Summary of Variational Bayesian Least Squares

Jo-Anne Ting
joanneti@usc.edu

December 12, 2006

1 Model

Given we have a training dataset $D = \{y_i, \mathbf{x}_i\}_{i=1}^N$ that consists of a scalar output y_i and d-dimensional inputs $\mathbf{x}_i$ (a d by 1 vector) for each of the N data samples, the model for Variational Bayesian Least Squares consists of the following distributions:

$$\begin{aligned} y_i|\mathbf{z}_i &\sim \text{Normal}\left(\mathbf{1}^T\mathbf{z}_i, \psi_y\right) \\ z_{im}|b_m, \alpha_m &\sim \text{Normal}\left(b_m x_{im}, \frac{\psi_{zm}}{\alpha_m}\right) \\ b_m|\alpha_m &\sim \text{Normal}\left(0, \frac{1}{\alpha_m}\right) \\ \alpha_m &\sim \text{Gamma}\left(a_{\alpha m}, b_{\alpha m}\right) \end{aligned} \tag{1}$$

where $\mathbf{1}$ is a d by 1 vector of ones. We can use the variational factorial assumption that the posterior over the unknown variables $\theta = \{\boldsymbol{\alpha}, \mathbf{b}, \mathbf{Z}\}$ factorizes as $Q(\alpha, \mathbf{b})Q(\mathbf{Z})$, where $\mathbf{Z}$ is a diagonal matrix with a diagonal vector of $\mathbf{z}$. The complete log evidence is:

$$\log p(\mathbf{y}, \theta|\mathbf{X}; \phi) = \sum_{i=1}^N \log p(y_i|\mathbf{z}_i) + \sum_{i=1}^N \sum_{m=1}^d \log p(z_{im}|b_m, \alpha_m) + \sum_{m=1}^d \log p(b_m|\alpha_m) + \sum_{m=1}^d \log p(\alpha_m)$$

where ϕ are the point-estimated parameters $\phi = \{\psi_y, \psi_z\}$. The EM update equations can now be derived using standard manipulations of Normal and Gamma distributions.

2 EM Update Equations

The final EM update equations are listed below:

E-step:

$$\boldsymbol{\Sigma}_z = \left(\frac{1}{\psi_y}\mathbf{1}\mathbf{1}^T + \boldsymbol{\Psi}_z^{-1}\langle\mathbf{A}\rangle\right)^{-1} = \boldsymbol{\Psi}_z\langle\mathbf{A}\rangle^{-1} - \frac{\boldsymbol{\Psi}_z\langle\mathbf{A}\rangle^{-1}\mathbf{1}\mathbf{1}^T\boldsymbol{\Psi}_z\langle\mathbf{A}\rangle^{-1}}{\psi_y + \mathbf{1}^T\boldsymbol{\Psi}_z\langle\mathbf{A}\rangle^{-1}\mathbf{1}} \tag{2}$$

$$\begin{aligned}\langle\mathbf{z}_i\rangle &= \boldsymbol{\Sigma}_z\left(\frac{1}{\psi_y}\mathbf{1}y_i + \boldsymbol{\Psi}_z^{-1}\langle\mathbf{A}\rangle\langle\mathbf{B}\rangle\mathbf{x}_i\right) \\ &= \left(\frac{\boldsymbol{\Psi}_z\langle\mathbf{A}\rangle^{-1}\mathbf{1}}{\psi_y + \mathbf{1}^T\boldsymbol{\Psi}_z\langle\mathbf{A}\rangle^{-1}\mathbf{1}}\right)y_i + \left(\langle\mathbf{B}\rangle - \frac{\boldsymbol{\Psi}_z\langle\mathbf{A}\rangle^{-1}\mathbf{1}\mathbf{1}^T\langle\mathbf{B}\rangle}{\psi_y + \mathbf{1}^T\boldsymbol{\Psi}_z\langle\mathbf{A}\rangle^{-1}\mathbf{1}}\right)\mathbf{x}_i\end{aligned} \tag{3}$$

$$\sigma_{b_m}^2 = \frac{\psi_{zm}}{\langle\alpha_m\rangle}\left(\sum_{i=1}^N x_{im}^2 + \psi_{zm}\right)^{-1} \tag{4}$$

$$\langle b_m\rangle = \left(\sum_{i=1}^N x_{im}^2 + \psi_{zm}\right)^{-1}\left(\sum_{i=1}^N\langle z_{im}\rangle x_{im}\right) \tag{5}$$

$$\hat{a}_\alpha = a_{\alpha 0} + \frac{N}{2} \tag{6}$$

$$\hat{b}_{\alpha m} = b_{\alpha m0} + \frac{1}{2\psi_{zm}}\left\{\sum_{i=1}^N\left\langle z_{im}^2\right\rangle - \left(\sum_{i=1}^N x_{im}^2 + \psi_{zm}\right)^{-1}\left(\sum_{i=1}^N\langle z_{im}\rangle x_{im}\right)^2\right\} \tag{7}$$

M-step:

$$\psi_y = \frac{1}{N}\sum_{i=1}^N\left(y_i - \mathbf{1}^T\langle\mathbf{z}_i\rangle\right)^2 + \mathbf{1}^T\boldsymbol{\Sigma}_z\mathbf{1} \tag{8}$$

$$\psi_{zm} = \frac{1}{N}\sum_{i=1}^N\langle\alpha_m\rangle\left(\langle z_{im}\rangle - \langle b_m\rangle x_{im}\right)^2 + \langle\alpha_m\rangle\sigma_{zm}^2 + \langle\alpha_m\rangle\sigma_{b_m}^2\left(\frac{1}{N}\sum_{i=1}^N x_{im}^2\right) \tag{9}$$

where $\langle\mathbf{A}\rangle$, $\langle\mathbf{B}\rangle$, $\boldsymbol{\Psi}_z$ are diagonal matrices with diagonal vectors of $\langle\alpha\rangle$, $\langle\mathbf{b}\rangle$, ψ_z, respectively; $a_{\alpha 0} = 10^{-8}$ and $b_{\alpha 0} = 10^{-8}\mathbf{1}$; and $\boldsymbol{\Sigma}_z$ is the covariance matrix of $\mathbf{z}$ that is a diagonal matrix with a diagonal vector of σ_z^2.

3 Monitoring the Incomplete Log Likelihood

To know when to stop iterating through the EM algorithm above, we should monitor the incomplete log likelihood and stop when the value appears to have converged. To calculate the incomplete log likelihood, we need to integrate out the variables α, $\mathbf{b}$, $\mathbf{Z}$ from the complete log likelihood expression. But since the calculation of the true posterior distribution $Q(\theta)$ is intractable, we cannot determine the true incomplete log likelihood. Hence, for the purpose of monitoring the incomplete log likelihood in the EM algorithm, we can monitor the lower bound of the incomplete log likelihood instead.

In the derivation of the EM algorithm, we reached an estimate of $Q(\theta)$, where $\theta = \{\alpha, \mathbf{b}, \mathbf{Z}\}$, to be $Q(\theta) = Q(\alpha, \mathbf{b})Q(\mathbf{Z})$. The lower bound to the incomplete log likelihood, where $\phi = \{\psi_y, \psi_z\}$

is:

$$\begin{aligned}\log p(\mathbf{y}|\mathbf{X};\phi) &\geq \int Q(\theta)\log\frac{p(\mathbf{y},\theta|\mathbf{X};\phi)}{Q(\theta)}d\theta \\ &= \int Q(\theta)\log p(\mathbf{y},\theta|\mathbf{X};\phi)d\theta - \int Q(\theta)\log Q(\theta)d\theta \\ &= \left\langle \log p(\mathbf{y},\theta|\mathbf{X};\phi)\right\rangle_{Q(\theta)} - \int Q(\theta)\log Q(\theta)d\theta\end{aligned} \tag{10}$$

where:

$$\begin{aligned}&\int Q(\theta)\log Q(\theta)d\theta \\ &= \int\int Q(\boldsymbol{\beta}|\boldsymbol{\alpha})\log Q(\boldsymbol{\beta}|\boldsymbol{\alpha})d\boldsymbol{\alpha}d\boldsymbol{\beta} + \int Q(\boldsymbol{\alpha})\log Q(\boldsymbol{\alpha})d\boldsymbol{\alpha} + \int Q(\mathbf{Z})\log Q(\mathbf{Z})d\mathbf{Z} \\ &= -\frac{1}{2}\sum_{m=1}^{d}\left\langle\alpha_m\right\rangle\left(\sigma_{b_m}^2+1\right) + \sum_{m=1}^{d}\log\hat{b}_{\alpha m} - \frac{1}{2}\log|\boldsymbol{\Sigma}_z| + \text{const}\end{aligned} \tag{11}$$

and:

$$\begin{aligned}&\left\langle\log p(\mathbf{y},\theta|\mathbf{X};\phi)\right\rangle_{Q(\theta)} \\ &= -\frac{N}{2}\log\psi_y - \frac{1}{2\psi_y}\sum_{i=1}^{N}\left(y_i - 2y_i\mathbf{1}^T\left\langle\mathbf{z}_i\right\rangle + \mathbf{1}^T\left\langle\mathbf{z}_i\mathbf{z}_i^T\right\rangle\mathbf{1}\right) \\ &\quad - \frac{N}{2}\sum_{m=1}^{d}\log\psi_{zm} + \frac{N}{2}\sum_{m=1}^{d}\left\langle\log\alpha_m\right\rangle - \sum_{m=1}^{d}\frac{\left\langle\alpha_m\right\rangle}{2\psi_{zm}}\sum_{i=1}^{N}\left(\left\langle z_{im}^2\right\rangle - 2\left\langle b_m\right\rangle x_{im} + \left\langle b_m^2\right\rangle x_{im}^2\right) \\ &\quad + \frac{1}{2}\sum_{m=1}^{d}\left\langle\log\alpha_m\right\rangle - \frac{1}{2}\sum_{m=1}^{d}\left\langle\alpha_m\right\rangle\left\langle b_m^2\right\rangle + \sum_{m=1}^{d}\left(\hat{a}_{\alpha m}-1\right)\left\langle\log\alpha_m\right\rangle \\ &\quad - \sum_{m=1}^{d}\hat{b}_{\alpha m}\left\langle\alpha_m\right\rangle + \sum_{m=1}^{d}\hat{a}_{\alpha m}\log\hat{b}_{\alpha m} + \text{const} \\ &= -\frac{N}{2}\log\psi_y - \frac{1}{2\psi_y}\sum_{i=1}^{N}\left(y_i - 2y_i\mathbf{1}^T\left\langle\mathbf{z}_i\right\rangle + \mathbf{1}^T\left\langle\mathbf{z}_i\mathbf{z}_i^T\right\rangle\mathbf{1}\right) \\ &\quad - \frac{N}{2}\sum_{m=1}^{d}\log\psi_{zm} - \sum_{m=1}^{d}\frac{\left\langle\alpha_m\right\rangle}{2\psi_{zm}}\sum_{i=1}^{N}\left(\left\langle z_{im}^2\right\rangle - 2\left\langle b_m\right\rangle x_{im} + \left\langle b_m^2\right\rangle x_{im}^2\right) \\ &\quad + \sum_{m=1}^{d}\left(\hat{a}_{\alpha m} + \frac{N+1}{2} - 1\right)\left\langle\log\alpha_m\right\rangle - \sum_{m=1}^{d}\left(\hat{b}_{\alpha m} + \frac{\left\langle b_m^2\right\rangle}{2}\right)\left\langle\alpha_m\right\rangle \\ &\quad + \sum_{m=1}^{d}\hat{a}_{\alpha m}\log\hat{b}_{\alpha m} + \text{const}\end{aligned}$$

Since $\langle \log \alpha_m \rangle = \Psi(\hat{a}_{\alpha m}) - \log \hat{b}_{\alpha m}$:

$$
\begin{aligned}
&\langle \log p(\mathbf{y}, \theta | \mathbf{X}; \phi) \rangle_{Q(\theta)} \\
&= -\frac{N}{2} \log \psi_y - \frac{1}{2\psi_y} \sum_{i=1}^{N} \left(y_i - 2 y_i \mathbf{1}^T \langle \mathbf{z}_i \rangle + \mathbf{1}^T \left\langle \mathbf{z}_i \mathbf{z}_i^T \right\rangle \mathbf{1} \right) \\
&\quad - \frac{N}{2} \sum_{m=1}^{d} \log \psi_{zm} - \sum_{m=1}^{d} \frac{\langle \alpha_m \rangle}{2\psi_{zm}} \sum_{i=1}^{N} \left(\left\langle z_{im}^2 \right\rangle - 2 \langle b_m \rangle x_{im} + \left\langle b_m^2 \right\rangle x_{im}^2 \right) \\
&\quad + \sum_{m=1}^{d} \left(\hat{a}_{\alpha m} + \frac{N+1}{2} - 1 \right) \left(\boldsymbol{\Psi}(\hat{a}_{\alpha m}) - \log \hat{b}_{\alpha m} \right) - \sum_{m=1}^{d} \left(\hat{b}_{\alpha m} + \frac{\left\langle b_m^2 \right\rangle}{2} \right) \langle \alpha_m \rangle \\
&\quad + \sum_{m=1}^{d} \hat{a}_{\alpha m} \log \hat{b}_{\alpha m} + \text{const} \\
&= -\frac{N}{2} \log \psi_y - \frac{1}{2\psi_y} \sum_{i=1}^{N} \left(y_i - 2 y_i \mathbf{1}^T \langle \mathbf{z}_i \rangle + \mathbf{1}^T \left\langle \mathbf{z}_i \mathbf{z}_i^T \right\rangle \mathbf{1} \right) \\
&\quad - \frac{N}{2} \sum_{m=1}^{d} \log \psi_{zm} - \sum_{m=1}^{d} \frac{\langle \alpha_m \rangle}{2\psi_{zm}} \sum_{i=1}^{N} \left(\left\langle z_{im}^2 \right\rangle - 2 \langle b_m \rangle x_{im} + \left\langle b_m^2 \right\rangle x_{im}^2 \right) \\
&\quad - \sum_{m=1}^{d} \left(\hat{b}_{\alpha m} + \frac{\left\langle b_m^2 \right\rangle}{2} \right) \langle \alpha_m \rangle + \sum_{m=1}^{d} \left(\hat{a}_{\alpha m} + \frac{N+1}{2} - 1 \right) \Psi(\hat{a}_{\alpha m}) \\
&\quad - \sum_{m=1}^{d} \left(\frac{N+1}{2} - 1 \right) \log \hat{b}_{\alpha m} + \text{const}
\end{aligned} \tag{12}
$$

We can put Eqs. (11) and (12) together so that Eq. (10) becomes:

$$
\begin{aligned}
&\log p(\mathbf{y} | \mathbf{X}; \phi) \geq \\
&\quad - \frac{N}{2} \log \psi_y - \frac{1}{2\psi_y} \sum_{i=1}^{N} \left(y_i - 2 y_i \mathbf{1}^T \langle \mathbf{z}_i \rangle + \mathbf{1}^T \left\langle \mathbf{z}_i \mathbf{z}_i^T \right\rangle \mathbf{1} \right) \\
&\quad - \frac{N}{2} \sum_{m=1}^{d} \log \psi_{zm} - \sum_{m=1}^{d} \frac{\langle \alpha_m \rangle}{2\psi_{zm}} \sum_{i=1}^{N} \left(\left\langle z_{im}^2 \right\rangle - 2 \langle b_m \rangle x_{im} + \left\langle b_m^2 \right\rangle x_{im}^2 \right) \\
&\quad - \sum_{m=1}^{d} \langle \alpha_m \rangle \left(\hat{b}_{\alpha m} + \frac{\left\langle b_m^2 \right\rangle}{2} \right) + \sum_{m=1}^{d} \left(\hat{a}_{\alpha m} + \frac{N+1}{2} - 1 \right) \Psi(\hat{a}_{\alpha m}) - \sum_{m=1}^{d} \left(\frac{N+1}{2} - 1 \right) \log \hat{b}_{\alpha m} \\
&\quad + \frac{1}{2} \sum_{m=1}^{d} \langle \alpha_m \rangle \left(\sigma_{b_m}^2 + 1 \right) - \sum_{m=1}^{d} \log \hat{b}_{\alpha m} + \frac{1}{2} \log |\boldsymbol{\Sigma}_z| + \text{const}
\end{aligned} \tag{13}
$$