Convex Approximation to Mixture Models Using Step Functions

Yi Xu, Yilin Zhu and Zhongfei Zhang

May 23, 2013

Abstract

The *parameter estimation* to mixture models has been shown as a local optimal solution for decades. In this paper, we propose a *functional estimation* to mixture models using step functions. We show that the proposed functional inference yields a convex formulation and consequently the mixture models are feasible for a global optimum inference. We further establish the asymptotic consistency and deduce the convergence rate of the proposed functional inference method. This rate turns out to be $\mathcal{O}(n^{-\frac{1}{2}})$ under a minor condition. The proposed approach further unifies the existing isolated exemplar-based clustering techniques, e.g. [8], at a higher level of generality: It provides a theoretical justification for the heuristics of the clustering by affinity propagation [8]; it reproduces [14]'s convex formulation as a special case under this step function construction.

1 Introduction

Mixture models are a technique to interpret an unknown distribution via an additive mixture of a (resonably assumed) component distribution function. While the family of mixture models provides convenient and meaningful interpretations to the data, to inference the mixture models poses non-trivial challenges. Current mainstream algorithms have formed three major directions: likelihood methods, Bayesian methods, and approaches based on the method of moments. The likelihood methods take their roots from the orthodox theory of parametric estimation; thorough studies in this direction have been carried on during the last few decades. Despite the solid mathematical foundation, due to the non-convex nature of the mixture likelihood, the likelihood based algorithms [16], such as the Expectation-Maximization algorithms [5, 15, 17, 27], have multiple local maxima or may fail to converge [16, 26].

To eliminate the dependency on the parameters that leads to the non-convexity in the likelihood methods, non-parametric Bayesian methods thrive in recent years [21, 20, 23, 10, 13]. In this line of formulation, a general prior (usually Dirichlet prior) is assigned to the parameter space, and the corresponding

Bayesian posteriors thus manage to abandon the parametric formulation as appeared in the likelihood methods. While conceptually appealing, the Bayesian methods are mostly analytically intractable due to the complexity of the posteriors; their solutions have to be estimated through approximation from the original formulation using the Markov Chain Monte Carlo (MCMC) techniques. Since MCMC methods also bear practical issues such as thinning the chain, burn-in, initial value selection, and label switching [2, 24], these technical tradeoffs and heuristic strategies increase the performance uncertainty of the Bayesian methods.

The method of moments initiates its idea by learning the mixture of two arbitrary Gaussians with a provable accuracy [4, 6, 12]. When generalizing this idea to learning mixture of many Gaussians, the applied techniques encounter a curse of dimensionality. [18] proposes a revised method of moments algorithm that is capable of learning mixtures of many Gaussians with a near optimal guarantee. This algorithm, however, depends exponentially on the number of Gaussians in the mixture and such a dependency is proven to be necessary [18].

The above brief review profiles the current challenges of mixture model inference. In this paper, we propose a new method on convex approximation to mixture models using step functions, called CAMS, that is different from the above three directions. In comparison with the existing methods, CAMS has the following confirmed properties. The objective functional is analytically tractable and convex; the estimation satisfies an asymptotic consistency and has a convergence rate of $(O(n^{-\frac{1}{2}}))$ under a minor condition.

The inference method of CAMS is a theoretic complement to that of the affinity propagation algorithm proposed in [7] and later in [8] for clustering, which is known for its simple exemplar-based message passing mechanism and superior empirical performance. Yet until now, doubts on this heuristic algorithm have never been fully clarified due to the lack of theoretical justification. Another closely related work is the convex clustering scheme proposed by [14]. However, [14] attempts to justify its intuition by relating to rate-distortion problems and fails to establish its theoretical relation to affinity propagation. CAMS succeeds in establishing a common theoretic foundation, as well as the desirable statistic properties, to all the related literature including [7] and [14].

The rest of the paper is organized as follows. In Section 2, we describe the mathematical procedure for CAMS and establish its global optimality. In Section 3, we discuss the statistic properties of CAMS. In Section 4, an empirical study is reported on CAMS, in comparison with affinity propagation [8], the popular EM algorithm for mixture models, and the MCMC based Bayesian algorithm for mixture models [1], with results verifying the theoretical promise of CAMS. In Section 5 we discuss the potential real-world applications of CAMS.

2 Density Estimation to Mixture Models Using Step Functions

The general form of additive mixture models is expressed as:

$$G(x|\mathrm{a}) = \int_{\Theta} g(x|\theta)\mathrm{a}(\theta)d\theta \tag{1}$$

In the mixture model Eq. (1), $G(x|\mathrm{a})$ is an unknown probability density distribution; $g(x|\theta)$ is a given component distribution function, usually chosen from the exponential family. θ is the location parameter, usually the first moment of g. $\mathrm{a}(\theta) > 0$ is an unknown non-negative weight function. Θ is the whole parameter space of θ. $d\theta$ is short for $d\theta_1 d\theta_2 ... d\theta_p$, where p is the dimension of Θ. In order to understand the unknown distribution $G(x|\mathrm{a})$, one needs to find the optimal estimation to $\mathrm{a}(\theta)$.

Notice that by assuming $\mathrm{a}(\theta)$ to be a delta function, $\mathrm{a}(\theta) = \sum_{j=1}^{k} \pi_j \delta(\theta - \theta_j)$, Eq. (1) is reduced to a more common form of mixture models [16].

$$G(x) = \sum_{j=1}^{k} \pi_j g(x|\theta_j) \tag{2}$$

In order to reconstruct $G(x)$, one needs to find the inference to k, $\{\theta\}_{j=1}^{k}$, and $\{\pi_j\}_{j=1}^{k}$.

By comparing the above two formulations of mixture models, Eq. (1) has the following advantages over Eq. (2). First, for any smooth $G(x)$, there is a solution $\mathrm{a}(\theta)$ existing to Eq. (1), given that $g(x|\theta)$ is Gaussian [11]. On the other hand, in most cases of a smooth $G(x)$, Eq. (2) does not have solutions at all, not even speaking to find a set of consistent estimations; even under the few rare cases (e.g., $\mathrm{a}(\theta)$ is a delta function) when Eq. (2) has solutions, solutions provided by the likelihood methods are always local optimum. Second, for Eq. (1) we only need to solve for one function variable $\mathrm{a}(\theta)$ as a solution while for Eq. (2) we need to solve for three types of parameters k, $\{\theta\}_{j=1}^{k}$, and $\{\pi_j\}_{j=1}^{k}$ as a solution; hence, solving Eq. (2) is more complicated and demanding. Third, given the fact that solutions to Eq. (1) are global optimum and solutions to Eq. (2) are local optimum, presumably it is possible for solutions to Eq. (1) to avoid the classic problems such as estimating the cluster number k and overfitting.

The conceptual simplicity of estimating the function $\mathrm{a}(\theta)$ has been considered in the non-parametric Bayesian methods [22, 20, 26]. The functional form Eq. (1) has appeared in the Bayesian inference literature. The strategy of Bayesian methods to estimate $\mathrm{a}(\theta)$ is mainly carried out by assigning Dirichlet prior to $\mathrm{a}(\theta)$ and then finding the maximal posterior of $\mathrm{a}(\theta)$ using techniques such as Gibbs sampling. As discussed in Section 1, the intractable formulations and the technical trade-offs make this line of methods infeasible for an optimal evaluation and a consistency proof. Rigorously speaking, Bayesian methods still aim to estimate Eq. (2) instead of Eq. (1), since $\mathrm{a}(\theta)$ is assumed as a delta function instead of an arbitrary non-negative function when an MCMC technique

is applied. Another thorough discussion on finding the solution of $\mathrm{a}(\theta)$ can be found in [11] where Jaynes mainly focuses on solving Eq. (1) as a mathematical problem of inversion, not an inference problem.

One subtle but non-trivial issue about the mixture models is the scale parameter of the component function $g(x|\theta)$. If g belongs to the exponential family, it usually involves a scaling parameter σ in addition to the location parameter θ. Some of the existing methods, such as those using EM or the method of moments, estimate σ, $\{\theta\}_{j=1}^{k}$, and $\{\pi_j\}_{j=1}^{k}$, given k as a parameter which is assumed known or estimated separately. If Eq. (1) is applied, depending on different values of σ, the corresponding global optimal estimations of $\mathrm{a}(\theta)$ can be found. In Section 3.3, we cite the conclusion in [11] that for a Gaussian mixture, the dependency on the scale parameter is necessary to uniquely determine $\mathrm{a}(\theta)$.

2.1 The Convex Likelihood Functional

In order to estimate an arbitrary non-negative function $\mathrm{a}(\theta)$, we consider approaching it using its step function approximation $\mathrm{a}_m(\theta)$:

$$\mathrm{a}(\theta) \doteq \mathrm{a}_m(\theta) = \sum_{j=1}^{m} \mathrm{a}(\boldsymbol{\theta}_j)\chi_{A_j}(\theta)$$
$$\text{where } \boldsymbol{\theta}_j \in A_j \text{is a constant} \tag{3}$$

χ_A is the indicator function:

$$\chi_{\mathrm{a}}(\theta) = \begin{cases} 1 & \text{if } \theta \in A \\ 0 & \text{otherwise} \end{cases}$$

$\{A_j | \cup_{j=1}^{m} A_j = \Theta \text{ and } A_i \cap A_j = \Phi, \forall i \neq j\}$ is an arbitrary finite partition to the parameter space Θ (See an one dimensional example in Fig. 1). The idea is that, by finding the optimal estimation of a_m defined on any partition $\{A_j\}_{j=1}^{m}$, as the partition becomes finer and finer, $\mathrm{a}(\theta)$ can be approximated at an arbitrary accuracy by its corresponding step function a_m.

Replacing $\mathrm{a}(\theta)$ with a_m in Eq. (1), the integral is approximated by the Riemann sum using the partition $\{A_j\}_{j=1}^{m}$:

$$\begin{aligned} G(x|\mathrm{a}) &= \tilde{G}(x|\mathrm{a}_m) + \epsilon \\ &= \sum_{j=1}^{m} g(x_i|\boldsymbol{\theta}_j)\mathrm{a}(\boldsymbol{\theta}_j)l(A_j) + \epsilon \qquad (4) \\ &= \sum_{j=1}^{m} g(x_i|\boldsymbol{\theta}_j)a_j + \epsilon \qquad (5) \end{aligned}$$

$l(A_j)$ is the volume of the subset A_j. From Eq. (4) to Eq. (5), a new variable $a_j = \mathrm{a}(\boldsymbol{\theta}_j)l(A_j)$ is introduced, since $\mathrm{a}(\boldsymbol{\theta}_j)$ and $l(A_j)$ act together as a whole. It is important to note that, different from the variable θ_j in Eq. (2), each $\boldsymbol{\theta}_j \in A_j$ in Eq. (4) is an arbitrary constant as defined in Eq. (3).

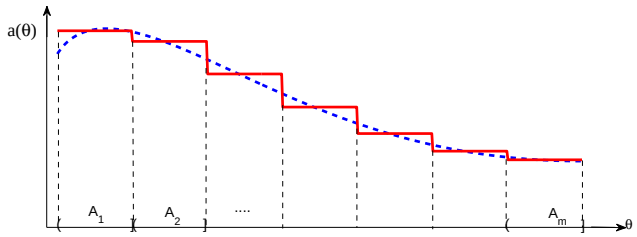


Figure 1: Using a step function to approximate a 1-dimension function

Consequently, given n observed samples $\{x_i\}_{i=1}^n$, the Fisher-Wald setting [25] of the log-likelihood functional for $\mathrm{a}(\theta)$ is approximated by:

$$\begin{aligned} L(\mathrm{a}) &= \sum_{i=1}^{n} \log G(x_i|\mathrm{a}) = \sum_{i=1}^{n} \log \tilde{G}(x|\mathrm{a}_m) + \epsilon' \\ &= \sum_{i=1}^{n} \log \sum_{j=1}^{m} g(x_i|\boldsymbol{\theta}_j) a_j + \epsilon' \\ &= L(\mathrm{a}_m) + \epsilon' \end{aligned} \tag{6}$$

In Section 3.1 we shall discuss how to construct a partition $\{A_j\}_{j=1}^m$ such that $|\epsilon'|$ can be arbitrarily small and how fast the optimal estimation to $\tilde{G}(x|\mathrm{a}_m)$ converges to $G(x|\mathrm{a})$. In this section we focus on the property of the likelihood functional $L(\mathrm{a}_m)$ and finding the optimal estimation to $\tilde{G}(x|\mathrm{a}_m)$.

Before we proceed to solve $L(\mathrm{a}_m)$, an l-1 norm constraint for $\{a_j|a_j \geq 0\}_{j=1}^m$ can be achieved by integrating over the density function $G(x|\mathrm{a})$:

$$\begin{aligned} 1 &= \int_{-\infty}^{+\infty} G(x|a)dx = \int_{-\infty}^{+\infty} \sum_{j=1}^{m} g(x|\boldsymbol{\theta}_j) a_j dx + \epsilon \\ &= \sum_{j=1}^{m} a_j + \epsilon \end{aligned} \tag{7}$$

To put Eq. (6) and Eq. (7) together, the maximum likelihood estimation to $\{a_j\}_{j=1}^m$ is achieved by maximizing the following objective functional:

$$\begin{aligned} Obj(\mathrm{a}_m) &= L(\mathrm{a}_m) - \lambda(\sum_{j=1}^{m} a_j - 1) \\ &= \sum_{i=1}^{n} \log \sum_{j=1}^{m} g(x_i|\boldsymbol{\theta}_j) a_j - \lambda(\sum_{j=1}^{m} a_j - 1) \end{aligned} \tag{8}$$

where λ is the Lagrange multiplier.

The following Lemma claims the convexity of the objective functional Eq. (8)[1].

[1] For the succinctness of the paper and the space limitation, all the proofs of the theorems and lemmas in this paper are provided as a supplementary file with this paper.

Lemma 1. *(***Convexity***) the likelihood functional Eq.* (8) *is concave down.*

2.2 Solutions

The variational solution to maximizing the likelihood functional Eq. (8) can be found by using various standard optimization techniques. Here we propose two options that relate closely to the exemplar-based techniques.

The first solution is by using the very simple gradient ascent method:

Lemma 2. *Using the standard gradient ascent method yields the following converging updating rules for* $\{a_j\}_{j=1}^m$*:*

$$a_j^{update} \leftarrow a_j + \Delta t \cdot (\sum_{i=1}^{n} r(i,j) - \lambda)$$

$$\lambda = \frac{1}{m} \sum_{j=1}^{m} \sum_{i=1}^{n} r(i,j) \tag{9}$$

where

$$r(i,j) = \frac{g(x_i|\boldsymbol{\theta}_j)}{\sum_{l=1}^{m} g(x_i|\boldsymbol{\theta}_l) a_l} \tag{10}$$

The gradient ascent method starts from a random initialization of $\{a_j\}_{j=1}^m$. Since Eq. (8) is convex, the final convergence of a_j's does not depend on the initialization. However, different initializations may yield different convergence paths with different speeds. The formulations of $r(i,j)$ and a_j in Lemma 2 are very simiar to the heuristic rules given in the affinity propagation algorithm by [7].

For the second solution, $L(\mathrm{a}_m)$ in Eq. (6) first needs to be expressed via the KL-divergence between $P(x) = 1/n$ and $G(x) = \sum_{j=1}^{m} a_j g(x|\boldsymbol{\theta}_j)$, where $x \in \{x_1, x_2 ... x_n\}$ for both P and G:

$$D(P|G) = -\sum_{i=1}^{n} P(x_i) \log G(x_i|\boldsymbol{\theta}_j) - H(P) = -L(\mathrm{a}_m) + const \tag{11}$$

Therefore, maximizing $L(\mathrm{a}_m)$ is equivalent to minimizing the KL-divergence $D(P|G)$. [3] proves that for this problem, the following non-negative iteration leads to the convergence for a_j's:

$$a_j^{update} = a_j \sum_{i=1}^{n} P(x_i) r(i,j) \tag{12}$$

where $r(i,j)$ is defined in Eq. (10). This solution, essentially the same as that proposed in [14], does not take into the consideration of the l-1 norm constraint Eq. (7). Therefore, during the implementation, one needs to manually normalize a_j's after each iteration, which may cause oscillations and cannot benefit from the sparsity encouraged by the l-1 norm.

Both algorithms have a time complexity of $\mathcal{O}(mn)$ for each iteration.

Algorithm 1 Clustering using CAMS

Input: $\{x_i\}_{i=1}^n$, $\{\boldsymbol{\theta}_j\}_{j=1}^m$, and g.
Output: Labels for $\{x_i\}_{i=1}^n$.
Method:

Compute $g(x_i|\boldsymbol{\theta}_j)$ for all i, j; initialize $a_j = 1/m$ for all $j = 1...m$
While $a_j, j = 1...m$ do not converge
  For $\{a_j\}_{j=1}^m$, if $a_j > 0$, update a_j using Eq. (9);
  For $\{x_i\}_{i=1}^n$, assign labels for x_i using Eq. (14);
  For $j = 1...m$, if no sample is labeled as j, set $a_j = 0$.
End while

2.3 Determining the Clustering Labels

In the problem of clustering, aside from fitting the mixture model, one must assign clustering labels to the observed samples. Given the partition $\{A_j\}_{j=1}^m$ to the parameter space, each x_i must be generated from one of the m partitions. Given the mixture model approximation $\tilde{G}(x|\mathrm{a}_m)$ defined on the partition $\{A_j\}_{j=1}^m$, the probability that x_i is generated from A_j can be computed via the following lemma:

Lemma 3. *Let $H_j^{x_i}$ denotes the event $\theta_i \in A_j$ where θ_i is the unknown parameter that generates x_i. Then we have:*

$$P(H_j^{x_i}|\tilde{G}(x|\mathrm{a}_m), x_i) = a_j r(i, j). \tag{13}$$

Based on Lemma 3, a natural labeling rule is to assign x_i to the most probable A_j:

$$x_i \text{ is labeled with } j \text{ if } a_j r(i, j) \geq a_l r(i, l), \forall 1 \leq l \leq m \tag{14}$$

The complete clustering algorithm proceeds as follows. First, a_j's are estimated using the updating rules in Eq. (9) until convergence. Then each sample x_i is labeled according to Eq. (14). This procedure thus automatically determines the cluster number.

To accelerate the convergence of a_j's, one can embed the labeling procedure Eq. (14) into the updating loop. More specifically, at any stage of the iteration to update a_j's, the label of x_i's can be identified by using Eq. (14). After each time the labels for x_i's identified, for those subsets A_j's that have no samples assigned, one directly sets the corresponding $a_j = 0$. This treatment is valid because for those A_j's that collect no samples, their a_j's have a negative increasing rate w.r.t. those a_j's that have samples assigned. Therefore, if we initialize all the a_j's with an equal value, those a_j's that have no samples assigned in the current iteration would never receive any samples in the subsequent iterations. Consequently, directly setting those a_j's as zeros cuts off unnecessary computation and accelerates the convergence. The complete clustering algorithm with this accelerating strategy is summarized in Alg. 1.

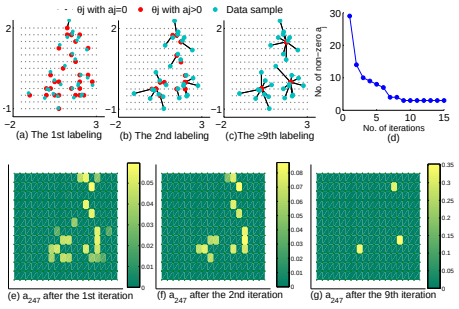


Figure 2: Dynamics of CAMS clustering

Fig. 2 showcases the dynamics of Alg. 1 applied to 30 2-dimension samples. The component distribution $g(x|\theta)$ is set as Gaussian with $\sigma^2 = 0.5$. The parameter space is discretized into a partition of 247 rectangles, as seen from Fig. 2 (a) and (e). Each $\cdot$ stands for a constant $\boldsymbol{\theta}_j$ in rectangle A_j. Fig. 2(e)-(g) visualize the step function a_{247} after the 1st, 2nd, and 9th iteration. Fig. 2(a)-(c) show the labeling results after the 1st, 2nd, and 9th iteration. After the 9th iteration, the values of a_j's stay stable and three clusters are found. Fig. 2(d) shows the converging process in terms of the number of non-zero a_j's after each iteration. One can observe that the converged a_{247} in Fig. 2(g) contains only three non-zero a_j's and all the other a_j's are converged to zero. The converged a_{247} in Fig. 2(g) can be understood as a rough approximation to a delta function $\mathrm{a}(\theta) = \sum_{j=1}^{3} a_j \delta(\theta - \boldsymbol{\theta}_j)$.

3 Statistic Properties of the Step Function Approximation

In this section we discuss the correctness of the proposed step function approximation method. We answer the following crucial questions:

1. Does there exist a finite partition, $\{A_j\}_{j=1}^m, m < \infty$, such that when the likelihood functions reach supreme values, $|\epsilon'|$ in Eq. (6) can be arbitrarily small? Or equivalently, does m have an upper bound?

2. Is $\hat{\tilde{G}}(x|\hat{\mathrm{a}}_m)$, where $\hat{\mathrm{a}}_m$ maximizes $Obj(\mathrm{a}_m)$, a consistent estimation to $G(x|\mathrm{a})$? If so, what is the convergence rate as the sample size $n \to \infty$?

3. How does the scale parameter σ in g influence the estimation?

For the clarity of the presentation, we define a matrix $\mathbf{G}$ as follows:

Definition 1. *Given n observations $\{x_i\}_{i=1}^n$ generated from $G(x|\mathrm{a})$ and m constants $\{\boldsymbol{\theta}_j\}_{j=1}^m$ from step function approximation $\tilde{G}(x|\mathrm{a}_m)$, define the matrix $\mathbf{G}^{m\times n}$ with $\mathbf{G}_{ji} = g(x_i|\boldsymbol{\theta}_j)$.*

Figure 3: The convergence of a_{30}

3.1 Upper Bound for m

By using a step function a_m to approximate $\mathrm{a}(\theta)$, an error ϵ' inevitably occurs between the supreme values of $L(\mathrm{a})$ and $L(\mathrm{a}_m)$ as indicated in Eq. (6). It appears that to reduce $|\epsilon'|$ to an infinitely small value, one must increase the partition number m to infinity. If this is true, the step function estimation becomes computationally infeasible. The good news is that, given a finite number of observations $\{x_i\}_{i=1}^n, n < \infty$, there exists an $m \leq n$, such that $|\epsilon'|$ between the supreme values of $L(\mathrm{a})$ and $L(\mathrm{a}_m)$ can be arbitrarily small. This conclusion is established in the following theorem:

Theorem 1. *Given n observations generated from $G(x|\mathrm{a})$, for any step function approximation $\tilde{G}(x|\mathrm{a}_m)$ with $m >> n$, there exists a step function approximation $\tilde{G}(x|\mathrm{a}_{m'})$, such that $m' \leq n$ and*

$$\begin{aligned} & \sup_{\{a_j|a_j\geq 0, j=1\ldots m; \sum_{j-1}^{m} a_j=1\}} \left\{L(\mathrm{a}_m)\right\} \\ = & \sup_{\{a'_j|a'_j\geq 0, j=1\ldots m'; \sum_{j-1}^{m'} a'_j=1\}} \left\{L(\mathrm{a}_{m'})\right\} \end{aligned} \tag{15}$$

Moreover, m' is the rank of $\mathbf{G}^{m\times n}$ from a_m. Let $\boldsymbol{\theta}_j, j = 1\ldots m'$ form the m' linear independent rows of $\mathbf{G}^{m\times n}$; The constants $\{\boldsymbol{\theta}'_j\}_{j=1}^{m'}$ in $\mathrm{a}_{m'}$ can be determined by setting $\boldsymbol{\theta}'_j = \boldsymbol{\theta}_j, j = 1\ldots m'$.

Theorem 1 claims that one can construct an arbitrarily fine approximation a_m with m as large as possible, so that $|\epsilon'|$ as small as possible; this a_m can be replaced with a coarser approximation $\mathrm{a}'_m, m' \leq n$ without increasing $|\epsilon'|$. This conclusion overrules the conjecture made in [19] that increasing the partitioning number always increases the supreme value of the likelihood function.

Here we demonstrate Theorem 1 using the earlier example in Fig. 2. Computation confirms that the 30 out of the total 247 $\boldsymbol{\theta}_j$'s that are closest to the 30 samples (in Fig. 3(a)) form 30 linear independent rows in $\mathbf{G}^{247\times 30}$. Fig. 3(b) showcases an imaginary step function a_{30} that is based on the 30 $\boldsymbol{\theta}_j$'s. By using the 30 $\boldsymbol{\theta}_j$'s in Fig. 3(a), we obtain the same three non-zero a_j's after convergence (in Fig. 3(c)) and consequently the same clustering result.

Fig. 3 (a) suggests that for Gaussian kernel g, a $\boldsymbol{\theta}_j$ that is closer to a sample x_j is more likely to contribute as a row vector in the basis of $\mathbf{G}$. This behavior inspires an efficient method to construct the partition $\{A_j\}_{j=1}^m$: Given

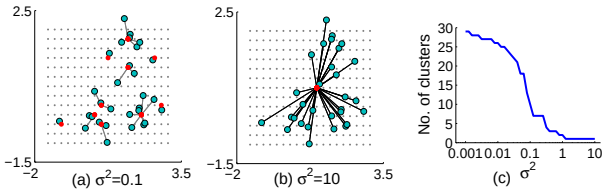


Figure 4: σ^2 vs. cluster number

n observations $\{x_i\}_{i=1}^n$, construct $\{A_j\}_{j=1}^n$ by setting $\boldsymbol{\theta}_i = x_i, i = 1...n$. The same strategy is applied as an intuitive choice in [7] and [14].The corresponding $\mathbf{G}^{n\times n}$ is symmetric and routinely called the similarity matrix in the clustering literature.

3.2 Consistency and Convergence Rate

Based on the convexity of the likelihood functional and Theorem 1, the asymptotic consistency, as well as the convergence rate of $\tilde{G}(x|\mathrm{a}_m)$, can be further established in the following theorem.

Theorem 2. *For any fixed x, the maximum likelihood step function estimator $\hat{\tilde{G}}(x|\hat{\mathrm{a}}_m)$, where $m = \mathcal{O}(n)$, is a consistent estimator of the mixture model $G(x|\mathrm{a})$, i.e.*

$$\hat{\tilde{G}}(x|\hat{\mathrm{a}}_m) \xrightarrow{p} G(x|\mathrm{a}), \text{ as } n \to \infty. \tag{16}$$

Suppose the whole parameter space Θ has a finite measure and $h(x,\theta) = g(x,\theta)\mathrm{a}(\theta)$ satisfies the Holder condition with $\alpha = 1/2$ respect to θ, that is, $\exists K > 0, s.t. |h(x,\theta_1) - h(x,\theta_2)| \leq K|\theta_1 - \theta_2|^{1/2}$, for any x and $\theta_1, \theta_2 \in \Theta$, the following convergence rate for $\hat{\tilde{G}}(x|\hat{\mathrm{a}}_m)$ holds:

$$\sup_{x\in\mathbb{R}} |\hat{\tilde{G}}(x|\hat{\mathrm{a}}_m) - G(x|\mathrm{a})| = \mathcal{O}_p(1/\sqrt{n}) \tag{17}$$

When the component distribution g is Gaussian, a stronger but more restricted convergence rate for maximum likelihood estimator (MLE) and posterior distribution in Bayesian density estimation problems has been obtained by [9]. The rate is seen to be $\mathcal{O}_p((\log n)^\kappa/\sqrt{n})$ in terms of the l-1 distance between $G(x|\mathrm{a})$ and its estimation, where $\kappa \geq 1$ is a constant. In MLE, κ depends on the type of the mixtures and the choice of the sieve; in Bayesian methods, it depends on the tail behavior of the base measure of the Dirichlet process. Given this stronger rate, the rate obtained in Theorem 2 is still meaningful, as $\mathcal{O}_p(1/\sqrt{n})$ can be achieved without specific assumptions on the component distribution g or the range of the sieve.

Name	A2	S2	D31	Face
Size	5250	5000	3100	900
Dimension	2	2	2	2500
Cluster No.	35	20	31	N/A

Table 1: Dataset information

3.3 The Scaling Parameter

When the component distribution $g(x|\theta)$ contains a scaling parameter σ, each different value of σ yields a different $\mathbf{G}$, leading to a unique maximum likelihood estimation to a_m, and consequently a different clustering result. Intuitively, in the problem of clustering, σ determines the size of a cluster. A larger value of σ results in fewer clusters; a smaller value of σ results in more clusters. By changing the value of σ, the clustering method using step function inference obtains different numbers of clusters.

Such a dependence on the scaling parameter for CAMS has an analytical explanation. In [11] (Chapter 7), Janeys obtained the analytical solution $\mathrm{a}(\theta)$ to the mixture equation Eq. (1), given $G(x) \in C^\infty$ and Gaussian component $g(x|\theta,\sigma)$:

$$\mathrm{a}(\theta) = \sum_{m=1}^{\infty} \frac{-1^m \sigma^{2m}}{2^m m!} \frac{d^{2m}}{dx^{2m}} G(\theta) \tag{18}$$

In this solution, when $G(x)$ is fixed, each value of σ results in a different $\mathrm{a}(\theta)$. Consistently, if $\mathrm{a}(\theta)$ needs to be inferred from observations generated from $G(x)$, a correctly behaved inference algorithm must deliver a dependence on σ, and each σ produces a unique estimation to $\mathrm{a}(\theta)$. Such a behavior of Gaussian mixture actually agrees well to the common sense, just like the fact that the earth can be described either in the scale of the solar system, or in the scale of the galaxy, or in the scale of the whole universe. By and large, it is natural and intuitive that different scales result in different judgments, each making its own sense.

Fig.4 demonstrates how σ^2 scales the inference result using the earlier examples. Fig.4(a) and (b) showcase the converged clustering results with $\sigma^2 = 0.1$ and $\sigma^2 = 10$, respectively, in comparison with $\sigma^2 = 0.5$ in Fig. 2 (c). The curve in Fig.4(c) further plots the decreasing cluster number from 30 to 1 as σ^2 increases from 0.001 to 10.

4 Empirical Evaluations

In this section we study empirical evaluations to CAMS. We compare the performance of Algorithm 1 with EM method (EM), variational Bayesian method (VB) and affinity propagation (AP). All the codes of the comparing methods are from the published open sources.[2] For CAMS, the $\{\boldsymbol{\theta}_j\}_{j=1}^m$ needed in Algorithm

[2]Open source codes:
www.psi.toronto.edu/affinitypropagation/software/apcluster.m.

	A2	S2	D31	Face
EM	2.1h	2.2h	5.2h	9.1h
VB	2.8h	6.9h	2.3h	5.3h
AF	109sec	208sec	79sec	27sec
CAMS	42sec	55sec	22sec	8.2sec

Table 2: CPU times

1 is constructed as $\boldsymbol{\theta}_i = x_i, i = 1...n$. In order to give a comprehensive evaluation, the clustering results under a wide range of cluster numbers are obtained using the four algorithms.

Four benchmark datasets from public domain [3] with different sizes, dimensions, and true cluster numbers are elected for the evaluation. The four datasets (summarized in Table 1) involve 3 synthetic Gaussian mixture datasets of different degrees of overlaps (A2, S2, and D31) and 1 real-world dataset, called Face, which is derived from olivettiface database and has been used for the evaluation of AP in [8].We use the mean square error (MSE) within clusters as the performance criterion for all the four algorithms.

The MSE evaluations are documented in Fig. 5. As for EM and variational Bayesian methods, since the performance depends on the initialization, 500 runs with different random initializations for each dataset are performed; all the MSE values are reported. The CPU time for all the four algorithms on the four datasets are recorded in Table 2.

Fig. 5(a) shows the MSE comparison between EM and CAMS. We observe that one run of CAMS outperforms 500 runs of EM in all the four datasets and all the cluster numbers. The gap between CAMS and EM becomes larger as the cluster number approaches to the ground truth.

Fig. 5(b) reports the MSE comparison between VB and CAMS. The ranges of cluster numbers depicted for VB in datasets A2, S2 and D31 are all the possible cluster numbers that can be obtained by VB. We observe that CAMS outperformes VB within the range of cluster numbers that VB can obtain. In the MSE graphs for A2 and D31, VB is not able to reach the ground truth cluster number 35 and 31, respectively.

Fig. 5(c) reports the MSE comparison between affinity propagation and CAMS. Since both methods share the same working mechanism with a slight difference in the updating equations, we expect the two algorithms to generate similar performance curves. As shown in the datasets A2, D31, and Face, the curves of the two algorithms indeed have a close match, indicating that both algorithms share the same charicteristics. However, in the graph of S2, the slight formulation difference results in a not so small performance gap when the cluster number is far from the ground truth.

www.mathworks.com/matlabcentral/fileexchange/26184-em-algorithm-for-gaussian-mixture-model

[3]Publicly accessible datasets:
cs.joensuu.fi/sipu/datasets/.
psi.toronto.edu/affinitypropagation/Faces.JPG

5 Last Comments

CAMS can be naturally applied to solving many real-world problems such as graph partitioning, finding information cascades in networks, and link prediction with a convex formulation. Given a graph $G(E, V)$, if each edge $e_{i,j} \in E$ indicates the strength of the relation between nodes i and j, one can then compute a_j and $r(i, j)$ for each node j. Using the labeling rule Eq. (14), each node i points to one of its neighbor node j, including itself; and immediately, these new directed relations reduce the original graph to a set of trees (one can prove that if Eq. (14) only holds when $i = j$, the directed graph is acyclic and each connected component forms a tree structure). These tree structures not only provide detected communities from the original graph, the hierarchical structure revealed in each tree can also be considered as the maximum likelihood information cascades detected from the graph. The roots of all the trees form a minimal seed set to propagate information. The tree structures further provide clues for link prediction; one can reasonably assume that a node i is more likely to link to its siblings and ancestors within the tree rather than to link to nodes from other trees. Consequently, by reducing the original graph to a set of trees with a convex formulation, CAMS connects solutions to all these independent problems under the same framework. These prospective arguments shall be our future work.

References

[1] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[2] H. Chung, E. Loken, and J. Schafer. Difficulties in drawing inferences with finite-mixture models: A simple example with a simple solution. *The American Statistician*, 58:152–158, 2004.

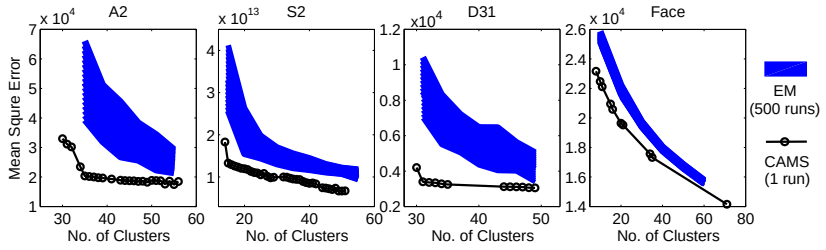
[3] I. Csiszar and P. Shields. Information theory and statistics: A tutorial. *Foundations and Trends in Communications and Information Theory*, 1:417–528, 2004.

[4] S. Dasgupta. Learning mixtures of gaussians. In *FOCS*, 1999.

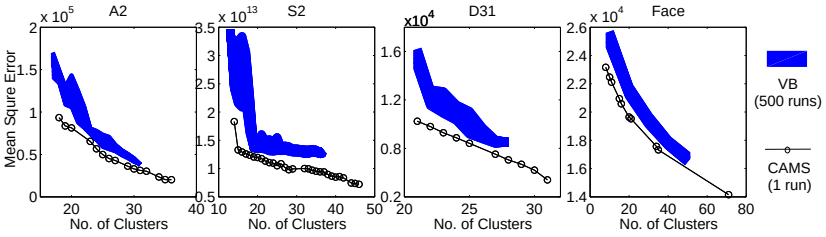
[5] S. Dasgupta and L. Schulman. A two-round variant of em for gaussian. In *Proc. 16th Conf. UAI*, pages 152–159, 2000.

[6] J. Feldman, R. Servedio, and R. O'Donnell. Pac learning axis-aligned mixture of gaussians with no separation assumption. In *COLT*, 2006.

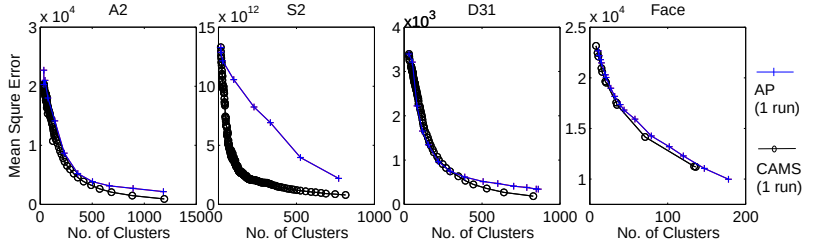
[7] B. Frey and D. Dueck. Mixture modelling by affinity propagation. In *In Advances in Neural Information Processing Systems 18*. MIT Press, 2005.



(a) CAMS v.s EM



(b) CAMS v.s VB



(c) CAMS v.s AP

Figure 5: Systematic evaluation results.

[8] B. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315:972–976, 2007.

[9] S. Ghosal and A. W. van Der Vaart. Entropies and rate of convergence for maximum likelihood and bayes estimation for mixtures of normal densities. *The annals of statistics*, 29:1233–1263, 2001.

[10] S. Jain and R. Neal. A split-merge markov chain monte carlo procedure for the dirichlet process mixture mode. *J. Comput. Graph. Stat.*, 13:158–182, 2004.

[11] E. T. Jaynes. *Probability Theory: The Logic of Science*. Unfinished manuscript, 1996.

[12] A. Kalai, A. Moitra, and G. Valiant. Efficiently learning mixture of two gaussians. In *STOC*, 2010.

[13] B. Kulis and M. Jordan. Revisiting k-means: New algorithms via bayesian nonparametrics. In *ICML*, 2012.

[14] D. Lashkari and P. Golland. Covex clustering with exemplar-based models. In *In Advances in Neural Information Processing Systems*. MIT Press, 2007.

[15] C. Liu and D. Rubin. The ecme algorithm: A simple extention of em and ecm with fast monotone convergence. *Biometrika*, 81:633–648, 1994.

[16] G. McLachlan and D. Peel. *Finite Mixture Models*. John Wiley and Sons, 2000.

[17] J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley, New York, 1997.

[18] A. Moitra and G. Valiant. Settling the polynomial learnability of mixture of gaussians. In *STOC*, 2010.

[19] S. Nowozin and G. Bakir. A decoupled approach to exemplar-based unsupervised learning. In *In ICML*, 2008.

[20] C. E. Rasmussen. The infinite gaussian mixture model. In *In Advances in Neural Information Processing Systems 12*, pages 554–560. MIT Press, 2000.

[21] S. Richardson and P. Green. On bayesian analysis of mixtures with an unknown nunmber of components. *Journal of the Royal Statistical Society*, 59:731–792, 1997.

[22] J. Sethuraman. A constructive definition of dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

[23] M. Stephens. Bayesian analysis of mixture models with an unknown number of components - an alternative to reversible jump methods. *The Annals of Statistics*, 28:40–74, 2000.

[24] M. Stephens. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society, Series B*, 62:795–809, 2000.

[25] V. N. Vapnik. *Statistical Learning Theory*. John Wiley & sons, 1998.

[26] D. Young. An overview of mixture models. *Statistics Surveys*, 0:1–24, 2008.

[27] Z. Zhang, B. T. Dai, and A. Tung. Estimating local optimums in em algorithm over gaussian mixture models. In *ICML*, 2008.

Supplemental File for Convex Approximation to Mixture Models Using Step Functions

February 10, 2013

$$
\begin{aligned}
Obj(\mathrm{a}_m) &= \{L(\mathrm{a}_m) - \lambda(\sum_{j=1}^{m} a_j - 1)\} \\
&= \{\sum_{i=1}^{n} \log \sum_{j=1}^{m} g(x_i|\boldsymbol{\theta}_j) a_j - \lambda(\sum_{j=1}^{m} a_j - 1)\} \qquad (1)
\end{aligned}
$$

Definition 1. *Given n observations $\{x_i\}_{i=1}^n$ generated from $G(x|a)$ and m constants $\{\boldsymbol{\theta}_j\}_{j=1}^m$ from step function approximation $\tilde{G}(x|a_m)$, define the matrix $\mathbf{G}^{m\times n}$ as $\mathbf{G}_{ji} = g(x_i|\boldsymbol{\theta}_j)$. Define the matrix $\mathbf{R}^{m\times n}$ as $\mathbf{R}_{ji} = \frac{g(x_i|\boldsymbol{\theta}_j)}{\sum_{l=1}^{m} g(x_i|\boldsymbol{\theta}_l)a_l}$*

Lemma 1. *(CONVEXITY) Objective functional Eq.* (1) *is concave down.*

Proof. To prove a function concave down, it is necessary and sufficient to prove that its Hessian (second derivative matrix) is negative semi definite. The second derivative matrix of Eq. (1) is

$$
\frac{\partial Obj}{\partial a_j \partial a_l} = -\sum_{i=1}^{n} \frac{g(x_i|\boldsymbol{\theta}_j) g(x_i|\boldsymbol{\theta}_l)}{(\sum_{k=1}^{m} g(x_i|\boldsymbol{\theta}_k) a_k)^2} = -(\mathbf{R}\mathbf{R}^T)_{jl}
$$

which means that the Hessian matrix $H = -\mathbf{R}\mathbf{R}^T$ is indeed negative semi definite. Therefore, Eq. (1) is a concave down function. □

Lemma 2. *Using the standard gradient ascent method yields the following converging updating rules for $\{a_j\}_{j=1}^m$:*

$$
\begin{aligned}
a_j^{update} &\leftarrow a_j + \Delta t \cdot (\sum_{i=1}^{n} r(i,j) - \lambda) \\
\lambda &= \frac{1}{m} \sum_{j=1}^{m} \sum_{i=1}^{n} r(i,j) \qquad (2)
\end{aligned}
$$

where

$$
r(i,j) = \frac{g(x_i|\boldsymbol{\theta}_j)}{\sum_{l=1}^{m} g(x_i|\boldsymbol{\theta}_l) a_l} \qquad (3)
$$

Proof. The derivative of the objective functional w.r.t. a_j is

$$\frac{\partial Obj}{\partial a_j} = \sum_{i=1}^{n} \frac{g(x_i|\boldsymbol{\theta}_j)}{\sum_{l=1}^{m} g(x_i|\boldsymbol{\theta}_l)a_l} - \lambda$$

The gradient ascent method then gives the converging updating rule:

$$\begin{aligned} a_j^{update} &\leftarrow a_j + \Delta t \cdot \frac{\partial F}{\partial a_j} \\ &= a_j + \Delta t \cdot (\sum_{i=1}^{n} r(i,j) - \lambda) \end{aligned} \tag{4}$$

where

$$r(i,j) = \frac{g(x_i|\boldsymbol{\theta}_j)}{\sum_{l=1}^{m} g(x_i|\boldsymbol{\theta}_l)a_l} \tag{5}$$

Since the objective functional F is convex, the gradient ascent method converges to the global optimal solution. By setting $\frac{\partial F}{\partial a_j} = 0$, we obtain λ:

$$\lambda = \sum_{i=1}^{n} r(i,j)$$

Since this equation holds for all a_j,$j = 1...m$, we finally achieve:

$$\lambda = \frac{1}{m} \sum_{j=1}^{m} \sum_{i=1}^{n} r(i,j)$$

□

Lemma 3. *Let $H_j^{x_i}$ denotes the event $\theta_i \in A_j$ where θ_i is the unknown parameter that generates x_i. Then we have:*

$$P(H_j^{x_i}|\tilde{G}(x|\mathrm{a}_m), x_i) = a_j r(i,j). \tag{6}$$

Proof. By using Bayesian rule, the probability $P(H_j^{x_i}|x_i, \tilde{G}(x|\mathrm{a}_m))$ can be expanded by:

$$\begin{aligned} &P(H_j^{x_i}|x_i, \tilde{G}(x|\mathrm{a}_m)) \\ &= \frac{P(x_i|H_j^{x_i}, \tilde{G}(x|\mathrm{a}_m))P(H_j^{x_i}|\tilde{G}(x|\mathrm{a}_m))}{P(x_i|\tilde{G}(x|\mathrm{a}_m))} \end{aligned} \tag{7}$$

where the denominator $P(x_i|\tilde{G}(x|\mathrm{a}_m)) = \sum_{j=1}^{m} P(x_i|H_j^{x_i}, \tilde{G}(x|\mathrm{a}_m))P(H_j^{x_i}|\{a_l\}_{l=1}^{m}, g)$.We compute the three terms on the right hand side one by one.

Since the hypothesis that x_j is generated by $\boldsymbol{\theta}_j$ implies $1 > a_j > 0$, we have

$$\begin{aligned} P(x_i|H_j^{x_i}, \tilde{G}(x|\mathrm{a}_m)) &= P(x_i|H_j^{x_i}, a_j, g) \\ &= P(x_i|H_j^{x_i}, g) = g(x_i|\boldsymbol{\theta}_j) \end{aligned} \tag{8}$$

Here we use the property of probability that if event A implies event B, that is $P(AB) = P(A)$ then $P(C|AB) = P(C|A)$.

$P(H_j^{x_i}|\tilde{G}(x|\mathrm{a}_m))$ is equal to $P(H_j^{x_i}|a_j, g)$, since if a_j is given, other a_l's give no more information to $H_j^{x_i}$. We first consider the two extreme cases. If $a_j = 1$, then with probability one we have $P(H_j^{x_i}|a_j, g) = 1$; if $a_j = 0$, then with probability one we have $P(H_j^{x_i}|a_j, g) = 0$. Moreover, the probability of the hypothesis $H_j^{x_i}$ should be proportional to the value of the given a_j. The only mathematical form of $P(H_j^{x_i}|a_j, g)$ that satisfies all these conditions is

$$P(H_j^{x_i}|a_j, g) = a_j \tag{9}$$

Computing the denominator $P(x_i|\tilde{G}(x|\mathrm{a}_m))$ now is easy using the above results.

$$\begin{aligned} & P(x_i|\tilde{G}(x|\mathrm{a}_m)) \\ &= \sum_{j=1}^{m} P(x_i, H_j^{x_i}|\tilde{G}(x|\mathrm{a}_m)) \\ &= \sum_{j=1}^{m} P(x_i|H_j^{x_i}, \tilde{G}(x|\mathrm{a}_m)) P(H_j^{x_i}|\tilde{G}(x|\mathrm{a}_m)) \\ &= \sum_{j=1}^{m} g(x_i|\boldsymbol{\theta}_j) a_j \end{aligned} \tag{10}$$

Putting the results of the three terms together we achieve:

$$\begin{aligned} P(H_j^{x_i}|x_i, \tilde{G}(x|\mathrm{a}_m)) &= \frac{g(x_i|\boldsymbol{\theta}_j) a_j}{\sum_{l=1}^{m} g(x_i|\boldsymbol{\theta}_l) a_l} \\ &= a_j r(i,j) \end{aligned} \tag{11}$$

This completes the proof.

□

Theorem 1. *Given n observations generated from $G(x|\mathrm{a})$, for any step function approximation $\tilde{G}(x|\mathrm{a}_m)$ with $m >> n$, there exists a step function approximation $\tilde{G}(x|\mathrm{a}_{m'})$, such that $m' \leq n$ and*

$$\begin{aligned} & \sup_{\{a_j | a_j \geq 0, j=1...m; \sum_{j=1}^{m} a_j = 1\}} \{L(\mathrm{a}_m)\} \\ = & \sup_{\{a'_j | a'_j \geq 0, j=1...m'; \sum_{j=1}^{m'} a'_j = 1\}} \{L(\mathrm{a}_{m'})\} \end{aligned} \tag{12}$$

Moreover, m' is the rank of $\mathbf{G}^{m\times n}$ from a_m. Let $\boldsymbol{\theta}_j, j = 1...m'$ form the m' linear independent rows of $\mathbf{G}^{m\times n}$; The constants $\{\boldsymbol{\theta}'_j\}_{j=1}^{m'}$ in $\mathrm{a}_{m'}$ can be determined by setting $\boldsymbol{\theta}'_j = \boldsymbol{\theta}_j, j = 1...m'$.

Proof. Given the convexity of the objective function Eq. (1), letting $\{a_j^*\}_{j=1}^m$ be the solution to maximizing Eq. (1), the following equation holds for $\{a_j^*\}_{j=1}^m$:

$$\frac{\partial Obj}{\partial a_j}\Big|_{a_j=a_j^*} = \sum_{i=1}^{n} \frac{g(x_i|\boldsymbol{\theta}_j)}{\sum_{l=1}^{m} g(x_i|\boldsymbol{\theta}_l)a_l^*} - \lambda \equiv 0 \tag{13}$$

To rewrite the above equation equivalently in a matrix form, we have:

$$\mathbf{R}^{*m\times n}\mathbf{1}^{n\times 1} \equiv \lambda \mathbf{1}^{m\times 1} \tag{14}$$

where $\mathbf{R}^*_{ji} = \frac{g(x_i|\boldsymbol{\theta}_j)}{\sum_{l=1}^{m} g(x_i|\boldsymbol{\theta}_l)a_l^*}$.

Since $m > n$, the m rows of $\mathbf{R}^*$ must be linearly dependent. Therefore, the rank m' of the row vectors in $\mathbf{R}^*$ must satisfy $m' \leq n$. One can assume that the first m' rows of $\mathbf{R}^*$ to be linearly independent, and we call them $\mathbf{R}_0^*$. Consequently, there exists a row transformation matrix $\mathbf{T}$ such that

$$\mathbf{T}^{m\times m}\mathbf{R}^{*m\times n} = \begin{bmatrix} \mathbf{R}_0^{*m'\times n} \\ \mathbf{0}^{(m-m')\times n} \end{bmatrix} \tag{15}$$

where $\mathbf{T}$ bears the form

$$\begin{bmatrix} \mathbf{I}^{m'\times m'} & \mathbf{0}^{m'\times(m-m')} \\ \mathbf{T}_0^{(m-m')\times m'} & -\mathbf{I}^{(m-m')\times(m-m')} \end{bmatrix}$$

By left multiplying $\mathbf{T}$ to both sides of Eq. (14), we obtain

$$\begin{bmatrix} \mathbf{R}_0^* \\ \mathbf{0} \end{bmatrix} \mathbf{1}^{n\times 1} \equiv \lambda \begin{bmatrix} \mathbf{1} \\ \mathbf{B} \end{bmatrix} \tag{16}$$

where $\mathbf{B} = \begin{bmatrix} \mathbf{T}_0 & \mathbf{I} \end{bmatrix} \times \mathbf{1}^{m\times 1}$

Since Eq. (14) must hold for $\{a_j^*\}_{j=1}^m$, $\mathbf{B} = \mathbf{0}$ must hold in Eq. (16), or equivalently,

$$\mathbf{T}_0\mathbf{1}^{m'\times 1} = \mathbf{1}^{(m-m')\times 1} \tag{17}$$

By omitting the zero rows on both sides of Eq. (16), we finally obtain

$$\mathbf{R}_0^{*m'\times n}\mathbf{1}^{n\times 1} \equiv \lambda \mathbf{1}^{m'\times 1} \tag{18}$$

Eq. (14) and Eq. (18) are equivalent. However, Eq. (18) can be understood as the necessary and sufficient condition for the solution to the objective functional Eq. (1) of some step function $\mathrm{a}_{m'}$. In the following, we are looking for the step function $\mathrm{a}_{m'}$ whose maixmum likelihood estimation $\{a_j'^*\}_{j=1}^{m'}$ yields $\mathbf{R}_0^*$ in Eq. (18).

Using E. (15) and Eq. (17), we have

$$\begin{aligned}
\sup_{\{a_j|a_j\geq 0\}_{j=1}^{m}}\{Obj(\mathrm{a}_m)\} &= \sum_{i=1}^{n}\log\sum_{j=1}^{m} g(x_i|\boldsymbol{\theta}_j)a_j^* - \lambda(\sum_{j=1}^{m} a_j^* - 1) \\
&= \sum_{i=1}^{n}\log\big(\sum_{j=1}^{m'} g(x_i|\boldsymbol{\theta}_j)a_j^* + \sum_{k=m'+1}^{m} a_k^*\sum_{j=1}^{m'} g(x_i|\boldsymbol{\theta}_j)\mathbf{T}_{kj}\big) \\
&+ \lambda\big(\sum_{j=1}^{m'} a_j^* + \sum_{k=m'+1}^{m} a_k^*\sum_{j=1}^{m'}\mathbf{T}_{kj} - 1\big) \\
&= \sum_{i=1}^{n}\log\sum_{j=1}^{m'} g(x_i|\boldsymbol{\theta}_j)\big(a_j^* + \sum_{k=m'+1}^{m}\mathbf{T}_{kj}a_k^*\big) \\
&+ \lambda\Big(\sum_{j=1}^{m'}\big(a_j^* + \sum_{k=m'+1}^{m}\mathbf{T}_{kj}a_k^*\big) - 1\Big) \qquad (19)
\end{aligned}$$

Let $a_j'^* = \big(a_j^* + \sum_{k=m'+1}^{m}\mathbf{T}_{kj}a_k^*\big), j = 1...m'$ and $\boldsymbol{\theta}_j' = \boldsymbol{\theta}_j, j = 1...m'$, we obtain that

$$\begin{aligned}
\sup_{\{a_j|a_j\geq 0, j=1...m; \sum_{j=1}^{m} a_j=1\}}\{L(\mathrm{a}_m)\} &= \sum_{i=1}^{n}\log\sum_{j=1}^{m} g(x_i|\boldsymbol{\theta}_j)a_j^* \\
&= \sum_{i=1}^{n}\log\sum_{j=1}^{m'} g(x_i|\boldsymbol{\theta}_j)\big(a_j^* + \sum_{k=m'+1}^{m}\mathbf{T}_{kj}a_k^*\big) \\
&= \sum_{i=1}^{n}\log\sum_{j=1}^{m'} g(x_i|\boldsymbol{\theta}_j')a_j'^* \\
&= \sup_{\{a_j'|a_j'\geq 0, j=1...m'; \sum_{j=1}^{m'} a_j'=1\}}\{L(\mathrm{a}_{m'})\} \qquad (20)
\end{aligned}$$

Obviously $\mathbf{R}^{*m'\times n}$ constructed by $a_j'^*$ and $\boldsymbol{\theta}_j'$, $j = 1...m'$, equals $\mathbf{R}_0^*$ in Eq. (18). This completes the proof of the theorem. □

Theorem 2. *For any fixed x, the maximum likelihood step function estimator $\hat{\tilde{G}}(x|\hat{\mathrm{a}}_m)$, where $m = \mathcal{O}(n)$, is a consistent estimator of the mixture model $G(x|\mathrm{a})$, i.e.,*

$$\hat{\tilde{G}}(x|\hat{\mathrm{a}}_m) \xrightarrow{p} G(x|\mathrm{a}), \ \textit{as } n\to\infty. \qquad (21)$$

Suppose the whole parameter space Θ has a finite measure and $h(x,\theta) = g(x,\theta)\mathrm{a}(\theta)$ satisfies the Holder condition with $\alpha = 1/2$ respect to θ, that is, $\exists K > 0, s.t. |h(x,\theta_1) - h(x,\theta_2)| \leq K|\theta_1 - \theta_2|^{1/2}$, for any x and $\theta_1, \theta_2 \in \Theta$, the following convergence rate for $\hat{\tilde{G}}(x|\hat{\mathrm{a}}_m)$ holds:

$$\sup_{x\in\mathbb{R}}|\hat{\tilde{G}}(x|\hat{\mathrm{a}}_m) - G(x|\mathrm{a})| = \mathcal{O}_p(1/\sqrt{n}) \qquad (22)$$

In order to prove Theorem 2, we first prove Theorem 3,Theorem 4 and Theorem 5

To avoid the cluster of mathematic symbols, below we use a to stand for a_m

Theorem 3. *The Maximum Likelihood Estimation* $(\hat{a}_1, \hat{a}_2, ..., \hat{a}_m)^T$ *is consistent, i.e.* $(\hat{a}_1, \hat{a}_2, ..., \hat{a}_m)^T \xrightarrow{a.s.} (a_1, a_2, ..., a_m)^T$. *Moreover, for any fixed* x, $\hat{\tilde{G}}(x|\hat{a})$ *is a consistent estimator of* $\tilde{G}(x|a)$.

Proof. In fact, we just need to prove the first part of the theorem. The second part follows obviously because $\hat{\tilde{G}}(x|\hat{a})$ is a finite linear combination of consistent estimators.
In order to prove that $(\hat{a}_1, \hat{a}_2, ..., \hat{a}_m)^T$ is a consistent estimator, we need to introduce the following lemma.

Lemma 4. *Let* $< \Omega, \mho, P >$ *and* $< X, \wp, F >$ *be two probalility spaces and* $X_1, X_2, ...$ *be measurable functions from* Ω *into* X *which are independent with the common distribution* F. *Let* Θ *be an open subset of* $\mathbb{R}^k$ *and* $\theta \in \Theta$. *Let* ψ *be a map from* $X \times \Theta$ *into* $\mathbb{R}$ *satisfying the following conditions:*
(A1) For every $\vartheta \in \Theta$, *the map* $x \to \psi(x, \vartheta)$ *is measuable.*
(A2) For each $x \in X$, *the map* $\vartheta \to \psi(x, \vartheta)$ *is continuous.*
(A3) For each $\vartheta \in \Theta$, *there is an* F*-integrable function* H_ϑ *and a positive* ϵ_ϑ *such that* $\psi(x, t) \leq H_\vartheta(x)$ *for all* $x \in X$ *and* $t \in \Theta$ *with* $\| t - \vartheta \| < \epsilon_\vartheta$.
(A4) The map μ *from* Θ *into* $[-\infty, \infty)$ *defined by* $\mu(\vartheta) = \int \psi(x, \vartheta) dF(x), \vartheta \in \Theta$ *is uniquely maximized by* θ.
(A5) The map $\vartheta \to \psi(x, \vartheta)$ *is concave down for every* $x \in X$. *Set*

$$\Psi_n(\vartheta) = \frac{1}{n} \sum_{j=1}^{n} \psi(X_j, \vartheta), \qquad \vartheta \in \Theta$$

$\langle \hat{\theta}_n \rangle$ *is a* Θ*-valued random variable such that* $\Psi(\hat{\theta}_n) = \sup_{\vartheta \in \Theta} \Psi_n(\vartheta)$.
Then, $\hat{\theta}_n \xrightarrow{a.s.} \theta$.

Let now v be a measure on $\wp$ and $\Im = \{\tilde{G}(x|\alpha) = \sum_{j=1}^{m} g(x|\theta_j)\alpha_j : \alpha_j \in (0, 1], j = 1, 2, ..., m, \sum_{j=1}^{m} \alpha_j = 1\}$ be a class of v-densities. Then we can easily see the identity of the density. That is, $v(f_{\alpha_1} \neq f_{\alpha_2}) > 0$ whenever $\alpha_1 \neq \alpha_2$.
Once applying the above lemma with $\psi(x, \alpha) = \log \tilde{G}(x|\alpha) - \log \tilde{G}(x|a)$ and $dF = \tilde{G}(x|a)dv$, where a denotes the true parameters, we can show that MLEs are consistent. Our job is to verify that conditions (A1)-(A5) are satisfied in our estimation.

First of all, it is easy to see that $\psi(x, \alpha) = \log \tilde{G}(x|\alpha) - \log \tilde{G}(x|a)$ satisfies (A1) and (A2).

For (A3), consider the function $H_\alpha(x) = |\log \tilde{G}(x|a)|$, which satisfies $E_a|H_a(x)| < \infty$ and $\psi(x, \alpha) = \log \tilde{G}(x|\alpha) - \log \tilde{G}(x|a) \leq H_a(x)$ for all appropriate a.

For (A4), by Dominated Convergence Theorem,

$$\frac{\partial^2 \mu(\alpha)}{\partial \alpha_i \partial \alpha_j} = -\int \frac{g(x|\theta_i)\, g(x|\theta_j)}{\left(\sum_{k=1}^{m} g(x|\theta_k)\alpha_k\right)^2} dF(x) \leq 0$$

Let $U = (\frac{g(x|\theta_1)}{(\sum_{k=1}^m g(x|\theta_k)\alpha_k)^2}, \frac{g(x|\theta_2)}{(\sum_{k=1}^m g(x|\theta_k)\alpha_k)^2}, ..., \frac{g(x|\theta_m)}{(\sum_{k=1}^m g(x|\theta_k)\alpha_k)^2})^T$, then the Hessian matrix of $\mu(\alpha)$, $H(\mu) = -\int U^T U$ is negative semi-definite. Therefore, $\mu(\alpha)$ is uniquely maximized by the MLE $\hat{a} = (\hat{a}_1, \hat{a}_2, ..., \hat{a}_m)^T$.

At last, (A5) can be verified by the convexity of $\psi(x, \alpha)$. It is necessary and sufficient to prove that its Hessian matrix is negative semi-definite.

$$H_{i,j} = \frac{\partial^2 \psi(\alpha)}{\partial \alpha_i \partial \alpha_j} = -\frac{g(x|\theta i)\, g(x|\theta j)}{(\sum_{k=1}^m g(x|\theta_k)\alpha_k)^2} \leq 0$$

We write $H(x, \alpha) = -U^T U$. Therefore, Hessian matrix $H(x, \alpha)$ is indeed negative semi-definite.

Thus, based on the previous lemma, the MLE $\hat{a} = (\hat{a}_1, \hat{a}_2, ..., \hat{a}_m)^T$ is a consistent estimator. Moreover,since $\tilde{G}(x|\alpha)$ is a linear combination of $\hat{a}_1, \hat{a}_2, ..., \hat{a}_m$ and $|\tilde{G}(x|\hat{a})| \leq 1$ is bounded, $\tilde{G}(x|\hat{a}) \xrightarrow{a.s.} \tilde{G}(x|a)$. This completes the proof of Theorem 3. □

Next, we want to identify the asymptotic distribution of MLE $\hat{a} = (\hat{a}_1, \hat{a}_2, ..., \hat{a}_m)^T$ and then induce the consistent rate. In oder to do this, we need the following lemma.

Lemma 5. *(Asymptotical efficiency of MLEs) Let $X_1, X_2, ...,$ be iid $f(x|\theta)$, let $\hat{\theta}$ denote the MLE of θ, and $\tau(\theta)$ be a continuous function of θ. Under the regularity conditions of $f(x|\theta)$ and, hence, $L(\theta|x)$,*

$$\sqrt{n}[\tau(\hat{\theta}) - \tau(\theta)] \longrightarrow N(0, v(\theta)). \tag{23}$$

where $v(\theta) = \dfrac{[\tau'(\theta)]^2}{E_\theta\left((\frac{\partial}{\partial\theta}(\log f(x|\theta))^2\right)}$ is the Cramér − Rao Lower Bound.

Proof. For details of the proof, see George Casella and Roger L. Berger (Statistical Inference, Chapter 10). □

With this Lemma, we can prove the following theorem.

Theorem 4. *Let $\hat{a}$ denote the MLE of a, and $\tilde{G}(x|a) = \sum_{j=1}^n g(x|a_j)a_j$ is a continous function of $a_1, a_2, ..., a_m$. Then, for any fixed x,*

$$\sqrt{n}[\hat{\tilde{G}}(x|\hat{a}) - \tilde{G}(x|a)] \longrightarrow N(0, g^T I^{-1}(a) g), \tag{24}$$

where $g = (g(x|\theta_1), g(x|\theta_2), ..., g(x|\theta_m))^T$. Therefore,

$$\sup_{x\in\mathbb{R}} |\hat{\tilde{G}}(x|\hat{a}) - \tilde{G}(x|a)| = \mathcal{O}_p(n^{-1/2}). \tag{25}$$

Proof. We know the density $\tilde{G}(x|a) = \sum_{j=1}^{n} g(x|a_j)a_j = g^T a$ is a linear cominations of $a_1, a_2, ..., a_m$. Obviously, $\tilde{G}(x|a)$ satisfies the regularity conditions in the Lemma. For details of regularity conditions see George Casella and Roger L. Berger (Statistical Inference, Chapter 10). From Lemma 5, we have

$$\sqrt{n}[\hat{\tilde{G}}(x|\hat{a}) - \tilde{G}(x|a)] \longrightarrow N(0, g_x^T I^{-1}(a) g_x), \tag{26}$$

for fixed x. Where $g_x = (g(x|a_1), g(x|a_2), ..., g(x|a_m))^T$ Thus, we get the rate $\sup_{x\in\mathbb{R}} |\hat{\tilde{G}}(x|\hat{a}) - \tilde{G}(x|a)| = \mathcal{O}_p(n^{-1/2})$. □

Theorem 5. *Suppose the whole parameter space Θ has a finite measure and $h(x,\theta) = g(x,\theta)a(\theta)$ satisfies the Holder condition with $\alpha = 1/2$ with respect to θ, that is, $\exists K > 0, s.t. |h(x,\theta_1) - h(x,\theta_2)| \leq K|\theta_1 - \theta_2|^{1/2}$, for any x and $\theta_1, \theta_2 \in \Theta$. Then, for any partition $\{A_1, A_2, ..., A_m\}$ of Θ, satisfying $\max_{1\leq i\leq m} l(A_i) = \mathcal{O}(1/m)$, we have*

$$\sup_{x\in\mathbb{R}} |\int_{\Theta} g(x|\theta)a(\theta)d\theta - \tilde{G}(x|a)| = \mathcal{O}(m^{-1/2}). \tag{27}$$

Proof. Let $\{A_1, A_2, ..., A_m\}$ be a partition such that $\max_{1\leq i\leq m} l(A_i) = \mathcal{O}(1/m)$.

$$\begin{aligned}
|\int_{\Theta} g(x|\theta)a(\theta)d\theta - \tilde{G}(x|a)| =& |\sum_{j=1}^{m} \int_{A_j} [g(x|\theta)a(\theta) - g(x|\theta_j)a(\theta_j)]d\theta| \\
\leq& \sum_{j=1}^{m} \int_{A_j} |g(x|\theta)a(\theta) - g(x|\theta_j)a(\theta_j)|d\theta \\
\leq& K * \sup_{1\leq i\leq m} l^{1/2}(A_i) * \mu(\Theta) \\
=& O(1/\sqrt{m})
\end{aligned}$$

Since the above inequality is true for all $x \in \mathbb{R}$, equation (25) is induced immediately. □

Now we are ready to prove Theorem 2.

Proof. Combining the results of Theorem 4 and Theorem5, we know that the difference between $G(x|\mathrm{a}(\theta)) = \int_{\Theta} g(x|\theta)\mathrm{a}(\theta)d\theta$ and $\tilde{G}(x|\mathrm{a}_m)$ can be controled by the number of partitions m. And the estimation error between the maximal likelihood estimator $\hat{\tilde{G}}(x|\hat{\mathrm{a}}_m)$ and $\tilde{G}(x|\mathrm{a}_m)$ can be controled by the sample size n. Thus,

$$\begin{aligned}
|G(x|\mathrm{a}) - \hat{\tilde{G}}(x|\hat{\mathrm{a}}_m)| \leq& |G(x|\mathrm{a}) - \tilde{G}(x|\mathrm{a}_m)| + |\tilde{G}(x|\mathrm{a}_m) - \hat{\tilde{G}}(x|\hat{\mathrm{a}}_m)| \\
=& \mathcal{O}(1/\sqrt{m}) + \mathcal{O}(1/\sqrt{n})
\end{aligned} \tag{28}$$

The rate of consistency between $G(x|\mathrm{a})$ and our MLE $\hat{\tilde{G}}(x|\hat{\mathrm{a}}_m)$ is $\mathcal{O}(1/\sqrt{m}) + \mathcal{O}(1/\sqrt{n})$. That makes sense, because the more partitions we made and the

more data we have,the better estimation we can have. Since $m = \mathcal{O}(n)$, we obtain

$$\sup_{x\in\mathbb{R}} |\hat{\tilde{G}}(x|\hat{\mathrm{a}}_m) - G(x|\mathrm{a})| = \mathcal{O}_p(1/\sqrt{n}) \tag{29}$$

□