Robust Estimation in Mixture Models and Small Area Estimation using Cross-Sectional Time Series models

by

Mi-Ja Woo

(Under the direction of Tharuvai N. Sriram and Gauri Datta)

Abstract

This dissertation considers robust estimation of unknown number of components, also known as the mixture complexity, in finite mixture models and cross-sectional time series modeling of civilian unemployment rate for all the states in the U.S..

We begin with the problem of finding the mixture with fewest possible components that provides a satisfactory fit of the data. Finite mixture models provide a natural way of modeling unobserved population heterogeneity, which is often encountered in data sets arising from biological, physical and social sciences. However, in many applications, it is unrealistic to expect that the component densities belong to some exact parametric family. The mixture of interest may even be contaminated, which causes the estimates such as based on KL distances to be unstable. To overcome this problem, we develop a robust estimator of mixture complexity based on the Minimum Hellinger Distance (MHD) when all other associated parameters are unknown. This estimator is considered in two cases, that is, when the random variables are continuous and discrete. For each case, an estimator of mixture complexity of mixture complexity is constructed as a by-product of minimizing a Hellinger Information Criterion, and this estimator is proved to be consistent for parametric family of mixtures. Via extensive simulations, our estimator is shown to be very competitive with several others in the literature when the model is correctly specified and to be robust under

symmetric departures from postulated component normality in terms of correctly identifying the true mixture complexity robustness.

Next, we consider the problem of modeling civilian unemployment rate for all the states in the U.S.. Unemployment rate estimates are published by the U.S. Bureau of the Labor Statistics (BLS) every month for the whole nation, 50 states and DC as well as other areas. In recent years, the demand for small area statistics has greatly increased. At the national level, The overall sample size for the Current Population Survey (CPS) is sufficient to produce reliable estimates of UE rate. However, for smaller domains, the effective sample sizes within a given domain are so small that standard design-based estimators are not precise enough. Therefore, there is a need to improve the efficiency for small areas. The overlaps in CPS samples over time and the availability of other states' records provide the development of reliable model-based unemployment rate estimators for the states. To improve the efficiency for small areas, we turn to explicit small area models that make specific allowance for between area variation, based on a Seasonal Autoregressive Integrated Moving Average (SARIMA) model. To carry out estimation of parameters in this random-effects version of time series model, a Bayesian inference methodology is constructed using Markov chain Monte Carlo methods. Through examining the model adequacy, and forecasting the last four observations for all the states, our model is shown to be reliable and efficient.

INDEX WORDS: Finite mixtures; Hellinger Information Criterion; Threshold; Consistency; Robustness; Adaptive Density Estimate; Symmetric Departures;Seasonal Autoregressive Moving Average Model; Bayesian Analysis; Gibbs Sampling; Metropolis-Hasting sampling; Forecasting; Model Adequacy

Robust Estimation in Mixture Models and Small Area Estimation using Cross-Sectional Time Series models

by

Mi-Ja Woo

B.S., Sookmyung Women's University, Korea, 1998

M.S., Sookmyung Women's University, Korea, 2000

A Dissertation Submitted to the Graduate Faculty

of The University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2005

Robust Estimation in Mixture Models and Small Area Estimation using Cross-Sectional Time Series models

by

Mi-Ja Woo

Approved:

Major Professors:	Tharuvai N. Sriram
	Gauri Datta
Committee:	William P. McCormick
	Jaxk Reeves
	Romdhane Rekaya
	XiangRong Yin

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2005

Table of Contents

Chapter 1

ROBUST ESTIMATOR IN MIXTURE MODELS

1.1 INTRODUCTION

Finite mixture models provide a natural way of modeling unobserved population heterogeneity, which is often encountered in data sets arising from biological, physical and social sciences. Over the last two decades or so, there has been a proliferation of literature on theory and applications of mixture models. A comprehensive account of statistical inference for mixture models with applications can be found in the books by Everitt and Hand (1981), Titterington, Smith and Makov (1985), and McLachlan and Basford (1988), while more recent developments and applications of the subject are documented in Lindsay (1995), Bohning (1999) and McLachlan and Peel (2000). For a comprehensive editorial on some of the recent developments in mixture models, see Bohning and Seidel (2003).

If the number of components in a finite mixture model is known, EM algorithm of Dempster, Laird and Rubin (1977) is undoubtedly a useful way to compute maximum likelihood estimates (MLE) of all the parameters. When there is no data contamination, use of good starting values for the EM algorithm do lead to ML estimates which overcome some well known shortcomings of MLE for mixtures. However, when there is a small perturbation in one of the component densities in the underlying parametric model, even with good starting values these ML estimates become highly unstable(Aitkin and Wilson 1980).

To address the issue of instability, a variety of minimum distance estimation methods possessing some degree of automatic robustness (see Donoho and Liu 1988) have been studied as alternative approaches for mixtures. When only the mixing proportions are unknown,

some of the distance based estimation methods discussed in the literature include the Wolfowitz distance (Choi and Bulgren 1968), the Levy distance (Yakowitz 1969), the Cramer-von Mises distance (Macdonald 1971), the squared L_2 distance (Clarke 1989; Clarke and Heathcote 1994) and the Hellinger distance (Woodward, Whitney and Eslinger 1994). For the general case of estimating all the unknown parameters, the methods considered include Wolfowitz distance (Choi 1969) the Cramer-von Mises distance (Woodward, Parr, Schucany and Lindsay 1984), the squared L_2 distance (Clarke and Heathcote 1994), the Kolmogorov distance (Deeley and Kruse 1968; Blum and Susarla 1977), the Hellinger distance (Cutler and Cordero-Braňa 1996; Karlis and Xekalaki 1998), a distance using kernel density estimate (Cao, Cuevas, Fraiman 1995) and a penalized minimum-distance (Chen and Kalbfleisch 1996). Other robust estimation approaches for mixtures are discussed in De Veaux and Krieger (1990), and Windham and Cutler (1994). Robust methods such as M-estimation are not easily adapted for mixtures and these generally achieve robustness at the cost of efficiency at the parametric model density.

One way to partially reconcile the conflicting concepts of robustness and efficiency is to use a density-based minimum Hellinger distance (MHD) estimator introduced by Beran (1977). Beran showed that MHD estimators achieve efficiency at the model density and simultaneously possess desirable robustness properties under gross-error contaminations. Tamura and Boos (1986) extended Beran's work to the multivariate setup, while Stather (1981) and Simpson (1987) studied the efficiency and robustness properties of MHD estimators in the discrete case. Recently, Sriram and Vidyashankar (2000) constructed an MHD estimator of the offspring mean in a supercritical Galton-Watson process and established its asymptotic efficiency and robustness properties. In a sequential sampling context, Lee, Sriram and Wei (2003) have shown that MHD method can be used to construct robust sequential fixed width confidence intervals for parametric models.

For finite mixtures with known number of components, Cutler and Cordero-Braňa (1996) developed a minimum Hellinger distance (MHD) estimator for all parameters when the exact

form of the component densities are unknown but are thought to be close to members of some parametric family. Cutler and Cordero-Braňa (1996) proposed a new computational algorithm, somewhat similar to the EM algorithm, and an adaptive density estimate to compute the MHD estimates. In addition to studying basic properties, they showed via simulations that their MHD estimates are also robust to certain departures from the parametric family. Furthermore, Cordero-Braňa and Cutler (1996) (also see Cordero-Braňa (1994)) established the consistency and asymptotic normality of these MHD estimators.

The estimation literature for finite mixture models described above assumes that there is sufficient *apriori* information about the number of components, known as the mixture complexity. In many situations, however, the mixture complexity is also unknown. In these cases, our objective is to find the mixture with fewest possible components that provides a satisfactory fit of the data. This is a challenging problem but examples of these scenarios are plentiful and are discussed in Bogardus et al (1989), McLaren (1991), Roeder (1994), McLachlan, McLaren and Matthews (1995), McLaren (1996) Richardson and Green (1997), McLachlan and Peel (1997, 2000). Due to the scope of applications, developing methods of estimation for mixture complexity has been an area of intense research in the recent years; see Henna (1985); McLachlan (1987); Roeder (1994); Escobar and West (1995); Chen and Kalbfleisch (1996); Dacunha-Castelle and Gassiat (1997, 1999); Roeder and Wasserman (1997); Keribin (2000); Priebe and Marchette (2000); and Ishwaran, James and Sun (2001).

Recently, James, Priebe and Marchette (2001) adopted a semi-parametric approach and constructed a consistent estimator of mixture complexity when the component densities are normal. Their estimator is based on Kullback-Leibler (KL) distance and relies on comparing KL distances between a normal kernel density estimator and the best parametric fit of a given complexity convolved with a normal density. James et al (2001) showed that their estimator of mixture complexity is consistent. Through extensive Monte Carlo simulations, James et al (2001) also assessed the performance of their estimator and showed that their method compares favorably with other available methods in the literature.

In many applications, it is unrealistic to expect that the component densities belong to some exact parametric family. The mixture of interest may even be contaminated, which causes the estimates based on KL distances to be unstable. We focus on the case when the exact form of the component densities are unknown but are postulated to be close to members of some parametric family. For this case, we develop a robust estimator of mixture complexity based on the MHD approach when all other associated parameters are unknown. Method of construction of our estimator is motivated by the work of James et al (2001) but is applicable more generally.

In this chapter, we consider the case when the random variables are continuous and propose an estimator of mixture complexity using the MHD estimation approach in section 1.2. We propose to establish the consistency of the estimator under certain the regularity conditions. This result is stated as Theorem 1 in section 1.3. In section 1.4, we give the details on the computation of our estimator. In the subsections 1.5.1 and 1.5.2, we list a variety of target densities for which we propose to carry out extensive Monte Carlo studies to compare the performance of our estimator of mixture complexity with those available in the literature. In section 1.6, we list contaminated mixtures of our estimator (via simulations) through which we propose to compare them with those obtained using the KL method of James et al (2001). Computations for our estimator are carried out using the HMIX algorithm due to Cutler and Cordero-Braňa (1996) and its details are given in section 1.4. In section 1.7, we propose to estimate the mixture complexity for an example concerning hypertension considered in Roeder (1994).

1.2 MHD ESTIMATION OF MIXTURE COMPLEXITY: CONTINUOUS CASE

Consider a parametric family of density functions $\mathcal{F}_m = \{f_{\boldsymbol{\theta}_m} : \boldsymbol{\theta}_m \in \Theta_m \subseteq R^p\}$ for each fixed $m < \infty$, such that $f_{\boldsymbol{\theta}_m}$ can be represented as a finite mixture of the form

$$f_{\boldsymbol{\theta}_m}(x) = \sum_{i=1}^{m} \pi_i f(x|\boldsymbol{\phi}_i), \quad x \in \mathcal{X} \subseteq \mathcal{R}, \tag{1.2.1}$$

where the component densities $f(x|\boldsymbol{\phi}_i) \geq 0$, $\int f(x|\boldsymbol{\phi}_i)dx = 1$, $\boldsymbol{\phi}_i \in \Phi \subseteq R^s$, the mixing proportions $\pi_i \geq 0$, $\sum_{i=1}^{m} \pi_i = 1$ for $i = 1, \ldots, m$ and $\boldsymbol{\theta}_m = (\pi_1, \ldots, \pi_{m-1}, \boldsymbol{\phi}_1^T, \ldots, \boldsymbol{\phi}_m^T)^T$. The class $\mathcal{F}_m \subseteq \mathcal{F}_{m+1}$ for all m and we denote $\mathcal{F} = \bigcup_{m=1}^{\infty} \mathcal{F}_m$.

For each fixed $m \geqslant 1$, a mixture is said to be *economically represented* if all components with nonzero mixing proportions are distinct (Redner and Walker 1984). The class $\mathcal{F}_m$ is identifiable if whenever two economically represented mixtures are identical, the two collections of components having nonzero probability are identical. Incidentally, identifiability of the model family does not imply identifiability of $\boldsymbol{\theta}_m$, because $f_{\boldsymbol{\theta}_m}(x)$ is invariant under permutations of the component labels. Finally, if a mixture with m components can be represented as a mixture with fewer than m components then it is said to be *degenerate* and can be represented in infinitely many ways.

Let $X_1, \ldots, X_n$ be independent and identically distributed random variables with an unknown density function g_0. For an arbitrary density g, define the *index of the economical representation of* g, relative to the family of mixtures defined above, as

$$m(g) = \min\{m : g \in \mathcal{F}_m\}.$$

If indeed g is a finite mixture then $m(g)$ is finite and denotes the true mixture complexity; otherwise $m(g) = \infty$. Note that $m(g)$ represents the most parsimonious mixture model representation for g.

We now describe a robust estimation procedure to estimate $m_0 = m(g_0)$. To this end, we follow the approach of Beran (1977) and define the Hellinger distance between two densities

f and g by

$$H^2(f,g) = ||f^{1/2} - g^{1/2}||_2^2 \tag{1.2.2}$$

where $||\cdot||_2$ is the L_2 norm. Let $\hat{g}_n$ be a kernel density estimator of g_0 of the form

$$\hat{g}_n(x) = \frac{1}{nc_n}\sum_{i=1}^{n} K(\frac{x - X_i}{c_n}) \tag{1.2.3}$$

where K is a density on $\Omega \subseteq R$ and the bandwidth $c_n = c_n(X_1, \ldots, X_n)$ satisfies certain regularity conditions. For each integer $m > 0$, define

$$\hat{g}^m = \arg \min_{f \in \mathcal{F}_m} H(\hat{g}_n, f)$$

and

$$g_0^m = \arg \min_{f \in \mathcal{F}_m} H(g_0, f) \tag{1.2.4}$$

where g_0 is the unknown underlying density. When $m > 0$ is known, the MHD estimator $\hat{\boldsymbol{\theta}}_{n,m}^{MHD}$ of $\boldsymbol{\theta}_m$ is defined as the value of a functional $T_m(g)$ at $\hat{g}_n$, where for any g, $T_m(g)$ is defined by

$$T_m(g) = \{\boldsymbol{\theta}_m \in \Theta_m : H(f_{\boldsymbol{\theta}_m}, g) = \min_{\boldsymbol{t}_m \in \Theta_m} H(f_{\boldsymbol{t}_m}, g)\}. \tag{1.2.5}$$

Here $T_m(g)$ is the set of solutions and a member of $T_m(g)$ is chosen arbitrarily when a solution is required. $T_m(g)$ is said to be essentially unique if $f_{\boldsymbol{\theta}_m}$ is nondegenerate for any $\boldsymbol{\theta}_m \in T_m(g)$, and any other element of $T_m(g)$ can be obtained from $\boldsymbol{\theta}_m \in T_m(g)$ by permuting the labels of the components. Note from (1.2.4) that $\hat{g}^m = f_{\hat{\boldsymbol{\theta}}_{n,m}^{MHD}}$ and $g_0^m = f_{T_m(g_0)}$.

Note that we can express m_0 as

$$\begin{aligned} m_0 &= \min\{m : H(g_0, g_0^m) - H(g_0, g_0^{m+1}) \leq 0\} \\ &= \min\{m : H(g_0, g_0^m) = 0\}, \end{aligned} \tag{1.2.6}$$

because $\mathcal{F}_m \subseteq \mathcal{F}_{m+1}$. Since the family of mixtures is nested, estimation of unknown number of components can be considered as a model selection problem, that is, selecting the model that fits a given dataset the best in some sense out of a candidate set of models. Poland and Shachter (1994) compare three approaches to model selection. Motivated by the classical

Akaike type of criterion for model selection and third approach of Poland and Shachter (1994) for model selection involving the Kellback-Leibler distance, a model selection criterion based on the Hellinger distance may be considered as the form

$$HIC = H^2(\hat{g}_n, \hat{g}^m) + n^{-1}b(n)\nu(m), \tag{1.2.7}$$

where $b(n)$ depends only on n and $\nu(m)$ is the number of parameters in the mixture model. Here, the value of m yielding the minimum HIC specifies the best model. In the context of minimum Hellinger distance estimation, the statistics $H^2(\hat{g}_n, \hat{g}^m)$ at (1.2.7) can be viewed as measuring goodness-of-fit of mixture models, and $n^{-1}b(n)\nu(m)$ as penalizing the goodness-of-statistics by a term proportional to the number of parameters in the mixture model. A simple heuristic to find the best model from a sequence of nested models is to try successive models, starting with the smallest, and stop with model when the HIC value for model m is less than that for model (m+1), that is,

$$H^2(\hat{g}_n, \hat{g}^m) + n^{-1}b(n)\nu(m) \leq H^2(\hat{g}_n, \hat{g}^{m+1}) + n^{-1}b(n)\nu(m+1),$$

or, equivalently,

$$H^2(\hat{g}_n, \hat{g}^m) - H^2(\hat{g}_n, \hat{g}^{m+1}) \leq n^{-1}b(n)[\nu(m+1) - \nu(m)]. \tag{1.2.8}$$

Hence, setting $\alpha_{n,m} = n^{-1}b(n)[\nu(m+1) - \nu(m)]$ in (1.2.8), an estimator of m_0 can be defined as

$$\hat{m}_n = \min\{m : H(\hat{g}_n, \hat{g}^m) \leq H(\hat{g}_n, \hat{g}^{m+1}) + \alpha_{n,m}\} \tag{1.2.9}$$

where $\hat{g}_n$ is the density estimator in (1.2.3) and $\{\alpha_{n,j}; j \geq 1\}$ are positive sequences of threshold values chosen in such a way they converge to zero as $n \to \infty$. We define $\hat{m}_n = \infty$ if the the minimum m in (1.2.9) does not exist for any n. Also, note that the estimator in (1.2.9), motivated by the HIC model selection criterion, is essentially a sample version of the representation in (1.2.6). Incidentally, a model selection criterion based on a Kullback-Leibler

goodness-of-fit statistics can be defined, and it motivates the estimator in James, Priebe and Marchette (2001). The KL distance between two densities g and f is defined by

$$KL(g,f) = \int g(x) ln\left(\frac{g(x)}{f(x)}\right) \mathrm{d}x.$$

In the next section, we establish the consistency of our estimator of mixture complexity.

1.3 CONSISTENCY OF $\hat{m}_n$

In this section, we establish the consistency of MHD estimator of mixture complexity defined in (1.2.9) as a theorem. The proof of the theorem is shown in the Appendix using four lemmas.

Throughout we will assume that, for each m, Θ_m can be embedded in a compact subset of R^p, the class $\mathcal{F}_m$ is identifiable for $\boldsymbol{\theta}_m \in \Theta_m$ and, for almost every x, $f(x|\boldsymbol{\phi})$ is continuous in $\boldsymbol{\phi}$. Under these conditions, Cutler and Cordero-Braňa (1996) established the existence, Fisher consistency, and continuity of the functional $T_m(g)$ with respect to the Hausdorff metric (Pollard 1981). Cordero-Braňa and Culter (1997) have shown that if $T_m(g_0)$ is essentially unique then $\hat{\theta}^{MHD}_{n,m} = T_m(\hat{g}_n)$ is consistent for $T_m(g_0)$ where $\hat{g}_n$ is as defined in (1.2.3) with c_n satisfying the condition $c_n + (nc_n)^{-1} \to 0$ almost surely (a.s.). Under more regularity conditions, they have also established the asymptotic normality of the MHD estimator when $g_0 = f_{\boldsymbol{\theta}_m}$. See Cordero-Braňa and Culter (1997) for details.

Theorem 1. Suppose $X_1, \ldots, X_n$ are independent and identically distributed random variables with a density function g_0. Suppose the bandwidth c_n in (1.2.3) satisfies $c_n + (nc_n)^{-1} \to 0$ a.s. as $n \to \infty$. If g_0 is a finite mixture with mixture complexity $m_0 \leq \infty$, then for any sequence $\alpha_{n,m} \to 0$, the estimator $\hat{m}_n$ defined in (1.2.9) is strongly consistent, i.e., as $n \to \infty$

$$\hat{m}_n \to m_0 \quad \text{a.s..} \tag{1.3.10}$$

1.4 COMPUTATIONAL DETAILS

Given a data set, computation of an estimate of mixture complexity using (1.2.9) is clearly an iterative procedure. The procedure starts by assuming that the data comes from a mixture with a single component ($m = 1$) whose form is known except for the parameter values. After fitting a nonparametric density estimator to the data, the MHD estimate of the parameter $\boldsymbol{\theta}_1$ is computed, which yields the best parametric fit $\hat{g}^1$. The Hellinger distance between the nonparametric density estimator and $\hat{g}^1$ is then computed. Next, another component density is added yielding a mixture of two components ($m = 2$), the best parametric fit $\hat{g}^2$ is computed using the MHD estimate of $\boldsymbol{\theta}_2$, and the Hellinger distance between the best parametric fit $\hat{g}^2$ and the density estimator is computed. The difference between the two Hellinger distances is compared with the threshold value $\alpha_{n,1}$. The above procedure of adding one more component to the previous mixture is repeated until the first value $m = k$ for which the difference between Hellinger distances computed at k and $k + 1$ as in (2.7) falls below the corresponding threshold value $\alpha_{n,k}$. At this time, the procedure terminates declaring k as an estimate of the number of components in the mixture.

There are several important computational details to consider in the course of implementing the above iterative algorithm, the first of which concerns the precise nature of the nonparametric density estimator. When all the mixture parameters are unknown, Cutler and Cordero-Braňa (1996) rightly point out that it is necessary to use some form of adaptive density estimate in order to avoid severe bias problems with the scale estimates. The bias occurs because components with small variance are smoothed too much and those with large variances are smoothed too little. In our computations, we propose to use the following adaptive density estimate proposed by Cutler and Cordero-Braňa (1992) which is a slight modification of the one due to Scott (1992):

$$\hat{g}_{n,m}(x) = n^{-1} \sum_{j=1}^{n} \sum_{i=1}^{m} [a_i(X_j)/c_{n,i}] K[(x - X_j)/c_{n,i}] \tag{1.4.11}$$

where

$$a_i(X_j) = \pi_i f(X_j|\boldsymbol{\phi}_i) / \sum_{l=1}^{m} \pi_l f(X_j|\boldsymbol{\phi}_l).$$

Note that $a_i(X_j)$ in (1.4.11) depends on unknown parameters and hence must be estimated using the current parameter estimates. Furthermore, as in Cutler and Cordero-Braňa (1996) we also use the Epanechnikov kernel (Scott 1992, p.140), and the value of $c_{n,i}$ is computed using the one-component empirically optimal formula given by Eslinger and Woodward (1991); namely, $c_{n,i} = 2.283 n^{-0.287} \hat{\sigma}_i$. The choice of bandwidth in our numerical studies is motivated by the simulation results in Table of Cutler and Cordero-Braňa (1996) where it is shown that the adaptive density estimate is considerabley less biased than the nonadaptive one. It can be verified that $\hat{g}_{n,m}$ in (1.4.11) is a density. Moreover, it is also possible to establish the almost sure convergence of the adaptive density estimator defined in (1.4.11) based on some regularity conditions on $\hat{\boldsymbol{\theta}}_m$.

Second issue is computation of MHD estimator which requires minimizing the Hellinger distance $||\tilde{g}_n^{1/2} - f_{\boldsymbol{\theta}_m}^{1/2}||_2$ with respect to $\boldsymbol{\theta}_m$ for a fixed m, subject to the constraint $\pi_i > 0$, $i = 1, \ldots, m$ and $\Sigma \pi_i = 1$, where $\tilde{g}_n$ is a nonparametric density estimate based on the data. We use the density estimator denfined in (1.4.11) and the HMIX algorithm proposed in section 4.1 of Cutler and Cordero-Braňa (1996). The HMIX algorithm is similar to the EM algorithm and it naturally leads to the use of the adaptive density estimate defined in (1.4.11).

Third issue concerns updating adaptive kernel density estimator in (1.4.11) at each iteration of the algorithm, which is somewhat similar to those in section 5 of James et al (2001). We exploit the nature of adaptive kernel density estimator in (1.4.11) by updating it at each step of the algorithm using the current MHD estimate, which in turn is used to obtain an updated MHD estimate at each step. For example, at stage $m = k$, compute the MHD estimate $\hat{\boldsymbol{\theta}}_{n,k} (= \arg\min H(\tilde{g}_{n,k-1}, f_{\boldsymbol{\theta}_k}))$, and use it to obtain a new bandwidth $c_{n,i}$ and $a_i(X_j)$ which in turn yields a updated adaptive density estimator $\tilde{g}_{n,k}$. Here, $\tilde{g}_{n,k-1}$ is the updated adaptive density estimator from the $(k-1)th$ step where $\tilde{g}_{n,0} = \hat{g}_{n,1}$. Now obtain a modified

best k component mixture

$$\tilde{g}^k = \arg\min_{f \in \mathcal{F}_k} H(\tilde{g}_{n,k}, f)$$

and calculate the Hellinger distance $H(\tilde{g}_{n,k}, \tilde{g}^k)$. Then, add a component and obtain $\hat{g}^{k+1} = \arg\min_{f \in \mathcal{F}_{k+1}} H(\tilde{g}_{n,k}, f)$. Now, compare the difference $H(\tilde{g}_{n,k}, \tilde{g}^k) - H(\tilde{g}_{n,k}, \hat{g}^{k+1})$ against the threshold value as in (1.2.9). This modification does not affect the theory but significantly improves the performance of the algorithm in simulation. More explicitly, the algorithm for finding $\hat{m}_n$ can be described as follows:

1. Step 1: Start with $m = 1$.

   - Compute $\hat{\theta}_{n,m}$, update $g_{n,\tilde{m-1}}$, and call it $g_{n,\tilde{m}}$.
   - Compute $\tilde{\theta}_{n,m}$ using $\tilde{g}_{n,m}$ and get $\tilde{g}^m = \arg\min_{f \in \mathcal{F}_m} H(\tilde{g}_{n,m}, f)$.
   - Calculate $HD^2(\tilde{g}_{n,m}, \tilde{g}^m)$.

2. Step 2: Add a component to m.

   - Compute $\hat{\theta}_{m+1}$ using $\tilde{g}_{n,m}$ and obtain $\hat{g}^{m+1} = \arg\min_{f \in \mathcal{F}_{m+1}} H(\tilde{g}_{n,m}, f)$.
   - Calculate $HD^2(\tilde{g}_{n,m}, \hat{g}^{m+1})$.
   - Compute the difference between $HD^2(\tilde{g}_{n,m}, \hat{g}^{m+1})$ and $HD^2(\tilde{g}_{n,m}, \tilde{g}^m)$ and compare it with a threshold value, $\alpha_{n,m}$.
   - If $HD^2(\tilde{g}_{n,m}, \tilde{g}^m) - HD^2(\tilde{g}_{n,m}, \hat{g}^{m+1}) \leq \alpha_{n,m}$ then stop, and let $\hat{m}_n = m$.
   - Otherwise, go to step 1 and repeat until the condition in (1.2.9) is satisfied.
   - Update $\hat{g}_{n,m+1}$.

Last one is the choice of threshold values $\alpha_{n,m}$ in (1.2.9) which is critical to the estimation of mixture complexity. It can be seen easily from (1.2.9) that threshold values have a direct effect on the $\hat{m}_n$ values, which increase as $\alpha_{n,m}$ values decrease. In our numerical studies, we assume normal mixture models with m univariate components which yields $\nu(m) = 3m - 1$ because each component has associated mixing proportion, mean and variance, and the

mixing proportions are constrained to sum to one. This, in view of (1.2.8), leads to $\alpha_{n,m} = 3b(n)/n$. Following Akaike Information Criterion, we set $b(n) = 1$ and obtain a threshold value of $\alpha_{n,m} = 3/n$, which is used in our numerical studies. Also, based on the Schwartz Bayesian Criterion, the set of $b(n) = ln(n)/2$ leads to $\alpha_{n,m} = (3/2)ln(n)/n$. Therefore, our choice of $\alpha_{n,m} = 3/n$ can be viewed as one based on an AIC criterion. Note that James et al (2001) choose $\alpha_{n,m} = 3/n$ in their algorithm based on the minimum description length (MDL) penalty of Rissanen (1978).

1.5 MONTE CARLO SIMULATIONS

In this section, we conduct a variety of simulations to assess the performance of our estimator of mixture complexity defined in (1.2.9) for moderate to large sample sizes. We carry out the theses studies in two different scenarios but in both instances the postulated model is a member of mixture family $\mathcal{F}_m$ where the component densities are normal. The first instance would examine the efficiency of our estimator when the model is correctly specified while the second would assess the robustness of our estimator against model misspecification.

For the first scenario, we perform the two simulation experiments discussed in James et al (2001)and compare our findings with five other algorithms for mixture complexity estimation available in the literature. The first is a Monte Carlo simulation demonstrating the performance on a target density, which is a three-component mixture of normal densities, over a variety of sample sizes. The second is a Monte Carlo simulation on target mixtures 2-10 from Marron and Wand (1992) for a fixed sample size.

For the second scenario, we perform four different simulation experiments to assess the robustness of our estimator under symmetric departures from postulated component normality. In these simulations, the samples are drawn from mixtures with two components where the component densities are those of scale and location transformations, respectively, of a Student's t random variable with two or four degrees of freedom, or a rescaled t random variable with three or four degrees of freedom. In addition, we consider varying degrees of

separation (or equivalently, overlap) between the two component densities. The setup for our robustness analysis is similar to those described in Woodward et al. (1984) and Markatou (2001); also see Woodward et al. (1995) and McCann and Sarkar (2000). In each of these simulations, robustness of our estimator of mixture complexity to model misspecification is also compared to with the estimator of mixture complexity defined in James et al. (2001).

1.5.1 THREE-COMPONENT MIXTURE

The first simulation demonstrates the performance of (1.2.9) for the target density given by

$$f(x) = (1/2)\phi(x|(0, 10)) + (1/4)\phi(x|(-0.3, 0.05)) + (1/4)\phi(x|(0.3, 0.05)), \qquad (1.5.12)$$

where ϕ denotes the normal density with respective mean and variance identified inside the parentheses. The first component has a large variance and the other two have small variances. We implement the computational algorithm described above for sample sizes $n = 50, 250, 500$ and 1000 drawn from (1.5.12). For each sample size, we perform 100 Monte Carlo replications of the algorithm, each yielding an estimate of mixture complexity. We then tallied the estimated number of components (out of 100 replications). These counts are reported for each sample size in Table 1.1 below where MHDE corresponds to the algorithm in (1.2.9). In addition, for comparison purposes, we also provide similar counts obtained via the NKE and MKE algorithm of James et al (2001, see Table 1); Bayesian algorithm of Roeder and Wasserman (1997) denoted by $R\&W$; Bootstrap algorithm of McLachlan (1987) denoted by *Bootstrap*; and the CDF method of Henna (1985) denoted by *Henna*. In this case the true mixture complexity is 3 and we denote only the highest percentage of correct identifications by an asterisk in (1.5.12).

The simulation results in Table 1.1 show that, for $n = 50$, only the $R\&W$ algorithm correctly identifies a large percentage of times, while all the other algorithms underestimate the true mixture complexity. For $n = 50$, it should be noted that the NKE, the MKE, the

Bootstrap and the Henna algorithms perform better than our MHDE algorithm, although all of them underestimate. For $n = 250$, the $R\&W$ and our MHDE algorithms correctly identify a larger percentage of times than all other procedures, with $R\&W$ performing better than our MHDE. For $n = 500$, our MHDE algorithm continues to correctly identify a high percentage of times and performs significantly better than the MKE and all the other algorithms. For $n = 1000$, the MHDE and MKE algorithms correctly identify the mixture complexity substantially higher percentage of times than all other algorithms, with MHDE performing better than the MKE. It is interesting to note that our MHDE algorithm seldom overestimates the mixture complexity, while all the others overestimate and this becomes rather severe in some cases as sample size increases. This is especially true for the $R\&W$ algorithm for all sample sizes. Finally, for $n = 1000$, the MHDE algorithm underestimates 26% of times, while the MKE underestimates 18% of times and overestimates 19% of the times. The Bootstrap and Henna algorithms incorrectly yield $\hat{m}_n = 2$ for all n. Overall, when the model is correctly specified, out MHDE algorithm is very competitive with all the other algorithms available in the literature.

1.5.2 MARRON AND WAND MIXTURES

Secondly, we propose to investigate the performance of our estimator of mixture complexity when the samples are drown from normal mixtures given if Table 1.2 below. These mixtures are considered in Marron and Wand(1992) and they exhibit a range of unimodal, skewed and multimodal densities appropriate for testing the performance of the above algorithms. The densities in Table 1.2 are graphed in as seen in Figure 2.1. As in James et al. (2001), we compare the performance of all the algorithms mentioned in Table 1.1 above based on percentage correct identification of the true mixture complexity. The sample size for this study is $n = 1000$. The true mixture complexity in each case is denoted by an asterisk in Table 1.3.

Table 1.1: Mixture Complexity Estimation results [Target mixture, display (1.5.12), has three components]

	Estimated number of components							
	1	2	3	4	5	6	7	8
$n = 50$								
MHDE	80	20						
NKE	44	56						
MKE	44	53	3					
R&W	22	7	59*	10	1	1		
Bootstrap	0	96	4					
Henna	25	68	6	1				
$n = 250$								
MHDE	16	39	45*					
NKE	0	99	1					
MKE	0	87	11	1	1			
R&W	0	0	60*	22	18			
Bootstrap	0	83	16	1				
Henna	0	90	10					
$n = 500$								
MHDE	0	35	65*					
NKE	0	97	3					
MKE	0	58	34	6	2			
R&W	0	0	22	12	61	5		
Bootstrap	0	74	20	6				
Henna	0	85	15					
$n = 1000$								
MHDE	0	26	74*					
NKE	0	86	14					
MKE	0	18	63*	10	2	3	1	3
R&W	0	0	0	1	89	10		
Bootstrap	0	79	15	4	2			
Henna	0	78	15	5	1	0	1	

When the true $m = 2$, as in mixtures 4 - 8, all the algorithms perform very well, except that R & W overestimates considerably in the case of mixture 4. In the case of mixture 2($m = 3$), the MHDE algorithm performs somewhat better than all other procedures although it does not correctly identify the true mixture complexity. In the case of mixture 9($m = 3$), the MKE and MHDE algorithms perform well but the *Bootstrap* algorithm performs the best. In the case of mixture 3($m = 8$), all the algorithms severely underestimate the mixture complexity. In the case of mixture 10($m = 6$), only MKE and $Henna$ identify correctly a higher proportion of times. As pointed out in James rt al. (2001, section 5.2), the *Bootstrap* does relatively well in many of these cases but is computationally quite intensive. These once again show that, when the model is correctly specified, the MHDE algorithm provides a useful way to estimate the mixture complexity for a variety of mixtures.

Table 1.2: The densities in Marron and Wand (1992)

Density	$f_\theta(x)$
1.Gauss	$N(0,1)$
2.Skewed unimodal	$\frac{1}{5}N(0,1) + \frac{1}{5}N(\frac{1}{2},(\frac{2}{3})^2) + \frac{3}{5}N(\frac{13}{15},(\frac{5}{9})^2)$
3. Strongly skewed	$\sum_{i=0}^{7}\frac{1}{8}N(3\{(\frac{2}{3})^i - 1\},(\frac{2}{3})^{2i})$
4.Kurtotic unimodal	$\frac{2}{3}N(0,1) + \frac{1}{3}N(0,(\frac{1}{10})^2)$
5.Outlier	$\frac{1}{10}N(0,1) + \frac{9}{10}N(0,(\frac{1}{10})^2)$
6.Bimodal	$\frac{1}{2}N(-1,(\frac{2}{3})^2) + \frac{1}{2}N(1,(\frac{2}{3})^2)$
7.Separateed bimodal	$\frac{1}{2}N(-\frac{3}{2},(\frac{1}{2})^2) + \frac{1}{2}N(\frac{3}{2},(\frac{1}{2})^2)$
8. Skewed bimodal	$\frac{3}{4}N(0,1) + \frac{1}{4}N(\frac{3}{2},(\frac{1}{3})^2)$
9.Trimodal	$\frac{9}{20}N(-\frac{6}{5},(\frac{3}{5})^2) + \frac{9}{20}N(\frac{6}{5},(\frac{3}{5})^2) + \frac{1}{10}N(0,(\frac{1}{4})^2)$
10.Claw	$\frac{1}{2}N(0,1) + \sum_{i=0}^{4}\frac{1}{10}N(i/2-1,(\frac{1}{10})^2)$

1.6 ROBUSTNESS

In this section we demonstrate the performance of the robustness of $\hat{m}_n$. To assess the robustness, we describe how much our $\hat{m}_n$ correctly identifies the true mixture complexity when the postulated mixture model is misspecified. Usually, the robustness of MHD estimators are examined by using $100\alpha\%$ gross-error contaminated mixture models and α-influenced functions defined in terms of Hellinger functionals (Beran 1997). To study robustness of MHD

Table 1.3: Mixture Complexity Estimation results for the Marron and Wand densities, 2-10

	Estimated number of components									
	1	2	3	4	5	6	7	8	9	10
Mixture2										
MHDE	0	78	22*							
NKE	0	99	1*							
MKE	0	99	1*							
R&W	3	96	1*							
Bootstrap	0	89	11*							
Henna	0	100	*							
Mixture3										
MHDE	0	13	12	49	1	24	1	*		
NKE	0	0	96	4				*		
MKE	0	1	54	37	8			*		
R&W	0	0	0	8	38	25	20	7*	2	
Bootstrap	0	0	0	17	59	21	2	1*		
Henna	0	0	26	74				*		
Mixture4										
MHDE	0	100*								
NKE	0	99*	1							
MKE	0	91*	6	3						
R&W	0	0*	0	0	75	18	5	2		
Bootstrap	0	95*	5							
Henna	0	88*	12							
Mixture5										
MHDE	0	100*								
NKE	0	96*	4							
MKE	0	91*	8	1						
R&W	0	55*	45							
Bootstrap	0	95*	5							
Henna	1	97*	1	0	0	0	0	0	1	

Table 1.3 (continued)

	Estimated number of components									
	1	2	3	4	5	6	7	8	9	10
Mixture6										
MHDE	0	100*								
NKE	0	100*								
MKE	0	98*	2							
R&W	0	100*								
Bootstrap	0	95*	5							
Henna	0	97*	3							
Mixture7										
MHDE	0	100*								
NKE	0	100*								
MKE	0	96*	4							
R&W	0	100*								
Bootstrap	0	93*	6	1						
Henna	0	96*	4							
Mixture8										
MHDE	2	97*	1							
NKE	0	100*								
MKE	0	97*	3							
R&W	0	80*	20							
Bootstrap	0	93*	7							
Henna	0	99*	1							
Mixture9										
MHDE	0	49	51*							
NKE	0	94	6*							
MKE	0	38	59*	2						
R&W	0	91	9*							
Bootstrap	0	13	75*	12						
Henna	0	82	18*							
Mixture10										
MHDE3	84	7	7	1	1	*				
NKE	33	51	15	1		*				
MKE	33	13	3	6	1	42*	2			
R&W	15	0	0	0	0	0*	39	28	17	1
Bootstrap	5	28	15	21	11	11*	5	4		
Henna	0	0	5	8	15	33*	14	9	10	6

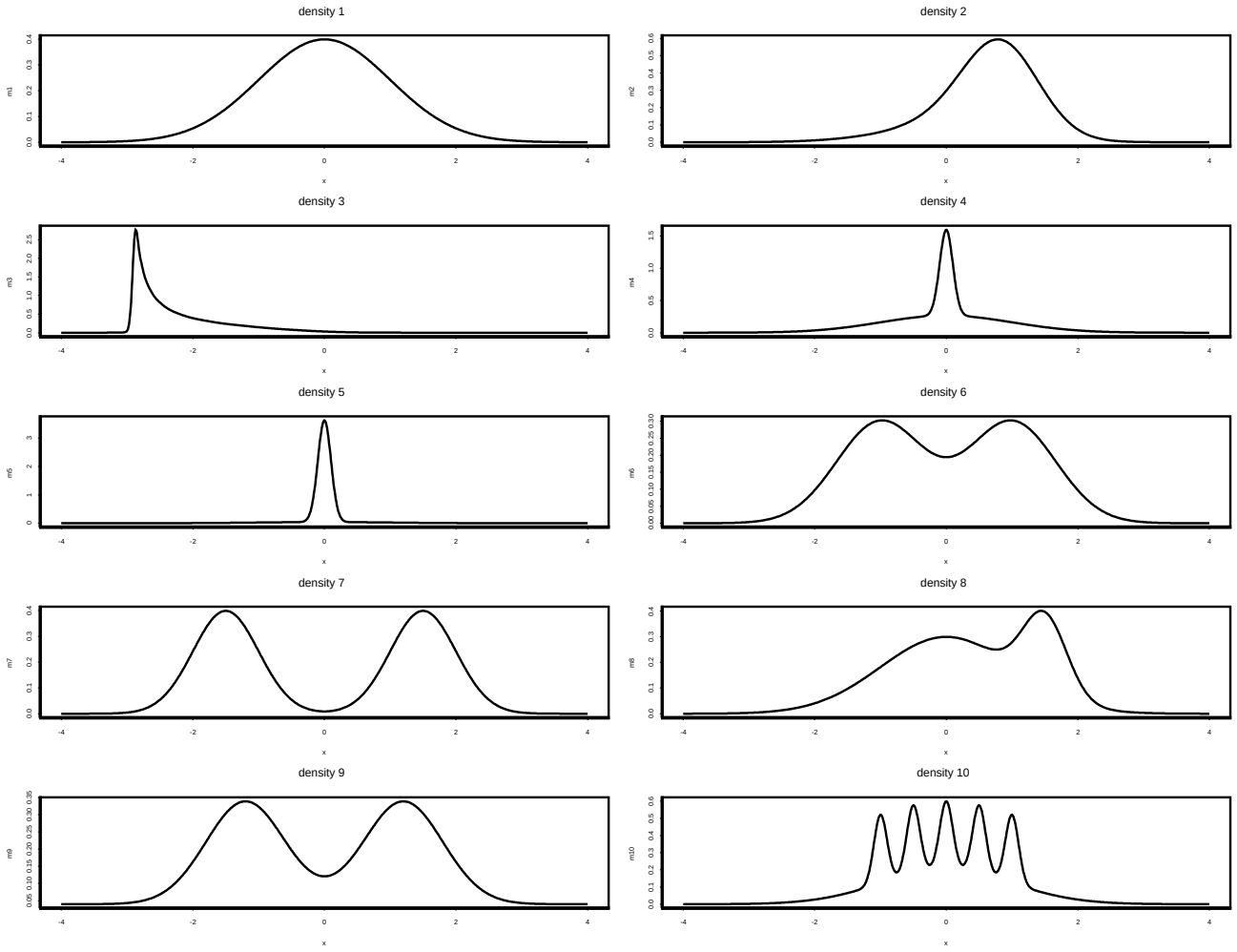

Figure 1.1: Mixtures 1-10 normal mixture densities from Marron and Wand(1992)

estimator, Cutler and Cordero-Braña (1996) postulated a two component normal mixture model $f_{\boldsymbol{\theta}_2}$, and showed that the performance of the MHD estimator of $\boldsymbol{\theta}_2$ remains unaffected even when the data are generated from a $100\alpha\%$ gross-error contaminated mixture model $(1-\alpha)f_{\boldsymbol{\theta}_2}(x) + \alpha\delta_z(x)$, where $\delta_z(x)$ is the normal density $N(8,1)$ and $\alpha = 0.01$. However, in terms of MHD estimator $\hat{m}_n$, such an approach would be inappropriate. By virtue of its consistency, our estimator $\hat{m}_n$ would (correctly) identify (for sufficiently large n) the number of components in the mixture from which data are generated, which in the above example

would be 3 instead of 2. Also, there is no Hellinger function representation of our estimator of mixture complexity which would facilitate the study of α-influence functions.

In view of these, we consider two different way of assessing the robustness of $\hat{m}_n$ when the postulated model is a mixture of normals but the data are generated from a mixture with symmetric departure from component normality. The first way is as described in Woodward et al. (1984) for the estimation of mixing proportions (also see Woodward et al. (1995) and McCann and Sarkar (2000)). The second setup is as described in Section 29.3.3 of Markatou (2001); also see section 4 of Markatou (2000). More specifically, for our simulation study, we consider a mixture with two components given by

$$f_{\boldsymbol{\theta}_m}(x) = pf_1(x) + (1-p)f_2(x), \tag{1.6.13}$$

where f_1 is the density associated with a random variable $X_1 = aY$ and f_2 is the density associated with a random variable $X_2 = Y + b$ for some $a > 0$ and $b > 0$. Here, the postulated distribution for Y is standard normal but, in the first setup, the samples are generated from the mixture in (1.6.13) when Y is a Student's $t(df)$-random variable with degrees of freedom $df = 2$ or 4. For our first setup, we set $p = .25, .50$ and $.75$, $a = 1$ and $\sqrt{2}$, and for each pair of (p, a) values, we choose the values of b so that the *overlap* (see Woodward et al. 1984 for definition) between the two t-component densities in (1.6.13) is either 0.10 or 0.03. The overlap is defined as the probability of misclassification using this rule: Classify an observation x as being from population 1 if $x < x_c$ and from population 2 if $x \geq x_c$, where x_c is the unique point between μ_1 and μ_2 such that $pf_1(x_c) = (1-p)f_2(x_c)$. These b values are given in Table 1.4 and Table 1.5 and they are referred to as t-overlap in Table 1.7 and Table 1.8. Note that the general shapes of such a two-component postulated (normal mixture) model and a two-component t-mixture model from which the data are generated are markedly different for some values of p, a and b. For instance, Figure 1.2 presents graphs for the case $p = 0.75$, $a = \sqrt{2}$, overlap=0.10, and $df = 4$ and reveals that the resulting mixture distributions have quite different shapes. (see, e.g., Figure 1 and Figure 2 in McCann and Sarkar (2000)). In

addition, the component densities in the sampling model have much heavier tail than those in the postulated (normal) mixture model.

Our second simulation setup differs slightly from the one above in that the samples are generated from the mixture in (1.6.13) when Y is a rescaled Student's $t(df)$-random variable with degrees of freedom $df = 3$ or 4. As in Markatou (2001), by a rescaled Student's $t(df)$ we mean a $t(df)$-random variable that is rescaled to have variance 1. Also, for each pair of (p, a) values given above, we choose the values of b so that the overlap between the two normal-component densities in (1.6.13) is either 0.10 or 0.03. That is, we use the b values that are given in Table 1.6. We will refer to these b values as N-overlap in Table 1.9 and Table 1.10 below.

The sample size for this study is $n = 1000$ and we performed 100 Monte Carlo replications of our MHDE algorithm and the MKE algorithm of James et al. (2001), both with $\alpha_{m,n} = 3/n$. Table 1.7 to Table 1.10 give a tally of estimated number of components for the MHDE and MKE algorithms, for each choice of a, p and b given above. In all these cases the true mixture complexity is 2 and we denote the highest percentage of correct identifications by an asterisk in Table 1.7 to Table 1.10.

The simulations presented here span over a variety of moderate to more extreme symmetric departures from component normality along with two different types and amounts of separation between the component densities. In all, there are 40 different cases of model misspecifications considered here, of which our MHDE algorithm significantly outperforms the MKE algorithm in about 36 cases but the MKE algorithm performs well only in 9 cases in terms of correctly identifying the true mixture complexity $m_0 = 2$. Only when the $t(2)$ components are poorly separated (t-overlap=0.10) and in the following three cases, (p, a)=$(0.5, 1)$,$(0.5, \sqrt{2})$ and $(0.75, \sqrt{2})$,Table 1.10 shows that the MKE outperforms our MHDE algorithm where the latter severely underestimates by practically ignoring an equal proportion $(1 - p = 0.5)$ or small proportion $(1 - p = 0.25)$of second (heavy tail) component located at values b = 3.771,4.517 and 4.401, respectively. This suggests that in these three

cases in Table 1.10, the performance of MHDE is affected by poor separation between the two $t(2)$ components and its inherent tendency to protect against heavy tail distributions,while it is possible that correct identification by the MKE may be partly due its tendency to fit a component to extreme values. It is interesting to note from Table 1.10, however, that even in the extreme departure case with $t(2)$ components, our MHDE algorithm outperforms the MKE algorithm when the components are well separated (t-overlap=0.03) and both the algorithms perform well when $p = 0.25$ and t-overlap=0.10.

In Table 1.1 and Table 1.3 of sections 1.5.1 and 1.5.2, respectively, we noticed that our MHDE algorithm seldom overestimates the true mixture complexity. However,Table 1.7 to Table 1.10 show that our MHDE algorithm overestimates rather mildly in some instances but rather severely in the two cases (p, a)=$(0.25,\sqrt{2})$ and (0.5,1), when sampling from mixtures with rescaled $t(3)$ components. We do not observe much underestimation with the MHDE algorithm. However, Table 1.7 to Table 1.10 show that in many instances the MKE algorithm rather severely overestimates or underestimates the true mixture complexity. Given the extreme nature of symmetric departures from component normality considered in our simulations, the results in Table 1.7 to Table 1.10 serve as a testament that our MHDE algorithm is highly robust, while the MKE algorithm is highly unstable.

Table 1.4: Parameter Values for Simulations: t(4)

		b	
p	a	Overlap=.10	Overlap=.03
.25	1	2.821	4.965
.50	1	3.066	5.202
.25	$\sqrt{2}$	3.175	5.777
.50	$\sqrt{2}$	3.672	6.249
.75	$\sqrt{2}$	3.570	6.151

Table 1.5: Parameter Values for Simulations: t(2)

		b	
p	a	Overlap=.10	Overlap=.03
.25	1	3.492	7.385
.50	1	3.771	7.793
.25	$\sqrt{2}$	3.957	8.546
.50	$\sqrt{2}$	4.517	9.354
.75	$\sqrt{2}$	4.401	9.185

Table 1.6: Parameter Values for Simulations: Normal

		b	
p	a	Overlap=.10	Overlap=.03
.25	1	2.319	3.603
.50	1	2.563	3.762
.25	$\sqrt{2}$	2.573	4.203
.50	$\sqrt{2}$	3.066	4.522
.75	$\sqrt{2}$	2.964	4.456

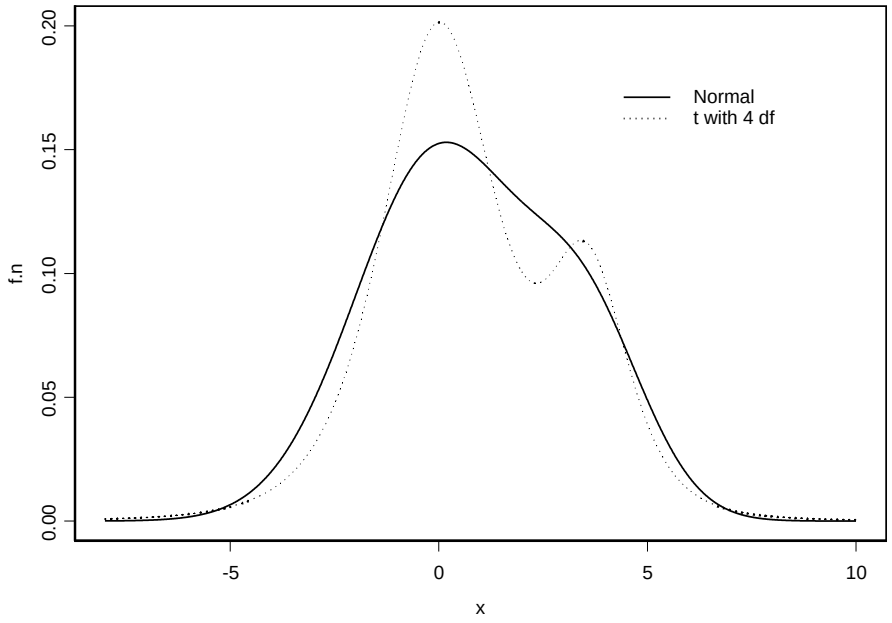

Figure 1.2: Comparison of t_4 and normal components when the means and variances are equal and set at the levels for t_4 with an overlap of 0.10, $a = \sqrt{2}$, and $p = 0.75$.

Table 1.7: Mixture Complexity Estimation results for $t(4)$ components

			t-overlap=.10					t-overlap=.03				
			Estimated number of components					Estimated number of components				
p	a		1	2	3	4	5	1	2	3	4	5
.25	1	MHDE	0	100*				0	100*			
		MKE	33	60*	7			2	23	75		
.25	1	MHDE	0	92*	8			0	100*			
		MKE	0	74*	26			0	35	64	1	
.50	1	MHDE	0	95*	5			0	100*			
		MKE	97	3				100				
.50	1	MHDE	0	100*				0	100*			
		MKE	94	4	2			99	1			
.75	1	MHDE	0	100*				0	100*			
		MKE	80	19	1			61	8	31		

Table 1.8: Mixture Complexity Estimation results for $t(2)$ components

			t-overlap=.10					t-overlap=.03				
			Estimated number of components					Estimated number of components				
p	a		1	2	3	4	5	1	2	3	4	5
.25	1	MHDE	3	97*				0	98*	2		
		MKE	6	91*	2	1		72	24	4		
.25	1	MHDE	0	100*				0	99*	1		
		MKE	8	89*	1	1	1	79	21			
.50	1	MHDE	89	11*				0	100*			
		MKE	9	77*	14			59	40	1		
.50	1	MHDE	77	23				0	100*			
		MKE	15	76*	9			88	12			
.75	1	MHDE	63	35	2			0	100*			
		MKE	9	86*	2	3		75	24	1		

Table 1.9: Mixture Complexity Estimation results for Rescaled $t(3)$ components

			N-overlap=.10					N-overlap=.03				
			Estimated number of components					Estimated number of components				
p	a		1	2	3	4	5	1	2	3	4	5
.25	1	MHDE	0	97*	3			0	100*			
		MKE	45	41	14			14	41	45		
.25	1	MHDE	0	60*	40			0	100*			
		MKE	10	63*	20	2		14	44*	38	4	
.50	1	MHDE	0	69*	31			0	97*	3		
		MKE	99	1				97	3			
.50	1	MHDE	0	91*	9			0	96*	4		
		MKE	98	2				98	2			
.75	1	MHDE	1	91*	8			0	100*			
		MKE	80	18	1	1		66	17	17		

Table 1.10: Mixture Complexity Estimation results for Rescaled $t(4)$ components

			N-overlap=.10					N-overlap=.03				
			Estimated number of components					Estimated number of components				
p	a		1	2	3	4	5	1	2	3	4	5
.25	1	MHDE	0	99*				0	100*			
		MKE	35	34	31			0	26	74		
.25	1	MHDE	0	88*	12			0	100*			
		MKE	55	44*	1			0	34*	64	2	
.50	1	MHDE	2	98*				0	99*	1		
		MKE	100					100				
.50	1	MHDE	1	99*	9			0	100*			
		MKE	100					99	1			
.75	1	MHDE	23	77*				0	100*			
		MKE	91	9				56	10	34		

1.7 ROEDER'S EXAMPLE

Here, we revisit Roeder (1994)'s example concerning mixture models for red blood cell sodium-lithium countertransport (SLC) activity data collected from 190 individuals. Geneticists are interested in SLC because it is correlated with blood pressure and hence may be an important cause of hypertension. For this data, Roeder (1994)'s methods based on graphical techniques and tests supported a three-component normal mixture. Roeder (1994)'s three-component fit was based on ML estimates, which were calculated assuming equal component variances. Roeder (1994) also noted that a square-root transformation of the SLC data pulls in large values and supports a two-component mixture.

Incidentally, for the SLC data, we assumed normal mixture models with unknown means, unequal variances and mixing proportions, and also obtained a mixture complexity estimate of $\hat{m}_n = 2$ using the MKE procedure of James et al. (2001) with threshold value $\alpha_{n,m} = 3/n$. For comparison sake, we computed the best fitting two component normal mixture density using updated MKE estimates, as done in our case above. These MKE parameter estimates are given in Table 1.11 under MKE($m = 2$).

It is well known that large values have little impact on MHD estimates. In view of this and the latter note of Roeder mentioned above, it may be natural to use our MHD algorithm to determine an estimate of mixture complexity for the SLC data. We use our MHDE algorithm to estimate mixture complexity for the SLC data. Here, we assume normal mixture models with unknown means, (unequal) variances and mixing proportions, and use our MHDE algorithm with threshold value $\alpha_{n,m} = 3/n$. Our analysis yields an estimate $\hat{m}_n = 2$ of the mixture complexity for the SLC data.

The SLC example was also discussed in Cutler and Cordero-Braňa (1996), where it was concluded that a three-component mixture based on MHD estimate provides a better fit than the one based on ML estimates. For comparison sake, values from Table1 of Cutler and Cordero-Braňa (1996) are given in Table 1.11 under MHDE($m = 3$) and MLE($m = 3$). Cutler and Cordero-Braňa (1996) showed that the fitted density based on ML estimates has

Table 1.11: Hypertension Parameter Estimates

	ϕ_1	ϕ_2	ϕ_3	μ_1	μ_2	μ_3	σ_1	σ_2	σ_3
MHDE($m = 2$)	.695	.305		.222	.352		.060	.106	
MKE ($m = 2$)	.754	.246		.225	.378		.060	.102	
MHDE($m = 3$)	.399	.485	.116	.199	.277	.424	.046	.078	.113
MKE ($m = 3$)	.568	.417	.015	.211	.322	.612	.051	.090	.014

a tiny component for some extreme data values, whereas the one based on MHD estimate largely ignores these data. Figure 1.3 below shows that our two-component normal mixture fit given by MHDE($m = 2$) and the one given by MKE($m = 2$) provide as good a fit as the three-component normal mixture fit given by MHDE($m = 3$).

In a recent article, Ishwaran, James and Sun (2001) adopted a Bayesian approach to estimating mixture complexity and proposed two algorithms called the *generalized weighted Chinese restaurant* (GWCR) and *blocked Gibbs sampler*. Their analysis of SLC data showed that GWCR supported a three component mixture while the blocked Gibbs sampler based on Bayes Information Criterion penalty supported a two-component mixture. Note that the latter result agrees with our answer. All these make a compelling case that our two-component mixture density based on the MHD (or the MKE) estimates provides a good and parsimonious fit of the SLC data.

1.8 SUMMARY AND CONCLUSIONS

An information criterion approach based on minimum Hellinger distances is used to construct an estimator of unknown number of components in finite mixtures, when the form of component densities are unknown but are postulated to be members of some parametric family. This estimator is consistent for parametric family of finite mixture models. When the postulated normal mixture model is same as the model from which samples are drawn, simulations show that our estimator competes well with other procedures available in the

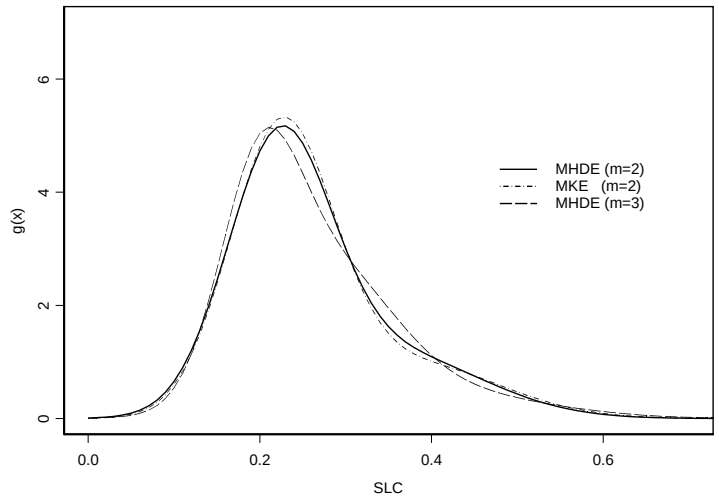

Figure 1.3: Fitted three-component normal mixture based on MHDE from Cutler and Cordero-Braña (1996) and fitted two-component normal mixture based on MHD and MKE algorithms for SLC data

literature, and particularly well against an estimator based on Kullback-Leibler distance introduced by James et al. (2001). The most distinguishing feature of our estimator is that it continues to identify the mixture complexity correctly even when the sampling model is a (moderate to more extreme) symmetric departure from postulated component normality, while the estimator of James et al. (2001) becomes highly unstable in these situations. The HMIX algorithm and the adaptive density estimator of Cutler and Cordero-Braña (1996) are crucial computational tools in our numerical studies. Updating the adaptive kernel density estimator at each step of our iterative procedure using the best fitted density further improves the performance of our estimator. Choice of threshold values $\alpha_{n,m}$ undoubtedly has an impact on the final estimate of the unknown mixture complexity. In our numerical studies we motivate our choice of $\alpha_{n,m} = 3/n$ based on the AIC criterion. More work remains to be done on the choice of $\alpha_{n,m}$ for our estimator, which is both consistent and robust. For an example concerning hypertension, our estimator and the estimator of James et al. (2001)

yield a parsimonious mixture model that provides good a fit of the data.

1.9 REFERENCES

Aitkin, M., and Wilson, G. T. (1980), "Mixture Models, Outliers, and the EM Algorithm," *Technometrics*, 22, 325-331.

Beran, R. (1977), "Minimum Hellinger distance estimates for parametric models," *The Annals of Statistics*, 5, 445-463.

Bogardus, C., Lillioja, S., Nyomba, B. L., Zurlo, F., Swinburn, B., Puente, A. E. -D., Knowler, W. C., Ravussin, E., Mott, D. M., and Bennett, P. H. (1989), "Distribution of in vivo insulin action in Pima-Indians as mixture of 3 normal-distributions", *Diabetes* 38, 1423-1432.

Böhning, D. (1999), *Computer-Assisted Analysis of Mixtures and Applications: Meta-Analysis, Disease Mapping and Others*, New York: Chapman & Hall/CRC.

Böhning, D., and Seidel, W. (2003), "Editorial: Recent Developments in Mixture Models," *Computational Statistics and Data Analysis*, 41, 349-257.

Chen, J. and Kalbfleisch, J. D. (1996), "Penalized minimum distance estimates in finite mixture models," *Canadian Journal of Statistics*, 24, 167-175.

Cordero-Braňa, O. I. and Cutler, A., (1997), "On the asymptotic properties of the minimum Hellinger distance estimator in the case of mixture model," unpublished manuscript, Utah State University.

Cutler, A., and Cordero-Braňa, O. I. (1996), "Minimum Hellinger distance estimation for finite mixture models," *Journal of the American Statistical Association* 91, 1716-1723.

Dacunha-Castelle, D. and Gassiat, E. (1997), "The estimation of the order of a mixture model," *Bernoulli*, 3, 279-299.

Dacunha-Castelle, D. and Gassiat, E. (1999), "Testing the order of a model using locally conic parameterization: population mixtures and stationary ARMA processes," *The Annals of Statistics*, 27, 1178-1209.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum-Likelihood From Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, Ser. B, 39, 1-38.

DeVeaux, R. D., and Krieger, A. M. (1990), "Robust Estimation of a Normal Mixture," *Statistics and Probability Letters*, 10, 1-7.

Devroye, L. P., and Györfi, L. (1985), *Nonparametric Density Estimation: The L_1 View*, New York: Wiley.

Donoho, D. L., and Liu, R. C. (1988), "The 'Automatic' Robustness of Minimum Distance Functionals," *The Annals of Statistics*, 16, 552-586.

Escobar, M. D. and West, M. (1995), "Bayesian density estimation and inference using mixtures," *Journal of the American Statistical Association*, 90, 577-588.

Eslinger, P. W., and Woodward, W. A. (1991), "Minimum Hellinger Distance Estimation for Normal Models," *Journal of Statistical Computation and Simulation*, 39, 95-113.

Everitt, B. S. and Hand, D. J. (1981), *Finite Mixture Distributions*, London: Chapman and Hall.

George, E. I., and Foster, D. P., (2000), "Calibration and Empirical Bayes Variable Selection," *Biometrika*, 87, 4, 731-747.

Henna, J. (1985), "On estimating of the number of constituents of a finite mixture of continuous distributions," *Annals of the Institute of Statistical Mathematics*, 37, 235-240.

Ishwaran, H., James, L. F., and Sun, J. (2001),"Bayesian Model Selection in Finite Mixtures by Marginal Density Decompositions," *Journal of the American Statistical Association*, 96, 1316-1332.

James, L. F., Priebe, C. E., and Marchette, D. J. (2001), "Consistent Estimation of Mixture Complexity," *The Annals of Statistics*, 29, 1281-1236.

Keribin, C. (2000), "Consistent estimation of the order of mixture models," *Sankhyā, Ser. A* 62, 49-62.

Leroux, B. G. (1992), "Consistent estimation of a mixing distribution," *The Annals of Statistics*, 20, 1350-1360.

Lindsay, B. G., (1995),"Mixture Models: Theory, Geometry, and Applications," NSF-CBMS Regional Conference Series in Probability and Statistics, Vol. 5, Institute of Mathematical Statistics, Hayward.

Markatou, M. (2000), "Mixture models, robustness and the weighted likelihood methodology", *Biometrics*, 56, 483-486.

Markatou, M. (2001), "A closer look at the weighted likelihood in the context of mixtures", *Probability and Statistical Models with Applications*, Charalambides, C.A., Koutras, M.V. and Balakrishnan, N. (eds), Chapman and Hall/CRC, 4447-467.

Markatou, M., Basu, A., and Lindsay, B. G. (1998), "Weighted likelihood estimating equations with a bootstrap root search", *Journal of the American Statistical Association*, 93, 740-750.

Marron, J. S. and Wand, M. P. (1992), "Exact mean integrated squared error," *The Annals of Statistics*, 20, 712-736.

McCann, M. and Sarkar, S. (2000), "Minimum Negative Exponential Disparity Estimation of Mixture Proportions," *Journal of Statistical Planning and Inference* 87, 187-197.

McLachlan, G. J. (1987), "On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture," *Journal of the Royal Statistical Society, Ser. C (Applied Statistics)* 36, 318-324.

McLachlan, G. J. and Basford, K. E. (1988), *Mixture Models: Inference and Applications to Clustering*, New York: Marcel Dekker.

McLachlan, G. J., McLaren, C. E., and Matthews, D. (1995), "An algorithm for the likelihood ratio test of one versus two components in a mixture model fitted to grouped and truncated data," *Communications in Statistics – Simulation and Computation*, 24, 965-985.

McLachlan, G. J. and Peel, D. (1997), "On a resampling approach to choosing the number of components in normal mixture models," in *Computing Science and Statistics*, Vol. 28, eds. L. Billard and N. I. Fisher (Eds.). Fairfax Station, Virginia: Interface Foundation of North America, pp. 260-266.

McLachlan, G. J. and Peel, D. (2000), "Mixtures of factor analyzers", in *Proceedings of the Seventeenth International Conference on Machine Learning*, San Francisco: Morgan Kaufmann, pp. 599-606.

McLaren, C. E. (1996), "Mixture models in haematology: a series of case studies," *Statistical Methods in Medical Research*, 5, 129-153.

McLaren, C. E., Wagstaff, M., Brittenham, G. M., and Jacobs, A. (1991),"Detection of Two Component Mixtures of Lognormal Distributions in Grouped Doubly-truncated Data: Analysis of Red Blood Cell Volume Distributions," *Biometrics*, 47, 607-622.

Poland, W. B., and Shachter, R. D. (1994), "Three approaches to probability model selection", *In Uncertainty in Artificial Intelligence: Proceedings of the Tenth Conference*, San Mateo, CA: Morgan Kaufmann, 478-483.

Pollard, D. (1981), "Strong Consistency of K-Means Clustering," *The Annals of Statistics*, 9, 135-140.

Priebe, C. E. and Marchette, D. J. (2000), "Alternating kernel and mixture density estimates," *Computational Statistics and Data Analysis*, 35, 43-65.

Redner, R. A. and Walker, H. F. (1984), "Mixture densities, maximum likelihood and the EM algorithm", *SIAM Review*, 26, 195-239.

Richardson, S. and Green, P. J. (1997), "On Bayesian analysis of mixtures with an unknown number of components (with discussion)," *Journal of the Royal Statistical Society Ser. B*, 59, 731-792. Correction (1998). *Journal of the Royal Statistical Society Ser. B*, 60, 661.

Rissanen, J. (1978), "Modeling by Shortest Data Description," *Automatica*, 14, 465-471.

Roeder, K. (1994), " A graphical technique for determining the number of components in a mixture of normals," *Journal of the American Statistical Association*, 89, 487-495.

Roeder, K. and Wasserman, L. (1997), "Practical Bayesian density estimation using mixtures of normals," *Journal of the American Statistical Association*, 92, 894-902.

Scott, D. W. (1992), *Multivariate Density Estimation*, New York: Wiley.

Simpson, D. G. (1987), "Minimum Hellinger distance estimation for the analysis of count data," *Journal of the American Statistical Association*, 82, 802-807.

Stather, G. R. (1981), "Robust statistical inference using Hellinger distance methods," unpublished Ph.D. dissertation, LaTrobe University, Australia, Department of Mathematical Statistics.

Tamura, R. N. and Boos, D. D. (1986), "Minimum Hellinger distance estimation for multivariate location and covariance," *Journal of the American Statistical Association*, 81, 223-229.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985), *Statistical Analysis of Finite Mixture Distributions*, New York: Wiley.

Windham, M. P. and Cutler, A. (1994), "Mixture Analysis with Noisy Data," in *New Approaches in Classification and Data Analysis*, eds. E. Diday, Y. Lechevallier, M. Schader, P. Bertrand, and B. Burtschy, Berlin: Springer-Verlag.

Woodward, W. A., Parr, W. C., Schucany, W. R., and Lindsay, H. (1984), "A Comparison of Minimum Distance and Maximum Likelihood Estimation of a Mixture Proportion," *Journal of the American Statistical Association*, 79, 590-598.

Woodward, W. A., Whitney, P., and Eslinger, P. (1995), "Minimum Hellinger Distance Estimation of Mixture Proportions," *Journal of Statistical Planning and Inference*, 48, 303-319.

Chapter 2

ROBUST ESTIMATOR OF MIXTURE COMPLEXITY: DISCRETE CASE

2.1 INTRODUCTION

Data consisting of counts often occur in areas such as public health, epidemiology, economics, sociology, psychology, engineering and agriculture. However, there are many instances where count data do not conform to simple mean variance relationships implied in using Binomial, Poisson or multinomial models. For example, a Poisson model often underestimates the observed dispersion. This phenomenon, called *overdispersion*, occurs because a single Poisson parameter λ is often insufficient to describe the population. In fact, in many cases it can be suspected that *population heterogeneity* which has not been accounted for is causing this overdispersion. One approach to this problem is to assume that the heterogeneity involved in the data can be adequately described by a mixed distribution.

Finite mixtures, in particular Poisson mixtures, have played a very useful role in modeling overdispersed count data arising in disease mapping and risk assessment (Schlattmann and Böhning 1993, Böhning 1999, Lawson et al 1999), mutation research and genetics (Beyers and Shenton 1999), health-care utilization (Mullahy 1997, Deb and Trivedi 1997) and health planning and management (Xiao, Lee and Vemuri 1999), to name a few. For more examples of discrete and continuous cases, a detailed description of estimation approaches and extensions to finite mixtures of Poisson regressions for count data, see Everitt and Hand (1981), Titterington, Smith and Makov (1985), and McLachlan and Basford (1988), Lindsay (1995), Böhning (1999) and McLachlan and Peel (2000).

EM algorithm of Dempster, Laird and Rubin (1977) is a widely used method to compute maximum likelihood estimates (MLE) of all the parameters in finite mixture models when

the number of components is assumed to be known. Recently, for Poisson mixtures, Karlis and Xekalaki (1998) developed a minimum Hellinger distance (MHD) estimator, which is an appealing alternative to the MLE, especially when the postulated mixture model is incorrect. Karlis and Xekalaki (1998) also developed an iterative algorithm which facilitates computation of MHD estimates of Poisson mixture parameters. Furthermore, they showed that the MHD estimators achieve efficiency at the model density and simultaneously possess desirable robustness properties under gross-error contaminations, thus reconciling the conflicting concepts of robustness and efficiency. Lu, Hui and Lee (2003) considered MHD estimation for finite mixtures of Poison regression for count data.

Typically, in practice where the mixture model is being used to handle overdispersion in count data, the number of components in the mixture has to be inferred from the data. From now on, we will refer to the number of components in a mixture as *mixture complexity*. Estimation of mixture complexity is a rather fundamental, yet challenging problem. Correct identification of mixture complexity followed by an efficient estimation of mixture parameters would lead to finding the mixture with fewest possible components that provides a satisfactory fit of the count data.

A survey of literature shows that, in the continuous and discrete cases, developing methods to determine mixture complexity has been an area of intense research for many years. In the continuous case, a variety of approaches for determining the mixture complexity have been discussed in the literature. See Henna (1985); McLachlan (1987); Roeder (1994); Escobar and West (1995); Chen and Kalbfleisch (1996); Dacunha-Castelle and Gassiat (1997, 1999); Roeder and Wasserman (1997); Keribin (2000); Priebe and Marchette (2000); James, Preibe and Marchette (2001); Ishwaran, James and Sun (2001); Woo and Sriram (2004), and references therein. For instance, James, Preibe and Marchette (2001) used Kullback-Liebler (KL) distance to construct a consistent estimator of mixture complexity, when the component densities are assumed to be normal.

Chapter 1 developed MHD estimation method to determine mixture complexity in the continuous case. It constructed an estimator of mixture complexity as a by-product of minimizing a Hellinger Information Criterion (HIC) defined in (1.2.7). When the mixture model is correctly specified, they showed that their estimator of mixture complexity is consistent and competes well against several other procedures in terms of correctly identifying the mixture complexity, for a variety of target densities. An attractive features of MHD estimator (1.2.8) in chapter 1 is that it is robust against model misspecification. That is, it continues to identify the mixture complexity correctly even when the sampling mixture model is a (moderate to more extreme) symmetric departure from postulated component normality, while the estimator of James et al. (2001) becomes highly unstable in these situations.

For count data, Schlattmann and Böhning (1993) used the resampling approach of McLachlan (1987) to decide on mixture complexity in their application of Poisson mixtures to disease mapping. Also, Pauler et al (1996) used this method to determine the mixture complexity in their modeling of anticipatory saccade counts from schizophrenic patients and controls. Karlis and Xekalaki (1999) determined the mixture complexity using a sequential testing procedure based on likelihood ratio test (LRT) that utilizes a resampling technique. Via simulations for a variety of target Poison mixtures and examples, Karlis and Xekalaki (1999) also illustrated the ability of their method to correctly determine the mixture complexity. Dellaportas, Karlis and Xekalaki (1997) used Bayesian analysis to infer the mixture complexity in Poisson mixtures and applied their results to a financial data. Recently, Karlis and Xekalaki (2001) developed diagnostics based on Hellinger gradient function in order to examine the presence of a mixture and obtain a semiparametric MHD estimate of number of components in Poison mixtures.

In many applications, however, it is unrealistic to expect that the component distributions are Poisson. More importantly, if one misspecifies the mixture model, which happens when one postulates a mixture model with Poisson components but the data are generated from, say, a mixture model with negative binomial components, then the mixture complexity

estimate based on sequential likelihood ratio tests of Karlis and Xekalaki (1999) may be unstable. This instability may become more severe if the overdispersion in negative binomial components becomes more severe.

Clearly, the above scenarios necessitate the development of an estimator of mixture complexity for the count data which performs well whether or not the postulated mixture model is correct. This chapter uses the methods developed in chapter 1 to construct an estimator of mixture complexity for count data which is consistent when the exact form of the component distributions are unknown but are postulated to be members of some parametric family and simultaneously robust against model misspecification. Note that our goal is not merely restricted to finding a consistent and robust estimator of the mixture complexity, but rather use this approach in fitting the best possible mixture distribution based on MHD estimates, which are inherently less influenced by small proportions of extreme data values.

In section 2.2, we present the basic framework and propose an estimator of mixture complexity using a Hellinger information criterion. The main theorem concerning the consistency of the estimator is stated and proved in section 2.3. Computational details concerning our estimator are given in section 2.4. In section 2.5.1, we carry out extensive Monte Carlo studies for a variety of correctly specified 2-,3- and 4- component Poisson mixtures and , in each case, compare the ability of our estimator in correctly determining the mixture complexity with those given in Karlis and Xekalaki (1999). In section 2.5.2, we examine the robustness of our estimator through extensive simulations, when postulated mixture model is incorrect. In section 2.6, we estimate the mixture complexity for three count datasets with overdispersion, two of which with possible zero-inflation. Our estimate of the number of components to use in these examples are compared with those in the literature. Overall summary and conclusions are given in section 2.7. We begin with some basic notations and definitions.

2.2 MHD ESTIMATOR OF MIXTURE COMPLEXITY

Consider a parametric family of probability mass functions (p.m.f.'s) $\mathcal{F}_m = \{f_{\boldsymbol{\theta}_m} : \boldsymbol{\theta}_m \in \Theta_m \subseteq R^p\}$ concentrated on $\mathcal{X} = \{0, 1, 2, \dots\}$ for each fixed $m > 0$, such that $f_{\boldsymbol{\theta}_m}$ can be represented as a finite mixture of the form

$$f_{\boldsymbol{\theta}_m}(x) = \sum_{i=1}^{m} \pi_i f(x|\phi_i), \quad x \in \mathcal{X}, \tag{2.2.1}$$

where $f(x|\phi_i)$ is the component p.m.f., $\phi_i \in R$, the mixing proportions $\pi_i \geq 0$, $\sum_{i=1}^{m} \pi_i = 1$ for $i = 1, \dots, m$ and $\boldsymbol{\theta}_m = (\pi_1, \dots, \pi_{m-1}, \phi_1, \dots, \phi_m)$. The class $\mathcal{F}_m \subseteq \mathcal{F}_{m+1}$ for all m and we denote $\mathcal{F} = \bigcup_{m=1}^{\infty} \mathcal{F}_m$.

Let $X_1, \dots, X_n$ be independent random variables taking values in $\mathcal{X}$ with an unknown p.m.f. $f_0 \in \Gamma$, where Γ denotes the set of all p.m.f.'s defined on $\mathcal{X}$. For an arbitrary p.m.f. $f \in \Gamma$, define the *index of the economical representation* of f, relative to the family of mixtures defined above, as

$$m(f) = \min\{m : f \in \mathcal{F}_m\}.$$

If indeed f is a finite mixture then $m(f)$ is finite and denotes the true mixture complexity; otherwise $m(f) = \infty$. Note that $m(f)$ represents the most parsimonious mixture model representation for f. We now describe a robust estimation procedure to estimate $m_0 = m(f_0)$.

To this end, we follow the approach of Simpson (1987) and define the Hellinger distance between two p.m.f.'s $f, g \in \Gamma$ by

$$\begin{aligned} H^2(f, g) &= \sum_{k=0}^{\infty} |f^{1/2}(x) - g^{1/2}(x)|^2 \\ &= 2 - 2\sum_{k=0}^{\infty} f^{1/2}(x) g^{1/2}(x). \end{aligned} \tag{2.2.2}$$

see Simpson (1987), for example. Let $\hat{f}_n$ be the empirical mass function which defines the nonparametric estimator of f_0:

$$\hat{f}_n(x) = n^{-1} \sum_{i=1}^{n} I_{\{X_i = x\}}, \quad x = 0, 1, \dots, \tag{2.2.3}$$

where I_A is the indicator of set A. When $m > 0$ is known, the MHD estimator $\hat{\boldsymbol{\theta}}_{n,m}^{MHD}$ of $\boldsymbol{\theta}_m$ is defined as the value of a functional $T_m(f)$ at $\hat{f}_n$, where for any f, $T_m(f)$ is defined by

$$T_m(f) = \{\boldsymbol{\theta}_m \in \Theta_m : H(f_{\boldsymbol{\theta}_m}, f) = \min_{\boldsymbol{t}_m \in \Theta_m} H(f_{\boldsymbol{t}_m}, f)\}. \tag{2.2.4}$$

Here $T_m(f)$ is the set of solutions, if one exits, and a member of $T_m(f)$ is chosen arbitrarily when a solution is required. For each integer $m > 0$, define

$$\hat{f}^m = \arg \min_{f \in \mathcal{F}_m} H(\hat{f}_n, f)$$

and

$$f_0^m = \arg \min_{f \in \mathcal{F}_m} H(f_0, f) \tag{2.2.5}$$

where f_0 is the underlying mass function. Then, note that $\hat{f}^m = f_{\hat{\boldsymbol{\theta}}_{n,m}^{MHD}}$ and $f_0^m = f_{T_m(f_0)}$.

Note that we can express m_0 as

$$\begin{aligned} m_0 &= \min\{m : H(f_0, f_0^m) - H(f_0, f_0^{m+1}) \leq 0\} \\ &= \min\{m : H(f_0, f_0^m) = 0\}, \end{aligned} \tag{2.2.6}$$

because $\mathcal{F}_m \subseteq \mathcal{F}_{m+1}$. Since the family of mixtures is nested, estimation of unknown number of components can be considered as a model selection problem, that is, selecting the model that fits a given dataset the best in some sense out of a candidate set of models. Poland and Shachter (1994) compare three approaches to model selection. Motivated by the classical Akaike type of criterion for model selection and third approach of Poland and Shachter (1994) for model selection involving the Kellback-Leibler distance, a model selection criterion based on the Hellinger distance may be considered as the form

$$HIC = H^2(\hat{g}_n, \hat{g}^m) + n^{-1}b(n)\nu(m), \tag{2.2.7}$$

where $b(n)$ depends only on n and $\nu(m)$ is the number of parameters in the mixture model. Here, the value of m yielding the minimum HIC specifies the best model. In the context of minimum Hellinger distance estimation, the statistics $H^2(\hat{g}_n, \hat{g}^m)$ at (2.2.7) can be viewed as

measuring goodness-of-fit of mixture models, and $n^{-1}b(n)\nu(m)$ as penalizing the goodness-of-statistics by a term proportional to the number of parameters in the mixture model. A simple heuristic to find the best model from a sequence of nested models is to try successive models, starting with the smallest, and stop with model when the HIC value for model m is less than that for model (m+1), that is,

$$H^2(\hat{g}_n, \hat{g}^m) + n^{-1}b(n)\nu(m) \leq H^2(\hat{g}_n, \hat{g}^{m+1}) + n^{-1}b(n)\nu(m+1),$$

or, equivalently,

$$H^2(\hat{g}_n, \hat{g}^m) - H^2(\hat{g}_n, \hat{g}^{m+1}) \leq n^{-1}b(n)[\nu(m+1) - \nu(m)]. \tag{2.2.8}$$

Hence, setting $\alpha_{n,m} = n^{-1}b(n)[\nu(m+1) - \nu(m)]$ in (2.2.8) naturally leads to the following estimator of m_0 defined by

$$\hat{m}_n = \min\{m : H(\hat{g}_n, \hat{g}^m) \leq H(\hat{g}_n, \hat{g}^{m+1}) + \alpha_{n,m}\} \tag{2.2.9}$$

where $\hat{g}_n$ is the density estimator in (2.2.3) and $\{\alpha_{n,j}; j \geq 1\}$ are positive sequences of threshold values chosen in such a way they converge to zero as $n \to \infty$. We define $\hat{m}_n = \infty$ if the the minimum m in (2.2.9) does not exist for any n. Also, note that the estimator in (2.2.9), motivated is essentially a sample version of the representation of m_0 in (2.2.6).

The equation (2.2.9) actually defines a class of (competing) estimators since the threshold value $\alpha_{n,m}$ has not been specified precisely yet. It can be seen easily from (2.2.9) that threshold values directly impact the $\hat{m}_n$ values, which increase as $\alpha_{n,m}$ values decrease. Since an $\hat{m}_n$ value determines the final mixture model for a dataset, choice of $\alpha_{n,m}$ values may also be viewed as a model selection problem. In all our numerical studies we assume Poisson mixture models with m univariate components, in which case the number of unknown parameters is $\nu(m) = 2m - 1$. By the definition of $\alpha_{n,m}$ above, this leads to $\alpha_{n,m} = 2b(n)/n$. Following Akaike Information Criterion (AIC), if we set $b(n) = 1$ then this leads to a threshold value $\alpha_{n,m} = 2/n$. Schwarz Bayesian Criterion (SBC) which sets $b(n) = ln(n)/2$ leads to $\alpha_{n,m} = ln(n)/n$. These two threshold values are used in all our numerical studies below,

where, unsurprisingly, we observe that the threshold $ln(n)/n$ based on SBC has a tendency to select a more parsimonious model than that based on AIC.

2.3 CONSISTENCY OF $\hat{m}_n$

In this section, we establish the consistency of MHD estimator of mixture complexity defined in (2.2.9) as a theorem. Since the proof of the theorem is very similar to that of the Theorem 1 in chapter 1, we only give a brief sketch of it.

For the consistency result, we will assume the following regularity conditions (see Simpson(1987) and Karlis and Xekalaki (1998)). Let $\tilde{\Gamma} \subset \Gamma$ denote the sub-class of p.m.f.'s defined on $\mathcal{X}$ for which the following condition holds (see Simpson (1987), equation (3.3)): For each m, there is a compact set $C_m \subseteq \Theta_m$ such that for every $f \in \tilde{\Gamma}$,

$$\inf_{\mathbf{t_m} \in \Theta_m - C_m} H(f_{\mathbf{t_m}}, f) > H(f_{\boldsymbol{\theta}_m^*}, f), \tag{2.3.10}$$

for some $\boldsymbol{\theta}_m^* \in C_m$. If, for each m, Θ_m is compact then $C_m = \Theta_m$. For each m, we will assume that $f_{\boldsymbol{\theta}_m}(x)$ is continuous in $\boldsymbol{\theta}_m$ for each $x \in \mathcal{X}$ and the class $\mathcal{F}_m$ is identifiable (see Teicher (1960,1961)). Under these conditions, Theorem 1 and Corollary of Simpson (1987) imply the following: For each $f \in \tilde{\Gamma}$, $T_m(f)$ exists, and if $T_m(f)$ is unique, then $H(\hat{f}_n, f) \to 0$ implies that $T_m(\hat{f}_n) \to T_m(f)$, as $n \to \infty$. In particular, $\hat{\theta}_{n,m}^{MHD} = T_m(\hat{f}_n)$ is consistent for $T_m(f_0)$. See Karlis and Xekalaki (1998) for a detailed account of MHD estimation for Poisson mixtures.

Theorem Assume that $X_1, \ldots, X_n$ are independent and identically distributed random variables with a p.m.f. $f_0 \in \tilde{\Gamma}$ and that all the regularity conditions stated above are satisfied. Then, for any sequence $\alpha_{n,m} \to 0$,

$$\hat{m}_n \to m_0 \quad \text{a.s.}$$

as $n \to \infty$, for $\hat{m}_n$ and m_0 defined in (2.2.9) and (2.2.6), respectively. If f_0 is not a finite mixture, then $\hat{m}_n \to \infty$.

Proof. First note that $H^2(\hat{f}_n, f_0) \leq \sum_{x=0}^{\infty} |\hat{f}_n(x) - f(x)|$. By the strong law of large numbers, $\hat{f}_n \to f_0$ almost surely (a.s.) as $n \to \infty$. Therefore, by Glick's theorem (Devroye and Györfi

(1985), p.10) we have that $\sum_{x=0}^{\infty} |\hat{f}_n(x) - f(x)| \to 0$ a.s., which implies that $H^2(\hat{f}_n, f_0) \to 0$. The rest of the proof follows using arguments exactly as for the Theorem in chapter 1, by replacing integrals with infinite sums.

2.4 COMPUTATIONAL DETAILS

In order to numerically assess the performance of our estimator of mixture complexity to correctly determine the number of components in a mixture and compare them with those available in the literature, we restrict our attention to Poisson mixture models. We assume $f_{\boldsymbol{\theta}_m}(x)$ in (2.2.1) denotes an m-component Poisson mixture with component means $\lambda_i, i = 1, \ldots, m$, where $\boldsymbol{\theta}_m = (\pi_1, \ldots, \pi_{m-1}, \lambda_1, \ldots, \lambda_m)$ with $0 \leq \lambda_1 < \ldots < \lambda_m$. Incidentally, Poisson mixtures with increasing component means satisfy the assumptions of our main theorem (see Teicher (1960,1961) and Karlis and Xekalaki (1998,1999)).

Computation of an estimate of mixture complexity using (2.2.9) is clearly an iterative procedure. The procedure starts by assuming that the data comes from a mixture with single Poisson component ($m = 1$) whose mean is unknown. Using the empirical mass function in (2.2.3), an MHD estimate of the parameter $\boldsymbol{\theta}_1$ is computed, which yields the best parametric fit $\hat{f}^1$. The Hellinger distance between the empirical mass function and $\hat{f}^1$ is then computed. Next, another Poisson component is added yielding a Poisson mixture with two components ($m = 2$), the best parametric fit $\hat{f}^2$ is computed using the MHDE estimate of $\boldsymbol{\theta}_2$, and the Hellinger distance between the best parametric fit $\hat{f}^2$ and the empirical mass function is computed. The difference between the two Hellinger distances is compared with the threshold value $\alpha_{n,1}$. The above procedure of adding one more component to the previous mixture is repeated until the first value $m = k$ for which the difference between Hellinger distances computed at k and $k + 1$ as in (2.2.9) falls below the corresponding threshold value $\alpha_{n,k}$. At this time, the procedure terminates and declares k as an estimate of the number of components in the mixture. Note that, at this stage, our procedure automatically provides the best parametric fit determined by the MHD estimate of $\boldsymbol{\theta}_k$. For all the datasets considered

in section 6, we compute the best parametric fit based on the MHD estimate provided at the termination of our (mixture complexity) computational algorithm.

At each step of iterative procedure, MHD estimates need to be computed. Computation of MHD estimate of $\boldsymbol{\theta}_m$, for each fixed m, requires minimizing the Hellinger distance $\sum_{x=0}^{\infty} |\hat{f}_n^{1/2}(x) - f_{\boldsymbol{\theta}_m}^{1/2}(x)|^2$, which is equivalent to maximizing $\sum_{x=0}^{\infty} \hat{f}_n^{1/2}(x) f_{\boldsymbol{\theta}_m}^{1/2}(x)$ from the equality in (2.2.2) with respect to $\boldsymbol{\theta}_m$, subject to the constraint $\pi_i > 0, i = 1, \ldots, m$, and $\sum \pi_i = 1$, for $\hat{f}_n$ defined in (2.2.3). Then, system of estimating equations can be written in the form

$$\sum_{i=0}^{\infty} \left(\frac{\hat{f}_n(x)}{f_{\boldsymbol{\theta}_m}(x)}\right)^{1/2} (f(x|\lambda_j) - f(x|\lambda_k)) = 0, j = 1, 2, \ldots, m-1, \tag{2.4.11}$$

$$\sum_{i=0}^{\infty} \left(\frac{\hat{f}_n(x)}{f_{\boldsymbol{\theta}_m}(x)}\right)^{1/2} p_j (f(x-1|\lambda_j) - f(x|\lambda_j)) = 0, j = 1, 2, \ldots, m-1, \tag{2.4.12}$$

where $f(x|\lambda) = \frac{exp(-\lambda)\lambda^x}{x!}$. An analytical solution of the above system of equations is not feasible, hence numerical methods are required to solve it. Here, we use the numerical algorithm, known as HELMIX, due to Karlis and Xekalaki (1998) which facilitates the MHD estimation procedure. See section 4 of Karlis and Xekalaki (1998) for more details on the HELMIX algorithm. Finally, as motivated at the end of section 2.2, we set the threshold values $\alpha_{n,m} = 2/n$ and $ln(n)/n$ and numerically study the performance of these two thresholds.

2.5 MONTE CARLO SIMULATIONS

In this section, we conduct two simulation studies to assess the performance of our estimator of mixture complexity defined in (2.2.9) for moderate to large sample sizes. In both of these studies the postulated model is a Poisson mixture. The first numerical study examines the efficiency of our estimator when the model is correctly specified, that is, the data are also generated from a Poisson mixture model. The second study examines the robustness of our estimator against model misspecification, that is, the data are generated from a mixture model where the component distributions are negative binomial with moderate to more

extreme overdispersion. These are done in the following subsections.

2.5.1 SIMULATIONS FOR CORRECTLY SPECIFIED POISSON MIXTURES

In order to compare the performance of our estimator with that of Karlis and Xekalaki (1999), the data is drawn from the 2-,3-, and 4-component Poisson mixtures studied in their article. As mentioned earlier, Karlis and Xekalaki (1999) used a sequential testing procedure based on LRT along with bootstrapping to determine the number of components in these mixtures. More specifically, we consider the following 2-,3- and 4-component Poisson mixtures given in Table 2.1.

Table 2.1: Poisson mixtures from Karlis and Xekalaki(1998)

Mixture Complexity	$f_{\boldsymbol{\theta}}(x)$
2	$.5P(1) + .5P(9)$
	$.8P(1) + .2P(9)$
	$.5P(1) + .5P(1.1)$
	$.95P(1) + .5P(10)$
3	$.45P(1) + .45P(5) + .1P(10)$
	$.4P(1) + .4P(3) + .2P(3.1)$
	$.33P(1) + .33P(5) + .33P(10)$
4	$.3P(1) + .4P(5) + .25P(9) + .05P(15)$
	$.3P(1) + .3P(1.2) + .2P(5) + .2P(9)$
	$.25P(1) + .25P(5) + .25P(10) + .25P(15)$

These Poisson mixtures include models with well separated components and poorly separated ones (in terms of component means), and models that result in skewed distributions. For each target mixture, we implemented our computational algorithm described in section 2.4 for four sample sizes $n = 50, 100, 500, 1000$ and using the two threshold values $\alpha_{n,m} = 2/n$, and $ln(n)$. For each sample size and threshold value, we performed 500 Monte Carlo replications of our algorithm, each yielding an estimate of mixture complexity. Tables 2.2 to 2.4 below correspond to the above mentioned 2-,3-, and 4-component mixtures, respectively. Each gives the relative frequencies (out of 500 replications) of the number of components

determined by our method for each sample size and threshold value. For comparative purposes, Tables 2.2 to 2.4 also list the relative frequencies from Table 1 in section 4 of Karlis and Xekalaki (1999), denoted by LRT. Note that we do not list the relative frequencies for LRT for $n = 1000$ because this case is not considered in Karlis and Xekalaki (1999). In Tables 2.2 to 2.4, the percentage (50% above) of correct identifications is given in bold with an asterisk beside it.

For the three well separated 2-component Poisson mixtures, Table 2.2 shows that our method correctly determines the number of components for both threshold values and for all the sample sizes, except in few small sample situations. In the case of $.8P(1) + .2P(9)$ and $n = 50$, only the SBC threshold value $\alpha_{n,m} = ln(n)/n$ incorrectly determines the true mixture complexity. This may be because the SBC based method, in general, has a tendency to select a more parsimonious model. In the case of $.95P(1) + .5P(10)$ and small sample sizes $n = 50$, and 100, both the threshold values $\alpha_{n,m} = 2/n$, and $ln(n)$ essentially ignore the second component, which has a very large mean and a small mixing proportion of .05. The latter may be attributable to the tendency of MHD to ignore the presence of a component with a very large mean and a small mixing proportion, especially for small samples. Note that the LRT method of Karlis and Xekalaki (1999) does will in all the three well separated cases. For the poorly separated 2-component Poisson distribution(i.e., $.5P(1) + .5P(1.1)$), our method incorrectly chooses a simple Poisson distribution for both the thresholds and for all sample sizes, which incidentally happens with the LRT based procedure of Karlis and Xekalaki (1999) as well. Overall, the performance of our mixture complexity estimator is as good as the LRT method of Karlis and Xekalaki (1999), except in few small sample situations.

For the two well separated 3-component Poisson mixtures, Table 2.3 shows that our method correctly determines the number of components for large sample sizes ($n =$ 500,1000), except in the case of $.45P(1) + .45P(5) + .1P(10)$ and sample size $n = 500$, where the SBC threshold value $\alpha_{n,m} = ln(n)/n$ incorrectly determines a 2-component

Poisson mixture. When $n = 50$, and 100, both the threshold values incorrectly determine a 2-component Poisson mixture for the case $.45P(1) + .45P(5) + .1P(10)$. For the cases $.45P(1) + .45P(5) + .1P(10)$ and $.33P(1) + .33P(5) + .33P(10)$, it should be noted that LRT method of Karlis and Xekalaki (1999) correctly determines the number of components when $n = 100$. For the poorly separated 3-component Poisson mixture (i.e., $.4P(1) + .4P(3) + .2P(3.1)$), neither our method nor the LRT based procedure of Karlis and Xekalaki (1999) is able to correctly determine the number of components for any sample sizes. For the 4-component Poisson mixtures, Table 2.4 shows that neither our method nor the LRT based procedure of Karlis and Xekalaki (1999) is able to correctly determine the number of components in these cases for any sample sizes.

In conclusion, when the postulated model is correct, our MHD based method is competitive with the LRT method of Karlis and Xekalaki (1999) in that it is very successful in correctly determining the mixture complexity if the model is well separated and sample sizes are large enough.

2.5.2 ROBUSTNESS OF $\hat{m}_n$ UNDER MODEL MISSPECIFICATION

Here, we describe an approach to assess the robustness of $\hat{m}_n$ in terms of its ability to correctly identify the true mixture complexity, when the postulated Poisson mixture model is incorrect. Generally, one examines the robustness of MHD estimatos against $100\alpha\%$ gross-error contaminated mixture models and using α-influence functions defined in terms of Hellinger functionals (Beran 1977). Karlis and Xekalaki (1998; see section 6.3) postulated a 2-component Poisson mixture model, $f_{\theta_2}(x) = \pi f(x|\lambda_1) + (1-\pi) f(x|\lambda_2)$, where f denotes a Poisson p.m.f. and $\theta_2 = (\pi, \lambda_1, \lambda_2)$ and showed via simulations that the performance of their MHD estimator of θ_2 remains unaffected even when the data are generated from a $100\alpha\%$ gross-error contaminated Poisson mixture model defined by

$$f_{\theta_{2,\epsilon,\lambda_3}}(x) = (1-\epsilon) f_{\theta_2}(x) + \epsilon f(x|\lambda_3), \tag{2.5.13}$$

Table 2.2: Relative frequencies of estimated number of components based on 500 replications [Target mixture: 2-component Poisson].

$\boldsymbol{\theta}_2 = (0.5, 1, 9)$																
Sample Size	$n = 50$				$n = 100$				$n = 500$				$n = 1000$			
α_n	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
$2/n$	0	***1.00**			0	***.998**	.002		0	***1.00**			0	***1.00**		
$ln(n)/n$	0	***1.00**			0	***1.00**			0	***1.00**			0	***1.00**		
LRT	0	***.95**	.05		0	***.95**	.05		0	***.96**	.04					
$\boldsymbol{\theta}_2 = (0.8, 1, 9)$																
Sample Size	$n = 50$				$n = 100$				$n = 500$				$n = 1000$			
α_n	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
$2/n$	.302	***.698**			0	***.998**	.002		0	***1.00**			0	***1.00**		
$ln(n)/n$	.818	.182			.002	***.998**			0	***1.00**			0	***1.00**		
LRT	0	***.92**	.08		0	***.95**	.05		0	***.96**	.04					
$\boldsymbol{\theta}_2 = (0.5, 1, 1.1)$																
Sample Size	$n = 50$				$n = 100$				$n = 500$				$n = 1000$			
α_n	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
$2/n$	1.00				.998	.002			1.00				1.00			
$ln(n)/n$	1.00				1.00				1.00				1.00			
LRT	.96	.04			.93	.07			.94	.05	.01					
$\boldsymbol{\theta}_2 = (0.95, 1, 10)$																
Sample Size	$n = 50$				$n = 100$				$n = 500$				$n = 1000$			
α_n	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
$2/n$	.992	.008			.616	.384			0	***1.00**			0	***1.00**		
$ln(n)/n$	1.00				.946	.054			.006	***.994**			0	***1.00**		
LRT	.11	***.83**	.06		0	***.93**	.07		0	***.95**	.05					

Table 2.3: Relative frequencies of estimated number of components based on 500 replications [Target mixture: 3-component Poisson].

$\boldsymbol{\theta}_3 = (0.45, 0.45, 1, 5, 10)$																
Sample Size	$n = 50$				$n = 100$				$n = 500$				$n = 1000$			
α_n	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
$2/n$	.108	.89	.002		0	.966	.034		0	.162	***.838**		0	.002	***.998**	
$ln(n)/n$	.606	.394			0	1.00			0	.846	.154		0	.26	***.74**	
LRT	0	.62	.36	.01	0	.39	***.58**	.02	0	0	***.94**	.06				
$\boldsymbol{\theta}_3 = (0.4, 0.4, 1, 3, 3.1)$																
Sample Size	$n = 50$				$n = 100$				$n = 500$				$n = 1000$			
α_n	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
$2/n$	.988	.012			.61	.39			0	1.00			0	.996	.004	
$ln(n)/n$	1.00				.95	.05			.138	.862			.002	.998		
LRT	.42	.56	.01		.14	.82	.03		0	.96	.04					
$\boldsymbol{\theta}_3 = (0.33, 0.33, 1, 5, 10)$																
Sample Size	$n = 50$				$n = 100$				$n = 500$				$n = 1000$			
α_n	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
$2/n$	.004	.994	.002		0	.84	.16		0	.018	***.982**		0	0	***1.00**	
$ln(n)/n$	.17	.83			0	.988	.012		0	.462	***.538**		0	.026	***.974**	
LRT	0	.54	.44	.01	0	.30	***.66**	.03	0	0	***.94**	.06				

Table 2.4: Relative frequencies of estimated number of components based on 500 replications [Target mixture: 4-component Poisson].

$\boldsymbol{\theta}_4 = (0.3, 0.4, 0.25, 1, 5, 9, 15)$																	
Sample Size	$n = 50$				$n = 100$				$n = 500$					$n = 1000$			
α_n	1	2	3	4	1	2	3	4	1	2	3	4	5	1	2	3	4
$2/n$	.066	.934			0	.718	.282		0	0	.956	.044		0	0	.866	.134
$ln(n)/n$	.466	.534			0	.962	.038		0	.06	.94			0	0	.998	.002
LRT	0	.31	.61	.08	0	.09	.78	.13	0	0	.59	.38	.03				
$\boldsymbol{\theta}_4 = (0.3, 0.3, 0.2, 1, 1.2, 5, 9)$																	
Sample Size	$n = 50$				$n = 100$				$n = 500$					$n = 1000$			
α_n	1	2	3	4	1	2	3	4	1	2	3	4	5	1	2	3	4
$2/n$	.052	.948			0	.998	.002		0	.614	.386			0	.18	.818	.002
$ln(n)/n$	.386	.614			0	1.00			0	.990	.010			0	.922	.078	
LRT	0	.78	.21	.01	0	.68	.31	.01	0	.17	.78	.03	.02				
$\boldsymbol{\theta}_4 = (0.25, 0.25, 0.25, 1, 5, 10, 15)$																	
Sample Size	$n = 50$				$n = 100$				$n = 500$					$n = 1000$			
α_n	1	2	3	4	1	2	3	4	1	2	3	4	5	1	2	3	4
$2/n$	.002	.944	.054		0	.182	.812	.006	0	0	.924	.076		0	0	.794	.206
$ln(n)/n$	.046	.954			0	.718	.282		0	0	1.00			0	0	.998	.002
LRT	0	.17	.76	.07	0	.02	.86	.12	0	0	.59	.40	.01				

where ϵ is the proportion of contamination and the (contaminating) value λ_3 is large compared to those of ϵ_1 and ϵ_2.

While it is a common practice to study robustness of MHD estimators against gross-error contamination models, such an approach would be inappropriate in our context because, by virtue of its consistency, our estimator $\hat{m}_n$ would (correctly) identify (for sufficiently large n) the number of components in the mixture from which data are generated, which in the above setup would be 3 instead of 2. Notice also that there is no Hellinger function representation of our estimator of mixture complexity which would facilitate the study of α-influence functions.

In view of these, we assess the robustness of $\hat{m}_n$ when the postulated model is a Poisson mixture but the data are generated from a Negative Binomial mixture. More precisely, to assess the robustness, we perform extensive simulation studies when the postulated model is a 2-component Poisson mixture $f_{\theta_2}(x)$ defined above with λ_1 and λ_2 as its component means, but the data are generated from a 2-component Negative Binomial mixture given by

$$f(x) = \pi f_1(x) + (1 - \pi) f_2(x), \tag{2.5.14}$$

where, for $i = 1, 2$, $f_i(x) = \binom{r+x-1}{x} p_i^r (1-p_i)^x$, $x = 0, 1, \ldots$. Let f_1 and f_2 be the p.m.f.s associated with random variables, say, X_1 and X_2, respectively. Then, it can be easily shown that the component means and the variances are $E(x_i) = r(1-p_i)/p_i$ and $Var(X_i) = r(1-p_i)/p_i^2$, for $i = 1, 2$. It is also well known that if, for each $i = 1, 2$, $r \to \infty$ and $p_i \to 1$ such that $r(1-p_i) \to \lambda_i$, then $E(X_i) \to \lambda_i$ and $Var(X_i) \to \lambda_i$, which agrees with the component Poisson means and variances. In fact, under these conditions, it can be shown that the negative binomial family of distributions includes the Poisson distribution as a limiting case.

The hallmark of the (postulated) Poisson distribution is that the mean is equal to the variance. However, the component variance of a negative binomial mixture from which the count data is generated may be more than can be expected on the basis of the postulated model. This phenomenon, known as overdispersion, in the negative binomial components in 2.5.14 may also be moderate to extreme depending on the values of r and p_i, for $i = 1, 2$.

In our simulation studies, we consider two scenarios. In both the scenarios, we set the component mean of the sampling model to be the same as that of the postulated model, that is, $r(1-p_i)/p_i = \lambda_i$ for $i = 1, 2$. In the first scenario, we set $r = 10$ and $\lambda_1 = 1$, but vary the values of λ_2 from 2 to 7 with a unit increment. The values of p_i, $i = 1, 2$ can be obtained from the equation $p_i = r/(\lambda_i + r)$ for $i = 1, 2$. This setting yields $E(X_1) = 1$ and $Var(X_1) = 1.1$ and the values of $E(X_2) = \lambda_2$ and $Var(X_2)$ are listed in the following table.

λ_2	2	3	4	5	6	7
$Var(X_2)$	2.4	3.9	5.6	7.5	9.6	11.9

Notice that the $Var(X_2)$ values in the above table are progressively much larger compared to the corresponding values of $E(X_2)(= \lambda_2)$, thus creating a moderate to more extreme overdispersion in the second negative binomial component.

In the second scenario, we set $\lambda_1 = 1$ and $\lambda_2 = 10$ (these set $E(X_1) = 1$ and $E(X_2) = 10$), but vary the values of r from 10 to 45 with an increment of 5. The following table lists the values of $Var(X_1)$ and $Var(X_2)$:

r	10	15	20	25	30	35	40	45
$Var(X_1)$	1.1	1.066	1.050	1.040	1.029	1.025	1.022	
$Var(X_2)$	20	16.667	15	14	13.333	12.857	12.222	

Note that, as the values of r decrease, the values of $Var(X_1)$ stay close to $E(X_1) = 1$, but the values of $Var(X_2)$ become much larger compared to $E(X_2) = 10$, thus one again creating a moderate to more extreme overdispersion in the second negative binomial component. Finally, in each of these two scenarios, we set $\pi = 0.25, 0.5, 0.75$ in 2.5.14.

For each of the above set of parameter values in each scenario, count data is generated from the appropriate negative binomial mixture in 2.5.14, but the computational algorithm described in section 2.4 is implemented under the assumption that the class $\mathcal{F}_m$ defined in section 2.2 is a family of Poisson mixtures. Here, we perform simulation studies for three sample sizes $n = 100, 500, 1000$ using the two threshold values $\alpha_{n,m} = 2/n$ and $ln(n)/n$. As before, for each sample size and threshold value, we performed 500 Monte Carlo replications of our algorithm, each yielding an estimate of mixture complexity. Tables 2.5 to 2.7 below give the relative frequencies (out of 500 replications) of the number of components determined by our method for the first scenario where $r = 10$ and $\lambda_1 = 1$ but λ_2 values vary, and Tables 2.8 to 2.10 give similar results for the second scenario, where $\lambda_1 = 1$ and $\lambda_2 = 10$ but the values of r vary. In all these tables, the percentage (50% or above) of correct identifications is given in bold with an asterisk beside it.

The simulations carried out here span over a variety of moderate to more extreme departures from component Poisson assumption along with two different scenarios of overdispersion in the second component. In all, the results in Tables 2.5 to 2.10 cover 126 different cases of model misspecification, which span over small to large sample sizes with moderate to extreme overdispersion. Of these 126 cases, Tables 2.5 to 2.10 show that our procedure based

on the SBC threshold value $\alpha_{n,m} = ln(n)/n$ correctly determines the number of components in 104 cases and our procedure based on the AIC threshold value $\alpha_{n,m} = 2/n$ correctly determines the number of components in 86 cases.

More specifically, Tables 2.5 to 2.7 show that, for some small sample sizes and small values of λ_2, our procedure based on the SBC threshold value $\alpha_{n,m} = ln(n)/n$ underestimates the true mixture complexity ($m_0 = 2$) but it correctly identifies the true mixture complexity a large majority of times for small and large samples, even as the overdispersion in the second component in 2.5.14 increases. These tables show that the situation is somewhat reversed for our procedure based on the AIC threshold value $\alpha_{n,m} = 2/n$ in that it overestimates the true mixture complexity in (some) instances where the sample size is large and the overdispersion is also large.

Tables 2.8 to 2.10 show that our procedure based on the SBC and the AIC threshold values perform well when $n = 100$ and for all values of r, including the smaller sample size $n = 500$, Table 2.10 shows that our procedure based on the SBC and AIC continue to correctly identify even for small values of r, whereas Tables 2.8 and 2.9 show that our procedure based on the SBC performs better than the one based on the AIC. Overall, given the extreme nature of departure from component Poisson assumption, the results in Tables 2.5 to 2.10 serve as a testament that our MHD based estimate of the mixture complexity is highly robust under model misspecification.

For the model misspecification setup described in this section, it is possible to calculate the percentage of correct identification of mixture complexity using the sequential testing procedure of Karlis and Xekalaki (1999) based on LRT that utilizes resampling techniques. However, as observed in chapter 1 for the continuous case, we believe that the mixture complexity estimate based on the sequential LRT of Karlis and Xekalaki (1999) will be highly unstable, and this instability may be exasperated when the overdispersion in the second negative binomial component becomes more severe.

Table 2.5: Relative frequencies of estimated number of components based on 500 replications. Sample from 2 component Negative Binomial mixture with $\lambda_1 = 1$,$\pi = .25$, and $r = 10$.

						$\lambda_2 = 2$						
Sample Size		$n = 100$				$n = 500$				$n = 1000$		
α_n	1	2	3	4	1	2	3	4	1	2	3	4
$2/n$	.908	.092			.218	***.782**			.008	***.992**		
$ln(n)/n$	.998	.002			.888	.112			.474	***.526**		
						$\lambda_2 = 3$						
Sample Size		$n = 100$				$n = 500$				$n = 1000$		
α_n	1	2	3	4	1	2	3	4	1	2	3	4
$2/n$	.456	***.546**			0	***1.00**			0	***1.00**		
$ln(n)/n$	.91	.09			.02	***.98**			0	***1.00**		
						$\lambda_2 = 4$						
Sample Size		$n = 100$				$n = 500$				$n = 1000$		
α_n	1	2	3	4	1	2	3	4	1	2	3	4
$2/n$	.066	***.93**	.004		0	***.958**	.042		0	***.856**	.144	
$ln(n)/n$	.45	***.55**			0	***1.00**			0	***.998**	.002	
						$\lambda_2 = 5$						
Sample Size		$n = 100$				$n = 500$				$n = 1000$		
α_n	1	2	3	4	1	2	3	4	1	2	3	4
$2/n$	.006	***.978**	.016		0	***.724**	.276		0	.252	.748	
$ln(n)/n$	.078	***.922**			0	***.996**	.004		0	***.948**	.052	
						$\lambda_2 = 6$						
Sample Size		$n = 100$				$n = 500$				$n = 1000$		
α_n	1	2	3	4	1	2	3	4	1	2	3	4
$2/n$	0	***.956**	.044		0	.292	.708		0	.022	.978	
$ln(n)/n$	.006	***.994**			0	***.936**	.064		0	***.576**	.424	
						$\lambda_2 = 7$						
Sample Size		$n = 100$				$n = 500$				$n = 1000$		
α_n	1	2	3	4	1	2	3	4	1	2	3	4
$2/n$	0	***.916**	.084		0	.068	.932		0	0	1.00	
$ln(n)/n$	.002	***.992**	.006		0	***.718**	.284		0	.112	.888	

Table 2.6: Relative frequencies of estimated number of components based on 500 replications. Sample from 2 component Negative Binomial mixture with $\lambda_1 = 1$,$\pi = .5$, and $r = 10$.

						$\lambda_2 = 2$						
Sample Size		$n = 100$				$n = 500$				$n = 1000$		
α_n	1	2	3	4	1	2	3	4	1	2	3	4
$2/n$	.882	.118			.14	***.86**			.002	***.998**		
$ln(n)/n$	.994	.006			.802	.198			.276	***.724**		
						$\lambda_2 = 3$						
Sample Size		$n = 100$				$n = 500$				$n = 1000$		
α_n	1	2	3	4	1	2	3	4	1	2	3	4
$2/n$	.264	***.736**			0	***.996**	.004		0	***.998**	.002	
$ln(n)/n$	.784	.216			0	***1.00**			0	***1.00**		
						$\lambda_2 = 4$						
Sample Size		$n = 100$				$n = 500$				$n = 1000$		
α_n	1	2	3	4	1	2	3	4	1	2	3	4
$2/n$	.016	***.98**	.004		0	***.956**	.044		0	***.862**	.138	
$ln(n)/n$	.182	***.818**			0	***1.00**			0	***1.00**		
						$\lambda_2 = 5$						
Sample Size		$n = 100$				$n = 500$				$n = 1000$		
α_n	1	2	3	4	1	2	3	4	1	2	3	4
$2/n$	0	***.992**	.008		0	***.808**	.192		0	.40	.60	
$ln(n)/n$	.006	***.992**	.002		0	***1.00**			0	***.984**	.016	
						$\lambda_2 = 6$						
Sample Size		$n = 100$				$n = 500$				$n = 1000$		
α_n	1	2	3	4	1	2	3	4	1	2	3	4
$2/n$	0	***.974**	.026		0	***.52**	.48		0	.094	.906	
$ln(n)/n$	0	***1.00**			0	***.982**	.018		0	***.842**	.158	
						$\lambda_2 = 7$						
Sample Size		$n = 100$				$n = 500$				$n = 1000$		
α_n	1	2	3	4	1	2	3	4	1	2	3	4
$2/n$	0	***.946**	.054		0	.234	.766		0	.006	.994	
$ln(n)/n$	0	***.998**	.002		0	***.914**	.086		0	.474	.526	

Table 2.7: Relative frequencies of estimated number of components based on 500 replications. Sample from 2 component Negative Binomial mixture with $\lambda_1 = 1$,$\pi = .75$, and $r = 10$.

						$\lambda_2 = 2$						
Sample Size		$n = 100$				$n = 500$				$n = 1000$		
α_n	1	2	3	4	1	2	3	4	1	2	3	4
$2/n$	.936	.064			.286	***.714**			.02	***.98**		
$ln(n)/n$	1.00				.94	.06			.604	.396		
						$\lambda_2 = 3$						
Sample Size		$n = 100$				$n = 500$				$n = 1000$		
α_n	1	2	3	4	1	2	3	4	1	2	3	4
$2/n$	.486	***.514**			0	***1.00**			0	***.998**	.002	
$ln(n)/n$	.932	.068			.022	***.978**			0	***1.00**		
						$\lambda_2 = 4$						
Sample Size		$n = 100$				$n = 500$				$n = 1000$		
α_n	1	2	3	4	1	2	3	4	1	2	3	4
$2/n$	.098	***.898**	.004		0	***.98**	.02		0	***.926**	.074	
$ln(n)/n$	.508	.492			0	***1.00**			0	***1.00**		
						$\lambda_2 = 5$						
Sample Size		$n = 100$				$n = 500$				$n = 1000$		
α_n	1	2	3	4	1	2	3	4	1	2	3	4
$2/n$	.014	***.98**	.006		0	***.934**	.066		0	***.75**	.25	
$ln(n)/n$	.14	***.86**			0	***1.00**			0	***1.00**		
						$\lambda_2 = 6$						
Sample Size		$n = 100$				$n = 500$				$n = 1000$		
α_n	1	2	3	4	1	2	3	4	1	2	3	4
$2/n$	0	***.988**	.012		0	***.82**	.18		0	.474	.526	
$ln(n)/n$	.018	***.982**			0	***.998**	.002		0	***.984**	.016	
						$\lambda_2 = 7$						
Sample Size		$n = 100$				$n = 500$				$n = 1000$		
α_n	1	2	3	4	1	2	3	4	1	2	3	4
$2/n$	0	***.974**	.026		0	***.624**	.376		0	.216	.784	
$ln(n)/n$	.002	***.998**			0	***.994**	.006		0	***.948**	.052	

Table 2.8: Relative frequencies of estimated number of components based on 500 replications. Sample from 2 component Negative Binomial mixture with $\lambda_1 = 1$,$\lambda_2 = 10$, and $\pi = .25$

							$n = 100$									
n		$r = 10$				$r = 15$				$r = 20$				$r = 25$		
α_n	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
$2/n$	.002	***.594**	.404		0	***.80**	.20		0	***.92**	.08		0	***.942**	.058	
$ln(n)/n$	.002	***.942**	.056		0	***.986**	.014		0	***.996**	.004		0	***.998**	.002	
n		$r = 30$				$r = 35$				$r = 40$				$r = 45$		
α_n	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
$2/n$	0	***.978**	.022		0	***.982**	.018		0	***.99**	.01		0	***.988**	.012	
$ln(n)/n$	0	***1.00**			0	***1.00**			0	***1.00**			0	***1.00**		
							$n = 500$									
r		$r = 10$				$r = 15$				$r = 20$				$r = 25$		
α_n	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
$2/n$	0	0	.984	.016	0	.018	.982		0	.124	.876		0	.296	.704	
$ln(n)/n$	0	.022	.978		0	.382	.618		0	***.782**	.218		0	***.946**	.054	
r		$r = 30$				$r = 35$				$r = 40$				$r = 45$		
α_n	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
$2/n$	0	.476	.524		0	***.67**	.33		0	***.766**	.234		0	***.82**	.18	
$ln(n)/n$	0	***.96**	.04		0	***.994**	.006		0	***.998**	.002		0	***.998**	.002	
							$n = 1000$									
r		$r = 10$				$r = 15$				$r = 20$				$r = 25$		
α_n	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
$2/n$	0	0	.926	.074	0	0	1.00		0	.004	.996		0	.01	.99	
$ln(n)/n$	0	0	1.00		0	.01	.99		0	.192	.808		0	***.542**	.458	
r		$r = 30$				$r = 35$				$r = 40$				$r = 45$		
α_n	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
$2/n$	0	.07	.93		0	.194	.806		0	.282	.716	.002	0	.448	.552	
$ln(n)/n$	0	***.776**	.224		0	***.93**	.07		0	***.966**	.034		0	***.992**	.008	

Table 2.9: Relative frequencies of estimated number of components based on 500 replications. Sample from 2 component Negative Binomial mixture with $\lambda_1 = 1$,$\lambda_2 = 10$, and $\pi = .5$

							$n = 100$									
r		$r = 10$				$r = 15$				$r = 20$				$r = 25$		
α_n	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
$2/n$	0	***.76**	.236	.004	0	***.888**	.112		0	***.94**	.06		0	***.984**	.016	
$ln(n)/n$	0	***.984**	.016		0	***.996**	.004		0	***.998**	.002		0	***1.00**		
r		$r = 30$				$r = 35$				$r = 40$				$r = 45$		
α_n	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
$2/n$	0	***.982**	.018		0	***.992**	.008		0	***.99**	.01		0	***.99**	.01	
$ln(n)/n$	0	***1.00**			0	***1.00**			0	***1.00**			0	***.998**	.002	
							$n = 500$									
r		$r = 10$				$r = 15$				$r = 20$				$r = 25$		
α_n	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
$2/n$	0	.006	.984	.01	0	.128	.872		0	.386	.614		0	***.59**	.41	
$ln(n)/n$	0	.252	.748		0	***.788**	.212		0	***.942**	.058		0	***.986**	.014	
r		$r = 30$				$r = 35$				$r = 40$				$r = 45$		
α_n	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
$2/n$	0	***.71**	.29		0	***.824**	.176		0	***.896**	.104		0	***.91**	.09	
$ln(n)/n$	0	***.992**	.008		0	***1.00**			0	***1.00**			0	***.998**	.001	
							$n = 1000$									
r		$r = 10$				$r = 15$				$r = 20$				$r = 25$		
α_n	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
$2/n$	0	0	.962	.038	0	0	.996	.004	0	.04	.96		0	.126	.872	.002
$ln(n)/n$	0	0	1.00		0	.182	.818		0	***.622**	.378		0	***.874**	.126	
r		$r = 30$				$r = 35$				$r = 40$				$r = 45$		
α_n	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
$2/n$	0	.292	.706	.002	0	.488	.512		0	***.608**	.392		0	***.706**	.294	
$ln(n)/n$	0	***.97**	.03		0	***.988**	.012		0	***.994**	.012		0	***.996**	.004	

Table 2.10: Relative frequencies of estimated number of components based on 500 replications. Sample from 2 component Negative Binomial mixture with $\lambda_1 = 1$,$\lambda_2 = 10$, and $\pi = .75$

							$n = 100$									
r		$r = 10$				$r = 15$				$r = 20$				$r = 25$		
α_n	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
$2/n$	0	***.918**	.082		0	***.978**	.022		0	***.984**	.016		0	***.994**	.006	
$ln(n)/n$	0	***.998**	.002		.002	***.996**	.002		.002	***.998**			0	***1.00**		
r		$r = 30$				$r = 35$				$r = 40$				$r = 45$		
α_n	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
$2/n$	0	***.994**	.006		0	***1.00**			0	***.99**	.01		0	***.996**	.004	
$ln(n)/n$	0	***1.00**			0	***1.00**			0	***1.00**			0	***1.00**		
							$n = 500$									
r		$r = 10$				$r = 15$				$r = 20$				$r = 25$		
α_n	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
$2/n$	0	.212	.782	.006	0	***.57**	.43		0	***.762**	.238		0	***.864**	.136	
$ln(n)/n$	0	***.89**	.11		0	***.984**	.016		0	***.996**	.004		0	***1.00**		
r		$r = 30$				$r = 35$				$r = 40$				$r = 45$		
α_n	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
$2/n$	0	***.902**	.098		0	***.946**	.054		0	***.97**	.03		0	***.972**	.028	
$ln(n)/n$	0	***1.00**			0	***1.00**			0	***1.00**			0	***1.00**		
							$n = 1000$									
r		$r = 10$				$r = 15$				$r = 20$				$r = 25$		
α_n	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
$2/n$	.008	.006	.962	.024	.002	.10	.894	.004	0	.366	.632	.002	0	***.59**	.41	
$ln(n)/n$	.008	.376	.616		.002	***.848**	.15		0	***.982**	.018		0	***.99**	.01	
r		$r = 30$				$r = 35$				$r = 40$				$r = 45$		
α_n	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
$2/n$	0	***.74**	.26		.002	***.814**	.184		0	***.896**	.104		0	***.916**	.084	
$ln(n)/n$	0	***1.00**			.002	***.996**	.002		0	***1.00**			0	***1.00**		

2.6 THREE EXAMPLES

Here, we consider three overdispersed count datasets which have been modeled using Poisson mixtures in Karlis and Xekalaki (1998, 1999 and 2001), respectively. For these three data sets, we first estimate the number of components using our MHD based estimator $\hat{m}_n$ in (2.2.9) for the SBC and the AIC thresholds. When our MHD based computational algorithm stops and reports a value $\hat{m}_n = k$, say, it automatically provides the MHD estimate of all the parameters in the k-component mixture. In all our examples, we use these MHD estimates to obtain the best fitting k-component mixture.

The first example concerns the number of accidents incurred by 414 machinists over a period of three months. This count data (see Table 2.13 below) is taken from the classical

paper of Greenwood and Yule (1920) and has been analyzed by several authors including Karlis and Xekalaki (1999). Greenwood and Yule noted that the fit provided by single Poisson distribution to this data is very poor. Using a sequential testing procedure based on likelihood ratio test (LRT) that utilizes a resampling approach, Karlis and Xekalaki (1999) determined that a 3-component Poisson mixture provides a better fit to the data. Observe from Table 2.13 below that this data contains excessive number of zeros, indicating that a (Poisson) mixture model that simultaneously addresses the excess zeros and overdispersion, referred here as a zero-inflated Poisson (ZIP) mixture model (see definition below), may also be appropriate for this data. We briefly discuss the estimation of mixture complexity for zero-inflated (finite) Poisson mixtures below and then fit an appropriate ZIP mixture to this data.

The second example concerns the number of environmental complaints placed by phone in an environmental station in Netherlands for the year 1985. A simple Poisson model is clearly inappropriate for this data (see Table 3 in Karlis and Xekalaki (1998)), since the mean is 22.11 while the variance is 324.08, which is about 15 times greater than the mean. Moreover, the data is highly skewed with a very long tail. Karlis and Xekalaki (1998) analyzed this data and illustrated the superiority of the MHD method of estimation over the ML estimation method. Specifically, for this data they fitted a 3-component Poisson mixture using MHD estimates as well as ML estimates and showed that the fit based on MHD estimates, compared to the one based on ML estimates, was much less influenced by relatively few extreme observations. More precisely, they showed that the fit based on MHD has a smoother right tail whereas the fit based on MLE has a bump, indicating a stronger influence of few high values on the ML method.

The third example concerns the number of defaulted installments in a financial institution in Spain (see Table 2.18 below), a data originally considered in Dionne, Artis and Guillen (1996). Due to the presence of overdispersion in the data, Karlis and Xekalaki (2001) concluded that a Poisson mixture would be plausible for modeling this data. Based on plots of

Hellinger gradient function for different values of mixture complexity, Karlis and Xekalaki (2001) concluded that a semiparametric MHD estimate of the mixing distribution supports a 6-component Poisson mixture model for the data. Their plots (see Figure 1 of Karlis and Xekalaki (2001)) also show that their MHD estimate of the mixing distribution cannot support a Poisson mixture with more than 6 components. In addition to overdispersion, once again, observe from Table 2.18 below that this data contains excessive number of zeros, indicating that a zero-inflated Poisson mixture (ZIP mixture) model may also be appropriate for the data. We discuss fitting an appropriate ZIP mixture to this data as well.

For the count data on the number of accidents incurred by 414 machinists, we determined an estimate of the number components m assuming two slightly different finite mixture models; namely, (1) the usual m-component Poisson mixture model with means $0 \leq \lambda_1 < .... < \lambda_m$ and (2) a m-component ZIP mixture model defined by $f_{\tilde{\boldsymbol{\theta}}_m}(x) = \pi_1 \delta_{\{0\}}(x) + \sum_{i=2}^{m} \pi_i f(x|\lambda_i)$, where π_1 is the unknown proportion of zero count, $\delta_{\{0\}}$ is a point mass at 0, $f(x|\lambda_i)$ is a Poisson p.m.f. with mean λ_i satisfying $0 \leq \lambda_2 < .... < \lambda_m$, $\pi_i \geq 0$ for $i = 1, \ldots, m$, and $\sum_{i=1}^{m} \pi_i = 1$. Note that the ZIP mixture models belong to the general family $\tilde{\mathcal{F}}_m = \{f_{\tilde{\boldsymbol{\theta}}_m} : \tilde{\boldsymbol{\theta}}_m \in \tilde{\Theta}_m \subseteq R^d\}$ such that $f_{\tilde{\boldsymbol{\theta}}_m}(x) = \pi_1 \delta_{\{0\}}(x) + \sum_{i=2}^{m} \pi_i f(x|\phi_i)$, where the m component p.m.f.s are not necessarily Poisson and ϕ_i's are as in (2.2.1). As in section 2.2, for an arbitrary p.m.f. $f \in \Gamma$, we can once again define the *index of the economical representation* of f, relative to the family of mixtures $\tilde{\mathcal{F}}_m$ and also modify those in equations (2.2.4) to (2.2.9) for the family $\tilde{\mathcal{F}}_m$. For the sake of clarity, we will denote the estimator of mixture complexity for the ZIP mixtures as $\tilde{m}_n$. Note that the conclusions of our main theorem would still hold for $\tilde{m}_n$. For computations in the case of ZIP mixtures, we will also appropriately modify the details given in section 2.4.

For model (1), our analysis using the MHD computational algorithm yielded an estimate of mixture complexity $\hat{m}_n = 2$ for the thresholds $\alpha_{n,m} = 2/n$ and $ln(n)/n$. For the ZIP mixture model (2), our analysis using MHD computational algorithm yielded an estimate of $\tilde{m}_n = 2$ for the threshold $\alpha_{n,m} = ln(n)/n$, while it yielded an estimate of $\tilde{m}_n = 3$ for

the thresholds $\alpha_{n,m} = 2/n$. We used the MHD estimate of the parameters automatically provided at the termination of our algorithm and obtained the best fitting 2-component Poisson mixture and the 3-component ZIP mixture. For comparison purposes, we also used the HELMIX algorithm and computed the MHD estimate of the parameters in a 3-component Poisson mixture, which Karlis and Xekalaki (1999) determined to be the best fit. In Table 2.11 below, we give the MHD estimates of the parameters corresponding to the 2- and 3-component Poisson mixtures along with the values of $H_m^2 = H^2(\hat{f}_n, \hat{f}^m)$, which can be used to assess each of these fitted Poisson mixtures. In Table 2.12 we give the MHD estimate of the parameters and the H_m^2 values corresponding to the best fitting 3-component ZIP mixture. Finally, in Table 2.13, we compare the observed frequencies with the expected frequencies provided by each of the fitted Poisson mixtures given in Tables 2.11 and 2.12.

Notice from Tables 2.11 and 2.12 that the 3-component ZIP mixture fit is essentially the same as the 3-component Poisson mixture fit. We conclude from the H_m^2 values in Tables 2.11 and 2.12, and the nature of fit given in Table 2.13 that our 3-component ZIP mixture fit (suggested by our AIC threshold) and the 3-component Poisson mixture fit based on MHD estimates provide the best fit to the data. However, from the point of view of slight parsimony (because λ_1 is set to 0 in the 3-component ZIP mixture), we would prefer the 3-component ZIP mixture fit (based on MHD estimates) for the data. We also computed expected frequencies based on a 3-component ZIP mixture using ML estimates and found that they also provided a very good fit to the data.

Table 2.11: MHD estimate of parameters in 2- and 3-component Poisson mixtures for data on the number of accidents incurred by 414 machinists over a period of three months.

m	H_m^2	π_1	π_2	π_3	λ_1	λ_2	λ_3
2	.00396	.8796	.1204		.22749	2.1859	
3	.00283	.42072	.52822	.05105	.0000111	.58567	3.0424

Table 2.12: MHD estimate of parameters in 3-component ZIP mixture for number of accidents data.

m	H_m^2	π_1	π_2	π_3	λ_1	λ_2	λ_3
3	.00284	.42335	.52580	.05084	0	.5896	3.0449

Table 2.13: Observed and expected frequencies (based on MHD estimates) of fitted 2- and 3-component Poisson mixtures, and 3-component ZIP mixture for number of accidents data.

X	0	1	2	3	4	≥ 5
$Frequency$	296	74	26	8	4	6
$m = 2$	295.66	78.23	20.89	10.32	5.36	3.54
$m = 3$	296.93	74.38	25.55	8.81	4.20	4.13
$ZIP(m = 3)$	297.02	74.20	25.62	8.84	4.19	4.13

The second is Karlis and Xekalaki(1998)'s example concerning the number of environmental complaints placed by phone in an environmental station in Netherlands for the year 1985, and it is summarized in Table 2.14 below. For this data, our analysis using MHD algorithm yielded an estimate of mixture complexity $\hat{m}_n = 4$ for the AIC and the SBC thresholds. Note that Karlis and Xekalaki (1998) fitted a 3-component Poisson mixture to this data, whereas our algorithm points to a 4-component Poisson mixture. It should be mentioned here that Karlis and Xekalaki (1998)'s decision to fit a 3-component Poisson mixture model for this data was not based on any prior testing procedure to determine the mixture complexity for this data, as done in Karlis and Xekalaki (1999).

In Table 2.15 below, we give the MHD estimate of the parameters corresponding to our 4-component Poisson mixture along with the H_4^2 value, and the the MHD estimates corresponding to the 3-component Poisson mixture from Table 4 of Karlis and Xekalaki (1998) along with the H_3^2 value. It is possible to compare our 4-component Poisson mixture

fit with the 3-component Poisson mixture fit of Karlis and Xekalaki (1998) in terms of observed frequencies and the expected frequencies, as done in Table 3 of Karlis and Xekalaki (1998). Instead, we compare the plot of the two fitted mixtures with the histogram of observed frequencies for this data in Figure 2.1 below (also see Figure 5 in Karlis and Xekalaki (1998)).

We conclude from the nature of fit in Figure 2.1 and the H_m^2 values in Table 2.15 that our MHD based 4-component Poisson mixture provides a better fit than the 3-component Poisson fit in Karlis and Xekalaki (1998). As noted in Karlis and Xekalaki (1998), we also found that the 4-component Poisson mixture fit based on MHD estimates largely ignores the extreme values, while the 4-component Poisson mixture fit based on MLE is very much influenced by extreme values.

Table 2.14: The number of environmental complaints placed in an environmental station in 1985

x	Frequency	x	Frequency	x	Frequency
0-4	37	30-34	21	60-64	7
5-9	67	35-39	13	65-69	2
10-14	69	40-44	13	70-79	3
15-19	56	45-49	13	80-89	1
20-24	28	50-54	3	90-99	2
25-29	23	55-59	3	≥ 100	6

Table 2.15: MHD estimates of parameters in 3-component Poisson mixture taken from Table 4 of Karlis and Xekalaki (1998) and MHD estimates of parameters in our 4-component Poisson mixture for the number of environmental complaints data.

m	H_m^2	π_1	π_2	π_3	π_4	λ_1	λ_2	λ_3	λ_4
3	.13602	.39	.418	.197		7.136	17.331	37.676	
4	.10772	.2524	.4287	.2033	.1156	5.3405	13.4953	26.0726	43.3879

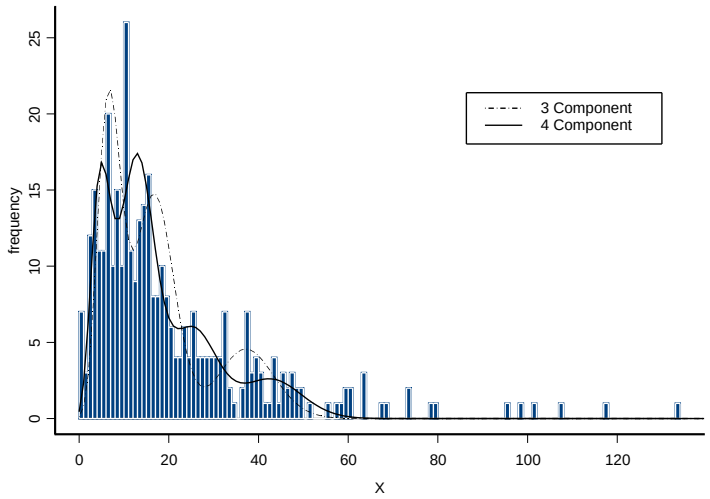

Figure 2.1: Histogram of the observed frequencies, the fitted 3- and 4-component Poisson mixtures for the number of environmental complaints data.

Finally, we revisit the data on the number of defaulted installments in a financial institution in Spain discussed in Karlis and Xekalaki (2001). Due to the presence of excessive number of zeros (see Table 2.18 below) and overdispersion, as in example 1 above, we determined an estimate of the number components m assuming a m-component Poisson mixture with means $0 \leq \lambda_1 < .... < \lambda_m$, as well as a m-component ZIP mixture defined above. For the m-component Poisson mixture model assumption, our analysis using the MHD computational algorithm yielded an estimate of mixture complexity $\hat{m}_n = 3, 4$ for thresholds $\alpha_{n,m} = ln(n)/n$ and $2/n$, respectively. Note that our answers differ drastically from Karlis and Xekalaki (2001)'s answer, which supports a 6-component Poisson mixture for this data.

In Table 2.16 below, we give the MHD estimates of parameters corresponding to 3-, 4-, 5- and 6-component Poisson mixtures along with the H_m^2 values. Although a 5-component mixture is not suggested either by our method or by Karlis and Xekalaki (2001), we fit it for the sake of completeness. In Table 2.16, note that the fit based on the 4-component Poisson

mixture has a tiny fourth (Poisson) component with mean 23.18354, the 5-component fit has tiny fourth and fifth (Poisson) components with means 14.31078 and 28.25839, respectively, and the 6-component fit has tiny fourth, fifth and sixth (Poisson) components with means 8.90081, 14.32837 and 28.2210445, respectively. Moreover, the 6-component Poisson mixture fit is not well separated since the 3^{rd} and 4^{th} estimated component means are almost the same. All these observations and the values of H_m^2 seem to suggest that the 3-component Poisson mixture determined by our SBC threshold provides a better fit to the data.

For the ZIP mixture model assumption, our analysis using the MHD computational algorithm yielded an estimate of $\tilde{m}_n = 4, 5$ for the thresholds $\alpha_{n,m} = ln(n)/n$ and $2/n$, respectively. Table 2.17 gives MHD estimates of parameters corresponding to 4-, 5- and 6-component ZIP mixtures along with the H_m^2 values. Once again, note in Table 2.17 that the fit based on the 5-component ZIP mixture has a tiny fifth (Poisson) component with mean 26.0526, and the 6-component ZIP mixture has tiny fifth and sixth (Poisson) components with means 14.4140 and 28.1969, respectively. Table 2.18 gives the observed frequencies and the expected frequencies (based on MHD estimates) corresponding to our fitted 4-, 5- and 6-component ZIP mixtures for this data. From the above observations and the H_m^2 values in Table 2.17, we suggest that a 4-component ZIP mixture determined by our SBC threshold also provides a good fit to the data. In view of the presence of high proportion of zeros and overdispersion, we conclude that the 4-component ZIP mixture provides a better fit for this data.

2.7 SUMMARY AND CONCLUSIONS

For count data, an information criterion based on minimum Hellinger distances is shown to naturally yield an estimator of the unknown number of components in finite mixtures, when the exact form of the component distributions are unknown but are postulated to be members of some parametric family. This estimator is consistent for parametric family of finite mixture models. The HELMIX numerical algorithm of Karlis and Xekalaki (1998) provides a useful

Table 2.16: MHD estimates of parameters in 3-, 4-, 5- and 6-component Poisson mixtures for the data on number of defaulted installments in a Spanish bank.

m	H_m^2	π_1	π_2	π_3	π_4	π_5	π_6
		λ_1	λ_2	λ_3	λ_4	λ_5	λ_6
3	.00434	.74422	.207518	.048256			
		.152897	4.26022	10.94475			
4	.00384	.741979	.204086	.052698	.0012357		
		.150406	4.15435	10.43217	23.18354		
5	.00349	.739226	.193037	.056636	.0105585	.00054105	
		.1473698	3.958507	8.893600	14.31078	28.25839	
6	.00349	.7391723	.193098	.0463737	.0102867	.0105206	.0005486
		.1473174	3.958116	8.893728	8.90081	14.32837	28.2210445

Table 2.17: MHD estimates of parameters in 4-, 5- and 6-component ZIP mixtures for the data on number of defaulted installments in a Spanish bank.

m	H_m^2	π_1	π_2	π_3	π_4	π_5	π_6
		λ_1	λ_2	λ_3	λ_4	λ_5	λ_6
4	.00412	.37260	.38536	.19851	.04352		
		0	.36241	4.51862	11.26306		
5	.00361	.32410	.42879	.19875	.04752	.00084	
		0	.30613	4.3824	10.8486	26.0526	
6	.00338	.29174	.45646	.19001	.05159	.00964	.00055
		0	.272787	4.1516	9.2107	14.4140	28.1969

Table 2.18: Comparison of observed frequencies and expected frequencies of fitted 4-component, 5-component and 6-component ZIP mixtures for defaulted installments in a Spanish bank.

x	Observed frequencies	Expected Frequencies MHDE			x	Observed frequencies	Expected Frequencies MHDE		
		$m=4$ ZIP	$m=5$ ZIP	$m=6$ ZIP			$m=4$ ZIP	$m=5$ ZIP	$m=6$ ZIP
0	3002	3016.21	3013.02	3012.68	18	8	3.484	3.023	3.689
1	502	501.894	504.492	503.127	19	6	2.065	1.798	2.584
2	187	186.455	181.522	182.579	20	3	1.163	1.070	1.783
3	138	166.742	171.423	175.990	21	0	0.624	0.669	1.216
4	233	179.042	182.085	181.329	22	1	0.319	0.469	0.828
5	160	163.424	162.387	157.702	23	0	0.156	0.379	0.574
6	107	127.480	124.431	120.595	24	1	.073	0.342	0.415
7	80	89.449	86.961	86.804	25	0	.033	0.326	0.322
8	59	60.605	59.923	62.936	26	0	.014	0.315	0.268
9	53	43.044	43.988	47.812	27	0	.006	0.298	0.237
10	41	33.659	35.344	37.829	28	1	.002	0.276	0.217
11	28	28.376	29.928	30.225	29	1	.00093	0.24769	0.20015
12	34	24.341	25.258	23.769	30	1	.00035	0.21422	0.18291
13	10	20.292	20.475	18.213	31	1	.00013	0.17994	0.16395
14	13	16.068	15.685	13.616	32	0	.00004	0.14646	0.14337
15	11	11.987	11.303	9.992	33	0	.00002	0.11562	0.12203
16	4	8.417	7.672	7.241	34	1	0	.08859	0.10100
17	5	5.570	4.929	5.196					

tool for the computation of our estimator, which is clearly an iterative procedure. Monte Carlo simulations for a wide variety of Poisson mixtures show that our estimator is able to correctly determine the number of components when the postulated mixture model is the same as the one from which samples are generated. These establish the efficiency of our estimation method, making it an attractive competitor to other existing methods in the literature. A distinguishing feature of our estimator is that it continues to correctly determine the number of components even when the mixture model from which samples are generated is a moderate to more extreme departure from the postulated mixture model. The basic construction, being firmly rooted in the minimum Hellinger distance approach, enables our estimator to naturally inherit the property of robustness under model misspecification without losing any efficiency when the model is correctly specified.

Simulations and data analysis carried out in this chapter involve implementation of our algorithm using two different threshold values, selected based on the well known Akaike and Schwarz information criterion. Numerical studies and data analysis presented here show that the estimates of mixture complexity provided by these two thresholds agree in some cases and do not agree in others. Overall, both the AIC and the SBC thresholds provide sensible and justifiable estimates of mixture complexity for the three overdispersed real datasets (two with zero-inflation) analyzed in this article, with the SBC threshold generally providing a more parsimonious fit. When the model is misspecified and/or when there is extreme overdispersion, we observed that the SBC threshold performs better than the AIC threshold. The fact that the value of our mixture complexity estimate depends on the choice of threshold values prompts us to look deeper into the issue of selection of other threshold values. More research is underway on this subject. In conclusion, it is shown here that our consistent and robust estimator of mixture complexity can effectively provide guidance in the search for the best mixture model for a given dataset.

2.8 REFERENCES

Beran, R. (1977), "Minimum Hellinger distance estimates for parametric models," *The Annals of Statistics*, 5, 445-463.

Beyers, R. H. and Shenton, L. R. (1999), "Sister chromatid exchange data fit with a mixture of Poisson distributions," *Mutation Research-Fundamental and Molecular Mechanisms of Mutagenesis*, 427, 157-162.

Böhning, D. (1999), *Computer-Assisted Analysis of Mixtures and Applications: Meta-Analysis, Disease Mapping and Others*, New York: Chapman & Hall/CRC.

Chen, J. and Kalbfleisch, J. D. (1996), "Penalized minimum distance estimates in finite mixture models," *Canadian Journal of Statistics*, 24, 167-175.

Dacunha-Castelle, D. and Gassiat, E. (1997), "The estimation of the order of a mixture model," *Bernoulli*, 3, 279-299.

Dacunha-Castelle, D. and Gassiat, E. (1999), "Testing the order of a model using locally conic parameterization: population mixtures and stationary ARMA processes," *The Annals of Statistics*, 27, 1178-1209.

Deb, P., and Trivedi, P. K. (1997), "Demand for medical care by the elderly: a finite mixture approach," *Journal of Applied Econometrics*, 12, 313-336.

Dellaportas, P., Karlis, D and Xekalaki, E. (1997), " Bayesian analysis of finite Poisson mixtures," Technical Report, Department of Statistics, Athens University of Economics and Business.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum-Likelihood From Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, Ser. B, 39, 1-38.

Devroye, L. P., and Györfi, L. (1985), *Nonparametric Density Estimation: The L_1 View*, New York: Wiley.

Dionne, G. Artis, M., and Guillen, M., (1996), " Count data models for a credit scoring system," *Journal of Empirical Finance*, 3, 303-325.

Escobar, M. D. and West, M. (1995), "Bayesian density estimation and inference using mixtures," *Journal of the American Statistical Association*, 90, 577-588.

Everitt, B. S. and Hand, D. J. (1981), *Finite Mixture Distributions*, London: Chapman and Hall.

Greenwood, M., and Yule, G., (1920), "An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents," *Journal of Royal Statistical Society, Ser. A* , 83, 255-279.

Henna, J. (1985), "On estimating of the number of constituents of a finite mixture of continuous distributions," *Annals of the Institute of Statistical Mathematics*, 37, 235-240.

Ishwaran, H., James, L. F., and Sun, J. (2001),"Bayesian Model Selection in Finite Mixtures by Marginal Density Decompositions," *Journal of the American Statistical Association*, 96, 1316-1332.

James, L. F., Priebe, C. E., and Marchette, D. J. (2001), "Consistent Estimation of Mixture Complexity," *The Annals of Statistics*, 29, 1281-1236.

Karlis, D. and Xekalaki, E. (1998), "Minimum Hellinger distance estimation for finite Poisson mixtures," *Computational Statistics and Data Analysis*, 29, 81-103.

Karlis, D. and Xekalaki, E. (1999), "On testing for the number of components in a mixed Poisson model," *Annals of Institute of Statistical Mathematics*, 51, 149-162.

Karlis, D. and Xekalaki, E. (2001), "Robust inference for finite Poisson mixtures," *Journal of Statistical Planning and Inference*, 93, 93-115.

Keribin, C. (2000), "Consistent estimation of the order of mixture models," *Sankhyā, Ser. A* 62, 49-62.

Lawson, A., Biggeri, A., Böhning, D., Lesaffre, E., Viel, J. F., Bertollini, R. (Eds.)(1999), *Disease mapping and risk assessment for Public Health decision making* Wiley, Chichester.

Lindsay, B. G., (1995),"Mixture Models: Theory, Geometry, and Applications," NSF-CBMS Regional Conference Series in Probability and Statistics, Vol. 5, Institute of Mathematical Statistics, Hayward.

Lu, Z., Hui, Y. V. and Lee, A. H. (2003)," Minimum Hellinger Distance Estimation for Finite Mixtures of Poisson Regression Models and its Applications," *Biometrics* 59, 1016-1026.

McLachlan, G. J. (1987), "On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture," *Journal of the Royal Statistical Society, Ser. C (Applied Statistics)* 36, 318-324.

McLachlan, G. J. and Basford, K. E. (1988), *Mixture Models: Inference and Applications to Clustering*, New York: Marcel Dekker.

McLachlan, G. J. and Peel, D. (2000), "Mixtures of factor analyzers", in *Proceedings of the Seventeenth International Conference on Machine Learning*, San Francisco: Morgan Kaufmann, pp. 599-606.

Mullahy, J. (1997),"Heterogeneity, excess zeros, and the structure of count data models," *Journal of Applied econometrics*, 12, 337-350.

Pauler, D. K., Escobar, M. D., Sweeney, J. A. and Greenhouse, J. (1996), "Mixture models for eye-tracking data: A case study," *Statistics in Medicine*, 15, 1365-1376.

Priebe, C. E. and Marchette, D. J. (2000), "Alternating kernel and mixture density estimates," *Computational Statistics and Data Analysis*, 35, 43-65.

Roeder, K. (1994), "A graphical technique for determining the number of components in a mixture of normals," *Journal of the American Statistical Association*, 89, 487-495.

Roeder, K. and Wasserman, L. (1997), "Practical Bayesian density estimation using mixtures of normals," *Journal of the American Statistical Association*, 92, 894-902.

Schlattmann, P. and Böhning, D. (1993), "Mixture models and disease mapping," *Statistics in Medicine*, 12, 943-950.

Simpson, D. G. (1987), "Minimum Hellinger distance estimation for the analysis of count data," *Journal of the American Statistical Association*, 82, 802-807.

Teicher, H. (1960), "On the mixture of distributions," *Annals of Mathematical Statistics*, 31, 55-73.

Teicher, H. (1961), " Identifiability of mixtures," *Annals of Mathematical Statistics*, 32, 244-248.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985), *Statistical Analysis of Finite Mixture Distributions*, New York: Wiley.

Woo, Mi-Ja, and Sriram, T. N. (2004), "Robust estimation of mixture complexity," Submitted.

Xiao, J., Lee, A. H., and Vemuri, S.R. (1999), "Mixture distribution analysis of length of hospital stay for efficient funding," *Socio-Economic Planning Sciences*, 33, 39-59.

Chapter 3

MODELING US UNEMPLOYMENT RATES

3.1 INTRODUCTION

In this chapter, we consider the problem of forecasting the civilian unemployment rate (UE) for each state and the District of Columbia in the United States. Among the important economic data developed by the Bureau of Labor Statistics (BLS), UE rates for states and local areas are viewed as key indicators of local economic conditions. These estimates are produced by state workforce agencies under the Federal-State cooperative Local Area Unemployment Statistics (LAUS) program. The U.S. civilian unemployment rate represents the percent of labor force that is unemployed. This rate is published every month by the BLS for the whole nation, as well as for its various geographic and demographic sub-domains. For example, the UE rate estimates are reported for all states and Washington DC, all metropolitan statistical areas, all counties (cities and towns in New England), and cities with a population of 25,000 or more. The unemployment rates are used in regional planning and fund allocation for states under various federal assistance programs.

The statistical models used for developing statewide LAUS estimates have been replaced with new, third-generation models, featuring real-time benchmarking to monthly national Current Population Survey (CPS) employment and unemployment totals. The models produce seasonally adjusted estimates within the estimation model, as well as non-seasonally adjusted estimates, and measures of error. The benchmark changes from annual state-level CPS estimates of employment and unemployment to monthly national-level CPS estimates. As part of the redesign, the historical series have been revised back to 1976 for various geographic and demographic sub-domains.

Modeling of the UE rate, including parameter estimation, may be done using the class of Box-Jenkins models (Box and Jenkins, 1976). Montgomery *et al.* (1998) presents a comparison of forecasting performance for a variety of linear and nonlinear time series models using the U.S. unemployment rate. They regard the seasonal ARIMA $(1,1,0)(1,0,1)_4$ model as a statistical model for the U.S. quarterly UE rate. The overall sample size for the CPS is sufficient to produce reliable estimates of UE rate at the national level that satisfy certain pre-specified precision requirements. Each state has been classified as a direct-use state or an indirect-use state with respect to the available sample size for each state. For states such as California, Florida, Illinois, Massachusetts, Michigan, New Jersey, New York, North Carolina, Ohio, Pennsylvania, and Texas, the sample sizes available are large enough to provide reliable estimates, and those states are classified as direct-use states. However, the remaining states and the District of Columbia which are classified as indirect-use states don't provide adequate samples, so that standard design-based estimators are not precise enough. Therefore, there is a need to improve efficiency for the states whose sample sizes are not sufficient.

In order to increase the accuracy of the state-level UE rate estimates, the current BLS method, developed by Tiller (1992) represents the observed CPS sample estimates $Y_{i,t}$ as $Y_{i,t} = \theta_{i,t} + \varepsilon_{i,t}$ for $i = 1, \ldots, m$, and $t = 1, \ldots, T$, where $\theta_{i,t}$ is the true UE rate for domain i at time t , and $\varepsilon_{i,t}$ is the sampling error. The BLS models $\theta_{i,t}$'s using structural time series with explanatory variables and the $\varepsilon_{i,t}$ as an Autoregressive Moving average (ARMA) model in order to capture the autocorrelations. Clearly, the BLS method does not utilize the information across states and does not provide uncertainty measures of the state estimates.

The focus of our research is to produce reliable forecasts of UE rates for each state by borrowing strength across states. To this end, it is assumed that each individual series arises from a distinct stochastic model but that all of these series share the same time-dependence structure. We also allow for anticipated dependence across the states, and such modeling allows for improved parameter estimation and forecasting. That is, we directly model the

sample estimates $Y_{i,t}$, for state i at time t, as a Seasonal Autoregressive Integrated Moving Average (SARIMA) model. In order to use the information across states, we assume that each parameter corresponding to state i in the above model is randomly distributed with a common mean and variance. To carry out estimation of parameters in this random-effects version of panel time series data, we use the Bayesian approach.

Seasonal Autoregressive Moving Average (SARMA) models have been widely used in economics and statistics. There is a considerable literature on inferences for these models using frequentist approaches, such as least squares or maximum likelihood methods (see Anderson (1978), and Azzanlini (1981)). A Bayesian modeling framework has the advantage of being able to incorporate available prior information in a natural way. Recently, Bayesian inference has been facilitated by the emergence of Markov Chain Monte Carlo (MCMC) simulation methods such as the Gibbs sampler (see Tanner and Wong (1987), Gelfand and Smith (1990)) and Metropolis-Hastings (MH) algorithms (Metroplolis et al. (1953), Hastings (1970), and Tierney (1993)). These methods are powerful tools for simulating intractable joint distribution of interest. A sample of draws is the output of the simulation, and it can be used for various purposes such as computing posterior moments and quantiles.

The Bayesian inference for time series regression, especially with autoregressive processes conditioned on initial observations, was recognized early by Chib (1993), McCulloch and Tsay (1993) and Albert and Chib (1993). The Bayesian work on ARMA models was spurred by the approach of Monahan (1983), and Broemeling and Shaarway (1984). Marriott *et. al* (1992) discussed an approach to the estimation of ARMA models that is based on sampling functions of the partial autocorrelations. Chib and Greenberg (1994) developed methods of analyzing ARMA regression error models in a Bayesian framework by using Gibbs sampling and Metropolis-Hastings algorithms.

Although a Bayesian perspective for time series has been actively pursued, a full treatment for SARMA model is not available. In section 3.2, we briefly explain the Bayesian method Chib and Greenberg (1994) developed. In section 3.3, we present a Bayesian infer-

ence methodology for SARMA processes using MCMC method. We combine Markov chain strategies, as has been done by Chib and Greenberg (1994), but with a different class of candidate-generating densities. In section 3.4, we describe our modeling of the UE rates for all the states following a multiplicative seasonal ARIMA model. Also, Bayesian fitting and inference through the Gibbs sampler is discussed. In section 3.4, we perform parameter estimation for the UE rates data using the framework discussed in the section 3.4, including examination of the model adequacy, and forecast of the last four unused observations for all the states. Our estimates and forecasts are compared with those in the univariate SARIMA model. Overall summary and conclusions are given in section 3.6. We begin with the Bayesian analysis for ARMA model.

3.2 LITERATURE REVIEW

Chib and Greenberg (1994) developed the procedure to analyzing $ARMA(p,q)$ regression error models in a Bayesian framework via the Gibbs sampling and Metropolis-Hastings algorithms. They consider the following model where the observation at time t, y_t, is generated by

$$y_t = \boldsymbol{x}_t'\beta + \varepsilon_t, \quad t = 1, \ldots, n$$

$$\varepsilon_t = \phi_1\varepsilon_{t-1} + \cdots + \phi_p\varepsilon_{t-p} + u_t + \theta_1 u_{t-1} + \cdots + \theta_q u_{t-q}, \tag{3.2.1}$$

where $\boldsymbol{x}_t$ is a $k \times 1$ vector of covariates, β is the $k \times 1$ vector of regression parameters, ε_t is a random error, $\phi_p \neq 0$, $\theta_q \neq 0$, $u_t \sim iid\ N(0, \sigma^2), \sigma^2 > 0$, N denotes the normal distribution, and ε_t follows an $ARMA(p,q)$ process. Then, the model in (3.2.1) is equivalent to follows in state space form (see Harvey, 1981):

$$y_t = \boldsymbol{x}_t'\beta + \boldsymbol{z}'\alpha_t \tag{3.2.2}$$

$$\alpha_t = G\alpha_{t-1} + fu_t, \tag{3.2.3}$$

where $\boldsymbol{z} = (1, 0, \ldots, 0)'$, $\alpha_t = (\alpha_{1t}, \ldots, \alpha_{mt})'$, $m = max(p,\ q+1)$,

$$G = \begin{vmatrix} \phi_1 & \vdots & & & \\ \phi_2 & \vdots & & & \\ \phi_3 & \vdots & & I_{m-1} & \\ \vdots & \vdots & & & \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \phi_m & \vdots & 0 & \cdots & 0 \end{vmatrix}$$

and $f = (1, \theta_1, \ldots, \theta_q)'$. In a Bayesian analysis, they determine moments and other features of the posterior distribution of $\boldsymbol{\Psi} = (\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\theta}, \sigma^2)$ under certain assumptions. Let $\pi(\boldsymbol{\Psi})$ be the prior density and $f(\boldsymbol{y}|\boldsymbol{\Psi})$ the likelihood function. Then, by Bayes theorem the posterior density is given by $f(\boldsymbol{\Psi}|\boldsymbol{y}) \propto \pi(\boldsymbol{\Psi}) f(\boldsymbol{y}|\boldsymbol{\Psi})$, where the likelihood function is dependent on the pre-sample errors $\lambda = (\varepsilon_0, \ldots, \varepsilon_{-p+1}, u_0, \ldots, u_{-q+1})$. By a consequence of the state space form of the $ARMA$ model, they showed that the conditional likelihood can be expressed in terms of only m pre-sample variables $\boldsymbol{\alpha}_0$, not all the $p+q$ elements of λ. They included $\boldsymbol{\beta}$, $\boldsymbol{\phi}$, $\boldsymbol{\theta}$, σ^2, and $\boldsymbol{\alpha}_0$ as elements in their MCMC algorithm. To perform a Bayesian analysis, they make the following assumption as the prior distributions.

Assumption (Prior distributions):

$$\begin{aligned} [\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\theta}, \sigma^2, \boldsymbol{\alpha}_0] &= [\boldsymbol{\beta}]\ [\boldsymbol{\phi}]\ [\boldsymbol{\theta}]\ [\sigma^2]\ [\boldsymbol{\alpha}_0|\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\theta}, \sigma^2] \\ &= N_k(\boldsymbol{\beta}|\boldsymbol{\beta}_0, B_\beta^{-1}),\ N_p(\boldsymbol{\phi}|\boldsymbol{\phi}_0, B_\phi^{-1})\ N_q(\boldsymbol{\theta}|\boldsymbol{\theta}_0, B_\theta^{-1}) \\ &\quad \times \mathcal{IG}(\sigma^2|v_0/2, \delta_0/2)\ [\boldsymbol{\alpha}_0|\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\theta}, \sigma^2], \end{aligned}$$

where notation, $[X]$ is a distribution of a random variable X, $[X, Y]$ is a joint distribution of random variables X and Y, $[X|Y]$ is a conditional distribution of random variables X given Y, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)$, $N_s(\cdot|\mu, \sigma^2)$ is the s-variate normal distribution with mean μ and variance σ^2, and $\mathcal{IG}(.)$ is the inverse gamma distribution. The hyperparameters

$\boldsymbol{\beta}_0$, ϕ_0, θ_0, B_β, B_ϕ, , B_θ, v_0, and δ_0 are known. Also, each parameter is assumed to be independently distributed with each other. For the initial state vector, the stationarity assumption implies that $\boldsymbol{\alpha}_0$, conditioned $\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\theta}$, and σ^2, has a normal distribution with $E(\boldsymbol{\alpha}_0) = 0$ and $V(\boldsymbol{\alpha}_0\boldsymbol{\alpha}_0') = \Omega$, where $vec(\Omega) = \sigma^2(I - G \otimes G)^{-1}vec(ff')$.

In many cases, the posterior density is analytically intractable, therefore a sampling-based approach to estimate the model parameters has been used. The Gibbs sampling approach involves sampling from the complete conditional distribution of each parameter in a systematic manner, conditional on the previous sample values of the other parameters. Therefore, they included $\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\theta}, \sigma^2$, and $\boldsymbol{\alpha}_0$ as elements in their MCMC algorithm, and simulated these parameters from the conditional densities $\pi(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{\Psi}_{-\beta}, \boldsymbol{\alpha}_0)$, $\pi(\boldsymbol{\phi}|\boldsymbol{y}, \boldsymbol{\Psi}_{-\phi}, \boldsymbol{\alpha}_0)$, $\pi(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{\Psi}_{-\theta}, \boldsymbol{\alpha}_0)$, $\pi(\sigma^2|\boldsymbol{y}, \boldsymbol{\Psi}_{-\sigma^2}, \boldsymbol{\alpha}_0)$, and $\pi(\boldsymbol{\alpha}_0|\boldsymbol{y}, \boldsymbol{\Psi})$, where $\boldsymbol{\Psi}_{-\phi}$ denotes all parameters in $\boldsymbol{\Psi}$ but for $\boldsymbol{\phi}$. In order to simulate these in the regression model with $ARMA(p, q)$ errors, the full conditional distributions are used. The following two results are essential to inducing the full conditional distributions given by the *Proposition 1.*

Definition 1. Let the scalars $y_s = y_s^* = 0$, and the vectors $\boldsymbol{x}_s = \boldsymbol{x}_s^* = 0$, $s \leq 0$, and let $\alpha_{r0} = 0$, $r > m$, where α_{r0} is the rth element of $\boldsymbol{\alpha}_0$. For $t = 1, \ldots, n$, define

$$y_t^* = y_t - \sum_{s=1}^{p} \phi_s y_{t-s} - \sum_{i=1}^{q} \theta_i y_{t-i}^* - \phi_t \alpha_{10} - \alpha_{t+1,0}$$

,

$$\boldsymbol{x}_t^* = \boldsymbol{x}_t - \sum_{s=1}^{p} \phi_s \boldsymbol{x}_{t-s} - \sum_{i=1}^{q} \theta_i \boldsymbol{x}_{t-i}^*$$

.

With *Definition 1*, the following lemma can be shown:

Lemma 1. Let $\boldsymbol{y}^*$ be the $n \times 1$ vector of the y_t^* and let X^* be the matrix with $\boldsymbol{x}_t^*$ as its ith row. Then,

$$f(\boldsymbol{y}^*|\boldsymbol{\Psi}, \boldsymbol{\alpha}_0) = (2\pi\sigma^2)^{-n/2} exp\left[-\frac{1}{2\sigma^2} (\boldsymbol{y}^* - X^*\boldsymbol{\beta})^T (\boldsymbol{y}^* - X^*\boldsymbol{\beta}) \right]. \tag{3.2.4}$$

Definition 2. Let the scalars $y_s = \bar{y}_s = \bar{\boldsymbol{x}}_s = 0$, $s \leq 0$, and the vectors $\boldsymbol{x}_s = 0$, and let $\alpha_{r0} = 0$, $r > m$. For $t = 1, \ldots, n$, define

$$\bar{y}_t = y_t - \boldsymbol{x}_t'\boldsymbol{\beta} - \sum_{i=1}^{q} \theta_i \bar{y}_{t-i} - \alpha_{t+1,0},$$

$$\bar{\boldsymbol{x}}_t = y_t - \boldsymbol{x}_t'\boldsymbol{\beta} - \sum_{i=1}^{q} \theta_i \bar{\boldsymbol{x}}_{t-i}.$$

With *Definition 2*, the following lemma can be proved:

Lemma 2. Let $\bar{\boldsymbol{y}}$ be the $n \times 1$ column vector of the $\bar{y}_t$ and let $\bar{X}$ be the matrix given by

$$\bar{X} = \begin{vmatrix} \alpha_{10} & 0 & \cdots & \cdots & 0 \\ \bar{x}_1 & \alpha_{10} & 0 & \cdots & 0 \\ \bar{x}_2 & \bar{x}_1 & \alpha_{10} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \bar{x}_{p-1} & \bar{x}_{p-2} & \cdots & \cdots & \alpha_{10} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \bar{x}_{n-1} & \bar{x}_{n-2} & \bar{x}_{n-3} & \cdots & \bar{x}_{n-p} \end{vmatrix}$$

Then,

$$f(\bar{\boldsymbol{y}}|\boldsymbol{\Psi}, \boldsymbol{\alpha}_0) = (2\pi\sigma^2)^{-n/2} exp\left[-\frac{1}{2\sigma^2} (\bar{\boldsymbol{y}} - \bar{X}\boldsymbol{\phi})^T (\bar{\boldsymbol{y}} - \bar{X}\boldsymbol{\phi}) \right]. \tag{3.2.5}$$

Based on two lemmas above, the full conditional distributions can be showed as the *Proposition 1.*

Proposition 1: Full conditional distribution

(i) $\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{\Psi}_{-\beta}, \boldsymbol{\alpha}_0 \sim N_k(B_n^{-1}(B_\phi \boldsymbol{\beta}_0 + \sigma^{-2} X^{*T} \boldsymbol{y}^*), B_n^{-1})$,

(ii) $\boldsymbol{\phi}|\boldsymbol{y}, \boldsymbol{\Psi}_{-\phi}, \boldsymbol{\alpha}_0 \sim p_1(\phi) \times N_p(V_n^{-1}(B_\phi \boldsymbol{\phi}_0 + \sigma^{-2} \bar{X}^T \bar{\boldsymbol{y}}), V_n^{-1})$,

(iii) $\sigma^2|\boldsymbol{y}, \boldsymbol{\Psi}_{-\sigma^2}, \boldsymbol{\alpha}_0 \sim \mathcal{IG}((v_0 + n)/2, (\delta_0 + d_1 + d_2)/2)$,

(vi) $\pi(\boldsymbol{\alpha}_0|\boldsymbol{y}, \boldsymbol{\Psi}) \propto N_m(\hat{\boldsymbol{\alpha}}_{0|n}, R_{0|n})$,

(v) $\pi(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{\Psi}) \propto p_2(\theta) \times \prod_{t=1}^{n} exp[-(1/2\sigma^2) u_t(\boldsymbol{\theta})^2] \times exp[-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T B_\theta (\boldsymbol{\theta} - \boldsymbol{\theta}_0)]$.

where we let $B_n = B_\phi + \sigma^{-2}X^{*T}X^*$, $V_n = B_\theta + \sigma^{-2}\bar{X}^T\bar{X}$, and define the function $p\ (\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma^2) = (\sigma^{-2})^{-m/2}|\ \Omega(\boldsymbol{\phi}, \boldsymbol{\theta})|^{-1/2} exp\ [-(1/2\sigma^2)\boldsymbol{\alpha}_0^T\Omega(\boldsymbol{\phi}, \boldsymbol{\theta})\boldsymbol{\alpha}_0]$, which is the prior density $\pi(\boldsymbol{\alpha}_0|\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\theta}, \sigma^2)$. For a given value of $(\boldsymbol{\theta}, \sigma^2)$, the latter function is denoted as $p_1(\boldsymbol{\phi})$, and for a given value of $(\boldsymbol{\phi}, \sigma^2)$, it is denoted as $p_2(\boldsymbol{\theta})$. Also, let $d_1 = ||\boldsymbol{y}^* - X^*\boldsymbol{\beta}||^2$ and $d_2 = \boldsymbol{\alpha}_0^T\Omega(\boldsymbol{\phi}, \boldsymbol{\theta})^{-1}\boldsymbol{\alpha}_0$. Finally, $\hat{\boldsymbol{\alpha}}_{0|n}$ and $R_{0|n}$ are the mean and covariance of the full conditional distribution of $\boldsymbol{\alpha}_0$, which are obtained from the recursion (see Harvey, 1981).

From the *Proposition 1*, they show that the full conditional distributions of $\boldsymbol{\beta}$, σ^2, and $\boldsymbol{\alpha}_0$ are straightforward to compute and belong to standard families of distributions, but those of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ are more intricate.

3.3 BAYESIAN ANALYSIS FOR SARIMA MODEL

In this section, we consider analyzing a Seasonal ARMA model in a Bayesian framework via the Gibbs sampling and Metropolis- Hasting algorithms. Let v_t denote a univariate time series generated by a Seasonal ARMA $(p, q)(P, Q)_s$ process

$$\boldsymbol{\Phi}_P(B^s)\boldsymbol{\phi}_p(B)v_t = \boldsymbol{\theta}_q(B)\boldsymbol{\Theta}_Q(B^s)a_t, t = 1, \ldots, n, \tag{3.3.6}$$

where $\boldsymbol{\Phi}_P(B^s) = (1 - \Phi_1 B^s - \cdots - \Phi_P B^{Ps})$, $\boldsymbol{\phi}_p(B) = (1 - \phi_1 B - \cdots - \phi_p B^p)$, $\boldsymbol{\Theta}_Q(B^s) = (1 + \Theta_1 B^s + \cdots + \Theta_Q B^{Qs})$, and $\boldsymbol{\theta}_p(B) = (1 + \theta_1 B + \cdots + \theta_q B^q)$ are autoregressive and moving average polynomials. Here, we assume that $\boldsymbol{\Phi}_P(B^s)$,$\boldsymbol{\phi}_p(B)$, $\boldsymbol{\theta}_q(B)$ and $\boldsymbol{\Theta}_Q(B^s)$ obey the usual stationarity and invertibility conditions. Let $\mathbf{v_0} = (v_0, \ldots, v_{1-p-s})^T$ denote the history of the data process, and $\mathbf{a_0} = (a_0, \ldots, a_{1-q-s})^T$ denote the history of the error process. For simplicity, we also assume $P \leq 1$, $Q \leq 1$, and $\mathbf{v_0} = (v_0, \ldots, v_{-p+1})^T = \mathbf{0}$. Then, equivalently, (3.3.6) can be expressed as:

$$\begin{aligned} v_t = \phi_1 v_{t-1} + \cdots + \phi_p v_{t-p} - \Phi(\phi_1 v_{t-1-s} + \cdots + \phi_p v_{t-p-s}) + \\ a_t + \theta_1 a_{t-1} + \cdots + \theta_q a_{t-q} + \Theta(\theta_1 a_{t-1-s} + \cdots + \theta_q a_{t-q-s}), \end{aligned} \tag{3.3.7}$$

where $\phi_p \neq 0$, $\theta_q \neq 0$, $a_t \sim i.i.d.N(0, \sigma^2)$, $\sigma^2 > 0$, and N denotes the normal distribution. Also, setting $\Phi = 0$, and $\Theta = 0$ would lead to ARMA models. Let $\boldsymbol{\phi} = (\phi_1, \cdots, \phi_p)^T$, $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_q)^T$, $\boldsymbol{v} = (v_1, \cdots, v_n)^T$, and $\boldsymbol{a} = (a_1, \cdots, a_n)^T$.

Given the data, along with the parameter vector $\boldsymbol{\Psi} = (\boldsymbol{\phi}, \boldsymbol{\theta}, \Phi, \Theta, \sigma^2)$ and pre-sample errors $\mathbf{a_0}$, the Bayesian model specification requires a likelihood function $f(\boldsymbol{v}|\boldsymbol{\Psi}, \mathbf{a_0})$ and a prior density $\pi(\boldsymbol{\Psi}, \mathbf{a_0})$. By the Bayes theorem, we obtain the posterior density as: $\pi(\boldsymbol{\Psi}, \mathbf{a_0}|\boldsymbol{v})$ $\propto$ $f(\boldsymbol{v}|\boldsymbol{\Psi}, \mathbf{a_0})\pi(\boldsymbol{\Psi}, \mathbf{a_0})$. Given the parameter vector $\boldsymbol{\Psi} = (\boldsymbol{\phi}, \boldsymbol{\theta}, \Phi, \Theta, \sigma^2)$ and pre-sample errors $\mathbf{a_0}$, the density of $\boldsymbol{v}$ can be expressed as

$$\begin{aligned} f(\mathbf{v}|\boldsymbol{\Psi}, \mathbf{a_0}) &= \prod_{t=1}^{n} (2\pi\sigma^2)^{-1/2} exp\left[-\frac{1}{2\sigma^2}\ a_t^2 \right] \\ &= \prod_{t=1}^{n} (2\pi\sigma^2)^{-1/2} exp\left[-\frac{1}{2\sigma^2}\ (v_t - \hat{v}_{t|t-1})^2 \right], \end{aligned}$$

where $\hat{v}_{t|t-1} = \phi_1 v_{t-1} + \cdots + \phi_p v_{t-p} + \Phi v_{t-s} - \Phi(\phi_1 v_{t-1-s} + \cdots + \phi_p v_{t-p-s}) + \theta_1 a_{t-1} + \cdots + \theta_q a_{t-q} + \Theta a_{t-s} + \Theta(\theta_1 a_{t-1-s} + \cdots + \theta_q a_{t-q-s})$ is the one-step-ahead predictor of v_t given information up to time $t-1$. For the prior distribution, we make the following assumption.

Assumption (Prior distributions):

$$\begin{aligned} [\boldsymbol{\phi}, \boldsymbol{\theta}, \Phi, \Theta, \sigma^2, \mathbf{a_0}] &= [\boldsymbol{\phi}]\ [\boldsymbol{\theta}]\ [\Phi]\ [\Theta]\ [\sigma^2]\ [\mathbf{a_0}] \\ &= N_p(\boldsymbol{\phi}|\boldsymbol{\phi}_0, B_\phi^{-1})\ N_q(\boldsymbol{\theta}|\boldsymbol{\theta}_0, B_\theta^{-1})\ N(\Phi|\Phi_0, V_\Phi^{-1})\ N(\Theta|\Theta_0, V_\Theta^{-1}) \\ &\quad \times \mathcal{IG}(\sigma^2|v_0/2, \delta_0/2), \end{aligned}$$

that is,

$$f(\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\Phi}, \boldsymbol{\Theta}, \sigma^2, \mathbf{a_0}|\boldsymbol{\phi}_0, B_\phi, \Phi_0, B_\theta, \boldsymbol{\Phi}_0, V_\Phi, \boldsymbol{\Theta}_0, V_\Theta, v_0, \delta_0)$$

$$\propto |B_\phi|^{1/2} e^{\{-1/2(\boldsymbol{\phi} - \boldsymbol{\phi_0})^T B_\phi (\boldsymbol{\phi} - \boldsymbol{\phi_0})\}} \times |B_\theta|^{1/2} e^{\{-1/2(\boldsymbol{\theta} - \boldsymbol{\theta_0})^T B_\theta (\boldsymbol{\theta} - \boldsymbol{\theta_0})\}}$$

$$\times V_\Phi|^{1/2} e^{\{-1/2(\Phi - \Phi_0)^T V_\Phi (\Phi - \Phi_0)\}} \times |V_\Theta|^{1/2} e^{\{-1/2(\Theta - \Theta_0)^T V_\Theta (\Theta - \Theta_0)\}} \times \sigma^{-v_0/2+1} e^{-\delta_0/(2\sigma^2)}$$

where the hyperparameters ϕ_0, B_ϕ, θ_0, B_θ,Φ_0, V_Φ, Θ_0, V_Θ, v_0, and δ_0 are known, and each parameter is assumed to be independently distributed with each other.

It should be noted that the usual normal-inverted gamma distribution has been assumed for σ^2, while those for $\boldsymbol{\phi}$, Φ, $\boldsymbol{\theta}$ and Θ are normal. A vague prior information has been assumed for pre-sample errors $\mathbf{a_0}$.

In our Bayesian analysis, we will determine moments and other features of the posterior distribution of $\boldsymbol{\Psi}$ under the assumptions described above. To estimate the model parameters, we sample those from the complete conditional distribution of each parameter via the Gibbs sampling. As shown in (3.3.7), two autoregressive polynomials, $\Phi_P(B^s)$ and $\phi_p(B)$ are expressed in multiplicative form in the seasonal $ARMA(p,q)(P,Q)_s$ model. Also, the expression of two moving average polynomials, $\Theta_Q(B^s)$ and $\theta_q(B)$ is multiplicative. To induce a relationship for $\boldsymbol{\phi}$ and Φ, we set $v_t^* = v_t - \Phi v_{t-s}$, $v_t^+ = v_t - \phi_1 v_{t-1} - \cdots - \phi_p v_{t-p}$, and $\boldsymbol{\theta}^* = (\theta_1^*, \ldots, \theta_{q+s}^*) = (\theta_1, \ldots, \theta_q, 0, \ldots, 0, \Theta, \theta_1\Theta, \ldots, \theta_q\Theta)$. Then, setting v_t^* makes the seasonal $ARMA(p,q)(P,Q)_s$ process of v_t be the $ARMA(p, Qs+q)$ process of v_t^*, and setting v_t^+ be the $ARMA(P, Qs+q)$ process of v_t^+, respectively. Using these notations of v_t^* and v_t^+ above, we obtain the complete conditional distribution of each parameter, following the framework that Chib and Greenberg (1994) presented as in section 3.2. Two results below are central to our Bayesian analysis for the seasonal $ARMA(p,q)(P,Q)_s$, ($P \leq 1$,$Q \leq 1$), and these simple recursive transformations of the data yield a relationship for $\boldsymbol{\phi}$ and Φ.

Definition 3. Let $v_t^* = v_t - \Phi v_{t-s}$. For $l < 0$, let the scalars $\bar{y}_l = a_l$, and $\bar{x}_l = 0$, and let $\boldsymbol{\theta}^* = (\theta_1^*, \ldots, \theta_{q+s}^*) = (\theta_1, \ldots, \theta_q, 0, \ldots, 0, \Theta, \theta_1\Theta, \ldots, \theta_q\Theta)$, where $s > q$, and the number of zeros is $s - q - 1$. For $t = 1, \ldots, n$, define

$$\bar{y}_t = v_t^* - \sum_{i=1}^{q+s} \theta_i^* \bar{y}_{t-i}$$

$$\bar{x}_t = v_t^* - \sum_{i=1}^{q+s} \theta_i^* \bar{x}_{t-i}$$

With this definition, we can show the following by verifying that $\bar{y}_1 - \bar{\boldsymbol{x}}_1' \boldsymbol{\phi} = a_1$, and proceeding by induction, where $\bar{\boldsymbol{x}}_1'$ is the first row of $\bar{X}$:

$$f(\bar{\mathbf{y}}|\boldsymbol{\Psi}, \mathbf{a_0}) = (2\pi\sigma^2)^{-n/2} exp\left[-\frac{1}{2\sigma^2} (\bar{\mathbf{y}} - \bar{X}\boldsymbol{\phi})^T (\bar{\mathbf{y}} - \bar{X}\boldsymbol{\phi}) \right], \tag{3.3.8}$$

where $\bar{\mathbf{y}}$ is the $n \times 1$ column vector of the $\bar{y}_t$, and $\bar{X}$ is $n \times p$ matrix given by

$$\bar{X} = \begin{vmatrix} 0 & 0 & \cdots & \cdots & 0 \\ \bar{x}_1 & 0 & 0 & \cdots & 0 \\ \bar{x}_2 & \bar{x}_1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \bar{x}_{n-1} & \bar{x}_{n-2} & \bar{x}_{n-3} & \cdots & \bar{x}_{n-p} \end{vmatrix}$$

Definition 4. Let $v_t^+ = v_t - \phi_1 v_{t-1} - \cdots - \phi_p v_{t-p}$. For $l < 0$, let the scalars $\tilde{y}_l = a_l$, and $\tilde{x}_l = 0$. For $t = 1, \ldots, n$, define

$$\tilde{y}_t = v_t^+ - \sum_{i=1}^{q+s} \theta_i^* \tilde{y}_{t-i}$$

$$\tilde{x}_t = v_t^+ - \sum_{i=1}^{q+s} \theta_i^* \tilde{x}_{t-i}$$

With this definition, we can show the following:

$$f(\tilde{\mathbf{y}}|\boldsymbol{\Psi}, \mathbf{a_0}) = (2\pi\sigma^2)^{-n/2} exp\left[-\frac{1}{2\sigma^2} \, (\tilde{\mathbf{y}} - \tilde{\mathbf{x}}\Phi)^T (\tilde{\mathbf{y}} - \tilde{\mathbf{x}}\Phi) \right], \tag{3.3.9}$$

where $\tilde{\mathbf{y}}$ is the $n \times 1$ column vector of the $\tilde{y}_t$, and $\tilde{\mathbf{x}} = (0, \ldots, 0, \tilde{x}_1, \ldots, \tilde{x}_{n-s})^T$, here the number of zeros is s.

Based on the two results above, followings are the full conditional distributions that are used in our data analysis using a Seasonal ARMA $(p, q)(P, Q)_s$ ($P \leq 1$,$Q \leq 1$) model:

Proposition 2: Full conditional distribution

(i) $\quad \boldsymbol{\phi}|\mathbf{v}, \boldsymbol{\Psi}_{-\phi}, \mathbf{a}_0 \sim N_p(B_1^{-1}(B_\phi \boldsymbol{\phi}_0 + \sigma^{-2} \bar{X}^T \bar{\mathbf{y}}), B_1^{-1})$,

(ii) $\quad \Phi|\mathbf{v}, \Psi_{-\Phi}, \mathbf{a}_0 \sim N(B_2^{-1}(B_\Phi \Phi_0 + \sigma^{-2} \tilde{\mathbf{x}}^T \tilde{\mathbf{y}}), B_2^{-1})$,

(iii) $\quad \boldsymbol{\pi}(\boldsymbol{\theta}, \Theta|\mathbf{v}, \Psi_{-(\boldsymbol{\theta},\Theta)}, \mathbf{a}_0) \propto \prod_{t=1}^n exp[-(1/2\sigma^2) a_t(\boldsymbol{\theta}, \Theta)^2]$

$$\times exp[-\tfrac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T B_\theta (\boldsymbol{\theta} - \boldsymbol{\theta}_0)] exp[-\tfrac{1}{2} B_\Theta (\Theta - \Theta_0)^2],$$

(iv) $\sigma^2|\mathbf{v}, \mathbf{\Psi}_{-\sigma^2}, \mathbf{a}_0 \sim \mathcal{IG}((v_0+n)/2, (\delta_0+d_1)/2)$,

(v) $\pi(\mathbf{a}_0|\mathbf{v}, \mathbf{\Psi}) \propto \prod_{t=1}^{n} exp[-(1/2\sigma^2)a_t(\mathbf{a}_0)^2]$,

where we let $B_1 = B_\phi + \sigma^{-2}\bar{X}^T\bar{X}$, $B_2 = B_\Phi + \sigma^{-2}\tilde{\mathbf{x}}^T\tilde{\mathbf{x}}$, and $d_1 = ||\bar{\mathbf{y}} - \bar{X}\boldsymbol{\phi}||^2$.

Proof. (i), (ii), and (iv) follow from assumption, (3.3.8) and (3.3.9). (iii) and (v) follow from the definition of the full conditional distribution, that is, the form $\pi(\mathbf{\Psi}, \mathbf{a_0}|\boldsymbol{v}) \propto f(\boldsymbol{v}|\mathbf{\Psi}, \mathbf{a_0})\pi(\mathbf{\Psi}, \mathbf{a_0})$ (see Gelfand and Smith(1990), for details).

Notice that our full conditional distributions are similar to those of Chib and Greenberg (1994). First thing we note is that our model does not include the parameter vector β, while their model has the regression parameter vector. Secondly, our model has other parameters Φ and Θ in multiplicative form. Thirdly, in our Bayesian analysis, the presample errors $(u_0, \ldots, u_{-p-Ps+1})$ are set to be zeros, and presample errors $(a_0, \ldots, a_{-q-Qs+1})$ are assumed to be vague prior information, whereas in their analysis the initial state vector $\boldsymbol{\alpha}_0$ is used as the history of data, and it is assumed to be normally distributed. Setting $(u_0, \ldots, u_{-p-Ps+1}) = \mathbf{0}$ makes computations of the full conditional distributions simplified. In addition, this makes it possible to calculate v_t^* and v_t^+ for $t < (p+Ps)$ since they are dependent of $(u_0, \ldots, u_{-p-Ps+1})$. Also, the vague prior information of $\boldsymbol{a}_0$ makes computation easier since it does not depend on other parameters, $\mathbf{\Psi}$, while the prior of $\boldsymbol{\alpha}_0$ depends on $\mathbf{\Psi}$.

In our Bayesian analysis, we will determine the mean and standard deviation of posterior distribution of $\mathbf{\Psi}$ under the assumption described above. We estimate $\mathbf{\Psi}$ by the posterior means, and measure the uncertainty in the estimates of $\mathbf{\Psi}$ by the posterior standard deviation. For these computations, we use the Gibbs sampler (Gelman and Rubin (1992), Gelfand and Smith (1990)).

The Gibbs sampler is a Monte Carlo Markovian updating scheme that gives the marginal, conditional, and joint distributions of the random variables. The Gibbs sampling algorithm requires sampling from the complete conditional distributions associated elements of ($\mathbf{\Psi}$, $\mathbf{a_0}$) in some systematic order. In our application, the full conditional distributions of $\boldsymbol{\phi}$, Φ and

σ^2 are standard distributions that can be easily sampled, while those of $\boldsymbol{\theta}$, Θ and $\mathbf{a_0}$ are more intricate, and therefore they are computed in an analogous manner.

To implement the sampling algorithm, we block the parameters into the following groups: ($\boldsymbol{\phi}$), (Φ), (σ^2), ($\boldsymbol{\theta}$, Θ), and ($\mathbf{a_0}$). Within each Gibbs iteration, we draw samples of ($\boldsymbol{\phi}$), (Φ) and (σ^2) from the Normal distributions and the inverse Gamma distribution described above, and use draws based on a Metropolis-Hastings for the groups of ($\boldsymbol{\theta}$, Θ) and ($\mathbf{a_0}$). For ($\boldsymbol{\theta}$, Θ) and ($\mathbf{a_0}$), we use the random walk version of the Metropolis-Hastings (MH) algorithm with appropriate Gaussian proposals to obtain samples from the required stationary distributions (Hastings, 1970). In the random walk version, new candidates are chosen by drawing from a distribution conditioned on the current parameter value, i.e. by drawing a step away from the current parameter value. More specifically, let U denote the current value of ($\boldsymbol{\theta}$,Θ). We draw V from the proposal centered at U. We calculate the ratio $\alpha(U,V) = f(\boldsymbol{v}|\boldsymbol{\phi},\Phi,V,\mathbf{a_0})\pi(V)/f(\boldsymbol{v}|\boldsymbol{\phi},\Phi,U,\mathbf{a_0})\pi(U)$. Then, we move U to V with the probability $\alpha(U,V)$. Similarly, let U' denote the current value of ($\mathbf{a_0}$). We draw V' from the proposal centered at U'. We calculate the ratio $\alpha(U',V') = f(\boldsymbol{v}|\boldsymbol{\phi},\Phi,\boldsymbol{\theta},\Theta,V')\pi(V')/f(\boldsymbol{v}|\boldsymbol{\phi},\Phi,\boldsymbol{\theta},\Theta,U')\pi(U')$. Then, we move U' to V' with the probability $\alpha(U',V')$.

Clearly, successful implementation of the MH algorithm requires a suitable proposal density. Note that, in a Bayesian analysis of ARMA model, Chib and Greenberg (1994) showed the full conditional distributions of $\boldsymbol{\phi}$, σ_2, and $\boldsymbol{\alpha}_0$ belong to standard families of distributions, and to implement the MH algorithm of $\boldsymbol{\theta}$, they proposed a candidate-generating density as the truncated normal approximation by expanding $u_t(\boldsymbol{\theta})$ around $\boldsymbol{\theta}^+$ as $u_t(\boldsymbol{\theta}) \approx u_t(\boldsymbol{\theta}^+) - w_t'(\boldsymbol{\theta}-\boldsymbol{\theta}^+)$, where $\boldsymbol{\theta}^+$ denotes the nonlinear least squares estimate of $\boldsymbol{\theta}$ and w_t' is the ith row of $W(\boldsymbol{\theta}^+) = (\partial u(\boldsymbol{\theta})/\partial\boldsymbol{\theta}')|\boldsymbol{\theta} = \boldsymbol{\theta}^+$. Also note that in a Bayesian analysis of AR or MA model, Marin *at el.*(2005) represented the polynomials as the factorized quantity $\boldsymbol{\phi}_p(B) = \prod_{i=1}^{p}(1-\lambda_i B)$ or $\boldsymbol{\theta}_q(B) = \prod_{i=1}^{q}(1-\lambda_i B)$, and used a reversible jump algorithm that distinguishes between the number of complex roots in the inverse roots, λ_i. Then, they

simulated $\boldsymbol{\phi}$ or $\boldsymbol{\theta}$ from a proposal based on a simple random walk. Our Bayesian analysis in seasonal ARMA model combines and expands two different methods, Chib and Greenberg (1994) and Marin *at el.*(2005).

3.4 BAYESIAN MODELING OF UE RATES DATA

The data consists of 29 years of quarterly observations (1976-2004) on the UE rates for 52 different states including the 51 states and the District of Columbia. The data for state i at time t is denoted by $z_{i,t}$, $t = 1, \ldots, 116$, $i = 1, \ldots, 52$. We model these series to share the same time-dependent structure. However, we expect the parameter estimates to be quite different across states, and the data for each state to be correlated with those for many other state. More precisely, following the Box-Jenkins framework of SARIMA model fitting, we fit a seasonal ARMA $(1,1)(1,1)_4$ model to the differenced series $v_{i,t} = z_{i,t} - z_{i,t-1}$, given by

$$\begin{aligned} v_{i,t} &= \phi_{i,1} v_{i,t-1} + \Phi_i v_{i,t-4} - \phi_{i,1}\Phi_i v_{i,t-5} \\ &\quad + a_{i,t} + \theta_{i,1} a_{i,t-1} + \Theta_i a_{i,t-4} + \theta_{i,1}\Theta_i a_{i,t-5}, \end{aligned} \tag{3.4.10}$$

for $t = 1, \ldots, n$, $i = 1, \ldots, I$, and $a_{i,t} \sim \mathcal{N}(0, \sigma_i^2)$, where n is the number of time points and I is the number of states.

For most of states, the seasonal ARIMA $(1,1,1)(1,0,1)_4$ model is fitted well as a statistical model for the U.S. quarterly UE rate. Note that the seasonal ARIMA $(1,1,0)(1,0,1)_4$ model was suggested for the U.S. quarterly UE rate data at the national level by Montgomery et. al (1998).

To incorporate dependence between the series, we assume that each of the parameters corresponding to state i in the above model is randomly distributed with a common mean and variance. That is, parameters are assumed to be from common probability distributions given by

$$\phi_{i,1} \sim \mathcal{N}(\phi_{10}, r_{\phi_1}^{-1}),\ \Phi_i \sim \mathcal{N}(\Phi_0, r_{\Phi}^{-1}),$$
$$\theta_{i,1} \sim \mathcal{N}(\theta_{10}, r_{\theta_1}^{-1}), \text{ and } \Theta_i \sim \mathcal{N}(\Theta_0, r_{\Theta}^{-1}).$$

Let $\mathbf{v}_{i0} = (v_{i,0}, \ldots, v_{i,1-p-s})$ and $\mathbf{a}_{i0} = (a_{i,0}, \ldots, a_{i,1-q-s})$ denote the relevant historical data for the ith series. Let $\mathbf{v}_i = (v_{i,1}, \ldots, v_{i,n})$ and $\mathbf{v} = (\mathbf{v}_1, \ldots, \mathbf{v}_I)$. Also let $\boldsymbol{\phi} = (\phi_{1,1}, \ldots, \phi_{I,1})$, $\boldsymbol{\Phi} = (\Phi_1, \ldots, \Phi_I)$, $\boldsymbol{\theta} = (\theta_{1,1}, \ldots, \theta_{I,1})$, $\boldsymbol{\Theta} = (\Theta_1, \ldots, \Theta_I)$, $\boldsymbol{\sigma^2} = (\sigma_1^2, \ldots, \sigma_I^2)$, and $\boldsymbol{a}_0 = (\boldsymbol{a}_{1,0}, \ldots, \boldsymbol{a}_{I,0})$. For simplicity, we set $\mathbf{y}_{i0} = \mathbf{0}$.

Given the data, along with the parameter vector $\boldsymbol{\Psi} = (\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\Phi}, \boldsymbol{\Theta}, \boldsymbol{\sigma^2}, \boldsymbol{a}_0, \phi_{10}, r_{\phi_1}, \Phi_0, r_\Phi, \theta_{10}, r_{\theta_1}, \Theta_0, r_\Theta)$, and pre-sample errors $\boldsymbol{a}_0$, the Bayesian model specification requires a likelihood $f(\mathbf{v}|\boldsymbol{\Psi}, \boldsymbol{a}_0)$ and a prior $\pi(\boldsymbol{\Psi}, \boldsymbol{a}_0)$. Given the parameters $\boldsymbol{\Psi}$ and pre-sample errors $\boldsymbol{a}_0$, the density of $\mathbf{v}$ can be expressed as:

$$f(\mathbf{v}|\boldsymbol{\Psi}, \boldsymbol{a}_0) = \prod_{i=1}^{I}\prod_{t=1}^{n}(2\pi\sigma_i^2)^{-1/2} exp\left[-\frac{1}{2\sigma_i^2}\ a_{i,t}^2\ \right] \tag{3.4.11}$$

$$= \prod_{i=1}^{I}\prod_{t=1}^{n}(2\pi\sigma_i^2)^{-1/2} exp\left[-\frac{1}{2\sigma_i^2}\ (v_{i,t} - \hat{v}_{i,t|t-1})^2\ \right],$$

where $\hat{v}_{i,t|t-1} = \phi_{i,1}v_{i,t-1} + \Phi_i v_{i,t-4} - \phi_{i,1}\Phi_i v_{i,t-5} + \theta_{i,1}a_{i,t-1} + \Theta_i a_{t-4} + \theta_{i,1}\Theta_i a_{i,t-5}$ is the one-step-ahead predictor of $v_{i,t}$ given information up to time $t-1$. We make the following assumption for the hyperparameters:

$$f(\phi_{10}, r_{\phi_1}, \Phi_0, r_\Phi, \theta_{10}, r_{\theta_1}, \Theta_0, r_\Theta, \sigma^2, \boldsymbol{a}_0) \tag{3.4.12}$$

$$\propto r_{\phi_1}^{1/2b-1} e^{-1/2ar_{\phi_1}} \times r_\Phi^{1/2d-1} e^{-1/2cr_\Phi}$$

$$\times r_{\theta_1}^{1/2f-1} e^{-1/2er_{\theta_1}} \times r_\Theta^{1/2h-1} e^{-1/2gr_\Theta}$$

$$\times \prod_{i=1}^{I} \sigma_i^{-v_0/2+1} e^{-\delta_0/(2\sigma_i^2)},$$

where the improper prior distributions are assumed for the hyperparameters $\phi_{10}, \Phi_0, \theta_{10}, \Theta_0$, and $\boldsymbol{a}_0$ and the values a, b, c, d, e, f, g, h, v_0, and δ_0 are known. A vague prior information can be entertained by setting b, d, f, h, and v_0 small positive numbers for r_{ϕ_1}, r_ϕ,r_{θ_1}, and r_θ. Note that the hyperparameters are assumed to be independent of each other.

To carry out the Gibbs sampling, it is necessary to sample from the full conditional distributions. As in the previous section, we present the full conditional distribution that

are used in simulation for the seasonal ARMA models introduced above. These give the conditional distribution for $\phi_{i,1}, \Phi_i, \theta_{i,1}, \Theta_i$ and σ_i^2 for the ith state under the model shown above. Hence, to carry out Gibbs sampling in our Bayesian setting across series, it is enough to compute the full conditional distribution of the hyperparemeters ϕ_{10}, Φ_0, θ_{10}, Θ_0, r_ϕ, r_Φ, r_θ, and r_Θ, in addition to conditional distribution for the $\phi_{i,1}$, Φ_i, $\theta_{i,1}$, Θ_i and σ_i^2, $i = 1, \ldots, I$. *Proposition 2* to obtain the full conditional distributions is used in the simulation for the cross-sectional modeling of UE rates. These are given in proposition 3.

Proposition 3: Full conditional distributions are given by the following: For $i = 1, \ldots, I$,

$$(i) \quad \phi_{i,1}|\boldsymbol{v}, \boldsymbol{\Psi}_{-\phi_{i,1}}, \boldsymbol{a}_0 \sim N(B_{i,1}^{-1}(B_{\phi_{i,1}}\phi_{10} + \sigma_i^{-2}\bar{\boldsymbol{x}}_i^T\bar{\boldsymbol{y}}_i), B_{i,1}^{-1}),$$

$$(ii) \quad \Phi_i|\boldsymbol{v}, \boldsymbol{\Psi}_{-\Phi_i}, \boldsymbol{a}_0 \sim N(B_{i,2}^{-1}(B_{\Phi_i}\Phi_0 + \sigma_i^{-2}\tilde{\boldsymbol{x}}_i^T\tilde{\boldsymbol{y}}_i), B_{i,2}^{-1}),$$

$$(iii) \quad \pi(\theta_{i,1}, \Theta_i|\boldsymbol{v}, \boldsymbol{\Psi}_{-(\theta_{i,1},\Theta_i)}, \boldsymbol{a}_{i,0}) \propto \prod_{t=1}^{n} exp[-(1/2\sigma_i^2)a_{i,t}(\theta_{i,1}, \Theta_i)^2]$$
$$\times exp[-\frac{1}{2}B_{\theta_{i,1}}(\theta_{i,1} - \theta_{10})^2]exp[-\frac{1}{2}B_{\Theta_{i,1}}(\Theta_i - \Theta_0)^2],$$

$$(iv) \quad \sigma_i^2|\boldsymbol{v}, \boldsymbol{\Psi}_{-\sigma_i^2}, \boldsymbol{a}_0 \sim \mathcal{IG}((v_0 + n)/2, (\delta_0 + d_{i,1})/2),$$

$$(v) \quad \pi(\boldsymbol{a}_{i,0}|\boldsymbol{v}, \boldsymbol{\Psi}) \propto \prod_{t=1}^{n} exp[-(1/2\sigma_i^2)a_{i,t}(\boldsymbol{a}_{i,0})^2].$$

For the hyperparemeters, we have

$$(vi) \quad \phi_{10}|\boldsymbol{v}, \boldsymbol{\Psi}_{-\phi_{10}}, \boldsymbol{a}_0 \sim N(\frac{1}{I}\sum_{i=1}^{I}\phi_{i,1}, r_\phi/I),$$

$$(vii) \quad \Phi_0|\boldsymbol{v}, \boldsymbol{\Psi}_{-\Phi_0}, \boldsymbol{a}_0 \sim N(\frac{1}{I}\sum_{i=1}^{I}\Phi_i, r_\Phi/I),$$

$$(vi) \quad \theta_{10}|\boldsymbol{v}, \boldsymbol{\Psi}_{-\theta_{10}}, \boldsymbol{a}_0 \sim N(\frac{1}{I}\sum_{i=1}^{I}\theta_{i,1}, r_\theta/I),$$

$$(vii) \quad \Theta_0|\boldsymbol{v}, \boldsymbol{\Psi}_{-\Theta_0}, \boldsymbol{a}_0 \sim N(\frac{1}{I}\sum_{i=1}^{I}\Theta_i, r_\Theta/I),$$

$$(viii) \quad r_\phi|\boldsymbol{v}, \boldsymbol{\Psi}_{-r_\phi}, \boldsymbol{a}_0 \sim \mathcal{IG}((b + I)/2, (a + \sum_{i=1}^{I}(\phi_{i,1} - \phi_{10})^2/2),$$

$$(ix) \quad r_\Phi|\boldsymbol{v}, \boldsymbol{\Psi}_{-r_\Phi}, \boldsymbol{a}_0 \sim \mathcal{IG}((d+I)/2, (c+\sum_{i=1}^{I}(\Phi_i - \Phi_0)^2/2),$$

$$(x) \quad r_\theta|\boldsymbol{v}, \boldsymbol{\Psi}_{-r_\theta}, \boldsymbol{a}_0 \sim \mathcal{IG}((f+I)/2, (e+\sum_{i=1}^{I}(\theta_{i,1} - \theta_{10})^2/2),$$

$$(xi) \quad r_\Theta|\boldsymbol{v}, \boldsymbol{\Psi}_{-r_\Theta}, \boldsymbol{a}_0 \sim \mathcal{IG}((h+I)/2, (g+\sum_{i=1}^{I}(\Theta_i - \Theta_0)^2/2),$$

where we let $B_{i,1} = B_{\phi_i} + \sigma_i^{-2}\bar{\boldsymbol{x}}_i^T\bar{\boldsymbol{x}}_i$, $B_{i,2} = B_{\Phi_i} + \sigma_i^{-2}\tilde{\boldsymbol{x}}_i^T\tilde{\boldsymbol{x}}_i$, and $d_{i,1} = ||\bar{\boldsymbol{y}}_i - \bar{\boldsymbol{x}}_i\phi_{i,1}||^2$, $i = 1, \ldots, I$. The quantities of $\bar{\boldsymbol{x}}_i$, $\tilde{\boldsymbol{x}}_i$, $\bar{\boldsymbol{y}}_i$, and $\tilde{\boldsymbol{y}}_i$ are defined above; see *D*efinitions 1 and 2.

Proof. (i)-(v) follow from *Proposition 2*, and (vi)-(xi) follow from (3.4.11), (3.4.12), and the definition of the full conditional distribution.

As mentioned above, we estimate the vector $\boldsymbol{\Psi}$ defined in section 3.3 by the posterior means, and measure the uncertainty in the estimates of $\boldsymbol{\Psi}$ by the posterior standard deviation. To generate samples from the posterior distribution using the MCMC method via Gibbs sampling, we block the parameters into the following groups: $(\boldsymbol{\phi})$, $(\boldsymbol{\Phi})$, (σ^2), $(\boldsymbol{\theta}, \Theta)$, $(\mathbf{a_0})$, and the hyperparameters (ϕ_{10}), (r_{ϕ_1}), (Φ_0), (r_Φ), (θ_{10}), (r_{θ_1}), (Θ_0), and (r_Θ). Within each Gibbs iteration, we draw samples of $(\phi_{1,i})$, (Φ_i) and (σ_i^2) from the Normal distribution and the inverse Gamma distribution described above, and use draws based on a Metropolis-Hastings for the groups of $(\theta_{i,1}, \Theta_i)$ and $(\boldsymbol{a}_{i,0})$ for each ith state. For (θ_i, Θ_i) and $(\boldsymbol{a}_{i,0})$, we use a Gaussian proposal as in the previous section. Conditioned on the currently drawn parameter values of $(\boldsymbol{\phi})$, $(\boldsymbol{\Phi})$, (σ^2), $(\boldsymbol{\theta}, \Theta)$, and $(\mathbf{a_0})$, we generate samples of the hyperparameters (ϕ_{10}), (Φ_0), (θ_{10}), and (Θ_0) from the normal distribution and (r_{ϕ_1}), (r_Φ), (r_{θ_1}), and (r_Θ) from the inverse Gamma distribution.

3.5 DATA ANALYSIS OF UE RATES DATA

In this section, we fit the Bayesian SARMA model proposed in section 3.4, which includes parameter estimation, prediction, model adequacy and forecasting of UE rates data.

Posterior features

In order to estimate the parameter space ($\boldsymbol{\Psi}$, $\boldsymbol{a}_0$) and measure the uncertainty in the estimates, we use the posterior means and standard deviations, respectively, computed through Gibbs sampler. All the full conditional distributions are described in *Proposition 3*, and their implementation is explained at the end of section 3.4.

To conduct the Gibbs sampler, we choose the values of a, b, c, d, e, f, g, h, v_0, and δ_0 to be $b = d = f = h = 1$, $a = c = d = g = 1$, $v_0 = 0$, and $\delta_0 = 0$. We also tried other values for these quantities and found that the results remain unchanged. Also, in our computation, we considered 10,000 replications. To reduce the effects of initial values on the final results, we delete the first 500 replications as "burning-in" samples. Moreover, to diminish the serial correlation of the run, we retain every 10th sample out of the remaining replications.

Estimating values of the hyperparameters ϕ_{10}, r_{ϕ_1}, Φ_0, r_Φ, θ_{10}, r_{θ_1}, Θ_0, and r_Θ are presented in Table 3.1, and those for parameters $\boldsymbol{\Psi}$ for each state are summarized in Table 3.2 to Table 3.6. We fit the proposed model for the first 112 observations of UE rates for each state. The last four observations which are not part of model fitting are used for subsequent forecast evaluation.

Table 3.1: Summary of fitted hyperparameters

	Posterior distribution				
hyperparemeter	Mean	Std. Dev.	Median	Lower 95% limit	Upper 95% limit
ϕ_{10}	0.5053009	0.0431541	0.5052150	0.4320100	0.5750100
r_ϕ	13.9587299	3.1702190	13.6671100	9.1450600	19.4364500
Φ_0	0.0214690	0.0359463	0.0209400	-0.0352200	0.0817800
r_Φ	23.9892970	5.2850984	23.3241700	16.1864600	33.1407800
θ_{10}	0.1616375	0.0389555	0.1619150	0.0984400	0.2271300
r_θ	18.9767827	4.6651724	18.4747850	12.2883400	27.2483100
Θ_0	-0.0244263	0.0297262	-0.0218400	-0.0752100	0.0208600
r_Θ	36.5458224	9.1574134	36.1650500	21.8570700	53.0459700

Model adequacy

Here, we assess the model adequacy for the unemployment rates data. From the Bayesian

Table 3.2: Summary of fitted parameter $\phi_{1,i}$

	Posterior distribution				
State	Mean	Std. Dev.	Median	Lower 95% limit	Upper 95% limit
Alabama	0.49590	0.10891	0.50406	0.29662	0.66623
Alaska	0.36324	0.13267	0.360245	0.1441	0.58781
Arizona	0.53694	0.09554	0.543025	0.36694	0.68675
Arkansas	0.51687	0.12139	0.522585	0.31348	0.70773
California	0.71216	0.07578	0.71627	0.57611	0.82972
Colorado	0.32236	0.12805	0.326865	0.10453	0.52154
Connecticut	0.50828	0.12648	0.507945	0.29774	0.70968
Delaware	0.41945	0.11929	0.42102	0.22534	0.60543
District of Columbia	0.86094	0.05243	0.86421	0.76482	0.94348
Florida	0.42107	0.14949	0.434575	0.15673	0.65359
Georgia	0.28458	0.11984	0.28356	0.08827	0.48051
Hawaii	0.37667	0.19327	0.385705	0.03659	0.66993
Idaho	0.85649	0.05228	0.859305	0.76927	0.94062
Illinois	0.54510	0.10007	0.55068	0.3736	0.7065
Indiana	0.29019	0.17081	0.299825	-0.005	0.5473
Iowa	0.56921	0.14635	0.580115	0.30451	0.78809
Kansas	0.42691	0.11895	0.431705	0.22137	0.61377
Kentucky	0.55402	0.10108	0.56075	0.38135	0.71098
Louisiana	0.46825	0.10673	0.47131	0.28484	0.63827
Maine	0.19314	0.13287	0.194885	-0.02387	0.40489
Maryland	0.42169	0.14027	0.42211	0.18706	0.6567
Massachusetts	0.81690	0.05790	0.815975	0.72065	0.91311
Michigan	0.53447	0.10034	0.539755	0.36877	0.69396
Minnesota	0.53672	0.10627	0.53933	0.35757	0.70101
Mississippi	0.34549	0.12021	0.350625	0.12954	0.53174
Missouri	0.52824	0.09377	0.5327	0.36322	0.67622
Montana	0.76367	0.06664	0.76475	0.65318	0.87578
Nebraska	0.20881	0.15759	0.207035	-0.04686	0.46981
Nevada	0.88749	0.046143	0.88719	0.81073	0.96107
New Hampshire	0.74002	0.08002	0.74101	0.60214	0.86269
New Jersey	0.61734	0.10238	0.61615	0.4483	0.77896
New Mexico	0.84781	0.05014	0.84839	0.76424	0.93033
New York	0.57878	0.10636	0.58665	0.39188	0.73945
North Carolina	0.39870	0.14213	0.407445	0.1537	0.62157
North Dakota	0.25252	0.15928	0.25398	-0.00155	0.50874
Ohio	0.45864	0.12072	0.46397	0.25567	0.64327
Oklahoma	0.25687	0.12246	0.2526	0.06272	0.45544
Oregon	0.35882	0.12723	0.362885	0.13837	0.5515
Pennsylvania	0.61735	0.10033	0.62471	0.44657	0.77199
Puerto Rico	0.11984	0.21750	0.09943	-0.21917	0.48399
Rhode Island	0.90959	0.03943	0.911085	0.84368	0.97234
South Carolina	0.45707	0.13223	0.464385	0.2335	0.67106
South Dakota	0.28624	0.14877	0.29245	0.02806	0.51394
Tennessee	0.91504	0.04009	0.917145	0.84826	0.97717
Texas	0.68325	0.07368	0.6873	0.56149	0.80019
Utah	0.57725	0.09949	0.582015	0.40392	0.73603
Vermont	0.25123	0.11417	0.249095	0.0727	0.43986
Virginia	0.40976	0.09917	0.416205	0.22876	0.5652
Washington	0.55921	0.10563	0.561615	0.37264	0.72441
West Virginia	0.58996	0.09370	0.592175	0.43027	0.73669
Wisconsin	0.34739	0.13434	0.350865	0.1099	0.55874
Wyoming	0.29993	0.12360	0.30274	0.08662	0.51082

Table 3.3: Summary of fitted parameter $\phi_{4,i}$

	Posterior distribution				
State	Mean	Std. Dev.	Median	Lower 95% limit	Upper 95% limit
Alabama	0.02005	0.10443	0.01653	-0.14986	0.1981
Alaska	0.08094	0.09718	0.0822	-0.07432	0.23826
Arizona	-0.16780	0.08842	-0.16655	-0.31459	-0.02174
Arkansas	0.03534	0.10787	0.029925	-0.13788	0.21574
California	0.04691	0.11114	0.047175	-0.13168	0.22597
Colorado	0.06594	0.10622	0.06528	-0.10116	0.24177
Connecticut	0.12062	0.10656	0.11624	-0.0542	0.30874
Delaware	0.06582	0.11917	0.06006	-0.13067	0.27045
District of Columbia	0.01736	0.11531	0.01657	-0.1706	0.20454
Florida	0.09342	0.09037	0.09419	-0.05561	0.24371
Georgia	0.02354	0.09956	0.02474	-0.13241	0.19055
Hawaii	0.11008	0.10844	0.106305	-0.05831	0.29053
Idaho	-0.16214	0.10933	-0.161475	-0.34249	0.01038
Illinois	-0.12089	0.11212	-0.127295	-0.28457	0.06643
Indiana	0.06022	0.10797	0.055275	-0.10671	0.24491
Iowa	0.08535	0.12451	0.08504	-0.1178	0.28831
Kansas	-0.08655	0.12246	-0.087955	-0.287	0.11742
Kentucky	-0.03561	0.09903	-0.038265	-0.19474	0.12874
Louisiana	-0.00775	0.10407	-0.008315	-0.17522	0.16377
Maine	0.24474	0.09015	0.24593	0.08942	0.39345
Maryland	0.09001	0.10567	0.090615	-0.0898	0.25956
Massachusetts	0.13386	0.08664	0.1314	0.00032	0.27556
Michigan	0.01193	0.10764	0.00798	-0.1646	0.19431
Minnesota	-0.00867	0.10888	-0.0134	-0.18133	0.17366
Mississippi	0.03415	0.10048	0.035495	-0.13016	0.20705
Missouri	0.11949	0.11100	0.12125	-0.06173	0.29831
Montana	0.03784	0.11416	0.037605	-0.14143	0.218
Nebraska	0.18182	0.10184	0.182175	0.02228	0.34897
Nevada	-0.22885	0.12013	-0.22873	-0.42865	-0.032
New Hampshire	-0.02493	0.09696	-0.025235	-0.18195	0.14473
New Jersey	-0.08125	0.11099	-0.080125	-0.2605	0.10272
New Mexico	-0.06304	0.10639	-0.066785	-0.23775	0.12659
New York	0.03927	0.10053	0.033895	-0.11487	0.20793
North Carolina	-0.13606	0.09979	-0.140685	-0.30002	0.02822
North Dakota	-0.00353	0.10349	-0.00141	-0.17682	0.16077
Ohio	0.06163	0.09314	0.058915	-0.08827	0.21025
Oklahoma	-0.02661	0.09735	-0.027465	-0.17892	0.14059
Oregon	-0.01513	0.09191	-0.015375	-0.16554	0.13638
Pennsylvania	0.06341	0.08889	0.06308	-0.08099	0.20689
Puerto Rico	-0.04433	0.09956	-0.04023	-0.20215	0.10672
Rhode Island	-0.02517	0.10482	-0.02738	-0.18984	0.14796
South Carolina	-0.08415	0.09299	-0.08701	-0.23351	0.06553
South Dakota	-0.17164	0.12199	-0.168485	-0.36481	0.03167
Tennessee	-0.05971	0.09391	-0.0678	-0.20274	0.10931
Texas	-0.02824	0.12354	-0.02731	-0.24128	0.17371
Utah	0.08518	0.10651	0.08248	-0.08243	0.26742
Vermont	0.36707	0.08529	0.36608	0.22872	0.50643
Virginia	0.15370	0.09751	0.15362	-0.01013	0.31309
Washington	0.09098	0.11684	0.08273	-0.10229	0.27389
West Virginia	0.04726	0.11839	0.045075	-0.14616	0.24841
Wisconsin	-0.03340	0.10857	-0.032095	-0.21269	0.14442
Wyoming	0.26834	0.10716	0.27281	0.09593	0.43293

Table 3.4: Summary of fitted parameter $\theta_{1,i}$

	Posterior distribution				
State	Mean	Std. Dev.	Median	Lower 95% limit	Upper 95% limit
Alabama	0.24267	0.12208	0.23855	0.038	0.44963
Alaska	0.09508	0.13888	0.099065	-0.13152	0.30431
Arizona	0.20830	0.10100	0.211145	0.0344	0.36637
Arkansas	0.07339	0.14638	0.07127	-0.15546	0.31452
California	0.18792	0.09450	0.18275	0.04732	0.359
Colorado	0.23060	0.12939	0.23174	0.01021	0.43378
Connecticut	0.07678	0.14226	0.092865	-0.18075	0.27925
Delaware	0.21777	0.12927	0.21906	-0.0101	0.41911
District of Columbia	0.17197	0.08761	0.172875	0.02403	0.31287
Florida	-0.04460	0.15558	-0.051355	-0.28507	0.22236
Georgia	0.24743	0.11607	0.254445	0.0447	0.43333
Hawaii	-0.15431	0.20863	-0.16564	-0.4802	0.20686
Idaho	0.33174	0.07567	0.336005	0.20226	0.44604
Illinois	0.18985	0.11263	0.191985	-0.0009	0.37642
Indiana	0.04953	0.18216	0.033065	-0.21563	0.3718
Iowa	-0.15973	0.14772	-0.15971	-0.40647	0.08677
Kansas	0.09887	0.11185	0.10232	-0.08017	0.28849
Kentucky	0.10609	0.10284	0.10461	-0.058	0.28179
Louisiana	0.13051	0.10062	0.132835	-0.03659	0.29566
Maine	0.34869	0.14937	0.35177	0.10549	0.57985
Maryland	-0.05223	0.13446	-0.03958	-0.29545	0.14715
Massachusetts	0.36581	0.07823	0.36945	0.23361	0.48815
Michigan	0.04775	0.09451	0.043665	-0.09644	0.20404
Minnesota	0.04455	0.11018	0.04469	-0.13537	0.23374
Mississippi	0.21240	0.12650	0.214875	0.01318	0.41569
Missouri	0.12893	0.08869	0.132515	-0.02088	0.27207
Montana	0.30606	0.09298	0.31309	0.14667	0.45366
Nebraska	0.19918	0.17350	0.21874	-0.10433	0.46156
Nevada	0.38138	0.06482	0.386255	0.27049	0.48399
New Hampshire	0.16496	0.10711	0.16277	-0.00661	0.35017
New Jersey	0.22483	0.13979	0.229925	-0.01587	0.44748
New Mexico	0.32451	0.06871	0.32564	0.21128	0.43637
New York	0.12508	0.12131	0.124295	-0.0748	0.32654
North Carolina	0.24962	0.16445	0.24547	-0.01592	0.52093
North Dakota	-0.00057	0.16625	0.00417	-0.27123	0.26038
Ohio	0.08159	0.12931	0.078115	-0.13156	0.29412
Oklahoma	0.40873	0.14202	0.41683	0.163	0.62153
Oregon	0.24621	0.12308	0.241995	0.05248	0.45469
Pennsylvania	0.00803	0.09955	0.01143	-0.15883	0.17242
Puerto Rico	0.01165	0.19366	0.04108	-0.32525	0.29503
Rhode Island	0.22016	0.06104	0.221525	0.12006	0.32149
South Carolina	0.16102	0.14968	0.14713	-0.0728	0.43094
South Dakota	0.17779	0.15579	0.175735	-0.07277	0.44464
Tennessee	0.31086	0.06889	0.31093	0.19465	0.42603
Texas	0.44782	0.07968	0.45131	0.30857	0.57333
Utah	0.09270	0.10076	0.09871	-0.07054	0.24507
Vermont	0.43144	0.11171	0.441475	0.2398	0.59848
Virginia	0.08228	0.07148	0.07713	-0.01994	0.20563
Washington	0.02144	0.11076	0.021405	-0.16755	0.19172
West Virginia	-0.04306	0.08936	-0.045225	-0.18647	0.09756
Wisconsin	0.14729	0.14340	0.14158	-0.07841	0.4027
Wyoming	0.22101	0.12644	0.22087	0.00747	0.42821

Table 3.5: Summary of fitted parameter $\theta_{4,i}$

	Posterior distribution				
State	Mean	Std. Dev.	Median	Lower 95% limit	Upper 95% limit
Alabama	-0.02181	0.06964	-0.010625	-0.16561	0.06793
Alaska	0.00203	0.06113	0.005215	-0.08713	0.08572
Arizona	0.00237	0.01068	0.00151	-0.00197	0.00939
Arkansas	0.00178	0.07048	0.004435	-0.1279	0.11018
California	-0.20070	0.09058	-0.19471	-0.35912	-0.07034
Colorado	0.00451	0.07264	-0.00654	-0.10652	0.13194
Connecticut	-0.07653	0.07390	-0.06834	-0.21146	0.03185
Delaware	0.00817	0.09801	0.01286	-0.15264	0.16311
District of Columbia	-0.04360	0.10528	-0.041255	-0.21115	0.12425
Florida	-0.01105	0.03773	-0.006825	-0.03661	0.02047
Georgia	-0.00085	0.06406	-0.00308	-0.0898	0.11637
Hawaii	-0.03185	0.07867	-0.010245	-0.21453	0.03692
Idaho	-0.11548	0.10191	-0.101775	-0.29424	0.03352
Illinois	-0.01886	0.08671	0.004	-0.20846	0.06259
Indiana	0.00096	0.07569	0.003165	-0.14698	0.10706
Iowa	0.01032	0.09577	0.011275	-0.15113	0.16135
Kansas	-0.04279	0.08697	-0.039215	-0.19353	0.0981
Kentucky	-0.00530	0.04847	-0.00253	-0.05153	0.04402
Louisiana	-0.00644	0.06123	-0.00802	-0.09338	0.08476
Kentucky	-0.00530	0.04847	-0.00253	-0.05153	0.04402
Louisiana	-0.00644	0.06123	-0.00802	-0.09338	0.08476
Maine	0.01105	0.04188	0.007925	-0.05164	0.08587
Maryland	0.01756	0.09084	0.014695	-0.1376	0.18011
Massachusetts	0.00959	0.04856	-0.00375	-0.01072	0.12422
Michigan	0.01009	0.07145	0.01267	-0.10446	0.11521
Minnesota	-0.10706	0.07394	-0.09425	-0.22897	-0.01467
Mississippi	-0.01134	0.06604	-0.00379	-0.15638	0.07949
Missouri	-0.07278	0.09157	-0.068025	-0.22791	0.06835
Montana	-0.04216	0.09732	-0.043945	-0.19626	0.11338
Nebraska	-0.02893	0.06180	-0.01714	-0.1517	0.04616
Nevada	0.05127	0.10092	0.051625	-0.10845	0.21951
New Hampshire	0.00802	0.04432	0.00457	-0.05609	0.09065
New Jersey	-0.03065	0.07829	-0.024225	-0.16531	0.07847
New Mexico	-0.28551	0.10661	-0.27988	-0.4668	-0.11593
New York	-0.02802	0.07473	-0.0093	-0.21802	0.01633
North Carolina	-0.02935	0.05521	-0.02013	-0.13297	0.02464
North Dakota	0.01222	0.07175	0.018625	-0.11004	0.13834
Ohio	-0.00391	0.04675	-0.004065	-0.08096	0.07263
Oklahoma	-0.01441	0.05076	-0.00674	-0.09783	0.04437
Oregon	0.00523	0.02965	0.001475	-0.03728	0.06372
Pennsylvania	-0.00317	0.02890	-0.009525	-0.03207	0.05799
Puerto Rico	0.00599	0.07439	-0.01448	-0.07515	0.17298
Rhode Island	-0.10036	0.09226	-0.09117	-0.2597	0.03296
South Carolina	-0.00897	0.03384	-0.001025	-0.05864	0.00804
South Dakota	0.06857	0.08687	0.067635	-0.07072	0.20878
Tennessee	-0.14749	0.07315	-0.1433	-0.26936	-0.03474
Texas	-0.01697	0.10213	-0.013555	-0.18275	0.14664
Utah	-0.04113	0.07245	-0.017015	-0.19523	0.03189
Vermont	0.00350	0.02543	0.00372	-0.03056	0.03854
Virginia	-0.06662	0.06227	-0.05691	-0.18433	0.01764
Washington	0.08076	0.10046	0.083615	-0.09553	0.23353
West Virginia	0.04682	0.08746	0.04089	-0.08579	0.19301
Wisconsin	-0.05356	0.07361	-0.037035	-0.19355	0.03566
Wyoming	0.01721	0.05938	0.00529	-0.06784	0.11409

Table 3.6: Summary of fitted parameter σ_i^2

	Posterior distribution				
State	Mean	Std. Dev.	Median	Lower 95% limit	Upper 95% limit
Alabama	0.09479	0.01316	0.09302	0.07601	0.11967
Alaska	0.11432	0.01693	0.112845	0.08895	0.14551
Arizona	0.15937	0.02121	0.157625	0.12814	0.19687
Arkansas	0.03637	0.00533	0.035915	0.02836	0.0459
California	0.04154	0.00591	0.040895	0.0329	0.05235
Colorado	0.16001	0.02239	0.158295	0.12694	0.20006
Connecticut	0.07983	0.01097	0.078865	0.06349	0.09984
Delaware	0.03869	0.00540	0.038195	0.03116	0.04858
District of Columbia	0.01421	0.00196	0.014055	0.01128	0.01766
Florida	0.09544	0.01267	0.094685	0.07637	0.11897
Georgia	0.07640	0.01108	0.0756	0.05969	0.09631
Hawaii	0.08606	0.01187	0.08509	0.0683	0.10747
Idaho	0.00627	0.00086	0.00619	0.00501	0.0077
Illinois	0.09299	0.01321	0.091535	0.07413	0.11688
Indiana	0.20843	0.02885	0.20578	0.16619	0.26018
Iowa	0.03867	0.00521	0.03823	0.03108	0.04805
Kansas	0.05623	0.00745	0.055655	0.04552	0.06942
Kentucky	0.07460	0.01039	0.07354	0.05887	0.09327
Louisiana	0.14789	0.01999	0.14653	0.11634	0.18236
Maine	0.12322	0.01661	0.121655	0.09818	0.15305
Maryland	0.07091	0.00990	0.070355	0.05608	0.08788
Massachusetts	0.01669	0.00228	0.016515	0.01331	0.02081
Michigan	0.18843	0.02622	0.18564	0.15067	0.23612
Minnesota	0.06051	0.00848	0.05957	0.04844	0.07638
Mississippi	0.18744	0.02576	0.18525	0.15077	0.23235
Missouri	0.07929	0.01126	0.07839	0.06261	0.0986
Montana	0.00499	0.00068	0.00492	0.00395	0.00619
Nebraska	0.07090	0.01008	0.070065	0.05519	0.0889
Nevada	0.01062	0.00146	0.01048	0.00851	0.0133
New Hampshire	0.03650	0.00509	0.03605	0.0291	0.04557
New Jersey	0.04912	0.00657	0.04834	0.03934	0.06117
New Mexico	0.00894	0.00129	0.0088	0.00702	0.0112
New York	0.04863	0.00680	0.048	0.03836	0.06017
North Carolina	0.11562	0.01603	0.114285	0.09195	0.14399
North Dakota	0.06434	0.00886	0.063625	0.05124	0.07924
Ohio	0.14781	0.02030	0.145665	0.11816	0.18409
Oklahoma	0.13604	0.01996	0.133755	0.10814	0.17252
Oregon	0.11860	0.01577	0.117025	0.09451	0.14716
Pennsylvania	0.07307	0.00979	0.072395	0.05875	0.09079
Puerto Rico	0.56225	0.07748	0.55387	0.44721	0.69676
Rhode Island	0.01008	0.00135	0.0099	0.00809	0.01242
South Carolina	0.12841	0.01782	0.12679	0.1031	0.16061
South Dakota	0.02528	0.00338	0.024975	0.02013	0.03134
Tennessee	0.00798	0.00112	0.00785	0.00638	0.01001
Texas	0.02380	0.00339	0.02334	0.01905	0.02991
Utah	0.06126	0.00870	0.0606	0.04808	0.07629
Vermont	0.10327	0.01448	0.101695	0.08214	0.12935
Virginia	0.06454	0.00917	0.0636	0.05124	0.08019
Washington	0.07289	0.00994	0.07208	0.05811	0.09055
West Virginia	0.19273	0.02659	0.19171	0.15157	0.23759
Wisconsin	0.14817	0.01995	0.147185	0.11785	0.18314
Wyoming	0.21072	0.02943	0.208745	0.16721	0.26426

perspective, model assessment may be carried out using predictive distributions, which amounts to comparing the predicted values based on the model with the actual observation. In general, prediction proceeds through the predictive density (Aitkin, 1991). Particularly, let $\boldsymbol{v}_{obs}$ and $\boldsymbol{v}_{new}$ be the observed and the generated data. Then the predictive density for $\boldsymbol{v}_{new}$ is

$$f(\boldsymbol{v}_{new}|\boldsymbol{v}_{obs}) = \int f(\boldsymbol{v}_{new}|\boldsymbol{v}_{obs}, \boldsymbol{\Psi}, \boldsymbol{a}_0)\pi(\boldsymbol{\Psi}, \boldsymbol{a}_0|\boldsymbol{v}_{obs})d\boldsymbol{\Psi}, d\boldsymbol{a}_0. \tag{3.5.13}$$

Note that the expression (3.5.13) fits natually within our sampling based approach. Let Ψ_l and $\boldsymbol{a}_{0,l}$ denote draws from the posterior $\pi(\boldsymbol{\Psi}, \boldsymbol{a}_0|\boldsymbol{v}_{obs})$ for $l = 1, \ldots, B$, where B is the total number of Gibbs iterations. Then a sample of predictions from (3.5.13) is created by drawing a $\boldsymbol{v}_{new}^{(l)}$ from $f(\boldsymbol{v}_{new}|\boldsymbol{v}_{obs}, \boldsymbol{\Psi}_l, \boldsymbol{a}_{0,l})$ for each $\boldsymbol{\Psi}_l$ and $\boldsymbol{a}_{0,l}$, for $l = 1, \ldots, B$.

To check how well a given model fits the present data, we compare $\boldsymbol{v}_{old}$ and $\boldsymbol{v}_{new}$ by using a discrepancy measure denoted by $d(\boldsymbol{v}, \hat{\boldsymbol{v}})$ (Shinha and Dey (1997)). A discrepancy measure is a scalar summary of parameters and data that is used as a standard when comparing data to predictive simulations. If a model adequately fits the observed data, then the generated new data should be similar to the observed data, and the two values of the discrepancy measure, $d(\boldsymbol{v}_{old}, \hat{\boldsymbol{v}})$ and $d(\boldsymbol{v}_{new}, \hat{\boldsymbol{v}})$ are similar. In other words, if the model adequately fits the data, the posterior predictive p value, $P\{d(\boldsymbol{v}_{old}, \hat{\boldsymbol{v}}) \geq d(\boldsymbol{v}_{new}, \hat{\boldsymbol{v}})|\boldsymbol{v}_{old}\}$ is expected to be around 0.5. Hence, the p-value is defined as the probability that the generated data could be more extreme than the observed data, as measured by the discrepancy measure.

To implement the posterior predictive p value using the MCMC output, we generate a $\boldsymbol{v}_{new}^{(l)}$ from $f(\boldsymbol{v}_{new}^{(l)}|\boldsymbol{v}_{obs}, \boldsymbol{\Psi}_l, \boldsymbol{a}_{0,l})$ for each $\boldsymbol{\Psi}_l$ and $\boldsymbol{a}_{0,l}$, and compute $d(\boldsymbol{v}_{old}, \hat{\boldsymbol{v}})$ and $d(\boldsymbol{v}_{new}^{(l)}, \hat{\boldsymbol{v}})$ for $l = 1, \ldots, B$. Then, these calculated samples can be used to approximate $P\{d(\boldsymbol{v}_{old}, \hat{\boldsymbol{v}}) \geq d(\boldsymbol{v}_{new}, \hat{\boldsymbol{v}})|\boldsymbol{y}_{old}\}$ by $B^{-1}\sum_{l=1}^{B} I\{d(\boldsymbol{v}_{old}, \hat{\boldsymbol{v}}) \geq d(\boldsymbol{v}_{new}^{(l)}, \hat{\boldsymbol{v}})\}$, where $I(.)$ is an indicator function. A value close to 0.5 indicates an adequate model, while for an inadequate model, this value is near 0 or 1. The discrepancy measure we use for our model is given by

$$d(\boldsymbol{v}, \hat{\boldsymbol{v}}) = \sum_{i=1}^{52}\sum_{t=1}^{112} \sigma_i^{-1}(v_{i,t} - \hat{v}_{i,t}). \tag{3.5.14}$$

The estimated value for our proposed model is 0.5055, which indicates the adequacy of the model.

Forecasting

Forecasting proceeds via the predictive density. Let $\boldsymbol{v}_{i,F} = (v_{i,n+1}, \ldots, v_{i,n+L})$ denote the L steps-ahead-forecast for the ith state and $\boldsymbol{v}_F = (\boldsymbol{v}_{1,F}, \ldots, \boldsymbol{v}_{I,F})$. Then, the predictive density for $\boldsymbol{v}_F$ is

$$f(\boldsymbol{v}_F|\boldsymbol{v}_{obs}) = \int f(\boldsymbol{v}_F|\boldsymbol{v}_{obs}, \boldsymbol{\Psi}, \boldsymbol{a}_0)\pi(\boldsymbol{\Psi}, \boldsymbol{a}_0|\boldsymbol{v}_{obs})d\boldsymbol{\Psi}, d\boldsymbol{a}_0. \tag{3.5.15}$$

where, by integration with respect to $\boldsymbol{a}_0$, we mean integration with respect to the elements that define $\boldsymbol{a}_0$ under the chosen model. Here, $f(\boldsymbol{v}_F|\boldsymbol{v}_{obs}, \boldsymbol{\Psi}, \boldsymbol{a}_0) = f(\boldsymbol{v}^{(n+1)}|\boldsymbol{v}_{obs}\ , \boldsymbol{\Psi}, \boldsymbol{a}_0)\cdot f(\boldsymbol{v}^{(n+2)}|\boldsymbol{v}^{(n+1)}, \boldsymbol{v}_{obs}, \boldsymbol{\Psi}, \boldsymbol{a}_0)\cdots\ f(\boldsymbol{v}^{(n+L)}|\boldsymbol{v}^{(n+1)}, \ldots, \boldsymbol{v}^{(n+L-1)}, \boldsymbol{v}_{obs}, \boldsymbol{\Psi}, \boldsymbol{a}_0)$.

In practice, if $\boldsymbol{\Psi}_{(l)}$ and $\boldsymbol{a}_{0,(l)}$ denote draws from the posterior, $\pi(\boldsymbol{\Psi}, \boldsymbol{a}_0|\boldsymbol{v}_{obs})$, $l = 1, \ldots, B$, then a sample of forecasts from equation (3.5.15) is created by drawing a $\boldsymbol{v}_{F,l}$ from $f(\boldsymbol{v}_F|\boldsymbol{v}_{obs}, \boldsymbol{\Psi}_{(l)}, \boldsymbol{a}_{0,(l)})$.

Returning to the unemployment rates, we look at the posteriors for the last four observations which are held out for forecasting, as described above, that is, $L = 4$, and $n = 112$. Figure 1.1 to Figure 1.3 present the forecasting features of the future data $\boldsymbol{v}_F$ under our proposed model. The 90% predictive intervals are shown for each state for the last four observations $v_{i,113}, \ldots, v_{i,116}$. The intervals capture the last four observations, and the forecasts up to four steps ahead are quite good.

As one way of model adequacy using forecasts, we recommend seeing the data we would obtain if the experiment that produced $\boldsymbol{v}_{obs}$ today were replicated with the same model and the same value of parameters that produced the observed data. That is, we compare the last four observations to what we expect to be observed under the same model and the estimates of $\boldsymbol{\Psi}$ and $\boldsymbol{a}_0$, which are used as the estimates given in Table 3.2 to Table 3.6. We simulate the 500 four-steps ahead forecasts for all the states under the same model and the estimates

of $\boldsymbol{\Psi}$ and $\boldsymbol{a}_0$. If the model adequately fits the data, the last four observations are expected not to be extreme around the generated forecasts.

For many problems, it is useful to examine graphical comparisons of summaries of the data to those from simulations, as in Figures below. Figure 3.4 to Figure 3.6 display 52 histograms of one-step ahead forecasts, each of which represents 500 draws under our proposed model and the estimates of $\boldsymbol{\Psi}$ and $\boldsymbol{a}_0$ for each state. Figure 3.7 to Figure 3.9, Figure 3.10 to Figure 3.12, and Figure 3.13 to Figure 3.15 show those of two-, three-, and four-step ahead, respectively. For comparison, the last four observations are indicated as the dot mark, and they are mostly centered in the histogram. Hence, our proposed model is adequate enough to provide reliable forecasts over all states.

Model Improvement

To evaluate the improvement of our proposed model on performance, an independent series seasonal ARIMA $(1,1,1)(1,0,1)_4$ model can be compared for each state. This model does not combine cross-sectional information across states. To perform the independent series model, we use the Bayesian algorithm given in section 3.3. Table 3.7 presents the mean and standard deviation of the marginal posterior distributions for the parameters and forecasts through the Gibbs sampler. For the evaluation of improvement, two different models are considered, that is, the independent series model and our proposed model, for two states Alabama (classified as an indirect-use state) and Massachusetts (classified as a direct-use state).

We note that, under our proposed model, the posterior means of ϕ_4 and θ_4 are more toward zero for both states. Also, for Alabama, the posterior distributions for all the parameters have smaller standard errors, while for Massachusetts, those for ϕ_1 and θ_1 have larger ones. With respect to forecasting, the posterior distributions have smaller standard errors under our proposed model than under the independent series for both states. As expected,

incorporating dependencies across states improves inferences and forecasts more for Alabama than for Massachusetts, because Alabama has fewer sample available than Massachusetts.

Table 3.7: Estimation Results of Alabama and Massachusetts Under the Independent Model and Our Suggested Model

		Posterior distribution			
		Independent Model		Our Suggested Model	
State	Parameter	Mean	Std. Dev.	Mean	Std. Dev.
Alabama	ϕ_1	0.5195	0.1219	0.49591	0.10891
	ϕ_4	0.01436	0.13519	0.02005	0.10443
	θ_1	0.20424	0.14224	0.24267	0.12208
	θ_4	0.20827	0.14351	-0.02181	0.06964
	σ^2	0.09564	0.01359	0.09479	0.0131
	One-step(5.6667)	5.91937	0.77027	5.8494	0.7360
	Two-step(5.6000)	5.93609	1.06841	5.8205	1.0020
	Three-step(5.5667)	5.95061	1.32009	5.7878	1.237
	Four-step(5.4000)	5.96133	1.54103	5.7515	1.453
Massa-chusetts	ϕ_1	0.8224562	0.0554	0.8169	0.5791
	ϕ_4	0.16834	0.09044	0.13386	0.08664
	θ_1	0.39182	0.06681	0.36581	0.07823
	θ_4	0.39514	0.06241	0.00959	0.04856
	σ^2	0.01685	0.00239	0.01669	0.002282
	One-step(5.8900)	5.88287	0.70606	5.890	0.6473
	Two-step(5.2000)	5.89454	1.04381	5.8937	0.961
	Three-step(5.0000)	5.90732	1.37125	5.8966	1.267
	Four-step(4.7667)	5.91934	1.68882	5.9050	1.562

3.6 CONCLUSION

In order to improve parameter estimation and forecasting of the state-level UE rate estimates, especially for the states which do not provide adequate sample sizes, we model the sample UE rate estimates as a SARIMA model. More specifically, we assume that each parameter corresponding to a state is randomly distributed with a common mean and variance. Our Bayesian algorithm which is obtained by combining and expanding the Bayesian analysis of Chib and Greenberg (1994) in ARIMA model and those of Marin *et al.* (2005) in AR model and MA model provides a useful way for fitting our proposed model.

Data analysis is carried out for the first 112 observations of UE rates for each state using our Bayesian algorithm based on our proposed model. The last four observations which are not part of modeling fitting are used for subsequent forecast evaluation. This Bayesian analysis presented here shows that the estimation of parameters fit nicely. To check how well our proposed model fits the UE rates, we compare the observed and generated data using a discrepancy measure. The estimated value, 0.5055 indicates that our proposed model is adequate. As one way of model adequacy, we see the actual last four observations with what we would obtain if the experiment was replicated with our proposed model and our estimated values of parameters that produced the observed data. These simulated 500 four-steps ahead forecasts are shown not to be extreme around the actual four observations. Once again, the 90% intervals for forecasting the last four observations for each state capture the actual values. Overall, our proposed model is adequate enough to provide reliable forecasts over all states. Comparison of an independent series SARIMA model and our proposed model on performance shows that incorporating dependencies across states improves inferences and forecasts more for an indirect-use state such as Alabama than a direct-use state such as Massachusetts because an indirect-use state has fewer sample sizes available than a direct-use state. It is shown here that our proposed model provides nice fitting of the UE rates for all states, and our Bayesian algorithm is useful to perform our proposed model.

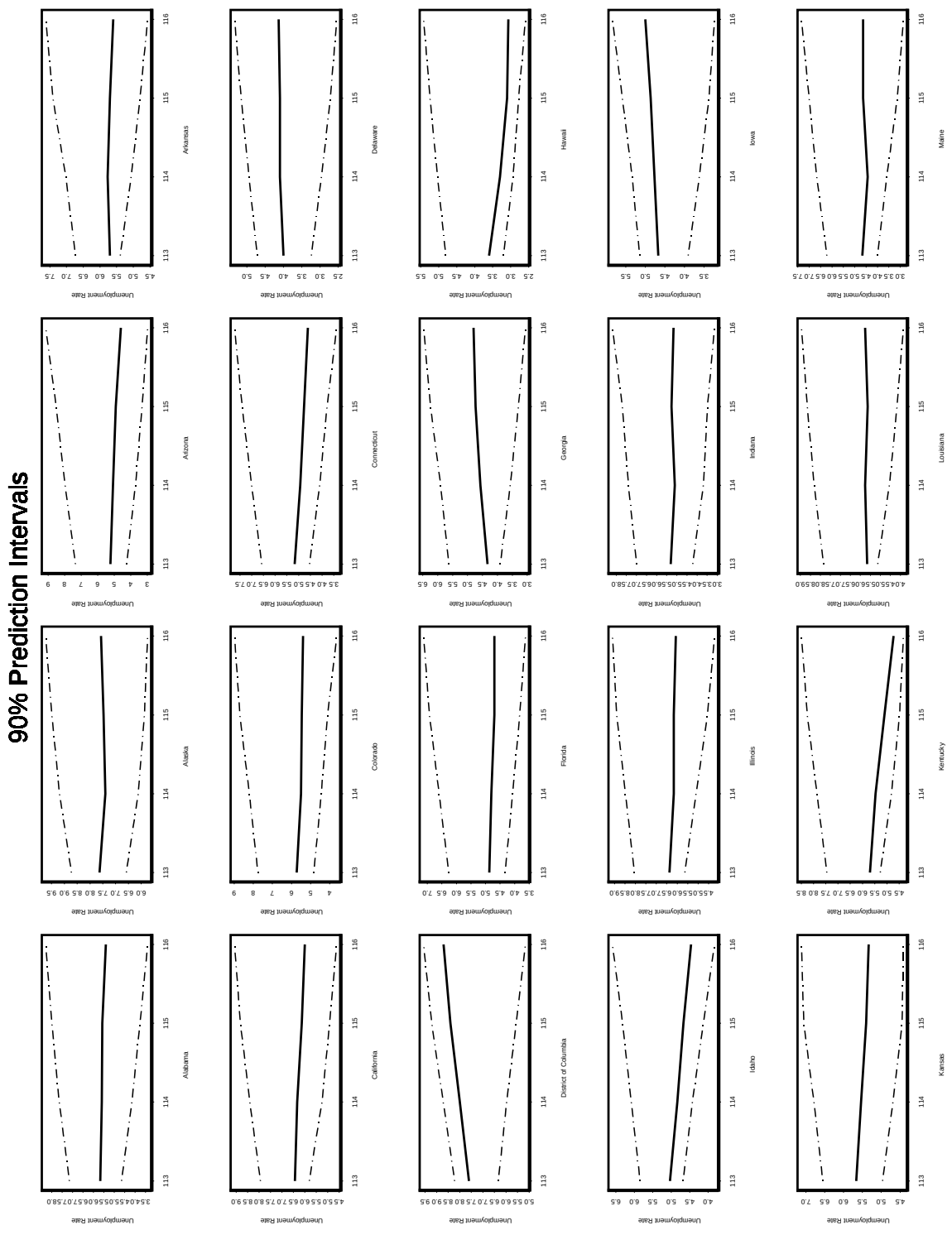

Figure 3.1: Four-quarter-ahead 90% forecast intervals compared with observed (solid line) from Alabama to Maine

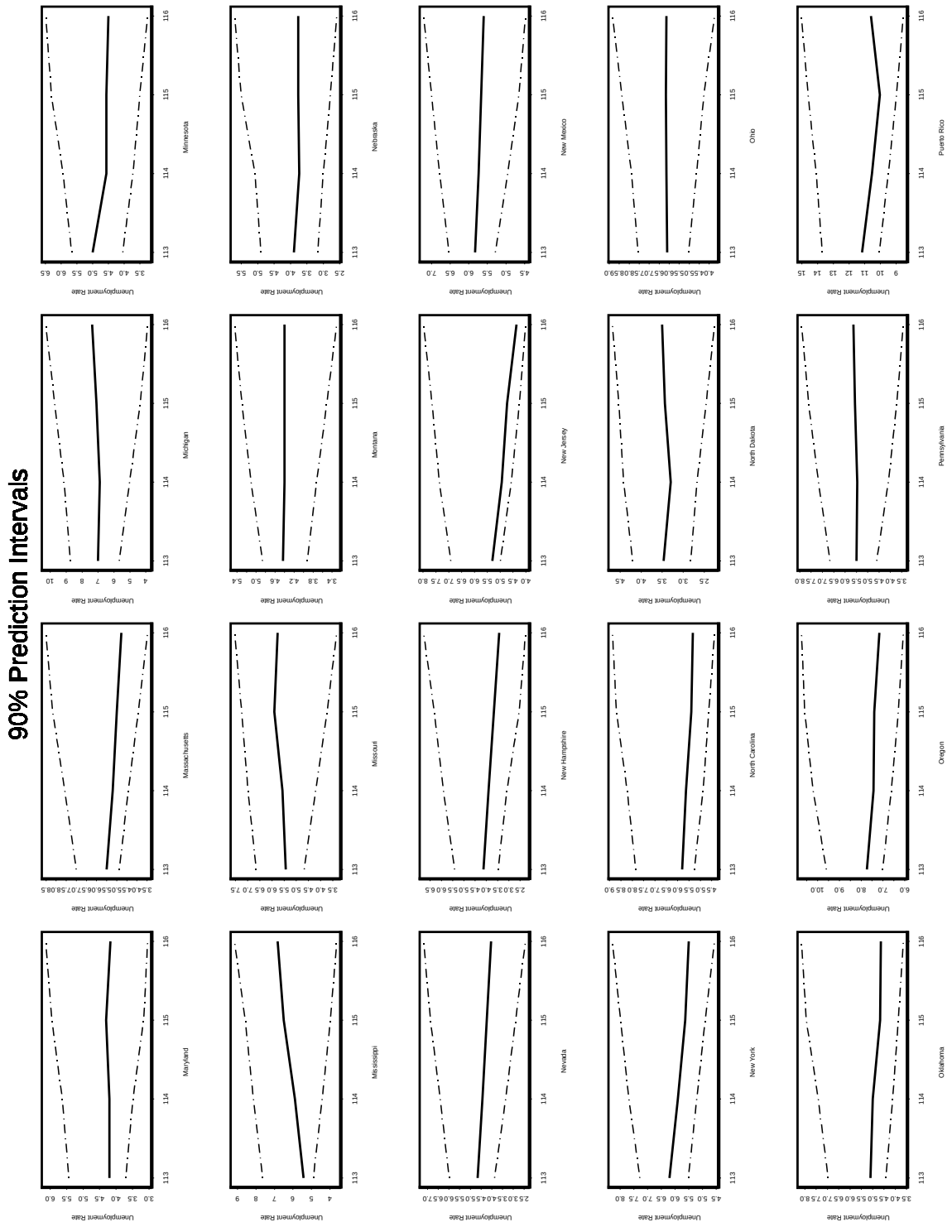

Figure 3.2: Four-quarter-ahead 90% forecast intervals compared with observed (solid line) from Maryland to PuertoRico

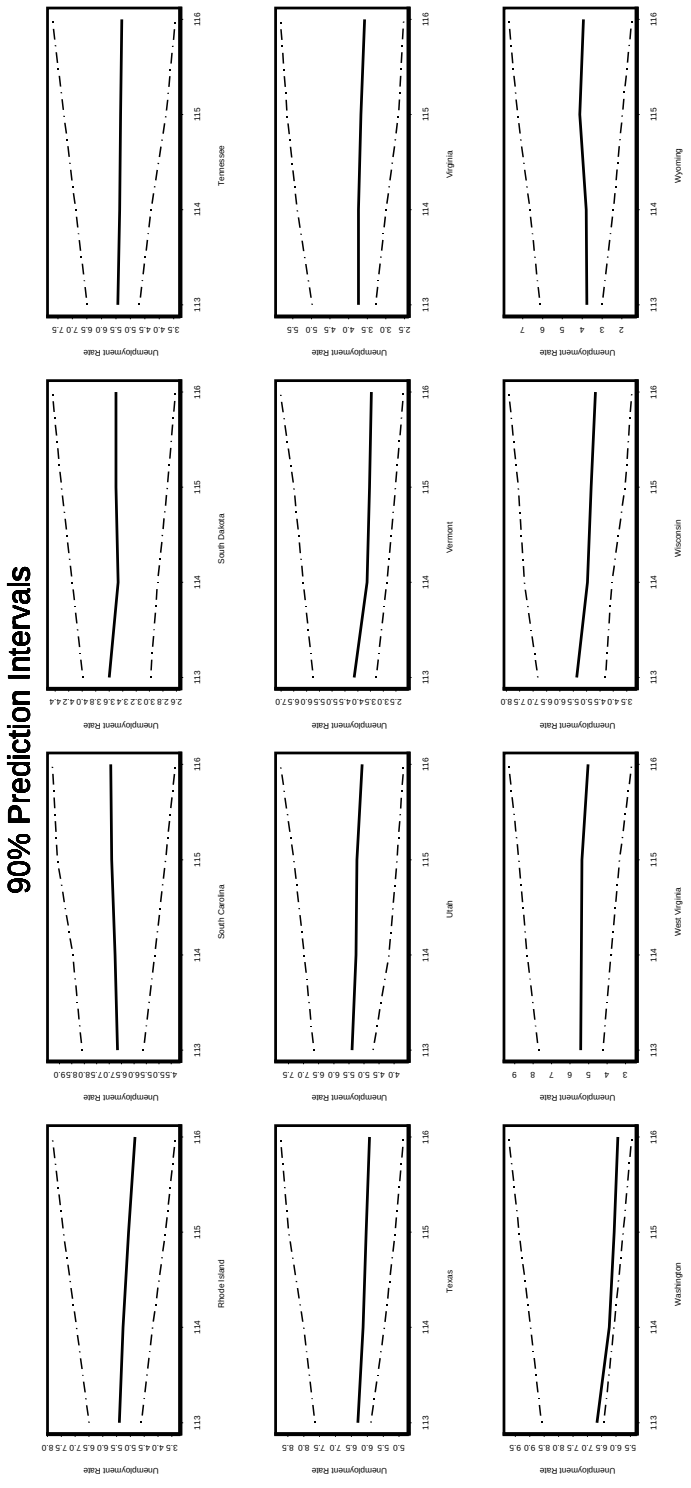

Figure 3.3: Four-quarter-ahead 90% forecast intervals compared with observed (solid line) from Rhode Island to Wyoming

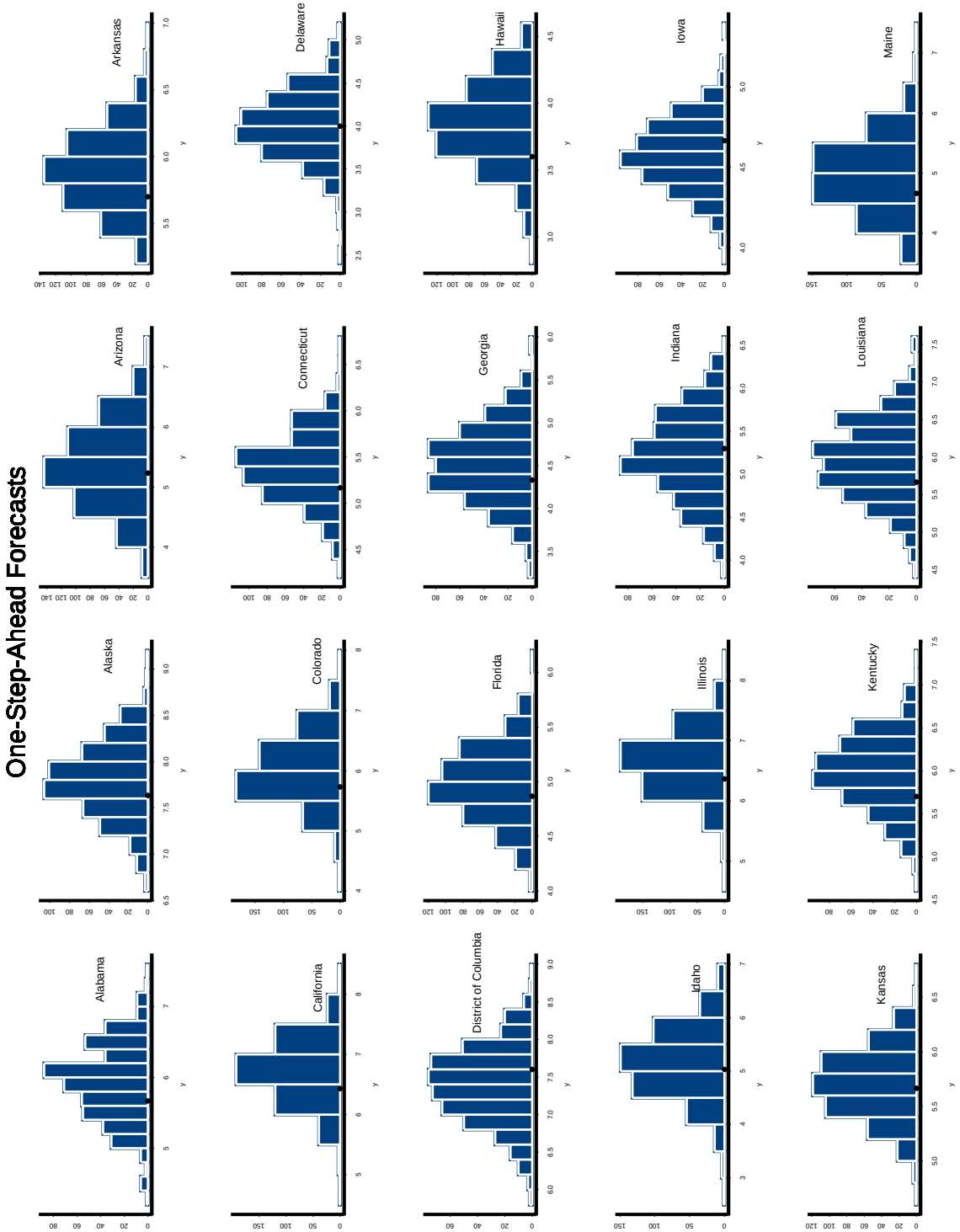

Figure 3.4: Histogram of five hundred replications of the one-quarter-ahead forecast compared with observed (dot point) from Alabama to Maine

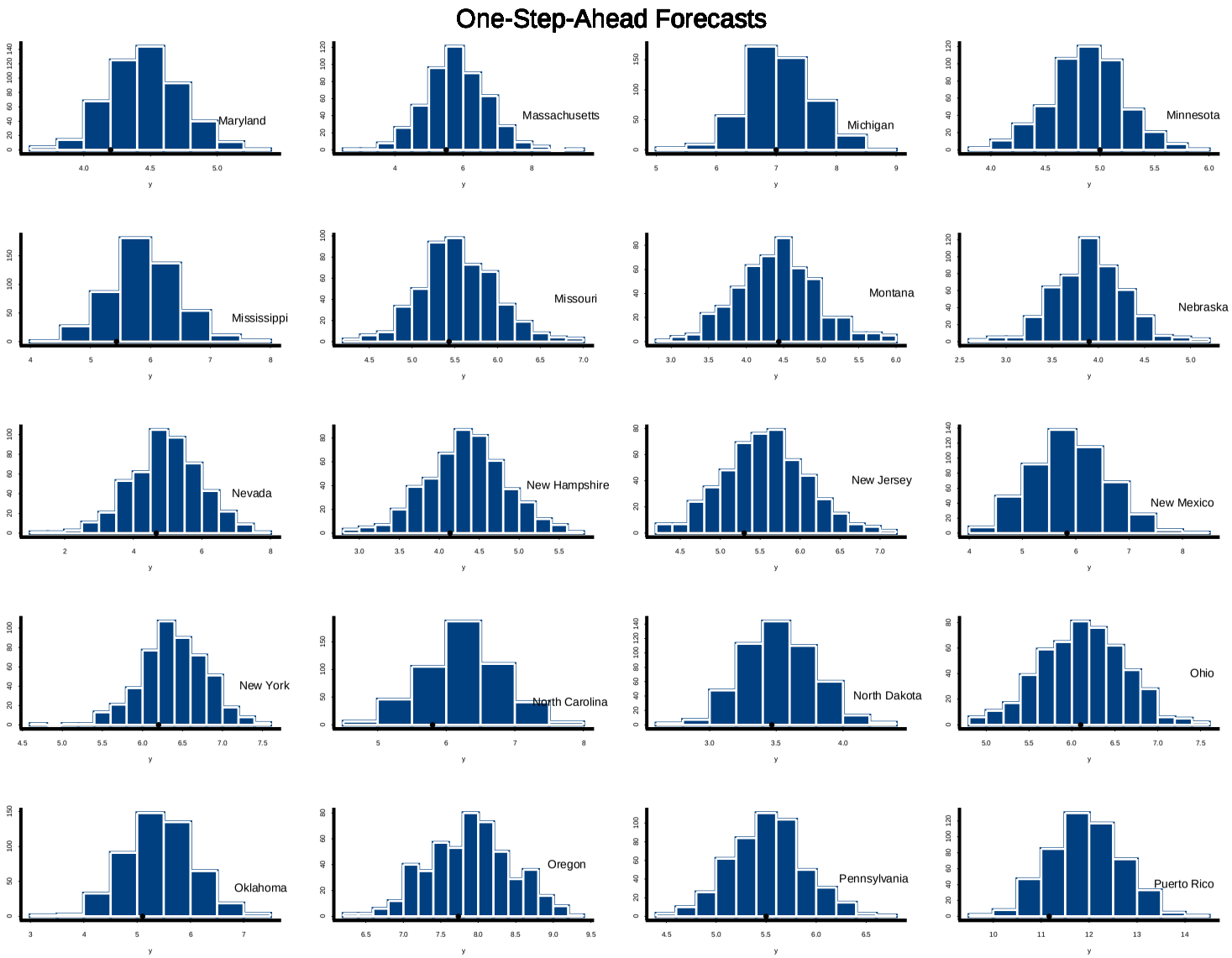

Figure 3.5: Histogram of five hundred replications of the one-quarter-ahead forecast compared with observed (dot point) from Maryland to Puerto Rico

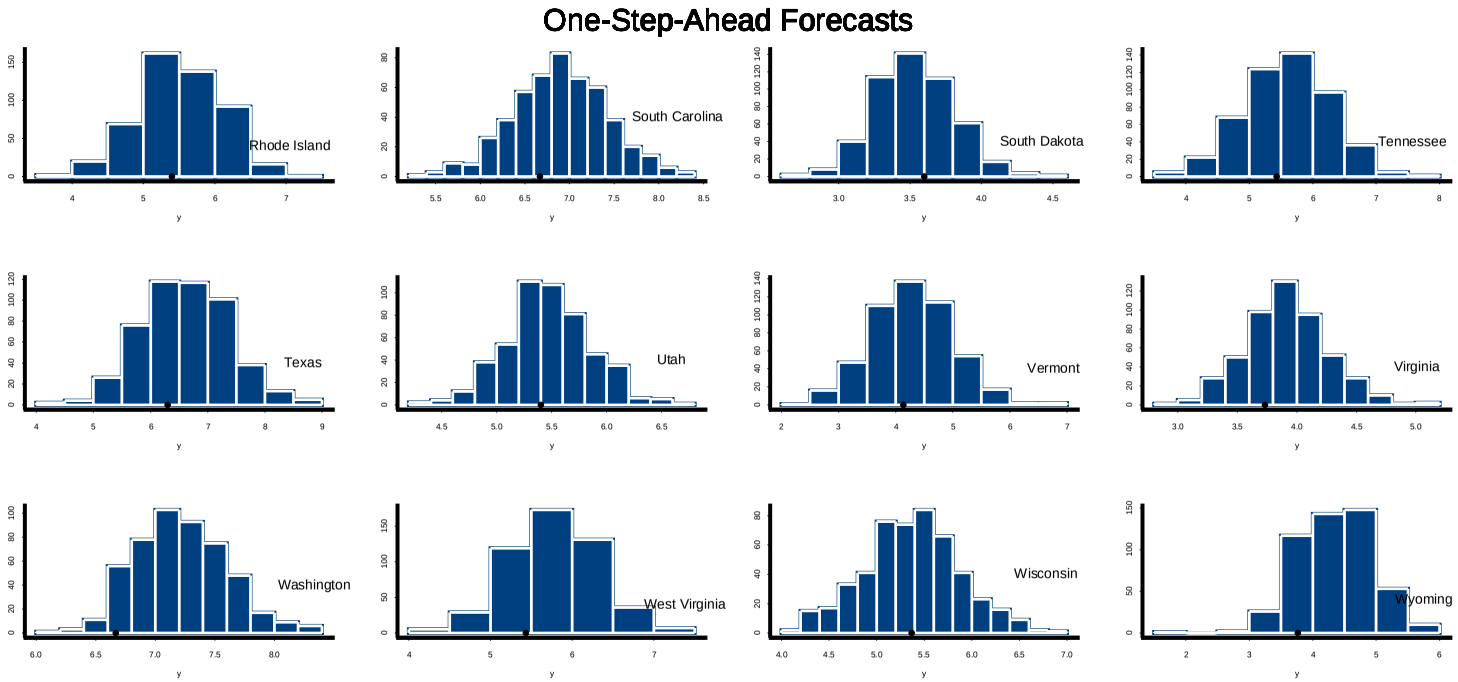

Figure 3.6: Histogram of five hundred replications of the one-quarter-ahead forecast compared with observed (dot point) from Rhode Island to Wyoming

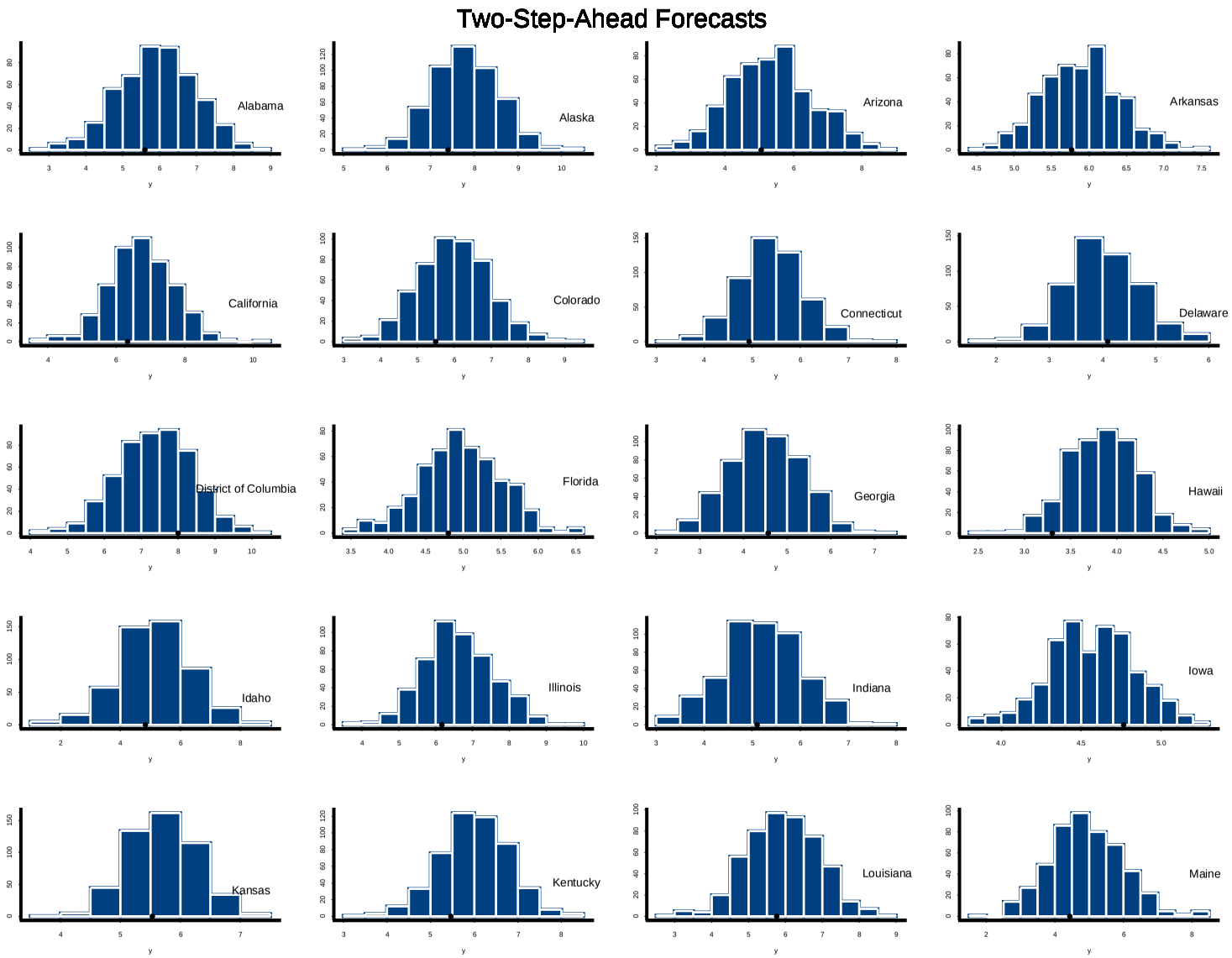

Figure 3.7: Histogram of five hundred replications of the two-quarter-ahead forecast compared with observed (dot point) from Alabama to Maine

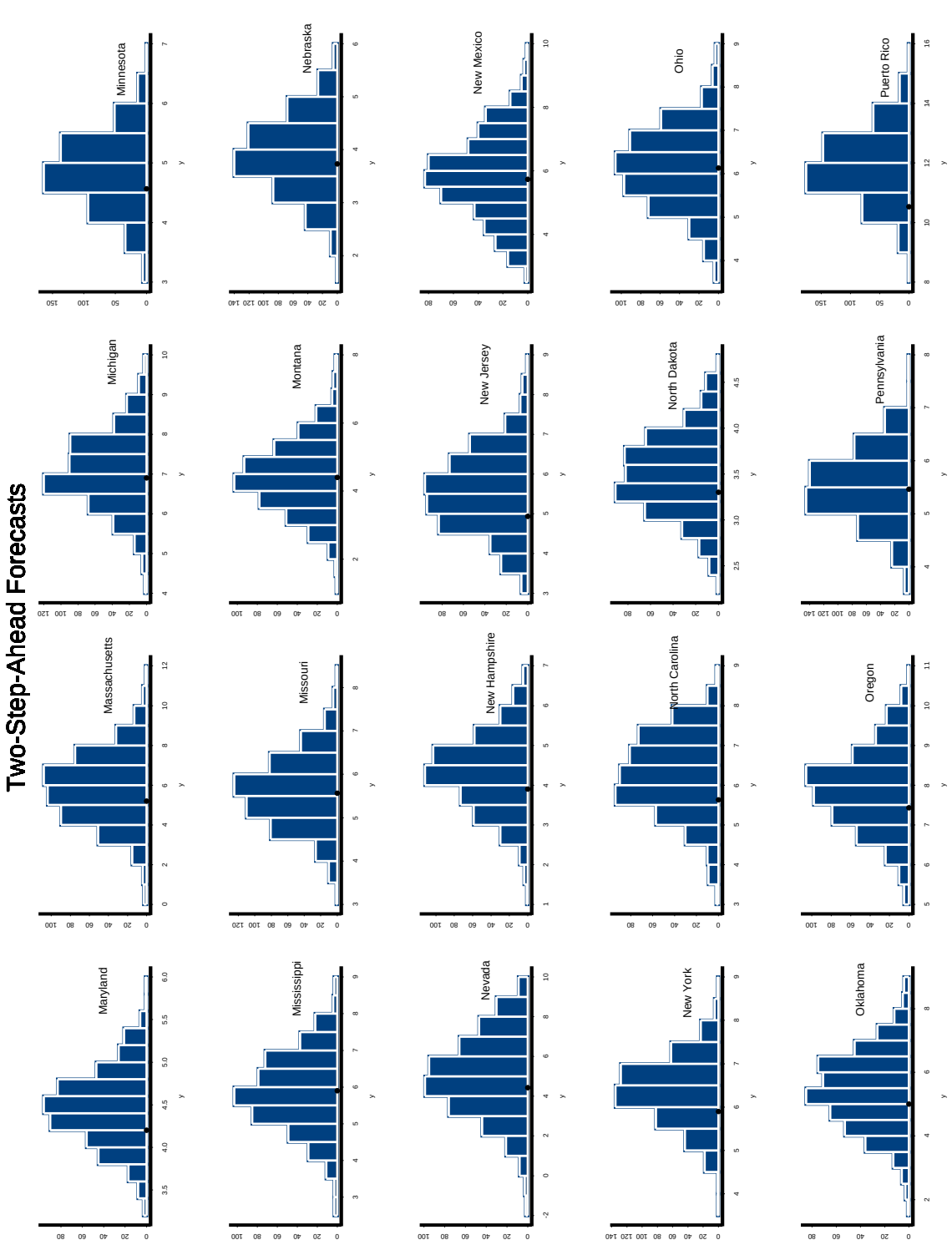

Figure 3.8: Histogram of five hundred replications of the two-quarter-ahead forecast compared with observed (dot point) from Maryland to Puerto Rico

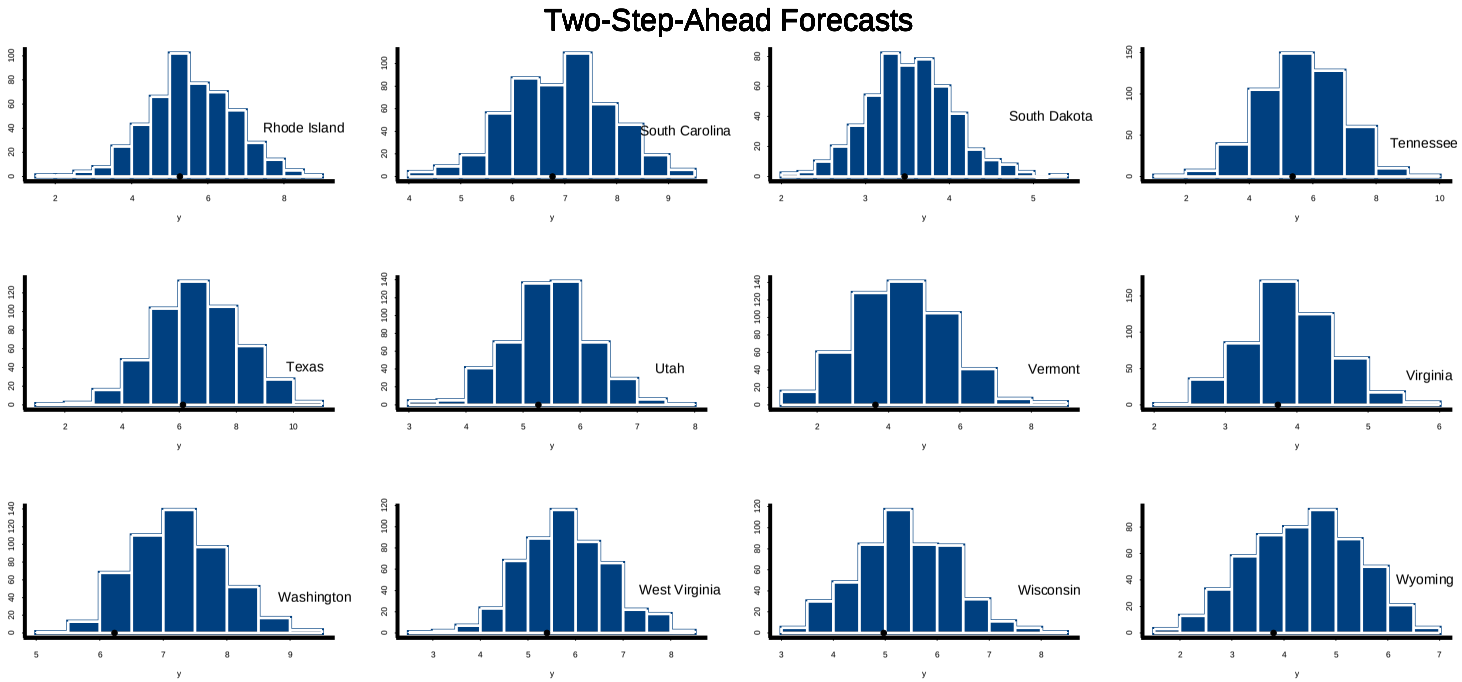

Figure 3.9: Histogram of five hundred replications of the two-quarter-ahead forecast compared with observed (dot point) from Rhode Island to Wyoming

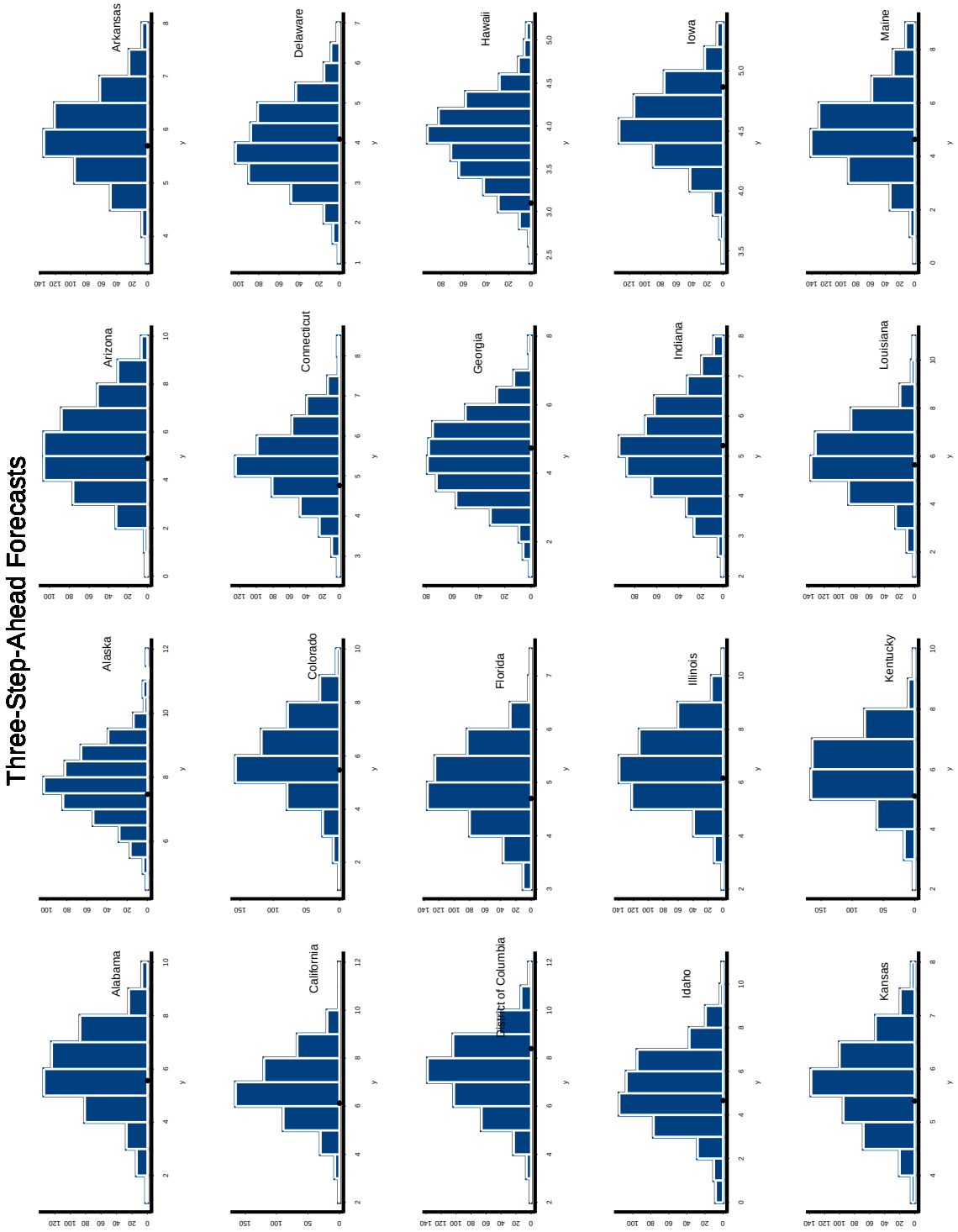

Figure 3.10: Histogram of five hundred replications of the three-quarter-ahead forecast compared with observed (dot point) from Alabama to Maine

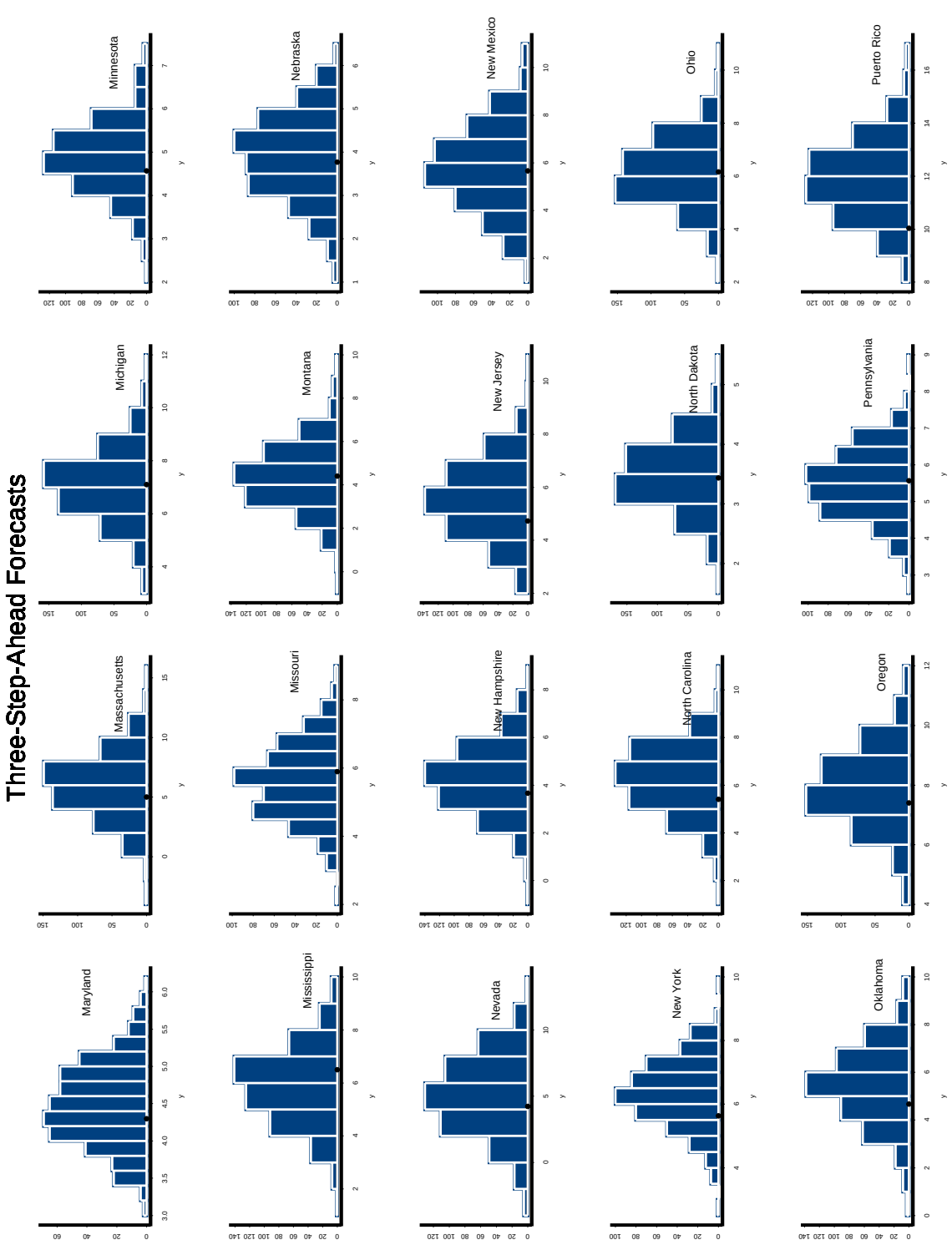

Figure 3.11: Histogram of five hundred replications of the three-quarter-ahead forecast compared with observed (dot point) from Maryland to Puerto Rico

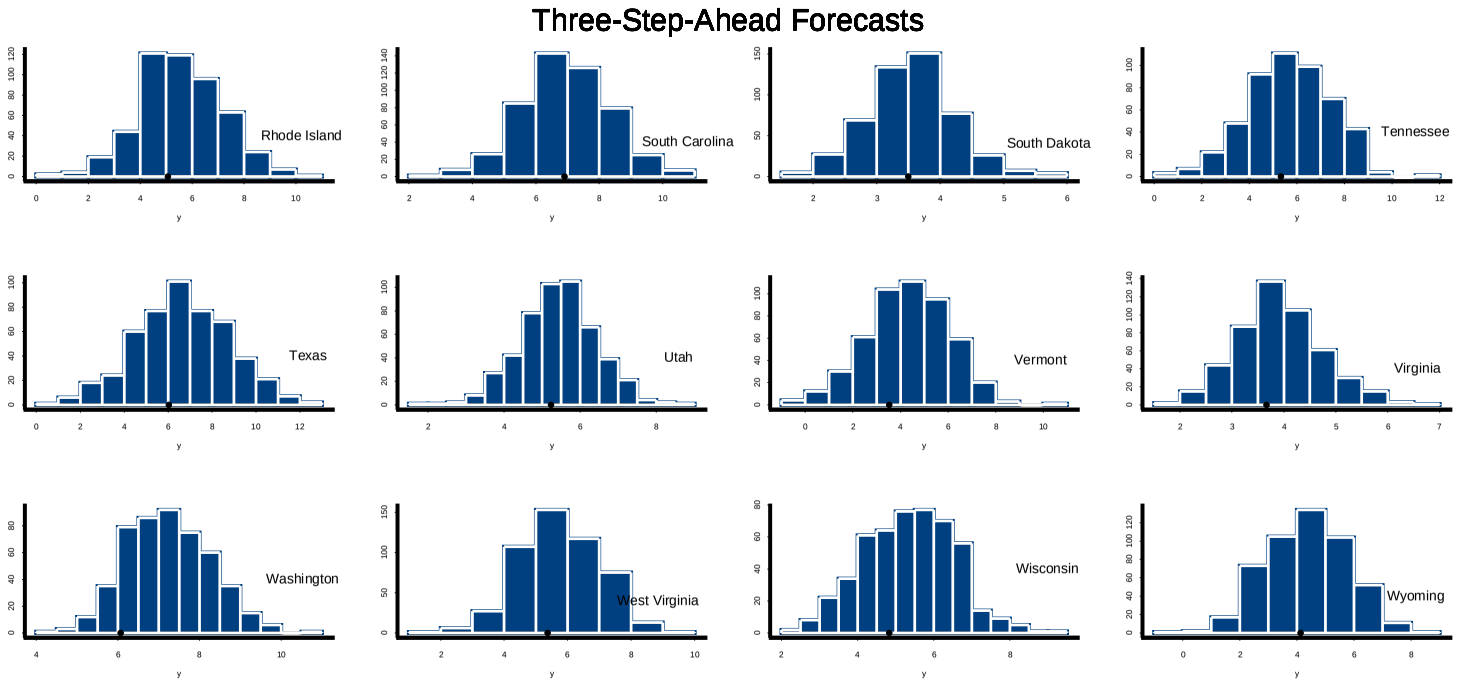

Figure 3.12: Histogram of five hundred replications of the three-quarter-ahead forecast compared with observed (dot point) from Rhode Island to Wyoming

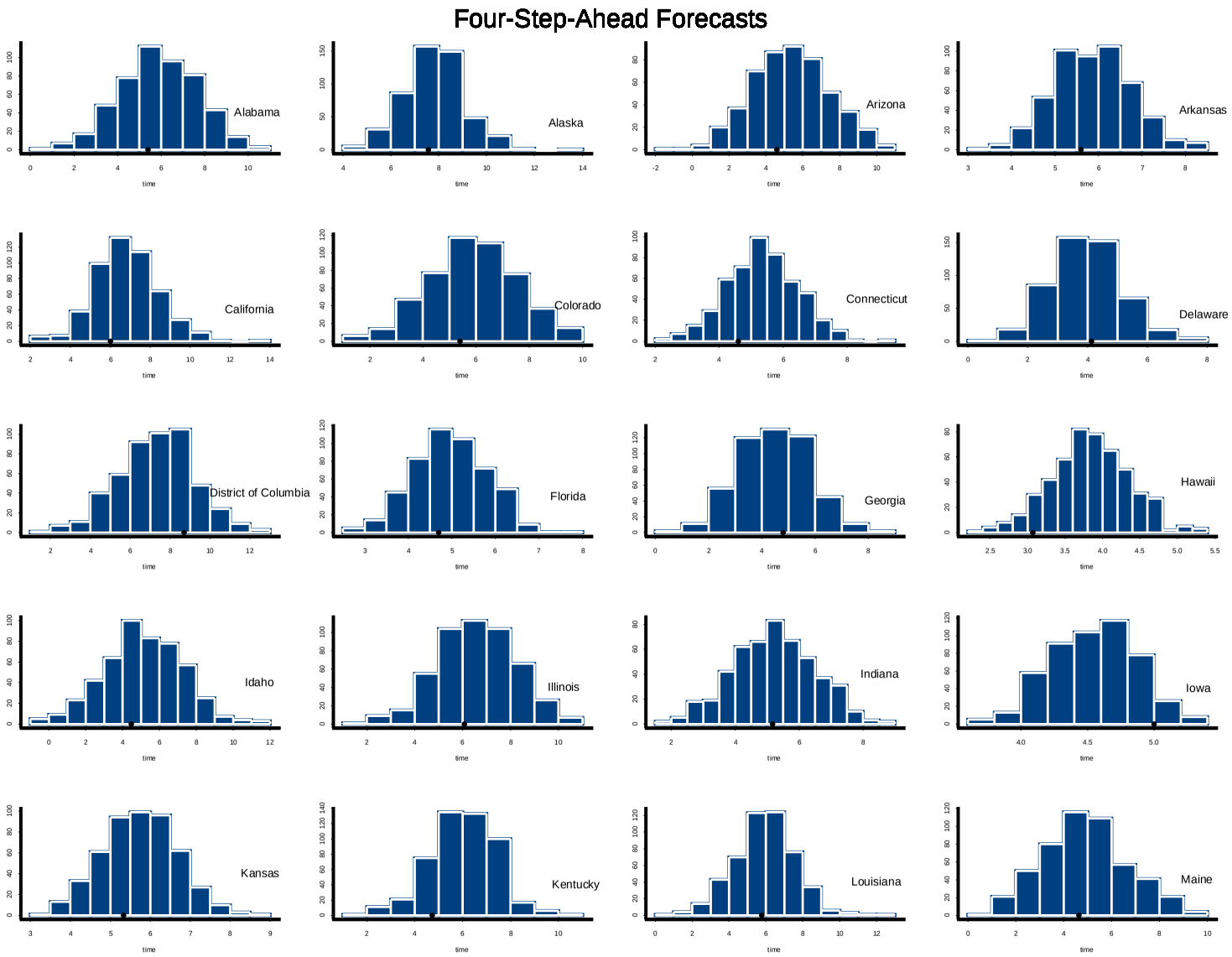

Figure 3.13: Histogram of five hundred replications of the four-quarter-ahead forecast compared with observed (dot point) from Alabama to Maine

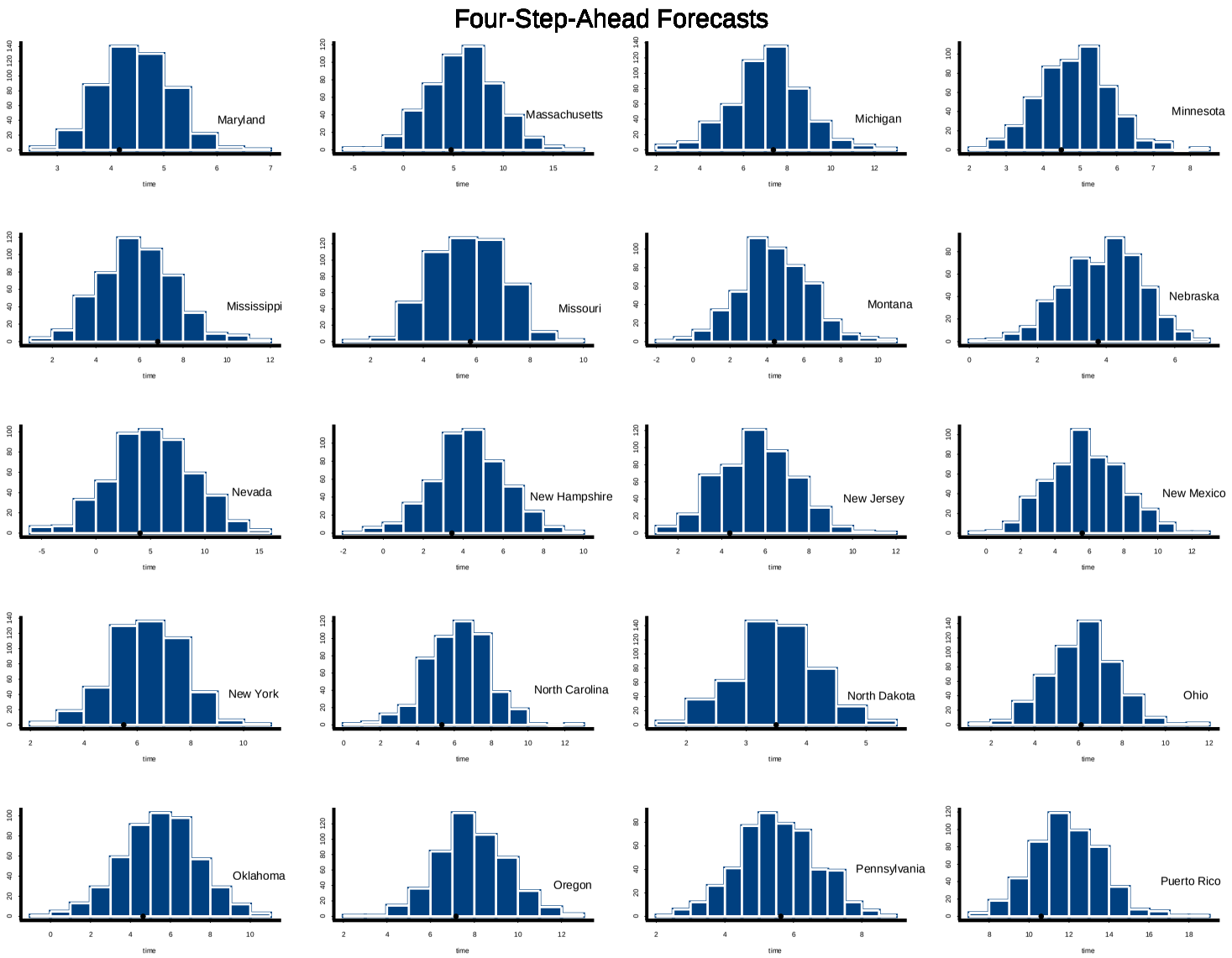

Figure 3.14: Histogram of five hundred replications of the four-quarter-ahead forecast compared with observed (dot point) from Maryland to Puerto Rico

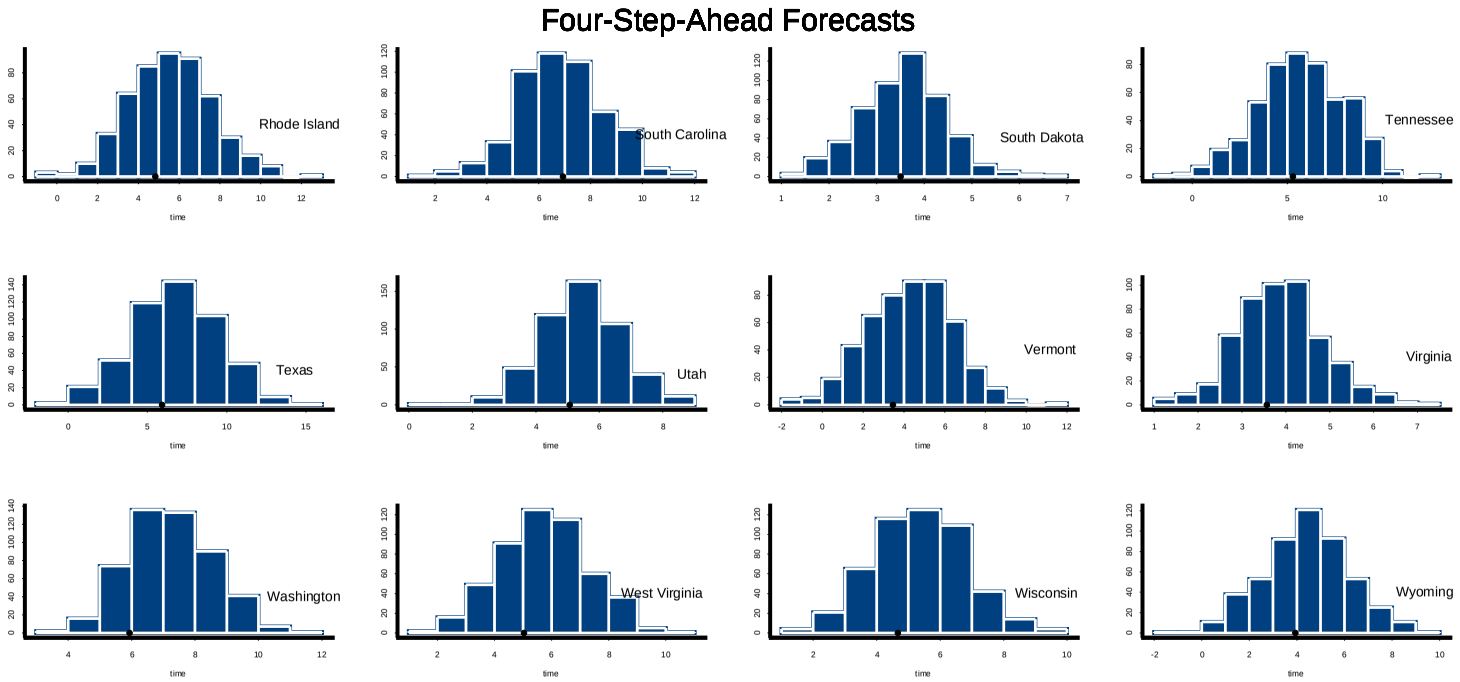

Figure 3.15: Histogram of five hundred replications of the four-quarter-ahead forecast compared with observed (dot point) from Rhode Island to Wyoming

3.7 REFERENCES

Albert, J. and Chib, S. (1993), "Bayesian inference for autoregressive time series with mean and variance subject to Markov jumps," *Journal of Business and Economic Statistics*, 11, 1-15.

Anderson, T. W. (1978)," Repeated measurements on autoregressive process," *Journal of American Statistical Association*, 73, 371-378.

Azzalini, A. (1981)," Replicated observations of low order autoregressive time series," *Journal of Time Series Analysis*, 2, 63-70.

Box, G. E. P. and Jenkins, G. M.(1976), *Time series analysis: Forecasting and control* , San Fransico: Hilden-Day.

Broemeling, L. and Shaaraway, S. (1986), *A Bayesian Analysis of Time Series, in Bayesian Inference and decision Techniques*, P. Goel and A. Zellner, eds., Elsevier, New York.

Chib, S. and Greenberg, E. (1994), "Bayes inference in regression models with ARMA(p,q) errors," *Journal of Econometrics*, 64, 183-206.

Dacunha-Castelle, D. and Gassiat, E.(1999). Testing the order of a model using locally conic parameterization: population mixtures and stationary ARMA processes.*Ann. Statist.*, **27**, 1178–1209.

Davidon, W. C.(1959). Variable Metric Method for Minimization, Research and Development Report ANL 5990, Argonne National Laboratory, Chicago.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood From Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society*,Ser. B **39** , 1–38.

Gelfand, A. E., and Smith, A. F. M. (1990),"Sampling-Based Approaches to calculating Marginal Densities," *Journal of American Statistical Association*, 85, 398-409.

Gelman, A., and Rubin, D. B. (1992), "Inference from iterative simulation using multiple seqences," (with discussion) *Statistical Science*, 7, 457-511.

Gelman, A., Carlin, J., Stern, H. and Rubin, D. B. (1995), "Bayesian Data Analysis," *New York: Chapman and Hall.*

Hastings, W. K. (1970), " Monte Carlo sampling methods using Markov chains and its applications," *Biometrika*, 57, 97109.

Marriott, J., Ravishanker, N., Gelfand, A. E. and Pai, J. (1995), " Bayesian analysis of ARMA processes: Complete sampling based inference under exact likelihoods," *Bayesian Statistics and Econometrics: Essays In Honor of Arnold Zellner, D. Berry, K. Chaloner and J. Geweke, Eds.*,J. Wiley and Sons, New York, 241-256.

McCulloch, R. E., and Tsay, R. S. (1994), "Bayesian Analysis of Threshold Autoregressive Processes with a Random Number of Regimes," *Journal of Time Series Analysis*, 15, 235-250.

Metropolis, N. and Rosenbluth, A. and Rosenbluth, R. and Teller, A. and Teller, E. (1953)," Equation of state calculations by fast computing machines," *Journal of Chemical Physics*, 21, 1087-1092

Monahan, JF. (1983), " Fully Bayesian analysis of ARMA time series models," *Journal of Econometrics*, 21, 307331.

Montgomery, A. L., Zarnowitz, V., Tsay, R. S., and Tiao, G. C. (1998), "Forecasting the U.S. Unemployment Rate," *Journal of American Statistical Association*, 93, 478-493.

Shinha, D., and Dey, D. (1997), "Semiparametric Bayesian Analysis of Survival Data," *Journal of American Statistical Association*, 92, 1195-1212.

Tanner, M. A., and Wong, W. H. (1987),“ The calculation of posterior distributions by data augmentation,” *Journal of American Statistical Association*, 82, 528-540.

Tierney L.(1993),“ Markov chains for exploring posterior distributions,” *Annals of Statistics*, 22, 1701-1762.

Tiller, R. (1992),“ Time series modeling of sample survey data from the U.S. Current Population Survey,” *Journal of Official Statistics*, 8, 149-166.

The proof of the Theorem stated in section 3 depends on the following four lemmas. The conclusions of Lemmas 1 to 3 below hold for any unknown density g_0.

Lemma 1. Under the conditions of the Theorem, $H^2(\hat{g}_n, g_0) \to 0$ a.s. as $n \to \infty$.

Proof. Since $\hat{g}_n$ and g_0 are densities such that $\hat{g}_n \to g_0$ a.s., by Devroye and Györfi (1985) we have that $||\hat{g}_n - g_0||_1 = \int |\hat{g}_n(x) - g_0(x)|dx \to 0$ a.s. The required result now follows from the inequality $H^2(\hat{g}_n, g_0) \leq ||\hat{g}_n - g_0||_1$. □

Lemma 2. Under the conditions of the Theorem, $\sup_{f \in \mathcal{F}} |H^2(\hat{g}_n, f) - H^2(g_0, f)| \to 0$ a.s. as $n \to \infty$, where $\mathcal{F}$ is the class of densities defined in Section 2.

Proof. Note that $H^2(g, f) = 2 - 2\int g^{1/2}(x) f^{1/2}(x)dx$. From this and the Cauchy-Schwarz inequality

$$\sup_{f \in \mathcal{F}} |H^2(\hat{g}_n, f) - H^2(g_0, f)| \leq 2H(\hat{g}_n, g_0) \to 0 \quad \text{a.s.}$$

as $n \to \infty$ by Lemma 1. □

Lemma 3. Under the conditions of the Theorem, for each $m > 0$, $H^2(\hat{g}_n, \hat{g}^m) - H^2(\hat{g}_n, g_0^m) \to 0$ a.s. as $n \to \infty$, where $\hat{g}^m$ and g_0^m are as defined in (2.4).

Proof. Write $H^2(\hat{g}_n, \hat{g}^m) - H^2(\hat{g}_n, g_0^m)$

$$= \{H^2(\hat{g}_n, \hat{g}^m) - H^2(g_0, \hat{g}^m)\} + \{H^2(g_0, \hat{g}^m) - H^2(g_0, g_0^m)\} + \{H^2(g_0, g_0^m) - H^2(\hat{g}_n, g_0^m)\}$$

$$= (i) + (ii) + (iii).$$

The terms (i) and (iii) above converge to zero a.s. by Lemma 2. By (2.4) and that $g_0^m \in \mathcal{F}_m$ and $\hat{g}^m \in \mathcal{F}_m$, it follows that $H^2(\hat{g}_n, \hat{g}^m) - H^2(\hat{g}_n, g_0^m) \leq 0$ and $(ii) \geq 0$. Therefore,

$$(i) + (iii) \leq H^2(\hat{g}_n, \hat{g}^m) - H^2(\hat{g}_n, g_0^m) \leq 0.$$

Since (i) and (iii) converge to zero a.s. we have the required result. □

Lemma 4. Let f and g be any two density functions. Then,

$$\int f^{3/2}(x)(g(x))^{-1/2}dx \geq 1.$$

Proof. Let $K(f,g)$ denote the Kullback-Leibler distance between two densities f and g defined by $K(f,g) = \int f(x)ln(f(x)/g(x))dx$. Observe that

$$\begin{aligned} 0 \leq K(f,g) &= 2\int f(x)ln(f^{1/2}(x)/g^{1/2}(x))dx \\ &\leq 2\int f(x)[(f^{1/2}(x)/g^{1/2}(x)) - 1]dx, \end{aligned}$$

which implies the required result. □

Proof of the Theorem. Let $d_m = H^2(g_0, g_0^m) - H^2(g_0, g_0^{m+1})$ for $m > 0$, where g_0 is any unknown density. Note that, for each $m > 0$, $d_m \geq 0$ by (2.4) and that $\mathcal{F}_m \subseteq \mathcal{F}_{m+1}$. We will first show that for each $m > 0$

$$H^2(\hat{g}_n, \hat{g}^m) - H^2(\hat{g}_n, \hat{g}^{m+1}) \to d_m \quad \text{a.s.} \tag{A.1}$$

as $n \to \infty$. To this end, write

$$H^2(\hat{g}_n, \hat{g}^m) - H^2(\hat{g}_n, \hat{g}^{m+1}) - d_m$$

$$\begin{aligned} &= \{H^2(\hat{g}_n, \hat{g}^m) - H^2(\hat{g}_n, g_0^m)\} + \{H^2(\hat{g}_n, g_0^{m+1}) - H^2(\hat{g}_n, \hat{g}^{m+1})\} \\ &+ H^2(\hat{g}_n, g_0^m) - H^2(g_0, g_0^m)\} + \{H^2(g_0, g_0^{m+1}) - H^2(\hat{g}_n, g_0^{m+1})\}. \end{aligned} \tag{A.2}$$

The first two terms on the right side of (A.2) converge to zero a.s. by Lemma 3 and the last two terms on the right side of (A.2) converge to zero a.s. by Lemma 2. Hence we have (A.1).

Suppose g_0 is *not* a finite mixture then $m_0 = \infty$. This and (2.6) imply that $d_m > 0$ for all $m > 0$. Therefore, by (A.1) and (2.9) it follows that $\hat{m}_n \to \infty$ a.s.

Suppose g_0 is a finite mixture. Then, the required result for $\hat{m}_n$ would follow once we show that $d_m > 0$ for $m < m_0$ and $d_m = 0$ for $m \geq m_0$.

Case $m \geq m_0$: Recall that $g_0 \in \mathcal{F}_{m_0} \subseteq \mathcal{F}_j$ for all $j \geq m_0$. Therefore, by (2.4), for each $j \geq m_0$, $0 \leq H(\hat{g}_n, \hat{g}^j) \leq H(\hat{g}_n, g_0) \to 0$ a.s. by Lemma 1. It now follows from (A.1) that $d_m = 0$ for $m \geq m_0$.

Case $m < m_0$: From (2.6) observe that $g_0 \in \mathcal{F}_{m_0}$ and $g_0 \notin \mathcal{F}_m$ for $m < m_0$. Following the arguments in Leroux (1992) [see proof of (ii), Lemma 3] we will show by a contradiction argument that $d_m > 0$.

Suppose $d_m = 0$ for $m < m_0$, that is, $H^2(g_0, g_0^m) = H^2(g_0, g_0^{m+1})$. This, (2.4) and $H^2(\cdot, \cdot) \geq 0$ implies that for all $f \in \mathcal{F}_{m+1}$

$$H^2(g_0, g_0^m) \leq H^2(g_0, f). \tag{A.3}$$

Since $g_0^m \in \mathcal{F}_m$, by (2.1) we have that $g_0^m(x) = f_{\boldsymbol{\theta}_m^0}(x)$ for some $\boldsymbol{\theta}_m^0$. For an arbitrary $\epsilon \in (0, 1)$ and $\boldsymbol{\phi} \in \Phi \subseteq R^s$, let $\tilde{f}(x) = (1 - \epsilon) g_0^m(x) + \epsilon f(x|\boldsymbol{\phi})$, where $f(x|\boldsymbol{\phi})$ is a density function. Then, $\tilde{f} \in \mathcal{F}_{m+1}$ and the inequality in (A.3) holds for the mixture density $\tilde{f}$. Therefore, by (A.3) and the definition of Hellinger distance

$$2 \int g_0^{1/2}(x) \{[(1 - \epsilon) g_0^m(x) + \epsilon f(x|\boldsymbol{\phi})]^{1/2} - (g_0^m(x))^{1/2}\} dx \leq 0 \tag{A.4}$$

for all $\epsilon \in (0, 1)$. Dividing both sides of (A.4) by ϵ and applying Fatou's lemma (as $\epsilon \to 0$) to the resulting expression yields

$$\int g_0^{1/2}(x) [f(x|\boldsymbol{\phi}) (g_0^m(x))^{-1/2} - (g_0^m(x))^{1/2}] dx \leq 0$$

which implies that

$$\int g_0^{1/2}(x) (g_0^m(x))^{1/2} dx \geq \int g_0^{1/2}(x) f(x|\boldsymbol{\phi}) (g_0^m(x))^{-1/2} dx \tag{A.5}$$

for every $\boldsymbol{\phi} \in \Phi$. Since $g_0 \in \mathcal{F}_{m_0}$ we can write $g_0(x) = \sum_{i=1}^{m_0} \pi_i^0 f(x|\boldsymbol{\phi}_i^0)$ and (A.5) holds for each $\boldsymbol{\phi} = \boldsymbol{\phi}_i^0$, $i = 1, \ldots, m_0$. Since $\sum_{i=1}^{m_0} \pi_i^0 = 1$, from (A.5)

$$\begin{aligned} \int g_0^{1/2}(x)(g_0^m(x))^{1/2}dx &\geq \int g_0^{3/2}(g_0^m(x))^{-1/2}dx \\ &\geq 1, \end{aligned} \tag{A.6}$$

by an application of Lemma 4 with $f = g_0$ and $g = g_0^m$. This implies that $H^2(g_0, g_0^m) = 0$, which contradicts the fact that $g_0 \notin \mathcal{F}_m$ for $m < m_0$. Hence the Theorem. □