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

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

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

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

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 $\boldsymbol{Z}_1, \dots, \boldsymbol{Z}_m \sim \mathcal{N}_n(\boldsymbol{0}, \boldsymbol{I}_n)$. Then we get

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

and hence

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

which implies that

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

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

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

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

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

Then the above optimization problem becomes

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

For simplicity, we define

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

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

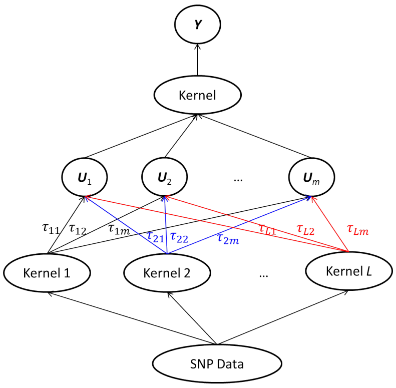
$$
\begin{aligned}
\frac{\partial \boldsymbol{A}}{\partial \lambda_i} &= -\frac{1}{2}e^{\lambda_i}\mathbb{E}\left[\left(e^{\lambda}\phi^{-1}\boldsymbol{U}\boldsymbol{U}^T + \boldsymbol{I}_n\right)^{-1}\mathrm{tr}\left((e^{\lambda_i}\boldsymbol{\Sigma}' + \boldsymbol{I}_n)^{-1}\boldsymbol{\Sigma}'\right)\right] + \\
&\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 $\boldsymbol{U}_1, \ldots, \boldsymbol{U}_m$,



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

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

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

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

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

where $\tilde{\tau} = \phi^{-1}\tau$. Still define $\boldsymbol{U} = [\boldsymbol{U}_1, \ldots, \boldsymbol{U}_m]$, then if $\boldsymbol{U}$ is given and $\Sigma(\boldsymbol{U}_1, \ldots, \boldsymbol{U}_m)$ 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

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

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

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

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

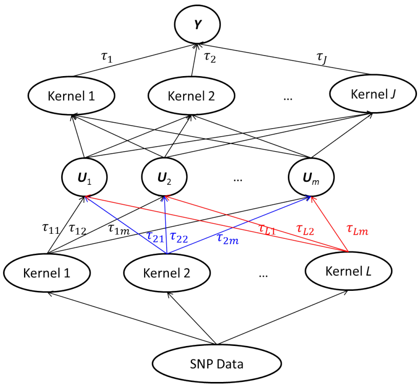
$$\begin{aligned}
\frac{\partial \boldsymbol{A}^2}{\partial \lambda} &= \frac{\partial \boldsymbol{A}}{\partial \lambda}\boldsymbol{A} + \boldsymbol{A}\frac{\partial \boldsymbol{A}}{\partial \lambda} \\
&= -e^{\lambda}\mathbb{E}\left[\left(e^{\lambda}\boldsymbol{D}\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{D}^T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

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:

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

* Single Layer Model with Multiple Kernel Inputs (Modification)

Consider a more flexible structure of the single layer model with multiple kernels, which is shown in the following figure.



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

$$\boldsymbol{Y} = \boldsymbol{a} + \boldsymbol{\epsilon}, \quad \text{where } \boldsymbol{a} \sim \mathcal{N}_n \left(\boldsymbol{0}, \sum_{j=1}^{J} \tau_j \boldsymbol{K}_j (\boldsymbol{U}_1, \ldots, \boldsymbol{U}_n) \right)$$

The latent variables $\boldsymbol{U}_i$ is modeled using another random effect model

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

The best predictor for $\boldsymbol{a}$ and hence for $\boldsymbol{Y}$ is given by

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

The prediction error is then

$$R = \boldsymbol{Y}^T \left(\mathbb{E} \left[\left(\sum_{j=1}^{J} \tilde{\tau}_j \boldsymbol{K}_j (\boldsymbol{U}) + \boldsymbol{I}_n \right)^{-1} \right] \right)^2 \boldsymbol{Y}.$$

- Sampling Method for Calculating Derivatives

  It is easy to see that if we let $\boldsymbol{A} = \mathbb{E}\left[\left(\sum_{j=1}^{J} \tilde{\tau}_j \boldsymbol{K}_j(\boldsymbol{U}) + \boldsymbol{I}_n\right)^{-1}\right]$, $\tilde{\tau}_j = e^{\lambda_j}, \phi = e^{\varphi}$ and $\tilde{\tau}_{li} = e^{\lambda_{li}}$, we can get

$$
\begin{aligned}
\frac{\partial \boldsymbol{A}}{\partial \lambda_j} &= -e^{\lambda_j}\mathbb{E}\left[\left(\sum_{j=1}^{J} e^{\lambda_j}\boldsymbol{K}_j(\boldsymbol{U}) + \boldsymbol{I}_n\right)^{-1} \boldsymbol{K}_j(\boldsymbol{U}) \left(\sum_{j=1}^{J} e^{\lambda_j}\boldsymbol{K}_j(\boldsymbol{U}) + \boldsymbol{I}_n\right)^{-1}\right] \\
\frac{\partial \boldsymbol{A}}{\partial \lambda_{li}} &= -\frac{1}{2}e^{\lambda_{li}}\mathbb{E}\left[\left(\sum_{j=1}^{J} e^{\lambda_j}\boldsymbol{K}_j(\boldsymbol{U}) + \boldsymbol{I}_n\right)^{-1} \operatorname{tr}\left(\left(\sum_{l=1}^{L} e^{\lambda_{li}}\boldsymbol{K}_l + \boldsymbol{I}_n\right)^{-1}\boldsymbol{K}_l\right)\right] \\
&\quad + \frac{1}{2e^{\varphi}}e^{\lambda_{li}}\mathbb{E}\left[\left(\sum_{j=1}^{J} e^{\lambda_j}\boldsymbol{K}_j(\boldsymbol{U}) + \boldsymbol{I}_n\right)^{-1} \boldsymbol{U}_i^T \left(\sum_{l=1}^{L} e^{\lambda_{li}}\boldsymbol{K}_l + \boldsymbol{I}_n\right)^{-1} \boldsymbol{K}_l \left(\sum_{l=1}^{L} e^{\lambda_{li}}\boldsymbol{K}_l + \boldsymbol{I}_n\right)^{-1} \boldsymbol{U}_i\right] \\
&= -\frac{1}{2}e^{\lambda_{li}}\mathbb{E}\left[\left(\sum_{j=1}^{J} e^{\lambda_j}\boldsymbol{K}_j(\boldsymbol{U}) + \boldsymbol{I}_n\right)^{-1} \operatorname{tr}\left(\left(\sum_{l=1}^{L} e^{\lambda_{li}}\boldsymbol{K}_l + \boldsymbol{I}_n\right)^{-1}\boldsymbol{K}_l\right)\right] \\
&\quad + \frac{1}{2e^{\varphi}}e^{\lambda_{li}}\mathbb{E}\left[\left(\sum_{j=1}^{J} e^{\lambda_j}\boldsymbol{K}_j(\boldsymbol{U}) + \boldsymbol{I}_n\right)^{-1} \operatorname{tr}\left(\left(\sum_{l=1}^{L} e^{\lambda_{li}}\boldsymbol{K}_l + \boldsymbol{I}_n\right)^{-1} \boldsymbol{K}_l \left(\sum_{l=1}^{L} e^{\lambda_{li}}\boldsymbol{K}_l + \boldsymbol{I}_n\right)^{-1} \boldsymbol{U}_i\boldsymbol{U_i}^T\right)\right] \\
&= -\frac{1}{2}e^{\lambda_{li}}\mathbb{E}\left[\left(\sum_{j=1}^{J} e^{\lambda_j}\boldsymbol{K}_j(\boldsymbol{U}) + \boldsymbol{I}_n\right)^{-1}\right. \\
&\qquad \left.\operatorname{tr}\left(\left(\sum_{l=1}^{L} e^{\lambda_{li}}\boldsymbol{K}_l + \boldsymbol{I}_n\right)^{-1} \boldsymbol{K}_l \left(\sum_{l=1}^{L} e^{\lambda_{li}}\boldsymbol{K}_l + \boldsymbol{I}_n\right)^{-1} \left(\sum_{l=1}^{L} e^{\lambda_{li}}\boldsymbol{K}_l + \boldsymbol{I}_n - e^{-\varphi}\boldsymbol{U}_i\boldsymbol{U}_i^T\right)\right)\right] \\
\frac{\partial \boldsymbol{A}}{\partial \varphi} &= -\frac{mn}{2}\boldsymbol{A} + \frac{1}{2}e^{-\varphi}\mathbb{E}\left[\left(\sum_{j=1}^{J} e^{\lambda_j}\boldsymbol{K}_j(\boldsymbol{U}) + \boldsymbol{I}_n\right)^{-1} \boldsymbol{U}_i^T \left(\sum_{l=1}^{L} e^{\lambda_{li}}\boldsymbol{K}_l + \boldsymbol{I}_n\right)^{-1} \boldsymbol{U}_i\right] \\
&= -\frac{1}{2}\mathbb{E}\left[\left(\sum_{j=1}^{J} e^{\lambda_j}\boldsymbol{K}_j(\boldsymbol{U}) + \boldsymbol{I}_n\right)^{-1} \left(mn - e^{-\varphi}\sum_{i=1}^{m}\operatorname{tr}\left(\left(\sum_{l=1}^{L} e^{\lambda_{li}}\boldsymbol{K}_l + \boldsymbol{I}_n\right)^{-1} \boldsymbol{U}_i\boldsymbol{U}_i^T\right)\right)\right]
\end{aligned}
$$

- Some theoretical approximations

  In particular, if $J = 1$, $\tau_{li} = \xi_l$ for $i = 1, \ldots, m$ and $K_1(\boldsymbol{U}) = \boldsymbol{U}\boldsymbol{U}^T$, then using the spectral decomposition of $\boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{G}\boldsymbol{L}\boldsymbol{G}^T$ to get

$$
\begin{aligned}
\mathbb{E}\left[\left(\tilde{\tau}\boldsymbol{U}\boldsymbol{U}^T + \boldsymbol{I}_n\right)^{-1}\right] &= \mathbb{E}\left[\left(\tilde{\tau}\boldsymbol{G}\boldsymbol{L}\boldsymbol{G}^T + \boldsymbol{I}_n\right)^{-1}\right] \\
&= \mathbb{E}\left[\boldsymbol{G}\left(\tilde{\tau}\boldsymbol{L} + \boldsymbol{I}_n\right)^{-1}\boldsymbol{G}^T\right] \\
&= \sum_{i=1}^{n}\mathbb{E}\left[\frac{\boldsymbol{g}_i\boldsymbol{g}_i^T}{1 + \tilde{\tau}l_i}\right]
\end{aligned}
$$

  Now we can apply the following theorem

Theorem 1 (Theorem 13.5.1, Anderson (2003))**.** Suppose $n\boldsymbol{S} \sim \mathcal{W}_p(\boldsymbol{\Sigma}, n)$. Consider the spectral decomposition of $\boldsymbol{\Sigma}$ and $\boldsymbol{S}$

$$\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^T. \quad \boldsymbol{S} = \boldsymbol{G}\boldsymbol{L}\boldsymbol{G}^T,$$

where $\lambda_1 > \lambda_2 > \cdots > \lambda_p$, $l_1 \geq l_2 \geq \cdots \geq l_p$, $\gamma_{1i} \geq 0$, g_{1i}, $i = 1, \ldots, p$. Define $\boldsymbol{C} = \sqrt{n}(\boldsymbol{G} - \boldsymbol{\Gamma})$ and diagonal matrix $\boldsymbol{D} = \sqrt{n}(\boldsymbol{L} - \boldsymbol{\Lambda})$. Then the limiting distribution of $\boldsymbol{C}$ and $\boldsymbol{D}$ is normal with $\boldsymbol{C}$ and $\boldsymbol{D}$ independent and the diagonal elements of $\boldsymbol{D}$ are independent. The diagonal elements $d_i \xrightarrow{d} \mathcal{N}(0, 2\lambda_i^2)$. The covariance matrix of $\boldsymbol{c}_i$, in the limiting distribution of $\boldsymbol{C} = [\boldsymbol{c}_1, \ldots, \boldsymbol{c}_p]$ is

$$\text{Var}[\boldsymbol{c}_i] = \sum_{k=1, k\neq i}^{p} \frac{\lambda_i \lambda_k}{(\lambda_i - \lambda_k)^2} \boldsymbol{\gamma}_k \boldsymbol{\gamma}_k^T.$$

In our case, we know that $\boldsymbol{U}_1, \ldots, \boldsymbol{U}_n \sim \mathcal{N}_n(\boldsymbol{0}, \sum_{l=1}^{L} \xi_l \boldsymbol{K}_l + \phi \boldsymbol{I}_n)$ and hence

$$\boldsymbol{U}\boldsymbol{U}^T \sim \mathcal{W}_n(\sum_{l=1}^{L} \xi_l \boldsymbol{K}_l + \phi \boldsymbol{I}_n, n).$$

Define $\boldsymbol{\Psi} = \sum_{l=1}^{L} \xi_l \boldsymbol{K}_l + \phi \boldsymbol{I}_n$ and consider the spectral decomposition of $\boldsymbol{\Psi}$, $\boldsymbol{\Psi} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^T$. Then Theorem 1 shows that

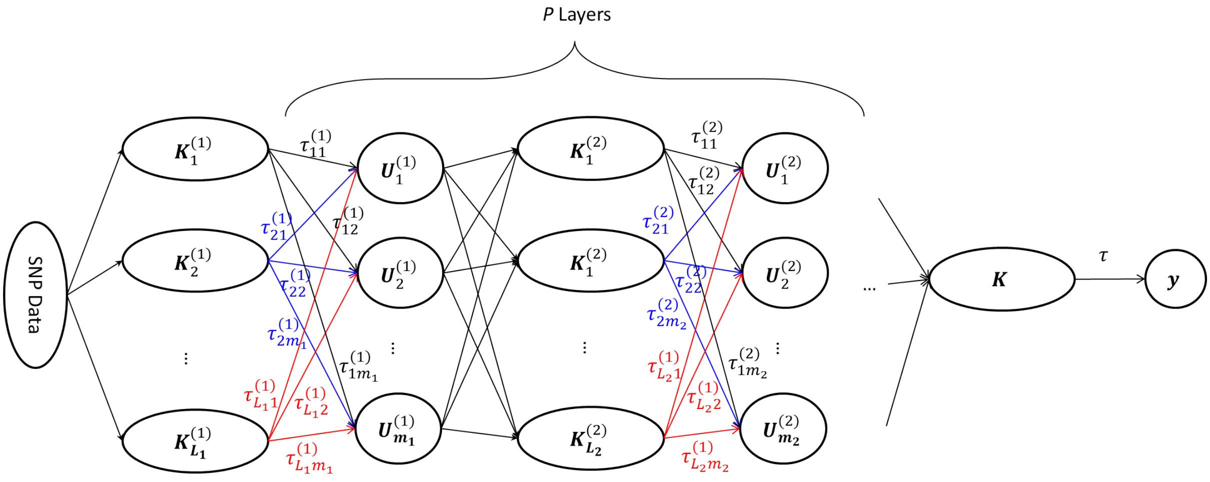
$$\sqrt{n}(l_i - \lambda_i) \xrightarrow{d} \mathcal{N}(0, 2\lambda_i^2)$$
$$\sqrt{n}(\boldsymbol{g}_i - \boldsymbol{\gamma}_i) \xrightarrow{d} \mathcal{N}_n\left(\boldsymbol{0}, \boldsymbol{\Delta}_i\right),$$

where

$$\boldsymbol{\Delta}_i = \sum_{k=1, k\neq i}^{n} \frac{\lambda_i \lambda_k}{(\lambda_i - \lambda_k)^2} \boldsymbol{\gamma}_k \boldsymbol{\gamma}_k^T.$$

3 Multiple Layer Model with Multiple Kernel Inputs

We further extend the model to the case in which there are multiple layers and in each layer, there are multiple kernels, either from the input data or from the hidden units in the previous layer. The basic structure of the model is shown below



Suppose that there are P hidden layers and let the phenotype $\boldsymbol{Y}$ be modeled as a random effect model, given $\boldsymbol{U}_1^{(P-1)}, \ldots, \boldsymbol{U}_{m_{P-1}}^{(P-1)}$,

$$\boldsymbol{Y} = \boldsymbol{a}^{(P)} + \boldsymbol{\epsilon}, \quad \text{where } \boldsymbol{a}^{(P)} \sim \mathcal{N}_n\left(\boldsymbol{0}, \tau\boldsymbol{\Sigma}\left(\boldsymbol{U}_1^{(P-1)}, \ldots, \boldsymbol{U}_{m_{P-1}}^{(P-1)}\right)\right),$$

where the latent variables $\boldsymbol{U}_i^{(k)}$'s are modeled by other random effect models. Specifically, we define $\boldsymbol{U}^{(k)} = [\boldsymbol{U}_1^{(k)}, \ldots, \boldsymbol{U}_{m_k}^{(k)}]$ and $\mathcal{F}_k = \sigma\left\{\boldsymbol{U}_1^{(k)}, \ldots, \boldsymbol{U}_{m_k}^{(k)}\right\}$ for $k = 1, \ldots, P-1$, then for each $k \in \{2, \ldots, P-1\}$, given $\mathcal{F}_{k-1}$, $\boldsymbol{U}_i^{(k)}$ is modeled as follow:

$$\boldsymbol{U}_i^{(k)} = \boldsymbol{a}_i^{(k)} + \boldsymbol{\epsilon}, \quad \text{where } \boldsymbol{a}_i^{(k)} \sim \mathcal{N}_n\left(\boldsymbol{0}, \sum_{l=1}^{L_k} \tau_{li}^{(k)} \boldsymbol{K}_l^{(k)}\left(\boldsymbol{U}_1^{(k-1)}, \ldots, \boldsymbol{U}_{m_{k-1}}^{(k-1)}\right)\right), \quad i = 1, \ldots, m_i;$$

and $\boldsymbol{U}_i^{(1)}$ is modeled by

$$\boldsymbol{U}_i^{(1)} = \boldsymbol{a}_i^{(1)} + \boldsymbol{\epsilon}, \quad \text{where } \boldsymbol{a}_i^{(1)} \sim \mathcal{N}_n\left(\boldsymbol{0}, \sum_{l=1}^{L_1} \tau_{li}^{(1)} \boldsymbol{K}_l^{(1)}\right), \quad i = 1, \ldots, m_1.$$

When the product kernel is used for the final stage, the best prediction for $\boldsymbol{Y}$ is then given by

$$\hat{\boldsymbol{Y}} = \hat{\boldsymbol{a}}^{(P)} = \mathbb{E}\left[\boldsymbol{a}^{(P)}|\boldsymbol{Y}\right] = \mathbb{E}\left[\mathbb{E}\left[\boldsymbol{a}^{(P)}|\boldsymbol{Y}, \mathcal{F}_{P-1}\right]\right] = \mathbb{E}_{\boldsymbol{U}^{(P-1)}}\left[\tilde{\tau}\boldsymbol{U}^{(P-1)}\boldsymbol{U}^{(P-1)^T}\left(\tilde{\tau}\boldsymbol{U}^{(P-1)}\boldsymbol{U}^{(P-1)^T} + \boldsymbol{I}_n\right)^{-1}\right]\boldsymbol{Y}.$$

For notation simplicity, we denote $\boldsymbol{K}_l^{(k)}\left(\boldsymbol{U}_1^{(k-1)}, \ldots, \boldsymbol{U}_{m_{k-1}}^{(k-1)}\right)$ as $\boldsymbol{K}_l^{(k)}(\boldsymbol{U}^{(k-1)})$. Based on the discussion, we know that

$$
\begin{aligned}
\boldsymbol{U}_i^{(k)}|\mathcal{F}_{k-1} &\sim \mathcal{N}_n\left(\boldsymbol{0}, \phi\left(\boldsymbol{I}_n + \sum_{l=1}^{L_k} \tilde{\tau}_{li}^{(k)} \boldsymbol{K}_l^{(k)}\left(\boldsymbol{U}^{(k-1)}\right)\right)\right), \quad k=2,\ldots,P-1;\ i=1,\ldots,m_k,\\
\boldsymbol{U}_i^{(1)} &\sim \mathcal{N}_n\left(\boldsymbol{0}, \phi\left(\boldsymbol{I}_n + \sum_{l=1}^{L_1} \tilde{\tau}_{li}^{(1)} \boldsymbol{K}_l^{(1)}\right)\right), \quad i=1,\ldots,m_1.
\end{aligned}
$$

To learn all the parameters, we need to minimize the prediction error, which is given by

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

For simplicity, we define $\boldsymbol{A} = \mathbb{E}_{\boldsymbol{U}^{(P-1)}}\left[\left(e^{\lambda}\phi^{-1}\boldsymbol{U}^{(P-1)}\boldsymbol{U}^{(P-1)^T}+\boldsymbol{I}_n\right)^{-1}\right]$. Obviously, it is difficult or even impossible to obtain the analytic expression for the matrix $\boldsymbol{A}$.

$$
\begin{aligned}
\boldsymbol{U}_i^{(k)} &\overset{d}{=} \phi^{\frac{1}{2}}\left(\boldsymbol{I}_n + \sum_{l=1}^{L_k} e^{\lambda_{li}^{(k)}}\boldsymbol{K}_l^{(k)}(\boldsymbol{U}^{(k-1)})\right)^{\frac{1}{2}}\boldsymbol{Z}_i^{(k)} := \phi^{\frac{1}{2}}\boldsymbol{D}_i^{(k)}\boldsymbol{Z}_i^{(k)}, \quad k=2,\ldots,P-1\\
\boldsymbol{U}_i^{(1)} &\overset{d}{=} \phi^{\frac{1}{2}}\left(\boldsymbol{I}_n + \sum_{l=1}^{L_1} e^{\lambda_{li}^{(1)}}\boldsymbol{K}_l^{(1)}\right)^{\frac{1}{2}}\boldsymbol{Z}_i^{(1)} = \phi^{\frac{1}{2}}\boldsymbol{D}_i^{(1)}\boldsymbol{Z}_i^{(1)}.
\end{aligned}
$$

Define

$$
\boldsymbol{D}^{(k)} = \begin{bmatrix}\boldsymbol{D}_1^{(k)} & \cdots & \boldsymbol{D}_{m_k}^{(k)}\end{bmatrix}, \qquad \boldsymbol{Z}^{(k)} = \begin{bmatrix}\boldsymbol{Z}_1^{(k)} & & \\ & \ddots & \\ & & \boldsymbol{Z}_{m_k}^{(k)}\end{bmatrix},
$$

then, we have $\boldsymbol{U}^{(k)} = \phi^{\frac{1}{2}}\boldsymbol{D}^{(k)}\boldsymbol{Z}^{(k)}$ for $k=1,\ldots,P-1$. For the kth ($k \geq 2$) hidden layer in the network, the hidden units $\boldsymbol{U}_1^{(k)},\ldots,\boldsymbol{U}_{m_k}^{(k)}$ are obtained through the best prediction based on the kernel matrices constructed using $\boldsymbol{U}_1^{(k-1)},\ldots,\boldsymbol{U}_{m_{k-1}}^{(k-1)}$, that is, given $\mathcal{F}_{k-1}$,

$$
\hat{\boldsymbol{U}}_i^{(k)} = \mathbb{E}\left[\boldsymbol{a}_i^{(k)}|\boldsymbol{U}_i^{(k)}\right]
$$

$$
\frac{\partial \boldsymbol{A}}{\partial \lambda}
$$