Rebuilding Factorized Information Criterion: Asymptotically Accurate Marginal Likelihood

Kohei Hayashi\textsuperscript{1,2}  Shin-ichi Maeda\textsuperscript{3}  Ryohei Fujimaki\textsuperscript{4}

\textsuperscript{1}National Institute of Informatics
\textsuperscript{2}Kawarabayashi Large Graph Project, ERATO, JST
\textsuperscript{3}Kyoto University
\textsuperscript{4}NEC Knowledge Discovery Laboratories

September 18, 2015
Introduction

Factorized asymptotic Bayesian inference (FAB)

- Recently-developed approximate Bayesian method

✓ Accurate and tractable

✗ Limited to binary latent variable models (LVMs)
Introduction

Factorized asymptotic Bayesian inference (FAB)

- Recently-developed approximate Bayesian method
  ✔️ Accurate and tractable
  ✖️ Limited to binary latent variable models (LVMs)

Our contributions:

- Extend FAB to general LVMs (e.g. PCA)
- Analyze theoretical properties that are unclear in the previous studies
1. Revisiting FAB
2. Generalization of FAB
Bayesian Inference for Binary LVMs

**Binary LVM:**

\[
p(X, Z, \Pi \mid K) = p(\Pi) p(X, Z \mid \Pi, K)
\]

**Assumptions:**
- \(X\) and \(Z\) are jointly i.i.d.
- The prior doesn't depend on \(N\)
- \(\ln p(\Pi) = O(1)\)
- "Flat" prior
Bayesian Inference for Binary LVMs

Binary LVM:

\[
p(\underbrace{X, Z, \Pi}_{\text{data, LVs, params}}, K) = p(\Pi)p(\underbrace{X, Z | \Pi, K}_{\text{joint likelihood}})\]

Assumptions:

- \(X\) and \(Z\) are jointly i.i.d.

\[
p(X, Z | \Pi, K) = \prod_{n=1}^{N} p(x_n, z_n | \Pi, K)
\]

- The prior doesn’t depend on \(N\)
  - \(\ln p(\Pi) = O(1)\)
  - “Flat” prior
Goal: To obtain

- the marginal likelihood:

\[ p(X|K) = \int p(X, Z, \Pi|K) dZd\Pi \]
Goal: To obtain

- the marginal likelihood:

\[ p(X|K) = \int p(X, Z, \Pi|K) dZ d\Pi \]

- the marginal posteriors:

\[ p(Z|X, K) = \int p(X, Z, \Pi|K) d\Pi / p(X|K) \]
\[ p(\Pi|X, K) = \int p(X, Z, \Pi|K) dZ / p(X|K) \]
Goal: To obtain

- the marginal likelihood:

\[ p(X|K) = \int p(X, Z, \Pi|K) dZ d\Pi \]

- the marginal posteriors:

\[ p(Z|X, K) = \int p(X, Z, \Pi|K) d\Pi / p(X|K) \]
\[ p(\Pi|X, K) = \int p(X, Z, \Pi|K) dZ / p(X|K) \]

Problem: The marginalizations are intractable
Key idea: Use

- the variational representation for $\int dZ$
- Laplace’s method for $\int d\Pi$
**Key idea:** Use
- the variational representation for $\int d\mathbf{Z}$
- Laplace’s method for $\int d\Pi$

**Factorized information criterion (FIC)**

$$
\text{FIC}(K) \equiv \max_{q} \mathbb{E}_q \left[ \max_{\Pi} \ln p(\mathbf{X}, \mathbf{Z}|\Pi, K) \right] \\
- \mathbb{E}_q \left[ \frac{D_{\Pi}}{2} \sum_k \ln \sum_n \z_{nk} \right] + H(q) + O(\ln N)
$$

- $q(\mathbf{Z})$: trial distribution
- $H(q)$: entropy
Accuracy of FIC

✔ Asymptotically equivalent to the marginal likelihood
Accuracy of FIC

✔ Asymptotically equivalent to the marginal likelihood

Theorem 3 of [Fujimaki+ 12a]

In mixture models, under mild conditions,

\[ \text{FIC}(K) = \ln p(X|K) + O(1) \approx \ln p(X|K) \]
Accuracy of FIC

✔ Asymptotically equivalent to the marginal likelihood

Theorem 3 of [Fujimaki+ 12a]

In mixture models, under mild conditions,

\[
FIC(K) = \ln p(X|K) + O(1) \\
\approx \ln p(X|K)
\]

Similar results are obtained for:

- HMMs [Fujimaki+ 12b]
- Latent feature models [KH+ 13]
- Mixture of experts [Eto+ 14]
- Factorial relational models [Liu+ yesterday]
Optimizing FIC

Computation of FIC is difficult

\[
\max_q \mathbb{E}_q \left[ \max_{\Pi} \ln p(X, Z|\Pi, K) \right] - \frac{D_{\Pi}}{2} \sum_k \mathbb{E}_q \left[ \ln \sum_n z_{nk} \right] + H(q)
\]
Optimizing FIC

Computation of FIC is difficult

\[
\max_q \mathbb{E}_q \left[ \max_{\Pi} \ln p(X, Z|\Pi, K) \right] - \frac{D_{\Pi}}{2} \sum_k \mathbb{E}_q \left[ \ln \sum_n z_{nk} \right] + H(q)
\]

\[
\geq \max_{q \in Q} \mathbb{E}_q \left[ \max_{\Pi} \ln p(X, Z|\Pi, K) \right] - \frac{D_{\Pi}}{2} \sum_k \mathbb{E}_q \left[ \ln \sum_n z_{nk} \right] + H(q)
\]

Mean-field approx. \((Q \equiv \{q(Z)|q(Z) = \prod_n q(z_n)\})\)
Optimizing FIC

Computation of FIC is difficult

$$\max_q \mathbb{E}_q \left[ \max_{\Pi} \ln p(X, Z|\Pi, K) \right] - \frac{D_{\Pi}}{2} \sum_k \mathbb{E}_q \left[ \ln \sum_n z_{nk} \right] + H(q)$$

$$\geq \max_{q \in Q} \mathbb{E}_q \left[ \max_{\Pi} \ln p(X, Z|\Pi, K) \right] - \frac{D_{\Pi}}{2} \sum_k \mathbb{E}_q \left[ \ln \sum_n z_{nk} \right] + H(q)$$

Mean-field approx. ($$Q \equiv \{ q(Z) | q(Z) = \prod_n q(z_n) \}$$)

$$\geq \max_{q \in Q, \Pi} \mathbb{E}_q \left[ \ln p(X, Z|\Pi, K) \right] - \frac{D_{\Pi}}{2} \sum_k \ln \sum_n \mathbb{E}_q [z_{nk}] + H(q)$$

Jensen’s ineq.

$$\equiv \text{FIC}(K)$$
Algorithm

Optimization problem:

\[
\max_{q \in Q, \Pi} \mathbb{E}_q \left[ \ln p(X, Z|\Pi, K) \right] - \frac{D_{\Pi}}{2} \sum_k \ln \sum_n \mathbb{E}_q [z_{nk}] + H(q)
\]
Algorithm

Optimization problem:

$$\max_{q \in Q, \Pi} \mathbb{E}_q [\ln p(\mathbf{X}, \mathbf{Z}|\Pi, K)] - \frac{D_{\Pi}}{2} \sum_k \ln \sum_n \mathbb{E}_q [z_{nk}] + H(q)$$

Can be solved by EM-like alternating updates:

1. Initialize $q$ and $\Pi$
2. Update $q$ (Fix $\Pi$)
3. Update $\Pi$ (Fix $q$)
4. Repeat step 2 and 3 until convergence
Model Pruning

The FAB algorithm eliminates irrelevant components automatically.
Model Pruning

The FAB algorithm eliminates irrelevant components automatically

\[
\mathbb{E}_q \left[ \ln p(X, Z|\Pi, K) \right] - \frac{D_{\Pi}}{2} \sum_k \ln \sum_n \mathbb{E}_q[z_{nk}] + H(q)
\]

• The penalty term introduces group sparsity to \(Z\)

\[K=6\]

\[Z\]

\[K=6\]

\[K=3\]

\[update\]

\[update\]
Model Pruning

The FAB algorithm eliminates irrelevant components automatically.

\[
\mathbb{E}_q \left[ \ln p(\mathbf{X}, \mathbf{Z}|\Pi, K) \right] - \frac{D_{\Pi}}{2} \sum_k \ln \sum_n \mathbb{E}_q[\mathbf{z}_{nk}] + H(q)
\]

- The penalty term introduces group sparsity to \( \mathbf{Z} \)

\[ \begin{array}{cccc}
    & & & \\
    & \text{update} & \text{update} & \\
K=6 & \rightarrow & \rightarrow & = \\
\end{array} \]

\[ \begin{array}{cccc}
    & & & \\
    & & & \\
    & & & \\
    & & & \\
    & & & \\
    & & & \\
K=6 & & & K=3 \\
\end{array} \]
Summary of FIC/FAB

✔ Asymptotically equivalent to the marginal likelihood
  • Fits to "Big Data" situations
Summary of FIC/FAB

- Asymptotically equivalent to the marginal likelihood
  - Fits to “Big Data” situations
- Performs parameter inference and model selection simultaneously
  - EM-like updates of $q$ and $\Pi$
  - ARD-like model pruning
Summary of FIC/FAB

✓ Asymptotically equivalent to the marginal likelihood
  • Fits to “Big Data” situations
✓ Performs parameter inference and model selection simultaneously
  • EM-like updates of $q$ and $\Pi$
  • ARD-like model pruning
✓ Doesn’t depend on the choice of $p(\Pi)$
  • More frequentist than Bayesian
Summary of FIC/FAB

- Asymptotically equivalent to the marginal likelihood
  - Fits to “Big Data” situations
- Performs parameter inference and model selection simultaneously
  - EM-like updates of $q$ and $\Pi$
  - ARD-like model pruning
- Doesn’t depend on the choice of $p(\Pi)$
  - More frequentist than Bayesian
- Works in many binary LVMs
Limitations of FIC/FAB

- Limited to binary LVMs
  - In real $Z$, $\sum_n z_{nk}$ can be negative
  - $-\ln \sum_n z_{nk}$ may diverge
Limitations of FIC/FAB

❖ Limited to binary LVMs
  • $\ln \sum_n z_{nk}$ can be negative
  • $-\ln \sum_n z_{nk}$ may diverge
❖ Missing relations to EM and VB
  • Similar approaches, but which are better?
Limitations of FIC/FAB

❌ Limited to binary LVMs
  • In real $Z$, $\sum_n z_{nk}$ can be negative
  • $-\ln \sum_n z_{nk}$ may diverge

❌ Missing relations to EM and VB
  • Similar approaches, but which are better?

❌ Unclear legitimacy of optimizing **FIC**
  • e.g. tightness
Revisiting FAB

Generalization of FAB
• Now $Z$ can take general values (e.g. $Z \in \mathbb{R}^{N \times K}$)
• Now $Z$ can take general values (e.g. $Z \in \mathbb{R}^{N \times K}$)

• Consider separating the parameters:
  \[ \Pi = \{ \Theta, \Xi \} \]
  - $\Theta$: $k$-\textit{independent} params
  - $\Xi = \{ \xi_k \}_{k=1}^K$: $k$-\textit{dependent} params (e.g. mixing coefficients)
Generalized FIC (gFIC)

Definition

\[
gFIC(K) \equiv \mathbb{E}_{q^*} \left[ \max_{\Pi} \ln p(X, Z|\Pi, K) - \frac{1}{2} \ln |\mathbf{F}_\Xi| \right] + H(q) + O(\ln N)
\]

- \( q^*(Z) \equiv p(Z|X, K) \): marginal posterior
- \( \mathbf{F}_\Xi \): Hessian of \(-\ln p(X, Z|\Pi, K)/N\) (i.e. empirical Fisher information)
- In PCA, \( \mathbf{F}_\Xi = Z^\top Z \)
Generalized FIC (gFIC)

**Definition**

\[
gFIC(K) \equiv \mathbb{E}_{q^*} \left[ \max_{\Pi} \ln p(X, Z|\Pi, K) - \frac{1}{2} \ln |F_\Xi| \right] + H(q) + O(\ln N)
\]

- \( q^*(Z) \equiv p(Z|X, K) \): marginal posterior
- \( F_\Xi \): Hessian of \(-\ln p(X, Z|\Pi, K)/N\) (i.e. empirical Fisher information)
- In PCA, \( F_\Xi = Z^\top Z \)

<table>
<thead>
<tr>
<th></th>
<th>FIC</th>
<th>gFIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Applicable class</td>
<td>Binary LVMs</td>
<td>General LVMs</td>
</tr>
<tr>
<td>Penalty term</td>
<td>(- \sum_k \ln \sum_n z_{nk})</td>
<td>(- \ln</td>
</tr>
<tr>
<td>Regularization</td>
<td>Group sparsity</td>
<td>&quot;Low-rank&quot;</td>
</tr>
</tbody>
</table>
Generalized FAB (gFAB)

✔ Use the same technique as FAB

\[ E_{q^*} \left[ \max_{\Pi} \ln p(X, Z|\Pi, K) \right] - \frac{1}{2} E_{q^*} \left[ \ln |F_\Xi| \right] + H(q^*) \]
Generalized FAB (gFAB)

✔️ Use the same technique as FAB

\[
E_{q^*} \left[ \max_{\Pi} \ln p(X, Z|\Pi, K) \right] - \frac{1}{2} E_{q^*} \left[ \ln |F_{\Xi}| \right] + H(q^*) \\
\geq \max_{q \in Q} E_q \left[ \max_{\Pi} \ln p(X, Z|\Pi, K) \right] - \frac{1}{2} E_q \left[ \ln |F_{\Xi}| \right] + H(q)
\]

Mean-field approx.

Jensen's ineq. \equiv gFIC(K)
Generalized FAB (gFAB)

✓ Use the same technique as FAB

\[
\mathbb{E}_{q^*} \left[ \max_{\Pi} \ln p(X, Z|\Pi, K) \right] - \frac{1}{2} \mathbb{E}_{q^*} [\ln |F_{\Xi}|] + H(q^*)
\]

\[
\geq \max_{q \in Q} \mathbb{E}_q \left[ \max_{\Pi} \ln p(X, Z|\Pi, K) \right] - \frac{1}{2} \mathbb{E}_q [\ln |F_{\Xi}|] + H(q)
\]

Mean-field approx.

\[
\geq \max_{q \in Q, \Pi} \mathbb{E}_q [\ln p(X, Z|\Pi, K)] - \frac{1}{2} \ln \mathbb{E}_q [|F_{\Xi}|] + H(q)
\]

Jensen’s ineq.

\equiv gFIC(K)
Generalized FAB (gFAB)

✔️ Use the same technique as FAB

\[
\mathbb{E}_{q^*} \left[ \max_{\Pi} \ln p(X, Z|\Pi, K) \right] - \frac{1}{2} \mathbb{E}_{q^*} [\ln |F_\Xi|] + H(q^*)
\]

\[
\geq \max_{q \in Q} \mathbb{E}_q \left[ \max_{\Pi} \ln p(X, Z|\Pi, K) \right] - \frac{1}{2} \mathbb{E}_q [\ln |F_\Xi|] + H(q)
\]

Mean-field approx.

\[
\geq \max_{q \in Q, \Pi} \mathbb{E}_q [\ln p(X, Z|\Pi, K)] - \frac{1}{2} \ln \mathbb{E}_q [|F_\Xi|] + H(q)
\]

Jensen’s ineq.

\[\equiv gFIC(K)\]

- Able to solve by alternating updates of \(q\) and \(\Pi\)
Comparison with EM and VB

✓ gFAB asymp. approx. $\ln p(X|K)$ for all $K$, whereas EM and VB don’t
Comparison with EM and VB

✓ gFAB asymp. approx. \( \ln p(X|K) \) for all \( K \), whereas EM and VB don’t

Theorem 2 & Corollary 5

Let \( K' \) be the “true” model of \( X \), then

\[
\begin{align*}
\text{gFIC}(K') & \approx \ln p(X|K) \quad \text{for } K > K' \\
\text{gFIC}(K) & \approx \ln p(X|K) \quad \text{for } K \leq K'
\end{align*}
\]

- \( K' \) can be obtained by model pruning
Comparison with EM and VB

✓ gFAB asymp. approx. \( \ln p(X|K) \) for all \( K \), whereas EM and VB don’t

Theorem 2 & Corollary 5

Let \( K' \) be the “true” model of \( X \), then

\[
\begin{align*}
gFIC(K') & \approx \ln p(X|K) \quad \text{for } K > K' \\
gFIC(K) & \approx \ln p(X|K) \quad \text{for } K \leq K'
\end{align*}
\]

- \( K' \) can be obtained by model pruning

Proposition 10+

❌ EM \( +O(\ln N) \) \( \approx \ln p(X|K) \) only for \( K \leq K' \)

❌ VB \( \approx \ln p(X|K) \) only for \( K \leq K' \)
Asymptotic Behavior of $gFIC$

✓ $gFIC(K) \approx gFIC(K)$ in some cases
Asymptotic Behavior of $gFIC$

$gFIC(K) \approx gFIC(K)$ in some cases

Proposition 6

$q^*$ is asymptotically mutually independent.

Justify mean-field approximation
Asymptotic Behavior of $\text{gFIC}$

✓ $\text{gFIC}(K) \approx \text{gFIC}(K)$ in some cases

Proposition 6

$q^*$ is asymptotically mutually independent.
✓ Justify mean-field approximation

Proposition 7

If $q$ is not degenerated and $\ln p(X, Z | \Pi, K)$ is smooth and concave w.r.t. $\Pi$,

$$\mathbb{E}_q[\max_{\Pi} \ln p(X, Z | \Pi, K)] \xrightarrow{p} \max_{\Pi} \mathbb{E}_q[\ln p(X, Z | \Pi, K)].$$

✓ Justify Jensen’s inequality
Experiments: Bayesian PCA

**Task:** model selection

- Choose $K$ that maximizes the objective
Experiments: Bayesian PCA

Task: model selection

- Choose $K$ that maximizes the objective

Results:

- gFAB: Successfully obtain true $K = 10$ w/ skipping $K = 10, \ldots, 29$
- EM: Always overestimates $K$ (as suggested in Prop. 10+)
- VB1: Select true $K$ but need to compute all $K = 1, \ldots, 30$
Conclusion

Summary of this talk:

- **FAB**: Tractable Bayesian method for *binary* LVMs
- Proposed **gFAB** for *general* LVMs (e.g. PCA)
- **Theoretical Analysis**
  - Showing the desirable properties of gFAB
Conclusion

Summary of this talk:

- FAB: Tractable Bayesian method for binary LVMs
- Proposed gFAB for general LVMs (e.g. PCA)
- Theoretical Analysis
  - Showing the desirable properties of gFAB

At the poster session (right after):
We will explain more details such as

- Full derivation of gFIC
- “High-level” mechanism of model pruning
- ...
Future work

- Potentially applicable to a wide class of LVMs
  - factor analysis, CCA, partial membership, linear dynamical systems, ...

- If you are interested in, let’s collaborate!
Future work

- Potentially applicable to a wide class of LVMs
  - factor analysis, CCA, partial membership, linear dynamical systems, ...
- If you are interested in, let’s collaborate!

Thank you!

References


