Summary of Bayesian Linear Regression for High-Dimensional Data with Input Noise

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 high-dimensional Bayesian linear regression with input noise consists of the following distributions:

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

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 $\theta = \{\boldsymbol{\alpha}, \mathbf{w}_z, \mathbf{w}_x, \mathbf{Z}, \mathbf{T}\}$, $Q(\theta)$, factorizes as $Q(\alpha)Q(\mathbf{w}_z)Q(\mathbf{w}_x)Q(\mathbf{Z}, \mathbf{T})$, where $\mathbf{Z}$ and $\mathbf{T}$ are diagonal matrices with diagonal vectors of $\mathbf{z}$ and $\mathbf{t}$, respectively. The complete log evidence is:

$$\begin{aligned}
\log p(\mathbf{y}, \theta|\mathbf{X}; \phi) = &\sum_{i=1}^N \log p(y_i|\mathbf{z}_i) + \sum_{i=1}^N\sum_{m=1}^d \log p(z_{im}|w_{zm}, t_{im}) + \sum_{i=1}^N\sum_{m=1}^d \log p(x_{im}|w_{xm}, t_{im}) \\
&+ \sum_{i=1}^N\sum_{m=1}^d \log p(t_{im}) + \sum_{m=1}^d \log\left[p(w_{zm}|\alpha_m)p(\alpha_m)\right] + \sum_{m=1}^d \log\left[p(w_x|\alpha_m)p(\alpha_m)\right]
\end{aligned}$$

where ϕ are the point-estimated parameters $\phi = \{\psi_y, \psi_z, \psi_x\}$. 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:

$$\boldsymbol{\Sigma}_{zz} = \mathbf{M} - \frac{\mathbf{M}\mathbf{1}\mathbf{1}^T\mathbf{M}}{\psi_y + \mathbf{1}^T\mathbf{M}\mathbf{1}} \tag{2}$$

$$\boldsymbol{\Sigma}_{tt} = \mathbf{K}^{-1} + \mathbf{K}^{-1}\left\langle \mathbf{W}_z \right\rangle^T \boldsymbol{\Psi}_z^{-1} \boldsymbol{\Sigma}_{zz} \boldsymbol{\Psi}_z^{-1} \left\langle \mathbf{W}_z \right\rangle \mathbf{K}^{-1} \tag{3}$$

$$\boldsymbol{\Sigma}_{zt} = -\boldsymbol{\Sigma}_{zz}\left\langle \mathbf{W}_z \right\rangle \boldsymbol{\Psi}_z^{-1}\mathbf{K}^{-1} \tag{4}$$

$$\boldsymbol{\Sigma}_{tz} = \boldsymbol{\Sigma}_{zt}^T \tag{5}$$

$$\left\langle \mathbf{z}_i \right\rangle = \boldsymbol{\Sigma}_{zz}\mathbf{1}\frac{y_i}{\psi_y} + \boldsymbol{\Sigma}_{tz}\boldsymbol{\Psi}_x^{-1}\left\langle \mathbf{W}_x \right\rangle^T \mathbf{x}_i \tag{6}$$

$$\left\langle \mathbf{t}_i \right\rangle = \mathbf{K}^{-1}\boldsymbol{\Psi}_x^{-1}\left\langle \mathbf{W}_z \right\rangle \boldsymbol{\Sigma}_{zz}\mathbf{1}\frac{y_i}{\psi_y} + \boldsymbol{\Sigma}_{tt}\boldsymbol{\Psi}_x^{-1}\left\langle \mathbf{W}_x \right\rangle^T \mathbf{x}_i \tag{7}$$

$$\sigma^2_{w_{zm}} = \frac{1}{\frac{1}{\psi_{zm}}\sum_{i=1}^N \left\langle t_{im}^2 \right\rangle + \left\langle \alpha_m \right\rangle} \tag{8}$$

$$\left\langle w_{zm} \right\rangle = \frac{\sigma^2_{w_{zm}}}{\psi_{zm}}\left(\sum_{i=1}^N \left\langle z_{im}t_{im} \right\rangle\right) \tag{9}$$

$$\sigma^2_{w_{xm}} = \frac{1}{\frac{1}{\psi_{xm}}\sum_{i=1}^N \left\langle t_{im}^2 \right\rangle + \left\langle \alpha_m \right\rangle} \tag{10}$$

$$\left\langle w_{xm} | \alpha_m \right\rangle = \frac{\sigma^2_{w_{xm}}}{\psi_{xm}}\left(\sum_{i=1}^N x_{im}\left\langle t_{im} \right\rangle\right) \tag{11}$$

$$\hat{a}_{\alpha m} = a_{\alpha m0} + 1 \tag{12}$$

$$\hat{b}_{\alpha m} = b_{\alpha m0} + \frac{\left\langle w_{xm}^2 \right\rangle + \left\langle w_{zm}^2 \right\rangle}{2} \tag{13}$$

M-step:

$$\psi_y = \frac{1}{N}\sum_{i=1}^N \left(y_i^2 - 2\mathbf{1}y_i\left\langle \mathbf{z}_i \right\rangle + \mathbf{1}^T\left\langle \mathbf{z}_i\mathbf{z}_i^T \right\rangle \mathbf{1}\right) \tag{14}$$

$$\psi_{zm} = \frac{1}{N}\sum_{i=1}^N \left(\left\langle z_{im}^2 \right\rangle - 2\left\langle w_{zm} \right\rangle \left\langle z_{im}t_{im} \right\rangle + \left\langle w_{zm}^2 \right\rangle \left\langle t_{im}^2 \right\rangle\right) \tag{15}$$

$$\psi_{xm} = \frac{1}{N}\sum_{i=1}^N \left(x_{im}^2 - 2\left\langle w_{xm} \right\rangle \left\langle t_{im} \right\rangle x_{im} + \left\langle w_{xm}^2 \right\rangle \left\langle t_{im}^2 \right\rangle\right) \tag{16}$$

where

$$\mathbf{K} = \mathbf{I} + \left\langle \mathbf{W}_x^T\mathbf{W}_x \right\rangle \boldsymbol{\Psi}_x^{-1} + \left\langle \mathbf{W}_z^T\mathbf{W}_z \right\rangle \boldsymbol{\Psi}_z^{-1}$$

$$\mathbf{M} = \boldsymbol{\Psi}_z + \left\langle \mathbf{W}_z \right\rangle \left(\mathbf{I} + \left\langle \mathbf{W}_x^T\mathbf{W}_x \right\rangle \boldsymbol{\Psi}_x^{-1} + \boldsymbol{\Sigma}_{\mathbf{w}_z}\boldsymbol{\Psi}_z^{-1}\right)^{-1}\left\langle \mathbf{W}_z \right\rangle^T$$

$$\sigma^2_{\mathbf{z}} = \operatorname{diag}\left(\boldsymbol{\Sigma}_{zz}\right), \sigma^2_{\mathbf{t}} = \operatorname{diag}\left(\boldsymbol{\Sigma}_{t}\right), \operatorname{cov}\left(z, t\right) = \operatorname{diag}\left(\boldsymbol{\Sigma}_{zt}\right)$$

where $\left\langle \mathbf{A} \right\rangle$, $\left\langle \mathbf{W}_z \right\rangle$, $\left\langle \mathbf{W}_x \right\rangle$, $\boldsymbol{\Psi}_z$, $\boldsymbol{\Psi}_x$, $\boldsymbol{\Sigma}_{\mathbf{w}_z}$ are diagonal matrices with diagonal vectors of $\left\langle \alpha \right\rangle$, $\left\langle \mathbf{w}_z \right\rangle$, $\left\langle \mathbf{w}_x \right\rangle$, ψ_z, ψ_x, $\sigma^2_{\mathbf{w}_z}$ respectively; and $a_{\alpha 0} = b_{\alpha 0} = 10^{-8}\mathbf{1}$. Note that the diagonal matrices

$\left\langle \mathbf{W}_z^T \mathbf{W}_z \right\rangle$ and $\left\langle \mathbf{W}_x^T \mathbf{W}_x \right\rangle$ have diagonal vectors each composed of elements $\left\langle w_{zm}^2 \right\rangle$ and $\left\langle w_{xm}^2 \right\rangle$, respectively.

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, $\mathbf{w}_z$, $\mathbf{w}_x$, $\mathbf{Z}$, $\mathbf{T}$ 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 $\theta = \{\alpha, \mathbf{w}_z, \mathbf{w}_x, \mathbf{Z}, \mathbf{T}\}$, to be $Q(\alpha)Q(\mathbf{w}_z)Q(\mathbf{w}_x)Q(\mathbf{Z}, \mathbf{T})$. The lower bound to the incomplete log likelihood, where $\phi = \{\psi_y, \psi_z, \psi_x\}$ is:

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

And this simplifies to:

$$\begin{aligned}
&\log p(\mathbf{y}|\mathbf{X};\phi) \geq \\
&\quad -\frac{N}{2}\log\psi_y - \frac{1}{2\psi_y}\sum_{i=1}^{N}\left(y_i - 2y_i\mathbf{1}^T\langle\mathbf{z}_i\rangle + \mathbf{1}^T\left\langle\mathbf{z}_i\mathbf{z}_i^T\right\rangle\mathbf{1}\right) \\
&\quad -\frac{N}{2}\sum_{m=1}^{d}\log\psi_{zm} - \sum_{m=1}^{d}\frac{1}{2\psi_{zm}}\sum_{i=1}^{N}\left(\left\langle z_{im}^2\right\rangle - 2\langle w_{zm}\rangle\langle z_{im}t_{im}\rangle + \left\langle w_{zm}^2\right\rangle\left\langle t_{im}^2\right\rangle\right) \\
&\quad -\frac{N}{2}\sum_{m=1}^{d}\log\psi_{xm} - \sum_{m=1}^{d}\frac{1}{2\psi_{xm}}\sum_{i=1}^{N}\left(x_{im}^2 - 2\langle w_{xm}\rangle\langle t_{im}\rangle x_{im} + \left\langle w_{xm}^2\right\rangle\left\langle t_{im}^2\right\rangle\right) \\
&\quad -\frac{1}{2}\sum_{m=1}^{d}\sum_{i=1}^{N}\left\langle t_{im}^2\right\rangle - \frac{1}{2}\sum_{m=1}^{d}\langle\alpha_m\rangle\left\langle w_{zm}^2\right\rangle - \frac{1}{2}\sum_{m=1}^{d}\langle\alpha_m\rangle\left\langle w_{xm}^2\right\rangle \\
&\quad -\sum_{m=1}^{d}a_{\alpha m0}\log\hat{b}_{\alpha m} - \sum_{m=1}^{d}b_{\alpha m0}\langle\alpha_m\rangle \\
&\quad -\frac{1}{2}\log|\mathbf{\Sigma}_{z,t}^{-1}| - \sum_{m=1}^{d}\log\hat{b}_{\alpha m} + \frac{1}{2}\sum_{m=1}^{d}\log\sigma_{w_{zm}}^2 + \frac{1}{2}\sum_{m=1}^{d}\log\sigma_{w_{xm}}^2 + \text{Constant}
\end{aligned} \tag{17}$$

4 Inferring the Regression Coefficient

Once the EM algorithm has converged and we have the final values for the unknown variables and point-estimated parameters, we still need to infer what the regression coefficient if we want to make predictions.

Given a noiseless test input $\mathbf{t}^q$, we would like to predict what the corresponding noiseless output y^q is, and we can do this since $\langle y^q | \mathbf{t}^q \rangle = \hat{b}_{true}^T \mathbf{t}^q$, where:

$$\hat{b}_{true}^T = \frac{\psi_y \mathbf{1}^T \mathbf{C}^{-1}}{\psi_y - \mathbf{1}^T \mathbf{C}^{-1} \mathbf{1}} \mathbf{\Psi}_z^{-1} \langle \mathbf{W}_z \rangle^T \langle \mathbf{W}_x \rangle^{-1} \tag{18}$$

where $\mathbf{C}^{-1} = \left(\frac{\mathbf{1}\mathbf{1}^T}{\psi_y} + \mathbf{\Psi}_z^{-1} \right)^{-1} = \mathbf{\Psi}_z - \frac{\mathbf{\Psi}_z \mathbf{1}\mathbf{1}\mathbf{\Psi}_z}{\psi_y + \mathbf{1}\mathbf{\Psi}_z \mathbf{1}}$.