Improved Variational Approximation for Bayesian PCA

Shivani Agarwal and Christopher M. Bishop

August 7, 2003

Abstract

As with most non-trivial models, an exact Bayesian treatment of the probabilistic PCA model (under a meaningful prior) is analytically intractable. Various approximations have therefore been proposed in the literature; these include approximations based on type-II maximum likelihood as well as variational approximations. In this document, we describe an improved variational approximation for Bayesian PCA. This is achieved by defining a more general prior over the model parameters that has stronger conjugacy properties, thereby allowing for a more accurate variational approximation to the true posterior.

1 Introduction

The probabilistic PCA model is defined by[1]

$$p(\mathbf{t}|\mathbf{x}, \boldsymbol{\mu}, \mathbf{W}, \tau) = \mathcal{N}(\mathbf{t}|\mathbf{W}\mathbf{x} + \boldsymbol{\mu}, \tau^{-1}\mathbf{I}_d), \tag{1}$$

where $\mathbf{t} \in \mathbb{R}^d$ is the observed variable, $\mathbf{x} \in \mathbb{R}^q$ ($q < d$) is a latent variable with prior distribution

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mathbf{0}, \mathbf{I}_q), \tag{2}$$

and $\boldsymbol{\mu} \in \mathbb{R}^d$, $\mathbf{W} \in \mathbb{R}^{d \times q}$ and $\tau \in \mathbb{R}$ constitute the model parameters.

Given a set of observed data points $D = \{\mathbf{t}_n\}_{n=1}^N \subset \mathbb{R}^d$, the Bayesian approach consists in defining a suitable prior $p(\boldsymbol{\mu}, \mathbf{W}, \tau)$ over the model parameters and finding the posterior $p(\boldsymbol{\mu}, \mathbf{W}, \tau|D)$ under this prior. The predictive density is then obtained by averaging over the model parameters according to the posterior:

$$p(\mathbf{t}|D) = \int\int\int p(\mathbf{t}|\boldsymbol{\mu}, \mathbf{W}, \tau) p(\boldsymbol{\mu}, \mathbf{W}, \tau|D) d\boldsymbol{\mu} d\mathbf{W} d\tau. \tag{3}$$

It has been observed that by setting q to its maximum possible value ($q = d - 1$) and choosing a hierarchical prior that effectively promotes sparseness in the columns of $\mathbf{W}$, the Bayesian approach can be used to automatically determine the number of principal components (number of non-zero columns in $\mathbf{W}$) supported by the observed data. Evaluation of the true posterior under such a prior being intractable, various approximations have been proposed; these include approximations based on type-II maximum likelihood [1] and variational approximations [2, 3]. While variational approximations generally provide a more accurate Bayesian treatment than type-II maximum likelihood methods, the variational approximations proposed so far are not completely satisfactory. Indeed, the variational approximation of [2] assumes complete factorization in both the prior and the variational posterior. An attempt to define a more general prior and obtain

[1] For $\mathbf{z} \in \mathbb{R}^d$, $\mathcal{N}(\mathbf{z}|\mathbf{m}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}} |\boldsymbol{\Sigma}|^{-1/2} \exp\{-\frac{1}{2}(\mathbf{z} - \mathbf{m})^T \boldsymbol{\Sigma}^{-1} (\mathbf{z} - \mathbf{m})\}$.

a more accurate variational approximation was recently made in [3]; however this contains some technical errors and falls short of achieving this goal.

Here we formulate a more general hierarchical prior which, while simulating a sparseness-promoting prior, has some desirable conjugacy properties. As we show, this allows for a more accurate variational approximation to the true posterior.

2 Prior

We define a hierarchical prior $p(\boldsymbol{\mu}, \mathbf{W}, \tau|\boldsymbol{\alpha})$ over the model parameters $\boldsymbol{\mu} \in \mathbb{R}^d$, $\mathbf{W} \in \mathbb{R}^{d\times(d-1)}$ and $\tau \in \mathbb{R}$, governed by a vector of hyper-parameters $\boldsymbol{\alpha} \in \mathbb{R}^{d-1}$, as follows:

$$p(\boldsymbol{\mu}, \mathbf{W}, \tau|\boldsymbol{\alpha}) = p(\boldsymbol{\mu}|\mathbf{W}, \tau)p(\mathbf{W}|\tau, \boldsymbol{\alpha})p(\tau), \tag{4}$$

with[2]

$$p(\boldsymbol{\mu}|\mathbf{W}, \tau) = \mathcal{N}(\boldsymbol{\mu}|\mathbf{W}\mathbf{s}_0 + \mathbf{m}_0, (\beta_0\tau)^{-1}\mathbf{I}_d), \tag{5}$$

$$p(\mathbf{W}|\tau, \boldsymbol{\alpha}) = \prod_{i=1}^{d-1} \mathcal{N}(\mathbf{w}_i|\mathbf{0}, (\alpha_i\tau)^{-1}\mathbf{I}_d), \tag{6}$$

$$p(\tau) = \mathcal{G}(\tau|a_0, b_0), \tag{7}$$

where $\mathbf{w}_i$ refers to the ith column of $\mathbf{W}$. The prior over the hyper-parameters $\boldsymbol{\alpha}$ is defined as

$$p(\boldsymbol{\alpha}) = \prod_{i=1}^{d-1} \mathcal{G}(\alpha_i|c_0, d_0). \tag{8}$$

In the above, $\mathbf{s}_0 \in \mathbb{R}^{d-1}$, $\mathbf{m}_0 \in \mathbb{R}^d$ and $\beta_0, a_0, b_0, c_0, d_0 \in \mathbb{R}$ are (fixed) hyper-parameters that can be chosen to give suitable priors.

3 Variational Posterior

We find a variational approximation to the true posterior $p(\boldsymbol{\mu}, \mathbf{W}, \tau|D)$ under the above prior that maximizes a rigorous lower bound on the log likelihood of the data.

Denoting all the parameters and latent variables in the model as $\boldsymbol{\theta}$, we have for all distributions $q(\boldsymbol{\theta})$,

$$\begin{aligned} \ln p(D) &= \ln \int p(D, \boldsymbol{\theta}) d\boldsymbol{\theta} \\ &= \ln \int q(\boldsymbol{\theta}) \frac{p(D, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta} \\ &\geq \int q(\boldsymbol{\theta}) \ln \frac{p(D, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta} \\ &= \mathcal{L}(q), \end{aligned} \tag{9}$$

where we have applied Jensen's inequality. It is easily shown that maximizing the lower bound $\mathcal{L}(q)$ is equivalent to minimizing the Kullback-Leibler divergence between $q(\boldsymbol{\theta})$ and the true posterior $p(\boldsymbol{\theta}|D)$.

[2] For $\tau \in \mathbb{R}$, $\mathcal{G}(\tau|a, b) = \frac{1}{\Gamma(a)} b^a \tau^{a-1} \exp\{-b\tau\}$.

Here $\boldsymbol{\theta} = (\boldsymbol{\mu}, \mathbf{W}, \tau, \boldsymbol{\alpha}, X)$, where $X = \{\mathbf{x}_n\}_{n=1}^N$ is the set of latent variables corresponding to the observed data D. We consider maximizing the lower bound $\mathcal{L}(q)$ subject to the following factorization constraint[3]:

$$q(\boldsymbol{\mu}, \mathbf{W}, \tau, \boldsymbol{\alpha}, X) = q(\boldsymbol{\mu}, \mathbf{W}, \tau) q(\boldsymbol{\alpha}) q(X). \tag{10}$$

As shown in Appendix A, the components of the resulting variational distribution $q^*(\boldsymbol{\mu}, \mathbf{W}, \tau, \boldsymbol{\alpha}, X)$ have the following forms[4]:

$$\begin{aligned}
q^*(\boldsymbol{\mu}, \mathbf{W}, \tau) &= q^*(\boldsymbol{\mu}|\mathbf{W}, \tau) q^*(\mathbf{W}|\tau) q^*(\tau), && (11)\\
q^*(\boldsymbol{\mu}|\mathbf{W}, \tau) &= \mathcal{N}(\boldsymbol{\mu}|\mathbf{W}\mathbf{s}_{\boldsymbol{\mu}} + \mathbf{m}_{\boldsymbol{\mu}}, (\beta_{\boldsymbol{\mu}}\tau)^{-1}\mathbf{I}_d), && (12)\\
q^*(\mathbf{W}|\tau) &= \prod_{k=1}^{d} \mathcal{N}(\tilde{\mathbf{w}}_k|\mathbf{m}_{\mathbf{w}}^{(k)}, (\tau\boldsymbol{\Lambda}_{\mathbf{w}})^{-1}), && (13)\\
q^*(\tau) &= \mathcal{G}(\tau|a_\tau, b_\tau), && (14)\\
q^*(\boldsymbol{\alpha}) &= \prod_{i=1}^{d-1} \mathcal{G}(\alpha_i|c_{\boldsymbol{\alpha}}, d_{\boldsymbol{\alpha}}^{(i)}), && (15)\\
q^*(X) &= \prod_{n=1}^{N} \mathcal{N}(\mathbf{x}_n|\mathbf{m}_{\mathbf{x}}^{(n)}, \boldsymbol{\Sigma}_{\mathbf{x}}), && (16)
\end{aligned}$$

where $\tilde{\mathbf{w}}_k$ denotes a column vector corresponding to the kth row of $\mathbf{W}$, and the various parameters are given by

$$\begin{aligned}
\beta_{\boldsymbol{\mu}} &= \beta_0 + N, && (17)\\
\mathbf{s}_{\boldsymbol{\mu}} &= \beta_{\boldsymbol{\mu}}^{-1}\left(\beta_0\mathbf{s}_0 - \sum_{n=1}^{N}\langle\mathbf{x}_n\rangle\right), && (18)\\
\mathbf{m}_{\boldsymbol{\mu}} &= \beta_{\boldsymbol{\mu}}^{-1}\left(\beta_0\mathbf{m}_0 + \sum_{n=1}^{N}\mathbf{t}_n\right), && (19)\\
\boldsymbol{\Lambda}_{\mathbf{w}} &= \mathrm{diag}\langle\boldsymbol{\alpha}\rangle + \beta_0\mathbf{s}_0\mathbf{s}_0^T - \beta_{\boldsymbol{\mu}}\mathbf{s}_{\boldsymbol{\mu}}\mathbf{s}_{\boldsymbol{\mu}}^T + \sum_{n=1}^{N}\langle\mathbf{x}_n\mathbf{x}_n^T\rangle, && (20)\\
\mathbf{m}_{\mathbf{w}}^{(k)} &= \boldsymbol{\Lambda}_{\mathbf{w}}^{-1}\left(\sum_{n=1}^{N} t_{nk}\langle\mathbf{x}_n\rangle - \beta_0 m_{0k}\mathbf{s}_0 + \beta_{\boldsymbol{\mu}} m_{\boldsymbol{\mu}k}\mathbf{s}_{\boldsymbol{\mu}}\right), && (21)\\
a_\tau &= a_0 + \frac{Nd}{2}, && (22)\\
b_\tau &= b_0 + \frac{1}{2}\sum_{n=1}^{N}\mathbf{t}_n^T\mathbf{t}_n + \frac{\beta_0}{2}\mathbf{m}_0^T\mathbf{m}_0 - \frac{\beta_{\boldsymbol{\mu}}}{2}\mathbf{m}_{\boldsymbol{\mu}}^T\mathbf{m}_{\boldsymbol{\mu}} - \frac{1}{2}\sum_{k=1}^{d}\mathbf{m}_{\mathbf{w}}^{(k)T}\boldsymbol{\Lambda}_{\mathbf{w}}\mathbf{m}_{\mathbf{w}}^{(k)}, && (23)\\
c_{\boldsymbol{\alpha}} &= c_0 + \frac{d}{2}, && (24)\\
d_{\boldsymbol{\alpha}}^{(i)} &= d_0 + \frac{\langle\tau\|\mathbf{w}_i\|^2\rangle}{2}, && (25)\\
\boldsymbol{\Sigma}_{\mathbf{x}} &= (\mathbf{I}_{d-1} + \langle\tau\mathbf{W}^T\mathbf{W}\rangle)^{-1}, && (26)\\
\mathbf{m}_{\mathbf{x}}^{(n)} &= \boldsymbol{\Sigma}_{\mathbf{x}}(\langle\tau\mathbf{W}\rangle^T\mathbf{t}_n - \langle\tau\mathbf{W}^T\boldsymbol{\mu}\rangle). && (27)
\end{aligned}$$

[3]Note that the factorization assumed in the variational distribution is weaker than that in [2], leading to a more accurate variational posterior.

[4]Note that due to conjugacy properties of the prior, the variational posterior $q^*(\boldsymbol{\mu}, \mathbf{W}, \tau)$ has the same form as $p(\boldsymbol{\mu}, \mathbf{W}, \tau|\boldsymbol{\alpha})$.

In the above, diag$\langle\boldsymbol{\alpha}\rangle$ denotes a diagonal matrix whose diagonal elements are given by $\langle\alpha_i\rangle$. The required moments can easily be computed as

$$\begin{aligned}
\langle\mathbf{x}_n\rangle &= \mathbf{m}_{\mathbf{x}}^{(n)}, && (28)\\
\langle\mathbf{x}_n\mathbf{x}_n^T\rangle &= \boldsymbol{\Sigma}_{\mathbf{x}} + \mathbf{m}_{\mathbf{x}}^{(n)}\mathbf{m}_{\mathbf{x}}^{(n)T}, && (29)\\
\langle\alpha_i\rangle &= \frac{c_{\boldsymbol{\alpha}}}{d_{\boldsymbol{\alpha}}^{(i)}}, && (30)\\
\langle\tau\mathbf{W}\rangle &= \frac{a_\tau}{b_\tau}\mathbf{M}_{\mathbf{w}}^T, && (31)\\
\langle\tau\mathbf{W}^T\mathbf{W}\rangle &= d\boldsymbol{\Lambda}_{\mathbf{w}}^{-1} + \frac{a_\tau}{b_\tau}\mathbf{M}_{\mathbf{w}}\mathbf{M}_{\mathbf{w}}^T, && (32)\\
\langle\tau\|\mathbf{w}_i\|^2\rangle &= d(\boldsymbol{\Lambda}_{\mathbf{w}}^{-1})_{ii} + \frac{a_\tau}{b_\tau}\|\tilde{\mathbf{M}}_{\mathbf{w}i}\|^2, && (33)\\
\langle\tau\mathbf{W}^T\boldsymbol{\mu}\rangle &= d\boldsymbol{\Lambda}_{\mathbf{w}}^{-1}\mathbf{s}_{\boldsymbol{\mu}} + \frac{a_\tau}{b_\tau}\mathbf{M}_{\mathbf{w}}(\mathbf{M}_{\mathbf{w}}^T\mathbf{s}_{\boldsymbol{\mu}} + \mathbf{m}_{\boldsymbol{\mu}}), && (34)
\end{aligned}$$

where $\mathbf{M}_{\mathbf{w}} \in \mathbb{R}^{(d-1)\times d}$ is a matrix whose kth column is $\mathbf{m}_{\mathbf{w}}^{(k)}$, and $\tilde{\mathbf{M}}_{\mathbf{w}i}$ denotes a column vector corresponding to the ith row of $\mathbf{M}_{\mathbf{w}}$. Eqs. (28)-(30) above follow directly from the properties of the Normal and Gamma distributions; Eqs. (31)-(34) are derived in Appendix B.

4 Lower Bound

The variational lower bound on the log likelihood of the data is given by

$$\begin{aligned}
\mathcal{L}(q^*) &= \langle\ln p(D|X,\boldsymbol{\mu},\mathbf{W},\tau)\rangle + \langle\ln p(X)\rangle + \langle\ln p(\boldsymbol{\mu}|\mathbf{W},\tau)\rangle + \langle\ln p(\mathbf{W}|\tau,\boldsymbol{\alpha})\rangle + \langle\ln p(\tau)\rangle + \langle\ln p(\boldsymbol{\alpha})\rangle\\
&\quad - \langle\ln q^*(X)\rangle - \langle\ln q^*(\boldsymbol{\mu}|\mathbf{W},\tau)\rangle - \langle\ln q^*(\mathbf{W}|\tau)\rangle - \langle\ln q^*(\tau)\rangle - \langle\ln q^*(\boldsymbol{\alpha})\rangle, \qquad (35)
\end{aligned}$$

where the various terms in the bound are given by

$$\begin{aligned}
\langle\ln p(D|X,\boldsymbol{\mu},\mathbf{W},\tau)\rangle &= \sum_{n=1}^{N}\left\{\frac{d}{2}\langle\ln\tau\rangle - \frac{d}{2}\ln(2\pi) - \frac{1}{2}\left(\langle\tau\rangle\mathbf{t}_n^T\mathbf{t}_n + \langle\tau\mathbf{x}_n^T\mathbf{W}^T\mathbf{W}\mathbf{x}_n\rangle + \langle\tau\boldsymbol{\mu}^T\boldsymbol{\mu}\rangle\right.\right.\\
&\quad \left.\left. - 2\mathbf{t}_n^T\langle\tau\boldsymbol{\mu}\rangle - 2\mathbf{t}_n^T\langle\tau\mathbf{W}\rangle\langle\mathbf{x}_n\rangle + 2\langle\mathbf{x}_n\rangle^T\langle\tau\mathbf{W}^T\boldsymbol{\mu}\rangle\right)\right\}, \qquad (36)\\
\langle\ln p(X)\rangle &= \sum_{n=1}^{N}\left\{-\frac{(d-1)}{2}\ln(2\pi) - \frac{1}{2}\langle\mathbf{x}_n^T\mathbf{x}_n\rangle\right\}, \qquad (37)\\
\langle\ln p(\boldsymbol{\mu}|\mathbf{W},\tau)\rangle &= \frac{d}{2}\ln(\beta_0\langle\tau\rangle) - \frac{d}{2}\ln(2\pi) - \frac{\beta_0}{2}\left\{\langle\tau\boldsymbol{\mu}^T\boldsymbol{\mu}\rangle + \mathbf{s}_0^T\langle\tau\mathbf{W}^T\mathbf{W}\rangle\mathbf{s}_0\right.\\
&\quad \left. + \langle\tau\rangle\mathbf{m}_0^T\mathbf{m}_0 - 2\mathbf{s}_0^T\langle\tau\mathbf{W}^T\boldsymbol{\mu}\rangle - 2\mathbf{m}_0^T\langle\tau\boldsymbol{\mu}\rangle + 2\mathbf{s}_0^T\langle\tau\mathbf{W}\rangle^T\mathbf{m}_0\right\}, \qquad (38)\\
\langle\ln p(\mathbf{W}|\tau,\boldsymbol{\alpha})\rangle &= \sum_{i=1}^{d-1}\left\{\frac{d}{2}\ln(\langle\alpha_i\rangle\langle\tau\rangle) - \frac{d}{2}\ln(2\pi) - \frac{\langle\alpha_i\rangle\langle\tau\|\mathbf{w}_i\|^2\rangle}{2}\right\}, \qquad (39)\\
\langle\ln p(\tau)\rangle &= a_0\ln b_0 - \ln\Gamma(a_0) + (a_0-1)\langle\ln\tau\rangle - b_0\langle\tau\rangle, \qquad (40)\\
\langle\ln p(\boldsymbol{\alpha})\rangle &= \sum_{i=1}^{d-1}\left\{c_0\ln d_0 - \ln\Gamma(c_0) + (c_0-1)\langle\ln\alpha_i\rangle - d_0\langle\alpha_i\rangle\right\}, \qquad (41)\\
\langle\ln q^*(X)\rangle &= \sum_{n=1}^{N}\left\{-\frac{1}{2}\ln|\boldsymbol{\Sigma}_{\mathbf{x}}| - \frac{(d-1)}{2}\ln(2\pi) - \frac{1}{2}\mathrm{Tr}(\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\langle\mathbf{x}_n\mathbf{x}_n^T\rangle)\right.
\end{aligned}$$

$$- \frac{1}{2}\mathbf{m}_{\mathbf{x}}^T\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbf{m}_{\mathbf{x}} + \langle\mathbf{x}_n\rangle^T\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbf{m}_{\mathbf{x}}\Big\}, \tag{42}$$

$$\begin{aligned}\langle\ln q^*(\boldsymbol{\mu}|\mathbf{W},\tau)\rangle &= \frac{d}{2}\ln(\beta_{\boldsymbol{\mu}}\langle\tau\rangle) - \frac{d}{2}\ln(2\pi) - \frac{\beta_{\boldsymbol{\mu}}}{2}\Big\{\langle\tau\boldsymbol{\mu}^T\boldsymbol{\mu}\rangle + \mathbf{s}_{\boldsymbol{\mu}}^T\langle\tau\mathbf{W}^T\mathbf{W}\rangle\mathbf{s}_{\boldsymbol{\mu}} \\ &\quad + \langle\tau\rangle\mathbf{m}_{\boldsymbol{\mu}}^T\mathbf{m}_{\boldsymbol{\mu}} - 2\mathbf{s}_{\boldsymbol{\mu}}^T\langle\tau\mathbf{W}^T\boldsymbol{\mu}\rangle - 2\mathbf{m}_{\boldsymbol{\mu}}^T\langle\tau\boldsymbol{\mu}\rangle + 2\mathbf{s}_{\boldsymbol{\mu}}^T\langle\tau\mathbf{W}\rangle^T\mathbf{m}_{\boldsymbol{\mu}}\Big\}, \end{aligned} \tag{43}$$

$$\begin{aligned}\langle\ln q^*(\mathbf{W}|\tau)\rangle &= \frac{d(d-1)}{2}\langle\ln\tau\rangle + \frac{d}{2}\ln|\boldsymbol{\Lambda}_{\mathbf{w}}| - \frac{d(d-1)}{2}\ln(2\pi) \\ &\quad - \frac{1}{2}\sum_{k=1}^{d}\Big\{\mathrm{Tr}(\boldsymbol{\Lambda}_{\mathbf{w}}\langle\tau\tilde{\mathbf{w}}_k\tilde{\mathbf{w}}_k^T\rangle) - 2\mathbf{m}_{\mathbf{w}}^{(k)T}\boldsymbol{\Lambda}_{\mathbf{w}}\langle\tau\tilde{\mathbf{w}}_k\rangle + \langle\tau\rangle\mathbf{m}_{\mathbf{w}}^{(k)T}\boldsymbol{\Lambda}_{\mathbf{w}}\mathbf{m}_{\mathbf{w}}^{(k)}\Big\}, \end{aligned} \tag{44}$$

$$\langle\ln q^*(\tau)\rangle = a_\tau\ln b_\tau - \ln\Gamma(a_\tau) + (a_\tau-1)\langle\ln\tau\rangle - b_\tau\langle\tau\rangle, \tag{45}$$

$$\langle\ln q^*(\boldsymbol{\alpha})\rangle = \sum_{i=1}^{d-1}\Big\{c_{\boldsymbol{\alpha}}\ln d_{\boldsymbol{\alpha}}^{(i)} - \ln\Gamma(c_{\boldsymbol{\alpha}}) + (c_{\boldsymbol{\alpha}}-1)\langle\ln\alpha_i\rangle - d_{\boldsymbol{\alpha}}^{(i)}\langle\alpha_i\rangle\Big\}. \tag{46}$$

The additional moments appearing in the above expressions are given by

$$\langle\ln\tau\rangle = \psi(a_\tau) - \ln b_\tau, \tag{47}$$

$$\langle\tau\rangle = \frac{a_\tau}{b_\tau}, \tag{48}$$

$$\langle\mathbf{x}_n^T\mathbf{x}_n\rangle = \mathrm{Tr}(\boldsymbol{\Sigma}_{\mathbf{x}}) + \mathbf{m}_{\mathbf{x}}^{(n)T}\mathbf{m}_{\mathbf{x}}^{(n)}, \tag{49}$$

$$\langle\ln\alpha_i\rangle = \psi(c_{\boldsymbol{\alpha}}) - \ln d_{\boldsymbol{\alpha}}^{(i)}, \tag{50}$$

$$\langle\tau\boldsymbol{\mu}\rangle = \frac{a_\tau}{b_\tau}\left(\mathbf{M}_{\mathbf{w}}^T\mathbf{s}_{\boldsymbol{\mu}} + \mathbf{m}_{\boldsymbol{\mu}}\right), \tag{51}$$

$$\langle\tau\boldsymbol{\mu}^T\boldsymbol{\mu}\rangle = d(\beta_{\boldsymbol{\mu}}^{-1} + \mathbf{s}_{\boldsymbol{\mu}}^T\boldsymbol{\Lambda}_{\mathbf{w}}^{-1}\mathbf{s}_{\boldsymbol{\mu}}) + \frac{a_\tau}{b_\tau}\|\mathbf{M}_{\mathbf{w}}^T\mathbf{s}_{\boldsymbol{\mu}} + \mathbf{m}_{\boldsymbol{\mu}}\|^2, \tag{52}$$

$$\langle\tau\tilde{\mathbf{w}}_k\rangle = \frac{a_\tau}{b_\tau}\mathbf{m}_{\mathbf{w}}^{(k)}, \tag{53}$$

$$\langle\tau\tilde{\mathbf{w}}_k\tilde{\mathbf{w}}_k^T\rangle = \boldsymbol{\Lambda}_{\mathbf{w}}^{-1} + \frac{a_\tau}{b_\tau}\mathbf{m}_{\mathbf{w}}^{(k)}\mathbf{m}_{\mathbf{w}}^{(k)T}, \tag{54}$$

where $\psi(\cdot)$ is the digamma function defined by

$$\psi(a) = \frac{d}{da}\ln\Gamma(a). \tag{55}$$

Eqs. (47)-(50) follow from the properties of the Normal and Gamma distributions; the remaining moments are derived in Appendix B.

5 Predictive Density

The predictive density obtained by approximating the true posterior $p(\boldsymbol{\mu},\mathbf{W},\tau|D)$ with the variational posterior $q^*(\boldsymbol{\mu},\mathbf{W},\tau)$ has the form of a continuous mixture of Student t-distributions:

$$\tilde{p}(\mathbf{t}|D) = \int\tilde{p}(\mathbf{t}|\mathbf{x},D)p(\mathbf{x})d\mathbf{x}, \tag{56}$$

where[5]

$$\tilde{p}(\mathbf{t}|\mathbf{x},D) = \mathcal{S}(\mathbf{t}|\mathbf{m}_{\mathbf{t}}(\mathbf{x}),\beta_{\mathbf{t}}(\mathbf{x})^{-1}\mathbf{I}_d,\nu_{\mathbf{t}}), \tag{57}$$

[5]For $\mathbf{z}\in\mathbb{R}^d$, $\mathcal{S}(\mathbf{z}|\mathbf{m},\boldsymbol{\Lambda},\nu) = \frac{\Gamma((\nu+d)/2)}{\Gamma(\nu/2)}\frac{|\boldsymbol{\Lambda}|^{1/2}}{(\nu\pi)^{d/2}}\left\{1+\frac{1}{\nu}(\mathbf{z}-\mathbf{m})^T\boldsymbol{\Lambda}(\mathbf{z}-\mathbf{m})\right\}^{-(\nu+d)/2}$.

with

$$\mathbf{m_t}(\mathbf{x}) = \mathbf{M_w}^T\mathbf{x} - (\mathbf{M_w}^T\mathbf{s}_{\boldsymbol{\mu}} + \mathbf{m}_{\boldsymbol{\mu}}), \tag{58}$$

$$\beta_{\mathbf{t}}(\mathbf{x}) = \frac{b_\tau}{a_\tau}(1 + \beta_{\boldsymbol{\mu}}^{-1} + (\mathbf{x} - \mathbf{s}_{\boldsymbol{\mu}})^T\boldsymbol{\Lambda}_{\mathbf{w}}^{-1}(\mathbf{x} - \mathbf{s}_{\boldsymbol{\mu}})), \tag{59}$$

$$\nu_{\mathbf{t}} = 2a_\tau, \tag{60}$$

and

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mathbf{0}, \mathbf{I}_{d-1}). \tag{61}$$

The above form is derived in Appendix C.

References

[1] Christopher M. Bishop. Bayesian PCA. In *Advances in Neural Information Processing Systems*, volume 11, 1999.

[2] Christopher M. Bishop. Variational principal components. In *International Conference on Artificial Neural Networks*, 1999.

[3] Shigeyuki Oba, Masa-aki Sato, and Shin Ishii. Prior hyperparameters in Bayesian PCA. In *International Conference on Artificial Neural Networks*, 2003.

A Form of Variational Distribution

The lower bound $\mathcal{L}(q)$ for a distribution $q(\boldsymbol{\mu}, \mathbf{W}, \tau, \boldsymbol{\alpha}, X)$ that satisfies the factorization constraint in Eq. (10) is given by

$$\begin{aligned}\mathcal{L}(q) &= \langle \ln p(D|X, \boldsymbol{\mu}, \mathbf{W}, \tau)\rangle + \langle \ln p(X)\rangle + \langle \ln p(\boldsymbol{\mu}|\mathbf{W}, \tau)\rangle + \langle \ln p(\mathbf{W}|\tau, \boldsymbol{\alpha})\rangle + \langle \ln p(\tau)\rangle + \langle \ln p(\boldsymbol{\alpha})\rangle \\ &\quad - \langle \ln q(X)\rangle - \langle \ln q(\boldsymbol{\mu}, \mathbf{W}, \tau)\rangle - \langle \ln q(\boldsymbol{\alpha})\rangle. \end{aligned} \tag{62}$$

The following sections derive the forms of the components $q^*(\boldsymbol{\mu}, \mathbf{W}, \tau)$, $q^*(\boldsymbol{\alpha})$ and $q^*(X)$ of the variational distribution $q^*(\boldsymbol{\mu}, \mathbf{W}, \tau, \boldsymbol{\alpha}, X)$ that maximizes $\mathcal{L}(q)$ subject to this factorization constraint.

A.1 Form of $q^*(\boldsymbol{\mu}, \mathbf{W}, \tau)$

We have,

$$\ln q^*(\boldsymbol{\mu}, \mathbf{W}, \tau) = \langle \ln p(D|X, \boldsymbol{\mu}, \mathbf{W}, \tau)\rangle_X + \ln p(\boldsymbol{\mu}|\mathbf{W}, \tau) + \langle \ln p(\mathbf{W}|\tau, \boldsymbol{\alpha})\rangle_{\boldsymbol{\alpha}} + \ln p(\tau) + \text{const.} \tag{63}$$

$$= \sum_{n=1}^{N}\left\{\frac{d}{2}\ln\tau - \frac{\tau}{2}\langle\|\mathbf{t}_n - \mathbf{W}\mathbf{x}_n - \boldsymbol{\mu}\|^2\rangle_{\mathbf{x}_n}\right\} + \frac{d}{2}\ln(\beta_0\tau) - \frac{\beta_0\tau}{2}\|\boldsymbol{\mu} - \mathbf{W}\mathbf{s}_0 - \mathbf{m}_0\|^2 + \sum_{i=1}^{d-1}\langle\frac{d}{2}\ln(\alpha_i\tau) - \frac{\alpha_i\tau}{2}\|\mathbf{w}_i\|^2\rangle_{\alpha_i} + (a_0 - 1)\ln\tau - b_0\tau + \text{const.} \tag{64}$$

$$\begin{aligned} &= \left\{(N+d)\frac{d}{2} + a_0 - 1\right\}\ln\tau - \left\{b_0 + \frac{1}{2}\sum_{n=1}^{N}\mathbf{t}_n^T\mathbf{t}_n + \frac{\beta_0}{2}\mathbf{m}_0^T\mathbf{m}_0\right\}\tau \\ &\quad - \frac{\tau}{2}\left\{\sum_{n=1}^{N}\langle\mathbf{x}_n^T\mathbf{W}^T\mathbf{W}\mathbf{x}_n\rangle + \beta_0\mathbf{s}_0^T\mathbf{W}^T\mathbf{W}\mathbf{s}_0 + \sum_{i=1}^{d-1}\langle\alpha_i\rangle\|\mathbf{w}_i\|^2\right\} \\ &\quad + \tau\left\{\sum_{n=1}^{N}\mathbf{t}_n^T\mathbf{W}\langle\mathbf{x}_n\rangle - \beta_0\mathbf{m}_0^T\mathbf{W}\mathbf{s}_0\right\} \\ &\quad - \frac{\tau}{2}(\beta_0 + N)\boldsymbol{\mu}^T\boldsymbol{\mu} + \tau\boldsymbol{\mu}^T\left\{\sum_{n=1}^{N}(\mathbf{t}_n - \mathbf{W}\langle\mathbf{x}_n\rangle) + \beta_0(\mathbf{W}\mathbf{s}_0 + \mathbf{m}_0)\right\} + \text{const.} \end{aligned} \tag{65}$$

$$\begin{aligned} &= \left\{(N+d-1)\frac{d}{2} + a_0 - 1\right\}\ln\tau - \left\{b_0 + \frac{1}{2}\sum_{n=1}^{N}\mathbf{t}_n^T\mathbf{t}_n + \frac{\beta_0}{2}\mathbf{m}_0^T\mathbf{m}_0\right\}\tau \\ &\quad - \frac{\tau}{2}\left\{\sum_{n=1}^{N}\langle\mathbf{x}_n^T\mathbf{W}^T\mathbf{W}\mathbf{x}_n\rangle + \beta_0\mathbf{s}_0^T\mathbf{W}^T\mathbf{W}\mathbf{s}_0 + \sum_{i=1}^{d-1}\langle\alpha_i\rangle\|\mathbf{w}_i\|^2\right\} \\ &\quad + \tau\left\{\sum_{n=1}^{N}\mathbf{t}_n^T\mathbf{W}\langle\mathbf{x}_n\rangle - \beta_0\mathbf{m}_0^T\mathbf{W}\mathbf{s}_0\right\} \\ &\quad + \ln q^*(\boldsymbol{\mu}|\mathbf{W}, \tau) + \frac{\beta_{\boldsymbol{\mu}}\tau}{2}(\mathbf{W}\mathbf{s}_{\boldsymbol{\mu}} + \mathbf{m}_{\boldsymbol{\mu}})^T(\mathbf{W}\mathbf{s}_{\boldsymbol{\mu}} + \mathbf{m}_{\boldsymbol{\mu}}) + \text{const.} \end{aligned} \tag{66}$$

where

$$q^*(\boldsymbol{\mu}|\mathbf{W}, \tau) = \mathcal{N}(\boldsymbol{\mu}|\mathbf{W}\mathbf{s}_{\boldsymbol{\mu}} + \mathbf{m}_{\boldsymbol{\mu}}, (\beta_{\boldsymbol{\mu}}\tau)^{-1}\mathbf{I}_d), \tag{67}$$

with

$$\beta_{\boldsymbol{\mu}} = \beta_0 + N, \tag{68}$$

$$\mathbf{s}_{\boldsymbol{\mu}} = \beta_{\boldsymbol{\mu}}^{-1}\left(\beta_0\mathbf{s}_0 - \sum_{n=1}^{N}\langle\mathbf{x}_n\rangle\right), \tag{69}$$

$$\mathbf{m}_{\boldsymbol{\mu}} = \beta_{\boldsymbol{\mu}}^{-1}\left(\beta_0\mathbf{m}_0 + \sum_{n=1}^{N}\mathbf{t}_n\right). \tag{70}$$

From Eq. (66), we have

$$\begin{aligned}
&\ln q^*(\boldsymbol{\mu},\mathbf{W},\tau) - \ln q^*(\boldsymbol{\mu}|\mathbf{W},\tau) = \\
&\quad\left\{(N+d-1)\frac{d}{2} + a_0 - 1\right\}\ln\tau - \left\{b_0 + \frac{1}{2}\sum_{n=1}^{N}\mathbf{t}_n^T\mathbf{t}_n + \frac{\beta_0}{2}\mathbf{m}_0^T\mathbf{m}_0 - \frac{\beta_{\boldsymbol{\mu}}}{2}\mathbf{m}_{\boldsymbol{\mu}}^T\mathbf{m}_{\boldsymbol{\mu}}\right\}\tau \\
&\quad - \frac{\tau}{2}\left\{\sum_{n=1}^{N}\langle\mathbf{x}_n^T\mathbf{W}^T\mathbf{W}\mathbf{x}_n\rangle + \beta_0\mathbf{s}_0^T\mathbf{W}^T\mathbf{W}\mathbf{s}_0 - \beta_{\boldsymbol{\mu}}\mathbf{s}_{\boldsymbol{\mu}}^T\mathbf{W}^T\mathbf{W}\mathbf{s}_{\boldsymbol{\mu}} + \sum_{i=1}^{d-1}\langle\alpha_i\rangle\|\mathbf{w}_i\|^2\right\} \\
&\quad + \tau\left\{\sum_{n=1}^{N}\mathbf{t}_n^T\mathbf{W}\langle\mathbf{x}_n\rangle - \beta_0\mathbf{m}_0^T\mathbf{W}\mathbf{s}_0 + \beta_{\boldsymbol{\mu}}\mathbf{m}_{\boldsymbol{\mu}}^T\mathbf{W}\mathbf{s}_{\boldsymbol{\mu}}\right\} + \text{const.}
\end{aligned} \tag{71}$$

$$\begin{aligned}
&= \left\{(N+d-1)\frac{d}{2} + a_0 - 1\right\}\ln\tau - \left\{b_0 + \frac{1}{2}\sum_{n=1}^{N}\mathbf{t}_n^T\mathbf{t}_n + \frac{\beta_0}{2}\mathbf{m}_0^T\mathbf{m}_0 - \frac{\beta_{\boldsymbol{\mu}}}{2}\mathbf{m}_{\boldsymbol{\mu}}^T\mathbf{m}_{\boldsymbol{\mu}}\right\}\tau \\
&\quad - \frac{\tau}{2}\sum_{k=1}^{d}\tilde{\mathbf{w}}_k^T\left\{\sum_{n=1}^{N}\langle\mathbf{x}_n\mathbf{x}_n^T\rangle + \beta_0\mathbf{s}_0\mathbf{s}_0^T - \beta_{\boldsymbol{\mu}}\mathbf{s}_{\boldsymbol{\mu}}\mathbf{s}_{\boldsymbol{\mu}}^T + \text{diag}\langle\boldsymbol{\alpha}\rangle\right\}\tilde{\mathbf{w}}_k \\
&\quad + \tau\sum_{k=1}^{d}\tilde{\mathbf{w}}_k^T\left\{\sum_{n=1}^{N}\mathbf{t}_{nk}\langle\mathbf{x}_n\rangle - \beta_0 m_{0k}\mathbf{s}_0 + \beta_{\boldsymbol{\mu}}m_{\boldsymbol{\mu}k}\mathbf{s}_{\boldsymbol{\mu}}\right\} + \text{const.}
\end{aligned} \tag{72}$$

where $\tilde{\mathbf{w}}_k$ denotes a column vector corresponding to the kth row of $\mathbf{W}$, and we have made use of the (easily proved) identities

$$\mathbf{s}_1^T\mathbf{W}^T\mathbf{W}\mathbf{s}_2 = \sum_{k=1}^{d}\tilde{\mathbf{w}}_k^T(\mathbf{s}_1\mathbf{s}_2^T)\tilde{\mathbf{w}}_k \tag{73}$$

and

$$\mathbf{m}^T\mathbf{W}\mathbf{s} = \sum_{k=1}^{d}m_k(\tilde{\mathbf{w}}_k^T\mathbf{s}). \tag{74}$$

This gives

$$\begin{aligned}
&\ln q^*(\boldsymbol{\mu},\mathbf{W},\tau) - \ln q^*(\boldsymbol{\mu}|\mathbf{W},\tau) = \\
&\quad\left\{\frac{Nd}{2} + a_0 - 1\right\}\ln\tau - \left\{b_0 + \frac{1}{2}\sum_{n=1}^{N}\mathbf{t}_n^T\mathbf{t}_n + \frac{\beta_0}{2}\mathbf{m}_0^T\mathbf{m}_0 - \frac{\beta_{\boldsymbol{\mu}}}{2}\mathbf{m}_{\boldsymbol{\mu}}^T\mathbf{m}_{\boldsymbol{\mu}}\right\}\tau \\
&\quad + \ln q^*(\mathbf{W}|\tau) + \frac{\tau}{2}\sum_{k=1}^{d}\mathbf{m}_{\mathbf{w}}^{(k)T}\boldsymbol{\Lambda}_{\mathbf{w}}\mathbf{m}_{\mathbf{w}}^{(k)} + \text{const.}
\end{aligned} \tag{75}$$

where

$$q^*(\mathbf{W}|\tau) = \prod_{k=1}^{d}\mathcal{N}(\tilde{\mathbf{w}}_k|\mathbf{m}_{\mathbf{w}}^{(k)},(\tau\boldsymbol{\Lambda}_{\mathbf{w}})^{-1}), \tag{76}$$

with

$$\mathbf{\Lambda}_{\mathbf{w}} = \text{diag}\langle\boldsymbol{\alpha}\rangle + \beta_0 \mathbf{s}_0 \mathbf{s}_0^T - \beta_{\boldsymbol{\mu}} \mathbf{s}_{\boldsymbol{\mu}} \mathbf{s}_{\boldsymbol{\mu}}^T + \sum_{n=1}^{N} \langle \mathbf{x}_n \mathbf{x}_n^T \rangle, \tag{77}$$

$$\mathbf{m}_{\mathbf{w}}^{(k)} = \mathbf{\Lambda}_{\mathbf{w}}^{-1} \left(\sum_{n=1}^{N} t_{nk} \langle \mathbf{x}_n \rangle - \beta_0 m_{0k} \mathbf{s}_0 + \beta_{\boldsymbol{\mu}} m_{\boldsymbol{\mu}k} \mathbf{s}_{\boldsymbol{\mu}} \right). \tag{78}$$

From Eq. (75), we have

$$\ln q^*(\boldsymbol{\mu}, \mathbf{W}, \tau) - \ln q^*(\boldsymbol{\mu}|\mathbf{W}, \tau) - \ln q^*(\mathbf{W}|\tau) = \tag{79}$$

$$\left\{ \frac{Nd}{2} + a_0 - 1 \right\} \ln \tau - \left\{ b_0 + \frac{1}{2} \sum_{n=1}^{N} \mathbf{t}_n^T \mathbf{t}_n + \frac{\beta_0}{2} \mathbf{m}_0^T \mathbf{m}_0 - \frac{\beta_{\boldsymbol{\mu}}}{2} \mathbf{m}_{\boldsymbol{\mu}}^T \mathbf{m}_{\boldsymbol{\mu}} - \frac{1}{2} \sum_{k=1}^{d} \mathbf{m}_{\mathbf{w}}^{(k)T} \mathbf{\Lambda}_{\mathbf{w}} \mathbf{m}_{\mathbf{w}}^{(k)} \right\} \tau$$

$$+ \text{const.} \tag{80}$$

$$= \ln q^*(\tau) + \text{const.} \tag{81}$$

where

$$q^*(\tau) = \mathcal{G}(\tau | a_\tau, b_\tau), \tag{82}$$

with

$$a_\tau = a_0 + \frac{Nd}{2}, \tag{83}$$

$$b_\tau = b_0 + \frac{1}{2} \sum_{n=1}^{N} \mathbf{t}_n^T \mathbf{t}_n + \frac{\beta_0}{2} \mathbf{m}_0^T \mathbf{m}_0 - \frac{\beta_{\boldsymbol{\mu}}}{2} \mathbf{m}_{\boldsymbol{\mu}}^T \mathbf{m}_{\boldsymbol{\mu}} - \frac{1}{2} \sum_{k=1}^{d} \mathbf{m}_{\mathbf{w}}^{(k)T} \mathbf{\Lambda}_{\mathbf{w}} \mathbf{m}_{\mathbf{w}}^{(k)}. \tag{84}$$

Thus, from Eq. (81), we see that

$$q^*(\boldsymbol{\mu}, \mathbf{W}, \tau) = q^*(\boldsymbol{\mu}|\mathbf{W}, \tau) q^*(\mathbf{W}|\tau) q^*(\tau), \tag{85}$$

where the forms of the components $q^*(\boldsymbol{\mu}|\mathbf{W}, \tau)$, $q^*(\mathbf{W}|\tau)$ and $q^*(\tau)$ are given by Eqs. (67), (76) and (82) respectively.

A.2 Form of $q^*(\boldsymbol{\alpha})$

We have,

$$\begin{aligned} \ln q^*(\boldsymbol{\alpha}) &= \langle \ln p(\mathbf{W}|\tau, \boldsymbol{\alpha}) \rangle_{\mathbf{W},\tau} + \ln p(\boldsymbol{\alpha}) + \text{const.} \\ &= \sum_{i=1}^{d-1} \left\{ \langle \frac{d}{2} \ln(\alpha_i \tau) - \frac{\alpha_i \tau}{2} \|\mathbf{w}_i\|^2 \rangle_{\mathbf{W},\tau} + (c_0 - 1) \ln \alpha_i - d_0 \alpha_i \right\} + \text{const.} \\ &= \sum_{i=1}^{d-1} \left\{ (c_0 + \frac{d}{2} - 1) \ln \alpha_i - \left(d_0 + \frac{\langle \tau \|\mathbf{w}_i\|^2 \rangle}{2} \right) \alpha_i \right\} + \text{const.} \end{aligned} \tag{86}$$

Hence

$$q^*(\boldsymbol{\alpha}) = \prod_{i=1}^{d-1} \mathcal{G}(\alpha_i | c_{\boldsymbol{\alpha}}, d_{\boldsymbol{\alpha}}^{(i)}), \tag{87}$$

with

$$c_{\boldsymbol{\alpha}} = c_0 + \frac{d}{2}, \tag{88}$$

$$d_{\boldsymbol{\alpha}}^{(i)} = d_0 + \frac{\langle \tau \|\mathbf{w}_i\|^2 \rangle}{2}. \tag{89}$$

A.3 Form of $q^*(X)$

We have,

$$\begin{aligned}
\ln q^*(X) &= \langle \ln p(D|X, \mathbf{W}, \boldsymbol{\mu}, \tau) \rangle_{\boldsymbol{\mu}, \mathbf{W}, \tau} + \ln p(X) + \text{const.} \\
&= -\frac{1}{2} \sum_{n=1}^{N} \left\{ \mathbf{x}_n^T \mathbf{x}_n + \langle \tau \| \mathbf{t}_n - \mathbf{W}\mathbf{x}_n - \boldsymbol{\mu} \|^2 \rangle_{\boldsymbol{\mu}, \mathbf{W}, \tau} \right\} + \text{const.} \\
&= -\frac{1}{2} \sum_{n=1}^{N} \left\{ \mathbf{x}_n^T (\mathbf{I}_{d-1} + \langle \tau \mathbf{W}^T \mathbf{W} \rangle) \mathbf{x}_n - 2\mathbf{x}_n^T \langle \tau \mathbf{W}^T (\mathbf{t}_n - \boldsymbol{\mu}) \rangle \right\} + \text{const.} \qquad (90)
\end{aligned}$$

Hence

$$q^*(X) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{x}_n | \mathbf{m}_{\mathbf{x}}^{(n)}, \boldsymbol{\Sigma}_{\mathbf{x}}), \qquad (91)$$

with

$$\begin{aligned}
\boldsymbol{\Sigma}_{\mathbf{x}} &= (\mathbf{I}_{d-1} + \langle \tau \mathbf{W}^T \mathbf{W} \rangle)^{-1}, \qquad (92) \\
\mathbf{m}_{\mathbf{x}}^{(n)} &= \boldsymbol{\Sigma}_{\mathbf{x}} (\langle \tau \mathbf{W} \rangle^T \mathbf{t}_n - \langle \tau \mathbf{W}^T \boldsymbol{\mu} \rangle). \qquad (93)
\end{aligned}$$

B Computation of Moments

The following sections derive the non-trivial moments in Sections 3 and 4.

B.1 Derivation of $\langle \tau \mathbf{W} \rangle$

$$\begin{aligned}
\langle \tau \mathbf{W} \rangle &= \int \int \tau \mathbf{W} q^*(\mathbf{W}|\tau) q^*(\tau) d\mathbf{W} d\tau \\
&= \int \tau \left[\int \mathbf{W} q^*(\mathbf{W}|\tau) d\mathbf{W} \right] q^*(\tau) d\tau \\
&= \int \tau \mathbf{M}_{\mathbf{w}}^T q^*(\tau) d\tau \\
&= \frac{a_\tau}{b_\tau} \mathbf{M}_{\mathbf{w}}^T. \qquad (94)
\end{aligned}$$

B.2 Derivation of $\langle \tau \mathbf{W}^T \mathbf{W} \rangle$

$$\begin{aligned}
\langle \tau \mathbf{W}^T \mathbf{W} \rangle &= \int \int \tau \mathbf{W}^T \mathbf{W} q^*(\mathbf{W}|\tau) q^*(\tau) d\mathbf{W} d\tau \\
&= \int \tau \left[\int \mathbf{W}^T \mathbf{W} q^*(\mathbf{W}|\tau) d\mathbf{W} \right] q^*(\tau) d\tau \\
&= \int \tau \left[d(\tau \boldsymbol{\Lambda}_{\mathbf{w}})^{-1} + \mathbf{M}_{\mathbf{w}} \mathbf{M}_{\mathbf{w}}^T \right] q^*(\tau) d\tau \\
&= d\boldsymbol{\Lambda}_{\mathbf{w}}^{-1} + \frac{a_\tau}{b_\tau} \mathbf{M}_{\mathbf{w}} \mathbf{M}_{\mathbf{w}}^T. \qquad (95)
\end{aligned}$$

B.3 Derivation of $\langle \tau \|\mathbf{w}_i\|^2 \rangle$

From Eq. (95) above, it follows directly that

$$\langle \tau \|\mathbf{w}_i\|^2 \rangle = d(\mathbf{\Lambda}_{\mathbf{w}}^{-1})_{ii} + \frac{a_\tau}{b_\tau} \|\tilde{\mathbf{M}}_{\mathbf{w}i}\|^2. \tag{96}$$

$$\tag{97}$$

B.4 Derivation of $\langle \tau \mathbf{W}^T \boldsymbol{\mu} \rangle$

$$\begin{aligned}
\langle \tau \mathbf{W}^T \boldsymbol{\mu} \rangle &= \int\int\int \tau \mathbf{W}^T \boldsymbol{\mu} q^*(\boldsymbol{\mu}|\mathbf{W},\tau) q^*(\mathbf{W}|\tau) q^*(\tau) d\boldsymbol{\mu} d\mathbf{W} d\tau \\
&= \int\int \tau \mathbf{W}^T \left[\int \boldsymbol{\mu} q^*(\boldsymbol{\mu}|\mathbf{W},\tau) d\boldsymbol{\mu} \right] q^*(\mathbf{W}|\tau) q^*(\tau) d\mathbf{W} d\tau \\
&= \int\int \tau \mathbf{W}^T \left[\mathbf{W}\mathbf{s}_{\boldsymbol{\mu}} + \mathbf{m}_{\boldsymbol{\mu}} \right] q^*(\mathbf{W}|\tau) q^*(\tau) d\mathbf{W} d\tau \\
&= \langle \tau \mathbf{W}^T \mathbf{W} \rangle \mathbf{s}_{\boldsymbol{\mu}} + \langle \tau \mathbf{W} \rangle^T \mathbf{m}_{\boldsymbol{\mu}} \\
&= d\mathbf{\Lambda}_{\mathbf{w}}^{-1} \mathbf{s}_{\boldsymbol{\mu}} + \frac{a_\tau}{b_\tau} \mathbf{M}_{\mathbf{w}} (\mathbf{M}_{\mathbf{w}}^T \mathbf{s}_{\boldsymbol{\mu}} + \mathbf{m}_{\boldsymbol{\mu}}).
\end{aligned} \tag{98}$$

B.5 Derivation of $\langle \tau \boldsymbol{\mu} \rangle$

$$\begin{aligned}
\langle \tau \boldsymbol{\mu} \rangle &= \int\int\int \tau \boldsymbol{\mu} q^*(\boldsymbol{\mu}|\mathbf{W},\tau) q^*(\mathbf{W}|\tau) q^*(\tau) d\boldsymbol{\mu} d\mathbf{W} d\tau \\
&= \int\int \tau \left[\int \boldsymbol{\mu} q^*(\boldsymbol{\mu}|\mathbf{W},\tau) d\boldsymbol{\mu} \right] q^*(\mathbf{W}|\tau) q^*(\tau) d\mathbf{W} d\tau \\
&= \int\int \tau \left[\mathbf{W}\mathbf{s}_{\boldsymbol{\mu}} + \mathbf{m}_{\boldsymbol{\mu}} \right] q^*(\mathbf{W}|\tau) q^*(\tau) d\mathbf{W} d\tau \\
&= \langle \tau \mathbf{W} \rangle \mathbf{s}_{\boldsymbol{\mu}} + \langle \tau \rangle \mathbf{m}_{\boldsymbol{\mu}} \\
&= \frac{a_\tau}{b_\tau} (\mathbf{M}_{\mathbf{w}}^T \mathbf{s}_{\boldsymbol{\mu}} + \mathbf{m}_{\boldsymbol{\mu}}).
\end{aligned} \tag{99}$$

B.6 Derivation of $\langle \tau \boldsymbol{\mu}^T \boldsymbol{\mu} \rangle$

$$\begin{aligned}
\langle \tau \boldsymbol{\mu}^T \boldsymbol{\mu} \rangle &= \int\int\int \tau \boldsymbol{\mu}^T \boldsymbol{\mu} q^*(\boldsymbol{\mu}|\mathbf{W},\tau) q^*(\mathbf{W}|\tau) q^*(\tau) d\boldsymbol{\mu} d\mathbf{W} d\tau \\
&= \int\int \tau \left[\int \boldsymbol{\mu}^T \boldsymbol{\mu} q^*(\boldsymbol{\mu}|\mathbf{W},\tau) d\boldsymbol{\mu} \right] q^*(\mathbf{W}|\tau) q^*(\tau) d\mathbf{W} d\tau \\
&= \int\int \tau \left[d(\beta_{\boldsymbol{\mu}}\tau)^{-1} + (\mathbf{W}\mathbf{s}_{\boldsymbol{\mu}} + \mathbf{m}_{\boldsymbol{\mu}})^T (\mathbf{W}\mathbf{s}_{\boldsymbol{\mu}} + \mathbf{m}_{\boldsymbol{\mu}}) \right] q^*(\mathbf{W}|\tau) q^*(\tau) d\mathbf{W} d\tau \\
&= d\beta_{\boldsymbol{\mu}}^{-1} + \mathbf{s}_{\boldsymbol{\mu}}^T \langle \tau \mathbf{W}^T \mathbf{W} \rangle \mathbf{s}_{\boldsymbol{\mu}} + 2\mathbf{s}_{\boldsymbol{\mu}}^T \langle \tau \mathbf{W} \rangle^T \mathbf{m}_{\boldsymbol{\mu}} + \langle \tau \rangle \mathbf{m}_{\boldsymbol{\mu}}^T \mathbf{m}_{\boldsymbol{\mu}} \\
&= d(\beta_{\boldsymbol{\mu}}^{-1} + \mathbf{s}_{\boldsymbol{\mu}}^T \mathbf{\Lambda}_{\mathbf{w}}^{-1} \mathbf{s}_{\boldsymbol{\mu}}) + \frac{a_\tau}{b_\tau} \|\mathbf{M}_{\mathbf{w}}^T \mathbf{s}_{\boldsymbol{\mu}} + \mathbf{m}_{\boldsymbol{\mu}}\|^2.
\end{aligned} \tag{100}$$

B.7 Derivation of $\langle \tau \tilde{\mathbf{w}}_k \rangle$

$$\begin{aligned}
\langle \tau \tilde{\mathbf{w}}_k \rangle &= \int\int \tau \tilde{\mathbf{w}}_k q^*(\tilde{\mathbf{w}}_k|\tau) q^*(\tau) d\tilde{\mathbf{w}}_k d\tau \\
&= \int \tau \left[\int \tilde{\mathbf{w}}_k q^*(\tilde{\mathbf{w}}_k|\tau) d\tilde{\mathbf{w}}_k \right] q^*(\tau) d\tau \\
&= \int \tau \mathbf{m}_{\mathbf{w}}^{(k)} q^*(\tau) d\tau \\
&= \frac{a_\tau}{b_\tau} \mathbf{m}_{\mathbf{w}}^{(k)}. \qquad (101)
\end{aligned}$$

B.8 Derivation of $\langle \tau \tilde{\mathbf{w}}_k \tilde{\mathbf{w}}_k^T \rangle$

$$\begin{aligned}
\langle \tau \tilde{\mathbf{w}}_k \tilde{\mathbf{w}}_k^T \rangle &= \int\int \tau \tilde{\mathbf{w}}_k \tilde{\mathbf{w}}_k^T q^*(\tilde{\mathbf{w}}_k|\tau) q^*(\tau) d\tilde{\mathbf{w}}_k d\tau \\
&= \int \tau \left[\int \tilde{\mathbf{w}}_k \tilde{\mathbf{w}}_k^T q^*(\tilde{\mathbf{w}}_k|\tau) d\tilde{\mathbf{w}}_k \right] q^*(\tau) d\tau \\
&= \int \tau \left[(\tau \mathbf{\Lambda}_{\mathbf{w}})^{-1} + \mathbf{m}_{\mathbf{w}}^{(k)} \mathbf{m}_{\mathbf{w}}^{(k)T} \right] q^*(\tau) d\tau \\
&= \mathbf{\Lambda}_{\mathbf{w}}^{-1} + \frac{a_\tau}{b_\tau} \mathbf{m}_{\mathbf{w}}^{(k)} \mathbf{m}_{\mathbf{w}}^{(k)T} \qquad (102)
\end{aligned}$$

C Derivation of (Variational) Predictive Density

The true predictive density is given by

$$\begin{aligned}
p(\mathbf{t}|D) &= \int\int\int p(\mathbf{t}|\boldsymbol{\mu}, \mathbf{W}, \tau) p(\boldsymbol{\mu}, \mathbf{W}, \tau|D) d\boldsymbol{\mu} d\mathbf{W} d\tau \\
&= \int\int\int\int p(\mathbf{t}|\mathbf{x}, \boldsymbol{\mu}, \mathbf{W}, \tau) p(\mathbf{x}) p(\boldsymbol{\mu}, \mathbf{W}, \tau|D) d\boldsymbol{\mu} d\mathbf{W} d\tau d\mathbf{x}. \qquad (103)
\end{aligned}$$

Approximating the true posterior $p(\boldsymbol{\mu}, \mathbf{W}, \tau|D)$ above with the variational posterior $q^*(\boldsymbol{\mu}, \mathbf{W}, \tau)$, we get

$$\begin{aligned}
p(\mathbf{t}|D) &\approx \int\int\int\int p(\mathbf{t}|\mathbf{x}, \boldsymbol{\mu}, \mathbf{W}, \tau) p(\mathbf{x}) q^*(\boldsymbol{\mu}, \mathbf{W}, \tau) d\boldsymbol{\mu} d\mathbf{W} d\tau d\mathbf{x} \\
&= \tilde{p}(\mathbf{t}|D). \qquad (104)
\end{aligned}$$

Now, we can write

$$\tilde{p}(\mathbf{t}|D) = \int\int\int\int p(\mathbf{t}|\mathbf{x}, \boldsymbol{\mu}, \mathbf{W}, \tau) p(\mathbf{x}) q^*(\boldsymbol{\mu}|\mathbf{W}, \tau) q^*(\mathbf{W}|\tau) q^*(\tau) d\boldsymbol{\mu} d\mathbf{W} d\tau d\mathbf{x}. \qquad (105)$$

Integrating out $\boldsymbol{\mu}$, this gives

$$\tilde{p}(\mathbf{t}|D) = \int\int\int \tilde{p}(\mathbf{t}|\mathbf{x}, \mathbf{W}, \tau, D) p(\mathbf{x}) q^*(\mathbf{W}|\tau) q^*(\tau) d\mathbf{W} d\tau d\mathbf{x}, \qquad (106)$$

where

$$\begin{aligned}
\tilde{p}(\mathbf{t}|\mathbf{x},\mathbf{W},\tau,D) &= \int p(\mathbf{t}|\mathbf{x},\boldsymbol{\mu},\mathbf{W},\tau)q^*(\boldsymbol{\mu}|\mathbf{W},\tau)d\boldsymbol{\mu} \\
&= \frac{\tau^{d/2}}{(2\pi)^{d/2}}\frac{(\beta_{\boldsymbol{\mu}}\tau)^{d/2}}{(2\pi)^{d/2}}\int \exp\left\{-\frac{\tau}{2}\|\mathbf{t}-\mathbf{W}\mathbf{x}-\boldsymbol{\mu}\|^2 - \frac{\beta_{\boldsymbol{\mu}}\tau}{2}\|\boldsymbol{\mu}-\mathbf{W}\mathbf{s}_{\boldsymbol{\mu}}-\mathbf{m}_{\boldsymbol{\mu}}\|^2\right\}d\boldsymbol{\mu} \\
&= \frac{\tau^{d/2}}{(2\pi)^{d/2}}\frac{(\beta_{\boldsymbol{\mu}}\tau)^{d/2}}{(2\pi)^{d/2}}\exp\left\{-\frac{\tau}{2}\left(\|\mathbf{t}-\mathbf{W}\mathbf{x}\|^2+\beta_{\boldsymbol{\mu}}\|\mathbf{W}\mathbf{s}_{\boldsymbol{\mu}}+\mathbf{m}_{\boldsymbol{\mu}}\|^2\right)\right\}\times \\
&\quad \int \exp\left\{-\frac{\tau}{2}(\beta_{\boldsymbol{\mu}}+1)\left(\boldsymbol{\mu}^t\boldsymbol{\mu}-2\boldsymbol{\mu}^T\frac{1}{(\beta_{\boldsymbol{\mu}}+1)}(\mathbf{t}-\mathbf{W}\mathbf{x}+\beta_{\boldsymbol{\mu}}(\mathbf{W}\mathbf{s}_{\boldsymbol{\mu}}+\mathbf{m}_{\boldsymbol{\mu}}))\right)\right\}d\boldsymbol{\mu} \\
&= \frac{\tau^{d/2}}{(2\pi)^{d/2}}\left(\frac{\beta_{\boldsymbol{\mu}}}{\beta_{\boldsymbol{\mu}}+1}\right)^{d/2}\times \\
&\quad \exp\left\{-\frac{\tau}{2}\left(\|\mathbf{t}-\mathbf{W}\mathbf{x}\|^2+\beta_{\boldsymbol{\mu}}\|\mathbf{W}\mathbf{s}_{\boldsymbol{\mu}}+\mathbf{m}_{\boldsymbol{\mu}}\|^2-\frac{1}{(\beta_{\boldsymbol{\mu}}+1)}\|\mathbf{t}-\mathbf{W}\mathbf{x}+\beta_{\boldsymbol{\mu}}(\mathbf{W}\mathbf{s}_{\boldsymbol{\mu}}+\mathbf{m}_{\boldsymbol{\mu}})\|^2\right)\right\} \\
&= \frac{\tau^{d/2}}{(2\pi)^{d/2}}\left(\frac{\beta_{\boldsymbol{\mu}}}{\beta_{\boldsymbol{\mu}}+1}\right)^{d/2}\exp\left\{-\left(\frac{\beta_{\boldsymbol{\mu}}}{\beta_{\boldsymbol{\mu}}+1}\right)\frac{\tau}{2}\|\mathbf{t}-\mathbf{W}\mathbf{x}+\mathbf{W}\mathbf{s}_{\boldsymbol{\mu}}+\mathbf{m}_{\boldsymbol{\mu}}\|^2\right\} \\
&= \mathcal{N}(\mathbf{t}|\mathbf{W}\mathbf{x}-(\mathbf{W}\mathbf{s}_{\boldsymbol{\mu}}+\mathbf{m}_{\boldsymbol{\mu}}),(\tilde{\beta}\tau)^{-1}\mathbf{I}_d), \qquad (107)
\end{aligned}$$

with

$$\tilde{\beta} = \frac{\beta_{\boldsymbol{\mu}}}{\beta_{\boldsymbol{\mu}}+1}. \qquad (108)$$

Next, we integrate out $\mathbf{W}$ from Eq. (106). This gives

$$\tilde{p}(\mathbf{t}|D) = \int\int \tilde{p}(\mathbf{t}|\mathbf{x},\tau,D)p(\mathbf{x})q^*(\tau)d\tau d\mathbf{x}, \qquad (109)$$

where

$$\begin{aligned}
\tilde{p}(\mathbf{t}|\mathbf{x},\tau,D) &= \int \tilde{p}(\mathbf{t}|\mathbf{x},\mathbf{W},\tau,D)q^*(\mathbf{W}|\tau)d\mathbf{W} \\
&= \frac{(\tilde{\beta}\tau)^{d/2}}{(2\pi)^{d/2}}\frac{\tau^{d(d-1)/2}|\boldsymbol{\Lambda}_{\mathbf{w}}|^{d/2}}{(2\pi)^{d(d-1)/2}}\times \\
&\quad \int \exp\left\{-\frac{\tilde{\beta}\tau}{2}\|\mathbf{t}-\mathbf{W}\mathbf{x}+\mathbf{W}\mathbf{s}_{\boldsymbol{\mu}}+\mathbf{m}_{\boldsymbol{\mu}}\|^2-\frac{\tau}{2}\sum_{k=1}^{d}(\tilde{\mathbf{w}}_k-\mathbf{m}_{\mathbf{w}}^{(k)})^T\boldsymbol{\Lambda}_{\mathbf{w}}(\tilde{\mathbf{w}}_k-\mathbf{m}_{\mathbf{w}}^{(k)})\right\}d\mathbf{W} \\
&= \frac{(\tilde{\beta}\tau)^{d/2}}{(2\pi)^{d/2}}\frac{\tau^{d(d-1)/2}|\boldsymbol{\Lambda}_{\mathbf{w}}|^{d/2}}{(2\pi)^{d(d-1)/2}}\exp\left\{-\frac{\tilde{\beta}\tau}{2}\|\mathbf{t}+\mathbf{m}_{\boldsymbol{\mu}}\|^2-\frac{\tau}{2}\sum_{k=1}^{d}\mathbf{m}_{\mathbf{w}}^{(k)T}\boldsymbol{\Lambda}_{\mathbf{w}}\mathbf{m}_{\mathbf{w}}^{(k)}\right\}\times \\
&\quad \int \exp\left\{-\frac{\tilde{\tau}}{2}\sum_{k=1}^{d}\left(\tilde{\mathbf{w}}_k^T(\boldsymbol{\Lambda}_{\mathbf{w}}+\tilde{\beta}(\mathbf{x}-\mathbf{s}_{\boldsymbol{\mu}})(\mathbf{x}-\mathbf{s}_{\boldsymbol{\mu}})^T)\tilde{\mathbf{w}}_k-2\tilde{\mathbf{w}}_k^T(\boldsymbol{\Lambda}_{\mathbf{w}}\mathbf{m}_{\mathbf{w}}^{(k)}+\tilde{\beta}(t_k+m_{\boldsymbol{\mu}k})(\mathbf{x}-\mathbf{s}_{\boldsymbol{\mu}}))\right)\right\}d\mathbf{W} \\
&= \frac{(\tilde{\beta}\tau)^{d/2}}{(2\pi)^{d/2}}|\boldsymbol{\Lambda}_{\mathbf{w}}\tilde{\boldsymbol{\Lambda}}(\mathbf{x})^{-1}|^{d/2}\times \\
&\quad \exp\left\{-\frac{\tilde{\beta}\tau}{2}\|\mathbf{t}+\mathbf{m}_{\boldsymbol{\mu}}\|^2-\frac{\tau}{2}\sum_{k=1}^{d}\left(\mathbf{m}_{\mathbf{w}}^{(k)T}\boldsymbol{\Lambda}_{\mathbf{w}}\mathbf{m}_{\mathbf{w}}^{(k)}-\tilde{\mathbf{m}}_k(\mathbf{t},\mathbf{x})^T\tilde{\boldsymbol{\Lambda}}(\mathbf{x})\tilde{\mathbf{m}}_k(\mathbf{t},\mathbf{x})\right)\right\}, \qquad (110)
\end{aligned}$$

with

$$\begin{aligned}
\tilde{\boldsymbol{\Lambda}}(\mathbf{x}) &= \boldsymbol{\Lambda}_{\mathbf{w}}+\tilde{\beta}(\mathbf{x}-\mathbf{s}_{\boldsymbol{\mu}})(\mathbf{x}-\mathbf{s}_{\boldsymbol{\mu}})^T, \qquad (111) \\
\tilde{\mathbf{m}}_k(\mathbf{t},\mathbf{x}) &= \tilde{\boldsymbol{\Lambda}}(\mathbf{x})^{-1}\left(\boldsymbol{\Lambda}_{\mathbf{w}}\mathbf{m}_{\mathbf{w}}^{(k)}+\tilde{\beta}(t_k+m_{\boldsymbol{\mu}k})(\mathbf{x}-\mathbf{s}_{\boldsymbol{\mu}})\right). \qquad (112)
\end{aligned}$$

Now,

$$\tilde{\mathbf{\Lambda}}(\mathbf{x})^{-1} = \mathbf{\Lambda}^{-1} - \left(\frac{1}{\tilde{\beta}^{-1} + (\mathbf{x} - \mathbf{s}_{\boldsymbol{\mu}})^T \mathbf{\Lambda}_{\mathbf{w}}^{-1} (\mathbf{x} - \mathbf{s}_{\boldsymbol{\mu}})} \right) \mathbf{\Lambda}^{-1} (\mathbf{x} - \mathbf{s}_{\boldsymbol{\mu}})(\mathbf{x} - \mathbf{s}_{\boldsymbol{\mu}})^T \mathbf{\Lambda}^{-1}. \quad (113)$$

With some algebraic manipulation, Eqs. (110-113) above give

$$\tilde{p}(\mathbf{t}|\mathbf{x}, \tau, D) = \mathcal{N}\left(\mathbf{t}|\mathbf{M}_{\mathbf{w}}^T \mathbf{x} - (\mathbf{M}_{\mathbf{w}}^T \mathbf{s}_{\boldsymbol{\mu}} + \mathbf{m}_{\boldsymbol{\mu}}), (\tilde{\beta}^{-1} + (\mathbf{x} - \mathbf{s}_{\boldsymbol{\mu}})^T \mathbf{\Lambda}_{\mathbf{w}}^{-1} (\mathbf{x} - \mathbf{s}_{\boldsymbol{\mu}}))\tau^{-1} \mathbf{I}_d\right). \quad (114)$$

Finally, integrating out τ from Eq. (109), we get

$$\tilde{p}(\mathbf{t}|D) = \int \tilde{p}(\mathbf{t}|\mathbf{x}, D) p(\mathbf{x}) d\mathbf{x}, \quad (115)$$

where

$$\begin{aligned} \tilde{p}(\mathbf{t}|\mathbf{x}, D) &= \int \tilde{p}(\mathbf{t}|\mathbf{x}, \tau, D) q^*(\tau) d\tau \\ &= \mathcal{S}(\mathbf{t}|\mathbf{m}_{\mathbf{t}}(\mathbf{x}), \beta_{\mathbf{t}}(\mathbf{x})^{-1} \mathbf{I}_d, \nu_{\mathbf{t}}), \end{aligned} \quad (116)$$

with

$$\mathbf{m}_{\mathbf{t}}(\mathbf{x}) = \mathbf{M}_{\mathbf{w}}^T \mathbf{x} - (\mathbf{M}_{\mathbf{w}}^T \mathbf{s}_{\boldsymbol{\mu}} + \mathbf{m}_{\boldsymbol{\mu}}), \quad (117)$$

$$\beta_{\mathbf{t}}(\mathbf{x}) = \frac{b_\tau}{a_\tau} (1 + \beta_{\boldsymbol{\mu}}^{-1} + (\mathbf{x} - \mathbf{s}_{\boldsymbol{\mu}})^T \mathbf{\Lambda}_{\mathbf{w}}^{-1} (\mathbf{x} - \mathbf{s}_{\boldsymbol{\mu}})), \quad (118)$$

$$\nu_{\mathbf{t}} = 2a_\tau. \quad (119)$$