An Expectation-Maximization algorithm for Learning the Latent Gaussian Model with Gaussian Likelihood

Mohammad Emtiyaz Khan
CS, UBC

April 22, 2011

Abstract

In this note, we derive an expectation-maximization (EM) algorithm for a latent Gaussian model with Gaussian likelihood. This model contains many popular models as a special case, such as factor analysis and linear regression. Our derived EM algorithm is general, and contains almost all the updates required for the special cases. We also describe modification of the algorithm in the presence of missing variables and with regularization priors.

1 Latent Gaussian Model with Gaussian Likelihood

In this section, we introduce the model. We denote the visible data vectors by $\mathbf{y}_n$ and the latent vectors by $\mathbf{z}_n$. In general, $\mathbf{y}_n$ and $\mathbf{z}_n$ will have dimensions D and L respectively with $\mathbf{y}_n \in \mathbb{R}^D$ and $\mathbf{z}_n \in \mathbb{R}^L$. $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ denote the mean and covariance matrix of the latent Gaussian as seen in Eq. 1. Mean parameter of the observation is obtained by transforming $\mathbf{z}_n$ using a weight matrix $\mathbf{W} \in \mathbb{R}^{D\times L}$ and an offset vector $\mathbf{w}_0$, as shown in Eq. 2. The final likelihood is defined in Eq. 3, where $\boldsymbol{\Psi}$ is a diagonal matrix.

$$p(\mathbf{z}_n) = \mathcal{N}(\mathbf{z}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{1}$$

$$\boldsymbol{\eta}_n = \mathbf{W}\mathbf{z}_n + \mathbf{w}_0 \tag{2}$$

$$p(\mathbf{y}_n|\boldsymbol{\eta}_n, \boldsymbol{\Psi}) = \mathcal{N}(\mathbf{y}_n|\boldsymbol{\eta}_n, \boldsymbol{\Psi}) \tag{3}$$

The parameter set for the latent variables is $\boldsymbol{\theta}_z = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$, while for the likelihood is $\boldsymbol{\theta}_y = \{\mathbf{W}, \mathbf{w}_0, \boldsymbol{\Psi}\}$. We denote the complete parameter set by $\boldsymbol{\theta} = \{\boldsymbol{\theta}_z, \boldsymbol{\theta}_y\}$.

A variety of models can be written as a special case of this model. For example, if $\boldsymbol{\mu} = 0$ and $\boldsymbol{\Sigma} = \mathbf{I}_L$, then we get a factor analysis model. When $n = 1$ and $\mathbf{W}$ is an observed feature matrix, we get a linear regression model with $\mathbf{z}$ as the weight vector. Learning algorithms for these special cases can be found in the literature, for example see [TB99], [GH97] for PCA and factor analysis, [Bis06] for regression models. In this note, we describe a general expectation-maximization (EM) algorithm for learning the parameters of this model. We derive updates for posterior distribution and all the parameters. These updates include other learning algorithm as special cases (for example, EM algorithm for a factor analysis model is obtained by only updating $\boldsymbol{\theta}_y$). In what follows, we denote d'th row of $\mathbf{W}$ by $\mathbf{w}_d$, and d'th element of a vector $\mathbf{x}$ by x_d.

2 An EM algorithm

We begin with the marginal log-likelihood $\mathcal{L}(\boldsymbol{\theta})$ given in Equation 4 and introduce a variational posterior distribution $q_n(\mathbf{z}|\boldsymbol{\gamma}_n)$ for each data case. We use a full covariance Gaussian posterior with mean $\mathbf{m}_n$ and covariance $\mathbf{V}_n$. The full set of variational parameters are thus $\boldsymbol{\gamma}_n = [\mathbf{m}_n, \mathbf{V}_n]$. We apply Jensen's inequality to obtain a lower bound $\mathcal{L}_J(\boldsymbol{\theta}, \boldsymbol{\gamma})$, as shown in Equation 5. The second and third terms in $\mathcal{L}_J(\boldsymbol{\theta}, \boldsymbol{\gamma})$ are easily seen to be the negative of the Kullback-Leibler divergence from the variational Gaussian posterior $q_n(\mathbf{z}|\mathbf{m}_n, \mathbf{V}_n)$ to the Gaussian prior distribution

$p(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, which has a well-known closed-form expression, as seen in Equation 7. Note that the Jensen's bound is tight when $q(\mathbf{z}|\boldsymbol{\gamma}_n) = p(\mathbf{z}|\mathbf{y}_n, \boldsymbol{\theta})$.

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{n=1}^{N} \log \int p(\mathbf{y}_n|\mathbf{z}, \boldsymbol{\theta}_y) p(\mathbf{z}|\boldsymbol{\theta}_z) d\mathbf{z} = \frac{1}{N}\sum_{n=1}^{N} \log \int \frac{q_n(\mathbf{z}|\boldsymbol{\gamma}_n)}{q_n(\mathbf{z}|\boldsymbol{\gamma}_n)} p(\mathbf{y}_n|\mathbf{z}, \boldsymbol{\theta}_y) p(\mathbf{z}|\boldsymbol{\theta}_z) d\mathbf{z} \tag{4}$$

$$\mathcal{L}_J(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}_{q_n(\mathbf{z}|\boldsymbol{\gamma}_n)} \left[\log p(\mathbf{y}_n|\mathbf{z}, \boldsymbol{\theta}_y)\right] + \mathbb{E}_{q_n(\mathbf{z}|\boldsymbol{\gamma}_n)} \left[\log p(\mathbf{z}|\boldsymbol{\theta}_z)\right] - \mathbb{E}_{q_n(\mathbf{z}|\boldsymbol{\gamma})} \left[\log q_n(\mathbf{z}|\boldsymbol{\gamma}_n)\right] \tag{5}$$

$$= \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}_{q_n(\mathbf{z}|\boldsymbol{\gamma}_n)} \left[\log p(\mathbf{y}_n|\mathbf{z}, \boldsymbol{\theta}_y)\right] - D_{KL}(q_n(\mathbf{z}|\boldsymbol{\gamma}_n)||p(\mathbf{z}|\boldsymbol{\theta}_z)) \tag{6}$$

$$D_{KL}(q_n(\mathbf{z}|\boldsymbol{\gamma}_n)||p(\mathbf{z}|\boldsymbol{\theta}_z)) = \frac{1}{2}\left(\log|\boldsymbol{\Sigma}| - \log|\mathbf{V}_n| + \text{tr}(\mathbf{V}_n\boldsymbol{\Sigma}^{-1}) + (\mathbf{m}_n - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{m}_n - \boldsymbol{\mu}) - L\right) \tag{7}$$

We now expand the likelihood term in $\mathcal{L}_J(\boldsymbol{\theta}, \boldsymbol{\gamma})$.

$$\mathbb{E}_{q_n(\mathbf{z}|\boldsymbol{\gamma}_n)} \left[\log p(\mathbf{y}_n|\mathbf{z}, \boldsymbol{\theta})\right] = \mathbb{E}_{q_n(\mathbf{z}|\boldsymbol{\gamma}_n)} \left[-\frac{1}{2}(\mathbf{y}_n - \mathbf{w}_0 - \mathbf{W}\mathbf{z})^T\boldsymbol{\Psi}^{-1}(\mathbf{y}_n - \mathbf{w}_0 - \mathbf{W}\mathbf{z}) - \frac{1}{2}\log|2\pi\boldsymbol{\Psi}|\right] \tag{8}$$

$$= -\frac{1}{2}[\mathbf{m}_n^T\mathbf{W}^T\boldsymbol{\Psi}^{-1}\mathbf{W}\mathbf{m}_n - 2(\mathbf{y}_n - \mathbf{w}_0)^T\boldsymbol{\Psi}^{-1}\mathbf{W}\mathbf{m}_n + (\mathbf{y}_n - \mathbf{w}_0)^T\boldsymbol{\Psi}^{-1}(\mathbf{y}_n - \mathbf{w}_0)$$
$$+ \text{trace}(\mathbf{W}^T\boldsymbol{\Psi}^{-1}\mathbf{W}\mathbf{V}_n) + \log|2\pi\boldsymbol{\Psi}|] \tag{9}$$

In E-step, we optimize $\mathcal{L}_J(\boldsymbol{\theta}, \boldsymbol{\gamma})$ with respect to $\mathbf{m}_n, \mathbf{V}_n$. Differentiating and setting the derivative to zero, gives us following expressions,

$$\mathbf{V}_n = \left(\mathbf{W}^T\boldsymbol{\Psi}^{-1}\mathbf{W} + \boldsymbol{\Sigma}^{-1}\right)^{-1} \tag{10}$$

$$\mathbf{m}_n = \mathbf{V}_n[\mathbf{W}^T\boldsymbol{\Psi}^{-1}(\mathbf{y}_n - \mathbf{w}_0) + \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}] \tag{11}$$

Note that $\mathbf{V}_n$ is same for n, hence we will refer to it by $\mathbf{V}$.

In M-step, we optimize with respect to $\boldsymbol{\theta}$. We first describe updates for $\boldsymbol{\theta}_z$. Only the second term in $\mathcal{L}_J(\boldsymbol{\theta}, \boldsymbol{\gamma})$ depends on $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. We first differentiate this term with respect to $\boldsymbol{\mu}$ and set to zero to obtain an update for $\boldsymbol{\mu}$.

$$\frac{\partial\mathcal{L}_J(\boldsymbol{\theta}, \boldsymbol{\gamma})}{\partial\boldsymbol{\mu}} = -\frac{1}{2N}\sum_{n=1}^{N} 2\boldsymbol{\Sigma}^{-1}(\mathbf{m}_n - \boldsymbol{\mu}) \tag{12}$$

$$\boldsymbol{\mu}^* = \frac{1}{N}\sum_{n=1}^{N} \mathbf{m}_n \tag{13}$$

Similarly, we take derivative with respect to $\boldsymbol{\Sigma}^{-1}$ to obtain the gradient in Eq. 14. We substitute the updated value of $\boldsymbol{\mu}^*$ in the gradient and set it to zero. Simplification in Eq. 15-19 gives us the update of $\boldsymbol{\Sigma}$ in Eq. 19. Note that we could

have used the value of $\boldsymbol{\mu}$ from the previous iteration but using the updated value usually leads to faster convergence.

$$\frac{\partial \mathcal{L}_J(\boldsymbol{\theta}, \boldsymbol{\gamma})}{\partial \boldsymbol{\Sigma}^{-1}} = -\frac{1}{2N} \sum_{n=1}^{N} \left[-\boldsymbol{\Sigma} + \mathbf{V}_n + (\mathbf{m}_n - \boldsymbol{\mu})(\mathbf{m}_n - \boldsymbol{\mu})^T \right] \tag{14}$$

$$\boldsymbol{\Sigma}^* = \mathbf{V} + \frac{1}{N} \sum_{n=1}^{N} (\mathbf{m}_n - \boldsymbol{\mu}^*)(\mathbf{m}_n - \boldsymbol{\mu}^*)^T \tag{15}$$

$$= \mathbf{V} + \frac{1}{N} \sum_{n=1}^{N} (\mathbf{m}_n \mathbf{m}_n^T - 2\mathbf{m}_n \boldsymbol{\mu}^{*T} + \boldsymbol{\mu}^* \boldsymbol{\mu}^{*T}) \tag{16}$$

$$= \mathbf{V} + \boldsymbol{\mu}^* \boldsymbol{\mu}^{*T} - 2\boldsymbol{\mu}^* \frac{1}{N} \sum_{n=1}^{N} \mathbf{m}_n^T + \frac{1}{N} \sum_{n=1}^{N} \mathbf{m}_n \mathbf{m}_n^T \tag{17}$$

$$= \mathbf{V} + \boldsymbol{\mu}^* \boldsymbol{\mu}^{*T} - 2\boldsymbol{\mu}^* \boldsymbol{\mu}^{*T} + \frac{1}{N} \sum_{n=1}^{N} \mathbf{m}_n \mathbf{m}_n^T \tag{18}$$

$$= \mathbf{V} - \boldsymbol{\mu}^* \boldsymbol{\mu}^{*T} + \frac{1}{N} \sum_{n=1}^{N} \mathbf{m}_n \mathbf{m}_n^T \tag{19}$$

To obtain updates for $\{\boldsymbol{\theta}_y\}$, note that these parameters appear only in the likelihood term (see Eq. 5). We first describe the update for $\mathbf{W}$. We collect all the terms depending on $\mathbf{W}$ from $\mathcal{L}_J(\boldsymbol{\theta}, \boldsymbol{\gamma})$; this is shown in Eq. 20. We simplify this to get Eq. 21. We obtain the gradient with respect to $\mathbf{W}$ in Eq. 22 using the fact that the derivatives of trace$(\mathbf{X}^T\mathbf{B}\mathbf{X}\mathbf{C})$ and trace$(\mathbf{B}\mathbf{X}\mathbf{A})$ with respect to $\mathbf{X}$ are $2\mathbf{B}\mathbf{X}\mathbf{C}$ and $\mathbf{B}^T\mathbf{A}^T$, when $\mathbf{B}$ and $\mathbf{C}$ are symmetric. We set the gradient to 0 and obtain the update of $\mathbf{W}$ in Eq. 23.

$$\mathcal{L}_J'(\mathbf{W}) = -\frac{1}{2N} \sum_{n=1}^{N} \left[\mathbf{m}_n^T \mathbf{W}^T \boldsymbol{\Psi}^{-1} \mathbf{W} \mathbf{m}_n - 2(\mathbf{y}_n - \mathbf{w}_0)^T \boldsymbol{\Psi}^{-1} \mathbf{W} \mathbf{m}_n + \text{trace}(\mathbf{W}^T \boldsymbol{\Psi}^{-1} \mathbf{W} \mathbf{V}) \right] \tag{20}$$

$$= -\frac{1}{2N} \text{trace} \left[\mathbf{W}^T \boldsymbol{\Psi}^{-1} \mathbf{W} \left(N\mathbf{V} + \sum_{n=1}^{N} \mathbf{m}_n \mathbf{m}_n^T \right) - 2\boldsymbol{\Psi}^{-1} \mathbf{W} \sum_{n=1}^{N} \mathbf{m}_n (\mathbf{y}_n - \mathbf{w}_0)^T \right] \tag{21}$$

$$\frac{\partial \mathcal{L}_J'(\mathbf{W})}{\partial \mathbf{W}} = -\frac{1}{2N} \left[2\boldsymbol{\Psi}^{-1} \mathbf{W} \left(N\mathbf{V} + \sum_{n=1}^{N} \mathbf{m}_n \mathbf{m}_n^T \right) - 2\boldsymbol{\Psi}^{-1} \sum_{n=1}^{N} (\mathbf{y}_n - \mathbf{w}_0) \mathbf{m}_n^T \right] \tag{22}$$

$$\mathbf{W}^* = \left[\sum_{n=1}^{N} (\mathbf{y}_n - \mathbf{w}_0) \mathbf{m}_n^T \right] \left[N\mathbf{V} + \sum_{n=1} \mathbf{m}_n \mathbf{m}_n^T \right]^{-1} \tag{23}$$

We repeat this procedure for $\mathbf{w}_0$ but substitute the updated value of $\mathbf{W}$.

$$\frac{\partial \mathcal{L}_J(\boldsymbol{\theta}, \boldsymbol{\gamma})}{\partial \mathbf{w}_0} = -\frac{1}{2N} \sum_{n=1}^{N} \left[2\boldsymbol{\Psi}^{-1} \mathbf{W} \mathbf{m}_n + 2\boldsymbol{\Psi}^{-1} (\mathbf{w}_0 - \mathbf{y}_n) \right] \tag{24}$$

$$\mathbf{w}_0^* = \frac{1}{N} \sum_{n=1}^{N} (\mathbf{y}_n - \mathbf{W}^* \mathbf{m}_n) \tag{25}$$

We obtain the update for $\boldsymbol{\Psi}$ in a similar fashion. The objective function and gradient for Ψ_{dd} are shown in Eq. 26 and 27. We set the gradient to zero to get an update in Eq. 28. We expand the second term in Eq. 29 and Eq. 30, and use Eq. 25 to substitute the value of w_{id}^* which leads to a simplification in Eq. 31. We simplify further in Eq. 32 and Eq.

33, and use the update for $\mathbf{w}_d^*$ from Eq. 23 to get Eq. 34. We simplify further to get the final update in Eq. 36.

$$\mathcal{L}_J^{'}(\mathbf{\Psi}) = -\frac{1}{2N}\sum_{n=1}^{N}\left[(\mathbf{y}_n - \mathbf{W}\mathbf{m}_n - \mathbf{w}_0)^T\mathbf{\Psi}^{-1}(\mathbf{y}_n - \mathbf{W}\mathbf{m}_n - \mathbf{w}_0) + \text{trace}(\mathbf{W}^T\mathbf{\Psi}^{-1}\mathbf{W}\mathbf{V}) + \log|2\pi\mathbf{\Psi}|\right] \tag{26}$$

$$\frac{\partial\mathcal{L}_J^{'}(\mathbf{\Psi})}{\partial\Psi_{dd}} = \frac{1}{2N\Psi_{dd}^2}\sum_{n=1}^{N}(y_{dn} - \mathbf{w}_d^T\mathbf{m}_n - \mathbf{w}_{0d})^2 + \frac{1}{2\Psi_{dd}^2}\mathbf{w}_d^T\mathbf{V}\mathbf{w}_d - \frac{1}{2\Psi_{dd}} \tag{27}$$

$$\Psi_{dd}^* = \mathbf{w}_d^{*T}\mathbf{V}\mathbf{w}_d^* + \frac{1}{N}\sum_{n=1}^{N}(y_{dn} - \mathbf{w}_d^{*T}\mathbf{m}_n - w_{0d}^*)^2 \tag{28}$$

$$= \mathbf{w}_d^{*T}\mathbf{V}\mathbf{w}_d^* + \frac{1}{N}\sum_{n=1}^{N}\left[(y_{dn} - \mathbf{w}_d^{*T}\mathbf{m}_n)^2 - 2(y_{dn} - \mathbf{w}_d^{*T}\mathbf{m}_n)w_{0d}^* + w_{0d}^{*2}\right] \tag{29}$$

$$= \mathbf{w}_d^{*T}\mathbf{V}\mathbf{w}_d^* + \frac{1}{N}\sum_{n=1}^{N}(y_{dn} - \mathbf{w}_d^{*T}\mathbf{m}_n)^2 - 2w_{0d}^*\frac{1}{N}\sum_{n=1}^{N}(y_{dn} - \mathbf{w}_d^{*T}\mathbf{m}_n) + w_{0d}^{*2} \tag{30}$$

$$= \mathbf{w}_d^{*T}\mathbf{V}\mathbf{w}_d^* + \frac{1}{N}\sum_{n=1}^{N}(y_{dn} - \mathbf{w}_d^{*T}\mathbf{m}_n)^2 - w_{0d}^{*2} \tag{31}$$

$$= \mathbf{w}_d^{*T}\mathbf{V}\mathbf{w}_d^* + \frac{1}{N}\sum_{n=1}^{N}(y_{dn}^2 - 2y_{dn}\mathbf{w}_d^{*T}\mathbf{m}_n + \mathbf{m}_n^T\mathbf{w}_d^*\mathbf{w}_d^{*T}\mathbf{m}_n)^2 - w_{0d}^{*2} \tag{32}$$

$$= \frac{1}{N}\sum_{n=1}^{N}(y_{dn}^2 - 2y_{dn}\mathbf{w}_d^{*T}\mathbf{m}_n) + \frac{1}{N}\text{trace}\left[\mathbf{w}_d^*\mathbf{w}_d^{*T}(N\mathbf{V} + \sum_{n=1}^{N}\mathbf{m}_n\mathbf{m}_n^T)\right] - w_{0d}^{*2} \tag{33}$$

$$= \frac{1}{N}\sum_{n=1}^{N}(y_{dn}^2 - 2y_{dn}\mathbf{w}_d^{*T}\mathbf{m}_n) + \frac{1}{N}\text{trace}\left[\mathbf{w}_d^*\sum_{n=1}^{N}(y_{dn} - w_{0d})\mathbf{m}_n^T\right] - w_{0d}^{*2} \tag{34}$$

$$= \frac{1}{N}\sum_{n=1}^{N}(y_{dn}^2 - 2y_{dn}\mathbf{w}_d^{*T}\mathbf{m}_n + (y_{dn} - w_{0d})\mathbf{w}_d^{*T}\mathbf{m}_n) - w_{0d}^{*2} \tag{35}$$

$$= \frac{1}{N}\sum_{n=1}^{N}(y_{dn}^2 - y_{dn}\mathbf{w}_d^{*T}\mathbf{m}_n - w_{0d}\mathbf{w}_d^{*T}\mathbf{m}_n - w_{0d}^{*2}) \tag{36}$$

2.1 Learning in Presence of the Missing Data

Let us denote the observed part of n'th data vector by $\mathbf{y}_{obs(n)}$ and the missing part by $\mathbf{y}_{miss(n)}$, so that $\mathbf{y}_n = [\mathbf{y}_{obs(n)}\mathbf{y}_{miss(n)}]$. Here, $obs(n)$ contains the indices of the dimensions that are observed in the n'th data vector. Since the likelihood factorizes given the latent variables, missing dimensions can simply be ignored. This is easy to see in the following. The marginal likelihood of the observed part of n'th data vector can be obtained by integrating over the missing part as in Eq. 37. We express the marginal likelihood of full data vector as integral over the latent variable in Eq. 38. Finally, we interchange the integrals and integrate the missing variable part to 1, which gives us the desired results in Eq. 39. Here $\mathbf{W}_{obs(n)}$ is obtained by selecting rows of $\mathbf{W}$ corresponding to observed dimensions. Similarly, $\mathbf{w}_{0,obs(n)}$ is a vector containing elements corresponding to observed dimensions, and $\mathbf{\Psi}_{obs(n)}$ is obtained

by selecting observed rows and columns of $\mathbf{\Psi}$.

$$p(\mathbf{y}_{obs(n)}|\boldsymbol{\theta}) = \int_{y_{miss(n)}} p(\mathbf{y}_{obs(n)}, \mathbf{y}_{miss(n)}|\boldsymbol{\theta}) d\mathbf{y}_{miss(n)} \tag{37}$$

$$= \int_{y_{miss(n)}} \int_{\mathbf{z}} p(\mathbf{y}_n|\mathbf{z}, \mathbf{W}, \mathbf{w}_0, \mathbf{\Psi}) p(\mathbf{z}|\boldsymbol{\theta}) d\mathbf{z} d\mathbf{y}_{miss(n)} \tag{38}$$

$$= \int_{\mathbf{z}} p(\mathbf{y}_{obs(n)}|\mathbf{z}, \mathbf{W}_{obs(n)}, \mathbf{w}_{0,obs(n)}, \mathbf{\Psi}_{obs(n)}) p(\mathbf{z}|\boldsymbol{\theta}) d\mathbf{z} \tag{39}$$

Using this fact, it is easy to show that the posterior mean and covariance are given as follows,

$$\mathbf{V}_n^{-1} = \mathbf{W}_{obs(n)}^T \mathbf{\Psi}_{obs(n)}^{-1} \mathbf{W}_{obs(n)} + \mathbf{\Sigma}^{-1} \tag{40}$$

$$\mathbf{m}_n = \mathbf{V} \left[\mathbf{W}_{obs(n)}^T \mathbf{\Psi}_{obs(n)}^{-1} (\mathbf{y}_{obs(n)} - \mathbf{w}_{0,obs(n)}) + \mathbf{\Sigma}^{-1} \boldsymbol{\mu} \right] \tag{41}$$

Note that the posterior covariance depends on n now, and hence needs to be recomputed for every data point (and we need to invert it for every data point). This increases the computational overhead. The M-step updates are different for $\boldsymbol{\theta}_y$ in that only observed dimensions contribute to the estimates as in Eq. 42-44. Here, $obs(d)$ are the data indices where the d'th dimension is observed and $N_{obs(d)}$ is the total number of those indices.

$$\mathbf{w}_d^{*T} = \left[\sum_{n \in obs(d)} (y_{dn} - w_{od}) \mathbf{m}_n^T \right] \left[\sum_{n \in obs(d)} \left(\mathbf{V}_n + \mathbf{m}_n \mathbf{m}_n^T \right) \right]^{-1} \tag{42}$$

$$\mathbf{w}_0^* = \frac{1}{N_{obs(d)}} \sum_{n \in obs(d)} (\mathbf{y}_n - \mathbf{W}^* \mathbf{m}_n) \tag{43}$$

$$\Psi_{dd}^* = \frac{1}{N_{obs(d)}} \sum_{n \in obs(d)} (y_{dn}^2 - y_{dn} \mathbf{w}_d^{*T} \mathbf{m}_n - w_{0d} \mathbf{w}_d^{*T} \mathbf{m}_n - w_{0d}^{*2}) \tag{44}$$

2.2 Parameter Regularization

When the data size is small, we might need to regularize the parameters. We now describe regularization prior and the corresponding parameter updates. We assume a normal-(inverse)-Wishart prior for $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ as defined in Eq. 45. Here, $\lambda_\mu > 0$, $\mathbf{S}_0$ is a positive-definite matrix, and $\nu_0 > -L - 1$ (mode of inverse-Wishart distribution is well-defined for these values). Setting $\lambda_\mu = 0$, $\mathbf{S}_0 = 0$ and $\nu_0 = -L - 1$ is equivalent to no prior. For $\boldsymbol{\theta}_y$, we use normal-(inverse)-gamma prior. For $\mathbf{W}$, we assume a normal distribution defined in Eq. 46 with $\lambda_{wd} > 0$ (note that it is conditioned on Ψ_{dd}). Similarly, we assume a normal prior for each element of $\mathbf{w}_0$ as shown in Eq. 47 with $\lambda_0 > 0$. Finally, we have an inverse-gamma prior on the elements of $\mathbf{\Psi}$ as in Eq. 48 with $a > -1$ and $b > 0$. Setting $\lambda_w = 0$, $\lambda_0 = 1$, $a = -1$, and $b = 0$ is equivalent to no prior.

$$p(\boldsymbol{\mu}|\lambda_\mu, \mathbf{\Sigma}) p(\mathbf{\Sigma}|\mathbf{S}_0, \boldsymbol{\nu}_0) \propto \mathcal{N}(\boldsymbol{\mu}|0, \lambda_\mu^{-1} \mathbf{\Sigma}) |\mathbf{\Sigma}|^{-(L+1+\nu_0)/2} \exp\left[-\text{trace}\left(\mathbf{S}_0 \mathbf{\Sigma}^{-1} \right) / 2 \right] \tag{45}$$

$$p(\mathbf{W}|\lambda_w, \Psi_{dd}) = \prod_{d=1}^{D} \prod_{l=1}^{L} \mathcal{N}(w_{dl}|0, \lambda_{wd}^{-1} \Psi_{dd}) \tag{46}$$

$$p(\mathbf{w}_0|\lambda_0, \mathbf{\Psi}) = \mathcal{N}(\mathbf{w}_0|0, \lambda_0^{-1} \mathbf{\Psi}) \tag{47}$$

$$p(\Psi_{dd}|a, b) \propto \Psi_{dd}^{-(a+1)/2} \exp\left(-b \Psi_{dd}^{-1} / 2 \right) \tag{48}$$

We maximize the penalized lower bound $N\mathcal{L}_J(\boldsymbol{\theta}, \boldsymbol{\gamma}) + \log p(\boldsymbol{\theta})$, where the second term is the prior distribution defined as above. It is easy to show that we find the updates shown below in Eq. 49-53. In the update for $\boldsymbol{\mu}$ in Eq. 49, λ_μ acts as prior sample-size. Similarly, for the update for $\mathbf{\Sigma}$ in Eq. 50, $\nu_0 + L + 1$ is the prior sample-size with an additional 1 contributed from the prior for $\boldsymbol{\mu}$ (if λ_μ =0, this additional term should be removed). Also, $\mathbf{S}_0$ is the prior information

about the sufficient statistics of the latent vectors. We can have similar interpretation of the other parameter updates. Specially, in Eq. 53, $a+1$ is the prior sample-size, and the term $L+1$ is the contribution from the prior for $\mathbf{w}_d$ and $\mathbf{w}_0$. If $\lambda_w = 0$ then L should be removed, and if $\lambda_0 = 0$ then 1 should be removed.

$$\boldsymbol{\mu}^* = \frac{1}{N+\lambda_\mu} \sum_{n=1}^{N} \mathbf{m}_n \tag{49}$$

$$\boldsymbol{\Sigma}^* = \frac{\mathbf{S}_0 + N\mathbf{V} + \sum_{n=1}^{N} \mathbf{m}_n \mathbf{m}_n^T + N\boldsymbol{\mu}^* \boldsymbol{\mu}^{*T}}{N + \nu_0 + L + 1 + 1} \tag{50}$$

$$\mathbf{W}^* = \left[\sum_{n=1}^{N} \mathbf{m}_n (\mathbf{y}_n - \mathbf{w}_0)^T \right] \left[\lambda_w \mathbf{I}_L + N\mathbf{V} + \sum_{n=1} \mathbf{m}_n \mathbf{m}_n^T \right]^{-1} \tag{51}$$

$$\mathbf{w}_0^* = \frac{1}{N+\lambda_0} \sum_{n=1}^{N} (\mathbf{y}_n - \mathbf{W}^* \mathbf{m}_n) \tag{52}$$

$$\Psi_{dd}^* = \frac{1}{N+a+1+L+1} \left[b + \sum_{n=1}^{N} (y_{dn}^2 - y_{dn} \mathbf{w}_d^{*T} \mathbf{m}_n - w_{0d} \mathbf{w}_d^{*T} \mathbf{m}_n - w_{0d}^{*2}) \right] \tag{53}$$

References

[Bis06] C. Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[GH97] Z. Ghahramani and G.E. Hinton. The EM algorithm for mixtures of factor analyzers. *University of Toronto Technical Report CRG-TR-96-1*, 1997.

[TB99] M.E. Tipping and C.M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 61(3):611–622, 1999.