Factorized Asymptotic Bayesian Inference for Mixture Modeling

Ryohei Fujimaki
NEC Laboratories America
Department of Media Analytics
rfujimaki@sv.nec-labs.com

Satoshi Morinaga
NEC Corporation
Information and Media Processing Research Laboratories
morinaga@cw.jp.nec.com

Abstract

This paper proposes a novel Bayesian approximation inference method for mixture modeling. Our key idea is to factorize marginal log-likelihood using a variational distribution over latent variables. An asymptotic approximation, a factorized information criterion (FIC), is obtained by applying the Laplace method to each of the factorized components. In order to evaluate FIC, we propose factorized asymptotic Bayesian inference (FAB), which maximizes an asymptotically-consistent lower bound of FIC. FIC and FAB have several desirable properties: 1) asymptotic consistency with the marginal log-likelihood, 2) automatic component selection on the basis of an intrinsic shrinkage mechanism, and 3) parameter identifiability in mixture modeling. Experimental results show that FAB outperforms state-of-the-art VB methods.

1 Introduction

Model selection is one of the most difficult and important challenges in machine learning problems, in which we optimize a model representation as well as model parameters. Bayesian learning provides a natural and sophisticated way to address the issue [1, 3]. A central task in Bayesian model selection is evaluation of marginal log-likelihood. Since exact evaluation is often computationally and analytically intractable, a number of approximation algorithms have been studied.

One well-known precursor is Bayes information criterion (BIC) [18] , an asymptotic second-order approximation using the Laplace method. Since BIC assumes *regularity conditions* that ensure the asymptotic normality of the maximum likelihood (ML) estimator, it loses its theoretical justification for *non-regular* models, including mixture models, hidden Markov models, multi-layer neural networks, etc. Further, it is known that the ML estimation does not have a unique solution for *non-identifiable* models in which the mapping between parameters and functions is not one-to-one [22, 23] (such equivalent models are said to be in an *equivalent class*). For such non-singular and non-identifiable models, the generalization error of the ML estimator significantly degrades [24].

Among recent advanced Bayesian methods, we focus on variational Bayesian (VB) inference. A VB method maximizes a computationally-tractable lower bound of the marginal log-likelihood by applying variational distributions on latent variables and parameters. Previous studies have investigated VB methods for a variety of models [2, 6, 5, 12, 15, 16, 20] and have demonstrated their practicality. Further, theoretical studies [17, 21] have shown that VB methods can resolve the non-regularity and non-identifiability issues. One disadvantage, however, is that latent variables and model parameters are partitioned into sub-groups that are independent of each other in variational distributions. While this independence makes computation tractable, the lower bound can be loose since their dependency is essential in the true distribution.

This paper proposes a novel Bayesian approximation inference method for a family of mixture models which is one of the most significant example of non-regular and non-identifiable model families. Our ideas and contributions are summarized as follows:

Factorized Information Criterion We derive a new approximation of marginal log-likelihood and refer to it as a factorized information criterion (FIC). A key observation is that, using a variational distribution over latent variables, the marginal log-likelihood can

Appearing in Proceedings of the 15^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume XX of JMLR: W&CP XX.

be rewritten as a factorized representation in which the Laplace approximation is applicable to each of the factorized components. FIC is unique in the sense that it takes into account dependencies among latent variables and parameters, and FIC is asymptotically consistent with the marginal log-likelihood.

Factorized Asymptotic Bayesian Inference We propose a new approximation inference algorithm and refer to it as a factorized asymptotic Bayesian inference (FAB). Since FIC is defined on the basis of both observed and latent variables, the latter of which are not available, FAB maximizes a lower bound of FIC through an iterative optimization similar to the EM algorithm [8] and VB methods [2]. This FAB optimization is proved to monotonically increase the lower bound of FIC. It is worth noting that FAB provides a natural way to control model complexity not only in terms of the number of components but also in terms of the types of individual components (e.g., degrees of components in a polynomial curve mixture (PCM).)

FIC and FAB offer the following desirable properties:

1. The FIC lower bound is asymptotically consistent with FIC, and therefore with the marginal log-likelihood.
2. FAB automatically selects components on the basis of its intrinsic shrinkage mechanism different from prior-based regularization. FAB therefore mitigates overfitting even if it ignores the prior effect in the asymptotic sense, or even if we apply a non-informative prior [13].
3. FAB addresses the non-identifiability issue. In an equivalent class, FAB automatically selects the model which maximizes the entropy of distributions for latent variables.

2 Preliminaries

This paper considers mixture models $p(X|\boldsymbol{\theta}) = \sum_{c=1}^{C} \alpha_c p_c(X|\boldsymbol{\phi}_c)$ for a D-dimensional random variable X. C and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_C)$ are the number of components and the mixture ratio. $\boldsymbol{\theta}$ is $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_C)$. We allow different components $p_c(X|\boldsymbol{\phi}_c)$ to be different in their model representations from one another[1]. We assume that $p_c(X|\boldsymbol{\phi}_c)$ satisfies the regularity conditions[2] while $p(X|\boldsymbol{\theta})$ is non-regular (**A1**). Assumption **A1** is less stringent than the regularity conditions on $p(X|\boldsymbol{\theta})$, and many models (e.g., Gaussian mixture models (GMM), PCM, autoregressive mixture models) satisfy **A1**.

A model M of $p(X|\boldsymbol{\theta})$ is specified by C and models S_c of $p_c(X|\boldsymbol{\phi}_c)$, i.e., M as $M = (C, S_1, \ldots, S_C)$. Although the representations of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}_c$ depend on M and S_c, respectively, we omit them for notational simplicity.

Let us denote a latent variable of X as $Z = (Z_1, \ldots, Z_C)$. Z is a component assignment vector, and $Z_c = 1$ if X is generated from the c-th component, and $Z_c = 0$ otherwise. The marginal distribution on Z and the conditional distribution on X given Z can be described as $p(Z|\boldsymbol{\alpha}) = \prod_{c=1}^{C} \alpha_c^{Z_c}$ and $p(X|Z, \boldsymbol{\phi}_c) = \prod_{c=1}^{C} (p_c(X|\boldsymbol{\phi}_c))^{Z_c}$. The observed data and their latent variables are denoted as $\boldsymbol{x}^N = \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ and $\boldsymbol{z}^N = \boldsymbol{z}_1, \ldots, \boldsymbol{z}_N$, respectively, where $\boldsymbol{z}_n = (z_{n1}, \ldots, z_{nC})$ and $\boldsymbol{z}_c^N = (z_{1c}, \ldots, z_{Nc})$. We make another assumption that $\log P(X, Z|\boldsymbol{\theta}) < \infty$ holds (**A2**). Condition **A2** is discussed in Section 4.5.

3 Factorized Information Criterion for Mixture Models

A Bayesian selects the model which maximizes the model posterior $p(M|\boldsymbol{x}^N) \propto p(M)p(\boldsymbol{x}^N|M)$. With uniform model prior $p(M)$, we are particularly interested in $p(\boldsymbol{x}^N|M)$, which is referred to as marginal likelihood or Bayesian evidence. VB methods for latent variable models [2, 5, 6, 21] consider a lower bound of the marginal log-likelihood to be:

$$\log p(\boldsymbol{x}^N|M) \geq \sum_{\boldsymbol{z}^N} \int q(\boldsymbol{z}^N) q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \log \frac{p(\boldsymbol{x}^N|\boldsymbol{\theta}) p(\boldsymbol{\theta}|M)}{q(\boldsymbol{z}^N) q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} d\boldsymbol{\theta}, \tag{1}$$

where q and $q_{\boldsymbol{\theta}}$ are variational distributions on $\boldsymbol{z}^N$ and $\boldsymbol{\theta}$, respectively. On q and $q_{\boldsymbol{\theta}}$, $\boldsymbol{z}^N$ and $\boldsymbol{\theta}$ are assumed to be independent of each other in order to make the lower bound computationally and analytically tractable. This ignores, however, significant dependency between $\boldsymbol{z}^N$ and $\boldsymbol{\theta}$, and basically the equality does not hold.

In contrast to this, we consider the lower bound on $q(\boldsymbol{z}^N)$ to be:

$$\log p(\boldsymbol{x}^N|M) \geq \sum_{\boldsymbol{z}^N} q(\boldsymbol{z}^N) \log\Big(\frac{p(\boldsymbol{x}^N, \boldsymbol{z}^N|M)}{q(\boldsymbol{z}^N)}\Big) \tag{2}$$

$$\equiv \mathcal{VLB}(q, \boldsymbol{x}^N, M), \tag{3}$$

Lemma 1 [2] guarantees that $\max_q\{\mathcal{VLB}(q, \boldsymbol{x}^N, M)\}$ is exactly consistent with $\log p(\boldsymbol{x}^N|M)$.

Lemma 1 *The inequality (3) holds for an arbitrary distribution q on $\boldsymbol{z}^N$, and the equality is satisfied by $q(\boldsymbol{z}^N) = p(\boldsymbol{z}^N|M, \boldsymbol{x}^N)$.*

[1] So-called "heterogeneous mixture models" [11]. In the example of PCM, the degree of $p_1(X|\boldsymbol{\phi}_1)$ will be two while that of $p_2(X|\boldsymbol{\phi}_2)$ will one.

[2] The Fisher information matrix of $p_c(X|\boldsymbol{\phi}_c)$ is non-singular around the maximum likelihood estimator.

Since this paper handles mixture models, we further assume mutual independence of $\boldsymbol{z}^N$, i.e. $q(\boldsymbol{z}^N) = \prod_{c=1}^{C} q(\boldsymbol{z}_c^N)$ and $q(\boldsymbol{z}_c^N) = \prod_{n=1}^{N} q(z_{nc})^{z_{nc}}$. The lower bound (2) is the same with that the collapse variational Bayesian (CVB) method [15, 20] considers. A key difference is that FAB employs an asymptotic second approximation of (2), which produces several desirable properties of FAB which we have described in Section 1, while CVB methods employ second oder approximations of variational parameters in each iterative update step.

Note that the numerator of (3) has the form of the parameter integration of $p(\boldsymbol{x}^N, \boldsymbol{z}^N|M)$, as:

$$p(\boldsymbol{x}^N, \boldsymbol{z}^N|M) = \int p(\boldsymbol{z}^N|\boldsymbol{\alpha}) \prod_{c=1}^{C} p_c(\boldsymbol{x}^N|\boldsymbol{z}_c^N, \boldsymbol{\phi}_c) p(\boldsymbol{\theta}|M) d\boldsymbol{\theta}. \tag{4}$$

A key idea in FIC is to apply the Laplace method to the individual factorized distributions $p(\boldsymbol{z}^N|\boldsymbol{\alpha})$ and $p_c(\boldsymbol{x}^N|\boldsymbol{z}_c^N, \boldsymbol{\phi}_c)$ as follows[3]:

$$\log p(\boldsymbol{x}^N,\boldsymbol{z}^N|\boldsymbol{\theta}) \approx \log p(\boldsymbol{x}^N, \boldsymbol{z}^N|\bar{\boldsymbol{\theta}}) - \frac{N}{2}\Big[\bar{\mathcal{F}}_Z, (\boldsymbol{\alpha} - \bar{\boldsymbol{\alpha}})\Big] - \sum_{c=1}^{C} \frac{\sum_{n=1}^{N} z_{nc}}{2}\Big[\bar{\mathcal{F}}_c, (\boldsymbol{\phi}_c - \bar{\boldsymbol{\phi}}_c)\Big], \tag{5}$$

where $[A, a]$ represents the quadratic form $a^T A a$ for a matrix A and a vector a. Here we denote the ML estimator of $\log p(\boldsymbol{x}^N, \boldsymbol{z}^N|\boldsymbol{\theta})$ as $\bar{\theta} = (\bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{\phi}}_1, \ldots, \bar{\boldsymbol{\phi}}_C)$. We discuss application of the Laplace method around the maximum a priori (MAP) estimator in Section 4.5. $\bar{\mathcal{F}}_Z$ and $\bar{\mathcal{F}}_c$ are factorized sample approximations of Fisher information matrices defined as follows:

$$\bar{\mathcal{F}}_Z = -\frac{1}{N} \frac{\partial^2 \log p(\boldsymbol{z}^N|\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^T}\Big|_{\boldsymbol{\alpha}=\bar{\boldsymbol{\alpha}}}, \tag{6}$$

$$\bar{\mathcal{F}}_c = \frac{-1}{\sum_{n=1}^{N} z_{nc}} \frac{\partial^2 \log p_c(\boldsymbol{x}^N|\boldsymbol{z}_c^N, \boldsymbol{\phi}_c)}{\partial \boldsymbol{\phi}_c \partial \boldsymbol{\phi}_c^T}\Big|_{\boldsymbol{\phi}_c=\bar{\boldsymbol{\phi}}_c}.$$

The following lemma is satisfied for $\bar{\mathcal{F}}_Z$ and $\bar{\mathcal{F}}_c$.

Lemma 2 *With $N \to \infty$, $\bar{\mathcal{F}}_Z$ and $\bar{\mathcal{F}}_c$ respectively converge to the Fisher information matrices described as:*

$$\mathcal{F}_Z = -\int \frac{\partial^2 \log p(Z|\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^T} p(Z|\boldsymbol{\alpha}) dZ\Big|_{\boldsymbol{\alpha}=\bar{\boldsymbol{\alpha}}}, \tag{7}$$

$$\mathcal{F}_c = -\int \frac{\partial^2 \log p_c(X|\boldsymbol{\phi}_c)}{\partial \boldsymbol{\phi}_c \partial \boldsymbol{\phi}_c^T} p_c(X|\boldsymbol{\phi}_c) dX\Big|_{\boldsymbol{\phi}_c=\bar{\boldsymbol{\phi}}_c}. \tag{8}$$

The proof follows the law of large numbers by taking into account that the effective number of samples for the c-th component is $\sum_{n=1}^{N} z_{nc}$.

[3]The Laplace approximation is not applicable to $p(\boldsymbol{x}^N|\boldsymbol{\theta})$ because of its non-regularity.

Then, by applying a prior of $\log p(\boldsymbol{\theta}|M) = \mathcal{O}(1)$, $p(\boldsymbol{x}^N, \boldsymbol{z}^N|M)$ can be asymptotically approximated as:

$$p(\boldsymbol{x}^N, \boldsymbol{z}^N|M) \approx p(\boldsymbol{x}^N, \boldsymbol{z}^N|\bar{\boldsymbol{\theta}}) \frac{(2\pi)^{\mathcal{D}_{\boldsymbol{\alpha}}/2}}{N^{\mathcal{D}_{\boldsymbol{\alpha}}/2}|\bar{\mathcal{F}}_Z|^{1/2}} \times \prod_{c=1}^{C} \frac{(2\pi)^{\mathcal{D}_c/2}}{(\sum_{n=1}^{N} z_{nc})^{\mathcal{D}_c/2}|\bar{\mathcal{F}}_c|^{1/2}} \tag{9}$$

Here, $\mathcal{D}_{\boldsymbol{\alpha}} \equiv \mathcal{D}(\boldsymbol{\alpha}) = C - 1$ and $\mathcal{D}_c \equiv \mathcal{D}(\boldsymbol{\phi}_c)$, in which $\mathcal{D}(\bullet)$ is the dimensionality of $\bullet$.

On the basis of **A1** and Lemma 2, both $\log|\bar{\mathcal{F}}_c|^{1/2}$ and $\log|\bar{\mathcal{F}}_Z|^{1/2}$ are $\mathcal{O}(1)$. By substituting (9) into (3) and ignoring asymptotically small terms, we obtain an asymptotic approximation of $\log p(\boldsymbol{x}^N|M) = \max_q\{\mathcal{VLB}(q, \boldsymbol{x}^N, M)\}$ (i.e., FIC) as follows:

$$FIC(\boldsymbol{x}^N, M) = \max_q \Big\{\mathcal{J}(q, \bar{\boldsymbol{\theta}}, \boldsymbol{x}^N)\Big\} \tag{10}$$

$$\mathcal{J}(q, \bar{\boldsymbol{\theta}}, \boldsymbol{x}^N) = \sum_{\boldsymbol{z}^N} q(\boldsymbol{z}^N)\Big(\log p(\boldsymbol{x}^N, \boldsymbol{z}^N|\bar{\boldsymbol{\theta}}) - \frac{\mathcal{D}_{\boldsymbol{\alpha}}}{2}\log N - \sum_{c=1}^{C} \frac{\mathcal{D}_c}{2}\log(\sum_{n=1}^{N} z_{nc}) - \log q(\boldsymbol{z}^N)\Big)\Big\}$$

The following theorem justifies FIC as an approximation of the marginal log-likelihood.

Theorem 3 *$FIC(\boldsymbol{x}^N, M)$ is asymptotically consistent with $\log p(\boldsymbol{x}^N|M)$ under* **A1** *and $\log p(\boldsymbol{\theta}|M) = \mathcal{O}(1)$.*

Sketch of Proof 3 *Since both $p_c(\boldsymbol{x}^N|\boldsymbol{z}_c^N, \boldsymbol{\phi}_c)$ and $p(\boldsymbol{z}^N|\boldsymbol{\alpha})$ satisfy the regularity condition (***A1** *and* **A2***), their Laplace approximations appearing in the left side of (9) have asymptotic consistency (for the same reason as with BIC [7, 18, 19]). Therefore, their product (9) asymptotically agrees with $p(\boldsymbol{x}^N, \boldsymbol{z}^N|M)$. By applying Lemma 1, this theorem is proved.*

Despite our ignoring the prior effect, FIC still has the regularization term $\mathcal{D}_c \sum_{\boldsymbol{z}_c^N} q(\boldsymbol{z}_c^N) \log(\sum_{n=1}^{N} z_{nc})/2$. This term has several interesting properties. First, a dependency between $\boldsymbol{z}^N$ and $\boldsymbol{\theta}$, which most of VB methods ignore, explicitly appears. Like BIC, it is not the parameter or function representation of $p(X|\boldsymbol{\phi}_c)$ but only its dimensionality $\mathcal{D}_c$ that plays an important role in the bridge between the latent variables and the parameters. Second, the complexity of the c-th component is adjusted by the value of $\sum_{\boldsymbol{z}_c^N} q(\boldsymbol{z}_c^N) \log(\sum_{n=1}^{N} z_{nc})$. Therefore, components of different sizes are automatically regularized in their individual appropriate levels. Third, roughly speaking, with respect to the terms, it is preferable for the entropy of $q(\boldsymbol{z}^N)$ to be large. This makes the parameter estimation in FAB identifiable. More details are discussed in Sections 4.4 and 4.5.

4 Factorized Asymptotic Bayesian Inference Algorithm

4.1 Lower bound of FIC

Since $\bar{\boldsymbol{\theta}}$ is not available in practice, we cannot evaluate FIC itself. Instead, FAB maximizes an asymptotically-consistent lower bound for FIC. We firstly derive the lower bound as follows:

Lemma 4 *Let us define* $\mathcal{L}(a,b) \equiv \log b + (a-b)/b$. $FIC(\boldsymbol{x}^N, M)$ *is lower bounded as follows:*

$$FIC(\boldsymbol{x}^N, M) \geq \mathcal{G}(q, \tilde{q}, \theta, \boldsymbol{x}^N) \tag{11}$$
$$\equiv \sum_{\boldsymbol{z}^N} q(\boldsymbol{z}^N)\Big(\log p(\boldsymbol{x}^N, \boldsymbol{z}^N|\boldsymbol{\theta}) - \frac{\mathcal{D}_{\boldsymbol{\alpha}}}{2}\log N - \sum_{c=1}^{C}\frac{\mathcal{D}_c}{2}\mathcal{L}(\sum_{n=1}^{N} z_{nc}, \sum_{n=1}^{N}\tilde{q}(z_{nc})) - \log q(\boldsymbol{z}^N)\Big),$$

with arbitrary choices of q, $\boldsymbol{\theta}$ *and a distribution* $\tilde{q}$ *on* $\boldsymbol{z}^N$ $(\sum_{n=1}^{N}\tilde{q}(z_{nc}) > 0)$.

Sketch of Proof 4 *Since* $\bar{\boldsymbol{\theta}}$ *is the ML estimator of* $\log p(\boldsymbol{x}^N, \boldsymbol{z}^N|\boldsymbol{\theta})$, $\log p(\boldsymbol{x}^N, \boldsymbol{z}^N|\bar{\boldsymbol{\theta}}) \geq \log p(\boldsymbol{x}^N, \boldsymbol{z}^N|\boldsymbol{\theta})$ *is satisfied. Further, on the basis of the concavity of the logarithm function,* $\log(\sum_{n=1}^{N} z_{nc}) \leq \mathcal{L}(\sum_{n=1}^{N} z_{nc}, \sum_{n=1}^{N}\tilde{q}(z_{nc}))$ *is satisfied. By substituting these two inequalities into (10), we obtain (11).*

In addition to the replacement of $\boldsymbol{\theta}$ with the unavailable estimator $\bar{\boldsymbol{\theta}}$, a computationally intractable $\log(\sum_{n=1}^{N} z_{nc})$ is linearized around $\sum_{n=1}^{N}\tilde{q}(z_{nc})$ with a new parameter (distribution) $\tilde{q}$.

Now our problem is to solve the following maximization problem (note that $q, \boldsymbol{\theta}$ and $\tilde{q}$ are functions of M):

$$M^*, q^*, \boldsymbol{\theta}^*, \tilde{q}^* = \arg\max_{M,q,\boldsymbol{\theta},\tilde{q}} \mathcal{G}(q, \tilde{q}, \boldsymbol{\theta}, \boldsymbol{x}^N). \tag{12}$$

The following lemma gives us a guide for optimizing the newly introduced distribution $\tilde{q}$. We omit the proof because of space limitations.

Lemma 5 *If we fix* q *and* $\boldsymbol{\theta}$, *then* $\tilde{q} = q$ *maximizes* $\mathcal{G}(q, \tilde{q}, \theta, \boldsymbol{x}^N)$.

With a finite number of model candidates $\mathcal{M}$, we can use a two-stage algorithm in which we first solve the maximization of (12) for all candidates in $\mathcal{M}$, and then select the best model. When we optimize only the number of components C (e.g., as in a GMM), we need to solve the inner maximization $C_{\max}$ times (the maximum search number of components.) However, if we intend to optimize $S_1, \ldots, S_C$ (e.g., as in PCM or an autoregressive mixture), we must avoid a combinatorial scalability issue. FAB provides a natural way to do this because the objective function is separated into independent parts in terms of $S_1, \ldots, S_C$.

4.2 Iterative Optimization Algorithm

Let us first fix C, and consider the following optimization problem:

$$\boldsymbol{S}^*, q^*, \boldsymbol{\theta}^*, \tilde{q}^* = \arg\max_{\boldsymbol{S},q,\boldsymbol{\theta},\tilde{q}} \mathcal{G}(q, \tilde{q}, \boldsymbol{\theta}, \boldsymbol{x}^N), \tag{13}$$

where $\boldsymbol{S} = (S_1, \ldots, S_C)$. As the inner maximization, FAB solves (13) on the basis of iterations of two substeps (V-step and M-step). Let the superscription (t) represent the t-th iteration.

V step The V-step optimizes the variational distribution q as follows:

$$q^{(t)} = \arg\max_{q}\Big\{\mathcal{G}(q, \tilde{q} = q^{(t-1)}, \boldsymbol{\theta}^{(t-1)}, \boldsymbol{x}^N)\Big\}. \tag{14}$$

More specifically, $q^{(t)}$ is obtained as follows:

$$q^{(t)}(z_{nc}) \propto \alpha_c^{(t-1)} p(\boldsymbol{x}_n|\boldsymbol{\phi}_c^{(t-1)}) \exp(\frac{-\mathcal{D}_c}{2\alpha_c^{(t-1)}N}), \tag{15}$$

where we use $\sum_{n=1}^{N} q^{(t-1)}(z_{nc}) = \alpha_c^{(t-1)}N$. The regularization terms which we discussed in Section 3 appear here as exponentiated update terms, i.e., $\exp(\frac{-\mathcal{D}_c}{2\alpha_c^{(t-1)}N})$. Roughly speaking, (15) indicates that smaller and more complex components are likely to become smaller through the iterations. The V-step is different from the E-step of the EM algorithm in the term $\exp(\frac{-\mathcal{D}_c}{2\alpha_c^{(t-1)}N})$. This makes an essential difference between FAB inference and the ML estimation that employs the EM algorithm (Sections 4.4 and 4.5).

M step The M-step optimizes the models of individual components $\boldsymbol{S}$ and the parameter $\boldsymbol{\theta}$ as follows:

$$\boldsymbol{S}^{(t)}, \boldsymbol{\theta}^{(t)} = \arg\max_{\boldsymbol{S},\boldsymbol{\theta}} \mathcal{G}(q^{(t)}, \tilde{q} = q^{(t)}, \boldsymbol{\theta}, \boldsymbol{x}^N). \tag{16}$$

$\mathcal{G}(q, \tilde{q}, \theta, \boldsymbol{x}^N)$ has a significant advantage in its avoiding of the combinatorial scalability issue on $\boldsymbol{S}$ selection. Since $\mathcal{G}(q, \tilde{q}, \theta, \boldsymbol{x}^N)$ has no cross term between components (if we fix $\tilde{q}$), we can separately optimize $\boldsymbol{S}$ and $\boldsymbol{\theta}$ as follows:

$$\alpha_c^{(t)} = \sum_{n=1}^{N} q^{(t)}(z_{nc})/N, \tag{17}$$

$$S_c^{(t)}, \boldsymbol{\phi}_c^{(t)} = \arg\max_{S_c,\boldsymbol{\phi}_c} \mathcal{H}_c(q^{(t)}, q^{(t)}, \boldsymbol{\phi}_c, \boldsymbol{x}^N) \tag{18}$$

$$\mathcal{H}_c(q^{(t)}, q^{(t)}, \boldsymbol{\phi}_c, \boldsymbol{x}^N) = \sum_{n=1}^{N} q^{(t)}(z_{nc}) \log p(\boldsymbol{x}_n|\boldsymbol{\phi}_c) - \frac{\mathcal{D}_c}{2}\mathcal{L}(\sum_{n=1}^{N} q^{(t)}(z_{nc}), \sum_{n=1}^{N} q^{(t)}(z_{nc})).$$

$\mathcal{H}_c(q^{(t)}, q^{(t)}, \boldsymbol{\phi}_c, \boldsymbol{x}^N)$ is a part of $\mathcal{G}(q^{(t)}, q^{(t)}, \boldsymbol{\phi}_c, \boldsymbol{x}^N)$, which is related to the c-th component. With a finite set of component candidates, in (18), we first optimize $\boldsymbol{\phi}_c$ for each element of a fixed S_c and then select the optimal one by comparing them. Note that, if we use a single component type (e.g., GMM), our M-step eventually becomes the same as that in the EM algorithm.

4.3 Convergence and Consistency

After the t-th iteration, FIC is lower bounded as:

$$FIC(\boldsymbol{x}^N, M) \geq FIC_{LB}^{(t)}(\boldsymbol{x}^N, M) \equiv \mathcal{G}(q^{(t)}, q^{(t)}, \boldsymbol{\theta}^{(t)}, \boldsymbol{x}^N). \tag{19}$$

We have the following theorem which guarantees that FAB will monotonically increase $FIC_{LB}^{(t)}(\boldsymbol{x}^N, M)$ over the iterative optimization and also converge to the local minima under certain regularity conditions.

Theorem 6 *For the iteration of the V-step and the M-step, the following inequality is satisfied:*

$$FIC_{LB}^{(t)}(\boldsymbol{x}^N, M) \geq FIC_{LB}^{(t-1)}(\boldsymbol{x}^N, M). \tag{20}$$

Sketch of Proof 6 *The theorem is proved as follows:*

$$\mathcal{G}(q^{(t)}, q^{(t)}, \boldsymbol{\theta}^{(t)}, \boldsymbol{x}^N) \geq \mathcal{G}(q^{(t)}, q^{(t)}, \boldsymbol{\theta}^{(t-1)}, \boldsymbol{x}^N) \geq$$
$$\mathcal{G}(q^{(t)}, q^{(t-1)}, \boldsymbol{\theta}^{(t-1)}, \boldsymbol{x}^N) \geq \mathcal{G}(q^{(t-1)}, q^{(t-1)}, \boldsymbol{\theta}^{(t-1)}, \boldsymbol{x}^N).$$

The first and the third inequalities arise from (16) and (14), respectively. The second inequality arises from Lemma 5.

We then use the following stopping criterion for the FAB iteration steps:

$$FIC_{LB}^{(t)}(\boldsymbol{x}^N, M) - FIC_{LB}^{(t-1)}(\boldsymbol{x}^N, M) \leq \varepsilon, \tag{21}$$

where ε is an optimization tolerance parameter and is set to 1e-6 in Section 5.

In addition to the monotonic property of FAB, we present the theorem below, which asymptotic-theoretically supports FAB. Let us denote $M^{(t)} = (C, \boldsymbol{S}^{(t)})$ and let the superscriptions T and $\star$ represent the number of steps at convergence and the true model/parameters, respectively.

Theorem 7 $FIC_{LB}^{(T)}(\boldsymbol{x}^N, M^{(T)})$ *is asymptotically consistent with* $FIC(\boldsymbol{x}^N, M^{(T)})$.

Sketch of Proof 7 *1) In (15),* $\exp(\frac{-\mathcal{D}_c}{2\alpha_c^{(T)}N})$ *converges to one, and* $\boldsymbol{S}^{(T-1)} = \boldsymbol{S}^{(T)}$ *holds without loss of generality. Therefore, the FAB algorithm is asymptotically reduced to the EM algorithm. This means* $\boldsymbol{\theta}^{(T)}$ *converges to the ML estimator* $\hat{\boldsymbol{\theta}}$ *of* $\log p(\boldsymbol{x}^N|\boldsymbol{\theta})$. *Then,* $|\sum_{\boldsymbol{z}^N} q^{(T)}(\boldsymbol{z}^N)(\log p(\boldsymbol{x}^N, \boldsymbol{z}^N|\bar{\boldsymbol{\theta}}) - \log p(\boldsymbol{x}^N, \boldsymbol{z}^N|\boldsymbol{\theta}^{(T)}))|/N \to 0$ *holds. 2) On the basis of the law of large numbers,* $\sum_{n=1}^N z_{nc}/N$ *converges to* $\alpha_c^\star$. *Then,* $|\sum_{\boldsymbol{z}^N} q^{(T)}(\boldsymbol{z}^N) \sum_{c=1}^C (\log(\sum_{n=1}^N z_{nc}) - \log(\sum_{n=1}^N q^{(T)}(z_{nc})))|/N \to 0$ *holds. The substitution of 1) and 2) into (11) proves the theorem.*

The following theorem arises from Theorems 3 and 7 and guarantees that FAB will be asymptotically capable of evaluating the marginal log-likelihood itself:

Theorem 8 $FIC_{LB}^{(T)}(\boldsymbol{x}^N, M^{(T)})$ *is asymptotically consistent with* $\log p(\boldsymbol{x}^N|M^{(T)})$.

4.4 Shrinkage Mechanism

As has been noted in Sections 3 and 4.2, the term $\exp(\mathcal{D}_c/2\alpha_c^{(t)}N)$ in (15), which arises from $\mathcal{D}_c \sum_{\boldsymbol{z}_c^N} q(\boldsymbol{z}_c^N) \log(\sum_{n=1}^N z_{nc})/2$ in (10), plays a significant role in the control of model complexity.

Fig.1 illustrates the shrinkage effects of FAB in the case of of a GMM. For $D = 10$ and $N = 50$, for example, a component of $\alpha_c < 0.2$ is severely regularized, and such components are automatically shrunk (see the left graph.) With increasing N, the shrinkage effect decreases, and small components manage to survive in the FAB optimization process. The right graph shows the shrinkage effect in terms of the dimensionality. In a low dimensional space (e.g., $D = 1, 3$), small components are not strictly regularized. For $D = 20$ and $N = 100$, however, a component of $\alpha_c < 0.4$ may be shrunk, and only one component might survive because of the constraint $\sum_{c=1}^C \alpha_c$.

Let us consider the gradient of $\exp(\frac{-\mathcal{D}_c}{2\alpha_c N})$ which is derived as $\partial \exp(\frac{-\mathcal{D}_c}{2\alpha_c N})/\partial \alpha_c = \frac{\mathcal{D}_c}{\alpha_c^2 N} \exp(\frac{-\mathcal{D}_c}{2\alpha_c N}) \geq 0$. This indicates that the iteration of FAB accelerates the shrinkages in order of $\mathcal{O}(\alpha_c^{-2})$. More intuitively speaking, the smaller the component is, the more likely it is to be shrunk. The above observations explain the overfitting mitigation mechanism in FAB.

The shrinkage effect of FAB is different from the prior-based regularization of the MAP estimation. In fact, it appears despite our asymptotically ignoring of the prior effect, and even if a non-informative prior or a flat prior is applied.

In terms of practical implementation, α_c does not get to exactly zero because (15) takes an exponentiated update. Therefore, in our FAB procedure, we apply a thresholding-based shrinkage of small components, after the V-step, as follows:

$$q^{(t)}(z_{nc}) = \begin{cases} 0 & \text{if } \sum_{n=1}^N q^{(t)}(z_{nc}) < \delta \\ q^{(t)}(z_{nc})/Q_c^{(t)} & \text{otherwise} \end{cases} \tag{22}$$

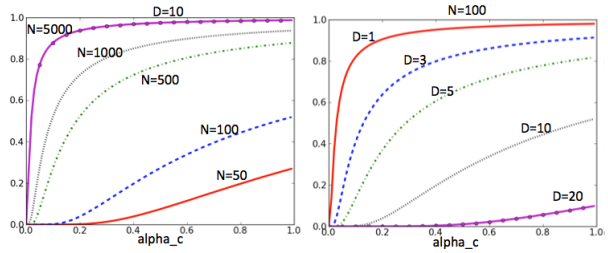

Figure 1: Shrinkage effects of FAB in GMM ($\mathcal{D}_c = D + D(D+1)/2$, where D and $\mathcal{D}_c$ are, respectively, the data dimensionality and the parameter dimensionality.) The horizontal and vertical axes represent α_c and $\exp(-\mathcal{D}_c/(\alpha_c N))$, respectively. The effect is evident when the number of data is not sufficient for the parameter dimensionality.

Algorithm 1 FAB_{two}: Two-stage FAB

input : $\boldsymbol{x}^N$, $C_{\max}, \mathcal{S}, \varepsilon$
output : C^*, $\boldsymbol{S}^*$, $\boldsymbol{\theta}^*, q^*(\boldsymbol{z}^N), FIC^*_{LB}$

$FIC^*_{LB} = -\infty$
for $C = 1, \ldots, C_{\max}$ **do**
  Calculate $FIC^{(T)}_{LB}, C, \boldsymbol{S}^{(T)}, \boldsymbol{\theta}^{(T)}$, and $q^{(T)}(\boldsymbol{z}^N)$ by $FAB_{shrink}(\boldsymbol{x}^N, C_{\max} = C, \mathcal{S}, \varepsilon, \delta = 0)$.
end for
Choose C^*, $\boldsymbol{S}^*$, $\boldsymbol{\theta}^*$ and $q^*(\boldsymbol{z}^N)$ by (12).

δ and $Q_c^{(t)}$ are a threshold value and a normalization constant for $\sum_{c=1}^{C} q^{(t)}(z_{nc}) = 1$, respectively. We used the threshold value $\delta = 0.01N$ in Section 5. With the shrinkage operation, FAB does not require the two-stage optimization in (12). By starting from a sufficient number of components $C_{\max}$, FAB iteratively and simultaneously optimizes all of $M = (C, \boldsymbol{S})$, $\boldsymbol{\theta}$, and q. Since FAB cannot revive shrunk components, it is a greedy algorithm, like the least angle regression method [9] which does not revive shrunk features.

The two-stage algorithm and the shrinkage algorithm are shown here as Algorithms 1 and 2. $\mathcal{S}$ is a set of component candidates (e.g., $\mathcal{S} = \{Gaussian\}$ for GMM. $\mathcal{S} = \{0, \ldots, K_{\max}\}$ for PCM where $K_{\max}$ is the maximum degree of curves.)

4.5 Identifiability

A well-known difficulty in ML estimation for mixture models is non-identiability, as we have noted in Section 1. Let us consider a simple mixture model, $p(X|a,b,c) = ap(X|b) + (1-a)p(X|c)$. All models $p(X|a,b,c=b)$ with arbitrary a are equivalent (i.e., in an equivalent class.) Therefore, the ML estimator is not unique. Theoretical studies have shown that Bayesian estimators and VB estimators avoid the above issue and significantly outperform the ML estimator in terms of generalization error [17, 21, 24].

Algorithm 2 FAB_{shrink}: Shrinkage FAB

input : $\boldsymbol{x}^N$, $C_{\max}, \mathcal{S}, \varepsilon, \delta$
output : $FIC^{(T)}_{LB}, C, \boldsymbol{S}^{(T)}, \boldsymbol{\theta}^{(T)}, q^{(T)}(\boldsymbol{z}^N)$

$t = 0, FIC^{(0)}_{LB} = -\infty$
randomly initialize $q^{(0)}(\boldsymbol{z}^N)$.
while convergence **do**
  Calculate $\boldsymbol{S}^{(t)}$ and $\boldsymbol{\theta}^{(t)}$ by (17) and (18).
  Check convergence by (21).
  Calculate $q^{(t+1)}$ by (15).
  Shrinkage components by (22).
  $t = t + 1$
end while
$T = t$

In FAB, at the stationary (convergence) point of (15), the following equality is satisfied as $\sum_{n=1}^{N} q^*(z_{nc}) = \alpha_c^* N = \sum_{n=1}^{N} \alpha_c^* p(\boldsymbol{x}_n|\boldsymbol{\phi}_c^*) \exp(\frac{-\mathcal{D}_c}{2\alpha_c^* N})/Z_n$, where $Z_n = \sum_{c=1}^{C} \alpha_c^* p(\boldsymbol{x}_n|\boldsymbol{\phi}_c^*) \exp(\frac{-\mathcal{D}_c}{2\alpha_c^* N})$ is the normalization constant. For all pairs of $\alpha_c^*, \alpha_{c'}^* > 0$, we have the following condition: $\sum_{n=1}^{N} p(\boldsymbol{x}_n|\boldsymbol{\phi}_c^*) \exp(\frac{-\mathcal{D}_c}{2\alpha_c^* N}) = \sum_{n=1}^{N} p(\boldsymbol{x}_n|\boldsymbol{\phi}_{c'}^*) \exp(\frac{-\mathcal{D}_{c'}}{2\alpha_{c'}^* N})$. Then, the necessary condition of $\boldsymbol{\phi}_c^* = \boldsymbol{\phi}_{c'}^*$ is $\alpha_c^* = \alpha_{c'}^*$. Therefore, FAB chooses the model, in an equivalent class, which maximizes the entropy of $\boldsymbol{q}$ and addresses the non-identifiability issue.

Unfortunately, FIC and FAB do not resolve another non-identifiability issue, i.e., divergence of the criterion. Without the assumption **A2**, (10) and (11) can diverge to infinity (e.g., a GMM with a zero-variance component[4].) We can address this issue by applying the Laplace approximation around the MAP estimator $\bar{\boldsymbol{\theta}}_{MAP}$ of $p(\boldsymbol{x}^N, \boldsymbol{z}^N|\boldsymbol{\theta})p(\boldsymbol{\theta})$ rather than around the ML estimator $\bar{\boldsymbol{\theta}}$. In such a case, a flat conjugate prior might be used in order to avoid the divergence. While such priors do not provide regularization effects, FAB has an overfitting mitigation mechanism, as has been previously discussed. Because of space limitations, we omit here a detailed discussion of FIC and FAB with prior distributions.

5 Experiments

5.1 Polynomial Curve Mixture

We first evaluated FAB_{shrink} for PCM to demonstrate how its model selection works. We used the settings $N = 300$ and $C_{\max} = K_{\max} = 10$ in this evaluation[5].

Let Y be a random dependent variable of X.

[4]Practically speaking, with the shrinkage (22), FAB does not have the divergence issue because small components are automatically removed.

[5]Larger values of $C_{\max}$ and $K_{\max}$ did not make a large difference in results, but they were hard to visualize.

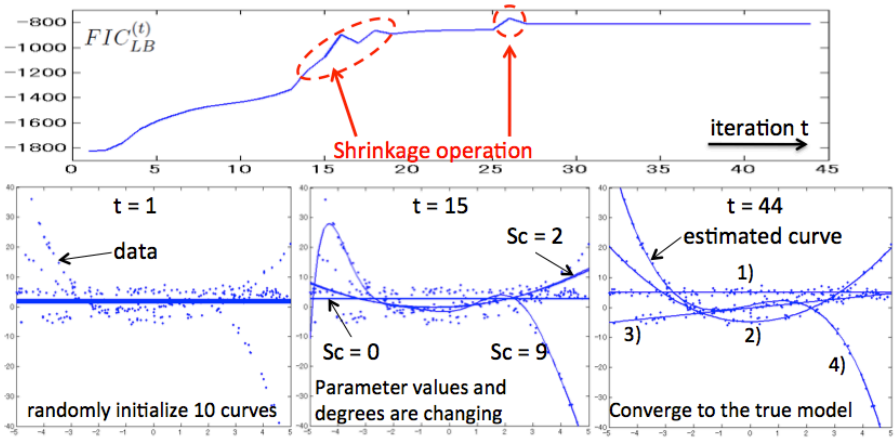

Figure 2: Model selection procedure for FAB_{shrink} in application to PCM. The true curves are : 1) $Y = 5 + \varepsilon$, 2) $Y = X^2 + 5 + \varepsilon$, 3) $Y = X + \varepsilon$, and 4) $Y = -0.5X^3 + 2X + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0, 1)$. The top shows the change in $FIC_{LB}^{(t)}$ over iteration t. The bottom three graphs are the estimated curves at $t = 1, 6$, and 44.

PCM has the mixed component $p_c(Y|X, \boldsymbol{\phi}_c) = \mathcal{N}(Y, X(S_c)\boldsymbol{\beta}_c, \sigma_c^2)$, where $X(S_c) = (1, X, \ldots, X^{S_c})$. The true model has four curves as shown in Fig. 2.We used the setting $\boldsymbol{\alpha} = (0.3, 0.2, 0.3, 0.2)$ for the curves 1) - 4). $\boldsymbol{x}^N$ was uniformly sampled from $[-5, 5]$.

Fig. 2 illustrates $FIC_{LB}^{(t)}$ (top) and intermediate estimation results (bottom). From the top figure, we are able to confirm that FAB monotonically increases $FIC_{LB}^{(t)}$, except for the points (dashed circles) of the shrinkage operation. In the initial state $t = 1$ (bottom left), $q^{(0)}(\boldsymbol{z}^N)$ is randomly initialized, and therefore the degrees of all ten curves are zero ($S_c = 0$). Over the iteration, FAB_{shrink} simultaneously searches C, $\boldsymbol{S}$, and $\boldsymbol{\theta}$. For example, at $t = 15$ (bottom center), we have one curve of $S_c = 9$, two curves of $S_c = 2$, and six curves of $S_c = 0$. Here two curves have already been shrunk. Finally, FAB_{shrink} selects the model consistent with the true model at $t = 44$ (bottom left). These results demonstrate the powerful simultaneous model selection procedure of FAB_{shrink}.

5.2 Gaussian Mixture Model

We applied FAB_{two} and FAB_{shrink} to GMM in application to artificial datasets and UCI datasets [10], and compared them with the variational Bayesian EM algorithm for GMM (VBEM) that is described in Section 10.2 of [3] and the variational Bayesian Dirichlet Process GMM (VBDP) [5]. The former and the latter use a Dirichlet prior and a Dirichlet process prior for $\boldsymbol{\alpha}$, respectively. We used the implementations by M. E. Khan[6] for VBEM and by K. Kurihara[7] for VBDP. We used the default hyperparameter values in the software. In the following experiments, $C_{\max}$ was set to $C_{\max} = 20$ for all methods.

5.2.1 Artificial Data

We generated the true model with $D = 15$, $C^\star = 5$ and $\boldsymbol{\phi}_c^\star = (\mu_c^\star, \Sigma_c^\star)$ (mean and covariance), where the superscription $\star$ denotes the true model. The parameter values were randomly sampled as $\alpha_c^\star \sim [0.4, 0.6]$ (before normalization), and $\mu_c^\star \sim [-5, 5]$. $\Sigma_c^\star$ were generated as $a_c \bar{\Sigma}_c^\star$, where $a_c \sim [0.5, 1.5]$ is a scale parameter and $\bar{\Sigma}_c^\star$ is generated using the "gallery" function in Matlab. The results are averages of ten runs.

The sufficient number of data is important measure to evaluate asymptotic methods, and this is usually measured in terms of the empirical convergences of their criteria. Table 1 shows how the values of $FIC_{LB}^{(T)}/N$ converge over N. Roughly speaking, $N = 1000 \sim 2000$ was sufficient for convergence (depending on data dimensionality D.) Our results indicate that FAB can be applied in actual practice to recent data analysis scenarios with large N values.

We next compared the four methods on the basis of three evaluation metrics: 1) the Kullback-Leibler divergence (KL) between the true distribution and the estimated distribution (estimation error), 2) $|C^* - C^\star|$ (model selection error), and 3) CPU time.

For $N \leq 500$, VBDP performed better than or comparably with FAB_{two} and FAB_{shrink} w.r.t. KL and $|C^* - C^\star|$ (left and middle graphs of Fig. 3). With increasing N, those of the FABs significantly decrease while that of VBDP does not. This might be because the lower bound in (1) generally does not reach the marginal log-likelihood as has been noted. VBEM was significantly inferior for both metrics. Another observation is that VBDP and VBEM were likely to re-

[6] http://www.cs.ubc.ca/~emtiyaz/software.html
[7] http://sites.google.com/site/kenichikurihara/.

Table 1: Convergence of FIC^*_{LB} over data size N. The standard deviations are in parenthesis.

N	50	100	250	500	1000	2000	3000
FAB_{two}	-10.31 (3.51)	-16.27 (2.51)	-15.46 (1.19)	-15.14 (1.05)	-14.63 (1.10)	-14.52 (0.50)	-14.34 (0.39)
FAB_{shrink}	-12.78 (2.61)	-17.22 (1.56)	-15.66 (1.24)	-15.16 (1.02)	-14.62 (0.85)	-14.68 (0.74)	-14.39 (0.39)

Table 2: Predictive likelihood per data. The standard deviations are in parenthesis. The maximum number of training samples was limited to 2000 (parenthesis in the N column), and the rest of them were used for test. The best results and those not significantly worse than them are highlighted in boldface (one-side t-test with 95% confidence.)

data	N	D	FAB_{two}	FAB_{shrink}	VBEM	VBDP
iris	150	4	-1.92 (0.76)	-2.11 (0.68)	-3.05 (0.64)	**-1.65 (0.57)**
yeast	1394	13	**18.79 (1.81)**	10.86 (3.38)	3.52 (3.80)	15.94 (0.72)
cloud	2048	10	7.89 (2.81)	**9.09 (2.62)**	5.47 (2.62)	2.67 (2.28)
wine quality	6497 (2000)	11	**-2.68 (0.22)**	**-2.68 (0.17)**	-10.49 (0.17)	-16.52 (0.06)
color moments	68040 (2000)	9	**-6.97 (0.20)**	**-6.95 (0.21)**	-7.58 (0.21)	-7.66 (0.11)
cooc texture	68040 (2000)	16	**14.45 (0.22)**	**14.45 (0.46)**	13.45 (0.13)	6.78 (0.28)
spoken arabic digits	350319 (2000)	13	**-12.94 (0.11)**	**-12.90 (0.09)**	-13.81 (0.09)	-14.34 (0.19)

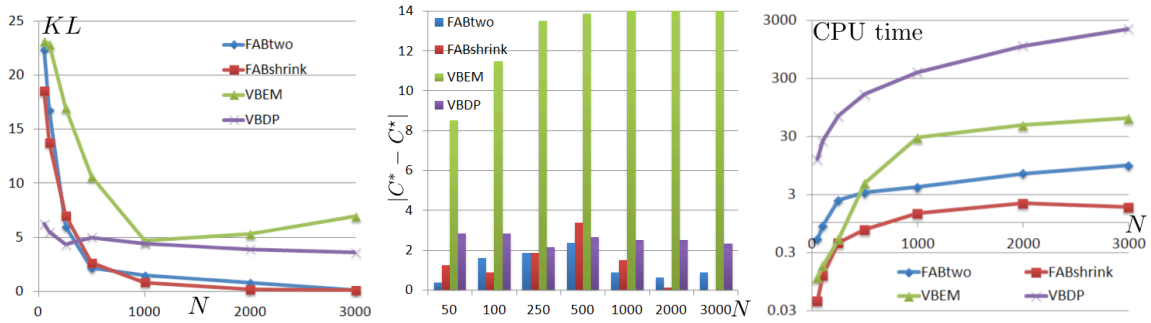

Figure 3: Comparisons of four methods in terms of KL (left), $|C^* - C^\star|$ (middle), and 3) CPU time (right).

spectively under-fit and to over-fit their models, while we did not find such biases in FABs. While KLs of FABs are similar, FAB_{shrink} performed better in terms of $|C^* - C^\star|$ because small components survived in FAB_{two}. Further, both FABs had a significant advantage over VBDP and VBEM in CPU time (right graph). In particular, while the CPU time of VBDP grew explosively over N, those of FABs remained in a practical range. Since FAB_{shrink} does not use a loop to optimize C, it was significantly faster than FAB_{two}.

5.2.2 UCI data

For the UCI datasets summarized in Table 2, we do not know their true distributions and therefore employed predictive log-likelihood as an evaluation metric. For large scale datasets, we randomly selected two thousands data for training and used the rest of the data for prediction because of the scalability issue in VBDP.

As shown in Table 2, FABs gave better performance then VBDP and VBEM with a sufficient number of data (the scale $\mathcal{O}(10^3)$ agrees with the results in the previous section). The trend of VBDP (under-fit) and VBEM (over-fit) was the same with the previous section. Also, while the predictive log-likelihood of FAB_{two} and FAB_{shrink} are competitive, FAB_{shrink} obtained more compact models (smaller C^* values) than FAB_{two}. Unfortunately, we have no space to show the results here.

6 Summary and Future Work

We have proposed approximation of marginal log-likelihood (FIC) and an inference method (FAB) for Bayesian model selection for mixture models, as an alternative to VB inference. We have given their justifications (asymptotic consistency, convergence, etc) and analyzed FAB mechanisms in terms of overfitting mitigation (shrinkage) and identifiability. Experimental results have shown that FAB outperforms state-of-the-art VB methods for a practical number of data in terms of both model selection performance and computational efficiency.

Our next step is extensions of FAB to more general non-regular and non-identifiable models which contain continuous latent variables (e.g., factor analyzer models [12] and matrix factorization models [17]), time-dependent latent variables (e.g., hidden Markov models [16]), hierarchical latent variables [14, 4], etc. We believe that our idea, factorized representation of the marginal log-likelihood in which the asymptotic approximation works, is widely applicable to them.

References

[1] T. Ando. *Bayesian Model Selection and Statistical Modeling*. Chapman and Hall/CRC, 2010.

[2] M. Beal. *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.

[3] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag, 2006.

[4] C. M. Bishop and M. E. Tipping. A hierarchical latent variable model for data visualization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):281–293, 1998.

[5] D. M. Blei and M. I. Jordan. Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1(1):121–144, 2006.

[6] A. Corduneanu and C. M. Bishop. Variational Bayesian model selection for mixture distributions. In *Proceedings of Artifficial Intelligence and Statistics*, pages 27–34, 2001.

[7] I. Csiszar and P. C. Shields. The consistency of the BIC Markov order estimator. *Annals of Statistics*, 28(6):1601–1619, 2000.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from imcomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, B39(1):1–38, 1977.

[9] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.

[10] A. Frank and A. Asuncion. UCI machine learning repository. `http://archive.ics.uci.edu/ml`, 2010.

[11] R. Fujimaki, Y. Sogawa, and S. Morinaga. Online heterogeneous mixture modeling with marginal and copula selection. In *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 645–653, 2011.

[12] Z. Ghahramani and M. Beal. Variational inference for Bayesian mixtures of factor analysers. In *Advances in Neural Information Processing Systems 12*, pages 449–455. MIT Press, 2000.

[13] H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 186:453–461, 1946.

[14] M. I. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6:181–214, 1994.

[15] K. Kurihara, M. Welling, and Y. W. Teh. Collapsed variational dirichlet process mixture models. In *Proceedings of Twentieth International Joint Conference on Artificial Intelligence*, pages 2796–2801, 2007.

[16] D. J. MacKay. Ensemble learning for hidden Markov models. Technical report, University of Cambridge, 1997.

[17] S. Nakajima and M. Sugiyama. Theoretical analysis of Bayesian matrix factorization. *Journal of Machine Learning Research*, 12:2579–2644, 2011.

[18] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[19] M. Stone. Comments on model selection criteria of Akaike and Schwarz. *Journal of the Royal Statistical Society: Series B*, 41(2):276–278, 1979.

[20] Y. W. Teh, D. Newman, and M. Welling. A collapsed variational bayesian inference algorithm for latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, pages 1378–1385, 2006.

[21] K. Watanabe and S. Watanabe. Stochastic complexities of Gaussian mixtures in variational Bayesian approximation. *Journal of Machine Learning Research*, 7:625–644, 2006.

[22] S. Watanabe. Algebraic analysis for nonidentiable learning machines. *Neural Computation*, 13:899–033, 2001.

[23] S. Watanabe. *Algebraic geometry and statistical learning*. UK: Cambridge University Press, 2009.

[24] K. Yamazaki and S. Watanabe. Mixture models and upper bounds of stochastic complexity. *Neural Networks*, 16:1029–1038, 2003.