Root Selection in Normal Mixture Models

Byungtae Seo\textsuperscript{a}, Daeyoung Kim\textsuperscript{a,b}

\textsuperscript{a}Department of Statistics, Sungkyunkwan University, Seoul 110-745, Korea
\textsuperscript{b}Department of Mathematics and Statistics, University of Massachusetts, Amherst, MA 01003, USA

Abstract

Finite mixtures of normal distributions are attractive in identifying the underlying group structure in the data. However, it is a challenging task to do statistical inference in normal mixture models using the method of maximum likelihood, due to the unbounded likelihood and the existence of multiple roots to the likelihood equation including a so-called spurious root. In this article we propose a new likelihood-based method for selecting a statistically reasonable root when there exist multiple roots of the likelihood equation for a finite normal mixture model. We first prove that our proposed methodology can choose a root to the mixture likelihood equation with consistency. We then show, by simulation studies and real examples, that the proposed methods can greatly reduce the risk of choosing problematic roots that have the same features as spurious roots.

Key words: Consistency, Maximum likelihood, Normal mixture, Singularity, Spurious local maximizer.

1. Introduction

It is well known that there are multiple roots of the likelihood equation for a finite normal mixture model. Hence, standard mixture analysis involves a strategy of using multiple starting values to search for a local maxima of the likelihood. Maximum likelihood philosophy then urges us to choose the root with the highest likelihood among the found multiple roots of the likelihood

*Corresponding author. Tel.:+1 413 577 0255
Email addresses: seobt@skku.edu (Byungtae Seo), daeyoung@math.umass.edu (Daeyoung Kim)
A first type of undesirable solutions is a singular solution due to an unbounded likelihood in a finite normal mixture model. For example, the likelihood of a normal mixture model with unequal variances goes to infinity if the location parameter of one of the component densities is equal to a sample observation and if the corresponding scale parameter tends to zero [19]. Consequently, the ML estimator always occurs on the boundary of the parameter space and is clearly inconsistent. Biernacki and Chrétien [1] showed that if mixture parameters are close to a singular solution, then the EM algorithm [7] converges towards it with an exponential speed, and numerical tolerance of the computer is reached very quickly.

One can avoid singular solutions in several ways, by either constraining the parameter space or modifying the likelihood function. Hathaway [14] suggested a constrained ML estimation imposing a lower bound to either the scale parameters or the ratios of the scale parameters. Note that one can use additional constraints on the component proportion parameters. This strategy can be implemented using the constrained EM algorithm developed by Hathaway [15] and Ingrassia and Rocci [17]. In a similar vein, Tanaka and Takemura [28] applied the sieve method [12, 11] using a sequentially constrained parameter space which grows to the whole parameter space as the sample size increases.

Ciuperca et al. [5], Chen et al. [4] and Chen and Tan [3] proposed a penalized maximum likelihood estimation method where a penalty term on the component variance (covariance) is added to the log likelihood function. From a Bayesian point of view, the use of their penalty function is equivalent to putting an inverse gamma (Wishart) distribution prior on the component variance (covariance). For more details, see Bayesian regularization method by Fraley and Raftery [10]. Seo and Lindsay [27] used a doubly-smoothed maximum likelihood estimation which modifies the likelihood using a smoothing technique. All of these approaches rely on different types of modifications (such as any prior information on the parameters) to remove the irregularity of the mixture likelihood. The consistency can be then obtained from the fact that the effects of those modifications disappear as the sample size increases. Note that these approaches may require a new (and complicated) type of EM algorithms.

Another type of undesirable solutions is a spurious solution that exists
in the interior of the parameter space. McLachlan and Peel [25, p. 99] characterized a spurious solution as a root to the mixture likelihood equation that “lies close to the boundary of the parameter space and has one or more components overfitting a small random localized pattern in the data rather than any underlying group structure”. One can observe this type of solution not only in mixtures of location-scale distributions (such as normal mixture models with unknown/known unequal covariances), but also in mixtures of exponential distributions [26].

When there exist multiple roots to the mixture likelihood equation, Seidel and Ševčíková [26] and McLachlan and Peel [25, pp.99–105] demonstrated by simulation studies and real data analysis that one needs to select the largest local maximizer in the mixture likelihood, after removing any spurious solutions. For example, in simulation studies where the true parameter values are known, the power of the likelihood ratio test of homogeneity can be greatly influenced by the spurious solutions whose likelihood values are higher than other solutions [26]. In addition, the existence of spurious solutions makes it difficult to interpret some numerical summaries such as the empirical bias and standard error calculated from simulation studies.

The problem becomes more challenging in real data analysis because the true value of the parameters is unknown and there is no obvious way to determine if a given solution is spurious without extra information on the parameters and their estimates. Yet, we still often observe solutions that have features similar to those of spurious solutions in the mixture analysis [25, pp.99–105].

Hathaway [14] showed that approach to constraining the parameters (such as ratio of scale parameters or proportion parameters), designed to remove singularities, might also lessen the problems associated the spurious solutions. However, as pointed out by McLachlan and Peel [25, p.96], the problem in practice is to choose tuning constant(s) in the parameter constraint(s).

In this article we propose a methodology designed to exclude all singular solutions and to reduce the risk of selecting spurious roots of the likelihood equation for a finite normal mixture model. The proposed method takes into account that the singular and spurious solutions overfit random localized pattern composed of few observations in the data set which exert unduly strong influence on the formation of the likelihood-based solution. We first create $k$-deleted likelihood free of the effect of $k$ overfitted observations, and propose using it as a root selection criterion. That is, after we delete the likelihood terms of such $k$ observations from the original likelihood, we compute
\( k \)-deleted likelihood at every found local maximizer and choose the local maximizer with the highest \( k \)-deleted likelihood. Furthermore, we prove that the root selected by our proposed criterion still preserves the property of consistency. Since our proposed method provides a method of root selection rather than a new type of likelihood, one can easily implement the proposed method in the standard EM algorithm.

Although the issues regarding the model selection and the existence of outliers would be important in mixture modelling, we investigate singular and spurious problems under the assumption that the mixture model is well specified (such as a proper choice of component densities and number of components, and data in general position). This is because the singular/spurious problems are still difficult to solve under even such circumstances.

This article is organized as follows: In Section 2, we give two real examples of singular and spurious solutions in normal mixture models. Section 3 proposes two simple criteria based on \( k \)-deleted likelihood to select a root among the multiple roots of the likelihood equation for a finite mixture model. In Section 4 we study the asymptotic property of the root of the mixture likelihood equation selected by our proposed criteria. Section 5 carries out simulation studies to evaluate the performance of the proposed criteria and revisits the examples in Section 2 to show how the proposed methods work. We then end this article with a discussion in Section 6.

2. Motivating examples

In this section we introduce two interesting examples concerning singular and spurious solutions. We fit finite normal mixture models to the two real data sets and use the EM algorithm to estimate the parameters in the models. For detailed information about the stopping rule and the initial values employed in the algorithm, see Section 5.1.

Example 1. The first example is a real data set concerning an acidity index measured in a sample of 155 lakes in the Northeastern United States [6]. Many authors have analyzed this data set as a mixture of normal distributions with two to six components. Based on the bootstrap likelihood ratio test for the number of components in McLachlan and Peel [25, p. 194], we fitted a univariate two-component normal mixture model with unequal variances to this data set.
Table 1: Four local maximizers in Example 1: $\ell(\hat{\theta})$ is a log likelihood at each local maximizer, $\hat{p}_1$ is an estimate of the first component mixing weight, $\hat{\mu}_j$ and $\hat{\sigma}^2_j$ are the estimates for the mean and variance parameters for component $j$ where $j=1,2$.

<table>
<thead>
<tr>
<th>$\hat{\theta}$</th>
<th>$\ell(\hat{\theta})$</th>
<th>$\hat{p}_1$</th>
<th>$\hat{\mu}_1$</th>
<th>$\hat{\mu}_2$</th>
<th>$\hat{\sigma}^2_1$</th>
<th>$\hat{\sigma}^2_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\theta}_1$</td>
<td>$\infty$</td>
<td>0.004</td>
<td>3.822</td>
<td>5.109</td>
<td>0</td>
<td>1.076</td>
</tr>
<tr>
<td>$\hat{\theta}_2$</td>
<td>-184.645</td>
<td>0.404</td>
<td>6.249</td>
<td>4.330</td>
<td>0.270</td>
<td>0.139</td>
</tr>
<tr>
<td>$\hat{\theta}_3$</td>
<td>-187.235</td>
<td>0.521</td>
<td>5.891</td>
<td>4.251</td>
<td>0.718</td>
<td>0.068</td>
</tr>
<tr>
<td>$\hat{\theta}_4$</td>
<td>-218.884</td>
<td>0.013</td>
<td>5.930</td>
<td>5.094</td>
<td>1.769e-08</td>
<td>1.083</td>
</tr>
</tbody>
</table>

We used the EM algorithm to compute the MLE for the parameters. In order to obtain local maxima in the mixture likelihood, we used 90 initial values for parameters from the two types of approach, random-start approach [25, p.55] (40 from the first random-start method and 40 from the second method described in Section 5.1) and clustering approach (5 from the $k$-means and 5 from Linde-Buzo-Gray (LBG) method [21]). For a stopping criterion of the EM algorithm we employed a criterion based on directional derivatives [2, 22, 26], with a level of accuracy $10^{-6}$.

From the EM algorithm with 90 starting values, we found two singular solutions with zero variance estimate for one component, and 12 nonsingular local maximizers. Table 1 shows one singular solution, $\hat{\theta}_1$, and the best three local maximizers among all found nonsingular solutions, $\hat{\theta}_2 - \hat{\theta}_4$. If one removes $\hat{\theta}_1$ due to singularity of the variance, one might choose the local maximizer with the largest log-likelihood value, $\hat{\theta}_2$, as the most desirable solution. However, there is no clear justification for the largest local likelihood criterion, as the information about the true clusters is not available in this data set.

**Example 2.** The second example concerns the real data analyzed by McLachlan and Peel [25, p. 103] in order to exemplify the existence of multiple local maximizers including spurious local maximizers in the mixture model application. The original data, collected by Habbema et al. [13], contained two variables, 100 log$_{10}$(anti-hemophilic factor (AHF)) and 100 log$_{10}$(AHF-like antigen). The goal of this study was to discriminate between normal women and hemophilia A carriers using both variables. The classification of 75 data points were known, 30 from the noncarriers and 45 from the carriers, respectively (see Figure 1).

Following McLachlan and Peel [25, p. 103], we ignored the known classification and fitted a two-component bivariate normal mixture model with
Figure 1: Hemophilia data with carriers group(○) and noncarriers group(●): the x axis and y axis represent the values of $100 \log_{10}$(anti-hemophilic factor (AHF)) and $100 \log_{10}$(AHF-like antigen), respectively.

Figure 2: Estimated densities for 14 local maximizers ordered by the log likelihood value (from left top to right bottom)
unequal covariance matrices to cluster 75 data points into two groups. We then used the EM algorithm to compute local maximizers of the likelihood from 100 starting values obtained in the same way as in Example 1: 88 initial values were from random-start approach, and 12 initial values were based on clustering approach. Note that we employed the same stopping rule as in Example 1.

From the EM algorithm with 100 initial values, we obtained 14 local maxima in the mixture likelihood. In addition, there were two singular solutions where the estimate of one of covariance matrices was singular. In order to check an early stopping of the algorithm, we ran the EM for additional 50,000 steps at each local maximum, and found that changes in the parameter estimates were very negligible. Figure 2 shows the estimated densities with the 14 local maxima, denoted as $\hat{\theta}_1 - \hat{\theta}_{14}$, in the order of their log likelihood values. Note that we found (or observed) that $\hat{\theta}_5$ (leftmost plot in the second row of Figure 2) was a local maximizer that produced the two estimated clusters closest to the known clusters among the 14 local maximizers found, but it had the fifth largest log likelihood value. We observe that each of the first four local maximizers, $\hat{\theta}_1 - \hat{\theta}_4$ (first row in Figure 2), formed a cluster with few observations in a lower dimensional subspace although such observations were not far away from the main data cloud. Note that the four observations, belonged in the second component with the smallest mixing weights of $\theta_3$, had the values of the second coordinate that were close to each other. The relatively smaller eigenvalue of the second estimated component covariance matrix in $\theta_3$ can confirm this finding (see Table 2).

Table 2 shows the log likelihood values, the estimates for mixing weights and the eigenvalues of the estimated component covariance matrices at 14 local maximizers shown in Figure 2. Note that McLachlan and Peel [25, p. 103] suggested using the eigenvalues of the component covariance matrices at each local maximizer as a useful guide to detect existence of spurious solution. We can see the imbalances between the eigenvalues ($\lambda_{21}, \lambda_{22}$) for the first four local maximizers, $\hat{\theta}_1 - \hat{\theta}_4$, compared to that of the fifth largest solution, $\hat{\theta}_5$.

If one uses the largest local likelihood criterion for root selection, one has to choose $\hat{\theta}_1$ with $\ell(\hat{\theta}_1) = -608.8689$. However, we can see that the clustering results based on $\hat{\theta}_5$ with $\ell(\hat{\theta}_5) = -613.739$ are most consistent with those based on the known classification. In general, one should not take for granted that, in a data set with known classes, the known classes are the ones that the largest local likelihood criterion in the normal mixture modelling should find
Table 2: 14 local maximizers in Example 2: \(\ell(\hat{\theta})\) is a log likelihood at each local maximizer, \(\hat{p}_1\) is an estimate of the first component mixing weight, \(\lambda_{1j}\) and \(\lambda_{2j}\) are the eigenvalues of the j-th component estimated covariance matrix where j=1,2.

<table>
<thead>
<tr>
<th>(\hat{\theta}_i)</th>
<th>(\ell(\hat{\theta}))</th>
<th>(\hat{p}_1)</th>
<th>(\lambda_{11})</th>
<th>(\lambda_{21})</th>
<th>(\lambda_{22})</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\hat{\theta}_1)</td>
<td>-608.8689</td>
<td>0.95</td>
<td>132.103</td>
<td>365.095</td>
<td>0.0005</td>
</tr>
<tr>
<td>(\hat{\theta}_2)</td>
<td>-612.1071</td>
<td>0.87</td>
<td>116.893</td>
<td>388.359</td>
<td>1.511</td>
</tr>
<tr>
<td>(\hat{\theta}_3)</td>
<td>-612.6143</td>
<td>0.95</td>
<td>134.705</td>
<td>399.659</td>
<td>0.026</td>
</tr>
<tr>
<td>(\hat{\theta}_4)</td>
<td>-612.8569</td>
<td>0.94</td>
<td>116.581</td>
<td>381.948</td>
<td>4.0875</td>
</tr>
<tr>
<td>(\hat{\theta}_5)</td>
<td>-613.7392</td>
<td>0.51</td>
<td>52.007</td>
<td>183.225</td>
<td>70.140</td>
</tr>
<tr>
<td>(\hat{\theta}_6)</td>
<td>-613.9792</td>
<td>0.82</td>
<td>69.407</td>
<td>420.325</td>
<td>21.048</td>
</tr>
<tr>
<td>(\hat{\theta}_7)</td>
<td>-614.2385</td>
<td>0.84</td>
<td>136.242</td>
<td>414.120</td>
<td>3.650</td>
</tr>
<tr>
<td>(\hat{\theta}_8)</td>
<td>-615.2544</td>
<td>0.92</td>
<td>124.697</td>
<td>347.007</td>
<td>1.781</td>
</tr>
<tr>
<td>(\hat{\theta}_9)</td>
<td>-615.3387</td>
<td>0.95</td>
<td>323.448</td>
<td>131.709</td>
<td>2.845</td>
</tr>
<tr>
<td>(\hat{\theta}_{10})</td>
<td>-615.6391</td>
<td>0.60</td>
<td>257.632</td>
<td>136.252</td>
<td>44.742</td>
</tr>
<tr>
<td>(\hat{\theta}_{11})</td>
<td>-615.7248</td>
<td>0.93</td>
<td>132.808</td>
<td>343.733</td>
<td>2.252</td>
</tr>
<tr>
<td>(\hat{\theta}_{12})</td>
<td>-616.3163</td>
<td>0.85</td>
<td>137.475</td>
<td>248.837</td>
<td>49.241</td>
</tr>
<tr>
<td>(\hat{\theta}_{13})</td>
<td>-616.4770</td>
<td>0.94</td>
<td>133.860</td>
<td>347.236</td>
<td>0.814</td>
</tr>
<tr>
<td>(\hat{\theta}_{14})</td>
<td>-616.5243</td>
<td>0.88</td>
<td>140.520</td>
<td>321.460</td>
<td>7.750</td>
</tr>
</tbody>
</table>

[16]. However, in the data set of Example 2, the bivariate normality assumption of the component densities is reasonable [24] and the mixture likelihood has a local maximizer that results in the estimated classes consistent with the known classes. Thus, we question the validity of the solution selected by the largest local likelihood criterion because there may exist spurious solutions, such as \(\hat{\theta}_1\) - \(\hat{\theta}_4\) in Example 2, that neither represent the feature of the true cluster nor have meaningful information for the data.

3. \(k\)-deleted likelihood criteria

In this section we propose two simple methods designed to avoid singularities and to reduce the risk of choosing spurious local maximizers of the likelihood for a finite mixture model. For an effective presentation, we here consider univariate normal mixtures with unequal variances although the proposed methods can be applied to any type of mixtures, regardless of the dimension of the data.

Let \(\phi(x)\) be the standard normal density and let \(f(x; \theta)\) be a \(m\)-component normal mixture density with a component density \(\phi((x - \mu)/\sigma)/\sigma\) :

\[
f(x; \theta) = \sum_{j=1}^{m} \frac{p_j}{\sigma_j} \phi \left( \frac{x - \mu_j}{\sigma_j} \right), \tag{3.1}
\]
where $\theta = \{(\mu_j, \sigma_j, p_j) : j = 1, \cdots, m\}$, $(\mu_j, \sigma_j, p_j) \in \mathbb{R} \times \mathbb{R}^+ \times \mathbb{R}(0, 1)$ for $j = 1, \cdots, m$ and $\sum_{j=1}^{m} p_j = 1$. We assume that no information on the parameters is available, and the component density $\phi(\cdot)$ and the number of components $m$ are correctly specified. Given a data set $x = \{x_1, \ldots, x_n\}$ from Eq (3.1), one can construct the ordinary log-likelihood function in the form

$$\ell(\theta, x) = \sum_{i=1}^{n} \ell_{i} = \sum_{i=1}^{n} \log f(x_i; \theta),$$

where $\ell_{i} \equiv \log f(x_i, \theta)$ is the log likelihood term at the observation $x_{i}$. Suppose that one has the likelihood equations by setting the gradient of the log likelihood in Eq (3.2) to zero vector, and obtains a set of solutions to the likelihood equations, denoted by $Q = \{\hat{\theta}_h; h = 1, \cdots, H\}$, using a strategy of multiple starting values for the EM algorithm.

The basic idea of the proposed methods is as follows. Detect a few, say $k$, log likelihood terms that are disproportionately influential on each solution $\hat{\theta}_h \in Q$, and compute a $k$-deleted log likelihood at $\hat{\theta}_h$ by removing the detected $k$ log likelihood terms from the ordinary log-likelihood at $\hat{\theta}_h$. One can then select the solution with the highest $k$-deleted log likelihood in $Q$. The crucial step here is to detect $k$ log likelihood terms which have unduly strong influence on generating singular and spurious solutions to the likelihood equations. In the Sections 3.1 and 3.2 we will propose two methods designed to identify such influential log likelihood terms.

### 3.1. Likelihood-based $k$-deleted likelihood criterion

Singularities in Eq (3.2) occur when there is at least one observation equals to one of $\mu_j$'s, and the corresponding $\sigma_j$ is zero. One might also have solutions with spurious and small clusters (corresponding $\hat{\theta}_j$s are very close to the boundary of the parameter space). In both cases it is likely that there are one or a few observations whose log likelihood terms are very large compared to those of the rest of the data, and these likelihood values dominate the whole likelihood in Eq (3.2). For example, for a given singular solution $\hat{\theta}$, the ordinary log-likelihood in Eq (3.2) is infinite because the likelihood extremely overfits an observation $x_i$ at $\hat{\theta}$: $\log f(x_i; \hat{\theta}) = \infty$ for a certain $x_i$.

To identify observations inappropriately influential on the existence of singular and spurious solutions, we suggest using the log likelihood of an individual observation. That is, we define the likelihood-based 1-deleted log-likelihood criterion.
likelihood by removing the largest likelihood term from the ordinary log likelihood,

$$\ell_{L-1}(\theta, x) = \ell(\theta; x) - \max\{\log(f(x_i; \theta)); i = 1, \ldots, n\}. \quad (3.3)$$

Note that $\ell_{L-1}(\theta, x)$ represents how well the rest of the data is fit after one removes the effect of a single overfitted observation. We can also construct the likelihood-based $k$-deleted likelihood, $\ell_{L-k}(\theta, x)$, by removing the $k$ largest likelihood terms at $\theta$. This is first proposed by Seo and Lindsay [27] as a tool to choose a reasonable bandwidth in their smoothed likelihood.

We then propose using Eq. (3.3) as a new criterion to choose a root to the mixture likelihood equation. Given the set of all found solutions, $Q = \{\hat{\theta}_h; h = 1, \ldots, H\}$, the likelihood-based $1$-deleted maximum likelihood estimator (denoted by MLE$_{L-1}$) is defined as

$$\hat{\theta}_{L-1} = \arg\max_{\theta \in Q} \{\ell_{L-1}(\theta, x)\}. \quad (3.4)$$

This does not require an extra burdensome computation as we have already calculated $\log f(x_i; \hat{\theta}_h)$ for each $\hat{\theta}_h \in Q$ and $i = 1, \ldots, n$. One can also define the likelihood-based $k$-deleted MLE, MLE$_{L-k}$ in a similar way.

For a given singular solution $\hat{\theta}$, $\ell_{L-1}(\hat{\theta})$ deletes a single divergent likelihood term and thus would likely be smaller than that of any nonsingular solution. In this case, all singular solutions are naturally removed and there will be no technical difficulty in defining MLE$_{L-1}$ unlike the MLE maximizing Eq (3.2). Clearly, the MLE in Eq (3.2) and MLE$_{L-1}$ in Eq (3.4) are same if the ordinary likelihood is unimodal. When there are multiple roots of the mixture likelihood equation and over-fitted observations are present at the MLE, however, the MLE$_{L-1}$ and MLE may be different.

One concern for MLE$_{L-1}$ would be whether or not the effect of deleting the largest likelihood term is so large that it might undermine the likelihood spirit. If one uses $\ell_{L-1}(\theta, x)$ in Eq. (3.3) as a new objective function, this could be a great concern. However, since the proposed method modifies only the root selection criterion, the concern would be less important. In fact MLE$_{L-1}$ still has the property of consistency, as will be shown in Section 4.

3.2. Score-based $k$-deleted likelihood criterion

Although the unbounded likelihood causes a clear theoretical problem, in practice, most numerical algorithms will easily remove singularities, as
they will stop running due to the zero estimated value of $\sigma_j$ for at least one component. A more important problem in practice is the presence of a spurious local maximizer in a mixture likelihood. The likelihood-based $k$-deleted likelihood criterion proposed in Section 3.1 might remove some spurious solutions if they have much higher likelihood values than other solutions due to a few too large log likelihood terms (see numerical results of Section 5).

However, these outlying likelihood terms are not the only reason for the existence of spurious solutions. To exemplify this, let us consider one of the spurious local maximizers found in the example of Section 2, $\hat{\theta}_4$ (the rightmost plot in the first row of Figure 2). From the estimated probability of belonging to the smaller component in $\hat{\theta}_4$, we found that $\hat{\theta}_4$ in fact gained its strength mainly through two data points (-23.75,28.76) and (-18.78,25.10). That is, without these two points, $\hat{\theta}_4$ looks much less plausible. In this case, the likelihood terms of these two points at $\hat{\theta}_4$ were very small as they were located on the tail of each fitted component density. This implies the largest likelihood term at a given solution is not necessarily the one having the highest impact on the formation of spurious solutions.

For better understanding of a spurious local maximizer in the mixture likelihood, we further investigate the score function, the gradient of the ordinary log-likelihood function, at a local maximizer, $\hat{\theta}$. Define $u_i(\theta) \equiv u(\theta;x_i)$ to be the vector of partial derivatives of $\ell_i = \log f(x_i,\theta)$ with respect to $\theta$. Then $u_i(\hat{\theta})$ is the score function of $\theta$ at the observation $x_i$ and $\sum_{i=1}^n u_i(\hat{\theta}) = 0$ at $\hat{\theta}$. If $\hat{\theta}$ is a spurious solution that includes at least one component with very small non-zero values of mixing proportion parameter and/or scale parameter, there may exist one or a few observations whose score vectors are in completely different directions than the others and thus the corresponding norms are very large compared to others. For instance, for a given spurious solution $\hat{\theta}$ with $\sum u_i(\hat{\theta}) = 0$, there may exist an observation $x_j$ satisfying $u_i(\hat{\theta})^T u_j(\hat{\theta}) < 0$ and $u_j(\hat{\theta})^T u_j(\hat{\theta}) > u_i(\hat{\theta})^T u_i(\hat{\theta})$ for all $i \neq j$. In this case, $u_j(\hat{\theta})$, score vector for $x_j$ at $\hat{\theta}$, has a large influence on $\sum_{i=1}^n u_i(\hat{\theta}) = 0$, and $x_j$ appears to be an unduly influential point for the solution $\hat{\theta}$. We here view 1-deleted likelihood function at $\hat{\theta}$, the ordinary log likelihood minus the log likelihood term of $x_j$, as a measure to help us see how a solution comes into being.

From this investigation, we propose a score-based 1-deleted log-likelihood as a method to avoid spurious solutions:

$$\ell_{-1}^S(\theta, x) = \ell(\theta, x) - \ell_t$$  (3.5)
and
\[ x_t = \arg \max_{1 \leq i \leq n} u_i(\theta)^T u_i(\theta). \] (3.6)

Similar to MLE\(^L_{-1}\) in Eq (3.4), we define the \textit{score-based 1-deleted maximum likelihood estimator}, denoted by MLE\(^S_{-1}\),
\[ \hat{\theta}^{S}_{-1} = \arg \max_{\theta \in Q} \ell^{S}_{-1}(\theta, x). \] (3.7)

A score-based \( k \)-deleted log-likelihood \( \ell^{S}_{-k}(\theta, x) \) and the score-based \( k \)-deleted MLE, MLE\(^S_{-k}\), can be defined in a similar way.

This score-based \( k \)-deleted likelihood is also free from the theoretical difficulty which occurs in the ordinary MLE because, for a singular solution \( \hat{\theta} \), \( \ell^{S}_{-k}(\theta, x) \) will not include the likelihood values of the observations \( x_t \) with \( u_t(\hat{\theta})^T u_t(\hat{\theta}) = \infty \). This means that the score-based \( k \)-deleted likelihood can also be used for identifying singular solutions. Note that, as will be shown in the simulation studies of Section 5, the score-based \( k \)-deleted log-likelihood \( \ell^{S}_{-k}(\theta, x) \) and the corresponding MLE\(^S_{-k}\) work reasonably well and even perform better than the MLE\(^L_{-k}\), especially in dealing with spurious solutions.

### 3.3. Guideline for the choice of \( k \)

The theoretical choice of \( k \) which guarantees non-singular MLE\(^L_{-k}\) and MLE\(^S_{-k}\) can be obtained as follows. Suppose, for simplicity, we consider a two-component univariate normal mixture and there is no tie in the data. If a solution is singular, there is only one divergent likelihood term. For a \( m \)-component mixture, theoretically, there will be at most \( m - 1 \) divergent likelihood terms, as a singular solution can have at most \( m - 1 \) singular components. Hence, \( k = m - 1 \) guarantees non-singular solutions in both MLE\(^L_{-k}\) and MLE\(^S_{-k}\).

For the \( p \)-dimensional data, we suggest using \( p(m - 1) \) as a value of \( k \) in MLE\(^L_{-k}\) and MLE\(^S_{-k}\) to completely remove singularities. To explain this, let us consider a two-component bivariate normal mixture (i.e., \( p=2 \) and \( m=2 \)). Assuming there is no three data points on a line, from the rotation invariance of the maximum likelihood estimator we can assume that there are two observations \((x_1, y_1)\) and \((x_2, y_2)\) with \( x_1 = x_2 = a \) without loss of generality. Then, one can see an unbounded likelihood by letting the mean of the first coordinate be equal to \( a \) and the corresponding variance go to
zero for both likelihood terms at \((x_1, y_1)\) and \((x_2, y_2)\). This implies that there exist at most two divergent likelihood terms at each singular component.

Regarding spurious solutions, the biggest challenge for the choice of \(k\) is that there is no clear mathematical definition of spurious solutions. If we use a large \(k\), there is less risk of choosing a spurious solution while we may have more risk of ignoring a potentially important local solution with small-sized cluster(s), and vice versa. To compromise with this trade-off, we recommend computing the MLE\(^L\)\(_{-k}\) and MLE\(^S\)\(_{-k}\) with the value of \(k\) equal to \(p, p + 1, \ldots, p(m - 1)\) as a simple guideline.

There are several reasons behind this recommendation. First, the spurious solution and singularities have common feature that they have small clusters overfitting a few data points, even though there is no clear mathematical definition of the spurious solution. Second, in many cases we observe that singular (spurious) solutions occur with one singular (spurious) component. But, we cannot exclude a possibility that a given singular (spurious) solution has multiple singular (spurious) components. Hence, \(p\) is the minimum value to safely remove solutions with one singular (spurious) component, and \(p(m - 1)\) is the maximum value to remove solutions with \(m - 1\) singular (spurious) components. Third, the computation of the MLE\(^L\)\(_{-k}\) and MLE\(^S\)\(_{-k}\) is straightforward and just a little amount of the computation time is needed, as it requires only the outputs from the final step of the EM algorithm (and the score function of the parameters for the MLE\(^S\)\(_{-k}\)). Last, we can examine the overall reliability of the obtained solution by checking if the MLE\(^L\)\(_{-k}\) and MLE\(^S\)\(_{-k}\) change over several values of \(k\). In Section 5 we will consider a range of \(k\) that is slightly wider than our recommendation in order to provide empirical justification for the proposed guideline.

4. Consistency of \(k\)-deleted maximum likelihood estimators

For the consistency of the \(k\)-deleted MLE’s, MLE\(^L\)\(_{-k}\) and MLE\(^S\)\(_{-k}\), we first show their strong consistency on a sequentially constrained parameter space \(\Gamma_n\). We then show that there is no local maximizer on \(\Gamma/\Gamma_n\) as \(n \to \infty\), where \(\Gamma\) is the whole parameter space. For simplicity we present the consistency results for univariate normal mixtures. The extension to multivariate mixtures, non-normal mixture, and \(k > 1\) can be made similarly.
Let \( f(x; \theta) \) be a \( m \)-component normal mixture density:
\[
f(x; \theta) = \sum_{j=1}^{m} \frac{p_j}{\sigma_j} \phi \left( \frac{x - \mu_j}{\sigma_j} \right)
\]
where \( \phi(x) \) is the standard normal density and \( \theta = \{ (\mu_j, \sigma_j, p_j) : j = 1, \cdots, m \} \) on the parameter space
\[
\Gamma = \{ (\mu_j, \sigma_j, p_j) : -\infty < \mu_j < \infty, \sigma_j > 0, p_j > 0, \sum p_j = 1, j = 1, \cdots, m \}\n\]
Let \( \Gamma_n \) be a sequence of a constrained parameter space growing to \( \Gamma \):
\[
\Gamma_n = \{ (\mu_j, \sigma_j, p_j) : -\infty < \mu_j < \infty, \sigma_j \geq Ce^{-nd}, p_j > 0, \sum p_j = 1, j = 1, \cdots, m \}\n\]
with fixed constant \( C > 0 \) and \( 0 < d < 1 \).

Suppose \( X_1, \ldots, X_n \) is a random sample from \( f(x; \theta_0) \), where \( \theta_0 \) is an interior point of \( \Gamma \). Suppose further that \( S_n = \{ \theta_n^{(1)}, \cdots, \theta_n^{(k_n)} \} \) is the set of all local maximizers of the ordinary likelihood in \( \Gamma_n \). Let \( \hat{\theta}_n \) be the sequentially constrained MLE and let \( \hat{\theta}'_n \) and \( \hat{\theta}''_n \) be the MLE\(_L^1\) and MLE\(_S^1\) on \( \Gamma_n \) respectively. That is,
\[
\hat{\theta}_n = \arg\max_{1 \leq s \leq k_n} \ell(\theta_n^{(s)}) \quad (4.1)
\]
\[
\hat{\theta}'_n = \arg\max_{1 \leq s \leq k_n} \left[ \ell(\theta_n^{(s)}) - \max_i \log f(x_i; \theta_n^{(s)}) \right] \quad (4.2)
\]
\[
\hat{\theta}''_n = \arg\max_{1 \leq s \leq k_n} \left[ \ell(\theta_n^{(s)}) - \log f(x^*_s; \theta_n^{(s)}) \right], \quad (4.3)
\]
where
\[
x^*_s = \arg\max_{x \in \mathcal{X}_n} ||u(\theta_n^{(s)}; x_i)||. \quad (4.4)
\]
\( \mathcal{X}_n = \{ x_1, \ldots, x_n \} \) and \( || \cdot || \) denotes \( L_2 \)-norm. The strong consistency of the sequentially constrained MLE on \( \Gamma_n \), \( \hat{\theta}_n \), can be found in [28] under some mild conditions.

**Theorem 1.** Under assumptions in Appendix, \( \hat{\theta}_n \rightarrow \theta_0 \) almost surely.

If \( \hat{\theta}_n \) and \( \hat{\theta}'_n \) are identical for every \( n \), it implies the consistency of \( \hat{\theta}'_n \) on \( \Gamma_n \) from Theorem 1. However, due to the existence of (spurious) multiple local maximizers, \( \hat{\theta}_n \) and \( \hat{\theta}'_n \) could be different in a finite sample. The following theorem shows that \( \hat{\theta}'_n \) is still consistent by utilizing the fact that deleting one likelihood term is negligible on \( \Gamma_n \) as \( n \rightarrow \infty \).
Theorem 2. Under assumptions in Appendix, $\hat{\theta}_n' \to \theta_0$ almost surely.

Proof See Appendix A.1.

When we apply the score-based deletion, the deleted observation in (4.4) falls into one of two categories: either (1) when $x^*_s$ is an outlying point far away from the rest of data or (2) when $||u(\hat{\theta}; x^*)||$ is large but $x^*$ is not far from the majority of the data. We call the first type of outliers traditional outlier and the second type score outlier. Establishing a valid consistency proof that covers both types of outliers is not a simple task. However, since our concern is not the traditional outlier but the score outlier in the presence of the multiple local maximizers, we slightly modify Eq (4.4) as

$$x^*_s = \arg\max_{x \in \mathcal{X}_n \cap \{M \geq ||x_i||\}} ||u(\theta^{(s)}_n, x_i)||,$$  (4.5)

where $M$ is a chosen bound. With this restriction, we can exclude the case where extreme traditional outliers going to infinity keep occurring as $n \to \infty$ so that the deleted likelihood term, $\log f(x^*_s, \theta^{(s)}_n)$, has a lower bound for fixed $\theta^{(s)}_n$.

In practice with a given finite sample, however, the bound $M$ does not much affect on the numerical calculation because it compactifies the sample space just for deletion. Now, before we prove the consistency of the $\hat{\theta}_n''$, we add the following assumption.

Assumption 1. For any solution $\theta \in S_n$, there exists at least one component $j$ such that $(\mu_j, \sigma_j, p_j) \in K \subset \mathbb{R} \times \mathbb{R}^+ \times \mathbb{R}(0,1)$ for a compact set $K$.

This assumption basically implies that there is no local maximizer whose all components are nearly singular. This additional mild condition can greatly simplify the proof of the following lemma.

Lemma 1. For any $\theta^{(s)}_n \in S_n$, $\{\log f(x^*_s; \theta^{(s)}_n)\}$ is bounded below.

Proof See Appendix A.2.

Theorem 3. Under assumptions in Appendix and Assumption 1, $\hat{\theta}_n'' \to \theta_0$ almost surely.
Proof See Appendix A.3.

Theorem 2 and 3 imply that \( \text{MLE}^L_{-1} \) and \( \text{MLE}^S_{-1} \) are strongly consistent on \( \Gamma_n \). However, these theorems still possess one concern; if 1-deleted MLE’s keep occurring on \( \Gamma/\Gamma_n \) and converge to a singular solution, \( \text{MLE}^L_{-1} \) and \( \text{MLE}^S_{-1} \) are not consistent on the whole parameter space \( \Gamma \). Thus a study on this type of estimators is required for the consistency.

When we have local maximizers on \( \Gamma/\Gamma_n \), we do not need to consider singular solutions for \( \text{MLE}^L_{-1} \) and \( \text{MLE}^S_{-1} \) because \( \text{MLE}^L_{-1} \) and \( \text{MLE}^S_{-1} \) can not be singular for any \( n \). Now, for all nonsingular solutions, one can prove there is no local likelihood solution on \( \Gamma/\Gamma_n \) as \( n \to \infty \). To explain this, suppose there is a sequence of nonsingular solutions, \( \{\theta_n\}_{n=1}^{\infty} = \{(\mu_{nj}, \sigma_{nj}, p_{nj})\}, j = 1 \cdots, m \} \) that occurs on \( \Gamma/\Gamma_n \). Without loss of generality, let us assume \( \sigma_{n1} \leq \sigma_{n2} \leq \cdots \leq \sigma_{nm} \). Then \( \sigma_{n1} \) should be less than \( Ce^{-nd} \) because \( \theta_n \in \Gamma/\Gamma_n \). Moreover, since \( \theta_n \) is the solution of the likelihood equation, \( \theta_n \) should satisfy

\[
0 = \frac{\partial \ell(\theta)}{\partial \sigma_{n1}^2} = \frac{1}{2\sigma_{n1}^4 \sqrt{\sigma_{n1}^2}} \sum_{i=1}^{n} \frac{[(x_i - \mu_{n1})^2 - \sigma_{n1}^2] \phi \left( \frac{x_i - \mu_{n1}}{\sigma_{n1}} \right)}{\sum_{j=1}^{m} p_{nj} \sigma_{nj} \phi \left( \frac{x_i - \mu_{nj}}{\sigma_{nj}} \right)}. \tag{4.6}
\]

To hold this equation, at least one \( x_i \) should satisfy \( (x_i - \mu_{n1})^2 - \sigma_{n1}^2 < 0 \), otherwise Eq (4.6) is always positive and cannot be zero. That is, \( \min_{i} |x_i - \mu_{n1}| < \sigma_{n1} \). Now we will show that this can not occur for infinitely many \( n \).

**Proposition 1.** \( P(\min_{i} |X_i - \mu_{n1}| < \sigma_{n1} \text{ i.o.}) = 0 \) where i.o. means infinitely often.

**Proof** See Appendix A.4.

This implies that it is sufficient to consider the likelihood solutions on \( \Gamma_n \) for the consistency of \( \text{MLE}^L_{-1} \) and \( \text{MLE}^S_{-1} \). Now, from Theorem 2-3 and Proposition 1, the \( \text{MLE}^L_{-1} \) and \( \text{MLE}^S_{-1} \) are consistent on the whole parameter space \( \Gamma \). A similar argument can be made for other location-scale component densities.

5. Numerical examples

In this section we examine the performance of the two proposed \( k \)-deleted ML estimators, \( \text{MLE}^L_{-k} \) and \( \text{MLE}^S_{-k} \), using simulation studies and real data
analysis, in the context of $p$-dimensional normal mixture models with $m$-components. For comparison purpose we also compute the two estimators, the ratified maximum likelihood estimator (RMLE) and the penalized maximum likelihood estimator (PMLE). The RMLE is the local maximizer with the largest likelihood among all found nonsingular solutions. The PMLE, maximizing the penalized log-likelihood function (consisting of the log likelihood function plus a penalty function), might be able to reduce the risk of selecting spurious solutions. This is because adding appropriate penalty terms smoothes the log-likelihood function, and thus the spurious roots might neither exist in the penalized log likelihood nor be likely to have higher penalized log likelihood values. Regarding the computation of the PMLE, we employ a penalty function proposed by Chen et al. [4] and Chen and Tan [3] that will be briefly introduced in Section 5.1. Note that all the computations in this paper were done in Matlab 7.5 using a computer with Intel Core 2 Duo CPU 2.80GHz.

5.1. Settings for simulation studies and data analysis

For a $p$-dimensional random vector $\mathbf{x}$, let $\phi(\mathbf{x}; \mu, \Sigma)$ be the $p$-dimensional normal density with $p \times 1$ mean vector $\mu$ and $p \times p$ covariance matrix $\Sigma$:

$$
\phi(\mathbf{x}; \mu, \Sigma) = (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \mu)^\prime \Sigma^{-1} (\mathbf{x} - \mu) \right\}.
$$

The density of a $p$-dimensional normal mixture distribution with $m$-components is given by

$$
f(\mathbf{x}; \theta) = \sum_{j=1}^{m} p_j \phi(\mathbf{x}; \mu_j, \Sigma_j)
$$

where $p_1, \ldots, p_m$ are mixing weights with $0 \leq p_j \leq 1$ and $\sum_{j=1}^{m} p_j = 1$, and $\theta = (p_1, \ldots, p_{m-1}, \mu_1', \ldots, \mu_m', \text{vech}(\Sigma_1)', \ldots, \text{vech}(\Sigma_m)')'$. Here $\text{vech}(A)$ means the half-vectorization of a symmetric matrix $A$ that vectorizes only the lower triangular part of $A$.

Given $n$ random samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ from Eq (5.1), we employed the EM algorithm for the estimation of $\theta$. With respect to a stopping criterion of the algorithm we used a stopping criterion based on directional derivatives [2, 22, 26], with a level of accuracy $10^{-6}$. That is, we stop the iteration at $\theta = \theta^t$ if

$$
\max \{ S(\mu_1^t, \text{vech}(\Sigma_1^t), \theta^t) - n, \ldots, S(\mu_m^t, \text{vech}(\Sigma_m^t), \theta^t) - n \} < \text{acc}
$$
where $\theta^t$ is a value of $\theta$ at the $t$-th iteration,

$$S(\mu^t_j, vech(\Sigma^t_j), \theta^t) = \sum_{i=1}^{n} \left[ \frac{\phi(x_i; \mu^t_j, \Sigma^t_j)}{f(x_i; \theta^t)} \right]$$

and $acc$ is a level of accuracy. Note that a stopping criterion based on directional derivatives has a better theoretical basis than other criteria such as the size of change in the parameters or the log likelihood [2, 22].

As to initial values for $\theta$ in the algorithm, we consider the three types of approach.

1. two perturbation approaches [3]
   1-1) We obtain the five initial values by using the true value for $\theta$ and the following perturbation process: get the $q$-th initial value for $\theta$ by adding $(q - 1) \times (-1)^j \times \Delta$ to the true value of $\mu_j$ where $\Delta = 1$ for $p = 1$, $\Delta = (1, -1)'$ for $p = 2$, $\Delta = (1, -1, 1, -1, 1)'$ for $p = 5$ and $q = 1, \ldots, 5$. Here, the initial values for $\Sigma_j$ and $p_j$ are equal to the true value of the corresponding parameters for all $j$s.
   1-2) Another five initial values are based on the sample mean vector and the sample covariance matrix of the data. To get the first initial value for $\theta$, we set the $j$-th component mean vector equal to the sample mean plus $0.5 \times j \times (-1)^j \times \delta$ where $\delta$ is the $p$-dimensional vector with all its elements equal to the trace of the sample covariance matrix. Then set the initial values of $p_j$ and $\Sigma_j$ are equal to $1/m$ and the sample covariance matrix, respectively, for all $j$s. For the other four initial values, we then apply the same perturbation process as above to the sample mean vector.

2. two random-start approaches [25, p. 55]
   2-1) We randomly partition the data into $m$ groups corresponding to the $m$ components of the mixture model.
   2-2) We obtain the initial values for the $m$ component means by randomly generating from a $p$-variate normal distribution with mean equal to the sample mean and covariance equal to the sample covariance matrix of the data. We then specify the sample covariance matrix and $1/m$ as the initial values for $\Sigma_j$ and $p_j$, respectively, for all $j$s.

3. two clustering approaches, $k$-means and Linde-Buzo-Gray (LBG) method [21]. The LBG method is similar to the $k$-means method for the purpose of data clustering, but it is known to achieve better performance in terms of minimizing the total within class distance.
In the simulation studies and data analysis we employed multiple initial values for the parameters from the three types of approach described above.

**Remark 1.** Approach 1-1 is not available in practice because this approach was based on the true parameter values. There are two reasons for employing Approach 1 in our article. First, for fair comparison with the PMLE proposed by Chen et al. [4] and Chen and Tan [3], we include Approach 1 (1-1 and 1-2) that was used in their papers as initial values of the algorithm. Second, we intended to confirm that existence of spurious solution cannot be avoided even when the true parameter value was used as a starting value [29, p. 97].

Regarding the computation of the PMLE, we considered minimum penalty on $\theta$ proposed in Chen et al. [4] and Chen and Tan [3]:

$$-a_n \sum_{j=1}^{m} \left\{ tr(S_x \Sigma_j^{-1}) + \log | \Sigma_j | \right\}$$

where $a_n = n^{-1}$, $S_x$ is the sample covariance matrix of the data and $tr(\cdot)$ is the trace function. Note that this penalty function leads to an explicit solution for $\theta$ in the M-step of the EM algorithm. As pointed out in Chen and Tan [3], applying the penalty described above is equivalent to putting a Wishart distribution prior on $\Sigma_j$.

5.2. Simulation studies

In the simulation study we consider the seven normal mixture models shown in Table 3. In order to generate sample from high-dimensional normal mixture models with larger numbers of clusters (i.e., $m > 3$ and $p > 2$), we employed a method proposed by Maitra and Melnykov [23] that was implemented in R-package “MixSim”. They first defined overlap between two normal clusters as the sum of their misclassification probabilities and developed an algorithm designed to simulate high-dimensional normal mixture components satisfying prespecified levels of average ($\bar{\omega}$) and/or maximum ($\tilde{\omega}$) overlap between every pair of components. We here choose two sets of values of $(\bar{\omega}, \tilde{\omega})$, (0.001, 0.004) and (0.05, 0.15), that generate two five-dimensional normal mixture models with six components, Model V and Model VI, respectively. The mixture components for Model V are much better separated than those for Model VI. Note that we did not include the true values for the parameters in Model V and VI due to limited space; they are available on the corresponding author’s website (http://www.math.umass.edu/~daeyoung/).
Table 3: Simulation models: \( m \)-component \( p \)-dimensional normal mixture model with unequal covariances (or variances for a case of \( p=1 \)). Note that \( n \) is the sample size.

<table>
<thead>
<tr>
<th>Model ((m, p, n))</th>
<th>( p_1=0.3, \mu_1=(0,0)), ( \Sigma_1=(0.5\ 1) )</th>
<th>( p_2=0.7, \mu_2=(0,0)), ( \Sigma_2=(0.6\ 1) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>I ((2,2,200))</td>
<td>(</td>
<td>\times</td>
</tr>
<tr>
<td>II ((2,2,300))</td>
<td>(</td>
<td>\times</td>
</tr>
<tr>
<td>III ((3,2,300))</td>
<td>(</td>
<td>\times</td>
</tr>
<tr>
<td>IV ((3,2,300))</td>
<td>(</td>
<td>\times</td>
</tr>
<tr>
<td>V ((6,5,625))</td>
<td>(</td>
<td>\times</td>
</tr>
<tr>
<td>VI ((6,5,625))</td>
<td>(</td>
<td>\times</td>
</tr>
<tr>
<td>VII ((2,1,50))</td>
<td>(</td>
<td>\times</td>
</tr>
</tbody>
</table>

We first generate 1000 data sets of size \( n \) for each model. In order to compute the RMLE, PMLE, MLE\(_L_k\) and MLE\(_S_k\) for \( \theta \) of Eq (5.1) at each generated dataset, we employ the EM algorithm with multiple starting values. Regarding Model I, II, III, IV and VII, we used 25 initial values: the first ten initial values are based on two perturbation approaches (five for each 1-1) and 1-2), respectively), the next ten values are from two random-start approaches (five for each 2-1) and 2-2), respectively) and the last five values are obtained by two clustering approaches, the \( k \)-means and the LBG method. With respect to Model V and VI, we used 50 initial values: the first twenty initial values are based on two perturbation approaches, the next twenty values are from two random-start approaches and the last ten values are obtained by two clustering approaches.

When we compute our proposed estimates, MLE\(_L_k\) and MLE\(_S_k\), for each data set, we use a range of \( k \) values between 1 and \( mp \) that include the values proposed in the guideline of Section 3.3 (i.e., the values between \( p \) and \( p(m-1) \)).

When one computes the RMLE, the EM algorithm starting from a certain initial value sometimes converges to a singular solution. Table 4 shows the number of occasions that the EM algorithm converges to a degenerate solution with singular covariance estimate(s) in our simulations. Note that the number of degeneracy is out of 25,000 (50,000) for each model, as we used 25 (50) starting values for each simulated data set.

For comparisons between the performances of the RMLE, PMLE, MLE\(_L_k\) and MLE\(_S_k\), one may want to see the empirical standard error and bias for
Table 4: Number of degeneracies in the EM algorithm

<table>
<thead>
<tr>
<th>Model I</th>
<th>Model II</th>
<th>Model III</th>
<th>Model IV</th>
<th>Model V</th>
<th>Model VI</th>
<th>Model VII</th>
</tr>
</thead>
<tbody>
<tr>
<td>4811</td>
<td>3607</td>
<td>8765</td>
<td>8405</td>
<td>12159</td>
<td>12990</td>
<td>5064</td>
</tr>
</tbody>
</table>

each individual parameter. However, in simulation studies concerning finite mixture models, one first needs to address the “label switching” problem, just as in Bayesian mixture model using Markov chain Monte Carlo (MCMC) computations [18, 31]. For example, for a given m-component mixture model and data set, there are m! parameters for θ whose mixture densities (and thus likelihood values) are exactly the same. Thus, the labels on the estimate for µ1 may switch back and forth across the data sets simulated from the same model. In our simulation we indeed observed that the labels on the estimates for the parameters were switched over different data sets simulated from the same model.

When mixture components are not completely separated, it is a very challenging task to relabel an estimate for the parameter. Thus, we consider the two label-invariant measures to compare performances of the four estimators. The first measure is the coverage probability of the likelihood ratio confidence region. That is, we count the proportion of cases where the true value of the parameters is included in the 95% likelihood-based confidence region for the parameter estimate. In this simulation study the number of components is fixed and known. Thus, the likelihood confidence region is constructed based on the likelihood ratio statistic evaluated at the true parameter value and parameter estimates: $R_{0.95} = \{ \theta : 2(\ell(\hat{\theta}, x) - \ell(\theta, x)) \leq c_{0.95} \}$ where $\hat{\theta}$ is the parameter estimate based on the random sample $x = \{x_1, \ldots, x_n\}$ generated from the true parameter value, $\theta$, and $c_{0.95}$ is the 5% upper quantile of the chi-squared distribution with degrees of freedom equal to the dimension of $\theta$.

In order to obtain more accurate upper quantile, one might try to use the empirical distribution of the likelihood ratio statistic. However, due to possible existence of spurious solutions with the largest likelihood, the empirical distribution based on such undesirable solutions might not provide valid upper quantile unless one completely removes them [26]. Note that for the PMLE we constructed the penalized log likelihood confidence region.

The second label-invariant measure is a graphical tool, the permutation scatter plot that is designed to picture the sampling distributions of the estimators [20, 31]. Since there are more than one possible label on the estimate of $\theta$, the permutation scatter plot contains each estimate in all
Table 5: Coverage probabilities for the four estimators, RMLE, MLE\(^L_{-k}\), MLE\(^S_{-k}\) and PMLE in the five simulation models. Note that the nominal level is 0.95.

<table>
<thead>
<tr>
<th>(k)</th>
<th>RMLE</th>
<th>MLE(^L_{-k})</th>
<th>MLE(^S_{-k})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model I</td>
<td>0.944</td>
<td>0.952</td>
<td>0.952</td>
</tr>
<tr>
<td>Model II</td>
<td>0.611</td>
<td>0.624</td>
<td>0.778</td>
</tr>
<tr>
<td>Model III</td>
<td>0.892</td>
<td>0.894</td>
<td>0.895</td>
</tr>
<tr>
<td>Model IV</td>
<td>0.446</td>
<td>0.466</td>
<td>0.513</td>
</tr>
<tr>
<td>Model VII</td>
<td>0.889</td>
<td>0.929</td>
<td>0.931</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(k)</th>
<th>PMLE</th>
<th>MLE(^S_{-k})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model I</td>
<td>0.954</td>
<td>0.952</td>
</tr>
<tr>
<td>Model II</td>
<td>0.797</td>
<td>0.712</td>
</tr>
<tr>
<td>Model III</td>
<td>0.896</td>
<td>0.899</td>
</tr>
<tr>
<td>Model IV</td>
<td>0.568</td>
<td>0.553</td>
</tr>
<tr>
<td>Model VII</td>
<td>0.961</td>
<td>0.932</td>
</tr>
</tbody>
</table>

Its permuted versions, one for each possible labelling. Kim [20] and Yao and Lindsay [31] illustrated that the permutation scatter plot is very useful for comparing the performance of different estimators in terms of sampling distributions. If MLE\(^L_{-k}\) and MLE\(^S_{-k}\) based on our proposed criteria are effective in avoiding spurious solutions including nearly singular solutions, one can expect that the proposed estimators will have fewer estimates for the parameters around the boundary of the parameter space than the RMLE. For space limitation we do not present the permutation scatter plots for Model V and VI. They are available upon request.

Table 5 shows the coverage probabilities of the likelihood ratio confidence region at the four estimators, RMLE, MLE\(^L_{-k}\), MLE\(^S_{-k}\) and PMLE, in the five simulation models, I, II, III, IV and VII. Note that the nominal level is 0.95. We can see that the PMLE, MLE\(^L_{-k}\) and MLE\(^S_{-k}\) show better coverage probabilities than the RMLE. When the components were relatively well-separated (Model I and III), the value of \(k\) did not change the coverage probabilities of the likelihood confidence region. Thus, the minimum value of \(k\), \(p=2\), is appropriate. Note that the proposed estimators, MLE\(^L_{-k}\) and MLE\(^S_{-k}\), showed the coverage probabilities comparable to those of the PMLE.

In the Model II, IV and VII where the components were poorly separated, both proposed estimates had much better coverage probabilities than the RMLE, regardless of the value of \(k\). In this case one can improve the
Figure 3: Plots of coverage probabilities as a function of $k$ for RMLE (solid line), $\text{MLE}^L_{-k}$ (cross), $\text{MLE}^S_{-k}$ (circle) and PMLE (dashed-dot line) in the Model V (left plot) and VI (right plot).

Performance of the proposed $k$-deleted MLE’s by increasing the value of $k$, and the degree of improvement appeared to become slow as $k$ increased. Note that in the Model II and IV the $\text{MLE}^S_{-k}$ had better coverage probabilities than the PMLE, even when $k$ is equal to 2(=p), the minimum value of $k$.

Figure 3 gives the plots of coverage probabilities as a function of $k$ for the four estimators, RMLE (solid line), $\text{MLE}^L_{-k}$ (cross), $\text{MLE}^S_{-k}$ (circle) and PMLE (dashed-dot line) in the Model V (left plot) and VI (right plot). The left plot for Model V shows that all coverage probabilities of the four estimators were similar, and the coverage probabilities for the $\text{MLE}^L_{-k}$ and $\text{MLE}^S_{-k}$ did not change as $k$ increased; the coverage probabilities for RMLE, $\text{MLE}^L_{-k}$, $\text{MLE}^S_{-k}$ and PMLE were 0.939, 0.942, 0.942 and 0.945, respectively. As to Model VI where the components were not well-separated relative to the sample size, we observe from the right plot in Figure 3 that the $\text{MLE}^S_{-k}$ performed significantly better than the other estimators even when $k$ was equal to the proposed minimal value, 5(=p). Note that $\text{MLE}^L_{-k}$ still showed better performance than RMLE, but was not better than PMLE even with a large value of $k$.

Figure 4 shows the permutation plots for $(p_1, \mu_{21})$ of the four estimators in Model I and II. Here $p_1$ and $\mu_{21}$ are the mixing weight for the first component and the second coordinate of the mean parameter vector for the first component in a permuted estimate, respectively. The marginal histograms for permuted estimates of $p_1$ and $\mu_{21}$ are on the horizontal and vertical axes.
of the plot, respectively. For limitation of space, we show the permutation plots for \( \text{MLE}^L_{-k} \) and \( \text{MLE}^S_{-k} \) when \( k=2 \); the permutation plots for other values of \( k \) and for other parameters are available upon request.

From Figure 4(a), we observe that the RMLE still selected spurious solutions near the boundary of \( p_1=0 \) or 1, even when the components were well-separated (Model I). Figure 4(c),(e) and (g) show that that \( \text{MLE}^L_{-2} \), \( \text{MLE}^S_{-2} \) and PMLE successfully chose solutions to the likelihood equations closer to the true value when the RMLE chose the spurious solutions.

Figure 4(b), (d), (f) and (h) show the permutation plots for \((p_1, \mu_{21})\) in Model II. We observe that one of the proposed estimators, \( \text{MLE}^S_{-2} \), shows much desirable behavior of the sampling distribution, as the number of estimates near the boundary of the parameter space for \( \text{MLE}^S_{-2} \) was much smaller than those for the RMLE and \( \text{MLE}^L_{-2} \). The shape of the marginal distribution of permuted \( \mu_{21} \) for \( \text{MLE}^S_{-2} \) is more reasonable than those for the other two estimates, as it clustered around the true values. Moreover, the marginal sampling distribution of permuted \( p_1 \) for \( \text{MLE}^S_{-2} \) appeared to be uniformly distributed over a wide range of a mixing weight, instead of being almost degenerate at 0 or 1. This appears to be sensible because the separations of components was so poor that the number of modes in the mixture density of Model II was one and this model did not contain enough information about \( p_1 \). Note that the (joint/marginal) sampling distributions of PMLE for \((p_1, \mu_{21})\) showed similar behavior as those of \( \text{MLE}^S_{-2} \).

Figure 5 shows the permutation plots for \((p_1, \mu_{21})\) of the four estimators in Model III and IV when \( k=2 \). Likewise, we have similar observations from Figure 5. For the case of relatively well-separated components (Model III), \( \text{MLE}^L_{-2} \), \( \text{MLE}^S_{-2} \) and PMLE were effective in removing spurious solutions, compared with the RMLE (see Figure 5(a), 5(c), 5(e), 5(g)). When the components were poorly separated (Model IV), the \( \text{MLE}^S_{-2} \) tended to choose solutions to the likelihood equations closer to the true value, especially for the mixing weight (see Figure 5(b), 5(d) and 5(f)). Note that there are many cases where the PMLE appeared to be spurious solutions in the sense that the minimums of the estimated mixing weights were close to zero even though the minimum value of the true mixing weights is not (see the marginal distribution of mixing weights in Figure 5(h)). In such cases one needs to use stronger penalty terms in the computation of the PMLE so that they can make the log-likelihood function much smoother [3].
5.3. Data analysis

In this subsection we provide three real examples to show application of the proposed methods. Note that the first two examples are ones introduced in Section 2. We employed the EM algorithm with several initial values from the second and third types of approach described in Section 2.

Example 1 (revisited). As shown in Table 1, the RMLE is \( \hat{\theta}_2 \) by the largest local likelihood criterion because \( \hat{\theta}_1 \) is a singular solution so that we remove it from a candidate of solutions to the likelihood equation. We here employed the two proposed criteria, likelihood-based and score-based \( k \) deleted likelihood to all found singular and nonsingular solutions, with the two values of \( k \): 1 and 2. Note that a value of \( k \) proposed by the guideline of Section 3.3 was 1 because the data is univariate and the number of components is two (i.e., \( p=1 \) and \( m=2 \)). For both proposed methods, \( \hat{\theta}_2 \) was selected as a final solution. That is, \( \text{MLE}^L_k=\text{MLE}^S_k=\hat{\theta}_2 \) for \( k=1 \) and 2. We also computed the PMLE, which was close to \( \hat{\theta}_2 \).

Example 2 (revisited). In this example we found 14 non-singular local maxima from 100 starting values. By the largest local likelihood criterion one needs to choose \( \hat{\theta}_1 \) as the RMLE because it has the largest log likelihood value among the 14 local maximizers. We here computed the estimates from the two proposed criteria, with the four values of \( k \): 1, 2, 3 and 4. Note that the minimum value of \( k \) proposed by the guideline of Section 3.3 was 2, as \( p=2 \) and \( m=2 \). In terms of the likelihood-based \( k \) deleted likelihood, \( \text{MLE}^L_k=\hat{\theta}_1 \) for \( k=1 \), \( \text{MLE}^L_k=\hat{\theta}_2 \) for \( k=2 \) and \( \text{MLE}^L_k=\hat{\theta}_4 \) for \( k=3 \) and 4. With regard to the score-based \( k \) deleted likelihood, on the other hand, \( \text{MLE}^S_k \) was \( \hat{\theta}_5 \) for \( k=2,3,4 \), and \( \hat{\theta}_7 \) for \( k=1 \). Note that the PMLE was very close to \( \hat{\theta}_5 \).

Example 3. The third example concerns Fisher’s iris data used by Fisher [9] to illustrate principles of discriminant analysis. Data contains four variables, Petal width, Petal length, Sepal width and Sepal length, collected on flowers of three iris species: Setosa, Virginica and Versicolor. Each species had 50 observations. We fitted a three component \( (m=3) \) trivariate \( (p=3) \) normal mixture model with unequal covariances to the observations of the first three variables (Petal width, Petal length, Sepal width) in order to cluster 150 data points into three groups, without using the known classification.
6. DISCUSSION

This article proposed a root selection method designed to choose a statistically reasonable root when there exist multiple roots of the likelihood.
Table 7: Runtime (seconds) involved in running the EM algorithm and in computing $\text{MLE}^S_{-k}$ for a sample of size $n$ generated from the models used in Section 5.2: the first, second and third values in each parenthesis are number of solutions found, runtime involved in the EM algorithm and runtime involved in $\text{MLE}^S_{-k}$, respectively.

<table>
<thead>
<tr>
<th>Model</th>
<th>n=100</th>
<th>n=200</th>
<th>n=400</th>
<th>n=800</th>
</tr>
</thead>
<tbody>
<tr>
<td>II</td>
<td>(5, 14.902, 0.011)</td>
<td>(4, 29.078, 0.025)</td>
<td>(5, 59.257, 0.116)</td>
<td>(3, 114.750, 0.277)</td>
</tr>
<tr>
<td>IV</td>
<td>(7, 7.849, 0.017)</td>
<td>(6, 15.975, 0.041)</td>
<td>(4, 30.622, 0.097)</td>
<td>(4, 132.816, 0.380)</td>
</tr>
<tr>
<td>VI</td>
<td>(28, 10.22, 0.848)</td>
<td>(30, 24.27, 1.495)</td>
<td>(25, 57.369, 2.329)</td>
<td>(11, 84.909, 2.790)</td>
</tr>
<tr>
<td>VII</td>
<td>(5, 1.150, 0.011)</td>
<td>(2, 0.944, 0.013)</td>
<td>(6, 1.728, 0.132)</td>
<td>(3, 2.464, 0.265)</td>
</tr>
</tbody>
</table>

equation for finite normal mixture models. We illustrated, through extensive simulation and data analysis, that the two proposed methods using a guideline for the choice of $k$, the likelihood-based and the score-based $k$-deleted likelihood, are guaranteed to eliminate the largest possible number of singularities and showed their ability to eliminate the spurious local maximizers. We also showed that the solution selected by the proposed methods is asymptotically consistent. We recommend using the score-based $k$-deleted likelihood criterion, especially for the case where the existence of spurious local maximizers is main concern.

One might concern the computation time involved in the score-based $k$-deletion method because it requires the calculation of the score function for all found solutions. Table 7 shows the runtime (seconds) involved in running the EM algorithm and in computing $\text{MLE}^S_{-k}$ for a sample of size $n$ generated from each of the four models used in Section 5.2. Note that these four models had poorly-separated components, and we used the same stopping rule and initial value strategy described in Section 5.1 and 5.2. We observe from Table 7 that the computation time required for computing $\text{MLE}^S_{-k}$ is not a critical issue, compared with the runtime of the EM algorithm, even though there were multiple solutions to the likelihood equations for a given model.

In a mixture model analysis, it is common that one uses a large number of initial values and an appropriate stopping rule in the algorithm. This is because one wants to ensure solutions to the mixture likelihood equation that detect reasonable cluster structure inherent in the data sets. Using such strategy will also help the estimators based on our proposed criteria perform better, as our proposed methods assumes that a set of all found non-singular solutions includes desirable solutions to the mixture likelihood equation.

It is in advance difficult to know how many starting values are needed and what are good starting values. However, one can estimate probability
of finding a new local maximum [8]. If the estimated probability is small enough, one can stop trying new initial values. We often also observe that some spurious solutions correspond to typically the “rare” solutions that need very specific initial values so as the algorithm maximizing the likelihood to be trapped there. Thus, one can always check, by a strategy of multiple starting values, whether or not there exist a few local maximizers occurring many times that are perhaps the reasonable ones.

According to the results from our simulation experiments and data examples where the model assumptions were appropriate, the $k$-deleted MLE’s, in particular the score-based MLE, were quite stable over $k$ when $k$ was larger than the minimum of the proposed range, $p$. However, as one of the reviewers pointed out, there may exist cases where the $k$-deleted MLE’s keep changing over $k$ for given data although we did not observe such cases in our article. If the model assumptions are acceptable and the proposed methods give different solutions over the values of $k$ in the proposed range, we suggest that one consider what information the roots selected by our proposed methods provides about the fit of different models to the data. This is because, instead of using a single root for explaining the data in hand, such an instability over $k$ (the presence of multiple roots selected by our proposed method) may indicate the presence of different meaningful interpretation of the data, even when the model assumptions are appropriate.

In simulation studies and data analysis, we assumed that the model assumptions are acceptable, for instance, non-existence of outliers far from the majority of the data and appropriate component densities. At this moment it is not clear how the violation of the model assumptions have the effect on the mixture likelihood surface and the presence of its local maximizers. Thus, a valuable extension of this research would be to study the mixture likelihood function when the model assumptions are not satisfied and investigate the performance of the $k$-deleted MLE’s if there still exist multiple local maximizers in the mixture likelihood.

The optimal choice of $k$ which works for every data/mixture model situation is still an open problem. The essential problem regarding the choice of $k$ is that there is no well-defined mathematical definition of the spurious root in a finite sample. If one could mathematically define spurious roots, the spurious problem would disappear at least in a practical sense, because one can just ignore such well-defined spurious roots from the set of all found roots. In this sense, we hope that our proposed methods can be used as a tool to define the spurious root. For example, one may define the RMLE as
a "spurious" root when the RMLE and the $k$-deleted MLE are not identical for some $k$.

Throughout this paper, we assumed that the number of components is fixed by external modelling considerations. Most likelihood-based methods to choose the number of components (such as the BIC) strongly rely on the likelihood value at a chosen solution among all the found solutions. Hence, it will be an important future task to investigate the performance of likelihood-based order selection criteria based on the $k$-deleted MLE in the presence of spurious solutions to the likelihood equation for a finite mixture model.

**Acknowledgement**

We thank the editor, an associate editor and three anonymous referees whose helpful suggestions and insightful comments greatly improved the quality of this paper. The authors also thank Professor Bruce G. Lindsay for his insightful comments and suggestions. This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology (2011-0014607).

**APPENDIX: PROOFS**

We first assume the following regularity conditions for Theorem 1. Note that these conditions are standard conditions assumed in discussing consistency of the MLE [28].

(A1) For $\theta \in \Gamma$ and any positive real number $\rho$, let

$$f(x; \theta, \rho) \equiv \sup_{\text{dist}(\theta', \theta) \leq \rho} f(x; \theta'),$$

where $\text{dist}(\theta, \theta')$ is the ordinary Euclidean distance between $\theta$ and $\theta'$. For each $\theta \in \Gamma$ and sufficiently small $\rho$, $f(x; \theta, \rho)$ is measurable.

(A2) For each $\theta \in \Gamma$, if $\lim_{n \to \infty} \theta_n = \theta$ then $\lim_{n \to \infty} f(x; \theta_n) = f(x; \theta)$ except on a set of probability zero and does not depend on the sequence $\{\theta_n\}_{n=1}^\infty$. This means that the map from the parameter space to the distribution space is continuous.

(A3)

$$\int |\log f(x; \theta_0)| f(x; \theta_0) dx < \infty.$$
In addition, [28] imposed the following condition to generalize their consistency result, which clearly holds for normal mixtures.

(A4) There exist real constants $v_0, v_1 > 0$ and $\beta > 1$ such that

$$\phi(x) \leq \min\{v_0, v_1|x|^{-\beta}\}.$$ 

**A.1. Proof of Theorem 2**

From Wald [30], it is sufficient to find a constant $B$ satisfying $\ell_n(\hat{\theta}'_n) - \ell_n(\theta_0) \geq B > -\infty$. Now Eq (4.1) and Eq (4.2) implies

$$\ell_n(\hat{\theta}'_n) \leq \ell_n(\hat{\theta}_n), \quad \ell_n(\theta_0) \leq \ell_n(\hat{\theta}_n), \quad (6.1)$$

and

$$\ell_n(\hat{\theta}'_n) - \max_{i}\{\log(f(x_i; \hat{\theta}'_n))\} \geq \ell_n(\hat{\theta}_n) - \max_{i}\{\log f(x_i; \hat{\theta}_n)\}, \quad (6.2)$$

where $\ell_n(\theta) = \sum_i \log(f(x_i, \theta))$. Since $\hat{\theta}_n$ almost surely converges to $\theta_0$ from Theorem 1 and $f(x; \theta_0)$ is bounded above, one can find a constant $A$ satisfying $\log f(x_i; \hat{\theta}_n) \leq A$ for all $n$ sufficiently large. Together this with Eq (6.1) and Eq (6.2), we have

$$\ell_n(\hat{\theta}'_n) - \max_{i}\{\log(f(x_i; \hat{\theta}'_n))\} \geq \ell_n(\theta_0) - A.$$ 

and

$$\ell_n(\hat{\theta}'_n) - \ell_n(\theta_0) \geq \max_{i}\{\log(f(x_i; \hat{\theta}'_n))\} - A \geq \frac{1}{n-1} \sum_{j \neq j^*} \log f(x_j; \hat{\theta}'_n) - A \geq \frac{1}{n-1} \sum_{j \neq j^{**}} \log f(x_j; \hat{\theta}_n) - A, \quad (6.3)$$

where $j^* = \arg\max_j \{\log(f(x_j; \hat{\theta}'_n))\}$ and $j^{**} = \arg\max_j \{\log(f(x_j; \hat{\theta}_n))\}$. Since $\log f(x_{j^*}; \hat{\theta}'_n)$ is the largest log-likelihood term at $\hat{\theta}'_n$, it should be greater than the average of the remaining log-likelihood terms. So the second inequality in Eq (6.3) holds and Eq (6.2) implies the last inequality in Eq (6.3). Now the first term of the last inequality in Eq (6.3) is again bounded below as $\hat{\theta}_n \to \theta_0$. 

30
A.2. Proof of Lemma 1
Let $K = [\mu_L, \mu_U] \times [\sigma_L, \sigma_U] \times [p_L, p_U]$. Assumption 1 implies there exist a component $j_n$ whose parameter is in $K$ for any given likelihood solution $\theta_n^{(s)}$. The lower bound of $\log f(x^*, \theta_n^{(s)})$ can be then obtained from

$$\log f(x^*, \theta_n^{(s)}) \geq \log \left( \frac{p_{nj_n}}{\sigma_{nj_n}} \phi \left( \frac{x_n^* - \mu_{nj_n}}{\sigma_{nj_n}} \right) \right) = \log \left( \frac{p_{nj_n}}{\sigma_{nj_n}} \right) + \log \phi \left( \frac{x_n^* - \mu_{nj_n}}{\sigma_{nj_n}} \right) \geq \log(p_L/\sigma_U) + \log \left[ \phi \left( \frac{\max(|M - \mu_L|, |M + \mu_U|)}{\sigma_L} \right) \right]$$

This implies $\log f(x^*; \theta_n) \geq K$, where

$$K = \log(p_L/\sigma_U) + \log \left[ \phi \left( \frac{\max(|M - \mu_L|, |M + \mu_U|)}{\sigma_L} \right) \right]$$

and $M$ is the constant in Eq (4.5).

A.3. Proof of Theorem 3
Similar to the proof of Theorem 2, it suffices to find constant $B$ such that $\ell_n(\hat{\theta}_n) - \ell_n(\theta_0) \geq B > -\infty$. Again, from Eq (4.1) and Eq (4.3), we have

$$\ell_n(\hat{\theta}_n^\prime) \leq \ell_n(\hat{\theta}_n), \quad \ell_n(\theta_0) \leq \ell_n(\hat{\theta}_n),$$

and

$$\ell_n(\hat{\theta}_n) - \log f(x^*; \hat{\theta}_n) \leq \ell_n(\hat{\theta}_n^\prime) - \log f(x^*; \hat{\theta}_n^\prime),$$

where

$$x^* = \text{argmax}_{x \in \Xi_n \cap \{x| M \geq ||x_i||\}} ||u_i(\hat{\theta}_n)|| \quad \text{and} \quad x^{**} = \text{argmax}_{x \in \Xi_n \cap \{x| M \geq ||x_i||\}} ||u_i(\hat{\theta}_n)||.$$

With these, we can see that

$$\ell_n(\hat{\theta}_n^\prime) - \ell_n(\theta_0) = \ell_n(\hat{\theta}_n^\prime) - \ell_n(\hat{\theta}_n) + \ell_n(\hat{\theta}_n) - \ell_n(\theta_0) \geq \log f(x^*; \hat{\theta}_n^\prime) - \log f(x^{**}; \hat{\theta}_n) + \ell_n(\hat{\theta}_n) - \ell_n(\theta_0) \geq K - \sup_x \log f(x; \hat{\theta}_n) + \ell_n(\hat{\theta}_n) - \ell_n(\theta_0),$$
where $K$ is the lower bound in Lemma 1. Since $\hat{\theta}_n$ almost surely converges to $\theta_0$ from Theorem 1, there exists $b$ satisfying $\ell_n(\hat{\theta}_n) - \ell_n(\theta_0) \geq b > -\infty$. Moreover, for any given $\epsilon > 0$, we have $\log f(x; \theta_0) - \epsilon < \log f(x; \hat{\theta}_n) < \log f(x; \theta_0) + \epsilon$ almost surely for a sufficiently large $n$. Note also that there exists $A$ such that $\log f(x; \theta_0) < A < \infty$ because $f(x; \theta_0)$ is bounded above as long as $\theta_0$ is an interior point of $\Gamma$. These imply

$$
\ell_n(\hat{\theta}_n) - \ell_n(\theta_0) \geq K - \sup_x \log f(x; \hat{\theta}_n) + \ell_n(\hat{\theta}_n) - \ell_n(\theta_0)
$$

$$
\geq K - \sup_x \log f(x; \theta_0) - \epsilon + b
$$

$$
\geq K - A - \epsilon + b > -\infty.
$$

This completes the proof.

A.4. Proof of Proposition 1

$$
P \left( \min_i |X_i - \mu_{n1}| < \sigma_{n1} \right) = 1 - \left[ P \left( |X_i - \mu_{n1}| > \sigma_{n1} \right) \right]^n
$$

$$
= 1 - \left[ F_0(\mu_{n1} - \sigma_{n1}) + 1 - F_0(\mu_{n1} + \sigma_{n1}) \right]^n
$$

$$
= 1 - \left[ 1 - (F_0(\mu_{n1} + \sigma_{n1}) - F_0(\mu_{n1} - \sigma_{n1})) \right]^n
$$

$$
= 1 - \left[ 1 - 2\sigma_{n1} \frac{(F_0(\mu_{n1} + \sigma_{n1}) - F_0(\mu_{n1} - \sigma_{n1}))}{2\sigma_{n1}} \right]^n,
$$

where $F_0$ is the true distribution of $X$. From the mean value theorem, there exists $a_n \in [\mu_{n1} - \sigma_{n1}, \mu_{n1} + \sigma_{n1}]$ such that

$$
\frac{F_0(\mu_{n1} + \sigma_{n1}) - F_0(\mu_{n1} - \sigma_{n1})}{2\sigma_{n1}} = f_0(a_n),
$$

where $f_0 = F_0'$. Then

$$
P \left( \min_i |x_i - \mu_{n1}| < \sigma_{n1} \right) = 1 - \left[ 1 - 2f_0(a_n)\sigma_{n1} \right]^n
$$

Since $\sigma_{n1} \leq Ce^{-n^d}$ and $f_0(a_n)$ is bounded by a constant $D$, we have

$$
\sum_{n=1}^{\infty} P \left( \min_i |x_i - \mu_{n1}| < \sigma_{n1} \right) \leq \sum_{n=1}^{\infty} \left[ 1 - \{1 - 2DCe^{-n^d}\} \right]^n.
$$

With some calculations, one can now show

$$
\left[ 1 - \{1 - 2DCe^{-n^d}\} \right]^n < \frac{DC}{n^{1+d}}
$$

for a sufficiently large $n$ and this implies $\sum_{n=1}^{\infty} P \left( \min_i |x_i - \mu_{n1}| < \sigma_{n1} \right) < \infty$. Applying Borel-Cantelli lemma then finishes the proof.
References


Figure 4: Permutation scatter plot with the marginal histograms for \((p_1, \mu_{21})\), in the RMLE (a,b), MLE\(_L\) (c,d), MLE\(_S\) (e,f) and PMLE (g,h): Model I (a,c,e,g) and Model II (b,d,f,h). Note that black square are the permuted true values for \((p_1, \mu_{21})\).
Figure 5: Permutation scatter plot with the marginal histograms for \((p_1, \mu_{21})\), in the RMLE (a, b), MLE\(_L^{-2}\) (c, d), MLE\(_S^{-2}\) (e, f) and PMLE (g, h): Model III(a, c, e, g) and Model IV(b, d, f, h). Note that black square are the permuted true values for \((p_1, \mu_{21})\).