{n}_0 - 1}{\sqrt{n}}. \tag{4.30}$$

Thus, the denominator of the limit in (4.29) is $\underline{\ell}^{\mathrm{T}}\Omega^+\underline{\ell} = \delta_1^{-2}a_1^2 = 4\mu_0^{-2}$, and we have

$$\frac{4}{\underline{\ell}^{\mathrm{T}}\Omega^+\underline{\ell}} = \mu_0^2 = \mu^2 + \frac{\sigma^2}{\tilde{n}_0} + \tau^2.$$

Under the null hypothesis $H_0 : \mathbb{P} \in LS_{n,d}$, we have $\mu^2 = 0$ and $\sigma^2 = \tau^2$. Thus,

$$\frac{4}{\underline{\ell}^{\mathrm{T}}\Omega^+\underline{\ell}} = \frac{\sigma^2}{\tilde{n}_0} + \sigma^2.$$

□

Proof of Theorem 14. We will prove the theorem by showing that $\Pr(\mathcal{J} \not\subset \mathcal{I})$ converges to 0 as $d \to \infty$. We first rewrite $\Pr(\mathcal{J} \not\subset \mathcal{I})$ as

$$\Pr\left(\mathcal{J} \not\subset \mathcal{I}\right) = \Pr\left(\mathcal{J} \not\subset \mathcal{I} | \mathcal{J} \cap \mathcal{I}_0 \neq \emptyset\right) \Pr\left(\mathcal{J} \cap \mathcal{I}_0 \neq \emptyset\right) \tag{4.31}$$

$$+ \Pr\left(\mathcal{J} \not\subset \mathcal{I} | \mathcal{J} \cap \mathcal{I}_0 = \emptyset\right) \Pr\left(\mathcal{J} \cap \mathcal{I}_0 = \emptyset\right). \tag{4.32}$$

For the right hand side of (4.31), the conditional probability is one as $\mathcal{J} \cap \mathcal{I}_0 = \emptyset$ is a necessary condition for the event $\mathcal{J} \subset \mathcal{I}$. Let us denote $\mathcal{E}$ by the event

$$\{DH(\mathbf{x}_j^o|\mathbf{Z}_0) \leq DH(\mathbf{x}_i|\mathbf{Z}_0), \text{ for some } \mathbf{x}_i \text{ and } \mathbf{x}_j^o\}.$$

Then the conditional probability in (4.32) is bounded above by $\Pr\{\mathcal{E}|\mathcal{J} \cap \mathcal{I}_0 = \emptyset\}$. Hence, we have

$$\Pr\left(\mathcal{J} \not\subset \mathcal{I}\right) \leq \Pr\left(\mathcal{J} \cap \mathcal{I}_0 \neq \emptyset\right) + \Pr\left(\mathcal{E} | \mathcal{J} \cap \mathcal{I}_0 = \emptyset\right). \tag{4.33}$$

The proof is then complete since the first and the second terms in the right side of (4.33) converge to 0 as d increases by Lemmas 6 and 5. Therefore, we have

$$\lim_{d\to\infty} \Pr(\mathcal{J} \not\subset \mathcal{I}) = 0.$$

□

Proof of Theorem 15 . Let $\mathbf{V}\mathbf{\Lambda}\mathbf{\Gamma}^{\mathrm{T}}$ be the singular value decomposition of $(I_n - n^{-1}J_{n,n})\mathbf{Y}$ in Lemma 7. The p-value in (4.21) is

$$\begin{aligned}\vartheta_t(\mathbf{Y}) &= \mathsf{E}_{\mathbf{R}}[1\{t(\mathbf{Y}) \le t(\mathbf{R}\mathbf{Y})\}|\mathbf{Y}] \\ &= \Pr\{t(\mathbf{Y}) \le t(\mathbf{R}\mathbf{Y})|\mathbf{Y}\}.\end{aligned}$$

Let $\underline{\ell}_{i'} = (\ell_{i',1}, \ldots, \ell_{i',n})^{\mathrm{T}}$ be the label vector such that $\ell_{i',i} = -1$ if $i' = i$, and $\ell_{i',i} = 1$, otherwise. Then,

$$t(\mathbf{R}\mathbf{Y}) = \max_{i'} \frac{4}{d\underline{\ell}_{i'}^{\mathrm{T}}\mathbf{V}\mathbf{O}\mathbf{\Lambda}^{-2}\mathbf{O}^{\mathrm{T}}\mathbf{V}^{\mathrm{T}}\underline{\ell}_{i'}},$$

where $\mathbf{O} \sim \mathrm{Unif}(\mathcal{O}_{\tilde{n}_0})$. Since the test statistic depends on $\mathbf{Y}$ only through $\mathbf{V}\mathbf{\Lambda}$, and $\mathbf{R} = M_0M_0^{\mathrm{T}} + M_1\mathbf{O}M_1^{\mathrm{T}}$, the p-value can be written as

$$\vartheta_t(\mathbf{X}_*) = \Pr\left(d\min_{i'} \|\mathbf{\Lambda}^{-1}\mathbf{O}^{\mathrm{T}}\mathbf{V}^{\mathrm{T}}\underline{\ell}_{i'}\|_2^2 \le d\|\mathbf{\Lambda}^{-1}\mathbf{V}^{\mathrm{T}}\underline{\ell}\|_2^2 \;\middle|\; \mathbf{\Lambda}, \mathbf{V} \right).$$

By Lemmas 6 and 7, as $d \to \infty$,

$$d\|\mathbf{\Lambda}^{-1}\mathbf{V}^{\mathrm{T}}\underline{\ell}\|_2^2 \xrightarrow{p} 4\mu_0^{-2}$$

and $d\mathbf{V}\mathbf{\Lambda}^{-2}\mathbf{V}^{\mathrm{T}} \xrightarrow{p} UD^{-2}U^{\mathrm{T}}$ under H_a. This implies

$$\vartheta_t(\mathbf{Y}) \xrightarrow{p} \Pr\left(\min_{i'} \|D^{-1}\mathbf{O}U^{\mathrm{T}}\underline{\ell}_{i'}\|_2^2 \le 4\mu_0^{-2}\right). \tag{4.34}$$

Now we show that the right-hand side of (4.34) is 0 by showing that $\min_{i'} \|D^{-1}OU^{\mathrm{T}}\underline{\ell}_i\|_2^2$ is bounded below by $4\mu_0^{-2}$ for any $O \in \mathcal{O}_{\tilde{n}_0}$. Let $U_0 = [U, \underline{u}_n]$ and $\underline{a} = (a_1, 0, \ldots, 0, a_n)^{\mathrm{T}}$, where $\underline{u}_n$, a_1 and a_n are defined in (4.30). Since $\underline{\ell} = a_1\underline{u}_1 + a_n\underline{u}_n$, we can express $\underline{\ell}$ and $\underline{\ell}_{i'}$ as

$$\underline{\ell} = U_0\underline{a} \quad \text{and} \quad \underline{\ell}_{i'} = \Pi_{i'}U_0\underline{a},$$

where $\Pi_{i'}$ is the permutation matrix that switches the i'th and the last components of $\underline{\ell}$. Let $\underline{a}_{i'} = U^{\mathrm{T}}\Pi_{i'}U_0\underline{a}$. Then $\min_{i'} \|D^{-1}OU^{\mathrm{T}}\underline{\ell}_{i'}\|_2^2$ is equivalent to $\min_{i'} \underline{a}_i^{\mathrm{T}}O^{\mathrm{T}}D^{-2}O\underline{a}_i$. Since $O^{\mathrm{T}} \in \mathcal{O}_{\tilde{n}_0}$, if $\underline{a}_i$ is proportional to the first column of O^{T}, we have

$$\underline{a}_i^{\mathrm{T}}O^{\mathrm{T}}D^{-2}O\underline{a}_i = \delta^{-2}\|\underline{a}_i\|_2^2 = 4\mu_0^{-2}.$$

and this is the lower bound of $\min_{i'} \|D^{-1}OU^{\mathrm{T}}\underline{\ell}_i\|_2^2$. This implies that

$$\Pr\left(\min_{i'} \|D^{-1}\mathbf{O}U^{\mathrm{T}}\underline{\ell}_{i'}\|_2^2 \leq 4\mu_0^{-2}\right) = \Pr\left(\min_{i'} \|D^{-1}\mathbf{O}U^{\mathrm{T}}\underline{\ell}_{i'}\|_2^2 = 4\mu_0^{-2}\right).$$

Note that every entry of $\mathbf{O}$ is a continuous random variable and $\mathbf{O}$ is of full rank with probability 1. Since $\underline{\ell}_{i'}$ takes only finite values, we have

$$\Pr\left(\min_{i'} \|D^{-1}\mathbf{O}U^{\mathrm{T}}\underline{\ell}_{i'}\|_2^2 = 4\mu_0^{-2}\right) = 0.$$

□

Algorithm 1 Two-Stage Outlier Detection Procedure

Stage I: Screening Candidate Outliers

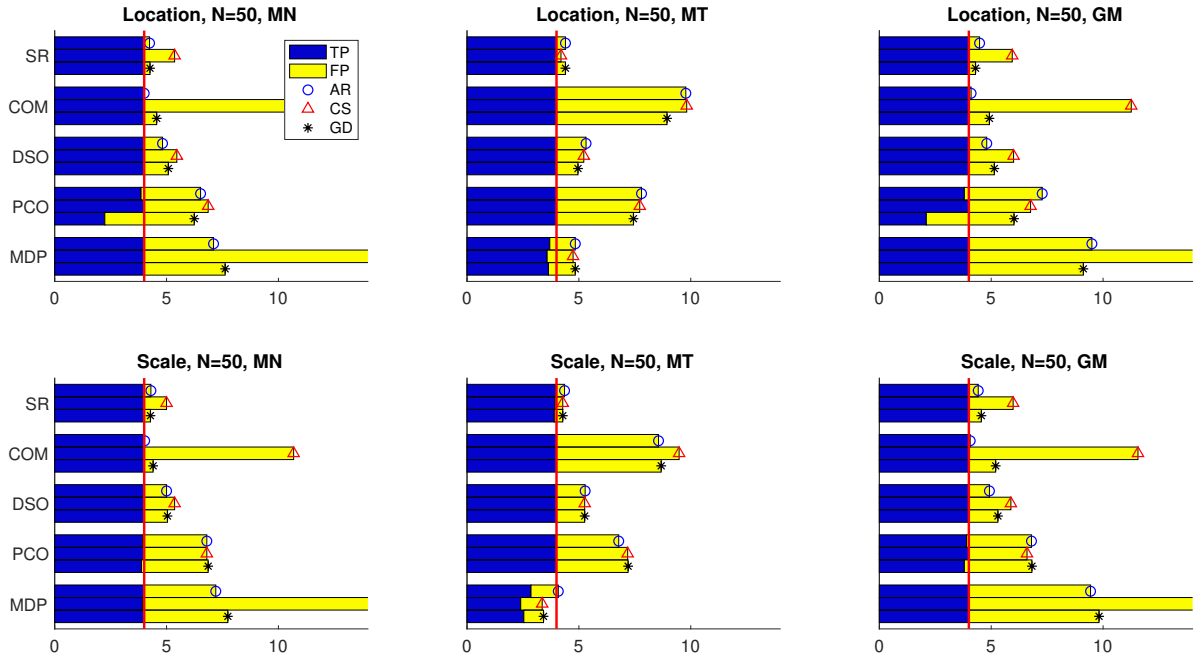
Input: S, $\tilde{n}_1$, α, K, $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_N]^{\mathrm{T}}$

for $i = 1$ to N **do**
  for $i' = \{1, \ldots, N\} \setminus \{i\}$ **do**
    $d_{i,i'} = \|\mathbf{x}_i - \mathbf{x}_{i'}\|_2$.
  end for
  Find $\mathrm{med}_i = \mathrm{median}\{d_{i,i'}\}_{i' \neq i}$
end for
Re-arrange med_i, $i = 1, \ldots, N$, so that $\mathrm{med}_{i_1} \leq \ldots \leq \mathrm{med}_{i_N}$
Set $\mathcal{I}_0 = \{i_1, \ldots, i_S\}$, $\mathbf{Z}_0 = [\mathbf{x}_{i_1}, \cdots, \mathbf{x}_{i_S}]^{\mathrm{T}}$, and $\mathbf{Z}_1 = [\mathbf{x}_{i_{S+1}}, \cdots, \mathbf{x}_{i_N}]^{\mathrm{T}}$.
for $s = S + 1$ to N **do**
  $D_{i_s} = DH(\mathbf{x}_{i_s} | \mathbf{Z}_0)$.
end for
Re-arrange $D_{i,s}$, $s = S + 1, \ldots, N$, so that $D_{i_{S+1}} \leq \ldots \leq D_{i_N}$
Set $\mathcal{I}$ by selecting $\tilde{n}_1$ indices corresponding to the largest D_{i_s}, $s = S + 1, \ldots, N$

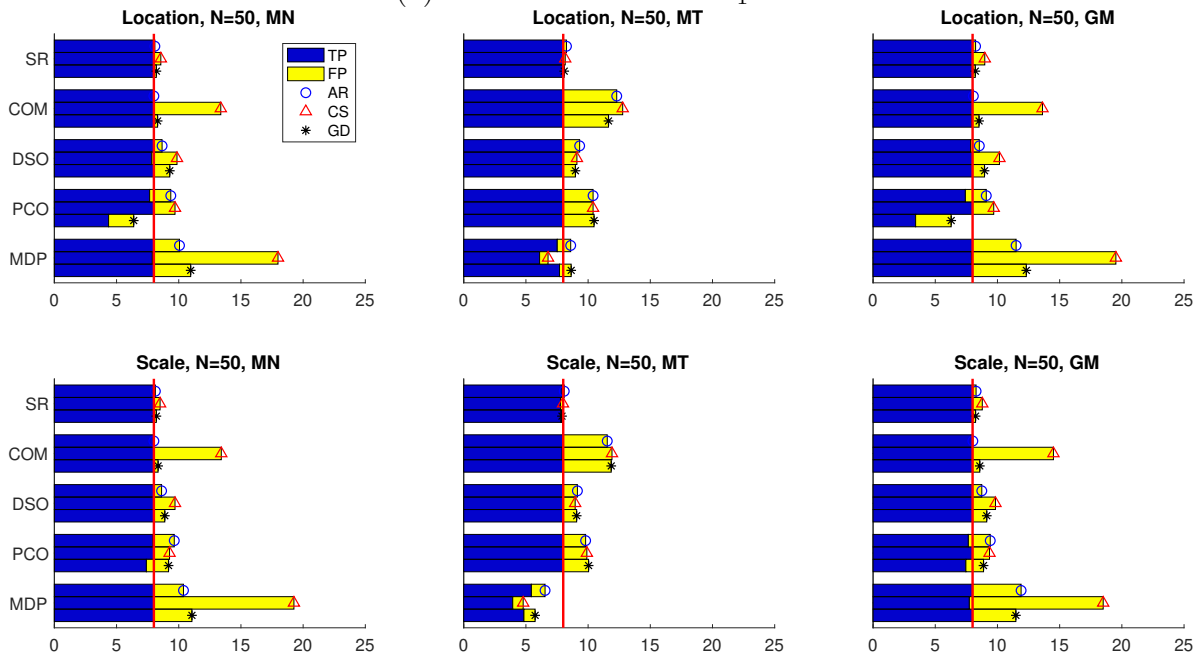
Stage II: Sequential SR Tests on Candidate Outliers

Initialize $\mathbf{X}_0$ and $\mathbf{X}_1$ with observations in $\mathcal{I}^c$ and $\mathcal{I}$, respectively.

for $j = 1$ to $\tilde{n}_1$ **do**
  Calculate $t_j = DH(\mathbf{x}_j^* | \mathbf{X}_0)$
end for
Rearrange $\mathbf{x}_j^*$ such that $t_1 \geq \ldots \geq t_{\tilde{n}_1}$.
for $j = 1$ to $\tilde{n}_1$ **do**
  Set $\mathbf{Y}_j = [\mathbf{X}_0^{\mathrm{T}}, \mathbf{x}_j^*]^{\mathrm{T}}$
  for $k = 1$ to K **do**
    Set $\mathbf{Y}_{j,k} = \mathbf{R}_k \mathbf{Y}_j$, where $\mathbf{R}_k \overset{iid}{\sim} \mathrm{Unif}(\mathcal{R})$
    Find the maximum LOO DH with $\mathbf{Y}_{j,k}$, and denote it by $t_{j,k}$
  end for
  Obtain the empirical SR distribution function $\hat{F}_{t|\mathbf{Y}_j}(z|\mathbf{Y}_j) = K^{-1} \sum_{k=1}^{K} 1(t_{j,k} \leq z)$.
  Set $\hat{c}_{\alpha,j} = \min\{z | \hat{F}_{t|\mathbf{Y}_j}(z|\mathbf{Y}_j) \geq 1 - \alpha\}$.
  if $t_j > \hat{c}_{\alpha,j}$ **then**
    Declare $\mathbf{x}_{j^*}$ as an outlier.
  else
    Declare $\mathbf{x}_1^*, \ldots, \mathbf{x}_{j-1}^*$ as outliers and terminate the procedure.
  end if
end for

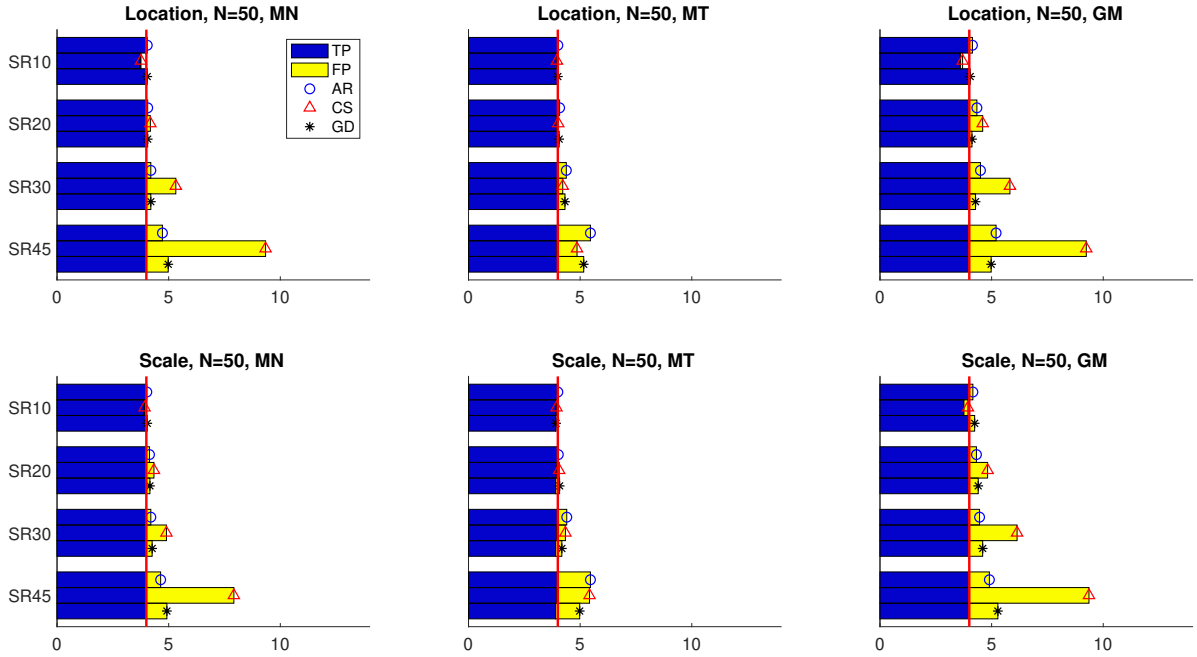


(a) Low contamination $n_1 = 4$

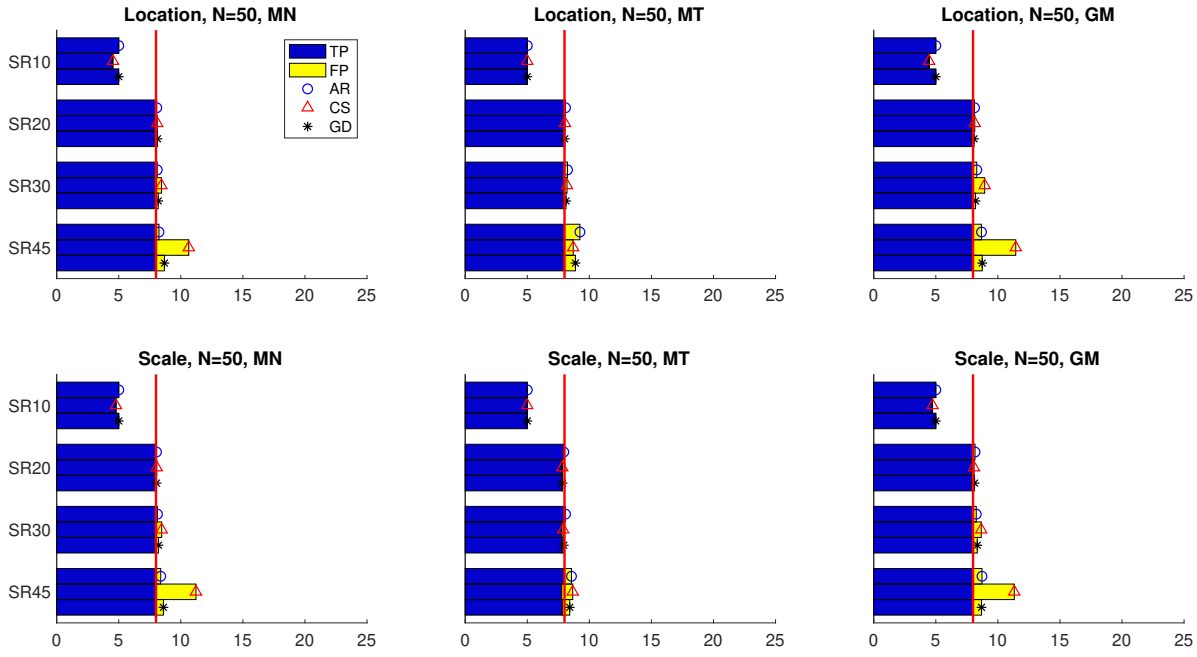


(b) High contamination $n_1 = 8$

Figure 4.11.1: The average numbers of detected outliers with simulated data based on 50 replications. The dimension $d = 1000$, the sample size $N = 50$ and the number of outliers, $n_1 = 4, 8$.

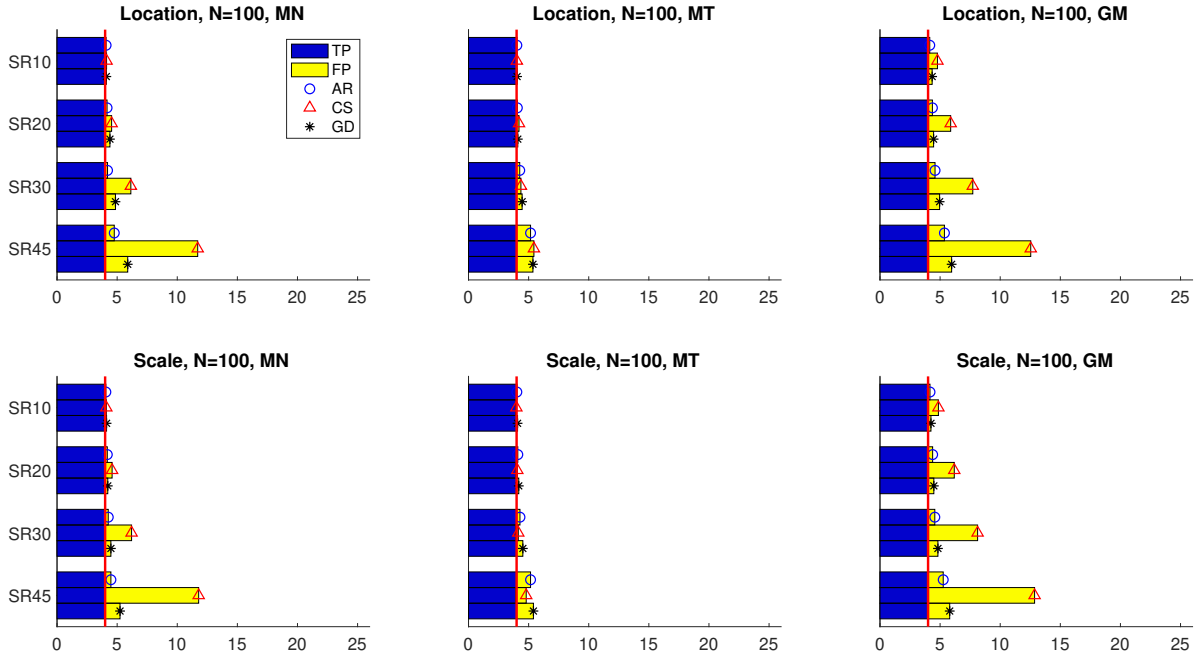


(a) Low contamination $n_1 = 4$

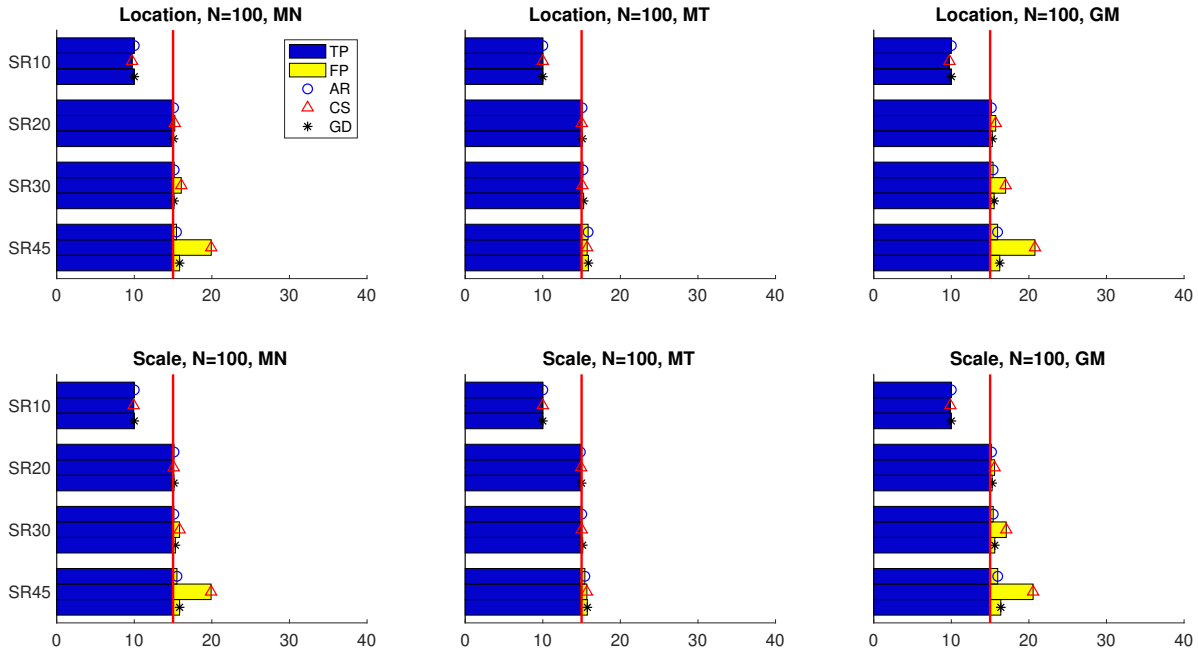


(b) High contamination $n_1 = 8$

Figure 4.11.2: Simulation results based on 50 replications with different screening proportions, $\eta = 0.1, 0.2, 0.3, 0.45$, each of which denoted by SR100η. The dimension $d = 1000$, the sample size $N = 50$ and the number of outliers $n_1 = 4, 8$.



(a) Low contamination $n_1 = 4$



(b) High contamination $n_1 = 15$

Figure 4.11.3: Simulation results based on 50 replications with different screening proportions, $\eta = 0.1, 0.2, 0.3, 0.45$, each of which denoted by SR100η. The dimension $d = 1000$, the sample size $N = 100$ and the number of outliers, $n_1 = 4, 15$.

Figure 4.11.4: The first ten cases of the face data outlier detection example. Each case is vertically arranged with ten images belonging to one individual, followed by three outlier images. Texts on the right side of an image indicate a method declaring it as an outlier.

Figure 4.11.5: The second ten cases of the face data outlier detection example. Each case is vertically arranged with ten images belonging to one individual, followed by three outlier images. Texts on the right side of an image indicate a method declaring it as an outlier.

Figure 4.11.6: The third ten cases of the face data outlier detection example. Each case is vertically arranged with ten images belonging to one individual, followed by three outlier images. Texts on the right side of an image indicate a method declaring it as an outlier.

Figure 4.11.7: The last ten cases of the face data outlier detection example. Each case is vertically arranged with ten images belonging to one individual, followed by three outlier images. Texts on the right side of an image indicate a method declaring it as an outlier.

Bibliography

Ahn, J., Lee, M. H., and Lee, J. A. (2019), "Distance-based outlier detection for high dimension, low sample size data," *Journal of Applied Statistics*, 46, 13–29.

Ahn, J., Lee, M. H., and Yoon, Y. J. (2012), "Clustering high dimension, low sample size data using the maximal data piling distance," *Statistica Sinica*, 22, 443–464.

Ahn, J. and Marron, J. S. (2010), "The maximal data piling direction for discrimination," *Biometrika*, 97, 254–259.

Ahn, J., Marron, J. S., Muller, K. E., and Chi, Y.-Y. (2007), "High dimension, low Sample size geometric representation holds under mild conditions," *Biometrika*, 3, 760–766.

Barnett, V. and Lewis, T. (1974), *Outliers in statistical data*, John Wiley & Sons.

Bradley, R. C. et al. (2005), "Basic properties of strong mixing conditions. A survey and some open questions," *Probability Surveys*, 2, 107–144.

Chikuse, Y. (2012), *Statistics on special manifolds*, vol. 174, Springer Science & Business Media.

Fang, K.-T. and Li, R. (1999), "Bayesian statistical inference on elliptical matrix distributions," *Journal of Multivariate Analysis*, 70, 66–85.

Filzmoser, P., Maronna, R., and Werner, M. (2008), "Outlier identification in high dimensions," *Computational Statistics and Data Analysis*, 52, 1694–1711.

Gupta, A. K. and Nagar, D. K. (1999), *Matrix variate distributions*, vol. 104, CRC Press.

Hall, P., Marron, J. S., and Neeman, A. (2005), "Geometric representation of high dimension, low sample size data," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67, 427–444.

Hoeffding, W. (1952), "The large-sample power of tests based on permutations of observations," *The Annals of Mathematical Statistics*, 23, 169–192.

James, A. T. (1954), "Normal multivariate analysis and the orthogonal group," *The Annals of Mathematical Statistics*, 25, 40–75.

Jung, S., Lee, M. H., and Ahn, J. (2018), "On the number of principal components in high dimensions," *Biometrika*, 105, 389–402.

Jung, S. and Marron, J. S. (2009), "PCA consistency in high dimension, low sample size context," *The Annals of Statistics*, 37, 4104–4130.

Kolmogorov, A. N. and Rozanov, Y. A. (1960), "On strong mixing conditions for stationary Gaussian processes," *Theory of Probability & Its Applications*, 5, 204–208.

Langsrud, o. (2005), "Rotation tests," *Statistics and Computing*, 15, 53–60.

Lee, S., Zou, F., and Wright, F. A. (2014), "Convergence of sample eigenvalues, eigenvectors, and principal component scores for ultra-high dimensional data," *Biometrika*, 101, 484–490.

Lehmann, E. L. and Romano, J. P. (2006), *Testing statistical hypotheses*, Springer Science & Business Media.

Muirhead, R. J. (2009), *Aspects of multivariate statistical theory*, vol. 197, John Wiley & Sons.

Perry, P. O. and Owen, A. B. (2010), "A rotation test to verify latent structure," *Journal of Machine Learning Research*, 11, 603–624.

Ro, K., Zou, C., Wang, Z., and Yin, G. (2015), "Outlier detection for high-dimensional data," *Biometrika*, 102, 589–599.

Sajesh, T. and Srinivasan, M. (2012), "Outlier detection for high dimensional data using the Comedian approach," *Journal of Statistical Computation and Simulation*, 82, 745–757.

Samaria, F. S. and Harter, A. C. (1994), "Parameterisation of a stochastic model for human face identification," in *Applications of Computer Vision, 1994., Proceedings of the Second IEEE Workshop on*, IEEE, pp. 138–142.

Simar, L. (2003), "Detecting outliers in frontier models: A simple approach," *Journal of Productivity Analysis*, 20, 391–424.

Solari, A., Finos, L., and Goeman, J. J. (2014), "Rotation-based multiple testing in the multivariate linear model," *Biometrics*, 70, 954–961.

Uhlig, H. et al. (1994), "On singular Wishart and singular multivariate beta distributions," *The Annals of Statistics*, 22, 395–405.