Globally Sparse Probabilistic PCA - Supplementary Material

1. Detailed Inference for the relaxed model

We denote $\boldsymbol{\theta} = (\mathbf{u}, \alpha, \sigma)$ the vector of parameters. In order to maximize the evidence $p(\mathbf{x_1}, ...\mathbf{x_n}|\boldsymbol{\theta})$, we adopt a variational approach. We view $\mathbf{y}_1, ...\mathbf{y}_n$ and $\mathbf{W}$ as latent variables.

Given a (variational) distribution q over the space of latent variables, the variational free energy is given by

$$\mathcal{F}_q(\mathbf{x_1}, ...\mathbf{x_n}|\boldsymbol{\theta}) = -\mathbb{E}_q[\ln p(\mathbf{X}, \mathbf{Y}, \mathbf{W}|\boldsymbol{\theta})] - H(q) \quad (1)$$

where H denotes the differential entropy, and is an upper bound to the negative log-evidence:

$$-\ln p(\mathbf{X}|\boldsymbol{\theta}) = \mathcal{F}_q(\mathbf{X}|\boldsymbol{\theta}) - \mathrm{KL}(q||p(\cdot|\boldsymbol{\theta})) \leq \mathcal{F}_q(\mathbf{X}|\boldsymbol{\theta}).$$

To be able to minimize $\mathcal{F}_q(\mathbf{X}|\boldsymbol{\theta})$, we make the following mean-field approximation on the variational distribution:

$$q(\mathbf{Y}, \mathbf{W}) = q(\mathbf{Y})q(\mathbf{W}) \quad (2)$$

We can now compute the variational posterior distribution q^* which minimizes the free energy. Note that two factorizations arise naturally. This will conveniently keep the size of the covariance matrices lower than d.

Proposition 1 *The variational posterior distribution of the latent variables which minimizes the free energy is given by*

$$q^*(\mathbf{Y}) = \prod_{i=1}^{n} \mathcal{N}(\mathbf{y}_i|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}) \quad (3)$$

and

$$q^*(\mathbf{W}) = \prod_{k=1}^{p} \mathcal{N}(\mathbf{w}_k|\boldsymbol{m}_k, \mathbf{S}_k) \quad (4)$$

where

$$\boldsymbol{\mu}_i = \frac{1}{\sigma}\boldsymbol{\Sigma}\mathbf{M}^T\mathbf{U}\mathbf{x}_i,\ \boldsymbol{\Sigma} = \left(\mathbf{I}_d + \frac{1}{\sigma^2}\mathbf{M}^T\mathbf{U}^2\mathbf{M} + \frac{1}{\sigma^2}\sum_{k=1}^{p} u_k^2\mathbf{S}_k\right)^{-1},\ \mathbf{M} = (\mathbf{m}_1, ...\mathbf{m}_n)$$

and

$$\mathbf{m}_k = \frac{u_k}{\sigma}\mathbf{S}_k\sum_{i=1}^{n} x_{i,k}\boldsymbol{\mu}_i,\ \mathbf{S}_k = \left(\frac{1}{\alpha}\mathbf{I}_d + \frac{nu_k^2}{\sigma^2}\boldsymbol{\Sigma} + \frac{u_k^2}{\sigma^2}\boldsymbol{\mathcal{M}}^T\boldsymbol{\mathcal{M}}\right)^{-1},\ \boldsymbol{\mathcal{M}} = (\boldsymbol{\mu}_1, ...\boldsymbol{\mu}_p).$$

Proof *Variational distribution of the latent vectors.* Using a standard result in variational mean-field approximations(Bishop, 2006, chap. 10), we can write

$$\ln q^*(\mathbf{y}) = \mathbb{E}_{q(\mathbf{W})}[\ln p(\mathbf{X}, \mathbf{Y}, \mathbf{W}|\boldsymbol{\theta})]$$

which leads to the factorization $q^*(\mathbf{y}) = \prod_{i \leq n} q^*(\mathbf{y}_i)$. Then, for each $i \leq n$, we can write

$$\ln q^*(\mathbf{y}_i) = \mathbb{E}_{q(\mathbf{W})}[\ln p(\mathbf{X}, \mathbf{Y}, \mathbf{W}|\boldsymbol{\theta})] = \mathbb{E}_{q(\mathbf{W})}[\frac{-1}{2\sigma^2}||\mathbf{x}_i - \mathbf{U}\mathbf{W}\mathbf{y}_i||_2^2] - \frac{1}{2}||\mathbf{y}_i||_2^2$$

therefore

$$\ln q^*(\mathbf{y}_i) = \frac{-1}{2\sigma^2}\mathbf{y_i}^T\mathbb{E}_{q(\mathbf{W})}[\mathbf{W}^T\mathbf{U}^2\mathbf{W}]\mathbf{y}_i + \frac{1}{\sigma^2}\mathbf{y}_i^T\mathbf{W}^T\mathbf{U}\mathbf{x}_i - \frac{1}{2}||\mathbf{y}_i||_2^2$$

which lead to the desired form.

Variational distribution of the loading matrix. Similarly,

$$\ln q^*(\mathbf{W}) = \frac{-1}{2\sigma^2}\sum_{i=1}^{n}\mathbb{E}_{q(\mathbf{y}_i)}[||\mathbf{x}_i - \mathbf{U}\mathbf{W}\mathbf{y}_i||_2^2] - \frac{1}{2\alpha}\sum_{i=1}^{n}||\mathbf{w}_i||_2^2$$

$$\ln q^*(\mathbf{W}) = \sum_{i=1}^{n}\left(\frac{-1}{2\sigma^2}\sum_{j=1}^{p}v_j^2\mathbf{w}_j^T\mathbf{E}_{q(\mathbf{y}_i)}[||\mathbf{y}_i||_2^2]\mathbf{w}_j + \frac{1}{\sigma^2}\sum_{j=1}^{p}\mathbf{w}_j^T x_{i,j}v_j\mathbb{E}_{q(\mathbf{y}_i)}(\mathbf{y}_i) - \frac{1}{\alpha}||\mathbf{w}_j||_2^2\right)$$

which leads to the factorization $q^*(\mathbf{W}) = \prod_{j \leq p} q^*(\mathbf{w}_i)$ and to the desired expression. ∎

We can now compute the value of the free energy.

Proposition 2 *Up to unnecessary additive constants, the entropy of the variational distribution is given by*

$$H(q) = \frac{n}{2}\ln|\boldsymbol{\Sigma}| + \frac{1}{2}\sum_{k=1}^{p}\ln|\mathbf{S}_k|. \quad (5)$$

Proof We have, using the factorizations of the former proposition

$$H(q) = -\mathbb{E}_q[\ln q(\mathbf{Y}, \mathbf{W})] = -\sum_{i=1}^{n}\mathbb{E}_{q(\mathbf{y}_i)}[\ln q(\mathbf{y}_i)] - \sum_{j=1}^{p}\mathbb{E}_{q(\mathbf{W})}[\ln q(\mathbf{W})]$$

which allows us to conclude. ∎

Proposition 3 *Up to unnecessary additive constants, the negative free energy is given by*

$$-\mathcal{F}_q(\mathbf{x_1}, ...\mathbf{x_n}|\boldsymbol{\theta}) = -np\ln\sigma - \frac{dp}{2}\ln\alpha - \frac{1}{2\sigma^2}\mathrm{Tr}(\mathbf{X}^T\mathbf{X}) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{k=1}^{p}u_k^2\mathrm{Tr}[(\boldsymbol{\Sigma} + \boldsymbol{\mu}_i\boldsymbol{\mu}_i^T)(\boldsymbol{S}_k + \boldsymbol{m}_i\boldsymbol{m}_i^T)]$$
$$+ \frac{1}{\sigma^2}\sum_{i=1}^{n}\mathbf{x}_i^T\mathbf{U}\mathbf{M}\boldsymbol{\mu}_i + \sum_{k=1}^{p} -\frac{1}{2\alpha}\mathrm{Tr}(\mathbf{S}_k + \mathbf{m}_k\mathbf{m}_k^T) - \sum_{i=1}^{n}\mathrm{Tr}(\boldsymbol{\Sigma} + \boldsymbol{\mu}_i\boldsymbol{\mu}_i^T)$$
$$+ \frac{n}{2}\ln|\boldsymbol{\Sigma}| + \frac{1}{2}\sum_{k=1}^{p}\ln|\mathbf{S}_k|. \quad (6)$$

Proof By definition, we have

$$-\mathcal{F}_q(\mathbf{x_1},...\mathbf{x_n}|\boldsymbol{\theta}) = -\mathbb{E}_q[\ln p(\mathbf{X},\mathbf{Y},\mathbf{W}|\boldsymbol{\theta})] - H(q)$$

therefore

$$\mathcal{F}_q(\mathbf{x_1},...\mathbf{x_n}|\boldsymbol{\theta}) = -np\ln\sigma - \frac{1}{2\sigma^2}\mathrm{Tr}(\mathbf{X}^T\mathbf{X}) - \frac{1}{2\sigma^2}\sum_{i=1}^n \mathbb{E}_q[\mathbf{y}_i\mathbf{W}^T\mathbf{U}^2\mathbf{W}\mathbf{y}_i] + \frac{1}{\sigma^2}\sum_{i=1}^n \mathbf{x}_i^T\mathbf{U}\mathbf{M}\boldsymbol{\mu}_i$$
$$+ \sum_{k=1}^p\left(-\frac{d}{2}\ln\alpha - \frac{1}{2\alpha}\mathbb{E}_q[\mathbf{w}_k^T\mathbf{w}_k]\right) - \sum_{i=1}^n \mathbb{E}_q[\mathbf{y}_i^t\mathbf{y}_i]$$
$$+ \frac{n}{2}\ln|\boldsymbol{\Sigma}| + \frac{1}{2}\sum_{k=1}^p \ln|\mathbf{S}_k|.$$

and computing the expectations leads to the desired expression. ■

This expression allows us to find the following M-step updates:

$$\alpha^* = \frac{1}{dp}\sum_{k=1}^p \mathrm{Tr}(\mathbf{S}_k + \mathbf{m}_k\mathbf{m}_k^T), \tag{7}$$

$$\sigma^* = \frac{\mathrm{Tr}(\mathbf{X}^T\mathbf{X})}{np} + \frac{1}{np}\sum_{i=1}^n\left(\sum_{k=1}^p u_k^2\mathrm{Tr}[(\boldsymbol{\Sigma}+\boldsymbol{\mu}_i\boldsymbol{\mu}_i^T)(\boldsymbol{S}_k+\boldsymbol{m}_i\boldsymbol{m}_i^T)] - 2\mathbf{x}_i^T\mathbf{U}\mathbf{M}\boldsymbol{\mu}_i\right), \tag{8}$$

and, for $k \in \{1,...,p\}$,

$$u_k^* = \mathrm{argmin}_{u\in[0,1]}\frac{u^2}{2\sigma^2}\sum_{i=1}^n \mathrm{Tr}[(\boldsymbol{\Sigma}+\boldsymbol{\mu}_i\boldsymbol{\mu}_i^T)(\boldsymbol{S}_k+\boldsymbol{m}_i\boldsymbol{m}_i^T)] - u\sum_{i=1}^n x_{i,k}\mathbf{m}_k^T\boldsymbol{\mu}_i \tag{9}$$

Note that the objective function of the optimization problem (9) is simply a univariate polynomial.

References

C. M. Bishop. *Pattern recognition and machine learning.* Springer, 2006.