1 Single Layer Model with One Input Kernel

Consider that the phenotype $\boldsymbol{Y}$ is modeled as a random effect model: given $\boldsymbol{U}_1, \ldots, \boldsymbol{U}_m$,

$$\boldsymbol{Y} = \boldsymbol{a} + \boldsymbol{\epsilon}, \quad \text{where } \boldsymbol{a} \sim \mathcal{N}_n(\boldsymbol{0}, \tau \boldsymbol{\Sigma}(\boldsymbol{U}_1, \ldots, \boldsymbol{U}_m)),$$

that is the covariance matrix of the random effect $\boldsymbol{a}$ depends on latent variables $\boldsymbol{U}_1, \ldots, \boldsymbol{U}_m$. Moreover, the latent variable $\boldsymbol{U}_i$ is modeled using another random effect model

$$\boldsymbol{U}_i = \boldsymbol{a}'_i + \boldsymbol{\epsilon}, \quad \text{where } \boldsymbol{a}'_i \sim \mathcal{N}_n(\boldsymbol{0}, \tau'_i \boldsymbol{\Sigma}')$$

The best predictor for $\boldsymbol{a}$ can be obtained as follows:

$$\left.\begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{Y} \end{bmatrix}\right| \boldsymbol{U}_1, \ldots, \boldsymbol{U}_m \sim \mathcal{N}_{2n}\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \tilde{\tau}\phi\boldsymbol{\Sigma} & \tilde{\tau}\phi\boldsymbol{\Sigma} \\ \tilde{\tau}\phi\boldsymbol{\Sigma} & \phi(\tilde{\tau}\boldsymbol{\Sigma} + \boldsymbol{I}_n) \end{bmatrix} \right),$$

where $\tilde{\tau} = \phi^{-1}\tau$ and the best predictor for $\boldsymbol{a}$ is given by

$$\hat{\boldsymbol{a}} = \mathbb{E}\left[\boldsymbol{a}|\boldsymbol{Y}\right] = \mathbb{E}_{\boldsymbol{U}_1,\ldots,\boldsymbol{U}_m}\left[\mathbb{E}\left(\boldsymbol{a}|\boldsymbol{Y}, \boldsymbol{U}_1, \ldots, \boldsymbol{U}_m\right)\right] = \mathbb{E}_{\boldsymbol{U}_1,\ldots,\boldsymbol{U}_m}\left[\tilde{\tau}\boldsymbol{\Sigma}(\tilde{\tau}\boldsymbol{\Sigma} + \boldsymbol{I}_n)^{-1}\boldsymbol{Y}\right]$$

Define $\boldsymbol{U} = [\boldsymbol{U}_1, \ldots, \boldsymbol{U}_m]$, then if $\boldsymbol{U}$ is given and $\boldsymbol{\Sigma}(\boldsymbol{U}_1, \ldots, \boldsymbol{U}_m)$ is defined using product kernel, we have

$$\boldsymbol{\Sigma}(\boldsymbol{U}_1, \ldots, \boldsymbol{U}_m) = \boldsymbol{U}\boldsymbol{U}^T.$$

Hence, the predicted response $\boldsymbol{Y}$ is given by

$$\hat{\boldsymbol{Y}} = \hat{\boldsymbol{a}} = \mathbb{E}_{\boldsymbol{U}_1,\ldots,\boldsymbol{U}_m}\left[\tilde{\tau}\boldsymbol{U}\boldsymbol{U}^T(\tilde{\tau}\boldsymbol{U}\boldsymbol{U}^T + \boldsymbol{I}_n)^{-1pred}\boldsymbol{Y}\right] = \tilde{\tau}\mathbb{E}_{\boldsymbol{U}_1,\ldots,\boldsymbol{U}_m}\left[\boldsymbol{U}\boldsymbol{U}^T(\tilde{\tau}\boldsymbol{U}\boldsymbol{U}^T + \boldsymbol{I}_n)^{-1}\right]\boldsymbol{Y}.$$

To learn the parameters $\tilde{\tau}, \tilde{\tau}'_1, \ldots, \tilde{\tau}'_m$, we need to minimize the prediction error, which is given by

$$(\boldsymbol{Y} - \hat{\boldsymbol{Y}})^T(\boldsymbol{Y} - \hat{\boldsymbol{Y}}) = \boldsymbol{Y}^T\left(\boldsymbol{I}_n - \mathbb{E}_{\boldsymbol{U}_1,\ldots,\boldsymbol{U}_m}[\tilde{\tau}\boldsymbol{U}\boldsymbol{U}^T(\tilde{\tau}\boldsymbol{U}\boldsymbol{U}^T + \boldsymbol{I}_n)^{-1}]\right)^T\left(\boldsymbol{I}_n - \mathbb{E}_{\boldsymbol{U}_1,\ldots,\boldsymbol{U}_m}[\tilde{\tau}\boldsymbol{U}\boldsymbol{U}^T(\tilde{\tau}\boldsymbol{U}\boldsymbol{U}^T + \boldsymbol{I}_n)^{-1}]\right)\boldsymbol{Y}.$$

Note that

$$\begin{aligned} \boldsymbol{I}_n - \mathbb{E}_{\boldsymbol{U}_1,\ldots,\boldsymbol{U}_m}[\tilde{\tau}\boldsymbol{U}\boldsymbol{U}^T(\tilde{\tau}\boldsymbol{U}\boldsymbol{U}^T + \boldsymbol{I}_n)^{-1}] &= \mathbb{E}_{\boldsymbol{U}_1,\ldots,\boldsymbol{U}_m}\left[\left(\tilde{\tau}\boldsymbol{U}\boldsymbol{U}^T + \boldsymbol{I}_n - \tilde{\tau}\boldsymbol{U}\boldsymbol{U}^T\right)\left(\tilde{\tau}\boldsymbol{U}\boldsymbol{U}^T + \boldsymbol{I}_n\right)^{-1}\right] \\ &= \mathbb{E}_{\boldsymbol{U}_1,\ldots,\boldsymbol{U}_m}\left[\left(\tilde{\tau}\boldsymbol{U}\boldsymbol{U}^T + \boldsymbol{I}_n\right)^{-1}\right], \end{aligned}$$

we get

$$(\boldsymbol{Y}-\hat{\boldsymbol{Y}})^T(\boldsymbol{Y}-\hat{\boldsymbol{Y}}) = \boldsymbol{Y}^T\left(\mathbb{E}_{\boldsymbol{U}_1,\ldots,\boldsymbol{U}_m}\left[\left(\tilde{\tau}\boldsymbol{U}\boldsymbol{U}^T+\boldsymbol{I}_n\right)^{-1}\right]\right)^2\boldsymbol{Y}.$$

Since $\boldsymbol{U}_i \sim \mathcal{N}_n(\boldsymbol{0}, \tau_i'\boldsymbol{\Sigma}'+\phi\boldsymbol{I}_n)$, we can know that $\boldsymbol{U}_i \overset{d}{=} \phi^{\frac{1}{2}}(\tilde{\tau}_i'\boldsymbol{\Sigma}'+\boldsymbol{I}_n)^{\frac{1}{2}}\boldsymbol{Z}_i$, where $\tilde{\tau}_i' = \phi^{-1}\tau_i'$ and $\boldsymbol{Z}_1,\ldots,\boldsymbol{Z}_m \sim \mathcal{N}_n(\boldsymbol{0},\boldsymbol{I}_n)$. Then we get

$$\boldsymbol{U} \overset{d}{=} \phi^{\frac{1}{2}}\begin{bmatrix}\left(\tilde{\tau}_1'\boldsymbol{\Sigma}'+\boldsymbol{I}_n\right)^{\frac{1}{2}} & \cdots & \left(\tilde{\tau}_m'\boldsymbol{\Sigma}'+\boldsymbol{I}_n\right)^{\frac{1}{2}}\end{bmatrix}\begin{bmatrix}\boldsymbol{Z}_1 & & \\ & \ddots & \\ & & \boldsymbol{Z}_m\end{bmatrix} := \phi^{\frac{1}{2}}\boldsymbol{D}(\tilde{\tau}_1',\ldots,\tilde{\tau}_m')\boldsymbol{Z}$$

and hence

$$\mathbb{E}_{\boldsymbol{U}_1,\ldots,\boldsymbol{U}_m}\left[\left(\tilde{\tau}\boldsymbol{U}\boldsymbol{U}^T+\boldsymbol{I}_n\right)^{-1}\right] = \mathbb{E}_{\boldsymbol{Z}_1,\ldots,\boldsymbol{Z}_m}\left[\left(\tilde{\tau}\phi\boldsymbol{D}\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{D}^T+\boldsymbol{I}_n\right)^{-1}\right] = \mathbb{E}_{\boldsymbol{Z}_1,\ldots,\boldsymbol{Z}_m}\left[\left(\tau\boldsymbol{D}\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{D}^T+\boldsymbol{I}_n\right)^{-1}\right]$$

which implies that

$$\begin{aligned}R(\tau,\tilde{\tau}_1',\ldots,\tilde{\tau}_m') &:= \boldsymbol{Y}^T\left(\mathbb{E}_{\boldsymbol{U}_1,\ldots,\boldsymbol{U}_m}\left[\left(\tilde{\tau}\boldsymbol{U}\boldsymbol{U}^T+\boldsymbol{I}_n\right)^{-1}\right]\right)^2\boldsymbol{Y}\\ &= \boldsymbol{Y}^T\left(\mathbb{E}_{\boldsymbol{Z}_1,\ldots,\boldsymbol{Z}_m}\left[\left(\tau\boldsymbol{D}\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{D}^T+\boldsymbol{I}_n\right)^{-1}\right]\right)^2\boldsymbol{Y}..\end{aligned}$$

Therefore, we need to solve the following optimization problem to learn $\tau,\tilde{\tau}_1',\ldots,\tilde{\tau}_m'$:

$$\begin{gathered}\text{minimize } R(\tau,\tilde{\tau}_1',\ldots,\tilde{\tau}_m')\\ \text{subject to } \tau>0,\quad \tilde{\tau}_i'>0,\quad i=1,\ldots,m.\end{gathered}$$

Since this is an optimization problem with inequality constraints, we reparameterize the problem to make it unconstrained. Let

$$\tau = e^{\lambda},\quad \tilde{\tau}_i' = e^{\lambda_i},\quad i=1,\ldots,m.$$

Then the above optimization problem becomes

$$\text{minimize } R(e^{\lambda},e^{\lambda_1},\ldots,e^{\lambda_m})$$

For simplicity, we define

$$\boldsymbol{A} = \mathbb{E}_{\boldsymbol{U}_1,\ldots,\boldsymbol{U}_m}\left[\left(e^{\lambda}\phi^{-1}\boldsymbol{U}\boldsymbol{U}^T+\boldsymbol{I}_n\right)^{-1}\right] = \mathbb{E}_{\boldsymbol{Z}_1,\ldots,\boldsymbol{Z}_m}\left[\left(e^{\lambda}\boldsymbol{D}\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{D}^T+\boldsymbol{I}_n\right)^{-1}\right]$$

Then we have

$$
\begin{aligned}
\frac{\partial \boldsymbol{A}^2}{\partial \lambda} &= \frac{\partial \boldsymbol{A}}{\partial \lambda}\boldsymbol{A} + \boldsymbol{A}\frac{\partial \boldsymbol{A}}{\partial \lambda} \\
&= -e^{\lambda}\mathbb{E}\left[\left(e^{\lambda}\boldsymbol{D}\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{D}^T + \boldsymbol{I}_n\right)^{-1}\boldsymbol{D}\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{D}^T\left(e^{\lambda}\boldsymbol{D}\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{D}^T + \boldsymbol{I}_n\right)^{-1}\right]\boldsymbol{A}- \\
&\qquad e^{\lambda}\boldsymbol{A}\mathbb{E}\left[\left(e^{\lambda}\boldsymbol{D}\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{D}^T + \boldsymbol{I}_n\right)^{-1}\boldsymbol{D}\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{D}^T\left(e^{\lambda}\boldsymbol{D}\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{D}^T + \boldsymbol{I}_n\right)^{-1}\right]; \\
\frac{\partial \boldsymbol{A}}{\partial \lambda_i} &= \int\cdots\int\left(e^{\lambda}\phi^{-1}\boldsymbol{U}\boldsymbol{U}^T + \boldsymbol{I}_n\right)^{-1}\frac{\partial}{\partial \lambda_i}\left(\prod_{i=1}^{m}(2\pi\phi)^{-\frac{n}{2}}|e^{\lambda_i}\boldsymbol{\Sigma}' + \boldsymbol{I}_n|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2\phi}\boldsymbol{u}_i^T(e^{\lambda_i}\boldsymbol{\Sigma}' + \boldsymbol{I}_n)^{-1}\boldsymbol{u}_i\right\}\right)\mathrm{d}\boldsymbol{u}_1\cdots\mathrm{d}\boldsymbol{u}_m
\end{aligned}
$$

We need to calculate the derivative in the integrand. First we denote

$$
\Delta(\lambda_1,\ldots,\lambda_m) = \prod_{i=1}^{m}(2\pi\phi)^{-\frac{n}{2}}|e^{\lambda_i}\boldsymbol{\Sigma}' + \boldsymbol{I}_n|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2\phi}\boldsymbol{u}_i^T(e^{\lambda_i}\boldsymbol{\Sigma}' + \boldsymbol{I}_n)^{-1}\boldsymbol{u}_i\right\}
$$

Then we have

$$
\begin{aligned}
\frac{\partial \Delta}{\partial \lambda_i} &= \frac{\partial}{\partial \lambda_i}\exp\left\{\sum_{i=1}^{m}\left(-\frac{n}{2}\log(2\pi\phi) - \frac{1}{2}\log|e^{\lambda_i}\boldsymbol{\Sigma}' + \boldsymbol{I}_n| - \frac{1}{2\phi}\boldsymbol{u}_i^T(e^{\lambda_i}\boldsymbol{\Sigma}' + \boldsymbol{I}_n)^{-1}\boldsymbol{u}_i\right)\right\} \\
&= \Delta\left(-\frac{1}{2}e^{\lambda_i}\mathrm{tr}\left[(e^{\lambda_i}\boldsymbol{\Sigma}' + \boldsymbol{I}_n)^{-1}\boldsymbol{\Sigma}'\right] + \frac{1}{2\phi}e^{\lambda_i}\boldsymbol{u}_i^T(e^{\lambda_i}\boldsymbol{\Sigma}' + \boldsymbol{I}_n)^{-1}\boldsymbol{\Sigma}'(e^{\lambda_i}\boldsymbol{\Sigma}' + \boldsymbol{I}_n)^{-1}\boldsymbol{u}_i\right)
\end{aligned}
$$

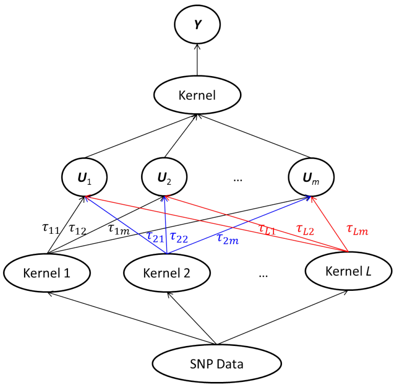
and hence

$$
\begin{aligned}
\frac{\partial \boldsymbol{A}}{\partial \lambda_i} &= -\frac{1}{2}e^{\lambda_i}\mathbb{E}\left[\left(e^{\lambda}\phi^{-1}\boldsymbol{U}\boldsymbol{U}^T + \boldsymbol{I}_n\right)^{-1}\mathrm{tr}\left((e^{\lambda_i}\boldsymbol{\Sigma}' + \boldsymbol{I}_n)^{-1}\boldsymbol{\Sigma}'\right)\right] + \\
&\qquad \frac{1}{2\phi}e^{\lambda_i}\mathbb{E}\left[\left(e^{\lambda}\phi^{-1}\boldsymbol{U}\boldsymbol{U}^T + \boldsymbol{I}_n\right)^{-1}\boldsymbol{U}_i^T(e^{\lambda_i}\boldsymbol{\Sigma}' + \boldsymbol{I}_n)^{-1}\boldsymbol{\Sigma}'(e^{\lambda_i}\boldsymbol{\Sigma}' + \boldsymbol{I}_n)^{-1}\boldsymbol{U}_i\right] \\
&= -\frac{1}{2}e^{\lambda_i}\mathbb{E}\left[\left(e^{\lambda}\boldsymbol{D}\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{D}^T + \boldsymbol{I}_n\right)^{-1}\mathrm{tr}\left((e^{\lambda_i}\boldsymbol{\Sigma}' + \boldsymbol{I}_n)^{-1}\boldsymbol{\Sigma}'\right)\right] + \\
&\qquad \frac{1}{2}e^{\lambda_i}\mathbb{E}\left[\left(e^{\lambda}\boldsymbol{D}\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{D}^T + \boldsymbol{I}_n\right)^{-1}\boldsymbol{Z}_i^T(e^{\lambda_i}\boldsymbol{\Sigma}' + \boldsymbol{I}_n)^{\frac{1}{2}}(e^{\lambda_i}\boldsymbol{\Sigma}' + \boldsymbol{I}_n)^{-1}\boldsymbol{\Sigma}'(e^{\lambda_i}\boldsymbol{\Sigma}' + \boldsymbol{I}_n)^{-1}(e^{\lambda_i}\boldsymbol{\Sigma}' + \boldsymbol{I}_n)^{\frac{1}{2}}\boldsymbol{Z}_i\right] \\
&= -\frac{1}{2}e^{\lambda_i}\mathbb{E}\left[\left(e^{\lambda}\boldsymbol{D}\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{D}^T + \boldsymbol{I}_n\right)^{-1}\mathrm{tr}\left((e^{\lambda_i}\boldsymbol{\Sigma}' + \boldsymbol{I}_n)^{-1}\boldsymbol{\Sigma}'\right)\right] + \\
&\qquad \frac{1}{2}e^{\lambda_i}\mathbb{E}\left[\left(e^{\lambda}\boldsymbol{D}\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{D}^T + \boldsymbol{I}_n\right)^{-1}\boldsymbol{Z}_i^T(e^{\lambda_i}\boldsymbol{\Sigma}' + \boldsymbol{I}_n)^{-\frac{1}{2}}\boldsymbol{\Sigma}'(e^{\lambda_i}\boldsymbol{\Sigma}' + \boldsymbol{I}_n)^{-\frac{1}{2}}\boldsymbol{Z}_i\right] \\
&= \frac{1}{2}e^{\lambda_i}\mathbb{E}\left[\left(e^{\lambda}\boldsymbol{D}\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{D}^T + \boldsymbol{I}_n\right)^{-1}\left(\boldsymbol{Z}_i^T(e^{\lambda_i}\boldsymbol{\Sigma}' + \boldsymbol{I}_n)^{-\frac{1}{2}}\boldsymbol{\Sigma}'(e^{\lambda_i}\boldsymbol{\Sigma}' + \boldsymbol{I}_n)^{-\frac{1}{2}}\boldsymbol{Z}_i - \mathrm{tr}\left((e^{\lambda_i}\boldsymbol{\Sigma}' + \boldsymbol{I}_n)^{-1}\boldsymbol{\Sigma}'\right)\right)\right]
\end{aligned}
$$

Therefore, the gradient of R with respect to λ and λ_i, $i = 1,\ldots,m$ can be obtained as follow:

$$
\begin{aligned}
\frac{\partial R}{\partial \lambda} &= \boldsymbol{Y}^T\left(\frac{\partial \boldsymbol{A}}{\partial \lambda}\boldsymbol{A} + \boldsymbol{A}\frac{\partial \boldsymbol{A}}{\partial \lambda}\right)\boldsymbol{Y} \\
\frac{\partial R}{\partial \lambda_i} &= \boldsymbol{Y}^T\left(\frac{\partial \boldsymbol{A}}{\partial \lambda_i}\boldsymbol{A} + \boldsymbol{A}\frac{\partial \boldsymbol{A}}{\partial \lambda_i}\right)\boldsymbol{Y}.
\end{aligned}
$$

2 Single Layer Model with Multiple Input Kernels



The basic structure of the single layer model with multiple kernels is shown in the following figure. The only difference here is that the covariance matrix of the latent variable $\boldsymbol{U}_i$ depends on several kernel matrices. Specifically, consider that the phenotype $\boldsymbol{Y}$ is modeled as a random effect model: given $\boldsymbol{U}_1, \dots, \boldsymbol{U}_m$,

$$\boldsymbol{Y} = \boldsymbol{a} + \boldsymbol{\epsilon}, \quad \text{where } \boldsymbol{a} \sim \mathcal{N}_n(\boldsymbol{0}, \tau \boldsymbol{\Sigma}(\boldsymbol{U}_1, \dots, \boldsymbol{U}_m)),$$

that is the covariance matrix of the random effect $\boldsymbol{a}$ depends on latent variables $\boldsymbol{U}_1, \dots, \boldsymbol{U}_m$. Moreover, the latent variable $\boldsymbol{U}_i$ is modeled using another random effect model

$$\boldsymbol{U}_i = \boldsymbol{a}'_i + \boldsymbol{\epsilon}, \quad \text{where } \boldsymbol{a}'_i \sim \mathcal{N}_n\left(\boldsymbol{0}, \sum_{j=1}^{L} \tau_{ji} \boldsymbol{K}_j\right)$$

Using the same arguments as in the single layer model with one input kernel, we can know that the best predictor for $\boldsymbol{a}$ is given by

$$\hat{\boldsymbol{a}} = \mathbb{E}[\boldsymbol{a}|\boldsymbol{Y}] = \mathbb{E}_{\boldsymbol{U}_1,\dots,\boldsymbol{U}_m}\left[\mathbb{E}\left(\boldsymbol{a}|\boldsymbol{Y}, \boldsymbol{U}_1, \dots, \boldsymbol{U}_m\right)\right] = \mathbb{E}_{\boldsymbol{U}_1,\dots,\boldsymbol{U}_m}\left[\tilde{\tau}\boldsymbol{\Sigma}(\tilde{\tau}\boldsymbol{\Sigma} + \boldsymbol{I}_n)^{-1}\right]\boldsymbol{Y},$$

where $\tilde{\tau} = \phi^{-1}\tau$. Still define $\boldsymbol{U} = [\boldsymbol{U}_1, \dots, \boldsymbol{U}_m]$, then if $\boldsymbol{U}$ is given and $\Sigma(\boldsymbol{U}_1, \dots, \boldsymbol{U}_m)$ is defined using product kernel, we have

$$\boldsymbol{\Sigma}(\boldsymbol{U}_1, \dots, \boldsymbol{U}_m) = \boldsymbol{U}\boldsymbol{U}^T.$$

Hence, the predicted response $\boldsymbol{Y}$ is given by

$$\hat{\boldsymbol{Y}} = \hat{\boldsymbol{a}} = \mathbb{E}_{\boldsymbol{U}_1,\dots,\boldsymbol{U}_m}\left[\tilde{\tau}\boldsymbol{U}\boldsymbol{U}^T\left(\tilde{\tau}\boldsymbol{U}\boldsymbol{U}^T + \boldsymbol{I}_n\right)^{-1}\right]\boldsymbol{Y}$$

and the loss function, i.e., the prediction error can be obtained similar as before:

$$R(\tau, \tilde{\tau}_{11}, \dots, \tilde{\tau}_{1m}, \dots, \tilde{\tau}_{L1}, \dots, \tilde{\tau}_{Lm}) = \boldsymbol{Y}^T \boldsymbol{A}^2 \boldsymbol{Y},$$

where $\tilde{\tau}_{ji} = \tau_{ji}\phi^{-1}$, $i = 1, \ldots, m$; $j = 1, \ldots, L$ and

$$\boldsymbol{A} = \mathbb{E}_{\boldsymbol{U}_1,\ldots,\boldsymbol{U}_m}\left[\left(\tilde{\tau}\boldsymbol{U}\boldsymbol{U}^T + \boldsymbol{I}_n\right)^{-1}\right] = \mathbb{E}_{\boldsymbol{Z}_1,\ldots,\boldsymbol{Z}_m}\left[\left(\tau\boldsymbol{D}\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{D}^T + \boldsymbol{I}_n\right)^{-1}\right]$$

with

$$\boldsymbol{D} = \left[\left(\sum_{j=1}^L \tilde{\tau}_{j1}\boldsymbol{K}_j + \boldsymbol{I}_n\right)^{\frac{1}{2}} \quad \cdots \quad \left(\sum_{j=1}^L \tilde{\tau}_{jm}\boldsymbol{K}_j + \boldsymbol{I}_n\right)^{\frac{1}{2}}\right], \quad \boldsymbol{Z} = \begin{bmatrix} \boldsymbol{Z}_1 & & \\ & \ddots & \\ & & \boldsymbol{Z}_m \end{bmatrix}$$

Due to the positive constraints on the parameters need to be learned, we similarly reparameterized the variance components as follows:

$$\tau = e^{\lambda}, \quad \tau_{ji} = e^{\lambda_{ji}}, \quad i = 1, \ldots, m; \quad j = 1, \ldots, L.$$

Hence, we need to solve the optimization problem:

$$\text{minimize } R(e^{\lambda}, e^{\lambda_{11}}, \ldots, e^{\lambda_{1m}}, \ldots, e^{\lambda_{L1}}, \ldots, e^{\lambda_{Lm}}) = \boldsymbol{Y}^T\boldsymbol{A}^2\boldsymbol{Y},$$

where

$$\boldsymbol{A} = \mathbb{E}_{\boldsymbol{U}_1,\ldots,\boldsymbol{U}_m}\left[\left(e^{\lambda}\phi^{-1}\boldsymbol{U}\boldsymbol{U}^T + \boldsymbol{I}_n\right)^{-1}\right] = \mathbb{E}_{\boldsymbol{Z}_1,\ldots,\boldsymbol{Z}_m}\left[\left(e^{\lambda}\boldsymbol{D}\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{D}^T + \boldsymbol{I}_n\right)^{-1}\right].$$

Then by similar reasoning as in the single layer model with one input kernel, we have

$$\begin{aligned}
\frac{\partial\boldsymbol{A}^2}{\partial\lambda} &= \frac{\partial\boldsymbol{A}}{\partial\lambda}\boldsymbol{A} + \boldsymbol{A}\frac{\partial\boldsymbol{A}}{\partial\lambda} \\
&= -e^{\lambda}\mathbb{E}\left[\left(e^{\lambda}\boldsymbol{D}\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{D}^T + \boldsymbol{I}_n\right)^{-1}\boldsymbol{D}\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{D}^T\left(e^{\lambda}\boldsymbol{D}\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{D}^T + \boldsymbol{I}_n\right)^{-1}\right]\boldsymbol{A} - \\
&\qquad e^{\lambda}\boldsymbol{A}\mathbb{E}\left[\left(e^{\lambda}\boldsymbol{D}\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{D}^T + \boldsymbol{I}_n\right)^{-1}\boldsymbol{D}\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{D}^T\left(e^{\lambda}\boldsymbol{D}\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{D}^T + \boldsymbol{I}_n\right)^{-1}\right]; \\
\frac{\partial\boldsymbol{A}}{\partial\lambda_{ji}} &= \frac{1}{2}e^{\lambda_{ji}}\mathbb{E}\left[\left(e^{\lambda}\boldsymbol{D}\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{D}^T + \boldsymbol{I}_n\right)^{-1}\boldsymbol{Z}_i^T\left(\sum_{j=1}^L e^{\lambda_{ji}}\boldsymbol{K}_j + \boldsymbol{I}_n\right)^{-\frac{1}{2}}\boldsymbol{K}_j\left(\sum_{j=1}^L e^{\lambda_{ji}}\boldsymbol{K}_j + \boldsymbol{I}_n\right)^{-\frac{1}{2}}\boldsymbol{Z}_i\right] \\
&\qquad - \frac{1}{2}e^{\lambda_{ji}}\mathbb{E}\left[\left(e^{\lambda}\boldsymbol{D}\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{D}^T + \boldsymbol{I}_n\right)^{-1}\operatorname{tr}\left(\left(\sum_{j=1}^L e^{\lambda_{ji}}\boldsymbol{K}_j + \boldsymbol{I}_n\right)^{-1}\boldsymbol{K}_j\right)\right]
\end{aligned}$$

Therefore, the gradient of R with respect to λ and λ_{ji}, $i = 1, \ldots, m$; $j = 1, \ldots, L$ can be obtained as follow:

$$\begin{aligned}
\frac{\partial R}{\partial \lambda} &= \boldsymbol{Y}^T \left(\frac{\partial \boldsymbol{A}}{\partial \lambda} \boldsymbol{A} + \boldsymbol{A} \frac{\partial \boldsymbol{A}}{\partial \lambda} \right) \boldsymbol{Y} \\
\frac{\partial R}{\partial \lambda_{ji}} &= \boldsymbol{Y}^T \left(\frac{\partial \boldsymbol{A}}{\partial \lambda_{ji}} \boldsymbol{A} + \boldsymbol{A} \frac{\partial \boldsymbol{A}}{\partial \lambda_{ji}} \right) \boldsymbol{Y}.
\end{aligned}$$