1. Recap: Linear CCA

Let $\underline{x}_1 \in \mathbb{R}^{d_1}$ and $\underline{x}_2 \in \mathbb{R}^{d_2}$ be two random vectors with covariance matrices Σ_{11}, Σ_{22} and cross-covariance matrix Σ_{12}. We wish to find the linear projection $\underline{w}_1 \in \mathbb{R}^{d^1}, \underline{w}_1 \in \mathbb{R}^{d^2}$ such that the Pearson correlation coefficient of the projected vectors is maximal:

$$\mathrm{corr}\left(\underline{w}_1^T \underline{x}_1, \underline{w}_2^T \underline{x}_2\right) = \frac{\underline{w}_1^T \Sigma_{12} \underline{w}_2}{\sqrt{\underline{w}_1^T \Sigma_{11} \underline{w}_1 \cdot \underline{w}_2^T \Sigma_{22} \underline{w}_2}}$$

to formulate as a maximization problem:

$$\left(\underline{w}_1^*, \underline{w}_2^*\right) = \underset{\underline{w}_1, \underline{w}_2}{\mathrm{argmax}} \frac{\underline{w}_1^T \Sigma_{12} \underline{w}_2}{\sqrt{\underline{w}_1^T \Sigma_{11} \underline{w}_1 \cdot \underline{w}_2^T \Sigma_{22} \underline{w}_2}} = \underset{\underline{w}_1, \underline{w}_2}{\mathrm{argmin}} - \frac{\underline{w}_1^T \Sigma_{12} \underline{w}_2}{\sqrt{\underline{w}_1^T \Sigma_{11} \underline{w}_1 \cdot \underline{w}_2^T \Sigma_{22} \underline{w}_2}}$$

The objective is invariant to scaling of both $\underline{w}_1, \underline{w}_2$, thus we can solve for $\underline{w}_1^T \Sigma_{11} \underline{w}_1 = \underline{w}_2^T \Sigma_{22} \underline{w}_2 = 1$

$$\left(\underline{w}_1^*, \underline{w}_2^*\right) = \underset{\underline{w}_1, \underline{w}_2}{\mathrm{argmin}} -\underline{w}_1^T \Sigma_{12} \underline{w}_2 \quad \mathrm{s.t.} \;\; \underline{w}_1^T \Sigma_{11} \underline{w}_1 = \underline{w}_2^T \Sigma_{22} \underline{w}_2 = 1$$

Which can be solved using Lagrange multipliers theorem (LMT). Constructing the Lagrangian:

$$\mathcal{L}\left(\underline{w}_1, \underline{w}_2, \mu_1, \mu_2\right) = -\underline{w}_1^T \Sigma_{12} \underline{w}_2 + \mu_1\left(\underline{w}_1^T \Sigma_{11} \underline{w}_1 - 1\right) + \mu_2\left(\underline{w}_2^T \Sigma_{22} \underline{w}_2 - 1\right)$$

Taking the derivatives:

$$(1)\ \frac{\partial \mathcal{L}}{\partial \underline{w}_1} = -\Sigma_{12} \underline{w}_2 + 2\mu_1 \Sigma_{11} \underline{w}_1 = 0 \rightarrow \Sigma_{12} \underline{w}_2 = 2\mu_1 \Sigma_{11} \underline{w}_1$$

$$(2)\ \frac{\partial \mathcal{L}}{\partial \underline{w}_2} = -\Sigma_{12}^T \underline{w}_1 + 2\mu_2 \Sigma_{22} \underline{w}_2 = 0 \rightarrow \Sigma_{12}^T \underline{w}_1 = 2\mu_2 \Sigma_{22} \underline{w}_2$$

$$(3)\ \frac{\partial \mathcal{L}}{\partial \mu_1} = 0 \rightarrow \underline{w}_1^T \Sigma_{11} \underline{w}_1 = 1$$

$$(4)\ \frac{\partial \mathcal{L}}{\partial \mu_2} = 0 \rightarrow \underline{w}_2^T \Sigma_{22} \underline{w}_2 = 1$$

Multiplying (1) by $\underline{w}_1^T$ and (2) by $\underline{w}_2^T$ yields:

$$\underline{w}_1^T \Sigma_{12} \underline{w}_2 = 2\mu_1 \underline{w}_1^T \Sigma_{11} \underline{w}_1 = 2\mu_1$$

$$\underline{w}_2^T \Sigma_{12}^T \underline{w}_1 = 2\mu_2 \underline{w}_2^T \Sigma_{22} \underline{w}_2 = 2\mu_2$$

Noting that $\underline{w}_2^T \Sigma_{12}^T \underline{w}_1 = \underline{w}_1^T \Sigma_{12} \underline{w}_2$ we conclude that

$$\mu_1 = \mu_2 = \mu$$

And our objective of minimizing $-\underline{w}_1^T \Sigma_{12} \underline{w}_2$ is attained by maximizing μ.

Assuming Σ_{11}, Σ_{22} are invertible (which implies they are both positive definite), let's reformulate (1) and (2) as follows:

$$\Sigma_{12} \overbrace{\Sigma_{22}^{-\frac{1}{2}} \Sigma_{22}^{\frac{1}{2}}}^{I} \underline{w_2} = 2\mu \overbrace{\Sigma_{11}^{\frac{1}{2}} \Sigma_{11}^{\frac{1}{2}}}^{\Sigma_{11}} \underline{w_1}$$

Denoting $\underline{w_1'} = \Sigma_{11}^{\frac{1}{2}} \underline{w_1}$, $\underline{w_2'} = \Sigma_{22}^{\frac{1}{2}} \underline{w_2}$ and $T = \Sigma_{22}^{-\frac{1}{2}} \Sigma_{12}^T \Sigma_{11}^{-\frac{1}{2}}$ we get:

$$\Sigma_{12} \Sigma_{22}^{-\frac{1}{2}} \underline{w_2'} = 2\mu \Sigma_{11}^{\frac{1}{2}} \underline{w_1'}$$

$$\Sigma_{11}^{-\frac{1}{2}} \Sigma_{12} \Sigma_{22}^{-\frac{1}{2}} \underline{w_2'} = 2\mu \underline{w_1'}$$

$$T \underline{w_2'} = 2\mu \underline{w_1'}$$

And similarly, by reformulating (2):

$$\Sigma_{22}^{-\frac{1}{2}} \Sigma_{12}^T \Sigma_{11}^{-\frac{1}{2}} \underline{w_1'} = 2\mu \underline{w_2'}$$

$$T^T \underline{w_1'} = 2\mu \underline{w_2'}$$

We conclude that $\underline{w_1'}$ and $\underline{w_2'}$ are left- and right-singular vectors of T with singular value 2μ. Thus, maximizing μ can be attained be choosing $\underline{w_1'}, \underline{w_2'}$ that corresponds to the largest singular value of T.

Recalling that $\underline{w_1'} = \Sigma_{11}^{\frac{1}{2}} \underline{w_1}$, $\underline{w_2'} = \Sigma_{22}^{\frac{1}{2}} \underline{w_2}$, the optimal solution is:

$$\left(\underline{w_1^*}, \underline{w_2^*}\right) = \left(\Sigma_{11}^{-\frac{1}{2}} \underline{w_1'}, \Sigma_{22}^{-\frac{1}{2}} \underline{w_2'}\right)$$

Generalizing to the multi-dimensional case, we search for subsequent projections $\left\{\left(\underline{w_1^i}, \underline{w_2^i}\right)\right\}_{i=1}^{K}$ we need to add the constraint that the projections are also uncorrelated, namely, $\underline{w_1^i}^T \Sigma_{11} \underline{w_1^j} = 0$ for $i < j$. Assembling the k-top projections we construct the matrices $A_1 \in \mathbb{R}^{d_1 \times k}, A_2 \in \mathbb{R}^{d_2 \times k}$ by placing the projection vectors $\underline{w_1^i}$'s as the columns of A_1 and similarly for A_2. The resulting optimization problem:

$$\underset{A_1, A_2}{\text{maximize}}\, \text{tr}(A_1^T \Sigma_{12} A_2) \;\; \text{st.} \;\; A_1^T \Sigma_{11} A_1 = I_{k \times k}, \;\; A_2^T \Sigma_{22} A_2 = I_{k \times k}$$

Repeating similar derivation, the optimal solution is:

$$(A_1^*, A_2^*) = \left(\Sigma_{11}^{-\frac{1}{2}} U_k, \Sigma_{22}^{-\frac{1}{2}} V_k\right)$$

Where, U_k, V_k are the first k left- and right-singular vectors of the matrix $T = \Sigma_{22}^{-\frac{1}{2}} \Sigma_{12}^T \Sigma_{11}^{-\frac{1}{2}}$

2. Deep CCA

* This derivation follows the proof outlines of the original paper 'Deep Canonical Correlation Analysis', ICML 2013

Given N pairs of samples $\left(\underline{x}_1^{(j)}, \underline{x}_2^{(j)}\right)$ $j = 1, \ldots, N$, construct the matrices $X_i \in \mathbb{R}^{d_i \times N}, i = 1,2$ where the k-th column is the k-th sample draws from distribution i. We wish to find two deep networks f_1, f_2 with output layers of dimension k and parametrized by $\underline{\theta}_1, \underline{\theta}_2$ such that:

$$H_i = f_i(X_i; \underline{\theta}_i) \in \mathbb{R}^{d_o \times N}$$

Note that the size of the output layers of both networks is identical. We further define the Centered Data Matrix:

$$\overline{H}_i = H_i - \frac{1}{N} H_i \cdot \mathbf{1} = H_i \left(I_{N\times N} - \frac{1}{N} \cdot \mathbf{1}_{N\times N}\right)$$

where, $\mathbf{1}$ is an $N \times N$ matrix filed with ones, i.e., the sample mean of each feature is reduced.

The empirical estimators for the autocorrelation and cross-correlation matrices are defined as follow:

$$\hat{\Sigma}_{ij} = \frac{1}{N-1} \overline{H}_i \cdot \overline{H}_j^T + \delta_{ij} r_i I_{d_o \times d_o} \in \mathbb{R}^{d_o \times d_o}$$

Where, $r_i I$ is a regularization term that guarantees $\hat{\Sigma}_{ii}$ are invertible. Define the matrix:

$$R = \hat{\Sigma}_{11}^{-1/2} \hat{\Sigma}_{12} \hat{\Sigma}_{22}^{-1/2} \in \mathbb{R}^{d_o \times d_o}$$

With SVD decomposition: $R = UDV^T$ where U, V are unitary matrices, i.e., $UU^T = VV^T = I$.

Note that $\mathrm{tr}(D)$ is the empirical correlation of the data represented by H_1, H_2:

$$f \triangleq \mathrm{corr}(H_1, H_2) = \mathrm{tr}(D)$$

However, computing D or the eigen-values of R using the SVD decomposition is non-differentiable.

3. Derivatives

Finding a differentiable representation for D and formulas for the derivatives of f (the correlation) w.r.t the network outputs H_1, H_2, i.e.,

$$\frac{\partial f}{\partial H_1}, \frac{\partial f}{\partial H_2}$$

3.1. Differentiable representation

To introduce a differentiable representation for $\mathrm{tr}(D)$, let's observe the trace norm of R:

$$\|R\|_{tr} = \mathrm{tr}\left(\sqrt{R^T R}\right) \overset{(1)}{=} \mathrm{tr}\left(\sqrt{VDU^T UDV^T}\right) \overset{(2)}{=} \mathrm{tr}\left(\sqrt{VDDV^T}\right) =$$

$$\overset{(3)}{\cong} \operatorname{tr}\left(\sqrt{(VDV^T)(VDV^T)}\right) \overset{(4)}{\cong} \operatorname{tr}(VDV^T) \overset{(5)}{\cong} \operatorname{tr}(DV^TV) = \operatorname{tr}(D)$$

1 – substituting SVD decomposition of R, 2 – unitarity of U, 3 – unitarity of V, 4- square root of matrix, 5 – circularity of trace

$$f \triangleq \operatorname{corr}(H_1, H_2) = \operatorname{tr}(D) = \operatorname{tr}\left(\sqrt{R^TR}\right)$$

3.2. Derivative w.r.t H_1

Applying the chain rule:

$$\frac{\partial f}{\partial (H_1)_{kl}} = \sum_i \sum_j \left[\frac{\partial f}{\partial \left(\hat{\Sigma}_{11}\right)_{ij}} \cdot \frac{\partial \left(\hat{\Sigma}_{11}\right)_{ij}}{\partial (H_1)_{kl}} + \frac{\partial f}{\partial \left(\hat{\Sigma}_{12}\right)_{ij}} \cdot \frac{\partial \left(\hat{\Sigma}_{12}\right)_{ij}}{\partial (H_1)_{kl}}\right]$$
$$= \sum_i \sum_j \left[(\nabla_{11})_{ij} \cdot \frac{\partial \left(\hat{\Sigma}_{11}\right)_{ij}}{\partial (H_1)_{kl}} + (\nabla_{12})_{ij} \cdot \frac{\partial \left(\hat{\Sigma}_{12}\right)_{ij}}{\partial (H_1)_{kl}}\right]$$

Where we denoted: $(\nabla_{kl})_{ij} \triangleq \frac{\partial f}{\partial \left(\hat{\Sigma}_{kl}\right)_{ij}}$

Solving for each element separately.

<u>First, the derivation of ∇_{12}:</u>

(1) Applying the chain rule:

$$(\nabla_{12})_{ij} \triangleq \frac{\partial f}{\partial \left(\hat{\Sigma}_{12}\right)_{ij}} = \sum_k \sum_l \frac{\partial f}{\partial R_{kl}} \cdot \frac{\partial R_{kl}}{\partial \left(\hat{\Sigma}_{12}\right)_{ij}}$$

(2) By lemma (2), $\frac{\partial f}{\partial R} = UV^T$ where U, V are SVD decomposition of R

$$(\nabla_{12})_{ij} = \sum_k \sum_l (UV^T)_{kl} \cdot \frac{\partial R_{kl}}{\partial \left(\hat{\Sigma}_{12}\right)_{ij}}$$

(3) Deriving $\frac{\partial R_{kl}}{\partial \left(\hat{\Sigma}_{12}\right)_{ij}}$ by substituting $R = \hat{\Sigma}_{11}^{-1/2}\hat{\Sigma}_{12}\hat{\Sigma}_{22}^{-1/2}$

$$R_{kl} = \left(\hat{\Sigma}_{11}^{-1/2}\hat{\Sigma}_{12}\hat{\Sigma}_{22}^{-1/2}\right)_{kl} = \sum_p \sum_q \left(\hat{\Sigma}_{11}^{-1/2}\right)_{kp} \left(\hat{\Sigma}_{12}\right)_{pq} \left(\hat{\Sigma}_{22}^{-1/2}\right)_{ql}$$

$$\frac{\partial R_{kl}}{\partial \left(\hat{\Sigma}_{12}\right)_{ij}} = \sum_p \sum_q \left(\hat{\Sigma}_{11}^{-1/2}\right)_{kp} \delta_{pi}\delta_{qj} \left(\hat{\Sigma}_{22}^{-1/2}\right)_{ql} = \left(\hat{\Sigma}_{11}^{-1/2}\right)_{ki} \left(\hat{\Sigma}_{22}^{-1/2}\right)_{jl}$$

(4) Substituting back to (2):

$$(\nabla_{12})_{ij} = \sum_k \sum_l (UV^T)_{kl} \cdot \left(\hat{\Sigma}_{11}^{-1/2}\right)_{ki} \left(\hat{\Sigma}_{22}^{-1/2}\right)_{jl}$$

(5) Using the symmetry of $\hat{\Sigma}_{ii}^{-1/2}$

$$(\nabla_{12})_{ij} = \sum_k \sum_l \cdot \left(\hat{\Sigma}_{11}^{-1/2}\right)_{ik} (UV^T)_{kl} \left(\hat{\Sigma}_{22}^{-1/2}\right)_{lj} = \left(\hat{\Sigma}_{11}^{-1/2} UV^T \hat{\Sigma}_{22}^{-1/2}\right)_{kl}$$

(6) It follows that:

$$\nabla_{12} \triangleq \frac{\partial f}{\partial \hat{\Sigma}_{12}} = \hat{\Sigma}_{11}^{-1/2} UV^T \hat{\Sigma}_{22}^{-1/2}$$

<u>Second, the derivation of</u> $\nabla_{11} \triangleq \frac{\partial f}{\partial \hat{\Sigma}_{11}}$

(1) Recall that $\frac{\partial f}{\partial R^T R} = \frac{\partial \mathrm{tr}\left(\sqrt{R^T R}\right)}{\partial R^T R}$ and using the chain rule:

$$(\nabla_{11})_{ij} \triangleq \frac{\partial f}{\partial \left(\hat{\Sigma}_{11}\right)_{ij}} = \sum_k \sum_l \frac{\partial \mathrm{tr}\left(\sqrt{R^T R}\right)}{\partial (R^T R)_{kl}} \cdot \frac{\partial (R^T R)_{kl}}{\partial \left(\hat{\Sigma}_{11}\right)_{ij}}$$

(2) Applying lemma 5 $\frac{\partial}{\partial X} \mathrm{tr}\left(X^{\frac{1}{2}}\right) = \frac{1}{2}\left(X^{-\frac{1}{2}}\right)^T$ and by the symmetry of $R^T R$:

$$(\nabla_{11})_{ij} = \frac{1}{2} \sum_k \sum_l \left(R^T R^{-1/2}\right)_{kl} \cdot \frac{\partial (R^T R)_{kl}}{\partial \left(\hat{\Sigma}_{11}\right)_{ij}}$$

(3) Calculating the second element by the chain rule:

$$\frac{\partial (R^T R)_{kl}}{\partial \left(\hat{\Sigma}_{11}\right)_{ij}} = \sum_r \sum_s \frac{\partial (R^T R)_{kl}}{\partial \left(\hat{\Sigma}_{11}^{-1}\right)_{rs}} \cdot \frac{\left(\hat{\Sigma}_{11}^{-1}\right)_{rs}}{\left(\hat{\Sigma}_{11}\right)_{ij}} = \sum_r \sum_s \frac{\partial (R^T R)_{kl}}{\partial \left(\hat{\Sigma}_{11}^{-1}\right)_{rs}} \cdot \frac{\left(\hat{\Sigma}_{11}^{-1}\right)_{rs}}{\left(\hat{\Sigma}_{11}\right)_{ij}}$$

(4) Recall that $R^T R = \hat{\Sigma}_{22}^{-1/2} \hat{\Sigma}_{21} \hat{\Sigma}_{11}^{-1/2} \hat{\Sigma}_{11}^{-1/2} \hat{\Sigma}_{12} \hat{\Sigma}_{22}^{-1/2} = \hat{\Sigma}_{22}^{-1/2} \hat{\Sigma}_{21} \hat{\Sigma}_{11}^{-1} \hat{\Sigma}_{12} \hat{\Sigma}_{22}^{-1/2}$

$$\frac{\partial (R^T R)_{kl}}{\partial \left(\hat{\Sigma}_{11}^{-1}\right)_{rs}} = \frac{\partial}{\partial \left(\hat{\Sigma}_{11}^{-1}\right)_{rs}} \left(\hat{\Sigma}_{22}^{-1/2} \hat{\Sigma}_{21} \hat{\Sigma}_{11}^{-1} \hat{\Sigma}_{12} \hat{\Sigma}_{22}^{-1/2}\right)_{kl} =$$

$$= \frac{\partial}{\partial \left(\hat{\Sigma}_{11}^{-1}\right)_{rs}} \left(\sum_p \sum_q \left(\hat{\Sigma}_{22}^{-1/2} \hat{\Sigma}_{21}\right)_{kp} \left(\hat{\Sigma}_{11}^{-1}\right)_{pq} \left(\hat{\Sigma}_{12} \hat{\Sigma}_{22}^{-1/2}\right)_{ql} \right) =$$

$$= \sum_p \sum_q \left(\hat{\Sigma}_{22}^{-1/2} \hat{\Sigma}_{21}\right)_{kp} \frac{\partial \left(\hat{\Sigma}_{11}^{-1}\right)_{pq}}{\partial \left(\hat{\Sigma}_{11}^{-1}\right)_{rs}} \left(\hat{\Sigma}_{12} \hat{\Sigma}_{22}^{-1/2}\right)_{ql} = \left(\hat{\Sigma}_{22}^{-1/2} \hat{\Sigma}_{21}\right)_{kr} \left(\hat{\Sigma}_{12} \hat{\Sigma}_{22}^{-1/2}\right)_{sl}$$

(5) By lemma 4 $\frac{\partial (X^{-1})_{ij}}{\partial X_{kl}} = -X_{ik}^{-1} X_{lj}^{-1}$, thus:

$$\frac{\left(\hat{\Sigma}_{11}^{-1}\right)_{rs}}{\left(\hat{\Sigma}_{11}\right)_{ij}} = -\left(\hat{\Sigma}_{11}^{-1}\right)_{ri} \left(\hat{\Sigma}_{11}^{-1}\right)_{sj}$$

(6) Substituting (4) and (5) to (3) and using the symmetry of $\hat{\Sigma}_{11}^{-1}$:

$$\frac{\partial(R^TR)_{kl}}{\partial(\hat{\Sigma}_{11})_{ij}}=\sum_r\sum_s\frac{\partial(R^TR)_{kl}}{\partial(\hat{\Sigma}_{11}^{-1})_{rs}}\cdot\frac{(\hat{\Sigma}_{11}^{-1})_{rs}}{(\hat{\Sigma}_{11})_{ij}}=-\sum_r\sum_s\left(\hat{\Sigma}_{22}^{-1/2}\hat{\Sigma}_{21}\right)_{kr}\left(\hat{\Sigma}_{12}\hat{\Sigma}_{22}^{-1/2}\right)_{sl}(\hat{\Sigma}_{11}^{-1})_{ri}(\hat{\Sigma}_{11}^{-1})_{sj} =$$

$$=-\sum_r\sum_s\left(\hat{\Sigma}_{22}^{-1/2}\hat{\Sigma}_{21}\right)_{kr}(\hat{\Sigma}_{11}^{-1})_{ri}(\hat{\Sigma}_{11}^{-1})_{js}^T\left(\hat{\Sigma}_{12}\hat{\Sigma}_{22}^{-1/2}\right)_{sl}=-\left(\hat{\Sigma}_{22}^{-1/2}\hat{\Sigma}_{21}\hat{\Sigma}_{11}^{-1}\right)_{ki}\left(\hat{\Sigma}_{11}^{-1}\hat{\Sigma}_{12}\hat{\Sigma}_{22}^{-1/2}\right)_{jl}$$

$$\frac{\partial(R^TR)_{kl}}{\partial(\hat{\Sigma}_{11})_{ij}}=-\left(R^T\hat{\Sigma}_{11}^{-1/2}\right)_{ki}\left(\hat{\Sigma}_{11}^{-1/2}R\right)_{jl}$$

(7) Substituting back in (2):

$$(\nabla_{11})_{ij}\triangleq\frac{\partial f}{\partial(\hat{\Sigma}_{11})_{ij}}=-\frac{1}{2}\sum_k\sum_l((R^TR)^{-1/2})_{kl}\left(R^T\hat{\Sigma}_{11}^{-1/2}\right)_{ki}\left(\hat{\Sigma}_{11}^{-1/2}R\right)_{jl}=$$

$$=-\frac{1}{2}\sum_k\sum_l\left(\hat{\Sigma}_{11}^{-1/2}R\right)_{ik}\left((R^TR)^{-1/2}\right)_{kl}\left(R^T\hat{\Sigma}_{11}^{-1/2}\right)_{lj}=-\frac{1}{2}\left(\hat{\Sigma}_{11}^{-1/2}R(R^TR)^{-1/2}R^T\hat{\Sigma}_{11}^{-1/2}\right)_{ij}$$

(8) Substituting $R=UDV^T$

$$\nabla_{11}=-\frac{1}{2}\hat{\Sigma}_{11}^{-1/2}R(R^TR)^{-1/2}R^T\hat{\Sigma}_{11}^{-1/2}=-\frac{1}{2}\hat{\Sigma}_{11}^{-1/2}UDV^T(VDU^TUDV)^{-1/2}VDU^T\hat{\Sigma}_{11}^{-1/2}=$$

$$=-\frac{1}{2}\hat{\Sigma}_{11}^{-1/2}UDV^T(VDDV)^{-1/2}VDU^T\hat{\Sigma}_{11}^{-1/2}=$$

$$=-\frac{1}{2}\hat{\Sigma}_{11}^{-1/2}UDV^T(VDV^TVDV)^{-1/2}VDU^T\hat{\Sigma}_{11}^{-1/2}=$$

$$=-\frac{1}{2}\hat{\Sigma}_{11}^{-1/2}UDV^T(VDV^T)^{-1}VDU^T\hat{\Sigma}_{11}^{-1/2}=-\frac{1}{2}\hat{\Sigma}_{11}^{-1/2}UDU^T\hat{\Sigma}_{11}^{-1/2}$$

$$\nabla_{11}=\frac{\partial f}{\partial\hat{\Sigma}_{11}}=-\frac{1}{2}\hat{\Sigma}_{11}^{-1/2}UDU^T\hat{\Sigma}_{11}^{-1/2}$$

<u>Continuing with the derivation of $\frac{\partial(\hat{\Sigma}_{11})_{ij}}{\partial(H_1)_{kl}}$</u>

In this section, subscripts of H are eliminated in parts for convenience, at all cases H refers to H_1.

(1) Recall that: $\bar{H}_1=H_1-\frac{1}{m}H_1\cdot\mathbf{1}$ and $\mathbf{1}\cdot\mathbf{1}=m\mathbf{1}$

$$\hat{\Sigma}_{11}=\frac{1}{m-1}\bar{H}_1\cdot\bar{H}_1^T+r_1I=\frac{1}{m-1}\left(H_1-\frac{1}{m}H_1\cdot\mathbf{1}\right)\left(H_1^T-\frac{1}{m}\mathbf{1}\cdot H_1^T\right)+r_1I=$$

$$=\frac{1}{m-1}\left(H_1H_1^T-\frac{2}{m}H_1\cdot\mathbf{1}\cdot H_1^T+\frac{1}{m^2}H_1\cdot\mathbf{1}\cdot\mathbf{1}\cdot H_1^T\right)+r_1I$$

$$=\frac{1}{m-1}\left(H_1H_1^T-\frac{1}{m}H_1\cdot\mathbf{1}\cdot H_1^T\right)+r_1I$$

(2) Calculating the derivative of the first element:

$$\frac{\partial(\hat{\Sigma}_{11})_{ij}}{\partial(H_1)_{kl}}=\frac{1}{m-1}\left[\frac{\partial(H_1H_1^T)_{ij}}{\partial(H_1)_{kl}}-\frac{1}{m}\frac{\partial(H_1\cdot\mathbf{1}\cdot H_1^T)_{ij}}{\partial(H_1)_{kl}}\right]$$

$$\frac{\partial (HH^T)_{ij}}{\partial (H)_{kl}} = \frac{\partial}{\partial (H_1)_{kl}}\left(\sum_r H_{ir}(H^T)_{rj}\right) = \frac{\partial}{\partial (H_1)_{kl}}\left(\sum_r H_{ir}H_{jr}\right) =$$
$$= \sum_r \left(\delta_{ik}\delta_{rl}H_{jr} + \delta_{jk}\delta_{rl}H_{ir}\right) = \delta_{ik}H_{jl} + \delta_{jk}H_{il}$$

(3) Calculating the derivative of the second element:

$$\frac{\partial (H_1 \cdot \mathbf{1} \cdot H_1^T)_{ij}}{\partial (H_1)_{kl}} = \frac{\partial}{\partial (H_1)_{kl}}\left(\sum_r\sum_s H_{ir}1_{rs}(H^T)_{sj}\right) = \frac{\partial}{\partial (H_1)_{kl}}\left(\sum_r\sum_s H_{ir}H_{js}\right)$$
$$= \sum_r\sum_s \left(\delta_{ik}\delta_{rl}H_{js} + \delta_{jk}\delta_{sl}H_{ir}\right) = \delta_{ik}\sum_s H_{js} + \delta_{jk}\sum_r H_{ir}$$

(4) Substituting (2) and (3) to (1):

$$\frac{\partial \left(\hat{\Sigma}_{11}\right)_{ij}}{\partial (H_1)_{kl}} = \frac{1}{m-1}\left[\delta_{ik}H_{jl} + \delta_{jk}H_{il} - \frac{1}{m}\delta_{ik}\sum_s H_{js} - \delta_{jk}\sum_r H_{ir}\right] =$$
$$= \frac{1}{m-1}\left[\delta_{ik}\left(H_{jl} - \frac{1}{m}\sum_s H_{js}\right) + \delta_{jk}\left(H_{il} - \frac{1}{m}\sum_r H_{ir}\right)\right] = \frac{1}{m-1}\left[\delta_{ik}\overline{H}_{jl} + \delta_{jk}\overline{H}_{il}\right]$$

(5) Overall:

$$\frac{\partial \left(\hat{\Sigma}_{11}\right)_{ij}}{\partial (H_1)_{kl}} = \frac{1}{m-1}\left[\delta_{ik}(\overline{H}_1)_{jl} + \delta_{jk}(\overline{H}_1)_{il}\right]$$

Last, the derivation of $\frac{\partial \left(\hat{\Sigma}_{12}\right)_{ij}}{\partial (H_1)_{kl}}$

In this section, subscripts of H are eliminated in parts for convenience, at all cases H refers to H_1.

(1) Recall that: $\overline{H}_1 = H_1 - \frac{1}{m}H_1 \cdot \mathbf{1}$ and $\mathbf{1} \cdot \mathbf{1} = m\mathbf{1}$

$$\hat{\Sigma}_{12} = \frac{1}{m-1}\overline{H}_1 \cdot \overline{H}_2^T = \frac{1}{m-1}\left(H_1 - \frac{1}{m}H_1 \cdot \mathbf{1}\right) \cdot \overline{H}_2^T = \frac{1}{m-1}H_1 \cdot \left(I - \frac{1}{m}\mathbf{1}\right) \cdot \overline{H}_2^T$$

(2) Calculating the derivative:

$$\frac{\partial \left(\hat{\Sigma}_{12}\right)_{ij}}{\partial (H_1)_{kl}} = \frac{1}{m-1}\frac{\partial}{\partial (H_1)_{kl}}\left(H_1 \cdot \left(I - \frac{1}{m}\mathbf{1}\right) \cdot \overline{H}_2^T\right) = \frac{1}{m-1}\frac{\partial}{\partial (H_1)_{kl}}\sum_r\sum_s (H_1)_{ir}\left(I - \frac{1}{m}\mathbf{1}\right)_{rs}(\overline{H}_2^T)_{sj} =$$
$$= \frac{1}{m-1}\sum_r\sum_s \delta_{ik}\delta_{rl}\left(I - \frac{1}{m}\mathbf{1}\right)_{rs}(\overline{H}_2^T)_{sj} = \frac{1}{m-1}\sum_r \delta_{ik}\delta_{rl}\left(\left(I - \frac{1}{m}\mathbf{1}\right) \cdot \overline{H}_2^T\right)_{rj} =$$
$$= \frac{1}{m-1}\delta_{ik}\left(\left(I - \frac{1}{m}\mathbf{1}\right) \cdot \overline{H}_2^T\right)_{lj} = \frac{1}{m-1}\delta_{ik}\left(\overline{H}_2 \cdot \left(I - \frac{1}{m}\mathbf{1}\right)\right)_{jl}$$

(3) Note that: $\left(I - \frac{1}{m}\mathbf{1}\right) \cdot \left(I - \frac{1}{m}\mathbf{1}\right) = \left(I - \frac{1}{m}\mathbf{1}\right)$ and thus: $\overline{H}_2 \cdot \left(I - \frac{1}{m}\mathbf{1}\right) = \overline{H}_2$

(4) It follows:

$$\frac{\partial\left(\hat{\Sigma}_{12}\right)_{ij}}{\partial(H_1)_{kl}} = \frac{1}{m-1}\delta_{ik}(\bar{H}_2)_{jl}$$

Finally, the gradient of $f = \mathrm{tr}(R^T R)$ w.r.t H_1:

(1) Substituting all the derivatives into the derivative formula:

$$\frac{\partial f}{\partial(H_1)_{kl}} = \sum_i \sum_j \left[(\nabla_{11})_{ij} \cdot \frac{\partial\left(\hat{\Sigma}_{11}\right)_{ij}}{\partial(H_1)_{kl}} + (\nabla_{12})_{ij} \cdot \frac{\partial\left(\hat{\Sigma}_{12}\right)_{ij}}{\partial(H_1)_{kl}}\right] =$$

$$= \sum_i \sum_j \left[(\nabla_{11})_{ij} \frac{1}{m-1}\left[\delta_{ik}(\bar{H}_1)_{jl} + \delta_{jk}(\bar{H}_1)_{il}\right] + (\nabla_{12})_{ij} \frac{1}{m-1}\delta_{ik}(\bar{H}_2)_{jl}\right]$$

$$= \frac{1}{m-1}\left[\sum_j (\nabla_{11})_{kj}(\bar{H}_1)_{jl} + \sum_i (\nabla_{11})_{ik}(\bar{H}_1)_{il} + \sum_j (\nabla_{12})_{kj}(\bar{H}_2)_{jl}\right] =$$

$$= \frac{1}{m-1}\left[(\nabla_{11} \cdot \bar{H}_1)_{kl} + (\nabla_{11}^T \cdot \bar{H}_1)_{kl} + (\nabla_{12} \cdot \bar{H}_2)_{kl}\right]$$

(2) By symmetry of ∇_{11}:

$$\frac{\partial f}{\partial(H_1)_{kl}} = \frac{1}{m-1}\left[2(\nabla_{11} \cdot \bar{H}_1)_{kl} + (\nabla_{12} \cdot \bar{H}_2)_{kl}\right]$$

(3) Overall:

$$\frac{\partial f}{\partial H_1} = \frac{1}{m-1}\left[2\nabla_{11} \cdot \bar{H}_1 + \nabla_{12} \cdot \bar{H}_2\right]$$

$$\nabla_{11} = \frac{\partial f}{\partial \hat{\Sigma}_{11}} = -\frac{1}{2}\hat{\Sigma}_{11}^{-1/2} U D U^T \hat{\Sigma}_{11}^{-1/2}$$

$$\nabla_{12} \triangleq \frac{\partial f}{\partial \hat{\Sigma}_{12}} = \hat{\Sigma}_{11}^{-1/2} U V^T \hat{\Sigma}_{22}^{-1/2}$$

$$R = UDV^T \ (SVD\ decomposition)$$

3.3. Derivative w.r.t H_2

For real valued random vectors: $corr(H_1, H_2) = corr(H_2, H_1)$, thus we could repeat the same derivation with $R \to R^T = \hat{\Sigma}_{22}^{-1/2}\hat{\Sigma}_{21}\hat{\Sigma}_{11}^{-1/2}$. Note that wherever the SVD decomposition of R was used, it should be replaced with the SVD decomposition of R^T. However, this can be done by interchanging between U and V since: $R = UDV^T \to R^T = VDU^T$, thus saving the computation of SVD decomposition of R^T.

$$\frac{\partial f}{\partial H_2} = \frac{1}{m-1}[2\nabla_{22} \cdot \bar{H}_2 + \nabla_{21} \cdot \bar{H}_1]$$

$$R = UDV^T \ (SVD\ decomposition)$$

$$\nabla_{22} = \frac{\partial f}{\partial \hat{\Sigma}_{22}} = -\frac{1}{2}\hat{\Sigma}_{22}^{-1/2}VDV^T\hat{\Sigma}_{22}^{-1/2}$$

$$\nabla_{21} \triangleq \frac{\partial f}{\partial \hat{\Sigma}_{12}} = \hat{\Sigma}_{22}^{-1/2}VU^T\hat{\Sigma}_{11}^{-1/2} = \nabla_{12}^T$$

Lemma 1: Let U be a unitary matrix ($UU^T = I$) and D a diagonal matrix with proper dimensions. Then,

$$\mathrm{tr}(U^T(dU)D) = \mathrm{tr}(D(dU)^T U) = 0$$

Proof:

(1) From unitarity: $U^T U = I$
(2) The differential: $d(U^T U) = (dU)^T U + U^T(dU) = dI = \mathbf{0}$
(3) It follows that: $(dU)^T U = -U^T dU = -((dU)^T U)^T$
(4) $(dU)^T U$ is anti-symmetric and thus its diagonal elements are all zeros, it also follows that its transpose $U^T(dU)$ is anti-symmetric
(5) $\mathrm{tr}((dU)^T UD) = \sum_i ((dU)^T UD)_{ii} = \sum_i \sum_j ((dU)^T U)_{ij} D_{ji} = \sum_i \sum_j ((dU)^T U)_{ij} D_{ji} \delta_{ji} = \sum_i ((dU)^T U)_{ii} D_{ii} = 0$

where the last two equalities since D is diagonal and the anti-symmetry.

Lemma 2: Let R be a matrix with SVD decomposition $R = UDV^T$, then:

$$\frac{\partial \mathrm{tr}\left(\sqrt{R^T R}\right)}{\partial R} = \frac{\partial \mathrm{tr}(D)}{\partial R} = UV^T$$

Proof:

(1) From linearity of trace and derivative: $d\left(\mathrm{tr}\left(\sqrt{R^T R}\right)\right) = d\big(\mathrm{tr}(D)\big) = \mathrm{tr}(dD)$
(2) Differential of R: $dR = (dU)DV^T + U(dD)V^T + UD(dV)^T$
(3) Left multiplication by U^T and right multiplication by V and solving for dD results:

$$dD = U^T(dR)V - U^T(dU)D - D(dV)^T V$$

(4) Taking the trace and using lemma 1:

$$\begin{aligned}\mathrm{tr}(dD) = \mathrm{tr}(U^T(dR)V) - \mathrm{tr}(U^T(dU)D) - \mathrm{tr}(D(dV)^T V) = \mathrm{tr}(U^T(dR)V) = \mathrm{tr}(VU^T dR) \\ = \langle UV^T, dR\rangle\end{aligned}$$

(5) It follows that:

$$\frac{\partial \mathrm{tr}\left(\sqrt{R^T R}\right)}{\partial R} = \frac{\partial \mathrm{tr}(D)}{\partial R} = UV^T$$

Lemma 3: Derivative of inverse matrix w.r.t a scalar p:

$$\frac{d(X^{-1})}{dp} = -X^{-1}\frac{\partial X}{\partial p}X^{-1}$$

Proof:

(1) By definition: $XX^{-1} = I$
(2) The differential: $d(XX^{-1}) = dX \cdot X^{-1} + X \cdot d(X^{-1}) = dI = \mathbf{0}$

(3) Multiplying by the inverse:

$$d(X^{-1}) = -X^{-1} \cdot dX \cdot X^{-1} = -X^{-1}\frac{\partial X}{\partial p}dp \cdot X^{-1} = -X^{-1}\frac{\partial X}{\partial p}X^{-1}dp$$

(4) It follows:

$$\frac{d(X^{-1})}{dp} = -X^{-1}\frac{\partial X}{\partial p}X^{-1}$$

Lemma 4: Derivative of the inverse matrix X^{-1} w.r.t element x_{kl}

$$\frac{\partial (X^{-1})_{ij}}{\partial X_{kl}} = -X_{ik}^{-1}X_{lj}^{-1}$$

Proof:

(1) applying lemma (5) with $p = x_{kl}$

$$\frac{\partial (X^{-1})_{ij}}{\partial X_{kl}} = -\left(X^{-1}\frac{\partial X}{\partial X_{kl}}X^{-1}\right)_{ij} = -\sum_r\sum_q X_{ir}^{-1}\frac{\partial X_{rq}}{\partial X_{kl}}X_{qj}^{-1} = -\sum_r\sum_q X_{ir}^{-1}\delta_{rk}\delta_{ql}X_{qj}^{-1} = -X_{ik}^{-1}X_{lj}^{-1}$$

Lemma 5: let X be a positive-definite matrix with eigen decomposition $X = UDU^T$, then,

$$\frac{\partial}{\partial X}\mathrm{tr}\left(X^{\frac{1}{2}}\right) = \frac{1}{2}\left(X^{-\frac{1}{2}}\right)^T$$

Proof:

(1) since X is positive definite $D = \sqrt{D}\sqrt{D}$ and thus

$$\mathrm{tr}\left(X^{\frac{1}{2}}\right) = \mathrm{tr}\left(\sqrt{UDU^T}\right) = \mathrm{tr}\left(\sqrt{U\sqrt{D}\sqrt{D}U^T}\right)$$

(2) Unitarity of U ($U^TU = I$)

$$\mathrm{tr}\left(X^{\frac{1}{2}}\right) = \mathrm{tr}\left(\sqrt{U\sqrt{D}\sqrt{D}U^T}\right) = \mathrm{tr}\left(\sqrt{(U\sqrt{D}U^T)(U\sqrt{D}U^T)}\right) = \mathrm{tr}(U\sqrt{D}U^T)$$

(3) Circularity of trace:

$$\mathrm{tr}\left(X^{\frac{1}{2}}\right) = \mathrm{tr}(\sqrt{D}U^TU) = \mathrm{tr}(\sqrt{D})$$

(4) The differential of $\mathrm{tr}\left(X^{\frac{1}{2}}\right)$, using the linearity of trace and derivative and the chain rule

$$d\mathrm{tr}\left(X^{\frac{1}{2}}\right) = d\mathrm{tr}(\sqrt{D}) = \mathrm{tr}\left(d\left(D^{\frac{1}{2}}\right)\right) = \mathrm{tr}\left(\frac{1}{2}D^{-\frac{1}{2}}dD\right)$$

(5) To compute $\mathrm{tr}\left(\frac{1}{2}D^{-\frac{1}{2}}dD\right)$, we will use the differential of X:

$$dX = d(UDU^T) = (dU)DU^T + U(dD)U^T + UD(dU)^T$$

(6) Multiplying by U^T from the left and U from the right

$$U^T dXU = U^T(dU)DU^T U + U^T U(dD)U^T U + U^T UD(dU)^T U = U^T(dU)D + dD + D(dU)^T U$$

(7) Multiplying by $D^{-1/2}$

$$D^{-1/2}U^T dXU = D^{-1/2}U^T(dU)D + D^{-1/2}dD + D^{-1/2}D(dU)^T U$$

(8) Applying trace operator trace:

$$\mathrm{tr}\left(D^{-1/2}U^T dXU\right) = \mathrm{tr}\left(D^{-1/2}U^T(dU)D\right) + \mathrm{tr}\left(D^{-1/2}dD\right) + \mathrm{tr}\left(D^{-1/2}D(dU)^T U\right)$$

(9) again $\mathrm{tr}\left(D^{-1/2}U^T(dU)D\right) = \mathrm{tr}\left(D^{-1/2}D(dU)^T U\right) = 0$ due to the anti-symmetry of $U^T(dU)$ and $(dU)^T U$ and symmetry of D and $D^{-1/2}$

$$\mathrm{tr}\left(D^{-1/2}dD\right) = \mathrm{tr}\left(D^{-1/2}U^T dXU\right)$$

(10) Substituting (9) to (4):

$$\mathrm{dtr}\left(X^{\frac{1}{2}}\right) = \frac{1}{2}\mathrm{tr}\left(D^{-\frac{1}{2}}dD\right) = \frac{1}{2}\mathrm{tr}\left(D^{-1/2}U^T dXU\right) = \frac{1}{2}\mathrm{tr}\left(UD^{-1/2}U^T dX\right) = \mathrm{tr}\left(\frac{1}{2}X^{-1/2}\,dX\right)$$
$$= \langle\left(\frac{1}{2}X^{-1/2}\right)^T, dX\rangle$$

$$\frac{\partial \mathrm{tr}\left(X^{\frac{1}{2}}\right)}{\partial X} = \frac{1}{2}\left(X^{-1/2}\right)^T$$