Rebuilding Factorized Information Criterion: Asymptotically Accurate Marginal Likelihood

Abstract

Factorized information criterion (FIC) is a recently developed approximation technique for the marginal log-likelihood, which provides an automatic model selection framework for a few latent variable models (LVMs) with tractable inference algorithms. This paper reconsiders FIC and fills theoretical gaps of previous FIC studies. First, we reveal the core idea of FIC that allows generalization for a broader class of LVMs, including continuous LVMs, in contrast to previous FICs, which are applicable only to binary LVMs. Second, we investigate the model selection mechanism of the generalized FIC. Our analysis provides a formal justification of FIC as a model selection criterion for LVMs and also a systematic procedure for pruning redundant latent variables that have been removed heuristically in previous studies. Third, we provide an interpretation of FIC as a variational free energy and uncover previously-unknown their relationship. A demonstrative study on Bayesian principal component analysis is provided and numerical experiments support our theoretical results.

1. Introduction

The marginal log-likelihood is a key concept of Bayesian model identification of latent variable models (LVMs), such as mixture models (MMs), probabilistic principal component analysis, and hidden Markov models (HMMs). Determination of dimensionality of latent variables is an essential task to uncover hidden structures behind the observed data as well as to mitigate overfitting. In general, LVMs are *singular* (i.e., mapping between parameters and probabilistic models is not one-to-one) and such classical information criteria based on the regularity assumption as the Bayesian information criterion (BIC) (Schwarz, 1978) are no longer justified. Since exact evaluation of the marginal log-likelihood is often not available, approxima-

tion techniques have been developed using sampling (i.e., Markov Chain Monte Carlo methods (MCMCs) (Hastings, 1970)), a variational lower bound (i.e., the variational Bayes methods (VB) (Attias, 1999; Jordan et al., 1999)), or algebraic geometry (i.e., the widely applicable BIC (WBIC) (Watanabe, 2013)). However, model selection using these methods typically requires heavy computational cost (e.g., a large number of MCMC sampling in a high-dimensional space, an outer loop for VB/WBIC.)

In the last few years, a new approximation technique and an inference method, factorized information criterion (FIC) and factorized asymptotic Bayesian inference (FAB), have been developed for some binary LVMs (Fujimaki & Morinaga, 2012; Fujimaki & Hayashi, 2012; Hayashi & Fujimaki, 2013; Eto et al., 2014). Unlike existing methods which evaluate approximated marginal log-likelihoods calculated for each latent variable dimensionality (and therefore need an outer loop for model selection), FAB finds an effective dimensionality via an EM-style alternating optimization procedure.

For example, let us consider a K-component MM for N observations $\mathbf{X}^\top = (\mathbf{x}_1, \dots, \mathbf{x}_N)$ with one-of-K coding latent variables $\mathbf{Z}^\top = (\mathbf{z}_1, \dots, \mathbf{z}_N)$, mixing coefficients $\boldsymbol{\beta} = (\beta_1, \dots, \beta_K)$, and $D_{\boldsymbol{\Xi}}$-dimensional component-wise parameters $\boldsymbol{\Xi} = \{\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_K\}$. FIC of MMs (Fujimaki & Morinaga, 2012) is defined by

$$\mathrm{FIC}_{\mathrm{MM}}(K) \equiv \max_q \mathbb{E}_q \left[\ln p(\mathbf{X}, \mathbf{Z} \mid \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Xi}}, K) \right.$$
$$\left. - \sum_k \frac{D_{\boldsymbol{\xi}_k}}{2} \ln \frac{\sum_n z_{nk}}{N} \right] + H(q) - \frac{D_{\boldsymbol{\Pi}}}{2} \ln N, \quad (1)$$

where q is the distribution of $\mathbf{Z}$, $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\Xi}}$ are the maximum joint-likelihood estimators (MJLEs), $D_{\boldsymbol{\Pi}} = D_{\boldsymbol{\Xi}} + K - 1$ is the total dimension of $\boldsymbol{\Xi}$ and $\boldsymbol{\beta}$, and $H(q)$ is the entropy of q. A key characteristic of FIC can be observed in the second term of Eq. (1), which gives the penalty in terms of model complexity. As we can see, the penalty term increases when $\sum_n z_{nk}$—the number of effective samples of the k-th component—is small, i.e., $\mathbf{Z}$ is degenerated. Therefore, through the optimization of q, the degenerate dimension is automatically pruned until a non-degenerated $\mathbf{Z}$ is found. This mechanism makes FAB a one-pass model selection algorithm and computationally more attractive than

the other methods. The validity of the penalty term has been confirmed for other binary LVMs, e.g., HMMs (Fujimaki & Hayashi, 2012), latent feature models (Hayashi & Fujimaki, 2013), and mixture of experts (Eto et al., 2014).

Despite FAB's practical success compared with BIC and VB, it is unclear that what conditions are actually necessary to guarantee that FAB yields the true latent variable dimensionality. In addition, the generalization of FIC for non-binary LVMs still remains an important open issue. In case that $\mathbf{Z}$ takes negative and/or continuous values, $\sum_n z_{nk}$ is no longer interpretable as the number of effective samples, and we loose the clue for finding the redundant dimension of $\mathbf{Z}$.

This paper proposes generalized FIC (gFIC), given by

$$\mathrm{gFIC}(K) \equiv \mathbb{E}_{q^*}[\mathcal{L}(\mathbf{Z}, \hat{\mathbf{\Pi}}, K)] + H(q), \qquad (2)$$

$$\mathcal{L}(\mathbf{Z}, \mathbf{\Pi}, K) = \ln p(\mathbf{X}, \mathbf{Z} \mid \mathbf{\Pi}, K) - \frac{1}{2} \ln |\mathbf{F}_{\Xi}| - \frac{D_{\mathbf{\Pi}}}{2} \ln N.$$

Here, $q^*(\mathbf{Z}) \equiv p(\mathbf{Z} \mid \mathbf{X}, K)$ is the marginal posterior and $\mathbf{F}_{\hat{\Xi}}$ is the Hessian matrix of $- \ln p(\mathbf{X}, \mathbf{Z} \mid \mathbf{\Pi}, K)/N$ with respect to Ξ. In gFIC, the penalty term is given by the volume of the (empirical) Fisher information matrix. It naturally penalizes model complexity even when the latent variable $\mathbf{Z}$ takes negative and/or continuous values. Accordingly, gFIC is applicable to a broader class of LVMs, such as Bayesian principal component analysis (BPCA) (Bishop, 1998).

Furthermore, we prove that FAB automatically prunes redundant dimensionality along with optimizing q, and gFIC for the optimized q asymptotically converges to the marginal log-likelihood with a constant order error under some reasonable assumptions. This justifies gFIC as a model selection criterion for LVMs and further a natural one-pass model "pruning" procedure is derived, which is performed heuristically in previous FIC studies. We also provide an interpretation of gFIC as a variational free energy and uncover previously-unknown their relationship. This interpretation gives formal conditions ensure that model selection by VB marginal log-likelihood is justified.

Finally, we demonstrate the validity of gFIC by applying it to BPCA. The experimental results agree with to the theoretical properties of gFIC.

2. LVMs and Degeneration

We first define the class of LVMs we deal with in this paper. Here, we consider LVMs that have K-dimensional latent variables $\mathbf{z}_n$ (including the MMs in the previous section), but now $\mathbf{z}_n$ can take not only binary but also real values. Given $\mathbf{X}$ and a model family (e.g., MMs), our goal is to determine K and we refer to this as a *model*. Note that we sometimes omit the notation K for the sake of clarity, if it is obvious from the context.

The LVMs have D_{Ξ}-dimensional local parameters $\Xi = \{\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_K\}$ and D_{Θ}-dimensional global parameters Θ, which can include hyperparameters of the prior of $\mathbf{Z}$. We abbreviate them as $\mathbf{\Pi} = \{\Theta, \Xi\}$ and assume that the dimension $D_{\mathbf{\Pi}} = D_{\Theta} + D_{\Xi}$ is finite. Then, we define the joint probability: $p(\mathbf{X}, \mathbf{Z}, \mathbf{\Pi}) = p(\mathbf{X} \mid \mathbf{Z}, \mathbf{\Pi}) p(\mathbf{Z} \mid \mathbf{\Pi}) p(\mathbf{\Pi})$ where $\ln p(\mathbf{X}, \mathbf{Z} \mid \mathbf{\Pi})$ is twice differentiable at $\mathbf{\Pi} \in \mathcal{P}$ and let $\mathbf{F}_{\mathbf{\Pi}} \equiv$

$$\begin{pmatrix} \mathbf{F}_{\Theta} & \mathbf{F}_{\Theta, \Xi} \\ \mathbf{F}_{\Theta, \Xi}^{\top} & \mathbf{F}_{\Xi} \end{pmatrix} = - \begin{pmatrix} \frac{\partial}{\partial \Theta^{\top}} \\ \frac{\partial}{\partial \Xi^{\top}} \end{pmatrix} \begin{pmatrix} \frac{\partial}{\partial \Theta} & \frac{\partial}{\partial \Xi} \end{pmatrix} \frac{\ln p(\mathbf{X}, \mathbf{Z} \mid \mathbf{\Pi})}{N}.$$

Note that the MJLE $\hat{\mathbf{\Pi}} \equiv \mathrm{argmax}_{\mathbf{\Pi}} \ln p(\mathbf{X}, \mathbf{Z} \mid \mathbf{\Pi})$ depends on $\mathbf{Z}$ (and $\mathbf{X}$). In addition, $\ln p(\mathbf{X}, \mathbf{Z} \mid \mathbf{\Pi})$ can have multiple maximizers, and $\hat{\mathbf{\Pi}}$ could be a set of solutions.

Model redundancy is a notable property of LVMs. Because the latent variable $\mathbf{Z}$ is unobservable, the pair $(\mathbf{Z}, \mathbf{\Pi})$ is not necessarily determined uniquely for a given $\mathbf{X}$. In other words, there could be pairs $(\mathbf{Z}, \mathbf{\Pi})$ and $(\tilde{\mathbf{Z}}, \tilde{\mathbf{\Pi}})$, whose likelihoods have the same value, i.e., $p(\mathbf{X}, \mathbf{Z} \mid \mathbf{\Pi}, K) = p(\mathbf{X}, \tilde{\mathbf{Z}} \mid \tilde{\mathbf{\Pi}}, K)$. Previous FIC studies address this redundancy by introducing a variational representation that enables treating $\mathbf{Z}$ as fixed, as we explain in the next section. However, even if $\mathbf{Z}$ is fixed, the redundancy still remains, namely, the case in which $\mathbf{Z}$ is "degenerated," and there exists an equivalent likelihood with a smaller model $K' < K$:

$$p(\mathbf{X}, \mathbf{Z} \mid \mathbf{\Pi}, K) = p(\mathbf{X}, \tilde{\mathbf{Z}}_{K'} \mid \tilde{\mathbf{\Pi}}_{K'}, K'). \qquad (3)$$

In this case, K is overcomplete for $\mathbf{Z}$, and $\mathbf{Z}$ lies on the subspace of the model K. As a simple example, let us consider a three-component MM for which $\mathbf{Z} = (\mathbf{z}, \mathbf{1} - \mathbf{z}, \mathbf{0})$. In this case, $\boldsymbol{\xi}_3$ is unidentifiable, because the third component is completely unused, and the $K' = 2$-component MM with $\tilde{\mathbf{Z}}_2 \equiv (\mathbf{z}, \mathbf{1} - \mathbf{z})$ and $\tilde{\mathbf{\Pi}}_2 \equiv (\Theta, (\boldsymbol{\xi}_1, \boldsymbol{\xi}_2))$ satisfies equivalence relation (3). The notion of degeneration is defined formally as follows.

Definition 1. *Given $\mathbf{X}$ and K, $\mathbf{Z}$ is degenerated if there are multiple MJLEs and any $\mathbf{F}_{\hat{\mathbf{\Pi}}}$ of the MJLEs are not positive definite. Similarly, $p(\mathbf{Z})$ is degenerated in distribution, if $\mathbb{E}_p[\mathbf{F}_{\hat{\mathbf{\Pi}}}]$ are not positive definite. Let $\kappa(\mathbf{Z}) \equiv \mathrm{rank}(\mathbf{F}_{\hat{\mathbf{\Pi}}})$ and $\kappa(p) \equiv \mathrm{rank}(\mathbb{E}_p[\mathbf{F}_{\hat{\mathbf{\Pi}}}])$.*

The idea of degeneration is conceptually understandable as an analogous of linear algebra. Namely, each component of a model is a "basis", $\mathbf{Z}$ are "coordinates", and $\kappa(\mathbf{Z})$ is the number of necessary components to represent $\mathbf{X}$, i.e., the "rank" of $\mathbf{X}$ in terms of the model family. The degeneration of $\mathbf{Z}$ is then the same idea of the "degeneration" in linear algebra, i.e., the number of components is too many and $\mathbf{\Pi}$ is not uniquely determined even if $\mathbf{Z}$ is fixed.

As discussed later, given a degenerated $\mathbf{Z}$ where $\kappa(\mathbf{Z}) = K'$, finding the equivalent parameters $\tilde{\mathbf{Z}}_{K'}$ and $\tilde{\mathbf{\Pi}}_{K'}$ that satisfy Eq. (3) is an important task. In order to analyze this, we assume **A1**): for any degenerated $\mathbf{Z}$ under a model $K \geq 2$ and $K' < K$, there exists a continuous onto mapping $(\mathbf{Z}, \mathbf{\Pi}) \to (\tilde{\mathbf{Z}}_{K'}, \tilde{\mathbf{\Pi}}_{K'})$ that satisfies Eq. (3), and $\tilde{\mathbf{Z}}_{K'}$ is not degenerated. Note that, if $\mathcal{P}$ is a subspace of $\mathbb{R}^{D_{\mathbf{\Pi}}}$, a linear projection $\mathbf{V} : \mathbb{R}^{D_{\mathbf{\Pi}}} \mapsto \mathbb{R}^{D_{\mathbf{\Pi}_{K'}}}$ satisfies **A1** where $\mathbf{V}$ is the top-$D_{\mathbf{\Pi}_{K'}}$ eigenvectors of $\mathbf{F}_{\mathbf{\Pi}}$. This is verified easily by the fact that, by using the chain rule, $\mathbf{F}_{\tilde{\mathbf{\Pi}}_{K'}} = \mathbf{V} \mathbf{F}_{\hat{\mathbf{\Pi}}} \mathbf{V}^{\top}$, which is a diagonal matrix whose elements are positive eigenvalues. Therefore, $\mathbf{F}_{\tilde{\mathbf{\Pi}}_{K'}}$ is positive definite and $\tilde{\mathbf{Z}}_{K'}$ is not degenerated.

Let us further introduce a few assumptions required to show the asymptotic properties of gFIC. Suppose **A2**) the joint distribution is mutually independent in sample-wise,

$$p(\mathbf{X}, \mathbf{Z} \mid \mathbf{\Pi}, K) = \prod_n p(\mathbf{x}_n, \mathbf{z}_n \mid \mathbf{\Pi}, K), \qquad (4)$$

and **A3**) $\ln p(\mathbf{\Pi} \mid K)$ is constant, i.e., $\lim_{N \to \infty} \ln p(\mathbf{\Pi} \mid K)/N = 0$. In addition, **A4**) $p(\mathbf{\Pi} \mid K)$ is continuous, not improper, and its support $\mathcal{P}$ is compact and the whole space. Note that for almost all $\mathbf{Z}$, we expect that $\hat{\mathbf{\Pi}} \in \mathcal{P}$ is uniquely determined and $\mathbf{F}_{\hat{\mathbf{\Pi}}}$ is positive definite, i.e., **A5**) if $\mathbf{Z}$ is not degenerated, then $\ln p(\mathbf{X}, \mathbf{Z} \mid \mathbf{\Pi}, K)$ is concave and $\det |\mathbf{F}_{\hat{\mathbf{\Pi}}}| < \infty$.

2.1. Examples of the LVM Class

The above definition covers a broad class of LVMs. Here, we show that, as examples, MMs and BPCA are included in that class. Note that **A2** does not allow correlation among samples and analysis of cases with sample correlation (e.g. time series models) remains as an open problem.

MMs In the same notation used in Section 1, the joint likelihood is given by $p(\mathbf{X}, \mathbf{Z}|\mathbf{\Pi}) = \prod_n \prod_k \{\beta_k p_k(\mathbf{x}_n|\boldsymbol{\xi}_k)\}^{z_{nk}}$ where p_k is the density of component k. If $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_K$ have no overlap, $\mathbf{F}_{\mathbf{\Xi}}$ is the block-diagonal matrix whose block is given by $\mathbf{F}_{\boldsymbol{\xi}_k} = -\sum_n \nabla\nabla \ln p_k(\mathbf{x}_n|\boldsymbol{\xi}_k) z_{nk}/N$. This shows that the MM is degenerated, when more than one column of $\mathbf{Z}$ is filled by zero. For that case, removing such columns and corresponding $\boldsymbol{\xi}_k$ suffices as $\tilde{\mathbf{Z}}_{K'}$ and $\tilde{\mathbf{\Pi}}_{K'}$ in **A1**. Note that if p_k is an exponential-family distribution $\exp(\mathbf{x}_n^{\top}\boldsymbol{\xi}_k - \psi(\boldsymbol{\xi}_k))$, $-\nabla\nabla \ln p_k(\mathbf{x}_n|\boldsymbol{\xi}_k) = \nabla\nabla\psi(\boldsymbol{\xi}_k) = \mathbf{C}$ does not depend on n and gFIC recovers the original formulation of FIC_{MM}, i.e., $\frac{1}{2}\ln|\mathbf{F}_{\hat{\boldsymbol{\xi}}_k}| = \frac{1}{2}\ln|\mathbf{C}(\frac{\sum_n z_{nk}}{N})| = \frac{D_{\boldsymbol{\xi}_k}}{2}\ln\frac{\sum_n z_{nk}}{N} + \text{const.}$

BPCA Suppose $\mathbf{X} \in \mathbb{R}^{N \times D}$ is centerized, i.e., $\sum_n \mathbf{x}_n = \mathbf{0}$. Then, the joint likelihood of $\mathbf{X}$ and $\mathbf{Z} \in \mathbb{R}^{N \times K}$ is given by $p(\mathbf{X}, \mathbf{Z}|\mathbf{\Pi}) = \prod_n N(\mathbf{x}_n|\mathbf{W}\mathbf{z}_n, \frac{1}{\lambda}\mathbf{I}) N(\mathbf{z}_n|\mathbf{0}, \mathbf{I})$,

where $\mathbf{\Xi} = \mathbf{W} = (\mathbf{w}_{\cdot 1}, \ldots, \mathbf{w}_{\cdot K})$ is a linear basis and $\mathbf{\Theta} = \lambda$ is the reciprocal of the noise variance. Note that the original study of BPCA (Bishop, 1998) introduces the additional priors $p(\mathbf{W}) = \prod_d N(\mathbf{w}_d|\mathbf{0}, \text{diag}(\boldsymbol{\alpha}^{-1}))$ and $p(\lambda) = \text{Gamma}(\lambda|a_\lambda, b_\lambda)$ and the hyperprior $p(\boldsymbol{\alpha}) = \prod_k \text{Gamma}(\alpha_k|a_{\boldsymbol{\alpha}}, b_{\boldsymbol{\alpha}})$. In this paper, however, we do not specify explicit forms of those priors but just treat them as $O(1)$ term.

Since there is no second-order interaction between $\mathbf{w}_i$ and $\mathbf{w}_{j \neq i}$, the Hessian $\mathbf{F}_{\mathbf{\Xi}}$ is a block-diagonal and each block is given by $\frac{\lambda}{N}\mathbf{Z}^{\top}\mathbf{Z}$. The penalty term is then given as

$$-\frac{1}{2}\ln|\mathbf{F}_{\mathbf{\Xi}}| = -\frac{D}{2}(K \ln\lambda + \ln|\frac{1}{N}\mathbf{Z}^{\top}\mathbf{Z}|), \qquad (5)$$

and $\mathbf{Z}$ is degenerated, if $\text{rank}(\mathbf{Z}) < K$. Suppose that $\mathbf{Z}$ is degenerated, let $K' = \text{rank}(\mathbf{Z}) < K$, and let the SVD be $\mathbf{Z} = \mathbf{U}\text{diag}(\boldsymbol{\sigma})\mathbf{V}^{\top} = (\mathbf{U}_{K'}, \mathbf{0})\text{diag}(\boldsymbol{\sigma}_{K'}, \mathbf{0})(\mathbf{V}_{K'}, \mathbf{0})^{\top}$, where $\mathbf{U}_{K'}$ and $\mathbf{V}_{K'}$ are K' non-zero singular vectors and $\boldsymbol{\sigma}_{K'}$ is K' non-zero singular values. From the definition of $\mathbf{F}_{\mathbf{\Xi}}$, the projection $\mathbf{V}$ removes the degeneration of $\mathbf{Z}$, i.e., by letting $\tilde{\mathbf{Z}} = \mathbf{Z}\mathbf{V}$ and $\tilde{\mathbf{W}} = \mathbf{W}\mathbf{V}$,

$$\ln p(\mathbf{X}, \mathbf{Z} \mid \mathbf{\Pi}, K) = -\frac{\lambda}{2}\|\mathbf{X} - \mathbf{Z}\mathbf{W}^{\top}\|_{\text{F}}^2 - \frac{1}{2}\|\mathbf{Z}\|_{\text{F}}^2 + \text{const.}$$
$$= -\frac{\lambda}{2}\|\mathbf{X} - \tilde{\mathbf{Z}}\tilde{\mathbf{W}}^{\top}\|_{\text{F}}^2 - \frac{1}{2}\|\tilde{\mathbf{Z}}\|_{\text{F}}^2 + \text{const.}$$
$$= \ln p(\mathbf{X}, \tilde{\mathbf{Z}}_{K'} \mid \{\lambda, \tilde{\mathbf{W}}_{K'}\}, K').$$

where $\|\mathbf{A}\|_{\text{F}}^2 = \sum_{ij} a_{ij}^2$ denotes the Frobenius norm, $\tilde{\mathbf{Z}}_{K'} = \mathbf{U}_{K'}\text{diag}(\boldsymbol{\sigma}_{K'})$, and $\tilde{\mathbf{W}}_{K'} = \mathbf{W}\mathbf{V}_{K'}$. $\mathbf{V}$ transforms $K - K'$ redundant components to $\mathbf{0}$-column vectors, and we can find the smaller model K' by removing the $\mathbf{0}$-column vectors from $\tilde{\mathbf{W}}$ and $\tilde{\mathbf{Z}}$, which satisfies **A1**.

3. Derivation of gFIC

To obtain $p(\mathbf{X} \mid K)$, we need to marginalize out two variables: $\mathbf{Z}$ and $\mathbf{\Pi}$. Let us consider the variational form for $\mathbf{Z}$, written as

$$\ln p(\mathbf{X}|K) = \mathbb{E}_q[\ln p(\mathbf{X}, \mathbf{Z}|K)] + H(q) + \text{KL}(q\|q^*) \quad (6)$$
$$= \mathbb{E}_{q^*}[\ln p(\mathbf{X}, \mathbf{Z}|K)] + H(q^*), \qquad (7)$$

where $\text{KL}(q\|p) = \int q(x)\ln q(x)/p(x)dx$ is the Kullback-Leibler (KL) divergence.

Variational representation (7) allows us to consider the cases of whether $\mathbf{Z}$ is degenerated or not separately. In particular, when $\mathbf{Z} \sim q^*(\mathbf{Z})$ is not degenerated, then **A5** guarantees that $p(\mathbf{X}, \mathbf{Z} \mid K)$ is regular, and standard asymptotic results such as Laplace's method are applicable. In contrast, if $q^*(\mathbf{Z})$ is degenerated, $p(\mathbf{X}, \mathbf{Z} \mid K)$ becomes singular and its asymptotic behavior is unclear.

In this section, we analyze the asymptotic behavior of the variational representation (7) in both cases and show that

gFIC is accurate even if $q^*(\mathbf{Z})$ is degenerated. Our main contribution is the following theorem.[1]

Theorem 2. *Let* $K' = \kappa(p(\mathbf{Z} \mid \mathbf{X}, K))$. *Then,*

$$\ln p(\mathbf{X} \mid K) = \text{gFIC}(K') + O(1). \tag{8}$$

We emphasize that the above theorem holds even if the model family does not include the true distribution of $\mathbf{X}$. To prove Theorem 2, we first investigate the asymptotic behavior of $\ln p(\mathbf{X} \mid K)$ for the non-degenerated case.

3.1. Non-degenerated Cases

Suppose K is fixed, and consider the marginalization $p(\mathbf{X}, \mathbf{Z}) = \int p(\mathbf{X}, \mathbf{Z} | \mathbf{\Pi}) p(\mathbf{\Pi}) d\mathbf{\Pi}$. If $p(\mathbf{Z}|\mathbf{X})$ is not degenerated, then $\mathbf{Z} \sim p(\mathbf{Z}|\mathbf{X})$ is not degenerated with probability one. This suffices to guarantee the regularity condition (**A5**) and hence to justify the application of Laplace's method, which approximates $p(\mathbf{X}, \mathbf{Z})$ in an asymptotic manner (Tierney & Kadane, 1986).

Lemma 3. *If* $\mathbf{Z}$ *is not degenerated,* $p(\mathbf{X}, \mathbf{Z}) =$

$$p(\mathbf{X}, \mathbf{Z}, \hat{\mathbf{\Pi}}) |\mathbf{F}_{\hat{\mathbf{\Pi}}}|^{-1/2} \left(\frac{2\pi}{N}\right)^{D_{\mathbf{\Pi}}/2} (1 + O(N^{-1})). \tag{9}$$

This result immediately yields the following relation:

$$\ln p(\mathbf{X}, \mathbf{Z}) = \mathcal{L}(\mathbf{Z}, \hat{\mathbf{\Pi}}, K) + O(1). \tag{10}$$

Substitution of Eq. (10) into Eq. (7) yields Eq. (8). Note that we drop the $O(1)$ terms: $\ln p(\hat{\mathbf{\Pi}})$ (see **A3**), $\frac{D_{\mathbf{\Pi}}}{2} \ln 2\pi$, and a term related to $\mathbf{F}_{\hat{\mathbf{\Theta}}}$ to obtain Eq. (10). We emphasize here that the magnitude of $\mathbf{F}_{\mathbf{\Pi}}$ (and $\mathbf{F}_{\mathbf{\Theta}}$ and $\mathbf{F}_{\mathbf{\Xi}}$) is constant by definition. Therefore, ignoring all of the information of $\mathbf{F}_{\hat{\mathbf{\Pi}}}$ in Eq. (9) just gives another $O(1)$ error and equivalence of gFIC (8) still holds. However, $\mathbf{F}_{\mathbf{\Xi}}$ contains important information of which component is effectively used to captures $\mathbf{X}$. Therefore, we use the relation $\ln|\mathbf{F}_{\mathbf{\Pi}}| = \ln|\mathbf{F}_{\mathbf{\Xi}}| + \ln|\mathbf{F}_{\mathbf{\Xi},\mathbf{\Pi}} \mathbf{F}_{\mathbf{\Pi}}^{-1} \mathbf{F}_{\mathbf{\Xi},\mathbf{\Pi}}^{\top}|$ and remain the first term in gFIC. In Section 4.3, we interpret the effect of $\mathbf{F}_{\mathbf{\Xi}}$ in more detail.

3.2. Degenerated Cases

If $p(\mathbf{Z} \mid \mathbf{X}, K)$ is degenerated, then the regularity condition does not hold, and we cannot use Laplace's method (Lemma 3) directly. In that case, however, **A1** guarantees the existence of a variable transformation $(\mathbf{Z}, \mathbf{\Pi}) \rightarrow (\tilde{\mathbf{Z}}_{K'}, \tilde{\mathbf{\Pi}}_{K'})$ that replaces the joint likelihood by the equivalent yet smaller "regular" model: $p(\mathbf{X}, \mathbf{Z} \mid K) =$

$$\int p(\mathbf{X}, \mathbf{Z} \mid \mathbf{\Pi}, K) p(\mathbf{\Pi} \mid K) d\mathbf{\Pi}$$

$$= \int p(\mathbf{X}, \tilde{\mathbf{Z}}_{K'} \mid \tilde{\mathbf{\Pi}}_{K'}, K') \tilde{p}(\tilde{\mathbf{\Pi}}_{K'} \mid K') d\tilde{\mathbf{\Pi}}_{K'}. \tag{11}$$

---

[1] A formal proof is given in supplemental material.

Since $\tilde{\mathbf{Z}}_{K'}$ is not degenerated in the model K', we can apply Laplace's method and obtain asymptotic approximation (10) by replacing K by K'. Note that the transformed prior $\tilde{p}(\mathbf{\Pi}_{K'} \mid K')$ would differ from the original prior $p(\mathbf{\Pi}_{K'} \mid K')$. However, since the prior does not depend on N (**A3**), the difference is at most $O(1)$, which is asymptotically ignorable.

Eq. (11) also gives us an asymptotic form of the marginal posterior.

Proposition 4.

$$p(\mathbf{Z} \mid \mathbf{X}, K) = p_K(\mathbf{Z})(1 + O(N^{-1})), \tag{12}$$

$$p_K(\mathbf{Z}) \equiv \begin{cases} \frac{p(\mathbf{Z}, \mathbf{X}|\hat{\mathbf{\Pi}}, K)|\mathbf{F}_{\hat{\mathbf{\Pi}}}|^{-1/2}}{C} & K = \kappa(\mathbf{Z}), \\ p_{\kappa(\mathbf{Z})}(\mathbf{T}_{\kappa(\mathbf{Z})}(\mathbf{Z})) & K > \kappa(\mathbf{Z}), \end{cases} \tag{13}$$

where $\mathbf{T}_{K'} : \mathbf{Z} \rightarrow \tilde{\mathbf{Z}}_{K'}$ *as Eq. (3) and* C *is the normalizing constant.*

The above proposition indicates that, if $\kappa(p(\mathbf{Z} \mid \mathbf{X}, K)) = K'$, $p(\mathbf{Z} \mid \mathbf{X}, K)$ is represented by the non-degenerated distribution $p(\mathbf{Z} \mid \mathbf{X}, K')$. Now, we see that the joint likelihood (11) and the marginal posterior (12) depend on K' rather than K. Therefore, putting these results into variational bound (7) leads to (8), i.e., $\ln p(\mathbf{X} \mid K)$ is represented by gFIC of the "true" model K'.

Theorem 2 indicates that, if the model K is overcomplete for the true model K', $\ln p(\mathbf{X} \mid K)$ takes the same value as $\ln p(\mathbf{X} \mid K')$.

Corollary 5. *For every* $K > K' = \kappa(p(\mathbf{Z} \mid \mathbf{X}))$,

$$\ln p(\mathbf{X} \mid K) = \ln p(\mathbf{X} \mid K') + O(1). \tag{14}$$

This implication is fairly intuitive in the sense that if $\mathbf{X}$ concentrates on the subspace of the model, then marginalization with respect to the parameters outside of the subspace contributes nothing to $\ln p(\mathbf{X} \mid K)$. Corollary 5 also justifies model selection of the LVMs on the basis of the marginal likelihood. According to Corollary 5, at $N \rightarrow \infty$ redundant models always take the same value of the marginal likelihood as that of the true model, and we can safely exclude them from model candidates.

4. The gFAB Inference

To evaluate gFIC (2), we need to solve several estimation problems. First, we need to estimate $p(\mathbf{Z} \mid \mathbf{X}, K)$ to minimize the KL divergence in Eq. (6). In addition, since $\ln p(\mathbf{X} \mid K)$ depends on the true model K' (Theorem 2), we need to check whether the current model is degenerated or not, and if it is degenerated, we need to estimate K'. This is paradoxical, because we would like to determine K' through model selection. However, by using the properties of gFIC, we can obtain K' efficiently by optimization.

4.1. Computation of gFIC

By applying Laplace's method to Eq. (11) and substituting it into the variational form (6), we obtain $\ln p(\mathbf{X} \mid K) =$

$$\mathbb{E}_q[\mathcal{L}(\mathbf{Z}, \hat{\mathbf{\Pi}}, \kappa(q))] + H(q) + \mathrm{KL}(q\|q^*) + O(1). \quad (15)$$

Since the KL divergence is non-negative, substituting this into Eq. (8) and ignoring the KL divergence gives a lower bound of gFIC(K'), i.e.,

$$\mathrm{gFIC}(K') \geq \mathbb{E}_q[\mathcal{L}(\mathbf{Z}, \hat{\mathbf{\Pi}}, \kappa(q))] + H(q). \quad (16)$$

This formulation allows us to estimate gFIC(K') via maximizing the lower bound. Moreover, we no longer need to know K'—if the initial dimension of q is greater than K', the maximum of lower bound (16) attains gFIC(K') and thus $\ln p(\mathbf{X} \mid K')$. Similarly to other variational inference problems, this optimization is solved by iterative maximization of q and $\mathbf{\Pi}$.

4.1.1. UPDATE OF q

As suggested in Eq. (15), the maximizer of lower bound (16) is $p(\mathbf{Z} \mid \mathbf{X})$ in which the asymptotic form is shown in Proposition 4. Unfortunately, we cannot use this as q, because the normalizing constant is intractable. One helpful tool is the mean-field approximation of q, i.e., $q(\mathbf{Z}) = \prod_n q_n(\mathbf{z}_n)$. Although the asymptotic marginal posterior (12) depends on n due to $\mathbf{F}_{\mathbf{\Pi}}$, this dependency eventually vanishes for $N \to \infty$, and the mean-field approximation still maintains the asymptotic consistency of gFIC.

Proposition 6. *Suppose $p(\mathbf{Z}|\mathbf{X}, K)$ is not degenerated in distribution. Then, $p(\mathbf{Z}|\mathbf{X}, K)$ converges to $p(\mathbf{Z}|\mathbf{X}, \hat{\mathbf{\Pi}}, K)$, and $p(\mathbf{Z}|\mathbf{X}, K)$ is asymptotically mutually independent for $\mathbf{z}_1, \ldots, \mathbf{z}_n$.*

In some models, such as MMs (Fujimaki & Morinaga, 2012), the mean-field approximation suffices to solve the variational problem. If it is still intractable, other approximations are necessary. For example, we restrict q as the Gaussian density $q(\mathbf{Z}) = \prod_n N(\mathbf{z}_n|\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$ for BPCA which we use in the experiments (Section 7).

4.1.2. UPDATE OF $\mathbf{\Pi}$

After obtaining q, we need to estimate $\hat{\mathbf{\Pi}}$ for each sample $\mathbf{Z} \sim q(\mathbf{Z})$, which is also intractable. Alternatively, we estimate the expected MJLE $\bar{\mathbf{\Pi}} = \mathrm{argmax}_{\mathbf{\Pi}} \mathbb{E}_q[\ln p(\mathbf{X}, \mathbf{Z} \mid \mathbf{\Pi})]$. Since the max operator has convexity, Jensen's inequality shows that replacing $\hat{\mathbf{\Pi}}$ by $\bar{\mathbf{\Pi}}$ introduces the following lower bound.

$$\mathbb{E}_q[\ln p(\mathbf{X}, \mathbf{Z} \mid \hat{\mathbf{\Pi}})] = \mathbb{E}_q[\max_{\mathbf{\Pi}} \ln p(\mathbf{X}, \mathbf{Z} \mid \mathbf{\Pi})]$$
$$\geq \mathbb{E}_q[\ln p(\mathbf{X}, \mathbf{Z} \mid \bar{\mathbf{\Pi}})] = \max_{\mathbf{\Pi}} \mathbb{E}_q[\ln p(\mathbf{X}, \mathbf{Z} \mid \mathbf{\Pi})].$$

Algorithm 1 The gFAB algorithm

> **Input:** data $\mathbf{X}$, initial model K, threshold δ
> **repeat**
> $\quad q \leftarrow \mathrm{argmax}_{q \in \mathcal{Q}} \mathbb{E}_q[\mathcal{L}(\mathbf{Z}, \bar{\mathbf{\Pi}}, \kappa(q))] + H(q)$
> $\quad$ **if** $\sigma_K(\mathbf{F}_{\mathbf{\Xi}}) \leq \cdots \leq \sigma_{K'}(\mathbf{F}_{\mathbf{\Xi}}) \leq \delta$ **then**
> $\quad\quad K \leftarrow K'$ and $(\mathbf{Z}, \bar{\mathbf{\Pi}}) \leftarrow (\tilde{\mathbf{Z}}_{K'}, \tilde{\mathbf{\Pi}}_{K'})$
> $\quad$ **end if**
> $\quad \bar{\mathbf{\Pi}} \leftarrow \mathrm{argmax}_{\mathbf{\Pi}} \mathbb{E}_q[\ln p(\mathbf{X}, \mathbf{Z} \mid \mathbf{\Pi}, K)]$
> **until** Convergence

Since $\bar{\mathbf{\Pi}}$ depends only on q, we now need to compute the parameter only once. Remarkably, $\hat{\mathbf{\Pi}}$ is consistent with $\bar{\mathbf{\Pi}}$ and the above equality holds asymptotically.

Proposition 7. *If $q(\mathbf{Z})$ is not degenerated in distribution, then $\hat{\mathbf{\Pi}} \xrightarrow{\mathrm{p}} \bar{\mathbf{\Pi}}$.*

Since $\mathbb{E}_q[\ln p(\mathbf{X}, \mathbf{Z} \mid \mathbf{\Pi})]$ is the average of the concave function (**A5**), $\mathbb{E}_q[\ln p(\mathbf{X}, \mathbf{Z} \mid \mathbf{\Pi})]$ itself is also concave and the estimation of $\bar{\mathbf{\Pi}}$ is relatively easy. If the expectations $\mathbb{E}_q[\ln p(\mathbf{X}, \mathbf{Z}|\mathbf{\Pi})]$ and $\mathbb{E}_q[\mathbf{F}_{\mathbf{\Pi}}]$ are analytically written, then gradient-based optimization suffices for the estimation. If these is no analytic form, then stochastic optimization, such as stochastic gradient assuming $\mathbf{Z} \sim q(\mathbf{Z})$ as a sample (Kingma & Welling, 2013), might help.

4.1.3. MODEL PRUNING

During the optimization of q, it can become degenerated or nearly degenerated. In such a case, by definition of objective (16), we need to change the form of $\mathcal{L}(\mathbf{Z}, \hat{\mathbf{\Pi}}, K)$ to $\mathcal{L}(\mathbf{Z}, \hat{\mathbf{\Pi}}, K')$. This can be accomplished by using the transformation $(\mathbf{Z}, \mathbf{\Pi}) \to (\tilde{\mathbf{Z}}_{K'}, \tilde{\mathbf{\Pi}}_{K'})$ and decreasing the current model from K to K', i.e., removing degenerated components. We refer to this operation as "model pruning". We practically verify the degeneration by the rank of $\mathbf{F}_{\mathbf{\Xi}}$, i.e., we perform model pruning if the eigenvalues are less than some threshold.

4.2. The gFAB Algorithm

Algorithm 1 summarizes the above procedures, solving the following optimization problem:

$$\max_{q \in \mathcal{Q}} \mathbb{E}_q[\mathcal{L}(\mathbf{Z}, \bar{\mathbf{\Pi}}(q), \kappa(q))] + H(q), \quad (17)$$

where $\mathcal{Q} = \{q(\mathbf{Z}) \mid q(\mathbf{Z}) = \prod_n q_n(\mathbf{z}_n)\}$. As shown in Propositions 6 and 7, the above objective is the lower bound of Eq. (16) and thus of gFIC(K'), and the equality holds asymptotically.

Corollary 8.

$$\begin{cases} \mathrm{gFIC}(K') = Eq.\ (17) & \text{for } N \to \infty, \\ \mathrm{gFIC}(K') \geq Eq.\ (17) & \text{for a finite } N > 0. \end{cases} \quad (18)$$

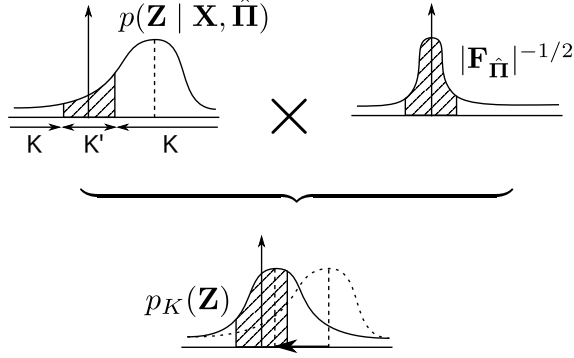Figure 1. The gFIC penalty $|\mathbf{F}_{\hat{\Xi}}|^{-1/2}$ changes the shape of the posterior $p(\mathbf{Z} \mid \mathbf{X}, \hat{\mathbf{\Pi}})$ as increasing the probability of degenerated $\mathbf{Z}$ (indicated by diagonal stripes).

The gFAB algorithm is the block coordinate ascent. Therefore, if the pruning threshold δ is sufficiently small, each step monotonically increases objective (17), and the algorithm stops at critical points.

A unique property of the gFAB algorithm is that it estimates the true model K' along with the updates of q and $\bar{\mathbf{\Pi}}$. If N is sufficiently large and the initial model $K_{\max}$ is larger than K', the algorithm learns $p_{K'}(\mathbf{Z})$ as q, according to Proposition 4. At the same time, model pruning removes degenerated $K - K'$ components. Therefore, if the solutions converge to the global optima, the gFAB algorithm returns K'.

4.3. How $\mathbf{F_\Pi}$ Works?

Proposition 4 shows that if the model is not degenerated, objective (17) is maximized at $q(\mathbf{Z}) = p_K(\mathbf{Z}) \propto p(\mathbf{Z}|\mathbf{X}, \hat{\mathbf{\Pi}})|\mathbf{F}_{\hat{\mathbf{\Pi}}}|^{-1/2}$, which is the product of the unmarginalized posterior $p(\mathbf{Z}|\mathbf{X}, \hat{\mathbf{\Pi}})$ and the gFIC penalty term $|\mathbf{F}_{\hat{\mathbf{\Pi}}}|^{-1/2}$. Since $|\mathbf{F}_{\hat{\mathbf{\Pi}}}|^{-1/2}$ has a peak where $\mathbf{Z}$ is degenerating, it changes the shape of $p(\mathbf{Z}|\mathbf{X}, \hat{\mathbf{\Pi}})$ and increases the probability that $\mathbf{Z}$ is degenerated. Figure 1 illustrates how the penalty term affects the posterior.

Note that, if the model family contains the true distribution of $\mathbf{X}$, then $\mathbf{F}_{\hat{\mathbf{\Pi}}}$ converges to the Fisher information matrix. From another viewpoint, $\mathbf{F}_{\hat{\mathbf{\Pi}}}$ is interpreted as the covariance matrix of the asymptotic posterior of $\mathbf{\Pi}$. As a result of applying the Bernstein-von Mises theorem, the asymptotic normality holds for the posterior $p(\mathbf{\Pi}|\mathbf{X}, \mathbf{Z})$ in which the covariance is given by $(N\mathbf{F}_{\hat{\mathbf{\Pi}}})^{-1}$.

Proposition 9. *Let* $\mathbf{\Omega} = \sqrt{N}(\mathbf{\Pi} - \hat{\mathbf{\Pi}})$. *Then, if* $\mathbf{Z}$ *is not degenerated,* $|p(\mathbf{\Omega} \mid \mathbf{X}, \mathbf{Z}) - N(\mathbf{0}, \mathbb{E}[\mathbf{F}_{\hat{\mathbf{\Pi}}}]^{-1})| \xrightarrow{\mathrm{p}} 0$.

This interpretation has the following implication. In maximizing the variational lower bound (7), we maximize $-\frac{1}{2} \ln |\mathbf{F}_{\hat{\Xi}}|$. In the gFAB algorithm, this is equivalent to

maximize the posterior covariance and pruning the components where those covariance diverge to infinity. Divergence of the posterior covariance means that there is insufficient information to determine those parameters, which are not necessary for the model and thus can reasonably be removed.

5. Relationship with VB

Similarly to FAB, VB alternatingly optimizes with respect to $\mathbf{Z}$ and $\mathbf{\Pi}$, whereas VB treats both of them as distributions. Suppose $K \leq K'$, i.e., the case when the posterior $p(\mathbf{Z} \mid \mathbf{X}, K')$ is not degenerated in distribution. Then, the marginal log-likelihood is written by the variational lower bound: $\ln p(\mathbf{X} \mid K') =$

$$\mathbb{E}_{q(\mathbf{Z},\mathbf{\Pi})}[\ln p(\mathbf{X}, \mathbf{Z}, \mathbf{\Pi} \mid K')] + H(q(\mathbf{Z}, \mathbf{\Pi}))$$
$$+ \mathrm{KL}(q(\mathbf{Z}, \mathbf{\Pi})\|p(\mathbf{Z}, \mathbf{\Pi} \mid \mathbf{X}, K'))$$
$$\geq \mathbb{E}_{q(\mathbf{Z},\mathbf{\Pi})}[\ln p(\mathbf{X}, \mathbf{Z}, \mathbf{\Pi} \mid K')] + H(q(\mathbf{Z}, \mathbf{\Pi}))$$
$$\geq \mathbb{E}_{q(\mathbf{Z})q(\mathbf{\Pi})}[\ln p(\mathbf{X}, \mathbf{Z}, \mathbf{\Pi} \mid K')] + H(q(\mathbf{Z})) + H(q(\mathbf{\Pi})),$$
$$(19)$$

where we use the mean-field approximation $q(\mathbf{Z}, \mathbf{\Pi}) = q(\mathbf{\Pi})q(\mathbf{Z})$ in the last line. Minimizing the KL divergence yields the maximizers of Eq. (19), given as

$$\tilde{q}(\mathbf{\Pi}) \propto \exp\left(\mathbb{E}_{q(\mathbf{Z})}[\ln p(\mathbf{X}, \mathbf{Z}, \mathbf{\Pi} \mid K')]\right), \quad (20)$$
$$\tilde{q}(\mathbf{Z}) \propto \exp\left(\mathbb{E}_{q(\mathbf{\Pi})}[\ln p(\mathbf{X}, \mathbf{Z}, \mathbf{\Pi} \mid K')]\right). \quad (21)$$

Here, we look inside the optimal distributions to see the relationship with the gFAB algorithm. Let us consider to restrict the density $q(\mathbf{\Pi})$ to be Gaussian. Since $\mathbb{E}_{q(\mathbf{Z})}[\ln p(\mathbf{X}, \mathbf{Z} \mid \mathbf{\Pi}, K')]$ increases proportional to N while $\ln p(\mathbf{\Pi})$ does not, $\tilde{q}(\mathbf{\Pi})$ attains its maximum around $\bar{\mathbf{\Pi}}$. Then, the second order expansion to $\ln \tilde{q}(\mathbf{\Pi})$ at $\bar{\mathbf{\Pi}}$ yields the solution $\tilde{q}(\mathbf{\Pi}) = N(\bar{\mathbf{\Pi}}, (N\mathbf{F}_{\bar{\mathbf{\Pi}}})^{-1})^2$. We remark that this solution can be seen as an empirical version of the asymptotic normal posterior given by Proposition 9. Then, if we further approximate $\ln p(\mathbf{X}, \mathbf{Z} \mid \mathbf{\Pi}, K')$ by the second order expansion at $\mathbf{\Pi} = \bar{\mathbf{\Pi}}$, the other expectation $\mathbb{E}_{q(\mathbf{\Pi})}[\ln p(\mathbf{X}, \mathbf{Z} \mid \mathbf{\Pi}, K')]$ appearing in $\tilde{q}(\mathbf{Z})$ is evaluated by $\ln p(\mathbf{X}, \mathbf{Z} \mid \bar{\mathbf{\Pi}}, K') - \frac{1}{2} \ln |\mathbf{F}_{\bar{\mathbf{\Pi}}}|$. Under these approximations, alternating updates of $\{\bar{\mathbf{\Pi}}, F_{\bar{\mathbf{\Pi}}}\}$ and $\tilde{q}(\mathbf{Z})$ coincide exactly with the gFAB algorithm[3], which justifies the VB lower bound as an asymptotic expansion of $p(\mathbf{X} \mid K')$.

Proposition 10. *Let* $L_{\mathrm{VB}}(K)$ *be the VB lower bound* (19) *with restricting* $q(\mathbf{\Pi})$ *to be Gaussian and approximating the expectation in* $\ln \tilde{q}(\mathbf{Z})$ *by the second order expansion. Then, for* $K \leq K'$, $\ln p(\mathbf{X} \mid K) = L_{\mathrm{VB}}(K) + O(1)$.

---

[2]The factor N is included in the covariance, because the negative Hessian of $\ln p(\mathbf{X}, \mathbf{Z} \mid \mathbf{\Pi}, K')$ is $N\mathbf{F_\Pi}$; see the definition of $\mathbf{F_\Pi}$.

[3]Note that model pruning is not necessary when $K \leq K'$.

It is worth noting that these second order expansions are valid only for the regular model, and the asymptotic behavior of $L_{\mathrm{VB}}(K)$ is unclear for $K > K'$.

Proposition 10 suggests that the mean-field approximation does not loose the consistency with $p(\mathbf{X} \mid K)$. This is because that the posterior of $\mathbf{\Pi}$ asymptotically converges to the point $\bar{\mathbf{\Pi}}$ rather than a distribution (for $K < K'$). As shown in Proposition 9, the posterior covariance is given by $(N\mathbf{F}_{\hat{\mathbf{\Pi}}})^{-1}$, which goes to 0 for $N \to \infty$. Therefore, mutual information in the posterior among $\mathbf{Z}$ and $\mathbf{\Pi}$ eventually diminishes, and the mean-field approximation is asymptotically justified. This observation also allows further employment of the mean-field approximation to $q(\mathbf{\Pi})$. For example, BPCA has two parameters $\mathbf{\Pi} = \{\mathbf{W}, \lambda\}$ (see Section 2.1), in which the joint distribution $\tilde{q}(\mathbf{W}, \lambda)$ has no analytical solution. However, the independence assumption $q(\mathbf{W}, \lambda) = q(\mathbf{W})q(\lambda)$ gives us analytical solutions of $\tilde{q}(\mathbf{W})$, $\tilde{q}(\lambda)$, and $\tilde{q}(\mathbf{Z})$ under suitable conjugate priors. As discussed above, since both $\mathbf{W}$ and λ converge to points, this approximation still maintains Proposition 10.

6. Related Work

The EM Algorithm Algorithm 1 looks quite similar to the EM algorithm, solving

$$\max_{q, \mathbf{\Pi}} \mathbb{E}_q[\ln p(\mathbf{X} \mid \mathbf{\Pi}, K)] + H(q). \tag{22}$$

We see that both gFAB and EM algorithms iteratively update the posterior-like distribution of $\mathbf{Z}$ and estimate $\mathbf{\Pi}$. The essential difference between them is that the EM algorithm infers the posterior $p(\mathbf{Z}|\mathbf{X}, \hat{\mathbf{\Pi}})$ in the E-step, but the gFAB algorithm infers the *marginal* posterior $p(\mathbf{Z}|\mathbf{X}) \simeq p(\mathbf{Z}|\mathbf{X}, \hat{\mathbf{\Pi}})|\mathbf{F}_{\hat{\mathbf{\Pi}}}|^{-1/2}$. As discussed in Section 4.3, the penalty term $|\mathbf{F}_{\hat{\mathbf{\Pi}}}|^{-1/2}$ increases the probability mass of the posterior, where $\mathbf{Z}$ is degenerating, enabling automatic model determination through model pruning. In contrast, the EM algorithm lacks such pruning mechanism, and always overfits to $\mathbf{X}$ as long as N is finite while $p(\mathbf{Z}|\mathbf{X})$ eventually converges to $p(\mathbf{Z}|\mathbf{X}, \hat{\mathbf{\Pi}})$ for $N \to \infty$ (see Proposition 6).

Note that Eq. (22) has $O(\ln N)$ error against $\ln p(\mathbf{X})$. Analogously to gFIC, this error is easily reduced to $O(1)$ by adding $-\frac{D_{\mathbf{\Pi}}}{2} \ln N$. This modification provides another information criterion, which we refer to as *BICEM*.

VB Methods The relationship between the VB and gFAB algorithms is discussed in the previous section.

Collapsed VB (CVB) (Teh et al., 2006) is a variation of VB. Similarly to FAB, CVB takes the variational bound after marginalizing out $\mathbf{\Pi}$ from the joint likelihood. In contrast to FAB, CVB approximates q in a non-asymptotic manner, such as the first-order Taylor expansion (Asuncion et al.,

2009). Although such approximation has been found to be accurate in practice, its asymptotic properties, such as consistency, have not been explored. Note that as one of those approximations, the mean-field assumption $q(\mathbf{Z}) \in \mathcal{Q}$ is used in the original paper on CVB (Teh et al., 2006), motivated by the intuition that the dependence among $\{\mathbf{z}_n\}$ is weak after marginalization. Proposition 6 justifies this intuition theoretically.

Several authors have studied about asymptotic behaviors of VB methods for LVMs. Wang & Titterington (2004) investigated the VB approximation for linear dynamical systems (a.k.a. Kalman filter) and showed the inconsistency of VB estimation with large observation noise. Watanabe & Watanabe (2006) derived an asymptotic variational lower bound of the Gaussian MMs and demonstrated its usefulness for the model selection. Recently, Nakajima et al. (2014) analyzed the VB learning on latent Dirichlet allocation (LDA) (Blei et al., 2003), who revealed conditions for the consistency and clarified its transitional behavior of the parameter sparsity. By comparing with these existing works, we have a contribution in terms of that our asymptotic analysis is valid for *general* LVMs, rather than individual models.

BIC and Extensions Let $\mathbf{Y} = (\mathbf{X}, \mathbf{Z})$ be a pair of non-degenerated $\mathbf{X}$ and $\mathbf{Z}$. By ignoring all the constant terms of Laplace's approximation (9), we obtain BIC (Schwarz, 1978) considering $\mathbf{Y}$ as an observation, which is given by the right-hand side of the following equation.

$$\ln p(\mathbf{Y} \mid K) = \ln p(\mathbf{Y} \mid \hat{\mathbf{\Pi}}, K) - \frac{D_{\mathbf{\Pi}}}{2} \ln N + O(1).$$

Unfortunately, the above relation does not hold for $p(\mathbf{X} \mid K)$. Since $p(\mathbf{X} \mid K) = \int p(\mathbf{Y} \mid K) d\mathbf{Z}$ mixes up degenerated and non-degenerated cases, $p(\mathbf{X} \mid K)$ always becomes singular, loosing the condition **A5** that Laplace's approximation holds.

There are several studies that extend BIC to be able to deal with singular models. Watanabe (Watanabe, 2009) evaluates $p(\mathbf{X} \mid K)$ with an $O(1)$ error for any singular models by using algebraic geometry. However, it requires an evaluation of the intractable rational number called the real log canonical threshold. Recent study (Watanabe, 2013) relaxes this intractable evaluation to the evaluation of criterion called WBIC at the expense of an $O_p(\sqrt{\ln N})$ error. Yet, the evaluation of WBIC needs an expectation with respect to a practically intractable distribution, which usually incurs heavy computation.

7. Numerical Experiments

We compare the performance of model selection for BPCA explained in Section 2.1 with the EM algorithm, BICEM in-

	EM	BICEM†	VB	CVB	FAB	gFAB†
Objective	Eq. (22)	Eq. (22)$-\frac{D_{\mathbf{\Pi}}}{2}\ln N$	Eq. (19)	Eq. (7)	Eq. (16)	
$\mathbf{\Pi}$	Point estimate		Posterior w/ MF	Marginalized out	Laplace approximation	
$q(\mathbf{Z})$	$= p(\mathbf{Z} \mid \mathbf{X}, \hat{\mathbf{\Pi}})$		$\simeq p(\mathbf{Z} \mid \mathbf{X})$		$\propto p(\mathbf{Z} \mid \mathbf{X})(1 + O(1))^\dagger$	
$\ln p(\mathbf{X}\mid K \le K')$	$O(\ln N)$	$O(1)^\dagger$	$O(1)^\dagger$		$O(1)^\dagger$	
$\ln p(\mathbf{X}\mid K > K')$	NA		NA		$O(1)^\dagger$	
Applicability	Many models		Many models	Binary LVMs	Binary LVMs	LVMs

Table 1. A comparison of approximated Bayesian methods. The symbol † highlights our contributions. "MF" stands for the mean-field approximation. Note that the asymptotic relations with $\ln p(\mathbf{X} \mid K)$ hold only for LVMs.
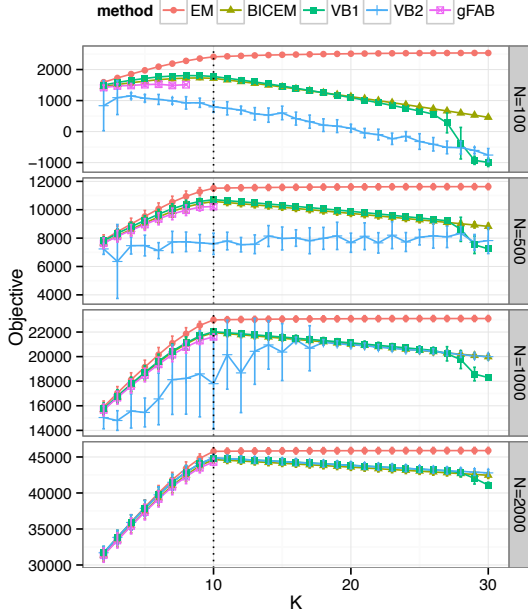

Figure 2. The objective function versus the model K. The error-bar shows the standard deviations over 10 different random seeds, which affect both data and initial values of the algorithms.

troduced in Section 6, simple VB (VB1), full VB (VB2), and the gFAB algorithm. VB2 had the priors for $\mathbf{W}, \lambda$, and $\boldsymbol{\alpha}$ described in Section 2.1 in which the hyperparameters were fixed as $a_\lambda = b_\lambda = a_{\boldsymbol{\alpha}} = b_{\boldsymbol{\alpha}} = 0.01$ by following (Bishop, 1999). VB1 is a simple variant of VB2, which fixed $\boldsymbol{\alpha} = \mathbf{1}$. In this experiments, We used the synthetic data $\mathbf{X} = \mathbf{Z}\mathbf{W}^\top + \mathbf{E}$ where $\mathbf{W} \sim \text{uniform}([0, 1])^4$, $\mathbf{Z} \sim N(\mathbf{0}, \mathbf{I})$, and $E_{nd} \sim N(0, \sigma^2)$. Under the data dimensionality $D = 30$ and the true model $K' = 10$, we generated data with $N = 100, 500, 1000$, and 2000. We stopped the algorithms if the relative error was less than 10^{-5} or the number of iterations was greater than 10^4.

Figure 2 depicts the objective functions after convergence

---

[4]This setting could be unfair because VB1 and VB2 assume the Gaussian prior for $\mathbf{W}$. However, we confirmed that data generated by $\mathbf{W} \sim N(0, 1)$ gave almost the same results.

for $K = 2, \ldots, 30$. Note that, we performed gFAB with $K = 30$ and it finally converged at $K \simeq 10$ owing to model pruning, which allowed us to skip the computation for $K \simeq 10, \ldots, 30$, and the objective values for those Ks are not drawn. We see that gFAB underestimated the model when the number of samples were small ($N < 500$), but it successfully chose $K = 10$ with sufficiently large sample sizes ($N \ge 500$). In contrast, the objective of EM slightly but monotonically increased with K, which means EM always chose the largest K as the best model. This is because EM maximizes Eq. (22), which does not impose the penalty on the model complexity brought by the marginalization of $\mathbf{\Pi}$. As our analysis suggested in Section 6, BICEM and VB1 are close to gFAB as N increasing and has a peak around $K' = 10$, meaning that BICEM and VB1 are adequate for model selection. However, in contrast to gFAB, both of them need to compute for all K. Interestingly, VB2 were unstable for $N < 2000$ and it gave the inconsistent model selection results. We observed that VB2 had very strong dependence on the initial values. This behavior is understandable because VB2 has the additional prior and hyperparameters to be estimated, which might produce additional local minima that make optimization difficult.

8. Conclusion

This paper provided an asymptotic analysis for the marginal log-likelihood on LVMs. As the main contribution, we proposed gFIC for model selection and showed its consistency with the marginal log-likelihood. Part of our analysis also provided insight into the EM and VB methods. Numerical experiments confirmed the validity of our analysis.

We remark that gFIC is potentially applicable to many other LVMs, including factor analysis, LDA, canonical correlation analysis, and partial membership models. Investigating the behavior of gFIC on these models is an important future research direction.

References

Asuncion, Arthur, Welling, Max, Smyth, Padhraic, and Teh, Yee Whye. On smoothing and inference for topic models. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI)*, 2009.

Attias, Hagai. Inferring Parameters and Structure of Latent Variable Models by Variational Bayes. In *Uncertainty in Artificial Intelligence (UAI)*, 1999.

Bishop, C. M. Bayesian pca. In *Advances in Neural Information Processing Systems (NIPS)*, 1998.

Bishop, C. M. Variational principal components. In *International Conference on Artificial Neural Networks (ICANN)*, 1999.

Blei, David M., Ng, Andrew Y., and Jordan, Michael I. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

Eto, Riki, Fujimaki, Ryohei, Morinaga, Satoshi, and Tamano, Hiroshi. Fully-automatic bayesian piecewise sparse linear models. In *AISTATS*, 2014.

Fujimaki, Ryohei and Hayashi, Kohei. Factorized asymptotic bayesian hidden markov model. In *International Conference on Machine Learning (ICML)*, 2012.

Fujimaki, Ryohei and Morinaga, Satoshi. Factorized asymptotic bayesian inference for mixture modeling. In *AISTATS*, 2012.

Hastings, W. K. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

Hayashi, Kohei and Fujimaki, Ryohei. Factorized asymptotic bayesian inference for latent feature models. In *27th Annual Conference on Neural Information Processing Systems (NIPS)*, 2013.

Jordan, Michael I., Ghahramani, Zoubin, Jaakkola, Tommi S., and Saul, Lawrence K. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.

Kingma, Diederik P. and Welling, Max. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013.

Nakajima, Shinichi, Sato, Issei, Sugiyama, Masashi, Watanabe, Kazuho, and Kobayashi, Hiroko. Analysis of variational bayesian latent dirichlet allocation: Weaker sparsity than MAP. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

Schwarz, Gideon. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464, 1978.

Teh, Yee W., Newman, David, and Welling, Max. A collapsed variational bayesian inference algorithm for latent dirichlet allocation. In *19th Annual Conference on Neural Information Processing Systems (NIPS)*, 2006.

Tierney, Luke and Kadane, Joseph B. Accurate Approximations for Posterior Moments and Marginal Densities. *Journal of the American Statistical Association*, 81 (393):82–86, 1986.

Wang, Bo and Titterington, D.M. Lack of consistency of mean field and variational bayes approximations for state space models. *Neural Processing Letters*, 20(3):151–170, 2004.

Watanabe, Kazuho and Watanabe, Sumio. Stochastic complexities of gaussian mixtures in variational bayesian approximation. *Journal of Machine Learning Research*, 7:625–644, 2006.

Watanabe, Sumio. *Algebraic Geometry and Statistical Learning Theory*. Cambridge University Press, 2009.

Watanabe, Sumio. A widely applicable bayesian information criterion. *Journal of Machine Learning Research*, 14(1):867–897, 2013.