Supplementary Material of "Rebuilding Factorized Information Criterion: Asymptotically Accurate Marginal Likelihood"

May 16, 2015

1 Proofs

Proof of Proposition 4. If $\mathbf{Z}$ is not degenerated, then Laplace's method yields Eq. (10). By collecting from Eq. (10) the terms that depend on $\mathbf{Z}$, we obtain

$$p(\mathbf{Z} \mid \mathbf{X}, K) \propto p(\mathbf{Z}, \mathbf{X} \mid \hat{\mathbf{\Pi}}, K)|\mathbf{F}_{\hat{\mathbf{\Pi}}}|^{-1/2}(1 + O(N^{-1})). \tag{101}$$

If $p(\mathbf{Z} \mid \mathbf{X}, K)$ is degenerated, we consider the transformation (11). Here, 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 mapping $\mathbf{\Pi} \to \tilde{\mathbf{\Pi}}_{K'}$ is onto **A1** and the prior is strictly positive in the whole space of $\mathbf{\Pi}$ **A4**, $\tilde{p}(\mathbf{\Pi} \mid K')$ is also strictly positive, including $\hat{\mathbf{\Pi}}_{K'} = \operatorname{argmax}_{\mathbf{\Pi}_{K'}} \ln p(\mathbf{X}, \tilde{\mathbf{Z}}_{K'} \mid \mathbf{\Pi}_{K'}, K')$. Consequently, we can again use Laplace's method for $\ln p(\mathbf{X}, \tilde{\mathbf{Z}}_{K'} \mid \hat{\mathbf{\Pi}}_{K'}, K')$, and by collecting the terms that depend on $\mathbf{Z}$, we obtain

$$p(\mathbf{X} \mid \mathbf{Z}, K) \propto p(\mathbf{X}, \tilde{\mathbf{Z}}_{K'} \mid \hat{\mathbf{\Pi}}_{K'}, K')|\mathbf{F}_{\hat{\mathbf{\Pi}}_{K'}}|^{-1/2}(1 + O(N^{-1})) \tag{102}$$

$$\propto p_{K'}(\tilde{\mathbf{Z}}_{K'}, K')(1 + O(N^{-1})). \tag{103}$$

This concludes the proof. □

Proof of Theorem 2. First, we prove the case that $p(\mathbf{Z} \mid \mathbf{X}, K)$ is not degenerated. In that case, Laplace's approximation yields Eq. (10) in probability, and substituting Eq. (10) into (7) gives (8).

If $\kappa(p(\mathbf{Z} \mid \mathbf{X}, K)) = K' < K$, Proposition 4 gives us that $p(\mathbf{Z} \mid \mathbf{X}, K) = p_{K'}(\mathbf{Z})(1 + O(N^{-1}))$. Since

$$\mathbb{E}_{p(\mathbf{Z}|\mathbf{X},K)}[\ln p(\mathbf{X}, \mathbf{Z} \mid K)] = \mathbb{E}_{p_{K'}}[\ln p(\mathbf{X}, \mathbf{Z} \mid K)] + O(1)$$

and

$$\begin{aligned} H(p(\mathbf{Z} \mid \mathbf{X}, K)) &= (1 + O(N^{-1}))H(p_{K'}) + (1 + O(N^{-1}))\ln(1 + O(N^{-1})) \\ &= H(p_{K'}) + O(1), \end{aligned}$$

$\ln p(\mathbf{X} \mid K)$ is rewritten by

$$\mathbb{E}_{p_{K'}}[\ln p(\mathbf{X}, \mathbf{Z} \mid K)] + H(p_{K'}) + O(1) \tag{104}$$

$$= \mathbb{E}_{p_{K'}}[\mathcal{L}(\hat{\mathbf{Z}}_{K'}, \tilde{\mathbf{\Pi}}_{K'}, K')] + H(p_{K'}) + O(1) \tag{105}$$

Here, since the projection $\mathbf{T}_{K'} : \mathbf{Z} \to \tilde{\mathbf{Z}}_{K'}$ is continuous and onto (**A1**), we can describe $p_{K'}(\mathbf{Z})$ as the density of $\mathbf{Z}_{K'}$ by using a change of variables, which we denote by $\tilde{p}_{K'}(\mathbf{Z}_{K'})$. Now, we can rewrite the first term as the integral over $\mathbf{Z}_{K'}$, i.e.,

$$\mathbb{E}_{p_{K'}}[\mathcal{L}(\tilde{\mathbf{Z}}_{K'}, \hat{\mathbf{\Pi}}_{K'}, K')] = \int \mathcal{L}(\mathbf{T}_{K'}(\mathbf{Z}), \hat{\mathbf{\Pi}}_{K'}, K') p_{K'}(\mathbf{T}_{K'}(\mathbf{Z})) \dot{\mathbf{Z}} \tag{106}$$

$$= \int \mathcal{L}(\mathbf{Z}_{K'}, \hat{\mathbf{\Pi}}_{K'}, K') \tilde{p}_{K'}(\mathbf{Z}_{K'}) \dot{\mathbf{Z}}_{K'}. \tag{107}$$

Similarly, $\mathrm{gFIC}(K')$ is rewritten using Proposition 4 as

$$\mathrm{gFIC}(K') = \mathbb{E}_{p_{K'}}[\mathcal{L}(\mathbf{Z}_{K'}, \hat{\mathbf{\Pi}}_{K'}, K')] + H(p_{K'}) + O(1) \tag{108}$$

Again, the first term is written as

$$\mathbb{E}_{p_{K'}}[\mathcal{L}(\mathbf{Z}_{K'}, \hat{\mathbf{\Pi}}_{K'}, K')] = \int \mathcal{L}(\mathbf{Z}_{K'}, \hat{\mathbf{\Pi}}_{K'}, K') p_{K'}(\mathbf{T}_{K'}(\mathbf{Z})) \dot{\mathbf{Z}}_{K'} \tag{109}$$

$$= \int \mathcal{L}(\mathbf{Z}_{K'}, \hat{\mathbf{\Pi}}_{K'}, K') \tilde{p}_{K'}(\mathbf{Z}_{K'}) \dot{\mathbf{Z}}_{K'} \tag{110}$$

Since Eq. (107) and (110) are the same, this concludes Eq. (8).

□

Proof of Proposition 6. Proposition 4 shows that, if $\mathbf{Z}$ is non-degenerated,

$$p(\mathbf{Z} \mid \mathbf{X}, K) \propto p(\mathbf{X}, \mathbf{Z} \mid \hat{\mathbf{\Pi}}) |\mathbf{F}_{\hat{\mathbf{\Pi}}}|^{-1/2} \tag{111}$$

$$\propto \prod_n p(\mathbf{x}_n, \mathbf{z}_n \mid \hat{\mathbf{\Pi}}) |\mathbf{F}_{\hat{\mathbf{\Pi}}}|^{-1/2N} \tag{112}$$

Since $\ln |\mathbf{F}_{\mathbf{\Pi}}| = O(1)$, $|\mathbf{F}_{\mathbf{\Xi}}|^{-1/2N}$ quickly diminishes to 1 for $N \to \infty$. □

Proof of Proposition 7. For technical reasons, we redefine the estimators as follows:

$$\hat{\mathbf{\Pi}} \equiv \operatorname*{argmax}_{\mathbf{\Pi}} g_N(\mathbf{\Pi}) = \operatorname*{argmax}_{\mathbf{\Pi}} \frac{1}{N} \ln p(\mathbf{X}, \mathbf{Z} | \mathbf{\Pi}), \tag{113}$$

$$\bar{\mathbf{\Pi}} \equiv \operatorname*{argmax}_{\mathbf{\Pi}} G_N(\mathbf{\Pi}) = \operatorname*{argmax}_{\mathbf{\Pi}} \mathbb{E}_q[\frac{1}{N} \ln p(\mathbf{X}, \mathbf{Z} | \mathbf{\Pi})]. \tag{114}$$

According to **A5**, $g_N(\mathbf{\Pi})$ is continuous and concave, and it uniformly converges to $G_N(\mathbf{\Pi})$, i.e.,

$$\sup_{\mathbf{\Pi} \in \mathcal{P}} |g_N(\mathbf{\Pi}) - G_N(\mathbf{\Pi})| \xrightarrow{\mathrm{p}} 0. \tag{115}$$

This suffices to show the consistency (for example, see Theorem 5.7 in van der Vaart (1998).) □

2 The gFAB Algorithm of BPCA

Let $\mathcal{Q}$ be restricted as Gaussian with mean field. Then, the form of q is determined as $q(\mathbf{Z}) = \prod_n N(\mathbf{z}_n \mid \boldsymbol{\mu}_n, \boldsymbol{\Omega})$. By substituting the BPCA's likelihood into Eq. (17), we obtain

$$
\begin{aligned}
&\max_{q\in\mathcal{Q}} \mathbb{E}_q\left[-\frac{\lambda}{2}\|\mathbf{X}-\mathbf{Z}\mathbf{W}^\top\|_{\mathrm{Fro}}^2 - \frac{1}{2}\|\mathbf{Z}\|_{\mathrm{Fro}}^2 - \frac{1}{2}\ln|\mathbf{F}|\right] + \frac{ND}{2}\ln\lambda + H(q) + \mathrm{const.}\\
=&\max_{q\in\mathcal{Q}} \mathbb{E}_q\left[-\frac{\lambda}{2}\|\mathbf{X}-\mathbf{Z}\mathbf{W}^\top\|_{\mathrm{Fro}}^2 - \frac{1}{2}\|\mathbf{Z}\|_{\mathrm{Fro}}^2 - \frac{D}{2}(\ln|\frac{1}{N}\mathbf{Z}^\top\mathbf{Z}| + K\ln\lambda)\right] + \frac{ND}{2}\ln\lambda + \frac{N}{2}\ln|\boldsymbol{\Omega}| + \mathrm{const.}\\
\geq&\max_{q\in\mathcal{Q}} \mathbb{E}_q\left[-\frac{\lambda}{2}\|\mathbf{X}-\mathbf{Z}\mathbf{W}^\top\|_{\mathrm{Fro}}^2 - \frac{1}{2}\|\mathbf{Z}\|_{\mathrm{Fro}}^2\right] - \frac{D}{2}\ln|\frac{1}{N}\mathbf{S}| + \frac{(N-K)D}{2}\ln\lambda + \frac{N}{2}\ln|\boldsymbol{\Omega}| + \mathrm{const.}\\
=&\max_{q\in\mathcal{Q}} -\frac{\lambda}{2}(\mathrm{tr}(\mathbf{W}^\top\mathbf{W}\mathbf{S} - \mathbf{X}\mathbf{W}\mathbb{E}[\mathbf{Z}]^\top)) - \frac{1}{2}\mathrm{tr}(\mathbf{S}) - \frac{D}{2}\ln|\mathbf{S}| + \frac{(N-K)D}{2}\ln\lambda + \frac{N}{2}\ln|\boldsymbol{\Omega}| + \mathrm{const.}
\end{aligned}
$$

where $\mathbf{S} \equiv \mathbb{E}[\mathbf{Z}^\top\mathbf{Z}] = N\boldsymbol{\Omega} + \sum_n \boldsymbol{\mu}_n\boldsymbol{\mu}_n^\top$. Note that the equality in the second-to-last line holds at $N \to \infty$.

Update q By setting the derivatives to zero, we obtain the following update rules:

$$\boldsymbol{\mu}_n^{\mathrm{new}} = \lambda\mathbf{x}_n\mathbf{W}\boldsymbol{\Omega}, \tag{116}$$

$$\boldsymbol{\Omega}^{\mathrm{new}} = (\mathbf{I} + \lambda\mathbf{W}^\top\mathbf{W} + D(\mathbf{S}+\mathbf{I})^{-1})^{-1}. \tag{117}$$

Note that we use $\mathbf{S}$ as an auxiliary variable.

Update $\boldsymbol{\Pi}$ Update rules of $\mathbf{W}$ and λ are almost the same as those of the EM algorithm, which are:

$$\mathbf{W}^{\mathrm{new}} = \mathbf{X}^\top[\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N]\mathbf{S}^{-1}, \tag{118}$$

$$\lambda^{\mathrm{new}} = \frac{(N-K)D}{\mathbb{E}[\|\mathbf{X}-\mathbf{Z}\mathbf{W}^\top\|_{\mathrm{Fro}}^2]}. \tag{119}$$

References

van der Vaart, A. W. *Asymptotic statistics*. Cambridge University Press, 1998.