Appendix A.1: Inference for MT-iLSVM

In this section, we provide the deviation of the inference algorithm for MT-iLSVM, which is outlined in Alg. 1 and detailed below.

For MT-iLSVM, the model $\mathcal{M}$ consists of all the latent variables $(\boldsymbol{\nu}, \mathbf{W}, \mathbf{Z}, \boldsymbol{\eta})$. Let $L_{mn}(p) \stackrel{\text{def}}{=} \mathbb{E}_p[\log p(\mathbf{x}_{mn}|\mathbf{Z}, \mathbf{w}_{mn}, \lambda_{mn}^2)]$ be the expected data likelihood. Then, under the truncated mean-field assumption (14), we have

$$L_{mn}(p) = -\frac{\mathbf{x}_{mn}^\top \mathbf{x}_{mn} - 2\mathbf{x}_{mn}^\top \mathbb{E}_p[\mathbf{Z}\mathbf{w}_{mn}] + \mathbb{E}_p[\mathbf{w}_{mn}^\top \mathbf{U} \mathbf{w}_{mn}]}{2\lambda_{mn}^2} - \frac{D\log(2\pi\lambda_{mn}^2)}{2},$$

where $\mathbf{x}_{mn}^\top \mathbb{E}_p[\mathbf{Z}\mathbf{w}_{mn}] = \sum_k \mathbf{x}_{mn}^\top \boldsymbol{\psi}_{.k}$; $\boldsymbol{\psi}_{.k} \stackrel{\text{def}}{=} (\psi_{1k} \cdots \psi_{Dk})^\top$ is the kth column of $\boldsymbol{\psi} = \mathbb{E}[\mathbf{Z}]$;

$$\mathbb{E}_p[\mathbf{w}_{mn}^\top \mathbf{U} \mathbf{w}_{mn}] = 2\sum_{j<k} \phi_{mn}^j \phi_{mn}^k \mathbf{U}_{jk} + \sum_k \mathbf{U}_{kk}(K\sigma_{mn}^2 + \Phi_{mn}^\top \Phi_{mn});$$

and $\mathbf{U} \stackrel{\text{def}}{=} \mathbb{E}[\mathbf{Z}^\top \mathbf{Z}]$ is a $K \times K$ matrix, whose element is

$$\mathbf{U}_{ij} = \begin{cases} \sum_d \psi_{di}, & \text{if } i = j \\ \sum_d \psi_{di}\psi_{dj}, & \text{otherwise.} \end{cases}$$

For the KL-divergence term, we have $\mathrm{KL}(p(\mathcal{M})\|\pi(\mathcal{M})) = \mathrm{KL}(p(\boldsymbol{\nu})\|\pi(\boldsymbol{\nu})) + \mathrm{KL}(p(\mathbf{W})\|\pi(\mathbf{W})) + \mathrm{KL}(p(\mathbf{Z})\|\pi(\mathbf{Z})) + \mathrm{KL}(p(\boldsymbol{\eta})\|\pi(\boldsymbol{\eta}))$, where the individual terms are

$$\begin{aligned}
\mathrm{KL}(p(\boldsymbol{\nu})\|\pi(\boldsymbol{\nu})) = & \sum_{k=1}^K \Big((\gamma_{k1} - \alpha)(\psi(\gamma_{k1}) - \psi(\gamma_{k1} + \gamma_{k2})) + (\gamma_{k2} - 1)(\psi(\gamma_{k2}) - \psi(\gamma_{k1} + \gamma_{k2})) \\
& - \log \frac{\Gamma(\gamma_{k1})\Gamma(\gamma_{k2})}{\Gamma(\gamma_{k1} + \gamma_{k2})} \Big) - K\log\alpha, \\
\mathrm{KL}(p(\mathbf{Z})\|\pi(\mathbf{Z})) = & \sum_{dk} \Big(-\psi_{dk} \sum_{j=1}^k \mathbb{E}_p[\log \nu_j] - (1 - \psi_{dk})\mathbb{E}_p[\log(1 - \prod_{j=1}^k \nu_j)] \\
& + \psi_{dk} \log \psi_{dk} + (1 - \psi_{dk}) \log(1 - \psi_{dk}) \Big) \\
\mathrm{KL}(p(\mathbf{W})\|\pi(\mathbf{W})) = & \sum_{mn} \Big(\frac{K\sigma_{mn}^2 + \Phi_{mn}^\top \Phi_{mn}}{2\sigma_{m0}^2} - \frac{K(1 + \log \frac{\sigma_{mn}^2}{\sigma_{m0}^2})}{2} \Big).
\end{aligned}$$

where $\psi(\cdot)$ is the digamma function and $\mathbb{E}_p[\log v_j] = \psi(\gamma_{j1}) - \psi(\gamma_{j1} + \gamma_{j2})$. For $\mathrm{KL}(p(\boldsymbol{\eta})\|\pi(\boldsymbol{\eta}))$, we do not need to write it explicitly, as we shall see. Finally, the effective discriminant function is

$$f_m(\mathbf{x}_{mn}; p(\mathbf{Z}, \boldsymbol{\eta})) = \boldsymbol{\eta}_m^\top \boldsymbol{\psi}^\top \mathbf{x}_{mn} = \sum_{k=1}^K \mathbb{E}_p[\eta_{mk}] \boldsymbol{\psi}_{.k}^\top \mathbf{x}_{mn}.$$

All the above terms can be easily computed, except the term $\mathbb{E}_p[\log(1 - \prod_{j=1}^k \nu_j)]$. Here, we adopt the multivariate lower bound [9]

$$\mathbb{E}_p[\log(1 - \prod_{j=1}^k \nu_j)] \geq \sum_{m=1}^k q_{km}\psi(\gamma_{m2}) + \sum_{m=1}^{k-1} (\sum_{n=m+1}^k q_{kn})\psi(\gamma_{m1}) - \sum_{m=1}^k (\sum_{n=m}^k q_{kn})\psi(\gamma_{m1} + \gamma_{m2}) + \mathcal{H}(q_{k.}),$$

where the variational parameters $q_{k.} = (q_{k1} \cdots q_{kk})^\top$ belong to the k-simplex, and $\mathcal{H}(q_{k.})$ is the entropy of $q_{k.}$. The tightest lower bound is achieved by setting $q_{k.}$ to be the optimum value

$$q_{km} = \frac{1}{Z_k} \exp\Big(\psi(\gamma_{m2}) + \sum_{n=1}^{m-1} \psi(\gamma_{n1}) - \sum_{n=1}^m \psi(\gamma_{n1} + \gamma_{n2}) \Big), \tag{17}$$

where Z_k is a normalization factor to make $q_{k.}$ be a distribution. We denote the tightest lower bound by $\mathcal{L}_k^\nu$. Replacing the term $\mathbb{E}_p[\log(1 - \prod_{j=1}^k \nu_j)]$ with its lower bound $\mathcal{L}_k^\nu$, we can have an upper bound of $\mathrm{KL}(p(\mathcal{M})\|\pi(\mathcal{M}))$ and we denote this upper bound by $\mathcal{L}(p)$.

Algorithm 1 Inference Algorithm of MT-iLSVM

Input: data $\mathcal{D} = \{(\mathbf{x}_{mn}, y_{mn})\}_{m,n\in\mathcal{I}^m_{\text{tr}}} \cup \{\mathbf{x}_{mn}\}_{m,n\in\mathcal{I}^m_{\text{tst}}}$, constants α and C
Output: distributions $p(\boldsymbol{\nu})$, $p(\mathbf{Z})$, $p(\mathbf{W})$, $p(\boldsymbol{\eta})$ and hyper-parameters σ^2_{m0} and λ^2_{mn}
Initialize $\gamma_{k1} = \alpha$, $\gamma_{k2} = 1$, $\psi_{dk} = 0.5 + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 0.001)$, $\Phi_{mn} = 0$, $\sigma^2_{mn} = \sigma^2_{m0} = 1$, $\boldsymbol{\mu}_m = 0$, λ^2_{mn} is computed from $\mathcal{D}$.
repeat
  repeat
    update $(\gamma_{k1}, \gamma_{k2})$ using Eq. (19), $\forall 1 \leq k \leq K$;
    update ϕ^k_{mn} and σ^2_{mn} using Eq. (18), $\forall m, \forall n, \forall 1 \leq k \leq K$;
    update ψ_{dk} using Eq. (20), $\forall 1 \leq d \leq D, \forall 1 \leq k \leq K$;
  until relative change of L is less than τ (e.g., $1e^{-3}$) or iteration number is T (e.g., 10)
  for $m = 1$ **to** M **do**
    solve the dual problem (21) using a binary SVM learner.
  end for
  update the hyper-parameters σ^2_{m0} using Eq. (22) and λ^2_{mn} using Eq. (23). (*Optional*)
until relative change of L is less than τ' (e.g., $1e^{-4}$) or iteration number is T' (e.g., 20)

With the above terms and the upper bound $\mathcal{L}(p)$, we use the Lagrangian method with the Lagrangian multipliers $\boldsymbol{\omega}$, one for each margin constraint, and $\mathbf{u}$ for the nonnegativity constraint of $\boldsymbol{\xi}$. We have the Lagrangian functional

$$L(p, \boldsymbol{\xi}, \boldsymbol{\omega}, \mathbf{u}) = \mathcal{L}(p) - \sum_{mn} L_{mn}(p) - \sum_{m,n\in\mathcal{I}^m_{\text{tr}}} \omega_{mn}\Big(y_{mn}(\mathbb{E}_p[\boldsymbol{\eta}_m]^\top \boldsymbol{\psi}^\top \mathbf{x}_{mn}) - 1 + \xi_{mn}\Big) - \mathbf{u}^\top \boldsymbol{\xi}.$$

Then, the inference procedure iteratively solves the following steps:

Infer $p(\boldsymbol{\nu})$, $p(\mathbf{Z})$ and $p(\mathbf{W})$: For $p(\mathbf{W})$, since both the prior $\pi(\mathbf{W})$ and $p(\mathbf{W})$ are Gaussian, we can easily derive the update rules, similar as in Gaussian mixture models

$$\begin{aligned} \phi^k_{mn} &= \frac{\sum_d x^d_{mn}\psi_{dk} - \sum_{j\neq k} \phi^j_{mn}\mathbf{U}_{kj}}{\lambda^2_{mn}} \Big(\frac{1}{\sigma^2_{m0}} + \frac{\sum_d \psi_{dk}}{\lambda^2_{mn}}\Big)^{-1} \\ \sigma^2_{mn} &= \Big(\frac{1}{\sigma^2_{m0}} + \frac{1}{K}\sum_k \frac{\mathbf{U}_{kk}}{\lambda^2_{mn}}\Big)^{-1} \end{aligned} \tag{18}$$

For $p(\boldsymbol{\nu})$, we have the update rules similar as in [9], that is,

$$\begin{aligned} \gamma_{k1} &= \alpha + \sum_{m=k}^{K}\sum_{d=1}^{D}\psi_{dm} + \sum_{m=k+1}^{K}(D - \sum_{d=1}^{D}\psi_{dm})(\sum_{i=k+1}^{m} q_{mi}) \\ \gamma_{k2} &= 1 + \sum_{m=k}^{K}(D - \sum_{d=1}^{D}\psi_{dm})q_{mk}. \end{aligned} \tag{19}$$

For $p(\mathbf{Z})$, we have the mean-field update equation as

$$\psi_{dk} = \frac{1}{1 + e^{-\vartheta_{dk}}}, \tag{20}$$

where

$$\begin{aligned} \vartheta_{dk} = \sum_{j=1}^{k}\mathbb{E}_p[\log v_j] - \mathcal{L}^\nu_k - \sum_{mn}\frac{1}{2\lambda^2_{mn}}\Big((K\sigma^2_{mn} + (\phi^k_{mn})^2) \\ -2x^d_{mn}\phi^k_{mn} + 2\sum_{j\neq k}\phi^j_{mn}\phi^k_{mn}\psi_{dj}\Big) + \sum_{m,n\in\mathcal{I}^m_{\text{tr}}} y_{mn}\mathbb{E}_p[\eta_{mk}]x^d_{mn}. \end{aligned}$$

Infer $p(\boldsymbol{\eta})$ and solve for $\boldsymbol{\omega}$ and $\boldsymbol{\xi}$: We can optimize L to solve for $q(\boldsymbol{\eta})$, which is

$$p(\boldsymbol{\eta}) \propto \pi(\boldsymbol{\eta})\exp\Big\{\sum_{m,n\in\mathcal{I}^m_{\text{tr}}} y_{mn}\omega_{mn}\boldsymbol{\eta}_m^\top\boldsymbol{\psi}^\top\mathbf{x}_{mn}\Big\} = \prod_{m=1}^{M}\pi(\boldsymbol{\eta}_m)\exp\Big\{\boldsymbol{\eta}_m^\top\big(\sum_{n\in\mathcal{I}^m_{\text{tr}}} y_{mn}\omega_{mn}\boldsymbol{\psi}^\top\mathbf{x}_{mn}\big)\Big\}.$$

Therefore, we can see that although we did not assume $p(\boldsymbol{\eta})$ is factorized, we can get the induced factorization form $p(\boldsymbol{\eta}) = \prod_m p(\boldsymbol{\eta}_m)$, where

$$p(\eta_m) \propto \pi(\boldsymbol{\eta}_m) \exp\Big\{\boldsymbol{\eta}_m^\top \big(\sum_{n \in \mathcal{I}_{\text{tr}}^m} y_{mn}\omega_{mn}\boldsymbol{\psi}^\top \mathbf{x}_{mn} \big) \Big\}.$$

Here, we assume $\pi(\boldsymbol{\eta}_m)$ is standard normal. Then, we have $p(\boldsymbol{\eta}_m) = \mathcal{N}(\boldsymbol{\eta}_m | \boldsymbol{\mu}_m, I)$, where

$$\boldsymbol{\mu}_m = \sum_{n \in \mathcal{I}_{\text{tr}}^m} y_{mn}\omega_{mn}\boldsymbol{\psi}^\top \mathbf{x}_{mn}.$$

Substituting the solution of $p(\boldsymbol{\eta})$ into the Lagrangian functional, we get the M independent dual problems

$$\max_{\boldsymbol{\omega}_m} \; -\frac{1}{2}\boldsymbol{\mu}_m^\top \boldsymbol{\mu}_m + \sum_{n \in \mathcal{I}_{\text{tr}}^m} \omega_{mn} \quad \text{s.t..: } 0 \leq \omega_{mn} \leq 1, \forall n \in \mathcal{I}_{\text{tr}}^m, \tag{21}$$

which (and its primal form) can be efficiently solved with a binary SVM solver, such as SVM-light.

As we have stated, the hyperparameters σ_0^2 and λ_{mn}^2 can be set a priori or estimated from the data. The empirical estimation can be easily done with closed form solutions. For MT-iLSVM, we have

$$\sigma_{m0}^2 = \frac{\sum_{n=1}^{N_m} (K\sigma_{mn}^2 + \Phi_{mn}^\top \Phi_{mn})}{KN_m} \tag{22}$$

$$\lambda_{mn}^2 = \frac{\mathbf{x}_{mn}^\top \mathbf{x}_{mn} - 2\mathbf{x}_{mn}^\top \mathbb{E}_p[\mathbf{Z}\mathbf{w}_{mn}] + \mathbb{E}_p[\mathbf{w}_{mn}^\top \mathbf{U}\mathbf{w}_{mn}]}{D}. \tag{23}$$

Appendix A.2: Inference for Infinite Latent SVM

In this section, we develop the inference algorithm for iLSVM based on the stick-breaking construction of the IBP prior. The algorithm is outlined in Alg. 2 and detailed below.

Similar as in the inference for MT-iLSVM, we make the additional constraint about the feasible distribution

$$p(\boldsymbol{\nu}, \mathbf{W}, \mathbf{Z}, \boldsymbol{\eta}) = p(\boldsymbol{\eta}) p(\mathbf{W}|\Phi, \Sigma) \prod_n \Big(\prod_{k=1}^K p(z_{nk}|\psi_{nk}) \Big) \prod_{k=1}^K p(\nu_k | \boldsymbol{\gamma}_k),$$

where K is the truncation level; $p(\mathbf{W}|\Phi, \Sigma) = \prod_k \mathcal{N}(\mathbf{W}_{\cdot k}|\Phi_{\cdot k}, \sigma_k^2 I)$; $p(z_{nk}|\phi_{nk}) = \text{Bernoulli}(\phi_{nk})$; and $p(\nu_k|\boldsymbol{\gamma}_k) = \text{Beta}(\gamma_{k1}, \gamma_{k2})$. Then, we solve the constrained problem using Lagrangian methods with Lagrangian multipliers being $\boldsymbol{\omega}$, one for each large-margin constraint, and $\mathbf{u}$ for the nonnegativity constraints of $\boldsymbol{\xi}$. Similarly, let $L_n(p) \stackrel{\text{def}}{=} \mathbb{E}_p[\log p(\mathbf{x}_n | \mathbf{z}_n, \mathbf{W})]$. We have

$$L_n(p) = -\frac{\mathbf{x}_n^\top \mathbf{x}_n - 2\mathbf{x}_n^\top \Phi \mathbb{E}_p[\mathbf{z}_n]^\top + \mathbb{E}_p[\mathbf{z}_n \mathbf{A} \mathbf{z}_n^\top]}{2\sigma_{n0}^2} - \frac{D \log(2\pi\sigma_{n0}^2)}{2}, \tag{24}$$

where $\mathbf{A} \stackrel{\text{def}}{=} \mathbb{E}_p[\mathbf{W}^\top \mathbf{W}]$ is a $K \times K$ matrix; $\mathbf{x}_n^\top \Phi \mathbb{E}_p[\mathbf{z}_n]^\top = 2\sum_k \psi_{nk}(\mathbf{x}_n^\top \Phi_{\cdot k})$; and

$$\mathbb{E}_p[\mathbf{z}_n \mathbf{A} \mathbf{z}_n^\top] = 2\sum_{j<k} \psi_{nj}\psi_{nk}\mathbf{A}_{jk} + \sum_k \psi_{nk}(D\sigma_k^2 + \mathbf{A}_{kk}).$$

The effective discriminant function is $f(y, \mathbf{x}_n) = \sum_k \mathbb{E}_p[\eta_y^k]\psi_{nk}$. Again, for computational tractability, we need the lower bound $\mathcal{L}_k^\nu$ of the term $\mathbb{E}_p[\log(1 - \prod_{j=1}^k v_j)]$. Using this lower bound, we can get an upper bound of the KL-divergence term, and we denote the Lagrangian functional by $L(p, \boldsymbol{\xi}, \boldsymbol{\omega}, \mathbf{u})$. Then, the inference procedure iteratively solves the following steps:

Infer $p(\boldsymbol{\nu})$, $p(\mathbf{Z})$ and $p(\mathbf{W})$: For $p(\mathbf{W})$, we have the update rules

$$\Phi_{\cdot k} = \sum_n \frac{\psi_{nk}}{\sigma_{n0}^2} \Big(\mathbf{x}_n - \sum_{j \neq k} \psi_{nj} \Phi_{\cdot j} \Big) \Big(1 + \sum_n \frac{\psi_{nk}}{\sigma_{n0}^2} \Big)^{-1} \tag{25}$$

$$\sigma_k^2 = \Big(1 + \sum_n \frac{\psi_{nk}}{\sigma_{n0}^2} \Big)^{-1}.$$

Algorithm 2 Inference Algorithm of iLSVM

Input: data $\mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{n \in \mathcal{I}_{\text{tr}}} \cup \{\mathbf{x}_n\}_{n \in \mathcal{I}_{\text{tst}}}$, constants α and C
Output: distributions $p(\boldsymbol{\nu})$, $p(\mathbf{Z})$, $p(\mathbf{W})$, $p(\boldsymbol{\eta})$ and hyper-parameters σ_0^2 and σ_{n0}^2
Initialize $\gamma_{k1} = \alpha$, $\gamma_{k2} = 1$, $\psi_{nk} = 0.5 + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 0.001)$, $\Phi_{\cdot k} = 0$, $\sigma_k^2 = \sigma_0^2 = 1$, $\boldsymbol{\mu} = 0$, σ_{n0}^2 is computed from $\mathcal{D}$.
repeat
  repeat
    update $(\gamma_{k1}, \gamma_{k2})$ using Eq. (26), $\forall 1 \leq k \leq K$;
    update $\Phi_{\cdot k}$ and σ_k^2 using Eq. (25), $\forall 1 \leq k \leq K$;
    update ψ_{nk} using Eq. (27), $\forall n \in \mathcal{I}_{\text{tr}}, \forall 1 \leq k \leq K$;
    update ψ_{nk} using Eq. (27), but ϑ_{nk} doesn't have the last term, $\forall n \in \mathcal{I}_{\text{tst}}, \forall 1 \leq k \leq K$;
  until relative change of L is less than τ (e.g., $1e^{-3}$) or iteration number is T (e.g., 10)
  solve the dual problem (28) (or its primal form) using a multi-class SVM learner.
  update the hyper-parameters σ_0^2 using Eq. (29) and σ_{n0}^2 using Eq. (30). (*Optional*)
until relative change of L is less than τ' (e.g., $1e^{-4}$) or iteration number is T' (e.g., 20)

For $p(\boldsymbol{\nu})$, we have the update rules similar as in [9], that is,

$$\gamma_{k1} = \alpha + \sum_{m=k}^{K}\sum_{n=1}^{N}\psi_{nm} + \sum_{m=k+1}^{K}(N - \sum_{n=1}^{N}\psi_{nm})(\sum_{i=k+1}^{m} q_{mi}) \tag{26}$$
$$\gamma_{k2} = 1 + \sum_{m=k}^{K}(N - \sum_{n=1}^{N}\psi_{nm}) q_{mk},$$

where $q_{\cdot k}$ is computed in the same way as in Eq. (17). For $p(\mathbf{Z})$, the mean-field update equation for ψ is

$$\psi_{nk} = \frac{1}{1 + e^{-\vartheta_{nk}}}, \tag{27}$$

where

$$\vartheta_{nk} = \sum_{j=1}^{k}\mathbb{E}_p[\log v_j] - \mathcal{L}_k^{\nu}(p) - \frac{1}{2\sigma_{n0}^2}(D\sigma_k^2 + \Phi_{\cdot k}^\top \Phi_{\cdot k})$$
$$+ \frac{1}{\sigma_{n0}^2}\Phi_{\cdot k}^\top\Big(\mathbf{x}_n - \sum_{j \neq k}\psi_{nj}\Phi_{\cdot j}\Big) + \sum_{y}\omega_n^y \mathbb{E}_p[\eta_{y_n}^k - \eta_y^k].$$

For testing data, ϑ_{nk} does not have the last term because of the absence of large-margin constraints.

Infer $p(\boldsymbol{\eta})$ and solve for $(\boldsymbol{\xi}, \boldsymbol{\omega}, \mathbf{u})$: We can optimize L to solve for $q(\boldsymbol{\eta})$, which is

$$p(\boldsymbol{\eta}) \propto \pi(\boldsymbol{\eta}) \exp\Big\{\boldsymbol{\eta}^\top(\sum_{n \in \mathcal{I}_{\text{tr}}}\sum_{y}\omega_n^y \mathbb{E}_p[\mathbf{g}(y_n, \mathbf{x}_n, \mathbf{z}_n) - \mathbf{g}(y, \mathbf{x}_n, \mathbf{z}_n)])\Big\}.$$

For the standard normal prior $\pi(\boldsymbol{\eta})$, we have that $q(\boldsymbol{\eta})$ is also normal, with mean

$$\boldsymbol{\mu} = \sum_{n \in \mathcal{I}_{\text{tr}}}\sum_{y}\omega_d^y \mathbb{E}_p[\mathbf{g}(y_n, \mathbf{x}_n, \mathbf{z}_n) - \mathbf{g}(y, \mathbf{x}_n, \mathbf{z}_n)]$$

and identity covariance matrix. Substituting the solution of $p(\boldsymbol{\eta})$ into the Lagrangian functional, we get the dual problem

$$\max_{\boldsymbol{\omega}} \; -\frac{1}{2}\boldsymbol{\mu}^\top\boldsymbol{\mu} + \sum_{n \in \mathcal{I}_{\text{tr}}}\sum_{y}\omega_n^y \quad \text{s.t.. : } 0 \leq \sum_{y}\omega_n^y \leq C, \forall n \in \mathcal{I}_{\text{tr}}, \tag{28}$$

which (and its primal form) can be efficiently solved with a multi-class SVM solver.

Similar as in MT-iLSVM, the hyperparameters σ_0^2 and σ_{n0}^2 can be set a priori or estimated from the data. The empirical estimation can be easily done with closed form solutions. For iLSVM, we have

$$\sigma_0^2 = \frac{\sum_{k=1}^{K}(D\sigma_k^2 + \Phi_{\cdot k}^\top\Phi_k)}{KD} \tag{29}$$
$$\sigma_{n0}^2 = \frac{\mathbf{x}_n^\top\mathbf{x}_n - 2\mathbf{x}_n^\top\Phi\mathbb{E}_p[\mathbf{z}_n]^\top + \mathbb{E}_p[\mathbf{z}_n\mathbf{A}\mathbf{z}_n^\top]}{D}. \tag{30}$$