# Supplementary Materials to "Transfer Learning for Ridge Regression with Random Coefficients in High-Dimensional Settings"

## Supplementary Materials

## Appendix A

The limiting values are $D, C$ and $F$ in Theorem **??** are given as

1. $D \to_{a.s.} \mathcal{D}$, where for $k = 1, \cdots K-1$,

$$\mathcal{D}_k = \rho_{kK}\sigma_k\sigma_K\alpha_k\alpha_K \left[1 - \frac{\lambda_k}{\gamma_k}\left\{\frac{1}{\lambda_k v_{F_{\gamma_k}}(-\lambda_k)}\right\}\right],$$

$$\mathcal{D}_K = \sigma_K^2\alpha_K^2 \left[1 - \frac{\lambda_k}{\gamma_K}\left\{\frac{1}{\lambda_K v_{F_{\gamma_K}}(-\lambda_K)}\right\}\right].$$

2. $C \to_{a.s.} \mathcal{C}$, where for $k = 1, \cdots, K$,

$$\mathcal{C}_{kk} = \sigma_k^2\alpha_k^2 \left[1 - 2\frac{\lambda_k}{\gamma_k}\left\{\frac{1}{\lambda_k v_{F_{\gamma_k}}(-\lambda_k)}\right\} + \frac{\lambda_k^2}{\gamma_k}\frac{v_{F_{\gamma_k}}(-\lambda_k) - \lambda_k v'_{F_{\gamma_k}}(-\lambda_k)}{[\lambda_k v_{F_{\gamma_k}}(-\lambda_k)]^2}\right],$$

and for $k, k' = 1, \cdots, K$ and $k \neq k'$,

$$\mathcal{C}_{kk'} = \rho_{kk'}\alpha_k\sigma_k\alpha_{k'}\sigma_{k'} \left[1 - \frac{\lambda_k}{\gamma_k}\left\{\frac{1}{\lambda_k v_{F_{\gamma_k}}(-\lambda_k)}\right\} - \frac{\lambda_{k'}}{\gamma_{k'}}\left\{\frac{1}{\lambda_{k'} v_{F_{\gamma_{k'}}}(-\lambda_{k'})}\right\} + \lambda_k\lambda_{k'}\mathcal{P}_{kk'}\right].$$

3. $F$ is a diagonal matrix, and $F \to_{a.s.} \mathcal{F}$, where

$$\mathcal{F}_{kk} = \sigma_k^2\gamma_k \left[\frac{1}{\gamma_k}\left\{\frac{1}{\lambda_k v_{F_{\gamma_k}}(-\lambda_k)} - 1\right\} - \lambda_k\frac{1}{\gamma_k}\frac{v_{F_{\gamma_k}}(-\lambda_k) - \lambda_k v'_{F_{\gamma_k}}(-\lambda_k)}{[\lambda_k v_{F_{\gamma_k}}(-\lambda_k)]^2}\right].$$

## Appendix B

### B.1 Technical Lemmas

**Lemma B.1.** *Under assumption **??** and assumption **??**; consider a $p \times p$ matrix $A^p$ whose entries $A^p_{ij}$ have finite $(8+\epsilon)$-th moment for some $\epsilon > 0$. Suppose further that $A^p$ is independent of $\beta_k$, $k, k' = 1, \cdots, K; k \neq k'$, we have as $p \to \infty$*

$$\beta_k^T A^p \beta_{k'} - \rho_{kk'}\alpha_k\alpha_{k'}\sigma_k\sigma_{k'}\,\mathrm{tr}(A^p)/p \to_{a.s.} 0$$

$$\beta_k^T A^p \beta_k - \alpha_k^2\sigma_k^2\,\mathrm{tr}(A^p)/p \to_{a.s.} 0$$

*Proof of Lemma B.1.* When $k = k'$, this lemma is the same as lemma C.3 in Dobriban & Wager (2018) and theorem 2 in Sheng & Dobriban (2020). When $k \neq k'$, the same results still holds trivially under the bounded moments condition of $A$. This result has already been used by theorem 3.1 and theorem 4.1 of Zhao & Zhu (2019). □

**Lemma B.2.** *Under the assumption* **??** *and* **??**, *recall the definition of sample covariance matrix* $\hat{\Sigma} = X^T X/n$ *and its companion* $\underline{\hat{\Sigma}} = XX^T/p$; *we have*

$$\begin{aligned}
\mathrm{tr}[(\hat{\Sigma} + \lambda I_p)^{-1}]/p &\to_{a.s.} m_{F_\gamma}(-\lambda) \\
\mathrm{tr}[(\underline{\hat{\Sigma}} + \lambda I_p)^{-1}]/n &\to_{a.s.} v_{F_\gamma}(-\lambda) \\
\mathrm{tr}[(\hat{\Sigma} + \lambda I_p)^{-2}]/p &\to_{a.s.} m'_{F_\gamma}(-\lambda) \\
\mathrm{tr}[(\underline{\hat{\Sigma}} + \lambda I_p)^{-2}]/n &\to_{a.s.} v'_{F_\gamma}(-\lambda) \\
\mathrm{tr}[(\hat{\Sigma} + \lambda I_p)^{-1}\Sigma]/p &\to_{a.s.} \frac{1}{\gamma}(\frac{1}{\lambda v_{F_\gamma}(-\lambda)} - 1) \\
\mathrm{tr}[(\hat{\Sigma} + \lambda I_p)^{-2}\Sigma]/p &\to_{a.s.} \frac{1}{\gamma}\frac{v_{F_\gamma}(-\lambda) - \lambda v'_{F_\gamma}(-\lambda)}{\lambda v_{F_\gamma}(-\lambda)}
\end{aligned}$$

*Proof of Lemma B.2.* The first four convergence statements follow from Marchenko & Pastur (1967) and Silverstein (1995). The convergence of last two trace terms are from lemma 2 of Ledoit & Péché (2011) and lemma 2.2. of Dobriban & Wager (2018). Their moment assumptions are satisfied given assumption **??**. □

### B.2 Proofs of Lemmas in the main text

*Proof of Lemma* **??**. In the first case, when $n_1 = \cdots = n_K = n$ so $\gamma_1 = \cdots = \gamma_K = \gamma$ and use $\lambda_1 = \cdots = \lambda_K = \lambda$, the limits of term $E_{kk'}$ has been found in the proof theorem 3 in Sheng & Dobriban (2020). When $\Sigma = I_p$, the term $E_{kk'}$ boils down to

$$E_{kk'} = \mathrm{tr}[(Z_k^T Z_k/n_k + \lambda_k I_p)^{-1}(Z_{k'}^T Z_{k'}/n_{k'} + \lambda_{k'} I_p)^{-1}]/p]$$

To prove the second case, we always have have

$$E_{kk'} \to_{a.s.} E_H \frac{1}{(x_k T + \lambda_k)(x_{k'} T + \lambda_{k'})}$$

Recall $H$ is the limiting population spectral distribution, and $x_k = x(\gamma_k, \lambda_k)$ is the fixed point solution to

$$1 - x_k = \gamma_k[1 - \lambda_k \int \frac{1}{x_k t + \lambda_k} dH(t)]$$

When $\Sigma = I_p$, $H$ only has a point mass on 1 so the expectation decomposes and

$$E_{kk'} \to_{a.s.} m_{F_{\gamma_k}}(-\lambda_k) m_{F_{\gamma_{k'}}}(-\lambda_{k'})$$

As an alternative proof when we have assumption **??**; we know $Z_k$ will be asymptoticly free from any bounded constant matrices, this is a standard result, ex. theorems 5.4.5 in Anderson et al. (2010). Further, we know sample covariances of the form $Z_k^T Z_k/n_k$ is asymptotically free from

$Z_{k'}^T Z_{k'}/n_{k'}$ Capitaine & Casalis (2004). Two arguments combined suggests that $(Z_k^T Z_k/n_k + \lambda_k I_p)^{-1}$ is asymptotically free from $(Z_{k'}^T Z_{k'}/n_{k'} + \lambda_{k'} I_p)^{-1}$ therefore,

$$E_{kk'} - m_{F_{\gamma_k}}(-\lambda_k) m_{F_{\gamma_{k'}}}(-\lambda_{k'}) \to_{a.s.} 0$$

For the third case, a slight generalization of Corollary 3.9 of Knowles & Yin (2017) tells us

$$\ell_1^T [(\hat{\Sigma}_k + \lambda_k I_p)^{-1} - \frac{1}{\lambda_k (1 + m_{F_{\gamma_k}}(-\lambda)\Sigma)}] \ell_2 \to_{a.s.} 0$$

where $\ell_1, \ell_2$ can be any continuous random vectors independent from $(\hat{\Sigma}_k + \lambda_k I_p)^{-1}$. We can decompose $E_{kk'}$ by

$$\underbrace{\operatorname{tr}[(\hat{\Sigma}_k + \lambda_k I_p)^{-1} (\hat{\Sigma}_{k'} + \lambda_{k'} I_p)^{-1}]/p}_{E_{kk'}} - \frac{1}{\lambda_k} \sum_{i=1}^{p} \ell_{1,i}^T (I_p + m_{F_{\gamma_k}}(-\lambda)\Sigma)^{-1} \ell_{2,i}/p \to_{a.s.} 0$$

where $\ell_{1,i}$ is $e_i$ with $1$ in its $i^{th}$ entry and $0$ else where; and $\ell_{2,i} := (\hat{\Sigma}_{k'} + \lambda_{k'} I_p)^{-1} e_i$. From now on, simplify the notation by using $m_k := m_{F_{\gamma_k}}(-\lambda_k)$ and $m_{k'} := m_{F_{\gamma_{k'}}}(-\lambda_{k'})$. Perform the similar trick to $\ell_{2,i}$, we have

$$E_{kk'} - \frac{1}{\lambda_k \lambda_{k'}} \operatorname{tr}((I_p + m_k \Sigma)^{-1} (I_p + m_{k'} \Sigma)^{-1})/p \to_{a.s.} 0$$

In addition

$$\begin{aligned} & \frac{1}{\lambda_k \lambda_{k'}} \operatorname{tr}((I_p + m_k \Sigma)^{-1} (I_p + m_{k'} \Sigma)^{-1})/p \\ = & \frac{1}{\lambda_k \lambda_{k'}} [1 - m_k \operatorname{tr}((I_p + m_k \Sigma)^{-1} \Sigma)/p - m_{k'} \operatorname{tr}((I_p + m_{k'} \Sigma)^{-1} \Sigma)/p \\ + & m_k m_{k'} \operatorname{tr}((I_p + m_k \Sigma)^{-1} \Sigma (I_p + m_{k'} \Sigma)^{-1} \Sigma)]/p \end{aligned}$$

where we used the matrix identity

$$(I_p + m_{F_{\gamma_k}}(-\lambda_k)\Sigma)^{-1} = I_p - m_{F_{\gamma_k}}(-\lambda_k)(I_p + m_k(-\lambda_k)\Sigma)^{-1}\Sigma$$

Each of the terms can be expressed in empirical quantities by

$$\operatorname{tr}((I_p + m_k \Sigma)^{-1} \Sigma)/p - E_H \frac{1}{m_k (1 + t m_k)} + E_H \frac{1}{m_k (1 + t m_k)} - \frac{1}{m_k} [1 - \lambda_k m_k] \to_{a.s.} 0$$

With the same techniques, we get

$$\begin{aligned} & \operatorname{tr}((I_p + m_k \Sigma)^{-1} \Sigma (I_p + m_{k'} \Sigma)^{-1} \Sigma)/p \\ \to_{a.s.} & \frac{\lambda_k m_k}{m_k (m_k - m_{k'})} - \frac{\lambda_{k'} m_{k'}}{m_{k'} (m_k - m_{k'})} + \frac{1}{m_k m_{k'}} \end{aligned}$$

Substitute these expressions back into the expressions for $E_{kk'}$ finishes the proof. □

*Proof of Lemma* **??**. The proof is similar to the proof of lemma **??**. In the first case where $n_1 = \cdots = n_K = n$ so $\gamma_1 = \cdots = \gamma_K = \gamma$ and use $\lambda_1 = \cdots = \lambda_K = \lambda$, we have

$$x_1 = \cdots = x_K := x$$

$$P_{kk'} \to_{a.s.} \int \frac{t}{(xt+\lambda)^2} dH(t)$$

When $\gamma$ is equal across all populations, we will use the shorter notation $m := m_{F_\gamma}(-\lambda), m' = m'_{F_\gamma}(-\lambda)$ in this proof. By definitions, we have

$$\int \frac{1}{xt+\lambda} dH(t) := m \quad \int \frac{x't+1}{(xt+\lambda)^2} dH(t) := m'$$

Here $x'$ is the derivative of $x$ with respect to $\lambda$. Then

$$m' = \int \frac{(xt+\lambda-\lambda)\frac{x'}{x}+1}{(xt+\lambda)^2} dH(t) = \frac{x'}{x} m + (1 - \frac{\lambda x'}{x}) \int \frac{1}{(xt+\lambda)^2} dH(t)$$

$$\int \frac{1}{(xt+\lambda)^2} dH(t) = \frac{xm' - x'm}{x - \lambda x'}$$

So the functional of interest is

$$\begin{aligned}
&\int \frac{t}{(xt+\lambda)^2} dH(t) \\
=&\frac{\int \frac{x't+1}{(xt+\lambda)^2} dH(t) - \int \frac{1}{(xt+\lambda)^2} dH(t)}{x'} \\
=&\frac{m' - \frac{xm'-x'm}{x-\lambda x'}}{x'} \\
=&\frac{\frac{m'x - \lambda m'x' - xm' + x'm}{x-\lambda x'}}{x'} \\
=&\frac{m - \lambda m'}{x - \lambda x'} \\
=&\frac{m - \lambda m'}{1 - \gamma + \gamma\lambda^2 m'}
\end{aligned}$$

For the second case, when $\Sigma = I_p$, $P_{kk'} = E_{kk'}$. So the proof follows from lemma **??**. For the third case, pulling the trick with results from Knowles & Yin (2017) again gives

$$\begin{aligned}
&\mathrm{tr}(\Sigma(\hat{\Sigma}_k + \lambda_k I_p)^{-1}(\hat{\Sigma}_{k'} + \lambda_{k'} I_p)^{-1})/p \\
\to_{a.s.}&\frac{1}{\lambda_k \lambda_{k'}} E_H \frac{t}{(1+tm_k)(1+tm_{k'})}
\end{aligned}$$

which can be consistently estimated by

$$\frac{1}{\lambda_k \lambda_{k'}} \frac{\lambda_k m_k - \lambda_{k'} m_{k'}}{m_{k'} - m_k}$$

as claimed by the lemma. □

### B.3 Proofs of the Theorems

*Proof of Theorem* **??**. We can express the estimation risk as

$$M(W) = W^T(A+R)W - 2\beta_K^T BW + ||\beta_K||^2$$

Where $A$, $R$ and $B$ are defined in Theorem **??**. Take the derivative w.r.t. $W$, we get

$$W_E^* = (A+R)^{-1}v$$

where $A, R, v$ have been defined in the main text. Substitute this $W_E^*$ into $M_K(W)$, we get

$$M(W_E^*) = \beta_K^T[I_p - B(B^TB+R)^{-1}B^T]\beta_K$$

as claimed. □

*Proof of Theorem* **??**. Under the assumptions of theorem **??**, given lemma B.1 and lemma B.2, firstly for $V$ and $k = 1, \cdots, K-1$, we have

$$\begin{aligned} V_k =& \beta_k Q_k \beta_K = \beta_k[1 - \lambda_k(\hat{\Sigma}_k + \lambda_k I_p)^{-1}]\beta_K \\ \to_{a.s.}& \rho_{kK}\sigma_k\sigma_K\alpha_k\alpha_K[1 - \lambda_k m_{F_{\gamma_k}}(-\lambda_k)] \end{aligned}$$

and similarly

$$\begin{aligned} V_K =& \beta_K Q_K \beta_K = \beta_k[1 - \lambda_K \operatorname{tr}(\hat{\Sigma}_K + \lambda_K I_p)^{-1}]\beta_K \\ \to_{a.s.}& \sigma_K^2\alpha_K[1 - \lambda_K m_{F_{\gamma_K}}(-\lambda_K)] \end{aligned}$$

We can then find the limits of diagonal term of $A$ and $R$; for $k = 1, \cdots, K$

$$\begin{aligned} A_{kk} =& \beta_k^T[I_p - 2\lambda_k(\hat{\Sigma}_k + \lambda_k I_p)^{-1} + \lambda_k^2(\hat{\Sigma}_k + \lambda_k I_p)^{-2}]\beta_k \\ \to_{a.s.}& \sigma_k^2\alpha_k^2[1 - 2\lambda_k m_{F_{\gamma_k}}(-\lambda_k) + \lambda_k^2 m'_{F_{\gamma_k}}(-\lambda_k)] \end{aligned}$$

$$\begin{aligned} R_{kk} &= n_k^{-1}\sigma_k^2 \operatorname{tr}[(\hat{\Sigma}_k + \lambda_k I_p)^{-2}\hat{\Sigma}_k] \\ &= \gamma\sigma_k^2 \operatorname{tr}[((\hat{\Sigma}_k + \lambda_k I_p)^{-1} - \lambda_k(\hat{\Sigma}_k + \lambda_k I_p)^{-2}]/p \\ &\to_{a.s.} \sigma_k^2[m_{F_{\gamma_k}}(-\lambda_k) - \lambda_k m'_{F_{\gamma_k}}(-\lambda_k)] \end{aligned}$$

For $k, k' = 1, \cdots K$ and $k \neq k'$;

$$\begin{aligned}A_{kk'} =&\beta_k^T[I_p - \lambda_k(\hat{\Sigma}_k + \lambda_k I_p)^{-1} - \lambda_{k'}(\hat{\Sigma}_{k'} + \lambda_{k'} I_p)^{-1}\\&+\lambda_k\lambda_{k'}(\hat{\Sigma}_k + \lambda_k I_p)^{-1}(\hat{\Sigma}_{k'} + \lambda_{k'} I_p)^{-1}]\beta_{k'}\\ \rightarrow_{a.s.}&\rho_{kk'}\sigma_k\sigma_{k'}\alpha_k\alpha_{k'}[1 - \lambda_k m_{F_{\gamma_k}}(-\lambda_k) - \lambda_{k'} m_{F_{\gamma_{k'}}}(-\lambda_{k'}) + \lambda_k\lambda_{k'}\mathcal{E}_{kk'}]\end{aligned}$$

where we used $\mathcal{E}_{ij}$ from lemma **??**. □

*Proof of Theorem* **??**.

$$\begin{aligned}r(W) =&\sigma_K^2 + E_{\epsilon_1,\cdots,\epsilon_K,x_0}(||(\sum_{k=1}^K W_k\hat{\beta}_k^T - \beta_K^T)x_0||^2)\\=&\sigma_K^2 + \sum_{k,k'=1}^K W_k W_{k'} E_{\epsilon_1,\cdots,\epsilon_K}(\hat{\beta}_k^T\Sigma\hat{\beta}_{k'}) - 2\sum_{k=1}^K W_k E_{\epsilon_1,\cdots,\epsilon_K}(\hat{\beta}_k^T)\Sigma\beta_K + \beta_K^T\Sigma\beta_K\\=&\sigma_K^2 + W^T(C+F)W - 2D^T W + \beta_K^T\Sigma\beta_K\end{aligned}$$

Where $C, F, D$ are defined in theorem **??**. Take derivative w.r.t. $W$ and equate it to $0$, we get

$$W_P^* = (C+F)^{-1}D$$

Substitute this solution into the expected prediction risk formula, we have

$$r_K(W_P^*) =\sigma_K^2 + \beta_K^T\Sigma\beta_K - D^T(C+F)^{-1}D$$

which is the claim in the theorem. □

*Proof of Theorem* **??**. This proof follows directly from assumption **??**, assumption **??**, lemma B.1 and lemma B.2,. We continue with the simplified notations using $m_k := m_{F_{\gamma_k}}(-\lambda_k)$, $m_{k'} := m_{F_{\gamma_{k'}}}(-\lambda_{k'})$, $v_k := v_{F_{\gamma_k}}(-\lambda_k)$ and $v_{k'} := v_{F_{\gamma_{k'}}}(-\lambda_{k'})$. Firstly note $\Sigma_{ii} = 1$ for $i = 1, \cdots, p$, $\mathrm{tr}(\Sigma)/p = 1$. Starting with the limit for $D_k, k = 1, \cdots, K-1$,

$$\begin{aligned}D_k &\rightarrow_{a.s.} \rho_{kK}\alpha_k\alpha_K\sigma_k\sigma_K \,\mathrm{tr}[\Sigma - \lambda_k(\hat{\Sigma}_k + \lambda_k I_p)^{-1}\Sigma]/p\\&\overset{(i)}{=} \rho_{kK}\alpha_k\alpha_K\sigma_k\sigma_K[1 - \frac{\lambda_k}{\gamma_k}(\frac{1}{\lambda_k v_k(-\lambda_k)})]\end{aligned}$$

where the first step follows from lemma B.1 and $(i)$ follows from lemma B.2. When $k = K$, we simply replace $\rho_{kK}\alpha_k\alpha_K\sigma_k\sigma_K$ with $\alpha_K^2\sigma_K^2$ in the first step. We now use the same set of techniques to find the limites for terms $C_{kk}, k = 1, \cdots, K$ and $F_{kk}, k = 1, \cdots, K$.

$$\begin{aligned}C_{kk} &= E_\epsilon(\hat{\beta}_k^T\Sigma\hat{\beta}_k)\\&= \beta_k^T[I_p - \lambda_k(\hat{\Sigma}_k + \lambda_k I_p)^{-1}]\Sigma[I_p - \lambda_k(\hat{\Sigma}_k + \lambda_k I_p)^{-1}]\beta_k^T\\&\rightarrow_{a.s.} \alpha_k^2\sigma_k^2\,\mathrm{tr}[\Sigma - \lambda_k(\hat{\Sigma}_k + \lambda_k I_p)^{-1}\Sigma - \lambda_k\Sigma(\hat{\Sigma}_k + \lambda_k I_p)^{-1}\\&+ \lambda_k^2(\hat{\Sigma}_k + \lambda_k I_p)^{-1}\Sigma(\hat{\Sigma}_k + \lambda_k I_p)^{-1}]/p\end{aligned}$$

$$
\begin{aligned}
&\to_{a.s.} \sigma_k^2\alpha_k^2[1 - 2\frac{\lambda_k}{\gamma_k}(\frac{1}{\lambda_k v_k(-\lambda_k)}) + \frac{\lambda_k^2}{\gamma_k}\frac{v_k(-\lambda_k) - \lambda_k v_k'(-\lambda_k)}{[\lambda_k v_k(-\lambda_k)]^2}] \\
F_{kk} &= \sigma_k^2 \operatorname{tr}[(\hat{\Sigma}_k + \lambda_k I_p)^{-1}\hat{\Sigma}(\hat{\Sigma}_k + \lambda_k I_p)^{-1}\Sigma]/n_k \\
&= \sigma_k^2\gamma_k[\operatorname{tr}[\Sigma(\hat{\Sigma}_k + \lambda I_p)^{-1}/p] - \lambda_k \operatorname{tr}[\Sigma(\hat{\Sigma}_k + \lambda I_p)^{-2}/p]] \\
&\to_{a.s.} \sigma_k^2\gamma_k[\frac{1}{\gamma_k}(\frac{1}{\lambda_k v_k(-\lambda_k)} - 1) - \lambda\frac{1}{\gamma}\frac{v_k(-\lambda_k) - \lambda_k v_k'(-\lambda_k)}{[\lambda_k v_k(-\lambda_k)]^2}]
\end{aligned}
$$

For $C_{kk'}$ when $k \neq k'$,

$$
\begin{aligned}
C_{kk'} &= \beta_k^T[I_p - \lambda_k(\hat{\Sigma}_k + \lambda_k I_p)^{-1}]\Sigma[I_p - \lambda_{k'}(\hat{\Sigma}_{k'} + \lambda_{k'} I_p)^{-1}]\beta_{k'}^T \\
&\to_{a.s.} \rho_{kk'}\alpha_k\sigma_k\alpha_{k'}\sigma_{k'} \operatorname{tr}[\Sigma - \lambda_k(\hat{\Sigma}_k + \lambda_k I_p)^{-1}\Sigma - \lambda_{k'}\Sigma(\hat{\Sigma}_{k'} + \lambda_{k'} I_p)^{-1} \\
&+ \lambda_k\lambda_{k'}(\hat{\Sigma}_k + \lambda_k I_p)^{-1}\Sigma(\hat{\Sigma}_{k'} + \lambda_{k'} I_p)^{-1}]/p \\
&\to_{a.s.} \rho_{kk'}\alpha_k\sigma_k\alpha_{k'}\sigma_{k'}[1 - \frac{\lambda_k}{\gamma_k}(\frac{1}{\lambda_k v_k(-\lambda_k)}) - \frac{\lambda_{k'}}{\gamma_{k'}}(\frac{1}{\lambda_{k'} v_{k'}(-\lambda_{k'})}) + \lambda_k\lambda_{k'}\mathcal{P}_{kk'}]
\end{aligned}
$$

The expression for $\mathcal{P}_{kk'}$ is given in lemma **??**. Substitute these limiting expressions into the optimal prediction risk and optimal prediction weights complete the proof. □

### B.4 Proofs of Corollaries

*Proof of Corollary* **??**. Assume without loss of generality that indices $1, \cdots, K-1$ are in $\mathcal{I}_2$, $K+1, \cdots, N$ indices are in $\mathcal{I}_1$; and the target population is put in the $K^{\text{th}}$ place. Then we have

$$
\begin{aligned}
\mathcal{W}_E^* &= (\mathcal{A} + \mathcal{R})^{-1}\mathcal{V} \\
&\overset{(i)}{=} \begin{pmatrix} (\mathcal{A}_K + \mathcal{R}_K) & 0 \\ 0 & (\mathcal{A}_N + \mathcal{R}_N) \end{pmatrix}^{-1} \begin{pmatrix} V_K \\ V_N \end{pmatrix} \\
&= \begin{pmatrix} C & D \\ E & F \end{pmatrix} \begin{pmatrix} V_K \\ V_N \end{pmatrix}
\end{aligned}
$$

where $\mathcal{A}_K, \mathcal{R}_K, V_K$ are defined as in theorem **??** for populations $1, \cdots, K$ and $\mathcal{A}_N, \mathcal{R}_N, V_N$ are for populations $K+1, \cdots, K+N$. $(i)$ follows since $\rho_{kk'} = 0$ for $k \in \mathcal{I}_2, k' \in \mathcal{I}_1$. By the block matrix inversion formula, we see both $D$ and $E$ will be all-zero matrix. Thus, the $K+1$ to $N$ entries of $W_E^*$ will be zero. Moreover, the first $K$ entries are $(\mathcal{A}_K + \mathcal{R}_K)^{-1}V_K$. We now prove the limiting weight for target population is $1$ when all other populations have correlation $0$. Under this scenario, from the results above we know

$$
\begin{aligned}
(W_E^*)_K &= \alpha^2\frac{V_K}{A_{KK} + R_{KK}} \\
&= \frac{\alpha^2[1 - \lambda m(-\lambda)]}{\alpha^2[1 - 2\lambda m(-\lambda) + \lambda^2 m'(-\lambda)] + \gamma[m(-\lambda) - \lambda m'(-\lambda)]} \\
&\overset{(i)}{=} \frac{\alpha^2[1 - \lambda m(-\lambda)]}{\alpha^2[1 - \lambda m(-\lambda)]} = 1
\end{aligned}
$$

The proofs for results when $\rho \to 1$ can be found in Theorem 8 of Sheng & Dobriban (2020). □

*Proof of Corollary* **??**. Define a length $K$ vector with constants $\tilde{V}_k = 1 - \lambda_k m(\lambda)$ and a diagonal matrix $D$ with $D_{kk} = R_{kk} + A_{kk} - \rho\alpha^2 \tilde{V}_k^2$. With $b_1 = R_{ii} + A_{ii} - \rho\alpha^2\tilde{V}_k^2$ we have

$$\begin{aligned}\mathcal{W}_E^* =& \frac{1}{\rho\alpha^2}(\tilde{V}\tilde{V}^T + \frac{b_1}{\rho\alpha^2} I_K)^{-1} V \\ =& \frac{1}{\rho\alpha^2}(\frac{\rho\alpha^2}{b_1} I_K - \frac{\frac{\rho^2\alpha^4\tilde{V}\tilde{V}^T}{b_1^2}}{1 + \frac{\rho\alpha^2}{b_1}\tilde{V}^T\tilde{V}})V \\ =& \frac{1}{\frac{b_1}{\rho\alpha^2} + \tilde{V}^T\tilde{V}}\Big[\frac{[1-\lambda m(-\lambda)]}{\rho}\mathbb{1}_\rho + \frac{K\alpha^2[1-\lambda m(-\lambda)]^3}{b_1}\mathbb{1}_\rho \\ & - \frac{\alpha^2(K-1)\rho[1-\lambda m(-\lambda)]^3 + [1-\lambda m(-\lambda)]^3}{b_1}\mathbb{1}\Big]\end{aligned}$$

here $\mathbb{1}$ is a length $K$ all one $p \times 1$ vector, and $\mathbb{1}_\rho$ is a length $K$ $p \times 1$ with first $K_1$ entries $\rho$ and $K^{th}$ entry 1. Note that $\alpha^2 \frac{[1-\lambda m(-\lambda)]^2}{b_1} = \frac{[1-\lambda m(-\lambda)]}{1-\rho[1-\lambda m(-\lambda)]} := C_{\gamma,\rho}$, so we further have

$$\mathcal{W}_E^* = \frac{1}{\frac{b_1}{\rho[1-\lambda m(-\lambda)]} + K[1-\lambda m(-\lambda)]}\Big[\frac{1}{\rho}\mathbb{1}_\rho + KC_{\gamma,\rho}\mathbb{1}_\rho - [(K-1)\rho + 1]C_{\gamma,\rho}\mathbb{1}\Big]$$

This gives us the final results

$$W^*(K)_K = \frac{1/\rho + \frac{(1-\rho)(K-1)[1-\lambda m(-\lambda)]}{1-\rho[1-\lambda m(-\lambda)]}}{[1-\rho[1-\lambda m(-\lambda)]] + K[1-\lambda m(-\lambda)]}$$

$$W^*(K)_{1,\cdots,K-1} = \frac{1 + \frac{(\rho-1)[1-\lambda m(-\lambda)]}{1-\rho[1-\lambda m(-\lambda)]}}{[1-\rho[1-\lambda m(-\lambda)]] + K[1-\lambda m(-\lambda)]}$$

□

*Proof of Corollary* **??**. This proof is purely algebraic. Recall that $\alpha^2 \frac{[1-\lambda m(-\lambda)]^2}{b_1} = \frac{[1-\lambda m(-\lambda)]}{1-\rho[1-\lambda m(-\lambda)]} := C_{\gamma,\rho}$.

$$\begin{aligned} & \alpha^2 - V^T(\mathcal{A} + \mathcal{R})^{-1}V \\ =& \frac{\frac{b_1}{\rho[1-\lambda m(-\lambda)]^2} + \alpha^2(1-\rho)[K - 1 - \frac{1}{\rho}] - \frac{1}{b_1}(K-1)\alpha^4(1-\rho)^2[1-\lambda m(-\lambda)]^2}{\frac{b_1}{\rho\alpha^2[1-\lambda m(-\lambda)]^2} + K} \\ =& \frac{\frac{\alpha^2}{C_{\gamma,\rho}} + \alpha^2(1-\rho)(K-1)[1 - (1-\rho)C_{\gamma,\rho} - \frac{1}{\rho(K-1)}]}{\frac{1}{\rho C_{\gamma,\rho}} + K} \\ =& \frac{\alpha^2 + \alpha^2(1-\rho)(K-1)C_{\gamma,\rho}[1 - (1-\rho)C_{\gamma,\rho} - \frac{1}{\rho(K-1)}]}{\frac{1}{\rho} + KC_{\gamma,\rho}} \\ =& \frac{\alpha^2[1-\rho[1-\lambda m(-\lambda)]] + \alpha^2[1-\lambda m(-\lambda)](1-\rho)(K-1)[1-(1-\rho)C_{\gamma,\rho} - \frac{1}{\rho(K-1)}]}{1/\rho + (K-1)[1-\lambda m(-\lambda)]}\end{aligned}$$

$$
\begin{aligned}
=&\frac{\alpha^2[1-\rho[1-\lambda m(-\lambda)]]}{1/\rho+(K-1)[1-\lambda m(-\lambda)]}\\
+&\frac{\alpha^2(1-\rho)(K-1)[1-\lambda m(-\lambda)][1-[1-\lambda m(-\lambda)]-\frac{1}{\rho(K-1)}+\frac{1}{(K-1)}[1-\lambda m(-\lambda)]]}{[\frac{1}{\rho[1-\lambda m(-\lambda)]}+(K-1)][\frac{1}{[1-\lambda m(-\lambda)]}+\rho]]}
\end{aligned}
$$

When $\rho = 1$, it indeed recovers the distributed ridge estimation risk. □

*Proof of Corollary* **??**. When $\Sigma = I$, lemma **??** states that

$$A_{kk'} \to \rho_{kk'}\sigma_k\sigma_{k'}\alpha_k\alpha_{k'}[1-\lambda_k m_k(\lambda_k)][1-\lambda_{k'}m_{k'}(\lambda_{k'})]$$

When $\lambda_1 = \cdots = \lambda_K = \lambda$, $n_k = \cdots = n_K$ so $\gamma_1 = \cdots = \gamma_K = \gamma$ and $\sigma_1 = \cdots = \sigma_K = 1$ Define

$$\Sigma_\beta := mat[\rho_{kk'}],\ \rho_{kk} = 1$$

w.l.o.g.

$$\mathcal{A} = c_1\Sigma_\beta + \lambda^2\alpha^2 D$$

Recall $\Sigma_\beta$ is defined in corollary **??**, and $c_1 = \alpha^2[1-2\lambda m(-\lambda)+\lambda^2 m(-\lambda)^2]$, $D = diag[m'(-\lambda)-m(-\lambda)^2]$. Then

$$\mathcal{A}+\mathcal{R} = c_1\Sigma_\beta + c_2 I_K$$

where $c_2 = \lambda^2\alpha^2[m'(-\lambda)-m(-\lambda)^2]+\gamma[m(-\lambda)-\lambda m'(-\lambda)]$. At the same time, we can control the largest eigenvalue of $\Sigma_\beta$ by the Gershgorin circle theorem. This implies the maximum eigenvalue of $\Sigma_\rho \leq 1+a$ and $\sum_{k=1,k\neq i}^K |\rho_{iK}| \leq a\ \forall i \in \{1,\cdots,K\}$. Choose $a$ so the inequality is tight, denote and arrange the eigenvalues for $\Sigma_\rho + \frac{c_2}{c_1}I_K$ as $\lambda_1 > \cdots > \lambda_K$. We have

$$
\begin{aligned}
\mathcal{M}_K^*(\mathcal{W}_E^*) &= \alpha^2 - V^T(\mathcal{A}+\mathcal{R})^{-1}V\\
&= \alpha^2 - \frac{1}{c_1}V^T(\Sigma_\rho + \frac{c_2}{c_1}I_K)^{-1}V\\
&\overset{(i)}{=} \alpha^2 - \frac{1}{c_1}V^T\Gamma diag(\frac{1}{\lambda_p+\frac{c_2}{c_1}},\cdots,\frac{1}{\lambda_1+\frac{c_2}{c_1}})\Gamma^T V\\
&\leq \alpha^2 - \frac{1}{c_1}V^T\Gamma diag(\frac{1}{\lambda_1+\frac{c_2}{c_1}},\cdots,\frac{1}{\lambda_1+\frac{c_2}{c_1}})\Gamma^T V\\
&\leq \alpha^2 - \frac{1}{c_1}V^T\Gamma diag(\frac{1}{1+a+\frac{c_2}{c_1}},\cdots,\frac{1}{1+a+\frac{c_2}{c_1}})\Gamma^T V\\
&= \alpha^2 - \frac{1}{c_1(1+a+\frac{c_2}{c_1})}V^T V\\
&= \alpha^2 - \frac{1}{1+\max_i\sum_{k,k\neq i}^K|\rho_{ik}|+\frac{c_2}{c_1}}\frac{\alpha^4}{c_1}[\sum_{i=1}^{K-1}\rho_{iK}^2+1][1-\lambda m(-\lambda)]^2
\end{aligned}
$$

where in step $(i)$ we performed the eigenvalue decomposition of $\Sigma_\rho + \frac{c_2}{c_1}I_K$. □

# Appendix C

### C.5 Demonstration of Corollary ??

Suppose the limiting estimation risks for individual ridge estimators are $0.2$, and we have a signal-to-noise ratio $0.5$. $1-\lambda m(-\lambda)$ is $0.4$. Taking derivative of $\mathcal{M}_K^*(\mathcal{W}_E^*)$, we get

$$\begin{aligned}
&\frac{d\mathcal{M}_K^*(\mathcal{W}_E^*)}{d\rho}\\
=&\frac{\alpha^2[[1-\rho[1-\lambda m(-\lambda)]]^2-K\rho^2[1-\lambda m(-\lambda)]^2]}{[(K-1)\rho[1-\lambda m(-\lambda)]+1]^2}\\
+&\frac{(K-2)(K-1)\rho^2[1-\lambda m(-\lambda)]^3-\rho[1-\lambda m(-\lambda)]^2(K(\rho-2)+2)}{\frac{1}{\alpha^2(1-\rho)(K-1)[1-\lambda m(-\lambda)]^3}(K-1)[\rho[1-\lambda m(-\lambda)]+1]^2[(K-1)\rho[1-\lambda m(-\lambda)]+1]^2}\\
+&\frac{\rho[1-\lambda m(-\lambda)](K(-K\rho+\rho+2)-4)-2K\rho+K+2(\rho+[1-\lambda m(-\lambda)])}{\frac{1}{\alpha^2(1-\rho)(K-1)[1-\lambda m(-\lambda)]^3}(K-1)[\rho[1-\lambda m(-\lambda)]+1]^2[(K-1)\rho[1-\lambda m(-\lambda)]+1]^2}
\end{aligned}$$

Substituting in the value of $1-\lambda m(-\lambda)$, it becomes a quadratic function of $\rho, K$. Simple calculations reveal that for $2 \le K \le 5$, this derivative is strictly negative for all $0<\rho<1$.

### C.6 Demonstration of Corollary ??

When $\arg\sup_{i\in\{1,\cdots,K\}}\sum_{k=1,k\neq i}^K|\rho_{ik}| \neq K$, the second term in the upper bound strictly increases with $\rho_{iK}$ for $i\in\{1,\cdots,K-1\}$. When $\arg\sup_{i\in\{1,\cdots,K\}}\sum_{k=1,k\neq i}^K|\rho_{ik}| = K$, $1+a^*=1+\sum_{i=1}^K\rho_{iK}$. Simple derivative analysis reveals that the sufficient condition for this upper bound to strictly decrease with $\rho_{iK}$ for $i\in\{1,\cdots,K-1\}$ is

$$\sum_{k\in\{1,\cdots,K-1\}\setminus i}\rho_{kK}^2+2\rho_{iK}>1$$

which holds for most $\rho_{iK}$, $k\in\{1,\cdots,K-1\}$.

# Appendix D

### D.7 Parameter Estimation for Analysis of the Colorectal Cancer Microbiome Data

We consider a Bayesian hierarchical regression framework for estimating the population parameters $\{\rho_{kk'},\sigma_k,\alpha_k\}$. Recall $X_k\in R^{n_k\times p}$ is the design matrix for the $k$-th population and $Y_k\in R^{n_k}$ is the corresponding response vector. Define the matrix of linear coefficients:

$$\boldsymbol{\beta}=\begin{pmatrix}\beta_1 & \beta_2 & \cdots & \beta_K\end{pmatrix},$$

where each $\beta_k\in\mathbb{R}^p$ is a vector of regression coefficients associated with the $k$-th population. The regression model is given by:

$$Y_k=X_k\beta_k+\epsilon_k,\quad \epsilon_k\sim N\big(0,\sigma_k^2 I_{n_k}\big)\quad \text{for } k=1,\dots,K.$$

For each predictor index $j = 1, \ldots, p$, we have:

$$\beta_j = (\beta_{1j}, \beta_{2j}, \ldots, \beta_{Kj})^\top \sim \mathcal{N}\big(0, \Sigma_\beta\big),$$

where $\Sigma_\beta$ is a $K \times K$ covariance matrix defined by:

$$(\Sigma_\beta)_{kk'} = \begin{cases} \alpha_k^2 \sigma_k^2 & \text{if } k = k', \\ \rho_{kk'} \alpha_k \alpha_{k'} \sigma_k \sigma_{k'} & \text{if } k \neq k'. \end{cases}$$

The hyperparameters $\sigma_k, \alpha_k$, and $\rho_{kk'}$ have the following priors:

$$\sigma_k \sim N_{>0}(0,1), \quad \alpha_k \sim N_{>0}(0,1), \quad \rho_{kk'} \sim \text{Uniform}(-1,1),$$

with $\rho_{kk'} = \rho_{k'k}$ ensuring symmetry of $\Sigma_\beta$. The definitions for the distribution are summarized below.

- $\mathcal{N}(0, \Sigma_\beta)$ denotes a multivariate normal distribution with mean vector $0$ and covariance matrix $\Sigma_\beta$.
- $N(0, \sigma_k^2)$ denotes a normal distribution with mean $0$ and variance $\sigma_k^2$.
- $N_{>0}(0,1)$ denotes a standard normal distribution truncated to $(0, \infty)$.
- Uniform$(-1,1)$ denotes a uniform distribution over $[-1,1]$.

Given this Bayesian framework, we perform 12,000 iterations of the No-U-Turn Sampler (NUTS) (Hoffman et al., 2014), setting an average proposal acceptance probability of 0.99 and using a maximum tree depth of 16. Implementations are conducted via `RStan` (McElreath, 2018). We use the posterior mean of $\alpha_k$, $\sigma_k$, and $\rho_{kk'}$ as the parameter estimates.

## References


ANDERSON, G. W., GUIONNET, A. & ZEITOUNI, O. (2010). *An introduction to random matrices*. No. 118. Cambridge university press.

CAPITAINE, M. & CASALIS, M. (2004). Asymptotic freeness by generalized moments for gaussian and wishart matrices. application to beta random matrices. *Indiana University mathematics journal* , 397–431.

DOBRIBAN, E. & WAGER, S. (2018). High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics* **46**, 247–279.

HOFFMAN, M. D., GELMAN, A. et al. (2014). The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.* **15**, 1593–1623.

KNOWLES, A. & YIN, J. (2017). Anisotropic local laws for random matrices. *Probability Theory and Related Fields* **169**, 257–352.

LEDOIT, O. & PÉCHÉ, S. (2011). Eigenvectors of some large sample covariance matrix ensembles. *Probability Theory and Related Fields* **151**, 233–264.

MARCHENKO, V. A. & PASTUR, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik* **114**, 507–536.

MCELREATH, R. (2018). *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC.

SHENG, Y. & DOBRIBAN, E. (2020). One-shot distributed ridge regression in high dimensions. In *Proceedings of the 37th International Conference on Machine Learning*, H. D. III & A. Singh, eds., vol. 119 of *Proceedings of Machine Learning Research*. PMLR.

SILVERSTEIN, J. W. (1995). Strong convergence of the empirical distribution of eigenvalues of large dimensional random matrices. *Journal of Multivariate Analysis* **55**, 331–339.

ZHAO, B. & ZHU, H. (2019). Cross-trait prediction accuracy of high-dimensional ridge-type estimators in genome-wide association studies. *arXiv preprint arXiv:1911.10142* .