Automated Model Selection (AMS) on Finite Mixtures: A Theoretical Analysis

on the AMS property of the BYY harmony function on the finite mixtures. In Section II, we introduce the BYY harmony learning theory and the harmony function on the finite mixtures. We then make an asymptotic analysis on the harmony function and prove that the global maximization of the harmony function leads to the AMS property if the average overlap measure between the actual components in the sample data is zero or becomes weak in Section III. We further analyze the deviation error of the maximum harmony estimates to the true parameters in Section IV. Finally, we conclude briefly in Section V.

II. BYY Harmony Learning and the Harmony Function

A BYY system describes each observation $x \in \mathcal{X} \subset R^n$ and its corresponding inner representation $y \in \mathcal{Y} \subset R^m$ via the two types of Bayesian decomposition of the joint density $p(x,y) = p(x)p(y|x)$ and $q(x,y) = q(x|y)q(y)$, called Yang machine and Ying machine, respectively. Here, y is limited to an integer variable, i.e., $y \in \mathcal{Y} = \{1, 2, \cdots, k\} \subset R$ with $m = 1$. Given a data set $D_x = \{x_t\}_{t=1}^N$, the task of learning on a BYY system consists of specifying all the aspects of $p(y|x), p(x), q(x|y), q(y)$ with a harmony learning principle implemented by maximizing the functional

$$H(p||q) = \int p(y|x)p(x) \ln [q(x|y)q(y)]dxdy - \ln z_q, \quad (1)$$

where z_q is a regularization term. Refer to [14] for details.

If both $p(y|x)$ and $q(x|y)$ are parametric, i.e., from a family of probability densities with a parameter θ, the BYY system is called to have a Bi-directional Architecture (or BI-Architecture for short). For the finite mixture modeling, we utilize the following specific BI-architecture of the BYY system: $q(j) = \alpha_j$ with $\alpha_j \geq 0$ and $\sum_{j=1}^k \alpha_j = 1$. Also, we ignore the regularization term z_q (i.e., set $z_q = 1$) and let $p(x)$ be the empirical density $p_0(x) = \frac{1}{N}\sum_{t=1}^N \delta(x - x_t)$, where $x \in \mathcal{X} = R^n$ and $\delta(\cdot)$ is a kind of kernel function (e.g., Gaussian function). Moreover, the BI-architecture is constructed with the following parametric form:

$$p(j|x) = p(y = j|x) = \frac{\alpha_j q(x|\theta_j)}{q(x|\Theta_k)}, \quad (2)$$

$$q(x|\Theta_k) = \sum_{j=1}^k \alpha_j q(x|\theta_j), \quad (3)$$

where $q(x|\theta_j) = q(x|y = j)$ with θ_j consisting of all its parameters and $\Theta_k = \{\alpha_j, \theta_j\}_{j=1}^k$. Substituting these component densities into Eq.(1) and letting the kernel functions tend to the delta functions, we have

$$\begin{aligned} & H(p||q) = J(\Theta_k) \\ = & \frac{1}{N}\sum_{t=1}^N \sum_{j=1}^k \frac{\alpha_j q(x_t|\theta_j)}{\sum_{i=1}^k \alpha_i q(x_t|\theta_i)} \ln [\alpha_j q(x_t|\theta_j)]. \end{aligned} \quad (4)$$

That is, $H(p||q)$ becomes a harmony function $J(\Theta_k)$ on the parameters Θ_k, i.e., the parameters of the finite mixture model $q(x, \Theta_k) = \sum_{j=1}^k \alpha_j q(x|\theta_j)$ for the observation x. Thus, the harmony learning on this BI-architecture of the BYY system reduces to the finite mixture modeling on a sample data set D_x.

Typically, we can let $q(x|\theta_j)$ be a Gaussian probability density function (pdf) given by

$$\begin{aligned} q(x|\theta_j) &= q(x|m_j, \Sigma_j) \\ &= \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma_j|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-m_j)^T \Sigma_j^{-1}(x-m_j)}, \end{aligned} \quad (5)$$

where m_j is the mean vector and Σ_j is the covariance matrix which is assumed to be positive definite. In this case, the BI-architecture of the BYY system contains the Gaussian mixture model $q(x, \Theta_k) = \sum_{j=1}^k \alpha_j q(x|m_j, \Sigma_j)$ which tries to model the underlying or true Gaussian mixture pdf of the sample data in D_x.

Under the BYY harmony learning principle [14], the maximization of $J(\Theta_k)$ should have the ability of AMS on the finite mixtures since it requires the least complexity of model structure. Indeed, the AMS property was demonstrated well via the gradient-type and iterative BYY learning algorithms [16]-[18] in the Gaussian mixture setting. However, this AMS property has not been proved mathematically. In the following, we try to analyze the harmony function asymptotically and prove this outstanding property.

III. Asymptotic Properties of the Harmony Function for Automated Model Section

A. Decomposition of the Harmony Function

We revisit the harmony function given in Eq.(4). In fact, it can be easily decomposed into two terms as follows.

$$\begin{aligned} J(\Theta_k) &= \frac{1}{N}\sum_{t=1}^N \sum_{j=1}^k p(j|x_t) \ln [\alpha_j q(x_t|\theta_j)] \\ &= \frac{1}{N}\sum_{t=1}^N \sum_{j=1}^k p(j|x_t) \ln p(j|x_t) \\ &\quad + \frac{1}{N}\sum_{t=1}^N \ln q(x_t|\Theta_k) \\ &= \frac{1}{N}\sum_{t=1}^N \ln q(x_t|\Theta_k) - \frac{1}{N}\sum_{t=1}^N I(x_t|\Theta_k), \end{aligned} \quad (6)$$

where

$$I(x_t|\Theta_k) = -\sum_{j=1}^k p(j|x_t) \ln p(j|x_t). \quad (7)$$

Clearly, the first term, i.e., $\frac{1}{N}\sum_{t=1}^N \ln q(x_t|\Theta_k)$, is just the log likelihood function on the finite mixture model with the sample data set D_x. The second term is a sum of the entropies of the posterior probabilities of the samples to k components.

As for the first term, i.e., the log likelihood function, there have been many investigations on its maximization that leads to the well-known maximum likelihood (ML) estimates of the parameters in the finite mixture. The EM

algorithm [2] is recognized as an efficient way to get the ML estimates, especially in the case of Gaussian mixture. However, the maximization of the log likelihood function, i.e., the maximum likelihood criterion, is incapable of model selection on the finite mixture. In fact, if we let $\mathcal{M}_k$ denote the class of all possible k-component mixtures built from a certain type of probability density functions (pdf's) (e.g., the pdf's of Gaussian mixtures):

$$q(x|\Theta_k) = \sum_{j=1}^{k} q(x|\theta_j), \tag{8}$$

it can be easily found that $\mathcal{M}_k \subset \mathcal{M}_{k+1}$. Thus, the maximized (log) likelihood is a nondecreasing function of k and the maximum likelihood criterion cannot detect the number of the components for a sample data set. That is, it has no ability to make model selection on the finite mixture.

In contrast, the second term of the harmony function must has the ability of model selection on the finite mixture if the harmony function really does. In fact, Roberts et al. [20] showed that the maximization of this part leads to the maximum certainty data partitioning that can allocate an appropriate number of clusters in the sample data. If each component in the mixture corresponds to a cluster in the sample data, this maximum certainty criterion can allocate an appropriate number k of components for the sample data. That is, it has the ability of model selection for the finite mixture model and we can use it as a model selection criterion on the finite mixture modeling. But its maximization may not lead to a result of AMS. As an illustration, it can always reach the maximum value 0 when we set one mixing proportion to be one and the others zeros. Therefore, its maximization cannot make model selection on the finite mixture.

However, since the harmony function combines these two terms together, the maximization of the harmony function may be able to make model selection on the finite mixture, which will be proved in the following subsections.

B. The AMS Property of the Harmony Function in the Well-Separated Case

To get rid of the randomness of the sample data, we consider the harmony function asymptotically. That is, we let $N \to \infty$. According to probability theory, we have

$$H(\Theta_k) = \lim_{N\to\infty} J(\Theta_k) = H_1(\Theta_k) + H_2(\Theta_k), \tag{9}$$

where

$$H_1(\Theta_k) = \int q(x|\Theta^*_{k^*}) \ln q(x|\Theta_k) d\mu; \tag{10}$$

$$H_2(\Theta_k) = -\int I(x|\Theta_k) q(x|\Theta^*_{k^*}) d\mu, \tag{11}$$

where μ is the appropriate underlying measure on R^n, and $\Theta^*_{k^*} = \{\alpha^*_j, \theta^*_j\}_{j=1}^{k^*}$ denotes the set of the parameters in the finite mixture pdf where the sample data come from. Specifically, k^* is the number of the actual components and $\Theta^*_{k^*}$ is the set of true parameters of the actual finite mixture pdf for the sample data. Here, we always assume that these actual components are different.

For convenience of analysis, we assume that all the components in the finite mixture have the same functional form (like Gaussian mixture). Moreover, the finite mixtures we consider are discriminant. That is, in the cases that all the components are different, $q(x|\Theta_k) = q(x|\Theta'_{k'})$ if and only if $\Theta_k = \Theta'_{k'}$ with $k' = k$ or $\Theta_k \subset \Theta'_{k'}$ with $k < k'$ and the mixing proportions of the other $k' - k$ extra components in $\Theta'_{k'}$ being zero (i.e., these components have no contribution to the finite mixture pdf.).

In the finite mixture model, the components may be well-separated in some special cases, i.e., each posterior probability $p(j|x)$ at a sample x is either 1 or 0. That is, each sample x is clearly belongs to one component. In this case, it is clear that $p(j|x) \ln p(j|x) = 0$ for all $x \in R^n$. We now investigate the AMS property of the harmony function with the components in the true (or actual) finite mixture being well-separated and have the following theorem.

Theorem 1. Suppose that the finite mixtures $q(x|\Theta_k)$ are discriminant. If the components in the true finite mixture $q(x|\Theta^*_{k^*})$ are well-separated, the asymptotic harmony function $H(\Theta_k)$ is globally maximized if and only if $\Theta_k = \Theta^*_{k^*}$ with $k = k^*$ or $\Theta^*_{k^*} \subset \Theta_k$ with $k > k^*$ and the mixing proportions of the other $k - k^*$ extra components in Θ_k being zeros.

Proof: According to the information theory, we have

$$H_1(\Theta_k) \leq H_1(\Theta^*_{k^*}); \tag{12}$$

$$H_2(\Theta_k) \leq 0. \tag{13}$$

Because the components in the true finite mixture $q(x|\Theta^*_{k^*})$ are well-separated, i.e., the posterior probability $p(j|x)$ at the parameters $\Theta^*_{k^*}$ are either 1 or 0, we thus have that $H_2(\Theta^*_{k^*}) = 0$. Therefore, $H(\Theta_k)$ is really globally maximized at $\Theta^*_{k^*}$.

On the other hand, suppose that $H(\Theta_k)$ is globally maximized. According to Eqs.(12)&(13), we must have that $H_1(\Theta_k) = H_1(\Theta^*_{k^*})$ and $H_2(\Theta_k) = 0$. From $H_1(\Theta_k) = H_1(\Theta^*_{k^*})$, we further have $q(x|\Theta_k) = q(x|\Theta^*_{k^*})$. Based on the discrimination of the finite mixtures, we consider the possible expressions for the parameters set Θ_k in the following three cases:

(i). $\Theta_k = \Theta^*_{k^*}$ with $k = k^*$;

(ii). $\Theta^*_{k^*} \subset \Theta_k$ with $k > k^*$ and the mixing proportions of the other $k - k^*$ extra components in Θ_k being zeros;

(iii). $k > k^*$ and there appears at least one repeating component parameter representation $\theta_j = \theta_{j'}$ in Θ_k with $\alpha_j > 0, \alpha_{j'} > 0$. In such a case, there are posterior probabilities $p(j|x)$ and $p(j'|x)$ being neither 1 nor 0 in a region with a positive measure. Thus, $H_2(\Theta_k) < 0$, which is contrary to $H_2(\Theta_k) = 0$. Thus, this case cannot happen for the global maximum of $H(\Theta_k)$.

Summing up the above the results, we have completed the proof. **Q. E. D.**

By Theorem 1, we have actually proved that the global maximization of the harmony function leads to the AMS

property on the finite mixture model in the well-separated case if we let $k > k^*$ and cancel the components with negligible mixing proportions. That is, in this case, if the model scale is actually defined by the number of positive mixing proportions in a finite mixture model, it will be equal to k^* via globally maximizing the harmony function. Thus, the true model scale can be correctly detected through the global maximization of the harmony function in this case. Moreover, in this special case, the global harmony maximum estimates of the parameters in the actual finite mixture is just those in $\Theta^*_{k^*}$, i.e., these maximum harmony estimates are unbiased.

From the above proof, we can also see that the global maximization of the asymptotic log likelihood function $H_1(\Theta_k)$ should also have the model selection property if we neglect the repeating component parameter representations if possible. Indeed, we can find such a phenomenon in the EM algorithm that some mixing proportions tend to zero in the case of $k > k^*$. However, it does not always happen. The reason is that the EM algorithm is conducted on a finite sample data set. In this situation, the complicated structure of the finite mixture tends to give a higher value of the log likelihood. But if we consider $H_2(\Theta_k)$ together, the situation may be changed considerably. Actually, it decreases greatly if there are a larger number of positive mixing proportions remained, even for a finite sample data set. That is, it is a strong penalty term for the model scale of the finite mixture. Therefore, the two terms in the harmony function play together to make AMS in the finite mixture more efficiently.

C. The AMS Property of the Harmony Function in the Weak-Separated Case

We further investigate the AMS property of the asymptotic harmony function in the weak-separated case where the average overlap among the actual components is low. That is, the actual components are overlapped in a weak mode such that most of the posterior probabilities are still either 1 or 0, or near 1 or 0, while the others remain within the interval (0,1). For mathematical analysis, we introduce the average overlap measure of the finite mixture which was defined in [21], [22].

We consider the posterior probabilities on the finite mixture at the true parameters $\Theta^*_{k^*}$:

$$p(j|x) = \frac{\alpha^*_j q(x|\theta^*_j)}{\sum_{i=1}^{k^*} \alpha^*_i q(x|\theta^*_i)}, \tag{14}$$

for $j = 1, \cdots, k^*$. We let

$$\gamma_{ij}(x) = (\delta_{ij} - p(i|x))p(j|x), \tag{15}$$

for $i, j = 1, \cdots, k^*$, where δ_{ij} is the Kronecker function. Then, we define a group of quantities on the overlap of component densities as follows:

$$e_{ij}(\Theta^*_{k^*}) = \int |\gamma_{ij}(x)| q(x|\Theta^*_{k^*}) d\mu, \tag{16}$$

for $i, j = 1, \cdots, k^*$, where $e_{ij}(\Theta^*_{k^*}) \leq 1$ since $|\gamma_{ij}(x)| \leq 1$.

We consider the worst case and define the average overlap measure of the finite mixture by

$$e(\Theta^*_{k^*}) = \max_{i,j} e_{ij}(\Theta^*_{k^*}). \tag{17}$$

In fact, for $i \neq j$, $e_{ij}(\Theta^*_{k^*})$ can be considered as a measure of the average overlap between the densities of components i and j in the finite mixture. In fact, when $q(x|\theta^*_i)$ and $q(x|\theta^*_j)$ have a high overlap at a point x, $p(i|x)p(j|x)$ takes a large value; otherwise, $p(i|x)p(j|x)$ takes a small value. When they are well separated at x, $p(i|x)p(j|x)$ becomes zero. Thus, the product $p(i|x)p(j|x)$ represents a degree of overlap between $q(x|\theta^*_i)$ and $q(x|\theta^*_j)$ at x in the mixture, and the above $e_{ij}(\Theta^*_{k^*})$ is an average overlap measure between the densities of components i and j in the mixture. On the other hand, $e_{ii}(\Theta^*_{k^*}) = \sum_{j \neq i} e_{ij}(\Theta^*_{k^*})$ which can be considered as the sum of the average overlap measures from component i to all the other components.

It can be easily found that in the well-separated case discussed above, each $\gamma_{ij}(x) = 0$ for all $x \in R^n$. Thus, the average overlap $e(\Theta^*_{k^*}) = 0$. In the following, we try to prove that the AMS property of the asymptotic harmony function still holds in the weak-separated case where the average overlap measure is very small. Actually, for the finite mixtures of densities from exponential families (including Gaussian densities), the average overlap $e(\Theta^*_{k^*})$ can be reduced to zero as an infinitesimal under some regular conditions [22]. We now give the variation of $H_2(\Theta^*_{k^*})$ with the average overlap measure $e(\Theta^*_{k^*})$ considering as an infinitesimal by the following theorem.

Theorem 2. Suppose that $e(\Theta^*_{k^*})$ tends to zero as an infinitesimal, we have

$$H_2(\Theta^*_{k^*}) \geq -\nu - O(e(\Theta^*_{k^*})), \tag{18}$$

where ν is a small positive number and $O(u)$ denotes the same order infinitesimal of an infinitesimal u.

Proof: According to Eq.(11), we have

$$\begin{aligned} |H_2(\Theta^*_{k^*})| &= \int I(x|\Theta^*_{k^*}) q(x|\Theta^*_{k^*}) d\mu \\ &= \sum_{j=1}^{k^*} \int |p(j|x) \ln p(j|x)| q(x|\Theta^*_{k^*}) d\mu \end{aligned} \tag{19}$$

Since $p(j|x) \in [0, 1]$, we consider it in two intervals $[0, \rho]$ and $(\rho, 1]$, where ρ is a small positive number. Because $\lim_{x \to 0^+} x \ln x = 0$, we can select ρ to be small enough to make $|x \ln x| \leq \nu / k^*$. On the other hand, it can be easily verified that there exists a positive number T such that $|x \ln x| \leq T|x(1-x)|$ in the interval $(\rho, 1]$. If we let $\mathcal{R}_1$ and $\mathcal{R}_2$ denote the regions of x for $p(j|x)$ in $[0, \rho]$ and

$(\rho, 1]$, respectively, we have

$$\begin{aligned}
& \int |p(j|x) \ln p(j|x)| q(x|\Theta^*_{k^*}) d\mu \\
= & \int_{\mathcal{R}_1} + \int_{\mathcal{R}_2} |p(j|x) \ln p(j|x)| q(x|\Theta^*_{k^*}) d\mu \\
= & \int_{\mathcal{R}_1} |p(j|x) \ln p(j|x)| q(x|\Theta^*_{k^*}) d\mu \\
& + \int_{\mathcal{R}_2} |p(j|x) \ln p(j|x)| q(x|\Theta^*_{k^*}) d\mu \\
\leq & \int_{\mathcal{R}_1} (\nu/k^*) q(x|\Theta^*_{k^*}) d\mu \\
& + T \int_{\mathcal{R}_2} |p(j|x)(1 - p(j|x))| q(x|\Theta^*_{k^*}) d\mu \\
\leq & \int \frac{\nu}{k^*} q(x|\Theta^*_{k^*}) d\mu \\
& + T \int |p(j|x)(1 - p(j|x))| q(x|\Theta^*_{k^*}) dx \\
= & \frac{\nu}{k^*} + T e_{jj}(\Theta^*_{k^*}) \\
= & \frac{\nu}{k^*} + O(e(\Theta^*_{k^*})). \qquad (20)
\end{aligned}$$

Substituting the above inequalities for $j = 1, \cdots, k^*$ into Eq.(19) and via $|H_2(\Theta^*_{k^*})| = -H_2(\Theta^*_{k^*})$, we finally have

$$H_2(\Theta^*_{k^*}) \geq -\nu - O(e(\Theta^*_{k^*})). \qquad (21)$$

The proof is completed.

Q. E. D.

According to Theorem 2, we further have

$$H(\Theta^*_{k^*}) \geq H_1(\Theta^*_{k^*}) - [\nu + O(e(\Theta^*_{k^*}))], \qquad (22)$$

which means that $H(\Theta^*_{k^*})$ is close to the upper bound of the asymptotic harmony function $H(\Theta_k)$, i.e., $H_1(\Theta^*_{k^*})$, when the average overlap measure between the actual components is very small (considering that ν is a very small number). However, $\Theta^*_{k^*}$ may not be the global maximum of the asymptotic harmony function. Although $H_1(\Theta_k)$ is globally maximized at $\Theta^*_{k^*}$, $H_2(\Theta_k)$ may be globally maximized at some point nearby $\Theta^*_{k^*}$. As a result, the global maximum of the asymptotic harmony function may has some deviation from $\Theta^*_{k^*}$. Clearly, this deviation is very small and the model scale of the finite mixture keeps k^*. Otherwise, the asymptotic harmony function will be decreased considerably and cannot be globally maximized. Therefore, in a similar way, the global maximization of the asymptotic harmony function also lead to the AMS property in the weak-separated case.

IV. Analysis of Deviation Error of the Maximum Harmony Estimates

In addition to the AMS property, it is also valuable to obtain good estimates of the parameters in the actual finite mixture via the global maximization of the harmony function. According to the previous analysis, the global maximum harmony estimates are unbiased in the well-separated cases. However, they may be biased in the overlap situation. In this section, we further analyze the deviation error of the maximum harmony estimates to the true parameters in the Gaussian mixture setting with help of the iterative learning algorithm constructed in [16].

A. The Iterative Learning Algorithm for Gaussian Mixtures with Automated Model Selection

We begin to introduce the iterative learning algorithm for maximizing the harmony function given in Eq.(4) where $q(x|\theta_i) = q(x|m_i$

analysis, we assume that this iterative learning algorithm is