Model Selection for Gaussian Mixture Models

Tao Huang, Heng Peng and Kun Zhang

Shanghai University of Finance and Economics,
Hong Kong Baptist University and
Carnegie Mellon University & Max Planck Institute for Intelligent Systems

S1 Example IV.

In this example, we generate a mixture of eight components with mixing probabilities $\pi_i = 0.125, i = 1, \cdots, 8$, mean vectors $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = [2, 2]^T$, $\boldsymbol{\mu}_3 = \boldsymbol{\mu}_4 = [-2, 2]^T$ $\boldsymbol{\mu}_5 = \boldsymbol{\mu}_6 = [2, -2]^T$, $\boldsymbol{\mu}_7 = \boldsymbol{\mu}_8 = [-2, -2]^T$, and covariance matrices

$$\boldsymbol{\Sigma}_i = \begin{bmatrix} 1.05 & 0.95 \\ 0.95 & 1.05 \end{bmatrix} \text{ for } i = 1, 3, 5, 7, \text{ and } \boldsymbol{\Sigma}_i = \begin{bmatrix} 1.05 & -0.95 \\ -0.95 & 1.05 \end{bmatrix} \text{ for } i = 2, 4, 6, 8.$$

In fact, all these eight components are obtained by rotating and shifting the Gaussian density $\mathcal{N}(0, \text{diag}(2, 0.1))$.

We consider three different sample sizes $n = 800, 1200, 1600$ and run our proposed methods for 300 times for each sample size. The maximum number of components M is set to be 20 or 50. The initial value for the modified EM algorithm is estimated by K-means clustering, and the tuning parameter λ is selected by our proposed BIC method. Figure 1 shows the evolution of the modified EM algorithm for (2.3) with the initial maximum number of components as 20 for one simulated data set. Table 1 shows that our proposed method can identify the number of components with high probability and performs much better than the AIC and BIC methods. Table 2 shows that the modified EM algorithm gives accurate estimates for both parameters and mixing probabilities.

S2 Real Data Analysis.

We apply our proposed method to an image segmentation data set at UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/
datasets/Image+Segmentation). This data set was created from a database of seven outdoor images (brickface, sky, foliage, cement, window, path, and grass). Each image was hand-segmented into instances of 3×3 regions, and 230 instances were drawn. For each instance, there are 19 attributes. We here only focus on four images, brickface, sky, foliage, and grass, and two attributes, extra red and extra green. Our objective is to estimate the joint probability density

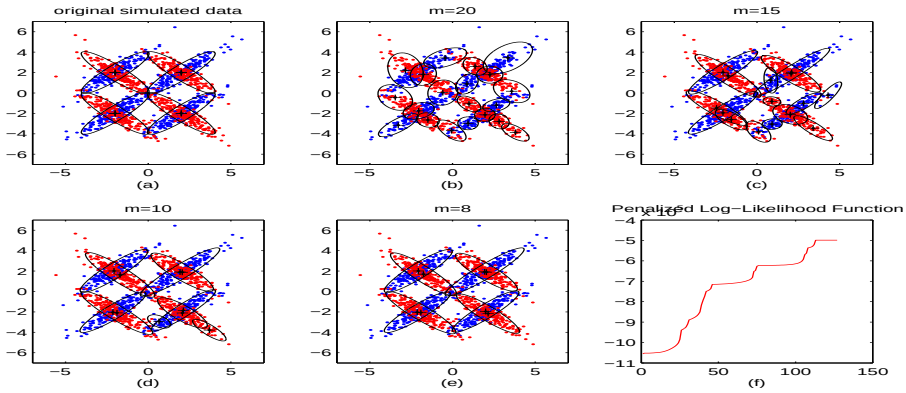

Figure 1: *One typical run of Example IV. (a) a simulated data set. (b) initialization for* $M = 20$ *components, (c-d) two intermediate estimates for* $M = 15, 10$*, respectively, (e) the final estimate for* $M = 8$*, (f) the penalized log likelihood function for this run.*

Table 1: Order selection for Example IV

		Penalized			AIC			BIC		
	SS	800	1200	1600	800	1200	1600	800	1200	1600
$M = 20$	$m = 8$	0.7733	0.9500	0.9733	0.1500	0.1267	0.1000	0.1500	0.1267	0.0933
	$\|m - 8\| = 1$	0.2200	0.0500	0.0233	0.1800	0.2100	0.1800	0.1800	0.2000	0.1767
	$\|m - 8\| \geq 2$	0.0067	0	0.0033	0.6700	0.6633	0.7200	0.6700	0.6733	0.7300
$M = 50$	$m = 8$	0.7333	0.9400	0.9300	0.1500	0.1267	0.1000	0.1500	0.1267	0.0933
	$\|m - 8\| = 1$	0.2633	0.0600	0.0667	0.1800	0.2100	0.1800	0.1800	0.2000	0.1767
	$\|m - 8\| \geq 2$	0.0033	0	0.0033	0.6700	0.6633	0.7200	0.6700	0.6733	0.7300

Table 2: Exp IV: Parameter estimation with standard deviation by maximizing (2.3)

		Mixing Probability	Mean		Covariance (eigenvalue)	
Component1	TRUE	0.125	2	2	0.1	2
$SS = 800$	$M = 20$	-.0010(.0151)	.0226(.1971)	.0223(.2016)	-.0040(.0196)	-.0135(.6148)
	$M = 50$	-.0003(.0149)	.0295(.1876)	.0267(.1930)	-.0040(.0199)	-.0477(.5675)
$SS = 1200$	$M = 20$	-.0003(.0088)	.0052(.1171)	.0067(.1176)	-.0018(.0162)	-.0007(.3863)
	$M = 50$	-.0003(.0088)	.0074(.1160)	.0088(.1156)	-.0013(.0160)	-.0134(.3691)
$SS = 1600$	$M = 20$	-.0007(.0072)	.0102(.0973)	.0102(.0954)	-.0027(.0141)	-.0094(.3183)
	$M = 50$	-.0005(.0073)	.0093(.0941)	.0084(.0927)	-.0027(.0142)	-.0078(.3205)
Component5	TRUE	0.125	2	2	0.1	2
$SS = 800$	$M = 20$	.0007(.0118)	-.0055(.1312)	.0009(.1279)	-.0026(.0218)	-.0429(.3535)
	$M = 50$	-.0002(.0125)	-.0081(.1351)	-.0008(.1356)	-.0026(.0235)	-.0443(.3644)
$SS = 1200$	$M = 20$	-.0008(.0078)	.0042(.0994)	.0032(.0994)	-.0026(.0154)	-.0193(.3082)
	$M = 50$	-.0008(.0080)	.0046(.1001)	.0034(.1002)	-.0027(.0154)	-.0231(.3091)
$SS = 1600$	$M = 20$	.0001(.0068)	-.0061(.0878)	.0007(.0871)	-.0016(.0146)	-.0003(.2550)
	$M = 50$	.0001(.0068)	-.0042(.0864)	.0021(.0852)	-.0015(.0147)	.0022(.2557)

Similar results for other components.

function of the two attributes (See Figure 2(a)) using a Gaussian mixture model with arbitrary covariance matrices. In other words, we implement our proposed method to identify the number of components, and to simultaneously estimate the unknown parameters of bivariate normal distributions and the mixing probabilities. Although we consider only four images, Figure 2(a) suggests that a five-component Gaussian mixture is more appropriate and the brickface image is better represented by two components.

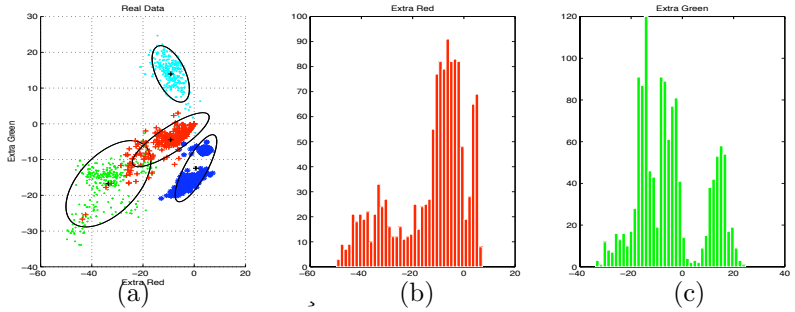

Figure 2: *(a) Scatter plot of scaled real data. Brickface (blue), Sky (green), Foliage (red), Grass (light blue); (b-c) Histograms of marginal density. (b) Extra red, and (c) Extra green.*

As in the simulation studies, we run our proposed method for 300 times. For each run, we randomly draw 200 instances. The maximum number of components is set to be 10, and the initial value of the modified EM algorithm is estimated by the K-means clustering. Because

there is little difference between numerical results of the two proposed methods (2.3) and (2.4), here we only show the numerical results obtained by maximizing (2.3). Figure 3 shows the evolution of the modified EM algorithm for one run. Figure 4 shows that our proposed method selects five components with high probability. For a five-component Gaussian mixture model, we summarize the estimation of parameters and mixing probabilities in Table 4.

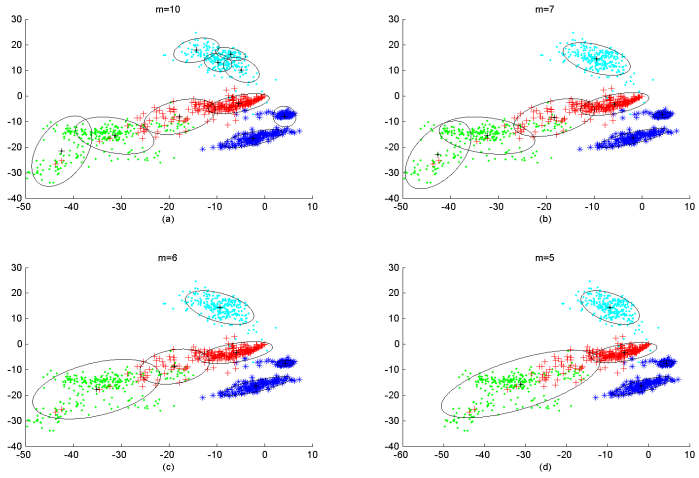

Figure 3: *One typical run. (a) Initialization with $M = 10$ components, (b) and (c) two intermediate estimates for $M = 7$ and $M = 6$, respectively, (d) the final estimate $M = 5$.*

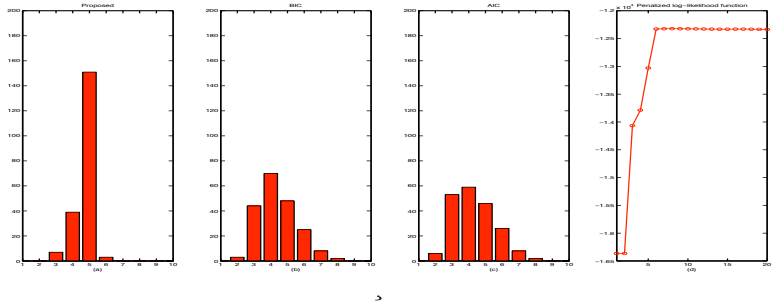

Figure 4: *Histogram of estimated numbers of components. (a) the proposed method (2.3), (b) BIC, (c) AIC. (d) The penalized log likelihood function for one typical run.*

Table 3: Parameter Estimation with standard deviation, $\hat{M} = 5$

Component	Underlying	Mixing Probability	Mean (Ex-red, Ex-green)		Standard Deviation (Ex-red, Ex-green)	
1	Sky & Foliage	.4153	-27.7689	-13.7343	12.0464	7.3739
		(0.0555)	(2.3336)	(1.1377)	(1.1726)	(0.3745)
2	Grass	.2617	-9.3724	13.4992	3.6393	3.5632
		(0.0523)	(0.8588)	(4.0740)	(1.2160)	(1.3567)
3	Foliage & Brickface	.1447	-4.9511	-3.3625	3.1584	1.6728
		(0.0425)	(1.3913)	(3.5537)	(0.7102)	(0.4004)
4	Brickface	.0824	-1.3474	-12.0906	2.5780	1.3302
		(0.0487)	(0.8417)	(7.2158)	(1.5227)	(0.7854)
5	Brickface	.0936	3.5476	-8.4996	1.5110	1.3293
		(0.0717)	(1.4703)	(2.1980)	(1.2288)	(1.6460)

S3 Proof of Theorem 2 and 3

Proof of Theorem 2: To prove Theorem 2, similar as the proof of Theorem 1, we first show that there exists a maximizer $(\theta, \boldsymbol{\beta})$ such that $\theta = O_p(1/\sqrt{n})$ when $\lambda = C/\sqrt{n}$. It is sufficient to show that, for a large constant C_1, $\ell(\theta, \boldsymbol{\beta}) < \ell(0, \boldsymbol{\beta})$ where $\theta = C_1/\sqrt{n}$ and $\boldsymbol{\beta}$ is in a local compact area $\Omega = \{\boldsymbol{\beta} : (\alpha_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \in \bigcup_{l=1}^{q} \Omega_l, i = 1, \ldots, M-q\}$ where Ω_l is defined in the proof of Proposition 1. Let $\theta = C_1/\sqrt{n}$, we then have

$$\begin{aligned} \ell_p(\theta, \boldsymbol{\beta}) - \ell_p(0, \boldsymbol{\beta}) &\leq \sum_{i=1}^{n} \{\log f(\mathbf{x}_i, \theta, \boldsymbol{\beta}) - \log g_0(\mathbf{x}_i)\} \\ &\quad -n\lambda D_f \sum_{m=M-q+1}^{M} [\log(\epsilon + \pi_m) - \log(\epsilon + \pi^0_{m-M+q})] \\ &\hat{=} I_1 + I_2. \end{aligned}$$

For I_2, because of $\theta = C_1/\sqrt{n}$ and by the restriction condition on $\rho_l, l = 1, \ldots, q$, we have $|\pi_m - \pi^0_{m-M+q}| \leq C_1/\sqrt{n}$ when $m > M - q$. By the property of the penalty function, we then have

$$\begin{aligned} |I_2| &= |-n\lambda D_f \sum_{m=M-q+1}^{M} [\log(\epsilon + \pi_m) - \log(\epsilon + \pi^0_{m-M+q})]| \\ &= \left| -n\lambda D_f \sum_{m=M-q+1}^{M} \left[\frac{(\pi_m - \pi^0_{m-M+q})}{\epsilon + \pi^0_{m-M+q}} \cdot (1 + o(1)) \right] \right| \\ &= O(\sqrt{n}) \cdot \frac{qC_1}{\sqrt{n}}(1 + o(1)) = O(C_1). \end{aligned}$$

For I_1, similar as the proof of Theorem 1, we have

$$I_1 = \frac{C_1}{\sqrt{n}} \cdot O_P(\sqrt{n}) - \frac{C_1^2}{n} \cdot O_p(n).$$

When C_1 is large enough, the second term of I_1 dominates I_2 and the other terms in the penalized likelihood ratio function. Then we have

$$\ell_p(\theta, \boldsymbol{\beta}) - \ell_p(0, \boldsymbol{\beta}) < 0$$

with probability tending to one. Hence there exists a maximizer $(\theta, \boldsymbol{\beta})$ with probability tending to one such that

$$\theta = O_p(\frac{1}{\sqrt{n}}).$$

Next we show that there exists a maximizer $(\hat{\theta}, \hat{\boldsymbol{\beta}})$ satisfying $\hat{\theta} = O_p(\frac{1}{\sqrt{n}})$ such that $\hat{q} = q$ or $\hat{\pi}_m = 0, m = 1, \ldots, M - q$.

First, we show that for any maximizer of $\ell_p(\theta^*, \boldsymbol{\beta}^*)$ with $|\theta^*| \leq C_1/\sqrt{n}$, if there is $k \leq M-q$ such that $C_1/\sqrt{n} \geq \pi_k^* > 1/(\sqrt{n}\log n)$, then there should exist another maximizer of $\ell_p(\theta, \boldsymbol{\beta})$ in the area of $|\theta| \leq C_1/\sqrt{n}$. It means that the extreme maximizer of $\ell_p(\theta, \boldsymbol{\beta})$ in the compact area of $|\theta| \leq C_1/\sqrt{n}$ should satisfy that $\pi_k < \frac{1}{\sqrt{n}\log n}$ for any $k < M - q + 1$. Hence it is equivalent to show that for any such kind maximizer of $\ell_p(\theta^*, \boldsymbol{\beta}^*)$ with $|\theta^*| \leq C_1/\sqrt{n}$, we always have $\ell_p(\theta^*, \boldsymbol{\beta}^*) < \ell_p(0, \beta^*)$ with probability tending to one. Similar to the analysis above, we have

$$\begin{aligned}
\ell_p(\theta^*, \boldsymbol{\beta}^*) - \ell_p(0, \boldsymbol{\beta}^*) \quad &\leq \quad \sum_{i=1}^{n} \{\log f(\mathbf{x}_i, \theta^*, \boldsymbol{\beta}^*) - \log g_0(\mathbf{x}_i)\} \\
&\quad -n\lambda D_f \sum_{m=M-q+1}^{M} [\log(\epsilon + \pi_m^*) - \log(\epsilon + \pi_{m-M+q}^0)] \\
&\quad -n\lambda D_f \sum_{k=1}^{M-q} \log \frac{\epsilon + \pi_k^*}{\epsilon} \\
&\hat{=} \quad I_1 + I_2 + I_3.
\end{aligned}$$

As shown before, we have $I_1 + I_2 = O_p(C_1^2)$. For I_3, because of $\epsilon = o(\frac{1}{\sqrt{n}\log n})$ we have

$$|I_3| = O(n \cdot C/\sqrt{n}) \cdot \log \frac{\epsilon + \pi_k^*}{\epsilon} = O(\sqrt{n}).$$

Notice that I_3 is always negative and dominates the terms in I_1 and I_2, and therefore we have $\ell_p(\theta^*, \boldsymbol{\beta}^*) < \ell_p(0, \beta^*)$.

In the following step, we need only consider the maximizer of $\ell_p(\hat{\theta}, \hat{\boldsymbol{\beta}})$ with $|\hat{\theta}| \leq C_1/\sqrt{n}$ and $\hat{\pi}_k < 1/(\sqrt{n}\log n)$ for $k < M - q + 1$.

A Lagrange multiplier β is taken into account for the constraint $\sum_{m=1}^{M} \hat{\pi}_m = 1$. It is then sufficient to show that

$$\frac{\partial \ell^*(\boldsymbol{\theta})}{\partial \hat{\pi}_m} < 0 \quad \text{for} \quad \hat{\pi}_m < \frac{1}{\sqrt{n}\log n} \tag{S3.1}$$

with probability tending to one for the maximizer $(\theta, \boldsymbol{\beta})$ where $\ell^*(\boldsymbol{\theta}) = \ell_p(\boldsymbol{\theta}) - \beta(\sum_{m=1}^{M} \pi_m - 1)$. To show the equation above, we consider the partial derivatives for $\hat{\pi}_m,\ m > M - q$, and have the following equations,

$$\frac{\partial \ell^*(\boldsymbol{\theta})}{\partial \hat{\pi}_m} = \sum_{i=1}^{n} \frac{\phi_m(\mu_m, \boldsymbol{\Sigma}_m)}{\sum_{i=1}^{M} \hat{\pi}_i \phi_i(\mu_i, \boldsymbol{\Sigma}_i)} - n\lambda D_f \frac{1}{\epsilon + \hat{\pi}_m} - \beta = 0. \tag{S3.2}$$

It is obvious that the first term in the equation above is of order $O_p(n)$ by the law of large numbers. If $m > M - q$ and $\theta = O_p(\frac{1}{\sqrt{n}})$, it is easy to know that $\hat{\pi}_m = \pi^0_{m-M+q} + O_p(1/\sqrt{n}) > \frac{1}{2} \cdot \min(\pi^0_1, \ldots, \pi^0_q)$, and then the second term should be $O_p(n\lambda) = o_p(n)$, and moreover $\beta = O_p(n)$.

Next, consider

$$\frac{\partial \ell^*(\boldsymbol{\theta})}{\partial \hat{\pi}_m} = \sum_{i=1}^{n} \frac{\phi_m(\mu_m, \boldsymbol{\Sigma}_m)}{\sum_{i=1}^{M} \hat{\pi}_i \phi_i(\mu_i, \boldsymbol{\Sigma}_i)} - n\lambda D_f \frac{1}{\epsilon + \hat{\pi}_m} - \beta, \tag{S3.3}$$

where $m \leq M - q$ and $\hat{\pi}_m < \frac{1}{\sqrt{n} \log n}$. It is obvious that the first term and the third term in (S3.2) are of order $O_p(n)$. For the second term, because $\hat{\pi}_m = O_p(\frac{1}{\sqrt{n} \log n})$, $\lambda = C/\sqrt{n}$ and $\epsilon = o(\frac{1}{\sqrt{n} \log n})$, we have

$$\left\{ n\lambda D_f \frac{1}{(\epsilon + \pi_m)} \right\} \Big/ n = \lambda D_f \frac{1}{\epsilon + \pi_m} = O_p(\lambda \cdot \sqrt{n} \log n) \to \infty,$$

with probability tending to one. Hence the second term in (S3.3) dominates the first and the third terms. Therefore we prove the equation (S3.1), or equivalently $\hat{\pi}_m = 0, m = 1, \ldots, M - q$ with probability tending to one when $n \to \infty$. □

Proof of Theorem 3: To prove this theorem, we follow similar steps as in the proof of Theorem 2 in Wang, Li and Tsai (2007).

First, given $\lambda^* = \sqrt{\frac{\log n}{n}}$, by Theorem 1, we know there exists a maximizer such that $\hat{q} = q$ with probability tending to one and that $\hat{\pi}_{M-q+m}, m = 1, \ldots q$ are the consistent estimates of $\pi^0_i, i = 1, \ldots, q$. Hence with probability tending to one, we have

$$\ell_P(\hat{\boldsymbol{\theta}}_{\lambda^*}) = \ell(\hat{\boldsymbol{\theta}}_{\lambda^*}) - n\lambda^* D_f \cdot q \cdot \log \frac{\epsilon + a\lambda^*}{\epsilon},$$

where $\hat{\boldsymbol{\theta}}_{\lambda^*}$ is the estimator of parameters of the multivariate Gaussian mixture model. On the other hand, when q is known, we know that the maximum likelihood estimate $\hat{\boldsymbol{\theta}}_{MLE}$ is consistent. Hence we have

$$\begin{aligned} \ell_P(\hat{\boldsymbol{\theta}}_{MLE}) &= \ell(\hat{\boldsymbol{\theta}}_{MLE}) - n\lambda^* D_f \sum_{m=1}^{q} [\log(\epsilon + p^*_\lambda(\hat{\pi}_{m,MLE})) - \log(\epsilon)] \\ &= \ell(\hat{\boldsymbol{\theta}}_{MLE}) - n\lambda^* D_f \cdot q \cdot \log \frac{\epsilon + a\lambda^*}{\epsilon} \geq \ell_P(\hat{\boldsymbol{\theta}}_{\lambda^*}), \end{aligned}$$

where $\hat{\pi}_{m,MLE}$ is the maximum likelihood estimate of $\pi^0_m, m = 1, \ldots, q$. Then by the definition of $\ell_P(\hat{\boldsymbol{\theta}}_{\lambda^*})$, when $\lambda^* = \sqrt{\frac{\log n}{n}}$, we have the oracle property, i.e., the penalized likelihood estimate $\hat{\boldsymbol{\theta}}_{\lambda^*}$ is equal to $\hat{\boldsymbol{\theta}}_{MLE}$ with probability tending to one.

Next we identify two different cases, that is, the under-fitted model and the over-fitted model.

Case 1: Under-fitted model, i.e., $\hat{q}_\lambda < q$. By the definition of BIC, we have

$$BIC_\lambda = \ell(\hat{\boldsymbol{\theta}}_\lambda) - \frac{1}{2} \hat{q}_\lambda D_f \log n \leq \ell(\hat{\boldsymbol{\theta}}_{\hat{q}_\lambda, MLE}) - \frac{1}{2} \hat{q}_\lambda D_f \log n,$$

where $\hat{\boldsymbol{\theta}}_{\hat{q},MLE}$ is the maximum likelihood estimate of the finite Gaussian mixture model when the number of the components is $\hat{q}_\lambda$. By the law of large numbers and some cumbersome calculations, we can show that

$$\frac{1}{n}\left\{\ell(\hat{\boldsymbol{\theta}}_{\hat{q}_\lambda,MLE}) - \ell(\hat{\boldsymbol{\theta}}_{q,MLE})\right\} = \frac{1}{n}\left\{\ell(\hat{\boldsymbol{\theta}}_{\hat{q}_\lambda,MLE}) - \ell(\hat{\boldsymbol{\theta}}_{\lambda^*})\right\} \to \inf_{g_\lambda \in \mathcal{G}_{\hat{q}_\lambda}} -K(g_0, g_\lambda) = -K(g_0, \mathcal{G}_{\hat{q}_\lambda})$$

where $-K(g_0, g_\lambda)$ is the Kullback distance between g_0, and g_λ and $\mathcal{G}_{\hat{q}_\lambda}$ is the finite Gaussian mixture model space with $\hat{q}$ mixture components. Then we have

$$\begin{aligned} BIC_\lambda - BIC_{\lambda^*} &\leq \ell(\hat{\boldsymbol{\theta}}_{\hat{q}_\lambda,MLE}) - \ell(\hat{\boldsymbol{\theta}}_{\hat{q}_{\lambda^*},MLE}) - \frac{1}{2}\hat{q}_\lambda D_f \log n + \frac{1}{2}\hat{q}_{\lambda^*} D_f \log n \\ &= \ell(\hat{\boldsymbol{\theta}}_{\hat{q}_\lambda,MLE}) - \ell(\hat{\boldsymbol{\theta}}_{q,MLE}) - \frac{1}{2}\hat{q}_\lambda D_f \log n + \frac{1}{2} q D_f \log n \\ &= -nK(g_0, \mathcal{G}_{\hat{q}_\lambda})(1 + o_p(1)) + \frac{1}{2}(q - \hat{q}_\lambda) D_f \log n \\ &< 0, \end{aligned}$$

This implies that

$$\Pr(\sup_{\lambda:\hat{q}_\lambda < q} BIC_\lambda > BIC_{\lambda^*}) \to 0. \tag{S3.4}$$

Case 2: Over-fitted model, i.e., $\hat{q}_\lambda > q$. If we can show that

$$\ell(\hat{\boldsymbol{\theta}}_{\hat{q}_\lambda,MLE}) - \ell(\hat{\boldsymbol{\theta}}_{q,MLE}) = O_p(1), \tag{S3.5}$$

then we have

$$\begin{aligned} BIC_\lambda - BIC_{\lambda^*} &\leq \ell(\hat{\boldsymbol{\theta}}_{\hat{q}_\lambda,MLE}) - \ell(\hat{\boldsymbol{\theta}}_{\hat{q}_{\lambda^*},MLE}) - \frac{1}{2}\hat{q}_\lambda D_f \log n + \frac{1}{2}\hat{q}_{\lambda^*} D_f \log n \\ &= \ell(\hat{\boldsymbol{\theta}}_{\hat{q}_\lambda,MLE}) - \ell(\hat{\boldsymbol{\theta}}_{q,MLE}) - \frac{1}{2}\hat{q}_\lambda D_f \log n + \frac{1}{2} q D_f \log n \\ &= O_p(1) + \frac{1}{2}(q - \hat{q}_\lambda) D_f \log n \\ &< 0, \end{aligned}$$

and this implies that

$$\Pr(\sup_{\lambda:\hat{q}_\lambda > q} BIC_\lambda > BIC_{\lambda^*}) \to 0. \tag{S3.6}$$

Therefore Theorem 3 follows (S3.4) and (S3.6).

To prove (S3.5), note that given the bounded $\hat{q}$, where $\hat{q} > q$, by conditions P1 and P2, and similar to the proof of Proposition 1, the class of functions $\log f(x, \theta, \boldsymbol{\beta})$ is P-Glivenko-Cantelli and P-Donsker. Hence we have

$$\sup_{(\theta,\boldsymbol{\beta})\in\Theta} \left|\frac{1}{n}\ell(\boldsymbol{\theta}) - \mathrm{E}\log f(x, \theta, \boldsymbol{\beta})\right| = \sup_{(\theta,\boldsymbol{\beta})\in\Theta} \left|\frac{1}{n}\sum_{i=1}^{n} \log f(\mathbf{x}_i, \theta, \boldsymbol{\beta}) - \mathrm{E}\log f(x, \theta, \boldsymbol{\beta})\right| \xrightarrow{P} 0,$$

where Θ is a compact parameter space which satisfies conditions P1 and P2 and contained in $\bigcup_{l=1}^{q} \Omega_l$ where Ω_l is defined in the proof of Proposition 1. Similar as the proof of Theorem 1, we know there is a maximizer in Θ.

First, notice for any $\boldsymbol{\beta}$ and by the properties of the likelihood function, we have

$$\mathrm{E}\log f(x,\theta,\boldsymbol{\beta}) < \mathrm{E}\log f(x,0,\boldsymbol{\beta}).$$

when $\theta > \varepsilon$ and ε is a small constant. Similar to the proof of Theorem 5.7 in van der Vaart (1998), we can then show that the maximum likelihood estimate of θ converges to 0 in probability.

Next, similar as the proof of Theorem 1 for I_1, we have

$$\mathrm{E}\{\log f(\mathbf{x},\theta,\boldsymbol{\beta}) - \log f(\mathbf{x},0,\boldsymbol{\beta})\} = -\frac{1}{2}\theta^2\mathrm{E}\left(\frac{f'(\mathbf{x},0,\boldsymbol{\beta})}{g_0(\mathbf{x})}\right)^2(1+o_p(1)).$$

According to conditions P1 and P2, and Proposition 1, $\boldsymbol{\beta}$ is in a compact space and $\left(\frac{f'(\mathbf{x},0,\boldsymbol{\beta})}{g_0(\mathbf{x})}\right)^2$ is a continuous function of $\boldsymbol{\beta}$; hence $\mathrm{E}\left(\frac{f'(\mathbf{x},0,\boldsymbol{\beta})}{g_0(\mathbf{x})}\right)^2$ can be bounded by a constant, and moreover, for a sufficiently small $\theta < \delta \to 0$,

$$\mathrm{E}\{\log f(\mathbf{x},\theta,\boldsymbol{\beta}) - \log f(\mathbf{x},0,\boldsymbol{\beta})\} \le -C\theta^2.$$

Let $G_n(\theta,\boldsymbol{\beta}) = 1/\sqrt{n}\sum_{i=1}^{n}\{\log f(\mathbf{x}_i,\theta,\boldsymbol{\beta}) - \mathrm{E}\log f(\mathbf{x}_i,\theta,\boldsymbol{\beta})\}$. Consider $G_n(\theta,\boldsymbol{\beta}) - G_n(0,\boldsymbol{\beta})$; for $|\theta| < \delta$ where δ is a sufficient small value, following the proof of Theorem 1, and by conditions P1 and P2, we have

$$\begin{aligned}
G_n(\theta,\boldsymbol{\beta}) - G_n(0,\boldsymbol{\beta}) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\log f(x_i,\theta,\boldsymbol{\beta}) - \log f(x_i,0,\boldsymbol{\beta})\} \\
&\quad -\sqrt{n}\mathrm{E}\{\log f(\mathbf{x},\theta,\boldsymbol{\beta}) - \log f(\mathbf{x},0,\boldsymbol{\beta})\} \\
&= \frac{1}{\sqrt{n}}\left\{\sum_{i=1}^{n}\theta\frac{f'(\mathbf{x}_i,0,\boldsymbol{\beta})}{g_0(\mathbf{x}_i)} - \frac{1}{2}\sum_{i=1}^{n}\theta^2\left(\frac{f'(\mathbf{x}_i,0,\boldsymbol{\beta})}{g_0(\mathbf{x}_i)}\right)^2\right\}(1+o_p(1)) \\
&\quad +\frac{\sqrt{n}}{2}\theta^2\mathrm{E}\left(\frac{f'(\mathbf{x},0,\boldsymbol{\beta})}{g_0(\mathbf{x})}\right)^2(1+o_p(1)).
\end{aligned}$$

By Proposition 1 and the properties of the Donsker class, for $\delta = C/\sqrt{n}$ with a large enough constant C, we have

$$\mathrm{E}\sup_{\theta<\delta,\boldsymbol{\beta}}|G_n(\theta,\boldsymbol{\beta}) - G_n(0,\boldsymbol{\beta})| < C_1\delta + \sqrt{n}C_2\delta^2 + \sqrt{n}C_3\delta^2 = O(\sqrt{n}\delta^2).$$

As shown before, the maximum likelihood estimate of θ is consistent under the constraint conditions P1 and P2 and Proposition 1. Then following the proof of Theorem 5.55 of van der Vaart (1998), we have

$$\hat{\theta}_{\hat{q}_\lambda,MLE} = O_p(1/\sqrt{n}),$$

and moreover, by straightforward calculations, we can show that

$$\begin{aligned}
\ell(\hat{\boldsymbol{\theta}}_{\hat{q}_\lambda,MLE}) - \ell(\hat{\boldsymbol{\theta}}_{q,MLE}) &= \sqrt{n}\{G_n(\hat{\theta}_{\hat{q}_\lambda,MLE},\hat{\boldsymbol{\beta}}_{\hat{q}_\lambda,MLE}) - G_n(\hat{\theta}_{q,MLE},\hat{\boldsymbol{\beta}}_{q,MLE})\} \\
&\quad +n\mathrm{E}\{\log f(\mathbf{x},\hat{\theta}_{\hat{q}_\lambda,MLE},\hat{\boldsymbol{\beta}}_{\hat{q}_\lambda,MLE}) - \log f(\mathbf{x},\hat{\theta}_{q,MLE},\hat{\boldsymbol{\beta}}_{q,MLE})\} \\
&= O_p\left(\sqrt{n}(C_4\delta_n + \sqrt{n}C_5\delta_n^2)\right) \\
&= O_p(1).
\end{aligned}$$

(S.11) has been proved and the proof of Theorem 3 is complete. □

References

Bunea, F., Tsybakov, A. B., Wegkamp, M. and Barbu, A. (2010). SPADES and mixture models. *The Annals of Statistics*, **38**, 2525-2558.

van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge University press, Cambridge, United Kingdom.

Wang, H., Li, R. and Tsai, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*. **94**, 553-568.

Shanghai University of Finance and Economics

E-mail: huang.tao@mail.shufe.edu.cn

Hong Kong Baptist University

E-mail: hpeng@math.hkbu.edu.hk

Carnegie Mellon University & Max Planck Institute for Intelligent Systems

E-mail: kunz1@cmu.edu