Small Area Estimation under Skew-Normal Nested Error Models

by

Mamadou Saliou Diallo

A Thesis submitted to
the Faculty of Graduate and Postdoctoral Affairs
in partial fulfilment of the requirements for the degree of
Doctor of Philosophy

Ottawa-Carleton Institute for Mathematics and Statistics

School of Mathematics and Statistics

Carleton University

Ottawa, Ontario, Canada

September 2014

Copyright ©

2014 - Mamadou Saliou Diallo

Abstract

In Small Area Estimation (SAE), the usual practice is to assume that the random components follow the normal distribution. A simulation study showed that under the one-fold nested error regression model assuming normal distribution may lead to large bias and significant increase of the mean squared error (MSE) of complex parameters (nonlinear function of the mean) estimation when the unit level errors' distribution is skewed. Hence, in this thesis, the assumption of normal distribution for the random components is relaxed by considering the skew-normal (SN) class of distributions. The SN class of distributions is very interesting because it contains the normal distribution family as a special case (when the skewness parameter is equal to zero).

For the one-fold nested error regression model with the random components following SN distribution, the empirical best linear unbiased prediction (EBLUP) and the empirical best (EB) estimators of linear parameters are provided. Under the same model, EB estimators of complex parameters are developed following the approach introduced by Molina and Rao (2010). A simpler conditional alternative method is compared to the previously mentioned method. The semi-parametric method developed by Elbers et al. (2003) is improved by correctly assigning the area effects. A parametric bootstrap approach for the MSE estimation of the EB estimator and a semi-parametric bootstrap method for the ELL method are specified. An HB method for estimating complex parameters is also presented. The HB method uses Monte Carlo (MC) simulations and sampling importance resampling (SIR) techniques no

Markov chain Monte Carlo (MCMC) is required.

For the two-fold nested error regression model with the random components following SN distribution, the empirical best linear unbiased prediction (EBLUP) and the empirical best (EB) for linear parameters for linear parameters are discussed. The best predictor of the area and sub-area random effects is provided under the general linear mixed model setup.

For repeated cross-sectional surveys, Rao and Yu (1992, 1994) used AR(1) series to model the area-by-time effects and doing so improved the estimation of the small areas linear parameters. The stability of the infinite AR(1) series requires the assumption of stationarity i.e. the autoregressive coefficient of correlation is assumed to be strictly smaller than 1 in absolute value. A separate random walk model used by Datta et al. (2002) forces the autoregressive coefficient of correlation to be equal to 1 under a finite series for the area-by-time effects. A unified approach under finite series for the area-by-time effects is developed. This unified approach has the advantage of letting the model “decide” whether the autoregressive coefficient of correlation is equal to 1 or not. Maximum likelihood (ML) and restricted maximum likelihood (REML) methods for estimating the parameters of the model and the equivalent MSE estimators are discussed.

Acknowledgments

I would like to express my deepest gratitude to my supervisor Professor J.N.K. Rao. His support, enthusiasm, and remarkable knowledge have made my PhD research journey extremely beneficial and instructive. But most importantly, he has made me passionate about research and I thank him for all that from the bottom of my heart. Also, I would like to thank the members of my thesis defense committee for their valuable comments.

I would like to extend my thanks to Statistics Canada for making it easy for me in the early stages of my PhD program to attend the required courses during working hours. I appreciate that my colleagues from Westat have shown genuine concern for me completing the PhD program. In particular I am very grateful to Dr. Robert E. Fay from Westat; I have learned so much these past few years working with him (must say for him) on the very interesting topic of estimating small area crime rates using the National Crime Victimization Survey (NCVS). I truly appreciate his kindness and positive personality. I also owe thanks to Dr. Judith Strenio for proofreading most of the chapters of this thesis. Since, I started working at Westat, Dr. Graham Kalton has consistently reminded me the value of completing the PhD program and pushed me towards that goal. I thank him for making my achievement his goal.

Last but not least, I would like to acknowledge the support of my family and friends. I dedicate this thesis to my wife Dalanda for being understanding and patient

during all these years I spent my nights and weekends with books and computers; to my new born son for bringing so much joy to my life; to my parents for all their sacrifices, commitment to my education, and their unconditional love; to my brothers and sisters for their positive attitude. To my friends, I say thank you for valuing our companionship and bringing so much to my life every day.

Table of Contents

Abstract	ii
Acknowledgments	iv
Table of Contents	vi
List of Tables	x
List of Figures	xi
1 Introduction	1
1.1 Small Area Estimation (SAE)	3
1.1.1 Area Level Models	4
1.1.2 Unit Level Models	6
1.2 Use of Skew-Normal (SN) distribution to Relax the normality assumption	8
1.3 Thesis Outline	10
2 Review of Skew-Normal (SN) Distributions	13
2.1 The Univariate SN Distribution	13
2.2 The Multivariate Skew-Normal (MSN) Distribution	17
2.3 The Closed Skew-Normal (CSN) Distribution	20
2.3.1 Generation of Random Vectors from the CSN Distribution . .	21
2.3.2 Moment Generating Function (mgf) of CSN	22

2.3.3	Linear Transformations	24
2.3.4	Marginal and Conditional Distributions	26
2.3.5	Joint distribution of Independent CSN Random Vectors	28
2.3.6	Sums of Independent CSN Random Vectors	30
2.4	Further Generalization	32
2.4.1	Skew-t Distribution	32
2.4.2	Skew-Elliptical Distribution	33
3	Parameter Estimation under the One-Fold SN Model	34
3.1	Small Area Estimation Model	34
3.2	Population and Sample Distributions	42
3.2.1	u_d and e_{dj} follow SN Distribution	43
3.2.2	u_d follows Normal and e_{dj} follows SN	45
3.2.3	u_d follows SN and e_{dj} follows Normal	45
3.2.4	Both u_d and e_{dj} follow normal distribution	46
3.3	Moment Generating Function and Moments	47
3.4	Parameters Estimation	52
3.4.1	Maximum Likelihood Estimation	52
3.4.2	Simulation Results for the ML method	56
3.4.3	Data Cloning (DC)	63
3.5	Appendix	66
3.5.1	Computing Multivariate Normal CDF - Special Case	66
3.5.2	Differentiation of the Likelihood Function	69
3.5.3	Marginal Distribution of Y_{ds}	73
4	Empirical Best (EB) Prediction under One-Fold SN Model	77
4.1	Prediction for Linear Parameters	78
4.1.1	Empirical Best Linear Unbiased Prediction (EBLUP)	79

4.1.2	Empirical Best (EB) Prediction	80
4.2	Prediction for Complex Parameters	83
4.3	Best Prediction for Complex Parameters under SN Random Errors	86
4.3.1	Conditional Distribution of $\mathbf{Y}_{dr s}$	86
4.3.2	Quasi-Univariate Generation for Big Data	88
4.3.3	Special Case 1: e_{dj} follows SN and u_d follows normal	94
4.3.4	Special Case 2: u_d follows SN and e_{dj} follows normal	96
4.4	Conditional Best Prediction for Complex Parameters	97
4.5	Prediction Based on the Half-Normal Representation for Complex Parameters	101
4.6	Normality Assumption: Molina-Rao Method	105
4.7	Nonparametric Approach: ELL Method	107
4.7.1	Traditional ELL Predictor	108
4.7.2	Modifications to the Traditional ELL Predictor	110
4.8	Prediction for Non Sampled Areas and Non Linkable samples	113
4.9	Simulation Study using FGT Measures	114
4.9.1	FGT Poverty Measures	114
4.9.2	Marginal and Conditional Best Prediction	121
4.9.3	Prediction based on the Half-Normal Representation	124
4.9.4	Molina-Rao Predictor under Skew-Normal Model	125
4.9.5	ELL Method under Skew-Normal Model	129
4.9.6	Comparison of the Different Predictors	132
4.10	MSE Estimation under Skew-Normal Errors	135
4.11	Appendix: Proof of Proposition 4.3	139
5	Hierarchical Bayes (HB) Prediction under One-Fold SN Model	142
5.1	Review of the HB Method for the Normal Model	144

5.1.1	Model	145
5.1.2	Derivation of the Posterior Densities	147
5.2	HB Method for the SN Model	150
5.2.1	Model	151
5.2.2	Model Fitting	153
5.2.3	Some Special Models	161
5.2.4	Prediction of Complex Parameters	165
5.3	Simulation Study	166
5.4	Appendix: Introduction to Sampling Importance Resampling (SIR) .	175
6	Two-Fold Model under Skew-Normal Errors	179
6.1	Two-Fold Nested Error Model	179
6.2	Prediction of Small Area Means	182
6.2.1	Best Linear Unbiased Prediction (BLUP)	184
6.2.2	Best Prediction (BP)	186
7	Cross-Sectional Time Series Data in Small Area Estimation	193
7.1	Rao-Yu Model Without Stationarity Assumption	195
7.2	Parameter Estimation	197
7.2.1	Maximum Likelihood (ML) Estimation	197
7.2.2	Restricted Maximum Likelihood (REML) Estimation	199
7.3	Mean Squared Error (MSE) Estimation	200
7.3.1	MSE Estimator of the EBLUP	200
7.3.2	Estimation of the MSE Estimator of the EBLUP	202
7.4	Simulation Results	203
8	Future Research	206
	Bibliography	211

List of Tables

3.1	Parameter estimation for the nested error model with SN errors using the ML method. We simulated 1,000 samples with $m = 80$ small areas and $n_d = 50$ units per area.	58
3.2	Parameters estimation for the nested error model with SN errors using the unconstrained ML method. We simulated 1,000 samples with $\lambda_u = 0$ i.e $u_d \sim N(0, \sigma_u^2)$	62
4.1	Parameter estimation from the Monte Carlo simulation of fitting the SN model ($\lambda_e = 3$ and $\lambda_u = 1$). We simulated 5,000 populations with $m = 80$ small areas and $n_d = 50$ units per area.	121
4.2	Parameter estimation from the Monte Carlo simulation when fitting SN using normal distribution. We simulated 5,000 population with $m = 30$ small areas and $n_d = 50$ units per area.	126
5.1	Distribution of the design effects of the importance weights across the 1,000 generated populations	174
7.1	ARBs for estimation of the coefficient ρ using ML method. We simulated 5,000 samples for each combination of $\sigma_u^2 = 0$ and σ^2	205
7.2	ARBs for estimation of the coefficient σ^2 using ML method. We simulated 5,000 samples for each combination of $\sigma_u^2 = 0$ and σ^2	205
7.3	ARBs for estimation of the coefficient σ_u^2 using ML method. We simulated 5,000 samples for each combination of $\sigma_u^2 = 0$ and σ^2	205

List of Figures

2.1	Density curves of two SN distributions.	16
2.2	Density surface of a bi-SN.	19
3.1	MC densities of the fixed effects' parameters for the nested error model with SN errors using the ML method. We simulated 1,000 samples with $d = 80$ small areas and $n_d = 50$ units per area.	59
3.2	MC densities of the random components' parameters for the nested error model with SN errors using the ML method. We simulated 1,000 samples with $d = 80$ small areas and $n_d = 50$ units per area.	59
3.3	MC boxplots of the parameters estimates for the nested error model with SN errors using the unconstrained ML method. We simulated 1,000 samples with $d = 80$ small areas and $n_d = 50$ units per area. . .	61
4.1	Pdf of SN and the pdf of the normals with the same mean and the standard deviation (sd).	118
4.2	Monte Carlo averages of the population poverty incidence, gap, and severity for each of the 80 areas; using the 5,000 populations with $n_d = 50$, $\lambda_u = 1$ and $\lambda_e = 3$	120
4.3	Reduction in percentage of the MSE obtained by EQB over the MSE of the direct estimator for all three poverty measures (incidence, gap, and severity) with $m = 80$ areas, $n_d = 50$ units, $\lambda_e = 3$, and $\lambda_u = 1$. .	122

4.4	Comparison of the empirical quasi-best predictor EQB to the conditional predictors C-EQB and C-EQBLUP in terms of bias and MSE (poverty gap, $\alpha=1$) with $m = 80$ areas, $n_d = 50$ units, $\lambda_e = 3$, and $\lambda_u = 1$	123
4.5	Comparison of the empirical quasi-best predictor EQB and the estimators using the half-normal representation in terms of bias and MSE (poverty gap, $\alpha=1$) with $m = 80$ areas, $n_d = 50$ units, $\lambda_e = 3$, and $\lambda_u = 1$.	124
4.6	Absolute and relative bias of the Molina-Rao normality-based estimator when both random errors follow SN distribution (poverty gap, $\alpha=1$).	126
4.7	MSE of the Molina-Rao normality-based estimator when at least one random error follows SN distribution (poverty gap, $\alpha=1$).	127
4.8	Ratio of bias squared to MSE of the Molina-Rao normality-based estimator (in percent) when at least one random error follows SN distribution (poverty gap, $\alpha=1$).	127
4.9	Ratio of MSE of the Molina-Rao normality-based estimator over the direct estimator when both random errors follow SN distribution (poverty gap, $\alpha=1$).	128
4.10	Bias of the three different ELL methods when both random errors follow SN distribution (poverty gap, $\alpha=1$) with $m = 80$ areas, $n_d = 50$ units, $\lambda_e = 3$, and $\lambda_u = 1$	129
4.11	MSE of the three different ELL methods when both random errors follow SN distribution (poverty gap, $\alpha=1$) with $m = 80$ areas, $n_d = 50$ units, $\lambda_e = 3$, and $\lambda_u = 1$	130
4.12	MSE of the three different ELL methods and Molina-Rao normality-based when both the random errors follow SN distribution (poverty gap, $\alpha=1$) with $m = 80$ areas, $n_d = 50$ units, $\lambda_e = 3$, and $\lambda_u = 1$	131

4.13	Ratio of the MSEs of the five estimators (best predictor, conditional best predictor, half-normal, Molina-Rao, and modified ELL using MOM) to the MSE of the direct estimator (poverty incidence, $\alpha=0$) with $m = 80$ areas, $n_d = 50$ units, $\lambda_e = 3$, and $\lambda_u = 1$	132
4.14	Ratio of the MSEs of the five estimators (best predictor, conditional best predictor, half-normal, Molina-Rao, and modified ELL using MOM) to the MSE of the direct estimator (poverty gap, $\alpha=1$) with $m = 80$ areas, $n_d = 50$ units, $\lambda_e = 3$, and $\lambda_u = 1$	133
4.15	Ratio of the MSEs of the five estimators (best predictor, conditional best predictor, half-normal, Molina-Rao, and modified ELL using MOM) to the MSE of the direct estimator (poverty severity, $\alpha=2$) with $m = 80$ areas, $n_d = 50$ units, $\lambda_e = 3$, and $\lambda_u = 1$	134
5.1	Average of HB point estimates compared to the average of the EB point estimates (incidence, gap, and severity) with $m = 40$ areas, $n_d = 50$ units, $\lambda_e = 3$, and $\lambda_u = 0$	169
5.2	Average of HB MC MSEs compared to the average of the EB MC MSE (incidence, gap, and severity) with $m = 40$ areas, $n_d = 50$ units, $\lambda_e = 3$, and $\lambda_u = 0$	170
5.3	HB coverage for both the equal tails and the HPD CIs (incidence, gap, and severity) with $m = 40$ areas, $n_d = 50$ units, $\lambda_e = 3$, and $\lambda_u = 0$. . .	171
5.4	HB widths for both the equal tails and the HPD CIs (incidence, gap, and severity) with $m = 40$ areas, $n_d = 50$ units, $\lambda_e = 3$, and $\lambda_u = 0$. . .	171
5.5	Bias of HB normality-based and HB SN methods (incidence, gap, and severity) with $m = 40$ areas, $n_d = 50$ units, $\lambda_e = 3$, and $\lambda_u = 0$	172
5.6	MSE estimates of HB normality-based and HB SN models (incidence, gap, and severity) with $m = 40$ areas, $n_d = 50$ units, $\lambda_e = 3$, and $\lambda_u = 0$	172

5.7	An example of the distribution of the importance weights from the SN-based HB method.	173
5.8	Proportion of the weights belonging to the interval $(0.00001, 0.001)$ for each of the 1,000 populations.	174
5.9	Proportion of the weights belonging to the interval $(0.000001, 0.01)$ for each of the 1,000 populations.	175

Chapter 1

Introduction

The preferred means of acquiring information about a large population is usually a sample survey rather than a census. Not only is a sample survey generally much less expensive than a census, it can also be completed more quickly, and non-sampling errors can be more closely controlled and thus minimized. Sample surveys are an important information-gathering tool for government, policy makers, and other organizations in order to inform their decision making. Traditional design-based theory uses weights obtained from the sampling scheme probability distribution to make population estimates. These are called direct estimators because they are based only on the sample information. A direct estimator may also use known auxiliary information such as the total of a variable x related to the characteristic of interest y . Well-known textbooks such as Cochran (1977), Sarndal et al. (2003), and Lohr (2010) provide the mathematical background for design-based theory.

Decision makers are increasingly requesting information for subpopulations (areas) as well as the population as a whole. These subpopulations can be geographical areas such as state, county, school district, etc. or socio-demographic groups such as age, gender, race, etc. or any subset of the population. Direct estimators can produce reliable estimates for subpopulations which have a large enough sample size. However,

as the sample design becomes more complex (with phases, stages, subsamples, etc.) and the desired number of local areas increases, the overall sample size required to support reliable direct estimates for all areas of interest becomes very large. Often reliable direct estimates are planned for only some main subpopulations due to limited budgets and the necessity of controlling costs, leaving many subpopulations with not enough sample to produce direct estimates with reasonable precision.

In response to this problem, a class of estimation known as Small Area Estimation (SAE) has been developed, and in recent years SAE methods have been used in many surveys to improve subpopulation estimates. The main idea of the SAE approach is to exploit the relationship among the areas (borrowing strength across areas) to improve the precision of a given small area. Sometimes, it may be advantageous to borrow strength across other dimensions such as time for repeated cross-sectional surveys or sub-small areas for example to better reflect stage sampling. Mixed models are often used to formulate the between area variation as well as the variation explained by auxiliary variables. In the next paragraphs, we present three examples of the use of SAE methods in surveys.

The Small Area Income and Poverty Estimates (SAIPE) program, conducted by the U.S. Census Bureau, provides more current estimates of selected income and poverty related statistics than the most recent decennial census. The estimates are produced for states, counties, and school districts. SAIPE's dependent variable is the estimates of poverty from American Community Survey (ACS), a continuous survey with people responding throughout a given year. Many data sources, including administrative data, postcensal population estimates, and the decennial census, are used to obtain auxiliary information. The SAIPE's SAE approach can be summarized in two steps. The first step consists of modeling the direct ACS poverty rates for defined age groups. In the second step, the modeled ratios are multiplied by the

population estimates to obtain the model-based estimates of number of people in poverty for the given age groups.

In 2003, the National Center for Education Statistics (NCES) conducted the National Assessment of Adult Literacy (NAAL) survey to measure the United States of America (USA) English literacy skills. The survey was designed to produce direct estimates for the USA at the national level and for some major subgroups (e.g regions, level of education, and race/ethnicity). However policymakers and educators were interested in estimates at the local levels such as state and county. Since NAAL was not designed to produce direct estimates at the state or county level, SAE methods were developed to satisfy policymakers' and educators' need for reliable estimates for their local areas (see Mohadjer et al. (2012)).

The World Bank has been producing internationally comparable and global poverty estimates since 1990. The World Bank uses surveys and censuses data on measures of income and consumption to derive poverty rates for developing countries. Survey data are rarely of sufficient size to yield reliable direct estimates for all countries and the censuses often lack income or consumption related information for the small areas or the information is not of good quality. Elbers et al. (2003) proposed an SAE method which uses auxiliary information from the census and the sample residuals to generate a complete synthetic census for the dependent variable. From the predicted census, any parameters including complex ones such as poverty indicators can be estimated.

1.1 Small Area Estimation (SAE)

The SAE models discussed in this thesis are based on the linear mixed model (LMM). LMMs and their generalizations are widely used in many areas of statistics where complex stochastic models are necessary to represent physical phenomena, natural

behaviour, social patterns, and many other events. The LMMs are composed of fixed effects and random effects. The fixed effects are used to model the variability explained by the auxiliary variables. The random effects, in the context of SAE, account for the between-area, as well as when applicable between-sub-area and across time variations. SAE techniques based on mixed models can be classified in two major groups: area level models and unit level models.

1.1.1 Area Level Models

This approach models characteristics of interest using known auxiliary information at some aggregated level. The auxiliary variables are usually population quantities obtained from censuses, major reference surveys such as ACS and Current Population Survey (CPS), or administrative sources. The first use of this model in SAE is found in Fay and Herriot (1979), who proposed the following two-level approach often referred to as the *Fay-Herriot model*:

$$\text{Sampling Model: } \hat{\theta}_d | \theta_d \stackrel{\text{ind}}{\sim} N(\theta_d, \psi_d), \quad d = 1, \dots, m \quad (1.1)$$

$$\text{Linking Model: } \theta_d | \boldsymbol{\beta}, \sigma_u^2 \stackrel{\text{ind}}{\sim} N(\mathbf{x}_d^T \boldsymbol{\beta}, \sigma_u^2), \quad (1.2)$$

where $\theta_d = g(\bar{Y}_d)$, the parameter of interest for area d , is a function of the small area mean $\bar{Y}_d$, $\hat{\theta}_d$ is the direct estimator of θ_d , and ψ_d is the sampling variance. A common representation of the Fay-Herriot model is:

$$\hat{\theta}_d = \theta_d + e_d = \mathbf{x}_d^T \boldsymbol{\beta} + u_d + e_d, \quad d = 1, \dots, m. \quad (1.3)$$

where $u_d \stackrel{\text{ind}}{\sim} N(0, \sigma_u^2)$ and $e_d \stackrel{\text{ind}}{\sim} N(0, \psi_d)$ are independent. The sampling variance ψ_d is assumed to be known in (1.1); in a real survey this quantity is unknown and must be estimated. The estimated sampling variances can be very unstable due to the small

effective sample sizes. Therefore a common approach is to first estimate the sampling variances using traditional domain estimation techniques or resampling methods then use a generalized variance function (GVF) (Wolker (1985)) to smooth the estimated sampling variances. At the end of this process, the smoothed variances are treated as known for the purpose of applying the Fay-Herriot model. The other parameters of the model, $\boldsymbol{\beta}$ and σ_u^2 are estimated using method of moment (MOM), maximum likelihood (ML), restricted maximum likelihood (REML), or any other suitable technique.

Under model (1.3), the best predictor (best in the sense of minimizing the mean squared error (MSE)) of θ_d is:

$$\hat{\theta}_d^B = (1 - B_d)\hat{\theta}_d + B_d\mathbf{x}_d^T\tilde{\boldsymbol{\beta}}, \quad d = 1, \dots, m \quad (1.4)$$

where $B_d = \psi_d / (\sigma_u^2 + \psi_d)$ and $\tilde{\boldsymbol{\beta}}$ is the best linear unbiased estimator of $\boldsymbol{\beta}$. If you replace the unknown parameter σ_u^2 by its estimator $\hat{\sigma}_u^2$ using a suitable method, then you obtain the empirical best (EB), or empirical bayes, predictor $\hat{\theta}_d^{EB} = (1 - \hat{B}_d)\hat{\theta}_d + \hat{B}_d\mathbf{x}_d^T\hat{\boldsymbol{\beta}}$. The EB estimator is clearly a weighted average of the direct estimator $\hat{\theta}_d$ and the regression predictor $\mathbf{x}_d^T\hat{\boldsymbol{\beta}}$. The weight $\hat{B}_d = \psi_d / (\hat{\sigma}_u^2 + \psi_d)$ is a function of the sampling variance ψ_d and when ψ_d is large meaning that the direct estimator is less precise then more weight is given to the regression predictor. On the other hand, smaller values of ψ_d put more weight to the direct estimator.

There are several extensions of the basic Fay-Herriot model. One of them, proposed by Rao and Yu (1992, 1994), handles time series and cross-sectional data by incorporating an area by time autoregressive process of order 1. In this document, we refer to this approach as the Rao-Yu model. They used the method of moments to estimate the unknown parameters of the model but they found it difficult to estimate the autoregressive coefficient in the presence of sampling errors. In particular, when

sampling errors are taken into account in the estimation, values of the autoregressive coefficient can fall outside the admissible interval $(-1, 1)$. And they concluded that “suitable modifications of $\hat{\rho}$ that can lead to more efficient estimators taking values in the admissible range are needed”. Recently Marhuenda et al. (2013) describe a restricted maximum likelihood approach for estimating the parameters of a extension of Rao-Yu model which they called the spatio-temporal Fay-Herriot model.

1.1.2 Unit Level Models

For this modeling approach, the characteristics of interest and the auxiliary data are available at the unit level and LMMs are used to represent the assumed stochastic relationship between the quantities. The unit level model which we refer to as the one-fold nested error model can be formally stated as follows:

$$Y_{dj} = \mathbf{x}_{dj}^T \boldsymbol{\beta} + u_d + e_{dj}, \quad j = 1, \dots, N_d, \quad d = 1, \dots, m. \quad (1.5)$$

$$u_d \stackrel{iid}{\sim} f_u, \quad \text{and} \quad e_{dj} \stackrel{iid}{\sim} f_e \quad (1.6)$$

where:

1. u_d and e_{dj} are independent,
2. f_u and f_e are probability distributions,
3. d designates the small-area and j designates the unit within the area d .

This unit level model was first used in the context of SAE by Battese et al. (1988) for prediction of corn and soybean crop areas for 12 counties in north-central Iowa using farm-interview data and satellite information. They assumed the normal distribution for the error terms that is $u_d \stackrel{iid}{\sim} N(0, \sigma_u^2)$, $e_{dj} \stackrel{iid}{\sim} N(0, \sigma_e^2)$. The model (1.5) may be

written in matrix form as:

$$\mathbf{y}_d = \mathbf{X}_d \boldsymbol{\beta} + u_d \mathbf{1}_{n_d} + \mathbf{e}_d, \quad d = 1, \dots, m, \quad (1.7)$$

where $\mathbf{y}_d$ is the $(n_d \times 1)$ -vector of sample observations y_{dj} from area d , $\mathbf{X}_d$ is the matrix of auxiliary variables and $\mathbf{e}_d$ is the vector of errors e_{dj} with $j = 1, \dots, n_d$. Also if we use the notation,

$$\text{Cov}(\mathbf{y}_d) = \mathbf{V}_d = \sigma_e^2 \mathbf{I}_{n_d} + \sigma_u^2 \mathbf{1}_{n_d} \mathbf{1}_{n_d}^T, \quad (1.8)$$

then the best linear unbiased predictor (BLUP) estimator of $\theta_d = \bar{\mathbf{X}}_d^T \boldsymbol{\beta} + u_d$ is:

$$\hat{\theta}_d^B = \bar{\mathbf{X}}_d^T \tilde{\boldsymbol{\beta}} + \gamma_d (\bar{y}_d - \bar{x}_d^T \tilde{\boldsymbol{\beta}}) \quad (1.9)$$

where $\gamma_d = \frac{\sigma_u^2}{\sigma_e^2 + n_d \sigma_u^2}$ and the estimator $\tilde{\boldsymbol{\beta}}$ is the BLUE of $\boldsymbol{\beta}$ that is $\tilde{\boldsymbol{\beta}} = (\sum_d \mathbf{X}_d^T \mathbf{V}_d^{-1} \mathbf{X}_d)^{-1} (\sum_d \mathbf{X}_d^T \mathbf{V}_d^{-1} \mathbf{y}_d)$.

The estimator (1.9) is well suited for parameters that are linear functions of the area means. However in many applications, the parameters are complex functions of the observations. For instance, many measures of poverty and inequality are nonlinear functions of unit level income or welfare related quantities. Elbers et al. (2003) proposed an empirical semi-parametric method to estimate small area poverty indices. This method, commonly called the ELL method, draws from the residuals (area-level and unit-level) to reconstitute the entire census. After predicting the census, any complex statistic is easily obtained. The ELL method has poor MSE performance in many situations even though the bias is usually small. Inspired by Elbers et al. (2003) work on measuring poverty and inequality in the developing countries, Molina and Rao (2010) proposed an EB approach that can handle non-linear complex parameters. This method is referred as the Molina-Rao method in this thesis. The idea behind the

Molina-Rao method is to assume the nested error model for the entire population and use the conditional distribution of the non-observed characteristics $\mathbf{y}_r$ (r designates the non-sampled units) given the sample $\mathbf{y}_s$ (s designates the sampled units) to predict the non-observed values. This procedure produces a complete census for the characteristics of interest by concatenating the sample values of interest to the predicted out-of-sample values. Hence, as for the ELL method, any complex function of $\mathbf{y}$ can be estimated.

Many large scale surveys are conducted in stages within small areas. For instance, primary sampling units (PSUs) may be selected within small areas then sub-PSUs at later stages. A class of models, called the two-fold models in this thesis, allows to model both the small areas and the PSUs in addition to the individual units. The two-fold model is defined as

$$Y_{djk} = \mathbf{x}_{djk}^T \boldsymbol{\beta} + u_d + v_{dj} + e_{djk}; \quad d = 1, \dots, m; j = 1, \dots, M_d; k = 1, \dots, N_{dj}, \quad (1.10)$$

where $u_d \stackrel{iid}{\sim} f_u$, $v_{dj} \stackrel{iid}{\sim} f_v$ and $e_{djk} \stackrel{iid}{\sim} f_e$ are independent, f_u , f_v and f_e are probability distributions, d designates the small-area, j designates the cluster within small area d , and k designates the unit within the area d and cluster j .

In an actual survey, assuming the normal distribution could be too restrictive; the SN distribution can be an alternative to the normal distribution assumption.

1.2 Use of Skew-Normal (SN) distribution to Relax the normality assumption

To my knowledge, the SN distribution has not yet been used in SAE. However, the SN distribution has been already used with LMMs (Arellano-Valle et al. (2005), Arellano-Valle et al. (2007), Lachos et al. (2007), Lin and Lee (2008), Lachos et al. (2010)).

Consider the following LMM

$$\mathbf{Y}_d = \mathbf{X}_d\boldsymbol{\beta} + \mathbf{Z}_d\mathbf{u}_d + \mathbf{e}_d, \quad d = 1, \dots, m. \quad (1.11)$$

where $\mathbf{u}_d$ and $\mathbf{e}_d$ are independent random vectors. If a random vector $\mathbf{W}$ follows a multivariate skew-normal (MSN), it can be shown that

$$W_j = \delta_j|V_0| + (1 - \delta_j^2)^{1/2}V_j, \quad j = 1, \dots, k \quad (1.12)$$

where the δ_j 's are constants and the V_j 's follow normal distributions. The expression (1.12) shows that the W_j 's are correlated because they are obtained using the same random variable V_0 . Arellano-Valle et al. (2005) applied this MSN to the random vector $\mathbf{e}_d$ resulting in a correlation of the unit level errors e_{dj} . Lachos et al. (2007) and Lin and Lee (2008) assumed the MSN for the term $\mathbf{u}_d$ and multivariate normal for the unit level errors $\mathbf{e}_d$. They did not consider skewed distributions for the errors e_{dj} . Lachos et al. (2010) used what they called skew-normal independent (SNI) distributions to model the random errors $\mathbf{u}_d$ and $\mathbf{e}_d$. Let us consider the random vector $\mathbf{W}$ following SNI. A stochastic representation of $\mathbf{W}$ is

$$\mathbf{W} = \mu + H^{-1/2}\mathbf{V} \quad (1.13)$$

where $\mathbf{V}$ follows a MSN as defined in Section 2.2 and H is a positive random variable independent of $\mathbf{V}$. For the error term $\mathbf{e}_d$, they only considered the case with $\mathbf{V}$ following a multivariate normal resulting in symmetrical distribution with heavy tails. Arellano-Valle et al. (2007) considered the linear mixed model (1.11) with the flexibility of independent e_{dj} if needed, however the inference was done under the hierarchical bayes (HB) framework.

In this thesis, we consider under the LMM in particular the one-fold nested error model (1.5) assuming independent e_{dj} following SN distributions. We only assume non-informative sampling designs. Under this setup, we extend the SAE techniques for estimating linear and complex parameters. We also consider some other extensions such as the HB method for estimating small area complex parameters under one-fold nested error model and the estimation of small area linear parameters under the two-fold nested error models with the errors following SN distributions. The two-fold nested error models have one random component to model the area effect, another random component to model the clusters within small areas, and an unit level random component (see precise definition in Chapter 5). Rao and Yu (1994) developed SAE model which takes advantage of correlation of estimates over time in repeated cross-sectional surveys to improve a given year estimate. Their model assumes an AR(1) series that is the stationarity assumption is required for the time series. We proposed an extension to the model by removing the stationarity assumption under finite series. In the following and throughout this thesis, we refer to the nested error linear model with the errors following SN distributions as the *SN model* and similarly we refer to the nested error linear model with the errors following normal distributions as the *normal model*.

1.3 Thesis Outline

In Chapter 2, we provide a brief introduction to the SN distribution. Survey statisticians are not very familiar with this distribution class so we discuss its main properties. The SN distribution has an asymmetric shape controlled by a shape parameter. When the parameter is equal to zero, it becomes the normal distribution. Therefore the SN distribution is a class of distribution which contains the normal distribution as a special case. When dealing with a distribution, it is nice to have closure properties

which ensure that basic operations such as joint distribution of independent variables, marginal distribution, and conditional distributions result in distributions belonging to the same stochastic family. A larger stochastic family, containing the SN, called closed-skew normal (CSN) distribution features those properties and is used in the next chapters to develop the predictors.

In Chapter 3, we provide the joint distribution of the population vector of interest under the SN model. To construct the empirical best prediction, it is necessary to estimate the parameters of the SN model. We provide a maximum likelihood (ML) method for estimating the parameters of the SN model. We also discuss using a data augmentation technique called data cloning for parameter estimation.

In Chapter 4, we derive several estimators for small area linear and complex parameters assuming the one-fold SN model. We show the best prediction (BP) and the BLUP estimators for small area linear parameters. Following the methodology developed by Molina and Rao (2010), we derive the EB estimator of small area complex parameters. We call this estimator the empirical quasi-best (EQB) estimator because it does not have a closed form and uses Monte Carlo approximation. Given the complexity of the EQB estimator, we propose the simpler conditional EQB which is based on the idea of predicting $\mathbf{Y}_r|\mathbf{y}_s$. The ELL method, as proposed by Elbers et al. (2003), does not assign correctly the area effects. We propose a modification to the ELL method to correctly assign the area effects and improve the MSEs' performance. We run a simulation study, using the FGT poverty measures as the complex parameters of interest, to assess the performance in terms of MSE of the best predictors compared to the Molina-Rao normality-based estimator, the ELL estimators, and the direct estimator.

In Chapter 5, we propose a hierarchical Bayes (HB) approach for the one-fold

SN model. HB methods have the advantage that they do not require the use of bootstrap. This is an important factor because the EB method requires predicting the characteristics of interest for most of the census (non-sampled units) and the prediction is repeated several times to derive a Monte Carlo approximation of the expected complex parameter. With that limitation in mind, Molina et al. (2014) developed a simple and fast hierarchical approach for the Molina-Rao model by avoiding the use of Markov Chain Monte Carlo (MCMC) techniques. Our HB approach is an extension of their method to the SN model.

In Chapter 6, we derive the best prediction and the best linear unbiased prediction (BLUP) estimators for linear parameters under the two-fold nested error model with the random components following SN distributions.

In Chapter 7, we present a more general version of the Rao-Yu model by dropping the stationarity assumption under a finite time series setup. This general Rao-Yu model does not suffer from the discontinuity issue when the autoregressive (AR) parameter is equal to 1 (random walk model). This is a unified approach that combines the AR(1) based method proposed by Rao and Yu (1992, 1994) and the random walk model proposed by Datta et al. (2002). ML and REML techniques are used to estimate the parameters of the model. We provide the MSE estimation by following Prasad and Rao (1990) approach.

Finally in Chapter 8, we conclude by discussing some possible extensions and directions for future research.

Chapter 2

Review of Skew-Normal (SN) Distributions

In this chapter we give a brief introduction of the skew-normal (SN) family of distributions. This class of distribution is not much used by the small area estimation researchers; nevertheless the SN distribution can be a good candidate for relaxing the assumption of normality. When the normality assumption is too restrictive to accommodate the asymmetry in the data, the SN class of distributions may be a good alternative to the normal family. One advantage of this class of distribution is the fact that it contains the normal family as a particular case. The second advantage is the closure properties of its generalization called the closed skew-normal (CSN) for the conditional densities, the marginal densities, the sum and joint distribution of independent random vectors. Genton (2004) provides a detailed review of the SN distributions and their applications.

2.1 The Univariate SN Distribution

The univariate SN density was first introduced by Roberts (1966) as an example of weighted densities. Later Azzalini (1985) formally introduces this distribution as a

class of probability density functions (pdf) and studied its properties. Azzalini (1985) defines a univariate SN density as follows:

Definition 2.1. *A random variable Z is said to follow SN distribution with parameter λ , written $Z \sim SN(\lambda)$, if its density function is*

$$f(z; \lambda) = 2\phi(z)\Phi(\lambda z) \quad \lambda, z \in \mathbb{R} \quad (2.1)$$

where $\phi(z)$ and $\Phi(z)$ denote respectively the $N(0, 1)$ probability density function (pdf) and the cumulative distribution function (cdf).

The parameter λ controls the skewness. Some elementary properties of the univariate distribution density given in (2.1) are:

- (a) If $\lambda = 0$ then $SN(\lambda = 0) \equiv N(0, 1)$,
- (b) If $Z \sim SN(\lambda)$ then $-Z \sim SN(-\lambda)$,
- (c) The moment generating function (mgf) of Z is $M(t) = 2\exp(t^2/2)\Phi(\delta t)$ where $\delta = \frac{\lambda}{\sqrt{1+\lambda^2}}$. The first and second central moments obtained from the mgf are:
 $\mathbb{E}(Z) = \delta\sqrt{\frac{2}{\pi}}, \quad \text{Var}(Z) = 1 - \frac{2}{\pi}\delta^2.$
- (d) If $Z \sim SN(\lambda)$ then $Z^2 \sim \chi_1^2$ for any value of λ .

The properties (a) and (b) are direct consequences of the definition 2.1. The mgf is obtained by noticing that if U is a $N(0, 1)$ random variable then $E\{\Phi(hU + k)\} = \Phi(k/\sqrt{1+h^2})$ for any real h, k .

The univariate SN family extends the normal distribution by introducing some skewness through the parameter λ . In other words, the SN distribution is a larger class of stochastic process which has the normal family as a special case. This property supports the use of the SN distribution as an alternative when the symmetric

property of the normal family is not appropriate. Throughout the book “Skew-elliptical distributions and their applications: A journey beyond normality”, the authors give numerous examples of applications using the SN distribution when the assumption of normality is not appropriate.

The skewness measure $\kappa = \mathbb{E} \left[\left(\frac{X-\mu}{\sigma} \right)^3 \right]$ is used to quantify the asymmetry of the distribution associated to the random variable X with mean μ and variance σ^2 . The skewness measure of the SN distribution (2.1) reduces to $\kappa = \text{sign}(\lambda) \frac{4-\pi}{2} \left(\frac{2\delta^2}{\pi-2\delta^2} \right)^{3/2}$. Note that the maximum absolute value of the skewness measure κ is approximately 0.995 and is achieved as the absolute value of δ tends to 1. Therefore, the SN distribution cannot accurately approximate a distribution with a measure of skewness much larger than 1. Some transformation may be necessary prior to approximating a highly skewed ($\kappa > 1$) distribution by a SN.

The SN distribution is fairly new and not very common in the statistical literature. Therefore the common statistical softwares do not include the SN among their standard probability distributions. To my knowledge, as of today, only the statistical software R provides at least one package that allows generation of SN distributions. The R package is called SN and was developed by Adelchi Azzalini (for more details visit the following website: <http://cran.r-project.org/web/packages/sn/index.html>). In this context, Proposition 2.1 below is very important for generating SN random variables because it defines SN using normal distributions. This transformation method represents the SN random variable as a weighted combination of a normal random variable and a half-normal random variable.

Proposition 2.1. *If X_0 and X_1 are independent $N(0, 1)$ normal variables and $\delta \in (-1, 1)$, then*

$$Z = \delta |X_0| + (1 - \delta^2)^{1/2} X_1 \quad (2.2)$$

is $SN(\lambda(\delta))$, where

$$\lambda(\delta) = \delta/\sqrt{1 - \delta^2} \quad \text{equivalently} \quad \delta(\lambda) = \lambda/\sqrt{1 + \lambda^2} \quad (2.3)$$

Note that λ can be any real numbers while δ varies in $(-1, 1)$. Because equation (2.3) is a one-on-one function, it is perfectly equivalent to define the SN using $\lambda(\delta)$ or $\delta(\lambda)$. For simplicity, we will refer to $\lambda(\delta)$ and $\delta(\lambda)$ as respectively λ and δ . The expression in Proposition 2.1 shows clearly that when $\lambda = 0$ the SN distribution simplifies to the normal. Higher absolute values of λ correspond to more departure from normality meaning less symmetrical distribution. A positive sign of λ indicates a skewness on the right while a negative sign leads to a skewness on the left. The graph 2.1 shows two SN densities corresponding to values of λ equal to 2 and 6.

Figure 2.1: Density curves of two SN distributions.

For a general use of the SN distribution, it is necessary to extend the definition 2.1 by adding a location and scale parameter. Proposition 2.2 provides the density

function of the SN distribution with location and scale parameter.

Proposition 2.2. *If $Z \sim SN(\lambda)$ and $Y = \mu + \sigma Z$, with $\mu \in \mathbb{R}$ and $\sigma > 0$ then the density function of the variable Y , written $Y \sim SN(\mu, \sigma^2, \lambda)$, is:*

$$f(y; \mu, \sigma, \lambda) = \frac{2}{\sigma} \phi\left(\frac{y - \mu}{\sigma}\right) \Phi\left(\lambda \frac{y - \mu}{\sigma}\right) \quad (2.4)$$

Proposition 2.2 is a direct consequence of the definition 2.1 and the transformation method (2.5). The method of transformation is widely used to find the probability distribution of $Y = h(X)$ when the probability distribution of X , f_X , is known and the function h is monotone (either increasing or decreasing). Let X have probability density function $f_X(x)$. If $h(x)$ is either increasing or decreasing in x , then $Y = h(X)$ has density function

$$f_Y(y) = f_X(h^{-1}(y)) \left| \frac{dh^{-1}(y)}{dy} \right|. \quad (2.5)$$

2.2 The Multivariate Skew-Normal (MSN) Distribution

In many applications, it is necessary to deal with more than one variable at the same time. For instance, joint distributions are frequently used for statistical inference. In this section, we present the multivariate extension of the univariate SN distribution discussed in the previous section. Azzalini (1985) proposed a multivariate extension of definition 2.1 but the resulting probability density function lacks some desirable properties such as closure under marginal operations. Later, Azzalini (1996) explored the nature of the SN family to develop a multivariate extension that has many desirable properties. He considered an k -dimensional multivariate random variable $\mathbf{Z}$ such that each component Z_j is univariate SN. To construct the multivariate extension from Proposition 2.1, consider k -dimensional normal random variable $X = (X_1, \dots, X_k)^T$

with a standardized marginal distribution, independent of $X_0 \sim N(0, 1)$, thus

$$\begin{pmatrix} X_0 \\ \mathbf{X} \end{pmatrix} \sim N_{k+1} \left\{ 0, \begin{bmatrix} 1 & 0 \\ 0 & \mathbf{\Psi} \end{bmatrix} \right\} \quad (2.6)$$

where $\mathbf{\Psi}$ is a $k \times k$ correlation matrix. If $(\delta_1, \dots, \delta_k)$ are in $(-1, 1)$, define

$$Z_j = \delta_j |X_0| + (1 - \delta_j^2)^{1/2} X_j, \quad j = 1, \dots, k \quad (2.7)$$

Given that $\mathbf{\Psi}$ is a correlation matrix (i.e. the diagonal is composed of 1's), it follows from Proposition 2.1 that Z_j follows a SN distribution. Therefore, Proposition 2.3 shows that $\mathbf{Z} = (Z_1, \dots, Z_k)^T$ has a multivariate SN distribution.

Proposition 2.3. *The k -dimensional random vector $\mathbf{Z} = (Z_1, \dots, Z_k)^T$, where Z_j is defined as in equation (2.7) with $j \in (1, \dots, k)$, follows a MSN distribution. The density function of $\mathbf{Z}$, written as $\mathbf{Z} \sim SN_k(\mathbf{\Psi}, \mathbf{\lambda})$, is*

$$f_{\mathbf{Z}}(\mathbf{z}; \mathbf{\Psi}, \mathbf{\lambda}) = 2\phi_k(\mathbf{z}; \mathbf{\Omega}) \Phi_1(\mathbf{\alpha}^T \mathbf{z}), \quad \mathbf{z} \in \mathbb{R} \quad (2.8)$$

where,

$$\mathbf{\alpha}^T = \frac{\mathbf{\lambda}^T \mathbf{\Psi}^{-1} \mathbf{\Delta}^{-1}}{(1 + \mathbf{\lambda}^T \mathbf{\Psi}^{-1} \mathbf{\lambda})^{1/2}}, \quad (2.9)$$

$$\mathbf{\Delta} = \text{diag}((1 - \delta_1^2)^{1/2}, \dots, (1 - \delta_k^2)^{1/2}),$$

$$\mathbf{\lambda} = (\lambda(\delta_1), \dots, \lambda(\delta_k))^T, \quad (2.10)$$

$$\mathbf{\Omega} = \mathbf{\Delta} (\mathbf{\Psi} + \mathbf{\lambda} \mathbf{\lambda}^T) \mathbf{\Delta}, \quad (2.11)$$

Note that it is equivalent to characterize the SN distribution using $(\mathbf{\Psi}, \mathbf{\lambda})$ or $(\mathbf{\Omega}, \mathbf{\alpha})$. For instance one can easily show, using equations from Proposition 2.3, that $SN_1(1, \lambda) \equiv SN_1(1, \alpha = \lambda)$. Arellano-Valle et al. (2005) considered a special case

of the fundamental MSN distribution that results from taking $\Psi = \mathbf{I}_k$, reflecting the situation when the Z'_j 's are uncorrelated. In the rest of this work, we will use the same special case $\Psi = \mathbf{I}_k$. For illustration, let's consider a bi-variate SN vector $\mathbf{Z} = (X, Y) \sim SN_2(\boldsymbol{\lambda})$ with $\boldsymbol{\lambda} = (-1, 3)^T$.

Figure 2.2: Density surface of a bi-SN.

2.3 The Closed Skew-Normal (CSN) Distribution

In the case of the normal distribution, the multivariate extension is widely used in part because of its closure properties. These closure properties ensure that many distributions resulting from basic operations such as the conditional, marginal, joint of independent variables, sum of independent variables remained in the same normal family distribution. When dealing with the normal distribution, these stability rules (closure properties) are very useful for deriving density functions. The MSN defined in Proposition 2.3 unfortunately lacks several important closure properties particularly under marginalization and conditioning. One more general multivariate distribution, the CSN Distribution, ensures more closure properties. In this section, we present some important closure properties of the CSN. The results presented here are based on the comprehensive review of the CSN distribution presented by Graciela Gonzales-Farias et al. in Chapter 2 of the book by Genton (2004).

Definition 2.2. Consider $p \geq 1$, $q \geq 1$, $\boldsymbol{\mu} \in \mathbb{R}^p$, $\boldsymbol{\nu} \in \mathbb{R}^q$, $\mathbf{D}$ an arbitrary $q \times p$ matrix, $\boldsymbol{\Sigma}$, and $\boldsymbol{\Gamma}$ positive definite matrices of dimensions $p \times p$ and $q \times q$, respectively. Then the probability density function of the CSN distribution is given by:

$$f_{p,q}(\mathbf{y}) = C \phi_p(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \Phi_q(\mathbf{D}(\mathbf{y} - \boldsymbol{\mu}); \boldsymbol{\nu}, \boldsymbol{\Gamma}), \quad \mathbf{y} \in \mathbb{R}^p, \quad (2.12)$$

with

$$C^{-1} = \Phi_q(\mathbf{0}; \boldsymbol{\nu}, \boldsymbol{\Gamma} + \mathbf{D}\boldsymbol{\Sigma}\mathbf{D}^T) \quad (2.13)$$

where ϕ_p and Φ_p are respectively the pdf and the cdf of the p -dimensional normal distribution. We denote this distribution by $\mathbf{y} \sim \text{CSN}_{p,q}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{D}, \boldsymbol{\nu}, \boldsymbol{\Gamma})$ and if $p = q$ we will denote it by $\text{CSN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{D}, \boldsymbol{\nu}, \boldsymbol{\Gamma})$.

The MSN defined in Proposition 2.3 is a special case of the more general CSN distribution. The MSN distribution is obtained by taking $\boldsymbol{\mu} = \mathbf{0}$, $\boldsymbol{\Sigma} = \boldsymbol{\Omega}$, $\boldsymbol{\alpha} =$

$\mathbf{\Gamma}^{-1/2}\mathbf{D}^T$, $\boldsymbol{\nu} = \mathbf{0}$, and $q = 1$. The generalization expands the shape parameter from a vector ($\boldsymbol{\lambda}$), in the case of the MSN to a matrix ($\mathbf{D}$) for the CSN. The generalization also adds two extra parameters $\boldsymbol{\nu}$ and $\mathbf{\Gamma}$. The parameter $\boldsymbol{\nu}$ ensures the closure property for conditional densities and $\mathbf{\Gamma}$ ensures the closure property for marginal densities. The constant C , which replaces the value 2 of the MSN, adds computational complexities when dealing with the CSN. Parameter estimation is much more complex in the context of the CSN distribution because of the extra parameters, the increase of the dimension of the shape parameter and the inclusion of the cdf of the multivariate normal distribution Φ_q .

2.3.1 Generation of Random Vectors from the CSN Distribution

In order to use simulations methods to study a distribution one must be able to generate a random variable from that distribution. Most statistical software provide standard procedures to generate many of the common probabilistic densities. However the CSN family has not yet been covered by most software. The open source software R addresses the univariate SN, the MSN, and skew-t distributions through a package called `sn` created by Azzalini (2011) but not the CSN distribution.

Copas and Li (1997) considers the following model:

$$\mathbf{W} = \boldsymbol{\mu} + \mathbf{U}_1$$

$$\mathbf{V} = -\boldsymbol{\nu} + \mathbf{D}\mathbf{U}_1 + \mathbf{U}_2$$

where $\mathbf{U}_1 \sim N_p(\mathbf{0}, \mathbf{\Sigma})$ and $\mathbf{U}_2 \sim N_q(\mathbf{0}, \mathbf{\Gamma})$ are independent random vectors, and $\mathbf{D}(q \times p)$ is an arbitrary matrix, $\boldsymbol{\mu} \in \mathbb{R}^p$, $\boldsymbol{\nu} \in \mathbb{R}^q$, and $\mathbf{\Gamma}(q \times q) > \mathbf{0}$. It is easy to see

that the joint distribution of $\mathbf{W}$ and $\mathbf{V}$ is:

$$\begin{pmatrix} \mathbf{W} \\ \mathbf{V} \end{pmatrix} \sim N_{p+q} \left\{ \begin{pmatrix} \boldsymbol{\mu} \\ -\boldsymbol{\nu} \end{pmatrix}, \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma}\mathbf{D}^T \\ \mathbf{D}\boldsymbol{\Sigma} & \boldsymbol{\Gamma} + \mathbf{D}\boldsymbol{\Sigma}\mathbf{D}^T \end{bmatrix} \right\}$$

Arellano-Valle et al. (2002) showed that

$$f_{\mathbf{W}|\mathbf{V} \geq \mathbf{0}}(\mathbf{w}|\mathbf{v} \geq \mathbf{0}) = \frac{f_{\mathbf{W}}(\mathbf{w})}{P(\mathbf{V} \geq \mathbf{0})} P(\mathbf{V} \geq \mathbf{0}|\mathbf{W} = \mathbf{w})$$

Thus the conditional density of $\mathbf{W}$ given $\mathbf{V} \geq \mathbf{0}$ is:

$$f_{\mathbf{W}|\mathbf{V} \geq \mathbf{0}}(\mathbf{w}|\mathbf{v} \geq \mathbf{0}) = C \phi_p(\mathbf{w}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \Phi_q(\mathbf{D}(\mathbf{w} - \boldsymbol{\mu}); \boldsymbol{\nu}, \boldsymbol{\Gamma}), \quad (2.14)$$

where C is given in equation (2.13). The density (2.14) is the same as (2.12); therefore the model by Copas and Li (1997) described above provides a natural way of generating random variables from CSN distributions.

2.3.2 Moment Generating Function (mgf) of CSN

Many properties of the CSN are proved using the mgf.

Proposition 2.4. *If $\mathbf{Y} \sim CSN_{p,q}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{D}, \boldsymbol{\nu}, \boldsymbol{\Gamma})$, then the mgf of $\mathbf{Y}$ is:*

$$M_{\mathbf{Y}}(\mathbf{t}) = \frac{\Phi_q(\mathbf{D}\boldsymbol{\Sigma}\mathbf{t}; \boldsymbol{\nu}, \boldsymbol{\Gamma} + \mathbf{D}\boldsymbol{\Sigma}\mathbf{D}^T)}{\Phi_q(\mathbf{0}; \boldsymbol{\nu}, \boldsymbol{\Gamma} + \mathbf{D}\boldsymbol{\Sigma}\mathbf{D}^T)} e^{\mathbf{t}^T \boldsymbol{\mu} + \frac{1}{2} \mathbf{t}^T \boldsymbol{\Sigma} \mathbf{t}}, \quad \mathbf{t} \in \mathbb{R}^p \quad (2.15)$$

Proof. From the definition of the mgf, we have:

$$M_{\mathbf{Y}}(\mathbf{t}) = \mathbb{E}(e^{\mathbf{t}^T \mathbf{y}}) = C \int_{\mathbb{R}^q} e^{\mathbf{t}^T \mathbf{y}} \phi_p(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \Phi_q(\mathbf{D}(\mathbf{y} - \boldsymbol{\mu}); \boldsymbol{\nu}, \boldsymbol{\Gamma}) d\mathbf{y},$$

The fact

$$e^{\mathbf{t}^T \mathbf{y}} \phi_p(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = e^{\mathbf{t}^T \boldsymbol{\mu} + (1/2) \mathbf{t}^T \boldsymbol{\Sigma} \mathbf{t}} \phi_p(\mathbf{y}; \boldsymbol{\mu} + \boldsymbol{\Sigma} \mathbf{t}, \boldsymbol{\Sigma}),$$

gives

$$\begin{aligned} M_{\mathbf{Y}}(\mathbf{t}) &= C e^{\mathbf{t}^T \boldsymbol{\mu} + (1/2) \mathbf{t}^T \boldsymbol{\Sigma} \mathbf{t}} \int_{\mathbb{R}^q} \phi_p(\mathbf{y}; \boldsymbol{\mu} + \boldsymbol{\Sigma} \mathbf{t}, \boldsymbol{\Sigma}) \Phi_q(\mathbf{D}(\mathbf{y} - \boldsymbol{\mu}); \boldsymbol{\nu}, \boldsymbol{\Gamma}) d\mathbf{y}, \\ &= C e^{\mathbf{t}^T \boldsymbol{\mu} + (1/2) \mathbf{t}^T \boldsymbol{\Sigma} \mathbf{t}} \int_{\mathbb{R}^q} \phi_p(\mathbf{y}; \boldsymbol{\mu} + \boldsymbol{\Sigma} \mathbf{t}, \boldsymbol{\Sigma}) \Phi_q(\mathbf{D}(\mathbf{y} - \boldsymbol{\mu} - \boldsymbol{\Sigma} \mathbf{t}); \boldsymbol{\nu} - \mathbf{D} \boldsymbol{\Sigma} \mathbf{t}, \boldsymbol{\Gamma}) d\mathbf{y}. \end{aligned}$$

Because (2.12) integrates to 1, we get:

$$\begin{aligned} &\int_{\mathbb{R}^q} \phi_p(\mathbf{y}; \boldsymbol{\mu} + \boldsymbol{\Sigma} \mathbf{t}, \boldsymbol{\Sigma}) \Phi_q(\mathbf{D}(\mathbf{y} - \boldsymbol{\mu} - \boldsymbol{\Sigma} \mathbf{t}); \boldsymbol{\nu} - \mathbf{D} \boldsymbol{\Sigma} \mathbf{t}, \boldsymbol{\Gamma}) d\mathbf{y} \\ &= \Phi_q(\mathbf{0}; \boldsymbol{\nu} - \mathbf{D} \boldsymbol{\Sigma} \mathbf{t}, \boldsymbol{\Gamma} + \mathbf{D} \boldsymbol{\Sigma} \mathbf{D}^T) \\ &= \Phi_q(\mathbf{D} \boldsymbol{\Sigma} \mathbf{t}; \boldsymbol{\nu}, \boldsymbol{\Gamma} + \mathbf{D} \boldsymbol{\Sigma} \mathbf{D}^T). \end{aligned}$$

□

The moments can be obtained from the mgf. However the expression of the derivatives obtained from equation (2.15) can be very complex. In practice, the expression of the first and second derivatives of the mgf are reasonable enough to allow deduction of the first and second moments of the CSN. The mathematical expressions of the higher moments of the CSN are too complex and therefore their practical use is very limited.

Corollary 2.1. *If $\mathbf{Y} \sim CSN_{p,q}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{D}, \boldsymbol{\nu}, \boldsymbol{\Gamma})$, then the mean and the variance of $\mathbf{Y}$*

are:

$$\mathbb{E}(\mathbf{Y}) = \boldsymbol{\mu} + \frac{\Phi_q^{(1)}(\mathbf{0}; \boldsymbol{\nu}, \boldsymbol{\Gamma} + \mathbf{D}\boldsymbol{\Sigma}\mathbf{D}^T)}{\Phi_q(\mathbf{0}; \boldsymbol{\nu}, \boldsymbol{\Gamma} + \mathbf{D}\boldsymbol{\Sigma}\mathbf{D}^T)} \quad (2.16)$$

$$\mathbb{V}ar(\mathbf{Y}) = \boldsymbol{\Sigma} + \frac{\Phi_q^{(2)}(\mathbf{0}; \boldsymbol{\nu}, \boldsymbol{\Gamma} + \mathbf{D}\boldsymbol{\Sigma}\mathbf{D}^T)}{\Phi_q(\mathbf{0}; \boldsymbol{\nu}, \boldsymbol{\Gamma} + \mathbf{D}\boldsymbol{\Sigma}\mathbf{D}^T)} - \mathbb{E}(\mathbf{Y} - \boldsymbol{\mu}) \mathbb{E}(\mathbf{Y} - \boldsymbol{\mu})^T \quad (2.17)$$

Proof. The proof is straightforward by noting that $\mathbb{E}(\mathbf{Y}^k) = \frac{\partial^k}{\partial \mathbf{t}^k} M_{\mathbf{Y}}(\mathbf{t})|_{\mathbf{t}=\mathbf{0}}$. The mgf $M_{\mathbf{Y}}(\mathbf{t})$ is provided by Proposition 2.4.

□

In the remaining part of this section, we will be showing some of the most important closure properties of the CSN.

2.3.3 Linear Transformations

It is straightforward to show that the CSN is closed under scalar multiplications and translations, using the mgf. Thus, If $\mathbf{y} \sim CSN_{p,q}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{D}, \boldsymbol{\nu}, \boldsymbol{\Gamma})$ then it follows that:

- ✓ If $\mathbf{a}$ is a real number then $\mathbf{a}\mathbf{y} \sim CSN_{p,q}(\mathbf{a}\boldsymbol{\mu}, \mathbf{a}^2\boldsymbol{\Sigma}, \mathbf{a}^{-1}\mathbf{D}, \boldsymbol{\nu}, \boldsymbol{\Gamma})$
- ✓ If $\mathbf{b} \in \mathbb{R}^p$ is a real number then $\mathbf{y} + \mathbf{b} \sim CSN_{p,q}(\boldsymbol{\mu} + \mathbf{b}, \boldsymbol{\Sigma}, \mathbf{D}, \boldsymbol{\nu}, \boldsymbol{\Gamma})$

In this section, linear transformation is generalized to the cases of full row rank and full column rank transformations. Closure under full row rank transformations is fundamental because it allows us to derive closure properties under marginal and conditional distributions.

Proposition 2.5. *Let $\mathbf{Y} \sim CSN_{p,q}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{D}, \boldsymbol{\nu}, \boldsymbol{\Gamma})$ and let $\mathbf{A}$ be an $n \times p$ ($n \leq p$) matrix of rank n . Then*

$$\mathbf{A}\mathbf{Y} \sim CSN_{n,q}(\boldsymbol{\mu}_{\mathbf{A}}, \boldsymbol{\Sigma}_{\mathbf{A}}, \mathbf{D}_{\mathbf{A}}, \boldsymbol{\nu}, \boldsymbol{\Gamma}_{\mathbf{A}}),$$

where

$$\boldsymbol{\mu}_A = \mathbf{A}\boldsymbol{\mu}, \quad \boldsymbol{\Sigma}_A = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T, \quad \mathbf{D}_A = \mathbf{D}\boldsymbol{\Sigma}\mathbf{A}^T\boldsymbol{\Sigma}_A^{-1},$$

$$\boldsymbol{\Gamma}_A = \boldsymbol{\Gamma} + \mathbf{D}\boldsymbol{\Sigma}\mathbf{D}^T - \mathbf{D}\boldsymbol{\Sigma}\mathbf{A}^T\boldsymbol{\Sigma}_A^{-1}\mathbf{A}\boldsymbol{\Sigma}\mathbf{D}^T.$$

Proof. For $\mathbf{t} \in \mathbb{R}^n$ the mgf of $\mathbf{A}\mathbf{Y}$ is given by:

$$\begin{aligned} M_{\mathbf{A}\mathbf{Y}}(\mathbf{t}) &= M_{\mathbf{A}\mathbf{Y}}(\mathbf{t}) \\ &= \frac{\Phi_q(\mathbf{D}\boldsymbol{\Sigma}\mathbf{A}^T\mathbf{t}; \boldsymbol{\nu}, \boldsymbol{\Gamma} + \mathbf{D}\boldsymbol{\Sigma}\mathbf{D}^T)}{\Phi_q(\mathbf{0}; \boldsymbol{\nu}, \boldsymbol{\Gamma} + \mathbf{D}\boldsymbol{\Sigma}\mathbf{D}^T)} e^{\mathbf{t}^T \mathbf{A}\boldsymbol{\mu} + \frac{1}{2} \mathbf{t}^T \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T \mathbf{t}}. \end{aligned}$$

Given that $\boldsymbol{\Sigma}$ is positive definite and $\text{rank}(\mathbf{A})=n$, we have that $\boldsymbol{\Sigma}_A = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T$ is a non-singular matrix and since:

$$\Phi_q(\mathbf{D}\boldsymbol{\Sigma}\mathbf{A}^T\mathbf{t}; \boldsymbol{\nu}, \boldsymbol{\Gamma} + \mathbf{D}\boldsymbol{\Sigma}\mathbf{D}^T) = \Phi_q(\mathbf{D}\boldsymbol{\Sigma}\mathbf{A}^T\boldsymbol{\Sigma}_A^{-1}\boldsymbol{\Sigma}_A\mathbf{t}; \boldsymbol{\nu}, \boldsymbol{\Gamma} + \mathbf{D}\boldsymbol{\Sigma}\mathbf{D}^T),$$

By noting that $\boldsymbol{\Gamma} + \mathbf{D}\boldsymbol{\Sigma}\mathbf{D}^T = \boldsymbol{\Gamma}_A + \mathbf{D}\boldsymbol{\Sigma}\mathbf{A}^T\boldsymbol{\Sigma}_A^{-1}\mathbf{A}\boldsymbol{\Sigma}\mathbf{D}^T$, we get:

$$M_{\mathbf{A}\mathbf{Y}}(\mathbf{t}) = \frac{\Phi_q(\mathbf{D}_A\boldsymbol{\Sigma}_A\mathbf{t}; \boldsymbol{\nu}, \boldsymbol{\Gamma}_A + \mathbf{D}_A\boldsymbol{\Sigma}_A\mathbf{D}_A^T)}{\Phi_q(\mathbf{0}; \boldsymbol{\nu}, \boldsymbol{\Gamma}_A + \mathbf{D}_A\boldsymbol{\Sigma}_A\mathbf{D}_A^T)} e^{\mathbf{t}^T \boldsymbol{\mu}_A + \frac{1}{2} \mathbf{t}^T \boldsymbol{\Sigma}_A \mathbf{t}}.$$

where $\boldsymbol{\mu}_A$, $\boldsymbol{\Sigma}_A$, $\mathbf{D}_A$, and $\boldsymbol{\Gamma}_A$ as defined in Proposition 2.5. □

One special case of Proposition 2.5 is linear combinations of the elements of the vector $\mathbf{Y}$. This special case corresponds to $n = 1$. Therefore, if $\mathbf{a} \neq \mathbf{0}$ is an arbitrary p-vector in $\mathbb{R}^p$, then

$$\mathbf{a}^T \mathbf{Y} \sim CSN_{1,q}(\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a, \mathbf{D}_a, \boldsymbol{\nu}, \boldsymbol{\Gamma}_a),$$

where

$$\boldsymbol{\mu}_a = \mathbf{a}^T \boldsymbol{\mu}, \quad \boldsymbol{\Sigma}_a = \mathbf{a}^T \boldsymbol{\Sigma} \mathbf{a}, \quad \mathbf{D}_a = \mathbf{D} \boldsymbol{\Sigma} \mathbf{a} \boldsymbol{\Sigma}_a^{-1},$$

$$\boldsymbol{\Gamma}_a = \boldsymbol{\Gamma} + \mathbf{D}\boldsymbol{\Sigma}\mathbf{D}^T - \mathbf{D}\boldsymbol{\Sigma}\mathbf{a}\mathbf{a}^T\boldsymbol{\Sigma}\mathbf{D}^T\boldsymbol{\Sigma}_a^{-1}.$$

Next, we consider linear transformations $\mathbf{A}\mathbf{Y}$ where $\mathbf{A}$ is an $n \times p$ matrix, with $n \geq p$. Then the matrix $\mathbf{A}$ is a full column rank n instead of a full row rank. An equivalent to Proposition 2.5 can be stated as follows:

Proposition 2.6. *Let $\mathbf{Y} \sim CSN_{p,q}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{D}, \boldsymbol{\nu}, \boldsymbol{\Gamma})$ and let $\mathbf{A}$ be an $n \times p$ ($n \geq p$) matrix of rank p . Then*

$$\mathbf{A}\mathbf{Y} \sim SSN_{n,q}(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T, \mathbf{D}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T, \boldsymbol{\nu}, \boldsymbol{\Gamma}),$$

where $SSN_{n,q}$ is used to denote a singular SN distribution.

Proof. For $\mathbf{t} \in \mathbb{R}^n$ the mgf of $\mathbf{A}\mathbf{Y}$ is given by:

$$\begin{aligned} M_{\mathbf{A}\mathbf{Y}}(\mathbf{t}) &= M_{\mathbf{A}\mathbf{Y}}(\mathbf{t}) \\ &= \frac{\Phi_q(\mathbf{D}\boldsymbol{\Sigma}\mathbf{A}^T\mathbf{t}; \boldsymbol{\nu}, \boldsymbol{\Gamma} + \mathbf{D}\boldsymbol{\Sigma}\mathbf{D}^T)}{\Phi_q(\mathbf{0}; \boldsymbol{\nu}, \boldsymbol{\Gamma} + \mathbf{D}\boldsymbol{\Sigma}\mathbf{D}^T)} e^{\mathbf{t}^T\mathbf{A}\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}^T\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T\mathbf{t}} \\ &= \frac{\Phi_q(\mathbf{DB}(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)\mathbf{t}; \boldsymbol{\nu}, \boldsymbol{\Gamma} + (\mathbf{DB})(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)(\mathbf{DB})^T)}{\Phi_q(\mathbf{0}; \boldsymbol{\nu}, \boldsymbol{\Gamma} + (\mathbf{DB})(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)(\mathbf{DB})^T)} e^{\mathbf{t}^T\mathbf{A}\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}^T\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T\mathbf{t}} \end{aligned}$$

where $\mathbf{B} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$. □

Note that $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T$ is singular, therefore the distribution of $\mathbf{A}\mathbf{Y}$ cannot have a density. We call this distribution the singular SN distribution and we use the notation $SSN_{n,q}$.

2.3.4 Marginal and Conditional Distributions

The closure properties of marginal and conditional distributions of the CSN are direct consequences of Proposition 2.5. These properties will be used to derive the distribution of the small-area predictors in the next chapter. Consider the following

partitions of $\mathbf{Y}$, $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, $\mathbf{D}$, $\boldsymbol{\nu}$, and $\boldsymbol{\Gamma}$:

$$\begin{aligned}\mathbf{Y} &= \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix}, \text{ with } \mathbf{Y}_1 \text{ and } \mathbf{Y}_2 \text{ respectively } k \times 1 \text{ and } (p-k) \times 1 \text{ vectors.} \\ \boldsymbol{\mu} &= \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \text{ with } \boldsymbol{\mu}_1 \text{ and } \boldsymbol{\mu}_2 \text{ respectively } k \times 1 \text{ and } (p-k) \times 1 \text{ vectors.} \\ \boldsymbol{\Sigma} &= \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}, \text{ with } \boldsymbol{\Sigma}_{11}, \boldsymbol{\Sigma}_{12}, \boldsymbol{\Sigma}_{21} \text{ and } \boldsymbol{\Sigma}_{22} \text{ respectively } k \times k, k \times (p-k), \\ &\quad (p-k) \times k \text{ and } (p-k) \times (p-k) \text{ matrices.} \\ \mathbf{D} &= \begin{bmatrix} \mathbf{D}_1 & \mathbf{D}_2 \end{bmatrix}, \text{ with } \mathbf{D}_1 \text{ and } \mathbf{D}_2 \text{ respectively } q \times k \text{ and } q \times (p-k) \text{ matrices.}\end{aligned}$$

Proposition 2.7 (Marginal Density). *Let $\mathbf{Y} \sim CSN_{p,q}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{D}, \boldsymbol{\nu}, \boldsymbol{\Gamma})$ and be partitioned as above. Then the marginal distribution of $\mathbf{Y}_1$ is:*

$$CSN_{k,q}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}, \mathbf{D}^*, \boldsymbol{\nu}, \boldsymbol{\Gamma}^*),$$

where $\mathbf{D}^* = \mathbf{D}_1 + \mathbf{D}_2 \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1}$, $\boldsymbol{\Gamma}^* = \boldsymbol{\Gamma} + \mathbf{D}_2 \boldsymbol{\Sigma}_{22.1} \mathbf{D}_2^T$, $\boldsymbol{\Sigma}_{22.1} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12}$, and $\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}, \boldsymbol{\Sigma}_{12}, \boldsymbol{\Sigma}_{21}, \boldsymbol{\Sigma}_{22}, \mathbf{D}_1, \mathbf{D}_2$ as defined in the partitions above.

Proof. Consider $A = [\mathbf{I}_k \quad \mathbf{0}]$, with $\mathbf{I}_k$ a $k \times k$ identity matrix and $\mathbf{0}$ a $k \times (p-k)$ zero matrix. Then by applying Proposition 2.5, we get:

$$\mathbf{Y}_1 = A\mathbf{Y} \sim CSN_{k,q}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}, \mathbf{D}^*, \boldsymbol{\nu}, \boldsymbol{\Gamma}^*),$$

where $\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}, \mathbf{D}^*, \boldsymbol{\Gamma}^*$ as defined in the Proposition 2.7. □

Proposition 2.8 (Conditional Density). *Let $\mathbf{Y} \sim CSN_{p,q}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{D}, \boldsymbol{\nu}, \boldsymbol{\Gamma})$ and be partitioned as above. Then the conditional distribution of $\mathbf{Y}_2$ given $\mathbf{Y}_1 = \mathbf{y}_1$ is:*

$$CSN_{p-k,q}(\boldsymbol{\mu}_{2|1}, \boldsymbol{\Sigma}_{2|1}, \mathbf{D}_2, \boldsymbol{\nu}_{2|1}, \boldsymbol{\Gamma}),$$

where $\boldsymbol{\mu}_{2|1} = \boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{y}_1 - \boldsymbol{\mu}_1)$, $\boldsymbol{\Sigma}_{2|1} = \boldsymbol{\Sigma}_{22.1}$, $\boldsymbol{\nu}_{2|1} = \boldsymbol{\nu} - \mathbf{D}^*(\mathbf{y}_1 - \boldsymbol{\mu}_1)$, $\mathbf{D}^*$ as in Proposition 2.7, and $\boldsymbol{\mu}_1$, $\boldsymbol{\Sigma}_{11}$, $\boldsymbol{\Sigma}_{12}$, $\boldsymbol{\Sigma}_{21}$, $\boldsymbol{\Sigma}_{22}$, $\mathbf{D}_1$, $\mathbf{D}_2$ as defined in the partitions above.

Proof. This can be shown by direct computation of the conditional density of $\mathbf{Y}_2$ given $\mathbf{Y}_1 = \mathbf{y}_1$. $\square$

Note that the dimension q is not changed for the marginal and the conditional distributions. This result is true for all linear transformations.

2.3.5 Joint distribution of Independent CSN Random Vectors

In this section, we show closure properties under joint operations of independent CSN vectors. In the following text, we will use symbols $\oplus$ and $\otimes$ to indicate kronecker matrix product and the matrix direct sum operator. For any two matrices $\mathbf{A}_{m \times n}$ and $\mathbf{B}_{n \times m}$:

✓ The Kronecker product for $\mathbf{A}$ and $\mathbf{B}$ is defined by:

$$\mathbf{A} \otimes \mathbf{B} = (a_{ij}\mathbf{B}) = \begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \cdots & a_{2n}\mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & a_{m2}\mathbf{B} & \cdots & a_{mn}\mathbf{B} \end{bmatrix}.$$

✓ The direct sum for $\mathbf{A}$ and $\mathbf{B}$ is defined by: $\mathbf{A} \oplus \mathbf{B} = \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{bmatrix}$.

Note that: $\bigoplus_{i=1}^k \mathbf{A}_i = \mathbf{A}_1 \oplus \cdots \oplus \mathbf{A}_k$ and $\bigoplus_{i=1}^k \mathbf{A} = \mathbf{I}_k \otimes \mathbf{A}$, where $\mathbf{I}_k$ is the identity matrix.

Proposition 2.9 (Joint Distribution). *Let $\mathbf{Y}_1, \dots, \mathbf{Y}_n$ be independent random vectors with $\mathbf{Y}_i \sim \text{CSN}_{p_i, q_i}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \mathbf{D}_i, \boldsymbol{\nu}_i, \boldsymbol{\Gamma}_i)$, $i = 1, \dots, n$, then the joint distribution of*

$\mathbf{Y}_1, \dots, \mathbf{Y}_n$ is

$$\mathbf{Y} = (\mathbf{Y}_1^T, \dots, \mathbf{Y}_n^T)^T \sim CSN_{p^\dagger, q^\dagger}(\boldsymbol{\mu}^\dagger, \boldsymbol{\Sigma}^\dagger, \mathbf{D}^\dagger, \boldsymbol{\nu}^\dagger, \boldsymbol{\Gamma}^\dagger),$$

where:

$$\begin{aligned} p^\dagger &= \sum_{i=1}^n p_i, & q^\dagger &= \sum_{i=1}^n q_i, & \boldsymbol{\mu}^\dagger &= (\boldsymbol{\mu}_1^T, \dots, \boldsymbol{\mu}_n^T), & \boldsymbol{\Sigma}^\dagger &= \bigoplus_{i=1}^n \boldsymbol{\Sigma}_i, \\ \mathbf{D}^\dagger &= \bigoplus_{i=1}^n \mathbf{D}_i, & \boldsymbol{\nu}^\dagger &= (\boldsymbol{\nu}_1^T, \dots, \boldsymbol{\nu}_n^T)^T, & \boldsymbol{\Gamma}^\dagger &= \bigoplus_{i=1}^n \boldsymbol{\Gamma}_i. \end{aligned}$$

Proof. For $\mathbf{Y} = (\mathbf{Y}_1^T, \dots, \mathbf{Y}_n^T)^T$ with $\mathbf{Y}_i \in \mathbb{R}^{p_i}$, the density function of $\mathbf{Y}$ is given by:

$$\begin{aligned} f_{p,q}(\mathbf{y}) &= \prod_{i=1}^n f_{p_i, q_i}(\mathbf{y}_i; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \mathbf{D}_i, \boldsymbol{\nu}_i, \boldsymbol{\Gamma}_i) \\ &= \prod_{i=1}^n \frac{\phi_{p_i}(\mathbf{y}_i; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \Phi_{q_i}(\mathbf{D}_i(\mathbf{y}_i - \boldsymbol{\mu}_i); \boldsymbol{\nu}_i, \boldsymbol{\Gamma}_i)}{\Phi_{q_i}(\mathbf{0}; \boldsymbol{\nu}_i, \boldsymbol{\Gamma}_i + \mathbf{D}_i \boldsymbol{\Sigma}_i \mathbf{D}_i^T)} \\ &= \frac{\prod_{i=1}^n \phi_{p_i}(\mathbf{y}_i; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \Phi_{q_i}(\mathbf{D}_i(\mathbf{y}_i - \boldsymbol{\mu}_i); \boldsymbol{\nu}_i, \boldsymbol{\Gamma}_i)}{\prod_{i=1}^n \Phi_{q_i}(\mathbf{0}; \boldsymbol{\nu}_i, \boldsymbol{\Gamma}_i + \mathbf{D}_i \boldsymbol{\Sigma}_i \mathbf{D}_i^T)} \\ &= \frac{\phi_{p^\dagger}(\mathbf{y}; \boldsymbol{\mu}^\dagger, \boldsymbol{\Sigma}^\dagger) \Phi_{q^\dagger}(\mathbf{D}^\dagger(\mathbf{y} - \boldsymbol{\mu}^\dagger); \boldsymbol{\nu}^\dagger, \boldsymbol{\Gamma}^\dagger)}{\prod_{i=1}^n \Phi_{q_i}(\mathbf{0}; \boldsymbol{\nu}_i, \boldsymbol{\Gamma}_i + \mathbf{D}_i \boldsymbol{\Sigma}_i \mathbf{D}_i^T)}, \end{aligned}$$

and by noting that:

$$\begin{aligned} \bigoplus_{i=1}^n (\boldsymbol{\Gamma}_i + \mathbf{D}_i \boldsymbol{\Sigma}_i \mathbf{D}_i^T) &= \bigoplus_{i=1}^n \boldsymbol{\Gamma}_i + \bigoplus_{i=1}^n \mathbf{D}_i \boldsymbol{\Sigma}_i \mathbf{D}_i^T \\ &= \bigoplus_{i=1}^n \boldsymbol{\Gamma}_i + \left(\bigoplus_{i=1}^n \mathbf{D}_i \right) \left(\bigoplus_{i=1}^n \boldsymbol{\Sigma}_i \right) \left(\bigoplus_{i=1}^n \mathbf{D}_i^T \right) \\ &= \boldsymbol{\Gamma}^\dagger + \mathbf{D}^\dagger \boldsymbol{\Sigma}^\dagger \mathbf{D}^{\dagger T} \end{aligned}$$

we get:

$$f_{p,q}(\mathbf{y}) = \frac{\phi_{p^\dagger}(\mathbf{y}; \boldsymbol{\mu}^\dagger, \boldsymbol{\Sigma}^\dagger) \Phi_{q^\dagger}(\mathbf{D}^\dagger(\mathbf{y} - \boldsymbol{\mu}^\dagger); \boldsymbol{\nu}^\dagger, \boldsymbol{\Gamma}^\dagger)}{\Phi_{q^\dagger}(\mathbf{0}; \boldsymbol{\nu}^\dagger, \boldsymbol{\Gamma}^\dagger + \mathbf{D}^\dagger \boldsymbol{\Sigma}^\dagger \mathbf{D}^{\dagger T})}.$$

□

Corollary 2.2 (special case of iid random vectors). *Let $\mathbf{Y}_1, \dots, \mathbf{Y}_n$ be independent and identically distributed (iid) random vectors from $CSN_{p,q}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{D}, \boldsymbol{\nu}, \boldsymbol{\Gamma})$, then $\mathbf{Y}_1, \dots, \mathbf{Y}_n$ has a CSN distribution with the parameters:*

$$\begin{aligned} p^\dagger &= np, & q^\dagger &= nq, & \boldsymbol{\mu}^\dagger &= \mathbf{1}_n \otimes \boldsymbol{\mu}, & \boldsymbol{\Sigma}^\dagger &= \mathbf{I}_n \otimes \boldsymbol{\Sigma}_i, \\ \mathbf{D}^\dagger &= \mathbf{I}_n \otimes \mathbf{D}, & \boldsymbol{\nu}^\dagger &= \mathbf{1}_n \otimes \boldsymbol{\nu}, & \boldsymbol{\Gamma}^\dagger &= \mathbf{I}_n \otimes \boldsymbol{\Gamma}. \end{aligned}$$

Proof. This is a direct consequence of Proposition 2.9 □

2.3.6 Sums of Independent CSN Random Vectors

The results in this section show that the sum of independent CSN belongs to the CSN family. These additive properties are also found in the normal family distribution.

Proposition 2.10 (Sum of independent CSNs). *Let $\mathbf{Y}_1, \dots, \mathbf{Y}_n$ be independent random vectors with $\mathbf{Y}_i \sim CSN_{p,q_i}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \mathbf{D}_i, \boldsymbol{\nu}_i, \boldsymbol{\Gamma}_i)$, $i = 1, \dots, n$, then*

$$\mathbf{Y} = \sum_{i=1}^n \mathbf{Y}_i \sim CSN_{p,q^\dagger}(\boldsymbol{\mu}^\dagger, \boldsymbol{\Sigma}^\dagger, \mathbf{D}^\dagger, \boldsymbol{\nu}^\dagger, \boldsymbol{\Gamma}^\dagger),$$

where:

$$\begin{aligned} q^\dagger &= \sum_{i=1}^n q_i, & \boldsymbol{\mu}^\dagger &= \sum_{i=1}^n \boldsymbol{\mu}_i, & \boldsymbol{\Sigma}^\dagger &= \sum_{i=1}^n \boldsymbol{\Sigma}_i, \\ \mathbf{D}^\dagger &= (\boldsymbol{\Sigma}_1 \mathbf{D}_1^T, \dots, \boldsymbol{\Sigma}_n \mathbf{D}_n^T)^T \left(\sum_{i=1}^n \boldsymbol{\Sigma}_i \right)^{-1}, & \boldsymbol{\nu}^\dagger &= (\boldsymbol{\nu}_1^T, \dots, \boldsymbol{\nu}_n^T)^T, \\ \boldsymbol{\Gamma}^\dagger &= \boldsymbol{\Gamma}^\dagger + \mathbf{D}^\dagger \boldsymbol{\Sigma}^\dagger \mathbf{D}^{\dagger T} - (\boldsymbol{\Sigma}_1 \mathbf{D}_1^T, \dots, \boldsymbol{\Sigma}_n \mathbf{D}_n^T)^T \left(\sum_{i=1}^n \boldsymbol{\Sigma}_i \right)^{-1} (\boldsymbol{\Sigma}_1 \mathbf{D}_1^T, \dots, \boldsymbol{\Sigma}_n \mathbf{D}_n^T). \end{aligned}$$

Proof. Let $\mathbf{Y} = (\mathbf{Y}_1^T, \dots, \mathbf{Y}_n^T)^T$ and $\mathbf{A} = \mathbf{1}_n^T \otimes \mathbf{I}_p$. Note that $\sum_{i=1}^n \mathbf{y}_i = \mathbf{A}\mathbf{y}$, where $\mathbf{y}$ is an np -vector and $\mathbf{A}$ is a $p \times np$ matrix of rank p . Then using propositions 2.5 and

2.9, it follows that:

$$\mathbf{A}\mathbf{Y} \sim CSN_{p, \sum_{i=1}^n q_i} \left(\boldsymbol{\mu}_A^\dagger, \boldsymbol{\Sigma}_A^\dagger, \mathbf{D}_A^\dagger, \boldsymbol{\nu}_A^\dagger, \boldsymbol{\Gamma}_A^\dagger \right),$$

where:

$$\begin{aligned} \boldsymbol{\mu}_A^\dagger &= \mathbf{A}\boldsymbol{\mu}^\dagger, \quad \boldsymbol{\Sigma}_A^\dagger = \mathbf{A}\boldsymbol{\Sigma}^\dagger\mathbf{A}^T, \quad \mathbf{D}_A^\dagger = \mathbf{D}^\dagger\boldsymbol{\Sigma}^\dagger\mathbf{A}^T (\mathbf{A}\boldsymbol{\Sigma}^\dagger\mathbf{A}^T)^{-1}, \\ \boldsymbol{\nu}_A^\dagger &= \boldsymbol{\nu}^\dagger, \quad \boldsymbol{\Gamma}_A^\dagger = \boldsymbol{\Gamma}^\dagger + \mathbf{D}^\dagger\boldsymbol{\Sigma}^\dagger\mathbf{D}^{\dagger T} - \mathbf{D}^\dagger\boldsymbol{\Sigma}^\dagger\mathbf{A}^T (\mathbf{A}\boldsymbol{\Sigma}^\dagger\mathbf{A}^T)^{-1} \mathbf{A}\boldsymbol{\Sigma}^\dagger\mathbf{D}^{\dagger T}, \end{aligned}$$

where $\boldsymbol{\mu}^\dagger$, $\boldsymbol{\Sigma}^\dagger$, $\mathbf{D}^\dagger$, $\boldsymbol{\nu}^\dagger$, and $\boldsymbol{\Gamma}^\dagger$ are defined in Proposition 2.9.

Therefore, it is not difficult to see that:

$$\begin{aligned} \mathbf{A}\boldsymbol{\mu}^\dagger &= \sum_{i=1}^n \boldsymbol{\mu}_i, \quad \mathbf{A}\boldsymbol{\Sigma}^\dagger\mathbf{A}^T = \sum_{i=1}^n \boldsymbol{\Sigma}_i, \\ \mathbf{D}^\dagger\boldsymbol{\Sigma}^\dagger\mathbf{A}^T &= \bigoplus_{i=1}^n (\mathbf{D}_i\boldsymbol{\Sigma}_i), \quad \mathbf{A}\boldsymbol{\Sigma}^\dagger\mathbf{D}^{\dagger T} = (\mathbf{D}^\dagger\boldsymbol{\Sigma}^\dagger\mathbf{A}^T)^T = \bigoplus_{i=1}^n (\boldsymbol{\Sigma}_i\mathbf{D}_i^T) \end{aligned}$$

□

Proposition 2.10 can be generalized to the singular skew-normal defined in Proposition 2.6. In which case the pdf is not defined, the mgfs is then used to prove the more general Proposition 2.10.

Corollary 2.3 (Special case of iid random vectors). *Let $\mathbf{Y}_1, \dots, \mathbf{Y}_n$ be independent and identically distributed (iid) random vectors with $\mathbf{Y}_i \sim CSN_{p,q}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{D}, \boldsymbol{\nu}, \boldsymbol{\Gamma})$, $i = 1, \dots, n$, then*

$$\mathbf{Y} = \sum_{i=1}^n \mathbf{Y}_i \sim CSN_{p,nq}(n\boldsymbol{\mu}, n\boldsymbol{\Sigma}, \mathbf{D}^\dagger, \boldsymbol{\nu}^\dagger, \boldsymbol{\Gamma}^\dagger),$$

where:

$$\mathbf{D}^\dagger = \frac{1}{n} \mathbf{1}_n \otimes \mathbf{D}, \quad \boldsymbol{\nu}^\dagger = \mathbf{1}_n \otimes \boldsymbol{\nu}, \quad \boldsymbol{\Gamma}^\dagger = \mathbf{I}_n \otimes (\boldsymbol{\Gamma} + \mathbf{D}\boldsymbol{\Sigma}\mathbf{D}^T) - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T \otimes (\mathbf{D}\boldsymbol{\Sigma}\mathbf{D}^T).$$

Proof. This is a direct consequence of Proposition 2.10 □

Note that the summation does not modify the dimension p but it does affect the dimension q .

2.4 Further Generalization

For completeness, we mention that many generalizations and other forms of skew distributions have been developed even though their literature is limited. Two distributions are of particular interest, the skew-t distribution and the skew-elliptical distributions (see Genton (2004) for more details).

2.4.1 Skew-t Distribution

Let $\mathbf{x}$ denote a p -dimensional random vector. It is taken to have a skew-t distribution if its density is of the form

$$g(\mathbf{x}) = 2t_p(\mathbf{x}; \nu) T_1 \left(\boldsymbol{\alpha}^T W^{-1}(\mathbf{x} - \boldsymbol{\mu}) \left[\frac{\nu + p}{\mathbf{Q} + \nu} \right]^{1/2}; \nu + p \right), \quad (2.18)$$

where t_p is the density of a p -dimensional t variate with ν degrees of freedom:

$$t_p(\mathbf{x}; \nu) = \frac{\Gamma((\nu + p)/2)}{|\boldsymbol{\Omega}|^{1/2} (\pi\nu)^{p/2} \Gamma(\nu/2)} (1 + \mathbf{Q}/\nu)^{-(\nu+p)/2}, \quad (2.19)$$

where

$$\mathbf{Q} = (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Omega}^{-1} (\mathbf{x} - \boldsymbol{\mu}). \quad (2.20)$$

The scalar parameter ν denotes the degree of freedom of the multivariate t distribution. The p -dimensional vector $\boldsymbol{\mu}$ is a locator parameter and $\boldsymbol{\Omega}$ is a $p \times p$ covariance matrix. The diagonal part of $\boldsymbol{\Omega}$ is denoted by W^2 and the correlation matrix is $\bar{\boldsymbol{\Omega}} = W^{-1}\boldsymbol{\Omega}W$. It can be shown that, the expected value of $\mathbf{x}$ is given by:

$$\mathbb{E}(\mathbf{x}) = \boldsymbol{\mu} + W\boldsymbol{\eta}, \quad (2.21)$$

and its variance by:

$$\mathbb{V}ar(\mathbf{x}) = \frac{\nu}{\nu - 2}\boldsymbol{\Omega} - W\boldsymbol{\eta}\boldsymbol{\eta}^T W, \quad (2.22)$$

where

$$\boldsymbol{\eta} = \left[\frac{\nu/\pi}{1 + \boldsymbol{\alpha}^T \bar{\boldsymbol{\Omega}} \boldsymbol{\alpha}} \right]^{1/2} \frac{\Gamma((\nu - 1)/2)}{\Gamma(\nu/2)} \bar{\boldsymbol{\Omega}} \boldsymbol{\alpha}. \quad (2.23)$$

2.4.2 Skew-Elliptical Distribution

A p -dimensional random vector $\mathbf{x}$ is said to follow the skew-elliptical distribution if its probability density function is equal to:

$$f(\mathbf{x}) = |\boldsymbol{\Omega}|^{-1/2} g^{(p)} \left((\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Omega}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right), \quad (2.24)$$

where $g^{(p)}(u)$ is a non-increasing function from $\mathbb{R}_+$ to $\mathbb{R}_+$ defined by

$$g^{(p)}(u) = \frac{\Gamma(p/2)}{\pi^{p/2}} \frac{g(u; p)}{\int_0^\infty r^{p/2-1} g(r; p) dr}, \quad (2.25)$$

and $g(u; p)$ is non-increasing function from $\mathbb{R}_+$ to $\mathbb{R}_+$ such that the integral $\int_0^\infty r^{p/2-1} g(r; p) dr$ exists. The function $g^{(p)}$ is often called the density generator of the random vector $\mathbf{x}$. Note that the function $g(u; p)$ provides the kernel of $\mathbf{x}$ and the other terms in $g^{(p)}$ constitute the normalizing constant of the density probability function f .

Chapter 3

Parameter Estimation under the One-Fold SN Model

In this chapter, we first derive the distribution of $\mathbf{Y}_d$ under the linear mixed model (LMM) with random components following CSN distributions. The one-fold nested error model with SN errors follows as a particular case of the LMM. We also provide the distribution of the sample vector $\mathbf{Y}_{ds}$ under the assumption of non-informative sampling. In the second part of this chapter, we present a maximum likelihood (ML) technique to estimate the parameters of the one-fold SN model (as a remainder the SN model refers to the nested error linear with the errors following SN distributions). We also present a data augmentation method called data cloning (DC) for estimating the parameters of the SN model.

3.1 Small Area Estimation Model

LMMs are widely used in biological and social sciences analyses where between-subjects variability is interest. The LMM has also become a reference approach for SAE (see Rao (2003) for an extensive review). The LMM with closed skew-normal (CSN)

random components is defined as follows:

$$\mathbf{Y}_d = \mathbf{X}_d \boldsymbol{\beta} + \mathbf{Z}_d \mathbf{u}_d + \mathbf{e}_d, \quad d = 1, \dots, m. \quad (3.1)$$

where $\mathbf{u}_d \sim CSN_q(\boldsymbol{\mu}_{u_d}, \boldsymbol{\Sigma}_{u_d}, \mathbf{D}_{u_d}, \boldsymbol{\nu}_{u_d}, \boldsymbol{\Gamma}_{u_d})$, $\mathbf{e}_d \sim CSN_{N_d}(\boldsymbol{\mu}_{e_d}, \boldsymbol{\Sigma}_{e_d}, \mathbf{D}_{e_d}, \boldsymbol{\nu}_{e_d}, \boldsymbol{\Gamma}_{e_d})$, and $\mathbf{u}_d$ and $\mathbf{e}_d$ are independent. $\mathbf{X}_d$ is a $N_d \times p$ matrix, $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects, $\mathbf{Z}_d$ is a $N_d \times q$ matrix, $\mathbf{u}_d$ is a $q \times 1$ vector of area-random effects, and $\mathbf{e}_d = (e_{d1}, \dots, e_{dN_d})^T$ is a $N_d \times 1$ vector of random errors.

The distribution of $\mathbf{Y}_d$ is derived using two approaches. The first approach is a direct derivation of the probability density function (pdf) using normal distribution properties (Theorem 3.1). The second approach uses the closure properties of the CSN to obtain the distribution of $\mathbf{Y}_d$ (Proposition 3.1).

Theorem 3.1. *The marginal distribution of $\mathbf{Y}_d$ as defined by (3.1) is:*

$$\begin{aligned} f_{\mathbf{Y}_d}(\mathbf{y}_d) &= C \phi_{N_d}(\mathbf{y}_d | \mathbf{X}_d \boldsymbol{\beta} + \mathbf{Z}_d \boldsymbol{\mu}_{u_d} + \boldsymbol{\mu}_{e_d}; \boldsymbol{\Sigma}_d) \\ &\quad \times \Phi_{N_d+q}(\boldsymbol{\mu}_{2d} - \mathbf{H}_d \boldsymbol{\mu}_{1d} | \boldsymbol{\mu}_{0d}; \mathbf{R}_d + \mathbf{H}_d \mathbf{G}_d \mathbf{H}_d^T), \end{aligned} \quad (3.2)$$

where:

$$\begin{aligned} C &= \Phi_{N_d+q}^{-1} \left(\mathbf{0} \middle| \begin{pmatrix} \boldsymbol{\nu}_{e_d} \\ \boldsymbol{\nu}_{u_d} \end{pmatrix}; \begin{pmatrix} \boldsymbol{\Gamma}_{e_d} + \mathbf{D}_{e_d} \boldsymbol{\Sigma}_{e_d} \mathbf{D}_{e_d}^T & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Gamma}_{u_d} + \mathbf{D}_{u_d} \boldsymbol{\Sigma}_{u_d} \mathbf{D}_{u_d}^T \end{pmatrix} \right) \\ \boldsymbol{\Sigma}_d &= \boldsymbol{\Sigma}_{e_d} + \mathbf{Z}_d \boldsymbol{\Sigma}_{u_d} \mathbf{Z}_d^T, \quad \boldsymbol{\mu}_{0d} = (\boldsymbol{\nu}_{e_d}^T, \boldsymbol{\nu}_{u_d}^T)^T \\ \boldsymbol{\mu}_{1d} &= \boldsymbol{\mu}_{u_d} + \mathbf{G}_d^T \mathbf{Z}_d^T \boldsymbol{\Sigma}_{e_d}^{-1} (\mathbf{y}_d - \mathbf{X}_d \boldsymbol{\beta} - \mathbf{Z}_d \boldsymbol{\mu}_{u_d} - \boldsymbol{\mu}_{e_d}), \\ \boldsymbol{\mu}_{2d} &= \begin{pmatrix} \mathbf{D}_{e_d} (\mathbf{y}_d - \mathbf{X}_d \boldsymbol{\beta} - \boldsymbol{\mu}_{e_d}) \\ -\mathbf{D}_{u_d} \boldsymbol{\mu}_{u_d} \end{pmatrix}, \\ \mathbf{R}_d &= \begin{bmatrix} \boldsymbol{\Gamma}_{e_d} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Gamma}_{u_d} \end{bmatrix}, \quad \mathbf{H}_d = \begin{bmatrix} \mathbf{D}_{e_d} \mathbf{Z}_d \\ -\mathbf{D}_{u_d} \end{bmatrix}, \quad \mathbf{G}_d = [\boldsymbol{\Sigma}_{u_d}^{-1} + \mathbf{Z}_d^T \boldsymbol{\Sigma}_{e_d}^{-1} \mathbf{Z}_d]^{-1}. \end{aligned}$$

Proof. To simplify the notations, we drop the subscript d in this proof.

$$\begin{aligned}
f_{\mathbf{y}}(\mathbf{y}) &= \int_{\mathbb{R}^q} f(\mathbf{y}|\mathbf{u})f(\mathbf{u})d\mathbf{u} \\
&= \int_{\mathbb{R}^q} C_u C_e \phi_n(\mathbf{y}|\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\mu}_u + \boldsymbol{\mu}_e; \boldsymbol{\Sigma}_e) \phi_q(\mathbf{u}|\boldsymbol{\mu}_u; \boldsymbol{\Sigma}_u) \\
&\quad \times \Phi_n(\mathbf{D}_e(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u} - \boldsymbol{\mu}_e)|\boldsymbol{\nu}_e, \boldsymbol{\Gamma}_e) \\
&\quad \times \Phi_q(\mathbf{D}_u(\mathbf{u} - \boldsymbol{\mu}_u)|\boldsymbol{\nu}_u; \boldsymbol{\Gamma}_u) d\mathbf{u}
\end{aligned} \tag{3.3}$$

where

$$C_u = \Phi_q^{-1}(\mathbf{0}|\boldsymbol{\nu}_u; \boldsymbol{\Gamma}_u + \mathbf{D}_u \boldsymbol{\Sigma}_u \mathbf{D}_u^T) \tag{3.4}$$

$$C_e = \Phi_N^{-1}(\mathbf{0}|\boldsymbol{\nu}_e; \boldsymbol{\Gamma}_e + \mathbf{D}_e \boldsymbol{\Sigma}_e \mathbf{D}_e^T) \tag{3.5}$$

We have the following lemma,

Lemma 3.1. *Let $\mathbf{Y}|\mathbf{X} = \mathbf{x} \sim N_p(\boldsymbol{\mu} + \mathbf{A}\mathbf{x}, \boldsymbol{\Sigma})$ and $X \sim N_q(\boldsymbol{\eta}, \boldsymbol{\Omega})$, then*

$$\begin{aligned}
\phi_p(\mathbf{y}|\boldsymbol{\mu} + \mathbf{A}\mathbf{x}; \boldsymbol{\Sigma}) \phi_q(\mathbf{x}|\boldsymbol{\eta}; \boldsymbol{\Omega}) &= \phi_p(\mathbf{y}|\boldsymbol{\mu} + \mathbf{A}\boldsymbol{\eta}; \boldsymbol{\Sigma} + \mathbf{A}\boldsymbol{\Omega}\mathbf{A}^T) \\
&\quad \times \phi_q(\mathbf{x}|\boldsymbol{\eta} + \mathbf{V}\mathbf{A}^T \boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu} - \mathbf{A}\boldsymbol{\eta}); \mathbf{V})
\end{aligned} \tag{3.6}$$

where $\mathbf{V} = (\boldsymbol{\Omega}^{-1} + \mathbf{A}^T \boldsymbol{\Sigma}^{-1} \mathbf{A})^{-1}$.

Using lemma 3.1 yields:

$$\begin{aligned}
\phi_n(\mathbf{y}|\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\mu}_e; \boldsymbol{\Sigma}_e) \phi_q(\mathbf{u}|\boldsymbol{\mu}_u; \boldsymbol{\Sigma}_u) &= \\
\phi_n(\mathbf{y}|\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\mu}_u + \boldsymbol{\mu}_e; \boldsymbol{\Sigma}_e + \mathbf{Z}\boldsymbol{\Sigma}_u \mathbf{Z}^T) &\times \phi_q(\mathbf{u}|\boldsymbol{\mu}_1; \mathbf{G})
\end{aligned} \tag{3.7}$$

Also we have,

Lemma 3.2. *If $\mathbf{W}_1 \sim N_p(\boldsymbol{\mu}_e, \boldsymbol{\Gamma}_e)$, $\mathbf{W}_2 \sim N_q(\boldsymbol{\mu}_u, \boldsymbol{\Gamma}_u)$, and are independent, then*

$$P(\mathbf{W}_1 \leq \mathbf{D}_e(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u} - \boldsymbol{\mu}_e)) P(\mathbf{W}_2 \leq \mathbf{D}_u(\mathbf{u} - \boldsymbol{\mu}_u)) = P(\mathbf{T} \leq \boldsymbol{\mu}_2) \quad (3.8)$$

where $\mathbf{T} = \mathbf{W} + \mathbf{H}\mathbf{u}$ with $\mathbf{w} = (\mathbf{W}_1, \mathbf{W}_2) \sim N_{n+q}(\boldsymbol{\mu}_0, \mathbf{R})$ and $\boldsymbol{\mu}_0, \boldsymbol{\mu}_2, \mathbf{H}, \mathbf{R}$ as in Theorem 3.1.

Since $\mathbf{T} \sim N_{n+q}(\boldsymbol{\mu}_0 + \mathbf{H}\mathbf{u}, \mathbf{R})$, it follows that:

$$P(\mathbf{T} \leq \boldsymbol{\mu}_2) = \Phi_{n+q}(\boldsymbol{\mu}_2 | \boldsymbol{\mu}_0 + \mathbf{H}\mathbf{u}; \mathbf{R}) = \Phi_{n+q}(-\boldsymbol{\mu}_0 - \mathbf{H}\mathbf{u} | -\boldsymbol{\mu}_2; \mathbf{R}) \quad (3.9)$$

Using lemma 3.2, we get the following result:

$$\begin{aligned} & \Phi_n(\mathbf{D}_e(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \boldsymbol{\mu}_e - \mathbf{Z}\boldsymbol{\mu}_u) | \boldsymbol{\nu}_e; \boldsymbol{\Gamma}_e) \Phi_q(\mathbf{D}_u(\mathbf{u} - \boldsymbol{\mu}_u) | \boldsymbol{\nu}_u; \boldsymbol{\Gamma}_u) \\ &= \Phi_{n+q}(\boldsymbol{\mu}_0 - \mathbf{H}\mathbf{u} | -\boldsymbol{\mu}_2; \mathbf{R}) \end{aligned} \quad (3.10)$$

where $\boldsymbol{\mu}_2, \mathbf{H}, \mathbf{R}$ as in Theorem 3.1. Results (3.7) and (3.10) lead to:

$$\begin{aligned} f_{\mathbf{y}}(\mathbf{y} | \boldsymbol{\theta}) &= \int_{\mathbb{R}^q} C_u C_e \phi_n(\mathbf{y} | \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\mu}_u + \boldsymbol{\mu}_e; \boldsymbol{\Sigma}) \times \phi_q(\mathbf{u} | \boldsymbol{\mu}_1; \mathbf{G}) \\ &\quad \times \Phi_{n+q}(-\boldsymbol{\mu}_0 - \mathbf{H}\mathbf{u} | -\boldsymbol{\mu}_2; \mathbf{R}) d\mathbf{u} \end{aligned} \quad (3.11)$$

$$\begin{aligned} &= C_u C_e \phi_n(\mathbf{y} | \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\mu}_u + \boldsymbol{\mu}_e; \boldsymbol{\Sigma}) \\ &\quad \times \int_{\mathbb{R}^q} \phi_q(\mathbf{u} | \boldsymbol{\mu}_1; \mathbf{G}) \Phi_{n+q}(-\boldsymbol{\mu}_0 - \mathbf{H}\mathbf{u} | -\boldsymbol{\mu}_2; \mathbf{R}) d\mathbf{u} \end{aligned} \quad (3.12)$$

Lemma 3.3. *Let $\mathbf{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then for any fixed k -dimensional vector $\mathbf{a}$, $k \times n$, matrix $\mathbf{B}$, k -dimensional vector $\boldsymbol{\eta}$, and $k \times k$ matrix $\boldsymbol{\Omega}$,*

$$\mathbb{E}[\Phi_k(\mathbf{a} + \mathbf{B}\mathbf{Y} | \boldsymbol{\eta}; \boldsymbol{\Omega})] = \Phi_k(\mathbf{a} | \boldsymbol{\eta} - \mathbf{B}\boldsymbol{\mu}; \boldsymbol{\Omega} + \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^T) \quad (3.13)$$

Lemma 3.4. Consider two independent multivariate normal vectors $\mathbf{W}_1 \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathbf{W}_2 \sim N_q(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, and matrices $\mathbf{A}$ and $\mathbf{B}$ with the appropriate dimensions, then

$$\Phi(\mathbf{A}\mathbf{w}_1|\boldsymbol{\mu}_1; \boldsymbol{\Sigma}_1) \Phi(\mathbf{B}\mathbf{w}_2|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) = \Phi(\mathbf{C}_w|\boldsymbol{\mu}; \boldsymbol{\Sigma}) \quad (3.14)$$

$$\text{where } \mathbf{C}_w = \begin{pmatrix} \mathbf{A}\mathbf{w}_1 \\ \mathbf{B}\mathbf{w}_2 \end{pmatrix}, \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_2 \end{pmatrix}.$$

Using lemmas 3.3 and 3.4, we get:

$$f_{\mathbf{y}}(\mathbf{y}|\boldsymbol{\theta}) = C\phi_n(\mathbf{y}|\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\mu}_u + \boldsymbol{\mu}_e; \boldsymbol{\Sigma}) \Phi_{n+q}(\boldsymbol{\mu}_2 - \mathbf{H}\boldsymbol{\mu}_1|\boldsymbol{\mu}_0; \mathbf{R} + \mathbf{H}\mathbf{G}\mathbf{H}^T)$$

□

We also can take advantage of the closure properties shown in Chapter 2 to derive the following result.

Proposition 3.1. Let $\mathbf{Y}_d$ be defined as in (3.1), then

$$\mathbf{Y}_d \sim CSN_{N_d, N_d+q}(\boldsymbol{\mu}_{Y_d}, \boldsymbol{\Sigma}_{Y_d}, \mathbf{D}_{Y_d}, \boldsymbol{\nu}_{Y_d}, \boldsymbol{\Gamma}_{Y_d}) \quad (3.15)$$

where:

$$\begin{aligned} \boldsymbol{\mu}_{Y_d} &= \mathbf{X}_d\boldsymbol{\beta} + \mathbf{Z}_d\boldsymbol{\mu}_{u_d} + \boldsymbol{\mu}_{e_d}, \quad \boldsymbol{\Sigma}_{Y_d} = \boldsymbol{\Sigma}_{e_d} + \mathbf{Z}_d\boldsymbol{\Sigma}_{u_d}\mathbf{Z}_d^T, \\ \mathbf{D}_{Y_d} &= \begin{bmatrix} \mathbf{D}_{e_d}\boldsymbol{\Sigma}_{e_d}(\boldsymbol{\Sigma}_{e_d} + \mathbf{Z}_d\boldsymbol{\Sigma}_{u_d}\mathbf{Z}_d^T)^{-1} \\ \mathbf{D}_{u_d}\boldsymbol{\Sigma}_{u_d}\mathbf{Z}_d^T(\boldsymbol{\Sigma}_{e_d} + \mathbf{Z}_d\boldsymbol{\Sigma}_{u_d}\mathbf{Z}_d^T)^{-1} \end{bmatrix}, \quad \boldsymbol{\nu}_{Y_d} = (\boldsymbol{\nu}_{e_d}, \boldsymbol{\nu}_{u_d}), \\ \boldsymbol{\Gamma}_{Y_d} &= \begin{bmatrix} \boldsymbol{\Gamma}_{11} & \boldsymbol{\Gamma}_{12} \\ \boldsymbol{\Gamma}_{21} & \boldsymbol{\Gamma}_{22} \end{bmatrix} \text{ with} \\ \boldsymbol{\Gamma}_{11} &= \boldsymbol{\Gamma}_{e_d} + \mathbf{D}_{e_d}\boldsymbol{\Sigma}_{e_d}\mathbf{D}_{e_d}^T - \mathbf{D}_{e_d}\boldsymbol{\Sigma}_{e_d}[\boldsymbol{\Sigma}_{e_d} + \mathbf{Z}_d\boldsymbol{\Sigma}_{u_d}\mathbf{Z}_d^T]^{-1}\boldsymbol{\Sigma}_{e_d}\mathbf{D}_{e_d}^T \\ \boldsymbol{\Gamma}_{22} &= \boldsymbol{\Gamma}_{u_d} + \mathbf{D}_{u_d}\boldsymbol{\Sigma}_{u_d}\mathbf{D}_{u_d}^T - \mathbf{D}_{u_d}\boldsymbol{\Sigma}_{u_d}\mathbf{Z}_d^T[\boldsymbol{\Sigma}_{e_d} + \mathbf{Z}_d\boldsymbol{\Sigma}_{u_d}\mathbf{Z}_d^T]^{-1}\mathbf{Z}_d\boldsymbol{\Sigma}_{u_d}\mathbf{D}_{u_d}^T \\ \boldsymbol{\Gamma}_{12} &= -\mathbf{D}_{e_d}\boldsymbol{\Sigma}_{e_d}[\boldsymbol{\Sigma}_{e_d} + \mathbf{Z}_d\boldsymbol{\Sigma}_{u_d}\mathbf{Z}_d^T]^{-1}\mathbf{Z}_d\boldsymbol{\Sigma}_{u_d}\mathbf{D}_{u_d}^T = \boldsymbol{\Gamma}_{21}^T \end{aligned}$$

Proof. Using Proposition 2.6, we get:

$$\mathbf{Z}_d \mathbf{u}_d \sim SSN_{N_d, q} (\mathbf{Z}_d \boldsymbol{\mu}_{u_d}, \mathbf{Z}_d \boldsymbol{\Sigma}_{u_d} \mathbf{Z}_d^T, \mathbf{D}_{u_d} (\mathbf{Z}_{u_d}^T \mathbf{Z}_{u_d})^{-1} \mathbf{Z}_{u_d}^T, \boldsymbol{\nu}_{u_d}, \boldsymbol{\Gamma}_{u_d})$$

Since $\mathbf{Z}_d \mathbf{u}_d$ and $\mathbf{e}_d$ are independent, Proposition 2.10 leads to:

$$\mathbf{e}_d + \mathbf{Z}_d \mathbf{u}_d \sim CSN_{N_d, N_d+q} (\boldsymbol{\mu}_{e_d} + \mathbf{Z}_d \boldsymbol{\mu}_{u_d}, \boldsymbol{\Sigma}_{Y_d}, \mathbf{D}_{Y_d}, \boldsymbol{\nu}_{Y_d}, \boldsymbol{\Gamma}_{Y_d})$$

where: $\boldsymbol{\Sigma}_{Y_d} = \boldsymbol{\Sigma}_{e_d} + \mathbf{Z}_d \boldsymbol{\Sigma}_{u_d} \mathbf{Z}_d^T,$

$$\begin{aligned} \mathbf{D}_{Y_d} &= \begin{bmatrix} \mathbf{D}_{e_d} \boldsymbol{\Sigma}_{e_d} (\boldsymbol{\Sigma}_{e_d} + \mathbf{Z}_d \boldsymbol{\Sigma}_{u_d} \mathbf{Z}_d^T)^{-1} \\ \mathbf{D}_{Z_d u_d} \boldsymbol{\Sigma}_{Z_d u_d} (\boldsymbol{\Sigma}_{e_d} + \mathbf{Z}_d \boldsymbol{\Sigma}_{u_d} \mathbf{Z}_d^T)^{-1} \end{bmatrix}, \\ &= \begin{bmatrix} \mathbf{D}_{e_d} \boldsymbol{\Sigma}_{e_d} (\boldsymbol{\Sigma}_{e_d} + \mathbf{Z}_d \boldsymbol{\Sigma}_{u_d} \mathbf{Z}_d^T)^{-1} \\ \mathbf{D}_{u_d} (\mathbf{Z}_d^T \mathbf{Z}_d)^{-1} (\mathbf{Z}_d^T \mathbf{Z}_d) \boldsymbol{\Sigma}_{u_d} \mathbf{Z}_d^T (\boldsymbol{\Sigma}_{e_d} + \mathbf{Z}_d \boldsymbol{\Sigma}_{u_d} \mathbf{Z}_d^T)^{-1} \end{bmatrix}, \end{aligned}$$

$$\boldsymbol{\nu}_{Y_d} = (\boldsymbol{\nu}_{e_d}^T, \boldsymbol{\nu}_{u_d}^T)^T, \quad \text{and} \quad \boldsymbol{\Gamma}_{Y_d} = \begin{bmatrix} \boldsymbol{\Gamma}_{11} & \boldsymbol{\Gamma}_{12} \\ \boldsymbol{\Gamma}_{21} & \boldsymbol{\Gamma}_{22} \end{bmatrix} \text{ with}$$

$$\begin{aligned} \boldsymbol{\Gamma}_{11} &= \boldsymbol{\Gamma}_{e_d} + \mathbf{D}_{e_d} \boldsymbol{\Sigma}_{e_d} \mathbf{D}_{e_d}^T - \mathbf{D}_{e_d} \boldsymbol{\Sigma}_{e_d} (\boldsymbol{\Sigma}_{e_d} + \mathbf{Z}_d \boldsymbol{\Sigma}_{u_d} \mathbf{Z}_d^T)^{-1} \boldsymbol{\Sigma}_{e_d} \mathbf{D}_{e_d}^T \\ \boldsymbol{\Gamma}_{22} &= \boldsymbol{\Gamma}_{u_d} + \mathbf{D}_{u_d} (\mathbf{Z}_d^T \mathbf{Z}_d)^{-1} (\mathbf{Z}_d^T \mathbf{Z}_d) \boldsymbol{\Sigma}_{u_d} (\mathbf{Z}_d^T \mathbf{Z}_d)^{-1} (\mathbf{Z}_d^T \mathbf{Z}_d) \mathbf{D}_{u_d}^T \\ &\quad - \mathbf{D}_{u_d} (\mathbf{Z}_d^T \mathbf{Z}_d)^{-1} (\mathbf{Z}_d^T \mathbf{Z}_d) \boldsymbol{\Sigma}_{u_d} \mathbf{Z}_d^T (\boldsymbol{\Sigma}_{e_d} + \mathbf{Z}_d \boldsymbol{\Sigma}_{u_d} \mathbf{Z}_d^T)^{-1} \\ &\quad \times \mathbf{Z}_d \boldsymbol{\Sigma}_{u_d} (\mathbf{Z}_d^T \mathbf{Z}_d) (\mathbf{Z}_d^T \mathbf{Z}_d)^{-1} \mathbf{D}_{u_d}^T \\ &= \boldsymbol{\Gamma}_{u_d} + \mathbf{D}_{u_d} \boldsymbol{\Sigma}_{u_d} \mathbf{D}_{u_d}^T - \mathbf{D}_{u_d} \boldsymbol{\Sigma}_{u_d} \mathbf{Z}_d^T (\boldsymbol{\Sigma}_{e_d} + \mathbf{Z}_d \boldsymbol{\Sigma}_{u_d} \mathbf{Z}_d^T)^{-1} \mathbf{Z}_d \boldsymbol{\Sigma}_{u_d} \mathbf{D}_{u_d}^T \\ \boldsymbol{\Gamma}_{12} &= -\mathbf{D}_{e_d} \boldsymbol{\Sigma}_{e_d} (\boldsymbol{\Sigma}_{e_d} + \mathbf{Z}_d \boldsymbol{\Sigma}_{u_d} \mathbf{Z}_d^T)^{-1} \mathbf{Z}_d \boldsymbol{\Sigma}_{u_d} (\mathbf{Z}_d^T \mathbf{Z}_d) (\mathbf{Z}_d^T \mathbf{Z}_d)^{-1} \mathbf{D}_{u_d}^T \\ &= -\mathbf{D}_{e_d} \boldsymbol{\Sigma}_{e_d} (\boldsymbol{\Sigma}_{e_d} + \mathbf{Z}_d \boldsymbol{\Sigma}_{u_d} \mathbf{Z}_d^T)^{-1} \mathbf{Z}_d \boldsymbol{\Sigma}_{u_d} \mathbf{D}_{u_d}^T \\ &= \boldsymbol{\Gamma}_{21}^T \end{aligned}$$

□

Proposition 3.2 (Reconciliation). *The function defined in Theorem 3.1 is the probability density function of the distribution $CSN_{N_d, N_d+q}(\boldsymbol{\mu}_{Y_d}, \boldsymbol{\Sigma}_{Y_d}, \mathbf{D}_{Y_d}, \boldsymbol{\nu}_{Y_d}, \boldsymbol{\Gamma}_{Y_d})$ with $\boldsymbol{\mu}_{Y_d}$, $\boldsymbol{\Sigma}_{Y_d}$, $\mathbf{D}_{Y_d}$, $\boldsymbol{\nu}_{Y_d}$, and $\boldsymbol{\Gamma}_{Y_d}$ as in Proposition 3.1.*

Proof. From Theorem 3.1, we have:

$$\begin{aligned}\boldsymbol{\mu}_{Y_d} &= \mathbf{X}_d \boldsymbol{\beta} + \mathbf{Z}_d \boldsymbol{\mu}_{u_d} + \boldsymbol{\mu}_{e_d}, \quad \boldsymbol{\Sigma}_{Y_d} = \boldsymbol{\Sigma}_{e_d} + \mathbf{Z}_d \boldsymbol{\Sigma}_{u_d} \mathbf{Z}_d^T, \\ \boldsymbol{\mu}_{2_d} - \mathbf{H}_d \boldsymbol{\mu}_{1_d} &= \begin{bmatrix} \mathbf{D}_{e_d} (\mathbf{y} - \mathbf{X}_d \boldsymbol{\beta} - \boldsymbol{\mu}_{e_d}) \\ -\mathbf{D}_{u_d} \boldsymbol{\mu}_{u_d} \end{bmatrix} - \begin{bmatrix} \mathbf{D}_{e_d} \mathbf{Z}_d \\ -\mathbf{D}_{u_d} \end{bmatrix} \boldsymbol{\mu}_{1_d} \\ &= \begin{bmatrix} \mathbf{D}_{e_d} (\mathbf{I} - \mathbf{Z}_d \mathbf{G}_d^T \mathbf{Z}_d^T \boldsymbol{\Sigma}_{e_d}^{-1}) (\mathbf{y}_d - \mathbf{X}_d \boldsymbol{\beta} - \mathbf{D}_{e_d} \mathbf{Z}_d - \boldsymbol{\mu}_{e_d}) \\ \mathbf{D}_{u_d} \mathbf{G}_d^T \mathbf{Z}_d^T \boldsymbol{\Sigma}_{e_d}^{-1} (\mathbf{y}_d - \mathbf{X}_d \boldsymbol{\beta} - \mathbf{Z}_d \boldsymbol{\mu}_{u_d} - \boldsymbol{\mu}_{e_d}) \end{bmatrix}\end{aligned}$$

From the result: $(\mathbf{P}^{-1} + \mathbf{B}^T \mathbf{R}^{-1} \mathbf{B})^{-1} \mathbf{B}^T \mathbf{R}^{-1} = \mathbf{P} \mathbf{B}^T (\mathbf{B} \mathbf{P} \mathbf{B}^T + \mathbf{R})^{-1}$, we get:

$$\begin{aligned}\mathbf{G}_d^T \mathbf{Z}_d^T \boldsymbol{\Sigma}_{e_d}^{-1} &= (\boldsymbol{\Sigma}_{u_d}^{-1} + \mathbf{Z}_d^T \boldsymbol{\Sigma}_{e_d}^{-1} \mathbf{Z}_d)^{-1} \mathbf{Z}_d^T \boldsymbol{\Sigma}_{e_d}^{-1} \\ &= \boldsymbol{\Sigma}_{u_d} \mathbf{Z}_d^T (\boldsymbol{\Sigma}_{e_d} + \mathbf{Z}_d \boldsymbol{\Sigma}_{u_d} \mathbf{Z}_d^T)^{-1}\end{aligned}$$

Thus,

$$\mathbf{D}_{u_d} \mathbf{G}_d^T \mathbf{Z}_d^T \boldsymbol{\Sigma}_{e_d}^{-1} = \mathbf{D}_{u_d} \boldsymbol{\Sigma}_{u_d} \mathbf{Z}_d^T (\boldsymbol{\Sigma}_{e_d} + \mathbf{Z}_d \boldsymbol{\Sigma}_{u_d} \mathbf{Z}_d^T)^{-1} \text{ and}$$

$$\begin{aligned}\mathbf{I} - \mathbf{Z}_d \mathbf{G}_d^T \mathbf{Z}_d^T \boldsymbol{\Sigma}_{e_d}^{-1} &= \mathbf{I} - \mathbf{Z}_d \boldsymbol{\Sigma}_{u_d}^{-1} \mathbf{Z}_d^T (\boldsymbol{\Sigma}_{e_d} + \mathbf{Z}_d \boldsymbol{\Sigma}_{u_d} \mathbf{Z}_d^T)^{-1} \\ &= (\boldsymbol{\Sigma}_{e_d} + \mathbf{Z}_d \boldsymbol{\Sigma}_{u_d} \mathbf{Z}_d^T) (\boldsymbol{\Sigma}_{e_d} + \mathbf{Z}_d \boldsymbol{\Sigma}_{u_d} \mathbf{Z}_d^T)^{-1} - \mathbf{Z}_d \boldsymbol{\Sigma}_{u_d}^{-1} \mathbf{Z}_d^T (\boldsymbol{\Sigma}_{e_d} + \mathbf{Z}_d \boldsymbol{\Sigma}_{u_d} \mathbf{Z}_d^T)^{-1} \\ &= [(\boldsymbol{\Sigma}_{e_d} + \mathbf{Z}_d \boldsymbol{\Sigma}_{u_d} \mathbf{Z}_d^T) - \mathbf{Z}_d \boldsymbol{\Sigma}_{u_d} \mathbf{Z}_d^T] (\boldsymbol{\Sigma}_{e_d} + \mathbf{Z}_d \boldsymbol{\Sigma}_{u_d} \mathbf{Z}_d^T)^{-1} \\ &= \boldsymbol{\Sigma}_{e_d} (\boldsymbol{\Sigma}_{e_d} + \mathbf{Z}_d \boldsymbol{\Sigma}_{u_d} \mathbf{Z}_d^T)^{-1}\end{aligned}$$

Also,

$$\mathbf{D}_{u_d} (\mathbf{I} - \mathbf{Z}_d \mathbf{G}_d^T \mathbf{Z}_d^T \boldsymbol{\Sigma}_{e_d}^{-1}) = \mathbf{D}_{u_d} \boldsymbol{\Sigma}_{e_d} (\boldsymbol{\Sigma}_{e_d} + \mathbf{Z}_d \boldsymbol{\Sigma}_{u_d} \mathbf{Z}_d^T)^{-1}$$

Therefore,

$$\boldsymbol{\mu}_{2_d} - \mathbf{H}_d \boldsymbol{\mu}_{1_d} = \begin{bmatrix} \mathbf{D}_{e_d} \boldsymbol{\Sigma}_{e_d} (\boldsymbol{\Sigma}_{e_d} + \mathbf{Z}_d \boldsymbol{\Sigma}_{u_d} \mathbf{Z}_d^T)^{-1} (\mathbf{y}_d - \mathbf{X}_d \boldsymbol{\beta} - \mathbf{Z}_d \boldsymbol{\mu}_{u_d} - \boldsymbol{\mu}_{e_d}) \\ \mathbf{D}_{u_d} \boldsymbol{\Sigma}_{u_d} \mathbf{Z}_d^T (\boldsymbol{\Sigma}_{e_d} + \mathbf{Z}_d \boldsymbol{\Sigma}_{u_d} \mathbf{Z}_d^T)^{-1} (\mathbf{y}_d - \mathbf{X}_d \boldsymbol{\beta} - \mathbf{Z}_d \boldsymbol{\mu}_{u_d} - \boldsymbol{\mu}_{e_d}) \end{bmatrix}$$

From the expression above, one can recognize,

$$\mathbf{D}_{Y_d} = \begin{bmatrix} \mathbf{D}_{e_d} \boldsymbol{\Sigma}_{e_d} (\boldsymbol{\Sigma}_{e_d} + \mathbf{Z}_d \boldsymbol{\Sigma}_{u_d} \mathbf{Z}_d^T)^{-1} \\ \mathbf{D}_{u_d} \boldsymbol{\Sigma}_{u_d} \mathbf{Z}_d^T (\boldsymbol{\Sigma}_{e_d} + \mathbf{Z}_d \boldsymbol{\Sigma}_{u_d} \mathbf{Z}_d^T)^{-1} \end{bmatrix}$$

The two last components are:

$$\boldsymbol{\nu}_{Y_d} = (\boldsymbol{\nu}_{e_d}^T, \boldsymbol{\nu}_{u_d}^T)^T, \quad \text{and}$$

$$\begin{aligned} \boldsymbol{\Gamma}_{Y_d} &= \mathbf{R}_d + \mathbf{H}_d \mathbf{G}_d \mathbf{H}_d^T \\ &= \begin{bmatrix} \boldsymbol{\Gamma}_{e_d} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Gamma}_{u_d} \end{bmatrix} + \mathbf{H}_d \mathbf{G}_d \mathbf{H}_d^T \text{ where } \mathbf{H}_d \mathbf{G}_d \mathbf{H}_d^T = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}, \text{ with} \end{aligned}$$

$$\begin{aligned} \mathbf{A}_{11} &= \mathbf{D}_{e_d} \mathbf{Z}_d \mathbf{G}_d \mathbf{Z}_d^T \mathbf{D}_{e_d}^T \\ &= \mathbf{D}_{e_d} \mathbf{Z}_d \boldsymbol{\Sigma}_{u_d} \mathbf{Z}_d^T (\boldsymbol{\Sigma}_{e_d} + \mathbf{Z}_d \boldsymbol{\Sigma}_{u_d} \mathbf{Z}_d^T)^{-1} \boldsymbol{\Sigma}_{e_d} \mathbf{D}_{e_d}^T \\ &= \mathbf{D}_{e_d} (\boldsymbol{\Sigma}_{Y_d} - \boldsymbol{\Sigma}_{e_d}) \boldsymbol{\Sigma}_{Y_d}^{-1} \boldsymbol{\Sigma}_{e_d} \mathbf{D}_{e_d}^T \\ &= \mathbf{D}_{e_d} \boldsymbol{\Sigma}_{e_d} \mathbf{D}_{e_d}^T - \mathbf{D}_{e_d} \boldsymbol{\Sigma}_{e_d} \boldsymbol{\Sigma}_{Y_d}^{-1} \boldsymbol{\Sigma}_{e_d} \mathbf{D}_{e_d}^T \\ &= \mathbf{D}_{e_d} \boldsymbol{\Sigma}_{e_d} \mathbf{D}_{e_d}^T - \mathbf{D}_{e_d} \boldsymbol{\Sigma}_{e_d} (\boldsymbol{\Sigma}_{e_d} + \mathbf{Z}_d \boldsymbol{\Sigma}_{u_d} \mathbf{Z}_d^T)^{-1} \boldsymbol{\Sigma}_{e_d} \mathbf{D}_{e_d}^T \end{aligned}$$

$$\begin{aligned} \mathbf{A}_{22} &= \mathbf{D}_{u_d} \mathbf{G}_d \mathbf{D}_{u_d}^T \\ &= \mathbf{D}_{u_d} \left(\boldsymbol{\Sigma}_{u_d} - \boldsymbol{\Sigma}_{u_d} \mathbf{Z}_d^T (\boldsymbol{\Sigma}_{e_d} + \mathbf{Z}_d \boldsymbol{\Sigma}_{u_d} \mathbf{Z}_d^T)^{-1} \mathbf{Z}_d \boldsymbol{\Sigma}_{u_d} \right) \mathbf{D}_{u_d}^T \\ &= \mathbf{D}_{u_d} \boldsymbol{\Sigma}_{u_d} \mathbf{D}_{u_d}^T - \mathbf{D}_{u_d} \boldsymbol{\Sigma}_{u_d} \mathbf{Z}_d^T (\boldsymbol{\Sigma}_{e_d} + \mathbf{Z}_d \boldsymbol{\Sigma}_{u_d} \mathbf{Z}_d^T)^{-1} \mathbf{Z}_d \boldsymbol{\Sigma}_{u_d} \mathbf{D}_{u_d}^T \end{aligned}$$

$$\begin{aligned} \mathbf{A}_{12} &= -\mathbf{D}_{e_d} \mathbf{Z}_d \mathbf{G}_d \mathbf{D}_{u_d}^T \\ &= -\mathbf{D}_{e_d} \mathbf{Z}_d \left(\boldsymbol{\Sigma}_{u_d} - \boldsymbol{\Sigma}_{u_d} \mathbf{Z}_d^T (\boldsymbol{\Sigma}_{e_d} + \mathbf{Z}_d \boldsymbol{\Sigma}_{u_d} \mathbf{Z}_d^T)^{-1} \mathbf{Z}_d \boldsymbol{\Sigma}_{u_d} \right) \mathbf{D}_{u_d}^T \\ &= -\mathbf{D}_{e_d} \mathbf{Z}_d \boldsymbol{\Sigma}_{u_d} \mathbf{D}_{u_d}^T + \mathbf{D}_{e_d} \mathbf{Z}_d \boldsymbol{\Sigma}_{u_d} \mathbf{Z}_d^T (\boldsymbol{\Sigma}_{e_d} + \mathbf{Z}_d \boldsymbol{\Sigma}_{u_d} \mathbf{Z}_d^T)^{-1} \mathbf{Z}_d \boldsymbol{\Sigma}_{u_d} \mathbf{D}_{u_d}^T \\ &= -\mathbf{D}_{e_d} \mathbf{Z}_d \boldsymbol{\Sigma}_{u_d} \mathbf{D}_{u_d}^T + \mathbf{D}_{e_d} (\boldsymbol{\Sigma}_{Y_d} - \boldsymbol{\Sigma}_{e_d}) (\boldsymbol{\Sigma}_{e_d} + \mathbf{Z}_d \boldsymbol{\Sigma}_{u_d} \mathbf{Z}_d^T)^{-1} \mathbf{Z}_d \boldsymbol{\Sigma}_{u_d} \mathbf{D}_{u_d}^T \\ &= -\mathbf{D}_{e_d} \boldsymbol{\Sigma}_{e_d} (\boldsymbol{\Sigma}_{e_d} + \mathbf{Z}_d \boldsymbol{\Sigma}_{u_d} \mathbf{Z}_d^T)^{-1} \mathbf{Z}_d \boldsymbol{\Sigma}_{u_d} \mathbf{D}_{u_d}^T \\ &= \mathbf{A}_{21}^T \end{aligned}$$

3.2 Population and Sample Distributions

Consider the partition $\mathbf{Y}_d = (\mathbf{Y}_{dr}^T, \mathbf{Y}_{ds}^T)^T$ where $\mathbf{Y}_{dr}$ is a $N_{dr} \times 1$ vector of the out-of-sample observations and $\mathbf{Y}_{ds}$ is a $n_d \times 1$ vector of the sample units, so that $N_{dr} + n_d = N_d$. Note that there is no loss of generality for ordering the elements of the population into out-of-sample and sample. In this section, we are interested in obtaining the distributions of the population vector $\mathbf{Y}_d$ and the sample data $\mathbf{Y}_{ds}$ assuming a non-informative sampling design. The distribution of $\mathbf{Y}_{ds}$ is used for estimating the parameters of the SN model using ML techniques. The closure properties of the CSN family are sufficient to get the distribution of $\mathbf{Y}_d$. The form of the vectors and matrices involved in the joint distribution of $\mathbf{Y}_d$ under the SN model are provided in the subsections below for the four situations (both e_{dj} and u_d follow SN distributions, e_{dj} follows SN distribution and u_d follows normal distribution, e_{dj} follows normal distribution and u_d follows SN distribution, and both e_{dj} and u_d follow normal distributions). The distribution of the sample characteristic vector $\mathbf{Y}_{ds}$ can be obtained in two ways. The first way is to assume the SN model for the units in the sample then derive the joint distribution of $\mathbf{Y}_{ds}$. The second way consists of first to obtain the joint population distribution of $\mathbf{Y}_d$ (assuming the SN model) then derive the distribution of $\mathbf{Y}_{ds}$ as a marginal. Under normality, these two ways of deriving the distribution of $\mathbf{Y}_{ds}$ give the same result. In appendix 3.5.3, we show that this result extends to the one fold SN model. Therefore, the distribution of $\mathbf{Y}_{ds}$ is the same as (3.15) where the population size N_d is replaced by the sample size n_d . Note that this is true under the one fold SN model but for a more general model the vector $\boldsymbol{\nu}_{Y_d}$ and the matrices $\mathbf{D}_{Y_d}$ and $\boldsymbol{\Gamma}_{Y_d}$ may not reduce. Typically if the errors e_{dj} are correlated or if the dimension of u_d is greater than 1, then the joint distribution of $\mathbf{Y}_{ds}$ obtained

using the first way may be different from the one from the marginal operation (second way). If these two distributions are different then the question is: which one should we use for parameter estimation?

3.2.1 u_d and e_{dj} follow SN Distribution

We derive in this section the joint distributions of $\mathbf{Y}_d$ and $\mathbf{Y}_{ds}$ for the model where both u_d and e_{dj} follow SN distributions.

$$\begin{aligned} Y_{dj} &= \mathbf{x}_{dj}^T \boldsymbol{\beta} + u_d + e_{dj}, \quad j = 1, \dots, N_d, \quad d = 1, \dots, m, \\ u_d &\overset{iid}{\sim} SN(\mu_u, \sigma_u^2, \lambda_u), \quad \text{and} \quad e_{dj} \overset{iid}{\sim} SN(\mu_e, \sigma_e^2, \lambda_e), \end{aligned} \quad (3.16)$$

where u_d and e_{dj} are independent for any d and j .

Remark. The model (3.16) is obtained from the general model (3.1) by letting $q = 1$ and $\mathbf{Z}_d = \mathbf{1}_{N_d}$. Many of the results below are given as a function of the quantity γ_x defined as:

$$\gamma_x = \frac{\sigma_u^2}{\sigma_e^2 + x\sigma_u^2} \quad (3.17)$$

where x is equal to the population size N_d or the sample size n_d for area d .

First we derive the distribution of the vectors $u_d \mathbf{1}_{N_d}$ and $\mathbf{e}_d$. From the discussion in Chapter 2, we know that the univariate SN is a special case of the CSN. In other words we have that $u_d \sim CSN_1(\mu_u, \sigma_u^2, \lambda_u \sigma_u^{-1}, 0, 1)$ and $e_{dj} \sim CSN_1(\mu_e, \sigma_e^2, \lambda_e \sigma_e^{-1}, 0, 1)$. Therefore the closure properties, presented in Chapter 2, give the distribution of the vectors $u_d \mathbf{1}_{N_d}$ and $\mathbf{e}_d$.

Using Proposition 2.6, it follows that:

$$u_d \mathbf{1}_{N_d} \sim SSN_1 \left(\mu_u \mathbf{1}_{N_d}, \sigma_u^2 \mathbf{1}_{N_d} \mathbf{1}_{N_d}^T, \frac{\lambda_u}{N_d \sigma_u} \mathbf{1}_{N_d}^T, 0, 1 \right).$$

Also, Proposition 2.2 leads to the following:

$$\mathbf{e}_d \sim CSN_{N_d} \left(\boldsymbol{\mu}_e, \sigma_e^2 \mathbf{I}_{N_d}, \frac{\lambda_e}{\sigma_e} \mathbf{I}_{N_d}, \mathbf{0}_{N_d}, \mathbf{I}_{N_d} \right).$$

Proposition 3.1 gives the general form of the distribution of $\mathbf{Y}_d$. The matrices $\mathbf{D}_{Y_d}$ and $\boldsymbol{\Gamma}_{Y_d}$ involve inverting matrices of the form $\boldsymbol{\Sigma}_1 + \mathbf{Z}\boldsymbol{\Sigma}_2\mathbf{Z}^T$ which reduce to $(a\mathbf{I} + b\mathbf{1}\mathbf{1}^T)^{-1}$ under the one-fold nested error. The inverse is obtained from the well-known result $(a\mathbf{I}_N + b\mathbf{1}_N\mathbf{1}_N^T)^{-1} = \frac{1}{a}\mathbf{I}_N - \frac{b}{a(a+nb)}\mathbf{1}_N\mathbf{1}_N^T$. Then after some algebra, we have:

$$\mathbf{Y}_d \sim CSN_{N_d, N_d+1} (\boldsymbol{\mu}_{Y_d}, \boldsymbol{\Sigma}_{Y_d}, \mathbf{D}_{Y_d}, \boldsymbol{\nu}_{Y_d}, \boldsymbol{\Gamma}_{Y_d}) \quad (3.18)$$

where,

$$\boldsymbol{\mu}_{Y_d} = \mathbf{X}_d \boldsymbol{\beta} + \mu_u \mathbf{1}_{N_d} + \boldsymbol{\mu}_e, \quad \boldsymbol{\Sigma}_{Y_d} = \sigma_e^2 \mathbf{I}_{N_d} + \sigma_u^2 \mathbf{1}_{N_d} \mathbf{1}_{N_d}^T, \quad \boldsymbol{\nu}_{Y_d} = \mathbf{0}_{N_d+1}$$

$$\mathbf{D}_{Y_d} = \begin{bmatrix} \frac{\lambda_e}{\sigma_e} [\mathbf{I}_{N_d} - \gamma_{N_d} \mathbf{1}_{N_d} \mathbf{1}_{N_d}^T] \\ \frac{\lambda_u}{\sigma_u} \gamma_{N_d} \mathbf{1}_{N_d}^T \end{bmatrix}, \text{ and}$$

$$\boldsymbol{\Gamma}_{Y_d} = \begin{bmatrix} \mathbf{I}_{N_d} + \lambda_e^2 \gamma_{N_d} \mathbf{1}_{N_d} \mathbf{1}_{N_d}^T & -\lambda_e \lambda_u \left(\frac{\sigma_e}{\sigma_u} \right) \gamma_{N_d} \mathbf{1}_{N_d} \\ -\lambda_e \lambda_u \left(\frac{\sigma_e}{\sigma_u} \right) \gamma_{N_d} \mathbf{1}_{N_d}^T & 1 + \lambda_u^2 (1 - N_d \gamma_{N_d}) \end{bmatrix}.$$

The sample data is used to estimate the parameters of the model (3.16) and the ML method shown in Section 3.4.1 requires the distribution of the sample characteristics $\mathbf{Y}_{ds}$. Using proposition 2.7 and after some matrix manipulation (detailed in appendix 3.5.3), we get that:

$$\mathbf{Y}_{ds} \sim CSN_{n_d, n_d+1} (\boldsymbol{\mu}_{Y_{ds}} = \boldsymbol{\mu}_{ds}, \boldsymbol{\Sigma}_{Y_{ds}} = \boldsymbol{\Sigma}_{dss}, \mathbf{D}_{Y_{ds}}, \boldsymbol{\nu}_{ds} = \boldsymbol{\nu}_{Y_d}, \boldsymbol{\Gamma}_{Y_{ds}}) \quad (3.19)$$

where, (3.19) is equivalent to the joint distribution (3.18) with N_d replaced by n_d .

3.2.2 u_d follows Normal and e_{dj} follows SN

We derive here results for the model where the e_{dj} 's follow a SN distribution and the u_d 's follow a normal distribution. Note that the normal distribution is a special case of the SN distribution where the shape parameter equals 0. Hence, we may write the distribution of u_d as a CSN and apply the closure properties.

$$Y_{dj} = \mathbf{x}_{dj}^T \boldsymbol{\beta} + u_d + e_{dj}, \quad j = 1, \dots, N_d, \quad d = 1, \dots, m.$$

$$u_d \stackrel{iid}{\sim} N(\mu_u, \sigma_u^2) \equiv CSN_1(\mu_u, \sigma_u^2, 0, 0, 1), \quad \text{and} \quad e_{dj} \stackrel{iid}{\sim} SN(\mu_e, \sigma_e^2, \lambda_e) \quad (3.20)$$

Using the same approach as for the more general model (3.2.1), that

$$\mathbf{Y}_d \sim CSN_{N_d, N_d+1}(\boldsymbol{\mu}_{Y_d}, \boldsymbol{\Sigma}_{Y_d}, \mathbf{D}_{Y_d}, \boldsymbol{\nu}_{Y_d}, \boldsymbol{\Gamma}_{Y_d}) \quad (3.21)$$

where,

$$\boldsymbol{\mu}_{Y_d} = \mathbf{X}_d \boldsymbol{\beta} + \mu_u \mathbf{1}_{N_d} + \boldsymbol{\mu}_e, \quad \boldsymbol{\Sigma}_{Y_d} = \sigma_e^2 \mathbf{I}_{N_d} + \sigma_u^2 \mathbf{1}_{N_d} \mathbf{1}_{N_d}^T, \quad \boldsymbol{\nu}_{Y_d} = \mathbf{0}_{N_d+1}$$

$$\mathbf{D}_{Y_d} = \begin{bmatrix} \frac{\lambda_e}{\sigma_e} (\mathbf{I}_{N_d} - \gamma_{N_d} \mathbf{1}_{N_d} \mathbf{1}_{N_d}^T) \\ \mathbf{0}_{(1 \times n_d)} \end{bmatrix}, \quad \text{and} \quad \boldsymbol{\Gamma}_{Y_d} = \begin{bmatrix} \mathbf{I}_{N_d} + \lambda_e^2 \gamma_{N_d} \mathbf{1}_{N_d} \mathbf{1}_{N_d}^T & \mathbf{0}_{(N_d \times 1)} \\ \mathbf{0}_{(1 \times N_d)} & 1 \end{bmatrix}.$$

The sample distribution of $\mathbf{Y}_{ds}$ is the same as (3.21) where N_d is replaced by n_d .

3.2.3 u_d follows SN and e_{dj} follows Normal

We derive here results for the model where the u_d 's follow a normal distribution and the e_{dj} 's follow a SN distribution. Again, the normal distribution is a special case of

the SN distribution where the shape parameter equals 0.

$$\begin{aligned} Y_{dj} &= \mathbf{x}_{dj}^T \boldsymbol{\beta} + u_d + e_{dj}, \quad j = 1, \dots, N_d, \quad d = 1, \dots, m. \\ u_d &\stackrel{iid}{\sim} SN(\mu_u, \sigma_u^2, \lambda_u), \quad \text{and} \quad e_{dj} \stackrel{iid}{\sim} N(\mu_e, \sigma_e^2) \end{aligned} \quad (3.22)$$

The probability density function of u_d is the same as in the general model 3.16 but e_{dj} is distributed as $CSN_1(\mu_e, \sigma_e, 0, 0, 1)$. It is easy to show, using the same approach as for the general model, that

$$\mathbf{Y}_d \sim CSN_{N_d, N_d+1}(\boldsymbol{\mu}_{Y_d}, \boldsymbol{\Sigma}_{Y_d}, \mathbf{D}_{Y_d}, \boldsymbol{\nu}_{Y_d}, \boldsymbol{\Gamma}_{Y_d}) \quad (3.23)$$

where,

$$\begin{aligned} \boldsymbol{\mu}_{Y_d} &= \mathbf{X}_d \boldsymbol{\beta} + \mu_u \mathbf{1}_{N_d} + \boldsymbol{\mu}_e, \quad \boldsymbol{\Sigma}_{Y_d} = \sigma_e^2 \mathbf{I}_{N_d} + \sigma_u^2 \mathbf{1}_{N_d} \mathbf{1}_{N_d}^T, \quad \boldsymbol{\nu}_{Y_d} = \mathbf{0}_{N_d+1} \\ \mathbf{D}_{Y_d} &= \begin{bmatrix} \mathbf{0}_{(N_d \times N_d)} \\ \frac{\lambda_u}{\sigma_u} \gamma_{N_d} \mathbf{1}_{N_d}^T \end{bmatrix}, \quad \text{and} \quad \boldsymbol{\Gamma}_{Y_d} = \begin{bmatrix} \mathbf{I}_{N_d} & \mathbf{0}_{(N_d \times 1)} \\ \mathbf{0}_{(1 \times N_d)} & 1 + \lambda_u^2 (1 - N_d \gamma_{N_d}) \end{bmatrix}. \end{aligned}$$

The sample distribution of $\mathbf{Y}_{ds}$ is the same as (3.23) where N_d is replaced by n_d .

3.2.4 Both u_d and e_{dj} follow normal distribution

This is the usual case covered in the textbooks and the theory is well developed (see McCulloch et al. (2008)). By letting $\lambda_u = 0$ and $\lambda_e = 0$ in the general model, we get the well know results:

$$\mathbf{Y}_d \sim CSN_{N_d, N_d+1}(\boldsymbol{\mu}_d, \boldsymbol{\Sigma}_d, \mathbf{0}, \mathbf{0}, \mathbf{I}) \equiv N_{N_d}(\boldsymbol{\mu}_d, \boldsymbol{\Sigma}_d) \quad (3.24)$$

and

$$\mathbf{Y}_{ds} \sim N_{n_d}(\boldsymbol{\mu}_{ds}, \boldsymbol{\Sigma}_{dss}) \quad (3.25)$$

3.3 Moment Generating Function and Moments

The method of moments and related estimation methods use the moments of the given distribution to estimate the parameters of the model. Here, we show how to derive the moments under the SN model. However, these expressions are so cumbersome that, in practice, they are not usable. The generating function of the random vector $\mathbf{Y}_{ds}$ is obtained by applying the general result (2.15). Because $\boldsymbol{\nu} = \mathbf{0}$, we have

$$\begin{aligned} \boldsymbol{\Gamma}_{Y_{ds}} + \mathbf{D}_{Y_{ds}} \boldsymbol{\Sigma}_{Y_{ds}} \mathbf{D}_{Y_{ds}}^T &= \mathbf{I}_{n_d+1} + \begin{bmatrix} \mathbf{D}_{e_d} \boldsymbol{\Sigma}_{e_d} \mathbf{D}_{e_d}^T & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_{u_d} \boldsymbol{\Sigma}_{u_d} \mathbf{D}_{u_d}^T \end{bmatrix} \\ &= \begin{bmatrix} 1 + \lambda_e^2 & \dots & 0 & 0 \\ \vdots & \ddots & \mathbf{0} & \mathbf{0} \\ 0 & \mathbf{0} & 1 + \lambda_e^2 & 0 \\ 0 & \mathbf{0} & 0 & 1 + \lambda_u^2 \end{bmatrix}, \text{ and} \end{aligned}$$

$$\mathbf{D}_{Y_{ds}} \boldsymbol{\Sigma}_{Y_{ds}} = \begin{bmatrix} \mathbf{D}_{e_d} \boldsymbol{\Sigma}_{e_d} (\boldsymbol{\Sigma}_{Y_{ds}})^{-1} \\ \mathbf{D}_{u_d} \boldsymbol{\Sigma}_{u_d} \mathbf{Z}_d^T (\boldsymbol{\Sigma}_{Y_{ds}})^{-1} \end{bmatrix} \boldsymbol{\Sigma}_{Y_{ds}} = \begin{bmatrix} \lambda_e \sigma_e \mathbf{I} \\ \lambda_u \sigma_u \mathbf{1}^T \end{bmatrix};$$

we also have,

$$\begin{aligned} \Phi_q(\mathbf{D}_{Y_{ds}} \boldsymbol{\Sigma}_{Y_{ds}} \mathbf{t}; \boldsymbol{\nu}_{Y_{ds}}, \boldsymbol{\Gamma}_{Y_{ds}} + \mathbf{D}_{Y_{ds}} \boldsymbol{\Sigma}_{Y_{ds}} \mathbf{D}_{Y_{ds}}^T) &= \prod_{j=1}^{n_d} \Phi_1(\lambda_e \sigma_e t_j; 0, 1 + \lambda_e^2) \times \\ &\quad \Phi_1(\lambda_u \sigma_u t_{n_d+1}; 0, 1 + \lambda_u^2) \end{aligned} \quad (3.26)$$

and

$$\Phi_q(0; \boldsymbol{\nu}_{Y_{ds}}, \boldsymbol{\Gamma}_{Y_{ds}} + \mathbf{D}_{Y_{ds}} \boldsymbol{\Sigma}_{Y_{ds}} \mathbf{D}_{Y_{ds}}^T) = \left(\frac{1}{2}\right)^{n_d+1} \quad (3.27)$$

Therefore, the moment generating function simplifies to:

$$M_{\mathbf{Y}_{ds}}(\mathbf{t}) = 2^{n_d+1} e^{\left(\mathbf{t}^T \boldsymbol{\mu}_{Y_{ds}} + \frac{1}{2} \mathbf{t}^T \boldsymbol{\Sigma}_{Y_{ds}} \mathbf{t}\right)} \Phi_1(\lambda_u \sigma_u \sum_{j=1}^{n_d} t_j; 0, 1 + \lambda_u^2) \prod_{j=1}^{n_d} \Phi_1(\lambda_e \sigma_e t_j; 0, 1 + \lambda_e^2) \quad (3.28)$$

The partial derivative of (3.28) with respect to t_j is equal to:

$$\begin{aligned} \frac{\partial}{\partial t_j} M_{\mathbf{Y}_{ds}}(\mathbf{t}) = & 2^{n_d+1} \left((\mu_j + [\boldsymbol{\Sigma}_{Y_{ds}} \mathbf{t}]_j) \Phi_1(\lambda_u \sigma_u \sum_{j=1}^{n_d} t_j; 0, 1 + \lambda_u^2) \Phi_1(\lambda_e \sigma_e t_j; 0, 1 + \lambda_e^2) \right. \\ & + \lambda_u \sigma_u \phi_1(\lambda_u \sigma_u \sum_{j=1}^{n_d} t_j; 0, 1 + \lambda_u^2) \Phi_1(\lambda_e \sigma_e t_j; 0, 1 + \lambda_e^2) \\ & \left. + \lambda_e \sigma_e \phi_1(\lambda_e \sigma_e t_j; 0, 1 + \lambda_e^2) \Phi_1(\lambda_u \sigma_u \sum_{j=1}^{n_d} t_j; 0, 1 + \lambda_u^2) \right) \\ & \times e^{\left(\mathbf{t}^T \boldsymbol{\mu}_{Y_{ds}} + \frac{1}{2} \mathbf{t}^T \boldsymbol{\Sigma}_{Y_{ds}} \mathbf{t}\right)} \prod_{i \neq j}^{n_d} \Phi_1(\lambda_e \sigma_e t_i; 0, 1 + \lambda_e^2) \end{aligned} \quad (3.29)$$

where $[\boldsymbol{\Sigma}_{Y_{ds}} \mathbf{t}]_j$ denotes the j^{th} element of the vector $\boldsymbol{\Sigma}_{Y_{ds}} \mathbf{t}$. It follows from the general result $\mathbb{E}(\mathbf{Y}^k) = \frac{\partial}{\partial^k \mathbf{t}} M_{\mathbf{Y}_{ds}}(\mathbf{t})|_{\mathbf{t}=\mathbf{0}}$, that the expectation of Y_{ds} is equal to:

$$\mathbb{E}(\mathbf{Y}_{ds}) = \boldsymbol{\mu}_{Y_{ds}} + (\sigma_u \delta_u + \sigma_e \delta_e) \sqrt{\frac{2}{\pi}} \mathbf{1}_{n_d} \quad (3.30)$$

It is possible to compute higher derivations of the moment generating function, however the expressions are very lengthy and cumbersome as one can see in the expressions (3.31) and (3.33) representing the second partial derivatives. The second

partial derivative with respect to t_j is:

$$\begin{aligned}
\frac{\partial}{\partial^2 t_j} M_{\mathbf{Y}_{ds}}(\mathbf{t}) = & 2^{n_d+1} \left((\sigma_e^2 + \sigma_u^2) \Phi_1(\lambda_u \sigma_u \sum_{j=1}^{n_d} t_j; 0, 1 + \lambda_u^2) \Phi_1(\lambda_e \sigma_e t_j; 0, 1 + \lambda_e^2) \right. \\
& + \lambda_u \sigma_u (\mu_j + [\Sigma_{Y_{ds}} \mathbf{t}]_j) \phi_1(\lambda_u \sigma_u \sum_{j=1}^{n_d} t_j; 0, 1 + \lambda_u^2) \Phi_1(\lambda_e \sigma_e t_j; 0, 1 + \lambda_e^2) \\
& + \lambda_e \sigma_e (\mu_j + [\Sigma_{Y_{ds}} \mathbf{t}]_j) \Phi_1(\lambda_u \sigma_u \sum_{j=1}^{n_d} t_j; 0, 1 + \lambda_u^2) \phi_1(\lambda_e \sigma_e t_j; 0, 1 + \lambda_e^2) \\
& - \lambda_u \sigma_u^3 \lambda_u^2 \sum_{j=1}^{n_d} t_j \phi_1(\lambda_u \sigma_u \sum_{j=1}^{n_d} t_j; 0, 1 + \lambda_u^2) \Phi_1(\lambda_e \sigma_e t_j; 0, 1 + \lambda_e^2) \\
& + \lambda_u \sigma_u \lambda_e \sigma_e \phi_1(\lambda_u \sigma_u \sum_{j=1}^{n_d} t_j; 0, 1 + \lambda_u^2) \phi_1(\lambda_e \sigma_e t_j; 0, 1 + \lambda_e^2) \\
& - \lambda_e \sigma_e^3 \lambda_e^2 t_j \Phi_1(\lambda_u \sigma_u \sum_{j=1}^{n_d} t_j; 0, 1 + \lambda_u^2) \phi_1(\lambda_e \sigma_e t_j; 0, 1 + \lambda_e^2) \\
& + \lambda_u \sigma_u \lambda_e \sigma_e \phi_1(\lambda_u \sigma_u \sum_{j=1}^{n_d} t_j; 0, 1 + \lambda_u^2) \phi_1(\lambda_e \sigma_e t_j; 0, 1 + \lambda_e^2) \\
& + ((\mu_j + [\Sigma_{Y_{ds}} \mathbf{t}]_j) \Phi_1(\lambda_u \sigma_u \sum_{j=1}^{n_d} t_j; 0, 1 + \lambda_u^2) \Phi_1(\lambda_e \sigma_e t_j; 0, 1 + \lambda_e^2) \\
& + \lambda_u \sigma_u \phi_1(\lambda_u \sigma_u \sum_{j=1}^{n_d} t_j; 0, 1 + \lambda_u^2) \Phi_1(\lambda_e \sigma_e t_j; 0, 1 + \lambda_e^2) \\
& + \lambda_e \sigma_e \phi_1(\lambda_e \sigma_e t_j; 0, 1 + \lambda_e^2) \Phi_1(\lambda_u \sigma_u \sum_{j=1}^{n_d} t_j; 0, 1 + \lambda_u^2)) (\mu_j + [\Sigma_{Y_{ds}} \mathbf{t}]_j) \Big) \\
& \times e^{(\mathbf{t}^T \boldsymbol{\mu}_{Y_{ds}} + \frac{1}{2} \mathbf{t}^T \Sigma_{Y_{ds}} \mathbf{t})} \prod_{i \neq j}^{n_d} \Phi_1(\lambda_e \sigma_e t_i; 0, 1 + \lambda_e^2)
\end{aligned} \tag{3.31}$$

From (3.31), we find:

$$\begin{aligned}
\mathbb{E}(Y_{dj}^2) &= \frac{\partial}{\partial^2 t} M_{\mathbf{Y}_{ds}}(\mathbf{t})|_{\mathbf{t}=\mathbf{0}} \\
&= \sigma_u^2 + \sigma_e^2 + \mu_j^2 + 2\sigma_u \sigma_u \sqrt{\frac{2}{\pi}} \mu_j + 2\sigma_e \sigma_e \sqrt{\frac{2}{\pi}} \mu_j + \frac{4}{\pi} \sigma_u \delta_u \sigma_e \delta_e
\end{aligned}$$

Using $\text{Var}(Y_{dj}) = \mathbb{E}(Y_{dj}^2) - \mathbb{E}(Y_{dj})^2$ gives:

$$\text{Var}(Y_{dj}) = \sigma_u^2(1 - \delta_u^2 \frac{2}{\pi}) + \sigma_e^2(1 - \delta_e^2 \frac{2}{\pi}) \quad (3.32)$$

To find the covariances, we derive the second partial derivative with respect to t_k :

$$\begin{aligned} \frac{\partial}{\partial t_j \partial t_k} M_{\mathbf{Y}_{ds}}(\mathbf{t}) = & 2^{n_d+1} \left(\left(\sigma_u^2 \Phi_1(\lambda_u \sigma_u \sum_{j=1}^{n_d} t_j; 0, 1 + \lambda_u^2) \Phi_1(\lambda_e \sigma_e t_j; 0, 1 + \lambda_e^2) \right. \right. \\ & + \lambda_u \sigma_u (\mu_j + [\Sigma_{Y_{ds}} \mathbf{t}]_j) \phi_1(\lambda_u \sigma_u \sum_{j=1}^{n_d} t_j; 0, 1 + \lambda_u^2) \Phi_1(\lambda_e \sigma_e t_j; 0, 1 + \lambda_e^2) \\ & - \lambda_u \sigma_u^3 \lambda_u^2 \sum_{j=1}^{n_d} t_j \phi_1(\lambda_u \sigma_u \sum_{j=1}^{n_d} t_j; 0, 1 + \lambda_u^2) \Phi_1(\lambda_e \sigma_e t_j; 0, 1 + \lambda_e^2) \\ & + \lambda_u \sigma_u \lambda_e \sigma_e \phi_1(\lambda_u \sigma_u \sum_{j=1}^{n_d} t_j; 0, 1 + \lambda_u^2) \phi_1(\lambda_e \sigma_e t_j; 0, 1 + \lambda_e^2) \\ & + ((\mu_j + [\Sigma_{Y_{ds}} \mathbf{t}]_j) \Phi_1(\lambda_u \sigma_u \sum_{j=1}^{n_d} t_j; 0, 1 + \lambda_u^2) \Phi_1(\lambda_e \sigma_e t_j; 0, 1 + \lambda_e^2) \\ & + \lambda_u \sigma_u \phi_1(\lambda_u \sigma_u \sum_{j=1}^{n_d} t_j; 0, 1 + \lambda_u^2) \Phi_1(\lambda_e \sigma_e t_j; 0, 1 + \lambda_e^2) \\ & \left. \left. + \lambda_e \sigma_e \phi_1(\lambda_e \sigma_e t_j; 0, 1 + \lambda_e^2) \Phi_1(\lambda_u \sigma_u \sum_{j=1}^{n_d} t_j; 0, 1 + \lambda_u^2)) (\mu_k + [\Sigma_{Y_{ds}} \mathbf{t}]_k) \right) \right) \\ & \times \Phi_1(\lambda_e \sigma_e t_k; 0, 1 + \lambda_e^2) + \lambda_e \sigma_e \phi_1(\lambda_e \sigma_e t_k; 0, 1 + \lambda_e^2) \\ & \times \left((\mu_j + [\Sigma_{Y_{ds}} \mathbf{t}]_j) \Phi_1(\lambda_u \sigma_u \sum_{j=1}^{n_d} t_j; 0, 1 + \lambda_u^2) \Phi_1(\lambda_e \sigma_e t_j; 0, 1 + \lambda_e^2) \right. \\ & + \lambda_u \sigma_u \phi_1(\lambda_u \sigma_u \sum_{j=1}^{n_d} t_j; 0, 1 + \lambda_u^2) \Phi_1(\lambda_e \sigma_e t_j; 0, 1 + \lambda_e^2) \\ & \left. \left. + \lambda_e \sigma_e \phi_1(\lambda_e \sigma_e t_j; 0, 1 + \lambda_e^2) \Phi_1(\lambda_u \sigma_u \sum_{j=1}^{n_d} t_j; 0, 1 + \lambda_u^2) \right) \right) \\ & \times e^{(\mathbf{t}^T \boldsymbol{\mu}_{Y_{ds}} + \frac{1}{2} \mathbf{t}^T \Sigma_{Y_{ds}} \mathbf{t})} \prod_{i \neq j, i \neq k}^{n_d} \Phi_1(\lambda_e \sigma_e t_i; 0, 1 + \lambda_e^2) \end{aligned} \quad (3.33)$$

Evaluating (3.33) at $\mathbf{t} = \mathbf{0}$, provides the expression of the expectation of the product $Y_{dj}Y_{dk}$ as:

$$\mathbb{E}(Y_{dj}Y_{dk}) = \sigma_u^2 + \frac{2}{\pi}\sigma_e^2\delta_e^2 + \frac{4}{\pi}\sigma_u\delta_u\sigma_e\delta_e + \mu_j\mu_k + (\mu_j + \mu_k)(\sigma_u\delta_u + \sigma_e\delta_e)\sqrt{\frac{2}{\pi}} \quad (3.34)$$

Using $\mathbb{C}ov(Y_{dj}, Y_{dk}) = \mathbb{E}(Y_{dj}Y_{dk}) - \mathbb{E}(Y_{dj})\mathbb{E}(Y_{dk})$ yields:

$$\mathbb{C}ov(Y_{dj}, Y_{dk}) = \sigma_u^2(1 - \delta_u^2\frac{2}{\pi}) \quad (3.35)$$

Combining results (3.32) and (3.35), we find the covariance matrix of Y_{ds} :

$$\mathbb{C}ov(\mathbf{Y}_{ds}) = \sigma_u^2(1 - \delta_u^2\frac{2}{\pi})\mathbf{1}_{n_d}\mathbf{1}_{n_d}^T + \sigma_e^2(1 - \delta_e^2\frac{2}{\pi})\mathbf{I}_{n_d} \quad (3.36)$$

The derivatives of the moment generating function can be used to derive method of moment estimates for the parameters of the SN model. For the SN model, at least 6 equations are needed to estimate μ_u , σ_u , λ_u , μ_e , σ_e , and λ_e . If the random errors $\mathbb{E}(u_d) = \mathbb{E}(e_{dj}) = 0$ then only 4 equations are needed since the location parameters μ_u and μ_e would be function of σ_u , λ_u , σ_e , and λ_e . In addition to those equations, extra equations may be necessary to estimate the fixed effects. Even though the method of moment is theoretically simple, in this case, the mathematical expressions are complex to handle. Flecher et al. (2009) proposed a method of moment method for estimating CSN distributions; however they mentioned the difficulty of deriving third and fourth moments even for the univariate distribution. In Section 3.4, we show ML and DC estimation methods for the parameters of the SN model.

3.4 Parameters Estimation

In the previous sections, we derived the density functions of $\mathbf{Y}_d$ and $\mathbf{Y}_{ds}$ which are function of the parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mu_u, \mu_e, \sigma_u^2, \sigma_e^2, \lambda_u, \lambda_e)$. These parameters are unknown and must be estimated. Note that if $\mathbb{E}(u_d) = \mathbb{E}(e_{dj}) = 0$ then only σ_u^2 , λ_u , σ_e^2 , and λ_e must be estimated. To estimate these parameters, we use the joint distribution of the sample vector $\mathbf{Y}_{ds}$. In the case of the normal model, several methods based on maximizing the likelihood function or some related functions are widely available in all major statistical software packages. For the SN model, the complexity of the estimation process is essentially due to the term $\Phi_q(\mathbf{D}(\mathbf{y} - \boldsymbol{\mu}); \boldsymbol{\nu}, \boldsymbol{\Gamma})$ in the pdf of the CSN (2.12) which does not have a closed form and requires multidimensional numerical integrals. Another difficulty is the weak identifiability of λ_u due to its small contribution to the likelihood i.e. the likelihood function remains nearly flat for different values of λ_u .

3.4.1 Maximum Likelihood Estimation

The likelihood function associated with the model (3.19) is:

$$L(\boldsymbol{\theta}|\mathbf{y}) = C\phi_n(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \Phi_{n+m}(\mathbf{D}(\mathbf{y} - \boldsymbol{\mu}); \boldsymbol{\nu}, \boldsymbol{\Gamma}), \quad \mathbf{y} \in \mathbb{R}^n, \quad (3.37)$$

where: $\mathbf{y} = (\mathbf{y}_1^T, \dots, \mathbf{y}_m^T)^T$, $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mu_u, \mu_e, \sigma_u^2, \sigma_e^2, \lambda_u, \lambda_e)^T$,
 $C^{-1} = \Phi_{n+m}(\mathbf{0}; \boldsymbol{\nu}, \boldsymbol{\Gamma} + \mathbf{D}\boldsymbol{\Sigma}\mathbf{D}^T)$,

$$\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} + \mu_u \mathbf{1}_n + \boldsymbol{\mu}_e, \boldsymbol{\Sigma} = \bigoplus_{d=1}^m \boldsymbol{\Sigma}_{Y_{ds}} = \begin{bmatrix} \boldsymbol{\Sigma}_{Y_{1s}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{\Sigma}_{Y_{ms}} \end{bmatrix},$$

$$\mathbf{D} = \bigoplus_{d=1}^m \mathbf{D}_{Y_{ds}} = \begin{bmatrix} \mathbf{D}_{Y_{1s}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{D}_{Y_{ms}} \end{bmatrix}, \quad \boldsymbol{\Gamma} = \bigoplus_{d=1}^m \boldsymbol{\Gamma}_{Y_{ds}} = \begin{bmatrix} \boldsymbol{\Gamma}_{Y_{1s}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{\Gamma}_{Y_{ms}} \end{bmatrix},$$

$$\boldsymbol{\nu} = \mathbf{0}, \text{ with } \boldsymbol{\Sigma}_{Y_{ds}}, \mathbf{D}_{Y_{ds}}, \boldsymbol{\nu}_{Y_{ds}}, \text{ and } \boldsymbol{\Gamma}_{Y_{ds}}, d = 1, \dots, m \text{ as in 3.19}$$

The block diagonal form of the matrices ($\boldsymbol{\Sigma}$, $\mathbf{D}$, and $\boldsymbol{\Gamma}$) is the consequence of the independence assumption between the m small areas. The block diagonal form simplifies the matrix calculation and reduces computer running times. The log-likelihood function corresponding to (3.37) is, up to a constant, equal to:

$$l(\boldsymbol{\theta}|\mathbf{y}) = \ln(C) - \frac{1}{2} \ln|\boldsymbol{\Sigma}| - \frac{1}{2} (\mathbf{y} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu}) + \ln(\Phi_{n+m}(\mathbf{D}(\mathbf{y} - \boldsymbol{\mu}); \boldsymbol{\nu}, \boldsymbol{\Gamma})) \quad (3.38)$$

We are interested in finding the value of $\boldsymbol{\theta}$ that minimizes $l(\boldsymbol{\theta}|\mathbf{y})$. In other words, the ML estimates are given by:

$$\boldsymbol{\theta} = \arg \max_{\boldsymbol{\theta} \in \Theta} l(\boldsymbol{\theta}|\mathbf{y}) \quad (3.39)$$

The expression (3.38) is basically the log-likelihood function of the distribution $N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ plus the terms $\ln(C)$ and $\ln(\Phi_{n+m}(\mathbf{D}(\mathbf{y} - \boldsymbol{\mu}); \boldsymbol{\nu}, \boldsymbol{\Gamma}))$. Not surprisingly, the difficulties of maximizing the log-likelihood (3.38) come from these two latter terms. The term $\ln(C)$ is difficult to compute in the general case, however for the one-fold SN model the analytical expression is easy to obtain because of the diagonal form of the following

matrix

$$\mathbf{\Gamma}_{Y_{ds}} + \mathbf{D}_{Y_{ds}} \mathbf{\Sigma}_{Y_{ds}} \mathbf{D}_{Y_{ds}}^T = \begin{bmatrix} 1 + \lambda_e^2 & \dots & 0 & 0 \\ \vdots & \ddots & \mathbf{0} & \mathbf{0} \\ 0 & \mathbf{0} & 1 + \lambda_e^2 & 0 \\ 0 & \mathbf{0} & 0 & 1 + \lambda_u^2 \end{bmatrix}$$

The diagonal form above is a consequence of the fact that $\mathbf{D}_{e_{ds}}$ and $\mathbf{\Sigma}_{e_{ds}}$ are diagonal matrices and $\mathbf{D}_{u_{ds}} \mathbf{\Sigma}_{u_{ds}} \mathbf{D}_{u_{ds}}^T$ is of size (1×1) . Therefore,

$$\ln(C) = -\ln(\Phi_{n+m}(\mathbf{0}; \boldsymbol{\nu}, \mathbf{\Gamma} + \mathbf{D}\mathbf{\Sigma}\mathbf{D}^T)) = \ln(2^{n+m})$$

The term $\ln(\Phi_{n+m}(\mathbf{D}(\mathbf{y} - \boldsymbol{\mu}); \boldsymbol{\nu}, \mathbf{\Gamma}))$ is more complex. First, there is no closed-form expression for $\ln(\Phi_{n+m}(\mathbf{D}(\mathbf{y} - \boldsymbol{\mu}); \boldsymbol{\nu}, \mathbf{\Gamma}))$ meaning that numerical approximations are necessary. Second, $\ln(\Phi_{n+m}(\mathbf{D}(\mathbf{y} - \boldsymbol{\mu}); \boldsymbol{\nu}, \mathbf{\Gamma}))$ requires evaluation of multi-dimensional integrals. However, due to the block-diagonal aspect of $\mathbf{\Gamma}$, we only have to evaluate the m terms $\ln(\Phi_{n_d+1}(\mathbf{D}_{Y_{ds}}(\mathbf{y}_{ds} - \boldsymbol{\mu}_{Y_{ds}}); \boldsymbol{\nu}_{Y_{ds}}, \mathbf{\Gamma}_{Y_{ds}}))$ separately and sum them up. This is important because it reduces the dimension of the integrals from $n + m$ to $n_d + 1$. Because of the small sample sizes involved in SAE problems, the $(n_d + 1)$ -dimensional integrals can be much more reasonable. Some statistical software provides routines to compute multivariate normal cumulative probabilities. For example, R has at least two functions for computing cumulative probabilities of multivariate normal. The function `PMVNORM`, from the `MVTNORM` package, computes the cumulative probabilities of the multivariate normal using algorithms based on the Genz and Bretz method. Genz and Bretz's method is stochastic therefore the results from `PMVNORM` are random. This can be a problem when using optimization routines because the randomness of the objective function makes it difficult to compute derivatives. The second function is `SADMVN` from the `MNORMT` package. `SADMVN` provides deterministic results which are more appropriate for

finding derivatives. However, the maximum dimension of the integral that SADSMVN can evaluate is 20. When $n_d + 1$ is greater than 20, users may need to write their own routines to evaluate the integrals. And even when $n_d + 1$ is less than 20 but close to 20, the estimation of the cumulative probabilities are unstable.

Appendix 3.5.1 provides details on how to take advantage of the special structure of the covariance matrix to simplify the calculation of the cumulative probabilities. For the nested error model, we are able to use result (3.68) to reduce the $(n_d + 1)$ -dimensional integral to $n_d + 1$ one-dimensional integrals. Not only does it greatly simplify the numerical evaluation of $\ln(\Phi_{n_d+1}(\mathbf{D}_{Y_{ds}}(\mathbf{y}_{Y_{ds}} - \boldsymbol{\mu}_{Y_{ds}}); \boldsymbol{\nu}_{Y_{ds}}, \boldsymbol{\Gamma}_{Y_{ds}}))$, but it also leads to simple expressions of the gradients (see appendix 3.5.2 for more details). Another advantage of the simplification is that any statistical software can be used to evaluate the one-dimensional integrals which are basically cumulative probabilities of the standard normal. The simplification leads to:

$$\Phi_{n_d+1}(\mathbf{D}_{Y_{ds}}(\mathbf{y}_{ds} - \boldsymbol{\mu}_{Y_{ds}}); \boldsymbol{\nu}_{Y_{ds}}, \boldsymbol{\Gamma}_{Y_{ds}}) = \int_{-\infty}^{\infty} \prod_{j=1}^{n_d+1} \Phi_1\left(\frac{z_j - c_j u_0}{\sqrt{1 - c_j^2}}\right) \phi(u_0) du_0 \quad (3.40)$$

where c_j and z_j are defined respectively in (3.76) and (3.77). The kernel of the log likelihood is:

$$\begin{aligned} l_k(\boldsymbol{\theta}|\mathbf{y}) = & -\frac{1}{2} \sum_{d=1}^m \ln|\boldsymbol{\Sigma}_{Y_{ds}}| - \frac{1}{2} \sum_{d=1}^m (\mathbf{y}_{ds} - \boldsymbol{\mu}_{Y_{ds}})^T \boldsymbol{\Sigma}_{Y_{ds}}^{-1} (\mathbf{y}_{ds} - \boldsymbol{\mu}_{Y_{ds}}) \\ & + \sum_{d=1}^m \ln(\Phi_{n_d+1}(\mathbf{D}_{Y_{ds}}(\mathbf{y}_{ds} - \boldsymbol{\mu}_{Y_{ds}}); \boldsymbol{\nu}_{Y_{ds}}, \boldsymbol{\Gamma}_{Y_{ds}})) \end{aligned} \quad (3.41)$$

With the partition of the parameters as follows: $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\lambda})$, the partial derivatives

of the log likelihood function $l(\boldsymbol{\theta})$ are:

$$\frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\theta}) = \boldsymbol{\mu}_{Y_{ds}(\boldsymbol{\beta})}^T \boldsymbol{\Sigma}_{Y_{ds}}^{-1} (\mathbf{y}_{ds} - \boldsymbol{\mu}_{Y_{ds}}) + \frac{\Phi_{(\boldsymbol{\beta})}(\mathbf{D}_{Y_{ds}}(\mathbf{y}_{ds} - \boldsymbol{\mu}_{Y_{ds}}); \boldsymbol{\nu}_{Y_{ds}}, \boldsymbol{\Gamma}_{Y_{ds}})}{\Phi(\mathbf{D}_{Y_{ds}}(\mathbf{y}_{ds} - \boldsymbol{\mu}_{Y_{ds}}); \boldsymbol{\nu}_{Y_{ds}}, \boldsymbol{\Gamma}_{Y_{ds}})} \quad (3.42)$$

$$\frac{\partial}{\partial \boldsymbol{\mu}} l(\boldsymbol{\theta}) = \boldsymbol{\mu}_{Y_{ds}(\boldsymbol{\mu})}^T \boldsymbol{\Sigma}_{Y_{ds}}^{-1} (\mathbf{y}_{ds} - \boldsymbol{\mu}_{Y_{ds}}) + \frac{\Phi_{(\boldsymbol{\mu})}(\mathbf{D}_{Y_{ds}}(\mathbf{y}_{ds} - \boldsymbol{\mu}_{Y_{ds}}); \boldsymbol{\nu}_{Y_{ds}}, \boldsymbol{\Gamma}_{Y_{ds}})}{\Phi(\mathbf{D}_{Y_{ds}}(\mathbf{y}_{ds} - \boldsymbol{\mu}_{Y_{ds}}); \boldsymbol{\nu}_{Y_{ds}}, \boldsymbol{\Gamma}_{Y_{ds}})} \quad (3.43)$$

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\sigma}} l(\boldsymbol{\theta}) = & -\frac{1}{2} \left[\text{tr}(\boldsymbol{\Sigma}_{Y_{ds}}^{-1} \boldsymbol{\Sigma}_{Y_{ds}}(\boldsymbol{\sigma})) - (\mathbf{y}_{ds} - \boldsymbol{\mu}_{Y_{ds}})^T \boldsymbol{\Sigma}_{Y_{ds}}^{-1} \boldsymbol{\Sigma}_{Y_{ds}}(\boldsymbol{\sigma}) \boldsymbol{\Sigma}_{Y_{ds}}^{-1} (\mathbf{y}_{ds} - \boldsymbol{\mu}_{Y_{ds}}) \right] \\ & + \frac{\Phi_{(\boldsymbol{\sigma})}(\mathbf{D}_{Y_{ds}}(\mathbf{y}_{ds} - \boldsymbol{\mu}_{Y_{ds}}); \boldsymbol{\nu}_{Y_{ds}}, \boldsymbol{\Gamma}_{Y_{ds}})}{\Phi(\mathbf{D}_{Y_{ds}}(\mathbf{y}_{ds} - \boldsymbol{\mu}_{Y_{ds}}); \boldsymbol{\nu}_{Y_{ds}}, \boldsymbol{\Gamma}_{Y_{ds}})} \end{aligned} \quad (3.44)$$

$$\frac{\partial}{\partial \boldsymbol{\lambda}} l(\boldsymbol{\theta}) = \frac{\Phi_{(\boldsymbol{\lambda})}(\mathbf{D}_{Y_{ds}}(\mathbf{y}_{ds} - \boldsymbol{\mu}_{Y_{ds}}); \boldsymbol{\nu}_{Y_{ds}}, \boldsymbol{\Gamma}_{Y_{ds}})}{\Phi(\mathbf{D}_{Y_{ds}}(\mathbf{y}_{ds} - \boldsymbol{\mu}_{Y_{ds}}); \boldsymbol{\nu}_{Y_{ds}}, \boldsymbol{\Gamma}_{Y_{ds}})} \quad (3.45)$$

where $\boldsymbol{\mu}_{Y_{ds}(\boldsymbol{\beta})}$, $\boldsymbol{\mu}_{Y_{ds}(\boldsymbol{\mu})}$, $\boldsymbol{\Sigma}_{Y_{ds}}(\boldsymbol{\sigma})$, $\Phi_{(\boldsymbol{\beta})}$, $\Phi_{(\boldsymbol{\mu})}$, $\Phi_{(\boldsymbol{\sigma})}$, and $\Phi_{(\boldsymbol{\lambda})}$ are the derivatives of $\boldsymbol{\mu}_{Y_{ds}}$, $\boldsymbol{\Sigma}_{Y_{ds}}$ and Φ with respect to $\boldsymbol{\beta}$, $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, and $\boldsymbol{\lambda}$. Expressions for these derivatives are provided in Appendix 3.5.2.

3.4.2 Simulation Results for the ML method

In this section, a limited simulation is conducted to assess the ML method. Consider the full model (3.16), which assumes that both the random effects u_d and the errors e_{dj} follow SN distributions. We use a setup similar to Molina and Rao (2010) with three fixed effects. The fixed effects β_0 (intercept), β_1 , and β_2 are chosen to be equal to 3, 0.03, and -0.04 respectively. The dispersion parameters σ_u and σ_e are set equal to 0.15 and 0.50 respectively. And the shape parameters λ_u and λ_e are set to be equal to 1 and 3 respectively. There are $m = 80$ small areas and in each area a population

of $N_d = 250$ units is generated from the model (3.16). In each area d , a sample of size $n_d = 50$ is selected with simple random sampling. In total, 1,000 Monte Carlo (MC) populations were simulated under this setup.

We use the log-likelihood (3.41) and its gradient developed in appendix 3.5.2 to do an unconstrained optimization. We use a quasi-Newton method called BroydenFletcherGoldfarbShanno (BFGS) algorithm from the R package optim. For the simulation, the initial values of the parameters β , σ_u and σ_e are their estimates assuming normal distributions for the errors terms u_d and e_{dj} in the nested error model. For λ_u and λ_e , we choose 0.5 as initial value. We multiplied λ_u by 50 to increase its contribution to the likelihood function.

Consider $\boldsymbol{\theta} = (\beta_0, \beta_1, \beta_2, \sigma_u, \sigma_e, \lambda_u, \lambda_e)$ the vector of the parameters to be estimated and note by θ_k the k^{th} element (parameter) of $\boldsymbol{\theta}$, where $k = 1, \dots, 7$. Let $\hat{\boldsymbol{\theta}}_k[i]$ denotes the estimate from the i^{th} generated population of θ_k ($\hat{\boldsymbol{\theta}}_k$ is a vector of 1,000 values). The standard error (SE) is defined as

$$SE = \sqrt{\frac{1}{1000} \sum_{k=1}^{1000} (\hat{\theta}_k[i] - \hat{\bar{\theta}}_k)^2}, \quad (3.46)$$

where $\hat{\bar{\theta}}_k$ is the average over the 1,000 estimates. The Bias is defined as

$$Bias = \frac{1}{1000} \sum_{k=1}^{1000} (\hat{\theta}_k[i] - \theta_k), \quad (3.47)$$

the absolute relative bias (ARB) is defined as

$$ARB = \left| \frac{Bias}{\theta_k} \right|, \quad (3.48)$$

the absolute bias ratio (ABR) is defined as

$$ABR = \frac{|Bias|}{SE}. \quad (3.49)$$

Table 3.1 shows results on estimating the parameters of the SN model defined in the beginning of this section. The standard error (SE) is the highest for λ_u . If we look at SE relative the the average estimates then the largest absolute relative values are 111% for λ_u followed by 32.1% and 29.3% for respectively β_1 and β_2 . This is not surprising, area level parameters (σ_u and λ_u) are more challenging to estimate due to the nature of the linear mixed model i.e. the number of areas is smaller than the number of units. As the number of areas increases, the estimation of area level related parameters improves. The estimates of the bias in Table 3.1 are generally small relative to the average estimates. This is shown by ARB which is less than 2% for all the parameters but λ_u with an ARB of 13.06%. The absolute bias ratio (ABR), which shows the bias relative to the standard error, is somewhat variable. However, only λ_u and σ_u have an ABR larger than 10% with respectively 10.42% and 27.67%. It is difficult to conclude on the quality of the bias using ABR because for a same given bias, estimates with smaller standard errors give higher bias ratios.

Table 3.1: Parameter estimation for the nested error model with SN errors using the ML method. We simulated 1,000 samples with $m = 80$ small areas and $n_d = 50$ units per area.

Parameters	Average	SE	Bias	ARB(%)	ABR(%)
$\beta_0(3)$	3.0153	0.0553	0.0153	0.51	27.67
$\beta_1(0.03)$	0.0299	0.0096	-0.0001	0.37	1.04
$\beta_2(-0.04)$	-0.0403	0.0118	0.0003	0.68	2.54
$\sigma_u(0.15)$	0.1477	0.0276	-0.0023	1.53	8.33
$\sigma_e(0.50)$	0.4999	0.0127	-0.0007	0.14	5.28
$\lambda_u(1)$	1.1306	1.2533	0.1306	13.06	10.42
$\lambda_e(3)$	3.0036	0.2386	0.0036	0.12	1.51

Estimation of the fixed effects is stable (see Figure 3.1) concentrated around the

true parameters with very few outliers.

Figure 3.1: MC densities of the fixed effects' parameters for the nested error model with SN errors using the ML method. We simulated 1,000 samples with $d = 80$ small areas and $n_d = 50$ units per area.

Figure 3.2 shows very long tails indicating few outliers estimates for the variance and shape parameters. As discussed earlier, estimation of the parameters associated with the area effects (σ_u and λ_u) is less stable.

Figure 3.2: MC densities of the random components' parameters for the nested error model with SN errors using the ML method. We simulated 1,000 samples with $d = 80$ small areas and $n_d = 50$ units per area.

Figure 3.3 shows fairly symmetric boxplots with few outliers for the fixed effects β . The ranges and the standard errors of the values of β_1 and β_2 are much smaller than the ones for β_0 however the CVs of the former are larger. The differential of the values of β_1 and β_2 with respect to β_0 is an additional source of instability for the estimation. The boxplots associated with σ_e and λ_e are very compact around the true parameters with very few outliers indicating reliable estimates. The estimation of σ_u is not very symmetric with some outliers. However the estimation of λ_u is less than desirable with a non-symmetric boxplot and many large outliers, the maximum estimated value been around 14. A constrained estimation can be applied to the parameter λ_u to reduce the variability and the outliers.

Estimation of λ_u is much more unstable compared to the other parameters. As n_d increases the skewness impact due to the area effect u_d vanishes. To see this, note that:

$$\mathbf{D}_{Y_{ds}} \rightarrow \mathbf{D}_\infty = \begin{bmatrix} \frac{\lambda_e}{\sigma_e} \mathbf{I}_\infty & \mathbf{0} \\ \mathbf{0} & 0 \end{bmatrix} \quad \text{as } n_d \rightarrow \infty \quad (3.50)$$

and

$$\mathbf{\Gamma}_{Y_{ds}} \rightarrow \mathbf{\Gamma}_\infty = \begin{bmatrix} \mathbf{I}_\infty & \mathbf{0} \\ \mathbf{0} & 0 \end{bmatrix} \quad \text{as } n_d \rightarrow \infty \quad (3.51)$$

where $\mathbf{D}_{Y_{ds}}$, $\mathbf{\Gamma}_{Y_{ds}}$ defined as in (3.20), and $\mathbf{I}_\infty$ is the identity matrix with infinite dimensions. Another way to see this vanishing skewness effect of $\mathbf{u}_d$ is to notice that:

$$\mathbf{D}_{\mathbf{u}_d} = \frac{\lambda_u}{n_d \sigma_u} \mathbf{1}^T \rightarrow 0 * \mathbf{1}_\infty^T = \quad \text{as } n_d \rightarrow \infty \quad (3.52)$$

meaning that

$$\mathbf{u}_d \rightarrow N_\infty(\mu_u \mathbf{1}_\infty, \sigma_u^2 \mathbf{1}_\infty \mathbf{1}_\infty^T) \quad \text{as } n_d \rightarrow \infty \quad (3.53)$$

In the context of survey sampling, let's consider a finite population with a large number of units N_d and a random vector $Y_{ds} \sim CSN_{n_d, n_d+1}(\boldsymbol{\mu}_{Y_{ds}}, \boldsymbol{\Sigma}_{Y_{ds}}, \mathbf{D}_{Y_{ds}}, \boldsymbol{\nu}_{Y_{ds}}, \mathbf{\Gamma}_{Y_{ds}})$

Figure 3.3: MC boxplots of the parameters estimates for the nested error model with SN errors using the unconstrained ML method. We simulated 1,000 samples with $d = 80$ small areas and $n_d = 50$ units per area.

with $\mu_{Y_{ds}}$, $\Sigma_{Y_{ds}}$, $D_{Y_{ds}}$, $\nu_{Y_{ds}}$, and $\Gamma_{Y_{ds}}$ defined as in (3.19). As n_d increases towards N_d , the quantity $\frac{\sigma_u^2}{\sigma_e^2 + n_d \sigma_u^2}$ decreases. In other words, the shape parameters associated with the random effect u_d gets smaller in absolute value as n_d increases. A common method to deal with this issue is to multiple the weak parameter by a constant in

order to make it contribute more to the likelihood function. Fortunately, as shown in the simulation study from the next chapter, estimation of λ_u does not matter very much since miss-specification on the distribution of u_d has very negligible impact on the accuracy of the small area estimators.

Table 3.2 shows results of the estimation of the parameters from model (3.20) which assumes $u_d \sim N(0, \sigma_u^2)$ i.e $\lambda_u = 0$. As expected, higher values of m and n_d give the best results. Only the scenario $m = 30$ and $n_d = 15$ shows ARB higher than 5% in absolute value for β_2 . The CVs for β_1 and β_2 are large for the four combination of m and n_d with the best values being around 30% for $m = 80$ and $n_d = 50$. The parameter σ_u show higher values of ABR in part because of its small values of SE.

Table 3.2: Parameters estimation for the nested error model with SN errors using the unconstrained ML method. We simulated 1,000 samples with $\lambda_u = 0$ i.e $u_d \sim N(0, \sigma_u^2)$.

m	n_d	Parameters	Average (SE)	Bias	ARB(%)	ABR(%)
30	15	$\beta_0(3)$	3.0001 (0.0438)	-0.0001	0.00	0.23
		$\beta_1(0.03)$	0.0293 (0.0301)	0.0007	2.34	2.33
		$\beta_2(-0.04)$	-0.0374 (0.0364)	-0.0026	6.51	4.06
		$\sigma_u(0.15)$	0.1433 (0.0256)	0.0067	4.44	26.17
		$\sigma_e(0.50)$	0.4984 (0.0275)	0.0016	0.33	5.81
		$\lambda_e(3)$	3.1169 (0.6718)	-0.1169	3.90	17.40
	50	$\beta_0(3)$	3.0012 (0.0333)	-0.0012	0.04	3.46
		$\beta_1(0.03)$	0.0295 (0.0161)	0.0005	1.70	3.16
		$\beta_2(-0.04)$	-0.0395 (0.0187)	-0.0005	1.17	2.50
		$\sigma_u(0.15)$	0.1453 (0.0211)	-0.0047	3.11	22.13
		$\sigma_e(0.50)$	0.4991 (0.0143)	0.0009	0.18	6.22
		$\lambda_e(3)$	3.0193 (0.3085)	-0.0193	0.64	6.24
80	15	$\beta_0(3)$	3.0001 (0.0258)	-0.0001	0.03	0.39
		$\beta_1(0.03)$	0.0288 (0.0176)	0.0012	4.03	6.81
		$\beta_2(-0.04)$	-0.0396 (0.0221)	-0.0004	1.02	1.81
		$\sigma_u(0.15)$	0.1482 (0.0151)	0.0018	1.23	11.92
		$\sigma_e(0.50)$	0.5002 (0.0168)	-0.0002	0.03	1.19
		$\lambda_e(3)$	3.0540 (0.3824)	-0.0540	1.80	14.12
	50	$\beta_0(3)$	3.0000 (0.0190)	0.0003	0.01	1.80
		$\beta_1(0.03)$	0.0300 (0.0097)	-0.0000	0.03	0.11
		$\beta_2(-0.04)$	-0.0398 (0.0118)	-0.0002	0.62	2.09
		$\sigma_u(0.15)$	0.1488 (0.0127)	0.0012	0.78	9.27
		$\sigma_e(0.50)$	0.4998 (0.0081)	0.0002	0.05	2.90
		$\lambda_e(3)$	3.0066 (0.1689)	0.0066	0.22	3.91

3.4.3 Data Cloning (DC)

Given the challenges associated with the ML estimation method, Lele et al. (2007) proposed an alternative approach for estimating the parameters called data cloning (DC). DC is particularly attractive because it requires neither optimization nor differentiation of a function and there is no need in our situation to evaluate multi-dimensional integrals. DC uses Bayesian formulation and computational techniques, therefore it is necessary to construct a full Bayesian setup of the nested-error model (1.5) with proper prior distributions for the unknown parameters. However, the inferences are based on the frequentist paradigm and do not depend on the choice of priors. The invariance of the inferences with respect to the priors addresses positively a major critique of the Bayesian methods.

Consider the observed data $\mathbf{y}$, the unobserved data $\mathbf{u}$, the likelihood function $l(\boldsymbol{\theta}|\mathbf{y})$, and the prior distribution $\pi(\boldsymbol{\theta})$. The nested error model (1.5) can be written as follows:

$$\begin{aligned}\mathbf{Y}_{dj}|\{U_d = u_d\} &\sim f_e(\mathbf{y}_{dj}|u_d, \boldsymbol{\theta}_e) \\ U_d &\sim f_u(u_d|\boldsymbol{\theta}_u)\end{aligned}$$

where $\boldsymbol{\theta}_e = (\mu_e, \sigma_e^2, \lambda_e^2)$ and $\boldsymbol{\theta}_u = (\mu_u, \sigma_u^2, \lambda_u^2)$. Then the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{y})$ is:

$$\begin{aligned}\pi(\boldsymbol{\theta}|\mathbf{y}) &= \frac{\{\int f_e(\mathbf{y}|\mathbf{u}, \boldsymbol{\theta}_e) f_u(\mathbf{u}|\boldsymbol{\theta}_u) d\mathbf{u}\} \pi(\boldsymbol{\theta})}{C(\mathbf{y})} \\ &= \frac{l(\boldsymbol{\theta}|\mathbf{y}) \pi(\boldsymbol{\theta})}{C(\mathbf{y})}\end{aligned}$$

where $C(\mathbf{y}) = \int l(\boldsymbol{\theta}|\mathbf{y}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}$ is the normalizing constant. Note that the integral in the denominator can be very challenging to evaluate. However, Markov Chain Monte

Carlo (MCMC) algorithms will generate random values from the posterior distribution without ever evaluating that integral since it is only a constant.

The DC technique is based on the idea of repeating the statistical process (control experiment, sampling, etc) to obtain the same information $\mathbf{y}$. Consider K sets of information ($\mathbf{y}_K$) obtained from K independent sampling process measuring the same information $\mathbf{y}$. Then the resulting maximum likelihood is $l(\boldsymbol{\theta}|\mathbf{y}_K) = [l(\boldsymbol{\theta}|\mathbf{y})]^K$. Note that, because the same information $\mathbf{y}$ is measure K independent times, the maximum likelihood estimates associated with $[l(\boldsymbol{\theta}|\mathbf{y})]^K$ are the same as for $l(\boldsymbol{\theta}|\mathbf{y})$. Also the Fisher information matrix based on $[l(\boldsymbol{\theta}|\mathbf{y})]^K$ is $K\mathbf{I}(\hat{\boldsymbol{\theta}})$ where $\mathbf{I}(\hat{\boldsymbol{\theta}})$ is the Fisher information matrix relative to $l(\boldsymbol{\theta}|\mathbf{y})$. This means that the maximum likelihood estimation based on the K independent samples leads to the same estimates as the ones based on a the single sample but with an improved accuracy (variance) by a factor of k . The posterior distribution of $\boldsymbol{\theta}$ based on the augmented data ($\mathbf{y}_K$) is:

$$\begin{aligned}\pi_K(\boldsymbol{\theta}|\mathbf{y}) &= \frac{\left(\int f_e(\mathbf{y}|\mathbf{u}, \boldsymbol{\theta}_e) f_u(\mathbf{u}|\boldsymbol{\theta}_u) d\mathbf{u}\right)^K \pi(\boldsymbol{\theta})}{C(\mathbf{y})} \\ &= \frac{(l(\boldsymbol{\theta}|\mathbf{y}))^K \pi(\boldsymbol{\theta})}{C(K, \mathbf{y})}\end{aligned}$$

where $C(K, \mathbf{y}) = \int (l(\boldsymbol{\theta}|\mathbf{y})\pi(\boldsymbol{\theta}))^K d\boldsymbol{\theta}$ is the normalizing constant. Under regularity conditions and if K is large then $\pi_K(\boldsymbol{\theta}|\mathbf{y})$ is approximately normal with mean $\hat{\boldsymbol{\theta}}$ and variance-covariance matrix equal to $\frac{1}{K}\mathbf{I}(\hat{\boldsymbol{\theta}})^{-1}$.

Unfortunately there are no K sets from independent sampling. However, $\pi_K(\boldsymbol{\theta}|\mathbf{y})$ can be looked at as any distribution defined over the parameter space Θ and $\pi_K(\boldsymbol{\theta}|\mathbf{y})$ is only a function of f_e , f_u , and π . Therefore the single sample $\mathbf{y}$ is sufficient for providing all the input information to $\pi_K(\boldsymbol{\theta}|\mathbf{y})$. To obtain K samples, we replicate the single sample K times to get the K -cloned dataset $\mathbf{y}_K = (\mathbf{y}, \dots, \mathbf{y})$. Even if, these

K clones are not independent, Lele et al. (2010) were able to show that if K is large enough then $\pi_K(\boldsymbol{\theta}|\mathbf{y})$ converges to a multivariate normal distribution with mean equal to $\hat{\boldsymbol{\theta}}$ and variance-covariance matrix equal to $\frac{1}{K}\mathbf{I}(\hat{\boldsymbol{\theta}})^{-1}$. In practice, to estimate the parameters from the DC method, we will pretend that the K samples were obtained from independent processes and use the MCMC method to generate random values from the posterior distribution $\pi_K(\boldsymbol{\theta}|\mathbf{y})$. If K is large, the MLE of $\boldsymbol{\theta}$ is the mean of the random values and the variance-covariance matrix associated with the MLE of $\boldsymbol{\theta}$ is equal to the inverse of the Fisher information matrix of the single sample divided by K .

As mentioned earlier, to use the DC method, we need to specify a full Bayesian method. To find the full Bayesian setup of the nested error model, we consider the conditional representation of the SN distribution. From Proposition 2.1, we know that if $Y \sim SN(\mu, \sigma^2, \lambda)$ then we have $Y = \mu + \sigma (\delta|X_0| + (1 - \delta^2)^{1/2}X_1)$ where X_0 and X_1 are independent $N(0, 1)$ and $\delta = \lambda/\sqrt{1 + \lambda^2}$. The full Bayesian model can be specified as follows:

$$Y_{dj}|u_d, \boldsymbol{\beta}, \sigma_e^2, \delta_e, t_e \sim N(\mu_{dj} + \sigma_e \delta_e t_e, (1 - \delta_e^2) \sigma_e^2) \quad (3.54)$$

$$t_e \sim N(0, 1)1(t_e > 0) \quad (3.55)$$

$$u_d|\sigma_d^2, \delta_u, t_u \sim N(\sigma_u \delta_u t_u, (1 - \delta_u^2) \sigma_u^2) \quad (3.56)$$

$$t_u \sim N(0, 1)1(t_u > 0) \quad (3.57)$$

$$\beta_i \sim N(\mu_{\beta_i}, \sigma_{\beta_i}^2), \quad i = 0, 1, \dots, p \quad (3.58)$$

$$\sigma_e^2 \sim IG\left(\frac{\tau_{1\sigma_e}}{2}, \frac{\tau_{2\sigma_e}}{2}\right) \quad (3.59)$$

$$\sigma_u^2 \sim IG\left(\frac{\tau_{1\sigma_u}}{2}, \frac{\tau_{2\sigma_u}}{2}\right) \quad (3.60)$$

$$\delta_e \sim Unif(\tau_{1\delta_e}, \tau_{2\delta_e}) \quad (3.61)$$

$$\delta_u \sim Unif(\tau_{1\delta_u}, \tau_{2\delta_u}) \quad (3.62)$$

where p is the number of fixed effects and the parameters $\mu_{\beta_i}, \sigma_{\beta_i}^2, \tau_{1\sigma_e}, \tau_{2\sigma_e}, \tau_{1\sigma_u}, \tau_{2\sigma_u}, \tau_{1\delta_e}, \tau_{2\delta_e}, \tau_{1\delta_u}$, and $\tau_{2\delta_u}$ can be chosen so that the priors are nearly non-informative. However DC estimates are invariant to the choice of the priors so better priors will only help convergence.

3.5 Appendix

This section provides details on two important features of the ML computation. The first sub-section lays out the technique for reducing the dimension of the integral involved in the maximum likelihood calculation by taking advantage of the special form of the variance-covariance matrix. The second sub-section provides the results of the differentiation of the likelihood function.

3.5.1 Computing Multivariate Normal CDF - Special Case

In this sub-section, we use the CDF based on the correlation matrix R and the standard values z . Note that there is no loss of generality because using result $\Sigma = \mathbf{V}^{1/2}\mathbf{R}\mathbf{V}^{1/2}$ where $\mathbf{V}$ is the diagonal matrix composed of the variances $\sigma_1^2, \dots, \sigma_k^2$, we have:

$$-\frac{1}{2}(\mathbf{y} - \mu)^T \Sigma^{-1}(\mathbf{y} - \mu) = -\frac{1}{2}\mathbf{Z}^T \mathbf{R}^{-1} \mathbf{Z} \quad (3.63)$$

and

$$|\Sigma| = |\mathbf{V}^{1/2}| |\mathbf{R}| |\mathbf{V}^{1/2}| = |\mathbf{V}| |\mathbf{R}| \quad (3.64)$$

where $\mathbf{z} = \mathbf{V}^{-1/2}(\mathbf{y} - \mu)$. From the two results (3.63), (3.64), and $d\mathbf{y} = \text{diag}_{1 \leq i \leq k}(\sigma_i) d\mathbf{z}$, we have:

$$\begin{aligned} & \int_{\Omega_{\mathbf{Y}}} \frac{1}{(2\pi)^{k/2} |\boldsymbol{\Sigma}|^{1/2}} \exp \left(-\frac{1}{2} (\mathbf{y} - \mu)^T \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \mu) \right) d\mathbf{y} \\ &= \int_{\Omega_{\mathbf{Z}}} \frac{1}{(2\pi)^{k/2} |\mathbf{R}|^{1/2} \sigma_1 \dots \sigma_n} \exp \left(-\frac{1}{2} \mathbf{z}^T \mathbf{R}^{-1} \mathbf{z} \right) \sigma_1 \dots \sigma_k dz_1 \dots dz_k \\ &= \int_{\Omega_{\mathbf{Z}}} \frac{1}{(2\pi)^{k/2} |\mathbf{R}|^{1/2}} \exp \left(-\frac{1}{2} \mathbf{z}^T \mathbf{R}^{-1} \mathbf{z} \right) d\mathbf{z}. \end{aligned}$$

Consider the standardized multivariate normal with pdf of the form:

$$\phi_{\mathbf{Z}}(z) = \frac{1}{(2\pi)^{k/2} |\mathbf{R}|^{1/2}} \exp \left(-\frac{1}{2} \mathbf{z}^T \mathbf{R}^{-1} \mathbf{z} \right), \quad (3.65)$$

where $\mathbf{Z}^T = (z_1, \dots, z_k)$, and $\mathbf{R}$ is the correlation matrix of $\mathbf{Z}$. The corresponding cumulative distribution function can be written as:

$$\Phi_k(\mathbf{z}, \mathbf{R}) = \frac{1}{(2\pi)^{k/2} |\mathbf{R}|^{1/2}} \int_{-\infty}^{z_1} \dots \int_{-\infty}^{z_k} \exp \left(-\frac{1}{2} \mathbf{z}^T \mathbf{R}^{-1} \mathbf{z} \right) dz_1 \dots dz_k. \quad (3.66)$$

For a general form of $\mathbf{R}$, there is no simple way (closed form) to evaluate the expression (3.65). Calculations can be feasible for small values of k (see some results from Kotz et al. (2000)), but as k increases, they become rapidly impracticable. For large values of k , numerical methods for evaluating integrals will be necessary. However, a special form of the correlation matrix $\mathbf{R}$ can lead to substantial simplification in the calculations. The matrix $\boldsymbol{\Gamma}_{\mathbf{Y}_d}$ from (3.18) has a special form which can be used to simplify and improve the numerical approximation.

When the correlation is of the form $\alpha \mathbf{I} + c^2 \mathbf{1}\mathbf{1}^T$ and if $\mathbb{E}(\mathbf{Z}) = \mathbf{0}$, we have:

$$\begin{aligned} \Phi_k(\mathbf{z}, \mathbf{R}) = & \frac{c}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}c^2 t^2\right) \int_{-\infty}^{z_1} \cdots \int_{-\infty}^{z_k} \frac{1}{(2\pi)^{k/2} |\alpha|^k} \times \\ & \exp\left(-\frac{1}{2} \sum_{i=1}^k \left(\frac{z_i - t}{\alpha}\right)^2\right) dz_1 \dots dz_k dt. \end{aligned} \quad (3.67)$$

This multiple integral is of $(k+1)^{th}$ order which is higher than the original order k but it is usually of simpler form. If the correlations can be expressed in the form $\rho_{ij} = c_i c_j$, $|c_i| \leq 1$ for all i and j , then we have $Z_i = c_i U_0 + \sqrt{1 - c_i^2} U_i$, $i = 1, \dots, k$ where $U_i \stackrel{ind}{\sim} N(0, 1)$. In this situation, the inequality $Z_i \leq z_i$ is equivalent to $U_i \leq (z_i - c_i U_0) / \sqrt{1 - c_i^2}$ and we get from Dunnett and Sobel (1955):

$$\Phi_k(\mathbf{z}, \mathbf{R}) = \int_{-\infty}^{\infty} \prod_{i=1}^k \Phi_1\left(\frac{z_i - c_i u_0}{\sqrt{1 - c_i^2}}\right) \phi(u_0) du_0. \quad (3.68)$$

If all the correlations are equal and positive ($\rho_{ij} = \rho \geq 0$ for all i and j), then we have:

$$\Phi_k(\mathbf{z}, \mathbf{R}) = \int_{-\infty}^{\infty} \prod_{i=1}^k \Phi_1\left(\frac{z_i - \sqrt{\rho} u_0}{\sqrt{1 - \rho}}\right) \phi(u_0) du_0. \quad (3.69)$$

If in addition $z_i = 0$ for all i then

$$\Phi_k(\mathbf{0}, \mathbf{R}) = \int_{-\infty}^{\infty} \left\{ \Phi_1\left(-\sqrt{\frac{\rho}{1 - \rho}}\right) \right\}^k \phi(u_0) du_0. \quad (3.70)$$

The integrals involved in (3.67), (3.68), and (3.69) are of order 1; therefore much simpler than the original k order multidimensional integral. All statistical software provides standard routines for computing the univariate normal pdf and cdf with a high level of accuracy. We can use result (3.68) to get $\ln(\Phi_{n_d+1}(\mathbf{D}_{ds}(\mathbf{y}_{ds} - \boldsymbol{\mu}_{ds}); \boldsymbol{\nu}_{ds}, \boldsymbol{\Gamma}_{ds}))$

using a much simpler single integral. Indeed, the nested error model (3.16) leads to:

$$\mathbf{\Gamma}_{Y_{ds}} = \begin{bmatrix} 1 + \frac{\lambda_e^2 \sigma_u^2}{\sigma_e^2 + n_d \sigma_u^2} & \frac{\lambda_e^2 \sigma_u^2}{\sigma_e^2 + n_d \sigma_u^2} & \cdots & \frac{\lambda_e^2 \sigma_u^2}{\sigma_e^2 + n_d \sigma_u^2} & -\frac{\lambda_e \lambda_u \sigma_e \sigma_u}{\sigma_e^2 + n_d \sigma_u^2} \\ \frac{\lambda_e^2 \sigma_u^2}{\sigma_e^2 + n_d \sigma_u^2} & 1 + \frac{\lambda_e^2 \sigma_u^2}{\sigma_e^2 + n_d \sigma_u^2} & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & \frac{\lambda_e^2 \sigma_u^2}{\sigma_e^2 + n_d \sigma_u^2} & \vdots \\ \frac{\lambda_e^2 \sigma_u^2}{\sigma_e^2 + n_d \sigma_u^2} & \cdots & \frac{\lambda_e^2 \sigma_u^2}{\sigma_e^2 + n_d \sigma_u^2} & 1 + \frac{\lambda_e^2 \sigma_u^2}{\sigma_e^2 + n_d \sigma_u^2} & -\frac{\lambda_e \lambda_u \sigma_e \sigma_u}{\sigma_e^2 + n_d \sigma_u^2} \\ -\frac{\lambda_e \lambda_u \sigma_e \sigma_u}{\sigma_e^2 + n_d \sigma_u^2} & \cdots & \cdots & -\frac{\lambda_e \lambda_u \sigma_e \sigma_u}{\sigma_e^2 + n_d \sigma_u^2} & 1 + \frac{\lambda_u^2 \sigma_e^2}{\sigma_e^2 + n_d \sigma_u^2} \end{bmatrix}. \quad (3.71)$$

The correlation matrix $\mathbf{R}_{Y_{ds}}$ associated with $\mathbf{\Gamma}_{Y_{ds}}$ is:

$$[\mathbf{R}_{Y_{ds}}]_{ij} = \begin{cases} 1 & \text{for all elements } i = j \\ \frac{\lambda_e^2 \sigma_u^2}{\sigma_e^2 + n_d \sigma_u^2 + \lambda_e^2 \sigma_u^2} & \text{if } i \neq j, \text{ both } \leq n_d \\ -\frac{\lambda_e \sigma_u}{\sqrt{\sigma_e^2 + n_d \sigma_u^2 + \lambda_e^2 \sigma_u^2}} \frac{\lambda_u \sigma_e}{\sqrt{\sigma_e^2 + n_d \sigma_u^2 + \lambda_u^2 \sigma_e^2}} & i = n_d + 1 \text{ or } j = n_d + 1. \end{cases} \quad (3.72)$$

When e_{dj} is SN ($\lambda_e \neq 0$) and u_d is normal ($\lambda_u = 0$), the equation (3.69) is applicable and provides even more simplification for computing the logarithm of the cumulative probabilities. In this situation, the $(n_d + 1)^{th}$ term is uncorrelated to the other terms and the correlation between the remaining terms are all equal to $\frac{\lambda_e^2 \sigma_u^2}{\sigma_e^2 + n_d \sigma_u^2 + \lambda_e^2 \sigma_u^2}$.

3.5.2 Differentiation of the Likelihood Function

In this section, the references to the sample and small areas are removed from the formulas to simplify the expressions. The partial derivatives of μ_{Y_d} for the unadjusted errors terms are:

$$\boldsymbol{\mu}_{(\beta)} = \mathbf{X}^T, \quad \boldsymbol{\mu}_{(\mu_u)} = \mathbf{1}_{n_d}, \quad \text{and} \quad \boldsymbol{\mu}_{(\mu_e)} = I_{n_d}.$$

The nested error model assumes that the errors terms u_d and e_{dj} have means equal 0.

For the adjusted model where $\mathbb{E}(u) = \mathbb{E}(e) = 0$, we have:

$$\begin{aligned}\mu_{(\sigma_u^2)} &= -\frac{1}{2}\left(\frac{\delta_u}{\sigma_u}\right)\sqrt{\frac{2}{\pi}}, & \mu_{(\sigma_e^2)} &= -\frac{1}{2}\left(\frac{\delta_e}{\sigma_e}\right)\sqrt{\frac{2}{\pi}}, \\ \mu_{(\delta_u)} &= -\frac{\sigma_u}{(1+\lambda_u^2)^{3/2}}\sqrt{\frac{2}{\pi}}, & \text{and} & \quad \mu_{(\sigma_e^2)} = -\frac{\sigma_e}{(1+\lambda_e^2)^{3/2}}\sqrt{\frac{2}{\pi}}.\end{aligned}$$

Also, the partial derivatives of the covariance matrix are:

$$\Sigma_{(\sigma_u^2)} = \mathbf{1}_{n_d}\mathbf{1}_{n_d}^T, \quad \Sigma_{(\sigma_e^2)} = I_{n_d}, \quad \Sigma_{(\sigma_u^2)}^{-1} = -\Sigma_{(\sigma_e^2)} \Sigma, \quad \text{and} \quad \Sigma_{(\sigma_e^2)}^{-1} = -\Sigma_{(\sigma_u^2)} \Sigma.$$

Consider $h_i(\theta) = \frac{z_i - c_i u_0}{\sqrt{1 - c_i^2}}$ then

$$\Phi_{n_d+1}(\mathbf{h}) = \int_{-\infty}^{\infty} \prod_{i=1}^{n_d+1} \Phi_1(h_i(\theta)) \phi(u_0) du_0. \quad (3.73)$$

The derivative of (3.73) with respect to θ is:

$$\Phi_{(\theta)}(\mathbf{h}) = \int_{-\infty}^{\infty} \sum_{j=1}^{n_d+1} \left(\left(\frac{\partial}{\partial \theta} h_j(\theta) \right) \phi(h_j(\theta)) \prod_{i=1, i \neq j}^{n_d+1} \Phi(h_i(\theta)) \right) \phi(u_0) du_0, \quad (3.74)$$

where

$$\frac{\partial}{\partial \theta} h_j(\theta) = \frac{1}{\sqrt{1 - c_j^2}} \left(\frac{1}{1 - c_j^2} (z_j c_j - u_0) c_{j(\theta)} + z_{j(\theta)} \right) \quad (3.75)$$

Using results from 3.72, we get the following expressions for c_j and z_j

$$c_j = \begin{cases} \frac{\lambda_e \sigma_u}{\sqrt{\sigma_e^2 + n_d \sigma_u^2 + \lambda_e^2 \sigma_u^2}} & \text{if } j \leq n_d \\ \frac{-\lambda_u \sigma_e}{\sqrt{\sigma_e^2 + n_d \sigma_u^2 + \lambda_u^2 \sigma_e^2}} & \text{if } j = n_d + 1 \end{cases} \quad (3.76)$$

and

$$z_j = \begin{cases} [\mathbf{D}(\mathbf{y} - \boldsymbol{\mu})]_j \left(\frac{\sigma_e^2 + n_d \sigma_u^2}{\sigma_e^2 + n_d \sigma_u^2 + \lambda_e^2 \sigma_u^2} \right)^{1/2} & \text{if } j \leq n_d \\ [\mathbf{D}(\mathbf{y} - \boldsymbol{\mu})]_j \left(\frac{\sigma_e^2 + n_d \sigma_u^2}{\sigma_e^2 + n_d \sigma_u^2 + \lambda_u^2 \sigma_e^2} \right)^{1/2} & \text{if } j = n_d + 1 \end{cases} \quad (3.77)$$

After some manipulations, we obtain the non-zero partial derivatives of c_j as follows:

$$c_{j(\sigma_e^2)} = \begin{cases} -\frac{1}{2} \frac{\lambda_e}{\sigma_u} (\sigma_u^2) (\sigma_e^2 + n_d \sigma_u^2 + \lambda_e^2 \sigma_u^2)^{-3/2} & \text{if } j \leq n_d \\ -\frac{1}{2} \frac{\lambda_u}{\sigma_e} (n_d \sigma_u^2) (\sigma_e^2 + n_d \sigma_u^2 + \lambda_u^2 \sigma_e^2)^{-3/2} & \text{if } j = n_d + 1 \end{cases} \quad (3.78)$$

$$c_{j(\sigma_u^2)} = \begin{cases} \frac{1}{2} \frac{\lambda_e}{\sigma_u} (\sigma_e^2) (\sigma_e^2 + n_d \sigma_u^2 + \lambda_e^2 \sigma_u^2)^{-3/2} & \text{if } j \leq n_d \\ \frac{1}{2} \frac{\lambda_u}{\sigma_e} (n_d \sigma_e^2) (\sigma_e^2 + n_d \sigma_u^2 + \lambda_u^2 \sigma_e^2)^{-3/2} & \text{if } j = n_d + 1 \end{cases} \quad (3.79)$$

$$c_{j(\lambda_e^2)} = \begin{cases} \sigma_u (\sigma_e^2 + n_d \sigma_u^2) (\sigma_e^2 + n_d \sigma_u^2 + \lambda_e^2 \sigma_u^2)^{-3/2} & \text{if } j \leq n_d \\ 0 & \text{if } j = n_d + 1 \end{cases} \quad (3.80)$$

$$c_{j(\lambda_u^2)} = \begin{cases} 0 & \text{if } j \leq n_d \\ -\sigma_e (\sigma_e^2 + n_d \sigma_u^2) (\sigma_e^2 + n_d \sigma_u^2 + \lambda_u^2 \sigma_e^2)^{-3/2} & \text{if } j = n_d + 1. \end{cases} \quad (3.81)$$

Let's consider

$$\gamma_j = \begin{cases} (\sigma_e^2 + n_d \sigma_u^2)^{1/2} (\sigma_e^2 + n_d \sigma_u^2 + \lambda_e^2 \sigma_u^2)^{-1/2} & \text{if } j \leq n_d \\ (\sigma_e^2 + n_d \sigma_u^2)^{1/2} (\sigma_e^2 + n_d \sigma_u^2 + \lambda_u^2 \sigma_e^2)^{-1/2} & \text{if } j = n_d + 1 \end{cases} \quad (3.82)$$

then, the expression 3.77 can be written as: $z_j = \gamma_j \mathbf{D}_j(\mathbf{y} - \boldsymbol{\mu})$ where $\mathbf{D}_j$ is the j^{th} row of the matrix $\mathbf{D}$. The partial derivatives of z_j can be found by using the following expression:

$$z_{j(\theta)} = \gamma_{j(\theta)} \mathbf{D}_j(\mathbf{y} - \boldsymbol{\mu}) + \gamma_j \mathbf{D}_{j(\theta)}(\mathbf{y} - \boldsymbol{\mu}) + \gamma_j \mathbf{D}_j(\mathbf{y} - \boldsymbol{\mu})_{(\theta)}. \quad (3.83)$$

The derivatives involved in (3.83) are:

$$\gamma_{j(\sigma_e^2)} = \begin{cases} \frac{1}{2} (\sigma_e^2 + n_d \sigma_u^2 + \lambda_e^2 \sigma_u^2)^{-1/2} \left(1 - \frac{\sqrt{\sigma_e^2 + n_d \sigma_u^2}}{\sigma_e^2 + n_d \sigma_u^2 + \lambda_e^2 \sigma_u^2} \right) & \text{if } j \leq n_d \\ \frac{1}{2} (\sigma_e^2 + n_d \sigma_u^2 + \lambda_u^2 \sigma_e^2)^{-1/2} \left(1 - \frac{(1 + \lambda_u^2) \sqrt{\sigma_e^2 + n_d \sigma_u^2}}{\sigma_e^2 + n_d \sigma_u^2 + \lambda_u^2 \sigma_e^2} \right) & \text{if } j = n_d + 1 \end{cases} \quad (3.84)$$

$$\gamma_{j(\sigma_u^2)} = \begin{cases} \frac{1}{2} (\sigma_e^2 + n_d \sigma_u^2 + \lambda_e^2 \sigma_u^2)^{-1/2} \left(n_d - \frac{(n_d + \lambda_e^2) \sqrt{\sigma_e^2 + n_d \sigma_u^2}}{\sigma_e^2 + n_d \sigma_u^2 + \lambda_e^2 \sigma_u^2} \right) & \text{if } j \leq n_d \\ \frac{1}{2} (\sigma_e^2 + n_d \sigma_u^2 + \lambda_u^2 \sigma_e^2)^{-1/2} \left(1 - \frac{\sqrt{\sigma_e^2 + n_d \sigma_u^2}}{\sigma_e^2 + n_d \sigma_u^2 + \lambda_u^2 \sigma_e^2} \right) & \text{if } j = n_d + 1 \end{cases} \quad (3.85)$$

$$\gamma_{j(\lambda_e)} = \begin{cases} -\lambda_e \sigma_u^2 (\sigma_e^2 + n_d \sigma_u^2) (\sigma_e^2 + n_d \sigma_u^2 + \lambda_e^2 \sigma_u^2)^{-3/2} & \text{if } j \leq n_d \\ 0 & \text{if } j = n_d + 1 \end{cases} \quad (3.86)$$

$$\gamma_{j(\lambda_u)} = \begin{cases} 0 & \text{if } j \leq n_d \\ -\lambda_u \sigma_e^2 (\sigma_e^2 + n_d \sigma_u^2) (\sigma_e^2 + n_d \sigma_u^2 + \lambda_u^2 \sigma_e^2)^{-3/2} & \text{if } j = n_d + 1 \end{cases} \quad (3.87)$$

$$\mathbf{D}_{j(\sigma_e^2)} = \begin{cases} -\frac{1}{2} \frac{\lambda_e}{\sigma_e} \frac{1}{\sigma_e^2} \left[\mathbf{I} + \sigma_u^2 (\sigma_e^2 + n_d \sigma_u^2)^{-1} \mathbf{1}_{n_d} \mathbf{1}_{n_d}^T \right]_j & \text{if } j \leq n_d \\ -\lambda_u \sigma_u (\sigma_e^2 + n_d \sigma_u^2)^{-1} \mathbf{1}_{n_d}^T & \text{if } j = n_d + 1 \end{cases} \quad (3.88)$$

$$\mathbf{D}_{j(\sigma_u^2)} = \begin{cases} -\frac{\lambda_e}{\sigma_e} (\sigma_e^2) (\sigma_e^2 + n_d \sigma_u^2)^{-2} \mathbf{1}_{n_d} \mathbf{1}_{n_d}^T & \text{if } j \leq n_d \\ \frac{1}{2} \frac{\lambda_u}{\sigma_u} (\sigma_e^2 - n_d \sigma_u^2) (\sigma_e^2 + n_d \sigma_u^2)^{-2} \mathbf{1}_{n_d}^T & \text{if } j = n_d + 1 \end{cases} \quad (3.89)$$

$$\mathbf{D}_{j(\lambda_e)} = \begin{cases} \frac{1}{\sigma_e} \left[\mathbf{I}_{n_d} - \frac{\sigma_u^2}{\sigma_e^2 + n_d \sigma_u^2} \mathbf{1}_{n_d} \mathbf{1}_{n_d}^T \right]_j & \text{if } j \leq n_d \\ 0 & \text{if } j = n_d + 1 \end{cases} \quad (3.90)$$

$$\mathbf{D}_{j(\lambda_u)} = \begin{cases} 0 & \text{if } j \leq n_d \\ \frac{\sigma_u}{\sigma_e^2 + n_d \sigma_u^2} \mathbf{1}_{n_d}^T & \text{if } j = n_d + 1. \end{cases} \quad (3.91)$$

The partial derivatives of $(\mathbf{y} - \boldsymbol{\mu})_{(\theta)}$ are already provided at the beginning of this Section 3.5.2. Note that in the expressions above, the notation $[\mathbf{A}]_j$ means the j^{th} row of the matrix $\mathbf{A}$.

3.5.3 Marginal Distribution of Y_{ds}

In the case of the multivariate normal distribution, if $\mathbf{Y}_d = (\mathbf{Y}_{dr}^T, \mathbf{Y}_{ds}^T)^T \sim N_{N_d}(\boldsymbol{\mu}_{Y_d}, \boldsymbol{\Sigma}_{Y_d})$, where $\boldsymbol{\mu}_{Y_d} = (\boldsymbol{\mu}_r^T, \boldsymbol{\mu}_s^T)^T$, and $\boldsymbol{\Sigma}_{Y_d} = \begin{bmatrix} \boldsymbol{\Sigma}_{Y_{dr}} & \boldsymbol{\Sigma}_{rs} \\ \boldsymbol{\Sigma}_{sr} & \boldsymbol{\Sigma}_{Y_{ds}} \end{bmatrix}$ then the marginal distribution of the sample vector $\mathbf{Y}_{ds}$ is $N_{n_d}(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_{ss})$. In this section, we derive the marginal distribution of $\mathbf{Y}_{ds}$ when $\mathbf{Y}_d$ follows the CSN defined in (3.18). Consider the following partition $\mathbf{D}_{Y_d} = \begin{bmatrix} \mathbf{D}_r & \mathbf{D}_s \end{bmatrix}$ then Proposition 2.7 shows that

$$\mathbf{Y}_{ds} \sim CSN_{n_d, N_d+1}(\boldsymbol{\mu}_{Y_{ds}}, \boldsymbol{\Sigma}_{Y_{ds}}, \mathbf{D}_{Y_{ds}}, \mathbf{0}, \boldsymbol{\Gamma}_{Y_{ds}}), \quad (3.92)$$

where $\mathbf{D}_{Y_{ds}} = \mathbf{D}_s + \mathbf{D}_r \boldsymbol{\Sigma}_{rs} \boldsymbol{\Sigma}_{Y_{ds}}^{-1}$, $\boldsymbol{\Gamma}_{Y_{ds}} = \boldsymbol{\Gamma}_{Y_d} + \mathbf{D}_r \boldsymbol{\Sigma}_{rr.s} \mathbf{D}_r^T$, and $\boldsymbol{\Sigma}_{rr.s} = \boldsymbol{\Sigma}_{Y_{dr}} - \boldsymbol{\Sigma}_{rs} \boldsymbol{\Sigma}_{Y_{ds}}^{-1} \boldsymbol{\Sigma}_{sr}$. To fully specify the distribution, we provide the simplified form of $\mathbf{D}_{Y_{ds}}$ and $\boldsymbol{\Gamma}_{Y_{ds}}$.

$$\mathbf{D}_{Y_{ds}} = \begin{bmatrix} \frac{\lambda_e}{\sigma_e} \left[\begin{pmatrix} \mathbf{0}_{N_{dr} \times n_d} \\ \mathbf{I}_{n_d} \end{pmatrix} - \gamma_{N_d} \mathbf{1}_{N_d} \mathbf{1}_{n_d}^T \right] \\ \frac{\lambda_u}{\sigma_u} \gamma_{N_d} \mathbf{1}_{n_d}^T \end{bmatrix} + \begin{bmatrix} \frac{\lambda_e}{\sigma_e} \left[\begin{pmatrix} \mathbf{I}_{N_{dr}} \\ \mathbf{0}_{n_d \times N_{dr}} \end{pmatrix} - \gamma_{N_d} \mathbf{1}_{N_d} \mathbf{1}_{N_{dr}}^T \right] \\ \frac{\lambda_u}{\sigma_u} \gamma_{N_d} \mathbf{1}_{N_{dr}}^T \end{bmatrix} \boldsymbol{\Sigma}_{Y_{dr}} \boldsymbol{\Sigma}_{Y_{ds}}^{-1}$$

where $\Sigma_{Y_{dr}} \Sigma_{Y_{ds}}^{-1} = \gamma_{n_d} \mathbf{1}_{N_{dr}} \mathbf{1}_{n_d}^T$. It follows after some matrix operations, we get

$$\begin{aligned}
\mathbf{D}_{Y_{ds}} &= \begin{bmatrix} \frac{\lambda_e}{\sigma_e} \left[\begin{pmatrix} \mathbf{0}_{N_{dr} \times n_d} \\ \mathbf{I}_{n_d} \end{pmatrix} - \gamma_{N_d} \mathbf{1}_{N_d} \mathbf{1}_{n_d}^T \right] \\ \frac{\lambda_u}{\sigma_u} \gamma_{N_d} \mathbf{1}_{n_d}^T \end{bmatrix} + \begin{bmatrix} \frac{\lambda_e}{\sigma_e} \begin{pmatrix} \gamma_{N_d} \mathbf{1}_{N_{dr}} \mathbf{1}_{N_{dr}}^T \\ -N_{dr} \gamma_{n_d} \gamma_{N_d} \mathbf{1}_{n_d} \mathbf{1}_{N_{dr}}^T \end{pmatrix} \\ \frac{\lambda_u}{\sigma_u} \gamma_{n_d} \gamma_{N_d} \mathbf{1}_{N_{dr}}^T \end{bmatrix} \\
&= \begin{bmatrix} \frac{\lambda_e}{\sigma_e} \begin{pmatrix} \mathbf{0}_{N_{dr} \times n_d} - \gamma_{N_d} \mathbf{1}_{N_{dr}} \mathbf{1}_{n_d}^T + \gamma_{N_d} \mathbf{1}_{N_{dr}} \mathbf{1}_{N_{dr}}^T \\ \mathbf{I}_{n_d} - \gamma_{N_d} \mathbf{1}_{n_d} \mathbf{1}_{n_d}^T - N_{dr} \gamma_{n_d} \gamma_{N_d} \mathbf{1}_{n_d} \mathbf{1}_{n_d}^T \end{pmatrix} \\ \frac{\lambda_u}{\sigma_u} (1 + N_{dr} \gamma_{n_d}) \gamma_{N_d} \mathbf{1}_{n_d}^T \end{bmatrix} \\
&= \begin{bmatrix} \frac{\lambda_e}{\sigma_e} \begin{pmatrix} \mathbf{0}_{N_{dr} \times n_d} \\ \mathbf{I}_{n_d} - \gamma_{n_d} \mathbf{1}_{n_d} \mathbf{1}_{n_d}^T \end{pmatrix} \\ \frac{\lambda_u}{\sigma_u} \gamma_{n_d} \mathbf{1}_{n_d}^T \end{bmatrix} = \begin{pmatrix} \mathbf{D}_{Y_{ds}}^r \\ \mathbf{D}_{Y_{ds}}^s \\ \mathbf{D}_{Y_{ds}}^u \end{pmatrix}
\end{aligned}$$

where $\mathbf{D}_{Y_{ds}}^r = \mathbf{0}_{N_{dr} \times n_d}$, $\mathbf{D}_{Y_{ds}}^s = \frac{\lambda_e}{\sigma_e} (\mathbf{I}_{n_d} - \gamma_{n_d} \mathbf{1}_{n_d} \mathbf{1}_{n_d}^T)$, and $\mathbf{D}_{Y_{ds}}^u = \frac{\lambda_u}{\sigma_u} \gamma_{n_d} \mathbf{1}_{n_d}^T$. Since $\mathbf{D}_{Y_{ds}}^r$ is a zero matrix, there is no skewness associated with those dimension i.e. $\mathbf{D}_{Y_{ds}}$ can be reduced to $\begin{bmatrix} \frac{\lambda_e}{\sigma_e} \begin{pmatrix} \mathbf{I}_{n_d} - \gamma_{n_d} \mathbf{1}_{n_d} \mathbf{1}_{n_d}^T \\ \frac{\lambda_u}{\sigma_u} \gamma_{n_d} \mathbf{1}_{n_d}^T \end{pmatrix} \end{bmatrix}$.

To simplify $\Gamma_{Y_{ds}}$, note first that

$\Sigma_{rr.s} = \Sigma_{Y_{dr}} - \Sigma_{rs} \Sigma_{Y_{ds}}^{-1} \Sigma_{sr} = \sigma_e^2 (\mathbf{I}_{N_{dr}} + \gamma_{n_d} \mathbf{1}_{N_{dr}} \mathbf{1}_{N_{dr}}^T)$. It follows that

$$\begin{aligned}
\mathbf{D}_r \Sigma_{rr.s} &= \begin{bmatrix} \frac{\lambda_e}{\sigma_e} \left[\begin{pmatrix} \mathbf{0}_{N_{dr} \times n_d} \\ \mathbf{I}_{n_d} \end{pmatrix} - \gamma_{N_d} \mathbf{1}_{N_d} \mathbf{1}_{n_d}^T \right] \\ \frac{\lambda_u}{\sigma_u} \gamma_{N_d} \mathbf{1}_{n_d}^T \end{bmatrix} \sigma_e^2 (\mathbf{I}_{N_{dr}} + \gamma_{n_d} \mathbf{1}_{N_{dr}} \mathbf{1}_{N_{dr}}^T) \\
&= \sigma_e^2 \begin{bmatrix} \frac{\lambda_e}{\sigma_e} \begin{pmatrix} (\mathbf{I}_{N_{dr}} - \gamma_{N_d} \mathbf{1}_{N_{dr}} \mathbf{1}_{N_{dr}}^T) (\mathbf{I}_{N_{dr}} + \gamma_{n_d} \mathbf{1}_{N_{dr}} \mathbf{1}_{N_{dr}}^T) \\ -\gamma_{N_d} \mathbf{1}_{N_d} \mathbf{1}_{n_d}^T (\mathbf{I}_{N_{dr}} + \gamma_{n_d} \mathbf{1}_{N_{dr}} \mathbf{1}_{N_{dr}}^T) \\ \frac{\lambda_u}{\sigma_u} \gamma_{N_d} \mathbf{1}_{n_d}^T (\mathbf{I}_{N_{dr}} + \gamma_{n_d} \mathbf{1}_{N_{dr}} \mathbf{1}_{N_{dr}}^T) \end{pmatrix} \end{bmatrix}
\end{aligned}$$

After some matrix manipulations and simplifications, we get

$$\mathbf{D}_r \Sigma_{rr.s} = \sigma_e^2 \begin{bmatrix} \frac{\lambda_e}{\sigma_e} \begin{pmatrix} \mathbf{I}_{N_{dr}} \\ -\gamma_{n_d} \mathbf{1}_{n_d} \mathbf{1}_{N_{dr}}^T \end{pmatrix} \\ \frac{\lambda_u}{\sigma_u} \gamma_{n_d} \mathbf{1}_{N_{dr}}^T \end{bmatrix}.$$

Then multiplying previous result by $\mathbf{D}_r^T$, we have

$$\begin{aligned} \mathbf{D}_r \boldsymbol{\Sigma}_{rr.s} \mathbf{D}_r^T &= \sigma_e^2 \begin{bmatrix} \frac{\lambda_e}{\sigma_e} \begin{pmatrix} \mathbf{I}_{N_{dr}} \\ -\gamma_{n_d} \mathbf{1}_{n_d} \mathbf{1}_{N_{dr}}^T \\ \frac{\lambda_u}{\sigma_u} \gamma_{n_d} \mathbf{1}_{N_{dr}}^T \end{pmatrix} \end{bmatrix} \begin{bmatrix} \frac{\lambda_e}{\sigma_e} \left[\begin{pmatrix} \mathbf{I}_{N_{dr}} \\ \mathbf{0}_{n_d \times N_{dr}} \end{pmatrix} - \gamma_{N_d} \mathbf{1}_{N_d} \mathbf{1}_{N_{dr}}^T \right] \\ \frac{\lambda_u}{\sigma_u} \gamma_{N_d} \mathbf{1}_{N_{dr}}^T \end{bmatrix}^T \\ &= \sigma_e^2 \begin{bmatrix} \frac{\lambda_e}{\sigma_e} \begin{pmatrix} \mathbf{I}_{N_{dr}} \\ -\gamma_{n_d} \mathbf{1}_{n_d} \mathbf{1}_{N_{dr}}^T \\ \frac{\lambda_u}{\sigma_u} \gamma_{n_d} \mathbf{1}_{N_{dr}}^T \end{pmatrix} \end{bmatrix} \begin{bmatrix} \frac{\lambda_e}{\sigma_e} \begin{pmatrix} (\mathbf{I}_{N_{dr}} - \gamma_{N_d} \mathbf{1}_{N_d} \mathbf{1}_{N_{dr}}^T) \\ -\gamma_{N_d} \mathbf{1}_{n_d} \mathbf{1}_{N_{dr}}^T \\ \frac{\lambda_u}{\sigma_u} \gamma_{N_d} \mathbf{1}_{N_{dr}}^T \end{pmatrix} \end{bmatrix}^T. \end{aligned}$$

Block-matrix multiplication leads to

$$\mathbf{D}_r \boldsymbol{\Sigma}_{rr.s} \mathbf{D}_r^T = \sigma_e^2 \begin{bmatrix} \mathbf{G}_{11} & \mathbf{G}_{12} & \mathbf{G}_{13} \\ \mathbf{G}_{21} & \mathbf{G}_{22} & \mathbf{G}_{23} \\ \mathbf{G}_{31} & \mathbf{G}_{32} & \mathbf{G}_{33} \end{bmatrix}$$

where $\mathbf{G}_{11} = (\frac{\lambda_e}{\sigma_e})^2 (\mathbf{I}_{N_{dr}} - \gamma_{N_d} \mathbf{1}_{N_d} \mathbf{1}_{N_{dr}}^T)$, $\mathbf{G}_{22} = (\frac{\lambda_e}{\sigma_e})^2 N_{dr} \gamma_{n_d} \gamma_{N_d} \mathbf{1}_{n_d} \mathbf{1}_{n_d}^T$, $\mathbf{G}_{33} = (\frac{\lambda_u}{\sigma_u})^2 N_{dr} \gamma_{n_d} \gamma_{N_d}$, $\mathbf{G}_{12} = -(\frac{\lambda_e}{\sigma_e})^2 \gamma_{N_d} \mathbf{1}_{N_d} \mathbf{1}_{N_{dr}}^T = \mathbf{G}_{21}^T$, $\mathbf{G}_{13} = \frac{\lambda_e}{\sigma_e} \frac{\lambda_u}{\sigma_u} \gamma_{N_d} \mathbf{1}_{N_d} = \mathbf{G}_{31}^T$, and $\mathbf{G}_{23} = (\frac{\lambda_e}{\sigma_e})(\frac{\lambda_u}{\sigma_u}) N_{dr} \gamma_{n_d} \gamma_{N_d} \mathbf{1}_{n_d} \mathbf{G}_{32}^T$. Finally, we get

$$\boldsymbol{\Gamma}_{Y_{ds}} = \boldsymbol{\Gamma}_{Y_d} + \mathbf{D}_r \boldsymbol{\Sigma}_{rr.s} \mathbf{D}_r^T = \begin{bmatrix} \boldsymbol{\Gamma}_{rr} & \boldsymbol{\Gamma}_{rs} = \mathbf{0} & \boldsymbol{\Gamma}_{ru} = \mathbf{0} \\ \boldsymbol{\Gamma}_{sr} = \mathbf{0} & \boldsymbol{\Gamma}_{ss} & \boldsymbol{\Gamma}_{su} \\ \boldsymbol{\Gamma}_{ur} = \mathbf{0} & \boldsymbol{\Gamma}_{us} & \boldsymbol{\Gamma}_u \end{bmatrix}$$

where $\boldsymbol{\Gamma}_{rr} = (1 + \lambda_e^2) \mathbf{I}_{N_{dr}}$, $\boldsymbol{\Gamma}_{ss} = \mathbf{I}_{n_d} + \lambda_e^2 \gamma_{n_d} \mathbf{1}_{n_d} \mathbf{1}_{n_d}^T$, $\boldsymbol{\Gamma}_{su} = -\lambda_e \lambda_u \left(\frac{\sigma_e}{\sigma_u} \right) \gamma_{n_d} \mathbf{1}_{n_d} = \boldsymbol{\Gamma}_{us}^T$, and $\boldsymbol{\Gamma}_u = 1 + \lambda_u^2 (1 - n_d \gamma_{n_d})$.

Using the definition of the CSN (2.12), we have:

$$\Phi_{N_{dr}}(\mathbf{D}_{Y_{ds}}^r(\mathbf{y}_{dr} - \boldsymbol{\mu}_{dr}); \mathbf{0}, \boldsymbol{\Gamma}_{rr}) = \left(\frac{1}{2} \right)^{N_{dr}}. \quad (3.93)$$

The term $(\frac{1}{2})^{N_{dr}}$ from equation (3.93) partially cancels out with denominator $\mathbf{C}^{-1} =$

$(\frac{1}{2})^{N_d+m}$. This later equality holds because of the special form the matrices due to the nested error model assumption. Hence, the distribution defined in (3.92) reduces to the distribution (3.18) with N_d replaced by n_d .

Chapter 4

Empirical Best (EB) Prediction under One-Fold SN Model

In this chapter, we consider the one-fold nested error model (1.5) where both u_d and e_{dj} follow skew-normal (SN) distributions such that $\mathbb{E}(u_d) = \mathbb{E}(e_{dj}) = 0$. The goal is to estimate the small area parameters η_d , $d = 1, \dots, m$. If η_d is a non-complex parameter i.e. any linear function of the small area mean then we follow the traditional unit-level small area estimation (SAE) approach to derive the empirical best linear unbiased predictor (BLUP) and empirical Bayes (EB) estimators (see Rao (2003) for details on these predictors). When η_d is a complex parameter, we extend the method proposed by Molina and Rao (2010) to the nested error model with the assumption of SN distributed random errors (area and/or unit level errors). Because the Molina-Rao approach requires predicting values for the entire target population, it is essential to be able to generate the data in an univariate manner. The univariate approach for the EB estimator under skew normal is presented but given its complexity a simpler conditional version of the EB estimator is proposed. Alternative estimators (the half-normal, the Molina-Rao normality-based approach, and the traditional ELL method) are compared to the best prediction. A simulation study is conducted to compare these different predictors.

Throughout this chapter, we assume that the sampling method is non-informative meaning the inclusion probabilities are independent of the characteristic of interest i.e. $Pr(i \in s|\mathbf{y}) = Pr(i \in s)$ where s designates the sample, for all units i in the population (Pfeffermann et al. (1998)). Non-informative sampling methods ensure that the sample density probability functions are not modified by the sampling strategy. As mentioned earlier, we consider the partition $\mathbf{Y}_d = \begin{pmatrix} \mathbf{Y}_{dr} \\ \mathbf{Y}_{ds} \end{pmatrix}$ where dr designates the non-sampled units and ds the sampled units. As a reminder, we refer to the nested error linear with the errors following SN distributions as the *SN model* and similarly we refer to the nested error linear with the errors following normal distributions as the *normal model*.

4.1 Prediction for Linear Parameters

A small area mean, defined as $\bar{\mathbf{Y}}_d = \bar{\mathbf{X}}_d^T \boldsymbol{\beta} + u_d + \bar{e}_d$, is involved in many applications. Since $\bar{e}_d = \sum_{j=1}^{N_d} e_{dj} \approx 0$, the linear parameter $\bar{\mathbf{X}}_d^T \boldsymbol{\beta} + u_d$ is used as an approximation of the small area mean $\bar{\mathbf{Y}}_d$. More generally we define a linear parameter η_d as a linear combination of the fixed and the random area effects that is

$$\eta_d = \boldsymbol{\ell}_d^T \boldsymbol{\beta} + m_d u_d. \quad (4.1)$$

In this section, we show the EBLUP and EB estimators of (4.1) under SN model. When the random errors follow the normal distribution, the EBLUP estimator is the same as the EB estimator. These two predictors are not the same under the SN model and the EB estimator must be used to achieve the best performance in terms of MSE.

4.1.1 Empirical Best Linear Unbiased Prediction (EBLUP)

The BLUP estimator $\hat{\eta}_d^{BLUP}$ of η_d , defined by (4.1), is given by

$$\hat{\eta}_d^{BLUP} = \ell_d^T \tilde{\beta} + m_d \hat{u}_d^{BLUP} \quad (4.2)$$

$\tilde{\beta} = (\sum_{d=1}^m \mathbf{X}_{ds}^T \mathbf{V}_{ds}^{-1} \mathbf{X}_{ds})^{-1} (\sum_{d=1}^m \mathbf{X}_{ds}^T \mathbf{V}_{ds}^{-1} \mathbf{y}_{ds})$ is the generalized least squares estimator of β and $\hat{u}_d^{BLUP}$ is the BLUP of the area random effect u_d . Note that expression (4.2) does not require any parametric distribution assumption. Only the mean and covariance matrix of the joint distribution of u_d and $\mathbf{y}_{ds}$ is needed to define the BLUP of u_d as

$$\hat{u}_d^{BLUP} = \mathbb{E}(u_d) + C_d \Sigma_d^{-1} (\mathbf{Y}_{ds} - \mathbb{E}(\mathbf{Y}_{ds})) \quad (4.3)$$

where $C_d = \text{Cov}(u_d, \mathbf{Y}_{ds})$ and $\Sigma_d^{-1} = [\text{Cov}(\mathbf{Y}_{ds})]^{-1}$. Since $\mathbb{E}(u_d) = 0$, $C_d = \sigma_u^2(1 - \frac{2}{\pi} \delta_u^2) \mathbf{1}_{Nd}^T$, $\Sigma_d^{-1} = \frac{1}{\sigma_e^2(1 - \frac{2}{\pi} \delta_e^2)} (\mathbf{I}_{n_d} - \frac{\sigma_u^2(1 - \frac{2}{\pi} \delta_u^2)}{\sigma_e^2(1 - \frac{2}{\pi} \delta_e^2) + n_d \sigma_u^2(1 - \frac{2}{\pi} \delta_u^2)} \mathbf{1}_{n_d} \mathbf{1}_{n_d}^T)$, and $\mu_{Y_d} = \mathbf{X}_d \beta$ we have that

$$\hat{u}_d^{BLUP} = \frac{\sigma_u^2(1 - \frac{2}{\pi} \delta_u^2)}{\sigma_e^2(1 - \frac{2}{\pi} \delta_e^2) + n_d \sigma_u^2(1 - \frac{2}{\pi} \delta_u^2)} \sum_{j \in s_d} (y_{dj} - \mathbf{x}_{dj}^T \tilde{\beta}). \quad (4.4)$$

Note that the errors are adjusted so that their mean is equal to zero. The estimator $\hat{\eta}_d^{BLUP}$ from (4.2) is a function of the unknown parameters $\theta = (\beta, \mu_u, \mu_e, \sigma_u^2, \sigma_e^2, \lambda_u, \lambda_e)$ of the nested error model. Replacing the unknown parameters by their suitable estimators $\hat{\beta}$ in (4.2) results in the EBLUP estimator $\hat{\eta}_d^{EBLUP}$ of the linear small area parameter η_d , that is

$$\hat{\eta}_d^{EBLUP} = \hat{\eta}_d^{BLUP}(\theta = \hat{\theta}) \quad (4.5)$$

If $\lambda_u = \lambda_e = 0$ then the area effect predictor (4.4) and the small area parameter estimator (4.2) reduce to the results for the traditional normal model.

4.1.2 Empirical Best (EB) Prediction

The best estimator of the linear parameter is given by the conditional expectation of η_d given $\mathbf{y}_d$ and $\boldsymbol{\theta}$ which reduces to:

$$\hat{\eta}_d^B = \boldsymbol{\ell}_d^T \boldsymbol{\beta} + m_d \hat{u}_d^B \quad (4.6)$$

where $\hat{u}_d^B = \mathbb{E}(u_d | \mathbf{y}_{ds}, \boldsymbol{\theta})$. Note that, expression (4.6) requires a distribution assumption to obtain the conditional expectation $\hat{u}_d^B$. Obtaining the best estimator $\hat{u}_d^B$ requires knowing the joint distribution of u_d and $\mathbf{Y}_{ds}$. This joint distribution may not be easy to derive when the errors of the nested error model are not normal. For the SN model, Proposition 4.1 provides a joint distribution corresponding to the marginal distributions for the random errors.

Proposition 4.1. *Let $\mathbf{T} = \begin{pmatrix} \mathbf{Y}_{ds} \\ u_d \end{pmatrix} \sim CSN_{n_d+1, n_d+1}(\boldsymbol{\mu}_T, \boldsymbol{\Sigma}_T, \mathbf{D}_T, \boldsymbol{\nu}_T, \boldsymbol{\Gamma}_T)$ where*

$$\boldsymbol{\mu}_T = \begin{pmatrix} \boldsymbol{\mu}_{Y_d} = \mathbf{X}_d \boldsymbol{\beta} + \mu_u \mathbf{1}_{n_d} + \boldsymbol{\mu}_{eds} \\ \mu_u \end{pmatrix}, \quad \boldsymbol{\Sigma}_T = \begin{bmatrix} \sigma_e^2 \mathbf{I}_{n_d} + \sigma_u^2 \mathbf{1}_{n_d} \mathbf{1}_{n_d}^T & \sigma_u^2 \mathbf{1}_{n_d} \\ \sigma_u^2 \mathbf{1}_{n_d}^T & \sigma_u^2 \end{bmatrix},$$

$$\mathbf{D}_T = \begin{bmatrix} -\frac{\lambda_e}{\sigma_e} \mathbf{1}_{n_d} & \frac{\lambda_e}{\sigma_e} \mathbf{I}_{n_d} \\ \frac{\lambda_u}{\sigma_u} & \mathbf{0} \end{bmatrix}, \quad \boldsymbol{\nu}_T = \mathbf{0}, \quad \text{and} \quad \boldsymbol{\Gamma}_T = \mathbf{I}_{n_d+1}$$

then $u_d \sim SN(\mu_u, \sigma_u^2, \lambda_u) \equiv CSN_1(\mu_u, \sigma_u^2, \lambda_u, 0, 1)$ and $\mathbf{Y}_{ds}$ as in (3.18) where N_d is replaced by n_d .

Proof. The proof is straightforward by applying Proposition 2.7. □

From the joint distribution define in Proposition 4.1, we obtain the conditional distribution of u_d given $\mathbf{y}_{ds}$ as follows:

Proposition 4.2. *Assuming the joint distribution in Proposition 4.1, then*

$$u_d | \mathbf{y}_{ds} \sim CSN_{1, n_d+1}(\mu_A, \sigma_A, \mathbf{D}_A, \boldsymbol{\nu}_A, \boldsymbol{\Gamma}_A)$$

where

$$\mu_A = \mu_u + \gamma_{n_d} \mathbf{1}_{n_d}^T (\mathbf{y}_{n_d} - (\mathbf{X}_d \boldsymbol{\beta} + \mu_u \mathbf{1}_{n_d} + \boldsymbol{\mu}_{eds})), \quad \sigma_A = \sigma_e^2 \gamma_{n_d}, \quad \mathbf{D}_A = \begin{pmatrix} -\frac{\lambda_e}{\sigma_e} \mathbf{1}_{n_d} \\ \frac{\lambda_u}{\sigma_u} \end{pmatrix},$$

$$\boldsymbol{\nu}_A = - \begin{bmatrix} \frac{\lambda_e}{\sigma_e} [\mathbf{I}_{n_d} - \gamma_{n_d} \mathbf{1}_{n_d} \mathbf{1}_{n_d}^T] \\ \frac{\lambda_u}{\sigma_u} \gamma_{n_d} \mathbf{1}_{n_d}^T \end{bmatrix} (\mathbf{y}_{n_d} - (\mathbf{X}_d \boldsymbol{\beta} + \mu_u \mathbf{1}_{n_d} + \boldsymbol{\mu}_{eds})), \text{ and } \boldsymbol{\Gamma}_A = \mathbf{I}_{n_d+1}$$

Proof. The proof is straightforward by applying Proposition 2.8. $\square$

The best predictor of the small area random effect is now obtained using Theorem 4.1.

Theorem 4.1. *For the nested error model (3.16) with adjusted errors so that $\mathbb{E}(u_d) = 0$ and $\mathbb{E}(\mathbf{e}_d) = \mathbf{0}$ and assuming the joint distribution in Proposition 4.1, the best predictor of u_d is:*

$$\hat{u}_d^B = \mu_u + \gamma_{n_d} \sum_{j \in s_d} \{y_{dj} - (\mathbf{x}_{dj}^T \boldsymbol{\beta} + \mu_u + \mu_e)\} + \frac{\Phi_{n_d+1}^{(1)}(\mathbf{0}; \boldsymbol{\nu}_A, \boldsymbol{\Gamma})}{\Phi_{n_d+1}(\mathbf{0}; \boldsymbol{\nu}_A, \boldsymbol{\Gamma})} \quad (4.7)$$

where

$$\boldsymbol{\Gamma} = \begin{bmatrix} \mathbf{I}_{n_d} + \lambda_e^2 \gamma_{n_d} \mathbf{1}_{n_d} \mathbf{1}_{n_d}^T & -\lambda_e \lambda_u \left(\frac{\sigma_e}{\sigma_u} \right) \gamma_{n_d} \mathbf{1}_{n_d} \\ -\lambda_e \lambda_u \left(\frac{\sigma_e}{\sigma_u} \right) \gamma_{n_d} \mathbf{1}_{n_d}^T & 1 + \lambda_u^2 (1 - n_d \gamma_{n_d}) \end{bmatrix}, \text{ and}$$

$$\Phi_{n_d+1}^{(1)}(\mathbf{0}; \boldsymbol{\nu}_A, \boldsymbol{\Gamma}) = \frac{\partial}{\partial t} \Phi_{n_d+1}(\mathbf{D}_A \sigma_A t; \boldsymbol{\nu}_A, \boldsymbol{\Gamma})_{t=0},$$

with σ_A , $\mathbf{D}_A$, $\boldsymbol{\nu}_A$ defined as in Proposition 4.2, and a simpler expression of $\Phi_{n_d+1}(\mathbf{D}_A \sigma_A t; \boldsymbol{\nu}_A, \boldsymbol{\Gamma})$ provided by expression (4.8) below.

Proof. Proposition 4.2 gives the distribution of $u_d | \mathbf{y}_{ds}$ as a closed-skew normal. Therefore the moment generating function is obtained from Proposition 2.4 as:

$$M_{u_d | \mathbf{y}_{ds}}(t) = \frac{\Phi_{n_d+1}(\mathbf{D}_A \sigma_A t; \boldsymbol{\nu}_A, \boldsymbol{\Gamma}_A + \sigma_A \mathbf{D}_A \mathbf{D}_A^T)}{\Phi_{n_d+1}(\mathbf{0}; \boldsymbol{\nu}_A, \boldsymbol{\Gamma}_A + \sigma_A \mathbf{D}_A \mathbf{D}_A^T)} e^{t \boldsymbol{\mu}_A + \frac{1}{2} \sigma_A t^T t}, \text{ where } \boldsymbol{\mu}_A, \sigma_A, \mathbf{D}_A, \boldsymbol{\nu}_A, \text{ and } \boldsymbol{\Gamma}_A$$

are defined as in Proposition 4.2. Then the best predictor is obtained as the expectation using $\mathbb{E}(X) = \frac{\partial}{\partial t} M(t)_{t=0}$. $\square$

The expression $\frac{\partial}{\partial t}\Phi_{n_d+1}(\mathbf{D}_A\sigma_A t; \boldsymbol{\nu}_A, \boldsymbol{\Gamma})$ does not have any known analytical closed form. However, given that t is of dimension one, it is easy to numerically evaluate the derivative using commonly available statistical packages. The first step is to numerically evaluate the $(n_d + 1)$ -dimensional integral $\Phi_{n_d+1}(\mathbf{D}_A\sigma_A t; \boldsymbol{\nu}_A, \boldsymbol{\Gamma})$ prior to the derivative. To simplify the problem, we can take advantage of the special form of the covariance matrix $\boldsymbol{\Gamma}$ to reduce the $n_d + 1$ multi-dimensional integral in product of 1-dimensional integrals. This reduction technique is discussed in Appendix 3.5.1.

Consider $\mathbf{R}_\Gamma$, the correlation matrix associated with $\boldsymbol{\Gamma}$, given by:

$$[\mathbf{R}_\Gamma]_{ij} = \begin{cases} 1 & \text{for all elements } i = j \\ \frac{\lambda_e^2 \sigma_u^2}{\sigma_e^2 + n_d \sigma_u^2 + \lambda_e^2 \sigma_u^2} & \text{if } i < n_d + 1, \text{ both } i \neq j \\ -\frac{\lambda_e \sigma_u}{\sqrt{\sigma_e^2 + n_d \sigma_u^2 + \lambda_e^2 \sigma_u^2}} \frac{\lambda_u \sigma_e}{\sqrt{\sigma_e^2 + n_d \sigma_u^2 + \lambda_u^2 \sigma_e^2}} & i = n_d + 1 \text{ or } i \neq j \end{cases}$$

and $\mathbf{z} = (\mathbf{D}_A\sigma_A t - \boldsymbol{\nu}_A)^T \mathbf{V}^{-1/2}$ with

$$\mathbf{V}^{-1/2} = \begin{bmatrix} (1 + \lambda_e^2 \gamma_{n_d})^{-1/2} \mathbf{I}_{n_d} & \mathbf{0} \\ \mathbf{0} & (1 + \lambda_u^2 (1 - n_d \gamma_{n_d}))^{-1/2} \end{bmatrix}, \sigma_A, \mathbf{D}_A, \text{ and } \boldsymbol{\nu}_A \text{ defined as in Proposition 4.2. Then we have:}$$

$$\Phi_{n_d+1}(\mathbf{D}_A\sigma_A t; \boldsymbol{\nu}_A, \boldsymbol{\Gamma}) = \int_{-\infty}^{\infty} \prod_{i=1}^{n_d+1} \Phi_1(h_i(t)) \phi(u_0) du_0, \quad (4.8)$$

where $h_i(t) = \frac{z_i(t) - c_i u_0}{\sqrt{1 - c_i^2}}$ with z_i being the i^{th} element of $\mathbf{z}$ above and

$$c_i = \begin{cases} \frac{\lambda_u \sigma_e}{\sqrt{\sigma_e^2 + n_d \sigma_u^2 + \lambda_u^2 \sigma_e^2}} & \text{if } i = 1 \\ \frac{-\lambda_e \sigma_u}{\sqrt{\sigma_e^2 + n_d \sigma_u^2 + \lambda_e^2 \sigma_u^2}} & \text{if } i > 1 \end{cases}$$

Note that when $\lambda_u = \lambda_e = 0$, we get $\Phi_{n_d+1}(\mathbf{0}; \boldsymbol{\nu}_A, \boldsymbol{\Gamma}) = (\frac{1}{2})^{n_d+1}$ and $\Phi_{n_d+1}^{(1)}(\mathbf{0}; \boldsymbol{\nu}_A, \boldsymbol{\Gamma}) = \frac{\partial}{\partial t}\Phi_{n_d+1}(\mathbf{D}_A\sigma_A t; \boldsymbol{\nu}_A, \boldsymbol{\Gamma})_{t=0} = 0$. Therefore the best estimator defined in Theorem 4.1 reduces to the traditional best estimator under normal random errors which is the same as the BLUP estimator. Replacing the unknown parameters in (4.7) by their

suitable estimator, we obtain the EB estimator.

4.2 Prediction for Complex Parameters

The EB method described in Section 4.1 is not appropriate for complex parameters. A complex parameter may be defined as $\eta_d = h(\mathbf{y}_d)$ where h is a nonlinear function of $\mathbf{y}_d$. In the context of a finite population, $\mathbf{y}_d$ represents the census data for the characteristic of interest and it can be partitioned into $\mathbf{y}_{ds}$ the sample data and $\mathbf{y}_{dr}$ the out-of-sample data. Instead of using only the sample data to estimate η_d , Elbers et al. (2003) proposed using the entire census by predicting the entire characteristic of interest $\mathbf{y}_d$. They developed a semi-parametric approach using the nested error model framework but they did not assume any distribution for the random components. They used the area-level and unit-level residuals to construct the predictor (see Section 4.7). The resulting estimator, say $\hat{\eta}_d^{res}$, has a small bias but large MSE as shown by Molina and Rao (2010). This approach is referred to as the ELL method.

Molina and Rao (2010) proposed a full parametric version of the ELL method. We refer to this approach as the Molina-Rao (MR) method. They assumed a normal distribution for the random components, however, their method is applicable to other probability density functions. The MR method has two main ideas: a) predict the non-sampled characteristic of interest $\mathbf{y}_{dr}$, using the conditional distribution of the non-observed characteristic given the sample data, to form the predicted census and b) numerically evaluate the conditional expectation of the complex parameter given the sample data using several predicted censuses. There are two generic difficulties. In point a) for a non-gaussian setup, the conditional distribution of the characteristic of interest given the sample may be challenging to obtain. The second difficulty is the dimension of the non-observed vector of characteristic of interest. This vector

may be so large that it is impossible to practically generate from the multivariate conditional distribution given the current computational capabilities. Therefore it is essential to be able to draw from the conditional distribution in a univariate manner. A univariate approximation to the multivariate conditional distribution may be necessary for some complex setups (non-gaussians or with sub-small area levels). The numerical evaluation in point b) above is only motivated by the fact that analytical expression of the conditional expectation is not known for most complex parameters.

The best estimator of η_d , in terms of minimizing the MSE, is the conditional expectation

$$\hat{\eta}_d^B = \mathbb{E}(\eta_d | \mathbf{y}_{ds}) = \int h(\mathbf{y}_d) f_{dr|s}(\mathbf{y}) d\mathbf{y} \quad (4.9)$$

where $f_{dr|s}$ is the conditional distribution of $\mathbf{Y}_{dr}$ given $\mathbf{y}_{ds}$. This conditional distribution is derived under the assumption that there is no sampling selection bias i.e the population model holds for the sample. The difficulty with the estimator (4.9) is the lack of closed-form expression in most situations due to the complexity of the function h .

To evaluate the expectation (4.9), the empirical Monte Carlo method proposed by Molina and Rao (2010) is used. It consists of generating a large number L of vectors from the conditional distribution of $\mathbf{Y}_{dr}$ given $\mathbf{y}_{ds}$. Let $\mathbf{y}_{dr}^{(\ell)}$ be the ℓ^{th} , $\ell = 1, \dots, L$, simulated vector for the out of sample observations obtained from the conditional distribution. The unknown vector $\mathbf{y}_{dr}$ is replaced by its predictor $\mathbf{y}_{dr}^{(\ell)}$ to form a complete censuses which we call the “predicted census”. Each of the L predicted census produces a predicted complex parameter $\hat{\eta}_d^{(\ell)}$, $\ell = 1, \dots, L$. A Monte Carlo approximation to the best predictor $\hat{\eta}_d^B$ of η_d , we call the quasi-best (QB) estimator, is then given by:

$$\hat{\eta}_d^{QB} = \frac{1}{L} \sum_{\ell=1}^L \hat{\eta}_d^{(\ell)} = \frac{1}{L} \sum_{\ell=1}^L h(\mathbf{y}_{dr}^{(\ell)}, \mathbf{y}_{ds}). \quad (4.10)$$

We call this estimator “quasi” because the expectation (4.9) is evaluated numerically as an approximation of the unknown analytical expression. Note that there are several methods that can be used to evaluate the integral (4.9) such as the midpoint Rule, the trapezoid rule, or the Simpson’s rule. In practice, the conditional probability distribution of $\mathbf{Y}_{dr}$ given $\mathbf{y}_{ds}$ is a function of unknown parameters $\boldsymbol{\theta}$ that need to be estimated using suitable methods such as the maximum likelihood (ML). The empirical quasi-best (EQB) estimator is the estimator (4.10) evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ where $\hat{\boldsymbol{\theta}}$ is the estimator of $\boldsymbol{\theta}$ using a suitable method.

In summary, the steps to obtain an empirical best estimator are the following:

1. Estimate the unknown parameter $\boldsymbol{\theta}$ using a suitable method such as ML estimation to obtain $\hat{\boldsymbol{\theta}}$. A ML approach for estimating the parameters is presented in Section 3.4.1.
2. Draw L out-of-sample vectors $\mathbf{y}_{dr}^{(\ell)}$, $\ell = 1, \dots, L$ from the predictive conditional distribution of $\mathbf{Y}_{dr}$ given $\mathbf{y}_{ds}$ evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. Often the dimension $N_d - n_d$ of the vector $\mathbf{y}_{dr}^{(\ell)}$ is very large and as a consequence a multivariate generation from $\mathbf{Y}_{dr}$ given $\mathbf{y}_{ds}$ is not feasible. Therefore, it is essential to be able to generate in a univariate manner the values $\mathbf{y}_{drj}^{(\ell)}$, $j = 1, \dots, N_d - n_d$ from the conditional distribution $\mathbf{Y}_{dr}$ given $\mathbf{y}_{ds}$.
3. Augment each of the L generated vector $\mathbf{y}_{dr}^{(\ell)}$ with the sample data to obtain the “predicted census” $\mathbf{y}_d^{(\ell)} = (\mathbf{y}_{dr}^{(\ell)T}, \mathbf{y}_{ds}^T)^T$.
4. Calculate the small area parameter of interest $\hat{\eta}_d^{(\ell)}(\hat{\boldsymbol{\theta}}) = h(\mathbf{y}_d^{(\ell)})$.
5. Obtain an approximation to the empirical best (EB) estimator as follows:

$$\hat{\eta}_d^{EQB} = \frac{1}{L} \sum_{\ell=1}^L \hat{\eta}_d^{(\ell)}(\hat{\boldsymbol{\theta}}) \quad (4.11)$$

We refer to the estimator (4.11) as the empirical quasi-best (EQB) predictor.

Note that this method only predicts the out-of-sample data $\mathbf{y}_{dr}$ unlike the ELL method (described in Section 4.7) which predicts the entire census including the sample data $\mathbf{y}_{ds}$. The univariate generation is not necessary trivial when the nested design is complex than one fold (for instance, clusters within small areas) or the random errors are not normally distributed.

4.3 Best Prediction for Complex Parameters under SN Random Errors

The best prediction discussed in 4.2 requires generating out-of-sample values from the conditional distribution of the random vector $\mathbf{Y}_{dr}$ given $\mathbf{y}_{ds}$ which we refer to as $\mathbf{Y}_{dr|s}$ for simplicity. We have already shown in Chapter 3, that $\mathbf{Y}_d$ follows a closed skew-normal (CSN) distribution when the random errors of the nested errors model follow SN distributions. In the subsections below, the distribution of $\mathbf{Y}_{dr|s}$ is provided as well as a quasi-univariate way of generating data from the predictor's distribution.

4.3.1 Conditional Distribution of $\mathbf{Y}_{dr|s}$

When both e_{dj} and u_d follow SN distributions, the vector $\mathbf{Y}_d$ follows the distribution defined by (3.18). Using Proposition 2.8, it follows that the conditional distribution of $\mathbf{Y}_{dr|s}$ is a member of the CSN family and we have:

$$\mathbf{Y}_{dr|s} \sim CSN_{N_{dr}, N_d+1}(\boldsymbol{\mu}_{dr|s}, \boldsymbol{\Sigma}_{dr|s}, \mathbf{D}_{dr|s}, \boldsymbol{\nu}_{dr|s}, \boldsymbol{\Gamma}_{dr|s} = \boldsymbol{\Gamma}_{Y_d}) \quad (4.12)$$

where,

$$\boldsymbol{\mu}_{dr|s} = \mathbf{X}_{dr}\boldsymbol{\beta} + \mu_u \mathbf{1}_{N_{dr}} + \boldsymbol{\mu}_{edr} + \gamma_{n_d} [\mathbf{1}_{n_d}^T (\mathbf{y}_{ds} - \boldsymbol{\mu}_{ds})] \mathbf{1}_{N_{dr}}$$

$$\boldsymbol{\Sigma}_{dr|s} = \sigma_e^2 [\mathbf{I}_{N_{dr}} + \gamma_{n_d} \mathbf{1}_{N_{dr}} \mathbf{1}_{N_{dr}}^T]$$

$$\mathbf{D}_{dr|s} = \mathbf{D}_{dr} = \begin{bmatrix} \frac{\lambda_e}{\sigma_e} \left[\begin{bmatrix} \mathbf{I}_{N_{dr}} \\ \mathbf{0}_{(n_d \times N_{dr})} \end{bmatrix} - \gamma_{N_d} \mathbf{1}_{N_d} \mathbf{1}_{N_{dr}}^T \right] \\ \frac{\lambda_u}{\sigma_u} \gamma_{N_d} \mathbf{1}_{N_d}^T \end{bmatrix}$$

$$\boldsymbol{\nu}_{dr|s} = - \begin{bmatrix} \mathbf{0}_{(N_{dr} \times n_d)} \\ \frac{\lambda_e}{\sigma_e} [\mathbf{I}_{n_d} - \gamma_{n_d} \mathbf{1}_{n_d} \mathbf{1}_{n_d}^T] \\ \frac{\lambda_u}{\sigma_u} \gamma_{n_d} \mathbf{1}_{n_d}^T \end{bmatrix} (\mathbf{y}_{ds} - \boldsymbol{\mu}_{ds}).$$

Similarly when e_{dj} follows SN distribution and u_d is normally distributed, the conditional distribution of $\mathbf{Y}_{dr|s}$ is:

$$\mathbf{Y}_{dr|s} \sim CSN_{N_{dr}, N_d+1} (\boldsymbol{\mu}_{dr|s}, \boldsymbol{\Sigma}_{dr|s}, \mathbf{D}_{dr|s}, \boldsymbol{\nu}_{dr|s}, \boldsymbol{\Gamma}_{dr|s} = \boldsymbol{\Gamma}_{Y_d}) \quad (4.13)$$

where $\boldsymbol{\Sigma}_{dr|s}$ is defined as in (4.12),

$$\boldsymbol{\mu}_{dr|s} = \mathbf{X}_{dr}\boldsymbol{\beta} + \boldsymbol{\mu}_{edr} + \gamma_{n_d} [\mathbf{1}_{n_d}^T (\mathbf{y}_{ds} - \mathbf{X}_{ds}\boldsymbol{\beta} - \boldsymbol{\mu}_{eds})] \mathbf{1}_{N_{dr}},$$

$$\mathbf{D}_{dr|s} = \mathbf{D}_{dr} = \begin{bmatrix} \frac{\lambda_e}{\sigma_e} \left[\begin{bmatrix} \mathbf{I}_{N_{dr}} \\ \mathbf{0}_{(n_d \times N_{dr})} \end{bmatrix} - \gamma_{N_d} \mathbf{1}_{N_d} \mathbf{1}_{N_{dr}}^T \right] \\ \mathbf{0}_{(1 \times N_{dr})} \end{bmatrix}, \text{ and}$$

$$\boldsymbol{\nu}_{dr|s} = - \begin{bmatrix} \mathbf{0}_{(N_{dr} \times n_d)} \\ \frac{\lambda_e}{\sigma_e} [\mathbf{I}_{n_d} - \gamma_{N_d} \mathbf{1}_{n_d} \mathbf{1}_{n_d}^T] \\ \mathbf{0}_{(1 \times n_d)} \end{bmatrix} (\mathbf{y}_{ds} - \mathbf{X}_{ds}\boldsymbol{\beta} - \boldsymbol{\mu}_{eds}).$$

For the situation where u_d follows SN distribution and e_{dj} is normally distributed, the conditional distribution of $\mathbf{Y}_{dr|s}$ is:

$$\mathbf{Y}_{dr|s} \sim CSN_{N_{dr}, N_d+1} (\boldsymbol{\mu}_{dr|s}, \boldsymbol{\Sigma}_{dr|s}, \mathbf{D}_{dr|s}, \boldsymbol{\nu}_{dr|s}, \boldsymbol{\Gamma}_{dr|s} = \boldsymbol{\Gamma}_{Y_d}) \quad (4.14)$$

where $\boldsymbol{\Sigma}_{dr|s}$ is defined as in (4.12),

$$\boldsymbol{\mu}_{dr|s} = \mathbf{X}_{dr}\boldsymbol{\beta} + \mu_u \mathbf{1}_{N_{dr}} + \gamma_{n_d} [\mathbf{1}_{n_d}^T (\mathbf{y}_{ds} - \mathbf{X}_{ds}\boldsymbol{\beta} - \mu_u \mathbf{1}_{n_d})] \mathbf{1}_{N_{dr}},$$

$$\mathbf{D}_{dr|s} = \mathbf{D}_{dr} = \begin{bmatrix} \mathbf{0}_{(N_d \times N_{dr})} \\ \frac{\lambda_u}{\sigma_u} \gamma_{n_d} \mathbf{1}_{N_d}^T \end{bmatrix}, \text{ and}$$

$$\boldsymbol{\nu}_{dr|s} = - \begin{bmatrix} \mathbf{0}_{(N_d \times n_d)} \\ \frac{\lambda_u}{\sigma_u} \gamma_{n_d} \mathbf{1}_{n_d}^T \end{bmatrix} (\mathbf{y}_{ds} - \mathbf{X}_{ds}\boldsymbol{\beta} - \mu_u \mathbf{1}_{n_d}).$$

4.3.2 Quasi-Univariate Generation for Big Data

Being able to generate the values in a univariate manner is fundamental because the method requires predicting the values of interest for a large portion of the target population. Molina and Rao (2010)'s method generates predicted values of the variables of interest for all non-sampled units while the ELL method generates those values for the entire population including the sample units. Many applications of these methods involve very large datasets. For instance, the ELL method has been used by the World Bank to estimate SAE poverty related measures for numerous countries. In these applications, the census information of the countries are used as the population data. It is easy to see how these census datasets can get very large especially in some developing countries like Indonesia with more than 200 million citizens. Given the current capabilities of computers and software, the multivariate models presented in this chapter can be very challenging to implement and even impossible to run for many of these countries. Therefore it is essential to be able to generate values using univariate schemes. The univariate approach translates into loop statements when using computer software. Loops can be significantly slower than matrix computations in R but are always feasible. Some other free use software like Julia and C/C++ are very efficient in running loops and can be considered when very large univariate runs are necessary.

The question is: how to generate values from the conditional distribution (4.12)

using a univariate scheme? Given the special forms of the matrices involved, it is possible to generate the data in a quasi-univariate manner meaning that all but a vector of size $n_d + 1$ can be generated using a univariate scheme. This is not a full univariate approach, as in the normal random errors case, but given the relatively small size n_d , this quasi-univariate approach is always feasible.

Arellano-Valle and Azzalini (2006) showed how to generate data from a CSN using a generalized version of result (2.1). They used a convolution mechanism instead of the conditioning argument. Begin by considering

$$\mathbf{V}_0 \sim LTN_q(\boldsymbol{\nu}, \mathbf{0}, \boldsymbol{\Psi}_0) \quad \text{and} \quad \mathbf{V}_1 \sim N_p(\mathbf{0}, \boldsymbol{\Psi}_1) \quad (4.15)$$

where $\boldsymbol{\Psi}_0$ and $\boldsymbol{\Psi}_1$ are covariance matrices and the notation $LTN_q(\mathbf{c}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a multivariate normal truncated below $\mathbf{c}$. Then consider the transformed random vector

$$\mathbf{V} = \boldsymbol{\mu} + \mathbf{B}_0 \mathbf{V}_0 + \mathbf{B}_1 \mathbf{V}_1 \quad (4.16)$$

where $\mathbf{B}_0 = (\mathbf{D}\boldsymbol{\Sigma})^T \boldsymbol{\Psi}_0^{-1}$ and $\mathbf{B}_1$ is a $p \times p$ matrix such that

$$\mathbf{B}_1 \boldsymbol{\Psi}_1 \mathbf{B}_1^T = \boldsymbol{\Sigma} - (\mathbf{D}\boldsymbol{\Sigma})^T \boldsymbol{\Psi}_0^{-1} (\mathbf{D}\boldsymbol{\Sigma}). \quad (4.17)$$

Arellano-Valle and Azzalini (2006) showed that the random vector $\mathbf{V}$ has the density function:

$$f_{p,q}(\mathbf{v}) = \phi_p(\mathbf{v}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \frac{\Phi_q(\mathbf{D}(\mathbf{v} - \boldsymbol{\mu}); \boldsymbol{\nu}, \boldsymbol{\Psi}_0 - \mathbf{D}\boldsymbol{\Sigma}^{-1}\mathbf{D}^T)}{\Phi_q(\mathbf{0}; \boldsymbol{\nu}, \boldsymbol{\Psi}_0)} \quad (4.18)$$

Note that if $p = N_{dr}$, $q = N_d + 1$, $\boldsymbol{\mu} = \boldsymbol{\mu}_{dr|s}$, $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{dr|s}$, $\mathbf{D} = \mathbf{D}_{dr|s}$, $\boldsymbol{\nu} = \boldsymbol{\nu}_{dr|s}$, and $\boldsymbol{\Psi}_0 = \boldsymbol{\Gamma}_{dr|s} + \mathbf{D}_{dr|s} \boldsymbol{\Sigma}_{dr|s} \mathbf{D}_{dr|s}^T$ then equation (4.18) defines the density function of the conditional distribution (4.12). In other words, if $\mathbf{V}_{dr|s}^0$ and $\mathbf{V}_{dr|s}^1$ are independent

random vectors such that:

$$\mathbf{V}_{dr|s}^0 \sim LTN_{N_d+1}(\boldsymbol{\nu}_{dr|s}; \mathbf{0}, \boldsymbol{\Psi}_{dr|s}^0 = \boldsymbol{\Gamma}_{dr|s} + \mathbf{D}_{dr|s} \boldsymbol{\Sigma}_{dr|s} \mathbf{D}_{dr|s}^T) \quad (4.19)$$

$$\mathbf{V}_{dr|s}^1 \sim N_{N_{dr}}(\mathbf{0}, \boldsymbol{\Psi}_{dr|s}^1 = \boldsymbol{\Sigma}_{dr|s} - (\mathbf{D}_{dr|s} \boldsymbol{\Sigma}_{dr|s})^T (\boldsymbol{\Psi}_{dr|s}^0)^{-1} \mathbf{D}_{dr|s} \boldsymbol{\Sigma}_{dr|s}), \quad (4.20)$$

then the random vector $\mathbf{Y}_{dr|s}$ defined in (4.12) can be obtained using the transformation:

$$\mathbf{Y}_{dr|s} = \boldsymbol{\mu}_{dr|s} + \mathbf{B}_{dr|s}^0 \mathbf{V}_{dr|s}^0 + \mathbf{V}_{dr|s}^1 \quad (4.21)$$

where the matrix $\mathbf{B}_{dr|s}^0$ is defined as follows

$$\mathbf{B}_{dr|s}^0 = (\mathbf{D}_{dr|s} \boldsymbol{\Sigma}_{dr|s})^T \boldsymbol{\Psi}_0^{-1} = \begin{bmatrix} \frac{\lambda_e \sigma_e}{1 + \lambda_e^2} \mathbf{I}_{N_{dr}} & b_{01} \mathbf{1}_{N_{dr}} \mathbf{1}_{n_d}^T & b_{02} \mathbf{1}_{N_{dr}} \end{bmatrix}, \quad (4.22)$$

with

$$b_{01} = \frac{-\gamma_{n_d} \lambda_e \sigma_e \sigma_u^2}{(1 + n_d \gamma_{n_d} \lambda_e^2) \sigma_u^2 + \gamma_{n_d} \lambda_u^2 \sigma_e^2} \quad (4.23)$$

$$b_{02} = \frac{\gamma_{n_d} \lambda_u \sigma_u \sigma_e^2}{(1 + n_d \gamma_{n_d} \lambda_e^2) \sigma_u^2 + \gamma_{n_d} \lambda_u^2 \sigma_e^2} \quad (4.24)$$

Using expression (4.21), the univariate generation of $\mathbf{Y}_{dr|s}$ reduces to the univariate generations of $\mathbf{V}_{dr|s}^0$ and $\mathbf{V}_{dr|s}^1$. The nested error model assumption leads to special forms of the matrices involved in the best predictor's distribution. Under this model we have:

$$\boldsymbol{\Psi}_{dr|s}^0 = \begin{bmatrix} (1 + \lambda_e^2) \mathbf{I}_{N_{dr}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{n_d} + \lambda_e^2 \gamma_{n_d} \mathbf{1}_{n_d} \mathbf{1}_{n_d}^T & -\lambda_e \lambda_u \left(\frac{\sigma_e}{\sigma_u} \right) \gamma_{n_d} \mathbf{1}_{n_d} \\ \mathbf{0} & -\lambda_e \lambda_u \left(\frac{\sigma_e}{\sigma_u} \right) \gamma_{n_d} \mathbf{1}_{n_d}^T & 1 + \lambda_u^2 \left(\frac{\sigma_e}{\sigma_u} \right)^2 \gamma_{n_d} \end{bmatrix}, \quad (4.25)$$

and

$$\boldsymbol{\Psi}_{dr|s}^1 = \alpha \mathbf{I}_{N_{dr}} + \beta \mathbf{1}_{N_{dr}} \mathbf{1}_{N_{dr}}^T \quad (4.26)$$

with α and β defined as follows

$$\alpha = \sigma_e^2 \left(1 - \frac{\lambda_e^2}{1 + \lambda_e^2} \right) = \sigma_e^2 (1 - \delta_e^2), \quad (4.27)$$

$$\beta = \frac{\gamma_{n_d} \sigma_u^2 \sigma_e^2}{(1 + n_d \gamma_{n_d} \lambda_e^2) \sigma_u^2 + \gamma_{n_d} \lambda_u^2 \sigma_e^2} \quad (4.28)$$

The random vector $\mathbf{V}_{dr|s}^1$ is easily generated in a univariate manner by:

$$(\mathbf{V}_{dr|s}^1)_j = v_{dj}^{11} + v_d^{10}, \quad j = 1, \dots, n_d \text{ and } d = 1, \dots, m, \quad (4.29)$$

where $v_{dj}^{11} \stackrel{ind}{\sim} N(0, \alpha)$ is independent of $v_d^{10} \stackrel{ind}{\sim} N(0, \beta)$ and α and β are defined as in (4.27, 4.28). Note that, for each small area d , we only generate one value v_d^{10} and n_d values v_{dj}^{11} , $j = 1, \dots, n_d$.

The random vector $\mathbf{V}_{dr|s}^0$ can be decomposed into two components $\mathbf{V}_{dr|s}^{0r}$ and $\mathbf{V}_{dr|s}^{0s}$ where

$$\mathbf{V}_{dr|s}^0 = \begin{pmatrix} \mathbf{V}_{dr|s}^{0r} \\ \mathbf{V}_{dr|s}^{0s} \end{pmatrix} \sim LT N_{N_d+1} \left(\begin{pmatrix} \mathbf{0} \\ \boldsymbol{\nu}_{dr|s}^{0s} \end{pmatrix}; \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{bmatrix} \boldsymbol{\Psi}_{dr|s}^{0r} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Psi}_{dr|s}^{0s} \end{bmatrix} \right) \quad (4.30)$$

and the components are defined as follows

$$\boldsymbol{\nu}_{dr|s}^{0s} = - \begin{bmatrix} \frac{\lambda_e}{\sigma_e} [\mathbf{I}_{n_d} - \gamma_{n_d} \mathbf{1}_{n_d} \mathbf{1}_{n_d}^T] \\ \frac{\lambda_u}{\sigma_u} \gamma_{n_d} \mathbf{1}_{n_d}^T \end{bmatrix} (\mathbf{y}_{d_s} - \boldsymbol{\mu}_{d_s}) \quad (4.31)$$

$$\boldsymbol{\Psi}_{dr|s}^{0r} = (1 + \lambda_e^2) \mathbf{I}_{N_{dr}} \quad (4.32)$$

$$\boldsymbol{\Psi}_{dr|s}^{0s} = \begin{bmatrix} \mathbf{I}_{N_{dr}} + \lambda_e^2 \gamma_{n_d} \mathbf{1}_{n_d} \mathbf{1}_{n_d}^T & -\lambda_e \lambda_u \left(\frac{\sigma_e}{\sigma_u} \right) \gamma_{n_d} \mathbf{1}_{n_d} \\ -\lambda_e \lambda_u \left(\frac{\sigma_e}{\sigma_u} \right) \gamma_{n_d} \mathbf{1}_{n_d}^T & 1 + \lambda_u^2 \left(\frac{\sigma_e}{\sigma_u} \right)^2 \gamma_{n_d} \end{bmatrix}. \quad (4.33)$$

$\mathbf{V}_{dr|s}^{0r}$ is simply the usual multivariate half-normal with a diagonal covariance matrix $\boldsymbol{\Psi}_{dr|s}^{0r} = (1 + \lambda_e^2) \mathbf{I}_{N_{dr}}$. This means that to obtain a realisation of $\mathbf{V}_{dr|s}^{0r}$, it is sufficient to generate univariate normal distributed values and take their absolute values. In

other words,

$$(\mathbf{V}_{dr|s}^{0r})_j = |v_{dj}^{0r}|, \quad j = 1, \dots, N_{dr} \text{ and } d = 1, \dots, m, \quad (4.34)$$

where $v_{dj}^{0r} \stackrel{\text{ind}}{\sim} N(0, 1 + \lambda_e^2)$.

The only vector left to be generated is $\mathbf{V}_{dr|s}^{0s}$. Unfortunately there is no known approach for generating the vector $\mathbf{V}_{dr|s}^{0s}$ in a univariate manner because $\boldsymbol{\nu}_{dr|s}^{0s}$ is not equal to the zero vector. However the length of this vector is $n_d + 1$ which is easily manageable in a multivariate way especially given the small sample sizes involved in SAE. The vector $\mathbf{V}_{dr|s}^{0s}$ will therefore be generated using the multivariate distribution.

The naive (acceptance/rejection) method for simulating a truncated multivariate normal $LTN(\mathbf{c}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ consists of generating a random vector from the equivalent multivariate $\mathbf{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ until $\mathbf{Y} > \mathbf{c}$. This naive method often requires a large number of simulations to get one random vector and can be extremely inefficient. Much more efficient alternative methods to draw from a multivariate truncated normal are based on Markov Chain Monte Carlo (MCMC) techniques such as Gibbs sampling. These methods can be much faster than the rejection approach but as any MCMC technique, they are approximations which may suffer from several problems: poor mixing, convergence issues, autocorrelation between generated samples, etc. Diagnostic checks can be very time consuming when applying MCMC methods.

Most of the major statistical software packages offer procedures for generating univariate truncated normals and some of them have routines for generating from multivariate truncated distributions. For the simulation study presented in Section 4.9, the R-package `tmvtnorm` (Truncated Multivariate Normal and Student t Distribution) was used to generate data from multivariate truncated normal distributions.

In summary the quasi-univariate procedure below shows how to generate the

predicted value $(y_{dr|s})_j$ for the non-sampled unit j in small area d :

1. Estimate the unknown parameters $\boldsymbol{\theta}$ of the model using a suitable method and the sample data $\mathbf{y}_s$.
2. Generate values v_{dj}^{11} from $N(0, \alpha)$ and v_d^{10} from $N(0, \beta)$ independently where α and β are respectively defined in (4.27) and (4.28) with $\boldsymbol{\theta}$ replaced by its estimator $\hat{\boldsymbol{\theta}}$. Note that only one v_d^{10} is generated for each small area d but a separate v_{dj}^{11} is draw for each unit j in small area d .
3. Generate value v_{dj}^{0r} from $N(0, 1 + \lambda_e^2)$ with λ_e^2 replaced by its estimator $\hat{\lambda}_e^2$.
4. Generate vector $\mathbf{v}_{dr|s}^{0s}$ from $LTN_{n_d+1}(\nu_{dr|s}^{0s}; \mathbf{0}, \boldsymbol{\Psi}_{dr|s}^{0s})$ where $\nu_{dr|s}^{0s}$ and $\boldsymbol{\Psi}_{dr|s}^{0s}$ are respectively defined in (4.31) and (4.33) with $\boldsymbol{\theta}$ replaced by $\hat{\boldsymbol{\theta}}$.
5. Create the vector $\mathbf{w}_{dr|s}^{0s}$ as follows

$$\mathbf{w}_{dr|s}^{0s} = (b_{01}\mathbf{1}_{N_{dr}}\mathbf{1}_{n_d}^T \quad b_{02}\mathbf{1}_{N_{dr}}) \mathbf{v}_{dr|s}^{0s} \quad (4.35)$$

where b_{01} and b_{02} are defined respectively in (4.23) and (4.24), with the parameters replaced by their estimators.

6. The element dj of the vector $y_{dr|s}$ is:

$$(y_{dr|s})_j = (\hat{\boldsymbol{\mu}}_{dr|s})_j + v_{dj}^{11} + v_d^{10} + \frac{\hat{\lambda}_e \hat{\sigma}_e}{1 + \hat{\lambda}_e^2} |v_{dj}^{0r}| + (w_{dr|s}^{0s})_j \quad (4.36)$$

$$\text{where } (\hat{\boldsymbol{\mu}}_{dr|s})_j = \mathbf{X}_{drj} \hat{\boldsymbol{\beta}} + \hat{\mu}_u + \hat{\mu}_e + \hat{\gamma}_{n_d} [\mathbf{1}_{n_d}^T (\mathbf{y}_{ds} - \hat{\boldsymbol{\mu}}_{ds})]$$

The necessity of generating from the multivariate truncated normal creates some complexity because even though the vector is small and easily manageable, the use of MCMC methods seems necessary. Section 4.4 provides a simpler approach for the

best prediction which is easily implemented using a full univariate scheme and does not require the use of MCMC methods.

In sections 4.3.3 and 4.3.4, we discuss the two special cases where one of the two random errors e_{dj} and u_d follows SN distribution and the other follows normal distribution. Note that these two situations are already covered in this Section 4.3.2, the next two sections only give more specifics for these two special cases. The special case for which both e_{dj} and u_d follow normal distribution is briefly presented in Section 4.6 and corresponds to the setup discussed by Molina and Rao (2010).

4.3.3 Special Case 1: e_{dj} follows SN and u_d follows normal

The conditional distribution for this special case is provided by (4.13). The transformation (4.21) is still valid where $\mathbf{B}_{dr|s}^0$ is defined with

$$b_{01} = \frac{-\gamma_{n_d}\lambda_e\sigma_e}{1 + n_d\gamma_{n_d}\lambda_e^2} \quad \text{and} \quad b_{02} = 0 \quad (4.37)$$

The vector $\mathbf{V}_{dr|s}^1$ from transformation (4.21) follows $N_{N_{dr}}(0, \mathbf{\Psi}_{dr|s}^1 = \alpha\mathbf{I}_{N_{dr}} + \beta\mathbf{1}_{N_{dr}}\mathbf{1}_{N_{dr}}^T)$ with α and β defined as follows

$$\alpha = \sigma_e^2 (1 - \delta_e^2) \quad \text{and} \quad \beta = \frac{\gamma_{n_d}\sigma_e^2}{1 + n_d\gamma_{n_d}\lambda_e^2} \quad (4.38)$$

The random vector $\mathbf{V}_{dr|s}^0$ can be decomposed into three components $\mathbf{V}_{dr|s}^{0r}$, $\mathbf{V}_{dr|s}^{0s(e)}$, and $\mathbf{V}_{dr|s}^{0s(u)}$ where

$$\mathbf{V}_{dr|s}^0 = \begin{pmatrix} \mathbf{V}_{dr|s}^{0r} \\ \mathbf{V}_{dr|s}^{0s(e)} \\ \mathbf{V}_{dr|s}^{0s(u)} \end{pmatrix} \sim LTN_{N_d+1} \left(\begin{pmatrix} \mathbf{0} \\ \boldsymbol{\nu}_{dr|s}^{0s(e)} \\ \mathbf{0} \end{pmatrix} ; \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{bmatrix} \boldsymbol{\Psi}_{dr|s}^{0r} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Psi}_{dr|s}^{0s(e)} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & 1 \end{bmatrix} \right)$$

with $\Psi_{dr|s}^{0r} = (1 + \lambda_e^2) \mathbf{I}_{N_{dr}}$ and

$$\nu_{dr|s}^{0s(e)} = -\frac{\lambda_e}{\sigma_e} [\mathbf{I}_{n_d} - \gamma_{n_d} \mathbf{1}_{n_d} \mathbf{1}_{n_d}^T] (\mathbf{y}_{d_s} - \boldsymbol{\mu}_{d_s}) \quad (4.39)$$

$$\Psi_{dr|s}^{0s(e)} = \mathbf{I}_{N_{dr}} + \lambda_e^2 \gamma_{n_d} \mathbf{1}_{n_d} \mathbf{1}_{n_d}^T \quad (4.40)$$

Note that because $b_{02} = 0$, the random component $\mathbf{V}_{dr|s}^{0s(u)}$ does not contribute to the transformation (4.21).

In summary the quasi-univariate procedure below shows how to generate the predicted value $(y_{dr|s})_{dj}$ for the non-sampled unit j in small area d :

1. Estimate the unknown parameters $\boldsymbol{\theta}$ of the model using a suitable method and the sample data $\mathbf{y}_s$.
2. Generate values v_{dj}^{11} from $N(0, \alpha)$ and v_d^0 from $N(0, \beta)$ independently where α and β are defined in (4.38) with $\boldsymbol{\theta}$ replaced by its estimator $\hat{\boldsymbol{\theta}}$. Note that only one v_d^{10} is generated for each small area d but a separate v_{dj}^{11} is draw for each unit j in small area d .
3. Generate value v_{dj}^{0r} from $N(0, 1 + \lambda_e^2)$ with λ_e^2 replaced by its estimator $\hat{\lambda}_e^2$.
4. Generate vector $\mathbf{v}_{dr|s}^{0s(e)}$ from $LTN_{n_d} \left(\nu_{dr|s}^{0s(e)}; \mathbf{0}, \Psi_{dr|s}^{0s(e)} \right)$ where $\nu_{dr|s}^{0s(e)}$ and $\Psi_{dr|s}^{0s(e)}$ are respectively defined in (4.39) and (4.40) with $\boldsymbol{\theta}$ replaced by $\hat{\boldsymbol{\theta}}$.
5. Create the vector $\mathbf{w}_{dr|s}^{0s}$ as follows

$$\mathbf{w}_{dr|s}^{0s} = b_{01} \mathbf{1}_{N_{dr}} \mathbf{1}_{n_d}^T \mathbf{v}_{dr|s}^{0s(e)} \quad (4.41)$$

where b_{01} is defined in (4.37) with $\boldsymbol{\theta}$ replaced by $\hat{\boldsymbol{\theta}}$.

6. The element dj of the vector $y_{dr|s}$ is:

$$(y_{dr|s})_{dj} = (\hat{\boldsymbol{\mu}}_{dr|s})_j + v_{dj}^{11} + v_d^{10} + \frac{\hat{\lambda}_e \hat{\sigma}_e}{1 + \hat{\lambda}_e^2} |v_{dj}^{0r}| + (w_{dr|s}^{0s})_j \quad (4.42)$$

$$\text{where } (\hat{\boldsymbol{\mu}}_{dr|s})_j = \mathbf{X}_{drj} \hat{\boldsymbol{\beta}} + \hat{\mu}_e + \hat{\gamma}_{n_d} [\mathbf{1}_{n_d}^T (\mathbf{y}_{ds} - \hat{\boldsymbol{\mu}}_{ds})]$$

4.3.4 Special Case 2: u_d follows SN and e_{dj} follows normal

The conditional distribution for this special case is provided by (4.14). The transformation (4.21) simplifies to

$$\mathbf{Y}_{dr|s} = \boldsymbol{\mu}_{dr|s} + b_{02} \mathbf{1}_{N_{dr}} v_{dr|s}^{0s(u)} + \mathbf{V}_{dr|s}^1 \quad (4.43)$$

where $b_{02} = \frac{\gamma_{n_d} \lambda_u \sigma_u \sigma_e^2}{\sigma_u^2 + \gamma_{n_d} \lambda_u^2 \sigma_e^2}$ and

$$v_{dr|s}^{0s(u)} = LTN_1 \left(\frac{\lambda_u}{\sigma_u} \gamma_{n_d} \mathbf{1}_{n_d}^T; 0, 1 + \lambda_u^2 \left(\frac{\sigma_e}{\sigma_u} \right)^2 \gamma_{n_d} \right) \quad (4.44)$$

$$\mathbf{V}_{dr|s}^1 = N_{N_{dr}} (0, \alpha \mathbf{I}_{N_{dr}} + \beta \mathbf{1}_{N_{dr}} \mathbf{1}_{N_{dr}}^T) \quad (4.45)$$

with α and β defined as follows

$$\alpha = \sigma_e^2 \quad \text{and} \quad \beta = \frac{\gamma_{n_d} \sigma_u^2 \sigma_e^2}{\sigma_u^2 + \gamma_{n_d} \lambda_u^2 \sigma_e^2} \quad (4.46)$$

In summary the quasi-univariate procedure below shows how to generate the predicted value $(y_{dr|s})_{dj}$ for the non-sampled unit j in small area d :

1. Estimate the unknown parameters $\boldsymbol{\theta}$ of the model using a suitable method and the sample data $\mathbf{y}_s$.
2. Generate values v_{dj}^{11} from $N(0, \alpha)$ and v_d^{10} from $N(0, \beta)$ independently where

α and β are defined in (4.46) with $\boldsymbol{\theta}$ replaced by $\hat{\boldsymbol{\theta}}$. Note that only one v_d^{10} is generated for each small area d but a separate v_{dj}^{11} is drawn for each unit j in small area d .

3. Generate value $v_{dr|s}^{0s(u)}$ from $LTN_1\left(\frac{\lambda_u}{\sigma_u}\gamma_{n_d}\mathbf{1}_{n_d}^T; 0, 1 + \lambda_u^2\left(\frac{\sigma_e}{\sigma_u}\right)^2\gamma_{n_d}\right)$ with $\boldsymbol{\theta}$ replaced by $\hat{\boldsymbol{\theta}}$.

4. The element j of the vector $y_{dr|s}$ is:

$$(y_{dr|s})_j = (\hat{\boldsymbol{\mu}}_{dr|s})_{dj} + v_{dj}^{11} + v_d^{10} + v_{dr|s}^{0s(u)} \quad (4.47)$$

$$\text{where } (\hat{\boldsymbol{\mu}}_{dr|s})_j = \mathbf{X}_{drj}\hat{\boldsymbol{\beta}} + \hat{\mu}_u + \hat{\gamma}_{n_d} [\mathbf{1}_{n_d}^T (\mathbf{y}_{ds} - \hat{\boldsymbol{\mu}}_{ds})]$$

4.4 Conditional Best Prediction for Complex Parameters

In Section 4.3, we developed the best predictor following Molina-Rao approach which consists of finding the conditional distribution of the out-of-sample vector of interest $\mathbf{Y}_{dr}$ given the sampled data $\mathbf{y}_{ds}$. When the errors follow SN distributions, this conditional distribution is fairly complex and it requires using an approximation procedure such as Gibbs sampling to generate values from the predictor distribution. A partial multivariate generation is necessary to obtain the predictor even though the multivariate approach is necessary only for a manageable vector of size $(n_d + 1)$. In this section, we propose a simpler conditional approach for the best prediction.

In SAE, the goal is to estimate the conditional parameter or area specific $\eta(u_d) = \mathbb{E}(h(\mathbf{Y}_d)|\mathbf{y}_{ds}, u_d, \boldsymbol{\theta})$ because only one population is “realized” and prediction is conducted conditionally on the given fixed population. The area effect is assumed to be random but for the given fixed population of interest a fixed number of

areas are involved. They can be interpreted as a realization set of areas from a much larger number of areas. The conditional distribution of $\mathbf{Y}_d$ given u_d is:

$$\mathbf{Y}_d|u_d \sim CSN_{N_d, N_d} \left(\mathbf{X}_d \boldsymbol{\beta} + u_d \mathbf{1}_{N_d} + \boldsymbol{\mu}_{ed}, \sigma_e^2 \mathbf{I}_{N_d}, \frac{\lambda_e}{\sigma_e} \mathbf{I}_{N_d}, \mathbf{0}_{N_d}, \mathbf{I}_{N_d} \right) \quad (4.48)$$

Note that the element of the vector $\mathbf{Y}_d|u_d$ are independent and we may write

$$(\mathbf{Y}_d|u_d)_j \sim SN \left(\mathbf{X}_{dj} \boldsymbol{\beta} + u_d + \mu_e, \sigma_e^2, \frac{\lambda_e}{\sigma_e} \right), \quad j \in s_d^c \quad (4.49)$$

where s_d^c indicates the out-of-sample units. The distribution in (4.49) is the distribution of the unit level error e_{dj} with a different location parameter. Therefore generating the out-of-sample values $(\mathbf{y}_{dr|s})_j = (\mathbf{y}_d|u_d)_j$, where $j \in s_d^c$, for the predicted census reduces to drawing from a univariate SN distribution. Unfortunately u_d is not observed. Therefore u_d is predicted by the best estimator $\hat{u}_d^B$ or the BLUP estimator $\hat{u}_d^{BLUP}$ and used to obtain predictions of $\eta(u_d)$ as respectively $\eta(\hat{u}_d^B)$ and $\eta(\hat{u}_d^{BLUP})$. In practice, the parameters of the model are estimated using a suitable method to obtain EB estimator $\hat{u}_d^{EB}$ and BLUP estimator $\hat{u}_d^{EBLUP}$, this leads to the empirical predictors $\hat{\eta}(\hat{u}_d^{EB})$ and $\hat{\eta}(\hat{u}_d^{EBLUP})$. We refer to these latter estimators as respectively $\hat{\eta}_d^{C-EQB}$ and $\hat{\eta}_d^{C-EQBLUP}$, where the C refers to conditional. The superscript $C - EQBLUP$ is a little confusing because $\hat{\eta}_d^{C-EQBLUP}$ is not linear nor unbiased; the subscript just refers to the way the area effect was estimated to distinguish between the two estimators. In practice, $\hat{\eta}_d^{C-EQB}$ should be preferred between the two because it approximates the conditional best predictor.

The univariate generation of the predicted value $(\mathbf{y}_{dr|s})_{dj}$ for the non-sampled unit j in small area d can be summarized as follows:

1. Estimate the unknown parameters $\boldsymbol{\theta}$ of the model using a suitable method and the sample data $\mathbf{y}_s$.

2. Predict the random effect u_d say $\hat{u}_d$ using the sample data $\mathbf{y}_s$ with $\boldsymbol{\theta}$ replaced by its estimator $\hat{\boldsymbol{\theta}}$.
3. Generate independently values v_{dj}^1 and v_{dj}^{0r} from $N(0, 1)$.
4. The element dj of the predictor $y_{dr|s}$ is:

$$(y_{dr|s})_j = \mathbf{X}_{drj}\hat{\boldsymbol{\beta}} + \hat{u}_d + \hat{\mu}_e + \hat{\sigma}_e(1 - \hat{\delta}_e^2)^{1/2}v_{dj}^1 + \hat{\sigma}_e\hat{\delta}_e|v_{dj}^{0r}| \quad (4.50)$$

To get $\hat{\eta}_d^{C-EQB}$ (respectively $\hat{\eta}_d^{C-EQBLUP}$), predict u_d using $\hat{u}_d^{EB}$ (respectively $\hat{u}_d^{EBLUP}$). Expression of $\hat{u}_d^{EB}$ is provided by Theorem 4.1 and $\hat{u}_d^{EBLUP}$ is given by (4.4).

An interesting question is; “how the best predictor described in Section 4.3 compares to the conditional best predictor discussed in this section?” To answer this question, it is essential to understand the difference between $\mathbf{Y}_{dr}|\mathbf{y}_{ds}$ and $\mathbf{Y}_{dr}|u_d$.

Proposition 4.3. *Under the one-fold SN model (3.16), we have that*

$$\begin{aligned} Y_{dj}|\mathbf{y}_{ds} &\stackrel{d}{=} (\mathbf{X}_{dj}\boldsymbol{\beta} + u_d + e_{dj})|\mathbf{y}_{ds} \\ &\stackrel{d}{=} \mathbf{X}_{dj}\boldsymbol{\beta} + u_d|\mathbf{y}_{ds} + e_{dj}, \quad j \in s_d^c \end{aligned} \quad (4.51)$$

where $\stackrel{d}{=}$ means same distribution and s_d^c indicates the out-of-sample units. Also, this result holds for the one-fold normal model as a special case of the one-fold SN model.

The proof of Proposition 4.3 is given in Appendix 4.11. The latter equality (4.51) shows that drawing from the conditional distribution of $Y_{drj}|\mathbf{y}_{ds}$ is equivalent to drawing independently two values from respectively the distributions of $u_d|\mathbf{y}_{ds}$ and e_{drj} and add the fixed term $\mathbf{X}_{drj}\boldsymbol{\beta}$. Similarly, from the conditional approach, we may write:

$$Y_{drj}|u_d \stackrel{d}{=} \mathbf{X}_{drj}\boldsymbol{\beta} + u_d + e_{drj}$$

Hence, the difference between the best prediction and the conditional best prediction is the estimation of the area effect u_d across the L censuses. The best prediction approach draws different $u_d^{(\ell)}$'s from the distribution of $u_d|\mathbf{y}_{ds}$, where $\ell = 1, \dots, L$. For a large L , the average of the area effects $u_d^{(\ell)}$ is approximately $\mathbb{E}(u_d|\mathbf{y}_{ds})$. On the contrary, the conditional best prediction uses the same value $\mathbb{E}(u_d|\mathbf{y}_{ds})$ across the L censuses. Drawing from the distribution of $u_d|\mathbf{y}_{ds}$ introduces additional variability for the best predictor comparatively to the conditional best predictor. The additional variability corresponds to $\text{Var}(u_d|\mathbf{y}_{ds})$. For the nested error model with the errors following SN distributions (3.16), we have:

$$\text{Var}(u_d|\mathbf{y}_{ds}) = \sigma_e^2 \gamma_{n_d} + \frac{\Phi_{n_d+1}^{(2)}(\mathbf{0}; \boldsymbol{\nu}_A, \boldsymbol{\Gamma})}{\Phi_{n_d+1}(\mathbf{0}; \boldsymbol{\nu}_A, \boldsymbol{\Gamma})} - \left(\frac{\Phi_{n_d+1}^{(1)}(\mathbf{0}; \boldsymbol{\nu}_A, \boldsymbol{\Gamma})}{\Phi_{n_d+1}(\mathbf{0}; \boldsymbol{\nu}_A, \boldsymbol{\Gamma})} \right)^2 \quad (4.52)$$

where $\boldsymbol{\nu}_A$ and $\boldsymbol{\Gamma}$ as in theorem (4.1). The variance (4.52) follows directly from Corollary 2.1 applied to the distribution of $u_d|\mathbf{y}_{ds}$ defined in Proposition 4.2. Note that if $\lambda_u = \lambda_e = 0$, as for Molina and Rao (2010), then $\text{Var}(u_d|\mathbf{y}_{ds}) = \sigma_e^2 \gamma_{n_d}$. Consequently, when γ_{n_d} is very small i.e. σ_u^2 small relative to σ_e^2 (small intraclass correlation coefficient) or n_d is large, the performances of the best predictor and the conditional best predictor are equivalent in terms of variability. Inversely, when γ_{n_d} is not small, the conditional best predictor outperforms the best predictor.

The question of which estimator between the best predictor and the conditional best predictor is more appropriate reduces to deciding whether the area effects across the L censuses should be estimated by drawing from their conditional distribution or by using the conditional expectation. Personally I prefer the latter approach for two reasons. First, the value of u_d is fixed for a given area d . Secondly, the L predicted censuses are only used to capture the distribution of the characteristics of interest within a small area d i.e. the distribution at the unit level for a fixed realized u_d .

4.5 Prediction Based on the Half-Normal Representation for Complex Parameters

The essence of this approach is the half-normal representation (2.2) of the SN distribution. The nested error model can be written in the conditional form as follows:

$$Y_{dj}|t_{dj}, t_{d0} = \mathbf{x}_{dj}^T \boldsymbol{\beta} + \mu_u + \sigma_u \delta_u t_{d0} + (1 - \delta_u^2)^{1/2} u_d^* + \mu_e + \sigma_e \delta_e t_{dj} + (1 - \delta_e^2)^{1/2} e_{dj}^*, \quad (4.53)$$

where:

1. $\delta_u = \frac{\lambda_u}{\sqrt{1+\lambda_u^2}}$ and $\delta_e = \frac{\lambda_e}{\sqrt{1+\lambda_e^2}}$
2. $u_d^* \stackrel{iid}{\sim} N(0, \sigma_u^2)$ and $e_{dj}^* \stackrel{iid}{\sim} N(0, \sigma_e^2)$ are independent, $j = 1, \dots, N_d, d = 1, \dots, m$
3. $t_{dk} \stackrel{iid}{\sim} HN(0, 1)$, $k = 0, 1, \dots, N_d$ with HN denoting the half-normal distribution.

Consider u_d and e_{dj} defined as follows:

$$u_d \equiv \mu_u + \sigma_u \delta_u t_{d0} + (1 - \delta_u^2)^{1/2} u_d^* \sim SN(\mu_u, \sigma_u^2, \lambda_u) \quad (4.54)$$

$$e_{dj} \equiv \mu_e + \sigma_e \delta_e t_{dj} + (1 - \delta_e^2)^{1/2} e_{dj}^* \sim SN(\mu_e, \sigma_e^2, \lambda_e) \quad (4.55)$$

Expressions (4.54) and (4.55) show that the model (4.53) is equivalent to the SN model defined by (3.16). The multivariate distribution of $\mathbf{Y}_d|t_d, t_{d0}$ where $\mathbf{Y}_d = (Y_{d1}, \dots, Y_{dN_d})^T$ and $\mathbf{t}_d = (t_{d0}, t_{d1}, \dots, t_{dN_d})^T$ is $N_{N_d}(\boldsymbol{\mu}_d^{t_d}, \boldsymbol{\Psi}_d)$ with

$$\boldsymbol{\mu}_d^{t_d} = \mathbf{X}_{dj} \boldsymbol{\beta} + \mu_u \mathbf{1}_{N_d} + \sigma_u \delta_u t_{d0} \mathbf{1}_{N_d} + \boldsymbol{\mu}_e + \sigma_e \delta_e \begin{pmatrix} t_{d1} \\ \vdots \\ t_{dN_d} \end{pmatrix} \quad (4.56)$$

$$\boldsymbol{\Psi}_d = \sigma_e^2 (1 - \delta_e^2) \mathbf{I}_{N_d} + \sigma_u^2 (1 - \delta_u^2) \mathbf{1}_{N_d} \mathbf{1}_{N_d}^T \quad (4.57)$$

Consider the following decompositions:

$$\begin{aligned}\mathbf{t}_d &= \begin{pmatrix} \mathbf{t}_{dr} \\ \mathbf{t}_{ds} \end{pmatrix} = \begin{pmatrix} (t_{d0}, t_{d1}, \dots, t_{dN_r})^T \\ (t_{d0}, t_{dN_r+1}, \dots, t_{dN_d})^T \end{pmatrix}, \\ \boldsymbol{\mu}_d^{t_d} &= \begin{pmatrix} \boldsymbol{\mu}_{dr}^{t_{dr}} \\ \boldsymbol{\mu}_{ds}^{t_{ds}} \end{pmatrix}, \quad \boldsymbol{\Psi}_d = \begin{bmatrix} \boldsymbol{\Psi}_{drd} & \boldsymbol{\Psi}_{drs} \\ \boldsymbol{\Psi}_{dsr} & \boldsymbol{\Psi}_{dss} \end{bmatrix} \\ \mathbf{Y}_d &= \begin{pmatrix} \mathbf{Y}_{dr} \\ \mathbf{Y}_{ds} \end{pmatrix}, \text{ and } \mathbf{Y}_d^{t_d} = \begin{pmatrix} \mathbf{Y}_{dr}^{t_{dr}} \\ \mathbf{Y}_{ds}^{t_{ds}} \end{pmatrix} = \begin{pmatrix} \mathbf{Y}_{dr} | \mathbf{t}_{dr} \\ \mathbf{Y}_{ds} | \mathbf{t}_{ds} \end{pmatrix}\end{aligned}$$

Using the well known result on the conditional multivariate normal distribution, we find that

$$\mathbf{Y}_{dr|s}^{t_d} \equiv \mathbf{Y}_{dr}^{t_d} | \mathbf{Y}_{ds}, \mathbf{t}_{ds} \sim N_{N_{dr}} \left(\boldsymbol{\mu}_{dr|s}^{t_d}, \boldsymbol{\Psi}_{dr|s} \right) \quad (4.58)$$

where

$$\boldsymbol{\mu}_{dr|s}^{t_d} = \boldsymbol{\mu}_{dr}^{t_{dr}} + \frac{\sigma_u^2(1 - \delta_u^2)}{\sigma_e^2(1 - \delta_e^2) + n_d\sigma_u^2(1 - \delta_u^2)} \mathbf{1}_{N_{dr}} \mathbf{1}_{n_d}^T (\mathbf{y}_{ds} - \boldsymbol{\mu}_{ds}^{t_{ds}}) \quad (4.59)$$

$$\boldsymbol{\Psi}_{dr|s} = \sigma_e^2(1 - \delta_e^2) \left[\mathbf{I}_{N_{dr}} + \frac{\sigma_u^2(1 - \delta_u^2)}{\sigma_e^2(1 - \delta_e^2) + n_d\sigma_u^2(1 - \delta_u^2)} \mathbf{1}_{N_{dr}} \mathbf{1}_{N_{dr}}^T \right] \quad (4.60)$$

Note that only the mean vector $\boldsymbol{\mu}_{dr|s}^{t_d}$ is a function of the truncated normal $\mathbf{t}_d$; the covariance matrix is free of $\mathbf{t}_d$. The conditional predictor (4.58) is easily generated using a univariate manner in the following way:

1. Generate t_{dj} from the standard half-normal $HN(0, 1)$, $j = 1, \dots, N_d$
2. Generate values v_{dj}^1 from $N(0, \alpha)$ and v_d^0 from $N(0, \beta)$ independently where

$$\alpha = \sigma_e^2 (1 - \delta_e^2) \quad (4.61)$$

$$\beta = \sigma_e^2 (1 - \delta_e^2) \frac{\sigma_u^2 (1 - \delta_u^2)}{\sigma_e^2 (1 - \delta_e^2) + n_d \sigma_u^2 (1 - \delta_u^2)} \quad (4.62)$$

Note that only one v_d^0 is generated for each small area d but a separate v_{dj}^1 is drawn for each unit j in small area d .

3. The element dj of the vector $y_{dr|s}^{t_d}$ is:

$$(y_{dr|s}^{t_d})_j = (\boldsymbol{\mu}_{dr|s}^{t_d})_j + v_{dj}^1 + v_d^0 \quad (4.63)$$

The difficulty of this conditional approach is the fact that the predictor (4.58) is a function of the unobserved variables $t_{dk}, k = 0, 1, \dots, N_d$. A natural approach to generate from (4.58) would be to first predict $\mathbf{t}_d$ given $\mathbf{y}_d$. Predicting $\mathbf{t}_{ds}$ conditionally on $\mathbf{y}_{ds}$ is possible. However, because $\mathbf{Y}_{dr}$ is also unobserved, conditioning on $\mathbf{y}_{dr}$ is not helpful for predicting $\mathbf{t}_{dr}$. Instead we can use results on conditional expectation to approximate the predictor. The first result is the following:

$$\mathbb{E}_{\mathbf{t}_d}(\mathbb{E}(\eta_d|\mathbf{y}_{ds}, \mathbf{t}_d)) = \mathbb{E}(\eta_d|\mathbf{y}_{ds}). \quad (4.64)$$

Therefore it is possible to predict the complex parameter by the conditional expectation on both $\mathbf{y}_{ds}$ and $\mathbf{t}_d$. Analytical expressions for this integral are not available due to the complexity of the parameters. Here we propose a Monte Carlo approach to approximate the conditional expectation. Instead of using the best predicted value for t_{dj} , we draw t_{dj} from $HN(0, 1)$; this approach is less efficient (in terms of MSE) than the best prediction method presented in section 4.2 since t_{dj} is not optimal. The steps of this first conditional approach are as follows:

1. Estimate the unknown parameters $\boldsymbol{\theta}$ of the model using a suitable method and the sample data $\mathbf{y}_s$.
2. Generate $t_{dj}^{(\ell)}$ from the standard half-normal $HN(0, 1)$, $j = 1, \dots, N_d$ and $\ell = 1, \dots, L$.
3. Generate out-of-sample predicted vectors $\mathbf{y}_{dr}^{(\ell)}$ using distribution in (4.58), $\ell = 1, \dots, L$ with $\boldsymbol{\theta}$ replaced by its estimator $\hat{\boldsymbol{\theta}}$. The univariate approach,

for generating the predictor described above in this subsection, should be used for large populations.

4. Augment each of the generated vector $\mathbf{y}_{dr}^{(\ell)}$ with the sample data to obtain $\mathbf{y}_d^{(\ell)} = (\mathbf{y}_{dr}^{(\ell)T}, \mathbf{y}_{ds}^T)^T$.
5. Obtain an approximation to the empirical best (EB) estimator as follows:

$$\hat{\eta}_d^{SC-HN} = \frac{1}{L} \sum_{\ell=1}^L \hat{\eta}_d^{(\ell)}(\hat{\boldsymbol{\theta}}) = \frac{1}{L} \sum_{\ell=1}^L h(\mathbf{y}_d^{(\ell)}) \quad (4.65)$$

We refer to the estimator (4.65) as the simple conditional half-normal (SC-HN) predictor.

The second approach for estimating the predictor uses the following result:

$$\mathbb{E}(\mathbb{E}(\eta_d | \mathbf{y}_{ds}, \mathbf{t}_d) | \mathbf{y}_{ds}) = \mathbb{E}(\eta_d | \mathbf{y}_{ds}). \quad (4.66)$$

Result (4.66) is a well known property of the conditional expectations. This approximation is more complex because it requires a double conditioning; it is also more variable than the best predictors because we draw values t_{dj} instead of using the best predictor of t_{dj} . However, it is easy to implement a Monte Carlo procedure as follows:

1. Estimate the unknown parameters $\boldsymbol{\theta}$ of the model using a suitable method and the sample data $\mathbf{y}_s$.
2. Generate $t_{dj}^{(\ell)}$ from the standard half-normal $HN(0, 1)$, $j = 1, \dots, N_d$ and $\ell = 1, \dots, L$.
3. For each $\mathbf{t}_d^{(\ell)}$, generate B_ℓ out-of-sample predicted vectors $\mathbf{y}_{dr}^{(\ell, b)}$ using distribution in (4.58), $\ell = 1, \dots, L$ and $b = 1, \dots, B_\ell$ with $\boldsymbol{\theta}$ replaced by its estimator $\hat{\boldsymbol{\theta}}$.

The univariate approach, for generating the predictor described above in this subsection, should be used for large populations.

4. Augment each of the generated vector $\mathbf{y}_{dr}^{(\ell,b)}$ with the sample data to obtain $\mathbf{y}_d^{(\ell,b)} = (\mathbf{y}_{dr}^{(\ell,b)T}, \mathbf{y}_{ds}^T)^T$.
5. Obtain an approximation to the empirical best (EB) estimator as follows:

$$\hat{\eta}_d^{DC-HN} = \frac{1}{L} \sum_{\ell=1}^L \frac{1}{B_\ell} \sum_{b=1}^{B_\ell} \hat{\eta}_d^{(\ell,b)}(\hat{\boldsymbol{\theta}}) = \frac{1}{L} \sum_{\ell=1}^L \frac{1}{B_\ell} \sum_{b=1}^{B_\ell} h(\mathbf{y}_d^{(\ell,b)}) \quad (4.67)$$

We refer to the estimator (4.67) as the double conditional half-normal (DC-HN) predictor.

4.6 Normality Assumption: Molina-Rao Method

Consider the nested error model (1.5) and assume that the distributions f_u and f_e in (1.6) are respectively $N(0, \sigma_u^2)$ and $N(0, \sigma_e^2)$. Note that this model is a special case of the SN model (3.16) where $\lambda_u = \lambda_e = 0$. Then the vectors $\mathbf{Y}_d$, $d = 1, \dots, m$ are independent with

$$\begin{aligned} \mathbf{Y}_d &\sim N_{N_d}(\boldsymbol{\mu}_d, \boldsymbol{\Sigma}_d), \quad \text{where} \\ \boldsymbol{\mu}_d &= \mathbf{X}_d \boldsymbol{\beta} \quad \text{and} \quad \boldsymbol{\Sigma}_d = \sigma_e^2 \mathbf{I}_{N_d} + \sigma_u^2 \mathbf{1}_{N_d} \mathbf{1}_{N_d}^T \end{aligned} \quad (4.68)$$

Using the same partitions as before and from the distribution of $\mathbf{Y}_d$ shown in (4.68), it is easy to obtain the conditional distribution of $\mathbf{Y}_{dr} | \mathbf{y}_{ds}$:

$$\mathbf{Y}_{dr} | \mathbf{y}_{ds} \sim N_{N_{dr}}(\boldsymbol{\mu}_{dr|s}, \boldsymbol{\Sigma}_{dr|s}), \quad (4.69)$$

where

$$\begin{aligned}\boldsymbol{\mu}_{dr|s} &= \mathbf{X}_{dr}\boldsymbol{\beta} + (\gamma_{n_d}\mathbf{1}_{n_d}^T(\mathbf{y}_{ds} - \mathbf{X}_{ds}\boldsymbol{\beta}))\mathbf{1}_{N_d-n_d} \\ \boldsymbol{\Sigma}_{dr|s} &= \sigma_e^2\mathbf{I}_{N_d-n_d} + \sigma_u^2(1 - n_d\gamma_{n_d})\mathbf{1}_{N_d-n_d}\mathbf{1}_{N_d-n_d}^T\end{aligned}$$

The multivariate form (4.69) of the predictor is a major limitation for many applications that requires generating values for large censuses. In fact, it can be very challenging and sometimes impossible to generate data for millions of records in multivariate fashion using current technology. Fortunately, one important feature of the MR model is the fact that the multivariate conditional distribution (4.69) is equivalent to the following univariate model:

$$Y_{dj} = \boldsymbol{\mu}_{dr|s} + u_d + \epsilon_{drj}, \quad j \in r_d, \quad (4.70)$$

where u_d and ϵ_{drj} are independent, $u_d \stackrel{iid}{\sim} N(0, \sigma_u^2(1 - n_d\gamma_{n_d}))$, $\epsilon_{drj} \stackrel{iid}{\sim} N(0, \sigma_e^2)$. Model (4.70) ensures that the predicted values can always be generated from univariate normal distributions and therefore the Molina-Rao method can be applied to a population of any size. Note that the predictor defined by (4.69) is obtained from the more general expression (4.21) by letting $\lambda_e = \lambda_u = 0$. The steps for estimating the complex parameter are the follow:

1. Estimate the unknown parameter $\boldsymbol{\theta}$ using the sample data $\mathbf{y}_s$ and a suitable method such as ML estimation to obtain $\hat{\boldsymbol{\theta}}$. Because the random components are assumed to follow normal distributions, there are numerous parameter estimation techniques available and all statistical software packages offer at least ML and REML options for estimating parameters of LMMs.
2. Draw L out-of-sample $\mathbf{y}_{dr}^{(\ell)}$, $\ell = 1, \dots, L$ from the conditional distribution (4.70) using the estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$.

3. Augment each of the L generated vector $\mathbf{y}_{dr}^{(\ell)}$ with the sample data to obtain the predicted census $\mathbf{y}_d^{(\ell)} = (\mathbf{y}_{dr}^{(\ell)T}, \mathbf{y}_{ds}^T)^T$.
4. Obtain an estimate of the complex parameter as follows:

$$\hat{\eta}_d^{MR-N} = \frac{1}{L} \sum_{\ell=1}^L \hat{\eta}_d^{(\ell)}(\hat{\boldsymbol{\theta}}) = \frac{1}{L} \sum_{\ell=1}^L h(\mathbf{y}_d^{(\ell)}) \quad (4.71)$$

The estimator (4.71) is the MR-N predictor meaning it follows the Molina-Rao normality-based approach by assuming that the errors are normality distributed. This estimator is not optimal when the errors follow a SN distribution and as shown in the simulations it is very sensitive to the departure of the unit level errors' (e_{dj} 's) distribution from normality.

4.7 Nonparametric Approach: ELL Method

The World Bank produces estimates of poverty and inequality measures for developing countries. Household surveys collecting income and consumption related information are used to estimate poverty and inequality measures for areas with large enough effective sample size. However these estimates are not reliable at the small area level. Elbers et al. (2003) proposed a clever and straightforward approach for obtaining estimates at the small area level; this approach is called the ELL method in this document. Their idea is to use the sample data residuals to estimate the distribution of the random components of the one-fold nested error model. Then use the estimated distribution of $\mathbf{y}$ to predict a larger sample or the census. The complex parameter η_d can therefore be estimated using the predicted census for any small area d .

4.7.1 Traditional ELL Predictor

Let us assume that $\mathbf{y}$ is defined by the nested error model (1.5). The ELL method does not require the assumption of any distribution for the distribution of the random components i.e. no assumption for f_u and f_e in (1.6). The approach consists of drawing from the empirical area and unit level residuals to reconstitute the census data. Different variants of the ELL method have been described in the literature. In this thesis, the unit level error variances are assumed homogeneous. The simplicity of the ELL method and the fact that it can be applied to any data without assumption on the errors' distribution are very attractive. The steps of the ELL method can be summarized as follows:

1. From the nested error model (1.5), calculate the total residuals $\hat{r}_{dj} = y_{dj} - \mathbf{x}_{dj}^T \hat{\boldsymbol{\beta}}_{OLS}$ where $\hat{\boldsymbol{\beta}}_{OLS}$ is the ordinary least square (OLS) estimate of $\boldsymbol{\beta}$.
2. The effect of small area d , u_d , is estimated as the empirical mean value of the total residuals $\hat{r}_{dj}$ over all the observations from the small area d .

$$\hat{u}_d = \frac{1}{n_d} \sum_{j=1}^{n_d} \hat{r}_{dj} \quad (4.72)$$

3. The unit level residuals e_{dj} are estimated as follows:

$$\hat{e}_{dj} = \hat{r}_{dj} - \hat{u}_d \quad (4.73)$$

These residuals are then mean-corrected to sum to zero across the small area, d .

4. Draw $\hat{\boldsymbol{\beta}}^{(\ell)}$, $\hat{u}_d^{(\ell)}$, and $\hat{e}_{dj}^{(\ell)}$, $\ell = 1, \dots, L$ from respectively $N\left(\hat{\boldsymbol{\beta}}_{OLS}, \text{Cov}(\hat{\boldsymbol{\beta}}_{OLS})\right)$, the empirical distributions of $\hat{u}_d$, and $\hat{e}_{dj}$.
5. Construct L predictors $y_{dj}^{(\ell)}$ as follows:

$$y_{dj}^{(\ell)} = \mathbf{x}_{dj}^T \hat{\boldsymbol{\beta}}^{(\ell)} + \hat{u}_d^{(\ell)} + \hat{e}_{dj}^{(\ell)} \quad (4.74)$$

6. Get an estimate of the complex parameter as follows:

$$\hat{\eta}_d^{ELL-TRAD} = \frac{1}{L} \sum_{\ell=1}^L \hat{\eta}_d^{(\ell)} = \frac{1}{L} \sum_{\ell=1}^L h(\mathbf{y}_d^{(\ell)}) \quad (4.75)$$

We refer to the estimator (4.75) as the traditional ELL (ELL-TRAD) predictor. Unfortunately, this traditional ELL method does not provide a correct prediction of the area effects u_d across the L censuses. To see this, consider the small area mean $\eta_d = \bar{\mathbf{X}}_d^T \boldsymbol{\beta} + u_d$. The traditional ELL estimator is equal to:

$$\begin{aligned} \hat{\eta}_d^{ELL-TRAD} &= \frac{1}{L} \sum_{\ell=1}^L \left(\bar{\mathbf{X}}_d^T \hat{\boldsymbol{\beta}}^{(\ell)} + \hat{u}_d^{(\ell)} + \bar{e}_d^{(\ell)} \right) \\ &= \frac{1}{L} \sum_{\ell=1}^L \bar{\mathbf{X}}_d^T \hat{\boldsymbol{\beta}}^{(\ell)} + \frac{1}{L} \sum_{\ell=1}^L \hat{u}_d^{(\ell)} + \frac{1}{L} \sum_{\ell=1}^L \bar{e}_d^{(\ell)} \end{aligned}$$

As expected, the last term $\bar{e}_d^{(\ell)}$ is approximatively equal to 0. However, the area effect $\hat{u}_d^{(\ell)}$ is selected from the empirical area level residuals $\hat{u}_d$ as shown in (4.72) that is $\frac{1}{L} \sum_{\ell=1}^L \hat{u}_d^{(\ell)} \approx 0$. Therefore, the traditional ELL predictor reduces approximately to the synthetic estimator $\frac{1}{L} \sum_{\ell=1}^L \left(\bar{\mathbf{X}}_d^T \hat{\boldsymbol{\beta}}^{(\ell)} \right)$.

If the parameter $\eta_d = h(\mathbf{y}_d)$ is complex, the traditional ELL predictor may not reduce to the synthetic estimator given the complexity of h . However each draw ℓ selects a different value $\hat{u}_d^{(\ell)}$ from the empirical area level residuals with mean approximately equals to 0. Across the L prediction cycles, the traditional ELL method does not attach a specific estimated random effect (area level residual) to the small area d . Instead the traditional ELL method uses a combination of estimated random effects from other areas to predict the complex parameter for small area d .

In the next section, we propose two modifications to the traditional ELL method. These modifications aim to attach a specific predicted random effect to the small areas

and reduce the MSE by not drawing a different area effect at each prediction cycle l .

4.7.2 Modifications to the Traditional ELL Predictor

Empirical studies of the traditional ELL method by Molina and Rao (2010) have shown very high MSEs compared to the MR estimator and even to the direct estimator under the normal model. In the simulation study in section 4.9 same results as in Molina and Rao (2010) are observed. In order to reduce the MSE, we propose two nonparametric adjustments to the original ELL method. The traditional ELL approach has two main problems:

1. The area effects are wrongly assigned to the small areas across the L censuses. In fact, the random area effect cancels out for linear parameters since the empirical average converges to zero ($\mathbb{E}(u_d) = 0$).
2. The variability of the empirical ELL is increased by drawing L different values from the empirical distributions of $N\left(\hat{\beta}_{OLS}, \text{Cov}(\hat{\beta}_{OLS})\right)$ and the area effects residuals $\hat{u}_d^{(\ell)}$, $\ell = 1, \dots, L$, for the same area given a fixed sample.

Therefore, to address those two issues, we estimate the fixed and the random area effects using the sample data. Then, for the given sample, the intra-area distribution (conditional on the area effects) is estimated by drawing from the unit level residuals. The fixed effects are estimated using OLS as previously; however we use two different nonparametric methods to obtain the area effect estimates.

The first method for estimating the area effects is the same as in the traditional approach. This consists of obtaining the area effects from the area level residuals. For a given sample, we do not bootstrap the area level residuals. The steps of the algorithm are as follows:

1. From the nested error model (1.5), estimate the fixed effects β using OLS.

2. Once $\hat{\beta}_{OLS}$ has been obtained, the effect of small area d is estimated as the empirical mean value of the total residuals $\hat{r}_{dj}$ over all the observations from the small area d .

$$\hat{u}_d = \frac{1}{n_d} \sum_{j=1}^{n_d} \hat{r}_{dj} \quad (4.76)$$

3. The unit level residuals e_{dj} are estimated as follows:

$$\hat{e}_{dj} = \hat{r}_{dj} - \hat{u}_d \quad (4.77)$$

These residuals are then mean-corrected to sum to zero across the small area.

4. Draw $\hat{e}_{dj}^{(\ell)}$, $\ell = 1, \dots, L$ from the empirical distribution of the $\hat{e}_{dj}$.
5. Construct L predictors $y_{dj}^{(\ell)}$ as follows:

$$y_{dj}^{(\ell)} = \mathbf{x}_{dj}^T \hat{\beta}_{OLS} + \hat{u}_d + \hat{e}_{dj}^{(\ell)} \quad (4.78)$$

Note that the same area level residual $\hat{u}_d$ is used for a given area d .

6. Obtain an estimate of the complex parameter as follows:

$$\hat{\eta}_d^{ELL-RES} = \frac{1}{L} \sum_{\ell=1}^L \hat{\eta}_d^{(\ell)} = \frac{1}{L} \sum_{\ell=1}^L h(\mathbf{y}_d^{(\ell)}) \quad (4.79)$$

The second method for estimating the area effects consists of using a combination of OLS and the method of moments to get estimators of σ_u^2 and σ_e^2 (Fuller and Battese (1973)). Then use the estimated variance components to get predictors of the area effects. Note that, as for the first method, no distributional assumption is needed. The steps of the algorithm are as follows:

1. Obtain an estimate of β using OLS.

2. Estimate σ_e^2 then σ_u^2 as follows:

$$\hat{\sigma}_e^2 = \frac{SSE(1)}{n - m - p_1}, \quad p_1 = \text{number of non-zero X-derivations} \quad (4.80)$$

where $SSE(1)$ is the residual sum of squares obtain by regressing $(y_{dj} - \bar{y}_{d.})$ on the non-zero X-derivations $(x_{dj} - \bar{x}_{d.})$ for areas with $n_d > 1$ with $\bar{y}_{d.} = \sum_{j=1}^{n_d} y_{dj}/n_d$ and $\bar{x}_{d.} = \sum_{j=1}^{n_d} x_{dj}/n_d$. Also, we have

$$\hat{\sigma}_{um}^2 = \frac{SSE(2) - (n - p)\hat{\sigma}_e^2}{\sum_{j=1}^{n_d} n_d \{1 - n_d \bar{x}_{d.} \left(\sum_{d=1}^m \sum_{j=1}^{n_d} x_{dj} x_{dj}^T \right)^{-1} \bar{x}_{d.}^T \}}, \quad (4.81)$$

where $SSE(2)$ is the residual sum of squares obtain by regressing y_{dj} on the non-zero x-derivations x_{dj} . Because the estimates can be negative, we truncate to get:

$$\hat{\sigma}_u^2 = \max(\hat{\sigma}_{um}^2, 0). \quad (4.82)$$

3. Compute the estimated small area effect $\hat{u}_d$ as follows:

$$\hat{u}_d = \hat{\sigma}_u^2 \mathbf{1}_{n_d}^T \hat{\mathbf{V}}_d^{-1} \left(\mathbf{y}_d - \mathbf{X}_d \hat{\boldsymbol{\beta}}_{OLS} \right) \quad (4.83)$$

where $\hat{\mathbf{V}}_d = (\hat{\sigma}_e^2 \mathbf{I}_{n_d} + \hat{\sigma}_u^2 \mathbf{1}_{n_d} \mathbf{1}_{n_d}^T)$.

4. Obtain unit level residuals as follows:

$$\tilde{e}_{dj} = y_{dj} - \mathbf{x}_{dj}^T \hat{\boldsymbol{\beta}}_{OLS} - \hat{u}_d \quad (4.84)$$

Adjust these residuals to sum to zero and obtain $\hat{e}_{dj}$. Draw $\hat{e}_{dj}^{(\ell)}$ from the empirical distribution of the $\hat{e}_{dj}$.

5. Construct L predictors $\hat{y}_{dj}^{(\ell)}$ as follows:

$$\hat{y}_{dj}^{(\ell)} = \mathbf{x}^T \hat{\boldsymbol{\beta}}_{OLS} + \hat{u}_d + \hat{e}_{dj}^{(\ell)}, \quad b = 1, \dots, B. \quad (4.85)$$

Note that the same area level residual $\hat{u}_d$ is used for a given area d .

6. Obtain an estimate of the complex parameter as follows:

$$\hat{\eta}_d^{ELL-MOM} = \frac{1}{L} \sum_{\ell=1}^L \hat{\eta}_d^{(\ell)} = \frac{1}{L} \sum_{\ell=1}^L h(\mathbf{y}_d^{(\ell)}) \quad (4.86)$$

4.8 Prediction for Non Sampled Areas and Non Linkable samples

A common issue in SAE is when some small areas have no sample/respondent. For a given not sampled area, conditioning on the total sample gives no information for predicting the area effect. For linear parameter, the customary approach is to use the synthetic part of the predictor, $\mathbf{X}_d \hat{\boldsymbol{\beta}}$, as the small area estimate. The analogue for estimating complex parameters would be to predict y_{dj} by $\mathbf{X}_d \hat{\boldsymbol{\beta}} + e_{dj}^*$ where e_{dj}^* is generated from the distribution of e_{dj} . This estimator preserves the within-area distribution however it results in a biased predictor of $\mathbf{X}_d \boldsymbol{\beta} + u_d$ since the area effect is essentially estimated to be 0. This approach is reasonable when the number of small areas with no sample is relatively small. Otherwise, extra efforts (pilot surveys, inexpensive surveys, phase-sampling, use of other surveys, administrative data, etc) should be taken to improve the coverage of the survey.

A second issue arise when it is not possible to link the sample data to the census auxiliary data then the best prediction described in Section 4.3 is not applicable. The reason is the conditional distribution requires knowing the partition of the auxiliary

variables $\begin{bmatrix} \mathbf{X}_{dr} \\ \mathbf{X}_{ds} \end{bmatrix}$ between the out-of-sample units and the sample. However the conditional method discussed in Section 4.4 is applicable. Indeed, the only use of the sample is to estimate the area effects u_d and the parameters of the model. Given that the small areas are clearly identifiable from the survey, one would use the survey to obtain $\hat{u}_d^{EB}$ (or $\hat{u}_d^{EBLUP}$) then use $\mathbf{X}_d\hat{\boldsymbol{\beta}} + \hat{u}_d^{EB} + e_{dj}^*$ (or $\mathbf{X}_d\hat{\boldsymbol{\beta}} + \hat{u}_d^{EBLUP} + e_{dj}^*$), where e_{dj}^* is generated from the distribution of e_{dj} , to predict y_{dj} . This is actually very similar to the Census EB discussed by Correa (2012).

4.9 Simulation Study using FGT Measures

The objectives of the following simulation are to study the relative performances of the predictors discussed in this chapter. FGT poverty measures are used as the complex parameters of interest.

4.9.1 FGT Poverty Measures

The parameters of interest considered in this simulation are the FGT poverty measures introduced by Foster et al. (1984). The FGT class is defined, for domain d , as:

$$F_{\alpha d} = \frac{1}{N_d} \sum_{j=1}^{N_d} \left(\frac{z - E_{dj}}{z} \right)^{\alpha} I(E_j < z), \quad j = 1, \dots, N_d, \alpha \geq 0, \quad (4.87)$$

where z is the poverty line, E_{dj} is a quantitative measure of welfare such as income or expenditure associated with individual j from domain d , and I is an indicator function. $I(E_j < z) = 1$ if $E_j < z$ meaning that the person j from area d is considered to be in poverty (welfare measure under poverty line) and similarly $I(E_j < z)$ is equal to 0 if $E_j \geq z$ (person j is not in poverty). The class $\alpha = 0$ yields the proportion of people in poverty for domain d and F_{0d} is called the poverty incidence. The statistic

resulting from $\alpha = 0$ is a simple area proportion and may not be considered as a complex parameter. The class $\alpha = 1$, called the poverty gap, uses the normalized gap $\frac{z-E_{dj}}{z}$ to differentiate among the poor. The poorer an individual is, the larger is their poverty gap; F_{1d} is the average poverty gap for area d . The class $\alpha = 2$, called poverty severity, squares the normalized gap and thus weights the gaps by the gaps. Therefore individuals with a larger poverty gap weigh more in the measure to indicate severity in poverty. As α tends to infinity the poorest of the poor weigh even more, however in practice, $\alpha > 2$ is not much used.

A direct estimator for small area d is obtained by using only the information in the sample associated to the given domain and is defined by:

$$\hat{F}_{\alpha d}^w = \frac{1}{\hat{N}_d} \sum_{j=1}^{n_d} w_{dj} F_{\alpha dj}, \quad j = 1, \dots, N_d, \alpha \geq 0, \quad (4.88)$$

where $F_{\alpha dj} = \left(\frac{z-E_{dj}}{z}\right)^\alpha I(E_j < z)$, n_d is the sample size for domain d , w_{dj} is the sampling weight for individual j from sampled domain d , and $\hat{N}_d = \sum_{j=1}^{n_d} w_{dj}$ is the design-unbiased estimator of population size N_d of the sampled domain d . Note that if simple random sampling (SRS) design is used then $w_{dj} = \frac{N_d}{n_d}$ is independent of unit j from area d . Therefore the direct estimator simplifies to

$$\hat{F}_{\alpha d}^{srs} = \frac{1}{n_d} \sum_{j=1}^{n_d} F_{\alpha dj}, \quad j = 1, \dots, n_d, \alpha \geq 0, \quad (4.89)$$

Unfortunately, when the effective sample size $n_d^* = n_d D_{eff}$ (D_{eff} is the design effect) is too small, the direct estimator is not precise enough to provide reliable estimates. Often, the coefficient of variation (CV), defined as the standard error of an estimate expressed as a ratio or a percent of the estimate, is used to decide whether an estimate is reliable or not. For instance, Statistics Canada follows the general rule which

considers an estimate with a coefficient of variation less than 15% to be reliable for general use while estimates with a coefficient of variation greater than 35% are deemed to be unreliable (unacceptable quality). Statistics Canada recommends not publishing unreliable estimates ($CV > 35\%$) and if published informing the public that the estimates are not reliable.

It is assumed that there exists a one-to-one transformation $Y_{dj} = T(E_{dj})$ of the welfare variables, E_{dj} , such that the vector $\mathbf{Y}_d = (y_{d1}, \dots, y_{dN_d})^T \sim N(\boldsymbol{\mu}_d, \boldsymbol{\Sigma}_d)$. Then the FGT measures $F_{\alpha dj}$ can be expressed as function of the random variables Y_{dj} :

$$F_{\alpha dj}(Y_{dj}) = \left(\frac{z - T(Y_{dj})^{-1}}{z} \right)^\alpha I(T(Y_{dj})^{-1} < z), \quad j = 1, \dots, N_d, \alpha \geq 0, \quad (4.90)$$

Expression (4.90) shows that the FGT poverty measures are complex nonlinear functions of the random vector $\mathbf{Y}_d$, especially when $\alpha > 0$, and the predictors discussed in Chapter 4 can be used to improve the small area estimates of the poverty measures relatively to the direct estimator (4.88).

As mentioned earlier, a similar setup to Molina and Rao (2010) is used for these simulations. They used $\mathbb{V}ar(u_d) = 0.15^2$ and $\mathbb{V}ar(e_{dj}) = 0.5^2$. From Chapter 2, we know that if $X \sim SN(\mu, \sigma^2, \lambda)$ then $\mathbb{E}(X) = \mu + \delta\sigma\sqrt{\frac{2}{\pi}}$ and $\mathbb{V}ar(X) = \sigma^2(1 - \frac{2\delta^2}{\pi})$. Using these results, the full SN nested error model used in the simulations is

$$\mathbf{Y}_{dj} = \mathbf{x}_{dj}^T \boldsymbol{\beta} + u_d + e_{dj}, j = 1, \dots, N_d = 250, d = 1, \dots, m = 80, \quad (4.91)$$

where

$$\begin{aligned}\boldsymbol{\beta} &= (3, 0.03, -0.04)^T, \delta_u = 1/\sqrt{1+1^2} \approx 0.7071, \delta_e = 3/\sqrt{1+3^2} \approx 0.9487, \\ u_d &\stackrel{iid}{\sim} SN(\mu_u = -\sigma_u \delta_u \sqrt{\frac{2}{\pi}} \approx -0.1025, \sigma_u^2 = 0.15^2 / (1 - \frac{2\delta_u^2}{\pi}) \approx 0.1817^2, \lambda_u = 1), \\ e_{dj} &\stackrel{iid}{\sim} SN(\mu_e = -\sigma_e \delta_e \sqrt{\frac{2}{\pi}} \approx -0.5792, \sigma_e^2 = 0.50^2 / (1 - \frac{2\delta_e^2}{\pi}) \approx 0.7651^2, \lambda_e = 3).\end{aligned}$$

This choice of parameters ensures that:

$$\mathbb{E}(u_d) = \mathbb{E}(e_{dj}) = 0, \quad \mathbb{V}ar(u_d) = 0.15^2, \quad \text{and} \quad \mathbb{V}ar(e_{dj}) = 0.50^2$$

In addition of the full SN nested error model, three other models were run in the simulations. These three models correspond to the combination of one of the two area or unit level errors being normally distributed ($\lambda_u = 0$ or $\lambda_e = 0$) or both being normally distributed ($\lambda_u = 0$ and $\lambda_e = 0$).

For the initial values, we fitted the nested error model assuming normal distribution and used the estimate of $\boldsymbol{\beta}$, $\mathbb{V}ar(u_d)$, and $\mathbb{V}ar(e_{dj})$ as the initial values of respectively $\boldsymbol{\beta}$, σ_u^2 , and σ_e^2 . The initial values of λ_u and λ_e were both chosen to be equal to 0.5.

Figure 4.1 shows comparisons of the probability density function (pdf) of the SN to the pdf of the normal with the same mean and standard deviation (sd).

Figure 4.1: Pdf of SN and the pdf of the normals with the same mean and the standard deviation (sd).

The shape parameter $\lambda_u = 1$ results in a minimal asymmetry and departure from the normal distribution with the same mean and variance. The skewness of the unit level distribution e_{dj} is moderate with $\lambda_e = 3$. Of course, because of the small sample sizes involved in the simulations, the actual empirical distributions are much less smooth than in Figure 4.1 and may show more departure from the assumed distributions.

There are two auxiliary variables $X_1 \in \{0, 1\}$ and $X_2 \in \{0, 1\}$ plus an intercept X_0 . The values of the two dummies X_1 and X_2 are generated from Bernoulli distributions with

$$P(X_1 = 1) = 0.3 + \frac{0.5d}{m}, \quad P(X_2 = 1) = 0.2, \quad d = 1, \dots, m = 80. \quad (4.92)$$

The probability of $X_1 = 1$ is higher for small areas with higher values of d . In other words, poverty reduces as d increases. The welfare variables E_{dj} are chosen to be exponential functions of the values y_{dj} , that is, the transformation $T(x) = \log(x)$. With this set of parameters choices, the overall poverty incidence (proportion of people under poverty) is around 15%. In each small area d , a sample s_d of $n_d = 50$ units is selected using simple random sampling. The total sample size is $n = 4,000$ selected from a total population size of $N = 20,000$. The Monte Carlo simulation consists of generating $I = 5,000$ populations, then for each generated population the SAE methods described in the previous sections (empirical best, conditional empirical best, half-normal, Molina-Rao, and ELL) are applied to obtain estimates of the complex parameter for the small areas. Figure 4.2 shows the Monte Carlo average of the poverty incidence, gap, and severity for all the 80 areas. The proportion of people under the poverty line (poverty incidence) on average goes from 15.69% to 14.35%; the areas with lower value of area tend to have higher poverty levels than the areas with higher value of area because of how X_2 was generated. Poverty gap and severity are also decreasing function of the area indicator. Note that in the graphs from Figure

4.2, the poverty gap is multiplied by 10 and the poverty severity is multiplied by 100.

Figure 4.2: Monte Carlo averages of the population poverty incidence, gap, and severity for each of the 80 areas; using the 5,000 populations with $n_d = 50$, $\lambda_u = 1$ and $\lambda_e = 3$.

The sampling design used is simple random sampling (SRS) with a fairly high sampling fraction of 20%. Empirical coefficients of variation (CVs) were obtained by drawing 50,000 SRS samples from a population of 80 areas with $n_d = 50$, $\lambda_u = 1$ and $\lambda_e = 3$. The empirical CVs for each domain were calculated as the ratio of the empirical standard error of the 50,000 poverty measure design-based estimates over the average of the 50,000 estimates. The results show that the average design-based CV over the 80 areas was about 36% for the poverty incidence, 45% for the poverty gap, and 59% for the poverty severity. Hence, the direct estimator (4.88) is not reliable even though the area sample size is $n_d = 50$. Further the CV seems to increase with the complexity of the parameter.

4.9.2 Marginal and Conditional Best Prediction

In the parametric approach, the first step is to use the sample to estimate the unknown parameters of the model (4.91). The ML method developed in section (3.4.1) is used to estimate the unknown parameters. Table 4.1 shows the Monte Carlo average estimates (Average), the standard errors (SE), the bias, the absolute relative bias (ARB) and the absolute bias ratio (ABR) for each estimated parameter in the model using the 5,000 populations. The statistics SE, ARB, and ABR are defined in (3.46), (3.47), (3.48), and (3.49) respectively. SE is large for λ_u since its value represents 110% of the average value. Similarly values of SE relative to the averages for β_1 and β_2 are large with respectively 49.0% and 44.9%. All the parameters have ARB smaller than 5%. The parameter σ_u is the only one to show an absolute ratio bias larger than 5% with a value of 28.1%.

Table 4.1: Parameter estimation from the Monte Carlo simulation of fitting the SN model ($\lambda_e = 3$ and $\lambda_u = 1$). We simulated 5,000 populations with $m = 80$ small areas and $n_d = 50$ units per area.

Parameters	Average (SE)	Bias	ARB(%)	ABR (%)
$\beta_0(3)$	3.0003 (0.0207)	-0.0003	0.01	1.36
$\beta_1(0.03)$	0.0300 (0.0146)	0.0001	0.48	0.98
$\beta_2(-0.04)$	-0.0397 (0.0178)	-0.0003	0.65	1.46
$\sigma_u(0.18)$	0.1729 (0.0312)	0.0088	4.83	28.08
$\sigma_e(0.77)$	0.7650 (0.0132)	0.0001	0.01	0.68
$\lambda_u(1)$	0.9713 (1.0663)	0.0287	2.87	2.69
$\lambda_e(3)$	3.0079 (0.1794)	-0.0079	0.26	4.42

Figure 4.3 shows the reduction in MSE obtained by the best predictor over the direct estimator. These gains range from about 25% to 40% for all three poverty measures (incidence, gap, and severity). And given that the bias of the best estimator is small, most of the decrease of the MSE results in a decrease of variability and therefore decrease of the coefficient of variation.

Figure 4.3: Reduction in percentage of the MSE obtained by EQB over the MSE of the direct estimator for all three poverty measures (incidence, gap, and severity) with $m = 80$ areas, $n_d = 50$ units, $\lambda_e = 3$, and $\lambda_u = 1$.

Figure 4.4 shows a comparison of the empirical quasi-best (EQB) and the conditional empirical predictors (C-EQB and C-EBLUP). The estimator EQB is nearly unbiased while C-EQB and C-EBLUP show a small bias. Two estimators EQB and C-EQB are equivalent in terms of MSE. As expected C-EQBLUP shows higher MSEs than the quasi-best predictors EQB and C-EQB.

Figure 4.4: Comparison of the empirical quasi-best predictor EQB to the conditional predictors C-EQB and C-EQBLUP in terms of bias and MSE (poverty gap, $\alpha=1$) with $m = 80$ areas, $n_d = 50$ units, $\lambda_e = 3$, and $\lambda_u = 1$.

4.9.3 Prediction based on the Half-Normal Representation

The estimators developed in Section 4.5, based on the half-normal representation of the SN, are more variable than the empirical best method because in the half-normal approach t_{dj} is drawn from its half-normal distribution instead of using the best predicted value of $\mathbf{t}_d$. Figure 4.5 shows the comparison of the quasi-best predictor $\hat{\eta}_d^{EQB}$ to the predictors $\hat{\eta}_d^{SC-HN}$ and $\hat{\eta}_d^{DC-HN}$ defined respectively by (4.65) and (4.67). Both half-normal based predictors show higher Monte Carlo bias and MSE than the empirical quasi-best predictor. The two half-normal based predictors have similar performance in terms of bias and MSE. The increase in MSE is quite high, in fact all the gain from the direct estimator is lost by using the half-normal based approaches.

Figure 4.5: Comparison of the empirical quasi-best predictor EQB and the estimators using the half-normal representation in terms of bias and MSE (poverty gap, $\alpha=1$) with $m = 80$ areas, $n_d = 50$ units, $\lambda_e = 3$, and $\lambda_u = 1$.

4.9.4 Molina-Rao Predictor under Skew-Normal Model

A normal distribution is often assumed for the random components when developing small area methods. In the use of these methods, this assumption is not always fully investigated using diagnostics and model checking. A common approach, when the outcome data is skewed, is to transform the data with some deterministic function to bring the outcome distribution closer to normal family and then apply usual techniques that assume normality. In this subsection, simulations are run to assess the performance of the MR normality-based estimator (which assumes normal distribution for the area effect and the errors) when the error terms are generated from a SN distribution. These simulation results give an idea of the bias and inefficiency of the Molina-Rao normality-based predictor when the errors follow SN distributions. In this sub-section, three different models were generated. The first model generates data where both u_d and e_{dj} are skew normal with $\text{Var}(u_d) = 0.15^2$ and $\text{Var}(e_{dj}) = 0.50^2$. The second model generates e_{dj} from a SN distribution with $\text{Var}(e_{dj}) = 0.50^2$ and u_d follows a normal distribution with $\text{Var}(u_d) = 0.15^2$. The last model assumes that u_d follows SN with $\text{Var}(u_d) = 0.15^2$ and $e_{dj} \sim N(0, 0.50^2)$. For all three models, the errors are centralized so that $\mathbb{E}(u_d) = \mathbb{E}(e_{dj}) = 0$.

Table 4.2 shows the estimated β as well as $\text{Var}(u_d)$ and $\text{Var}(e_{dj})$. Estimation of $\text{Var}(e_{dj})$ is stable and consistent across the three models. However $\text{Var}(u_d)$ is estimated correctly only when u_d is assumed normal; the two other models both understate the variability of the random area effect. The fixed effect is best estimated when only u_d follows a SN distribution.

Table 4.2: Parameter estimation from the Monte Carlo simulation when fitting SN using normal distribution. We simulated 5,000 population with $m = 30$ small areas and $n_d = 50$ units per area.

Parameters	u_d and $e_{dj} \sim SN$	Only $e_{dj} \sim SN$	Only $u_d \sim SN$
$\beta_0(3)$	3.0005 (0.0219)	3.0002 (0.0220)	3.0000 (0.0218)
$\beta_1(0.03)$	0.0296 (0.0189)	0.0300 (0.0194)	0.0300 (0.0190)
$\beta_2(-0.04)$	-0.0401 (0.0205)	-0.0397 (0.0209)	-0.0403 (0.0208)
$\text{Var}(u_d)(0.15)$	0.1495 (0.0149)	0.1493 (0.0146)	0.1488 (0.0146)
$\text{Var}(e_{dj})(0.50)$	0.5001 (0.0064)	0.4993 (0.0064)	0.5001 (0.0057)

Figure 4.6 shows the effect of the departure from normal on the bias. The bias is small when only u_d follows SN relative to the two other models. In fact, for the model “Only $u_d \sim SN$ ”, the relative bias is practically zero. However for the two other models, when only e_{dj} follows SN, the relative bias increase to be a little higher than 30%.

Figure 4.6: Absolute and relative bias of the Molina-Rao normality-based estimator when both random errors follow SN distribution (poverty gap, $\alpha=1$).

In terms of MSEs, shown on Figure 4.7, the model with only u_d following SN has much smaller values than the two other models.

Figure 4.7: MSE of the Molina-Rao normality-based estimator when at least one random error follows SN distribution (poverty gap, $\alpha=1$).

The contribution of the bias to the MSE, measured as the ratio of the bias squared to the MSE, is very small for the model with only u_d following SN. For the two other models, about half of the MSEs come from the bias (Figure 4.8).

Figure 4.8: Ratio of bias squared to MSE of the Molina-Rao normality-based estimator (in percent) when at least one random error follows SN distribution (poverty gap, $\alpha=1$).

When only the random area effect follows a SN distribution, the MR normality-based estimator is very efficient as shown in Figure 4.9. In this situation, MR normality-based estimator is about 50% more efficient than the direct estimator. However, when the unit level errors follow SN the MR normality-based estimator becomes inefficient and does even worse than the direct estimator.

Figure 4.9: Ratio of MSE of the Molina-Rao normality-based estimator over the direct estimator when both random errors follow SN distribution (poverty gap, $\alpha=1$).

4.9.5 ELL Method under Skew-Normal Model

In this section, simulation results are shown on the comparison of the original ELL, $\hat{\eta}_d^{ELL-TRAD}$, to the two alternatives, $\hat{\eta}_d^{ELL-RES}$ and $\hat{\eta}_d^{ELL-MOM}$, proposed in this thesis. The top graph from Figure 4.10 shows that $\hat{\eta}_d^{ELL-MOM}$ has a negative bias and is larger in absolute value than the two other ELL methods. In relative terms, bottom graph from Figure 4.10, one can see that the two methods based on the residuals, $\hat{\eta}_d^{ELL-TRAD}$ and $\hat{\eta}_d^{ELL-RES}$, have the smallest values around 2 to 3% on average. The relative bias of $\hat{\eta}_d^{ELL-MOM}$ is much larger with average value about 8%.

Figure 4.10: Bias of the three different ELL methods when both random errors follow SN distribution (poverty gap, $\alpha=1$) with $m = 80$ areas, $n_d = 50$ units, $\lambda_e = 3$, and $\lambda_u = 1$.

The MSEs, as shown in Figure 4.11, are very high for the original ELL, $\hat{\eta}_d^{ELL-TRAD}$, compared to the two proposed alternative methods. The bottom graph of Figure 4.11 shows the improvement in terms of MSE of the two alternatives over the original ELL represented by the ratio of the MSE of the alternative methods to the MSE of the original ELL minus one (improvement = MSE of alternative/MSE of original ELL - 1). The results show a very high gain for both alternative methods with $\hat{\eta}_d^{ELL-RES}$ achieving nearly 80% improvement over the original ELL.

Figure 4.11: MSE of the three different ELL methods when both random errors follow SN distribution (poverty gap, $\alpha=1$) with $m = 80$ areas, $n_d = 50$ units, $\lambda_e = 3$, and $\lambda_u = 1$.

Figure 4.12 shows how the three ELL methods compare to Molina-Rao. In terms of MSE, only “ELL single u_d ” does better than Molina-Rao and as observed in previous empirical studies the original ELL does worse than Molina-Rao.

Figure 4.12: MSE of the three different ELL methods and Molina-Rao normality-based when both the random errors follow SN distribution (poverty gap, $\alpha=1$) with $m = 80$ areas, $n_d = 50$ units, $\lambda_e = 3$, and $\lambda_u = 1$.

4.9.6 Comparison of the Different Predictors

In this section, the five estimators (EQB, C-EQB, SC-HN, ELL-MOM, and MR-N) are compared to the direct estimator. For each poverty measure, the ratio of the MSEs of the five estimators over the MSE of the direct estimator is computed. Values of the ratio under 1 show gain over the direct estimator while values over 1 indicate that the direct estimator is better in term of MSE. Results on poverty incidence ($\alpha = 0$), from Figure 4.13, show that all five estimators do better than the direct estimator. Among the five estimators, the marginal and conditional quasi-best estimators are equivalently the best and the modified ELL is marginally the worse. Note that MR normality-based estimator does barely better than both the Half-Normal and the ELL. The poverty incidence ($\alpha = 0$) is a simple proportion and therefore may not be considered as a truly complex parameter.

Figure 4.13: Ratio of the MSEs of the five estimators (best predictor, conditional best predictor, half-normal, Molina-Rao, and modified ELL using MOM) to the MSE of the direct estimator (poverty incidence, $\alpha=0$) with $m = 80$ areas, $n_d = 50$ units, $\lambda_e = 3$, and $\lambda_u = 1$.

Results on the poverty gap ($\alpha = 1$), from Figure 4.14, show that only the two quasi-best estimators (marginal and conditional) do better than the direct estimator with improvement about 25% for the lowest area indicator to about 33% for highest area indicator. The half-normal only a little worse than the direct estimator and ELL methods performs equivalently to the direct estimator. The MR normality-based estimator is much worse than the other four predictors and the director estimator. The performance of the MR normality-based estimator is worse for the poverty gap ($\alpha = 1$) than for the previous less complex poverty incidence ($\alpha = 0$). Relative to the direct estimator, the other four predictors are less affected by the extra complexity of the parameter of interest.

Figure 4.14: Ratio of the MSEs of the five estimators (best predictor, conditional best predictor, half-normal, Molina-Rao, and modified ELL using MOM) to the MSE of the direct estimator (poverty gap, $\alpha=1$) with $m = 80$ areas, $n_d = 50$ units, $\lambda_e = 3$, and $\lambda_u = 1$.

Results on poverty severity ($\alpha = 2$), from Figure 4.15, shows similar results to the ones discussed for poverty gap (Figure 4.14). One major difference is the fact that MR normality-based estimator performs significantly worse for poverty severity than for poverty gap. The ratio of its MSE to that of the direct estimator is about 2.70 on average across areas. As the parameter of interest get more complex from $\alpha = 1$ to $\alpha = 2$, the MR normality-based estimator's efficiency, relative to the direct estimator, deteriorates while the four other estimators keep approximately keep the same efficiency relative to the direct estimator.

Figure 4.15: Ratio of the MSEs of the five estimators (best predictor, conditional best predictor, half-normal, Molina-Rao, and modified ELL using MOM) to the MSE of the direct estimator (poverty severity, $\alpha=2$) with $m = 80$ areas, $n_d = 50$ units, $\lambda_e = 3$, and $\lambda_u = 1$.

4.10 MSE Estimation under Skew-Normal Errors

The MSE estimator must take account not only of the variability of the predictor but also the extra variability due to estimating the model parameters. Prasad and Rao (1990) derived a second-order approximation of the MSE for three special SAE models (nested error regression, random regression coefficient, and Fay-Herriot) assuming normal distribution for the random components. Harville and Jeske (1992) extend the Prasad-Rao estimator to the general LMM assuming unbiased estimation of the model parameters. Datta and Lahiri (2000) give a bias correction to the second-order approximation for ML-based estimators. Das et al. (2004) provide rigorous proofs to the second-order approximations under the general LMM with errors normally distributed both for REML and ML estimators. The Prasad-Rao estimator may not be applicable to complex parameters due to the lack of closed form expressions. Replication-based methods such as Jackknife (see Jiang et al. (2002)) and Bootstrap have been proposed for estimating MSE of complex parameters or non-normal SAE models.

The bootstrap technique is a popular resampling method used for estimating variance, MSE, or confidence intervals when analytical expressions are too difficult or impossible to derive (see Shao and Tu (1995) for an comprehensive discussion of the bootstrap method). In the recent years much literature was produced to propose nonparametric and parametric bootstrap methods for small area parameters. Hall and Maiti (2006a) proposed a nonparametric bootstrap approach method for estimating MSE of small area linear parameters. This method consisted in resampling from empirical approximations of the distributions of the random area effects u_d and the unit level errors e_{dj} that have the same moments up to the fourth one. Gonzalez-Manteiga et al. (2007) used a parametric bootstrap method to estimate MSE of small area linear parameters under logistic mixed model. Hall and Maiti (2006b) developed a double

bootstrap method to perform bias correction of parametric MSE estimators for small area parameters not necessary linear. Chatterjee et al. (2008) derived a parametric bootstrap method for estimating MSE focusing on the LMM. They considered more dependency in data than Hall and Maiti (2006b) and established results as total sample size increases on the joint variability of the random effects and the data, thus they obtained area specific estimators. Hence, they estimated the entire distribution of the EBLUP estimator and constructed CIs. Lahiri et al. (2007) proposed an bootstrap-based estimator of the MSE under the generalized (including nonlinear) mixed models with nonnormality of the random errors.

The parametric bootstrap procedure, described by Gonzalez-Manteiga et al. (2007), for estimating the MSE is as follows:

1. Fit the nested error model to the sample data $(\mathbf{y}_s, \mathbf{X}_s)$ and obtain $\hat{\boldsymbol{\beta}}, \hat{\sigma}_u^2, \hat{\sigma}_e^2, \hat{\lambda}_u,$ and $\hat{\lambda}_e$ using the ML method from Section 3.4.1 or any other suitable method.
2. Generate $u_d^{*(b)}$ and $e_{dj}^{*(b)}$ independently from respectively $SN(0, \hat{\sigma}_u^2, \hat{\lambda}_u)$ and $SN(0, \hat{\sigma}_e^2, \hat{\lambda}_e)$, $j = 1, \dots, N_d$ and $d = 1, \dots, m$.
3. Construct B independent and identically distributed bootstrap populations $y_{dj}^{*(b)} = \mathbf{x}_{dj}^T \hat{\boldsymbol{\beta}} + u_d^{*(b)} + e_{dj}^{*(b)}$ and calculate bootstrap population parameters $\eta_d^{*(b)} = h(\mathbf{y}_d^{*(b)})$, $b = 1, \dots, B$.
4. From each bootstrap population b generated in 3) take the sample with the same indices s as the initial sample. Calculate the bootstrap EB estimator $\hat{\eta}_d^{EQB*(b)}$ using the best prediction.
5. The bootstrap MSE estimator is:

$$mse_*(\hat{\eta}_d^{EQB}) = \frac{1}{B} \sum_{b=1}^B \left(\hat{\eta}_d^{EQB*(b)} - \eta_d^{*(b)} \right)^2 \quad (4.93)$$

Note that the best prediction in point 4) may be replaced by the conditional best prediction to obtain in point 5) the MSE estimator of the conditional predictor as follows:

$$mse_*(\hat{\eta}_d^{C-EQB}) = \frac{1}{B} \sum_{b=1}^B \left(\hat{\eta}_d^{C-EQB*(b)} - \eta_d^{*(b)} \right)^2. \quad (4.94)$$

The double bootstrap proposed by Hall and Maiti (2006b) may be used to improve the MSE estimators above. However this double bootstrap approach is computationally very intensive. The bias correction proposed by Lahiri et al. (2007) is computationally more friendly. Note that the method proposed by Chatterjee et al. (2008) does not apply since it assumes normal distributions for the random components.

To estimate the MSE of the ELL-MOM estimators, the bootstrap procedure is as follows:

1. Fit the nested error model to the sample data $(\mathbf{y}_s, \mathbf{X}_s)$ and obtain $\hat{\boldsymbol{\beta}}_{OLS}$ using OLS method, and $\hat{\sigma}_u^2$ and $\hat{\sigma}_e^2$ using the MOM method (4.80)-(4.82).
2. Generate $u_d^{*(b)}$ using the EBLUP estimator of the random effect (4.83) and $e_{dj}^{*(b)}$ from the unit level residuals (4.84).
3. Construct B independent and identically distributed bootstrap populations $y_{dj}^{*(b)} = \mathbf{x}_{dj}^T \hat{\boldsymbol{\beta}}_{OLS} + u_d^{*(b)} + e_{dj}^{*(b)}$ and calculate bootstrap population parameters $\eta_d^{*(b)} = h(\mathbf{y}_d^{*(b)})$, $b = 1, \dots, B$.
4. From each bootstrap population b generated in 3) take the sample with the same indices s as the initial sample. Calculate the bootstrap ELL-MOM estimator $\hat{\eta}_d^{ELL-MOM*(b)}$ by following the entire procedure to get the ELL-MOM estimator (4.86) for the bootstrap population $\mathbf{y}_d^{*(b)}$.

5. The bootstrap MSE estimator is:

$$mse_*(\hat{\eta}_d^{ELL-MOM}) = \frac{1}{B} \sum_{b=1}^B \left(\hat{\eta}_d^{ELL-MOM*(b)} - \eta_d^{*(b)} \right)^2. \quad (4.95)$$

Note that since we have estimated the variance components in the ELL-MOM method, we may replaced the OLS estimator $\hat{\beta}_{OLS}$ by the more efficient generalized least square (GLS) estimator $\hat{\beta}_{GLS}$.

To estimate the MSE of the ELL-RES estimators, the bootstrap procedure is as follows:

1. Fit the nested error model to the sample data $(\mathbf{y}_s, \mathbf{X}_s)$ and obtain $\hat{\beta}_{OLS}$ using OLS method, and $\hat{\sigma}_u^2$ and $\hat{\sigma}_e^2$ using the residual method (4.76)-(4.77).
2. Generate $u_d^{*(b)}$ and $e_{dj}^{*(b)}$ by drawing from their respective empirical distributions.
3. Construct B independent and identically distributed bootstrap populations $y_{dj}^{*(b)} = \mathbf{x}_{dj}^T \hat{\beta}_{OLS} + u_d^{*(b)} + e_{dj}^{*(b)}$ and calculate bootstrap population parameters $\eta_d^{*(b)} = h(\mathbf{y}_d^{*(b)})$, $b = 1, \dots, B$.
4. From each bootstrap population b generated in 3) take the sample with the same indices s as the initial sample. Calculate the bootstrap ELL-RES estimator $\hat{\eta}_d^{ELL-RES*(b)}$ by following the entire procedure to get the ELL-RES estimator (4.79) for the bootstrap population $\mathbf{y}_d^{*(b)}$.

5. The bootstrap MSE estimator is:

$$mse_*(\hat{\eta}_d^{ELL-RES}) = \frac{1}{B} \sum_{b=1}^B \left(\hat{\eta}_d^{ELL-RES*(b)} - \eta_d^{*(b)} \right)^2. \quad (4.96)$$

4.11 Appendix: Proof of Proposition 4.3

To proof this proposition, it is sufficient to show that the vector $Y_{dr}|\mathbf{y}_{ds}$ has the same distribution as the sum of vectors $\mathbf{X}_{dr}\boldsymbol{\beta} + (u_d|\mathbf{y}_{ds})\mathbf{1}_{N_{dr}} + \mathbf{e}_{dr}$. The distribution of $Y_{dr}|\mathbf{y}_{ds}$ is given in the expression (4.12). In the other side of the equality, the distribution of $u_d|\mathbf{y}_{ds}$ is provided in Proposition 4.2. from the SN model assumption that e_{dj} follows SN distribution and Proposition 2.2, we have:

$$\mathbf{e}_{dr} \sim CSN_{N_{dr}} \left(\boldsymbol{\mu}_1 = \boldsymbol{\mu}_e, \boldsymbol{\Sigma}_1 = \sigma_e^2 \mathbf{I}_{N_{dr}}, D_1 = \frac{\lambda_e}{\sigma_e} \mathbf{I}_{N_{dr}}, \boldsymbol{\nu}_1 = \mathbf{0}_{N_{dr}}, \boldsymbol{\Gamma}_1 = \mathbf{I}_{N_{dr}} \right).$$

Also, using Proposition 2.6, it follows that:

$$(u_d|\mathbf{y}_{ds})\mathbf{1}_{N_{dr}} \sim SSN_{1,n_d+1}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2, D_2, \boldsymbol{\nu}_2, \boldsymbol{\Gamma}_2 = \mathbf{I}_{n_d+1}),$$

where $\mu_2 = (\mu_u + \gamma_{n_d} \mathbf{1}_{n_d}^T (\mathbf{y}_{ds} - (\mathbf{X}_{ds}\boldsymbol{\beta} + \mu_u \mathbf{1}_{n_d} + \boldsymbol{\mu}_{eds}))) \mathbf{1}_{N_{dr}}$,

$$\boldsymbol{\Sigma}_2 = \sigma_e^2 \gamma_{n_d} \mathbf{1}_{N_{dr}} \mathbf{1}_{N_{dr}}^T, \quad \mathbf{D}_2 = \frac{1}{N_{dr}} \begin{pmatrix} -\frac{\lambda_e}{\sigma_e} \mathbf{1}_{n_d} \\ \frac{\lambda_u}{\sigma_u} \end{pmatrix} \mathbf{1}_{N_d}^T, \text{ and}$$

$$\boldsymbol{\nu}_2 = - \begin{bmatrix} \frac{\lambda_e}{\sigma_e} [\mathbf{I}_{n_d} - \gamma_{n_d} \mathbf{1}_{n_d} \mathbf{1}_{n_d}^T] \\ \frac{\lambda_u}{\sigma_u} \gamma_{n_d} \mathbf{1}_{n_d}^T \end{bmatrix} (\mathbf{y}_{ds} - (\mathbf{X}_{ds}\boldsymbol{\beta} + \mu_u \mathbf{1}_{n_d} + \boldsymbol{\mu}_{eds})).$$

Proposition 2.10 provides the distribution of the sum as follows

$$(u_d|\mathbf{y}_{ds})\mathbf{1}_{N_{dr}} + \mathbf{e}_{dr} \sim CSN_{N_{dr}, N_d+1}(\boldsymbol{\mu}^\dagger, \boldsymbol{\Sigma}^\dagger, \mathbf{D}^\dagger, \boldsymbol{\nu}^\dagger, \boldsymbol{\Gamma}^\dagger). \quad (4.97)$$

The parameters $\boldsymbol{\mu}^\dagger$, $\boldsymbol{\Sigma}^\dagger$, $\mathbf{D}^\dagger$, $\boldsymbol{\nu}^\dagger$, and $\boldsymbol{\Gamma}^\dagger$ are computed below. The location parameter is simply the some of the location parameters that is

$$\boldsymbol{\mu}^\dagger = (\mu_u + \gamma_{n_d} \mathbf{1}_{n_d}^T (\mathbf{y}_{n_d} - (\mathbf{X}_{ds}\boldsymbol{\beta} + \mu_u \mathbf{1}_{n_d} + \boldsymbol{\mu}_{eds}))) \mathbf{1}_{N_{dr}} + \boldsymbol{\mu}_{edr}. \quad (4.98)$$

The dispersion matrix $\Sigma^\dagger$ is also the sum of the two dispersion matrices that is

$$\Sigma^\dagger = \sigma_e^2 (\mathbf{I}_{N_{dr}} + \gamma_{n_d} \mathbf{1}_{N_{dr}} \mathbf{1}_{N_{dr}}^T). \quad (4.99)$$

The matrix $\mathbf{D}^\dagger$ is obtained by $\mathbf{D}^\dagger = [\Sigma_1 \mathbf{D}_1^T \quad \Sigma_2 \mathbf{D}_2^T]^T (\Sigma_2 + \Sigma_2)^{-1}$ and after some matrix manipulations, we have

$$\mathbf{D}^\dagger = \begin{bmatrix} \frac{\lambda_e}{\sigma_e} [\mathbf{I}_{N_{dr}} - \gamma_{N_d} \mathbf{1}_{N_{dr}} \mathbf{1}_{N_{dr}}^T] \\ -\frac{\lambda_e}{\sigma_e} \gamma_{N_d} \mathbf{1}_{N_d} \mathbf{1}_{N_{dr}}^T \\ \frac{\lambda_u}{\sigma_u} \gamma_{N_d} \mathbf{1}_{N_{dr}}^T \end{bmatrix}. \quad (4.100)$$

The vector $\boldsymbol{\nu}^\dagger$ is obtained by $\boldsymbol{\nu}^\dagger = (\boldsymbol{\nu}_i^T, \boldsymbol{\nu}_2^T)^T$, that is

$$\begin{aligned} \boldsymbol{\nu}^\dagger &= \begin{pmatrix} \mathbf{0}_{N_{dr}} \\ - \begin{bmatrix} \frac{\lambda_e}{\sigma_e} [\mathbf{I}_{n_d} - \gamma_{n_d} \mathbf{1}_{n_d} \mathbf{1}_{n_d}^T] \\ \frac{\lambda_u}{\sigma_u} \gamma_{n_d} \mathbf{1}_{n_d}^T \end{bmatrix} (\mathbf{y}_{n_d} - (\mathbf{X}_d \boldsymbol{\beta} + \mu_u \mathbf{1}_{n_d} + \boldsymbol{\mu}_{eds})) \end{pmatrix} \\ &= - \begin{pmatrix} \mathbf{0}_{(N_{dr} \times n_d)} \\ \frac{\lambda_e}{\sigma_e} [\mathbf{I}_{n_d} - \gamma_{n_d} \mathbf{1}_{n_d} \mathbf{1}_{n_d}^T] \\ \frac{\lambda_u}{\sigma_u} \gamma_{n_d} \mathbf{1}_{n_d}^T \end{pmatrix} (\mathbf{y}_{n_d} - (\mathbf{X}_d \boldsymbol{\beta} + \mu_u \mathbf{1}_{n_d} + \boldsymbol{\mu}_{eds})). \end{aligned} \quad (4.101)$$

And last, the matrix $\mathbf{\Gamma}^\dagger$ is obtained by

$$\mathbf{\Gamma}^\dagger = \begin{pmatrix} \mathbf{\Gamma}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{\Gamma}_2 \end{pmatrix} + \begin{pmatrix} \mathbf{D}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_2 \end{pmatrix} \begin{pmatrix} \Sigma_1 & \mathbf{0} \\ \mathbf{0} & \Sigma_2 \end{pmatrix} \begin{pmatrix} \mathbf{D}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_2 \end{pmatrix}^T \quad (4.102)$$

$$- [\Sigma_1 \mathbf{D}_1^T \quad \Sigma_n \mathbf{D}_n^T]^T (\Sigma_1 + \Sigma_2)^{-1} [\Sigma_1 \mathbf{D}_1^T \quad \Sigma_2 \mathbf{D}_2^T]. \quad (4.103)$$

That is (after some matrix manipulations), we have

$$\mathbf{\Gamma}^\dagger = \begin{bmatrix} \mathbf{I}_{N_d} + \lambda_e^2 \gamma_{N_d} \mathbf{1}_{N_d} \mathbf{1}_{N_d}^T & -\lambda_e \lambda_u \left(\frac{\sigma_e}{\sigma_u} \right) \gamma_{N_d} \mathbf{1}_{N_d} \\ -\lambda_e \lambda_u \left(\frac{\sigma_e}{\sigma_u} \right) \gamma_{N_d} \mathbf{1}_{N_d}^T & 1 + \lambda_u^2 (1 - N_d \gamma_{N_d}) \end{bmatrix}. \quad (4.104)$$

Hence the distribution of $\mathbf{X}_{dr}\boldsymbol{\beta} + (u_d|\mathbf{y}_{ds})\mathbf{1}_{N_{dr}} + \mathbf{e}_{dr}$ is the same as the distribution of $Y_{dr}|\mathbf{y}_{ds}$ given in (4.12). The special case of the normal model is covered in this proof by setting $\lambda_u = \lambda_e = 0$. This is also easily proved using the normal distribution properties by noting that under the normal model, we have

$$\mathbf{e}_{dr} \sim CSN_{N_{dr}}(\mathbf{0}, \sigma_e^2 \mathbf{I}_{N_{dr}}) \quad (4.105)$$

$$(u_d|\mathbf{y}_{ds})\mathbf{1}_{N_{dr}} \sim N_{n_d+1}(\gamma_{n_d}\mathbf{1}_{n_d}^T(\mathbf{y}_{ds} - \mathbf{X}_{ds}\boldsymbol{\beta})\mathbf{1}_{N_{dr}}, \sigma_e^2\gamma_{n_d}\mathbf{1}_{N_{dr}}\mathbf{1}_{N_{dr}}^T). \quad (4.106)$$

Given (4.105) and (4.106), it follows that $\mathbf{X}_{dr}\boldsymbol{\beta} + (u_d|\mathbf{y}_{ds})\mathbf{1}_{N_{dr}} + \mathbf{e}_{dr}$ has the same multivariate normal distribution as $Y_{dr}|\mathbf{y}_{ds}$.

Chapter 5

Hierarchical Bayes (HB) Prediction under One-Fold SN Model

The EB approach treats the unknown parameters θ_k 's of the nested error model, where k indicates a element of the vector $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma_u^2, \lambda_u, \delta_u, \sigma_e^2, \lambda_e, \delta_e)$, as fixed quantities. An alternative approach when constructing the small area predictor, called hierarchical Bayes (HB) method, consists of considering the model parameters as random variables and therefore assume a distribution for $\boldsymbol{\theta}$ called the prior distribution. Inference is based on the posterior distribution which is the distribution of the parameters given the sample. Using Bayes rule and the likelihood function $f(\mathbf{y}_s|\boldsymbol{\theta})$, where s designates the sample, the probability density function (pdf) of the posterior distribution is obtained by

$$\pi(\boldsymbol{\theta}|\mathbf{y}_s) = \frac{f(\mathbf{y}_s|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(\mathbf{y}_s)} \quad (5.1)$$

where $f(\mathbf{y}_s) = \int f(\mathbf{y}_s|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$. Since the marginal pdf $f(\mathbf{y}_s)$ is not function of $\boldsymbol{\theta}$, it is customary to just write $\pi(\boldsymbol{\theta}|\mathbf{y}_s) \propto f(\mathbf{y}_s|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$ where $\propto$ means proportional to. The goal is to estimate the parameter $\eta_d = h(\mathbf{Y}_d)$, where h is nonlinear function and $d = 1, \dots, m$ and m is the number of small areas. Under the HB framework, the complex parameter η_d is estimated by the posterior mean $\mathbb{E}_{\boldsymbol{\theta}|\mathbf{y}_s}(\eta_d)$ and its variability measured by $\text{Var}_{\boldsymbol{\theta}|\mathbf{y}_s}(\eta_d)$.

The HB approach is straightforward. However evaluating the posterior mean can be analytically impossible for many applications due to the intractable form of the density $\pi(\boldsymbol{\theta}|\mathbf{y}_s)$. These computational issues can be addressed in two ways: the use of conjugate priors or numerical approximations. The conjugate priors solution consists of choosing the likelihood and the priors which ensure closed-form expressions for the posterior pdf. This solution may not be appropriate because subjective priors should be chosen accordingly to the problem but not for the sole reason of ensuring tractable form of the posterior. The numerical approximations are often more appropriate and can be classified into two major categories: classical Monte Carlo Methods (see Appendix 5.4 for a very brief introduction on classical Monte Carlo methods) and Markov chain Monte Carlo (MCMC) methods. MCMC methods consist of approximating the posterior distribution by constructing Markov chains which have the desired distribution as the equilibrium distribution. The MCMC methods are very popular because they can be used in almost any application. However their performance varies widely depending on the complexity of the problem. Often when successive samples from the posterior are highly correlated, MCMC methods take a long time to converge.

HB models have been routinely used in small area estimation (SAE) applications; see Nandram and Choi (2005, 2010), Mohadjer et al. (2012), Ghosh and Steorts (2013), Chapter 10 of Rao (2003). Most of these applications have focussed on non-complex parameters such as proportions, means, and totals. Recently, Molina et al. (2014) proposed an HB method for the estimation of complex small area parameters. Using classical Monte Carlo methods their method avoids MCMC techniques. They assumed the nested error model with random components following normal distributions. In this chapter, we extend their method by assuming skew-normal (SN) distribution for the error terms. The literature on the linear mixed model with independent unit level

errors following SN distribution is very limited. To my knowledge, only Arellano-Valle et al. (2007) considered the nested error model with independent e_{dj} 's following SN distribution. However they used a different parametrization of the multivariate extension and they also used MCMC along with the half-normal representation of the SN, for the Bayesian inference. In an earlier paper, Arellano-Valle et al. (2005) considered the linear mixed model but they used correlated unit level errors e_{dj} 's. The model with correlated unit level errors leads to a simpler multivariate extension of the SN distribution.

In this chapter, we refer to the nested error model with the random components following normal distribution as the *normal model*. Similarly, the nested error model with at least one of the components following the SN distribution is referred to as the *SN model*.

5.1 Review of the HB Method for the Normal Model

This section is a quick review of the HB method proposed by Molina et al. (2014). The normal model (1.5) is assumed for the characteristics of interest $\mathbf{Y}$. The authors used a reparameterization of the model coupled with an appropriate choice of noninformative priors so that samples can be drawn directly from the posterior density. The avoidance of the MCMC method is a great saving of time since MCMC procedures require drawing a large number of correlated samples from the proxy distributions to the posterior and monitoring the convergence can be a tedious task. In the next two sections, we review the normal model specifications and the derivation of the joint posterior density.

5.1.1 Model

As mentioned earlier, the assumed nested error model is reparameterized to substitute the area level variance component σ_u^2 by a function of the intraclass correlation coefficient $\rho = \sigma_u^2 / (\sigma_u^2 + \sigma_e^2)$. This reparameterization and the choice of prior distributions shown below lead to tractable form of the posterior which avoids the need of MCMC procedure, see Toto and Nandram (2010) for another use of this reparameterization. The nested error model after reparameterization is given by:

$$Y_{dj} = \mathbf{x}_{dj}^T \boldsymbol{\beta} + u_d + e_{dj}, \quad j = 1, \dots, N_d, \quad d = 1, \dots, m, \quad (5.2)$$

$$u_d \stackrel{iid}{\sim} N\left(0, \frac{\rho}{1-\rho} \sigma_e^2\right), \quad \text{and} \quad e_{dj} \stackrel{iid}{\sim} N(0, \sigma_e^2) \quad (5.3)$$

where $\mathbf{x}$ is a q -vector of auxiliary information including an intercept with q being an integer, $\boldsymbol{\beta}$ is a q -vector, u_d and e_{dj} are independent, d designates the small-area and j designates the unit within the area d . To better show the hierarchical structure of the model given by (5.2)-(5.3), we may write the model in the conditional form as follows:

$$Y_{dj} | u_d, \boldsymbol{\beta}, \sigma_e^2 \sim N(\mathbf{x}_{dj}^T \boldsymbol{\beta} + u_d, \sigma_e^2), \quad (5.4)$$

$$u_d | \rho, \sigma_e^2 \stackrel{iid}{\sim} N\left(0, \frac{\rho}{1-\rho} \sigma_e^2\right). \quad (5.5)$$

One of the main criticisms of the Bayesian method is the choice of the prior distribution. Ideally the prior distribution is constructed based on the mechanism underlying the generation of $\boldsymbol{\theta}$. In theory, past or similar surveys (experiments) can be used to estimate the distribution of $\pi(\boldsymbol{\theta})$. Often, it is not possible to find similar enough surveys to estimate $\pi(\boldsymbol{\theta})$; Jeffreys (1961) provides an extensive criticism of the concept of “repeatability” of experiments. When no information or only partial information is available on the distribution of $\boldsymbol{\theta}$, several methods have been proposed for the choice of the prior distribution $\pi(\boldsymbol{\theta})$. Some of these methods for choosing the prior distribution

are:

- The marginal distribution of $\mathbf{y}$, $m(\mathbf{y}) = \int f(\mathbf{y}|\theta)\pi(\theta)d\theta$ can be used to derive information on $\pi(\theta)$ (see Berger (1985)).
- The *maximum entropy method*, under the finite setting $\{\theta_i\}_{i \in \Omega}$, defines the prior distribution $\pi(\theta)$ as the distribution that maximizes the entropy function defined as

$$\varepsilon(\pi) = - \sum_i \pi(\theta_i) \log(\pi(\theta_i)) \quad (5.6)$$

under the constraints $\mathbb{E}_\pi(g_j(\theta)) < \infty$, $j = 1, \dots, J$, where $g_j(\theta)$ is a characteristic of the distribution $\pi(\theta)$ such as moments or quantiles.

- Using a parametric setup, the prior $\pi(\theta)$ can be chosen among the parametric pdf. This parametric approach reduces to estimating the parameters of distribution e.g. the mean and the variance for the normal distribution.
- If there is no prior information or the choice of the priors may lead to controversy, then prior distribution with minimal influence on the inference should be used. Such priors are called noninformative priors. Kass and Wasserman (1996) gave two interpretations of noninformative priors that summarize their usage. They stated that *1) Noninformative priors are formal representations of ignorance. 2) There is no objective, unique prior that represents ignorance, instead noninformative priors are chosen by public agreement much like units of length and weight.*

Noninformative Jeffreys' priors, introduced by Jeffreys (1946, 1961), are often used in practice. The Jeffreys' prior is based on the Fisher information and is given by

$$\pi(\theta) \propto I^{1/2}(\theta), \quad (5.7)$$

where θ is a scalar parameter and $I(\theta)$ is the Fisher information defined under some regularity conditions as

$$I(\theta) = -\mathbb{E}_{\theta}\left(\frac{\partial^2 \log f(\mathbf{y}|\theta)}{\partial^2 \theta}\right) \quad (5.8)$$

Note that the Jeffreys' prior only uses the information from the data and no other subjective information is used. One important property of the Jeffreys' prior is its invariance to transformation. Consider an one-to-one transformation h , we have that

$$I^{1/2}(\theta) = I^{1/2}(h(\theta))h'(\theta). \quad (5.9)$$

From result (5.9), it follows that if $\pi(\theta) \propto I^{1/2}(\theta)$ then $\pi(h(\theta)) \propto I^{1/2}(h(\theta)) = I^{1/2}(\theta)(h'(\theta))^{-1}$ (invariance property to one-to-one transformation). The Jeffreys prior determination extends to the case of vector parameters.

Molina et al. (2014) used uniform priors for ρ and $\pi(\boldsymbol{\beta}) \propto 1$. Values of ρ have the same probability of realization over any closed interval of $(0, 1)$, that is, in $[\epsilon, 1 - \epsilon]$, $\epsilon > 0$. For the dispersion parameter σ_e^2 , they used Jeffreys' reference prior $\pi(\sigma_e^2) \propto 1/\sigma_e^2$, $\sigma_e^2 > 0$. Thus for the nested error model given by (5.4)-(5.5), they considered for the unknown parameters $\boldsymbol{\beta}$, σ_e^2 , and ρ joint the noninformative prior

$$\pi(\boldsymbol{\beta}, \sigma_e^2, \rho) \propto \frac{1}{\sigma_e^2}, \quad \epsilon \leq \rho \leq 1 - \epsilon, \boldsymbol{\beta} \in \mathbb{R}^q. \quad (5.10)$$

5.1.2 Derivation of the Posterior Densities

In the derivation of the posterior distribution, Molina et al. (2014) treated the case where some of the small areas do not have any sample. Assume that there are m^* small areas with a sample size strictly positive. Without loss of generality, we may reorder the small areas so that $n_d > 0$ for $d \leq m^*$ and $n_d = 0$ for $d \in [m^* + 1, m]$. Under the nested error model (5.4)-(5.5), Bayes' theorem leads to the posterior distribution

given by

$$\begin{aligned} \pi(\mathbf{u}, \boldsymbol{\beta}, \sigma_e^2, \rho | \mathbf{y}_s) &\propto \left(\prod_{d=m^*+1}^m \pi(u_d | \sigma_e^2, \rho) \right) \left(\frac{1-\rho}{\rho} \right)^{m^*/2} \sigma_e^{-2((m^*+n)/2+1)} \\ &\times \exp \left(-\frac{1}{2\sigma_e^2} \sum_{d=1}^{m^*} \left[\sum_{j \in s_d} (y_{dj} - \mathbf{x}_{dj}^T \boldsymbol{\beta} - u_d)^2 + \frac{1-\rho}{\rho} u_d^2 \right] \right), \end{aligned} \quad (5.11)$$

where $\pi(u_d | \sigma_e^2, \rho)$ is the normal prior of u_d given in (5.5). Using the chain rule of probability, the joint posterior density (5.11) can be represented in product of univariate conditional probabilities as follows:

$$\pi(\mathbf{u}, \boldsymbol{\beta}, \sigma_e, \rho | \mathbf{y}_s) \propto \pi_1(\mathbf{u} | \boldsymbol{\beta}, \sigma_e^2, \rho, \mathbf{y}_s) \pi_2(\boldsymbol{\beta} | \sigma_e^2, \rho, \mathbf{y}_s) \pi_3(\sigma_e^2 | \rho, \mathbf{y}_s) \pi_4(\rho | \mathbf{y}_s) \quad (5.12)$$

Molina et al. (2014) showed that the posterior (5.12) is proper provided that the matrix $\mathbf{X} = \text{col}_{1 \leq d \leq m} \text{col}_{j \in s_d}(\mathbf{x}_{dj}^T)$ has full rank and ρ remains in the interior of $(0, 1)$ that is $\epsilon \leq \rho \leq 1 - \epsilon$, $\epsilon > 0$. The first individual posterior $\pi_1(\mathbf{u} | \boldsymbol{\beta}, \sigma_e^2, \rho, \mathbf{y}_s)$ is a product of univariate normal distributions defined as

$$\pi_1(\mathbf{u} | \boldsymbol{\beta}, \sigma_e^2, \rho, \mathbf{y}_s) \propto \left(\prod_{d=1}^{m^*} \pi(u_d | \boldsymbol{\beta}, \sigma_e^2, \rho, \mathbf{y}_s) \right) \left(\prod_{d=m^*+1}^m \pi(u_d | \sigma_e^2, \rho) \right) \quad (5.13)$$

where

$$u_d | \boldsymbol{\beta}, \sigma_e^2, \rho, \mathbf{y}_s \stackrel{\text{ind}}{\sim} N(\tau_d(\rho)(\bar{y}_{ds} - \bar{\mathbf{x}}_d^T \boldsymbol{\beta}), (1 - \tau_d(\rho)) \frac{\rho}{1 - \rho} \sigma_e^2) \quad (5.14)$$

with $\tau_d(\rho) = n_d(n_d + (1 - \rho)/\rho)^{-1}$, $d = 1, \dots, m^*$, $\bar{\mathbf{x}}_d = \frac{1}{n_d} \sum_{j \in s_d} \mathbf{x}_{dj}$, and $\bar{y}_d = \frac{1}{n_d} \sum_{j \in s_d} y_{dj}$. The posterior $\pi_2(\boldsymbol{\beta} | \sigma_e^2, \rho, \mathbf{y}_s)$ is the pdf of a normal distribution such that

$$\boldsymbol{\beta} | \sigma_e^2, \rho, \mathbf{y}_s \sim N(\hat{\boldsymbol{\beta}}(\rho) = \mathbf{Q}^{-1}(\rho) \mathbf{p}(\rho), \sigma_e^2 \mathbf{Q}^{-1}(\rho)), \quad (5.15)$$

where

$$\mathbf{Q}(\rho) = \sum_{d=1}^{m^*} \sum_{j \in s_d} (\mathbf{x}_{dj} - \bar{\mathbf{x}}_d)(\mathbf{x}_{dj} - \bar{\mathbf{x}}_d)^T + \frac{1-\rho}{\rho} \sum_{d=1}^{m^*} \tau_d(\rho) \bar{\mathbf{x}}_d \bar{\mathbf{x}}_d^T, \quad (5.16)$$

and

$$\mathbf{p}(\rho) = \sum_{d=1}^{m^*} \sum_{j \in s_d} (\mathbf{x}_{dj} - \bar{\mathbf{x}}_d)(y_{dj} - \bar{y}_d)^T + \frac{1-\rho}{\rho} \sum_{d=1}^{m^*} \tau_d(\rho) \bar{\mathbf{x}}_d \bar{y}_d. \quad (5.17)$$

The posterior density $\pi_3(\sigma_e^2 | \rho, \mathbf{y}_s)$ is the pdf of a gamma distribution such that

$$\sigma_e^{-2} | \rho, \mathbf{y}_s \sim \text{Gamma}\left(\frac{n-p}{2}, \frac{\gamma(\rho)}{2}\right), \quad (5.18)$$

where

$$\gamma(\rho) = \sum_{d=1}^{m^*} \sum_{j \in s_d} (y_{dj} - \bar{y}_d - (\mathbf{x}_{dj} - \bar{\mathbf{x}}_d)^T \hat{\boldsymbol{\beta}}(\rho))^2 + \frac{1-\rho}{\rho} \sum_{d=1}^{m^*} \tau_d(\rho) (\bar{y}_d - \bar{\mathbf{x}}_d^T \hat{\boldsymbol{\beta}}(\rho))^2. \quad (5.19)$$

The last posterior density $\pi_4(\rho | \mathbf{y}_s)$ is given by

$$\pi_4(\rho | \mathbf{y}_s) \propto \left(\frac{1-\rho}{\rho}\right)^{m^*/2} |\mathbf{Q}(\rho)|^{-1/2} \gamma(\rho)^{-(n-p)/2} \prod_{d=1}^{m^*} \tau_d(\rho)^{1/2}, \quad \epsilon \leq \rho \leq 1 - \epsilon. \quad (5.20)$$

The chain rule representation (5.12) allows drawing the parameters from the posterior distribution one at a time. Indeed, generation of $\boldsymbol{\theta} = (\mathbf{u}, \boldsymbol{\beta}, \sigma_e^2, \rho)$ jointly from $\pi(\mathbf{u}, \boldsymbol{\beta}, \sigma_e, \rho | \mathbf{y}_s)$ is equivalent to first generating ρ from $\pi_4(\rho | \mathbf{y}_s)$, then σ_e^2 from $\pi_3(\sigma_e^2 | \rho, \mathbf{y}_s)$, then $\boldsymbol{\beta}$ from $\pi_2(\boldsymbol{\beta} | \sigma_e^2, \rho, \mathbf{y}_s)$, and finally $\mathbf{u}$ from $\pi_1(\mathbf{u} | \boldsymbol{\beta}, \sigma_e^2, \rho, \mathbf{y}_s)$. The posterior distributions $\pi_1(\mathbf{u} | \boldsymbol{\beta}, \sigma_e^2, \rho, \mathbf{y}_s)$, $\pi_2(\boldsymbol{\beta} | \sigma_e^2, \rho, \mathbf{y}_s)$, and $\pi_3(\sigma_e^2 | \rho, \mathbf{y}_s)$ are simple with closed forms; drawing from these distributions does not carry any difficulty. However $\pi_4(\rho | \mathbf{y}_s)$ is complex and the authors used the grid method to draw ρ . To generate a set of parameters $\boldsymbol{\theta}^{(\ell)} = (\mathbf{u}^{(\ell)}, \boldsymbol{\beta}^{(\ell)}, \sigma_e^{2(\ell)}, \rho^{(\ell)})$, $\ell = 1, \dots, L$ from the posterior densities, the procedure is as follows:

1. Generation of the intraclass correlation coefficient $\rho^{(\ell)}$: take a grid of R points,

ρ_r , in the interval $[\epsilon, 1 - \epsilon]$ where $\epsilon < 1/(2R)$, $\rho_r = (r - 0.5)/R$ $r = 1, \dots, R$.

Consider the kernel of the posterior density of ρ as

$$k_4(\rho) = \left(\frac{1 - \rho}{\rho} \right)^{m^*/2} |\mathbf{Q}(\rho)|^{-1/2} \gamma(\rho)^{-(n-p)/2} \prod_{d=1} m^* \tau_d(\rho)^{1/2}.$$

Calculate $k_4(\rho_r)$, $r = 1, \dots, R$ and take $\pi_4(\rho_r) = \frac{k_4(\rho_r)}{\sum_{r=1}^R k_4(\rho_r)}$, $r = 1, \dots, R$. Then generate $\rho^{(\ell)}$ from the discrete distribution $\rho_r, \pi_4(\rho_r)_{r=1}^R$. Jitter each discrete value by adding to it a uniform random number in the interval $(0, 1/R)$.

2. Generation of the variance component σ_e^2 : draw $\sigma^{-2(\ell)}$ from the distribution

$$\sigma_e^{-2(\ell)} | \rho^{(\ell)}, \mathbf{y}_s \sim \text{Gamma}\left(\frac{n-p}{2}, \frac{\gamma(\rho^{(\ell)})}{2}\right),$$

and take $\sigma_e^{2(\ell)} = 1/\sigma_e^{-2(\ell)}$

3. Generation of regression coefficients: draw $\boldsymbol{\beta}^{(\ell)}$ from the distribution

$$\boldsymbol{\beta}^{(\ell)} | \sigma_e^{2(\ell)}, \rho^{(\ell)}, \mathbf{y}_s \sim N(\hat{\boldsymbol{\beta}}(\rho^{(\ell)}) = \mathbf{Q}^{-1}(\rho^{(\ell)}) \mathbf{p}(\rho^{(\ell)}), \sigma_e^{2(\ell)} \mathbf{Q}^{-1}(\rho^{(\ell)}))$$

4. Generation of random area effects: draw $u_d^{(\ell)}$, $d = 1, \dots, m^*$ from the distribution

$$u_d^{(\ell)} | \boldsymbol{\beta}^{(\ell)}, \sigma_e^{2(\ell)}, \rho^{(\ell)}, \mathbf{y}_s \stackrel{\text{ind}}{\sim} N(\tau_d(\rho^{(\ell)})(\bar{y}_{ds} - \bar{\mathbf{x}}_d^T \boldsymbol{\beta}^{(\ell)}), (1 - \tau_d(\rho^{(\ell)})) \frac{\rho^{(\ell)}}{1 - \rho^{(\ell)}} \sigma_e^{2(\ell)})$$

5.2 HB Method for the SN Model

As a reminder, the SN model refers to the nested error model with at least one of the random components following SN distribution. When both random components follow SN distribution, we may refer to the model as the *full SN model*. The normal distribution is a special case of SN distribution, therefore the normal model is a special

case of SN model. The pdf of the standard SN distribution, denoted $SN(0, 1, \lambda) \equiv SN(\lambda)$, is

$$f(z) = 2\phi(z)\Phi(\lambda z), \quad -\infty \leq z, \lambda \leq \infty$$

where λ is a shape parameter. If $y = \mu + \sigma z$ then $y \sim SN(\mu, \sigma^2, \lambda)$ where μ is the location parameter and σ^2 is the scale (dispersion) parameter. Note that the variance, say ϑ^2 , is a function of the dispersion parameter σ^2 and the shape parameter δ , that is

$$\vartheta^2 = \left(1 - \frac{2\delta^2}{\pi}\right)\sigma^2 \quad (5.21)$$

where $\delta = \lambda/\sqrt{1 + \lambda^2}$ hence $-1 < \delta < 1$. Since there is a one-to-one relationship between λ and δ , it is equivalent to specify the SN distribution using either λ or δ i.e. $SN(\mu, \sigma^2, \lambda(\delta)) \equiv SN(\mu, \sigma^2, \delta(\lambda))$.

In the following, we propose an extension of the HB method presented in Section 5.1 to the case of the SN model. The idea of the method is to approximate the SN model by a normal model to first draw the parameters $\mathbf{u}, \boldsymbol{\beta}, \vartheta_e^2$, and ρ using the method proposed by Molina et al. (2014) then use the sampling importance resampling (SIR) technique to improve the initial draws. Secondly, using the chain rule, the shape parameters δ_e and δ_u are drawn from their conditional posterior distributions. Finally, σ_e^2 is obtained using the relationship (5.21).

5.2.1 Model

Consider the nested error model where both the area random effect u_d and the unit level error e_{dj} follow SN distribution (full SN model); the full SN model can be written

as follows:

$$Y_{dj} = \mathbf{x}_{dj}^T \boldsymbol{\beta} + u_d + e_{dj}, \quad j = 1, \dots, N_d, \quad d = 1, \dots, m, \quad (5.22)$$

$$u_d \stackrel{iid}{\sim} N(-\delta_u \sigma_u \sqrt{\frac{2}{\pi}}, \sigma_u^2, \delta_u), \quad \text{and} \quad e_{dj} \stackrel{iid}{\sim} N(-\delta_e \sigma_e \sqrt{\frac{2}{\pi}}, \sigma_e^2, \delta_e) \quad (5.23)$$

where u_d is independent of e_{dj} for any $d = 1, \dots, m$ and j in area d . Note that the choice of the location parameters $-\delta_u \sigma_u \sqrt{\frac{2}{\pi}}$ and $-\delta_e \sigma_e \sqrt{\frac{2}{\pi}}$ ensures that $E(u_d) = E(e_{dj}) = 0$. Both Arellano-Valle et al. (2005) and Arellano-Valle et al. (2007) did not treat the specific case of adjusted zero-mean random errors in their SN linear mixed model applications. Following this adjusted choice of u_d and e_{dj} , the SN model can be written after reparameterization as follows:

$$y_{dj} - (x_{dj}^T \boldsymbol{\beta} + u_d) | u_d, \boldsymbol{\beta}, \sigma_e^2, \delta_e \stackrel{ind}{\sim} SN(-\delta_e \sigma_e \sqrt{\frac{2}{\pi}}, \sigma_e^2, \delta_e), \quad (5.24)$$

$$u_d | \sigma_e^2, \rho, \delta_u \stackrel{ind}{\sim} SN(-\delta_u \sigma_e \sqrt{\frac{2\rho\omega_e}{\pi(1-\rho)\omega_u}}, \frac{\rho}{1-\rho} \frac{\omega_e}{\omega_u} \sigma_e^2, \delta_u), \quad (5.25)$$

where $j = 1, \dots, n_d$, $d = 1, \dots, m$, $\rho = \frac{\vartheta_u^2}{\vartheta_u^2 + \vartheta_e^2}$ with $\vartheta_u^2 = \omega_u \sigma_u^2$, $\omega_u = (1 - \frac{2\delta_u^2}{\pi})$, $\vartheta_e^2 = \omega_e \sigma_e^2$, and $\omega_e = (1 - \frac{2\delta_e^2}{\pi})$. If both u_d and e_{dj} are normally distributed then $\omega_u = \omega_e = 1$ (since $\delta_u = \delta_e = 0$), ρ simplifies to the usual expressions, and the SN model (5.24)-(5.25) reduces to the case of normal model (5.4)-(5.5).

The choice of the priors is partially guided by the strategy we intend to follow for drawing from the posterior. As mentioned earlier, the plan is to take advantage of the easy draws from the posterior of the normal model to obtain an importance function. Here we use the sampling importance resampling (SIR). Therefore the first step is to approximate the SN distributions (5.24) and (5.25) by normal distributions with the same first and second moments. This means that the unit level distribution (5.24) is approximated by $N(0, \vartheta_e^2)$. Similarly the area level distribution (5.25) is

approximated by $N(0, \frac{\rho}{1-\rho}\vartheta_e^2)$. The HB method for the SN model, presented in this section, draws the parameters $(\vartheta_e^2, \delta_e)$ then uses the relationship (5.21) to derive σ_e^2 ; this is justified by the fact that knowing (σ_e^2, δ_e) is equivalent to knowing $(\vartheta_e^2, \delta_e)$ given the relationship (5.21). The prior is therefore attached to ϑ_e^2 rather than σ_e^2 . Following the model developed for the normal case in Section 5.1, the prior associated with the parameters β , ϑ_e^2 , and ρ is:

$$\pi(\beta, \vartheta_e^2, \rho) \propto \frac{1}{\vartheta_e^2}, \quad \vartheta_e^2 > 0. \quad (5.26)$$

For the parameters δ_u and δ_e , we choose the noninformative uniform prior over their support, that is

$$\delta_u \sim U(-1, 1) \quad \text{and} \quad \delta_e \sim U(-1, 1) \quad (5.27)$$

If only one of the two random errors follows SN and the other one is normally distributed, the corresponding “partial” SN model is easily obtained from (5.24)-(5.25) by setting $\delta_u = 0$ or $\delta_e = 0$ accordingly.

5.2.2 Model Fitting

Using Bayes’ theorem, the joint posterior density resulting from the SN model (5.24)-(5.24) and the choice of the prior distributions (5.26) and (5.27) is given by

$$\begin{aligned}
\pi(\mathbf{u}, \boldsymbol{\beta}, \vartheta_e^2, \rho, \delta_e, \delta_u | \mathbf{y}_d) &\propto \frac{1}{\vartheta_e^2} \frac{1}{2} \prod_{d=1}^m \prod_{j=1}^{n_d} \left(\frac{2}{\sqrt{2\pi\sigma_e^2}} \exp\left(-\frac{(y_{dj} - x_{dj}^T \boldsymbol{\beta} - u_d + \delta_e \sigma_e \sqrt{\frac{2}{\pi}})^2}{2\sigma_e^2}\right) \right. \\
&\quad \times \left. \Phi\left(\frac{\delta_e}{\sqrt{1-\delta_e^2}} \frac{(y_{dj} - x_{dj}^T \boldsymbol{\beta} - u_d + \delta_e \sigma_e \sqrt{\frac{2}{\pi}})}{\sigma_e}\right) \right) \\
&\quad \times \frac{1}{2} \prod_{d=1}^m \left(2\sqrt{\frac{(1-\rho)\omega_u}{2\pi\rho\sigma_e^2\omega_e}} \exp\left(-\frac{(1-\rho)\omega_u}{2\rho\sigma_e^2\omega_e} (u_d + \delta_u \sigma_e \sqrt{\frac{2\rho\omega_e}{\pi(1-\rho)\omega_u}})^2 \right) \right. \\
&\quad \times \left. \Phi\left(\frac{\delta_u}{\sqrt{1-\delta_u^2}} \sqrt{\frac{(1-\rho)\omega_u}{\rho\sigma_e^2\omega_e}} (u_d + \delta_u \sigma_e \sqrt{\frac{2\rho\omega_e}{\pi(1-\rho)\omega_u}}) \right) \right) \quad (5.28)
\end{aligned}$$

In the posterior density expression (5.28), σ_e^2 is not a fundamental parameter instead σ_e^2 is function of ϑ_e^2 and δ_e . Indeed, we could replace σ_e^2 by $\vartheta_e^2(1 - \frac{2\delta_e^2}{\pi})^{-1}$ in the posterior density (5.28). To simplify the expressions, we denote $\eta_{dj} = y_{dj} - x_{dj}^T \boldsymbol{\beta} - u_d$, $\mu_e = -\delta_e \sigma_e \sqrt{\frac{2}{\pi}}$ and $\mu_u = -\delta_u \sigma_e \sqrt{\frac{2\rho\omega_e}{\pi(1-\rho)\omega_u}}$. It follows that

$$(y_{dj} - x_{dj}^T \boldsymbol{\beta} - u_d + \delta_e \sigma_e \sqrt{\frac{2}{\pi}})^2 = \eta_{dj}^2 + \mu_e(\mu_e - 2\eta_{dj}) \quad (5.29)$$

and similarly

$$(u_d + \delta_u \sigma_e \sqrt{\frac{2\rho\omega_e}{\pi(1-\rho)\omega_u}})^2 = u_d^2 + \mu_u(\mu_u - 2u_d). \quad (5.30)$$

Using result (5.29) and the property $\exp(a+b) = \exp(a)\exp(b)$, the first exponential function in the joint density (5.28) can be written as follows:

$$\begin{aligned}
\exp\left(-\frac{(\eta_{dj} - \mu_e)^2}{2\sigma_e^2}\right) &= \exp\left(-\frac{(\eta_{dj} - \mu_e)^2}{2\vartheta_e^2} \frac{\vartheta_e^2}{\sigma_e^2}\right) \\
&= \exp\left(-\frac{(\eta_{dj} - \mu_e)^2}{2\vartheta_e^2} \frac{(1 - \frac{2\delta_e^2}{\pi})\sigma_e^2}{\sigma_e^2}\right) \\
&= \exp\left(-\frac{(\eta_{dj} - \mu_e)^2}{2\vartheta_e^2} + \frac{(\eta_{dj} - \mu_e)^2 \delta_e^2}{\pi \vartheta_e^2}\right) \\
&= \exp\left(-\frac{(\eta_{dj} - \mu_e)^2}{2\vartheta_e^2}\right) \exp\left(\frac{\delta_e^2}{\pi} \frac{(\eta_{dj} - \mu_e)^2}{\vartheta_e^2}\right) \\
&= \exp\left(-\frac{\eta_{dj}^2 + \mu_e(\mu_e - 2\eta_{dj})}{2\vartheta_e^2}\right) \exp\left(\frac{\delta_e^2}{\pi} \frac{(\eta_{dj} - \mu_e)^2}{\vartheta_e^2}\right) \\
&= \exp\left(-\frac{\eta_{dj}^2}{2\vartheta_e^2}\right) \exp\left(-\frac{\mu_e(\mu_e - 2\eta_{dj})}{2\vartheta_e^2}\right) \exp\left(\frac{\delta_e^2}{\pi} \frac{(\eta_{dj} - \mu_e)^2}{\vartheta_e^2}\right)
\end{aligned}$$

and similarly using result (5.30), we decompose the second exponential expression as follows:

$$\begin{aligned}
\exp\left(-\frac{(1-\rho)\omega_u(u_d - \mu_u)^2}{2\rho\omega_e\sigma_e^2}\right) &= \exp\left(-\frac{(u_d - \mu_u)^2}{2\sigma_u^2}\right) \\
&= \exp\left(-\frac{(u_d - \mu_u)^2}{2\vartheta_u^2} \frac{\vartheta_u^2}{\sigma_u^2}\right) \\
&= \exp\left(-\frac{(u_d - \mu_u)^2}{2\vartheta_u^2} \left(1 - \frac{2\delta_u^2}{\pi}\right)\right) \\
&= \exp\left(-\frac{(u_d - \mu_u)^2}{2\vartheta_u^2}\right) \exp\left(\frac{(u_d - \mu_u)^2 \delta_u^2}{\pi \vartheta_u^2}\right) \\
&= \exp\left(-\frac{u_d^2 + \mu_u(\mu_u - 2u_d)}{2\vartheta_u^2}\right) \exp\left(\frac{(u_d - \mu_u)^2 \delta_u^2}{\pi \vartheta_u^2}\right) \\
&= \exp\left(-\frac{(1-\rho)}{2\rho\vartheta_e^2} u_d^2\right) \exp\left(-\frac{(1-\rho)}{2\rho\vartheta_e^2} \mu_u(\mu_u - 2u_d)\right) \exp\left(\frac{\delta_u^2}{\pi} \frac{(1-\rho)}{\rho\vartheta_e^2} (u_d - \mu_u)^2\right).
\end{aligned}$$

In the next step, we express the cumulative probabilities in terms of the second moments ϑ_e^2 as follows:

$$\begin{aligned}
\Phi\left(\frac{\delta_e}{\sqrt{1-\delta_e^2}}\frac{(\eta_{dj}-\mu_e)}{\sigma_e}\right) &= \Phi\left(\frac{\delta_e}{\sqrt{1-\delta_e^2}}\frac{\omega_e^{1/2}(\eta_{dj}-\mu_e)}{\omega_e^{1/2}\sigma_e}\right) \\
&= \Phi\left(\frac{\delta_e}{\sqrt{1-\delta_e^2}}\frac{\omega_e^{1/2}(\eta_{dj}-\mu_e)}{\vartheta_e}\right) \\
&= \Phi\left(\frac{\delta_e}{\sqrt{1-\delta_e^2}}\sqrt{1-\frac{2\delta_e^2}{\pi}}\frac{(\eta_{dj}-\mu_e)}{\vartheta_e}\right) \\
&= \Phi\left(\frac{\delta_e\omega_e^{1/2}}{\sqrt{1-\delta_e^2}}\frac{(\eta_{dj}-\mu_e)}{\vartheta_e}\right)
\end{aligned}$$

and

$$\begin{aligned}
\Phi\left(\frac{\delta_u}{\sqrt{1-\delta_u^2}}\sqrt{\frac{(1-\rho)\omega_u}{\rho\omega_e\sigma_e^2}}(u_d-\mu_u)\right) &= \Phi\left(\frac{\delta_u}{\sqrt{1-\delta_u^2}}\sqrt{\frac{(1-\rho)(1-\frac{2\delta_u^2}{\pi})}{\rho\vartheta_e^2}}(u_d-\mu_u)\right) \\
&= \Phi\left(\frac{\delta_u}{\sqrt{1-\delta_u^2}}\sqrt{1-\frac{2\delta_u^2}{\pi}}\sqrt{\frac{1-\rho}{\rho\vartheta_e^2}}(u_d-\mu_u)\right) \\
&= \Phi\left(\frac{\delta_u\omega_u^{1/2}}{\sqrt{1-\delta_u^2}}\sqrt{\frac{1-\rho}{\rho\vartheta_e^2}}(u_d-\mu_u)\right).
\end{aligned}$$

The joint posterior distribution (5.28) can now be expressed as follows:

$$\begin{aligned}
\pi(\mathbf{u}, \boldsymbol{\beta}, \vartheta_e^2, \rho, \delta_e, \delta_u | \mathbf{y}_d) &\propto \frac{1}{\vartheta_e^2} \frac{1}{2} \prod_{d=1}^m \prod_{j=1}^{n_d} \frac{1}{\sqrt{2\pi\vartheta_e^2}} \exp\left(-\frac{\eta_{dj}^2}{2\vartheta_e^2}\right) \prod_{d=1}^m \sqrt{\frac{1-\rho}{2\pi\rho\vartheta_e^2}} \exp\left(-\frac{1-\rho}{2\rho\vartheta_e^2}u_d^2\right) \\
&\times \prod_{d=1}^m \prod_{j=1}^{n_d} 2\omega_e^{1/2} \exp\left(\frac{\delta_e^2}{\pi} \frac{(\eta_{dj}-\mu_e)^2}{\vartheta_e^2}\right) \exp\left(\frac{\mu_e(2\eta_{dj}-\mu_e)}{2\vartheta_e^2}\right) \\
&\times \Phi\left(\frac{\delta_e\omega_e^{1/2}}{\sqrt{1-\delta_e^2}} \frac{\eta_{dj}-\mu_e}{\vartheta_e}\right) \\
&\times \frac{1}{2} \prod_{d=1}^m 2\omega_u^{1/2} \exp\left(\frac{\delta_u^2}{\pi} \frac{(1-\rho)}{\rho\vartheta_e^2} (u_d-\mu_u)^2\right) \exp\left(\frac{(1-\rho)}{2\rho\vartheta_e^2} \mu_u(2u_d-\mu_u)\right) \\
&\times \Phi\left(\frac{\delta_u\omega_u^{1/2}}{\sqrt{1-\delta_u^2}} \sqrt{\frac{1-\rho}{\rho\sigma_e^2}} (u_d-\mu_u)\right)
\end{aligned} \tag{5.31}$$

The manipulations above on the the exponential terms and the cumulative functions are intended to isolate the posterior of the parameters associated with the normal model. The posterior pdf associated with the parameters $\mathbf{u}$, $\boldsymbol{\beta}$, ϑ_e^2 , ρ , say $\pi_1(\mathbf{u}, \boldsymbol{\beta}, \vartheta_e^2, \rho | \mathbf{y}_d)$, is obtained by integrating out δ_u and δ_e from the full prior expression (5.31). That is, $\pi_1(\mathbf{u}, \boldsymbol{\beta}, \vartheta_e^2, \rho | \mathbf{y}_d)$ is defined by

$$\begin{aligned}
\pi_1(\mathbf{u}, \boldsymbol{\beta}, \vartheta_e^2, \rho | \mathbf{y}_d) &\propto \frac{1}{\vartheta_e^2} \prod_{d=1}^m \prod_{j=1}^{n_d} \frac{1}{\sqrt{2\pi\vartheta_e^2}} \exp\left(-\frac{\eta_{dj}^2}{2\vartheta_e^2}\right) \prod_{d=1}^m \sqrt{\frac{1-\rho}{2\pi\rho\vartheta_e^2}} \exp\left(-\frac{1-\rho}{2\rho\vartheta_e^2} u_d^2\right) \\
&\times \frac{1}{2} \int_{-1}^1 \prod_{d=1}^m \prod_{j=1}^{n_d} 2\omega_e^{1/2} \exp\left(\frac{\delta_e^2 (\eta_{dj} - \mu_e)^2}{\pi \vartheta_e^2}\right) \exp\left(\frac{\mu_e (2\eta_{dj} - \mu_e)}{2\vartheta_e^2}\right) \\
&\times \Phi\left(\frac{\delta_e \omega_e^{1/2}}{\sqrt{1-\delta_e^2}} \frac{\eta_{dj} - \mu_e}{\vartheta_e}\right) d\delta_e \\
&\times \frac{1}{2} \int_{-1}^1 \prod_{d=1}^m 2\omega_u^{1/2} \exp\left(\frac{\delta_u^2 (1-\rho)}{\pi \rho \vartheta_e^2} (u_d - \mu_u)^2\right) \exp\left(\frac{(1-\rho)}{2\rho\vartheta_e^2} \mu_u (2u_d - \mu_u)\right) \\
&\times \Phi\left(\frac{\delta_u \omega_u^{1/2}}{\sqrt{1-\delta_u^2}} \sqrt{\frac{1-\rho}{\rho\sigma_e^2}} (u_d - \mu_u)\right) d\delta_u.
\end{aligned} \tag{5.32}$$

Hence, the posterior $\pi_1(\mathbf{u}, \boldsymbol{\beta}, \vartheta_e^2, \rho | \mathbf{y}_d)$ is decomposable as follows:

$$\pi_1(\mathbf{u}, \boldsymbol{\beta}, \vartheta_e^2, \rho | \mathbf{y}_d) \propto \pi_{1a}(\mathbf{u}, \boldsymbol{\beta}, \vartheta_e^2, \rho | \mathbf{y}_d) A(\mathbf{u}, \boldsymbol{\beta}, \vartheta_e^2, \rho | \mathbf{y}_d) \tag{5.33}$$

where

$$\pi_{1a}(\mathbf{u}, \boldsymbol{\beta}, \vartheta_e^2, \rho | \mathbf{y}_d) \propto \frac{1}{\vartheta_e^2} \prod_{d=1}^m \prod_{j=1}^{n_d} \frac{1}{\sqrt{2\pi\vartheta_e^2}} \exp\left(-\frac{\eta_{dj}^2}{2\vartheta_e^2}\right) \prod_{d=1}^m \sqrt{\frac{1-\rho}{2\pi\rho\vartheta_e^2}} \exp\left(-\frac{1-\rho}{2\rho\vartheta_e^2} u_d^2\right) \tag{5.34}$$

and

$$\begin{aligned}
A(\mathbf{u}, \boldsymbol{\beta}, \vartheta_e^2, \rho | \mathbf{y}_d) &= \frac{1}{2} \int_{-1}^1 \prod_{d=1}^m \prod_{j=1}^{n_d} 2\omega_e^{1/2} \exp\left(\frac{\delta_e^2 (\eta_{dj} - \mu_e)^2}{\pi \vartheta_e^2}\right) \exp\left(\frac{\mu_e (2\eta_{dj} - \mu_e)}{2\vartheta_e^2}\right) \\
&\times \Phi\left(\frac{\delta_e \omega_e^{1/2}}{\sqrt{1 - \delta_e^2}} \frac{\eta_{dj} - \mu_e}{\vartheta_e}\right) d\delta_e \\
&\times \frac{1}{2} \int_{-1}^1 \prod_{d=1}^m 2\omega_u^{1/2} \exp\left(\frac{\delta_u^2 (1 - \rho)}{\pi \rho \vartheta_e^2} (u_d - \mu_u)^2\right) \exp\left(\frac{(1 - \rho)}{2\rho \vartheta_e^2} \mu_u (2u_d - \mu_u)\right) \\
&\times \Phi\left(\frac{\delta_u \omega_u^{1/2}}{\sqrt{1 - \delta_u^2}} \sqrt{\frac{1 - \rho}{\rho \sigma_e^2}} (u_d - \mu_u)\right) d\delta_u.
\end{aligned} \tag{5.35}$$

It is interesting to note that $\pi_{1a}(\mathbf{u}, \boldsymbol{\beta}, \vartheta_e^2, \rho | \mathbf{y}_d)$ is exactly the posterior proposed by Molina et al. (2014) which we discussed in Section 5.1. The joint posterior pdf of the shape parameters δ_e and δ_u , say $\pi_2(\delta_e, \delta_u | \mathbf{u}, \boldsymbol{\beta}, \vartheta_e^2, \rho | \mathbf{y}_d)$, are obtained by dividing $\pi(\mathbf{u}, \boldsymbol{\beta}, \vartheta_e^2, \rho, \delta_e, \delta_u | \mathbf{y}_d)$ by $\pi_1(\mathbf{u}, \boldsymbol{\beta}, \vartheta_e^2, \rho | \mathbf{y}_d)$. It is easy to show that

$$\pi_2(\delta_e, \delta_u | \mathbf{u}, \boldsymbol{\beta}, \vartheta_e^2, \rho | \mathbf{y}_d) \propto \pi_{2e}(\delta_e | \mathbf{u}, \boldsymbol{\beta}, \vartheta_e^2, \rho, \mathbf{y}_d) \pi_{2u}(\delta_u | \mathbf{u}, \boldsymbol{\beta}, \vartheta_e^2, \rho, \mathbf{y}_d) \tag{5.36}$$

with

$$\begin{aligned}
\pi_{2e}(\delta_e | \mathbf{u}, \boldsymbol{\beta}, \vartheta_e^2, \rho | \mathbf{y}_d) &\propto \prod_{d=1}^m \prod_{j=1}^{n_d} 2\omega_e^{1/2} \exp\left(\frac{\delta_e^2 (\eta_{dj} - \mu_e)^2}{\pi \vartheta_e^2}\right) \exp\left(\frac{\mu_e (2\eta_{dj} - \mu_e)}{2\vartheta_e^2}\right) \\
&\times \Phi\left(\frac{\delta_e \omega_e^{1/2}}{\sqrt{1 - \delta_e^2}} \frac{\eta_{dj} - \mu_e}{\vartheta_e}\right)
\end{aligned} \tag{5.37}$$

and

$$\begin{aligned}
\pi_{2u}(\delta_u | \mathbf{u}, \boldsymbol{\beta}, \vartheta_e^2, \rho, \mathbf{y}_d) &\propto \prod_{d=1}^m 2\omega_u^{1/2} \exp\left(\frac{\delta_u^2 (1 - \rho)}{\pi \rho \vartheta_e^2} (u_d - \mu_u)^2\right) \\
&\times \exp\left(\frac{(1 - \rho)}{2\rho \vartheta_e^2} \mu_u (2u_d - \mu_u)\right) \Phi\left(\frac{\delta_u \omega_u^{1/2}}{\sqrt{1 - \delta_u^2}} \sqrt{\frac{1 - \rho}{\rho \sigma_e^2}} (u_d - \mu_u)\right).
\end{aligned} \tag{5.38}$$

In summary, the joint posterior pdf for the nested model with SN errors can be written using the following chain rule,

$$\begin{aligned} \pi(\mathbf{u}, \boldsymbol{\beta}, \vartheta_e^2, \rho, \delta_e, \delta_u | \mathbf{y}_d) &\propto \pi_{1a}(\mathbf{u}, \boldsymbol{\beta}, \vartheta_e^2, \rho | \mathbf{y}_d) A(\mathbf{u}, \boldsymbol{\beta}, \vartheta_e^2, \rho | \mathbf{y}_d) \\ &\quad \times \pi_{2e}(\delta_e | \mathbf{u}, \boldsymbol{\beta}, \vartheta_e^2, \rho, \mathbf{y}_d) \pi_{2u}(\delta_u | \mathbf{u}, \boldsymbol{\beta}, \vartheta_e^2, \rho, \mathbf{y}_d) \end{aligned} \quad (5.39)$$

where $\pi_{1a}(\mathbf{u}, \boldsymbol{\beta}, \vartheta_e^2, \rho | \mathbf{y}_d)$ is defined as in (5.34), $A(\mathbf{u}, \boldsymbol{\beta}, \vartheta_e^2, \rho | \mathbf{y}_d)$ as in (5.35), $\pi_{2e}(\delta_e | \mathbf{u}, \boldsymbol{\beta}, \vartheta_e^2, \rho, \mathbf{y}_d)$ as in (5.37), and $\pi_{2u}(\delta_u | \mathbf{u}, \boldsymbol{\beta}, \vartheta_e^2, \rho, \mathbf{y}_d)$ as in (5.38).

To draw the parameters from the posterior (5.39), we use the following strategy:

1. Draw a large sample of the set of parameters $(\mathbf{u}, \boldsymbol{\beta}, \vartheta_e^2, \rho)$ from posterior $\pi_{1a}(\mathbf{u}, \boldsymbol{\beta}, \vartheta_e^2, \rho | \mathbf{y}_d)$ following Molina et al. (2014). Let us denote the draws by $\mathbf{u}^{(h)}, \boldsymbol{\beta}^{(h)}, \vartheta_e^{2(h)}, \rho^{(h)}, h = 1, \dots, H$
2. Subsample without replacement (no more than 10% sampling rate) from the full sample in step (1) with probabilities proportional to $A(\mathbf{u}, \boldsymbol{\beta}, \vartheta_e^2, \rho | \mathbf{y}_d)$. This approach is referred to as Sampling Importance Resampling (SIR) and we give a very short introduction to SIR in Section 5.4. Note that expression (5.35) looks complex but it is composed of only one dimensional integrals of smooth functions; therefore it is easily evaluated using the grid method. Let's denote $\mathbf{u}^{(\ell)}, \boldsymbol{\beta}^{(\ell)}, \vartheta_e^{2(\ell)}, \rho^{(\ell)}, \ell = 1, \dots, L$ the final draws.
3. Draw samples of the shape parameter δ_e , of the same size as in step (2), from $\pi_{2e}(\delta_e | \mathbf{u}, \boldsymbol{\beta}, \vartheta_e^2, \rho, \mathbf{y}_d)$. We use a grid method to draw from these posteriors. Take a grid of R points, say $\delta_e^{(r)}$, in the interval $[-1 + \epsilon, 1 - \epsilon]$, where $\epsilon < 1/R$. For each value $\delta_e^{(r)}$, computes the probabilities $\pi_{2e}^{(r)} = \pi_{2e}(\delta_e^{(r)} | \mathbf{u}^{(\ell)}, \boldsymbol{\beta}^{(\ell)}, \vartheta_e^{2(\ell)}, \rho^{(\ell)}, \mathbf{y}_d)$ using (5.37). Then draw the value $\delta_e^{(\ell)}$ from the discrete distribution $\{\delta_e^{(r)}, \pi_{2e}^{(r)}\}_{r=1}^R$ and add a uniform random number from $[0, 1/R]$ to jitter the discrete values $\delta_e^{(\ell)}$.

4. Draw values $\delta_u^{(\ell)}$ using the same grid method as in step (3) with the posterior $\pi_{2u}(\delta_u|\mathbf{u}, \boldsymbol{\beta}, \vartheta_e^2, \rho, \mathbf{y}_d)$ defined by (5.38).
5. Use expression (5.21) to obtain σ_e^2 and σ_u^2 .

As discussed in Section 5.4, the importance weights, which are proportional to $A(\mathbf{u}^{(\ell)}, \boldsymbol{\beta}^{(\ell)}, \vartheta_e^{2(\ell)}, \rho^{(\ell)}|\mathbf{y}_d)$, must be bounded for the SIR method to work. The boundedness property of the weights is given by Lemma 5.1 below. In practice, we want the weights to be as closed to 1 as possible. Additional variability of the importance sampling weights increases the instability of the estimates resulting from the SIR procedure. Our choice of the importance function is the normal distribution with the same first and second moments as the SN distribution. Hence small values of the shape parameters λ_u and λ_e translate into a better approximation of the SN by the normal distribution with less disperse weights. Unfortunately, higher values of the shape parameters will produce more instability to our SIR method. Higher values of the shape parameters should be coupled with larger initial sample sizes and smaller subsampling rates (much less than 10%) to achieve similar efficiency than for smaller values of λ_u and λ_e .

Lemma 5.1. *The importance sampling weights $A(\mathbf{u}, \boldsymbol{\beta}, \vartheta_e^2, \rho|\mathbf{y}_d)$, as defined by (5.35), are bounded.*

Proof. The weight $A(\mathbf{u}, \boldsymbol{\beta}, \vartheta_e^2, \rho|\mathbf{y}_d)$ is a product of two integrals. If we show that both integrals are bounded, then $A(\mathbf{u}, \boldsymbol{\beta}, \vartheta_e^2, \rho|\mathbf{y}_d)$ is bounded as well. Denote the first integral I_e and the second integral I_u . Since $0 < \Phi\left(\frac{\delta_e \omega_e^{1/2}}{\sqrt{1-\delta_e^2}} \frac{\eta_{dj} - \mu_e}{\vartheta_e}\right) < 1$, for any d and j

then we have

$$\begin{aligned}
I_e &= \frac{1}{2} \int_{-1}^1 \prod_{d=1}^m \prod_{j=1}^{n_d} 2\omega_e^{1/2} \exp\left(\frac{\delta_e^2 (\eta_{dj} - \mu_e)^2}{\pi \vartheta_e^2}\right) \exp\left(\frac{\mu_e (2\eta_{dj} - \mu_e)}{2\vartheta_e^2}\right) \\
&\quad \times \Phi\left(\frac{\delta_e \omega_e^{1/2}}{\sqrt{1 - \delta_e^2}} \frac{\eta_{dj} - \mu_e}{\vartheta_e}\right) d\delta_e \\
&< \int_{-1}^1 \prod_{d=1}^m \prod_{j=1}^{n_d} 2\omega_e^{1/2} \exp\left(\frac{\delta_e^2 (\eta_{dj} - \mu_e)^2}{\pi \vartheta_e^2}\right) \exp\left(\frac{\mu_e (2\eta_{dj} - \mu_e)}{2\vartheta_e^2}\right) d\delta_e
\end{aligned} \tag{5.40}$$

The product of the exponential functions inside the integral from (5.40) is well defined, positive, and continuous in δ_e over $(-1, 1)$. Therefore the integrand has a maximum value, say I_{max} , and the integral is bounded by $2I_{max}$. Similarly, since $0 < \Phi\left(\frac{\delta_u \omega_u^{1/2}}{\sqrt{1 - \delta_u^2}} \sqrt{\frac{1 - \rho}{\rho \sigma_e^2}} (u_d - \mu_u)\right) < 1$ for any d , it follows that the second integral I_u is bounded. $\square$

5.2.3 Some Special Models

The SN model (5.24)-(5.25) is specified so that $E(e_{dj}) = E(u_d) = 0$. As mentioned earlier, Arellano-Valle et al. (2005) and Arellano-Valle et al. (2007) did not adjust the distributions of u_d and e_{dj} to get zero-mean errors. They basically assumed $u_d \sim SN(0, \sigma_u^2, \delta_u)$ and $e_{dj} \sim SN(0, \sigma_e^2, \delta_e)$. This particular choice of the distributions of u_d and e_{dj} is a special case of our model (5.24)-(5.25) which may be written as follows:

$$y_{dj} - (x_{dj}^T \boldsymbol{\beta} + u_d) | u_d, \boldsymbol{\beta}, \sigma_e^2, \delta_e \stackrel{ind}{\sim} SN(0, \sigma_e^2, \delta_e), \tag{5.41}$$

$$u_d | \sigma_e^2, \rho, \delta_u \stackrel{ind}{\sim} SN\left(0, \frac{\rho}{1 - \rho} \frac{\omega_e}{\omega_u} \sigma_e^2, \delta_u\right) \tag{5.42}$$

The same choice of priors as in (5.26) and (5.27) give a posterior with the same chain rule as (5.39) where $\pi_{1a}(\mathbf{u}, \boldsymbol{\beta}, \vartheta_e^2, \rho | \mathbf{y}_d)$ same as in (5.34),

$$\begin{aligned} A(\mathbf{u}, \boldsymbol{\beta}, \vartheta_e^2, \rho | \mathbf{y}_d) &= \frac{1}{2} \int_{-1}^1 \prod_{d=1}^m \prod_{j=1}^{n_d} 2\omega_e^{1/2} \exp\left(\frac{\delta_e^2 \eta_{dj}^2}{\pi \vartheta_e^2}\right) \Phi\left(\frac{\delta_e \omega_e^{1/2}}{\sqrt{1 - \delta_e^2}} \frac{\eta_{dj}}{\vartheta_e}\right) d\delta_e \\ &\times \frac{1}{2} \int_{-1}^1 \prod_{d=1}^m 2\omega_u^{1/2} \exp\left(\frac{\delta_u^2 (1 - \rho)}{\pi \rho \vartheta_e^2} u_d^2\right) \Phi\left(\frac{\delta_u \omega_u^{1/2}}{\sqrt{1 - \delta_u^2}} \sqrt{\frac{1 - \rho}{\rho \sigma_e^2}} u_d\right) d\delta_u, \end{aligned} \quad (5.43)$$

$$\pi_{2e}(\delta_e | \mathbf{u}, \boldsymbol{\beta}, \vartheta_e^2, \rho | \mathbf{y}_d) \propto \prod_{d=1}^m \prod_{j=1}^{n_d} 2\omega_e^{1/2} \exp\left(\frac{\delta_e^2 \eta_{dj}^2}{\pi \vartheta_e^2}\right) \Phi\left(\frac{\delta_e \omega_e^{1/2}}{\sqrt{1 - \delta_e^2}} \frac{\eta_{dj}}{\vartheta_e}\right), \quad (5.44)$$

and

$$\pi_{2u}(\delta_u | \mathbf{u}, \boldsymbol{\beta}, \vartheta_e^2, \rho, \mathbf{y}_d) \propto \prod_{d=1}^m 2\omega_u^{1/2} \exp\left(\frac{\delta_u^2 (1 - \rho)}{\pi \rho \vartheta_e^2} u_d^2\right) \Phi\left(\frac{\delta_u \omega_u^{1/2}}{\sqrt{1 - \delta_u^2}} \sqrt{\frac{1 - \rho}{\rho \sigma_e^2}} u_d\right) \quad (5.45)$$

Other interesting special cases are when e_{dj} or u_d follows a normal distribution i.e. respectively $\delta_e = 0$ or $\delta_u = 0$. When both δ_e and δ_u are equal to zero, we get the normal model discussed by Molina et al. (2014). Here we look at the case where $(\delta_e = 0, \delta_u \neq 0)$ or $(\delta_e \neq 0, \delta_u = 0)$. We know from the EB approach, developed in Chapter 4, that the Molina-Rao estimator based on normality assumption performs almost without loss of efficiency in terms of MSE when only u_d follows SN distribution $(\delta_e = 0, \delta_u \neq 0)$. However misspecification on the distribution of e_{dj} (assuming normal distribution for the e_{dj} 's when they follow SN distribution) leads to large bias and increases significantly the MSE.

Consider the case where e_{dj} follows SN distribution $(\delta_e \neq 0)$ and u_d is normally

distributed ($\delta_u = 0$). The model may be written as follows:

$$y_{dj} - (x_{dj}^T \boldsymbol{\beta} + u_d) | u_d, \boldsymbol{\beta}, \sigma_e^2, \delta_e \stackrel{ind}{\sim} SN(-\delta_e \sigma_e \sqrt{\frac{2}{\pi}}, \sigma_e^2, \delta_e), \quad (5.46)$$

$$u_d | \sigma_e^2, \rho, \delta_u \stackrel{ind}{\sim} N(0, \frac{\rho}{1-\rho} \omega_e \sigma_e^2), \quad (5.47)$$

The priors are:

$$\pi(\boldsymbol{\beta}, \vartheta_e^2, \rho) \propto \frac{1}{\vartheta_e^2}, \quad \vartheta_e^2 > 0 \quad \text{and} \quad \delta_e \sim U(-1, 1), \quad (5.48)$$

where $\rho = \sigma_u^2 / (\sigma_u^2 + \vartheta_e^2)$. Note that σ_u^2 is used in the definition of the ρ because it is the second moment of the normal distribution associated with u_d . The resulting posterior density can be written as follows:

$$\pi(\mathbf{u}, \boldsymbol{\beta}, \vartheta_e^2, \rho, \delta_e | \mathbf{y}_d) \propto \pi_{1a}(\mathbf{u}, \boldsymbol{\beta}, \vartheta_e^2, \rho | \mathbf{y}_d) A(\mathbf{u}, \boldsymbol{\beta}, \vartheta_e^2, \rho | \mathbf{y}_d) \pi_{2e}(\delta_e | \mathbf{u}, \boldsymbol{\beta}, \vartheta_e^2, \rho, \mathbf{y}_d) \quad (5.49)$$

where $\pi_{1a}(\mathbf{u}, \boldsymbol{\beta}, \vartheta_e^2, \rho | \mathbf{y}_d)$ and $\pi_{2e}(\delta_e | \mathbf{u}, \boldsymbol{\beta}, \vartheta_e^2, \rho, \mathbf{y}_d)$ defined as in respectively (5.35) and (5.37), and

$$\begin{aligned} A(\mathbf{u}, \boldsymbol{\beta}, \vartheta_e^2, \rho | \mathbf{y}_d) = & \frac{1}{2} \int_{-1}^1 \prod_{d=1}^m \prod_{j=1}^{n_d} 2\omega_e^{1/2} \exp\left(\frac{\delta_e^2 (\eta_{dj} - \mu_e)^2}{\pi \vartheta_e^2}\right) \exp\left(\frac{\mu_e (2\eta_{dj} - \mu_e)}{2\vartheta_e^2}\right) \\ & \times \Phi\left(\frac{\delta_e \omega_e^{1/2}}{\sqrt{1 - \delta_e^2}} \frac{\eta_{dj} - \mu_e}{\vartheta_e}\right) d\delta_e \end{aligned} \quad (5.50)$$

The importance weight defined by (5.50) is a little simpler since it only requires integrating out δ_e . The model, defined by (5.46)-(5.47), seems to be the most relevant because misspecifying the area effect u_d (assuming u_d follows normal distribution when in fact it is generated from a SN distribution) has little effect on MSE. Hence, one should avoid the extra complexity of the full SN model (5.24)-(5.25) and just use the simpler model (5.46)-(5.47) to address skewness in the distribution.

The last special case, for which the area effect u_d follows SN ($\lambda_u \neq 0$) and the unit level error e_{dj} follows normal ($\lambda_e = 0$), is defined as follows:

$$y_{dj} - (x_{dj}^T \boldsymbol{\beta} + u_d) | u_d, \boldsymbol{\beta}, \sigma_e^2, \stackrel{ind}{\sim} N(0, \sigma_e^2), \quad (5.51)$$

$$u_d | \sigma_e^2, \rho, \delta_u \stackrel{ind}{\sim} SN(-\delta_u \sigma_e \sqrt{\frac{2\rho}{\pi(1-\rho)\omega_u}}, \frac{\rho}{1-\rho} \frac{1}{\omega_u} \sigma_e^2, \delta_u), \quad (5.52)$$

The priors are:

$$\pi(\boldsymbol{\beta}, \sigma_e^2, \rho) \propto \frac{1}{\sigma_e^2}, \quad \sigma_e^2 > 0 \quad \text{and} \quad \delta_u \sim U(-1, 1). \quad (5.53)$$

Since e_{dj} follows a normal distribution σ_e^2 is the second moment and therefore the intraclass correlation coefficient is defined as $\rho = \vartheta_u^2 / (\vartheta_u^2 + \sigma_e^2)$. The posterior density is

$$\pi(\mathbf{u}, \boldsymbol{\beta}, \sigma_e^2, \rho, \delta_e | \mathbf{y}_d) \propto \pi_{1a}(\mathbf{u}, \boldsymbol{\beta}, \sigma_e^2, \rho | \mathbf{y}_d) A(\mathbf{u}, \boldsymbol{\beta}, \sigma_e^2, \rho | \mathbf{y}_d) \pi_{2u}(\delta_u | \mathbf{u}, \boldsymbol{\beta}, \sigma_e^2, \rho, \mathbf{y}_d) \quad (5.54)$$

where $\pi_{1a}(\mathbf{u}, \boldsymbol{\beta}, \sigma_e^2, \rho | \mathbf{y}_d)$ and $\pi_{2u}(\delta_u | \mathbf{u}, \boldsymbol{\beta}, \sigma_e^2, \rho, \mathbf{y}_d)$ are obtained by replacing ϑ_e^2 by σ_e^2 in respectively (5.35) and (5.38), and

$$\begin{aligned} A(\mathbf{u}, \boldsymbol{\beta}, \sigma_e^2, \rho | \mathbf{y}_d) &= \frac{1}{2} \int_{-1}^1 \prod_{d=1}^m 2\omega_u^{1/2} \exp\left(\frac{\delta_u^2 (1-\rho)}{\pi \rho \sigma_e^2} (u_d - \mu_u)^2\right) \\ &\times \exp\left(\frac{(1-\rho)}{2\rho \sigma_e^2} \mu_u (2u_d - \mu_u)\right) \Phi\left(\frac{\delta_u \omega_u^{1/2}}{\sqrt{1-\delta_u^2}} \sqrt{\frac{1-\rho}{\rho \sigma_e^2}} (u_d - \mu_u)\right) d\delta_u \end{aligned} \quad (5.55)$$

The model, defined by (5.51)-(5.52), is less interesting in practice because it does not differentiate enough from the normal model.

5.2.4 Prediction of Complex Parameters

As discussed in Section 4.2, we are interested in estimating nonlinear function of $\mathbf{y}_d$, say $\eta_d = h(\mathbf{y}_d)$. The HB predictor of η_d is given by the posterior mean

$$\hat{\eta}_d^{HB} = \mathbb{E}(\eta_d | \mathbf{y}_{ds}) = \int h(\mathbf{y}_d) f(\mathbf{y}_{dr} | \mathbf{y}_s) d\mathbf{y}_{dr} \quad (5.56)$$

where $\mathbf{y}_{ds}$ and $\mathbf{y}_{dr}$ designate respectively the sample and the out-of-sample values falling in domain d , and $f(\mathbf{y}_{dr} | \mathbf{y}_s)$ is the posterior density of $\mathbf{y}_{dr}$ given $\mathbf{y}_s$ that is

$$f(\mathbf{y}_{dr} | \mathbf{y}_s) = \int \prod_{j \in r_d} f(\mathbf{y}_{dj} | \boldsymbol{\theta}) \pi(\boldsymbol{\theta} | \mathbf{y}_s) d\boldsymbol{\theta}, \quad (5.57)$$

r_d designates the out-of-sample units from domain d and $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma_u^2, \lambda_u, \delta_u, \sigma_e^2, \lambda_e, \delta_e)$. The integral (5.56) is numerically evaluated using Monte Carlo methods. Following the introductory remarks in Section 5.4, we need to draw a large number of vectors $\mathbf{y}_{dr}$ from the posterior density $f(\mathbf{y}_{dr} | \mathbf{y}_s)$. Under the SN model, $f(\mathbf{y}_{dr} | \mathbf{y}_s)$ corresponds to the SN distribution defined by

$$Y_{dj} | \boldsymbol{\theta} \stackrel{ind}{\sim} SN(x_{dj}^T \boldsymbol{\beta} + u_d - \delta_e \sigma_e \sqrt{\frac{2}{\pi}}, \sigma_e^2, \delta_e), \quad j \in r_d. \quad (5.58)$$

Drawing from the SN distribution defined in (5.58) is easily done given proposition 2.1. We may rewrite (5.58) as follows:

$$\frac{Y_{dj} - (x_{dj}^T \boldsymbol{\beta} + u_d - \delta_e \sigma_e \sqrt{\frac{2}{\pi}})}{\sigma_e^2} | \boldsymbol{\theta} \stackrel{ind}{\sim} SN(\delta_e), \quad j \in r_d. \quad (5.59)$$

To draw a value from distribution $SN(\delta_e)$, a simple strategy is to draw two values x_0 and x_1 from two independent standard normals $N(0, 1)$; then the composite value $z = \delta_e |x_0| + (1 - \delta_e^2)^{1/2} x_1$ is considered to be a pick from $SN(\delta_e)$. A draw y_{dj} from

the distribution in (5.58) is then obtained by

$$y_{dj} = x_{dj}^T \boldsymbol{\beta} + u_d - \delta_e \sigma_e \sqrt{\frac{2}{\pi}} + \sigma_e^2 z. \quad (5.60)$$

Hence, the prediction of η_d can be summarized as follows:

1. Draw a large number, say L , of the set of parameters from the posterior $\pi(\boldsymbol{\theta}|\mathbf{y}_s)$ using the SIR method described in Section 5.2.2.
2. For each draw $\boldsymbol{\theta}^{(\ell)}$ from step (1), $\ell = 1, \dots, L$, generate $y_{dj}^{(\ell)}$ from the distribution of $Y_{dj}|\boldsymbol{\theta}$, $j \in r_d$ using (5.60).
3. Combine the out-of-sample generated values $y_{dj}^{(\ell)}$ to the sample vector $\mathbf{y}_{ds}$ to construct the full predicted census $\mathbf{y}_d^{(\ell)} = \left((\mathbf{y}_{dr}^{(\ell)})^T, \mathbf{y}_{ds}^T \right)^T$ for small area d . The complex parameter of interest η_d is estimated by $\hat{\eta}_d^{(\ell)} = h(\mathbf{y}_d^{(\ell)})$.
4. Obtain the HB estimator of η_d for small area d as the Monte Carlo average, that is

$$\hat{\eta}_d^{HB} = \frac{1}{L} \sum_{h=1}^L \hat{\eta}_d^{(\ell)} \quad (5.61)$$

The uncertainty associated with $\hat{\eta}_d^{HB}$ is estimated by the Monte Carlo approximation of the posterior variance given by

$$\hat{\mathbb{V}}ar_d(\eta_d|\mathbf{y}_s) = \frac{1}{L} \sum_{\ell=1}^L \left(\hat{\eta}_d^{(\ell)} - \hat{\eta}_d^{HB} \right)^2. \quad (5.62)$$

5.3 Simulation Study

For the purpose of evaluating the proposed HB method, we use the FGT poverty measures as the complex parameters of interest (see Section 4.9.1 for a short introduction

to FGT measures). As a reminder, the FGT class is defined, for domain d , as:

$$F_{\alpha d} = \frac{1}{N_d} \sum_{j=1}^{N_d} \left(\frac{z - E_{dj}}{z} \right)^{\alpha} I(E_j < z), \quad j = 1, \dots, N_d, \alpha \geq 0,$$

where z is the poverty line, E_{dj} is a quantitative measure of welfare such as income or expenditure associated with individual j from domain d , and I is an indicator function. We assume that there is a one-to-one relationship between the measure of welfare E_{dj} and y_{dj} i.e. $y_{dj} = T(E_{dj})$. It may be necessary to apply some transformation to the survey characteristics to reduce skewness even when using SN model since SN distributions have a skewness statistics of less than 1 in absolute value. For instance, Elbers et al. (2003) used a logarithm transformation of the unit level consumption and applied the nonparametric ELL method to the transformed variable.

The model (5.46)-(5.47) was considered for this simulation study. It assumes that $u_d \sim N(0, \sigma_u^2)$ and $e_{dj} \sim SN(-\delta_e \sigma_e \sqrt{\frac{2}{\pi}})$. This model is the most relevant since we have seen that skewness of u_d 's distribution has very little effect on the performance of the SAE estimators. The parameters of the model were, for the most part, chosen to be the same as for the simulation study in Molina et al. (2014). One difference is the number of small areas was reduced from 80 to 40 due to excessive simulation time. And the second main difference, was the use of SN distribution for the error terms e_{dj} instead of the normal. In summary, the specifications of the model for the simulation were:

- There were $m = 40$ small areas with $N_d = 250$ units in each area d , $d = 1, \dots, m$.
- Within each small area d , a sample of size $n_d = 50$ was selected using simple random sampling (SRS).
- Two dummy variables X_1 and X_2 were used as auxiliary variables plus an intercept X_0 . The population values of X_k , where $k = 1, 2$ were generated from

Bernoulli distributions $Ber(p_{kd})$ with probabilities of success $p_{1d} = 0.3 + 0.5d/m$ and $p_{2d} = 0.2$, $d = 1, \dots, m$. The auxiliary variables $\mathbf{X}$ are held fixed across the simulated populations.

- The fixed effects were $\boldsymbol{\beta} = (3, 0.03, -0.04)^T$.
- The random area effect u_d followed $N(0, \sigma_u^2)$ where $\sigma_u^2 = 0.15^2$.
- The unit level error e_{dj} followed $SN(-\delta_e \sigma_e \sqrt{\frac{2}{\pi}}, \sigma_e^2, \delta_e)$. The shape parameter λ_e was equal to 3 equivalently $\delta_e = 3/\sqrt{1+3^2} \approx 0.9487$. Note that because of the choice of the location parameter $\mathbb{E}(e_{dj}) = 0$; also σ_e^2 chosen so that $\text{Var}(e_{dj}) = 0.50^2$. The resulting ρ was approximately equals to 8%.
- Use of logarithm transformation i.e. $y_{dj} = \ln(E_{dj})$ equivalently $E_{dj} = \exp(y_{dj})$.
- The poverty line was equal to 12.

In this simulation study a total of $I = 1,000$ populations were generated using the specifications listed above and an SRS sample was selected from each population. For each generated population and using the corresponding sample, the HB estimates were computed using the following scheme:

1. Draw $L = 1,000$ independent samples from $\pi(\mathbf{u}, \boldsymbol{\beta}, \vartheta_e^2, \rho, \delta_e, \delta_u | \mathbf{y}_d)$, defined by (5.39), using the selection strategy laid out in Section 5.2.2 with $H = 10,000$, $R = 1,000$, and $\epsilon = 0.0001$.
2. Draw the out-of-sample values $y_{dj}^{(\ell)}$, $i \in r_d$, conditionally on $\mathbf{y}_s$, $u_1^{(\ell)}, \dots, u_m^{(\ell)}$, $\boldsymbol{\beta}^{(\ell)}$, $\vartheta_e^{2(\ell)}$, $\rho^{(\ell)}$ from the distribution of $y_{dj}^{(\ell)} | \mathbf{y}_s, u_d^{(\ell)}, \boldsymbol{\beta}^{(\ell)}, \vartheta_e^{2(\ell)}, \rho^{(\ell)}$ given by

$$SN(x_{dj}^T \boldsymbol{\beta}^{(\ell)} + u_d^{(\ell)} - \delta_e^{(\ell)} \sigma_e^{(\ell)} \sqrt{\frac{2}{\pi}}, \sigma_e^{2(\ell)}, \delta_e^{(\ell)}), \quad j \in r_d, d = 1, \dots, m.$$

3. Construct the census $y_d^{(\ell)} = ((y_{dr}^{(\ell)})^T, y_{ds}^T)^T$ by combining the out-of-sample predicted values $y_{dr}^{(\ell)}$ to the sample values y_{ds} for each domain $d = 1, \dots, m$. Compute an estimate of the complex parameter for each domain d as follows:

$$\hat{\eta}_d^{(\ell)} = h(\mathbf{y}_d^{(\ell)}).$$

The HB estimate is the Monte Carlo average of the L values $\hat{\eta}_d^{(\ell)}$. For comparison, we also computed the EB estimates for each of the 1,000 populations following the best prediction method described in Section 4.3. The specifications above resulted in an average poverty incidence of about 15% over the 1,000 populations. The small areas with low values of d had an average poverty incidence rates on average a little below 16% while the small areas with high values of d had poverty incidence rates on average a little above 14%.

Figure 5.1 provides a scatter plot of the average of the 1,000 HB estimates of the complex parameters versus the average of the 1,000 corresponding EB estimates across the small areas. The figure shows a strong correlation between the HB estimates and the EB estimates.

Figure 5.1: Average of HB point estimates compared to the average of the EB point estimates (incidence, gap, and severity) with $m = 40$ areas, $n_d = 50$ units, $\lambda_e = 3$, and $\lambda_u = 0$.

Similarly Figure 5.2 provides a scatter plot of the average of the 1,000 HB MC MSEs versus the average of the 1,000 EB MC MSE across the small areas. The figure shows again a strong correlation between the HB and EB MC MSEs even if the correlation seems a little less strong compared to the point estimates above.

Figure 5.2: Average of HB MC MSEs compared to the average of the EB MC MSE (incidence, gap, and severity) with $m = 40$ areas, $n_d = 50$ units, $\lambda_e = 3$, and $\lambda_u = 0$.

Both figures 5.1 and 5.2 suggest that the EB and HB methods have similar performances. HB confidence intervals, often called credible intervals (CI), are easily obtained from the empirical distribution of the posterior. Consider the sample of $L = 1,000$ parameters from the posterior, $\{\boldsymbol{\theta}^{(\ell)}\}_{\ell=1,\dots,L}$, and derive the sample of complex parameters $\{\eta_d^{(\ell)} = h(\boldsymbol{\theta}^{(\ell)})\}_{\ell=1,\dots,L}$ for $d = 1, \dots, m$. We may define two type of CIs. The first type is symmetrical CI which guarantees equal tails of $\alpha/2$ where $1 - \alpha$ is the level of confidence. The second type is a asymmetrical CI which provides the narrowest interval i.e. interval with the highest posterior density (HPD) for unimodal distributions. Figure 5.3 shows the coverage associated with the equal tails and the HPD CIs. Both CIs have coverages closed to 95%. For the poverty incidence, HPD is closer to 95% and the equal tails method seems to overestimate slightly. For poverty gap, the two methods seems to perform similarly for most of the small areas. For poverty severity, HPD seems to overestimate a little.

Figure 5.3: HB coverage for both the equal tails and the HPD CIs (incidence, gap, and severity) with $m = 40$ areas, $n_d = 50$ units, $\lambda_e = 3$, and $\lambda_u = 0$.

Figure 5.4 shows as expected that the HPD method produces the shortest CIs for all three poverty measures.

Figure 5.4: HB widths for both the equal tails and the HPD CIs (incidence, gap, and severity) with $m = 40$ areas, $n_d = 50$ units, $\lambda_e = 3$, and $\lambda_u = 0$.

Figure 5.5 shows a negligible bias of the SN-based HB method for all three poverty measures. However if the skewness of the data is not taking into account, the normality-based HB method is heavily biased in particular for poverty gap and severity.

Figure 5.5: Bias of HB normality-based and HB SN methods(incidence, gap, and severity) with $m = 40$ areas, $n_d = 50$ units, $\lambda_e = 3$, and $\lambda_u = 0$.

The MSE estimates are much smaller for the SN-based HB method comparatively to the normality-based HB method as shown in Figure 5.6. This result was expected as we saw that the SN-based HB method performs similarly to the best predictor under the frequentist paradigm.

Figure 5.6: MSE estimates of HB normality-based and HB SN models (incidence, gap, and severity) with $m = 40$ areas, $n_d = 50$ units, $\lambda_e = 3$, and $\lambda_u = 0$.

One important aspect of the SIR method is the distribution of the importance weights. A good SIR strategy should produce importance weights with minimum variability and avoid very few large weights dominating the others. To develop our

SN-based HB method, we did not try to find the best importance function. Instead we decided to use the normality-based posterior as the importance function and use the resulting importance weights $A(\mathbf{u}^{(\ell_0)}, \boldsymbol{\beta}^{(\ell_0)}, \vartheta_e^{2(\ell_0)}, \rho^{(\ell_0)} | \mathbf{y}_d^{(i)})$ where the sample ℓ_0 has the average design effect (Deff) of the weights. The Deffs were computed using the formula $1 + CV(A(\mathbf{u}, \boldsymbol{\beta}, \vartheta_e^2, \rho | \mathbf{y}_d))^2$ where $CV(A(\mathbf{u}, \boldsymbol{\beta}, \vartheta_e^2, \rho | \mathbf{y}_d))$ is the coefficient of variation of the importance weights.

Figure 5.7: An example of the distribution of the importance weights from the SN-based HB method.

There are few very large weights, many are large weights, and most of them are very small. The Deffs, as shown in Table 5.1 below, are very large. Obviously, the choice of the normal-based posterior as the importance function is far from being optimal. The mean of the Deffs is much higher than the median and the very large IQR indicate important differences in the distribution of the weights across the 1,000 generated population. The third line from Table 5.1 provides the corresponding effective sample sizes (Neffs). Half of the population have an effective sample size below 38. For these populations, if we use the estimator (5.68) then the 10,000 initial sample of parameters would have same variance as a SRS sample of less than 38 parameters. Hence in a real application much larger samples than 10,000 should be used and the Deff should be assessed when choosing the sample sizes.

Table 5.1: Distribution of the design effects of the importance weights across the 1,000 generated populations

	Min	P5	P25	P50	Mean	P75	P95	Max
Deff	16.17	59.19	134.25	258.06	569.89	570.15	2215.49	8309.13
Neff	618	168	74	38	17	17	4	1

In the perfect situation, all the weights would be equal. Hence, a good SIR strategy should produce normalized weights close to $1/10,000 = 0.0001$. For each of the 1,000 populations, we compute the proportion of the 10,000 weights in the interval $(0.00001, 0.001)$. Figure 5.8 shows that the proportion of the weights in the interval varies from less than 10% to close to 60% with an average of about 22% (red line).

Figure 5.8: Proportion of the weights belonging to the interval $(0.00001, 0.001)$ for each of the 1,000 populations.

If we widen the interval to $(0.000001, 0.01)$, the proportions as shown in Figure 5.9 increase to an average of about 48%.

In summary, even with the large importance weights, we saw from the simulation study that the HB estimator is performing similarly to the best empirical predictor which has a large gain over the direct estimator. Hence, the ultimate SAE goal is

Figure 5.9: Proportion of the weights belonging to the interval $(0.000001, 0.01)$ for each of the 1,000 populations.

achieved. However due to the large weights, we recommend for a real application drawing large samples, say 100,000 or more, with small subsample rate, say 5% or less.

5.4 Appendix: Introduction to Sampling Importance Resampling (SIR)

The objective is to obtain the expected value of the parameter $\eta(\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ follows the posterior distribution with pdf $\pi(\boldsymbol{\theta}|\mathbf{y})$. The expected value is defined by the following integral

$$\int \eta(\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} < \infty \quad (5.63)$$

The classical Monte Carlo method, introduced by Metropolis and Ulam (1949), consists of approximating the integral by taking advantage of the fact that $\pi(\boldsymbol{\theta}|\mathbf{y})$ is a pdf. Generate a large sample $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_L$ from $\pi(\boldsymbol{\theta}|\mathbf{y})$ and approximate the integral (5.63)

by the Monte Carlo average

$$\frac{1}{H} \sum_{h=1}^H \eta(\boldsymbol{\theta}_h). \quad (5.64)$$

This approximation is justified by the Law of Large Numbers (LLN) which ensures that the average (5.64) converges to the integral (5.63) when H goes to $+\infty$. Since the posterior $\pi(\boldsymbol{\theta}|\mathbf{y})$ is proportional to $f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$, the expectation calculation reduces to approximating the integral

$$\frac{\int \eta(\boldsymbol{\theta}) f(\mathbf{y}|\boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int f(\mathbf{y}|\boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}}. \quad (5.65)$$

The numerator of (5.65) can be approximated by the Monte Carlo average

$$\frac{1}{H} \sum_{h=1}^H \eta(\boldsymbol{\theta}_h) f(\mathbf{y}|\boldsymbol{\theta}_h). \quad (5.66)$$

where the $\boldsymbol{\theta}_h$'s are drawn from $\pi(\boldsymbol{\theta})$. If the integral (5.63) is unbounded then the posterior is said to be improper and inferences are invalid. The use of proper priors, $\pi(\boldsymbol{\theta})$, guarantees that the posterior, $\pi(\boldsymbol{\theta}|\mathbf{y})$, is proper. However, in many applications, noninformative priors such as Jeffrey's prior or flat priors over an unbounded interval, often improper, are of interest. Such improper priors may not lead to improper posteriors but the "properness" propriety of the posterior is no longer guaranteed and need to be checked. The HB inference is based on the fact that if the posterior variance $\text{Var}(\eta(\boldsymbol{\theta})|\mathbf{y})$ is finite then, according to the Central Theorem Limit (CTL), the Monte Carlo average (5.64) is asymptotically normal with variance $\text{Var}(\eta(\boldsymbol{\theta})|\mathbf{y})/H$.

For some applications, it is not easy to draw from the posterior $\pi(\boldsymbol{\theta}|\mathbf{y})$ or the prior $\pi(\boldsymbol{\theta})$. In such situations, a more general Monte Carlo method called *Sampling Importance Resampling* (SIR) can be applied. It consists of noting that the integral (5.63) can be written as

$$\int \eta(\boldsymbol{\theta}) \frac{\pi(\boldsymbol{\theta}|\mathbf{y})}{g(\boldsymbol{\theta})} g(\boldsymbol{\theta}) d\boldsymbol{\theta} \quad (5.67)$$

where the importance function $g(\boldsymbol{\theta})$, also called the sampling distribution, is a probability density function with $\text{supp}(g)$ including the support of the posterior pdf $\pi(\boldsymbol{\theta}|\mathbf{y})$. The quantity $w_h = w(\boldsymbol{\theta}^{(h)}) = \frac{\pi(\boldsymbol{\theta}^{(h)}|\mathbf{y})}{g(\boldsymbol{\theta}^{(h)})}$ is the weight to account for the difference between the posterior density $\pi(\boldsymbol{\theta}|\mathbf{y})$ and the importance function g . In the application of the SN model from Section 5.2.2, the quantity $A(\boldsymbol{\theta}^{(h)})$ defined by (5.35) is the equivalent of w_h defined in this paragraph. Also, $g(\boldsymbol{\theta})$ from this section is the equivalent of $\pi_{1a}(\boldsymbol{\theta})$ defined by (5.34) from the SN model in Section 5.2.2.

Following the original Monte Carlo method, we generate a large sample $\{\boldsymbol{\theta}^{(h)}\}_{h=1,\dots,H}$ from g and approximate the integral (5.67) by the Monte Carlo average

$$\frac{1}{H} \sum_{h=1}^H \eta(\boldsymbol{\theta}^{(h)}) w_h. \quad (5.68)$$

Also, an approximation of the expectation (5.63) is

$$\frac{\sum_{h=1}^H \eta(\boldsymbol{\theta}^{(h)}) w_h}{\sum_{h=1}^H w_h}, \quad (5.69)$$

since the numerator converges to $\int \eta(\boldsymbol{\theta}) f(\mathbf{y}|\boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}$ and the denominator converges to $\int f(\mathbf{y}|\boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}$. Note that the ratio (5.69) does not depend on the normalizing constants. The variance of estimator 5.68 is given by

$$\int \left(\eta(\boldsymbol{\theta}) w(\boldsymbol{\theta}) - \int \eta(\boldsymbol{\theta}) w(\boldsymbol{\theta}) g(\boldsymbol{\theta}) d\boldsymbol{\theta} \right)^2 g(\boldsymbol{\theta}) d\boldsymbol{\theta} = \mathbb{E}_g \left(\eta(\boldsymbol{\theta})^2 w(\boldsymbol{\theta})^2 \right) - \mathbb{E}_g^2 \left(\eta(\boldsymbol{\theta}) w(\boldsymbol{\theta}) \right). \quad (5.70)$$

A natural estimator of the variance is

$$\frac{1}{H} \sum_{h=1}^H \eta(\boldsymbol{\theta}^{(h)})^2 w_h^2 - \left(\frac{1}{H} \sum_{h=1}^H \eta(\boldsymbol{\theta}^{(h)}) w_h \right)^2. \quad (5.71)$$

From the left side of the equality (5.70), it follows that the optimal choice of $g(\boldsymbol{\theta})$,

resulting in a variance equal to 0, is

$$g^*(\boldsymbol{\theta}) = \frac{\eta(\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y})}{\mathbb{E}_{\pi(\boldsymbol{\theta}|\mathbf{y})}(\eta(\boldsymbol{\theta}))} = \eta(\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y}) / \int \eta(\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} \quad (5.72)$$

The optimal solution (5.72) is not useful in practice since it depends on the integral $\int \eta(\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}$ we are trying to approximate. Therefore, in the Monte Carlo setup, there is no ultimate best importance function. The goal is to choose an importance function g tailored to the approximation of the posterior $\pi(\boldsymbol{\theta}|\mathbf{y})$ by balancing ease to draw from g and closeness of g to $\eta(\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y})$.

In the implementation of the SIR approach for the simulation study from Section 5.3 we did not use estimator (5.68) instead, we computed (5.64) where the parameters $\boldsymbol{\theta}$ were drawn in two steps. First, a large sample of parameters were drawn using the importance function $g(\boldsymbol{\theta})$ then in a second step a smaller subsample of parameters were drawn proportionally to the importance weights $w(\boldsymbol{\theta}^{(h)})$. In survey sampling, this importance sampling method is called *two-phase sampling* or *double sampling*. We may write the variance of this estimator as follows

$$\mathbb{V}ar(\eta(\boldsymbol{\theta})) = \mathbb{V}ar_1(\mathbb{E}_2(\eta(\boldsymbol{\theta})|S_1)) + \mathbb{E}_1(\mathbb{V}ar_2(\eta(\boldsymbol{\theta})|S_1)) \quad (5.73)$$

where S_1 is the selected sample at the first phase. Expression (5.73) does not have a closed form because of the complexity of $\eta(\boldsymbol{\theta})$ but it can be numerically evaluated. The term $\mathbb{V}ar_1(\mathbb{E}_2(\eta(\boldsymbol{\theta})|S_1))$ is the variance if only the first phase was used to estimate the complex parameter. And the term $\mathbb{E}_1(\mathbb{V}ar_2(\eta(\boldsymbol{\theta})|S_1))$ is the addition due to the subsampling at the second phase. Highly unequal importance weights $w(\boldsymbol{\theta}^{(h)})$ will result in a very inefficient second phase sampling strategy and large values of $\mathbb{V}ar_2(\eta(\boldsymbol{\theta})|S_1)$.

Chapter 6

Two-Fold Model under Skew-Normal Errors

In this chapter, we consider the two-fold nested error regression model with random errors following skew-normal (SN) distributions for estimating small area means and linear function of the means (linear parameters). The description of the two-fold model is provided in Section 6.1 and the prediction of small area means (equivalently small area linear parameters) are shown in Section 6.2.

6.1 Two-Fold Nested Error Model

The model considered in Chapter 4 is referred to as the one-fold nested error model since only one aggregated level, the small area, is modeled. However, in many applications, it may be of interest to incorporate additional aggregated levels in the model to account for extra variability or to reflect the sampling design. In complex survey sampling, often to reduce cost, the sample is selected in stages where at the first level primary sampling units (PSUs) are selected, within each PSU secondary sampling units (SSUs) are selected and finally individual units are selected in SSUs. Fuller and Battese (1973) proposed a two-fold model that can be used to model data

from such complex design in order to capture variability from both the PSU and SSU levels. Later, Stukel and Rao (1997) extend the method to unbalanced data under normally distributed errors. In the following, we extend the model further to account for random errors following SN distributions. The two-fold nested error model is formally defined as

$$Y_{dj k} = \mathbf{x}_{dj k}^T \boldsymbol{\beta} + u_d + v_{dj} + e_{dj k}; \quad d = 1, \dots, m; j = 1, \dots, M_d; k = 1, \dots, N_{dj}. \quad (6.1)$$

The value $y_{dj k}$ is the observed characteristic of interest associated to unit k from cluster j within small area d . The auxiliary information $\mathbf{x}_{dj k}^T = (x_{dj k1}, \dots, x_{dj kp})$ is a $1 \times p$ vector of known variables and therefore $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown regression parameters. If an intercept is necessary then consider $x_{dj k1}$ to be equal to 1 for all the records. The three random errors u_d , v_{dj} , and $e_{dj k}$ are independent with $\mathbb{E}(u_d) = \mathbb{E}(v_{dj}) = \mathbb{E}(e_{dj k}) = 0$. We assume that u_d , v_{dj} , and $e_{dj k}$ follow SN distributions, that is

$$u_d \sim SN(\mu_u, \sigma_u^2, \lambda_u); \quad v_{dj} \sim SN(\mu_v, \sigma_v^2, \lambda_v); \quad \text{and} \quad e_{dj k} \sim SN(\mu_e, \sigma_e^2, \lambda_e). \quad (6.2)$$

Note that, as a general result, if $Z \sim SN(\mu, \sigma^2, \lambda)$ then $\mathbb{E}(Z) = \mu + \sigma \delta \sqrt{\frac{2}{\pi}}$ and $\mathbb{V}ar(Z) = \sigma^2 \left(1 - \frac{2}{\pi} \delta\right)^2$ where $\delta = \frac{\lambda}{\sqrt{1+\lambda^2}}$. Hence, to ensure that the means of the random errors are equal to zero, we define μ_u , μ_v , and μ_e from (6.2) as follows

$$\mu_u = -\sigma_u \delta_u \sqrt{\frac{2}{\pi}}; \quad \mu_v = -\sigma_v \delta_v \sqrt{\frac{2}{\pi}}; \quad \text{and} \quad \mu_e = -\sigma_e \delta_e \sqrt{\frac{2}{\pi}}. \quad (6.3)$$

The variances of the random errors are

$$\mathbb{V}ar(u_d) = \vartheta_u = \sigma_u^2 \left(1 - \frac{2}{\pi} \delta_u\right)^2, \quad (6.4)$$

$$\mathbb{V}ar(v_{dj}) = \vartheta_v = \sigma_v^2 \left(1 - \frac{2}{\pi} \delta_v\right)^2, \quad (6.5)$$

$$\mathbb{V}ar(e_{dj}) = \vartheta_e = \sigma_e^2 \left(1 - \frac{2}{\pi} \delta_e\right)^2. \quad (6.6)$$

The two-fold nested error model (6.1) is a special case of the linear mixed model (3.1) with

$$\mathbf{u}_d = \begin{pmatrix} u_d \\ v_{d1} \\ \vdots \\ v_{dM_d} \end{pmatrix} \quad \text{and} \quad \mathbf{Z}_d = (\mathbf{Z}_{du} | \mathbf{Z}_{dv}) = \begin{pmatrix} \mathbf{1}_{d1} & \mathbf{1}_{d1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{1}_{d2} & \mathbf{0} & \mathbf{1}_{d2} & \mathbf{0} & \mathbf{0} \\ \vdots & \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{1}_{dM_d} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1}_{dM_d} \end{pmatrix} \quad (6.7)$$

where $\mathbf{1}_{dj}$ indicates a column vector of N_{dj} ones. The covariance matrix of the random components is

$$\mathbb{V}ar \begin{pmatrix} \mathbf{u}_d \\ \mathbf{e}_d \end{pmatrix} = \begin{pmatrix} \vartheta_u & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \vartheta_v \mathbf{I}_{M_d} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \vartheta_e \mathbf{I}_{N_d} \end{pmatrix} \quad (6.8)$$

where $\mathbf{I}_{M_d}$ and $\mathbf{I}_{N_d}$ are identity matrices of order M_d and $N_d = \sum_j N_{dj}$ respectively.

By analogy to the notation adopted by Rao (2003), we define

$$\mathbf{G}_d = \begin{pmatrix} \vartheta_u & \mathbf{0} \\ \mathbf{0} & \vartheta_v \mathbf{I}_{M_d} \end{pmatrix} \quad \text{and} \quad \mathbf{R}_d = \vartheta_e \mathbf{I}_{N_d} \quad \text{such that} \quad \mathbf{V}_d = \mathbf{R}_e + \mathbf{Z}_d \mathbf{G}_d \mathbf{Z}_d^T \quad (6.9)$$

That is, the covariance matrix of the characteristic of interest $\mathbf{Y}_d$ is

$$\mathbb{V}ar(\mathbf{Y}_d) = \mathbf{V}_d = \vartheta_e \mathbf{I}_{N_d} + \vartheta_v \text{BlockDiag}(\mathbf{1}_{N_{dj}} \mathbf{1}_{N_{dj}}^T)_{1 \leq j \leq t_d} + \vartheta_u \mathbf{1}_{N_d} \mathbf{1}_{N_d}^T \quad (6.10)$$

Note that

$$\begin{aligned}
 \text{BlockDiag}(\mathbf{1}_{N_{dj}} \mathbf{1}_{N_{dj}}^T)_{1 \leq j \leq M_d} &= \begin{pmatrix} \mathbf{1}_{N_{d1}} \mathbf{1}_{N_{d1}}^T & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{1}_{N_d M_d} \mathbf{1}_{n_d M_d}^T \end{pmatrix} \\
 &= \bigoplus_{j=1}^{M_d} \mathbf{1}_{N_{dj}} \mathbf{1}_{N_{dj}}^T \\
 &= \mathbf{1}_{N_{d1}} \mathbf{1}_{N_{d1}}^T \oplus \cdots \oplus \mathbf{1}_{N_d M_d} \mathbf{1}_{N_d M_d}^T
 \end{aligned}$$

If the design is balanced (i.e. $N_{dj} = \bar{N}_d$ for all j in area d),

$$\text{BlockDiag}(\mathbf{1}_{N_{dj}} \mathbf{1}_{N_{dj}}^T)_{1 \leq j \leq M_d} = \bigotimes_{j=1}^{M_d} \mathbf{1}_{\bar{N}_d} \mathbf{1}_{\bar{N}_d}^T$$

The covariance matrix (6.10) can be rewritten as follow

$$\text{Cov}(Y_{dj k}, Y_{d' j' k'}) = \begin{cases} \vartheta_e^2 + \vartheta_v^2 + \vartheta_u^2, & d = d', j = j', k = k'; \\ \vartheta_v^2 + \vartheta_u^2, & d = d', j = j', k \neq k'; \\ \vartheta_u^2, & d = d', j \neq j', k \neq k'. \end{cases}$$

Note that the covariance between two units from two different small areas is equal to zero. Also when $\lambda_u = \lambda_v = \lambda_e = 0$, we get the usual results from the nested error model with normally distributed random errors.

6.2 Prediction of Small Area Means

As mentioned earlier, the parameter of interest is the small area mean (equivalently total) that is

$$\bar{Y}_d = \frac{1}{N_d} \sum_{j=1}^{M_d} \sum_{k=1}^{N_{dj}} Y_{dj k} = \bar{\mathbf{X}}_d^T \boldsymbol{\beta} + u_d + \bar{\mathbf{Z}}_{dv}^T \mathbf{v}_d + \bar{E}_d \quad (6.11)$$

where $\bar{\mathbf{X}}_d = \sum_{j=1}^{M_d} \sum_{k=1}^{N_{dj}} \mathbf{X}_{dj k}$ and $\bar{\mathbf{Z}}_{dv}^T = \left(\frac{N_{d1}}{\sum_j N_{dj}}, \frac{N_{d2}}{\sum_j N_{dj}}, \dots, \frac{N_{dM_d}}{\sum_j N_{dj}} \right)$. There are at least two possible approaches for estimating the small area mean (6.11).

The first approach consists of noting that, from the law of large numbers, it follows that $\bar{E}_d \approx 0$ hence $\bar{Y}_d \approx \bar{\mathbf{X}}_d \boldsymbol{\beta} + u_d + \bar{\mathbf{Z}}_{dv}^T \mathbf{v}_d$. This approximation holds even if the errors are not normally distributed but independent and identically distributed (i.i.d) with mean zero. It is then necessary to choose μ_u , μ_v , and μ_e from the SN errors so that the errors have means equal to zero. Therefore, estimating the small area mean reduces to estimating a linear combination of the fixed and random areas (both small area and sub-small areas) effects. That is the parameter of interest for prediction is η_d defined as

$$\eta_d = \bar{\mathbf{X}}_d \boldsymbol{\beta} + u_d + \bar{\mathbf{Z}}_{dv}^T \mathbf{v}_d. \quad (6.12)$$

Note that this parameter is a special case of the more general parameter $\mathbf{l}_d^T \boldsymbol{\beta} + \mathbf{m}_d^T \mathbf{u}_d$ discussed in Rao (2003) with $\mathbf{l}_d^T = \bar{\mathbf{X}}_d$ and $\mathbf{m}_d^T = (1, \bar{\mathbf{Z}}_{dv}^T)$. Following this first approach, the small area mean is estimated by the best predictor η_d^{BP} of η_d defined as

$$\eta_d^{BP} = \bar{\mathbf{X}}_d \boldsymbol{\beta} + \hat{u}_d^{BP} + \bar{\mathbf{Z}}_{dv}^T \hat{\mathbf{v}}_d^{BP} \quad (6.13)$$

where $\hat{u}_d^{BP} = \mathbb{E}(u_d | \mathbf{y}_d)$ and $\hat{\mathbf{v}}_d^{BP} = \mathbb{E}(\mathbf{v}_d | \mathbf{y}_d)$. If the number of clusters M_d in the population area d is large then $\bar{\mathbf{Z}}_{dv}^T \mathbf{v}_d \approx 0$ leading to $\eta_d \approx \bar{\mathbf{X}}_d \boldsymbol{\beta} + u_d$.

The second approach for estimating the small area mean consists of predicting the non-observed characteristic and use them to obtain the predicted means. Note that the mean of the characteristic of interest for the small area d can be written as

$$\bar{Y}_d = \frac{1}{N_d} \left[\sum_{j \in s_d} \sum_{k \in s_{dj}} y_{dj k} + \sum_{j \in s_d} \left(\sum_{k \in s_{dj}^c} y_{dj k} \right) + \sum_{j \in s_d^c} \left(\sum_{k=1}^{N_{dj}} y_{dj k} \right) \right]. \quad (6.14)$$

Finding the best predictor of the small area mean reduces to deriving the best predictors

of $y_{dj k}$. In expression (6.14), the pool of units were divided in three groups: sampled units ($j \in s_d$ and $k \in s_{dj}$), the non-sampled units in sampled clusters ($j \in s_d$ and $k \in s_{dj}^c$), and the non-sampled units in non-sampled clusters (all k from $j \in s_d^c$). The best predictor of the observed characteristic, from a sampled unit, is the characteristic itself. The goal then is to find the best predictors for the non-observed characteristic for the units in selected clusters ($j \in s_d$) as well as for the units in the non-selected clusters ($j \in s_d^c$). In this chapter, we focuss on this second approach for predicting small area means. Stukel and Rao (1999) provide the best linear unbiased prediction (BLUP) for the unbalanced two-fold nested model under normally distributed errors. However, under random errors following SN distribution, the BLUP and BP estimators are different. In the following, we derive both the BLUP and BP estimators for the two-fold SN errors model in the context where m_d of M_d clusters are sampled.

6.2.1 Best Linear Unbiased Prediction (BLUP)

Following Stukel and Rao (1999), it is straightforward to show that the BLUP estimator of the small area mean $\bar{Y}_d$ defined by (6.14), under errors following SN, is

$$\hat{\bar{Y}}_d^{BLUP} = \frac{1}{N_d} \left[\sum_{j \in s_d} \sum_{k \in s_{dj}} y_{dj k} + \sum_{j \in s_d} \left(\sum_{k \in s_{dj}^c} \hat{y}_{dj k}^{BLUP} \right) + \sum_{j \in s_d^c} \left(\sum_{k=1}^{N_{dj}} \hat{y}_{dj k}^{BLUP} \right) \right] \quad (6.15)$$

where s_{dj}^c is the set of nonsampled units in the j^{th} sampled clusters and s_d^c is the set of nonsampled sampled clusters in small area d . The predictors $\hat{y}_{dj k}^{BLUP}$ and $\hat{\bar{y}}_{dj k}^{BLUP}$, from (6.15), are defined as follows

$$\hat{y}_{dj k}^{BLUP} = \mathbf{x}_{dj k}^T \hat{\boldsymbol{\beta}}^{BLUP} + \hat{u}_d^{BLUP} + \hat{v}_{dj}^{BLUP} \quad (6.16)$$

and

$$\hat{y}_{djk}^{BLUP} = \mathbf{x}_{djk}^T \hat{\boldsymbol{\beta}}^{BLUP} + \hat{u}_d^{BLUP} \quad (6.17)$$

where $\hat{\boldsymbol{\beta}}^{BLUP} = (\sum_{d=1}^m \mathbf{X}_d^T \mathbf{V}_d^{-1} \mathbf{X}_d)^{-1} (\sum_{d=1}^m \mathbf{X}_d^T \mathbf{V}_d^{-1} \mathbf{y}_d)$ is the generalized least squares estimator of $\boldsymbol{\beta}$; also the random area and clusters effects are given by

$$\hat{\mathbf{u}}_d^{BLUP} = \begin{pmatrix} \hat{u}_d^{BLUP} \\ \hat{v}_{dj}^{BLUP} \end{pmatrix} = \mathbf{G}_d \mathbf{Z}_d^T \mathbf{V}_d^{-1} (\mathbf{y}_d - \mathbf{X}_d \hat{\boldsymbol{\beta}}^{BLUP}). \quad (6.18)$$

The proof of the “best” property among all the linear estimators, provided by Stukel and Rao (1999), does not assume any particular distribution. The same proof applies to the estimator (6.15) for proving its “best” property under SN distribution of the random terms.

To find the BLUP of the random area and clusters effects, defined by (6.18), one needs to invert the covariance matrix $\mathbf{V}_d$. Note that the matrix $\mathbf{V}_d$ can be written as $\mathbf{V}_d = \mathbf{A} + \mathbf{bc}^T$ where $\mathbf{A} = \vartheta_e \mathbf{I}_{n_d} + \vartheta_v \text{BlockDiag}(\mathbf{1}_{n_{dj}} \mathbf{1}_{n_{dj}}^T)_{1 \leq j \leq t_d}$, $\mathbf{b} = \vartheta_u \mathbf{1}_{n_d}$, and $\mathbf{c} = \mathbf{1}_{n_d}$. Since $1 + \mathbf{c}^T \mathbf{A}^{-1} \mathbf{b} \neq 0$, the Sherman-Morrison formula, $(\mathbf{A} + \mathbf{bc}^T)^{-1} = \mathbf{A}^{-1} - \frac{1}{1 + \mathbf{c}^T \mathbf{A}^{-1} \mathbf{b}} \mathbf{A}^{-1} \mathbf{bc}^T \mathbf{A}^{-1}$, leads to

$$\begin{aligned} \mathbf{V}_d^{-1} = & \frac{1}{\vartheta_e} \left(\mathbf{I}_{n_d} - \text{BlockDiag} \left(\frac{\gamma_{dj}}{n_{dj}} \mathbf{1}_{n_{dj}} \mathbf{1}_{n_{dj}}^T \right)_{1 \leq j \leq m_d} \right. \\ & \left. - \frac{\vartheta_e}{\vartheta_v} \left(\frac{\gamma_d}{\sum_j \gamma_{dj}} \right) \left[\begin{matrix} \text{col} \\ 1 \leq j \leq m_d \end{matrix} \left(\frac{\gamma_{dj}}{n_{dj}} \mathbf{1}_{n_{dj}} \right) \right] \left[\begin{matrix} \text{col} \\ 1 \leq j \leq m_d \end{matrix} \left(\frac{\gamma_{dj}}{n_{dj}} \mathbf{1}_{n_{dj}} \right) \right]^T \right) \end{aligned} \quad (6.19)$$

where n_{dj} is the sample size of cluster j in small area d , $\gamma_{dj} = \vartheta_v \left(\vartheta_v + \frac{\vartheta_e}{n_{dj}} \right)^{-1}$ and $\gamma_d = \vartheta_u \left(\vartheta_u + \frac{\vartheta_v}{\sum_j \gamma_{dj}} \right)^{-1}$. The resulting closed form expression of $\hat{\boldsymbol{\beta}}^{BLUP}$ is then given

by

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{BLUP} &= \left(\sum_d \sum_j \sum_k \mathbf{x}_{djk} \mathbf{x}_{djk}^T - \sum_d \sum_j \left(n_{dj} \gamma_{dj} + \frac{\vartheta_e}{\vartheta_v} \frac{\gamma_d}{\sum_j \gamma_{dj}} \right) \bar{\mathbf{x}}_{dj} \bar{\mathbf{x}}_{dj}^T \right)^{-1} \\ &\times \left(\sum_d \sum_j \sum_k \mathbf{x}_{djk} y_{djk} - \sum_d \sum_j \left(n_{dj} \gamma_{dj} + \frac{\vartheta_e}{\vartheta_v} \frac{\gamma_d}{\sum_j \gamma_{dj}} \right) \bar{\mathbf{x}}_{dj} \bar{y}_{dj} \right) \end{aligned} \quad (6.20)$$

where $\bar{\mathbf{x}}_{dj} = \sum_k \frac{\mathbf{x}_{djk}}{n_{dj}}$ and $\bar{y}_{dj} = \sum_k \frac{y_{djk}}{n_{dj}}$. Also, the predictors $\hat{u}_d^{BLUP}$ and $\hat{v}_{dj}^{BLUP}$ reduce to respectively

$$\hat{u}_d^{BLUP} = \gamma_d \sum_j \frac{\gamma_{dj}}{\sum_j \gamma_{dj}} \left(\bar{y}_{dj} - \bar{\mathbf{x}}_{dj}^T \hat{\boldsymbol{\beta}}^{BLUP} \right) \quad (6.21)$$

and

$$\hat{v}_{dj}^{BLUP} = \gamma_{dj} \left[\left(\bar{y}_{dj} - \bar{\mathbf{x}}_{dj}^T \hat{\boldsymbol{\beta}}^{BLUP} \right) - \hat{u}_d^{BLUP} \right]. \quad (6.22)$$

6.2.2 Best Prediction (BP)

Under the two-fold nested model with errors following SN, the BLUP estimator is only the “best” among the linear predictor; that is the BLUP is not the same as the BP. Theorem 6.1 gives the form of the BP which is similar to expression (6.15) where the BLUP predictors are replaced by their BP equivalent.

Theorem 6.1. *Under the two-fold nested error model (6.1,6.2), the best predictor (BP) of the small area mean $\bar{Y}_d$, $d=1, \dots, m$, is given by*

$$\hat{Y}_d^{BP} = \frac{1}{N_d} \left[\sum_{j \in s_d} \sum_{k \in s_{dj}} y_{djk} + \sum_{j \in s_d} \left(\sum_{k \in s_{dj}^c} \hat{y}_{djk}^{BP} \right) + \sum_{j \in s_d^c} \left(\sum_{k=1}^{N_{dj}} \hat{y}_{djk}^{BP} \right) \right] \quad (6.23)$$

where s_{dj}^c is the set of nonsampled units in the j^{th} sampled clusters and s_d^c is the set of nonsampled sampled clusters in small area d . Also, the predictors $\hat{y}_{djk}^{BP}$ and $\hat{y}_{djk}^{BP}$, from

(6.23), are defined as follows

$$\hat{y}_{djk}^{BP} = \mathbf{x}_{djk}^T \boldsymbol{\beta} + \hat{u}_d^{BP} + \hat{v}_{dj}^{BP} \quad (6.24)$$

and

$$\hat{y}_{djk}^{BP} = \mathbf{x}_{djk}^T \boldsymbol{\beta} + \hat{u}_d^{BP} \quad (6.25)$$

where $\hat{u}_d^{BP} = \mathbb{E}(u_d | \mathbf{y}_d)$ and $\hat{v}_{dj}^{BP} = \mathbb{E}(v_{dj} | \mathbf{y}_d)$

Proof. The BP estimator of $\bar{Y}_d$ is given by $\bar{Y}_d^{BP} = \mathbb{E}(\bar{Y}_d | \mathbf{y}_d)$ where $\mathbf{y}_d$ is the observed characteristic of interest. To see this, consider another estimator $\hat{\bar{Y}}_d$ of $\bar{Y}_d$ function of $\mathbf{y}_{ds}$ then

$$\begin{aligned} MSE(\hat{\bar{Y}}_d) &= \mathbb{E} \left(\hat{\bar{Y}}_d - \bar{Y}_d \right)^2 \\ &= \mathbb{E} \left(\hat{\bar{Y}}_d - \bar{Y}_d^{BP} + \bar{Y}_d^{BP} - \bar{Y}_d \right)^2 \\ &= \mathbb{E} \left(\hat{\bar{Y}}_d - \bar{Y}_d^{BP} \right)^2 + \mathbb{E} \left(\bar{Y}_d^{BP} - \bar{Y}_d \right)^2 + 2\mathbb{E} \left(\hat{\bar{Y}}_d - \bar{Y}_d^{BP} \right) \left(\bar{Y}_d^{BP} - \bar{Y}_d \right) \\ &= MSE(\bar{Y}_d^{BP}) + \mathbb{E} \left(\hat{\bar{Y}}_d - \bar{Y}_d^{BP} \right)^2 + 2\mathbb{E} \left(\hat{\bar{Y}}_d - \bar{Y}_d^{BP} \right) \left(\bar{Y}_d^{BP} - \bar{Y}_d \right). \end{aligned}$$

Since $\mathbb{E} \left(\hat{\bar{Y}}_d - \bar{Y}_d^{BP} \right)^2$ is positive, it suffices to show that $\mathbb{E} \left(\hat{\bar{Y}}_d - \bar{Y}_d^{BP} \right) \left(\bar{Y}_d^{BP} - \bar{Y}_d \right) = 0$. Note that $\hat{\bar{Y}}_d - \bar{Y}_d^{BP}$ is a function of $\mathbf{y}_d$ say $h(\mathbf{y}_d)$, it follows that

$$\begin{aligned} \mathbb{E} \left(\hat{\bar{Y}}_d - \bar{Y}_d^{BP} \right) \left(\bar{Y}_d^{BP} - \bar{Y}_d \right) &= \mathbb{E} \left[h(\mathbf{y}_d) \left(\bar{Y}_d^{BP} - \bar{Y}_d \right) \right] \\ &= \mathbb{E} \left[h(\mathbf{y}_d) \bar{Y}_d^{BP} \right] - \mathbb{E} \left[h(\mathbf{y}_d) \bar{Y}_d \right] \\ &= \mathbb{E} \left[h(\mathbf{y}_d) \mathbb{E}(\bar{Y}_d | \mathbf{y}_d) \right] - \mathbb{E} \left[h(\mathbf{y}_d) \bar{Y}_d \right] \\ &= \mathbb{E} \left[h(\mathbf{y}_d) \bar{Y}_d \right] - \mathbb{E} \left[h(\mathbf{y}_d) \bar{Y}_d \right] \\ &= 0. \end{aligned} \quad (6.26)$$

The equality (6.26) is the direct application of the following conditional expectation

property $\mathbb{E}[h(X)\mathbb{E}(Y|X)] = \mathbb{E}[h(X)Y]$. At this point, we showed that $MSE(\hat{\bar{Y}}_d) \geq MSE(\bar{Y}_d^{BP})$. This means that $\bar{Y}_d^{BP}$ is the “best”, in terms of MSE, among all the estimators (function of $\mathbf{y}_d$) of the small area mean $\bar{Y}_d$. Now, it remains to derive the form of the best predictor as shown in (6.23) with the expressions (6.24) and (6.25). The best predictor $\bar{Y}_d^{BP}$ can be written as:

$$\begin{aligned}\bar{Y}_d^{BP} &= \mathbb{E}(\bar{Y}_d|\mathbf{y}_d) \\ &= \frac{1}{N_d} E \left[\sum_{j \in s_d} \sum_{k \in s_{dj}} y_{dj k} + \sum_{j \in s_d} \sum_{k \in s_{dj}^c} y_{dj k} + \sum_{j \in s_d} \sum_{k=1}^{N_{dj}} y_{dj k} | \mathbf{y}_d \right] \\ &= \frac{1}{N_d} \left[\sum_{j \in s_d} \sum_{k \in s_{dj}} \mathbb{E}(y_{dj k} | \mathbf{y}_d) + \sum_{j \in s_d} \sum_{k \in s_{dj}^c} \mathbb{E}(y_{dj k} | \mathbf{y}_d) + \sum_{j \in s_d} \sum_{k=1}^{N_{dj}} \mathbb{E}(y_{dj k} | \mathbf{y}_d) \right].\end{aligned}$$

The expression of $\mathbb{E}(y_{dj k} | \mathbf{y}_d)$ is different for the three groups: sampled units ($j \in s_d$ and $k \in s_{dj}$), the non-sampled units in sampled clusters ($j \in s_d$ and $k \in s_{dj}^c$), and the non-sampled units in non-sampled clusters (all k from $j \in s_d^c$).

- For $j \in s_d$ and $k \in s_{dj}$, we get $\mathbb{E}(y_{dj k} | \mathbf{y}_d) = y_{dj k}$ since $y_{dj k}$ is observed.
- For $j \in s_d$ and $k \in s_{dj}^c$, we get $\mathbb{E}(y_{dj k} | \mathbf{y}_d) = \mathbf{x}_{dj k}^T \boldsymbol{\beta} + \mathbb{E}(u_d | \mathbf{y}_d) + \mathbb{E}(v_{dj} | \mathbf{y}_d) + \mathbb{E}(e_{dj k} | \mathbf{y}_d)$. Since $e_{dj k}$ is independent of $\mathbf{y}_d$, it follows that $\mathbb{E}(e_{dj k} | \mathbf{y}_d) = \mathbb{E}(e_{dj k}) = 0$. Therefore, for units in this group, we have $\mathbb{E}(y_{dj k} | \mathbf{y}_d) = \mathbf{x}_{dj k}^T \boldsymbol{\beta} + \hat{u}_d^{BP} + \hat{v}_{dj}^{BP}$.
- For the last group $j \in s_d^c$, we get $\mathbb{E}(y_{dj k} | \mathbf{y}_d) = \mathbf{x}_{dj k}^T \boldsymbol{\beta} + \mathbb{E}(u_d | \mathbf{y}_d) + \mathbb{E}(v_{dj} | \mathbf{y}_d) + \mathbb{E}(e_{dj k} | \mathbf{y}_d)$. Since $e_{dj k}$ and v_{dj} are both independent of $\mathbf{y}_d$, it follows that $\mathbb{E}(e_{dj k} | \mathbf{y}_d) = \mathbb{E}(e_{dj k}) = 0$ and $\mathbb{E}(v_{dj} | \mathbf{y}_d) = \mathbb{E}(v_{dj}) = 0$. Therefore, for units in this group, we have $\mathbb{E}(y_{dj k} | \mathbf{y}_d) = \mathbf{x}_{dj k}^T \boldsymbol{\beta} + \hat{u}_d^{BP}$.

□

To fully specify the best predictor (6.23) under errors following SN distribution, it is necessary to derive the conditional expectation of the random area and cluster effects given the observed characteristic that is $\mathbb{E}(\mathbf{u}_d|\mathbf{y}_d)$. Theorem 6.2 provides the best predictor of the random effects for the general linear mixed model.

Theorem 6.2. *Consider the linear mixed model with closed skew-normal (CSN) random components i.e. $\mathbf{Y}_d = \mathbf{X}_d\boldsymbol{\beta} + \mathbf{Z}_d\mathbf{u}_d + \mathbf{e}_d$, $d = 1, \dots, m$, where $\mathbf{u}_d \sim \text{CSN}_q(\boldsymbol{\mu}_{u_d}, \boldsymbol{\Sigma}_{u_d}, \mathbf{D}_{u_d}, \boldsymbol{\nu}_{u_d}, \boldsymbol{\Gamma}_{u_d})$, $\mathbf{e}_d \sim \text{CSN}_{N_d}(\boldsymbol{\mu}_{e_d}, \boldsymbol{\Sigma}_{e_d}, \mathbf{D}_{e_d}, \boldsymbol{\nu}_{e_d}, \boldsymbol{\Gamma}_{e_d})$, $\mathbf{u}_d$ and $\mathbf{e}_d$ are independent. $\mathbf{X}_d$ is a $N_d \times p$ matrix, $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects, $\mathbf{Z}_d$ is a $N_d \times q$ matrix, $\mathbf{u}_d$ is a $q \times 1$ vector of area-random effects, and $\mathbf{e}_d = (e_{d1}, \dots, e_{dN_d})^T$ is a $N_d \times 1$ vector of random errors. Then, the BP of the random effect is the expected value of $\mathbf{u}_d$ given the data $\mathbf{y}_d$ given by*

$$\hat{\mathbf{u}}_d^{BP} = \boldsymbol{\mu}_{u_d} + \boldsymbol{\Sigma}_{u_d} \mathbf{Z}_d^T \boldsymbol{\Sigma}_d^{-1} (\mathbf{y}_d - \boldsymbol{\mu}_{y_d}) + \frac{\Phi_{N_d+q}^{(1)}(\mathbf{0}; \mathbf{D}_{y_d} \boldsymbol{\Sigma}_d^{-1} (\mathbf{y}_d - \boldsymbol{\mu}_{y_d}), \boldsymbol{\Gamma}_{y_d})}{\Phi_{N_d+q}(\mathbf{0}; \mathbf{D}_{y_d} \boldsymbol{\Sigma}_d^{-1} (\mathbf{y}_d - \boldsymbol{\mu}_{y_d}), \boldsymbol{\Gamma}_{y_d})} \quad (6.27)$$

where

$$\begin{aligned} \boldsymbol{\mu}_{y_d} &= \mathbf{X}_d \boldsymbol{\beta} + \mathbf{Z}_d \boldsymbol{\mu}_{u_d} + \boldsymbol{\mu}_{e_d}, \boldsymbol{\Sigma}_d = \boldsymbol{\Sigma}_{e_d} + \mathbf{Z}_d \boldsymbol{\Sigma}_{u_d} \mathbf{Z}_d^T, \mathbf{D}_{y_d} = \begin{pmatrix} \mathbf{D}_{u_d} \boldsymbol{\Sigma}_{u_d} \mathbf{Z}_d^T \\ \mathbf{D}_{e_d} \boldsymbol{\Sigma}_{e_d} \end{pmatrix}, \text{ and} \\ \boldsymbol{\Gamma}_{y_d} &= \begin{pmatrix} \boldsymbol{\Gamma}_{u_d} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Gamma}_{e_d} \end{pmatrix} + \begin{pmatrix} \mathbf{D}_{u_d} \boldsymbol{\Sigma}_{u_d} \mathbf{D}_{u_d}^T & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_{e_d} \boldsymbol{\Sigma}_{e_d} \mathbf{D}_{e_d}^T \end{pmatrix} \\ &\quad - \begin{pmatrix} \mathbf{D}_{u_d} \boldsymbol{\Sigma}_{u_d} \mathbf{Z}_d^T \boldsymbol{\Sigma}^{-1} \mathbf{Z} \boldsymbol{\Sigma}_{u_d} \mathbf{D}_{u_d}^T & \mathbf{D}_{u_d} \boldsymbol{\Sigma}_{u_d} \mathbf{Z}_d^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma}_{e_d} \mathbf{D}_{e_d}^T \\ \mathbf{D}_{e_d} \boldsymbol{\Sigma}_{e_d} \boldsymbol{\Sigma}^{-1} \mathbf{Z} \boldsymbol{\Sigma}_{u_d} \mathbf{D}_{u_d}^T & \mathbf{D}_{e_d} \boldsymbol{\Sigma}_{e_d} \boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma}_{e_d} \mathbf{D}_{e_d}^T \end{pmatrix}. \end{aligned}$$

Proof. For notation simplification, we will drop the subscript d in this proof. The best predictor of the random effects is defined as the conditional expectation of $\mathbf{u}$ given $\mathbf{y}$

that is

$$\begin{aligned}
\mathbb{E}(\mathbf{u}|\mathbf{y}) &= \int_{\mathbb{R}^q} \mathbf{u} f_{\mathbf{u}|\mathbf{y}}(\mathbf{u}) d\mathbf{u} \\
&= \int_{\mathbb{R}^q} \mathbf{u} \frac{f_{\mathbf{y}|\mathbf{u}}(\mathbf{y}) f_{\mathbf{u}}(\mathbf{u})}{f_{\mathbf{y}}(\mathbf{y})} d\mathbf{u} \\
&= f_{\mathbf{y}}(\mathbf{y}) \int_{\mathbb{R}^q} \mathbf{u} C_e C_u \phi_n(\mathbf{y}|\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\mu}_u + \boldsymbol{\mu}_e; \boldsymbol{\Sigma}_e) \phi_q(\mathbf{u}|\boldsymbol{\mu}_u; \boldsymbol{\Sigma}_u) \\
&\quad \times \Phi_n(\mathbf{D}_e(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u} - \boldsymbol{\mu}_e)|\boldsymbol{\nu}_e, \boldsymbol{\Gamma}_e) \Phi_q(\mathbf{D}_u(\mathbf{u} - \boldsymbol{\mu}_u)|\boldsymbol{\nu}_u; \boldsymbol{\Gamma}_u) d\mathbf{u}
\end{aligned}$$

Using lemma 3.1, we get

$$\begin{aligned}
&\phi_n(\mathbf{y}|\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\mu}_e; \boldsymbol{\Sigma}_e) \phi_q(\mathbf{u}|\boldsymbol{\mu}_u; \boldsymbol{\Sigma}_u) = \\
&\phi_n(\mathbf{y}|\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\mu}_u + \boldsymbol{\mu}_e; \boldsymbol{\Sigma}_e + \mathbf{Z}\boldsymbol{\Sigma}_u\mathbf{Z}^T) \phi_q(\mathbf{u}|\boldsymbol{\mu}^*; \boldsymbol{\Sigma}^*)
\end{aligned}$$

where $\boldsymbol{\mu}^* = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\mu}_e$, $\boldsymbol{\Sigma}^* = (\boldsymbol{\Sigma}_u^{-1} + \mathbf{Z}^T \boldsymbol{\Sigma}_e^{-1} \mathbf{Z})^{-1} = \boldsymbol{\Sigma}_u - \boldsymbol{\Sigma}_u \mathbf{Z}^T \boldsymbol{\Sigma}_e^{-1} \mathbf{Z} \boldsymbol{\Sigma}_u$. Also, using lemma 3.2, we get

$$\Phi_n(\mathbf{D}_e(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \boldsymbol{\mu}_e - \mathbf{Z}\boldsymbol{\mu}_u)|\boldsymbol{\nu}_e; \boldsymbol{\Gamma}_e) \Phi_q(\mathbf{D}_u(\mathbf{u} - \boldsymbol{\mu}_u)|\boldsymbol{\nu}_u; \boldsymbol{\Gamma}_u) = \Phi_{n+q}(\mathbf{D}^*\mathbf{u}|\boldsymbol{\nu}_1; \boldsymbol{\Gamma}^*) \quad (6.28)$$

$$\text{where } \mathbf{D}^* = \begin{pmatrix} \mathbf{D}_u \\ -\mathbf{D}_e \mathbf{Z} \end{pmatrix}, \boldsymbol{\nu}_1 = \begin{pmatrix} \mathbf{D}_u \boldsymbol{\mu}_u \\ -\mathbf{D}_e(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \boldsymbol{\mu}_e) \end{pmatrix}, \boldsymbol{\Gamma}^* = \begin{pmatrix} \boldsymbol{\Gamma}_u & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Gamma}_e \end{pmatrix}.$$

We may rewrite the right-hand side of (6.28) as follows

$$\Phi_{n+q}(\mathbf{D}^*\mathbf{u}|\boldsymbol{\nu}_1; \boldsymbol{\Gamma}^*) = \Phi_{n+q}(\mathbf{D}^*(\mathbf{u} - \boldsymbol{\mu}^*)|\boldsymbol{\nu}^*; \boldsymbol{\Gamma}^*)$$

$$\text{where } \boldsymbol{\nu}^* = \begin{pmatrix} \mathbf{D}_u \boldsymbol{\Sigma}_u \mathbf{Z}^T \\ \mathbf{D}_e \boldsymbol{\Sigma} \end{pmatrix} \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\mu}_u - \boldsymbol{\mu}_e). \text{ The expectation of } \mathbf{u} \text{ given } \mathbf{y}$$

can be now written as

$$\begin{aligned}
\mathbb{E}(\mathbf{u}|\mathbf{y}) &= C_u C_e \frac{\phi_n(\mathbf{y}|\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\mu}_u + \boldsymbol{\mu}_e; \boldsymbol{\Sigma}_e)}{f_{\mathbf{y}}(\mathbf{y})} \\
&\quad \times \int_{\mathbb{R}^q} \mathbf{u} \phi_q(\mathbf{u}|\boldsymbol{\mu}^*; \boldsymbol{\Sigma}^*) \Phi_{n+q}(\mathbf{D}^*(\mathbf{u} - \boldsymbol{\mu}^*)|\boldsymbol{\nu}^*; \boldsymbol{\Gamma}^*) d\mathbf{u} \\
&= \frac{C_u C_e}{C_{u^*}} \frac{\phi_n(\mathbf{y}|\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\mu}_u + \boldsymbol{\mu}_e; \boldsymbol{\Sigma}_e)}{f_{\mathbf{y}}(\mathbf{y})} \\
&\quad \times \int_{\mathbb{R}^q} \mathbf{u} C_{u^*} \phi_q(\mathbf{u}|\boldsymbol{\mu}^*; \boldsymbol{\Sigma}^*) \Phi_{n+q}(\mathbf{D}^*(\mathbf{u} - \boldsymbol{\mu}^*)|\boldsymbol{\nu}^*; \boldsymbol{\Gamma}^*) d\mathbf{u} \\
&= \frac{C_u C_e}{C_{u^*}} \frac{\phi_n(\mathbf{y}|\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\mu}_u + \boldsymbol{\mu}_e; \boldsymbol{\Sigma}_e)}{f_{\mathbf{y}}(\mathbf{y})} \mathbb{E}(\mathbf{U}^*)
\end{aligned}$$

where $\mathbf{U}^* \sim CSN_{q,n+q}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*, \mathbf{D}^*, \boldsymbol{\nu}^*, \boldsymbol{\Gamma}^*)$ and $C_{u^*}^{-1} = \Phi(\mathbf{0}|\boldsymbol{\nu}^*; \boldsymbol{\Gamma}^* + \mathbf{D}^* \boldsymbol{\Sigma}^* \mathbf{D}^{*T})$. We note that

$$\begin{aligned}
\boldsymbol{\nu}^* &= \begin{pmatrix} \mathbf{D}_u \boldsymbol{\Sigma}_u \mathbf{Z}^T \\ \mathbf{D}_e \boldsymbol{\Sigma}_e \end{pmatrix} \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\mu}_u + \boldsymbol{\mu}_e) \\
&= \mathbf{D}_y (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\mu}_u + \boldsymbol{\mu}_e)
\end{aligned}$$

and

$$\begin{aligned}
\boldsymbol{\Gamma}^* + \mathbf{D}^* \boldsymbol{\Sigma}^* \mathbf{D}^{*T} &= \begin{pmatrix} \boldsymbol{\Gamma}_u & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Gamma}_e \end{pmatrix} + \begin{pmatrix} \mathbf{D}_u \\ -\mathbf{D}_e \mathbf{Z} \end{pmatrix} (\boldsymbol{\Sigma}_u^{-1} + \mathbf{Z}^T \boldsymbol{\Sigma}_e^{-1} \mathbf{Z})^{-1} \begin{pmatrix} \mathbf{D}_u \\ -\mathbf{D}_e \mathbf{Z} \end{pmatrix}^T \\
&= \begin{pmatrix} \boldsymbol{\Gamma}_u & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Gamma}_e \end{pmatrix} + \begin{pmatrix} \boldsymbol{\Gamma}_{11} & \boldsymbol{\Gamma}_{12} \\ \boldsymbol{\Gamma}_{21} & \boldsymbol{\Gamma}_{22} \end{pmatrix}
\end{aligned}$$

where

$$\begin{aligned}
\boldsymbol{\Gamma}_{11} &= \mathbf{D}_u (\boldsymbol{\Sigma}_u^{-1} + \mathbf{Z}^T \boldsymbol{\Sigma}_e^{-1} \mathbf{Z})^{-1} \mathbf{D}_u^T \\
&= \mathbf{D}_u (\boldsymbol{\Sigma}_u - \boldsymbol{\Sigma}_u \mathbf{Z}^T \boldsymbol{\Sigma}^{-1} \mathbf{Z} \boldsymbol{\Sigma}_u) \mathbf{D}_u^T \\
&= \mathbf{D}_u \boldsymbol{\Sigma}_u \mathbf{D}_u^T - \mathbf{D}_u \boldsymbol{\Sigma}_u \mathbf{Z}^T \boldsymbol{\Sigma}^{-1} \mathbf{Z} \boldsymbol{\Sigma}_u \mathbf{D}_u^T
\end{aligned}$$

$$\begin{aligned}
\Gamma_{22} &= \mathbf{D}_e \mathbf{Z} (\Sigma_u^{-1} + \mathbf{Z}^T \Sigma_e^{-1} \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{D}_e^T \\
&= \mathbf{D}_e \mathbf{Z} (\Sigma_u^{-1} + \mathbf{Z}^T \Sigma_e^{-1} \mathbf{Z})^{-1} \mathbf{Z}^T \Sigma_e^{-1} \Sigma_e \mathbf{D}_e^T \\
&= \mathbf{D}_e \mathbf{Z} \Sigma_u \mathbf{Z}^T \Sigma^{-1} \mathbf{Z}^T \Sigma_e^{-1} \Sigma_e \mathbf{D}_e^T \\
&= \mathbf{D}_e (\Sigma - \Sigma_e) \Sigma^{-1} \Sigma_e \mathbf{D}_e^T \\
&= \mathbf{D}_e \Sigma_e \mathbf{D}_e^T - \mathbf{D}_e \Sigma_e \Sigma^{-1} \Sigma_e \mathbf{D}_e^T
\end{aligned}$$

$$\begin{aligned}
\Gamma_{12} &= \mathbf{D}_u (\Sigma_u^{-1} + \mathbf{Z}^T \Sigma_e^{-1} \mathbf{Z})^{-1} \mathbf{Z}^T \Sigma_e^{-1} \Sigma_e \mathbf{D}_e^T \\
&= \mathbf{D}_u \Sigma_u \mathbf{Z}^T \Sigma^{-1} \Sigma_e \mathbf{D}_e^T \\
&= \Gamma_{21}^T.
\end{aligned}$$

Lemma 3.4 leads to

$$\begin{aligned}
C_e^{-1} C_u^{-1} &= \Phi_n(\mathbf{0}|\mathbf{0}; (1 + \lambda_e^2)^2 \mathbf{I}_n) \Phi_q(\mathbf{0}|\mathbf{0}; (1 + \lambda_u^2)^2 \mathbf{I}_q) \\
&= \Phi_{n+q} \left(\mathbf{0}|\mathbf{0}; \begin{pmatrix} (1 + \lambda_e^2)^2 \mathbf{I}_n & \mathbf{0} \\ \mathbf{0} & (1 + \lambda_u^2)^2 \mathbf{I}_q \end{pmatrix} \right) \\
&= C_y^{-1}
\end{aligned}$$

Hence, it follows $C_e^{-1} C_u^{-1} \phi_n(\mathbf{y}|\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\mu}_u + \boldsymbol{\mu}_e; \Sigma_e) C_u^{-1} = f_{\mathbf{y}}(\mathbf{y})$. The conditional expectation of $\mathbf{u}$ given $\mathbf{y}$ reduces to $\mathbb{E}(\mathbf{u}|\mathbf{y}) = \frac{f_{\mathbf{y}}(\mathbf{y})}{f_{\mathbf{y}}(\mathbf{y})} \mathbb{E}(\mathbf{U}^*) = \mathbb{E}(\mathbf{U}^*)$. Therefore, using corollary 2.1, we conclude that the conditional expectation is

$$\mathbb{E}(\mathbf{u}|\mathbf{y}) = \boldsymbol{\mu}^* + \frac{\Phi_{N_d+q}^{(1)}(\mathbf{0}; \boldsymbol{\nu}^*, \boldsymbol{\Gamma}^* + \mathbf{D}^* \Sigma^* \mathbf{D}^{*T})}{\Phi_{N_d+q}(\mathbf{0}; \boldsymbol{\nu}^*, \boldsymbol{\Gamma}^* + \mathbf{D}^* \Sigma^* \mathbf{D}^{*T})}$$

where $\Phi_{N_d+q}^{(1)}(\mathbf{0}; \boldsymbol{\nu}^*, \boldsymbol{\Gamma}^* + \mathbf{D}^* \Sigma^* \mathbf{D}^{*T}) = \frac{\partial}{\partial \mathbf{t}} \Phi_{N_d+q}(\mathbf{D}^* \Sigma^* \mathbf{t} | \boldsymbol{\nu}^*; \boldsymbol{\Gamma}^* + \mathbf{D}^* \Sigma^* \mathbf{D}^{*T})|_{\mathbf{t}=\mathbf{0}}$. $\square$

Chapter 7

Cross-Sectional Time Series Data in Small Area Estimation

For some cross-sectional surveys repeated periodically, it may be possible to take advantage of the correlation of the variables of interest over time to improve the efficiency of small area estimators. With the purpose of borrowing strength across time and across areas, many authors such as Pfeiffermann and Burck (1990) considered cross-sectional data and time series models in the context of small area estimation. However Rao and Yu (1992, 1994) were the first to propose a model that handles an arbitrary covariance matrix for the sampling errors. Rao and Yu (1992, 1994) proposed an extension to the basic Fay-Herriot model by incorporating area by time random components following an autoregressive process of order 1 ($AR(1)$).

Their model is defined as follows:

$$\hat{\theta}_{dt} = \mathbf{x}_{dt}^T \boldsymbol{\beta} + u_d + \nu_{dt} + e_{dt}, \quad t = 1, \dots, T, \quad d = 1, \dots, m, \quad (7.1)$$

where $\hat{\theta}_{dt}$ is the direct estimator of the parameter of interest, $\theta_{dt} = \mathbf{x}_{dt}^T \boldsymbol{\beta} + u_d + \nu_{dt}$, for area d at time t , $\mathbf{x}_{dt}^T = (x_{dt1}, \dots, x_{dtp})^T$ is a vector of p fixed auxiliary variables for area d at time t , the sampling errors e_{dt} have mean zero and known variance-covariance

matrix, $u_d \stackrel{iid}{\sim} N(0, \sigma_u^2)$ is the time independent area effect and the ν_{dt} 's are assumed to follow a first order autoregressive process within each area d ,

$$\nu_{dt} = \rho \nu_{d,t-1} + \epsilon_{dt}, \quad |\rho| < 1, \quad (7.2)$$

with $\epsilon_{dt} \stackrel{iid}{\sim} N(0, \sigma^2)$. The errors e_{dt} , u_d , and ϵ_{dt} are assumed to be independent of each other. The condition $|\rho| < 1$ ensures stationarity of the series defined by (7.2) in order to obtain an autoregressive process of order 1. The stationarity assumption leads to:

$$\mathbb{V}ar(\nu_{dt}) = \frac{\sigma^2}{1 - \rho^2}. \quad (7.3)$$

The variance of ν_{dt} is not defined for $\rho = 1$. This discontinuity issue makes it difficult to estimate the parameter ρ when it is close to 1. Rao and Yu (1992, 1994) used the method of moments to estimate the parameters σ_u^2 , σ^2 , and ρ . Their two estimators of ρ , the naive one and the one taking into account the distribution of the sampling errors e_{dt} , are unstable and often take values outside the admissible range $(-1, 1)$. Facing the issue of ρ very closed to 1, Datta et al. (2002) proposed the random walk model instead of $AR(1)$ for the time series (7.2). The random walk model forces the parameter ρ to be equal to 1 and therefore avoid the difficulty of its estimation. Also, unlike $AR(1)$, the random walk model leads to a divergent series (7.2) i.e. infinite variance of ν_{dt} (no stationarity). Hence, in practice, the random walk is only applicable to finite series of ν_{dt} by specifying an initial value ν_{d,t_0} . The initial value has to be assumed known to avoid identifiability issues (see Datta et al. (2002)).

In this chapter, we propose a unified approach that removes the stationarity assumption of the Rao-Yu and therefore the discontinuity at $\rho = 1$. We refer to this model as the general Rao-Yu model. The general Rao-Yu model allows the parameter ρ to take any finite value. This is achieved by considering only finite series for (7.2)

instead of theoretical infinite $AR(1)$ or random walk models. The main drawback of this approach is the fact that the variance of the series $(\nu_{dt})_t$ increases very fast when t increases and ρ is close to 1 or higher than 1. Therefore, in practice, this general Rao-Yu approach is applicable only to short and moderate length series. The main advantage is letting the model “decide” whether the coefficient ρ is equal to 1 or not, no need by the user to force ρ to be equal to 1 as in the random walk model or to be strictly smaller than 1 as in the original Rao-Yu model. Also, in theory, the coefficient can take any real value however the usual interpretation of ρ as correlation coefficient may not be valid for values strictly larger than 1 in absolute value.

7.1 Rao-Yu Model Without Stationarity Assumption

To obtain the general Rao-Yu model, the stationarity assumption is dropped by assuming finite series. The general Rao-Yu model has the same global model (7.1) but the time series part is now finite and defined as:

$$\nu_{dt} = \rho \nu_{d,t-1} + \epsilon_{dt}, \text{ for } t \geq 1 \text{ and } \nu_{d0} = 0, \quad (7.4)$$

with $\epsilon_{dt} \stackrel{iid}{\sim} N(0, \sigma^2)$. The errors e_{dt} , u_d , and ϵ_{dt} are assumed to be independent of each other. The area-time effect ν_{dt} can be written in terms of the time errors as follows:

$$\nu_{dt} = \sum_{t'=1}^t \rho^{t'-1} \epsilon_{d,t-(t'-1)}. \quad (7.5)$$

From expression (7.5), it easy to see that

$$\mathbb{V}ar(\nu_{dt}) = \sigma^2 \sum_{t'=1}^t \rho^{2(t'-1)} \quad (7.6)$$

and

$$\mathbb{C}ov(\nu_{d,t_1}, \nu_{d,t_2}) = \sigma^2 \rho^{|t_1-t_2|} \sum_{t=1}^{\min(t_1, t_2)} \rho^{2(t-1)}. \quad (7.7)$$

Expression (7.6) is the equivalent of (7.3) for the finite series model. It follows that the variance of ν_{dt} is always defined for any value of ρ avoiding the discontinuity issue at $\rho = 1$. Under this model, assuming that σ_u^2 , σ^2 , and ρ are known, the best linear unbiased predictor (BLUP) of the parameter of interest at time T is defined as:

$$\hat{\theta}_{dT}^{BLUP} = \mathbf{x}_{dT}^T \tilde{\boldsymbol{\beta}} + (\sigma^2 \boldsymbol{\gamma}_{\nu, T} + \sigma_u^2 \mathbf{1}_{n_d})^T \mathbf{V}_d^{-1} (\hat{\boldsymbol{\theta}}_d - \mathbf{X}_d \tilde{\boldsymbol{\beta}}) \quad (7.8)$$

where $\boldsymbol{\gamma}_{\nu, T}$ is the T^{th} column of $\boldsymbol{\Gamma}_\nu$ which is a $T \times T$ symmetric matrix with upper triangular elements defined as:

$$(\boldsymbol{\Gamma}_\nu)_{i,j} = \rho^{(j-i)} \sum_{i'=1}^i \rho^{2(i'-1)}, \quad 1 \leq i \leq j \leq T. \quad (7.9)$$

Also $\mathbf{V}_d = \sigma_u^2 \mathbf{1}_{n_d} \mathbf{1}_{n_d}^T + \sigma^2 \boldsymbol{\Gamma}_\nu + \boldsymbol{\Sigma}_d$, $\tilde{\boldsymbol{\beta}} = (\sum_{d=1}^m \mathbf{X}_d^T \mathbf{V}_d^{-1} \mathbf{X}_d)^{-1} (\sum_{d=1}^m \mathbf{X}_d^T \mathbf{V}_d^{-1} \hat{\boldsymbol{\theta}}_d)$ is the generalized least squares estimator of $\boldsymbol{\beta}$, and $\boldsymbol{\Sigma}_d$ is the known sampling covariance matrix. The BLUP estimator is a weighted sum of the direct estimator $\hat{\theta}_{dt}$, the synthetic estimator $\mathbf{x}_{dt}^T \tilde{\boldsymbol{\beta}}$, and the residuals $\hat{\theta}_{dt} - \mathbf{x}_{dt}^T \tilde{\boldsymbol{\beta}}$, $t = 1, \dots, T-1$. This can be seen by rewriting the BLUP as follows:

$$\hat{\theta}_{dT}^{BLUP} = \mathbf{w}_{dT} \hat{\theta}_{dT} + (1 - \mathbf{w}_{dT}) \mathbf{x}_{dT}^T \tilde{\boldsymbol{\beta}} + \sum_{t=1}^{T-1} \mathbf{w}_{dt} (\hat{\theta}_{dt} - \mathbf{x}_{dt}^T \tilde{\boldsymbol{\beta}}), \quad (7.10)$$

$\mathbf{w}_{dT} = (\sigma^2 \boldsymbol{\gamma}_{\nu, T} + \sigma_u^2 \mathbf{1}_{n_d})^T \mathbf{V}_d^{-1}$. By replacing the unknown parameters in (7.8) by their estimators, one get the empirical best linear unbiased predictor (EBLUP), that is:

$$\hat{\theta}_{dT}^{EBLUP} = \hat{\theta}_{dT}^{BLUP}(\sigma_u^2, \sigma^2, \rho) |_{\sigma_u^2 = \hat{\sigma}_u^2, \sigma^2 = \hat{\sigma}^2, \rho = \hat{\rho}} \quad (7.11)$$

7.2 Parameter Estimation

To obtain the EBLUP, it is necessary to estimate the parameters β , σ_u^2 , σ^2 , and ρ . Two methods for estimating the unknown parameters are described here: maximum likelihood (ML) and restricted maximum likelihood (REML). These two methods have shown much more stability than the method of moments (MOM) estimation. Also, with ML and REML, all the parameters are estimated simultaneously. In this section, the main parameters estimation steps are presented, we refer the reader to the book by McCulloch et al. (2008) for more details on linear mixed model parameter estimation.

7.2.1 Maximum Likelihood (ML) Estimation

The logarithm of the maximum likelihood (log likelihood) function associated with the dynamic model (7.1) is equal to

$$l(\beta, \varphi | \hat{\theta}) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \ln |\mathbf{V}| - \frac{1}{2} (\hat{\theta} - \mathbf{X}\beta)^T \mathbf{V}^{-1} (\hat{\theta} - \mathbf{X}\beta), \quad (7.12)$$

note that $\mathbf{V} = \mathbf{V}(\varphi)$ with $\varphi = (\sigma_u^2, \sigma^2, \rho)$. It is clear the covariance matrix is not a function of β but only of φ . The partial derivative of the log likelihood function with respect to φ is:

$$\frac{\partial l(\beta, \varphi | \hat{\theta})}{\partial \varphi_j} = -\frac{1}{2} \text{tr}(\mathbf{V}^{-1} \mathbf{V}_{(\varphi_j)}) - \frac{1}{2} (\hat{\theta} - \mathbf{X}\beta)^T \mathbf{V}^{-1} \mathbf{V}_{(\varphi_j)} \mathbf{V}^{-1} (\hat{\theta} - \mathbf{X}\beta) \quad (7.13)$$

where $\mathbf{V}_{(\varphi_j)} = \frac{\partial \mathbf{V}}{\partial \varphi_j}$. The matrix $\mathbf{V}$ has a block diagonal structure which is denoted $\mathbf{V} = \text{diag}_{1 \leq d \leq m}(\mathbf{V}_d)$. This block diagonal structure is important because it allows us to specify the model independently at the area level d for $d = 1, \dots, m$. However $\mathbf{V}_d$ does not have a simple form, because of the general structure of the sampling covariance Σ_d . The information matrix defined as the expectation of the second

derivative of $-l(\boldsymbol{\beta}, \boldsymbol{\varphi}|\hat{\boldsymbol{\theta}})$ is:

$$\mathbf{I}(\boldsymbol{\beta}, \boldsymbol{\varphi}) = \begin{pmatrix} \sum_{d=1}^m \mathbf{X}_d^T \mathbf{V}_d^{-1} \mathbf{X}_d & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{\boldsymbol{\varphi}} \end{pmatrix} \quad (7.14)$$

The sub-matrix $\mathbf{I}_{\boldsymbol{\varphi}} = \left[-\frac{1}{2} \sum_{d=1}^m \text{tr}(\mathbf{V}_d^{-1} \mathbf{V}_{d(\varphi_i)} \mathbf{V}_d^{-1} \mathbf{V}_{d(\varphi_j)}) \right]_{1 \leq i \leq j \leq 3}$ is a 3×3 information matrix associated with σ_u^2 , σ^2 , and ρ ; the elements are obtained from the differentiation of the appropriate components. Moreover the diagonal elements of $\mathbf{I}_{\boldsymbol{\varphi}}$ are:

$$\begin{aligned} \mathbf{I}_{uu}(\boldsymbol{\varphi}) &= \frac{1}{2} \sum_{d=1}^m \text{tr}(\mathbf{V}_d^{-1} \mathbf{V}_{d(\sigma_u^2)})^2 = \frac{1}{2} \sum_{d=1}^m \text{tr}(\mathbf{V}_d^{-1} \mathbf{1}_{n_d} \mathbf{1}_{n_d}^T)^2 \\ \mathbf{I}_{\nu\nu}(\boldsymbol{\varphi}) &= \frac{1}{2} \sum_{d=1}^m \text{tr}(\mathbf{V}_d^{-1} \mathbf{V}_{d(\sigma^2)})^2 = \frac{1}{2} \sum_{d=1}^m \text{tr}(\mathbf{V}_d^{-1} \boldsymbol{\Gamma}_{\nu})^2 \\ \mathbf{I}_{\rho\rho}(\boldsymbol{\varphi}) &= \frac{1}{2} \sum_{d=1}^m \text{tr}(\mathbf{V}_d^{-1} \mathbf{V}_{d(\rho)})^2 = \frac{1}{2} \sum_{d=1}^m \text{tr} \left(\mathbf{V}_d^{-1} \left(\sigma^2 \frac{\partial \boldsymbol{\Gamma}_{\nu}}{\partial \rho} \right) \right)^2 \end{aligned}$$

with

$$\left[\frac{\partial \boldsymbol{\Gamma}_{\nu}}{\partial \rho} \right]_{i,j} = \rho^{(j-i-1)} \sum_{i'=1}^i [j-i+2(i'-1)] \rho^{2(i'-1)}, \quad 1 \leq i \leq j \leq T \quad (7.15)$$

The off-diagonal elements of $\mathbf{I}_{\boldsymbol{\varphi}}$ are:

$$\begin{aligned} \mathbf{I}_{u\nu}(\boldsymbol{\varphi}) &= \frac{1}{2} \sum_{d=1}^m \text{tr}(\mathbf{V}_d^{-1} \mathbf{V}_{d(\sigma_u^2)} \mathbf{V}_d^{-1} \mathbf{V}_{d(\sigma^2)}) = \frac{1}{2} \sum_{d=1}^m \text{tr}(\mathbf{V}_d^{-1} \mathbf{1}_{n_d} \mathbf{1}_{n_d}^T \mathbf{V}_d^{-1} \boldsymbol{\Gamma}_{\nu}) \\ \mathbf{I}_{u\rho}(\boldsymbol{\varphi}) &= \frac{1}{2} \sum_{d=1}^m \text{tr}(\mathbf{V}_d^{-1} \mathbf{V}_{d(\sigma_u^2)} \mathbf{V}_d^{-1} \mathbf{V}_{d(\rho)}) = \frac{1}{2} \sum_{d=1}^m \text{tr} \left(\mathbf{V}_d^{-1} \mathbf{1}_{n_d} \mathbf{1}_{n_d}^T \mathbf{V}_d^{-1} \left(\sigma^2 \frac{\partial \boldsymbol{\Gamma}_{\nu}}{\partial \rho} \right) \right) \\ \mathbf{I}_{\nu\rho}(\boldsymbol{\varphi}) &= \frac{1}{2} \sum_{d=1}^m \text{tr}(\mathbf{V}_d^{-1} \mathbf{V}_{d(\sigma^2)} \mathbf{V}_d^{-1} \mathbf{V}_{d(\rho)}) = \frac{1}{2} \sum_{d=1}^m \text{tr} \left(\mathbf{V}_d^{-1} \boldsymbol{\Gamma}_{\nu} \mathbf{V}_d^{-1} \left(\sigma^2 \frac{\partial \boldsymbol{\Gamma}_{\nu}}{\partial \rho} \right) \right). \end{aligned}$$

The maximum-likelihood estimates that maximize $l(\boldsymbol{\beta}, \boldsymbol{\varphi}|\hat{\boldsymbol{\theta}})$ over the valid range of parameter values are obtained by solving $\frac{\partial l(\boldsymbol{\beta}, \boldsymbol{\varphi}|\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} = \mathbf{0}$. These equations are solved simultaneously and often iteratively. The Fisher scoring method, a variation of the

Newton-Raphson method, for solving these equations is equal at step $k + 1$ to

$$\boldsymbol{\varphi}_{k+1} = \boldsymbol{\varphi}_k + [\mathbf{I}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi}_k)]^{-1} \frac{\partial l(\tilde{\boldsymbol{\beta}}(\boldsymbol{\varphi}_k), \boldsymbol{\varphi}_k | \hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\varphi}}, \quad (7.16)$$

$$\boldsymbol{\beta}_{k+1} = \tilde{\boldsymbol{\beta}}(\boldsymbol{\varphi}_{k+1}). \quad (7.17)$$

At the convergence of the Fisher scoring iteration, we get the ML estimates $\hat{\boldsymbol{\varphi}}_{ML}$ and $\hat{\boldsymbol{\beta}}_{ML}$ of $\boldsymbol{\varphi}$ and $\boldsymbol{\beta}$, respectively. The asymptotic covariance matrices of $\hat{\boldsymbol{\beta}}_{ML}$ and $\hat{\boldsymbol{\varphi}}_{ML}$ are $\text{Var}(\hat{\boldsymbol{\beta}}_{ML}) = \sum_{d=1}^m \mathbf{X}_d^T \mathbf{V}_d^{-1} \mathbf{X}_d$ and $\text{Var}(\hat{\boldsymbol{\varphi}}_{ML}) = [\mathbf{I}_{\boldsymbol{\varphi}}]^{-1}$, respectively.

7.2.2 Restricted Maximum Likelihood (REML) Estimation

The ML estimation depends on the estimation of the fixed effects as shown in (7.16). REML estimation is a modification of the ML procedure that takes into account the loss in degrees of freedom resulting from estimating the fixed effects. The idea is to perform estimation on $\mathbf{A}^T \hat{\boldsymbol{\theta}}$ where $\mathbf{A}$ is defined so that $\mathbf{A}^T \mathbf{X} = \mathbf{0}$. The log likelihood function of the transformed data is:

$$l_{RE}(\boldsymbol{\varphi} | \hat{\boldsymbol{\theta}}) = - \left(\frac{n - \text{rank}(\mathbf{X})}{2} \right) \log(2\pi) - \frac{1}{2} \ln |\mathbf{A}^T \mathbf{V} \mathbf{A}| - \frac{1}{2} \hat{\boldsymbol{\theta}}^T \mathbf{A} (\mathbf{A}^T \mathbf{V} \mathbf{A})^{-1} \mathbf{A}^T \hat{\boldsymbol{\theta}}. \quad (7.18)$$

One important result shows that if $\mathbf{A}^T \mathbf{X} = \mathbf{0}$ where $\mathbf{A}^T$ has maximum row rank and $\mathbf{V}$ is positive definite then

$$\mathbf{A} (\mathbf{A}^T \mathbf{V} \mathbf{A})^{-1} \mathbf{A}^T = \mathbf{P}, \text{ with } \mathbf{P} \equiv \mathbf{V}^{-1} - \mathbf{V}^{-1} \mathbf{X} (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1}. \quad (7.19)$$

It follows that

$$\begin{aligned} \frac{\partial \ln |\mathbf{A}^T \mathbf{V} \mathbf{A}|}{\partial \varphi_j} &= \text{tr} \left[(\mathbf{A}^T \mathbf{V} \mathbf{A})^{-1} \frac{\partial (\mathbf{A}^T \mathbf{V} \mathbf{A})}{\partial \varphi_j} \right] \\ &= \text{tr} \left[\mathbf{A} (\mathbf{A}^T \mathbf{V} \mathbf{A})^{-1} \mathbf{A}^T \mathbf{V}_{(\varphi_j)} \right]. \end{aligned}$$

The partial derivative of $l_{RE}(\boldsymbol{\varphi}|\hat{\boldsymbol{\theta}})$ with respect to $\boldsymbol{\varphi}$ is then obtained by

$$\frac{\partial l_{RE}(\boldsymbol{\varphi}|\hat{\boldsymbol{\theta}})}{\partial \varphi_j} = -\frac{1}{2}tr(\mathbf{P}\mathbf{V}_{(\varphi_j)}) + \frac{1}{2}\hat{\boldsymbol{\theta}}^T \mathbf{P}\mathbf{V}_{(\varphi_j)} \mathbf{P}\hat{\boldsymbol{\theta}} \quad (7.20)$$

Note that $\mathbf{P}\hat{\boldsymbol{\theta}} = \mathbf{V}^{-1}(\hat{\boldsymbol{\theta}} - \mathbf{X}\tilde{\boldsymbol{\beta}}^T)$. The REML information matrix defined as the second derivative of $-l_{RE}(\boldsymbol{\varphi}|\hat{\boldsymbol{\theta}})$ is equal to:

$$\mathbf{I}_{RE}(\boldsymbol{\varphi}) = \left[\frac{1}{2}tr(\mathbf{P}\mathbf{V}_{(\varphi_i)}\mathbf{P}\mathbf{V}_{(\varphi_j)}) \right]_{1 \leq i \leq j \leq 3}. \quad (7.21)$$

A similar iteration procedure as for the Fisher scoring algorithm (7.16) is used replacing the ML quantities by the REML equivalents to obtain the REML estimates $\hat{\boldsymbol{\varphi}}_{RE}$. A REML estimator of $\boldsymbol{\beta}$ is obtained by $\hat{\boldsymbol{\beta}}_{RE} = \tilde{\boldsymbol{\beta}}(\hat{\boldsymbol{\varphi}}_{RE})$. Also, the covariance matrix of $\hat{\boldsymbol{\beta}}_{RE}$ and $\hat{\boldsymbol{\varphi}}_{RE}$ are asymptotically equivalent to the ML covariance matrices.

7.3 Mean Squared Error (MSE) Estimation

In this section, we provide the MSE to second-order accuracy, as a measure of uncertainty of the EBLUP (7.11). Also, an approximation of $MSE(\hat{\theta}_{dT}^{BLUP})$ is derived.

7.3.1 MSE Estimator of the EBLUP

The EBLUP estimator $\hat{\theta}_{dT}^{EBLUP}$ shown in (7.11) can be written in a more general form:

$$\hat{\theta}_{dT}^{EBLUP} = \mathbf{x}_{dT}^T \tilde{\boldsymbol{\beta}}(\hat{\boldsymbol{\varphi}}) + \mathbf{m}_d^T \hat{\mathbf{G}}_d \mathbf{Z}_d^T \hat{\mathbf{V}}_d^{-1}(\boldsymbol{\theta}_d - \mathbf{X}_d \tilde{\boldsymbol{\beta}}(\hat{\boldsymbol{\varphi}})) \quad (7.22)$$

where $\mathbf{m}_d^T = (0, \dots, 0, 1)$ with 1 at the T^{th} position, $\mathbf{G}_d = \hat{\sigma}^2 \boldsymbol{\Gamma}_\nu(\hat{\rho}) + \hat{\sigma}_u^2 \mathbf{1}_{n_d} \mathbf{1}_{n_d}^T$, $\mathbf{Z}_d^T = \mathbf{I}_{n_d}$. Following the decomposition $\hat{\theta}_{dT}^{EBLUP} - \theta_{dT} = (\hat{\theta}_{dT}^{BLUP} - \theta_{dT}) + (\hat{\theta}_{dT}^{EBLUP} - \hat{\theta}_{dT}^{BLUP})$ and noting that $\mathbb{E}((\hat{\theta}_{dT}^{BLUP} - \theta_{dT})(\hat{\theta}_{dT}^{EBLUP} - \hat{\theta}_{dT}^{BLUP})) = 0$, Kackar and Harville (1984)

showed that

$$MSE(\hat{\theta}_{dT}^{EBLUP}) = MSE(\hat{\theta}_{dT}^{BLUP}) + \mathbb{E}(\hat{\theta}_{dT}^{EBLUP} - \hat{\theta}_{dT}^{BLUP})^2. \quad (7.23)$$

Based on the linear mixed model, the first term $MSE(\hat{\theta}_{dT}^{BLUP})$ of (7.23) is equal to

$$MSE(\hat{\theta}_{dT}^{BLUP}) = g_{1dT}(\boldsymbol{\varphi}) + g_{2dT}(\boldsymbol{\varphi}) \quad (7.24)$$

where

$$\begin{aligned} g_{1dT}(\boldsymbol{\varphi}) &= \mathbf{m}_d^T (\mathbf{G}_d - \mathbf{G}_d \mathbf{Z}_d^T \mathbf{V}_d^{-1} \mathbf{Z}_d \mathbf{G}_d) \mathbf{m}_d \\ &= \sigma_u^2 + \sigma^2 \sum_{t=1}^{T-1} \rho^{2(t-1)} - (\sigma_u^2 \mathbf{1}_{n_d} + \sigma^2 \boldsymbol{\gamma}_{\nu,T})^T \mathbf{V}_d^{-1} (\sigma_u^2 \mathbf{1}_{n_d} + \sigma^2 \boldsymbol{\gamma}_{\nu,T}) \end{aligned} \quad (7.25)$$

with $\boldsymbol{\gamma}_{\nu,T}$ is the T^{th} column of $\boldsymbol{\Gamma}_\nu$, and $\boldsymbol{\gamma}_{u,T}$ is the T^{th} column of $\boldsymbol{\Gamma}_u$, and

$$g_{2dT}(\boldsymbol{\varphi}) = \mathbf{d}_d \left(\sum_{d=1}^m \mathbf{X}_d^T \mathbf{V}_d^{-1} \mathbf{X}_d \right)^{-1} \mathbf{d}_d \quad (7.26)$$

with $\mathbf{d}_d = \mathbf{x}_{dT}^T - \mathbf{b}_d \mathbf{X}_d^T$ and $\mathbf{b}_d = \mathbf{m}_d^T \hat{\mathbf{G}}_d \mathbf{Z}_d^T \hat{\mathbf{V}}_d^{-1}$. Thus,

$$\begin{aligned} g_{2dT}(\boldsymbol{\varphi}) &= \mathbf{x}_{dT}^T \left(\sum_{d=1}^m \mathbf{X}_d^T \mathbf{V}_d^{-1} \mathbf{X}_d \right)^{-1} \mathbf{x}_{dT} \\ &\quad + (\sigma_u^2 \mathbf{1}_{n_d} + \sigma^2 \boldsymbol{\gamma}_{\nu,T})^T \mathbf{V}_d^{-1} \mathbf{X}_d^T \left(\sum_{d=1}^m \mathbf{X}_d^T \mathbf{V}_d^{-1} \mathbf{X}_d \right)^{-1} \mathbf{X}_d \mathbf{V}_d^{-1} (\sigma_u^2 \mathbf{1}_{n_d} + \sigma^2 \boldsymbol{\gamma}_{\nu,T}) \\ &\quad - 2(\sigma_u^2 \mathbf{1}_{n_d} + \sigma^2 \boldsymbol{\gamma}_{\nu,T})^T \mathbf{V}_d^{-1} \mathbf{X}_d^T \left(\sum_{d=1}^m \mathbf{X}_d^T \mathbf{V}_d^{-1} \mathbf{X}_d \right)^{-1} \mathbf{x}_{dT}. \end{aligned} \quad (7.27)$$

The term $g_{2dT}(\boldsymbol{\varphi})$ accounts for the extra variability due to estimating $\boldsymbol{\beta}$ and is of order $O(m^{-1})$ while the leading term $g_{1dT}(\boldsymbol{\varphi})$ is of order $O(1)$.

For estimating the term $\mathbb{E}(\hat{\theta}_{dT}^{EBLUP} - \hat{\theta}_{dT}^{BLUP})^2$ from (7.23), Prasad and Rao (1990) proposed the following second-order approximation:

$$\mathbb{E}(\hat{\theta}_{dT}^{EBLUP} - \hat{\theta}_{dT}^{BLUP})^2 \approx g_{3dT}(\boldsymbol{\varphi}) = \text{tr}((\nabla \mathbf{b}) \mathbf{V}_d (\nabla \mathbf{b}) (\mathbf{I}_{\boldsymbol{\varphi}})^{-1}), \quad (7.28)$$

where $\nabla \mathbf{b} = \text{col}_{1 \leq j \leq 3}(\frac{\partial \mathbf{b}_d^T}{\partial \boldsymbol{\varphi}_j}) = (\frac{\partial \mathbf{b}_d^T}{\partial \sigma_u^2}, \frac{\partial \mathbf{b}_d^T}{\partial \sigma^2}, \frac{\partial \mathbf{b}_d^T}{\partial \rho})^T$ with

$$\begin{aligned} \frac{\partial \mathbf{b}_d^T}{\partial \sigma_u^2} &= \mathbf{1}_{n_d}^T \mathbf{V}_d^{-1} - (\sigma_u^2 \mathbf{1}_{n_d} + \sigma^2 \boldsymbol{\gamma}_{\nu, T})^T \mathbf{V}_d^{-1} \mathbf{1}_{n_d} \mathbf{1}_{n_d}^T \mathbf{V}_d^{-1} \\ \frac{\partial \mathbf{b}_d^T}{\partial \sigma^2} &= \boldsymbol{\gamma}_{\nu, T}^T \mathbf{V}_d^{-1} - (\sigma_u^2 \mathbf{1}_{n_d} + \sigma^2 \boldsymbol{\gamma}_{\nu, T})^T \mathbf{V}_d^{-1} \boldsymbol{\Gamma}_{\nu} \mathbf{V}_d^{-1} \\ \frac{\partial \mathbf{b}_d^T}{\partial \rho^2} &= (\sigma^2 \frac{\partial \boldsymbol{\gamma}_{\nu, T}}{\partial \rho})^T \mathbf{V}_d^{-1} - (\sigma_u^2 \mathbf{1}_{n_d} + \sigma^2 \boldsymbol{\gamma}_{\nu, T})^T \mathbf{V}_d^{-1} (\sigma^2 \frac{\partial \boldsymbol{\Gamma}_{\nu}}{\partial \rho})^T \mathbf{V}_d^{-1}. \end{aligned}$$

In summary, a second-order approximation of the EBLUP $\hat{\theta}_{dT}^{EBLUP}$ is given by

$$MSE(\hat{\theta}_{dT}^{EBLUP}) \approx g_{1dT}(\boldsymbol{\varphi}) + g_{2dT}(\boldsymbol{\varphi}) + g_{3dT}(\boldsymbol{\varphi}) \quad (7.29)$$

where $g_{1dT}(\boldsymbol{\varphi})$, $g_{2dT}(\boldsymbol{\varphi})$, and $g_{3dT}(\boldsymbol{\varphi})$ are defined respectively in (7.25), (7.27), and (7.28). The terms (7.27) and (7.28) account for the variability due to estimating $\boldsymbol{\beta}$ and $\boldsymbol{\varphi}$, respectively and are of lower order than (7.25).

7.3.2 Estimation of the MSE Estimator of the EBLUP

To obtain an actual estimator of the approximation (7.29), one replaces the parameter vector $\boldsymbol{\varphi}$ by its estimator $\hat{\boldsymbol{\varphi}}$. Because $g_{1dT}(\hat{\boldsymbol{\varphi}})$ is a biased estimator of $g_{1dT}(\boldsymbol{\varphi})$, Prasad and Rao (1990) proposed a correct estimator of $MSE(\hat{\theta}_{dT}^{EBLUP})$ to the desired order approximation as:

$$mse(\hat{\theta}_{dT}^{EBLUP}) = g_{1dT}(\hat{\boldsymbol{\varphi}}) + g_{2dT}(\hat{\boldsymbol{\varphi}}) + 2g_{3dT}(\hat{\boldsymbol{\varphi}}) \quad (7.30)$$

The estimator (7.30) holds for REML estimators meaning that all the quantities (parameter estimates, information matrix, etc) involved are obtained through the REML procedure. When ML is used to estimate φ an extra bias correction term is necessary and the MSE estimator is then given by:

$$mse_{ML}(\hat{\theta}_{dT}^{EBLUP}) = g_{1dT}(\hat{\varphi}_{ML}) + g_{2dT}(\hat{\varphi}_{ML}) + 2g_{3dT}(\hat{\varphi}_{ML}) - \mathbf{b}_{\hat{\varphi}_{ML}}(\hat{\varphi}_{ML})^T \nabla g_{1dT}(\hat{\varphi}_{ML}) \quad (7.31)$$

In expression (7.31), the information matrix is the same as $\mathbf{I}_{\varphi}$ in (7.14),

$$\nabla g_{1dT}(\hat{\varphi}_{ML}) = \left(\frac{\partial g_{1dT}(\hat{\varphi}_{ML})}{\partial \sigma_u^2}, \frac{\partial g_{1dT}(\hat{\varphi}_{ML})}{\partial \sigma^2}, \frac{\partial g_{1dT}(\hat{\varphi}_{ML})}{\partial \rho} \right)^T,$$

and

$$\mathbf{b}_{\hat{\varphi}_{ML}}(\hat{\varphi}_{ML})^T = \frac{1}{2m} \left((\mathbf{I}_{\beta}(\hat{\varphi}_{ML}))^{-1} col_{1 \leq j \leq 3} tr((\mathbf{I}_{\beta}(\hat{\varphi}_{ML}))^{-1} \frac{\partial \mathbf{I}_{\beta}(\hat{\varphi}_{ML})}{\partial \varphi}) \right)$$

where $\mathbf{I}_{\beta}$ is the information matrix associated with the estimation of β . Given that $\mathbf{I}_{\beta} = \sum_{d=1}^m \mathbf{X}_d^T \mathbf{V}_d^{-1} \mathbf{X}_d$, we get:

$$\begin{aligned} \mathbf{b}_{\hat{\varphi}_{ML}}(\hat{\varphi}_{ML})^T = & -\frac{1}{2m} \left((\mathbf{I}_{\beta}(\hat{\varphi}_{ML}))^{-1} col_{1 \leq j \leq 3} tr \left(\left(\sum_{d=1}^m \mathbf{X}_d^T \mathbf{V}_d^{-1} (\hat{\varphi}_{ML}) \mathbf{X}_d \right)^{-1} \right. \right. \\ & \left. \left. \left(\sum_{d=1}^m \mathbf{X}_d^T \mathbf{V}_d^{-1} (\hat{\varphi}_{ML}) \mathbf{V}_{d(\varphi_j)} (\hat{\varphi}_{ML}) \mathbf{V}_d^{-1} (\hat{\varphi}_{ML}) \mathbf{X}_d \right) \right) \right). \end{aligned} \quad (7.32)$$

7.4 Simulation Results

We conducted a limited simulation study by generating populations from the random walk model defined by (7.1) and (7.2) where $\rho = 1$. We used $m = 40$ small areas and considered only small ($T = 5$) and moderate ($T = 10$) time periods. We also considered nine combinations of the pair (σ_u^2, σ^2) by letting each variance component takes the

values 0.5, 1, 1.5. The sampling variance component on the diagonal of Σ_d were equal to 1. The rest of the covariance matrix was obtaining assuming an autoregressive correlation structure with coefficient of 0.90 that is $\text{Corr}(e_{dj}, e_{dj'}) = 0.9^{|j-j'|}$. As mentioned earlier, we do not estimate Σ_d and treat it as known. We generated an intercept X_0 and two other auxiliary variables such that $P(X_1 = 1) = 0.7$ and $P(X_2 = 1) = 0.2$. The auxiliary variables were held fixed across the simulated populations. The fixed effects were set to $\beta = (1, 1, 1)^T$. A total of 5,000 populations were generated for each set of parameters.

To estimate the parameters of the model, we used ML method essentially because of the simulation time. In a real application, we would prefer REML method because the bias of the REML estimators of variance components is of lower order than that of the ML estimators (see Datta and Lahiri (2000) and Das et al. (2004)). In the next three tables, the absolute relative bias (ARB) is computed to assess the ML parameter estimation under the general Rao-Yu. The ARB is defined as

$$ARB(\delta_k) = \frac{1}{5000} \sum_{i=1}^{5000} \frac{\hat{\delta}_k[i] - \delta_k}{\delta_k}, \quad k = 1, \dots, 3$$

where δ_k indicates one of the model parameters ρ , σ^2 , or σ_u^2 . Table 7.1 shows the ARBs of the coefficient ρ . For the most part, these ARBs are very low. The ARBs are higher, around 10%, for the highest value of $\sigma^2 = 1.5$ and $T = 10$.

The ARBs from Figure 7.2 show that estimation of σ^2 is less precise than for ρ . However the ARBs are all below 5% except for $\sigma^2 = 1.5$ and $T = 10$. At 16%, the ARBs are not very bad however they are moderately high.

Again, looking at the ARBs in Figure 7.3, the estimation is not much biased except for $\sigma^2 = 1.5$ and $T = 10$. This is not a surprise since estimation of area level parameters is always more difficult than sub-area level parameters. However, except

Table 7.1: ARBs for estimation of the coefficient ρ using ML method. We simulated 5,000 samples for each combination of $\sigma_u^2 = 0$ and σ^2 .

T		$\sigma^2 = 0.5$	$\sigma^2 = 1$	$\sigma^2 = 1.5$
5	$\sigma_u^2 = 0.5$	0.0076	0.0280	0.0563
	$\sigma_u^2 = 1$	-0.0074	0.0028	0.0196
	$\sigma_u^2 = 1.5$	-0.0111	-0.0038	0.0062
10	$\sigma_u^2 = 0.5$	0.0058	0.0207	0.1105
	$\sigma_u^2 = 1$	-0.0023	0.0144	0.0928
	$\sigma_u^2 = 1.5$	-0.0041	0.0158	0.0906

Table 7.2: ARBs for estimation of the coefficient σ^2 using ML method. We simulated 5,000 samples for each combination of $\sigma_u^2 = 0$ and σ^2 .

T		$\sigma^2 = 0.5$	$\sigma^2 = 1$	$\sigma^2 = 1.5$
5	$\sigma_u^2 = 0.5$	-0.0420	-0.0320	0.0209
	$\sigma_u^2 = 1$	-0.0347	-0.0266	0.0007
	$\sigma_u^2 = 1.5$	-0.0324	-0.0248	-0.0115
10	$\sigma_u^2 = 0.5$	-0.0130	-0.0054	0.1605
	$\sigma_u^2 = 1$	-0.0136	0.0122	0.1556
	$\sigma_u^2 = 1.5$	-0.0010	0.0231	0.1586

for ρ and σ^2 , the estimation of σ_u^2 is very unstable for $\sigma^2 = 1.5$, $\sigma_u^2=0.5$, and $T = 10$ with an ARB of 50%. This combination of parameters ($\sigma^2 = 1.5$, $\sigma_u^2=0.5$, and $T = 10$) is the most biased in all three tables.

Table 7.3: ARBs for estimation of the coefficient σ_u^2 using ML method. We simulated 5,000 samples for each combination of $\sigma_u^2 = 0$ and σ^2 .

T		$\sigma^2 = 0.5$	$\sigma^2 = 1$	$\sigma^2 = 1.5$
5	$\sigma_u^2 = 0.5$	-0.0286	0.0312	0.1003
	$\sigma_u^2 = 1$	-0.0521	-0.0399	-0.0199
	$\sigma_u^2 = 1.5$	-0.0462	-0.0494	-0.0369
10	$\sigma_u^2 = 0.5$	-0.0428	0.0114	0.5004
	$\sigma_u^2 = 1$	-0.0526	-0.0141	0.2142
	$\sigma_u^2 = 1.5$	-0.0472	-0.0158	0.1339

Chapter 8

Future Research

We have provided the empirical best linear unbiased prediction (EBLUP) and the empirical best (EB) estimators of small area linear parameters for both the one-fold and the two-fold nested model with errors following skew-normal (SN) distributions. Under the one-fold nested error model with errors following SN distributions, two empirical best prediction approaches (marginal and conditional on u_d , $d = 1, \dots, m$ where m is the number of small areas) and an HB method were developed to estimate complex parameters. Nevertheless some aspects of the methods can be improved. Some future research ideas are discussed below. Here we refer to the nested error model with the errors following SN distributions as the SN model and similarly the normal model refers to the nested error model with the errors following normal distributions.

In Chapter 3, we developed a maximum likelihood (ML) method for the one-fold SN model. Simulation study showed good estimates of the model parameters except for the shape parameter λ_u of the random effects. To improve estimation of the parameter λ_u , several methods can be explored such as restricted maximum likelihood (REML) and expectation-maximization (EM). REML would be particularly indicated in situations of very unbalanced sample sizes across small areas or large rank of the auxiliary matrix $\mathbf{X}$ relative to the sample size. The EM algorithm is based on the utilization of

unobserved data. We may think of at least two ways of specifying the EM algorithm. The first approach is to consider the random effects u_d , $d = 1, \dots, m$, as the unobserved data and use the joint probability density function (pdf) of $\mathbf{Y}_c = (\mathbf{Y}^T, \mathbf{u}^T)^T$ as the likelihood function for the EM method. The second approach consists of using the stochastic representation (4.3.2). This stochastic representation shows that any vector $\mathbf{Y}$ following closed skew-normal (CSN) can be expressed as $\mathbf{Y} = \boldsymbol{\mu} + \mathbf{B}_0 \mathbf{T}_0 + \mathbf{B}_1 \mathbf{T}_1$ where $\mathbf{T}_0$ follows a left truncated multivariate normal, $\mathbf{T}_1$ follows a multivariate normal, $\boldsymbol{\mu}$ is a vector, and $\mathbf{B}_0$ and $\mathbf{B}_1$ are matrices. The likelihood function for this second EM method would be the joint pdf of $\mathbf{Y}_c = (\mathbf{Y}^T, \mathbf{T}_0^T)^T$.

In Chapter 4, it would be interesting to compare the EB estimator based on the SN distribution family to the EB estimator based on the finite mixture of normals developed by Elbers and Weide (2014). The latter estimator uses a mixture of finite number of normals to approximate any well-behaved non-normal distribution followed by the random errors of the nested model. Implementation of mixtures of normals should be done with a lot of care and this is illustrated by the following quote from Chen and Li (2009): “*Contrary to intuition, of all the finite mixture models, the normal mixture models have the most undesirable mathematical properties. Their likelihood functions are unbounded unless the component variances are assumed equal or constrained, the Fisher information can be infinite and the strong identifiability condition is not satisfied.*”. Some of the challenges using mixtures of normals are:

- *More parameters to estimate.* For the normal model, there are two variances components to estimate and four parameters to estimate under the SN model. For the mixture-normal nested error model there are $(2k_u - 1) + (2k_e - 1)$ variance components to estimate, where k_u and k_e are the number of normals in the mixtures for modeling respectively u_d and e_{dj} . If $k_u = k_e = 2$ then six variance components need to be estimated, if $k_u = k_e = 3$ then 10 variance components

need to be estimated, and so on. Given the small sample sizes, if the number of small areas is not large then the estimation of the area level parameters may be difficult.

- *Issues on the boundary of the parameter space.* Consider the following mixture of normals $\phi(x; \omega_1, \dots, \omega_{K-1}, \mu_1, \dots, \mu_K, \sigma_1^2, \dots, \sigma_K^2) = \sum_{k=1}^K \omega_k \phi(x; \mu_k, \sigma_k^2)$, where ϕ is the pdf of the univariate normal distribution and $\omega_K = 1 - \sum_{k=1}^{K-1} \omega_k$. If $\omega_{k_0} = 0$, where $1 \leq k_0 \leq K$ then $\phi(x; \omega_1, \dots, \omega_{K-1}, \mu_1, \dots, \mu_K, \sigma_1^2, \dots, \sigma_K^2)$ is not identifiable since $\mu_{k_0}^1 \neq \mu_{k_0}^2$ implies $\phi(x; \dots, \mu_{k_0}^1, \dots) = \phi(x; \dots, \mu_{k_0}^2, \dots)$. Similarly if $\mu_1, \dots, \mu_K = \mu$ and $\sigma_1^2, \dots, \sigma_K^2 = \sigma^2$ then ϕ is not identifiable since different values of ω_k would give the same distribution $\phi(x; \omega_1, \dots, \omega_{K-1}, \mu_1, \sigma_1^2, \dots, \mu_K, \sigma_K^2) = \phi(x; \mu, \sigma^2)$. Other issues on the boundary of the parameter space are infinite likelihood and fisher information (see Chen and Li (2009)).

In Chapter 5, the hierarchical Bayes (HB) method uses sampling importance resampling (SIR). However the importance weights are more dispersed than desired. The importance weights are the consequence of approximating the SN model posterior by the normal model posterior. Therefore to improve the method, we may look for an importance function closer to being proportional to the SN model posterior and give up the idea of using the normal model based posterior. Another idea is to explore the half normal representation of the SN distribution in order to find a better HB model.

In Chapter 6, the estimators address only linear parameters under the two-fold SN model. The next step is to develop the EB estimators for complex parameters. The estimator based on the distribution of $\mathbf{Y}_{dr} | \mathbf{Y}_{ds} = \mathbf{y}_{ds}$ will be challenging specially the necessity of generating the values in a univariate manner. The covariances associated with $\mathbf{T}_0$ and $\mathbf{T}_1$ do not have a simple form which allows easy univariate draws. To solve this issues, we may investigate the Cholesky decomposition $\mathbf{LDL}^T$, where $\mathbf{L}$ is a

lower unit triangular (diagonal of 1's) matrix and $\mathbf{D}$ is a diagonal matrix. Consider a random vector $\mathbf{Y}$ such that $\mathbf{Y} \sim N_n(\mathbf{0}, \mathbf{\Sigma})$. If y_j , where $j = 1, \dots, n$, is not a linear combination of the other elements of $\mathbf{Y}$ then the Cholesky decomposition, say $\mathbf{\Sigma} = \mathbf{LDL}^T$, exists and is unique. We may write the two matrices $\mathbf{L}$ and $\mathbf{D}$ as follows:

$$\mathbf{L} = \begin{pmatrix} 1 & 0 & \dots & \dots & \dots & 0 \\ \ell_{2,1} & 1 & 0 & \dots & \dots & 0 \\ \vdots & \dots & \dots & \dots & \dots & 0 \\ \ell_{j,1} & \dots & \ell_{j,j-1} & 1 & \dots & 0 \\ \vdots & \dots & \dots & \dots & \dots & 0 \\ \ell_{n,1} & \dots & \dots & \dots & \ell_{n,n-1} & 1 \end{pmatrix}$$

and

$$\mathbf{D} = \text{diag}(d_{1,1}, d_{2,2}, \dots, d_{n,n})$$

To generate $\mathbf{Y}$ in a univariate manner, we first generate n values z_j , $j = 1, \dots, n$ from independent $N(0, 1)$ then we obtain the elements of $\mathbf{y}$ as follows:

$$y_1 = z_1 \sqrt{d_{1,1}}$$

$$y_j = z_j \sqrt{d_{j,j}} + \sum_{k=1}^{j-1} z_k \ell_{j,k} \sqrt{d_{k,k}}, \quad 1 < j \leq n$$

The only question left to answer to complete the univariate generation of the elements of $\mathbf{Y}$ is to be able to create the matrices $\mathbf{L}$ and $\mathbf{D}$ in an univariate manner. This is easily done by the following recursive relations

$$d_{j,j} = \Sigma_{(j,j)} - \sum_{k=1}^{j-1} \ell_{j,k}^2 d_{k,k}$$

$$\ell_{j,i} = \frac{1}{d_{j,j}} \left(\Sigma_{(j,i)} - \sum_{k=1}^{i-1} \ell_{j,k} \ell_{i,k} d_{k,k} \right), \quad i < j \leq n,$$

where $\Sigma_{(j,i)}$ is the (j, i) entry of the covariance matrix Σ . Note that the algorithm can be simplified even further by considering the decomposition $\mathbf{L}\mathbf{L}^T$. Of course if the matrix Σ is very large then these operations will take a lot of time but they are always possible regardless of the complexity of the covariance matrix. It may be possible to use the block structure of the covariance matrix to reduce the computational time. Unfortunately this decomposition may not solve the problem in the case of the two-fold SN model because of the involvement of the multivariate left truncated normal. A solution would be to use the conditional approach which consists of using the sample to predict the area and sub-area effects then draw unit level errors from the distribution of e_{dj} to construct the censuses. The prediction of u_d may require dealing with multivariate vectors of dimension $n_d + q$, where q is the number of area and sub-area effects combined. These are small vectors and easily manageable by current computational capabilities.

Another issue to solve under the two-fold SN nested error model is the parameter estimation. The reduction technique used for the one-fold SN model may not apply. The ML method would require an accurate numerical approximation of the multivariate normal cumulative distribution function Φ_{n_d} for small to moderate n_d .

In Chapter 7, the general Rao-Yu model may be extended to the multivariate case (jointly modeling several characteristics of interest) and to hierarchical Bayes.

Bibliography

- Arellano-Valle, R. B. and Azzalini, A. (2006). On the unification of the families of skew-normal distributions. *Scandinavian Journal of Statistics*, 33:516–574.
- Arellano-Valle, R. B., Bolfarine, H., and Lachos, V. H. (2005). Skew-normal linear mixed models. *Journal of Data Science*, 3:415–438.
- Arellano-Valle, R. B., Bolfarine, H., and Lachos, V. H. (2007). Bayesian inference for skew-normal linear mixed models. *Journal of Applied Statistics*, 34(6):663–682.
- Arellano-Valle, R. B., Del Pino, G., and San Martin, E. (2002). Definition and probabilistic properties of skew-distributions. *Statistics & Probability Letters*, 58:111–121.
- Azzalini, A. (1985). A class of distribution which includes the normal ones. *Scandinavian Journal of Statistics*, 12:171–178.
- Azzalini, A. (1996). The multivariate skew-normal distribution. *Biometrika*, 83:715–726.
- Azzalini, A. (2011). Cran package sn: The skew-normal and skew-t distributions. <http://azzalini.stat.unipd.it/SN/>.
- Azzalini, A. and Capitanio, A. (1999). Statistical applications of multivariate skew-normal distributions. *Journal of the Royal Statistical Society: Series B*, 61:579–602.

- Battese, G. E., Harter, R. M., and Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83:28–36.
- Bell, W. (2002). Jackknife in the Fay-Herriot model with an application: discussion. *Proceedings of the Seminar on Funding Opportunities in Survey Research*, pages 98–104. Washington DC: U.S Census Bureau.
- Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis (2nd edition)*. Springer-Verlag, New-York.
- Chatterjee, S., Lahiri, P., and Li, H. (2008). Parametric bootstrap approximation to the distribution of EBLUP and related prediction intervals in linear mixed models. *Journal of the Royal Statistical Society: Series B*, 59(3):1221–1245.
- Chen, J. and Li, P. (2009). Hypothesis test for normal mixture models: The em approach. *The Annals of Statistics*, 37:2523–2542.
- Cochran, W. G. (1977). *Sampling Techniques*. John Wiley and Sons, Hoboken, New Jersey.
- Copas, J. B. and Li, H. G. (1997). Inference for non-random samples (with discussion). *The Annals of Statistics*, 36:55–95.
- Correa, L. (2012). Comparison of methods for estimation of poverty indicators in small areas. Master’s final project, Universidad Carlos III de Madrid.
- Das, K., Jiang, J., and Rao, J. N. K. (2004). Mean squared error of empirical predictor. *The Annals of Statistics*, 32(2):818–840.
- 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., Lahiri, P., and Maiti, T. (2002). Empirical Bayes estimation of median income of four-person families by state using time series and cross-sectional data. *Journal of Statistical Planning and Inference*, 102:83–97.
- Dominguez-Molina, J. A., Gonzales-Farias, G., and Gupta, A. K. (2003). *The multivariate closed skew normal distribution*. Department of Mathematics and Statistics, Bowling Green State University, Technical Report, No. 03-12.
- Dunnett, C. W. and Sobel, M. (1955). Approximations of the probability integral and certain percentage points of a multivariate analogue of student's t-distribution. *Biometrika*, 42:258–260.
- Elbers, C., Lanjouw, J. O., and Lanjouw, P. (2003). Micro-level estimation of poverty and inequality. *Econometrica*, 71:355–364.
- Elbers, C. and Weide, R. (2014). Estimation of normal mixtures in a nested error model with an application to small area estimation of poverty and inequality. Technical report, Development Research Group at the World Bank.
- Fay, R. E. and Diallo, M. S. (2012). Small area estimation alternatives for the national crime victimization survey. In *JSM Proceedings, Survey Research Methods Section, American Statistical Association*, pages 3742–3756.
- Fay, R. E. and Herriot, R. A. (1979). Estimation of income from small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74:269–277.
- Flecher, C., Naveau, P., and Allard, D. (2009). Estimating the closed skew-normal distributions parameters using weighted moments. *Statistics and Probability Letters*, 79(19):1977–1984.

- Foster, J., Greer, J., and Thorbecke, E. (1984). A class of decomposable poverty measures. *Econometrica*, 52:761–766.
- Fuller, W. A. and Battese, G. (1973). Transformations for estimation of linear models with nested-error structure. *Journal of the American Statistical Association*, 68:626–632.
- Genton, M. G. (2004). *Skew-Elliptical Distributions And Their Applications, A Journey Beyond Normality*. Chapman and Hall/CRC, New York.
- Ghosh, M. and Steorts, R. (2013). Two-stage Bayesian benchmarking as applied to small area estimation. *TEST*, 22(4):670–687.
- Gonzalez-Manteiga, W., Lombardia, M., Molina, I., Morales, D., and Santa-Maria, L. (2007). Estimation of the mean squared error of predictors of small area linear parameters under logistic mixed model. *Computational Statistics and Data Analysis*, 51:2720–2733.
- Hall, P. and Maiti, T. (2006a). Nonparametric estimation of mean-squared prediction error in nested-error regression models. *The Annals of Statistics*, 34:1733–1750.
- Hall, P. and Maiti, T. (2006b). On parametric bootstrap methods for small area prediction. *Journal of the Royal Statistical Society: Series B*, 68:221–238.
- Harville, D. A. and Jeske, D. R. (1992). Mean squared error of estimation or prediction under general linear model. *Journal of the American Statistical Association*, 87:724–731.
- Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A*, 186:453–461.
- Jeffreys, H. (1961). *Theory of Probability (third edition)*. Oxford University Press, London.

- Jiang, J., Lahiri, P., and Wan, S. M. (2002). A unified jackknife theory for empirical best prediction with M-estimation. *The Annals of Statistics*, 30:1782–1810.
- Kackar, R. N. and Harville, D. A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association*, 79:853–862.
- Kass, R. E. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91:1343–1370.
- Kotz, S., Balakrishnan, N., and Johnson, N. L. (2000). *Continuous Multivariate Distributions, Volume 1: Models and Applications*. John Wiley and Sons, Hoboken, New Jersey.
- Lachos, V. H., Bolfarine, H., Arellano-Valle, R. B., and Montenegro, L. C. (2007). Likelihood based inference for multivariate skew-normal regression models. *Communications in Statistics: Theory and Methods*, 36:1769–1786.
- Lachos, V. H., Gosh, P., and Arellano-Valle, R. B. (2010). Likelihood based inference for skew-normal independent linear mixed models. *Statistica Sinica*, 20:303–322.
- Lahiri, S. N., Maiti, T., Katzoff, M., and Parsons, V. (2007). Resampling-based empirical prediction: an application to small area estimation. *Biometrika*, 94:469–485.
- Lele, S. R., Dennis, B., and Lutschers, F. (2007). Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov Chain Monte Carlo methods. *Ecology Letters*, 105:551–563.
- Lele, S. R., Nadeem, K., and Schmuland, B. (2010). Estimability and likelihood inference for generalized linear mixed models using data cloning. *Journal of the American Statistical Association*, 105:1617–1625.

- Lin, T. I. and Lee, J. C. (2008). Estimation and prediction in linear mixed models with skewnormal random effects for longitudinal data. *Statistics in Medicine*, 27:1490–1507.
- Lohr, S. L. (2010). *Sampling: design and analysis*. Cengage Learning.
- Marhuenda, Y., Molina, I., and Morales, D. (2013). Small area estimation with spatio-temporal Fay-Herriot models. *Computational Statistics and Data Analysis*, 58:308–325.
- McCulloch, C. E., Searle, S. R., and Neuhaus, J. M. (2008). *Generalized, Linear, and Mixed Models*. John Wiley and Sons, Hoboken, New Jersey.
- Metropolis, N. and Ulam, S. (1949). The Monte Carlo method. *Journal of the American Statistical Association*, 44:335–341.
- Mohadjer, L., Rao, J. N. K., Liu, B., Krenzke, T., and Van De Kerckhove, W. (2012). Hierarchical Bayes small area estimates of adult literacy using unmatched sampling and linking models. *Journal of Indian Society of Agricultural Statistics*, 66(1):55–63.
- Molina, I., Nandram, B., and Rao, J. N. K. (2014). Small area estimation of general parameters with application to poverty indicators: A hierarchical Bayes approach. *The Annals of Applied Statistics*, 8:852–885.
- Molina, I. and Rao, J. N. K. (2010). Small area estimation of poverty indicators. *The Canadian Journal of Statistics*, 38:369–385.
- Murray, J. D. (2002). *Mathematical Biology: I. An introduction*. Springer-Verlag, New York.
- Nandram, B. and Choi, J. W. (2005). Hierarchical Bayesian nonignorable nonresponse regression models for small areas: An application to the NHANES data. *Survey Methodology*, 31:73–84.

- Nandram, B. and Choi, J. W. (2010). A Bayesian analysis of body mass index data from small domains under nonignorable nonresponse and selection. *Journal of the American Statistical Association*, 105:120–135.
- Nandram, B., Sedransk, J., and Pickle, L. W. (2000). Bayesian analysis and mapping of mortality rates for chronic obstructive pulmonary disease. *Journal of the American Statistical Association*, 95:1110–1118.
- Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science*, 28:40–68.
- Pfeffermann, D. and Burck, L. (1990). Robust small area estimation combining time series and cross-sectional data. *Survey Methodology*, 16:217–237.
- Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., and Rasbash, J. (1998). Parametric distributions of complex survey data under informative probability sampling. *Statistica Sinica*, 8:1087–1114.
- 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. (2003). *Small Area Estimation*. John Wiley and Sons, Hoboken, New Jersey.
- Rao, J. N. K. and Yu, M. (1992). Small area estimation combining time series and cross-sectional data. In *JSM Proceedings, Survey Research Methods Section*, American Statistical Association, pages 1–9.
- Rao, J. N. K. and Yu, M. (1994). Small area estimation by combining time series and cross-sectional data. *Canadian Journal of Statistics*, 22:511–528.

- Robert, C. P. (2007). *The Bayesian Choice (Second Edition)*. Springer-Verlag, New-York.
- Roberts, C. (1966). A correlation model useful in the study of twins. *Journal of the American Statistical Association*, 61:1184–1190.
- Sahu, K., Dey, D. K., and Branco, M. D. (2003). A new class of multivariate skew distributions with applications to bayesian regression models. *The Canadian Journal of Statistics*, 31:129–150.
- Sarndal, C.-E., Swensson, B., and Wretman, J. H. (2003). *Model assisted survey sampling*. Springer-Verlag, New York.
- Shao, J. and Tu, D. (1995). *The Jackknife and Bootstrap*. Springer-Verlag, New York.
- Stukel, D. M. and Rao, J. N. K. (1997). Estimation of regression models with nested error structure and unequal error variances under two and three stage cluster sampling. *Statistics and Probability Letters*, 35:401–407.
- Stukel, D. M. and Rao, J. N. K. (1999). Small-area estimation under two-fold nested errors regression models. *Journal of Statistical Planning and Inference*, 78:131–147.
- Torabi, M. and Shokoohi, F. (2012). Likelihood inference in small area estimation by combining time-series and cross-sectional data. *Journal of Multivariate Analysis*, 111:213–221.
- Toto, M. C. S. and Nandram, B. (2010). A Bayesian predictive inference for small area means incorporating covariates and sampling weights. *Journal of Statistical Planning and Inference*, 140:2963–2979.
- Walker, A. M. (1969). On the asymptotic behavior of posterior distributions. *Journal of the Royal Statistical Society: Series B*, 31:80–88.

Wolker, K. M. (1985). *Introduction to Variance Estimation*. Springer-Verlag, New York.