Supplementary Material

In this extra material, we provide more details about the variational lower bound of "*Bayesian Efficient Multiple Kernel Learning*" and the modified Bayesian model for multiclass classification in Appendices A and B, respectively.

A. Variational Lower Bound of Bayesian Efficient Multiple Kernel Learning

The variational lower bound of our model can be written as

$$\mathcal{L} = \mathrm{E}_{q(\boldsymbol{\Theta},\boldsymbol{\Xi})}[\log p(\boldsymbol{y},\boldsymbol{\Theta},\boldsymbol{\Xi}|\{\mathbf{K}_m\}_{m=1}^P)] - \mathrm{E}_{q(\boldsymbol{\Theta},\boldsymbol{\Xi})}[\log q(\boldsymbol{\Theta},\boldsymbol{\Xi})]$$

where the joint likelihood and the factorable ensemble approximation are defined as

$$\begin{aligned} p(\boldsymbol{y},\boldsymbol{\Theta},\boldsymbol{\Xi}|\{\mathbf{K}_m\}_{m=1}^P) &= p(\boldsymbol{\lambda})p(\boldsymbol{a}|\boldsymbol{\lambda})p(\mathbf{G}|\boldsymbol{a},\{\mathbf{K}_m\}_{m=1}^P)p(\gamma)p(b|\gamma)p(\boldsymbol{\omega})p(\boldsymbol{e}|\boldsymbol{\omega})p(\boldsymbol{f}|b,\boldsymbol{e},\mathbf{G})p(\boldsymbol{y}|\boldsymbol{f}) \\ q(\boldsymbol{\Theta},\boldsymbol{\Xi}) &= q(\boldsymbol{\lambda})q(\boldsymbol{a})q(\mathbf{G})q(\gamma)q(\boldsymbol{\omega})q(b,\boldsymbol{e})q(\boldsymbol{f}). \end{aligned}$$

Using these definitions, the variational lower bound becomes

$$\begin{aligned} \mathcal{L} = {} & \mathrm{E}_{q(\boldsymbol{\lambda})}[\log p(\boldsymbol{\lambda})] + \mathrm{E}_{q(\boldsymbol{\lambda})q(\boldsymbol{a})}[\log p(\boldsymbol{a}|\boldsymbol{\lambda})] + \mathrm{E}_{q(\boldsymbol{a})q(\mathbf{G})}[\log p(\mathbf{G}|\boldsymbol{a},\{\mathbf{K}_m\}_{m=1}^P)] + \mathrm{E}_{q(\gamma)}[\log p(\gamma)] \\ & + \mathrm{E}_{q(\gamma)q(b,\boldsymbol{e})}[\log p(b|\gamma)] + \mathrm{E}_{q(\boldsymbol{\omega})}[\log p(\boldsymbol{\omega})] + \mathrm{E}_{q(\boldsymbol{\omega})q(b,\boldsymbol{e})}[\log p(\boldsymbol{e}|\boldsymbol{\omega})] + \mathrm{E}_{q(\mathbf{G})q(b,\boldsymbol{e})q(\boldsymbol{f})}[\log p(\boldsymbol{f}|b,\boldsymbol{e},\mathbf{G})] \\ & + \mathrm{E}_{q(\boldsymbol{f})}[\log p(\boldsymbol{y}|\boldsymbol{f})] - \mathrm{E}_{q(\boldsymbol{\lambda})}[\log q(\boldsymbol{\lambda})] - \mathrm{E}_{q(\boldsymbol{a})}[\log q(\boldsymbol{a})] - \mathrm{E}_{q(\mathbf{G})}[\log q(\mathbf{G})] - \mathrm{E}_{q(\gamma)}[\log q(\gamma)] - \mathrm{E}_{q(\boldsymbol{\omega})}[\log q(\boldsymbol{\omega})] \\ & - \mathrm{E}_{q(b,\boldsymbol{e})}[\log q(b,\boldsymbol{e})] - \mathrm{E}_{q(\boldsymbol{f})}[\log q(\boldsymbol{f})] \end{aligned}$$

where the exponential form expectations of the distributions in the joint likelihood can be calculated as

$$\begin{aligned} \mathrm{E}_{q(\boldsymbol{\lambda})}[\log p(\boldsymbol{\lambda})] &= \sum_{i=1}^{N}\left((\alpha_\lambda - 1)\widetilde{\log\lambda_i} - \frac{\widetilde{\lambda}_i}{\beta_\lambda} - \log\Gamma(\alpha_\lambda) - \alpha_\lambda\log\beta_\lambda\right) \\ \mathrm{E}_{q(\boldsymbol{\lambda})q(\boldsymbol{a})}[\log p(\boldsymbol{a}|\boldsymbol{\lambda})] &= -\frac{1}{2}\operatorname{tr}(\operatorname{diag}(\widetilde{\boldsymbol{\lambda}})\widetilde{\boldsymbol{a}\boldsymbol{a}^\top}) - \frac{1}{2}N\log 2\pi + \frac{1}{2}\log|\operatorname{diag}(\widetilde{\boldsymbol{\lambda}})| \\ \mathrm{E}_{q(\boldsymbol{a})q(\mathbf{G})}[\log p(\mathbf{G}|\boldsymbol{a},\{\mathbf{K}_m\}_{m=1}^P)] &= \sum_{i=1}^{N}\left(-\frac{1}{2}\widetilde{\boldsymbol{g}_i^\top\boldsymbol{g}_i} + \operatorname{tr}\left(\begin{bmatrix}\boldsymbol{k}_1^i\\ \vdots \\ \boldsymbol{k}_P^i\end{bmatrix}^\top\widetilde{\boldsymbol{g}_i\boldsymbol{a}^\top}\right) - \frac{1}{2}\operatorname{tr}\left(\begin{bmatrix}\boldsymbol{k}_1^i\\ \vdots \\ \boldsymbol{k}_P^i\end{bmatrix}^\top\begin{bmatrix}\boldsymbol{k}_1^i\\ \vdots \\ \boldsymbol{k}_P^i\end{bmatrix}\widetilde{\boldsymbol{a}\boldsymbol{a}^\top}\right) - \frac{1}{2}P\log 2\pi\right) \\ \mathrm{E}_{q(\gamma)}[\log p(\gamma)] &= (\alpha_\gamma - 1)\widetilde{\log\gamma} - \frac{\widetilde{\gamma}}{\beta_\gamma} - \log\Gamma(\alpha_\gamma) - \alpha_\gamma\log\beta_\gamma \\ \mathrm{E}_{q(\gamma)q(b,\boldsymbol{e})}[\log p(b|\gamma)] &= -\frac{1}{2}\widetilde{\gamma}\widetilde{b^2} - \frac{1}{2}\log 2\pi + \frac{1}{2}\log\widetilde{\gamma} \\ \mathrm{E}_{q(\boldsymbol{\omega})}[\log p(\boldsymbol{\omega})] &= \sum_{m=1}^{P}\left((\alpha_\omega - 1)\widetilde{\log\omega_m} - \frac{\widetilde{\omega_m}}{\beta_\omega} - \log\Gamma(\alpha_\omega) - \alpha_\omega\log\beta_\omega\right) \\ \mathrm{E}_{q(\boldsymbol{\omega})q(b,\boldsymbol{e})}[\log p(\boldsymbol{e}|\boldsymbol{\omega})] &= -\frac{1}{2}\operatorname{tr}(\operatorname{diag}(\widetilde{\boldsymbol{\omega}})\widetilde{\boldsymbol{e}\boldsymbol{e}^\top}) - \frac{1}{2}P\log 2\pi + \frac{1}{2}\log|\operatorname{diag}(\widetilde{\boldsymbol{\omega}})| \\ \mathrm{E}_{q(\mathbf{G})q(b,\boldsymbol{e})q(\boldsymbol{f})}[\log p(\boldsymbol{f}|b,\boldsymbol{e},\mathbf{G})] &= \sum_{i=1}^{N}\left(-\frac{1}{2}\widetilde{f_i^2} + (\widetilde{\boldsymbol{e}^\top}\widetilde{\boldsymbol{g}_i} + \widetilde{b})\widetilde{f_i} - \frac{1}{2}\left(\operatorname{tr}(\widetilde{\boldsymbol{e}\boldsymbol{e}^\top}\widetilde{\boldsymbol{g}_i\boldsymbol{g}_i^\top}) + 2\widetilde{b\boldsymbol{e}^\top}\widetilde{\boldsymbol{g}_i} + \widetilde{b^2}\right) - \frac{1}{2}\log 2\pi\right) \\ \mathrm{E}_{q(\boldsymbol{f})}[\log p(\boldsymbol{y}|\boldsymbol{f})] &= 0 \end{aligned}$$

and the negative entropies of the approximate posteriors in the ensemble are given as

$$\begin{aligned} \mathrm{E}_{q(\boldsymbol{\lambda})}[\log q(\boldsymbol{\lambda})] &= \sum_{i=1}^{N}(-\alpha(\lambda_i) - \log\beta(\lambda_i) - \log\Gamma(\alpha(\lambda_i)) - (1-\alpha(\lambda_i))\psi(\alpha(\lambda_i))) \\ \mathrm{E}_{q(\boldsymbol{a})}[\log q(\boldsymbol{a})] &= -\frac{1}{2}N(\log 2\pi + 1) - \frac{1}{2}\log|\Sigma(\boldsymbol{a})| \end{aligned}$$

$$\begin{aligned}
\mathrm{E}_{q(\mathbf{G})}[\log q(\mathbf{G})] &= \sum_{i=1}^{N}\left(-\frac{1}{2}P(\log 2\pi+1)-\frac{1}{2}\log|\Sigma(\boldsymbol{g}_i)|\right)\\
\mathrm{E}_{q(\gamma)}[\log q(\gamma)] &= -\alpha(\gamma)-\log\beta(\gamma)-\log\Gamma(\alpha(\gamma))-(1-\alpha(\gamma))\psi(\alpha(\gamma))\\
\mathrm{E}_{q(\boldsymbol{\omega})}[\log q(\boldsymbol{\omega})] &= \sum_{m=1}^{P}(-\alpha(\omega_i)-\log\beta(\omega_i)-\log\Gamma(\alpha(\omega_i))-(1-\alpha(\omega_i))\psi(\alpha(\omega_i)))\\
\mathrm{E}_{q(b,\boldsymbol{e})}[\log q(b,\boldsymbol{e})] &= -\frac{1}{2}(P+1)(\log 2\pi+1)-\frac{1}{2}\log|\Sigma(b,\boldsymbol{e})|\\
\mathrm{E}_{q(\boldsymbol{f})}[\log q(\boldsymbol{f})] &= \sum_{i=1}^{N}\left(-\frac{1}{2}(\log 2\pi+\Sigma(f_i))-\log\mathcal{Z}_i\right)
\end{aligned}$$

where $\Gamma(\cdot)$ denotes the gamma function and $\psi(\cdot)$ denotes the digamma function. The only nonstandard distribution we need to operate on is the truncated normal distribution used for the auxiliary variables. From our model definition, the truncation points for each auxiliary variable are defined as

$$(l_i,u_i)=\begin{cases}(-\infty,-\nu) & \text{if } y_i=-1\\ (+\nu,+\infty) & \text{otherwise}\end{cases}$$

where l_i and u_i denote the lower and upper truncation points, respectively. The normalization coefficient, the expectation, and the variance of the auxiliary variables can be calculated as

$$\begin{aligned}
\mathcal{Z}_i &= \Phi(\beta_i)-\Phi(\alpha_i)\\
\widetilde{f}_i &= \widetilde{\boldsymbol{e}^\top}\widetilde{\boldsymbol{g}}_i+\widetilde{b}+\frac{\phi(\alpha_i)-\phi(\beta_i)}{\mathcal{Z}_i}\\
\widetilde{f_i^2}-(\widetilde{f}_i)^2 &= 1+\frac{\alpha_i\phi(\alpha_i)-\beta_i\phi(\beta_i)}{\mathcal{Z}_i}-\frac{(\phi(\alpha_i)-\phi(\beta_i))^2}{\mathcal{Z}_i^2}
\end{aligned}$$

where $\Phi(\cdot)$ is the standardized normal cumulative distribution function, $\phi(\cdot)$ is the standardized normal probability density function, and $\{\alpha_i,\beta_i\}$ are defined as

$$\begin{aligned}
\alpha_i &= l_i-\widetilde{\boldsymbol{e}^\top}\widetilde{\boldsymbol{g}}_i-\widetilde{b}\\
\beta_i &= u_i-\widetilde{\boldsymbol{e}^\top}\widetilde{\boldsymbol{g}}_i-\widetilde{b}.
\end{aligned}$$

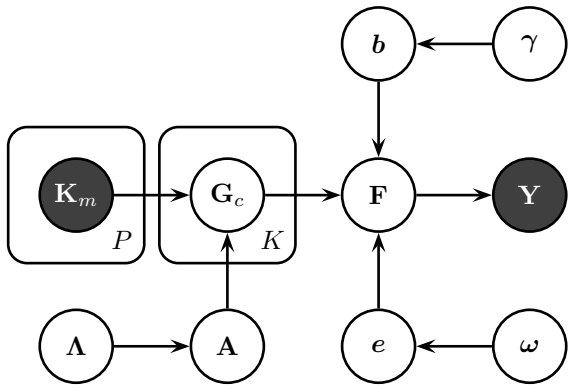
B. Bayesian Efficient Multiple Kernel Learning for Multiclass Classification

We modify the probabilistic model proposed for binary classification using an approach similar to one-versus-all idea but with a common set of kernel weights. Figure 1 illustrates the modified probabilistic model for multiclass classification with a graphical model and its distributional assumptions.

There are slight modifications to the notation but we explain them in detail for completeness. N, P, and K represent the numbers of training instances, input kernels, and classes, respectively. The $N\times N$ kernel matrices are denoted by $\mathbf{K}_m$, where the columns of $\mathbf{K}_m$ by $\boldsymbol{k}_{m,i}$ and the rows of $\mathbf{K}_m$ by $\boldsymbol{k}_m^i$. The $N\times K$ matrices of weight parameters a_c^i and their priors λ_c^i are denoted by $\mathbf{A}$ and $\mathbf{\Lambda}$, respectively, where the columns of $\mathbf{A}$ and $\mathbf{\Lambda}$ by $\boldsymbol{a}_c$ and $\boldsymbol{\lambda}_c$. The $P\times N$ matrices of intermediate outputs for each class $g_{c,i}^m$ are represented as $\mathbf{G}_c$, where the columns of $\mathbf{G}_c$ as $\boldsymbol{g}_{c,i}$ and the rows of $\mathbf{G}_c$ as $\boldsymbol{g}_c^m$. The vectors of bias parameters b_c and their priors γ_c are denoted by $\boldsymbol{b}$ and $\boldsymbol{\gamma}$, respectively. The $P\times 1$ vector of kernel weights e_m and their priors ω_m are denoted by $\boldsymbol{e}$ and $\boldsymbol{\omega}$, respectively. The $K\times N$ matrix of auxiliary variables f_i^c is represented as $\mathbf{F}$, where the columns of $\mathbf{F}$ as $\boldsymbol{f}_i$. The $K\times N$ matrix of associated class labels is represented as $\mathbf{Y}$, where each element $y_i^c\in\{-1,+1\}$ and there is only one $+1$ in each column (i.e., 1-of-K encoding). As short-hand notations, all priors in the model are denoted by $\mathbf{\Xi}=\{\boldsymbol{\gamma},\mathbf{\Lambda},\boldsymbol{\omega}\}$, where the remaining variables by $\mathbf{\Theta}=\{\mathbf{A},\boldsymbol{b},\boldsymbol{e},\mathbf{F},\{\mathbf{G}_c\}_{c=1}^K\}$ and the hyper-parameters by $\boldsymbol{\zeta}=\{\alpha_\gamma,\beta_\gamma,\alpha_\lambda,\beta_\lambda,\alpha_\omega,\beta_\omega\}$. Dependence on $\boldsymbol{\zeta}$ is again omitted for clarity throughout the rest.

We can write the factorable ensemble approximation of the required posterior as

$$p(\mathbf{\Theta},\mathbf{\Xi}|\{\mathbf{K}_m\}_{m=1}^P,\mathbf{Y})\approx q(\mathbf{\Theta},\mathbf{\Xi})=q(\mathbf{\Lambda})q(\mathbf{A})q(\{\mathbf{G}_c\}_{c=1}^K)q(\boldsymbol{\gamma})q(\boldsymbol{\omega})q(\boldsymbol{b},\boldsymbol{e})q(\mathbf{F})$$



$$
\begin{aligned}
\lambda_c^i &\sim \mathcal{G}(\lambda_c^i; \alpha_\lambda, \beta_\lambda) && \forall (i,c)\\
a_c^i|\lambda_c^i &\sim \mathcal{N}(a_c^i; 0, (\lambda_c^i)^{-1}) && \forall (i,c)\\
g_{c,i}^m|\boldsymbol{a}_c, \boldsymbol{k}_{m,i} &\sim \mathcal{N}(g_{c,i}^m; \boldsymbol{a}_c^\top \boldsymbol{k}_{m,i}, 1) && \forall (c,m,i)\\
\gamma_c &\sim \mathcal{G}(\gamma_c; \alpha_\gamma, \beta_\gamma) && \forall c\\
b_c|\gamma_c &\sim \mathcal{N}(b_c; 0, \gamma_c^{-1}) && \forall c\\
\omega_m &\sim \mathcal{G}(\omega_m; \alpha_\omega, \beta_\omega) && \forall m\\
e_m|\omega_m &\sim \mathcal{N}(e_m; 0, \omega_m^{-1}) && \forall m\\
f_i^c|b_c, \boldsymbol{e}, \boldsymbol{g}_{c,i} &\sim \mathcal{N}(f_i^c; \boldsymbol{e}^\top \boldsymbol{g}_{c,i} + b_c, 1) && \forall (c,i)\\
y_i^c|f_i^c &\sim \delta(f_i^c y_i^c > \nu) && \forall (c,i)
\end{aligned}
$$

Figure 1. Bayesian efficient MKL for multiclass classification.

and define each factor in the ensemble just like its full conditional distribution:

$$
\begin{aligned}
q(\boldsymbol{\Lambda}) &= \prod_{i=1}^{N}\prod_{c=1}^{K} \mathcal{G}(\lambda_c^i; \alpha(\lambda_c^i), \beta(\lambda_c^i))\\
q(\mathbf{A}) &= \prod_{c=1}^{K} \mathcal{N}(\boldsymbol{a}_c; \mu(\boldsymbol{a}_c), \Sigma(\boldsymbol{a}_c))\\
q(\{\mathbf{G}_c\}_{c=1}^K) &= \prod_{c=1}^{K}\prod_{i=1}^{N} \mathcal{N}(\boldsymbol{g}_{c,i}; \mu(\boldsymbol{g}_{c,i}), \Sigma(\boldsymbol{g}_{c,i}))\\
q(\boldsymbol{\gamma}) &= \prod_{c=1}^{K} \mathcal{G}(\gamma_c; \alpha(\gamma_c), \beta(\gamma_c))\\
q(\boldsymbol{\omega}) &= \prod_{m=1}^{P} \mathcal{G}(\omega_m; \alpha(\omega_m), \beta(\omega_m))\\
q(\boldsymbol{b}, \boldsymbol{e}) &= \mathcal{N}\left(\begin{bmatrix}\boldsymbol{b}\\ \boldsymbol{e}\end{bmatrix}; \mu(\boldsymbol{b}, \boldsymbol{e}), \Sigma(\boldsymbol{b}, \boldsymbol{e})\right)\\
q(\mathbf{F}) &= \prod_{c=1}^{K}\prod_{i=1}^{N} \mathcal{TN}(f_i^c; \mu(f_i^c), \Sigma(f_i^c), \rho(f_i^c)).
\end{aligned}
$$

We can bound the marginal likelihood using Jensen's inequality:

$$
\log p(\mathbf{Y}|\{\mathbf{K}_m\}_{m=1}^P) \geq \mathrm{E}_{q(\boldsymbol{\Theta},\boldsymbol{\Xi})}[\log p(\mathbf{Y}, \boldsymbol{\Theta}, \boldsymbol{\Xi}|\{\mathbf{K}_m\}_{m=1}^P)] - \mathrm{E}_{q(\boldsymbol{\Theta},\boldsymbol{\Xi})}[\log q(\boldsymbol{\Theta}, \boldsymbol{\Xi})]
$$

and optimize this bound by maximizing with respect to each factor separately until convergence. The approximate posterior distribution of a specific factor $\boldsymbol{\tau}$ can be found as

$$
q(\boldsymbol{\tau}) \propto \exp(\mathrm{E}_{q(\{\boldsymbol{\Theta},\boldsymbol{\Xi}\}\setminus\boldsymbol{\tau})}[\log p(\mathbf{Y}, \boldsymbol{\Theta}, \boldsymbol{\Xi}|\{\mathbf{K}_m\}_{m=1}^P)]).
$$

The approximate posterior distributions of the ensemble factors can be found as

$$
q(\boldsymbol{\Lambda}) = \prod_{i=1}^{N}\prod_{c=1}^{K} \mathcal{G}\left(\lambda_c^i; \alpha_\lambda + \frac{1}{2}, \left(\frac{1}{\beta_\lambda} + \frac{\widetilde{(a_c^i)^2}}{2}\right)^{-1}\right)
$$

$$q(\boldsymbol{\gamma}) = \prod_{c=1}^{K} \mathcal{G}\left(\gamma_c; \alpha_\gamma + \frac{1}{2}, \left(\frac{1}{\beta_\gamma} + \frac{\widetilde{b_c^2}}{2}\right)^{-1}\right)$$

$$q(\boldsymbol{\omega}) = \prod_{m=1}^{P} \mathcal{G}\left(\omega_m; \alpha_\omega + \frac{1}{2}, \left(\frac{1}{\beta_\omega} + \frac{\widetilde{e_m^2}}{2}\right)^{-1}\right)$$

$$q(\mathbf{A}) = \prod_{c=1}^{K} \mathcal{N}\left(\boldsymbol{a}_c; \Sigma(\boldsymbol{a}_c)\left(\sum_{m=1}^{P} \mathbf{K}_m \widetilde{\boldsymbol{g}_c^{m\top}}\right), \left(\operatorname{diag}(\widetilde{\boldsymbol{\lambda}_c}) + \sum_{m=1}^{P} \mathbf{K}_m \mathbf{K}_m\right)^{-1}\right)$$

$$q(\{\mathbf{G}_c\}_{c=1}^{K}) = \prod_{c=1}^{K}\prod_{i=1}^{N} \mathcal{N}\left(\boldsymbol{g}_{c,i}; \Sigma(\boldsymbol{g}_{c,i})\left(\begin{bmatrix}\boldsymbol{k}_1^i \\ \vdots \\ \boldsymbol{k}_P^i\end{bmatrix}\widetilde{\boldsymbol{a}_c} + \widetilde{f_i^c}\widetilde{\boldsymbol{e}} - \widetilde{b_c}\widetilde{\boldsymbol{e}}\right), \left(\mathbf{I} + \widetilde{\boldsymbol{e}\boldsymbol{e}^\top}\right)^{-1}\right)$$

$$q(\boldsymbol{b}, \boldsymbol{e}) = \mathcal{N}\left(\begin{bmatrix}\boldsymbol{b} \\ \boldsymbol{e}\end{bmatrix}; \Sigma(\boldsymbol{b}, \boldsymbol{e})\begin{bmatrix}\widetilde{\mathbf{F}}\mathbf{1} \\ \sum_{i=1}^{N} \widetilde{\mathbf{G}_{\cdot,i}}\widetilde{\boldsymbol{f_i}}\end{bmatrix}, \begin{bmatrix}\operatorname{diag}(\widetilde{\boldsymbol{\gamma}}) + N\mathbf{I} & \sum_{i=1}^{N} \widetilde{\mathbf{G}_{\cdot,i}^\top} \\ \sum_{i=1}^{N} \widetilde{\mathbf{G}_{\cdot,i}} & \operatorname{diag}(\widetilde{\boldsymbol{\omega}}) + \sum_{i=1}^{N} \widetilde{\mathbf{G}_{\cdot,i}\mathbf{G}_{\cdot,i}^\top}\end{bmatrix}^{-1}\right)$$

$$q(\mathbf{F}) = \prod_{c=1}^{K}\prod_{i=1}^{N} \mathcal{TN}(f_i^c; \widetilde{\boldsymbol{e}^\top}\widetilde{\boldsymbol{g}_{c,i}} + \widetilde{b_c}, 1, f_i^c y_i^c > \nu).$$

We can replace $p(\mathbf{A}|\{\mathbf{K}_m\}_{m=1}^P, \mathbf{Y})$ with its approximate posterior distribution $q(\mathbf{A})$ and obtain the predictive distribution of the intermediate outputs $\mathbf{G}_{\cdot,\star}$ for a new data point as

$$p(\mathbf{G}_{\cdot,\star}|\{\boldsymbol{k}_{m,\star}, \mathbf{K}_m\}_{m=1}^P, \mathbf{Y}) = \prod_{c=1}^{K}\prod_{m=1}^{P} \mathcal{N}(g_{c,\star}^m; \mu(\boldsymbol{a}_c)^\top \boldsymbol{k}_{m,\star}, 1 + \boldsymbol{k}_{m,\star}^\top \Sigma(\boldsymbol{a}_c)\boldsymbol{k}_{m,\star}).$$

The predictive distribution of the auxiliary variables $\boldsymbol{f}_\star$ can also be found by replacing $p(\boldsymbol{b}, \boldsymbol{e}|\{\mathbf{K}_m\}_{m=1}^P, \mathbf{Y})$ with its approximate posterior distribution $q(\boldsymbol{b}, \boldsymbol{e})$:

$$p(\boldsymbol{f}_\star|\mathbf{G}_{\cdot,\star}, \{\mathbf{K}_m\}_{m=1}^P, \mathbf{Y}) = \prod_{c=1}^{K} \mathcal{N}\left(f_\star^c; \mu(b_c, \boldsymbol{e})^\top \begin{bmatrix}1 \\ \boldsymbol{g}_{c,\star}\end{bmatrix}, 1 + \begin{bmatrix}1 & \boldsymbol{g}_{c,\star}\end{bmatrix}\Sigma(b_c, \boldsymbol{e})\begin{bmatrix}1 \\ \boldsymbol{g}_{c,\star}\end{bmatrix}\right)$$

and the predictive distribution of the class label $y_\star^c$ can be formulated using the auxiliary variable distribution:

$$p(y_\star^c = +1|\{\boldsymbol{k}_{m,\star}, \mathbf{K}_m\}_{m=1}^P, \mathbf{Y}) = (\mathcal{Z}_\star^c)^{-1}\Phi\left(\frac{\mu(f_\star^c) - \nu}{\Sigma(f_\star^c)}\right)$$

where $\mathcal{Z}_\star^c$ is the normalization coefficient calculated for class c. The test data point is assigned to the class that has the highest probability.