Robust Estimation for the Annual Survey of Public Employment & Payroll Using Mixture of Linear Mixed-Effect Models with the MCMC Procedure

Giang Trinh
Giang.B.Trinh@census.gov
Bac Tran
Bac.Tran@census.gov
U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233

Abstract

The Public Sector Sample Design and Estimation Branch uses Horvitz-Thompson, Empirical Best Linear Unbiased Prediction (EBLUP) and Bayesian approach to small area estimation (SAE) for the Annual Survey of Public Employment & Payroll (ASPEP). The EBLUP estimator is based on a linear mixed-effect model (LMM) with errors that are assumed to be normally distributed. In this study we provide a robust estimate for the total number of full-time employees in the ASPEP using Bayesian method for a LMM assuming errors governed by a mixture of normal distributions. We specify the Markov Chain Monte Carlo (MCMC) procedure in order to produce samples for the LMM's parameter space. We then compare our research method to the existing methods being used at the U.S. Census Bureau. The Census of Governments (CoG), Survey of Public Employment & Payroll data of 2007 and 2012 were used for the evaluation of this research.

Key Words: Linear mixed-effect models, Mixture models, Bayesian method, MCMC procedure.

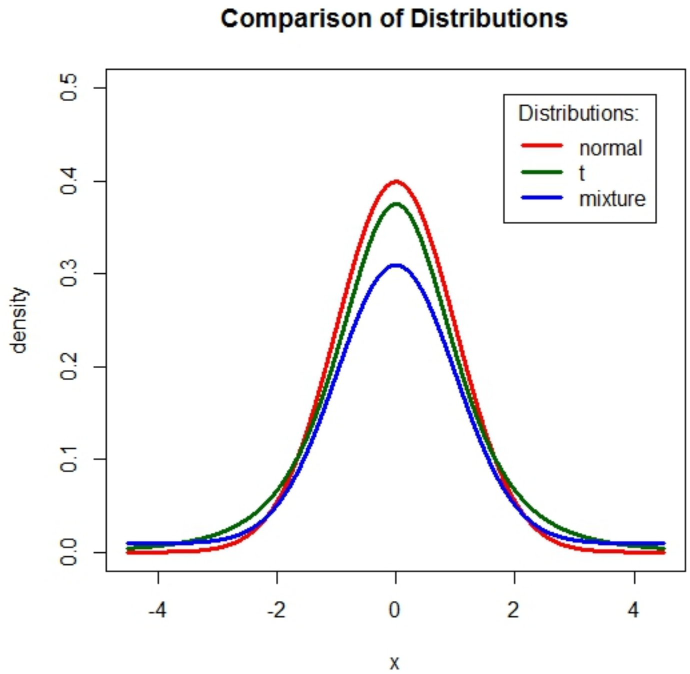
1 Introduction

The presence of outliers in the ASPEP data requires a robust regression approach when fitting linear mixed-effect models (LMMs). The error term in LMMs is assumed to be normally distributed. Outliers often appear when this assumption is violated. To accommodate outliers we may use error terms that follow a $t-$distribution that has relatively thicker tails (e.g. McLachlan and Basford [12] (1987), McLachlan and Peel [13] (2000), McLachlan, Ng, and Bean [14] (2006), Bell and Huang [2] (2006), Staudenmayer, Lake and Wand [17] (2009)). A more flexible way is to assume that the data are drawn from a finite mixture model, including the normally-distributed error models as a special case, in capturing a broader range of non-normal behaviors (see McLachlan and Peel [13] (2000), McLachlan and Basford [12] (1987), and McLachlan et al. [14] (2006)). Figure 1 shows the standard normal distribution, the $t-$distribution with 4 degrees of freedom, and a mixture of normal distributions. Notice that the mixture of normal distributions has relatively thicker tails among the three. In reality, real data often show more than one mode, and estimates using mixture models appear more reliable than the others.

− − − − − − − − − − − − − −

Any views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

Figure 1: $\mathcal{N}(0,1)$, $t_4(0,1)$ and Mixture $0.75\mathcal{N}(0,1) + 0.25\mathcal{N}(0,10)$



The mixture model parameters can be deduced by using Expectation-Maximization (EM) algorithm (e.g. McLachlan and Basford [12] (1987), De Veux and Krieger [5] (1990), McLachlan and Peel [13] (2000), Tadjudin and Landgrebe [18] (2000), Hall and Wang [10] (2005), McLachlan et al. [14] (2006), Gershunskaya and Lahiri [9] (2010), Trinh and Tran [19] (2016)), Gibbs Sampling or the MCMC procedure (e.g. Wand et al. [20] (1994), Woodworth [21] (2004), Bolstad [3, 4] (2007, 2010) using posterior sampling as indicated by Bayes' theorem:

$$P(\theta|\, y) \;=\; \frac{P(\theta)P(y|\,\theta)}{P(y)} \propto P(\theta)P(y|\,\theta) \tag{1.1}$$

where $P(y|\,\theta)$ is the likelihood function or the density of the data y given the parameters θ; $P(\theta)$ is the prior density of the parameters; and $P(\theta|\,y)$ is the posterior density of the parameters given the data. Once the likelihood and the prior density are specified, inferences on parameters can be made by using samples drawn from the posterior distribution produced by the MCMC simulation.

This report is organized as follows: In Section 2, three Hierarchical Bayes (HB) estimators (using normal, t, and mixture of normal distributions) are specified in order to estimate the total number of full-time employees for various government functions in the ASPEP data. Section 3 shows convergence diagnostics and test results. The 2012 ASPEP data (in Alabama, California, Georgia, Illinois, and Louisiana) are used, with the 2007 ASPEP data as auxiliary information. The performance of the three HB estimators is assessed by Relative Root Mean Square Errors (RRMSE).

2 Robust Estimations

To accommodate outliers in the ASPEP data, three LMMs will be used: normal, t (e.g. Lange, Little, and Taylor [11] (1989), Bell and Huang [2] (2006)), and mixture of normal distributions (e.g. De Veux and Krieger [5] (1990), Gershunskaya and Lahiri [9] (2010)). We will fit these models using Bayesian methods.

Let y_{mk} denote the value of the k^{th} unit within the m^{th} government function (area). We are interested in estimating the total $Y_m = \sum_{k=1}^{N_m} y_{mk}$ for $m = 1, ..., M$ (N_m: number of units of the m^{th} area; M: number of areas). An estimator of Y_m is given by:

$$\widehat{Y}_m = y_m + \widehat{Y}_{mr} \tag{2.1}$$

$$= N_m[f_m \overline{y}_m + (1 - f_m)\widehat{\overline{\boldsymbol{Y}}}_{\boldsymbol{mr}}] \tag{2.2}$$

where $y_m = \sum_{k=1}^{n_m} y_{mk}$: the sum of the sample values; $\widehat{Y}_{mr}$ is a predictor of the total of the non-sampled part of the m^{th} area; $\overline{y}_m = \frac{y_m}{n_m}$: the sample mean; $f_m = \frac{n_m}{N_m}$: the sampling rate; n_m: the sample size; and $\widehat{\overline{Y}}_{mr}$: a predictor for the mean of the non-sampled part of the m^{th} area. The predictor $\widehat{Y}_{mr}$ can be derived from a LMM or Fay-Herriot model (see Fay and Herriot [7] (1979)).

Hierarchical Bayesian (HB) Model

$$\log(y_{mk}) = \beta_1 + \boldsymbol{\beta}_2 \log(x_{mk}) + u_m + \varepsilon_{mk}, \tag{2.3}$$

where y_{mk} and x_{mk} are the number of full-time employees from the survey year and census year, respectively; $m = 1, 2, ..., 29$ denotes the m^{th} area; $k = 1, 2, ..., N_m$ the k^{th} unit; u_m is the random effect of the m^{th} area, ε_{mk} is the error term. The log-transformation is applied to x_{mk} and y_{mk} to make the predictor and response variables approximately conform to normality. Then y_{mk} ($m = 1, ..., M$; $k = n_m + 1, ..., N_m$) is predicted using the inverse transformation $\widehat{y}_{mk} = \exp(\widehat{\beta}_1 + \widehat{\boldsymbol{\beta}}_{\boldsymbol{2}} \log(x_{mk}) + \widehat{u}_m)$.

2.1 HB model assuming normally distributed errors (N-model) given by (2.3), (2.4)-(2.5):

$$u_m | \tau^2 \overset{\text{iid}}{\sim} \mathcal{N}(0, \tau^2), \tag{2.4}$$

$$\varepsilon_{mk} | \sigma^2 \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2). \tag{2.5}$$

MCMC specification for the N-model:

$$\text{Parms}: \quad \beta_1 = 0, \ \beta_2 = 1, \ \tau^2 = 1, \ \sigma^2 = 1 \tag{2.6}$$

$$\text{Priors}: \quad \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} \sim \text{BVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 50 & 0 \\ 0 & 50 \end{bmatrix} \right) \tag{2.7}$$

$$\tau^2, \ \sigma^2 \sim \text{igamma}(0.01, \ 0.01) \tag{2.8}$$

$$\text{Random}: \quad u_m \sim \mathcal{N}(0, \tau^2) \tag{2.9}$$

$$\text{Likelihood}: \quad \log(y_{mk}) | \, u_m \sim \mathcal{N}(\beta_1 + \beta_2 \log(x_{mk}) + u_m, \ \sigma^2) \tag{2.10}$$

2.2 HB model with t-distributed errors (t-model) given by (2.3), (2.11)-(2.12):

$$u_m|\,\tau^2 \overset{\text{iid}}{\sim} \mathcal{N}(0,\tau^2), \tag{2.11}$$
$$\varepsilon_{mk}|\,\sigma^2 \overset{\text{iid}}{\sim} t(0,\sigma^2,\nu). \tag{2.12}$$

MCMC specification for the t-model:

$$\text{Parms}: \quad \beta_1 = 0,\ \beta_2 = 1,\ \tau^2 = 1,\ \sigma^2 = 1 \tag{2.13}$$
$$\text{Priors}: \quad \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} \sim \text{BVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 50 & 0 \\ 0 & 50 \end{bmatrix} \right) \tag{2.14}$$
$$\tau^2,\ \sigma^2 \sim \text{igamma}(0.01,\ 0.01) \tag{2.15}$$
$$\text{Random}: \quad u_m \sim \mathcal{N}(0,\tau^2) \tag{2.16}$$
$$\text{Likelihood}: \quad \log(y_{mk})|\,u_m \sim t(\beta_1 + \beta_2 \log(x_{mk}) + u_m,\ \sigma^2,\ \nu = 4) \tag{2.17}$$

2.3 HB model assuming error terms follow mixture of normal distributions (MN-model) given by (2.3), (2.18)-(2.20):

$$u_m|\,\tau^2 \overset{\text{iid}}{\sim} \mathcal{N}(0,\tau^2), \tag{2.18}$$
$$\varepsilon_{mk}|\,\sigma_1^2, \sigma_2^2, s_p \overset{\text{iid}}{\sim} (1-s_p)\mathcal{N}(0,\sigma_1^2) + s_p\mathcal{N}(0,\sigma_2^2),\ \sigma_1 < \sigma_2, \tag{2.19}$$
$$s_p|\,p \overset{\text{iid}}{\sim} \text{Bin}(1;p). \tag{2.20}$$

MCMC specification for the MN-model:

$$\text{Parms}: \quad \beta_1 = 0,\ \beta_2 = 1,\ s_p = 0,\ \tau^2 = 0.01,\ \sigma_1^2 = 0.01,\ \sigma_2^2 = 1 \tag{2.21}$$
$$p = \frac{1}{1+\exp(-s_p)} \tag{2.22}$$
$$\text{Priors}: \quad \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} \sim \text{BVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 50 & 0 \\ 0 & 50 \end{bmatrix} \right) \tag{2.23}$$
$$\tau^2,\ \sigma_1^2,\ \sigma_2^2 \sim \text{igamma}(0.01,\ 0.01) \tag{2.24}$$
$$s_p \sim \text{Bin}(1,p) \tag{2.25}$$
$$\text{Random}: \quad u_m \sim \mathcal{N}(0,\tau^2) \tag{2.26}$$
$$\mu = \beta_1 + \beta_2 \log(x_{mk}) + u_m,\ z_1 = \frac{\log(y_{mk}) - \mu}{\sigma_1},\ z_2 = \frac{\log(y_{mk}) - \mu}{\sigma_2}$$
$$\text{Likelihood}: \quad \log(y_{mk})|\,u_m \sim \frac{p}{\sigma_1}\exp(-\frac{z_1^2}{2}) + \frac{1-p}{\sigma_2}\exp(-\frac{z_2^2}{2}) \tag{2.27}$$

Notice that, random effects are set up using random-effect distributions as specified in equations (2.9), (2.16), and (2.26). Non-informative priors (equations (2.7), (2.14), (2.23), (2.8), (2.15), (2.24)) would not influence the posterior distribution (see Ojo et al. [15] (2017) and Gelman [8] (2006)). The degree of freedom $\nu = 4$ (equation (2.17)) is recommended in problems with small sampling rates (e.g. Lange et al. [11] (1989)). The MCMC procedures discard the first 2,500 as *burn-in* and keep the next $12,500$ samples. The *thinning* rate of 5 is applied to produce $2,500$ *thinned* samples from the posterior distribution. Then u_m is predicted by the average of MCMC posterior estimates of u_m.

3 Application to ASPEP data: MCMC Diagnostics and Test results

The ASPEP survey is designed to produce estimates of statistics on the number of federal, state, and local government civilian employees and their gross payroll for the month of March at the national level and for large domains. The target population of approximately 90,000 government units includes 5 types: counties, cities, townships, special districts, and school districts. The ASPEP consists of three components: a census of select federal agencies, a census of 50 state governments, and a sample of about 10,000 local governments. Every five years, in years ending in "2" and "7," the Census Bureau conducts a CoG. The employment component of the CoG, known as CoG-E, collects public employment and payroll data. About two years after every CoG-E, the Census Bureau redesigns and selects a new sample of local governments. The sample design is a two-phase, stratified, systematic probability-proportional-to-size design where the measure of size depends on total pay. See Dumbacher and Hill [6] (2014) for more details on the description of the sample design for the ASPEP.

To produce reliable estimates on the total number of full-time employees in government function codes where sampling rates are relatively small, we are exploring the Small Area Estimation (SAE) methodology that borrows strengths from previous census data instead of collecting expensive additional data for small cells (we refer the reader to Rao [16] (2003) for a comprehensive account on SAE techniques). Two recent consecutive censuses are used in this study, the 2007 and 2012 CoG-E..

The ASPEP data on employment include the number of full-time, part-time employees and gross pay as well as hours paid for part-time employees. The parameter of interest in this study is the total number of full-time employees, Y_m, for each function code $m = 1, 2, ..., 29$. We use samples from posterior distribution, produced by the MCMC procedure, to predict the total number of non-sampled full-time employees Y_{mr}. Then the estimate of the total number of full-time employees for the m^{th} area would be $\widehat{Y}_m = y_m + \widehat{Y}_{mr}$.

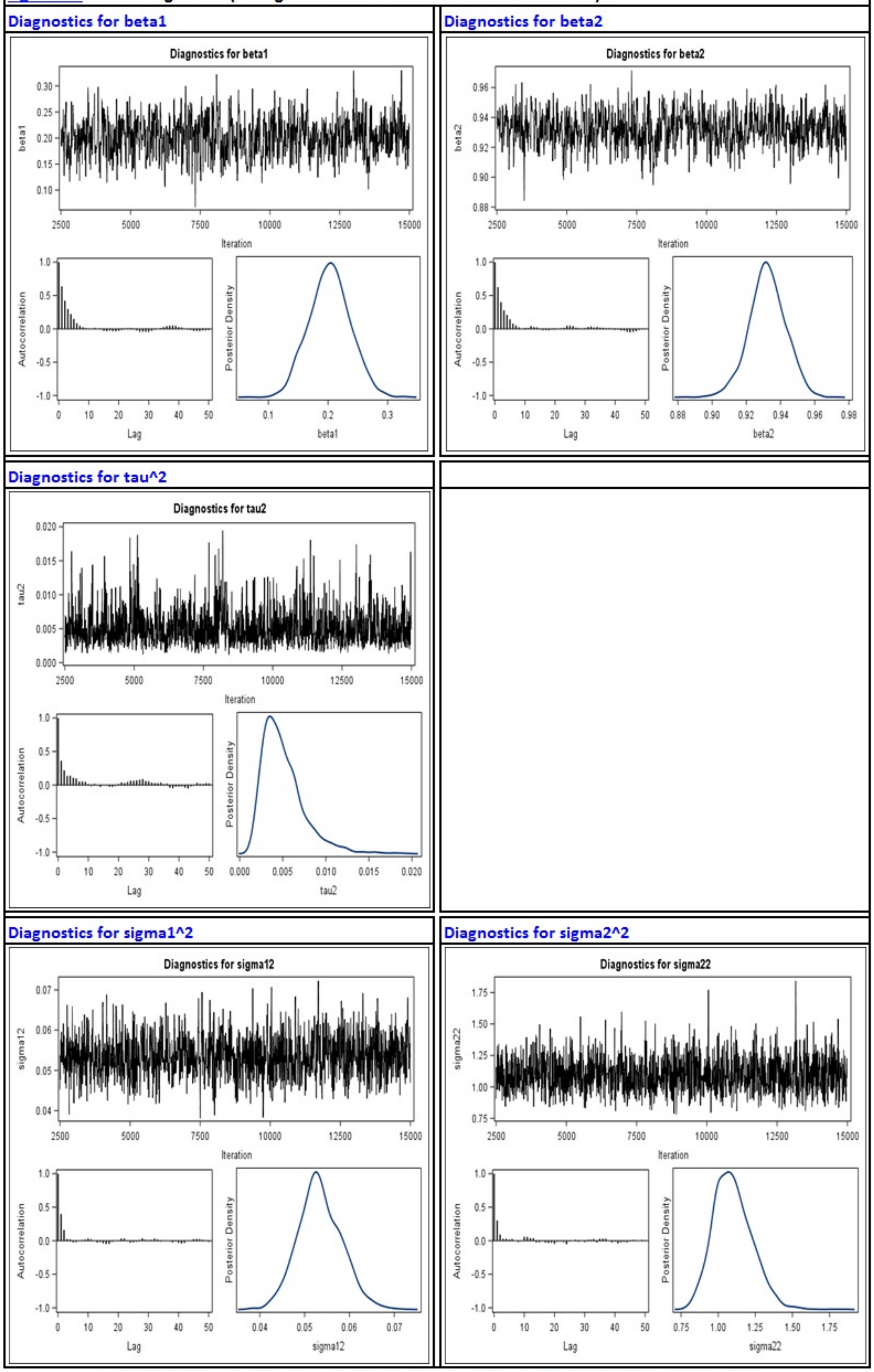
The convergence of the MCMC procedures can be assessed from Figures D1, D2, D3, D4, D5. The traceplots, of the *drawn value* of the parameters at each iteration against the number of MCMC iterations, all show good mixing of Markov chains. Samples from the posterior distribution can then be used to estimate the non-sampled total number of full-time employees for each function code.

The quality of the estimators is evaluated using the Relative Root Mean Square Error: $\text{RRMSE} = \sqrt{\frac{1}{\text{rep}}\sum_{i=1}^{\text{rep}}(\frac{\widehat{Y}_{m,i}-Y_m}{Y_m})^2}$ where $\widehat{Y}_{m,i}$ is an estimate of Y_m, $\text{rep} = 1,000$ is the number of replicate ASPEP samples selected from 2007 CoG data and used to estimate totals for 2012.

Let RRMSE_N, RRMSE_T, and RRMSE_{MN} (last three columns in each Table) be the RRMSEs of the HB estimates using normal, t, and mixture of normal distributions respectively. The following five tables show test results as applied to the ASPEP data of the states of Alabama, California, Georgia, Illinois, and Louisiana:

3.1 Application to Alabama ASPEP data - MCMC convergence diagnostics for $\theta = (\beta_1, \beta_2, \tau^2, \sigma_1^2, \sigma_2^2)$ and comparison of RRMSEs of the estimates

Figure D1: MCMC Diagnostics (Fitting Mixture Model to ALABAMA ASPEP data)



Data Source: U.S. Census Bureau 2007 and 2012 Census of Government - Employment

Table 1: RRMSEs of HB Estimates for Various Function Codes of Government (in Alabama)

FC	Y_m	med(f_m)	$RRMSE_N$	$RRMSE_T$	$RRMSE_{MN}$
001	448	6.00%	2.56%	1.30%	1.28%
005	3,196	24.90%	0.35%	0.48%	0.50%
012	64,006	24.24%	5.16%	2.58%	1.60%
023	4,120	25.77%	1.27%	1.10%	0.79%
024	5,521	19.22%	0.87%	1.62%	1.62%
025	1,369	21.63%	1.75%	0.78%	0.57%
029	2,350	19.20%	0.99%	3.32%	3.70%
032	5,151	21.93%	1.81%	1.28%	1.20%
040	20,817	11.27%	0.49%	0.43%	0.42%
044	5,973	24.70%	4.87%	2.39%	2.11%
050	1,888	7.74%	4.56%	2.02%	1.36%
052	885	14.37%	1.03%	0.56%	0.54%
059	422	5.08%	2.42%	1.19%	1.19%
061	2,834	21.98%	1.08%	0.72%	0.66%
062	10,075	23.77%	0.48%	1.23%	1.31%
079	865	11.74%	0.75%	0.88%	0.84%
080	1,387	16.49%	3.18%	1.30%	0.95%
081	3,197	23.80%	1.00%	1.09%	1.06%
089	4,339	0.58%	2.64%	1.97%	1.95%
091	3,749	22.82%	0.55%	1.53%	1.63%
092	1,062	11.45%	1.15%	1.28%	1.27%
093	845	9.60%	2.15%	1.15%	1.09%
094	359	8.21%	1.85%	1.26%	1.27%
112	30,366	5.13%	6.76%	6.13%	5.68%
124	271	23.53%	0.23%	0.23%	0.21%
162	2,169	14.73%	1.52%	1.18%	1.33%

FC: Function Code (see description in the Appendix)
med(f_m): median sampling rates out of 1,000 replicates

Figure T1: Comparing RRMSEs of HB LMMs using Normal, T, and Mixture of Normal distributions

Figure T1a: Count for Largest RRMSE across Function Codes

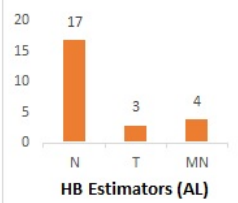


Figure T1b: Count for Smallest RRMSE across Function Codes

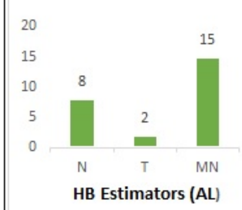
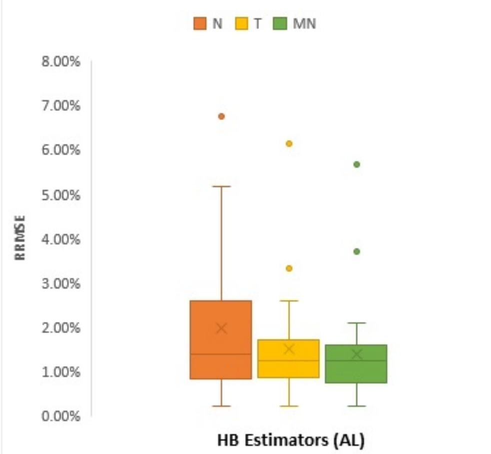


Figure T1c: Boxplots of RRMSE of HB LMMs using Normal, T, and Mixture of Normal distributions

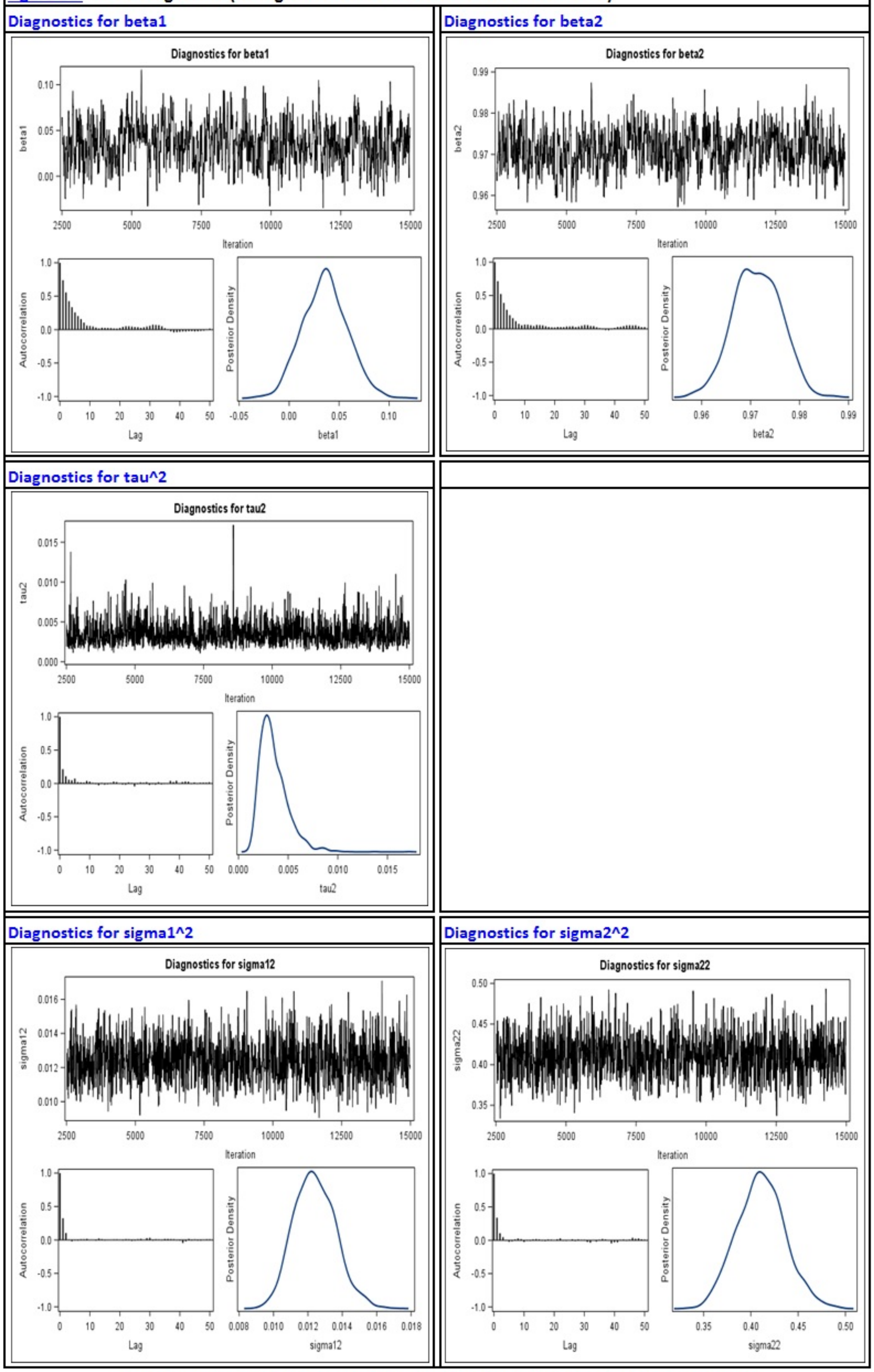


Data Source: U.S. Census Bureau 2007 and 2012 Census of Government - Employment

Figure T1a indicates the N-model is the least efficient. Figure T1b suggests the MN-model is the most efficient. Figure T1c shows that, on the average, the MN-model provides HB estimates with smallest RRMSEs and estimates using N-model have largest RRMSEs.

3.2 Application to California ASPEP data - MCMC convergence diagnostics for $\theta = (\beta_1, \beta_2, \tau^2, \sigma_1^2, \sigma_2^2)$ and comparison of RRMSEs of the estimates

Figure D2: MCMC Diagnostics (Fitting Mixture Model to CALIFORNIA ASPEP data)



Data Source: U.S. Census Bureau 2007 and 2012 Census of Government - Employment

Table 2: RRMSEs of HB Estimates for Various Function Codes of Government (in California)

FC	Y_m	med(f_m)	$RRMSE_N$	$RRMSE_T$	$RRMSE_{MN}$
001	6,832	19.81%	0.29%	0.30%	0.29%
005	29,177	21.51%	0.88%	0.71%	0.57%
012	344,113	21.16%	1.77%	1.70%	1.65%
016	27,817	7.83%	4.65%	3.79%	2.53%
018	19,578	7.85%	4.20%	3.00%	2.50%
023	24,341	23.09%	0.68%	0.77%	0.97%
024	27,646	13.97%	1.64%	1.08%	1.20%
025	36,257	33.94%	0.32%	0.49%	0.58%
029	24,521	22.63%	1.52%	1.71%	1.49%
032	40,691	17.63%	0.79%	0.34%	0.35%
040	55,640	20.73%	1.11%	0.87%	0.78%
044	20,826	19.74%	0.51%	0.42%	0.42%
050	11,306	22.18%	2.07%	1.12%	0.69%
052	7,377	14.24%	1.43%	1.10%	0.93%
059	7,031	7.61%	8.29%	3.70%	3.92%
061	17,471	18.34%	4.22%	2.81%	1.55%
062	62,868	26.05%	0.29%	0.32%	0.39%
079	65,826	23.29%	0.64%	0.32%	0.25%
080	11,243	14.29%	2.69%	1.60%	1.25%
081	8,213	23.64%	1.82%	1.41%	1.35%
087	2,193	7.28%	1.21%	1.16%	1.18%
089	26,665	11.61%	2.05%	2.73%	1.98%
091	22,677	11.59%	0.96%	0.89%	0.38%
092	11,583	13.51%	0.46%	0.50%	0.51%
093	224	1.54%	1.64%	1.76%	1.80%
094	30,868	17.03%	1.37%	1.11%	1.07%
112	156,040	19.09%	1.33%	0.97%	0.23%
124	2,895	16.27%	0.88%	2.66%	2.74%
162	23,996	25.44%	0.13%	0.43%	0.50%

Figure T2: Comparing RRMSEs of HBLMMs using Normal, T, and Mixture of Normal distributions

Figure T2a: Count for Largest RRMSE across Function Codes

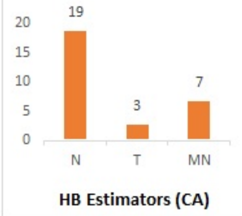


Figure T2b: Count for Smallest RRMSE across Function Codes

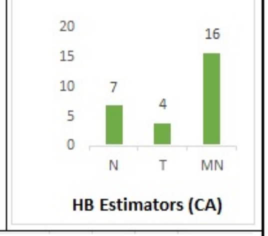
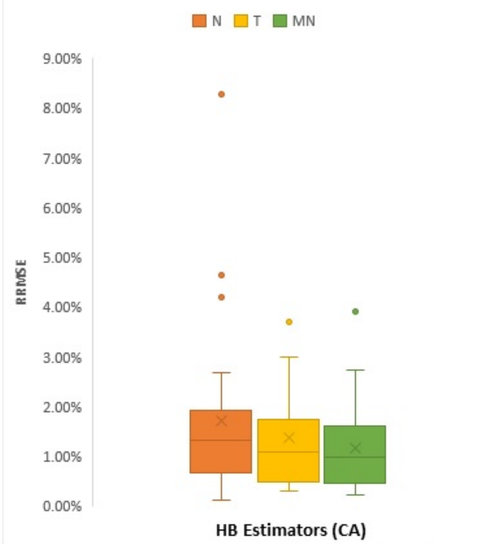


Figure T2c: Boxplots of RRMSE of HB LMMs using Normal, T, and Mixture of Normal distributions



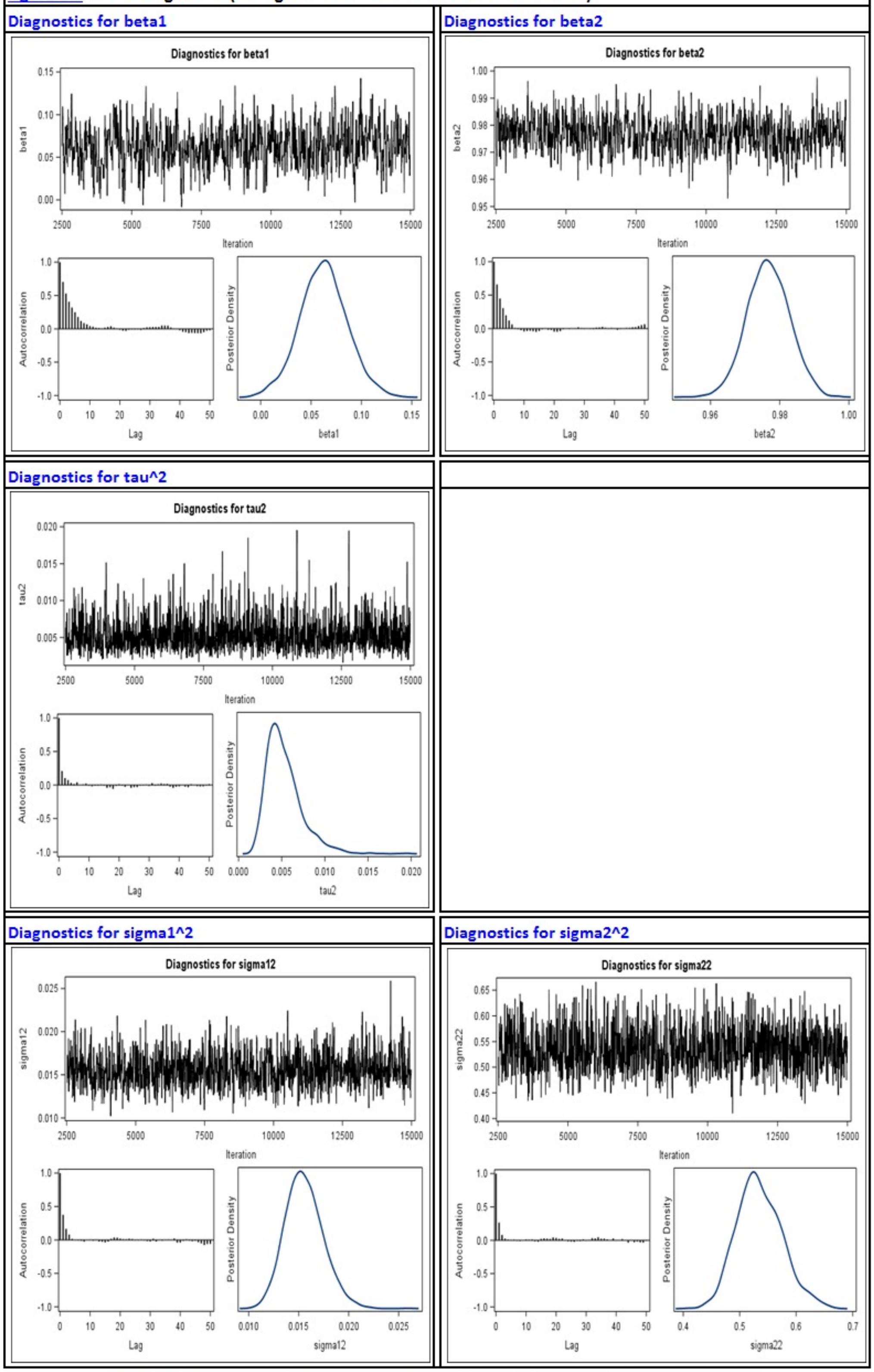
FC: Function Code (see description in the Appendix)
med(f_m): median sampling rates out of 1,000 replicates

Data Source: U.S. Census Bureau 2007 and 2012 Census of Government - Employment

Figure T2a indicates the N-model is the least efficient. Figure T2b suggests the MN-model is the most efficient. Figure T2c shows that, on the average, the MN-model provides HB estimates with smallest RRMSEs and estimates using N-model have largest RRMSEs.

3.3 Application to Georgia ASPEP data - MCMC convergence diagnostics for $\theta = (\beta_1, \beta_2, \tau^2, \sigma_1^2, \sigma_2^2)$ and comparison of RRMSEs of the estimates

Figure D3: MCMC Diagnostics (Fitting Mixture Model to GEORGIA ASPEP data)



Data Source: U.S. Census Bureau 2007 and 2012 Census of Government - Employment

Table 3: RRMSEs of HB Estimates for Various Function Codes of Government (in Georgia)

FC	Y_m	med(f_m)	$RRMSE_N$	$RRMSE_T$	$RRMSE_{MN}$
001	963	13.69%	1.34%	1.21%	1.14%
005	10,101	23.55%	2.21%	1.07%	1.23%
012	153,765	20.92%	3.20%	1.12%	0.66%
023	7,440	17.81%	1.30%	1.54%	1.91%
024	11,061	21.68%	0.11%	0.24%	0.86%
025	9,262	18.03%	0.78%	0.15%	0.22%
029	5,110	14.56%	0.65%	1.44%	0.68%
032	7,504	24.45%	3.00%	1.25%	0.96%
040	16,581	13.27%	1.78%	1.56%	1.54%
044	7,851	17.88%	1.04%	0.42%	0.14%
050	2,428	10.38%	4.63%	3.69%	3.54%
052	1,856	16.67%	0.45%	0.30%	0.17%
059	148	8.91%	3.05%	1.80%	1.83%
061	3,897	24.02%	4.27%	3.81%	2.89%
062	20,212	17.79%	0.30%	0.52%	0.81%
079	917	14.29%	1.34%	0.69%	0.45%
080	3,354	15.68%	0.90%	0.82%	0.23%
081	4,326	19.37%	2.00%	1.95%	1.79%
089	5,960	16.76%	5.59%	2.87%	3.50%
091	6,210	13.53%	0.68%	0.14%	0.22%
092	928	11.45%	1.09%	1.12%	0.73%
093	558	9.78%	3.65%	3.38%	2.78%
094	4,818	10.43%	0.19%	0.16%	0.19%
112	73,016	19.80%	3.05%	1.85%	0.80%
124	446	25.33%	1.70%	1.74%	1.71%
162	4,350	21.62%	2.83%	2.68%	2.43%

Figure T3: Comparing RRMSEs of HB LMMs using Normal, T, and Mixture of Normal distributions

Figure T3a: Count for Largest RRMSE across Function Codes

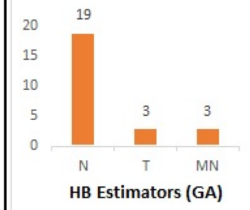


Figure T3b: Count for Smallest RRMSE across Function Codes

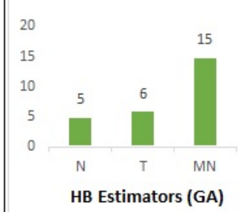
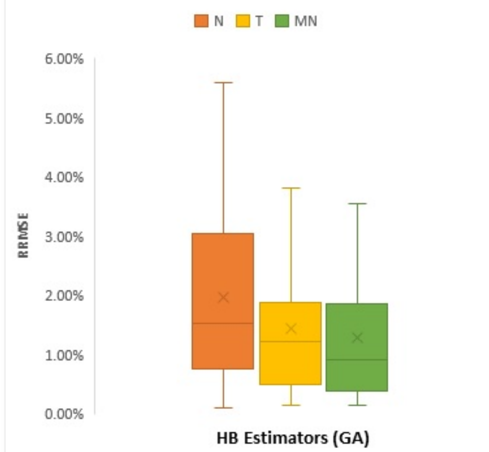


Figure T3c: Boxplots of RRMSE of HB LMMs using Normal, T, and Mixture of Normal distributions



FC: Function Code (see description in the Appendix)
med(f_m): median sampling rates out of 1,000 replicates

Data Source: U.S. Census Bureau 2007 and 2012 Census of Government - Employment

Figure T3a indicates the N-model is the least efficient. Figure T3b suggests the MN-model is the most efficient. Figure T3c shows that, on the average, the MN-model provides HB estimates with smallest RRMSEs and estimates using N-model have largest RRMSEs.

3.4 Application to Illinois ASPEP data - MCMC convergence diagnostics for $\theta = (\beta_1, \beta_2, \tau^2, \sigma_1^2, \sigma_2^2)$ and comparison of RRMSEs of the estimates

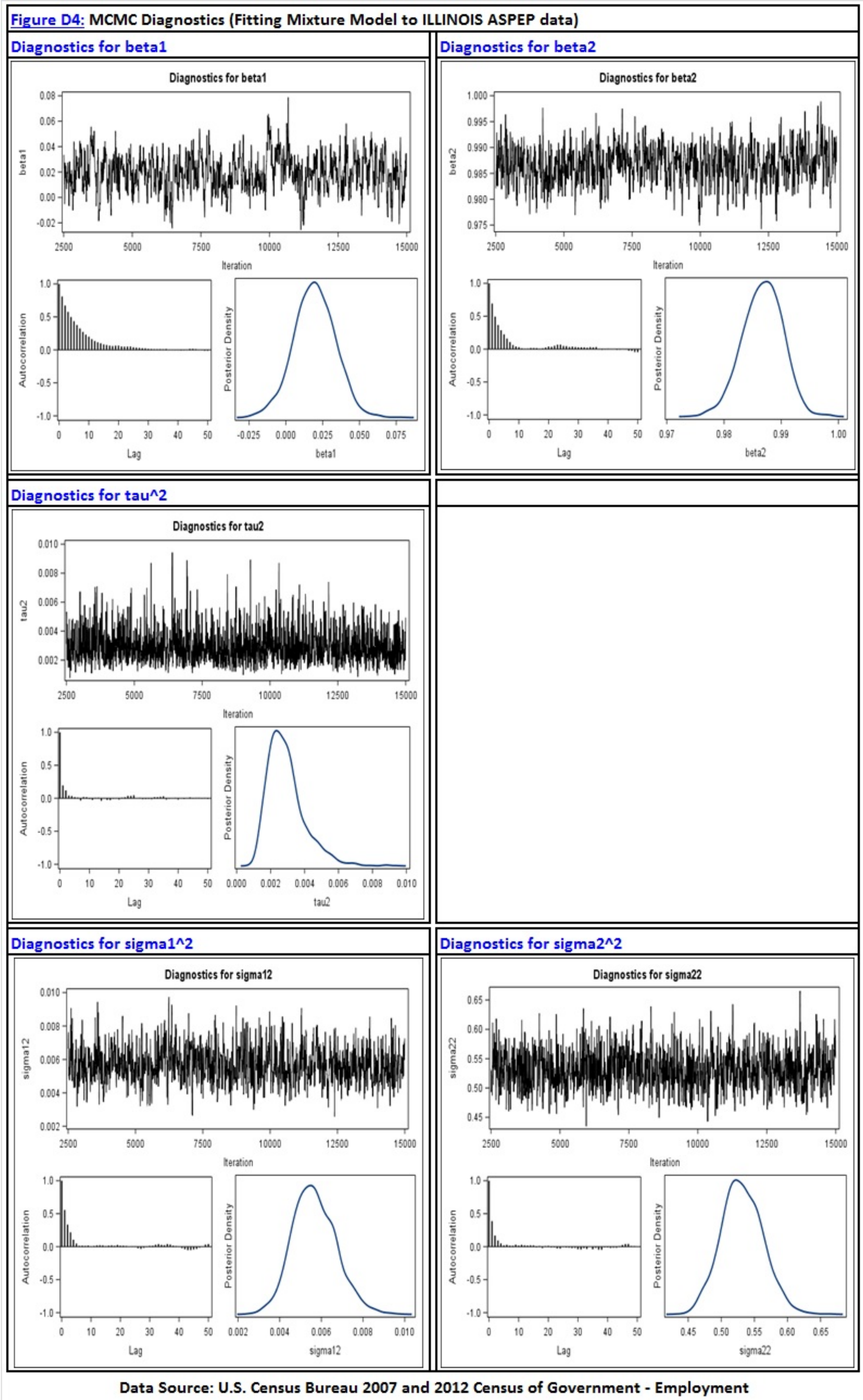


Table 4: RRMSEs of HB Estimates for Various Function Codes of Government (in Illinois)

FC	Y_m	med(f_m)	$RRMSE_N$	$RRMSE_T$	$RRMSE_{MN}$
001	898	1.49%	5.21%	2.67%	1.68%
005	10,086	4.22%	1.12%	1.27%	1.44%
012	186,958	9.13%	0.47%	1.42%	1.48%
016	9,905	4.17%	11.00%	6.49%	3.93%
018	5,455	4.18%	10.20%	6.83%	6.70%
023	6,144	9.08%	0.76%	0.35%	0.73%
024	10,243	7.64%	3.49%	2.61%	2.66%
025	12,277	8.45%	0.89%	0.84%	0.74%
029	6,166	7.25%	2.84%	2.53%	1.69%
032	4,981	7.62%	3.90%	2.49%	2.43%
040	10,247	2.12%	1.21%	0.90%	0.65%
044	8,510	8.47%	1.12%	2.12%	1.26%
050	2,963	8.09%	7.20%	2.87%	1.59%
052	4,631	6.24%	1.71%	1.21%	2.50%
059	929	2.49%	4.32%	3.30%	2.54%
061	8,537	7.65%	4.02%	2.23%	1.19%
062	34,630	13.74%	0.66%	0.13%	0.36%
079	4,239	5.51%	3.19%	1.77%	1.43%
080	4,807	9.56%	1.93%	1.75%	0.94%
081	1,355	6.48%	0.56%	0.32%	0.24%
087	57	0.11%	2.73%	2.81%	3.23%
089	5,568	9.97%	1.43%	0.88%	0.53%
091	4,003	8.91%	1.89%	0.74%	0.06%
092	1,103	1.27%	2.45%	2.60%	2.11%
093	108	0.35%	16.70%	6.21%	2.91%
094	13,753	2.51%	1.13%	1.08%	1.04%
112	67,659	8.98%	1.13%	1.38%	1.32%
124	1,061	7.44%	1.26%	0.98%	1.13%
162	6,226	15.43%	1.52%	0.71%	0.34%

Figure T4: Comparing RRMSEs of HB LMMs using Normal, T, and Mixture of Normal distributions

Figure T4a: Count for Largest RRMSE across Function Codes

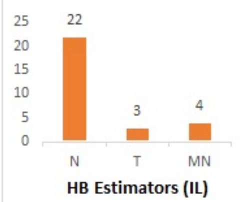


Figure T4b: Count for Smallest RRMSE across Function Codes

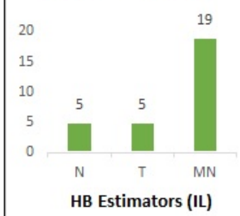
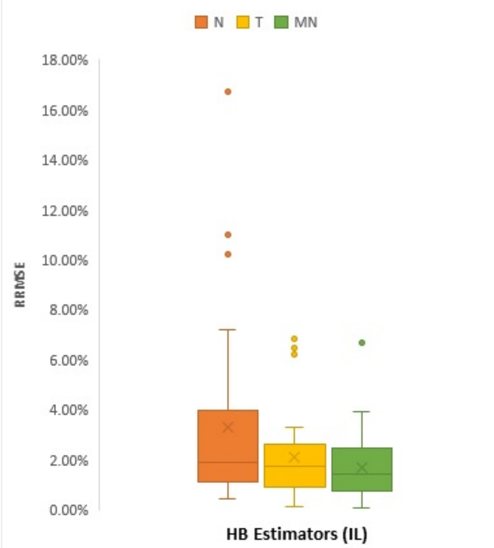


Figure T4c: Boxplots of RRMSE of HB LMMs using Normal, T, and Mixture of Normal distributions



FC: Function Code (see description in the Appendix)
med(f_m): median sampling rates out of 1,000 replicates

Data Source: U.S. Census Bureau 2007 and 2012 Census of Government - Employment

Figure T4a indicates the N-model is the least efficient. Figure T4b suggests the MN-model is the most efficient. Figure T4c shows that, on the average, the MN-model provides HB estimates with smallest RRMSEs and estimates using N-model have largest RRMSEs.

3.5 Application to Louisiana ASPEP data - MCMC convergence diagnostics for $\theta = (\beta_1, \beta_2, \tau^2, \sigma_1^2, \sigma_2^2)$ and comparison of RRMSEs of the estimates

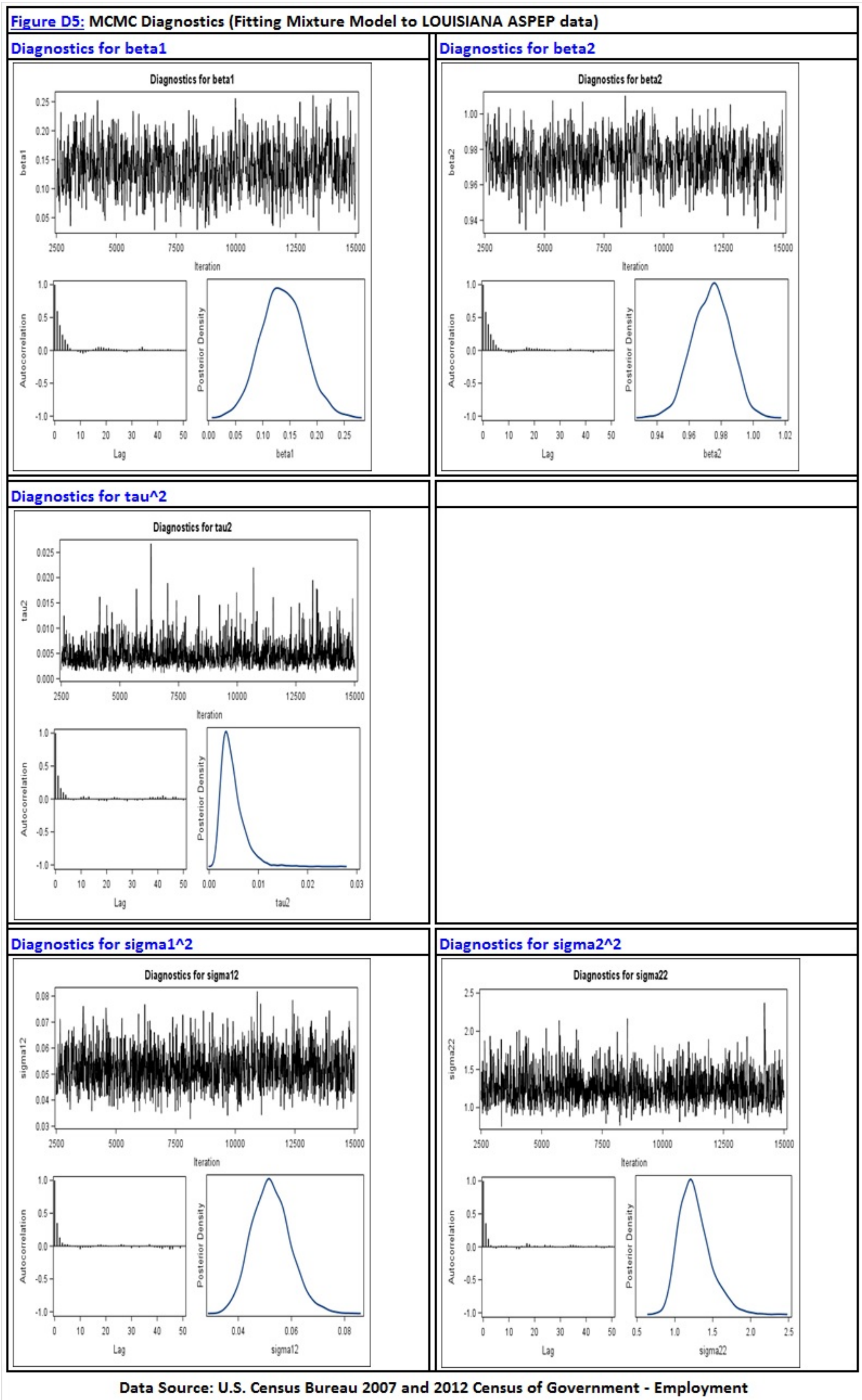


Table 5: RRMSEs of HB Estimates for Various Function Codes of Government (in Louisiana)

FC	Y_m	med(f_m)	$RRMSE_N$	$RRMSE_T$	$RRMSE_{MN}$
001	413	8.33%	0.21%	0.11%	0.10%
005	5,847	25.27%	5.51%	5.33%	5.16%
012	63,892	34.78%	5.68%	1.44%	0.26%
023	3,492	19.67%	3.10%	2.14%	2.08%
024	4,605	14.22%	3.02%	2.74%	2.66%
025	5,417	21.77%	3.07%	1.74%	1.49%
029	2,676	15.56%	5.33%	5.17%	4.91%
032	1,491	20.99%	5.90%	5.75%	5.79%
040	16,480	17.42%	0.93%	0.47%	0.50%
044	5,077	19.34%	5.61%	4.92%	4.86%
050	1,025	17.68%	0.73%	0.23%	0.11%
052	2,130	20.22%	2.01%	1.60%	1.51%
059	903	11.11%	0.47%	0.47%	0.46%
061	2,847	18.53%	0.86%	0.55%	0.57%
062	11,633	17.24%	4.55%	4.31%	4.13%
079	631	10.67%	0.13%	0.35%	0.39%
080	1,832	13.72%	0.95%	1.28%	1.41%
081	935	14.61%	0.54%	0.93%	0.93%
087	318	3.80%	0.93%	0.99%	1.14%
089	3,016	17.59%	2.41%	1.84%	1.92%
091	1,980	15.46%	4.79%	2.22%	2.43%
092	603	5.26%	5.53%	6.75%	7.04%
093	335	5.52%	2.41%	2.29%	2.46%
094	344	7.04%	1.35%	1.40%	1.40%
112	29,228	33.33%	4.13%	1.21%	0.61%
124	98	10.24%	6.86%	5.90%	5.42%
162	3,043	20.08%	1.65%	0.80%	0.69%

Figure T5: Comparing RRMSEs of HB LMMs using Normal, T, and Mixture of Normal distributions

Figure T5a: Count for Largest RRMSE across Function Codes

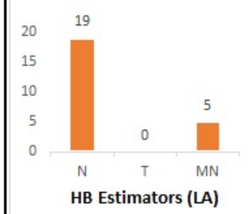


Figure T5b: Count for Smallest RRMSE across Function Codes

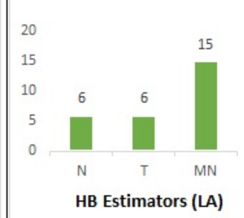
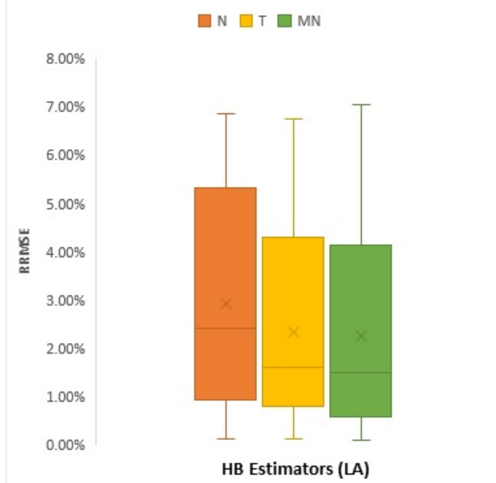


Figure T5c: Boxplots of RRMSE of HB LMMs using Normal, T, and Mixture of Normal distributions



FC: Function Code (see description in the Appendix)
med(f_m): median sampling rates out of 1,000 replicates

Data Source: U.S. Census Bureau 2007 and 2012 Census of Government - Employment

Figure T5a indicates the N-model is the least efficient. Figure T5b suggests the MN-model is the most efficient. Figure T5c shows that, on the average, the MN-model provides HB estimates with smallest RRMSEs and estimates using N-model have largest RRMSEs.

3.6 Conclusions

On the average, the MN-model provides HB estimates with smallest RRMSEs, estimates using N-model have largest RRMSEs. The N-model is the least efficient. The MN-model is the most efficient.

The performance of the estimators using the t or the mixture of normal distributions is significantly more reliable than the one using normal distribution. The mixture of normal model is better than the t-model in terms of RRMSE. Our future projects may focus on models using mixture of t-distributions or finite mixture of normal distributions with more than two mixing components.

Appendix

FC	Description
001	Air Transportation
005	Correction
012	Elementary and Secondary - Instruction
016	Higher Education - Other
018	Higher Education - Instructional
023	Financial Administration
024	Firefighters
025	Judicial & Legal
029	Other Government Administration
032	Health
040	Hospitals
044	Highways
050	Housing & Community Development
052	Libraries
059	Natural Resources
061	Parks & Recreation
062	Police Protection - Officers
079	Public Welfare
080	Sewerage
081	Solid Waste Management
087	Water Transport & Terminals
089	All Other & Unallocable
091	Water Supply
092	Electric Power
093	Gas Supply
094	Transit
112	Elementary & Secondary Schools - Other
124	Fire - Other
162	Police - Other

References

[1] 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*

[2] Bell, W. R. and Huang, T. E. (2006). Using the t-distribution to Deal with Outliers in Small Area Estimation. *Proceedings of Statistics Canada Symposium on Methodological Issues in Measuring Population Health. 2006.*

[3] Bolstad, W. (2007). Introduction to Bayesian Statistics, *John Wiley and Sons, second edition.*

[4] Bolstad, W. (2010). Understanding Computational Bayesian Statistics, *John Wiley and Sons.*

[5] De Veaux, R. and Krieger, A. (1990). Robust Estimation of Normal Mixture. *Statistics and Probability Letters 10 (1990): 1-7. North-Holland.*

[6] Dumbacher, B. and Hill, C. (2014). Description of the 2014 Sample Design for the Annual Survey of Public Employment and Payroll. *Internal U.S. Census Bureau Memorandum. February 14, 2014*

[7] Fay, R. E. and Herriot, R. A. (1979). Estimates of Income for Small Places: an Application of James-Stein Procedure to Census Data, *Journal of American Statistical Association, 74, 269- 277*

[8] Gelman, A. (2006). Prior Distributions for Variance Parameters in Hierarchical Models. *International Society for Bayesian Analysis. 2006, No 3, pp. 515-533.*

[9] Gershunskaya, J. and Lahiri, P. (2011). Robust Small Area Estimation Using a Mixture Model. *Int. Statistical Inst.: Proc. 58th World Statistical Congress, 2011, Dublin (Session IPS062)*

[10] Hall, D. and Wang, L. (2005). Two-component Mixtures of Generalized Linear Mixed-Effects Models for Cluster Correlated Data. *Statistical Modeling 5:21-37.*

[11] Lange, K.L., Little, R.J.A., and Taylor, J.M.G. (1989). Robust Statistical Modeling Using the t-distribution. *Journal of the American Statistical Association, 84(408), 881-896.*

[12] McLachlan, G. and Basford, K. (1988). Mixture Models: Inference and Applications to Clustering. *Vol. 84. Marcel Deker.*

[13] McLachlan, G. and Peel, D. (2000). Finite Mixture Models, *John Wiley and Sons.*

[14] McLachlan, G., Ng, S., and Bean, R. (2006). Robust Cluster Analysis via Mixture Models. *Austrian Journal of Statistics. Vol. 35 (2006), No 2-3, 157-174.*

[15] Ojo, O., Lougue, S., and Woldegerima, W. (2017). Bayesian Generalized Linear Mixed Modeling of Tuberculosis Using Informative Priors. *PLoS ONE, 12(3), e0172580. http://doi.org/10.1371/journal.pone.0172580*

[16] Rao, J.N.K. (2003), Small Area Estimation*, New-York, John Wiley and Sons, Inc.*

[17] Staudenmayer, J., Lake, E.E. and Wand, M.P. (2009). Robustness for general design mixed models using the t-distribution. *Statistical Modeling 9.3 (2009): 235-255.*

[18] Tadjudin, S. and Landgrebe, D. (2000). Robust Parameter Estimation for Mixture Model. *IEEE Transactions on Geoscience and Remote Sensing, Vol. 38, No. 1, pp. 439-445, January 2000.*

[19] Trinh, G. and Tran, B. (2016). Outliers in Annual Survey of Public Employment & Payroll Small-area Approach, *American Statistical Association, JSM2016 Conference Proceedings.*

[20] Wang, C., Rutledge, J., and Gianola, D. (1994). Bayesian Analysis of Mixed Linear Models via Gibbs Sampling with an Application to Litter Size in Iberian Pigs. *Genetics Selection Evolution, BioMed Central, 26, 91-115.*

[21] Woodworth, G. (2004). Biostatistics: a Bayesian Introduction. *New-Jersey, John Wiley and Sons, Inc.*