Regularized Parameter Estimation in High-Dimensional Gaussian Mixture Models

Lingyan Ruan
lruan@gatech.edu
Ming Yuan
ming.yuan@isye.gatech.edu
School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, U.S.A.

Hui Zou
hzou@stat.umn.edu
School of Statistics, University of Minnesota, Minneapolis, MN 55455, U.S.A.

Finite gaussian mixture models are widely used in statistics thanks to their great flexibility. However, parameter estimation for gaussian mixture models with high dimensionality can be challenging because of the large number of parameters that need to be estimated. In this letter, we propose a penalized likelihood estimator to address this difficulty. The ℓ_1-type penalty we impose on the inverse covariance matrices encourages sparsity on its entries and therefore helps to reduce the effective dimensionality of the problem. We show that the proposed estimate can be efficiently computed using an expectation-maximization algorithm. To illustrate the practical merits of the proposed method, we consider its applications in model-based clustering and mixture discriminant analysis. Numerical experiments with both simulated and real data show that the new method is a valuable tool for high-dimensional data analysis.

1 Introduction

In finite gaussian mixture models, a p-dimensional random vector $X = (X^{(1)}, \ldots, X^{(p)})$ is assumed to come from a mixture distribution,

$$\pi_1\mathcal{N}(\mu_1, \Sigma_1) + \pi_2\mathcal{N}(\mu_2, \Sigma_2) + \cdots + \pi_M\mathcal{N}(\mu_M, \Sigma_M), \tag{1.1}$$

where $\mathcal{N}(\mu, \Sigma)$ is a multivariate normal distribution with mean vector μ and covariance matrix Σ, and π_ks are nonnegative proportions such that $\pi_1 + \cdots + \pi_M = 1$. Gaussian mixture models are among the most popular statistical modeling tools and are routinely used for density estimation, clustering, are discriminant analysis among others (see, e.g., Fraley & Raftery, 2002; McLachlan & Peel, 2000).

Despite its great flexibility, the practical use of gaussian mixture models in modeling high-dimensional data is often hampered by difficulty in parameter estimation. The number of parameters required to specify a covariance matrix quickly grows with the dimensionality. The problem is exacerbated in mixture models where multiple covariance matrices are to be estimated. Without any parameter restriction, each cluster must have at least $(p + 1)$ observations to ensure the existence of the maximum likelihood estimate (MLE; Symons, 1981). As a result, it is well known that the usual MLE can be notoriously unstable, if well defined at all, when the data are of moderate or high dimensionality when compared with the sample size. To address this issue, a variety of parameter-reduction techniques have been developed. In particular, Banfield and Raftery (1993) suggest reparameterizing Σ_k through its eigenvalue decomposition and assume that through this parameterization, some parameters are shared across clusters. Extensive studies within the same framework can also be found in Celeux and Govaert (1995). Later work has since demonstrated that through parameter sharing across clusters, the problem of estimating a large number of parameters can be alleviated for data of moderate dimensions. The challenge, however, persists for high-dimensional data, as the number of parameters remains of the order of p^2 even if a common covariance matrix is assumed for all clusters.

In this letter, we propose a new technique to address this challenge. Built on recent advances in estimating covariance matrices of high-dimensional multivariate gaussian distributions, we propose a penalized likelihood estimate for high-dimensional gaussian mixture models. The ℓ_1 type of penalty we employ encourages sparsity of the inverse covariance matrices and therefore can help reduce the effective dimensionality of the problem. We show that the proposed estimate can be conveniently computed using an EM algorithm. Moreover, a BIC type of criterion is introduced to select the tuning parameter as well as the number of clusters. Numerical experiments, both simulated and real data examples, are presented to demonstrate the merits of the proposed method.

Our method could prove useful for a variety of statistical problems. For illustration purposes, we consider in particular model-based clustering (Fraley & Raftery, 2002) and mixture discriminant analysis (Hastie & Tibshirani, 1996), two notable methods that take advantage of the flexibility of finite gaussian mixture models in clustering and classification, respectively. We demonstrate that with the proposed ℓ_1 penalized estimator, both approaches can be substantially improved when dealing with high-dimensional problems.

Our investigation here is related to recent studies on parameter estimation in high-dimensional multivariate gaussian distribution, which can also be viewed as a special case of finite gaussian mixture models (see equation 1.1) with $M = 1$. A number of methods have been introduced in the past several years to estimate the covariances matrix with high-dimensional data (Ledoit & Wolf, 2004; Huang, Liu, Pourahmadi, & Liu, 2006; Li & Gui, 2006; Yuan & Lin, 2007; Banerjee, El Ghaoui, & d'Aspremont, 2008; Bickel & Levina, 2008a, 2008b; Rothman, Bickel, Levina, & Zhu, 2008; d'Aspremont, Banerjee, & El Ghaoui, 2008; El Karoui, 2008; Fan, Fan, & Lv, 2008; Friedman, Hastie, & Tibshirani, 2008; Lam & Fan, 2009; Levina, Rothman, & Zhu, 2008; Rothman, Bickel, Levina, & Zhu, 2008; Yuan, 2008; Deng & Yuan, 2009; Rothman, Levina, & Zhu, 2009, among others). A common strategy is to work with the sample covariance matrix, which is readily computable regardless of the dimensionality (see Bickel & Levina, 2008a). For the more general finite gaussian mixture model, we no longer have the luxury of such an initial estimate, and regularization as employed here becomes critical. We thus adopt the idea of a penalized likelihood estimate from Yuan and Lin (2007) and apply an ℓ_1-type penalty on the off-diagonal entries of the inverse covariance matrices.

The rest of the letter is organized as follows. In sections 2 and 3, we introduce the proposed penalized likelihood estimator and discuss how it can be efficiently computed in practice. Section 4 presents numerical studies to demonstrate the practical merits of the proposed method. Applications of the new method to model-based clustering and mixture discriminant analysis are discussed in section 5. We conclude with some comments and discussions in section 6.

2 Methodology

We start with the case when the number of clusters, M, is known a priori. In this case, the log likelihood for a sample $X_1, \ldots, X_n$ of n independent copies of X is given by

$$\mathrm{L}(\mathrm{data}|\Theta) = \sum_{i=1}^{n} \log\left(\sum_{k=1}^{M} \pi_k \phi(X_i|\mu_k, \Sigma_k)\right), \tag{2.1}$$

where $\Theta = \{(\pi_k, \mu_k, \Sigma_k) : k = 1, \ldots, M\}$ is the collection of all unknown parameters, and $\phi(\cdot|\mu, \Sigma)$ is the density function of a multivariate gaussian distribution with mean vector μ_k and covariance matrix Σ_k.

The usual MLE can be computed by maximizing $\mathrm{L}(\mathrm{data}|\Theta)$ with respect to Θ. Without any constraints on the parameter, Θ includes a total of $Mp(p+1)/2$ free parameters, which can be prohibitive from both a statistical and computational point of view when p is moderate or large when compared with the sample size n. To address this problem, we suggest exploiting potential sparsity in the covariance matrix. Sparsity can be found in multiple ways for covariance matrix estimation. In particular, we consider sparsity on the entries of the inverse covariance matrix. In the case of multivariate gaussian distribution, the inverse covariance matrix collects all the partial correlations, and a zero entry of the inverse covariance matrix corresponds to the conditional independence between the corresponding variables given the remaining ones. This relationship naturally connects with the so-called gaussian graphical models (Whittaker, 1990; Lauritzen, 1996) and makes this type of sparsity particularly suitable for many applications. Similar interpretation can also be given to the gaussian mixture models where each cluster can be viewed as instances of a particular gaussian graphical model. For this purpose, we suggest using the following penalized likelihood estimate for gaussian mixture models,

$$\hat{\Theta} := \underset{\mu_k, \Sigma_k \succ 0}{\operatorname{argmin}} \left\{ -\sum_{i=1}^{n} \log \left(\sum_{k=1}^{M} \pi_k \phi(X_i | \mu_k, \Sigma_k) \right) + \lambda \sum_{k=1}^{M} \| \Sigma_k^{-1} \|_{\ell_1} \right\}, \tag{2.2}$$

where $\Sigma \succ 0$ indicates that Σ is a symmetric and positive-definite matrix, $\lambda \geq 0$ is a tuning parameter, and $\|A\|_{\ell_1} = \sum_{i \neq j} |a_{ij}|$. Obviously when $M = 1$ the estimate defined above reduces to the so-called graph lasso estimate of Yuan and Lin (2007).

Thus far, we have treated the number of clusters M and the tuning parameter λ as fixed. In practice, the choice is critical in determining the performance of our method. A commonly used strategy to choose these parameters is the multifold cross-validation(CV). In CV, the data are first split into training and testing sets. For each pair of tuning parameters (M, λ), we compute the penalized likelihood estimate on the training data and then evaluate its performance on the testing data. The split, estimation, and evaluation are repeated many times to obtain a score for each pair of tuning parameters. The pair associated with the optimal score is then used for computing the final estimate based on all data. Despite its general applicability and competitive performance, a major drawback of CV is the intensive computation it requires. To overcome this problem, we suggest a BIC type of criterion as an alternative to the CV score.

Following Yuan and Lin (2007), the degrees of freedom for each estimated covariance matrix using the ℓ_1 type of regularization can be approximated by the number of nonzero entries in the upper half of the inverse covariance matrix. Therefore, the total number of degrees of freedom can be approximated by

$$\mathrm{df}(M, \lambda) = \sum_{k=1}^{M} \left(p + \sum_{i \leq j} I((\hat{\Sigma}_k^{-1})_{ij} \neq 0) \right), \tag{2.3}$$

where p represents the degrees of freedom associated with the unknown mean and $\hat{\Sigma}_k$ is the penalized likelihood estimate associated with tuning parameters (M, λ). For each pair of (M, λ), the corresponding BIC score function is defined as

$$\mathrm{BIC}(M, \lambda) = -\mathrm{L}(X|\hat{\Theta}(M, \lambda)) + \log(n)\,\mathrm{df}(M, \lambda). \tag{2.4}$$

Let $(\hat{M}, \hat{\lambda})$ be the pair with the smallest BIC score, and we let $\hat{\Theta}(\hat{M}, \hat{\lambda})$ be our final estimate.

3 Computation

Direct computation of $\hat{\Theta}$ as defined by equation 2.2 can be quite complicated because the objective function is nonconvex and the optimization problem is of rather high dimensionality. Fortunately, we show here that it can be efficiently done using an expectation-maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977). To this end, we consider the following "missing data" formulation. Let τ be a random variable indicating which cluster X comes from such that

$$X|\tau = k \sim \mathcal{N}(\mu_k, \Sigma_k) \tag{3.1}$$

and

$$P(\tau = k) = \pi_k, \qquad k = 1, \ldots, M. \tag{3.2}$$

If we can observe the "complete data" (X_i, τ_i), $i = 1, \ldots, n$, we can follow the same strategy as before and estimate Θ by the ℓ_1 penalized log likelihood, which can be given by

$$\begin{aligned}\mathrm{PL}(\Theta) &= \sum_{i=1}^{n} \log\left(f(X_i|\tau_i)P(\tau_i)\right) + \lambda \sum_{k=1}^{M} \|\Sigma_k^{-1}\|_{\ell_1} \\ &= \sum_{i=1}^{n} \log\left(\phi(X_i|\mu_{\tau_i}, \Sigma_{\tau_i})\pi_{\tau_i}\right) + \lambda \sum_{k=1}^{M} \|\Sigma_k^{-1}\|_{\ell_1}.\end{aligned} \tag{3.3}$$

Now that we can observe only X_is, we may treat τ_is as missing data, and the following EM algorithm can therefore be employed. We proceed in an iterative fashion. Each iteration consists of the E-step and the M-step. Let $\Theta^{(t)}$ be the estimate of Θ at the tth iteration. In the E-step, we compute the conditional expectation of τ_i given X_i and the current estimate of Θ. In particular, from Bayes' rule,

$$\gamma_{ik}^{(t)} := P(\tau_i = k|X_i; \Theta^{(t)}) = \frac{\pi_k^{(t)}\phi\left(X_i|\mu_k^{(t)}, \Sigma_k^{(t)}\right)}{\sum_{l=1}^{M} \pi_l^{(t)}\phi\left(X_i|\mu_l^{(t)}, \Sigma_l^{(t)}\right)}. \tag{3.4}$$

This leads to construction of the so-called Q function,

$$
\begin{aligned}
&Q(\Theta, \Theta^{(t)}) \\
&\quad = \sum_{k=1}^{M} \left\{ \sum_{i=1}^{n} \log(\pi_k)\gamma_{ik}^{(t)} + \sum_{i=1}^{n} \log(\phi(X_i | \mu_k^{(t)}, \Sigma_k^{(t)}))\gamma_{ik}^{(t)} + \lambda \| \Sigma_k^{-1} \|_{\ell_1} \right\} \\
&\quad =: \sum_{k=1}^{M} Q_k(\Theta_k, \Theta_k^{(t)}),
\end{aligned}
$$

where $\Theta_k = \{\pi_k, \mu_k, \Sigma_k\}$ and $\Theta^{(t)}$ is defined in a similar manner.

In the M-step, we update the estimate of Θ by maximizing the Q function, which can be done by maximizing Q_k with respect to Θ_k separately. More specifically, the updated value of Θ_k can be given by

$$\pi_k^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} \gamma_{ik}^{(t)} \tag{3.5}$$

and

$$\mu_k^{(t+1)} = \frac{\sum_{i=1}^{n} \gamma_{ik}^{(t)} X_i}{\sum_{i=1}^{n} \gamma_{ik}^{(t)}}. \tag{3.6}$$

Moreover,

$$\Sigma_k^{(t+1)} = \operatorname{argmin}_{\Sigma_k} \left\{ \log |\Sigma_k| + \operatorname{tr}\left(\Sigma_k^{-1} A_k^{(t)}\right) + \lambda \| \Sigma_k^{-1} \|_{\ell_1} \right\}, \tag{3.7}$$

where

$$A_k^{(t)} = \sum_{i=1}^{n} \left(\frac{\gamma_{ik}^{(t)}}{\sum_{j=1}^{n} \gamma_{jk}^{(t)}} \right) (X_i - \mu_k^{(t+1)})(X_i - \mu_k^{(t+1)})'.$$

The optimization problem of equation 3.7 is in a similar form as the graph lasso of Yuan and Lin (2007) and can be computed efficiently using a newly developed algorithm by Friedman et al. (2008).

To sum up, we have the following algorithm to compute $\hat{\Theta}$ as defined by equation 2.2:

Step 1: Initialize $\Theta^{(0)}$.
Step 2: For each iteration, update the estimate for each mixture component individually.

- E-step: Calculate the distribution of unknown variables by equation 3.4.
- M-step: update parameters by equations 3.6, 3.7, and 3.5.

Step 3: Go back to step 2 until a certain convergence criterion is met.

Following the same argument as that of Dempster et al. (1977), it is not hard to see that in each iteration, the objective function of equation 2.2 decreases. Furthermore, the algorithm converges, and to its minimizer, $\hat{\Theta}$. We note that the choice of a good initial value may greatly reduce the number of iterations. Our experience suggests that the estimators given by Banfield and Raftery (1993) are often a good starting point.

4 Simulation Studies

To assess the finite sample performance of the proposed method, we now conduct several sets of simulation studies.

We begin with the case where the number of clusters is known in advance. In particular, we fix $M = 2$ in the first set of simulations. The sample size is set to be 100, whereas the dimension p is set to be 30, 50, 100, or 300. The tuning parameter λ is determined by either five-fold CV or the BIC criterion defined by equation 2.4. For simplicity, we fix the mean vector of each mixture component to be 0_p and three consider covariance structures:

Model 1: The covariance matrix for both clusters follows an AR(1) model:

$$\Sigma_1(i, j) = 0.4^{|i-j|}; \qquad \Sigma_2(i, j) = 0.5 \times 0.8^{|i-j|}. \tag{4.1}$$

Model 2: Both covariance matrices are diagonal:

$$\Sigma_1(j, j) = \log(j + 1); \qquad \Sigma_2(j, j) = \log(p + 2 - j). \tag{4.2}$$

Model 3: The two covariance matrices follow the AR(1) and AR(2) models, respectively:

$$\Sigma_1^{-1}(i, j) = \begin{cases} 1 & \text{if } i = j \\ 0.2 & \text{if } |i - j| = 1 \\ 0 & \text{otherwise} \end{cases} \tag{4.3}$$

and

$$\Sigma_2^{-1}(i, j) = \begin{cases} 2 & \text{if } i = j \\ 0.25 & \text{if } |i - j| = 1 \\ 0.2 & \text{if } |i - j| = 2 \\ 0 & \text{otherwise} \end{cases}. \tag{4.4}$$

We compare the proposed estimate with the method of Banfield and Raftery (1993) and MLE if applicable. This method of Banfield and Raftery has been implemented in the R package mclust, and the MLE can be computed using EM algorithm (see McLachlan & Peel, 2000). We examine these estimate through several criteria: the averaged spectral norm of the difference between the estimating inverse covariance matrix and the truth

$$\mathrm{SL} = \frac{1}{M}\sum_{k=1}^{M} \|\hat{\Sigma}_k^{-1} - \Sigma_k^{-1}\|, \tag{4.5}$$

where $\|A\|$ is the largest singular value of matrix A; the averaged Frobenius norm of the difference

$$\begin{aligned} \mathrm{FL} &= \frac{1}{M}\sum_{k=1}^{M} \|\hat{\Sigma}_k^{-1} - \Sigma_k^{-1}\|_{\mathrm{F}} \\ &= \frac{1}{M}\sum_{k=1}^{M} \sqrt{\sum_{i,j}(\hat{\Sigma}_k^{-1}(i, j) - \Sigma_k^{-1}(i, j))^2}, \end{aligned} \tag{4.6}$$

and the average Kullback-Leibler (KL) loss,

$$\mathrm{KL} = \frac{1}{M}\sum_{k=1}^{M} \mathrm{KL}(\Sigma_k, \hat{\Sigma}_k), \tag{4.7}$$

where

$$\mathrm{KL}(\Sigma, \hat{\Sigma}) = \mathrm{tr}(\Sigma\hat{\Sigma}^{-1}) - \log|\Sigma\hat{\Sigma}^{-1}| - p. \tag{4.8}$$

The results, averaged over 100 runs for each case, are reported in Tables 1, 2, and 3 for the three models, respectively. It is clear from these results that the proposed method outperforms the other two methods for all three models. The superior performance becomes more evident when the dimension increases. We also note the similar behavior of the penalized likelihood estimates tuned with either CV or BIC. This observation is of

Table 1: Simulation Results for Model 1: Estimation Errors Measured by Different Metrics for Different Methods.

	Penalized Likelihood											
	CV			BIC			Mclust			MLE		
p	SL	FL	KL	SL	FL	KL	SL	FL	KL	SL	FL	KL
30	2.53	6.87	5.7	3.27	9.1	9.68	3.26	10.39	17.59	31.14	43.42	40.81
	(0.032)	(0.096)	(0.143)	(0.071)	(0.113)	(0.366)	(0.006)	(0.022)	(0.069)	(1.077)	(1.101)	(1.376)
50	2.81	10.59	11.44	3.14	11.56	15.7	3.38	13.94	29.25			
	(0.012)	(0.025)	(0.099)	(4.29)	(2.973)	(0.06)	(0.006)	(0.028)	(0.148)			
100	2.61	13.8	18.44	3.75	18.68	44.3	3.41	20	59.12			
	(0.006)	(0.018)	(0.093)	(4.479)	(3.165)	(0.052)	(0.002)	(0.012)	(0.008)			
300	2.7	24.27	70.42	3.52	33.13	149.63	3.42	34.86	179.38			
	(0.014)	(0.15)	(1.205)	(0.016)	(0.139)	(2.874)	(0.001)	(0.015)	(0.015)			

Notes: Averaged over 100 runs. The numbers in parentheses are the standard errors. Empty cells = not applicable.

Table 2: Simulation Results for Model 2: Estimation Errors Measured by Different Metrics for Different Methods.

	Penalized Likelihood											
	CV			BIC			Mclust			MLE		
p	SL	FL	KL	SL	FL	KL	SL	FL	KL	SL	FL	KL
30	0.53	0.84	1.15	0.53	0.84	1.15	1.05	1.29	1.64	10.72	13.72	52.42
	(0.021)	(0.019)	(0.03)	(0.022)	(0.02)	(0.03)	(0.001)	(0.001)	(0.003)	(0.744)	(0.728)	(1.897)
50	0.48	0.83	1.62	0.48	0.83	1.6	1.11	1.42	2.44			
	(0.018)	(0.016)	(0.034)	(0.046)	(0.234)	(0.018)	(0.001)	(0.001)	(0.004)			
100	0.39	0.82	2.57	0.39	0.82	2.57	1.16	1.59	4.01			
	(0.011)	(0.009)	(0.034)	(0.053)	(0.371)	(0.011)	(0)	(0.001)	(0.005)			
300	0.39	1.02	7.06	0.39	1.02	7.06	1.23	1.83	8.45			
	(0.013)	(0.009)	(0.05)	(0.013)	(0.009)	(0.05)	(0)	(0)	(0.006)			

Notes: Averaged over 100 runs. The numbers in parentheses are the standard errors. Empty cells = not applicable.

Table 3: Simulation Results for Model 3: Estimation Errors Measured by Different Metrics for Different Methods.

	Penalized Likelihood											
	CV			BIC			Mclust			MLE		
p	SL	FL	KL	SL	FL	KL	SL	FL	KL	SL	FL	KL
30	1.01	2.6	1.61	1	2.61	1.63	1.21	4.06	5.59	22.68	31.71	40.26
	(0.019)	(0.017)	(0.016)	(0.019)	(0.017)	(0.016)	(0.048)	(0.195)	(0.439)	(0.923)	(0.929)	(1.301)
50	1.09	3.39	2.73	1.09	3.38	2.73	1.2	5.07	7.47			
	(0.02)	(0.015)	(0.02)	(0.02)	(0.015)	(0.021)	(0.03)	(0.157)	(0.476)			
100	1.15	4.78	5.49	1.15	4.79	5.51	1.18	7	12.53			
	(0.016)	(0.015)	(0.028)	(0.016)	(0.015)	(0.028)	(0.002)	(0.017)	(0.074)			
300	1.38	8.34	16.61	1.38	8.34	16.63	1.18	12.17	37.86			
	(0.018)	(0.017)	(0.048)	(0.018)	(0.017)	(0.048)	(0.002)	(0.027)	(0.204)			

Notes: Averaged over 100 runs. The numbers in parentheses are the standard errors. Empty cells = not applicable.

great practical importance because BIC is much more efficient to compute than the CV. For this reason, we shall use BIC as the tuning criterion in the rest of the letter unless otherwise indicated.

To investigate the effect of the sample size, we repeat the experiment with sample sizes $n = 200$ and 400 and dimension fixed at $p = 100$. The estimation errors, again averaged over 100 runs, are reported in Table 4. It is clear from Table 4 that the increasing sample sizes leads to improved estimation for the proposed method under all metrics.

To further demonstrate the merits of the proposed method, we report in Table 5 the averaged percentage of zero off-diagonal entries of inverse covariance matrices and in Table 6 computing times for `Mclust` and the proposed method. The results from Table 5 show that the proposed method's superior performance may be attributed to its ability to exploit sparsity in the inverse covariance matrices. The EM algorithm generally converges very quickly, and we set the number of iterations to be limited by 20 throughout the simulations. We also used `Mclust` to generate initial estimates for the algorithm. The reported computing time for the proposed estimate does not include that for `Mclust` for comparison purposes. It is evident from Table 6 that the proposed method can be efficiently computed.

We now consider a more complicated setting where the number of clusters also needs to be selected. We consider the true number of clusters, M, to be either 2 or 3. The sample size n is fixed at 100, whereas the dimension p is set to be 50. When $M = 2$, we used model 3 as our data-generating mechanism. When $M = 3$, the last cluster has the same covariance matrix as the first cluster in model 2. The experiment was repeated 100 times for each value of M. The results are summarized in Table 7.

To gain further insight, Figure 1 shows the smallest BIC scores for each value of the number of clusters for one typical simulated data set with $M = 2$ and $M = 3$, respectively.

5 Applications

The proposed method for estimating high-dimensional gaussian mixture models could be useful for a variety of applications. For illustration purposes, we consider here model-based clustering and the mixture discriminant analysis.

5.1 Model-Based Clustering. Gaussian mixture models have been one of the more popular tools for clustering (Fraley & Raftery, 2002). They provide a principled statistical approach to the practical questions that arise in clustering. To demonstrate the potential of our method in clustering high-dimensional inputs, we apply it to handwritten digit data (LeCun et al., 1990). The data set, collected by the U.S. Postal Service, consists of scanned digits from handwritten ZIP codes on envelopes. Every handwritten digit image has been digitalized to a 16 × 16 image with the intensity value

Table 4: Effect of Sample Sizes: Estimation Errors for Models 1 to 3, with $p = 100$ and $n = 200$ and 400.

		Penalized Likelihood											
		CV			BIC			Mclust			MLE		
Model	n	SL	FL	KL	SL	FL	KL	SL	FL	KL	SL	FL	KL
1	200	2.02 (0.005)	10.04 (0.019)	9.75 (0.043)	2.77 (0.007)	15.23 (0.029)	20.49 (0.139)	3.42 (0.002)	20.06 (0.01)	59.08 (0.004)			
	400	1.96 (0.005)	9.97 (0.017)	6.38 (0.028)	1.96 (0.005)	9.97 (0.017)	6.38 (0.028)	3.43 (0.001)	20.09 (0.009)	59.07 (0.004)	17.29 (0.134)	42.37 (0.19)	61.3 (0.216)
2	200	0.25 (0.01)	0.55 (0.008)	1.2 (0.015)	0.25 (0.01)	0.55 (0.008)	1.19 (0.015)	1.16 (0.002)	1.61 (0.006)	4.54 (0.167)			
	400	0.17 (0.005)	0.36 (0.004)	0.56 (0.006)	0.17 (0.005)	0.36 (0.004)	0.56 (0.006)	1.17 (0)	1.59 (0)	3.97 (0.001)	3.46 (0.033)	7.42 (0.03)	74.21 (0.277)
3	200	0.88 (0.005)	4.15 (0.007)	4.02 (0.02)	2.77 (0.007)	15.23 (0.029)	20.49 (0.139)	1.19 (0.002)	7.09 (0.012)	12.92 (0.054)			
	400	0.79 (0.002)	3.65 (0.009)	2.86 (0.023)	1.96 (0.005)	9.97 (0.017)	6.38 (0.028)	1.2 (0.002)	7.12 (0.009)	13.07 (0.04)	12.56 (0.09)	31.04 (0.122)	64.5 (0.22)

Notes: Averaged over 100 runs. The numbers in parentheses are the standard errors. Empty cells = not applicable.

Table 5: Percentage of Zero Off-Diagonal Entries in the Estimated Inverse Covariance Matrices.

	Model 1		Model 2		Model 3	
n	CV	BIC	CV	BIC	CV	BIC
100	88% (0.12%)	94% (0.20%)	99% (0.01%)	99% (0.01%)	99% (0.02%)	88% (0.10%)
200	80% (0.11%)	90% (0.08%)	99% (0.02%)	99% (0.01%)	98% (0.13%)	99% (0.01%)
400	87% (0.09%)	87% (0.09%)	99% (0.01%)	99% (0.01%)	94% (0.32%)	99% (0.01%)

Notes: Averaged over 100 runs. Numbers in parentheses represent the standard errors.

Table 6: Computing time in Seconds for Mclust and the Proposed Method (PLE).

n	Method	Model 1	Model 2	Model 3
100	Mclust	0.235	0.188	0.186
	PLE	0.812	0.503	0.114
200	Mclust	0.824	0.845	0.828
	PLE	0.595	1.47	0.208
400	Mclust	3.667	3.635	3.655
	PLE	5.129	2.815	0.425

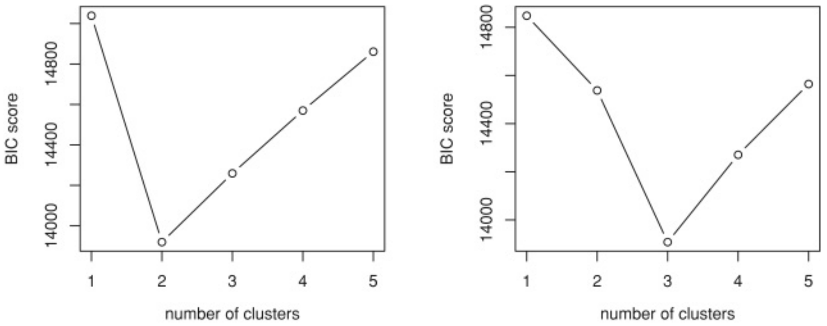
Notes: Averaged over 100 runs. Numbers in parentheses represent the standard errors.

Table 7: Frequency That Mclust and the Proposed Method (PLE) Identify the Correct Number of Clusters.

Method	$M = 2$	$M = 3$
Mclust	47%	56%
PLE	100%	100%

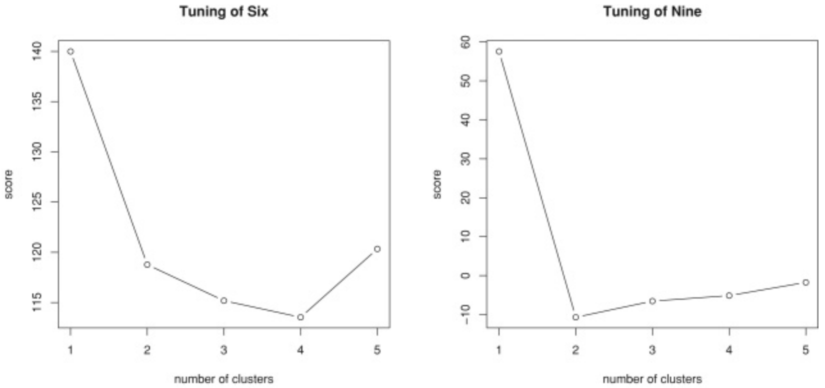
Note: Averaged over 100 runs.

of each pixel normalized to range from −1 and 1. We focus here on the digits 6 and 9. There are 834 images of digit 6 and 821 of digit 9. For each digit, we fix a gaussian mixture model using the proposed method with both the number of clusters and the tuning parameter λ jointly chosen by minimizing the BIC score. The minimal BIC score associated with each value of the number of clusters is given in Figure 2, which suggests that there are four clusters for digit 6 but only two for digit 9. The typical examples from each cluster in Figure 3 show that the clustering based on our method is indeed meaningful.



(a) M= 2 (b) M= 3

Figure 1: BIC score versus number of clusters.



(a) BIC score for digit 6 (b) BIC score for digit 9

Figure 2: Number of clusters selected for the handwritten digit data.

5.2 Mixture Discriminant Analysis. We now turn to classification, where the mixture discriminant analysis (MDA) introduced by Hastie and Tibshirani (1996) provides a much more flexible alternative to linear or quadratic discriminant analysis. The idea here is to model each class distribution using a gaussian mixture model and then classify an instance according to Bayes' rule. Unlike the usual linear or quadratic discriminant analysis, MDA is able to produce more general nonlinear classification boundaries. The main difficulty of using MDA in classification with high-dimensional inputs remains how to fit high-dimensional gaussian mixture models where our method could be a valuable tool. To demonstrate the merit of such practice, we apply this strategy to the handwritten digit data. We again focus on the digits 6 and 9 and investigate automatic classification

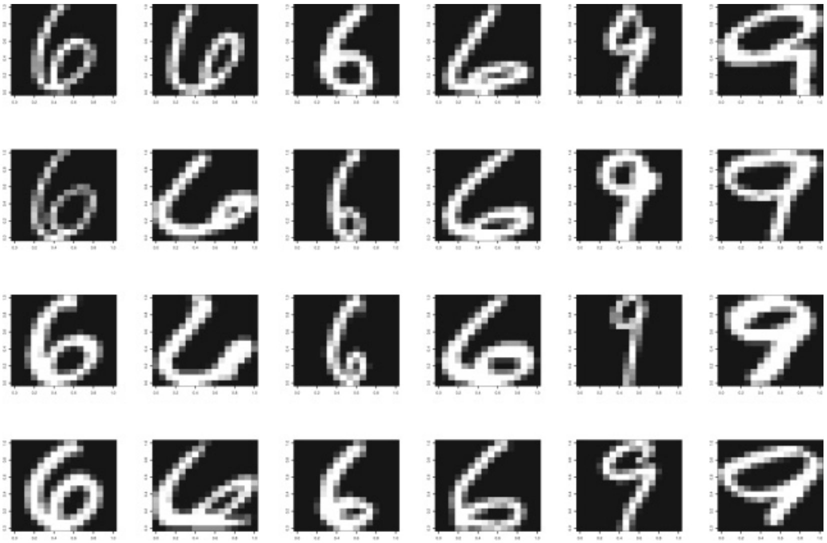


Figure 3: Clustering of the digits 6 and 9. Images from each column are randomly chosen from a particular cluster: the first four columns correspond to the four selected clusters of digit 6, and the last two correspond to digit 9.

between these two classes. For evaluation purposes, we randomly select 80% of the combined images as the training set and use the remaining 20% as the testing set. Gaussian mixture models were fit with tuning parameters determined by BIC for the digits 6 and 9, respectively, on the training set, and the resulting classifier is applied to the testing data to obtain a test error. This procedure was repeated—splitting the data, fitting the mixture model, and evaluating the test error—100 times. With the proposed method, the average test error rate is 0.26% with a standard error 0.029%. Note that the direct MLE as employed in the original mixture discriminant analysis is rather unstable for this example due to its high dimensionality. Generalization with the proposals of Banfield and Raftery (1993) has been investigated by Fraley and Raftery (2002, 2007) and implemented in R. For comparison purpose, we ran a similar analysis using this method as well. It yields an error rate of 0.42% with a standard error of 0.03%.

6 Discussion

We have developed a penalized likelihood estimator for high-dimensional gaussian graphical models. By imposing an ℓ_1 penalty on the inverse covariance matrices, the proposed estimator encourages sparsity and therefore could be useful for high-dimensional cases. We show that the estimate can be efficiently computed by an EM algorithm. Simulation studies show that the method is quite promising in extending the scope of the gaussian mixture model in handling high-dimensional data. Its usefulness is further assessed in the context of model-based clustering and mixture discriminant analysis. These empirical successes warrant a more thorough theoretical study of the proposed method, which we leave for future research.

The proposed methods can be easily extended in several directions. We have used a single tuning parameter λ in defining our estimator 2.2. In many applications, the sparsity of precision matrices may vary a lot from cluster to cluster. In these cases, one may adopt different tuning parameters for different precision matrices, leading to the following extension:

$$\hat{\Theta} := \operatorname*{argmin}_{\mu_k, \Sigma_k \succ 0} \left\{ -\sum_{i=1}^{n} \log \left(\sum_{k=1}^{M} \pi_k \phi(X_i | \mu_k, \Sigma_k) \right) + \sum_{k=1}^{M} \lambda_k \| \Sigma_k^{-1} \|_{\ell_1} \right\}.$$

Moreover, other penalty functions such as hard thresholding or SCAD (Fan & Li, 2001) could also be used in place of the ℓ_1 type of penalty.

Acknowledgments

This research was supported in part by NSF grants DMS-0706724 and DMS-0846234, NIH grant R01 GM076274-01, and a grant from Georgia Cancer Coalition.

References

Banfield, J. D., & Raftery, A. E. (1993). Model-based gaussian and non-gaussian clustering. *Biometrics*, *49*, 803–821.

Banerjee, O., El Ghaoui, L., & d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine Learning Research*, *9*, 485–516.

Bickel, P. J., & Levina, E. (2008a) Regularized estimation of large covariance matrices, *Annals of Statistics*, *36*(1), 199–227.

Bickel, P., & Levina, E. (2008b). Covariance regularization by thresholding. *Annals of Statistics*, *36*, 2577–2604.

Celeux, G., & Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition*, *28*(5), 781–793.

d'Aspremont, A., Banerjee, O., & El Ghaoui, L. (2008). First-order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and Its Applications*, *30*(1), 56–66.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society Series B*, *39*, 1–38.

Deng, X., & Yuan, M. (2009). Large Gaussian covariance matrix estimation with Markov structures. *Journal of Computational and Graphical Statistics*, *18*, 640–657.

El Karoui, N. (2008). Operator norm consistent estimation of large dimensional sparse covariance matrices. *Annals of Statistics*, *36*, 2717–2756.

Fan, J., Fan, Y., & Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Economics*, *147*, 186–197.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*, 1348–1360.

Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association*, *97*, 611–631.

Fraley, C., & Raftery, A. E. (2007). Bayesian regularization for normal mixture estimation and model-based clustering. *Journal of Classification*, *24*, 155–181.

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, *9*(3), 432–441.

Hastie, T., & Tibshirani, R. (1996). Discriminant analysis by gaussian mixtures. *Journal of the Royal Statistical Society Ser. B*, *58*, 155–176.

Huang, J., Liu, N., Pourahmadi, M., & Liu, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, *93*(1), 85–98.

Lam, C., & Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrices estimation. *Annals of Statistics*, *37*, 4254–4278.

Lauritzen, S. L. (1996). *Graphical models*. Oxford: Clarendon Press.

LeCun,Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., et al. (1990). Handwritten digit recognition with a back propagation network. In D. Touretzky (Ed.), *Advances in neural information processing systems*, *2*. San Francisco: Morgan Kaufmann.

Ledoit, O., & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, *88*, 365–411.

Levina, E., Rothman, A., & Zhu, J. (2008). Sparse estimation of large covariance matrices via a nested LASSO penalty. *Annals of Applied Statistics*, *2*(1), 245–263.

Li, H., & Gui, J. (2006). Gradient directed regularization for sparse gaussian concentration graphs, with applications to inference of genetic networks. *Biostatistics*, *7*(2), 302–317.

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. Hoboken, NJ: Wiley.

Rothman, A. J., Bickel P. J., Levina, E., & Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronical Journal of Statistics*, *2*, 494–515.

Rothman, A. J., Levina, E., & Zhu, J. (2009). Generalized thresholding of large covariance matrices. *Journal of American Statistical Association*, *104*(485), 177–186.

Symons, M. J. (1981). Clustering criteria and multivariate normal mixtures. *Biometrics*, *37*(1), 35–43.

Whittaker, J. (1990). *Graphical models in applied multivariate statistics*. Hoboken, NJ: Wiley.

Yuan, M. (2008). Efficient computation of the ℓ_1 regularized solution path in gaussian graphical models. *Journal of Computational and Graphical Statistics*, *17*, 809–826.

Yuan, M., & Lin, Y. (2007). Model selection and estimation in the gaussian graphical model. *Biometrica*, *94*(1), 19–35.

Received May 15, 2010; accepted October 29, 2010.