Modern Multivariate Statistical Techniques

Jinhong Du,15338039

January 8, 2019

Content

Chapter 1 Random Vectors and Matrices

1 Basic Matrix Operations

1.1 Partitioned Matrices

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}$$

then

$$\begin{aligned} |\mathbf{\Sigma}| &= |\mathbf{A}| \cdot |\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}| \\ &= |\mathbf{D}| \cdot |\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}| \end{aligned}$$

$$\begin{aligned} rank(AB) &= rank(A) \qquad \text{if } |B| \neq 0 \\ rank(AB) &\leq \min\{rank(A), rank(B)\} \end{aligned}$$

If $\mathbf{A}$ and $\mathbf{D}$ are symmetric, then

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^\top & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}(\mathbf{D} - \mathbf{B}^\top\mathbf{A}^{-1}\mathbf{B})\mathbf{B}^\top\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{B}(\mathbf{D} - \mathbf{B}^\top\mathbf{A}^{-1}\mathbf{B})^{-1} \\ -(\mathbf{D} - \mathbf{B}^\top\mathbf{A}^{-1}\mathbf{B})\mathbf{B}^\top\mathbf{A}^{-1} & (\mathbf{D} - \mathbf{B}^\top\mathbf{A}^{-1}\mathbf{B})^{-1} \end{pmatrix}$$

1.1.1 Woodbury Matrix Identity

$$(\mathbf{A} + \mathbf{UCV})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}\left(\mathbf{C}^{-1} + \mathbf{V}\mathbf{A}^{-1}\mathbf{U}\right)^{-1} VA^{-1}$$

in a special case, when $\mathbf{U} = \mathbf{u}$, $\mathbf{V} = \mathbf{v}^\top$ and $\mathbf{C} = 1$,

$$\left(\mathbf{A} + \mathbf{u}\mathbf{v}^\top\right)^{-1} = \mathbf{A}^{-1} - \frac{(\mathbf{A}^{-1}\mathbf{u})(\mathbf{v}^\top\mathbf{A}^{-1})}{1 + \mathbf{v}^\top\mathbf{A}^{-1}\mathbf{u}}$$

(1) Direct Proof

The formula can be proven by checking that $(A + UCV)$ times its alleged inverse on the right side of the Woodbury identity gives the identity matrix:

$$
\begin{aligned}
&(A+UCV)\left[A^{-1}-A^{-1}U\left(C^{-1}+VA^{-1}U\right)^{-1}VA^{-1}\right] \\
&=\left\{I-U\left(C^{-1}+VA^{-1}U\right)^{-1}VA^{-1}\right\}+\left\{UCVA^{-1}-UCVA^{-1}U\left(C^{-1}+VA^{-1}U\right)^{-1}VA^{-1}\right\} \\
&=\left\{I+UCVA^{-1}\right\}-\left\{U\left(C^{-1}+VA^{-1}U\right)^{-1}VA^{-1}+UCVA^{-1}U\left(C^{-1}+VA^{-1}U\right)^{-1}VA^{-1}\right\} \\
&=I+UCVA^{-1}-\left(U+UCVA^{-1}U\right)\left(C^{-1}+VA^{-1}U\right)^{-1}VA^{-1} \\
&=I+UCVA^{-1}-UC\left(C^{-1}+VA^{-1}U\right)\left(C^{-1}+VA^{-1}U\right)^{-1}VA^{-1} \\
&=I+UCVA^{-1}-UCVA^{-1} \\
&=I.
\end{aligned}
$$

(2) Algebraic Proof

First consider these useful identities,

$$
\begin{aligned}
U+UCVA^{-1}U &= UC\left(C^{-1}+VA^{-1}U\right)=(A+UCV)A^{-1}U \\
(A+UCV)^{-1}UC &= A^{-1}U\left(C^{-1}+VA^{-1}U\right)^{-1}
\end{aligned}
$$

Now,

$$
\begin{aligned}
A^{-1} &= (A+UCV)^{-1}(A+UCV)A^{-1} \\
&= (A+UCV)^{-1}\left(I+UCVA^{-1}\right) \\
&= (A+UCV)^{-1}+(A+UCV)^{-1}UCVA^{-1} \\
&= (A+UCV)^{-1}+A^{-1}U\left(C^{-1}+VA^{-1}U\right)^{-1}VA^{-1}.
\end{aligned}
$$

(3) Derivation via Blockwise Elimination

Deriving the Woodbury matrix identity is easily done by solving the following block matrix inversion problem

$$
\begin{bmatrix} A & U \\ V & -C^{-1} \end{bmatrix}\begin{bmatrix} X \\ Y \end{bmatrix}=\begin{bmatrix} I \\ 0 \end{bmatrix}
$$

Expanding, we can see that the above reduces to

$$
\begin{cases} AX+UY=I \\ VX-C^{-1}Y=0 \end{cases}
$$

which is equivalent to

$$
(A+UCV)X=I.
$$

Eliminating the first equation, we find that

$$
X=A^{-1}(I-UY),
$$

which can be substituted into the second to find

$$VA^{-1}(I-UY)=C^{-1}Y.$$

Expanding and rearranging, we have

$$VA^{-1}=(C^{-1}+VA^{-1}U)Y,$$

or

$$(C^{-1}+VA^{-1}U)^{-1}VA^{-1}=Y.$$

Finally, we substitute into our

$$AX+UY=I,$$

and we have

$$AX+U(C^{-1}+VA^{-1}U)^{-1}VA^{-1}=I,$$

Thus,

$$(A+UCV)^{-1}=X=A^{-1}-A^{-1}U\left(C^{-1}+VA^{-1}U\right)^{-1}VA^{-1}.$$

We have derived the Woodbury matrix identity.

1.2 Kronecker Product And Vectoring

The (left) Kronecker Product is given by

$$\mathbf{A}\otimes\mathbf{B}=\begin{pmatrix}\mathbf{A}B_{11} & \cdots & \mathbf{A}B_{1M}\\ \vdots & \ddots & \vdots\\ \mathbf{A}B_{L1} & \cdots & \mathbf{A}B_{LM}\end{pmatrix}$$

where $\mathbf{A}\in\mathbf{R}^{J\times K}$ and $\mathbf{B}\in\mathbf{R}^{L\times M}$.

$$\begin{aligned}(\mathbf{A}\otimes\mathbf{B})\otimes\mathbf{C}&=\mathbf{A}\otimes(\mathbf{B}\otimes\mathbf{C})\\ (\mathbf{A}\otimes\mathbf{B})(\mathbf{C}\otimes\mathbf{D})&=(\mathbf{AC})\otimes(\mathbf{BD})\\ (\mathbf{A}+\mathbf{B})\otimes\mathbf{C}&=(\mathbf{A}\otimes\mathbf{C})+(\mathbf{B}\otimes\mathbf{C})\\ (\mathbf{A}\otimes\mathbf{B})^{\top}&=\mathbf{A}^{\top}\otimes\mathbf{B}^{\top}\\ tr(\mathbf{A}\otimes\mathbf{B})&=tr(\mathbf{A})tr(\mathbf{B})\\ rank(\mathbf{A}\otimes\mathbf{B})&=rank(\mathbf{A})rank(\mathbf{B})\end{aligned}$$

If $\mathbf{A}\in\mathbf{R}^{J\times J}$, $\mathbf{B}\in\mathbf{R}^{K\times K}$, then

$$|\mathbf{A}\otimes\mathbf{B}|=|\mathbf{A}|^{K}|\mathbf{B}|^{J}$$

If $\mathbf{A}\in\mathbf{R}^{J\times K}$, $\mathbf{B}\in\mathbf{R}^{L\times M}$, then

$$\mathbf{A}\otimes\mathbf{B}=(\mathbf{A}\otimes\mathbf{I}_{L})(\mathbf{I}_{M}\otimes\mathbf{B})$$

If $\mathbf{A}$ and $\mathbf{B}$ are squared and nonsingular, then

$$(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}$$

Combine vectorizing with Kronecker products,

$$vec(\mathbf{ABC}) = (\mathbf{A} \otimes \mathbf{C}^{\top}) vec(\mathbf{B})$$

where the vectorizing operation $vec(\mathbf{A})$ denotes the $(JK \times 1)$-column vector formed by placing the columns of a $(J \times K)$-matrix $\mathbf{A}$ under one another successively.

1.3 Quadratic Form

$$\mathbb{E}(\boldsymbol{\varepsilon}^{\top} \boldsymbol{\Lambda} \boldsymbol{\varepsilon}) = tr(\boldsymbol{\Lambda}\Sigma) + \boldsymbol{\mu}^{\top} \boldsymbol{\Lambda} \boldsymbol{\mu}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ the expected value and variance-covariance matrix of $\boldsymbol{\varepsilon}$, respectively.

Proof.

$$\begin{aligned}
\mathbb{E}(\boldsymbol{\varepsilon}^{\top} \boldsymbol{\Lambda} \boldsymbol{\varepsilon}) &= \mathbb{E}[tr(\boldsymbol{\varepsilon}^{\top} \boldsymbol{\Lambda} \boldsymbol{\varepsilon})] \\
&= \mathbb{E}[tr(\boldsymbol{\Lambda} \boldsymbol{\varepsilon} \boldsymbol{\varepsilon}^{\top})] \\
&= tr[\mathbb{E}(\boldsymbol{\Lambda} \boldsymbol{\varepsilon} \boldsymbol{\varepsilon}^{\top})] \\
&= tr[\boldsymbol{\Lambda} \mathbb{E}(\boldsymbol{\varepsilon} \boldsymbol{\varepsilon}^{\top})] \\
&= tr[\boldsymbol{\Lambda}(\Sigma + \boldsymbol{\mu} \boldsymbol{\mu}^{\top})] \\
&= tr(\boldsymbol{\Lambda}\Sigma) + tr(\boldsymbol{\Lambda} \boldsymbol{\mu} \boldsymbol{\mu}^{\top}) \\
&= tr(\boldsymbol{\Lambda}\Sigma) + tr(\boldsymbol{\mu}^{\top} \boldsymbol{\Lambda} \boldsymbol{\mu}) \\
&= tr(\boldsymbol{\Lambda}\Sigma) + \boldsymbol{\mu}^{\top} \boldsymbol{\Lambda} \boldsymbol{\mu}
\end{aligned}$$

□

1.4 Matrix Calculus

1.4.1 Vector Derivatives

Let $\mathbf{x} = \begin{pmatrix} x_1 & \cdots & x_K \end{pmatrix}^{\top}$ and $\mathbf{y} = \begin{pmatrix} y_1(\mathbf{x}) & \cdots & y_J(\mathbf{x}) \end{pmatrix}^{\top}$, then

$$\frac{\partial \mathbf{f}}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_J}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_K} & \cdots & \frac{\partial y_J}{\partial x_K} \end{pmatrix}^{\top}$$

The Jacobian matrix is given by

$$\mathbf{J}_{\mathbf{x}}\mathbf{y} = \frac{\partial \mathbf{y}}{\partial \mathbf{x}^\top} = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_K} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_K} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_J}{\partial x_1} & \frac{\partial y_J}{\partial x_2} & \cdots & \frac{\partial y_J}{\partial x_K} \end{pmatrix}$$

If $\mathbf{A} \in \mathbf{R}^{J \times K}$, then

$$\begin{aligned} \frac{\partial (\mathbf{A}\mathbf{x})}{\partial \mathbf{x}^\top} &= \mathbf{A} \\ \frac{\partial (\mathbf{x}^\top \mathbf{A}\mathbf{x})}{\partial \mathbf{x}^\top} &= \mathbf{x}^\top (\mathbf{A} + \mathbf{A}^\top) \qquad (J = K) \end{aligned}$$

1.4.2 Matrix Derivatives

$$\frac{\partial \mathbf{A}}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial \mathbf{A}^\top}{\partial x_1} & \cdots & \frac{\partial \mathbf{A}^\top}{\partial x_K} \end{pmatrix}^\top$$

If $\alpha \in \mathbb{R}$, then

$$\begin{aligned} \frac{\partial (\alpha \mathbf{A})}{\partial \mathbf{x}} &= \alpha \frac{\partial \mathbf{A}}{\partial \mathbf{x}} \\ \frac{\partial (\mathbf{A} + \mathbf{B})}{\partial \mathbf{x}} &= \frac{\partial \mathbf{A}}{\partial \mathbf{x}} + \frac{\partial \mathbf{B}}{\partial \mathbf{x}} \\ \frac{\partial \mathbf{A}\mathbf{B}}{\partial \mathbf{x}} &= \frac{\partial \mathbf{A}}{\partial \mathbf{x}} \mathbf{B} + \mathbf{A} \frac{\partial \mathbf{B}}{\partial \mathbf{x}} \\ \frac{\partial (\mathbf{A} \otimes \mathbf{B})}{\partial \mathbf{x}} &= \left(\frac{\partial \mathbf{A}}{\partial \mathbf{x}} \otimes \mathbf{B} \right) + \left(\mathbf{A} \otimes \frac{\partial \mathbf{B}}{\partial \mathbf{x}} \right) \\ \frac{\partial \mathbf{A}^{-1}}{\partial \mathbf{x}} &= -\mathbf{A}^{-1} \left(\frac{\partial \mathbf{A}}{\partial \mathbf{x}} \right) \mathbf{A}^{-1} \end{aligned}$$

1.4.3 Gradient Vectors And Matrices

The gradient vector is given by

$$\nabla_{\mathbf{x}} y = \frac{\partial y}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial y}{\partial x_1} & \frac{\partial y}{\partial x_2} & \cdots & \frac{\partial y}{\partial x_K} \end{pmatrix}^\top = \left(\frac{\partial y}{\partial \mathbf{x}} \right)^\top = (\mathbf{J}_{\mathbf{x}} y)^\top$$

The gradient matrix is given by

$$\frac{\partial y}{\partial \mathbf{A}} = \begin{pmatrix} \frac{\partial y}{\partial A_{11}} & \frac{\partial y}{\partial A_{12}} & \cdots & \frac{\partial y}{\partial A_{1K}} \\ \frac{\partial y}{\partial A_{21}} & \frac{\partial y}{\partial A_{22}} & \cdots & \frac{\partial y}{\partial A_{2K}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y}{\partial A_{J1}} & \frac{\partial y}{\partial A_{12}} & \cdots & \frac{\partial y}{\partial A_{JK}} \end{pmatrix}$$

If $\mathbf{A} \in \mathbf{R}^{J \times J}$, then

$$\frac{\partial tr(\mathbf{A})}{\partial \mathbf{A}} = \mathbf{I}_J$$
$$\frac{\partial |\mathbf{A}|}{\partial \mathbf{A}} = |\mathbf{A}| \cdot (\mathbf{A}^\top)^{-1}$$

1.4.4 Hessian Matrices

$$\mathbf{H}_{\mathbf{x}} y = \frac{\partial}{\partial \mathbf{x}} \left(\frac{\partial y}{\partial \mathbf{x}} \right)^\top = \frac{\partial^2 y}{\partial \mathbf{x} \partial \mathbf{x}^\top} = \begin{pmatrix} \frac{\partial^2 y}{\partial x_1^2} & \frac{\partial^2 y}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 y}{\partial x_1 \partial x_K} \\ \frac{\partial^2 y}{\partial x_2 x_1} & \frac{\partial^2 y}{\partial x_2 \partial x_2} & \cdots & \frac{\partial^2 y}{\partial x_2 \partial x_K} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 y}{\partial x_K \partial x_1} & \frac{\partial^2 y}{\partial x_K \partial x_2} & \cdots & \frac{\partial^2 y}{\partial x_K \partial x_K} \end{pmatrix} = \nabla^2_{\mathbf{x}} y$$

2 Eigenvalue Inequalities

2.1 The Eckart–Young Theorem

Lemma Let $\mathbf{A}_1, \mathbf{A}_2 \in \mathbb{R}^{m\times n}$ be given and let $q = \min\{m,n\}$. With the decreasingly ordered singular values of $\mathbf{A}_1, \mathbf{A}_2$ and $\mathbf{A}_1+\mathbf{A}_2$, for $1 \le i,j \le q$ and $i+j \le q+1$,

$$\sigma_{i+j-1}(\mathbf{A}_1+\mathbf{A}_2) \le \sigma_i(\mathbf{A}_1) + \sigma_j(\mathbf{A}_2)$$

Proof. Let $\mathbf{A}_1 = \mathbf{U}_1\Sigma_1\mathbf{V}_1^\top$ and $\mathbf{A}_2 = \mathbf{U}_2\Sigma_2\mathbf{V}_2^\top$ be singular value decompositions of $\mathbf{A}_1$ and $\mathbf{A}_2$, where $\mathbf{U}_1 = \left(U_{1i}\right)_{1\le i\le m}$ and $\mathbf{U}_2 = \left(U_{2i}\right)_{1\le i\le m}$. Let $i,j \in \mathbb{Z}^+$ with $1 \le i,j \le q$ and $i+j \le q+1$.

Define $S_1 = span\{U_{1i}, U_{1(i+1)}, \cdots, U_{1n}\}$ and $S_2 = span\{U_{2j}, U_{2(j+1)}, \cdots, U_{2n}\}$. Since $\dim S_1 = n-i+1$ and $\dim S_2 = n-j+1$,

$$\begin{aligned}
v &= \dim(S_1 \cap S_2) \\
&= \dim S_1 + \dim S_2 - \dim(S_1 \cup S_2) \\
&= (q-i+1) + (q-j+1) - \dim(S_1 \cup S_2) \\
&\ge (q-i+1) + (q-j+1) - q \qquad (1) \\
&= q - (i+j-1) + 1 \\
&\ge 1
\end{aligned}$$

and from (1) we have

$$q - v + 1 \le i + j - 1$$

Therefore,

$$\begin{aligned}
\sigma_{i+j-1}(\mathbf{A}_1+\mathbf{A}_2) &\le \sigma_{q-v+1}(\mathbf{A}_1+\mathbf{A}_2) \\
&= \min_{\substack{S\subset\mathbb{C}^n \\ \dim S = v}} \max_{\substack{x\in S \\ \|x\|_2=1}} \|(\mathbf{A}_1+\mathbf{A}_2)x\|_2 \qquad \text{(Courant-Fischer Min-Max Theorem)} \\
&\le \max_{\substack{x\in S_1\cap S_2 \\ \|x\|_2=1}} \|(\mathbf{A}_1+\mathbf{A}_2)x\|_2 \\
&\le \max_{\substack{x\in S_1\cap S_2 \\ \|x\|_2=1}} \|\mathbf{A}_1 x\|_2 + \max_{\substack{x\in S_1\cap S_2 \\ \|x\|_2=1}} \|\mathbf{A}_2 x\|_2 \\
&\le \max_{\substack{x\in S_1 \\ \|x\|_2=1}} \|\mathbf{A}_1 x\|_2 + \max_{\substack{x\in S_2 \\ \|x\|_2=1}} \|\mathbf{A}_2 x\|_2 \\
&= \sigma_i(\mathbf{A}_1) + \sigma_j(\mathbf{A}_2)
\end{aligned}$$

□

If $\mathbf{A}$ and $\mathbf{B}_k$ are both $(J \times K)$-matrices, and we plan on using $\mathbf{B}_k$ with reduced rank $r(\mathbf{B}_k) = k$ to approximate $\mathbf{A}$ with full rank $r(\mathbf{A}) = \min(J,K)$, then the Eckart–Young (1936) Theorem states that

$$\lambda_j((\mathbf{A}-\mathbf{B}_k)(\mathbf{A}-\mathbf{B}_k)^\top) \geqslant \lambda_{j+k}(\mathbf{A}\mathbf{A}^\top)$$

with equality if

$$\mathbf{B}_k = \sum_{i=1}^{k} \lambda_i^{\frac{1}{2}} u_i v_i^\top$$

where $\sqrt{\lambda_i}$, $u_i, v_i (1 \leq i \leq k)$ are the first k components of the singular values and vectors of $\mathbf{A}$, with $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{r(\mathbf{A})}$.

Proof.

Let $A \in \mathbb{R}^{m \times n}$ be a real (possibly rectangular) matrix with $m \geq n$. Suppose that

$$A = U\Sigma V^\top$$

is the singular value decomposition of A. We claim that the best rank k approximation to A in the Frobenius norm, denoted by $\|\cdot\|_F$, is given by

$$A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^\top$$

where u_i and v_i denote the ith column of U and V, respectively.

First, note that we have

$$\begin{aligned}
\|A - A_k\|_F^2 &= \left\| \sum_{i=k+1}^{n} \sigma_i u_i v_i^\top \right\|_F^2 \\
&= \left\| U_{(k)} \Sigma_{(k)} V_{(k)}^\top \right\|_F^2 \\
&= trace[(U_{(k)}\Sigma_{(k)}V_{(k)}^\top)^\top U_{(k)}\Sigma_{(k)}V_{(k)}^\top] && \text{(Definition of F-norm)} \\
&= trace(V_{(k)}\Sigma_{(k)}^2 V_{(k)}^\top) \\
&= trace(\Sigma_{(k)}^2 V_{(k)}^\top V_{(k)}) && \text{(Cyclic Permutations Invariance of Trace)} \\
&= \sum_{i=k+1}^{n} \sigma_i^2
\end{aligned}$$

Therefore, we need to show that if $B_k = XY^\top$ where X and Y have k columns then

$$\|A - A_k\|_F^2 = \sum_{i=k+1}^{n} \sigma_i^2 \leq \|A - B_k\|_F^2$$

By the triangle inequality with the spectral norm, if $A = A' + A''$ then $\sigma_1(A) \leq \sigma_1(A') + \sigma_1(A'')$. Suppose A'_k and A''_k respectively denote the rank k approximation to A' and A'' by SVD method described above. Then, for any $i, j \geq 1$

$$\begin{aligned}
\sigma_i(A') + \sigma_j(A'') &\geq \sigma_{i+j-1}(A' + A'') \\
&= \sigma_{i+j-1}(A).
\end{aligned}$$

Since $\sigma_{k+1}(B_k) = 0$, when $A' = A - B_k$ and $A'' = B_k$ we conclude that for $i \geq 1, j = k+1$

$$\sigma_i(A - B_k) \geq \sigma_{k+i}(A).$$

$\because$

$$\begin{aligned}\lambda_i[(A-B_k)(A-B_k)^\top] &= \lambda_i[(U'\Sigma'V'^\top)(U'\Sigma'V'^\top)^\top] \\ &= \lambda_i(V'^\top\Sigma'^2V'^\top) \\ &= \lambda_i(\Sigma'^2) \\ &= \sigma_i^2(A-B_k)\end{aligned}$$

$\therefore$

$$\lambda_i[(A-B_k)(A-B_k)^\top] \geqslant \lambda_{j+k}(AA^\top) \qquad (j = 1,2,\cdots,r(A)-k)$$

$$\|A-B_k\|_F^2 = \sum_{i=1}^{n}\sigma_i^2(A-B_k) \geq \sum_{i=k+1}^{n}\sigma_i^2(A) = \|A-A_k\|_F^2$$

as required. □

2.2 The Courant–Fischer Min-Max Theorem

Let A be a $n \times n$ Hermitian matrix with eigenvalues $\lambda_1 \leq \ldots \leq \lambda_k \leq \ldots \leq \lambda_n$ then

$$\lambda_k = \min_U\{\max_x\{R_A(x) \mid x \in U \text{ and } x \neq 0\} \mid \dim(U) = k\}$$

and

$$\lambda_k = \max_U\{\min_x\{R_A(x) \mid x \in U \text{ and } x \neq 0\} \mid \dim(U) = n-k+1\}$$

in particular,

$$\lambda_1 \leq R_A(x) \leq \lambda_n \quad \forall x \in \mathbf{C}^n\backslash\{0\}$$

where

$$R_A(x) = \frac{(Ax,x)}{(x,x)},$$

and these bounds are attained when x is an eigenvector of the appropriate eigenvalues.

Also note that the simpler formulation for the maximal eigenvalue λ_n is given by:

$$\lambda_n = \max\{R_A(x) : x \neq 0\}.$$

Similarly, the minimal eigenvalue λ_1 is given by:

$$\lambda_1 = \min\{R_A(x) : x \neq 0\}.$$

Proof.

Since the matrix A is Hermitian, it is diagonalizable and we can choose an orthonormal basis of eigenvectors $\{u_1,\ldots,u_n\}$. That is, u_i is an eigenvector for the eigenvalue λ_i and such that $(u_i,u_i)=1$ and $(u_i,u_j)=0$ for all $i\neq j$.

If U is a subspace of dimension k then its intersection with the subspace $span\{u_k,\ldots,u_n\}$ isn't zero (by simply checking dimensions) and hence there exists a vector $v\neq 0$ in this intersection that we can write as

$$v=\sum_{i=k}^{n}\alpha_i u_i$$

and whose Rayleigh quotient is

$$R_A(v)=\frac{\sum_{i=k}^{n}\lambda_i\alpha_i^2}{\sum_{i=k}^{n}\alpha_i^2}\geq\lambda_k$$

(as all $\lambda_i\geq\lambda_k$ for $i=k,..,n$) and hence

$$\max_x\{R_A(x)\mid x\in U \text{ and } x\neq 0\}\geq\lambda_k$$

Since this is true for all U, we can conclude that

$$\min_U\{\max_x\{R_A(x)\mid x\in U \text{ and } x\neq 0\}\mid \dim(U)=k\}\geq\lambda_k$$

This is one inequality. To establish the other inequality, chose the specific k-dimension space $V=span\{u_1,\ldots,u_k\}$, for which

$$\max_x\{R_A(x)\mid x\in V \text{ and } x\neq 0\}\leq\lambda_k$$

because λ_k is the largest eigenvalue in V. Therefore, also

$$\min_U\{\max_x\{R_A(x)\mid x\in U \text{ and } x\neq 0\}\mid \dim(U)=k\}\leq\lambda_k$$

In the case where U is a subspace of dimension $n-k+1$, we proceed in a similar fashion: Consider the subspace of dimension k, $span\{u_1,\ldots,u_k\}$. Its intersection with the subspace U isn't zero (by simply checking dimensions) and hence there exists a vector v in this intersection that we can write as

$$v=\sum_{i=1}^{k}\alpha_i u_i$$

and whose Rayleigh quotient is

$$R_A(v)=\frac{\sum_{i=1}^{k}\lambda_i\alpha_i^2}{\sum_{i=1}^{k}\alpha_i^2}\leq\lambda_k$$

and hence

$$\min_x\{R_A(x)\mid x\in U\}\leq\lambda_k$$

Since this is true for all U, we can conclude that

$$\max_U\{\min_x\{R_A(x)\mid x\in U \text{ and } x\neq 0\}\mid \dim(U)=n-k+1\}\leq\lambda_k$$

Again, this is one part of the equation. To get the other inequality, note again that the eigenvector u of λ_k is contained in $U=span\{u_k,\ldots,u_n\}$ so that we can conclude the equality. □

2.3 The Hoffman–Wielandt Theorem

Suppose $\mathbf{A}$ and $\mathbf{B}$ are $(J \times J)$-matrices with $\mathbf{A}-\mathbf{B}$ symmetric. Suppose $\mathbf{A}$ and $\mathbf{B}$ have eigenvalues $\{\lambda_j(A)\}$ and $\{\lambda_j(B)\}$, respectively. Hoffman and Wielandt (1953) showed that

$$\sum_{j=1}^{J}(\lambda_j(\mathbf{A})-\lambda_j(\mathbf{B}))^2 \leq tr\{(\mathbf{A}-\mathbf{B})(\mathbf{A}-\mathbf{B})^\top\}.$$

Proof.

Symmetric matrices are orthogonally diagonalizable. Let $\mathbf{A}=\mathbf{U}\Lambda\mathbf{U}^\top$, $\mathbf{B}=\mathbf{V}\Sigma\mathbf{V}^\top$ where $\mathbf{U},\mathbf{V}$ are real orthogonal. The eigenvalues in the two diagonal matrices Λ, Σ are arranged in descending order. Let $\mathbf{W}=\mathbf{U}^\top\mathbf{V}$. Hence

$$\mathbf{A}-\mathbf{B}=\mathbf{U}\Lambda\mathbf{U}^\top-\mathbf{V}\Sigma\mathbf{V}^\top$$

$\because$

$$\Lambda\mathbf{U}^\top\mathbf{V}-\mathbf{U}^\top\mathbf{V}\Sigma=\mathbf{U}^\top(\mathbf{A}-\mathbf{B})\mathbf{V}$$

i.e.

$$\Lambda\mathbf{W}-\mathbf{W}\Sigma=\mathbf{U}^\top(\mathbf{A}-\mathbf{B})\mathbf{V}$$

$\therefore$

$$\begin{aligned}\|\Lambda\mathbf{W}-\mathbf{W}\Sigma\|_F^2 &= \|\mathbf{U}^\top(\mathbf{A}-\mathbf{B})\mathbf{V}\|_F^2\\ &= \|\mathbf{A}-\mathbf{B}\|_F^2\\ &= trace[(\mathbf{A}-\mathbf{B})^2]\end{aligned}$$

$\because$ $\mathbf{W}$ is an orthonormal matrix

$\therefore$

$$\|\Lambda\mathbf{W}-\mathbf{W}\Sigma\|_F^2=\sum_i\sum_j|\lambda_i(\mathbf{A})-\lambda_j(\mathbf{B})|^2|w_{ij}|^2$$

Now the matrix $\mathbf{P}$ with $\mathbf{P}_{ij}=|\lambda_i(\mathbf{A})-\lambda_j(\mathbf{B})|^2$ is real and from the ordering of the $\lambda_i(\mathbf{A})$ and $\lambda_i(\mathbf{B})$, its principal diagonal is minimal. Further, since $\mathbf{W}$ is unitary, the matrix $\mathbf{Z}$ with $Z_{ij}=|w_{ij}|^2$ is a doubly stochastic matrix. Hence by Birkhoff-von Neumann theorem, and equation

$$\begin{aligned}\sum_{j=1}^{J}(\lambda_j(\mathbf{A})-\lambda_j(\mathbf{B}))^2 &\leq \sum_i\sum_j|\lambda_i(\mathbf{A})-\lambda_j(\mathbf{B})|^2|q_{ij}|^2\\ &\leq tr\{(\mathbf{A}-\mathbf{B})(\mathbf{A}-\mathbf{B})^\top\}\end{aligned}$$

□

2.4 Poincaré Separation Theorem

Let $\mathbf{A}$ be a real symmetric $(J \times J)$-matrix and let $\mathbf{U}$ be a $(J \times k)$-matrix, $k \leq J$, such that $\mathbf{U}^\top \mathbf{U} = \mathbf{I}_k$. Then,

$$\lambda_{J-j+1}(\mathbf{A}) \leq \lambda_j(\mathbf{U}^\top \mathbf{A}\mathbf{U}) \leq \lambda_j(\mathbf{A})$$

with equality if the columns of $\mathbf{U}$ are the first k eigenvectors of $\mathbf{A}$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_J$.

Proof.

Let $1 \leq j \leq k \leq J$ and let $\mathbf{R}$ be a semi-orthonormal $n \times (j-1)$ matrix whose columns are eigenvectors of $\mathbf{A}$ associated with $\lambda_1, \lambda_2, \ldots, \lambda_{j-1}$. Let $\mathbf{T}$ be a semi-orthonormal $n \times (J-j+1)$ matrix whose columns are eigenvectors of $\mathbf{A}$ associated with $\lambda_j, \lambda_{j+1}, \ldots, \lambda_J$. Let $\mathbf{A} = \mathbf{V}\Lambda\mathbf{V}^\top$ and $\mathbf{U}^\top\mathbf{A}\mathbf{U} = \mathbf{W}\Sigma\mathbf{W}^\top$ be the spectral decompositions where $\mathbf{V} = \begin{pmatrix} \mathbf{R} & \mathbf{T} \end{pmatrix}$ and $\mathbf{W} = \begin{pmatrix} w_1 & \cdots & w_k \end{pmatrix}$.

Then, for $1 \leq j \leq k$,

$$\begin{aligned}
\lambda_j(\mathbf{A}) &= \max_{\mathbf{T}^\top x = 0} \frac{x^\top \mathbf{A} x}{x^\top x} && \text{(Courant-Fischer Min-Max Theorem)} \\
&\geq \max_{\substack{\mathbf{T}^\top x = 0 \\ x = \mathbf{U}y}} \frac{x^\top \mathbf{A} x}{x^\top x} \\
&= \max_{\mathbf{T}^\top \mathbf{U} y = 0} \frac{y^\top \mathbf{U}^\top \mathbf{A}\mathbf{U} y}{y^\top y} \\
&\geq \lambda_j(\mathbf{U}^\top \mathbf{A}\mathbf{U}) && (\text{Range}(\mathbf{U}^\top\mathbf{T}) \subset \text{Range}(\mathbf{T}))
\end{aligned}$$

with equlity if $\mathbf{U} = \mathbf{R}$.

Also, let $J-k+1 \leq J-k+j \leq J$ and let $\mathbf{R}'$ be a semi-orthonormal $n \times (J-j+1)$ matrix whose columns are eigenvectors of $\mathbf{A}$ associated with $\lambda_1, \lambda_2, \ldots, \lambda_{J-j}$. Let $\mathbf{T}'$ be a semi-orthonormal $n \times (j-1)$ matrix whose columns are eigenvectors of $\mathbf{A}$ associated with $\lambda_{J-j+1}, \lambda_{J-j+2}, \ldots, \lambda_J$.

$$\begin{aligned}
\lambda_{J-k+j}(\mathbf{A}) &= \min_{\mathbf{R}'^\top x = 0} \frac{x^\top \mathbf{A} x}{x^\top x} && \text{(Courant-Fischer Min-Max Theorem)} \\
&\leq \min_{\substack{\mathbf{R}'^\top x = 0 \\ x = \mathbf{U}y}} \frac{x^\top \mathbf{A} x}{x^\top x} \\
&= \min_{\mathbf{R}'^\top \mathbf{U} y = 0} \frac{y^\top \mathbf{U}^\top \mathbf{A}\mathbf{U} y}{y^\top y} \\
&\leq \lambda_j(\mathbf{U}^\top \mathbf{A}\mathbf{U}) && (\text{Range}(\mathbf{U}^\top\mathbf{R}') \subset \text{Range}(\mathbf{R}'))
\end{aligned}$$

with equlity if $\mathbf{U} = \mathbf{T}'$. □

3 Normal Distribution

3.1 Univariate Normal Distribution

1. Standard normal distribution

   The standard normal random variable X is defined to have probability density function

   $$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

   denoted by $X \sim N(0,1)$.

2. General normal distribution

   If $U \sim N(0,1)$ and $X = \sigma U + \mu$, then $X \sim N(\mu, \sigma^2)$, where $\sigma, \mu \in \mathbb{R}$.

3.2 Multivariate Normal Distribution

3.2.1 Definition

A random vector $\mathbf{Y} = \begin{pmatrix} Y_1 & Y_2 \dots & Y_m \end{pmatrix}^\top$ is said to have the m-dimension multivariate normal distribution if it satisfies the following equivalent conditions.

(1) There exists a random n-vector $\mathbf{X}$, whose components are independent standard normal random variables, a m-vector $\boldsymbol{\mu}$, and a $m \times n$ matrix $\mathbf{A}$, such that

$$\mathbf{Y} = \mathbf{A}\mathbf{X} + \boldsymbol{\mu}$$

Here the covariance matrix $\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}^\top$ and $\mathbf{Y} \sim N_m(\boldsymbol{\mu}, \mathbf{A}\mathbf{A}^\top)$.

(2) Every linear combination of its components $Z = \sum_{i=1}^{m} a_i Y_i$ is normally distributed. That is, for any constant vector $a \in \mathbb{R}^m$, the random variable $Z = a^\top \mathbf{Y}$ has a univariate normal distribution, where an univariate normal distribution with zero variance is a point mass on its mean.

(3) There is a m-vector $\boldsymbol{\mu}$ and a symmetric, positive semidefinite $m \times m$ matrix $\boldsymbol{\Sigma}$, such that the characteristic function of $\mathbf{Y}$ is

$$\varphi_{\mathbf{Y}}(\mathbf{u}) = e^{i\mathbf{u}^T\boldsymbol{\mu} - \frac{1}{2}\mathbf{u}^T\boldsymbol{\Sigma}\mathbf{u}}$$

The covariance matrix is allowed to be singular (in which case the corresponding distribution has no density).

3.2.2 Properties

(1) Suppose that $\mathbf{Y} \sim N_m(\boldsymbol{\mu}, \mathbf{A}\mathbf{A}^\top)$, $\mathbf{Z} = \mathbf{B}\mathbf{Y} + \boldsymbol{\nu}$ where $\mathbf{B} \in \mathbb{R}^{l \times m}$ and $\boldsymbol{\nu} \in \mathbb{R}^l$, then

$$\mathbf{Z} \sim N_l(\mathbf{B}\boldsymbol{\mu} + \boldsymbol{\nu}, \mathbf{B}\mathbf{A}\mathbf{A}^\top\mathbf{B}^\top)$$

(2) Suppose that $\mathbf{Y} \sim N_m(\boldsymbol{\mu}, \mathbf{A}\mathbf{A}^\top)$, $\mathbf{Y}_1, \boldsymbol{\mu}_1 \in \mathbf{R}^r$, $\mathbf{Y}_2, \boldsymbol{\mu}_1 \in \mathbf{R}^{m-r}$, $\mathbf{V}_{11} \in \mathbb{R}^{r\times r}$, $\mathbf{V}_{12} \in \mathbb{R}^{r\times(m-r)}$, $\mathbf{V}_{21} \in \mathbb{R}^{(m-r)\times r}$, $\mathbf{V}_{22} \in \mathbb{R}^{(m-r)\times(m-r)}$,

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{bmatrix} \qquad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \qquad \mathbf{A}\mathbf{A}^\top = \begin{bmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{bmatrix}$$

then

$$\mathbf{Y}_1 \sim N_r(\boldsymbol{\mu}_1, \mathbf{V}_{11}) \qquad \mathbf{Y}_2 \sim N_r(\boldsymbol{\mu}_2, \mathbf{V}_{22})$$

Proof.

From Property 1, let $\mathbf{B} = \begin{pmatrix} \mathbf{I}_r & \mathbf{0}_{r\times(m-r)} \end{pmatrix}$ and $\boldsymbol{\nu} = 0$, then $\mathbf{Y}_1 \sim N_r(\boldsymbol{\mu}_1, \mathbf{V}_{11})$.

Let $\mathbf{B} = \begin{pmatrix} \mathbf{0}_r & \mathbf{I}_{r\times(m-r)} \end{pmatrix}$ and $\boldsymbol{\nu} = 0$, then $\mathbf{Y}_2 \sim N_r(\boldsymbol{\mu}_2, \mathbf{V}_{22})$.

□

(3) $\mathbf{Y}$ is a m-dimension multivariate normal random vector if and only if $\forall\ \mathbf{a} \in \mathbb{R}^m$, $\mathbf{a}^\top\mathbf{Y}$ is an univariate normal random variable.

Proof.

$\Longrightarrow$

Suppose that $\mathbf{Y} \sim N_m(\boldsymbol{\mu}, \mathbf{A}\mathbf{A}^\top)$. From Property 1, $\forall\ \mathbf{a} \in \mathbb{R}^m$, $\mathbf{a}^\top\mathbf{Y} \sim N_1(\mathbf{a}^\top\boldsymbol{\mu}, \mathbf{a}^\top\mathbf{A}\mathbf{A}^\top\mathbf{a})$, where an univariate normal distribution with zero variance is a point mass on its mean.

$\Longleftarrow$

Suppose that $\mathbb{E}\mathbf{Y} = \boldsymbol{\mu}$, $Var\mathbf{Y} = \boldsymbol{\Sigma}$.

$\because$ $\forall\mathbf{a} \in \mathbb{R}^m$, $\mathbf{a}^\top\mathbf{Y}$ is an univariate normal random variable.

$$\begin{aligned} \mathbb{E}(\mathbf{a}^\top\mathbf{Y}) &= \mathbf{a}^\top\mathbb{E}(\mathbf{Y}) \\ &= \mathbf{a}^\top\boldsymbol{\mu} \\ Var(\mathbf{a}^\top\mathbf{Y}) &= \mathbf{a}^\top\mathbb{E}(\mathbf{Y})\mathbf{a} \\ &= \mathbf{a}^\top\boldsymbol{\Sigma}\mathbf{a} \end{aligned}$$

i.e. $X = \mathbf{a}^\top\mathbf{Y} \sim N(\mathbf{a}^\top\boldsymbol{\mu}, \mathbf{a}^\top\boldsymbol{\Sigma}\mathbf{a})$.

$\therefore$

$$\mathbb{E}e^{itX} = \varphi_X(t) = e^{it\mathbf{a}^\top\boldsymbol{\mu} - \frac{1}{2}t^2\mathbf{a}^\top\boldsymbol{\Sigma}\mathbf{a}}$$

$$\begin{aligned} \varphi_X(1) &= e^{i\mathbf{a}^\top\boldsymbol{\mu} - \frac{1}{2}\mathbf{a}^\top\boldsymbol{\Sigma}\mathbf{a}} \\ &= \mathbb{E}e^{iX} \\ &= \mathbb{E}e^{i\mathbf{a}^\top\mathbf{Y}} \end{aligned}$$

$$= \varphi_Y(\mathbf{a})$$

$\because \quad \forall\, \mathbf{a} \in \mathbb{R}^m$,

$$\varphi_Y(\mathbf{a}) = \mathbb{E}e^{i\mathbf{a}^\top \mathbf{Y}} = e^{i\mathbf{a}^\top \boldsymbol{\mu} - \frac{1}{2}\mathbf{a}^\top \boldsymbol{\Sigma} \mathbf{a}}$$

is the characteristic function of random vector with distribution $N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

$\therefore \quad \mathbf{Y}$ is a m-dimension normal random vecter $\square$

(4) If $\mathbf{X}$ and $\mathbf{Y}$ are independent m-dimension random vectors whose sum $\mathbf{X}+\mathbf{Y}$ is a m-dimension normal random vector, then both X and Y must be m-dimension normal random vectors.

Proof.

Lemma : If $\mathbf{X}$ is a random vector such that $\mathbb{E}e^{\lambda \mathbf{X}^\top \mathbf{X}} < \infty$ for some $\lambda > 0$ and $\phi_{\mathbf{X}}(\mathbf{z}) \neq 0$ for all $\mathbf{z} = \mathbf{x} + i\mathbf{y} \in \mathbb{C}^m$, then $\mathbf{X}$ is normal.

Since $f(\mathbf{z}) = \log \phi(\mathbf{z})$ is well defined,

$$\begin{aligned} Rf(\mathbf{z}) &= \log |\phi(\mathbf{z})| \\ &\leq \log \mathbb{E}e^{|\mathbf{y}^\top \mathbf{X}|} \\ &\leq \log \mathbb{E}\left[e^{\frac{1}{2}\left(\lambda \mathbf{X}^\top \mathbf{X} + \frac{\mathbf{y}^\top \mathbf{y}}{\lambda}\right)} \right] \\ &= \log \mathbb{E}(Ce^{\frac{\mathbf{y}^\top \mathbf{y}}{2\lambda}}) \\ &= C' + \frac{\mathbf{y}^\top \mathbf{y}}{2\lambda} \end{aligned}$$

From Liouville Theorem, $f(\mathbf{z})$ is quadratic polynomial in z, i.e. $f(\mathbf{z}) = A + B\mathbf{z} + C\mathbf{z}^2$. It is easy to see that $A = 0$, $B = i\mathbb{E}\mathbf{X}$, $C = -\frac{1}{2}Var(\mathbf{X})$. Therefore, $\mathbf{X}$ is normal.

Without loss of generality, suppose that

$$\begin{aligned} \mathbb{E}\mathbf{X} = \mathbb{E}\mathbf{Y} &= \mathbf{0} \\ \mathbb{E}(\mathbf{X}+\mathbf{Y}) &= 0 \\ Var\mathbf{X} = Var\mathbf{Y} &= \mathbf{I} \\ Var(\mathbf{X}+\mathbf{Y}) &= 2\mathbf{I} \end{aligned}$$

$\because$

$$\mathbb{E}e^{\lambda(\mathbf{X}+\mathbf{Y})^\top (\mathbf{X}+\mathbf{Y})} < \infty$$

$\therefore$

$$\begin{aligned} \mathbb{E}e^{\lambda \mathbf{X}^\top \mathbf{X}} &< \infty \\ \mathbb{E}e^{\lambda \mathbf{Y}^\top \mathbf{Y}} &< \infty \end{aligned}$$

$\because$ $\mathbf{X}$ and $\mathbf{Y}$ are independent

$\therefore$

$$\phi_{\mathbf{X}+\mathbf{Y}}(\mathbf{t}) = \phi_{\mathbf{X}}(\mathbf{t})\phi_{\mathbf{Y}}(\mathbf{t}) \neq 0 \qquad (\forall\ \mathbf{t} \in \mathbb{R}^m)$$

$\therefore$ From Lemma, $\mathbf{X}$ and $\mathbf{Y}$ are m-dimension normal random vectors

□

(5) Suppose that $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n$ are independent m-dimension random vectors, $\mathbf{a}, \mathbf{b} \in \mathbb{R}^m$ and

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 & \mathbf{Y}_2 & \cdots & \mathbf{Y}_n \end{pmatrix} \qquad \mathbf{Z}_1 = \mathbf{Y}\mathbf{a} \qquad \mathbf{Z}_2 = \mathbf{Y}\mathbf{b}$$

If $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n \overset{iid}{\sim} N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathbf{a}^\top\mathbf{b}$, then $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are independent.

If $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are independent and $a_i b_i \neq 0$, then $\mathbf{Y}_i$ is a m-dimension random vector.

3.2.3 Joint Normality

1. Normality and Independence

   If $\mathbf{X}$ and $\mathbf{Y}$ are normally distributed and independent, this implies they are "jointly normally distributed", i.e., the pair $(\mathbf{X}, \mathbf{Y})$ must have multivariate normal distribution.

   However, a pair of jointly normally distributed variables need not be independent (would only be so if uncorrelated, $\rho = 0$).

2. Marginal Normality and Jointly Bivariate Normality

   The fact that two random variables $\mathbf{X}$ and $\mathbf{Y}$ both have a normal distribution does not imply that the pair $(\mathbf{X}, \mathbf{Y})$ has a joint normal distribution. A simple example is one in which $\mathbf{X} \sim N(0,1)$, and

   $$\mathbf{Y} = \begin{cases} \mathbf{X} & , |\mathbf{X}| > c \\ -\mathbf{X} & , |\mathbf{X}| < c \end{cases}$$

   where $c > 0$. There are similar counterexamples for more than two random variables. In general, they sum to a mixture model.

3. Correlations and Independence

   In general, random variables may be uncorrelated but statistically dependent. But if a random vector has a multivariate normal distribution then any two or more of its components that are uncorrelated are independent. This implies that any two or more of its components that are pairwise independent are independent.

   But, as pointed out just above, it is not true that two random variables that are (separately, marginally) normally distributed and uncorrelated are independent.

3.3 Independence And Correlation

3.4 Maximum Likelihood Estimator

$$L(\boldsymbol{\mu},\mathbf{V}) = \prod_{i=1}^{n} \frac{2}{(2\pi)^{\frac{m}{2}}|\mathbf{V}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{X}_i-\boldsymbol{\mu})^\top \mathbf{V}^{-1}(\mathbf{X}_i-\boldsymbol{\mu})}$$

3.5 James-Stein Estimator

3.5.1 Bayes Rule

Suppose that $\mathbf{z}|\boldsymbol{\mu} \sim N_n(\boldsymbol{\mu},\mathbf{I})$, $\boldsymbol{\mu} \sim N_n(0,A\mathbf{I})$, where $A \in \mathbb{R}$, then

$$\mathbf{z} \sim N_n(0,(A+1)\mathbf{I})$$
$$\boldsymbol{\mu}|\mathbf{z} \sim N_n(B\mathbf{z},B\mathbf{I})$$

where $B = \dfrac{A}{A+1}$.

The MLE of $\boldsymbol{\mu}$ is given by

$$\hat{\boldsymbol{\mu}}_{MLE} = \mathbf{z}$$

The Bayes estimator of $\boldsymbol{\mu}$ is given by

$$\hat{\boldsymbol{\mu}}_{Bayes} = B\mathbf{z} = \left(1 - \frac{1}{A+1}\right)\mathbf{z}$$

Proof.

$\because$ $\mathbf{z}|\boldsymbol{\mu} \sim N_n(\boldsymbol{\mu},\mathbf{I}), \boldsymbol{\mu} \sim N_n(0,A\mathbf{I})$

$\therefore$

$$f_{\mathbf{z}|\boldsymbol{\mu}}(z|\mu) = \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}(z-\mu)^\top(z-\mu)}$$
$$f_{\boldsymbol{\mu}}(\mu) = \frac{1}{(2\pi)^{\frac{n}{2}}|A\mathbf{I}|^{\frac{1}{2}}} e^{-\frac{1}{2}\mu^\top(A\mathbf{I})^{-1}\mu} = \frac{1}{(2\pi A)^{\frac{n}{2}}} e^{-\frac{1}{2A}\mu^\top\mu}$$

$\therefore$

$$\begin{aligned} f_{\mathbf{z},\boldsymbol{\mu}}(z,\mu) &= f_{\mathbf{z}|\boldsymbol{\mu}}(z|\mu) f_{\boldsymbol{\mu}}(\mu) \\ &= \frac{1}{(2\pi)^n A^{\frac{n}{2}}} e^{-\frac{1}{2}(z-\mu)^\top(z-\mu)-\frac{1}{2A}\mu^\top\mu} \\ &= \frac{1}{(2\pi)^n A^{\frac{n}{2}}} e^{-\frac{1}{2}(\sqrt{\frac{A}{A+1}}z-\sqrt{\frac{A+1}{A}}\mu)^\top(\sqrt{\frac{A}{A+1}}z-\sqrt{\frac{A+1}{A}}\mu)-\frac{1}{2(A+1)}z^\top z} \end{aligned}$$

$$
\begin{aligned}
f_{\mathbf{z}}(z) &= \iint_{\mathbb{R}^n} f_{\mathbf{z},\boldsymbol{\mu}}(z,\mu)\mathrm{d}\mu \\
&\overset{y=\sqrt{\frac{A+1}{A}}\mu}{=\!=\!=} \frac{1}{(2\pi)^n A^{\frac{n}{2}}} e^{-\frac{1}{2(A+1)}z^\top z}\left(\frac{A}{A+1}\right)^{\frac{n}{2}} \iint_{\mathbb{R}^n} e^{-\frac{1}{2}(\sqrt{\frac{A}{A+1}}z-y)^\top(\sqrt{\frac{A}{A+1}}z-y)}\mathrm{d}y \\
&= \frac{1}{(2\pi)^{\frac{n}{2}}(A+1)^{\frac{n}{2}}} e^{-\frac{1}{2(A+1)}z^\top z} \frac{1}{(2\pi)^{\frac{n}{2}}} \iint_{\mathbb{R}^n} e^{-\frac{1}{2}y^\top y}\mathrm{d}y \\
&= \frac{1}{(2\pi)^{\frac{n}{2}}(A+1)^{\frac{n}{2}}} e^{-\frac{1}{2(A+1)}z^\top z} \frac{1}{(2\pi)^{\frac{n}{2}}} \cdot (2\pi)^{\frac{n}{2}} \\
&= \frac{1}{[2\pi(A+1)]^{\frac{n}{2}}} e^{-\frac{1}{2(A+1)}z^\top z} \\
f_{\boldsymbol{\mu}|\mathbf{z}}(\mu|z) &= \frac{f_{\mathbf{z},\boldsymbol{\mu}}(z,\mu)}{f_{\mathbf{z}}(z)} \\
&= \frac{(A+1)^{\frac{n}{2}}}{(2\pi)^n A^{\frac{n}{2}}} e^{-\frac{1}{2}(\sqrt{\frac{A}{A+1}}z-\sqrt{\frac{A+1}{A}}\mu)^\top(\sqrt{\frac{A}{A+1}}z-\sqrt{\frac{A+1}{A}}\mu)} \\
&= \frac{(A+1)^{\frac{n}{2}}}{(2\pi)^n A^{\frac{n}{2}}} e^{-\frac{A+1}{2A}(\frac{A}{A+1}z-\mu)^\top(\frac{A}{A+1}z-\mu)}
\end{aligned}
$$

$\therefore$

$$\mathbf{z} \sim N_n(0,(A+1)\mathbf{I})$$

$$\boldsymbol{\mu}|\mathbf{z} \sim N_n(B\mathbf{z},B\mathbf{I})$$

where $B = \dfrac{A}{A+1}$ □

3.5.2 Loss and Risk

We use total squared error loss to measure the error of estimating $\boldsymbol{\mu}$ by $\hat{\boldsymbol{\mu}}$,

$$L(\boldsymbol{\mu},\hat{\boldsymbol{\mu}}) = \|\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}\|_2^2$$

with the corresponding risk function being the expected value of $L(\boldsymbol{\mu},\hat{\boldsymbol{\mu}})$ for a given $\boldsymbol{\mu}$,

$$R(\boldsymbol{\mu}) = \mathbb{E}_{\mathbf{z}|\boldsymbol{\mu}}[L(\boldsymbol{\mu},\hat{\boldsymbol{\mu}})] = \mathbb{E}_{\mathbf{z}|\boldsymbol{\mu}}[\|t(z)-\boldsymbol{\mu}\|_2^2]$$

where $\hat{\boldsymbol{\mu}} = t(\mathbf{z})$ and the expectation is taken with respect to $\mathbf{z}|\boldsymbol{\mu}$.

We have

$$R_{MLE}(\boldsymbol{\mu}) = n$$

$$R_{Bayes}(\boldsymbol{\mu}) = (1-B)^2\|\boldsymbol{\mu}\|^2 + nB^2$$

and overall Bayes risk is given by

$$\mathbb{E}_{\boldsymbol{\mu}}[R_{Bayes}(\boldsymbol{\mu})] = nB$$

Proof.

Since

$$\mathbf{z}|\boldsymbol{\mu} \sim N(\boldsymbol{\mu}, \mathbf{I}_n)$$

we have

$$\begin{aligned}
R_{MLE}(\boldsymbol{\mu}) &= \mathbb{E}_{\mathbf{z}|\boldsymbol{\mu}}[\|\boldsymbol{\mu} - \mathbf{z}\|_2^2] \\
&= \mathbb{E}_{\mathbf{z}|\boldsymbol{\mu}}(\mathbf{z}^\top \mathbf{z}) - 2\boldsymbol{\mu}^\top \mathbb{E}_{\mathbf{z}|\boldsymbol{\mu}}(\mathbf{z}) + \boldsymbol{\mu}^\top \boldsymbol{\mu} \\
&= \mathbb{E}_{\mathbf{z}|\boldsymbol{\mu}}[tr(\mathbf{z}^\top \mathbf{z})] - 2\boldsymbol{\mu}^\top \boldsymbol{\mu} + \boldsymbol{\mu}^\top \boldsymbol{\mu} \\
&= \mathbb{E}_{\mathbf{z}|\boldsymbol{\mu}}[tr(\mathbf{z}\mathbf{z}^\top)] - \boldsymbol{\mu}^\top \boldsymbol{\mu} \\
&= tr[\mathbb{E}_{\mathbf{z}|\boldsymbol{\mu}}(\mathbf{z}\mathbf{z}^\top)] - \boldsymbol{\mu}^\top \boldsymbol{\mu} \\
&= tr(I + \boldsymbol{\mu}\boldsymbol{\mu}^\top) - \boldsymbol{\mu}^\top \boldsymbol{\mu} \\
&= N \\
R_{Bayes}(\boldsymbol{\mu}) &= \mathbb{E}_{\mathbf{z}|\boldsymbol{\mu}}[\|\boldsymbol{\mu} - B\mathbf{z}\|_2^2] \\
&= B^2 \mathbb{E}_{\mathbf{z}|\boldsymbol{\mu}}(\mathbf{z}^\top \mathbf{z}) - 2B\boldsymbol{\mu}^\top \mathbb{E}_{\mathbf{z}|\boldsymbol{\mu}}(\mathbf{z}) + \boldsymbol{\mu}^\top \boldsymbol{\mu} \\
&= B^2 \mathbb{E}_{\mathbf{z}|\boldsymbol{\mu}}[tr(\mathbf{z}^\top \mathbf{z})] - 2B^2 \boldsymbol{\mu}^\top \boldsymbol{\mu} + \boldsymbol{\mu}^\top \boldsymbol{\mu} \\
&= B^2 \mathbb{E}_{\mathbf{z}|\boldsymbol{\mu}}[tr(\mathbf{z}\mathbf{z}^\top)] + (1 - 2B^2)\boldsymbol{\mu}^\top \boldsymbol{\mu} \\
&= B^2 tr[\mathbb{E}_{\mathbf{z}|\boldsymbol{\mu}}(\mathbf{z}\mathbf{z}^\top)] + (1 - 2B^2)\boldsymbol{\mu}^\top \boldsymbol{\mu} \\
&= B^2 tr(I + B^2 \boldsymbol{\mu}\boldsymbol{\mu}^\top) + (1 - 2B^2)\boldsymbol{\mu}^\top \boldsymbol{\mu} \\
&= (1 - B)^2 \|\boldsymbol{\mu}\|_2^2 + nB^2 \\
\mathbb{E}_{\boldsymbol{\mu}}[R_{Bayes}(\boldsymbol{\mu})] &= (1 - B)^2 \mathbb{E}[tr(\boldsymbol{\mu}^\top \boldsymbol{\mu})] + nB^2 \\
&= (1 - B)^2 \mathbb{E}[tr(\boldsymbol{\mu}\boldsymbol{\mu}^\top)] + nB^2 \\
&= (1 - B)^2 tr(A\mathbf{I}) + nB^2 \\
&= (1 - B)^2 An + nB^2 \\
&= n\frac{A + A^2}{(A+1)^2} \\
&= nB
\end{aligned}$$

□

3.5.3 James-Stein Estimator

By

$$S = \mathbf{z}^\top \mathbf{z} \sim (A+1)\chi_n^2$$

we have

$$\mathbb{E}\left(\frac{n-2}{S}\right) = \frac{1}{A+1}$$

So the J-S estimator is given by

$$\hat{\boldsymbol{\mu}}_{JS} = \left(1 - \frac{n-2}{S}\right)\mathbf{z}$$

Proof.

$\because \quad X = \dfrac{S}{A+1} \sim \chi_n^2 = \Gamma\left(\dfrac{n}{2}, 2\right)$

$\therefore$

$$f_X(x) = \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} \mathbb{1}_{[0,\infty)}$$

$\therefore$

$$\begin{aligned}
\mathbb{E}\left(\frac{n-2}{S}\right) &= \frac{n-2}{A+1}\mathbb{E}\left(\frac{A+1}{S}\right) \\
&= \frac{n-2}{A+1}\mathbb{E}\left(\frac{1}{X}\right) \\
&= \frac{n-2}{A+1}\int_0^\infty \frac{1}{x} f_X(x)\,\mathrm{d}x \\
&= \frac{n-2}{A+1}\int_0^\infty \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} x^{\frac{n-2}{2}-1} e^{-\frac{x}{2}}\,\mathrm{d}x \\
&= \frac{n-2}{A+1}\cdot\frac{1}{2^{\frac{2}{2}}\frac{n-2}{2}}\int_0^\infty \frac{1}{2^{\frac{n-2}{2}}\Gamma\left(\frac{n-2}{2}\right)} x^{\frac{n-2}{2}-1} e^{-\frac{x}{2}}\,\mathrm{d}x \\
&= \frac{n-2}{A+1}\cdot\frac{1}{n-2} \\
&= \frac{1}{A+1}
\end{aligned}$$

□

For $n \geq 3$, the James–Stein estimator everywhere dominates the MLE $\hat{\boldsymbol{\mu}}_{MLE}$ in terms of expected total squared error, i.e.,

$$R_{JS}(\boldsymbol{\mu}) < R_{MLE}(\boldsymbol{\mu})$$

for every choice of $\boldsymbol{\mu}$. Besides,

$$R_{JS}(\boldsymbol{\mu}) = n\frac{A}{A+1} + \frac{2}{A+1}$$

Proof.

Since

$$(\hat{\mu}_{JS,i} - \mu_i)^2 = (\hat{\mu}_{JS,i} - z_i)^2 - (z_i - \mu_i)^2 + 2(\hat{\mu}_{JS,i} - \mu_i)(z_i - \mu_i)$$

by summing it over $i = 1, 2, \cdots, n$, we have

$$R_{JS}(\boldsymbol{\mu}) = \mathbb{E}_{\mathbf{z}|\boldsymbol{\mu}}[\|\hat{\boldsymbol{\mu}}_{JS} - \mathbf{z}\|_2^2] - \mathbb{E}_{\mathbf{z}|\boldsymbol{\mu}}[\|\mathbf{z} - \boldsymbol{\mu}\|_2^2] + 2\sum_{i=1}^n Cov_{\mathbf{z}|\boldsymbol{\mu}}(\hat{\mu}_{JS,i}, z_i)$$

Integration by parts involving the multivariate normal density function

$$Cov_{\mathbf{z}|\boldsymbol{\mu}}(\hat{\mu}_{JS,i}, z_i) = \mathbb{E}_{\mathbf{z}|\boldsymbol{\mu}}\left(\frac{\partial \hat{\mu}_{JS,i}}{\partial z_i}\right)$$

as long as $\hat{\mu}_{JS,i}$ is continuously differentiable in $\mathbf{z}$.

$$\begin{aligned}
Cov_{\mathbf{z}|\boldsymbol{\mu}}(\hat{\mu}_{JS,i}, z_i) &= \mathbb{E}_{\mathbf{z}|\boldsymbol{\mu}}[(\hat{\mu}_{JS,i} - \mathbb{E}\hat{\mu}_{JS,i})(z_i - \mu_i)] \\
&= \iint_{\mathbb{R}^n} (\hat{\mu}_{JS,i} - \mathbb{E}\hat{\mu}_{JS,i})(z_i - \mu_i) \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\sum_{i=1}^n (z_i-\mu_i)^2} \mathrm{d}\mathbf{z} \\
&= -\iint_{\mathbb{R}^n} (\hat{\mu}_{JS,i} - \mathbb{E}\hat{\mu}_{JS,i}) \frac{\partial}{\partial z_i}\left[\frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\sum_{i=1}^n (z_i-\mu_i)^2}\right] \mathrm{d}\mathbf{z} \\
&= -\iint_{\mathbb{R}^n} (\hat{\mu}_{JS,i} - \mathbb{E}\hat{\mu}_{JS,i}) \mathrm{d}\left[\frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\sum_{i=1}^n (z_i-\mu_i)^2}\right] \\
&= -(\hat{\mu}_{JS,i} - \mathbb{E}\hat{\mu}_{JS,i}) \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\sum_{i=1}^n (z_i-\mu_i)^2}\Big|_{(-\infty,\infty)^n} + \iint_{\mathbb{R}^n} \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\sum_{i=1}^n (z_i-\mu_i)^2} \mathrm{d}(\hat{\mu}_{JS,i} - \mathbb{E}\hat{\mu}_{JS,i}) \\
&= \iint_{\mathbb{R}^n} \frac{\partial(\hat{\mu}_{JS,i} - \mathbb{E}\hat{\mu}_{JS,i})}{\partial z_i} \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\sum_{i=1}^n (z_i-\mu_i)^2} \mathrm{d}\mathbf{z} \\
&= \mathbb{E}_{\mathbf{z}|\boldsymbol{\mu}}\left(\frac{\partial \hat{\mu}_{JS,i}}{\partial z_i}\right)
\end{aligned}$$

Since

$$\begin{aligned}
\mathbb{E}_{\mathbf{z}|\boldsymbol{\mu}}[\|\hat{\boldsymbol{\mu}}_{JS} - \mathbf{z}\|_2^2] &= \mathbb{E}_{\mathbf{z}|\boldsymbol{\mu}}\left[\left\|\frac{n-2}{S}\mathbf{z}\right\|_2^2\right] \\
&= \mathbb{E}_{\mathbf{z}|\boldsymbol{\mu}}\left[\frac{(n-2)^2}{S}\right] \\
\mathbb{E}_{\mathbf{z}|\boldsymbol{\mu}}[\|\mathbf{z} - \boldsymbol{\mu}\|_2^2] &= n \\
\sum_{i=1}^n \mathbb{E}_{\mathbf{z}|\boldsymbol{\mu}}\left(\frac{\partial \hat{\mu}_{JS,i}}{\partial z_i}\right) &= \sum_{i=1}^n \mathbb{E}_{\mathbf{z}|\boldsymbol{\mu}}\left[\frac{\partial}{\partial z_i}\left(1 - \frac{n-2}{\sum_{i=1}^n z_i}\right) z_i\right] \\
&= \sum_{i=1}^n \mathbb{E}_{\mathbf{z}|\boldsymbol{\mu}}\left[1 - \frac{n-2}{S} + \frac{2(n-2)}{S} z_i^2\right] \\
&= \mathbb{E}_{\mathbf{z}|\boldsymbol{\mu}}\left[n\left(1 - \frac{n-2}{S}\right) + \frac{2(n-2)}{S}\right] \\
&= n - \mathbb{E}_{\mathbf{z}|\boldsymbol{\mu}}\left[\frac{(n-2)^2}{S}\right]
\end{aligned}$$

we have when $n \geq 3$,

$$\begin{aligned}
R_{JS}(\boldsymbol{\mu}) &= n - \mathbb{E}_{\mathbf{z}|\boldsymbol{\mu}}\left[\frac{(n-2)^2}{S}\right] \\
&= n - (n-2)\frac{1}{A+1} \\
&= n\frac{A}{A+1} + \frac{2}{A+1} \\
&\leq n \\
&= R_{MLE}(\boldsymbol{\mu})
\end{aligned}$$

□

Chapter 2 Multiple Regression

4 Model Assessment and Selection

4.1 Criteria for Best Model

In multiple regression, we have several models and we want to define which is the best model. The general criteria is as follow

$$\underbrace{\frac{SSE_p}{n}}_{\text{Training Error}} + \underbrace{\lambda p \frac{\hat{\sigma}^2}{n}}_{\text{Penalization}}$$

where SSE_p is sum of square error of the p-th model, $\hat{\sigma}^2$ is the estimated variance of ε, n is the number of data points, p is the number of predictors and λ is the penalized coefficient.

When $\lambda = 2$, it turns out to be the Akaike Information Criterion (AIC).

When $\lambda = \ln n$, Bayesian Information Criterion (BIC)

And another useful criteria is Mallow's C_p

$$C_p = \frac{SSE_p}{\hat{\sigma}^2} - n + 2p$$

4.2 Bias–Variance Tradeoff

Suppose that we have a training set consisting of a set of points $x_1, \ldots, x_n$ and real values y_i associated with each point x_i. We assume that there is a function with noise $y = f(x) + \varepsilon$, where the noise, ε, has zero mean and variance σ^2. Then

$$\mathrm{E}\left[\left(y - \hat{f}(x)\right)^2\right] = \mathrm{Bias}\left[\hat{f}(x)\right]^2 + \mathrm{Var}\left[\hat{f}(x)\right] + \sigma^2$$

Proof. The derivation of the bias–variance decomposition for squared error proceeds as follows. For notational convenience, abbreviate $f = f(x)$ and $\hat{f} = \hat{f}(x)$. First, recall that, by definition, for any random variable X, we have

$$\mathrm{Var}[X] = \mathrm{E}[X^2] - \mathrm{E}[X]^2$$

Rearranging, we get:

$$\mathrm{E}[X^2] = \mathrm{Var}[X] + \mathrm{E}[X]^2$$

Since f is deterministic

$$\mathrm{E}[f] = f$$

This, given $y = f + \varepsilon$ and $\mathrm{E}[\varepsilon] = 0$, implies $\mathrm{E}[y] = \mathrm{E}[f + \varepsilon] = \mathrm{E}[f] = f$.

Also, since $\mathrm{Var}[\varepsilon] = \sigma^2$

$$\text{Var}[y] = \text{E}[(y - \text{E}[y])^2] = \text{E}[(y - f)^2] = \text{E}[(f + \varepsilon - f)^2] = \text{E}[\varepsilon^2] = \text{Var}[\varepsilon] + \text{E}[\varepsilon]^2 = \sigma^2$$

Thus, since ε and $\hat{f}$ are independent, we can write

$$\begin{aligned}
\text{E}\left[(y - \hat{f})^2\right] &= \text{E}[y^2 + \hat{f}^2 - 2y\hat{f}] \\
&= \text{E}[y^2] + \text{E}[\hat{f}^2] - \text{E}[2y\hat{f}] \\
&= \text{Var}[y] + \text{E}[y]^2 + \text{Var}[\hat{f}] + \text{E}[\hat{f}]^2 - 2f\,\text{E}[\hat{f}] \\
&= \text{Var}[y] + \text{Var}[\hat{f}] + (f^2 - 2f\,\text{E}[\hat{f}] + \text{E}[\hat{f}]^2) \\
&= \text{Var}[y] + \text{Var}[\hat{f}] + (f - \text{E}[\hat{f}])^2 \\
&= \sigma^2 + \text{Var}[\hat{f}] + \text{Bias}[\hat{f}]^2
\end{aligned}$$

□

4.3 Subset Selection Methods

1. Forward-stepwise
2. Backward-stepwise
3. Stepwise-regression
4. Stagewise regression

Algorithm 1 Forward Stagewise Regression Algorithm

Require: $\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_2 & \cdots & \mathbf{X}_p \end{bmatrix}$, $\mathbf{Y}$, ε

Ensure: $\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_1 & \hat{\beta}_2 & \cdots & \hat{\beta}_p \end{bmatrix}$

1: Start with the residual $r = Y$ and $\hat{\beta} = \mathbf{0}$. X_j and Y are standardized to have mean zero and unit norm.
2: while r is correlated with some predictors X_j do
3: Find X_j most correlated with r.
4: $\delta_j \leftarrow \varepsilon \cdot sign(< X_j, r >)$
5: $\hat{\beta}_j \leftarrow \hat{\beta}_j + \delta_j$
6: $r \leftarrow r - \delta_j X_j$
7: end while

5 Shrinkage Methods

$$\hat{\beta} = \arg\min_{\beta}[\|Y - X\beta\|_2^2 + \lambda p(\beta)]$$

where λ is a regularization parameter and $p(\beta)$ is a penalized function.

When $\lambda = 0$, $\hat{\beta}$ is the ordinary least squares estimator.

When $\lambda \neq 0$ and $p(\beta) = \|\beta\|_0 = \sum_{i=1}^{p} \mathbb{1}_{\{\beta_i \neq 0\}}$, it equals to the best subset selection method.

5.1 Ridge Regression

When $\lambda > 0$ and $p(\beta) = \|\beta\|_2^2$, the ridge estimator is defined by

$$\hat{\beta}_{Ridge} = \arg\min_{\beta}[\|Y - X\beta\|_2^2 + \lambda \|\beta\|_2^2]$$

Equivalently,

$$\hat{\beta}_{Ridge} = \arg\min_{\beta} \|Y - X\beta\|_2^2$$
$$\text{subject to } \sum_{i=1}^{p} \beta_i^2 \leq t$$

The solution is

$$\hat{\beta}_{Ridge} = (X^\top X + \lambda I)^{-1} X^\top Y$$

solved by deviating.

5.2 Least Absolute Shrinkage and Selection Operator

When $\lambda > 0$ and $p(\beta) = \|\beta\|_1 = \sum_{i=1}^{p} |\beta_i|$, the LASSO estimator is defined by

$$\hat{\beta}_{Ridge} = \arg\min_{\beta}[\|Y - X\beta\|_2^2 + \lambda \sum_{i=1}^{p} |\beta_i|]$$

Equivalently,

$$\hat{\beta}_{Ridge} = \arg\min_{\beta} \|Y - X\beta\|_2^2$$
$$\text{subject to } \sum_{i=1}^{p} |\beta_i| \leq t$$

The solution is solved by coordinate descending.

Algorithm 2 Coordinate Descending Algorithm

Require: $\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_2 & \cdots & \mathbf{X}_p \end{bmatrix}$, $\mathbf{Y}$

Ensure: $\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_1 & \hat{\beta}_2 & \cdots & \hat{\beta}_p \end{bmatrix}$

1: Start with the residual $\hat{\boldsymbol{\beta}} = \mathbf{0}$. $\mathbf{X}_j$ and $\mathbf{Y}$ are standardized to have mean zero and unit norm.
2: while $\hat{\boldsymbol{\beta}}$ is not converge do
3: for $j = 1, 2, \ldots, p$ do
4: $\mathbf{r}_{(j)} \leftarrow \mathbf{Y} - \mathbf{X}_{(j)}\hat{\beta}_{(j)}$
5: $\hat{\beta}_j^* \leftarrow \hat{\beta}_{(j)} + \frac{1}{N}\mathbf{X}^\top \mathbf{r}_{(j)}$
6: $\hat{\beta}_j \leftarrow sign(\hat{\beta}_j^*)(|\hat{\beta}_j^*| - \lambda)_+$
7: end for
8: end while

5.3 Least Angle Regression

Algorithm 3 Least Angle Regression Algorithm

Require: $\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_2 & \cdots & \mathbf{X}_p \end{bmatrix}$, $\mathbf{Y}$

Ensure: $\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_1 & \hat{\beta}_2 & \cdots & \hat{\beta}_p \end{bmatrix}$

1: Standardize the predictors to have mean zero and unit norm. Start with the residual $\mathbf{r} = \mathbf{Y} - \overline{\mathbf{Y}}$, $\hat{\boldsymbol{\beta}} = \mathbf{0}$.
2: Find the predictor $\mathbf{X}_j$ most correlated with $\mathbf{r}$.
3: Move β_j from 0 towards its least-squares coefficient $< \mathbf{X}_j, \mathbf{r} >$, until some other competitor $\mathbf{X}_k$ has as much correlation with the current residual as does $\mathbf{X}_j$.
4: Move β_j and β_k in the direction defined by their joint least squares coefficient of the current residual on $(\mathbf{X}_j, \mathbf{X}_k)$, until some other competitor $\mathbf{X}_l$ has as much correlation with the current residual.
5: Continue in this way until all p predictors have been entered. After $\min(N-1, p)$ steps, we arrive at the full least-squares solution.

Consider a regression problem with all variables and response having mean zero and standard deviation one. Suppose also that each variable has identical absolute correlation with the response:

$$\frac{1}{N}| < x_j, y > | = \lambda, \quad j = 1, \ldots, p$$

Let $\hat{\beta}$ be the least-squares coefficient of y on $\mathbf{X}$, and let $u(\alpha) = \alpha X \hat{\beta}$ for $\alpha \in [0, 1]$ be the vector that moves a fraction α toward the least squares fit u. Let RSS be the residual sum-of-squares from the full least squares fit.

1. Show that

$$\frac{1}{N}| < x_j, y - u(\alpha) > | = (1 - \alpha)\lambda, \quad j = 1, \ldots, p$$

and hence the correlations of each x_j with the residuals remain equal in magnitude as we progress toward u.

Proof.

Let

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top$$

$\because$

$$\mathbf{H} = \mathbf{H}^\top$$
$$\mathbf{H}\mathbf{X} = \mathbf{X}$$

$\therefore \quad \forall\ j \in \{1, \cdots, p\}$,

$$\begin{aligned} x_j^\top \mathbf{H} &= (\mathbf{H}x_j)^\top \\ &= x_j^\top \end{aligned}$$

$\therefore \quad \forall\ j \in \{1, \cdots, p\}$,

$$\begin{aligned} \frac{1}{N}|< x_j, y - u(\alpha) >| &= \frac{1}{N}|< x_j, y - \alpha\mathbf{X}\hat{\beta} >| \\ &= \frac{1}{N}|< x_j, y > -\alpha < x_j, \mathbf{H}y >| \\ &= \frac{1}{N}|< x_j, y > -\alpha x_j^\top \mathbf{H}y| \\ &= \frac{1}{N}|< x_j, y > -\alpha x_j^\top y| \\ &= (1-\alpha)\frac{1}{N}|< x_j, y >| \\ &= (1-\alpha)\lambda \end{aligned}$$

□

2. Show that these correlations are all equal to

$$\lambda(\alpha) = \frac{1-\alpha}{\sqrt{(1-\alpha)^2 + \frac{\alpha(2-\alpha)}{N} \cdot RSS}}\lambda$$

and hence they decrease monotonically to zero.

Proof.

The absolute correlations (not covariances) would be given by

$$\frac{\frac{1}{N}|< x_j, y - u(\alpha) >|}{\sqrt{\frac{1}{N} < x_j, x_j > \cdot \frac{1}{N} < y - u(\alpha), y - u(\alpha) >}} = \frac{(1-\alpha)\lambda}{\sqrt{\frac{1}{N} < y - u(\alpha), y - u(\alpha) >}}$$

As a first step we have

$$\begin{aligned} < y - u(\alpha), y - u(\alpha) > &= < y - \alpha \mathbf{X}\hat{\beta}, y - \alpha \mathbf{X}\hat{\beta} > \\ &= y^\top y - 2\alpha y^\top \mathbf{X}\hat{\beta} + \alpha^2 \hat{\beta}^\top \mathbf{X}^\top \mathbf{X}\hat{\beta} \end{aligned}$$

Now recall the normal equations for linear regression

$$\begin{aligned} \mathbf{X}^\top (y - \mathbf{X}\hat{\beta}) &= \mathbf{X}^\top (y - \mathbf{H}y) \\ &= \mathbf{X}^\top (\mathbf{I} - \mathbf{H}) y \\ &= 0 \\ \mathbf{X}^\top y &= \mathbf{X}\mathbf{X}\hat{\beta} \end{aligned}$$

Using this we can write

$$\begin{aligned} < y - u(\alpha), y - u(\alpha) > &= y^\top y - 2\alpha y^\top \mathbf{X}\hat{\beta} + \alpha^2 \hat{\beta}^\top \mathbf{X} y \\ &= y^\top y + \alpha(\alpha - 2) y^\top \mathbf{X}\hat{\beta} \end{aligned}$$

If $\alpha = 1$ the left-hand-side is the *RSS*. This means that

$$\begin{aligned} RSS &= y^\top y - y^\top \mathbf{X}\hat{\beta} \\ y^\top \mathbf{X}\hat{\beta} &= y^\top y - RSS \end{aligned}$$

Using this we have that

$$\begin{aligned} < y - u(\alpha), y - u(\alpha) > &= y^\top y + \alpha(\alpha - 2)(y^\top y - RSS) \\ &= (1 - \alpha)^2 y^\top y + \alpha(2 - \alpha) RSS \end{aligned}$$

As y has a mean zero and a standard deviation of one means that $\frac{1}{N} y^\top y = 1$ so the above becomes

$$\frac{1}{N} < y - u(\alpha), y - u(\alpha) > = (1 - \alpha)^2 + \frac{\alpha(2 - \alpha)}{N} RSS$$

Therefore, these absolute correlation equal to

$$\begin{aligned} \lambda(\alpha) &= \frac{1 - \alpha}{\sqrt{(1 - \alpha)^2 + \frac{\alpha(2-\alpha)}{N} \cdot RSS}} \lambda \\ \lambda'(\alpha) &< 0 \\ \lambda(1) &= 0 \end{aligned}$$

and hence they decrease monotonically to zero. □

3. Use these results to show that the LAR algorithm in Section 3.4.4 keeps the correlations tied and monotonically decreasing, as claimed in (3.55).

Proof.

From the given expression derived in Part (a) and (b) one sees that when $\alpha = 0$ we have $\lambda(0) = \lambda$, when $\alpha = 1$ we have that $\lambda(1) = 0$, where all correlations are tied and decrease from λ to zero as α moves from 0 to 1. □

6 Projection Related Methods

6.1 Pricipal Components Regression

From optimization viewpoint,

$$\begin{aligned} & \max Var(\mathbf{X}\mathbf{b}_i) & \\ s.t. \quad & \mathbf{b}_j^\top \mathbf{b}_j = 1 & j = 1, 2, \cdots, t \\ & \mathbf{b}_i^\top \mathbf{\Sigma}_{\mathbf{XX}} \mathbf{b}_j = 0 & i < j \end{aligned}$$

Algorithm 4 Pricipal Components Regression

Require: $\mathscr{X} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_2 & \cdots & \mathbf{X}_p \end{bmatrix} \in \mathbb{R}^{n \times p}$, $\mathbf{Y} \in \mathbb{R}^n$

Ensure: $\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 & \hat{\beta}_2 & \cdots & \hat{\beta}_p \end{bmatrix}$

1: Centering $\mathscr{X}$ and $\mathscr{Y}$ to $\mathscr{X}_c$ and $\mathscr{Y}_c$.

2: Compute the sample covariance matrix $\mathbf{\Sigma} = \mathscr{X}\mathscr{X}^\top$

3: Compute the spectral decomposition of $\mathbf{\Sigma} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$, where $\mathbf{V} = \begin{bmatrix} \mathbf{V}_1 & \mathbf{V}_2 & \cdots & \mathbf{V}_p \end{bmatrix}$, $\mathbf{V}^\top\mathbf{V} = \mathbf{I}$, $\mathbf{\Lambda} = diag\{\lambda_1, \lambda_2, \cdots, \lambda_p\}$ and $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$.

4: Compute the ith principal component of $\mathbf{X}$ - $\mathbf{Z}_i = \mathscr{X}\mathbf{V}_i$

5: Regress $\mathbf{Y}$ on the first m ($m \leq p$) principal components,

$$\hat{\mathbf{Y}} = \overline{\mathbf{Y}} + \sum_{i=1}^{m} \frac{< \mathbf{Z}_i, \mathscr{Y}_c >}{< \mathbf{Z}_i, \mathbf{Z}_i >} \mathbf{Z}_i$$

Because the extraction of the principal components is accomplished without any reference to the output variable $\mathbf{Y}$, we have no reason to expect $\mathbf{Y}$ to be highly correlated with any of the principal components, in particular those having the largest eigenvalues.

6.2 Partial Least-Squares Regression

From optimization viewpoint,

$$\begin{aligned} & \max Corr^2(\mathbf{Y}, \mathbf{X}\mathbf{b}_i) Var(\mathbf{X}\mathbf{b}_i) \\ s.t. \quad & \mathbf{b}_j^\top \mathbf{b}_j = 1 \qquad\qquad j = 1, 2, \cdots, t \\ & \mathbf{b}_i^\top \mathbf{\Sigma}_{\mathbf{XX}} \mathbf{b}_j = 0 \qquad\quad i < j \end{aligned}$$

Let $\mathbf{X} \in \mathbb{R}^p$ be a random vector. $\mathbf{b}_i$ $(i = 1, 2 \cdots, r)$ are chosen in a sequential manner so that the variances of the derived variables

$$Corr^2(\mathbf{Y}, \mathbf{X}\mathbf{b}_i) Var(\mathbf{X}\mathbf{b}_i) = \frac{(\mathbf{b}_i^\top \Sigma_{\mathbf{XY}})^2}{\Sigma_{\mathbf{YY}}}$$

are arranged in descending order subject to the normalizations

$$\begin{aligned} \mathbf{b}_i^\top \mathbf{b}_i &= 1 \\ \mathbf{b}_i^\top \Sigma_{\mathbf{XX}} \mathbf{b}_j &= 0 \qquad (i < j) \end{aligned}$$

i.e., they are uncorrelated with previously chosen derived variables.

(1) For the first component, we want

$$\begin{aligned} & \max (\mathbf{b}_1^\top \Sigma_{\mathbf{XY}})^2 \\ s.t. \quad & \mathbf{b}_1^\top \mathbf{b}_1 = 1 \end{aligned}$$

Since $\Sigma_{\mathbf{XY}} \in \mathbb{R}^p$, by Cauchy–Schwarz Inequality

$$< \mathbf{b}_1, \Sigma_{\mathbf{XY}} >^2 \leq \|\mathbf{b}_1\|_2^2 \|\Sigma_{\mathbf{XY}}\|_2^2$$

with equality holds iff

$$\mathbf{b}_1 \propto \Sigma_{\mathbf{XY}}$$

Therefore

$$\mathbf{b}_1 = \frac{\Sigma_{\mathbf{XY}}}{\|\Sigma_{\mathbf{XY}}\|_2}$$

(2) For the second component, we want

$$\begin{aligned} & \max (\mathbf{b}_2^\top \Sigma_{\mathbf{XY}})^2 \\ s.t. \quad & \mathbf{b}_2^\top \mathbf{b}_2 = 1 \\ & \mathbf{b}_1^\top \Sigma_{\mathbf{XX}} \mathbf{b}_2 = 0 \end{aligned}$$

From condition 2 we have

$$\frac{\Sigma_{\mathbf{YX}}\Sigma_{\mathbf{XX}}\mathbf{b}_2}{\|\Sigma_{\mathbf{XY}}\|_2} = 0$$

Therefore, we can maximize $< \mathbf{b}_2, \Sigma_{\mathbf{XY}} - c\Sigma_{\mathbf{XX}}\Sigma_{\mathbf{XY}} >^2$. And this gives the solution

$$\mathbf{b}_2 = \frac{\Sigma_{\mathbf{XY}} - c\Sigma_{\mathbf{XX}}\Sigma_{\mathbf{XY}}}{\|\Sigma_{\mathbf{XY}} - c\Sigma_{\mathbf{XX}}\Sigma_{\mathbf{XY}}\|_2}$$

So we need to determine c.

From condition 2, we have

$$\Sigma_{\mathbf{YX}}\Sigma_{\mathbf{XX}}(\Sigma_{\mathbf{XY}} - c\Sigma_{\mathbf{XX}}\Sigma_{\mathbf{XY}}) = 0$$

Thus

$$c = \frac{\Sigma_{\mathbf{YX}}\Sigma_{\mathbf{XX}}\Sigma_{\mathbf{XY}}}{\Sigma_{\mathbf{YX}}\Sigma_{\mathbf{XX}}^2\Sigma_{\mathbf{XY}}}$$

(3) Similarly, we can obtain the remaining sets of coefficients for the components.

Algorithm 5 Partial Least-squares Regression

Require: $\mathscr{X} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_2 & \cdots & \mathbf{X}_p \end{bmatrix} \in \mathbb{R}^{n\times p}$, $\mathbf{Y}$

Ensure: $\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 & \hat{\beta}_2 & \cdots & \hat{\beta}_p \end{bmatrix}$

1: Standardize each $\mathbf{X}_i$ to have mean zero and variance one. Set $\hat{\mathbf{Y}}^{(0)} = \overline{\mathbf{Y}}\mathbf{1}$ and $\mathbf{X}_i^{(0)} = \mathbf{X}_i$, $i = 1, 2, \cdots, p$

2: for $m = 1, 2, \cdots, p$ do

3: $\quad \mathbf{Z}_m = \sum_{j=1}^{p} < \mathbf{X}_j^{(m-1)}, \mathbf{Y} > \mathbf{X}_j^{(m-1)}$

4: $\quad \hat{\mathbf{Y}}^{(m)} = \hat{\mathbf{Y}}^{(m-1)} + \frac{< \mathbf{Z}_m, \mathbf{Y} >}{< \mathbf{Z}_m, \mathbf{Z}_m >}\mathbf{Z}_m$

5: $\quad$ Orthogonalize each $\mathbf{X}_j^{(m-1)}$ with respect to $\mathbf{Z}_m$: $\mathbf{X}_j^{(m)} = \mathbf{X}_j^{(m-1)} - \frac{< \mathbf{Z}_m, \mathbf{X}_j^{(m-1)} >}{< \mathbf{Z}_m, \mathbf{Z}_m >}\mathbf{Z}_m$

6: end for

7: Output the sequence of fitted vectors $\{\hat{\mathbf{Y}}^{(m)},\ m = 1, 2\cdots, p\}$. Since the $\{\mathbf{Z}_m,\ m = 1, 2\cdots, p\}$ are linear in the original $\mathbf{X}_j$, so is $\hat{\mathbf{Y}}^{(m)} = \mathscr{X}\hat{\beta}_{pls}$. These linear coefficients can be recovered from the sequence of PLS transformations.

7 Generalized Least-Squares Regression

Suppose the error component $\mathbf{e}$ of the linear regression model has mean $\mathbf{0}$, but now has $Var(\mathbf{e}) = \sigma^2\mathbf{V}$, where $\mathbf{V}$ is a known $(n \times n)$ positive-definite symmetric matrix and $\sigma^2 > 0$ may not be necessarily known. Let $\hat{\boldsymbol{\beta}}_{gls}$ denote the generalized least-squares (GLS) estimator:

$$\hat{\boldsymbol{\beta}}_{gls} = \arg\min_{\beta}(\mathbf{Y} - \mathbf{Z}\boldsymbol{\beta})^\top \mathbf{V}^{-1}(\mathbf{Y} - \mathbf{Z}\boldsymbol{\beta})$$

Show that

$$\hat{\boldsymbol{\beta}}_{gls} = (\mathbf{Z}^\top \mathbf{V}^{-1}\mathbf{Z})^{-1}\mathbf{Z}^\top \mathbf{V}^{-1}\mathbf{Y}$$

has expectation $\boldsymbol{\beta}$ and covariance matrix $Var(\hat{\boldsymbol{\beta}}_{gls}) = \sigma^2(\mathbf{Z}^\top \mathbf{V}^{-1}\mathbf{Z})^{-1}$.

Proof.

Method 1

Let

$$\begin{aligned} f(\boldsymbol{\beta}) &= (\mathbf{Y} - \mathbf{Z}\boldsymbol{\beta})^\top \mathbf{V}^{-1}(\mathbf{Y} - \mathbf{Z}\boldsymbol{\beta}) \\ &= \mathbf{Y}^\top \mathbf{V}^{-1}\mathbf{Y} - 2\boldsymbol{\beta}^\top \mathbf{Z}^\top \mathbf{V}^{-1}\mathbf{Y} + \boldsymbol{\beta}^\top \mathbf{Z}^\top \mathbf{V}^{-1}\mathbf{Z}\boldsymbol{\beta} \end{aligned}$$

From Matrix Cookbook, for vecter $\mathbf{x} \in \mathbb{R}^n$, matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$,

$$\begin{aligned} \frac{\mathrm{d}(\mathbf{x}^\top \mathbf{A})}{\mathrm{d}\mathbf{x}} &= \frac{\mathrm{d}(\mathbf{A}^\top \mathbf{x})}{\mathrm{d}\mathbf{x}} \\ &= \mathbf{A} \\ \frac{\mathrm{d}(\mathbf{x}^\top \mathbf{A}\mathbf{x})}{\mathrm{d}\mathbf{x}} &= (\mathbf{A} + \mathbf{A}^\top)\mathbf{x} \end{aligned}$$

Therefore,

$$\begin{aligned} \frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{\beta}} &= -2\mathbf{Z}^\top \mathbf{V}^{-1}\mathbf{Y} + 2\mathbf{Z}^\top \mathbf{V}^{-1}\mathbf{Z}\boldsymbol{\beta} \\ &= 2\mathbf{Z}^\top \mathbf{V}^{-1}(\mathbf{Z}\boldsymbol{\beta} - \mathbf{Y}) \\ \frac{\mathrm{d}^2 f}{\mathrm{d}\boldsymbol{\beta}\mathrm{d}\boldsymbol{\beta}^\top} &= 2\mathbf{Z}^\top \mathbf{V}^{-1}\mathbf{Z} \end{aligned}$$

Since $\mathbf{Z}^\top \mathbf{V}^{-1}\mathbf{Z}$ is positive-definite, by setting $\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{\beta}} = \mathbf{0}$, we get the global minima

$$\hat{\boldsymbol{\beta}}_{gls} = (\mathbf{Z}^\top \mathbf{V}^{-1}\mathbf{Z})^{-1}\mathbf{Z}^\top \mathbf{V}^{-1}\mathbf{Y}$$

Therefore,

$$\begin{aligned} \mathbb{E}\hat{\boldsymbol{\beta}}_{gls} &= \mathbb{E}[(\mathbf{Z}^\top \mathbf{V}^{-1}\mathbf{Z})^{-1}\mathbf{Z}^\top \mathbf{V}^{-1}\mathbf{Y}] \\ &= (\mathbf{Z}^\top \mathbf{V}^{-1}\mathbf{Z})^{-1}\mathbf{Z}^\top \mathbf{V}^{-1}\mathbb{E}(\mathbf{Y}) \end{aligned}$$

$$
\begin{aligned}
&= (\mathbf{Z}^\top \mathbf{V}^{-1}\mathbf{Z})^{-1}\mathbf{Z}^\top \mathbf{V}^{-1}\mathbb{E}(\mathbf{Z}\boldsymbol{\beta}+\mathbf{e}) \\
&= (\mathbf{Z}^\top \mathbf{V}^{-1}\mathbf{Z})^{-1}(\mathbf{Z}^\top \mathbf{V}^{-1}\mathbf{Z})\boldsymbol{\beta} \\
&= \boldsymbol{\beta} \\
Var\hat{\boldsymbol{\beta}}_{gls} &= Var[(\mathbf{Z}^\top \mathbf{V}^{-1}\mathbf{Z})^{-1}\mathbf{Z}^\top \mathbf{V}^{-1}\mathbf{Y}] \\
&= (\mathbf{Z}^\top \mathbf{V}^{-1}\mathbf{Z})^{-1}\mathbf{Z}^\top \mathbf{V}^{-1}Var(\mathbf{Y})\mathbf{V}^{-1}\mathbf{Z}(\mathbf{Z}^\top \mathbf{V}^{-1}\mathbf{Z})^{-1} \\
&= (\mathbf{Z}^\top \mathbf{V}^{-1}\mathbf{Z})^{-1}\mathbf{Z}^\top \mathbf{V}^{-1}Var(\mathbf{Z}\boldsymbol{\beta}+\mathbf{e})\mathbf{V}^{-1}\mathbf{Z}(\mathbf{Z}^\top \mathbf{V}^{-1}\mathbf{Z})^{-1} \\
&= \sigma^2(\mathbf{Z}^\top \mathbf{V}^{-1}\mathbf{Z})^{-1}\mathbf{Z}^\top \mathbf{V}^{-1}\mathbf{V}\mathbf{V}^{-1}\mathbf{Z}(\mathbf{Z}^\top \mathbf{V}^{-1}\mathbf{Z})^{-1} \\
&= \sigma^2(\mathbf{Z}^\top \mathbf{V}^{-1}\mathbf{Z})^{-1}(\mathbf{Z}^\top \mathbf{V}^{-1}\mathbf{Z})(\mathbf{Z}^\top \mathbf{V}^{-1}\mathbf{Z})^{-1} \\
&= \sigma^2(\mathbf{Z}^\top \mathbf{V}^{-1}\mathbf{Z})^{-1}
\end{aligned}
$$

Method 2

Since $\mathbf{V}$ is positive-definite, by Cholesky decomposition,

$$\mathbf{V} = \mathbf{C}\mathbf{C}^\top$$

where $\mathbf{C}$ is a lower triangular matrix with real and positive diagonal entries.

Let

$$
\begin{aligned}
\mathbf{Y}^* &= \mathbf{C}^{-1}\mathbf{Y} \\
\mathbf{Z}^* &= \mathbf{C}^{-1}\mathbf{Z} \\
\mathbf{e}^* &= \mathbf{C}^{-1}\mathbf{e}
\end{aligned}
$$

Then

$$
\begin{aligned}
\mathbf{Y}^* &= \mathbf{Z}^*\boldsymbol{\beta} + \mathbf{e} \\
\mathbf{e} &\sim N(\mathbf{0}, \sigma^2\mathbf{I})
\end{aligned}
\tag{1}
$$

Since

$$
\begin{aligned}
(\mathbf{C}^\top)^{-1}\mathbf{C}^\top &= \mathbf{I} \\
(\mathbf{C}^{-1})^\top\mathbf{C}^\top &= (\mathbf{C}\mathbf{C}^{-1})^\top = \mathbf{I}
\end{aligned}
$$

i.e.

$$(\mathbf{C}^\top)^{-1} = (\mathbf{C}^{-1})^\top$$

Therefore,

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}_{gls} &= \arg\min_{\boldsymbol{\beta}}(\mathbf{Y}-\mathbf{Z}\boldsymbol{\beta})^\top \mathbf{V}^{-1}(\mathbf{Y}-\mathbf{Z}\boldsymbol{\beta}) \\
&= \arg\min_{\boldsymbol{\beta}}(\mathbf{Y}^*-\mathbf{Z}^*\boldsymbol{\beta})^\top(\mathbf{Y}^*-\mathbf{Z}^*\boldsymbol{\beta}) \\
&= \hat{\boldsymbol{\beta}}^*_{ols} \\
&= (\mathbf{Z}^{*\top}\mathbf{Z}^*)^{-1}\mathbf{Z}^{*\top}\mathbf{Y}^*
\end{aligned}
$$

$$= (\mathbf{Z}^\top \mathbf{V}^{-1} \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{V}^{-1} \mathbf{Y}$$

where $\hat{\boldsymbol{\beta}}^*_{ols}$ is the ordinary least squares estimator of (1)

$\mathbb{E}\hat{\boldsymbol{\beta}}_{gls}$ and $Var\hat{\boldsymbol{\beta}}_{gls}$ are abtained the same as method 1. □

What would be the consequences of incorrectly using the ordinary least-squares estimator $\hat{\boldsymbol{\beta}}_{ols} = (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{Y}$, of $\boldsymbol{\beta}$ when $Var(\mathbf{e}) = \sigma^2 \mathbf{V}$?

Proof.

In the case of $Var(\mathbf{e}^*) = \sigma^2 \mathbf{I}$, the Gauss–Markov theorem applies $\hat{\boldsymbol{\beta}}^*_{ols}$ is the best linear unbiased estimator (BLUE) for $\boldsymbol{\beta}$. And therefore, $\hat{\boldsymbol{\beta}}_{gls}$ is the BULE for $\boldsymbol{\beta}$.

However, if incorrectly using the ordinary least-squares estimator $\hat{\boldsymbol{\beta}}_{ols} = (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{Y}$, of $\boldsymbol{\beta}$ when $Var(\mathbf{e}) = \sigma^2 \mathbf{V}$, although it is still unbiased, its varaince will larger than $\hat{\boldsymbol{\beta}}_{gls}$ as following.

Since both $\hat{\boldsymbol{\beta}}_{ols}$ and $\hat{\boldsymbol{\beta}}_{ols}$ are linear w.r.t. $\mathbf{Y}$,

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{ols} &= \hat{\boldsymbol{\beta}}_{gls} + \mathbf{A}\mathbf{Y} + \mathbf{b} \\
\mathbb{E}\hat{\boldsymbol{\beta}}_{ols} &= \mathbb{E}\hat{\boldsymbol{\beta}}_{gls} + \mathbf{A}\mathbf{Z}\boldsymbol{\beta} + \mathbf{b} \\
\boldsymbol{\beta} &= \boldsymbol{\beta} + \mathbf{A}\mathbf{Z}\boldsymbol{\beta} + \mathbf{b}
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbf{A}\mathbf{Z} &= \mathbf{0} \\
\mathbf{b} &= \mathbf{0}
\end{aligned}$$

Therefore,

$$\begin{aligned}
Var\hat{\boldsymbol{\beta}}_{ols} &= Var(\hat{\boldsymbol{\beta}}_{gls} + \mathbf{A}\mathbf{Y} + \mathbf{b}) \\
&= Var\{[(\mathbf{Z}^\top \mathbf{V}^{-1} \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{V}^{-1} + \mathbf{A}]\mathbf{Y}\} \\
&= [(\mathbf{Z}^\top \mathbf{V}^{-1} \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{V}^{-1} + \mathbf{A}] Var(\mathbf{Y}) [(\mathbf{Z}^\top \mathbf{V}^{-1} \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{V}^{-1} + \mathbf{A}]^\top \\
&= \sigma^2 [(\mathbf{Z}^\top \mathbf{V}^{-1} \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{V}^{-1} + \mathbf{A}] \mathbf{V} [(\mathbf{Z}^\top \mathbf{V}^{-1} \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{V}^{-1} + \mathbf{A}]^\top \\
&= Var\hat{\boldsymbol{\beta}}_{gls} + \sigma^2 \mathbf{A}\mathbf{V}\mathbf{A}^\top \\
&= Var\hat{\boldsymbol{\beta}}_{gls} + \sigma^2 (\mathbf{A}\mathbf{C})(\mathbf{A}\mathbf{C})^\top
\end{aligned}$$

Since $(\mathbf{AC})(\mathbf{AC})^\top$ is a positive semi-definite matrix, $Var\hat{\boldsymbol{\beta}}_{ols}$ exceeds $Var\hat{\boldsymbol{\beta}}_{gls}$ by a positive semi-definite matrix. □

Chapter 3 Multivariate Regression

8 The Fiexed-X Case

8.1 Regression Model

Let $\mathbf{Y} = \begin{pmatrix} Y_1 & \cdots & Y_s \end{pmatrix}^\top \in \mathbb{R}^s$ be a random vector and $\mathbf{X} = \begin{pmatrix} X_1 & \cdots & X_r \end{pmatrix}^\top \in \mathbb{R}^r$ be a fixed vector. Suppose that we observe n replications $(\mathbf{X}_j^\top, \mathbf{Y}_j^\top)^\top \in \mathbb{R}^{r+s}$, $j = 1, 2, \cdots, n$. We define

$$
\begin{aligned}
\underset{r\times n}{\mathscr{X}} &= \begin{pmatrix} \mathbf{X}_1 & \cdots & \mathbf{X}_n \end{pmatrix} \\
\underset{s\times n}{\mathscr{Y}} &= \begin{pmatrix} \mathbf{Y}_1 & \cdots & \mathbf{Y}_n \end{pmatrix} \\
\underset{r\times 1}{\overline{\mathbf{X}}} &= \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i \\
\underset{s\times 1}{\overline{\mathbf{Y}}} &= \frac{1}{n}\sum_{i=1}^n \mathbf{Y}_i \\
\underset{r\times n}{\overline{\mathscr{X}}} &= \begin{pmatrix} \overline{\mathbf{X}} & \cdots & \overline{\mathbf{X}} \end{pmatrix} \\
\underset{s\times n}{\overline{\mathscr{Y}}} &= \begin{pmatrix} \overline{\mathbf{Y}} & \cdots & \overline{\mathbf{Y}} \end{pmatrix} \\
\underset{r\times n}{\mathscr{X}_c} &= \mathscr{X} - \overline{\mathscr{X}} \\
\underset{s\times n}{\mathscr{Y}_c} &= \mathscr{Y} - \overline{\mathscr{Y}}
\end{aligned}
$$

Consider the multivariate linear regression model

$$
\underset{s\times n}{\mathscr{Y}} = \underset{s\times n}{\boldsymbol{\mu}} + \underset{s\times r}{\boldsymbol{\Theta}}\,\mathscr{X} + \underset{s\times n}{\boldsymbol{\varepsilon}}
$$

where $\boldsymbol{\mu} = \underset{s\times 1}{\boldsymbol{\mu}_0}\mathbf{1}_n^\top$ is an matrix of unknown constants, $\boldsymbol{\Theta} = (\theta_{jk})$ is a matrix of unknown regression coefficients, and $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, \cdots, \boldsymbol{\varepsilon}_n)$ is the error matrix whose columns are each random s-vectors with mean 0 and the same unknown nonsingular $(s \times s)$ error covariance matrix $\Sigma_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}}$, and pairs of column vectors, $(\boldsymbol{\varepsilon}_j, \boldsymbol{\varepsilon}_k)$, $j \neq k$, are uncorrelated with each other.

If we set $\boldsymbol{\mu} = \overline{\mathscr{Y}} - \boldsymbol{\Theta}\overline{\mathscr{X}}$, we get

$$
\underset{s\times n}{\mathscr{Y}_c} = \underset{s\times r}{\boldsymbol{\Theta}}\,\mathscr{X}_c + \underset{s\times n}{\boldsymbol{\varepsilon}}
$$

The least-squares estimator is given by

$$
\begin{aligned}
\hat{\boldsymbol{\Theta}} &= \mathscr{Y}_c\mathscr{X}_c^\top(\mathscr{X}_c\mathscr{X}_c^\top)^{-1} \\
\hat{\boldsymbol{\mu}} &= \overline{\mathscr{Y}} - \hat{\boldsymbol{\Theta}}\overline{\mathscr{X}}
\end{aligned}
$$

Proof. By vectorizing, we have

$$
\underset{sn\times 1}{vec(\mathscr{Y})_c} = (\underset{sn\times sr}{\mathbf{I}_s \otimes \mathscr{X}_c^\top})\underset{sr\times 1}{vec(\hat{\boldsymbol{\Theta}})} + \underset{sn\times 1}{vec(\boldsymbol{\varepsilon})}
$$

which turns to a multiple regression problem.

We have

$$\begin{aligned}
\mathbb{E}vec(\boldsymbol{\varepsilon}) &= \mathbf{0} \\
Cov(vec(\boldsymbol{\varepsilon})) &= \mathbb{E}[(vec(\boldsymbol{\varepsilon}))(vec(\boldsymbol{\varepsilon}))^\top] \\
&= \mathbb{E}\left[\begin{pmatrix}\boldsymbol{\varepsilon}_1^\top & \cdots & \boldsymbol{\varepsilon}_n^\top\end{pmatrix}^\top \begin{pmatrix}\boldsymbol{\varepsilon}_1^\top & \cdots & \boldsymbol{\varepsilon}_n^\top\end{pmatrix}\right] \\
&= \begin{pmatrix}\mathbb{E}(\boldsymbol{\varepsilon}_1^\top\boldsymbol{\varepsilon}_1) & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbb{E}(\boldsymbol{\varepsilon}_2^\top\boldsymbol{\varepsilon}_2) & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{E}(\boldsymbol{\varepsilon}_n^\top\boldsymbol{\varepsilon}_n)\end{pmatrix} \\
&= \Sigma_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}} \otimes \mathbf{I}_n
\end{aligned}$$

since for $i \neq j$,

$$\mathbb{E}(\boldsymbol{\varepsilon}_i^\top \boldsymbol{\varepsilon}_j) = \mathbf{0}$$

The generalized least-squares estimator of $vec(\boldsymbol{\Theta})$ is given by

$$\begin{aligned}
vec(\hat{\boldsymbol{\Theta}}) &= [(\mathbf{I}_s \otimes \mathscr{X}_c^\top)^\top(\Sigma_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}} \otimes \mathbf{I}_n)^{-1}(\mathbf{I}_s \otimes \mathscr{X}_c^\top)]^{-1}(\mathbf{I}_s \otimes \mathscr{X}_c^\top)^\top(\Sigma_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}} \otimes \mathbf{I}_n)^{-1}vec(\mathscr{Y}_c) \\
&= [(\mathbf{I}_s \otimes \mathscr{X}_c)(\Sigma_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}}^{-1} \otimes \mathbf{I}_n)(\mathbf{I}_s \otimes \mathscr{X}_c^\top)]^{-1}(\mathbf{I}_s \otimes \mathscr{X}_c)(\Sigma_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}}^{-1} \otimes \mathbf{I}_n)vec(\mathscr{Y}_c) \\
&= [(\Sigma_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}}^{-1} \otimes \mathscr{X}_c)(\mathbf{I}_s \otimes \mathscr{X}_c^\top)]^{-1}(\Sigma_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}}^{-1} \otimes \mathscr{X}_c)vec(\mathscr{Y}_c) \\
&= (\Sigma_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}}^{-1} \otimes \mathscr{X}_c\mathscr{X}_c^\top)^{-1}(\Sigma_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}}^{-1} \otimes \mathscr{X}_c)vec(\mathscr{Y}_c) \\
&= [\Sigma_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}} \otimes (\mathscr{X}_c\mathscr{X}_c^\top)^{-1}](\Sigma_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}}^{-1} \otimes \mathscr{X}_c)vec(\mathscr{Y}_c) \\
&= [(\mathbf{I}_s \otimes (\mathscr{X}_c\mathscr{X}_c^\top)^{-1}\mathscr{X}_c)]vec(\mathscr{Y}_c)
\end{aligned}$$

By un-vectorizing, it follwos that

$$\begin{aligned}
\hat{\boldsymbol{\Theta}} &= \mathscr{Y}_c\mathscr{X}_c^\top(\mathscr{X}_c\mathscr{X}_c^\top)^{-1} \\
\hat{\boldsymbol{\mu}} &= \overline{\mathscr{Y}} - \hat{\boldsymbol{\Theta}}\overline{\mathscr{X}} \\
\hat{\boldsymbol{\mu}} &= \overline{\mathbf{Y}} - \hat{\boldsymbol{\Theta}}\overline{\mathbf{X}}
\end{aligned}$$

□

8.2 Gauss–Markov Theorem

Consider the problem of estimating arbitrary linear combinations of the $\{\theta_{jk}\}$,

$$tr(\mathbf{A}\boldsymbol{\Theta}) = \sum_{j,k} A_{jk}\theta_{jk}$$

where $\mathbf{A} = (A_{jk})$ is an arbitrary matrix of constants.

Under the above conditions and if $\mathscr{X}_c\mathscr{X}_c^\top$ is nonsingular, then the minimum-variance linear unbiased estimator of $tr(\mathbf{A}\boldsymbol{\Theta})$ is given by $tr(\mathbf{A}\hat{\boldsymbol{\Theta}})$. This is the multivariate form of the Gauss–Markov theorem.

8.3 Relationship with OLS

Since

$$\mathscr{Y}_c^\top = \mathscr{X}_c^\top \boldsymbol{\Theta}^\top + \boldsymbol{\varepsilon}^\top$$

let $\mathbf{Z} = \mathscr{Y}_c^\top$, $W = \mathscr{X}_c^\top$, $\boldsymbol{\beta} = \boldsymbol{\Theta}^\top$ and $\mathbf{E} = \boldsymbol{\varepsilon}^\top$, then

$$\underset{n\times s}{\mathbf{Z}} = \underset{n\times r}{\mathbf{W}}\underset{r\times s}{\boldsymbol{\beta}} + \underset{n\times s}{\mathbf{E}}$$

Each columns will give a ordinary least-squares problem,

$$\mathbf{Z}_i = \mathbf{W}\boldsymbol{\beta}_i + \mathbf{E}_i$$

The OLS estimator is given by

$$\hat{\boldsymbol{\beta}}_i = (\mathbf{W}^\top \mathbf{W})^{-1}\mathbf{W}^\top \mathbf{Z}_i$$

Thus the ith row of $\hat{\boldsymbol{\Theta}}$ is given by

$$\hat{\boldsymbol{\Theta}}_{(i)} = \mathscr{Y}_{c(i)}\mathscr{X}_c^\top(\mathscr{X}_c\mathscr{X}_c^\top)^{-1}$$

where $\mathscr{Y}_{c(i)}$ is the ith row of $\mathscr{Y}_c$.

That means that even though the variables in $\mathscr{Y}$ may be correlated, perhaps even heavily correlated, the LS estimator $\hat{\boldsymbol{\Theta}}$ of $\boldsymbol{\Theta}$ does not contain any reference to that correlation.

8.4 Properties of $\hat{\boldsymbol{\Theta}}$

8.4.1 Covariance Matrix

Since

$$\begin{aligned}
\hat{\boldsymbol{\Theta}} &= (\boldsymbol{\Theta}\mathscr{X}_c + \boldsymbol{\varepsilon})\mathscr{X}_c^\top(\mathscr{X}_c\mathscr{X}_c^\top)^{-1} \\
&= \boldsymbol{\Theta} + \boldsymbol{\varepsilon}\mathscr{X}_c^\top(\mathscr{X}_c\mathscr{X}_c^\top)^{-1} \\
vec(\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}) &= vec(\boldsymbol{\varepsilon}\mathscr{X}_c^\top(\mathscr{X}_c\mathscr{X}_c^\top)^{-1}) \\
&= (\mathbf{I}_s \otimes (\mathscr{X}_c\mathscr{X}_c^\top)^{-1}\mathscr{X}_c)vec(\boldsymbol{\varepsilon})
\end{aligned}$$

we have

$$\begin{aligned}
\mathbb{E}\hat{\boldsymbol{\Theta}} &= \boldsymbol{\Theta} \\
Cov(vec(\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})) &= \mathbb{E}\{vec(\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})[vec(\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})]^\top\} \\
&= (\mathbf{I}_s \otimes (\mathscr{X}_c\mathscr{X}_c^\top)^{-1}\mathscr{X}_c)\mathbb{E}[vec(\boldsymbol{\varepsilon})vec(\boldsymbol{\varepsilon})^\top](\mathbf{I}_s \otimes (\mathscr{X}_c\mathscr{X}_c^\top)^{-1}\mathscr{X}_c)^\top \\
&= (\mathbf{I}_s \otimes (\mathscr{X}_c\mathscr{X}_c^\top)^{-1}\mathscr{X}_c)(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}} \otimes \mathbf{I}_n)(\mathbf{I}_s \otimes \mathscr{X}_c^\top(\mathscr{X}_c\mathscr{X}_c^\top)^{-1}) \\
&= \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}} \otimes (\mathscr{X}_c\mathscr{X}_c^\top)^{-1}
\end{aligned}$$

8.4.2 Fitted Values and Multivariate Residuals

8.5 Separate and Multivariate Ridge Regressions

Since

$$\begin{aligned} vec(\hat{\mathbf{\Theta}}) &= [(\mathbf{I}_s \otimes (\mathscr{X}_c \mathscr{X}_c^\top)^{-1} \mathscr{X}_c)] vec(\mathscr{Y}_c) \\ &= [\mathbf{I}_s \otimes (\mathscr{X}_c \mathscr{X}_c^\top)^{-1}](\mathbf{I}_s \otimes \mathscr{X}_c) vec(\mathscr{Y}_c) \\ &= (\mathbf{I}_s \otimes \mathscr{X}_c \mathscr{X}_c^\top)^{-1}(\mathbf{I}_s \otimes \mathscr{X}_c) vec(\mathscr{Y}_c) \end{aligned}$$

Introducing a positive-definite $(s \times s)$ ridge matrix $\mathbf{K}$ so that a multivariate ridge regression estimator of $vec(\mathbf{\Theta})$ is given by

$$vec(\hat{\mathbf{\Theta}}(\mathbf{K})) = [(\mathbf{I}_s \otimes \mathscr{X}_c \mathscr{X}_c^\top) + (\mathbf{K} \otimes \mathbf{I}_r)]^{-1}(\mathbf{I}_s \otimes \mathscr{X}_c) vec(\mathscr{Y}_c)$$

8.6 Linear Constraints on the Regression Coefficients

The general set of linear constraints is given by

$$\underset{m \times s}{\mathbf{K}} \; \underset{s \times r}{\mathbf{\Theta}} \; \underset{r \times u}{\mathbf{L}} = \underset{m \times u}{\mathbf{\Gamma}}$$

where the matrix $\mathbf{K}$ $(m \leq s)$ and the matrix $\mathbf{L}$ $(u \leq r)$ are full-rank matrices of known constants, and $\mathbf{\Gamma}$ is a matrix of parameters (known or unknown). We often take $\mathbf{\Gamma} = \mathbf{0}$.

Consider the problem of finding $\hat{\mathbf{\Theta}}^*$ that solves the following constrained minimization problem:

$$\begin{aligned} \hat{\mathbf{\Theta}}^* &= \underset{\mathbf{K\Theta L}=\mathbf{\Gamma}}{\arg\min}\, tr\{(\mathscr{Y}_c - \mathbf{\Theta}\mathscr{X}_c)^\top (\mathscr{Y}_c - \mathbf{\Theta}\mathscr{X}_c)\} \\ &= \underset{\mathbf{K\Theta L}=\mathbf{\Gamma}}{\arg\min}\, \|\mathscr{Y}_c - \mathbf{\Theta}\mathscr{X}_c\|_F \end{aligned}$$

8.6.1 Normal Equation

The normal equation is given by

$$\begin{aligned} \hat{\mathbf{\Theta}}^* \mathscr{X}_c \mathscr{X}_c^\top + \mathbf{K}^\top \mathbf{\Lambda} \mathbf{L}^\top &= \mathscr{Y}_c \mathscr{X}_c^\top \\ \mathbf{K}\hat{\mathbf{\Theta}}^* \mathbf{L} &= \mathbf{\Gamma} \end{aligned}$$

Method 1

Proof. Let $\mathbf{\Lambda} = (\lambda_{ij})$ be a matrix of Lagrangian coefficients.

$$\begin{aligned} f[vec(\mathbf{\Theta}), vec(\mathbf{\Lambda})] &= [vec(\mathscr{Y}_c - \mathbf{\Theta}\mathscr{X}_c)]^\top [vec(\mathscr{Y}_c - \mathbf{\Theta}\mathscr{X}_c)] - 2[vec(\mathbf{\Lambda})]^\top [vec(\mathbf{K\Theta L} - \mathbf{\Gamma})] \\ &= [vec(\mathscr{Y}_c)]^\top [vec(\mathscr{Y}_c)] - 2[vec(\mathscr{Y}_c)]^\top (\mathbf{I}_s \otimes \mathscr{X}_c) vec(\mathbf{\Theta}) \\ &\quad [vec(\mathbf{\Theta})]^\top (\mathbf{I}_s \otimes \mathscr{X}_c)^\top (\mathbf{I}_s \otimes \mathscr{X}_c) vec(\mathbf{\Theta}) - 2[vec(\mathbf{\Lambda})]^\top [(\mathbf{K} \otimes \mathbf{L}^\top) vec(\mathbf{\Theta}) - vec(\mathbf{\Gamma})] \end{aligned}$$

Let

$$\frac{\partial f[vec(\mathbf{\Theta}), vec(\mathbf{\Lambda})]}{\partial vec(\mathbf{\Theta})} = -2(\mathbf{I}_s \otimes \mathscr{X}_c)^\top vec(\mathscr{Y}_c) + 2(\mathbf{I}_s \otimes \mathscr{X}_c^\top \mathscr{X}_c) vec(\mathbf{\Theta}) - 2(\mathbf{K}^\top \otimes \mathbf{L}) vec(\mathbf{\Lambda}) = 0$$

we have

$$(\mathbf{I}_s \otimes \mathscr{X}_c^\top \mathscr{X}_c) vec(\Theta) + (\mathbf{K}^\top \otimes \mathbf{L}) vec(\mathbf{\Lambda}) = (\mathbf{I}_s \otimes \mathscr{X}_c)^\top vec(\mathscr{Y}_c)$$

By un-vectorizing, we have

$$\hat{\mathbf{\Theta}} \mathbf{X}_c \mathbf{X}_c^\top + \mathbf{K}^\top \Lambda \mathbf{L}^\top = \mathbf{Y}_c \mathbf{X}_c^\top$$

Therefore, $\hat{\mathbf{\Theta}}^*$ satisfies the normal equation

$$\hat{\Theta}^* \mathscr{X}_c \mathscr{X}_c^\top + \mathbf{K}^\top \Lambda \mathbf{L}^\top = \mathscr{Y}_c \mathscr{X}_c^\top$$
$$\mathbf{K} \hat{\Theta}^* \mathbf{L} = \Gamma$$

Method 2

$$f(\mathbf{\Theta}, \mathbf{\Lambda}) = tr\{(\mathscr{Y}_c - \mathbf{\Theta} \mathscr{X}_c)^\top (\mathscr{Y}_c - \mathbf{\Theta} \mathscr{X}_c)\} - 2tr\{\mathbf{\Lambda}^\top (\mathbf{K} \mathbf{\Theta} \mathbf{L} - \mathbf{\Gamma})\}$$

Since

$$\frac{\partial tr\{F(\mathbf{X})\}}{\partial \mathbf{X}} = F_X(\mathbf{X})^\top$$
$$\frac{\partial tr\{\mathbf{X}^\top \mathbf{A}\}}{\partial \mathbf{X}} = \mathbf{A}$$
$$\frac{\partial \mathbf{a}^\top \mathbf{X} \mathbf{b}}{\partial \mathbf{X}} = \mathbf{b}\mathbf{a}^\top$$
$$\frac{\partial \mathbf{b}^\top \mathbf{X}^\top \mathbf{X} \mathbf{c}}{\partial \mathbf{X}} = \mathbf{X}(\mathbf{b}\mathbf{c}^\top + \mathbf{c}\mathbf{b}^\top)$$

Let

$$\frac{\partial f}{\partial \Theta} = 2\mathscr{Y}_c \mathscr{X}_c^\top - 2\hat{\mathbf{\Theta}} \mathscr{X}_c \mathscr{X}_c^\top - 2\mathbf{K}^\top \mathbf{\Lambda} \mathbf{L}^\top = 0$$
$$\frac{\partial f}{\partial \mathbf{\Lambda}} = \mathbf{K} \mathbf{\Theta} \mathbf{L} - \mathbf{\Gamma} = 0$$

Therefore,

$$\hat{\mathbf{\Theta}}^* \mathscr{X}_c \mathscr{X}_c^\top + \mathbf{K}^\top \mathbf{\Lambda} \mathbf{L}^\top = \mathscr{Y}_c \mathscr{X}_c^\top$$
$$\mathbf{K} \hat{\mathbf{\Theta}}^* \mathbf{L} = \mathbf{\Gamma}$$

□

8.6.2 Solution

Proof. Since

$$\hat{\boldsymbol{\Theta}}^{*}\mathscr{X}_c\mathscr{X}_c^{\top}+\mathbf{K}^{\top}\boldsymbol{\Lambda}\mathbf{L}^{\top}=\mathscr{Y}_c\mathscr{X}_c^{\top}$$

we have

$$\begin{aligned}\hat{\boldsymbol{\Theta}}^{*}+\mathbf{K}^{\top}\boldsymbol{\Lambda}\mathbf{L}^{\top}(\mathscr{X}_c\mathscr{X}_c^{\top})^{-1}&=\mathscr{Y}_c\mathscr{X}_c^{\top}(\mathscr{X}_c\mathscr{X}_c^{\top})^{-1}\\&=\hat{\boldsymbol{\Theta}}\end{aligned}$$

Premultiplicating $\mathbf{K}$ and postmultiplicating $\mathbf{L}$, we have

$$\begin{aligned}\mathbf{K}\hat{\boldsymbol{\Theta}}^{*}\mathbf{L}+\mathbf{K}\mathbf{K}^{\top}\boldsymbol{\Lambda}\mathbf{L}^{\top}(\mathscr{X}_c\mathscr{X}_c^{\top})^{-1}\mathbf{L}&=\mathbf{K}\hat{\boldsymbol{\Theta}}\mathbf{L}\\\boldsymbol{\Gamma}+\mathbf{K}\mathbf{K}^{\top}\boldsymbol{\Lambda}\mathbf{L}^{\top}(\mathscr{X}_c\mathscr{X}_c^{\top})^{-1}\mathbf{L}&=\mathbf{K}\hat{\boldsymbol{\Theta}}\mathbf{L}\\\mathbf{K}\mathbf{K}^{\top}\boldsymbol{\Lambda}\mathbf{L}^{\top}(\mathscr{X}_c\mathscr{X}_c^{\top})^{-1}\mathbf{L}&=\mathbf{K}\hat{\boldsymbol{\Theta}}\mathbf{L}-\boldsymbol{\Gamma}\\\boldsymbol{\Lambda}&=(\mathbf{K}\mathbf{K}^{\top})^{-1}(\mathbf{K}\hat{\boldsymbol{\Theta}}\mathbf{L}-\boldsymbol{\Gamma})[\mathbf{L}^{\top}(\mathscr{X}_c\mathscr{X}_c^{\top})^{-1}\mathbf{L}]^{-1}\end{aligned}$$

Therefore,

$$\hat{\boldsymbol{\Theta}}^{*}=\hat{\boldsymbol{\Theta}}-\mathbf{K}^{\top}(\mathbf{K}\mathbf{K}^{\top})^{-1}(\mathbf{K}\hat{\boldsymbol{\Theta}}\mathbf{L}-\boldsymbol{\Gamma})[\mathbf{L}^{\top}(\mathscr{X}_c\mathscr{X}_c^{\top})^{-1}\mathbf{L}]^{-1}\mathbf{L}^{\top}(\mathscr{X}_c\mathscr{X}_c^{\top})^{-1}$$

□

9 The Random-X Case

9.1 Regression Model

Suppose $\mathbf{Y}$ is related to $\mathbf{X}$ by the following multivariate linear model:

$$\underset{s\times 1}{\mathbf{Y}}=\underset{s\times 1}{\boldsymbol{\mu}}+\underset{s\times r}{\boldsymbol{\Theta}}\underset{r\times 1}{\mathbf{X}}+\underset{s\times 1}{\boldsymbol{\varepsilon}}$$

where $\boldsymbol{\mu}=\underset{s\times 1}{\boldsymbol{\mu}_0}\mathbf{1}_n^{\top}$ is an matrix of unknown constants, $\boldsymbol{\Theta}=(\theta_{jk})$ is a matrix of unknown regression coefficients, and $\boldsymbol{\varepsilon}=(\boldsymbol{\varepsilon}_1,\boldsymbol{\varepsilon}_2,\cdots,\boldsymbol{\varepsilon}_n)$ is the error matrix independent to $\mathbf{X}$ whose columns are each random s-vectors with mean 0 and the same unknown nonsingular $(s\times s)$ error covariance matrix $\Sigma_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}}$, and pairs of column vectors, $(\boldsymbol{\varepsilon}_j,\boldsymbol{\varepsilon}_k)$, $j\neq k$, are uncorrelated with each other.

We are interested in finding the s-vector $\boldsymbol{\mu}$ and $(s\times r)$-matrix $\boldsymbol{\Theta}$ that minimize the $(s\times s)$-matrix,

$$W(\boldsymbol{\mu},\boldsymbol{\Theta})=\mathbb{E}\{(\mathbf{Y}-\boldsymbol{\mu}-\boldsymbol{\Theta}\mathbf{X})(\mathbf{Y}-\boldsymbol{\mu}-\boldsymbol{\Theta}\mathbf{X})^{\top}\}$$

We have

$$\boldsymbol{\mu}^*,\ \boldsymbol{\Theta}^* = \arg\min_{\boldsymbol{\mu},\boldsymbol{\Theta}} W(\boldsymbol{\mu},\boldsymbol{\Theta})$$

where

$$\begin{aligned}\boldsymbol{\mu}^* &= \boldsymbol{\mu}_{\mathbf{Y}} - \boldsymbol{\Theta}^*\boldsymbol{\mu}_{\mathbf{X}} \\ \boldsymbol{\Theta}^* &= \Sigma_{\mathbf{YX}}\Sigma_{\mathbf{XX}}^{-1}\end{aligned}$$

Proof. Set $\mathbf{Y}_c = \mathbf{Y} - \boldsymbol{\mu}_{\mathbf{Y}}$ and $\mathbf{X}_c = \mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}}$, and assume that $\Sigma_{\mathbf{XX}}$ is nonsingular. We get that

$$\begin{aligned}W(\boldsymbol{\mu},\boldsymbol{\Theta}) &= \mathbb{E}\{(\mathbf{Y}_c - \boldsymbol{\Theta}\mathbf{X}_c + \boldsymbol{\mu}_{\mathbf{Y}} - \boldsymbol{\mu} - \boldsymbol{\Theta}\boldsymbol{\mu}_{\mathbf{X}})(\mathbf{Y}_c - \boldsymbol{\Theta}\mathbf{X}_c + \boldsymbol{\mu}_{\mathbf{Y}} - \boldsymbol{\mu} - \boldsymbol{\Theta}\boldsymbol{\mu}_{\mathbf{X}})^\top\} \\ &= \mathbb{E}\{\mathbf{Y}_c\mathbf{Y}_c^\top - \mathbf{Y}_c\mathbf{X}_c^\top\boldsymbol{\Theta}^\top - \boldsymbol{\Theta}\mathbf{X}_c\mathbf{Y}_c^\top + \boldsymbol{\Theta}\mathbf{X}_c\mathbf{X}_c^\top\boldsymbol{\Theta}^\top\} \\ &\quad + (\boldsymbol{\mu} - \boldsymbol{\mu}_{\mathbf{Y}} + \boldsymbol{\Theta}\boldsymbol{\mu}_{\mathbf{X}})(\boldsymbol{\mu} - \boldsymbol{\mu}_{\mathbf{Y}} + \boldsymbol{\Theta}\boldsymbol{\mu}_{\mathbf{X}})^\top \\ &= \Sigma_{\mathbf{YY}} - \Sigma_{\mathbf{YX}}\boldsymbol{\Theta}^\top - \boldsymbol{\Theta}\Sigma_{\mathbf{XY}} + \boldsymbol{\Theta}\Sigma_{\mathbf{XX}}\boldsymbol{\Theta}^\top \\ &\quad + (\boldsymbol{\mu} - \boldsymbol{\mu}_{\mathbf{Y}} + \boldsymbol{\Theta}\boldsymbol{\mu}_{\mathbf{X}})(\boldsymbol{\mu} - \boldsymbol{\mu}_{\mathbf{Y}} + \boldsymbol{\Theta}\boldsymbol{\mu}_{\mathbf{X}})^\top \\ &= (\Sigma_{\mathbf{YY}} - \Sigma_{\mathbf{YX}}\Sigma_{\mathbf{XX}}^{-1}\Sigma_{\mathbf{XY}}) \\ &\quad + (\Sigma_{\mathbf{YX}}\Sigma_{\mathbf{XX}}^{-\frac{1}{2}} - \boldsymbol{\Theta}\Sigma_{\mathbf{XX}}^{\frac{1}{2}})(\Sigma_{\mathbf{YX}}\Sigma_{\mathbf{XX}}^{-\frac{1}{2}} - \boldsymbol{\Theta}\Sigma_{\mathbf{XX}}^{\frac{1}{2}})^\top \\ &\quad + (\boldsymbol{\mu} - \boldsymbol{\mu}_{\mathbf{Y}} + \boldsymbol{\Theta}\boldsymbol{\mu}_{\mathbf{X}})(\boldsymbol{\mu} - \boldsymbol{\mu}_{\mathbf{Y}} + \boldsymbol{\Theta}\boldsymbol{\mu}_{\mathbf{X}})^\top \\ &\geq \Sigma_{\mathbf{YY}} - \Sigma_{\mathbf{YX}}\Sigma_{\mathbf{XX}}^{-1}\Sigma_{\mathbf{XY}}\end{aligned}$$

with equality when

$$\begin{aligned}\boldsymbol{\mu}^* &= \boldsymbol{\mu}_{\mathbf{Y}} - \boldsymbol{\Theta}^*\boldsymbol{\mu}_{\mathbf{X}} \\ \boldsymbol{\Theta}^* &= \Sigma_{\mathbf{YX}}\Sigma_{\mathbf{XX}}^{-1}\end{aligned}$$

□

9.2 Multivariate Reduced-Rank Regression

Through a rank condition on the matrix of regression coefficients we can constrain a multivariate regression model.

Consider the multivariate linear regression model given by

$$\underset{s\times 1}{\mathbf{Y}} = \underset{s\times 1}{\boldsymbol{\mu}} + \underset{s\times r}{\mathbf{C}}\ \underset{r\times 1}{\mathbf{X}} + \underset{s\times 1}{\boldsymbol{\varepsilon}}$$

where $\boldsymbol{\mu}$ and $\mathbf{C}$ are unknown regression parameters, and the unobservable error variate $\boldsymbol{\varepsilon}$ of the model has mean $\mathbb{E}(\boldsymbol{\varepsilon}) = 0$ and covariance matrix $Cov(\boldsymbol{\varepsilon}) = \mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \Sigma_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}}$, and is distributed independently of $\mathbf{X}$. We allow the possibility that the rank of the regression coefficient matrix $\mathbf{C}$ is deficient; that is,

$$rank(\mathbf{C}) = t \leq \min\{r, s\}$$

When $\mathbf{C}$ has reduced-rank t, then, there exist two (nonunique) full-rank matrices, an $(s \times t)$ matrix $\mathbf{A}$ and a $(t \times r)$ matrix $\mathbf{B}$, such that $\mathbf{C} = \mathbf{AB}$. The model can now be written as

$$\mathbf{Y} = \boldsymbol{\mu} + \mathbf{ABX} + \boldsymbol{\varepsilon}$$

Given a sample, $(\mathbf{X}_1^\top, \mathbf{Y}_1^\top)^\top, \cdots, (\mathbf{X}_n^\top, \mathbf{Y}_n^\top)^\top$ of observations on $(\mathbf{X}^\top, \mathbf{Y}^\top)^\top$, our goal is to estimate the parameters $\boldsymbol{\mu}$, $\mathbf{A}$ and $\mathbf{B}$ (and, hence, $\mathbf{C}$) in some optimal manner.

One of the primary aspects of reduced-rank regression is to assess the unknown value of the metaparameter t, which we call the effective dimensionality of the multivariate regression. Given that $\boldsymbol{\Gamma}$ is a positive-definite symmetric $(s \times s)$-matrix of weights, the minimizing criterion becomes:

Proof.

$$\begin{aligned}
W(t) &= \mathbb{E}\{(\mathbf{Y} - \boldsymbol{\mu} - \mathbf{CX})^\top \boldsymbol{\Gamma} (\mathbf{Y} - \boldsymbol{\mu} - \mathbf{CX})\} \\
&= \mathbb{E}\{\mathbf{Y}_c^\top \boldsymbol{\Gamma} \mathbf{Y}_c - \mathbf{Y}_c^\top \boldsymbol{\Gamma} \mathbf{C} \mathbf{X}_c - \mathbf{X}_c^\top \mathbf{C}^\top \boldsymbol{\Gamma} \mathbf{Y}_c + \mathbf{X}_c^\top \mathbf{C}^\top \boldsymbol{\Gamma} \mathbf{C} \mathbf{X}_c\} \\
&\quad + (\boldsymbol{\mu} - \boldsymbol{\mu}_\mathbf{Y} + \mathbf{C}\boldsymbol{\mu}_\mathbf{X})^\top \boldsymbol{\Gamma} (\boldsymbol{\mu} - \boldsymbol{\mu}_\mathbf{Y} + \mathbf{C}\boldsymbol{\mu}_\mathbf{X}) \\
&\geq \mathbb{E}\{\mathbf{Y}_c^\top \boldsymbol{\Gamma} \mathbf{Y}_c - \mathbf{Y}_c^\top \boldsymbol{\Gamma} \mathbf{C} \mathbf{X}_c - \mathbf{X}_c^\top \mathbf{C}^\top \boldsymbol{\Gamma} \mathbf{Y}_c + \mathbf{X}_c^\top \mathbf{C}^\top \boldsymbol{\Gamma} \mathbf{C} \mathbf{X}_c\} \\
&= tr\{\Sigma^*_{\mathbf{YY}} - \mathbf{C}^* \Sigma^*_{\mathbf{XY}} - \Sigma^*_{\mathbf{YX}} \mathbf{C}^{*\top} + \mathbf{C}^* \Sigma^*_{\mathbf{XX}} \mathbf{C}^{*\top}\} \qquad \text{(Property of Quadratic Forms)} \\
&= tr\{(\Sigma^*_{\mathbf{YY}} - \Sigma^*_{\mathbf{YX}} \Sigma^{*-1}_{\mathbf{XX}} \Sigma^*_{\mathbf{XY}}) \\
&\quad + (\Sigma^*_{\mathbf{YX}} \Sigma^{*-\frac{1}{2}}_{\mathbf{XX}} - \mathbf{C}^* \Sigma^{*\frac{1}{2}}_{\mathbf{XX}})(\Sigma^*_{\mathbf{YX}} \Sigma^{*-\frac{1}{2}}_{\mathbf{XX}} - \mathbf{C}^* \Sigma^{*\frac{1}{2}}_{\mathbf{XX}})^\top\}
\end{aligned}$$

where

$$\begin{aligned}
\Sigma^*_{\mathbf{XX}} &= \Sigma_{\mathbf{XX}} \\
\Sigma^*_{\mathbf{YY}} &= \boldsymbol{\Gamma}^{\frac{1}{2}} \Sigma_{\mathbf{YY}} \boldsymbol{\Gamma}^{\frac{1}{2}} \\
\Sigma^*_{\mathbf{XY}} &= \Sigma_{\mathbf{XY}} \boldsymbol{\Gamma}^{\frac{1}{2}} \\
\mathbf{C}^* &= \boldsymbol{\Gamma}^{\frac{1}{2}} \mathbf{C}
\end{aligned}$$

with equality when

$$\boldsymbol{\mu} = \boldsymbol{\mu}_\mathbf{Y} - \mathbf{C}\boldsymbol{\mu}_\mathbf{X}$$

From Eckart–Young Theorem, the last expression is minimized by setting

$$\mathbf{C}^* \Sigma^{\frac{1}{2}}_{\mathbf{XX}} = \sum_{j=1}^{t} \lambda_j^{\frac{1}{2}} \mathbf{v}_j \mathbf{w}_j^\top$$

where $\mathbf{v}_j$ and $\mathbf{w}_j$ are the jth left and right singular vectors of $\Sigma^*_{\mathbf{YX}} \Sigma^{*-\frac{1}{2}}_{\mathbf{XX}}$ respectively, i.e.,

$$\Sigma^*_{\mathbf{YX}} \Sigma^{*-\frac{1}{2}}_{\mathbf{XX}} = \sum_{j=1}^{s} \lambda_j^{\frac{1}{2}} \mathbf{v}_j \mathbf{w}_j^\top$$

Equivalently, $\mathbf{v}_j$ is the eigenvector associated with the jth largest eigenvalue λ_j of the matrix

$$\Sigma^*_{YX} \Sigma^{*-1}_{XX} \Sigma^*_{XY} = \boldsymbol{\Gamma}^{\frac{1}{2}} \Sigma_{YX} \Sigma^{-1}_{XX} \Sigma_{XY} \boldsymbol{\Gamma}^{\frac{1}{2}}$$

and

$$\mathbf{w}_j = \lambda_j^{-\frac{1}{2}}(\Sigma^*_{\mathbf{YX}}\Sigma^{*-\frac{1}{2}}_{\mathbf{XX}})^\top \mathbf{v}_j = \lambda_j^{-\frac{1}{2}}\Sigma^{-\frac{1}{2}}_{\mathbf{XX}}\Sigma_{\mathbf{XY}}\mathbf{\Gamma}^{\frac{1}{2}}\mathbf{v}_j$$

Thus, the minimizing $\mathbf{C}$ with reduced-rank t is given by

$$\begin{aligned}
\mathbf{C}^{(t)} &= \mathbf{\Gamma}^{-\frac{1}{2}} \sum_{j=1}^{t} \lambda_j^{\frac{1}{2}} \mathbf{v}_j \mathbf{w}_j^\top \\
&= \mathbf{\Gamma}^{-\frac{1}{2}} \sum_{j=1}^{t} \lambda_j^{\frac{1}{2}} \mathbf{v}_j (\lambda_j^{-\frac{1}{2}} \Sigma^{-\frac{1}{2}}_{\mathbf{XX}} \Sigma_{\mathbf{XY}} \mathbf{\Gamma}^{\frac{1}{2}} \mathbf{v}_j)^\top \\
&= \mathbf{\Gamma}^{-\frac{1}{2}} \left(\sum_{j=1}^{t} \mathbf{v}_j \mathbf{v}_j^\top \right) \mathbf{\Gamma}^{\frac{1}{2}} \Sigma_{\mathbf{YX}} \Sigma^{-1}_{\mathbf{XX}}
\end{aligned}$$

Therefore,

$$\begin{aligned}
\boldsymbol{\mu}^{(t)} &= \boldsymbol{\mu}_{\mathbf{Y}} - \mathbf{C}^{(t)} \boldsymbol{\mu}_{\mathbf{X}} \\
\mathbf{A}^{(t)} &= \mathbf{\Gamma}^{-\frac{1}{2}} \mathbf{V}_t \\
\mathbf{B}^{(t)} &= \mathbf{V}_t^\top \mathbf{\Gamma}^{\frac{1}{2}} \Sigma_{\mathbf{YX}} \Sigma^{-1}_{XX}
\end{aligned}$$

where

$$\mathbf{V}_t = \begin{pmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_t \end{pmatrix}$$

which gives the minimum value of $W(t)$,

$$\begin{aligned}
W_{\min}(t) &= tr\{(\Sigma^*_{\mathbf{YY}} - \Sigma^*_{\mathbf{YX}} \Sigma^{*-1}_{\mathbf{XX}} \Sigma^*_{\mathbf{XY}}) \\
&\qquad + (\Sigma^*_{\mathbf{YX}} \Sigma^{*-\frac{1}{2}}_{\mathbf{XX}} - \mathbf{C}^{(t)} \Sigma^{*\frac{1}{2}}_{\mathbf{XX}})(\Sigma^*_{\mathbf{YX}} \Sigma^{*-\frac{1}{2}}_{\mathbf{XX}} - \mathbf{C}^{(t)} \Sigma^{*\frac{1}{2}}_{\mathbf{XX}})^\top\} \\
&= tr\{\Sigma_{\mathbf{YY}} \mathbf{\Gamma}\} - tr\{\Sigma^*_{\mathbf{YX}} \Sigma^{*-1}_{\mathbf{XX}} \Sigma^*_{\mathbf{XY}}\} + \sum_{j=t+1}^{s} \lambda_j \\
&= tr\{\Sigma_{\mathbf{YY}} \mathbf{\Gamma}\} - \sum_{j=1}^{t} \lambda_j
\end{aligned}$$

□

9.3 Sample Estimation

$$\begin{aligned}
\hat{\boldsymbol{\mu}}_{\mathbf{X}} &= \overline{X} \\
\hat{\boldsymbol{\mu}}_{\mathbf{Y}} &= \overline{Y} \\
\hat{\Sigma}_{\mathbf{XX}} &= \frac{1}{n} \mathscr{X}_c \mathscr{X}_c^\top \\
\hat{\Sigma}_{\mathbf{YX}} &= \frac{1}{n} \mathscr{Y}_c \mathscr{X}_c^\top = \hat{\Sigma}^\top_{\mathbf{XY}} \\
\hat{\Sigma}_{\mathbf{YY}} &= \frac{1}{n} \mathscr{Y}_c \mathscr{Y}_c^\top
\end{aligned}$$

Chapter 4 Linear Dimensionality Reduction

10 Principal Component Analysis

10.1 Interpretation by RRR

Let $\mathbf{B} = \begin{pmatrix} \mathbf{b_1} & \cdots & \mathbf{b}_r \end{pmatrix}^\top$ be a $(t \times r)$-matrix of weights $(t \leq r)$. The linear projections can be written as a t-vector,

$$\boldsymbol{\xi} = \mathbf{BX}$$

where $\boldsymbol{\xi} = \begin{pmatrix} \xi_1 & \cdots & \xi_t \end{pmatrix}^\top$

Consider the multivariate linear model

$$\mathbf{X} = \boldsymbol{\mu} + \mathbf{ABX} + \boldsymbol{\varepsilon}$$

where $\mathbb{E}\boldsymbol{\varepsilon} = \mathbf{0}$ and $Cov(\boldsymbol{\varepsilon}) = \Sigma_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}}$, i.e. $\mathbf{Y} = \mathbf{X}$ $(r = s)$ and $\boldsymbol{\Gamma} = I_r$ in RRR.

We want to find an r-vector $\boldsymbol{\mu}$ and an $(r \times t)$-matrix A such that the projections $\boldsymbol{\xi}$ have the property that $\mathbf{X}\boldsymbol{\mu} + \mathbf{A}\boldsymbol{\xi}$ in some least-squares sense. We use the least-squares error criterion,

$$\min_{\boldsymbol{\mu},\mathbf{A},\mathbf{B}} \mathbb{E}[(\mathbf{X} - \boldsymbol{\mu} - \mathbf{A}\boldsymbol{\xi})^\top (\mathbf{X} - \boldsymbol{\mu} - \mathbf{A}\boldsymbol{\xi})]$$

Therefore, for $t < r$,

$$\begin{aligned} \mathbf{A}^{(t)} &= \mathbf{V}_t \\ \mathbf{B}^{(t)} &= \mathbf{V}_t^\top \\ \mathbf{C}^{(t)} &= \mathbf{V}\mathbf{V}^\top \\ \boldsymbol{\mu}^{(t)} &= (\mathbf{I}_t - \mathbf{V}\mathbf{V}^\top)\boldsymbol{\mu}_{\mathbf{X}} \end{aligned}$$

where $\mathbf{V}_t = \begin{pmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_t \end{pmatrix}$ and $\mathbf{v}_i$ is the eigenvector associated with the ith largest eigenvalue of $\Sigma_{\mathbf{XX}}$.

The best rank-t approximation to the original $\mathbf{X}$ is given by

$$\begin{aligned} \hat{\mathbf{X}}^{(t)} &= (\mathbf{I}_t - \mathbf{V}\mathbf{V}^\top)\boldsymbol{\mu}_{\mathbf{X}} + \mathbf{V}\mathbf{V}^\top\mathbf{X} \\ &= \boldsymbol{\mu}_{\mathbf{X}} + \mathbf{V}\mathbf{V}^\top(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}}) \end{aligned}$$

Then the first t principle components of $\mathbf{X}$ are given by $\boldsymbol{\xi}_t = \begin{pmatrix} \xi_1 & \cdots & \xi_t \end{pmatrix}^\top$ where

$$\xi_i = \mathbf{v}_i^\top \mathbf{X}$$

10.2 Interpretation by Optimization

$\mathbf{b}_i$ $(i = 1, 2 \cdots, r)$ are chosen in a sequential manner so that the variances of the derived variables

$$Var\xi_i = \mathbf{b}_i^\top \Sigma_{\mathbf{XX}} \mathbf{b}_i$$

are arranged in descending order subject to the normalizations

$$\begin{aligned} \mathbf{b}_i^\top \mathbf{b}_i &= 1 \\ \mathbf{b}_i^\top \Sigma_{\mathbf{XX}} \mathbf{b}_j &= 0 \qquad (i < j) \end{aligned}$$

i.e., they are uncorrelated with previously chosen derived variables.

(1) For the first principal component, we want

$$\begin{aligned} & \max Var(\mathbf{b}_1^\top \mathbf{X}) \\ s.t. \qquad & \mathbf{b}_1^\top \mathbf{b}_1 = 1 \end{aligned}$$

We form the function

$$f(\mathbf{b}_1) = \mathbf{b}_1^\top \Sigma_{\mathbf{XX}} \mathbf{b}_1 - \lambda_1 (1 - \mathbf{b}_1^\top \mathbf{b}_1)$$

where λ_1 is a Lagrangian multiplier.

Let

$$\frac{\partial f(\mathbf{b}_1)}{\partial \mathbf{b}_1} = 2(\Sigma_{\mathbf{XX}} - \lambda_1 \mathbf{I}_r)\mathbf{b}_1 = 0$$

If $\mathbf{b}_1 \neq \mathbf{0}$, then λ_1 must be chosen to satisfy the determinantal equation

$$|\Sigma_{\mathbf{XX}} - \lambda_1 \mathbf{I}_r| = 0$$

Thus, λ_1 has to be the largest eigenvalue of $\Sigma_{\mathbf{XX}}$, and $\mathbf{b}_1$ the eigenvector, $\mathbf{v}_1$, associated with λ_1.

(2) For the second principal component, we want

$$\begin{aligned} & \max Var(\mathbf{b}_2^\top \mathbf{X}) \\ s.t. \qquad & \mathbf{b}_2^\top \mathbf{b}_2 = 1 \\ & \mathbf{v}_1^\top \Sigma_{\mathbf{XX}} \mathbf{b}_2 = 0 \end{aligned}$$

We form the function

$$f(\mathbf{b}_2) = \mathbf{b}_2^\top \Sigma_{\mathbf{XX}} \mathbf{b}_2 - \lambda_2 (1 - \mathbf{b}_2^\top \mathbf{b}_2) + \mu_2 \mathbf{v}_1^\top \Sigma_{\mathbf{XX}} \mathbf{b}_2$$

where λ_2 and μ_2 are Lagrangian multipliers.

Let

$$\frac{\partial f(\mathbf{b}_2)}{\partial \mathbf{b}_2} = 2(\Sigma_{\mathbf{XX}} - \lambda_2 \mathbf{I}_r)\mathbf{b}_2 + \mu_2 \Sigma_{\mathbf{XX}} \mathbf{v}_1 = 0$$

Premultiplying by $\mathbf{v}_1^\top \Sigma_{\mathbf{XX}}$ gives

$$\begin{aligned} 2\mathbf{v}_1^\top \Sigma_{\mathbf{XX}} (\Sigma_{\mathbf{XX}} - \lambda_2 \mathbf{I}_r)\mathbf{b}_2 + \mu_2 \mathbf{v}_1^\top \Sigma_{\mathbf{XX}} \Sigma_{\mathbf{XX}} \mathbf{v}_1 &= 0 \\ 2\mathbf{v}_1^\top \Sigma_{\mathbf{XX}}^2 \mathbf{b}_2 + \mu_2 \mathbf{v}_1^\top \Sigma_{\mathbf{XX}}^2 \mathbf{v}_1 &= 0 \end{aligned}$$

Since

$$\begin{aligned}(\Sigma_{\mathbf{XX}} - \lambda_1 \mathbf{I})\mathbf{v}_1 &= 0\\ \mathbf{b}_2^\top \Sigma_{\mathbf{XX}}(\Sigma_{\mathbf{XX}} - \lambda_1 \mathbf{I})\mathbf{v}_1 &= 0\\ \mathbf{b}_2^\top \Sigma_{\mathbf{XX}}^2 \mathbf{v}_1 &= 0\end{aligned}$$

we have

$$\begin{aligned}2\mathbf{v}_1^\top \Sigma_{\mathbf{XX}}^2 \mathbf{b}_2 &= 0\\ \mu_2 \mathbf{v}_1^\top \Sigma_{\mathbf{XX}}^2 \mathbf{v}_1 &= 0\\ \mu_2 &= 0\end{aligned}$$

Therefore,

$$\frac{\partial f(\mathbf{b}_2)}{\partial \mathbf{b}_2} = 2(\Sigma_{\mathbf{XX}} - \lambda_2 \mathbf{I}_r)\mathbf{b}_2 = 0$$

Similarly, λ_2 has to be the largest eigenvalue of $\Sigma_{\mathbf{XX}}$, and $\mathbf{b}_2$ the eigenvector, $\mathbf{v}_2$, associated with λ_2.

(3) Similarly, we can obtain the remaining sets of coefficients for the principal components.

11 Probabilistic Principal Component Analysis

Suppose that $\mathbf{X} \sim N_s(0, \mathbf{I}_s)$, $\mathbf{Y}|\mathbf{X} \sim N_r(\mathbf{CX} + \boldsymbol{\mu}, \sigma^2 \mathbf{I})$. We want to estimate $\mathbf{C}$ and σ^2 by samples $(\mathbf{X}_1, \mathbf{Y}_1)$, $\cdots$, $(\mathbf{X}_n, \mathbf{Y}_n)$. When $\sigma^2 \to 0$, PPCA problem becomes PCA.

Proof.

$$\begin{aligned}f_{\mathbf{XY}}(\mathbf{x},\mathbf{y}) &= f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) f_{\mathbf{X}}(\mathbf{x})\\ &= \frac{1}{(2\pi)^{\frac{r}{2}}|\sigma^2\mathbf{I}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{y}-\mathbf{Cx}-\boldsymbol{\mu})^\top(\sigma^{-2}\mathbf{I})(\mathbf{y}-\mathbf{Cx}-\boldsymbol{\mu})} \cdot \frac{1}{(2\pi)^{\frac{s}{2}}} e^{-\frac{1}{2}\mathbf{x}^\top\mathbf{x}}\\ &= \frac{1}{(2\pi\sigma^2)^{\frac{r}{2}}(2\pi)^{\frac{s}{2}}} e^{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{Cx}-\boldsymbol{\mu})^\top(\mathbf{y}-\mathbf{Cx}-\boldsymbol{\mu})-\frac{1}{2}\mathbf{x}^\top\mathbf{x}}\\ &= \frac{1}{(2\pi\sigma^2)^{\frac{r}{2}}(2\pi)^{\frac{s}{2}}} e^{-\frac{1}{2}\left[\mathbf{x}^\top(\sigma^{-2}\mathbf{C}^\top\mathbf{C}+\mathbf{I}_s)\mathbf{x}-2\sigma^{-2}(\mathbf{y}-\boldsymbol{\mu})^\top\mathbf{Cx}+\sigma^{-2}(\mathbf{y}-\boldsymbol{\mu})^\top(\mathbf{y}-\boldsymbol{\mu})\right]}\\ &= \frac{1}{(2\pi\sigma^2)^{\frac{r}{2}}(2\pi)^{\frac{s}{2}}} e^{-\frac{1}{2}\left\|(\sigma^{-2}\mathbf{C}^\top\mathbf{C}+\mathbf{I}_s)^{\frac{1}{2}}\mathbf{x}-(\sigma^{-2}\mathbf{C}^\top\mathbf{C}+\mathbf{I}_s)^{-\frac{1}{2}}\mathbf{C}^\top(\mathbf{y}-\boldsymbol{\mu})\sigma^{-2}\right\|_2^2}\\ &\quad \cdot e^{\frac{1}{2\sigma^4}(\mathbf{y}-\boldsymbol{\mu})^\top\mathbf{C}(\sigma^{-2}\mathbf{C}^\top\mathbf{C}+\mathbf{I}_s)^{-1}\mathbf{C}^\top(\mathbf{y}-\boldsymbol{\mu})-\frac{1}{2\sigma^2}(\mathbf{y}-\boldsymbol{\mu})^\top(\mathbf{y}-\boldsymbol{\mu})}\\ &= \frac{1}{(2\pi\sigma^2)^{\frac{r}{2}}(2\pi)^{\frac{s}{2}}} e^{-\frac{1}{2}\left\|(\sigma^{-2}\mathbf{C}^\top\mathbf{C}+\mathbf{I}_s)^{\frac{1}{2}}\mathbf{x}-(\sigma^{-2}\mathbf{C}^\top\mathbf{C}+\mathbf{I}_s)^{-\frac{1}{2}}\mathbf{C}^\top(\mathbf{y}-\boldsymbol{\mu})\sigma^{-2}\right\|_2^2}\\ &\quad \cdot e^{\frac{1}{2\sigma^2}(\mathbf{y}-\boldsymbol{\mu})^\top[\sigma^{-1}\mathbf{C}(\sigma^{-2}\mathbf{C}^\top\mathbf{C}+\mathbf{I}_s)^{-1}\sigma^{-1}\mathbf{C}^\top-\mathbf{I}_s](\mathbf{y}-\boldsymbol{\mu})}\\ &= \frac{1}{(2\pi\sigma^2)^{\frac{r}{2}}(2\pi)^{\frac{s}{2}}} e^{-\frac{1}{2}\left\|(\sigma^{-2}\mathbf{C}^\top\mathbf{C}+\mathbf{I}_s)^{\frac{1}{2}}\mathbf{x}-(\sigma^{-2}\mathbf{C}^\top\mathbf{C}+\mathbf{I}_s)^{-\frac{1}{2}}\mathbf{C}^\top(\mathbf{y}-\boldsymbol{\mu})\sigma^{-2}\right\|_2^2}\end{aligned}$$

$$\cdot e^{-\frac{1}{2\sigma^2}(\mathbf{y}-\boldsymbol{\mu})^\top(\mathbf{I}_r+\sigma^{-2}\mathbf{C}\mathbf{C}^\top)^{-1}(\mathbf{y}-\boldsymbol{\mu})} \qquad \text{(Woodbury Matrix Inequality)}$$

$$= \frac{1}{(2\pi\sigma^2)^{\frac{r}{2}}(2\pi)^{\frac{s}{2}}} e^{-\frac{1}{2}\left\|(\sigma^{-2}\mathbf{C}^\top\mathbf{C}+\mathbf{I}_s)^{\frac{1}{2}}\mathbf{x}-(\sigma^{-2}\mathbf{C}^\top\mathbf{C}+\mathbf{I}_s)^{-\frac{1}{2}}\mathbf{C}^\top(\mathbf{y}-\boldsymbol{\mu})\sigma^{-2}\right\|_2^2}$$

$$\cdot e^{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^\top(\sigma^2\mathbf{I}_r+\mathbf{C}\mathbf{C}^\top)^{-1}(\mathbf{y}-\boldsymbol{\mu})}$$

$$f_{\mathbf{Y}}(\mathbf{y}) = \iint_{\mathbb{R}^s} f_{\mathbf{X}\mathbf{Y}}(\mathbf{x},\mathbf{y})\mathbf{x}$$

$$= \frac{1}{(2\pi)^{\frac{r}{2}}|\sigma^2\mathbf{I}_r+\mathbf{C}\mathbf{C}^\top|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^\top(\sigma^2\mathbf{I}_s+\mathbf{C}\mathbf{C}^\top)^{-1}(\mathbf{y}-\boldsymbol{\mu})}$$

i.e.

$$\mathbf{y} \sim N(\boldsymbol{\mu}, \sigma^2\mathbf{I}_r + \mathbf{C}\mathbf{C}^\top)$$

□

12 Canonical Variate and Correlation Analysis

12.0.1 Definition

Canonical variate and correlation analysis (CVA or CCA) is a method for studying linear relationships between two vector variates, which we denote by $\mathbf{X} = \begin{pmatrix} X_1 & \cdots & X_r \end{pmatrix}^\top$ and $\mathbf{Y} = \begin{pmatrix} Y_1 & \cdots & Y_s \end{pmatrix}^\top$

We assume that

$$\mathbb{E}\left[\begin{pmatrix}\mathbf{X}\\ \mathbf{Y}\end{pmatrix}\right] = \begin{pmatrix}\boldsymbol{\mu}_{\mathbf{X}}\\ \boldsymbol{\mu}_{\mathbf{Y}}\end{pmatrix}$$

$$Var\left[\begin{pmatrix}\mathbf{X}\\ \mathbf{Y}\end{pmatrix}\right] = \begin{pmatrix}\Sigma_{\mathbf{X}\mathbf{X}} & \Sigma_{\mathbf{X}\mathbf{Y}}\\ \Sigma_{\mathbf{Y}\mathbf{X}} & \Sigma_{\mathbf{Y}\mathbf{Y}}\end{pmatrix}$$

CVA seeks to replace the two sets of correlated variables, $\mathbf{X}$ and $\mathbf{Y}$, by t pairs of new variables,

$$(\xi_i, \omega_i) \qquad i = 1, 2, \cdots, t,\ t < \min\{r, s\}$$

where

$$\xi_i = \mathbf{g}_i^\top \mathbf{X}$$

$$\omega_i = \mathbf{h}_i^\top \mathbf{Y}$$

$$\mathbf{g}_i = \begin{pmatrix} g_{1i} & \cdots & g_{ri} \end{pmatrix}^\top$$

$$\mathbf{h}_i = \begin{pmatrix} h_{1i} & \cdots & h_{ri} \end{pmatrix}^\top$$

The jth pair of coefficient vectors, $\mathbf{g}_j$ and $\mathbf{h}_j$ are chosen so that

1. the pairs $\{(\xi_j, \omega_j),\ j = 1,2,\cdots,t\}$ are ranked in importance through their correlations,

$$\rho_j = Corr(\xi_j, \omega_j) = \frac{\mathbf{g}_j^\top \Sigma_{\mathbf{XX}} \mathbf{h}_j}{\sqrt{(\mathbf{g}_j^\top \Sigma_{\mathbf{YY}} \mathbf{g}_j)(\mathbf{h}_j^\top \Sigma_{\mathbf{XY}} \mathbf{h}_j)}}$$

   which are listed in descending order of magnitude: $\rho_1 \geq \rho_2 \geq \cdots \geq \rho_t$.

2. ξ_j is uncorrelated with all previously derived ξ_i $(i < j)$

$$Cov(\xi_i, \xi_j) = \mathbf{g}_i^\top \Sigma_{\mathbf{XX}} \mathbf{g}_j = 0$$

3. ω_j is uncorrelated with all previously derived ω_i $(i < j)$

$$Cov(\omega_i, \omega_j) = \mathbf{h}_i^\top \Sigma_{\mathbf{YY}} \mathbf{h}_j = 0$$

The pairs $\{(\xi_j, \omega_j),\ j = 1,2,\cdots,t\}$ are known as the first t pairs of canonical variates of $\mathbf{X}$ and $\mathbf{Y}$ and their correlations $\{\rho_j,\ j = 1,2,\cdots,t\}$ as the t largest canonical correlations.

12.1 Interpretation by LS

Let the $(t \times r)$-matrix $\mathbf{G}$ and the $(t \times s)$-matrix $\mathbf{H}$, with $1 \leq t \leq \min\{r,s\}$, be such that $\mathbf{X}$ and $\mathbf{Y}$ are linearly projected into new vector variates,

$$\boldsymbol{\xi} = \mathbf{GX}$$
$$\boldsymbol{\omega} = \mathbf{HY}$$

Consider the problem of finding $\boldsymbol{\nu}$, $\mathbf{G}$ and $\mathbf{H}$ so that

$$\mathbf{HY} = \boldsymbol{\nu} + \mathbf{GX} + \boldsymbol{\varepsilon}$$

To

$$\min_{\boldsymbol{\nu},\mathbf{G},\mathbf{H}} \mathbb{E}[(\mathbf{HY} - \boldsymbol{\nu} - \mathbf{GX})^\top (\mathbf{HY} - \boldsymbol{\nu} - \mathbf{GX})]$$

where we assume that the covariance matrix of $\boldsymbol{\omega}$ is

$$\Sigma_{\boldsymbol{\omega}\boldsymbol{\omega}} = \mathbf{H}\Sigma_{\mathbf{YY}}\mathbf{H}^\top = \mathbf{I}$$

12.2 Interpretation by RRR

Let

$$\boldsymbol{\Gamma} = \Sigma_{\mathbf{YY}}^{-1}$$

we have

$$\boldsymbol{\mu}^{(t)}, \mathbf{A}^{(t)}, \mathbf{B}^{(t)} = \min_{\boldsymbol{\mu},\mathbf{A},\mathbf{B}} \mathbb{E}[(\mathbf{Y} - \boldsymbol{\mu} - \mathbf{ABX})^\top \boldsymbol{\Gamma} (\mathbf{Y} - \boldsymbol{\mu} - \mathbf{ABX})]$$

Then

$$\mathbf{H}^{(t)}\mathbf{A}^{(t)}\mathbf{H}^{(t)} = \mathbf{H}^{(t)}$$
$$\mathbf{A}^{(t)}\mathbf{H}^{(t)}\mathbf{A}^{(t)} = \mathbf{A}^{(t)}$$

Thus

$$\mathbf{H}^{(t)} = \mathbf{A}^{(t)-}$$
$$\mathbf{G}^{(t)} = \mathbf{B}^{(t)}$$
$$\mathbf{C}^{(t)} = \Sigma_{\mathbf{YY}}^{\frac{1}{2}}\mathbf{V}_t\mathbf{V}_t^\top\Sigma_{\mathbf{YY}}^{-\frac{1}{2}}\Sigma_{\mathbf{YX}}\Sigma_{\mathbf{XX}}^{-1}$$

where $\mathbf{V}_t$ is the first t eigenvectors of $\Gamma^{\frac{1}{2}}\Sigma_{\mathbf{YX}}\Sigma_{\mathbf{XX}}^{-1}\Sigma_{\mathbf{XY}}\Gamma^{\frac{1}{2}}$.

Thus, in a least-squares sense,

$$\mathbf{A}^{(t)-}\mathbf{Y} = \mathbf{A}^{(t)-}\boldsymbol{\mu}^{(t)} + \mathbf{B}^{(t)}\mathbf{X} + \mathbf{A}^{(t)-}\boldsymbol{\varepsilon}$$

12.3 Interpretation by Optimization

To simplify the problem, we assume

$$\mathbb{E}\mathbf{X} = 0$$
$$\mathbb{E}\mathbf{Y} = 0$$
$$Var(\xi_i) = \mathbf{g}_i^\top\Sigma_{\mathbf{XX}}\mathbf{g}_i = 1$$
$$Var(\omega_i) = \mathbf{h}_i^\top\Sigma_{\mathbf{YY}}\mathbf{h}_i = 1$$

(1) For the first pair,

$$\begin{aligned} & \max Corr(\xi,\omega) = Cov(\xi,\omega) = \mathbf{g}^\top\Sigma_{\mathbf{XY}}\mathbf{h} \\ s.t. \quad & \mathbf{g}^\top\Sigma_{\mathbf{XX}}\mathbf{g} = 1 \\ & \mathbf{h}^\top\Sigma_{\mathbf{YY}}\mathbf{h} = 1 \end{aligned}$$

Proof. Let

$$f(\mathbf{g},\mathbf{h}) = \mathbf{g}^\top\Sigma_{\mathbf{XY}}\mathbf{h} - \frac{1}{2}\lambda(\mathbf{g}^\top\Sigma_{\mathbf{XX}}\mathbf{g}-1) - \frac{1}{2}\mu(\mathbf{h}^\top\Sigma_{\mathbf{YY}}\mathbf{h}-1)$$

where λ and μ are Lagrangian multipliers.

Let

$$\frac{\partial f}{\partial \mathbf{g}} = \Sigma_{\mathbf{XY}}\mathbf{h} - \lambda\Sigma_{\mathbf{XX}}\mathbf{g} = 0$$
$$\frac{\partial f}{\partial \mathbf{h}} = \Sigma_{\mathbf{YX}}\mathbf{g} - \mu\Sigma_{\mathbf{YY}}\mathbf{h} = 0$$

Multiplying on the left by $\mathbf{g}^\top$ and $\mathbf{h}^\top$, we obtain

$$\mathbf{g}^\top\Sigma_{\mathbf{XY}}\mathbf{h} - \lambda\mathbf{g}^\top\Sigma_{\mathbf{XX}}\mathbf{g} = 0$$

$$\mathbf{h}^\top \Sigma_{\mathbf{YX}} \mathbf{g} - \mu \mathbf{h}^\top \Sigma_{\mathbf{YY}} \mathbf{h} = 0$$

Since $\mathbf{g}^\top \Sigma_{\mathbf{XX}} \mathbf{g} = \mathbf{h}^\top \Sigma_{\mathbf{YY}} \mathbf{h} = 1$, we have

$$\mathbf{g}^\top \Sigma_{\mathbf{XY}} \mathbf{h} = \lambda = \mu$$

Therefore, we only need to maximize λ.

$$\frac{\partial f}{\partial \mathbf{g}} = \Sigma_{\mathbf{XY}} \mathbf{h} - \lambda \Sigma_{\mathbf{XX}} \mathbf{g} = 0$$
$$\frac{\partial f}{\partial \mathbf{h}} = \Sigma_{\mathbf{YX}} \mathbf{g} - \lambda \Sigma_{\mathbf{YY}} \mathbf{h} = 0$$

Multiplying on the two equations by $\Sigma_{\mathbf{YX}} \Sigma_{\mathbf{XX}}^{-1}$ and λ respectively, then adding them up gives

$$(\Sigma_{\mathbf{YX}} \Sigma_{\mathbf{XX}}^{-1} \Sigma_{\mathbf{XY}} - \lambda^2 \Sigma_{\mathbf{YY}}) \mathbf{h} = 0$$
$$(\Sigma_{\mathbf{YY}}^{-\frac{1}{2}} \Sigma_{\mathbf{YX}} \Sigma_{\mathbf{XX}}^{-1} \Sigma_{\mathbf{XY}} \Sigma_{\mathbf{YY}}^{-\frac{1}{2}} - \lambda^2 \mathbf{I}_s) \Sigma_{\mathbf{YY}}^{\frac{1}{2}} \mathbf{h} = 0$$

When $\mathbf{h} \neq \mathbf{0}$, we have

$$|\Sigma_{\mathbf{YY}}^{-\frac{1}{2}} \Sigma_{\mathbf{YX}} \Sigma_{\mathbf{XX}}^{-1} \Sigma_{\mathbf{XY}} \Sigma_{\mathbf{YY}}^{-\frac{1}{2}} - \lambda^2 \mathbf{I}_s| = 0$$

Suppose that $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_s$ are the uint eigenvectors of $\Sigma_{\mathbf{YY}}^{-\frac{1}{2}} \Sigma_{\mathbf{YX}} \Sigma_{\mathbf{XX}}^{-1} \Sigma_{\mathbf{XY}} \Sigma_{\mathbf{YY}}^{-\frac{1}{2}}$ corresponding to eigenvalues $\lambda_1^2 \geq \cdots \lambda_s^2 \geq 0$. Then

$$\lambda = \lambda_1$$
$$\mathbf{h}_1 = \Sigma_{\mathbf{YY}}^{-\frac{1}{2}} \mathbf{v}_1$$
$$\mathbf{g}_1 = \frac{1}{\lambda_1} \Sigma_{\mathbf{XX}}^{-1} \Sigma_{\mathbf{XY}} \Sigma_{\mathbf{YY}}^{-\frac{1}{2}} \mathbf{v}_1$$
$$\mathbf{g}_1^\top \Sigma_{\mathbf{XY}} \mathbf{h}_1 = \lambda_1$$

Therefore,

$$\xi_1 = \mathbf{g}_1^\top \mathbf{X}$$
$$\omega_1 = \mathbf{h}_1^\top \mathbf{Y}$$

□

(2) Suppose that we have calculated the first $j-1$ $(j \geq 2)$ canonical variates of $\mathbf{X}$ and $\mathbf{Y}$, then to find the jth pair is to

$$\max Corr(\xi, \omega) = Cov(\xi, \omega) = \mathbf{g}^\top \Sigma_{\mathbf{XY}} \mathbf{h}$$
$$s.t. \quad \mathbf{g}^\top \Sigma_{\mathbf{XX}} \mathbf{g} = 1$$
$$\mathbf{h}^\top \Sigma_{\mathbf{YY}} \mathbf{h} = 1$$
$$\mathbf{g}_i^\top \Sigma_{\mathbf{XX}} \mathbf{g} = 0$$
$$\mathbf{h}_i^\top \Sigma_{\mathbf{YY}} \mathbf{h} = 0$$

$$\mathbf{g}_i^\top \Sigma_{\mathbf{XY}} \mathbf{h} = 0$$
$$\mathbf{h}_i^\top \Sigma_{\mathbf{YX}} \mathbf{g} = 0 \qquad i < j$$

For $j = 2$, the proof is given as belows:

Proof. Let

$$f(\mathbf{g}, \mathbf{h}) = \mathbf{g}^\top \Sigma_{\mathbf{XY}} \mathbf{h} - \frac{1}{2}\lambda(\mathbf{g}^\top \Sigma_{\mathbf{XX}} \mathbf{g} - 1) - \frac{1}{2}\mu(\mathbf{h}^\top \Sigma_{\mathbf{YY}} \mathbf{h} - 1) + \eta \mathbf{g}_1^\top \Sigma_{\mathbf{XX}} \mathbf{g} + \nu \mathbf{h}_1^\top \Sigma_{\mathbf{YY}} \mathbf{h}$$

where λ, μ, η and ν are Lagrangian multipliers.

Let

$$\frac{\partial f}{\partial \mathbf{g}} = \Sigma_{\mathbf{XY}} \mathbf{h} - \lambda \Sigma_{\mathbf{XX}} \mathbf{g} + \eta \Sigma_{\mathbf{XX}} \mathbf{g}_1 = 0$$
$$\frac{\partial f}{\partial \mathbf{h}} = \Sigma_{\mathbf{YX}} \mathbf{g} - \mu \Sigma_{\mathbf{YY}} \mathbf{h} + \nu \Sigma_{\mathbf{YY}} \mathbf{h}_1 = 0$$

Multiplying on the left by $\mathbf{g}_1^\top$ and $\mathbf{h}_1^\top$, we obtain

$$\mathbf{g}_1^\top \Sigma_{\mathbf{XY}} \mathbf{h} + \eta = \eta = 0$$
$$\mathbf{h}_1^\top \Sigma_{\mathbf{YX}} \mathbf{g} + \nu = \nu = 0$$

Multiplying on the left by $\mathbf{g}^\top$ and $\mathbf{h}^\top$, we obtain

$$\mathbf{g}^\top \Sigma_{\mathbf{XY}} \mathbf{h} - \lambda \mathbf{g}^\top \Sigma_{\mathbf{XX}} \mathbf{g} = 0$$
$$\mathbf{h}^\top \Sigma_{\mathbf{YX}} \mathbf{g} - \mu \mathbf{h}^\top \Sigma_{\mathbf{YY}} \mathbf{h} = 0$$

since $\mathbf{g}_1^\top \Sigma_{\mathbf{XX}} \mathbf{g} = \mathbf{h}_1^\top \Sigma_{\mathbf{YY}} \mathbf{h} = 0$.

Since $\mathbf{g}^\top \Sigma_{\mathbf{XX}} \mathbf{g} = \mathbf{h}^\top \Sigma_{\mathbf{YY}} \mathbf{h} = 1$, we have

$$\mathbf{g}^\top \Sigma_{\mathbf{XY}} \mathbf{h} = \lambda = \mu$$

Therefore, we only need to maximize λ.

$$\frac{\partial f}{\partial \mathbf{g}} = \Sigma_{\mathbf{XY}} \mathbf{h} - \lambda \Sigma_{\mathbf{XX}} \mathbf{g} = 0$$
$$\frac{\partial f}{\partial \mathbf{h}} = \Sigma_{\mathbf{YX}} \mathbf{g} - \lambda \Sigma_{\mathbf{YY}} \mathbf{h} = 0$$

Multiplying on the two equations by $\Sigma_{\mathbf{YX}} \Sigma_{\mathbf{XX}}^{-1}$ and λ respectively, then adding them up gives

$$(\Sigma_{\mathbf{YX}} \Sigma_{\mathbf{XX}}^{-1} \Sigma_{\mathbf{XY}} - \lambda^2 \Sigma_{\mathbf{YY}}) \mathbf{h} = 0$$
$$(\Sigma_{\mathbf{YY}}^{-\frac{1}{2}} \Sigma_{\mathbf{YX}} \Sigma_{\mathbf{XX}}^{-1} \Sigma_{\mathbf{XY}} \Sigma_{\mathbf{YY}}^{-\frac{1}{2}} - \lambda^2 \mathbf{I}_s) \Sigma_{\mathbf{YY}}^{\frac{1}{2}} \mathbf{h} = 0$$

When $\mathbf{h} \neq \mathbf{0}$, we have

$$|\Sigma_{\mathbf{YY}}^{-\frac{1}{2}} \Sigma_{\mathbf{YX}} \Sigma_{\mathbf{XX}}^{-1} \Sigma_{\mathbf{XY}} \Sigma_{\mathbf{YY}}^{-\frac{1}{2}} - \lambda^2 \mathbf{I}_s| = 0$$

Suppose that $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_s$ are the uint eigenvectors of $\Sigma_{\mathbf{YY}}^{-\frac{1}{2}} \Sigma_{\mathbf{YX}} \Sigma_{\mathbf{XX}}^{-1} \Sigma_{\mathbf{XY}} \Sigma_{\mathbf{YY}}^{-\frac{1}{2}}$ corresponding to eigenvalues $\lambda_1^2 \geq \cdots \lambda_s^2 \geq 0$. Then

$$\lambda = \lambda_2$$

$$\mathbf{h}_2 = \Sigma_{\mathbf{YY}}^{-\frac{1}{2}} \mathbf{v}_2$$
$$\mathbf{g}_2 = \frac{1}{\lambda_2} \Sigma_{\mathbf{XX}}^{-1} \Sigma_{\mathbf{XY}} \Sigma_{\mathbf{YY}}^{-\frac{1}{2}} \mathbf{v}_2$$
$$\mathbf{g}_2^\top \Sigma_{\mathbf{XY}} \mathbf{h}_2 = \lambda_2$$

Therefore,

$$\xi_2 = \mathbf{g}_2^\top \mathbf{X}$$
$$\omega_2 = \mathbf{h}_2^\top \mathbf{Y}$$

□

Chapter 5 Discriminant Analysis

13 Linear Discriminant Analysis

13.1 Bayes's Rule Classifier

Let

$$\mathbb{P}\{\mathbf{X} \in \Pi_i\} = \pi_i \qquad i = 1, 2$$

be the prior probabilities that a randomly selected observation $\mathbf{X} = \mathbf{x}$ belongs to either Π_1 or Π_2. Suppose also that the conditional multivariate probability density of $\mathbf{X}$ for the ith class is

$$\mathbb{P}\{\mathbf{X} = \mathbf{x} | \mathbf{X} \in \Pi_i\} = f_i(\mathbf{x}) \qquad i = 1, 2$$

Then Bayes's theorem yields the posterior probability,

$$p(\Pi_i|\mathbf{x}) = \mathbb{P}\{\mathbf{X} \in \Pi_i | \mathbf{X} = \mathbf{x}\} = \frac{f_i(\mathbf{x})\pi_i}{f_1(\mathbf{x})\pi_1 + f_2(\mathbf{x})\pi_2}$$

that the observed $\mathbf{x}$ belongs to Π_i, $i = 1, 2$.

Bayes's rule classifier assign $\mathbf{x}$ to Π_1 if

$$\frac{p(\Pi_1|\mathbf{x})}{p(\Pi_2|\mathbf{x})} > 1$$

i.e.

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} > \frac{\pi_2}{\pi_1}$$

On the boundary $\left\{\mathbf{x} \in \mathbb{R}^r \middle| \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} = \frac{\pi_2}{\pi_1}\right\}$, we randomize between assigning $\mathbf{x}$ to either Π_1 or Π_2.

13.2 Gaussian Linear Discriminant Analysis

Suppose $\mathbb{E}\mathbf{X}_1 = \mu_1$, $Var\mathbf{X}_1 = \Sigma$ and $\mathbb{E}\mathbf{X}_2 = \mu_2$, $Var\mathbf{X}_2 = \Sigma$ are independently distributed. Consider the statistic

$$\frac{\{\mathbb{E}(\mathbf{a}^\top \mathbf{X}_1) - \mathbb{E}(\mathbf{a}^\top \mathbf{X}_2)\}^2}{Var(\mathbf{a}^\top \mathbf{X}_1 - \mathbf{a}^\top \mathbf{X}_2)}$$

as a function of $\mathbf{a}$. Show that $\mathbf{a} \propto \Sigma^{-1}(\mu_1 - \mu_2)$ maximizes the statistic by using a Lagrange multiplier approach.

Proof. Since $\forall\ k \in \mathbb{R}$,

$$\frac{\{\mathbb{E}(k\mathbf{a}^\top \mathbf{X}_1) - \mathbb{E}(k\mathbf{a}^\top \mathbf{X}_2)\}^2}{Var(k\mathbf{a}^\top \mathbf{X}_1 - k\mathbf{a}^\top \mathbf{X}_2)} = \frac{\{\mathbb{E}(\mathbf{a}^\top \mathbf{X}_1) - \mathbb{E}(\mathbf{a}^\top \mathbf{X}_2)\}^2}{Var(\mathbf{a}^\top \mathbf{X}_1 - \mathbf{a}^\top \mathbf{X}_2)}$$

we can restrict $\mathbf{a}^\top \Sigma \mathbf{a} = 1$, then

$$L(\mathbf{a}, \lambda) = \mathbf{a}^\top (\mu_1 - \mu_2)(\mu_1 - \mu_2)^\top \mathbf{a} - \lambda(\mathbf{a}^\top \Sigma \mathbf{a} - 1)$$

$$\frac{\partial L}{\partial \mathbf{a}} = 2(\mu_1 - \mu_2)(\mu_1 - \mu_2)^\top \mathbf{a} - 2\lambda \Sigma \mathbf{a} = 0 \tag{1}$$
$$\frac{\partial L}{\partial \lambda} = -\mathbf{a}^\top \Sigma \mathbf{a} + 1 = 0 \tag{2}$$

The maximizor $\mathbf{a}^*$ should satisfy

$$(\mu_1 - \mu_2)(\mu_1 - \mu_2)^\top \mathbf{a}^* = \lambda \Sigma \mathbf{a}^*$$
$$\mathbf{a}^* = \frac{1}{\lambda} \Sigma^{-1} (\mu_1 - \mu_2)(\mu_1 - \mu_2)^\top \mathbf{a}^*$$

Since $\frac{1}{\lambda}(\mu_1 - \mu_2)^\top \mathbf{a}^* = c$ is a scalar, so we can write as

$$\mathbf{a}^* = c\Sigma^{-1}(\mu_1 - \mu_2)$$

And therefore,

$$\mathbf{a}^* = \frac{c}{\lambda} \Sigma^{-1}(\mu_1 - \mu_2)$$
$$\propto \Sigma^{-1}(\mu_1 - \mu_2)$$

□

14 Quadratic Discriminant Analysis

Chapter 6 Recursive Partitioning and Tree-Based Methods

15 Classification Trees

15.1 Tree-Growing Procedure

1. Choose the Boolean conditions for splitting at each node.
2. Use a criterion to split a parent node into its two daughter node.
3. Decide when a node become a terminal node, stop splitting.
4. Assign a class to a terminal node.

Splitting Strategies Let $\Pi_1, \cdots, \Pi_K$ be the $K \geq 2$ classes. For node τ, we define the **node impurity function** $i(\tau)$ as

$$i(\tau) = \phi(p(1|\tau), \cdot, p(K|\tau))$$

where $p(k|\tau)$ is an estimate of $\mathbb{P}\{\mathbf{X} \in \Pi_k|\tau\}$, the conditional probability that an observation $\mathbf{X}$ is in Π_k given that it falls into node τ.

1. Misclassification rate
$$i(\tau) = 1 - \max_k p(k|\tau)$$
2. Entropy function
$$i(\tau) = -\sum_{i=1}^{K} p(k|\tau) \ln p(k|\tau)$$
3. Gini diversity index
$$i(\tau) = \sum_{k \neq k'} p(k|\tau)p(k'|\tau) = 1 - \sum_{k=1}^{K} [p(k|\tau)]^2$$

The **goodness of split** s **at node** τ is given by the reduction in impurity gained by splitting the parent node τ into its daughter nodes, τ_R and τ_L,

$$\Delta i(s, \tau) = i(\tau) - p_L i(\tau_L) - p_R i(\tau_R)$$

The best split for the single variable X_j is the one that has the largest value of $\Delta i(s, \tau)$ over all $s \in S_j$, the set of possible distinct splits for X_j.

15.2 Estimating the Misclassification Rate

The **resubstitution estimate of the misclassification rate** $R(\tau)$ of an observation in node τ is

$$r(\tau) = 1 - \max_k p(k|\tau)$$

Let T be the tree classifier and let $\tilde{T} = \{\tau_1, \tau_2, \cdots, \tau_L\}$ denote the set of all terminal nodes of T. We can now estimate the true misclassification rate,

$$R(T) = \sum_{\tau \in \tilde{T}} R(\tau)\mathbb{P}(\tau) = \sum_{l=1}^{L} R(\tau_l)\mathbb{P}(\tau_l)$$

for T, where $\mathbb{P}(\tau)$ is the probability that an observation falls into node τ. If we estimate $\mathbb{P}(\tau_l)$ by the proportion $p(\tau_l)$ of all observations that fall into node τ_l , then, the resubstitution estimate of $R(T)$ is

$$R^{re}(T) = \sum_{l=1}^{L} r(\tau_l)p(\tau_l) = \sum_{l=1}^{L} R^{re}(\tau_l)$$

where

$$R^{re}(\tau) = r(\tau)p(\tau)$$

15.3 Pruning the Tree

The pruning algorithm is as follows:

1. Grow a large tree, say, $T_{\max}$, where we keep splitting until the nodes each contain fewer than $n_{\min}$ observations;
2. Compute an estimate of $R(\tau)$ at each node $\tau \in T_{\max}$;
3. Prune $T_{\max}$ upwards toward its root node so that at each stage of pruning, the estimate of $R_\alpha(T)$ is minimized where

$$\begin{aligned} R_\alpha(T) &= R^{re} + \alpha \\ R_\alpha(T) &= \sum_{l=1}^{L} R_\alpha(\tau_l) \\ &= R^{re}(T) + \alpha|\tilde{T}| \end{aligned}$$

For each α, we then choose that subtree $T(\alpha)$ of $T_{\max}$ that minimizes $R_\alpha(T)$:

$$R_\alpha(T(\alpha)) = \min_T R_\alpha(T)$$

then $T(\alpha)$ is called a minimizing subtree (or an optimally pruned subtree) of $T_{\max}$.

We call $T(\alpha)$ the smallest minimizing subtree if it is a minimizing subtree and satisfies the following condition: if $R_\alpha(T) = R_\alpha(T(\alpha))$, then $T(\alpha)$ is a subtree of T and hence has fewer terminal nodes than T.

As we increase α, the minimizing subtrees $T(\alpha)$ will have fewer and fewer terminal nodes. When α is very large, we will have pruned the entire tree $T_{\max}$, leaving only the root node. Note that although α is defined on the interval $[0,\infty)$, the number of subtrees of T is finite.

For each α, to prune $T_{\max}$,

1. From $T_{\max}$ to T_1. Suppose the node τ in the tree $T_{\max}$ has daughter nodes τ_L and τ_R, both of which are terminal nodes. Then, we have

$$R^{re}(\tau) \geq R^{re}(\tau_L) + R^{re}(\tau_R)$$

with equality if

$$\max_k p(k|\tau) = \max_k p(k|\tau_L) = \max_k p(k|\tau_R)$$

Proof.

$$\begin{aligned}
r(\tau) &= 1 - \max_k p(k|\tau) \\
r(\tau_L) &= 1 - \max_k p(k|\tau_L) \\
r(\tau_R) &= 1 - \max_k p(k|\tau_R) \\
R^{re}(\tau) &= r(\tau)p(\tau) \\
&= r(\tau)[p(\tau_L) + p(\tau_R)] \\
R^{re}(\tau_L) &= r(\tau_L)p(\tau_L) \\
R^{re}(\tau_R) &= r(\tau_R)p(\tau_R)
\end{aligned}$$

$\because$

$$\begin{aligned}
p(k|\tau) &= \frac{p(\tau_L)p(k|\tau_L) + p(\tau_R)p(k|\tau_R)}{p(\tau)} \\
p(\tau)\max_k p(k|\tau) &\leq p(\tau_L)\max_k p(k|\tau_L) + p(\tau_R)\max_k p(k|\tau_R) \qquad (1)
\end{aligned}$$

$\therefore$

$$\begin{aligned}
p(\tau) - p(\tau)\max_k p(k|\tau) &\geq [p(\tau_L) - p(\tau_L)\max_k p(k|\tau_L)] + [p(\tau_R) - p(\tau_R)\max_k p(k|\tau_R)] \\
r(\tau)p(\tau) &\geq r(\tau_L)p(\tau_L) + r(\tau_R)p(\tau_R) \\
R^{re}(\tau) &\geq R^{re}(\tau_L) + R^{re}(\tau_R)
\end{aligned}$$

The equality holds if $\max_k p(k|\tau) = \max_k p(k|\tau_L) = \max_k p(k|\tau_R)$ from (1). □

If equality occurs at node τ, then prune the terminal nodes τ_L and τ_R from the tree. Continue this pruning strategy until no further pruning of this type is possible. The resulting tree is T_1.

2. Find T_2. Let τ be any nonterminal node of T_1, let T_τ be the subtree whose root node is τ, and let $\tilde{T}_\tau = \{\tau'_1, \tau'_2, \cdots, \tau'_{L_\tau}\}$ be the set of terminal nodes of T_τ.

Let

$$R^{re}(T_\tau) = \sum_{\tau' \in \tilde{T}_\tau} R^{re}(\tau') = \sum_{l'=1}^{L_\tau} R^{re}(\tau'_{l'})$$

then

$$R^{re}(\tau) > R^{re}(T_\tau)$$

Proof.

If $T_\tau = \tau$, then it holds.

Else, let A_1 be the set of nodes whose left and right nodes are in $\tilde{T}_\tau$ and B_1 be the set of these left and right nodes of nodes in A, then $A_1, B_1 \neq \varnothing$ and

$$\sum_{\tau' \in A_1} R^{re}(\tau') > \sum_{\tau' \in B_1} R^{re}(\tau')$$

since we have pruned before.

Therefore,

$$\sum_{\tau' \in A_1 \cup (\tilde{T}_\tau \setminus B_1)} R^{re}(\tau') > \sum_{\tau' \in \tilde{T}_\tau} R^{re}(\tau')$$

Treat $C_1 = A_1 \cup (\tilde{T}_\tau \setminus B_1)$ as the set of terminal nodes and define A_2 and B_2 similarly, then we have

$$\sum_{\tau' \in A_2 \cup (C_1 \setminus B_2)} R^{re}(\tau') \geq \sum_{\tau' \in C_1} R^{re}(\tau') > \sum_{\tau' \in B_1} R^{re}(\tau')$$

Repeat this process, we have

$$R^{re}(\tau) > R^{re}(T_\tau)$$

□

Set

$$R_\alpha(T_\tau) = R^{re}(T_\tau) + \alpha |\tilde{T}_\tau|$$

then as long as $R_\alpha(\tau) > R_\alpha(T_\tau)$, i.e., when

$$\alpha < \frac{R^{re}(\tau) - R^{re}(T_\tau)}{|\tilde{T}_\tau| - 1} = g(\tau)$$

we don't prune the subtree T_τ,

3. Increase α and use cross validation or test to determine the optimal value of α.

Chapter 7 Committee Machines

16 Bagging

16.1 Definition

Denote the learning set of n observations by

$$\mathscr{L} = \{(\mathbf{X}_i, Y_i), i = 1, \cdots, n\}$$

where $\{Y_i\}$ are continuous responses (a regression problem) or unordered class labels (a classification problem).

The bootstrap aggregating (Bagging) procedure starts by drawing B bootstrap samples from $\mathscr{L}$.

Algorithm 6 Bootstrap Sampler

Require: $\mathscr{L} = \{(\mathbf{X}_i, Y_i), i = 1, \cdots, n\}$

Ensure: $\mathscr{L}^{*b}$, $b = 1, 2, \cdots, B$

1: for $i = 1 : B$ do

2: Draw $\mathscr{L}^{*b} = \{(\mathbf{X}_i^{*b}, Y_i^{*b}), i = 1, 2, \cdots, n\}$ from $\mathscr{L}$ with probability $p_i = \frac{1}{n}$ on the ith observation $(\mathbf{X}_i, Y_i)$ and with replacement.

3: end for

Out-of-bag (OOB) observations for the ith boopstrap sampler is the observations that in $\mathscr{L} \setminus \mathscr{L}^{*b}$. For the ith observation $(\mathbf{X}_i, Y_i)$ in $\mathscr{L}$, we denote

$$\mathscr{N}_i = \{b | (\mathbf{X}_i, Y_i) \notin \mathscr{L}^{*b}, b = 1, 2, \cdots, B\}$$

The out-of-bag observations can be use to estimate generalization error.

16.2 Bagging Tree-Based Classifiers

In the classification case, $Y_i \in \{1, 2, \cdots, K\}$ is a class label attached to X_i.

Algorithm 7 Bootstrap Tree-Based Classifiers

Require: $\mathscr{L} = \{(\mathbf{X}_i, Y_i), i = 1, \cdots, n\}$, B, $\mathbf{X}_{new}$

Ensure: $\hat{Y}_{new}$

1: Draw boopstrap samples $\{\mathscr{L}^{*b}, b = 1, 2, \cdots, B\}$.

2: Grow classification trees $\{T^{*b}, b = 1, 2, \cdots, B\}$ with respect to bootstrap samples $\{L^b, b = 1, 2, \cdots, B\}$ without pruning to reduce bias.

3: Drop X_{new} down each of the B boopstrap trees and get B votes $\{\hat{Y}_{new}^b, b = 1, 2, \cdots, B\}$.

4: By the majority-vote rule, choose the mode of $\{\hat{Y}_{new}^b, b = 1, 2, \cdots, B\}$ as $\hat{Y}_{new}$, the predicted label of $\mathbf{X}_{new}$.

The predicted label for the ith observation based on OOB approach is given by

$$mode(\{\hat{Y}_i^b,\ b \in \mathcal{N}_i\})$$

The OOB misclassification rate is given by

$$PE_{bag} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{mode(\{\hat{Y}_i^b,\ b\in\mathcal{N}_i\})\neq Y_i\}}$$

16.3 Bagging Tree-Based Regressers

Algorithm 8 Bootstrap Tree-Based Regressers

Require: $\mathscr{L} = \{(\mathbf{X}_i, Y_i), i = 1, \cdots, n\}$, B, $\mathbf{X}_{new}$

Ensure: $\hat{Y}_{new}$

1: Draw boopstrap samples $\{\mathscr{L}^{*b},\ b = 1, 2, \cdots, B\}$.

2: Grow regression trees $\{T^{*b},\ b = 1, 2, \cdots, B\}$ with respect to bootstrap samples $\{L^b,\ b = 1, 2, \cdots, B\}$ without pruning to reduce bias.

3: Drop X_{new} down each of the B boopstrap trees and get B votes $\{\hat{Y}_{new}^b,\ b = 1, 2, \cdots, B\}$.

4: By the majority-vote rule, choose the mean of $\{\hat{Y}_{new}^b,\ b = 1, 2, \cdots, B\}$ as $\hat{Y}_{new}$, the predicted label of $\mathbf{X}_{new}$.

The predicted label for the ith observation based on OOB approach is given by

$$\frac{1}{|\mathcal{N}_i|}\sum_{b\in\mathcal{N}_i} \hat{Y}_i^b$$

The OOB misclassification rate is given by

$$PE_{bag} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{|\mathcal{N}_i|}\sum_{b\in\mathcal{N}_i} \hat{Y}_i^b - Y_i\right)^2$$

16.4 Random Forests

In bagging, randomization is used only in selecting the data set on which to grow each tree. An extension of this idea is random forests by introducing random split selection and random input selection.

There are only two tuning parameters for a random forest: the number m of variables randomly chosen as a subset at each node and the number B of bootstrap samples. A good starting point is to take m as $\sqrt{r}$ for classification problem and $\frac{1}{3}r$ for regression problem where r is the dimension of features.

Algorithm 9 Random Forests

Require: $\mathscr{L} = \{(\mathbf{X}_i, Y_i), i = 1, \cdots, n\}$, m, B, $\mathbf{X}_{new}$

Ensure: $\hat{Y}_{new}$

1: Draw boopstrap samples $\{\mathscr{L}^{*b},\ b = 1, 2, \cdots, B\}$.

2: Grow classification (or regression) trees $\{T^{*b},\ b = 1, 2, \cdots, B\}$ with respect to bootstrap samples $\{L^b,\ b = 1, 2, \cdots, B\}$ without pruning to reduce bias. In each node, randomly select a subset m of the r input variables, and, using only the m selected variables, determine the best split at that node.

3: Drop X_{new} down each of the B boopstrap trees and get B votes $\{\hat{Y}^b_{new},\ b = 1, 2, \cdots, B\}$.

4: By the majority-vote rule, choose the mode (or the mean) of $\{\hat{Y}^b_{new},\ b = 1, 2, \cdots, B\}$ as $\hat{Y}_{new}$, the predicted label of $\mathbf{X}_{new}$.

Variable importance

17 Boosting

17.1 Definition

For binary classification problem ($Y_i \in \{-1, 1\}$), boosting is to create a strong classifier by substantially improve weak (or base) classifiers. Boosting algorithms combine M base classifiers $\{C_1, C_2, \cdots, C_M\}$ in the following way. For an observation X, the boosted classifier is given by

$$C_\alpha(\mathbf{X}) = sign\{f_\alpha(\mathbf{X})\}$$

where

$$f_\alpha(\mathbf{X}) = \sum_{j=1}^{M} \left(\frac{\alpha_j}{\sum_{i=1}^{M} \alpha_i} \right) C_j(\mathbf{X})$$

and

$$\alpha = \begin{pmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_M \end{pmatrix}^\top$$

17.2 AdaBoost

17.2.1 Boosting by Reweighting

Adaptive boosting (AdaBoost) is a method of boosting by Reweighting.

Algorithm 10 AdaBoost Binary Classifier

Require: $\mathscr{L} = \{(\mathbf{X}_i, Y_i),\ Y_i \in \{-1,1\},\ i = 1,\cdots,n\}$, $\mathscr{C} = \{C_1,\cdots,C_M\}$, m, T, $\mathbf{X}_{new}$

Ensure: $\hat{Y}_{new}$

1: Initialize the weight vector for the samples $\mathbf{w}_1 = \begin{pmatrix} w_{11} & \cdots & w_{n1} \end{pmatrix}^\top$ where $w_{i1} = \frac{1}{n}$, $i = 1,2,\cdots,n$.

2: for $t = 1 : T$ do

3: Select a weak classifier $C_{j_t}(x) \in \{-1,+1\}$ from $\mathscr{C}$, $j_t \in \{1,2,\cdots,M\}$, and train it on the learning set $\mathscr{L}$, where the ith observation $(\mathbf{X}_i, Y_i)$ has (normalized) weight w_{it}, $i = 1,2,\cdots,n$.

4: Compute the weighted prediction error:

$$PE_t = PE(\mathbf{w}_t) = \mathbb{E}_{\mathbf{w}}[\mathbb{1}_{\{Y_i \neq C_{j_t}(\mathbf{X}_i)\}}] = \left(\frac{\mathbf{w}_t^\top}{\mathbf{1}_n^\top \mathbf{w}_t}\right)\mathbf{e}_t$$

where $\mathbb{E}_{\mathbf{w}}$ indicates taking expectation with respect to the probability distribution of $\mathbf{w}_t = (w_{1t},\cdots,w_{nt})^\top$, and $\mathbf{e}_t$ is an n-vector with ith entry equals $\mathbb{1}_{\{Y_i \neq C_{j_t}(\mathbf{X}_i)\}}$.

5: Set $\beta_t = \frac{1}{2}\log\left(\frac{1-PE_t}{PE_t}\right)$.

6: Update weights:

$$w_{i,t+1} = \frac{w_{it}}{W_t} e^{2\beta_t \mathbb{1}_{\{Y_i \neq C_{j_t}(\mathbf{X}_i)\}}}, \qquad i = 1,2,\cdots,n$$

where W_t is a normalizing constant needed to ensure that the vector $\mathbf{w}_{t+1} = (w_{1,t+1},\cdots,w_{n,t+1})^\top$ represents a true weight distribution over $\mathscr{L}$; that is, $\mathbf{1}_n^\top \mathbf{w}_{t+1} = 1$.

7: end for

8: Let $\hat{Y}_{new} = sign\{f(x)\}$, where $f(x) = \sum_{t=1}^{T} \beta_t C_{j_t}(x) = \sum_{j=1}^{M} \alpha_j C_j(x)$, and $\alpha_j = \sum_{t=1}^{T} \beta_t \mathbb{1}_{\{j_t = j\}}$.

17.2.2 A Statistical Interpretation of AdaBoost

Derivation

Suppose we have computed classifiers $C_1, \cdots, C_{m-1}$ along with their corresponding weights $\beta_1, \cdots, \beta_{m-1}$ and we want to compute the next classifier C_m along with its weight β_m. The output of our model so far is $f_{m-1}(\mathbf{X}) = \sum_{i=1}^{m-1} \beta_i C_i(\mathbf{X})$, and we want to minimize the risk:

$$\arg\min_{\beta, C} \sum_{i=1}^{n} L(Y_i, f_{m-1}(\mathbf{X}_i) + \beta C(\mathbf{X}_i))$$

We can write the AdaBoost optimization problem with exponential loss $L(Y, h(\mathbf{X})) = e^{-Yh(\mathbf{X})}$ as follows:

$$\begin{aligned}\beta_m, C_m &= \arg\min_{\beta, C} \sum_{i=1}^{n} e^{-Y_i[f_{m-1}(\mathbf{X}_i) + \beta C(\mathbf{X}_i)]} \\ &= \arg\min_{\beta, C} \sum_{i=1}^{n} e^{-Y_i f_{m-1}(\mathbf{X}_i)} e^{-Y_i \beta C(\mathbf{X}_i)}\end{aligned}$$

The term $w_i^{(m)} = e^{-Y_i f_{m-1}(\mathbf{X}_i)}$ is a constant with respect to our optimization variables. We can split out this sum into the components with correctly classified points and incorrectly classified points:

$$\begin{aligned}\beta_m, C_m &= \arg\min_{\beta, C} \left[\sum_{Y_i = C(\mathbf{X}_i)} w_i^{(m)} e^{-\beta} + \sum_{Y_i \neq C(\mathbf{X}_i)} w_i^{(m)} e^{\beta} \right] \\ &= \arg\min_{\beta, C} \left[(e^{\beta} - e^{-\beta}) \sum_{Y_i \neq C(\mathbf{X}_i)} w_i^{(m)} + e^{-\beta} \sum_{i=1}^{n} w_i^{(m)} \right]\end{aligned} \tag{6.2.1}$$

For a fixed value of β, the second term does not depend on C. Thus we can see that the best choice of $C_m(\mathbf{X})$ is the classifier that minimizes the error given the weights $w_i^{(m)}$. Let

$$e_m = \frac{\sum_{Y_i \neq C_m(\mathbf{X}_i)} w_i^{(m)}}{\sum_{i=1}^{n} w_i^{(m)}}$$

Once we have obtained C_m, we can solve for β_m: by dividing 6.2.1 by the constant $\sum_{i=1}^{n} w_i^{(m)}$, we obtain

$$\begin{aligned}\beta_m &= \arg\min_{\beta} [(1 - e_m) e^{-\beta} + e_m e^{\beta}] \\ &= \frac{1}{2} \ln \left(\frac{1 - e_m}{e_m} \right)\end{aligned}$$

by setting

$$\frac{\partial}{\partial \beta} [(1 - e_m) e^{-\beta} + e_m e^{\beta}] = -(1 - e_m) e^{-\beta} + e_m e^{\beta} = 0$$

From the optimal β_m, we can derive the weights:

$$w_i^{(m+1)} = e^{-Y_i F_m(\mathbf{X}_i)}$$

$$
\begin{aligned}
&= e^{-Y_i[F_{m-1}(\mathbf{X}_i)+\beta_m C_m(\mathbf{X}_i)]} \\
&= w_i^{(m)} e^{-Y_i \beta_m C_m(\mathbf{X}_i)} \\
&= w_i^{(m)} e^{\ln\left[\left(\frac{1-e_m}{e_m}\right)^{-\frac{1}{2}Y_i C_m(\mathbf{X}_i)}\right]} \\
&= w_i^{(m)} \left(\frac{1-e_m}{e_m}\right)^{-\frac{1}{2}Y_i C_m(\mathbf{X}_i)}
\end{aligned}
$$

Since

$$
-Y_i C_m(\mathbf{X}_i) = 2\mathbb{1}_{\{Y_i \neq C_m(\mathbf{X}_i)\}} - 1
$$

we have

$$
\begin{aligned}
w_i^{(m+1)} &= w_i^{(m)} e^{-Y_i \beta_m C_m(\mathbf{X}_i)} \\
&= w_i^{(m)} e^{2\beta_m \mathbb{1}_{\{Y_i \neq C_m(\mathbf{X}_i)\}}} e^{-\beta_m}
\end{aligned}
$$

Normalize and get

$$
w_i^{(m+1)} = \frac{w_i^{(m)}}{W_m} e^{2\beta_m \mathbb{1}_{\{Y_i \neq C_m(\mathbf{X}_i)\}}}
$$

17.3 Gradient Boosting

In gradient boosting, we approximate the negative gradient $-g_t(x)$ by a parametric function $h(\mathbf{X};\boldsymbol{\theta}_t)$, with parameter vector $\boldsymbol{\theta}_t$.

Algorithm 11 Gradient Boosting

Require: $\mathscr{L} = \{(\mathbf{X}_i, Y_i), i = 1, \cdots, n\}$, T

Ensure: $\hat{f}(\mathbf{X})$

1: Initialize $f_0(x) = \arg\min_{\rho} \sum_{i=1}^{n} L(Y_i, \rho)$.

2: for $t = 1 : T$ do

3: Compute $\tilde{Y}_i = -g_t(\mathbf{X}_i) = -\dfrac{\partial L(Y_i, f(\mathbf{X}_i))}{\partial f(\mathbf{X}_i)}\Big|_{f(\mathbf{X}_i)=f_{t-1}(\mathbf{X}_i)}$, $i = 1, 2, \cdots, n$.

4: Compute $(\boldsymbol{\theta}_t, \beta_t) = \arg\min_{\boldsymbol{\theta},\beta} \sum_{i=1}^{n} [\tilde{Y}_i - \beta h(\mathbf{X}_i;\boldsymbol{\theta})]^2$.

5: Compute $\rho_t = \arg\min_{\rho} \sum_{i=1}^{n} L(Y_i, f_{t-1}(\mathbf{X}_i) + \rho h(\mathbf{X}_i;\boldsymbol{\theta}))$.

6: Set $f_t(\mathbf{X}) = f_{t-1}(\mathbf{X}) + \rho_t h(\mathbf{X};\boldsymbol{\theta}_t)$.

7: end for

8: Set $\hat{f}(\mathbf{X}) = f_T(\mathbf{X}) = \sum_{i=1}^{T} \rho_t h(\mathbf{X};\boldsymbol{\theta}_t)$.

Chapter 8 Artifical Neural Network

18 Single-Layer Perceptrons

$\mathbf{X} = \begin{pmatrix} X_1 & \cdots & X_r \end{pmatrix}^\top$ represents a random r-vector of input. The lth linear activation function is given by

$$U_l = \beta_{0l} + \mathbf{X}^\top \boldsymbol{\beta}_l$$

In matrix notation,

$$\mathbf{U} = \boldsymbol{\beta}_0 + \mathbf{B}\mathbf{X}$$

where $\mathbf{U} = \begin{pmatrix} U_1 & \cdots & U_s \end{pmatrix}^\top$, $\boldsymbol{\beta}_0 = \begin{pmatrix} \beta_{01} & \cdots & \beta_{0s} \end{pmatrix}^\top$ and $\mathbf{B} = \begin{pmatrix} \boldsymbol{\beta}_1 & \cdots & \boldsymbol{\beta}_s \end{pmatrix}^\top$.

The activation values are then each filtered through a nonlinear threshold activation function $f(U_l)$. In matrix form,

$$\mathbf{f}(\mathbf{U}) = \mathbf{f}(\boldsymbol{\beta}_0 + \mathbf{B}\mathbf{X})$$

19 Multilayer Perceptrons

Multilayer percetrons with only one hidden layer. Suppose we have a two-layer network with r input nodes $(X_m, m = 1,2,\cdots,r)$, a single layer $(L = 1)$ of t hidden nodes $(Z_j, j = 1,2,\cdots,t)$, and s output nodes $(Y_k, k = 1,2,\cdots,s)$. Let β_{mj} be the weight of the connection $X_m \to Z_j$ with bias β_{0j} and let α_{jk} be the weight of the connection $Z_j \to Y_k$ with bias α_{0k}.

Let $\mathbf{X} = \begin{pmatrix} X_1 & \cdots & X_r \end{pmatrix}^\top$, $\mathbf{Z} = \begin{pmatrix} Z_1 & \cdots & Z_r \end{pmatrix}^\top$, $\boldsymbol{\alpha}_k = \begin{pmatrix} \alpha_{1k} & \cdots & \alpha_{tk} \end{pmatrix}$, $\boldsymbol{\beta}_j = \begin{pmatrix} \beta_{1j} & \cdots & \beta_{rj} \end{pmatrix}$, $U_j = \beta_{0j} + \mathbf{X}^\top \boldsymbol{\beta}_j$ and $V_k = \alpha_{0k} + \mathbf{Z}^\top \boldsymbol{\alpha}_k$. Then,

$$\begin{aligned} Z_j &= f_j(U_j) \qquad j = 1,2,\cdots,t \\ \mu_k(\mathbf{X}) &= g_k(U_k) \\ &= g_k\left[\alpha_{0k} + \sum_{j=1}^{t} \alpha_{jk} f_j\left(\beta_{0j} + \sum_{m=1}^{r} \beta_{mj} X_m\right)\right] \qquad k = 1,2,\cdots,s \\ Y_k &= \mu_k(\mathbf{X}) + \varepsilon_k \end{aligned}$$

When $f_j(x) = \sigma(x) = \dfrac{1}{1+e^{-x}}$, $j = 1,2,\cdots,t$ and $g_k(x) = x$, $k = 1,2,\cdots,s$, then the network is equivalent to a single-layer perceptron.

We can express the model in matrix notation as follows:

$$\boldsymbol{\mu}(\mathbf{X}) = \mathbf{g}[\boldsymbol{\alpha}_0 + \mathbf{A}\mathbf{f}(\boldsymbol{\beta}_0 + \mathbf{B}\mathbf{X})]$$

20 Related Statistical Methods

20.1 Projection-Pursuit Regression

The regression function is taken to be

$$\mu(\mathbf{X}) = \alpha_0 + \sum_{j=1}^{t} f(\beta_{0j} + \mathbf{X}^\top \boldsymbol{\beta}_j)$$

where α_0, $\{\beta_{0j}\}$, $\{\boldsymbol{\beta}_j = \begin{pmatrix} \beta_{1j} & \cdots & \beta_{rj} \end{pmatrix}^\top\}$, and the $\{f_j(\cdot)\}$ are the unknown parameters of the model.

20.2 Generalized Additive Models

Chapter 9 Support Vector Machines

21 Linear Support Vector Machines

21.1 The Linearly Separable Case

Assume we have available a learning set of data, $\mathscr{L} = \{(\mathbf{x}_i, y_i) : i = 1, 2, \cdots, n\}$ where $\mathbf{x}_i \in \mathbb{R}^r$ and $y_i \in \{1, -1\}$. The binary classification problem is to use $\mathscr{L}$ to construct a function $f : \mathbb{R}^r \to \mathbf{R}$ so that

$$\begin{aligned} C(\mathbf{x}) &= sign(f(\mathbf{x})) \\ &= sign(\beta_0 + \mathbf{x}^\top \boldsymbol{\beta}) \end{aligned}$$

is a classifier.

If $\mathscr{L}$ is linearly separable, then the optimization problem is given by

$$\begin{aligned} &\min_{\beta_0, \boldsymbol{\beta}} \ \frac{1}{2}\|\boldsymbol{\beta}\|^2 \\ &s.t. \quad y_i(\beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta}) \geq 1, \qquad i = 1, 2, \cdots, n \end{aligned}$$

Proof. By using Lagrangian multipliers, the primal function is given by

$$F_P(\beta_0, \boldsymbol{\beta}, \boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{\beta}\|^2 + \sum_{i=1}^{n} \alpha_i [1 - y_i(\beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta})]$$

where $\boldsymbol{\alpha} = \begin{pmatrix} \alpha_1 & \cdots & \alpha_n \end{pmatrix}^\top \succeq \mathbf{0}$ is the Lagrangian coefficients. So the primal problem is equivalent to

$$\begin{aligned} &\min_{\beta_0, \boldsymbol{\beta}} \max_{\boldsymbol{\alpha}} \ F_P(\beta_0, \boldsymbol{\beta}, \boldsymbol{\alpha}) \\ &s.t. \quad \boldsymbol{\alpha} \succeq \mathbf{0} \end{aligned}$$

The Karush–Kuhn–Tucker conditions give necessary and sufficient conditions for a solution to a constrained optimization problem:

$$\begin{aligned} \frac{\partial F_P(\beta_0, \boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \beta_0} &= -\sum_{i=1}^{n} \alpha_i y_i = 0 \\ \frac{\partial F_P(\beta_0, \boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \boldsymbol{\beta}} &= \boldsymbol{\beta} - \sum_{i=1}^{n} \alpha_i \mathbf{x}_i = 0 \\ 1 - y_i(\beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta}) &\leq 0 \\ \alpha_i &\geq 0 \\ \alpha_i [1 - y_i(\beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta})] &= 0 \end{aligned}$$

KKT conditions yields

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

$$\boldsymbol{\beta} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$$

and β_0 is implicitly determined by the KKT complementarity condition, by choosing any i for which $\alpha_i \neq 0$ and computing β_0 (note that it is numerically safer to take the mean value of β_0 resulting from all such equations).

Applying KKT conditions to the primal function simplifies the dual function

$$\begin{aligned} F_D(\boldsymbol{\alpha}) &= \min_{\beta_0, \boldsymbol{\beta}} F_P(\beta_0, \boldsymbol{\beta}, \boldsymbol{\alpha}) \\ &= \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j - \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j + \sum_{i=1}^{n} \alpha_i \\ &= -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j + \sum_{i=1}^{n} \alpha_i \\ &= -\frac{1}{2} \boldsymbol{\alpha}^\top \mathbf{H} \boldsymbol{\alpha} + \mathbf{1}_n^\top \boldsymbol{\alpha} \end{aligned}$$

where $\mathbf{H} = \Big(\langle y_i \mathbf{x}_i, y_j \mathbf{x}_j \rangle \Big)$.

When KKT conditions are satisfied, the primal problem is equivalent to the dual problem

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \min_{\beta_0, \boldsymbol{\beta}} \; & F_P(\beta_0, \boldsymbol{\beta}, \boldsymbol{\alpha}) \\ s.t. \quad & \boldsymbol{\alpha} \succeq \mathbf{0} \end{aligned}$$

i.e.

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \; & F_D(\boldsymbol{\alpha}) \\ s.t. \quad & \boldsymbol{\alpha} \succeq \mathbf{0} \\ & \boldsymbol{\alpha}^\top \mathbf{y} = 0 \end{aligned}$$

If $\boldsymbol{\alpha}^*$ solves this optimization problem, then

$$\boldsymbol{\beta}^* = \sum_{i=1}^{n} \alpha_i^* y_i \mathbf{x}_i$$

By using KKT complementarity condition,

$$\boldsymbol{\beta}^* = \sum_{i \in SV} \alpha_i^* y_i \mathbf{x}_i$$

where $SV \subset \{1, 2, \cdots, n\}$ is the set of supporting vectors.

Then

$$\beta_0^* = \frac{1}{|SV|} \sum_{i \in SV} \frac{1 - y_i \mathbf{x}_i^\top \boldsymbol{\beta}^*}{y_i}$$

Since

$$\begin{aligned} \|\boldsymbol{\beta}\|^2 &= \sum_{i=1}^{n} \alpha_i^* y_i \mathbf{x}_i^\top \boldsymbol{\beta} \\ &= \sum_{i=1}^{n} \alpha_i^* [1 - y_i \beta_0] \end{aligned}$$

$$= \sum_{i=1}^{n} \alpha_i^*$$

the maximum margin is given by $\frac{2}{\|\boldsymbol{\beta}\|} = \frac{2}{\sqrt{\sum_{i=1}^{n} \alpha_i^*}}$. □

21.2 The Linearly Nonseparable Case

When $\mathscr{L}$ is linearly nonseparable, which means it is either nonlinearly separable or non separable, we search for the soft-margin solution. First we introduce the concept of slack variables $\xi_i \geq 0$, $i = 1, 2, \cdots, n$. The constants becomes $y_i(\beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta}) + \xi_i \geq 1$, $i = 1, 2, \cdots, n$.

Usually, we add constants of some simply functions of slack variables, such as $g_\sigma(\boldsymbol{\xi}) = \sum_{i=1}^{n} \xi_i^\sigma$. The 1-norm soft-margin and 2-norm soft-margin optimization problem correspond to $\sigma = 1$ and $\sigma = 2$ respectively.

21.2.1 1-Norm Soft-Margin SVM Classification

The 1-norm soft-margin optimization problem is given by

$$\begin{aligned} \min_{\beta_0, \boldsymbol{\beta}, \boldsymbol{\xi}} \ & \frac{1}{2}\|\boldsymbol{\beta}\|_2^2 + C\sum_{i=1}^{n} \xi_i \\ s.t. \quad & \boldsymbol{\xi} \succeq 0 \\ & y_i(\beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta}) + \xi_i \geq 1, \ i = 1, 2, \cdots, n \end{aligned}$$

where $C > 0$ is a regularization parameter.

Proof. By using Lagrangian multipliers, the primal function is given by

$$F_P(\beta_0, \boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\eta}) = \frac{1}{2}\|\boldsymbol{\beta}\|^2 + C\sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} \alpha_i[1 - \xi_i - y_i(\beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta})] + \sum_{i=1}^{n} \eta_i \xi_i$$

where $\boldsymbol{\alpha} = \begin{pmatrix} \alpha_1 & \cdots & \alpha_n \end{pmatrix}^\top \succeq \mathbf{0}$ and $\boldsymbol{\eta} = \begin{pmatrix} \eta_1 & \cdots & \eta_n \end{pmatrix}^\top$ are the Lagrangian coefficients. So the primal problem is equivalent to:

$$\begin{aligned} \min_{\beta_0, \boldsymbol{\beta}, \boldsymbol{\xi}} \max_{\boldsymbol{\alpha}, \boldsymbol{\eta}} \ & F_P \\ s.t. \quad & \boldsymbol{\alpha} \succeq 0 \\ & \boldsymbol{\eta} \succeq 0 \end{aligned}$$

When the KKT conditions are satisfied, the primal problem is equivalent to the dual problem:

$$\begin{aligned} \max_{\boldsymbol{\alpha}, \boldsymbol{\eta}} \ & F_D(\boldsymbol{\alpha}, \boldsymbol{\eta}) \\ s.t. \quad & \boldsymbol{\alpha} \succeq 0 \end{aligned}$$

$$\boldsymbol{\eta} \succeq 0$$

where the dual function is given by

$$F_D(\boldsymbol{\alpha}, \boldsymbol{\eta}) = \min_{\beta_0, \boldsymbol{\beta}, \boldsymbol{\xi}} F_P$$

By using KKT conditions, differentiate F_P with respect to β_0, $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$ and set to 0:

$$\begin{aligned}
\frac{\partial F_P}{\partial \beta_0} &= -\sum_{i=1}^{n} \alpha_i y_i = 0 \\
\frac{\partial F_P}{\partial \boldsymbol{\beta}} &= \boldsymbol{\beta} - \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i = \mathbf{0} \\
\frac{\partial F_P}{\partial \boldsymbol{\xi}} &= C\mathbf{1}_n - \boldsymbol{\alpha} + \boldsymbol{\eta} = \mathbf{0}
\end{aligned}$$

Therefore,

$$\begin{aligned}
F_D(\boldsymbol{\alpha}, \boldsymbol{\eta}) &= \frac{1}{2}\boldsymbol{\alpha}^\top \mathbf{H}\boldsymbol{\alpha} + (C\mathbf{1}_n - \boldsymbol{\alpha} + \boldsymbol{\eta})^\top \boldsymbol{\xi} + \mathbf{1}_n^\top \boldsymbol{\alpha} - \boldsymbol{\alpha}^\top \mathbf{H}\boldsymbol{\alpha} \\
&= -\frac{1}{2}\boldsymbol{\alpha}^\top \mathbf{H}\boldsymbol{\alpha} + \mathbf{1}_n^\top \boldsymbol{\alpha}
\end{aligned}$$

In addition, from the constraints $C - \alpha_i - \eta_i = 0$ and $\eta_i \geq 0$, the dual problem becomes

$$\begin{aligned}
\max_{\boldsymbol{\alpha}, \boldsymbol{\eta}} \quad & -\frac{1}{2}\boldsymbol{\alpha}^\top \mathbf{H}\boldsymbol{\alpha} + \mathbf{1}^\top \boldsymbol{\alpha} \\
s.t. \quad & \boldsymbol{\alpha}^\top \mathbf{y} = 0 \\
& \mathbf{0} \preceq \boldsymbol{\alpha} \preceq C\mathbf{1}_n
\end{aligned}$$

From the KKT complementarity condition, $\alpha_i[1 - \xi_i - y_i(\beta_0 + \mathbf{x}_j^\top \boldsymbol{\beta})] = 0$, if $\boldsymbol{\alpha}^*$ solves this optimization problem. then

$$\begin{aligned}
\boldsymbol{\beta}^* &= \sum_{i=1}^{n} \alpha_i^* y_i \mathbf{x}_i \\
&= \sum_{i \in SV} \alpha_i^* y_i \mathbf{x}_i \\
\beta_0^* &= \frac{1}{|\{i | 0 < a_i^* < C\}|} \sum_{i \in \{i | 0 < a_i^* < C\}} \frac{1 - y_i \mathbf{x}_i^\top \boldsymbol{\beta}^*}{y_i}
\end{aligned}$$

where $SV = \{i | a_i > 0\}$. □

21.2.2 2-Norm Soft-Margin SVM Classification

The 2-norm soft-margin optimization problem is given by

$$\begin{aligned}
\min_{\beta_0, \boldsymbol{\beta}, \boldsymbol{\xi}} \quad & \frac{1}{2}\|\boldsymbol{\beta}\|_2^2 + \frac{C}{2}\sum_{i=1}^{n} \xi_i^2 \\
s.t. \quad & \boldsymbol{\xi} \succeq 0
\end{aligned}$$

$$y_i(\beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta}) + \xi_i \geq 1,\ i = 1, 2, \cdots, n$$

where $C > 0$ is a regularization parameter.

Proof. Noet that if $\xi_i < 0$ violates the constraint, then there exists $\xi_i = 0$ satisfies the constraint with a even smaller value of the objective function. Therefore, $\xi_i < 0$ is never a solution and $\boldsymbol{\xi} \succeq \mathbf{0}$ is redundant.

By using Lagrangian multipliers, the primal function is given by

$$F_P(\beta_0, \boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{\beta}\|^2 + \frac{C}{2}\sum_{i=1}^{n}\xi_i^2 + \sum_{i=1}^{n}\alpha_i[1 - \xi_i - y_i(\beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta})]$$

where $\boldsymbol{\alpha} = \begin{pmatrix} \alpha_1 & \cdots & \alpha_n \end{pmatrix}^\top \succeq \mathbf{0}$ are the Lagrangian coefficients. So the primal problem is equivalent to:

$$\begin{aligned} &\min_{\beta_0, \boldsymbol{\beta}, \boldsymbol{\xi}} \max_{\boldsymbol{\alpha}} F_P \\ &s.t. \quad \boldsymbol{\alpha} \succeq 0 \end{aligned}$$

When the KKT conditions are satisfied, the primal problem is equivalent to the dual problem:

$$\begin{aligned} &\max_{\boldsymbol{\alpha}} F_D(\boldsymbol{\alpha}) \\ &s.t. \quad \boldsymbol{\alpha} \succeq 0 \end{aligned}$$

where the dual function is given by

$$F_D(\boldsymbol{\alpha}) = \min_{\beta_0, \boldsymbol{\beta}, \boldsymbol{\xi}} F_P$$

By using KKT conditions, differentiate F_P with respect to β_0, $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$ and set to 0:

$$\begin{aligned} \frac{\partial F_P}{\partial \beta_0} &= -\sum_{i=1}^{n}\alpha_i y_i = 0 \\ \frac{\partial F_P}{\partial \boldsymbol{\beta}} &= \boldsymbol{\beta} - \sum_{i=1}^{n}\alpha_i y_i \mathbf{x}_i = \mathbf{0} \\ \frac{\partial F_P}{\partial \boldsymbol{\xi}} &= C\boldsymbol{\xi} - \boldsymbol{\alpha} = \mathbf{0} \end{aligned}$$

Therefore,

$$\begin{aligned} F_D(\boldsymbol{\alpha}) &= \frac{1}{2}\boldsymbol{\alpha}^\top \mathbf{H}\boldsymbol{\alpha} + \left(\frac{C}{2}\boldsymbol{\xi} - \boldsymbol{\alpha}\right)^\top \boldsymbol{\xi} + \mathbf{1}_n^\top \boldsymbol{\alpha} - \boldsymbol{\alpha}^\top \mathbf{H}\boldsymbol{\alpha} \\ &= -\frac{1}{2}\boldsymbol{\alpha}^\top \mathbf{H}\boldsymbol{\alpha} - \frac{1}{2C}\boldsymbol{\alpha}^\top \boldsymbol{\alpha} + \mathbf{1}_n^\top \boldsymbol{\alpha} \\ &= -\frac{1}{2}\boldsymbol{\alpha}^\top \left(\mathbf{H} - \frac{1}{C}\mathbf{I}\right)\boldsymbol{\alpha} - \mathbf{1}_n^\top \boldsymbol{\alpha} \end{aligned}$$

The dual problem becomes

$$\max_{\boldsymbol{\alpha}, \boldsymbol{\eta}} \ -\frac{1}{2}\boldsymbol{\alpha}^\top \left(\mathbf{H} - \frac{1}{C}\mathbf{I}\right)\boldsymbol{\alpha} + \mathbf{1}^\top \boldsymbol{\alpha}$$

$$s.t. \quad \boldsymbol{\alpha}^\top \mathbf{y} = 0$$
$$\boldsymbol{\alpha} \succeq \mathbf{0}$$

If $\boldsymbol{\alpha}^*$ solves this optimization problem. then

$$\begin{aligned}
\boldsymbol{\beta}^* &= \sum_{i=1}^{n} \alpha_i^* y_i \mathbf{x}_i \\
&= \sum_{i \in SV} \alpha_i^* y_i \mathbf{x}_i \\
\beta_0^* &= \frac{1}{|SV|} \sum_{i \in SV} \frac{1 - \frac{a_i^*}{C} y_i \mathbf{x}_i^\top \boldsymbol{\beta}^*}{y_i}
\end{aligned}$$

Since

$$\begin{aligned}
\|\boldsymbol{\beta}^*\|^2 &= \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i^\top \boldsymbol{\beta}^* \\
&= \sum_{i=1}^{n} \alpha_i (1 - \xi_i - y_i \beta_0) \\
&= \sum_{i=1}^{n} \alpha_i - \frac{1}{C} \sum_{i=1}^{n} \alpha_i^2
\end{aligned}$$

the maximum margin is given by $\frac{2}{\|\boldsymbol{\beta}^*\|}$. □

22 Nonlinear Support Vector Machines

When $\mathscr{L}$ is linear nonseparable, we can first transform observations to spaces with higher dimensions.

Suppose we transform each observation, $\mathbf{x}_i \in \mathbb{R}^r$, in $\mathscr{L}$ using some nonlinear mapping $\Phi : \mathbb{R}^r \to \mathscr{H}$, where $\mathscr{H}$ is an $N_{\mathscr{H}}$-dimensional feature space. The nonlinear map Φ is generally called the feature map and the space $\mathscr{H}$ is called the feature space. The space $\mathscr{H}$ may be very high-dimensional, possibly even infinite dimensional. We will generally assume that $\mathscr{H}$ is a Hilbert space of real-valued functions on with inner product $\langle \cdot, \cdot \rangle$ and norm $\| \cdot \|$.

Let

$$\Phi(\mathbf{x}_i) = \begin{pmatrix} \phi_1(\mathbf{x}_i) & \cdots & \phi_{N_{\mathscr{H}}}(\mathbf{x}_i) \end{pmatrix}^\top, \quad i = 1, 2, \cdots, n$$

The transformed sample is then $\{\Phi(\mathbf{x}_i), y_i\}$. The difficulty is how to compute the inner products $\langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle = \Phi(\mathbf{x}_i)^\top \Phi(\mathbf{x}_j)$ in high-dimensional space $\mathscr{H}$.

22.1 Kernel

A kernel K is a function $K : \mathbb{R}^r \times \mathbb{R}^r \to \mathbb{R}$ such that $\forall\ \mathbf{x}, \mathbf{y} \in \mathbb{R}^r$,

$$K(\mathbf{x}, \mathbf{y}) = \langle \Phi(\mathbf{x}), \Phi(\mathbf{y}) \rangle$$

The kernel function is designed to compute inner-products in H by using only the original input data, which helps speed up the computations.

The properties of a kernel are

1. Symmetric
$$K(\mathbf{x}, \mathbf{y}) = K(\mathbf{y}, \mathbf{x})$$
2. Cauchy–Schwarz inequality
$$[K(\mathbf{x}, \mathbf{y})]^2 \leq K(\mathbf{x}, \mathbf{x}) K(\mathbf{y}, \mathbf{y})$$
3. If $\forall\ \mathbf{x} \in \mathbb{R}^r$, $K(\mathbf{x}, \mathbf{x}) = 1$, then $\|\Phi\|_{\mathscr{H}} = 1$.

If K is a specific Mercer kernel, i.e. nonnegtive kernel, on $\mathbb{R}^r \times \mathbb{R}^r$, then we can always construct a unique Hilbert space $\mathscr{H}_K$ of real-valued functions for which K is its reproducing kernel.

22.2 The Linearly Separable Case in the Feature Space

We replace $\mathbf{H}$ by $H_{ij} = y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$. If $\boldsymbol{\alpha}^*$ is the solution, then

$$\boldsymbol{\beta}^* = \sum_{i=1}^{n} \alpha_i^* y_i \Phi(\mathbf{x}_i)$$

The SVM decision rule is

$$sign\{\beta_0^* + \sum_{i \in SV} \alpha_i^* y_i K(\mathbf{x}, \mathbf{x}_i)\}$$

22.3 The Linearly Nonseparable Case in the Feature Space

22.3.1 1-Norm Soft-Margin SVM Classification

22.3.2 2-Norm Soft-Margin SVM Classification

23 Support Vector Regression

We define a loss function that ignores errors associated with points falling within a certain distance of the true linear regression function,

$$\mu(\mathbf{x}) = \beta_0 + \mathbf{x}^\top \boldsymbol{\beta}$$

In other words, if the point $(\mathbf{x}, y)$ is such that $|y - \mu(\mathbf{x})| \leq \varepsilon$, then the loss is taken to be zero; if, on the other hand, $|y - \mu(\mathbf{x})| > \varepsilon$, then we take the loss to be $|y - \mu(\mathbf{x})| - \varepsilon$. Thus, we can define the following two types of loss function, linear ε-insensitive loss function and quadratic ε-insensitive loss function:

$$L_1^{\varepsilon}(y, \mu(\mathbf{x})) = \max\{0, |y - \mu(\mathbf{x})| - \varepsilon\}$$
$$L_2^{\varepsilon}(y, \mu(\mathbf{x})) = \max\{0, [y - \mu(\mathbf{x})]^2 - \varepsilon\}$$

23.1 Linear ε-Insensitive SVR

We define slack variables ξ_i and ξ_i' in the following way. If the point $(\mathbf{x}_i, y_i)$ lies above the ε-tube, then $\xi_i' = y_i - \mu(\mathbf{x}_i) - \varepsilon \geq 0$, whereas if the point $(\mathbf{x}_i, y_i)$ lies below the ε-tube, then $\xi_i = \mu(\mathbf{x}_i) - y_i - \varepsilon \geq 0$. For points that fall outside the ε-tube, the values of the slack variables depend upon the shape of the loss function; for points inside the ε-tube, the slack variables have value zero.

The primal optimization problem is givem by

$$\begin{aligned} \min_{\beta_0, \boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\xi}'} \quad & \frac{1}{2}\|\boldsymbol{\beta}\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i') \\ s.t. \quad & y_i - (\beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta}) \leq \varepsilon + \xi_i' \\ & (\beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta}) - y_i \leq \varepsilon + \xi_i \\ & \boldsymbol{\xi} \succeq \mathbf{0} \\ & \boldsymbol{\xi}' \succeq \mathbf{0} \end{aligned}$$

Proof. Form the primal Lagrangian,

$$\begin{aligned} F_P = & \frac{1}{2}\|\boldsymbol{\beta}\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i') + \sum_{i=1}^{n} a_i[y_i - (\beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta}) - \varepsilon - \xi_i'] \\ & + \sum_{i=1}^{n} b_i[(\beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta}) - y_i - \varepsilon - \xi_i] - \sum_{i=1}^{n} c_i \xi_i' - \sum_{i=1}^{n} d_i \xi_i \end{aligned}$$

where $\mathbf{a} = \begin{pmatrix} a_1 & \cdots & a_n \end{pmatrix}^\top$, $\mathbf{b} = \begin{pmatrix} b_1 & \cdots & b_n \end{pmatrix}^\top$, $\mathbf{a} = \begin{pmatrix} c_1 & \cdots & c_n \end{pmatrix}^\top$, $\mathbf{a} = \begin{pmatrix} d_1 & \cdots & d_n \end{pmatrix}^\top \succeq \mathbf{0}$ are the Lagrange multipliers. So the primal problem is equivalent to

$$\begin{aligned} \min_{\beta_0, \boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\xi}'} \max_{\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}} \quad & F_P \\ s.t. \quad & \mathbf{a} \succeq \mathbf{0} \\ & \mathbf{b} \succeq \mathbf{0} \\ & \mathbf{c} \succeq \mathbf{0} \\ & \mathbf{d} \succeq \mathbf{0} \end{aligned}$$

When the KKT conditions are satisfied, the primal problem is equivalent to the dual problem:

$$\max_{\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}} F_D(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d})$$

$$
\begin{aligned}
s.t. \quad & \mathbf{a} \succeq \mathbf{0} \\
& \mathbf{b} \succeq \mathbf{0} \\
& \mathbf{c} \succeq \mathbf{0} \\
& \mathbf{d} \succeq \mathbf{0}
\end{aligned}
$$

where the dual function is given by

$$F_D(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}) = \min_{\beta_0, \boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\xi}'} F_P$$

By using KKT conditions, differentiate F_P with respect to β_0, $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$ and set to 0:

$$
\begin{aligned}
\frac{\partial F_P}{\partial \beta_0} &= -\sum_{i=1}^{n} a_i + \sum_{i=1}^{n} b_i = 0 \\
\frac{\partial F_P}{\partial \boldsymbol{\beta}} &= \boldsymbol{\beta} - \sum_{i=1}^{n} a_i \mathbf{x}_i + \sum_{i=1}^{n} b_i \mathbf{x}_i = \mathbf{0} \\
\frac{\partial F_P}{\partial \boldsymbol{\xi}'} &= C\mathbf{1}_n - \mathbf{a} - \mathbf{c} = \mathbf{0} \\
\frac{\partial F_P}{\partial \boldsymbol{\xi}} &= C\mathbf{1}_n - \mathbf{b} - \mathbf{d} = \mathbf{0}
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
F_D(\boldsymbol{\alpha}, \boldsymbol{\eta}) &= \frac{1}{2}(\mathbf{a}-\mathbf{b})^\top \mathbf{H}(\mathbf{a}-\mathbf{b}) + (C\mathbf{1}_n - \mathbf{a} - \mathbf{c})^\top \boldsymbol{\xi}' + (C\mathbf{1}_n + \mathbf{b} - \mathbf{d})^\top \boldsymbol{\xi} \\
&\quad + (\mathbf{a}-\mathbf{b})^\top \mathbf{y} - (\mathbf{a}-\mathbf{b})^\top \mathbf{H}(\mathbf{a}-\mathbf{b}) \\
&= -\frac{1}{2}(\mathbf{a}-\mathbf{b})^\top \mathbf{H}(\mathbf{a}-\mathbf{b}) + (\mathbf{a}-\mathbf{b})^\top \mathbf{y}
\end{aligned}
$$

where $\mathbf{H} = \Big(\langle y_i \mathbf{x}_i, y_j \mathbf{x}_j \rangle\Big)$.

The dual problem becomes

$$
\begin{aligned}
\max_{\boldsymbol{\alpha}, \boldsymbol{\eta}} \quad & -\frac{1}{2}(\mathbf{a}-\mathbf{b})^\top \mathbf{H}(\mathbf{a}-\mathbf{b}) + (\mathbf{a}-\mathbf{b})^\top \mathbf{y} \\
s.t. \quad & (\mathbf{a}-\mathbf{b})^\top \mathbf{1}_n = 0 \\
& \mathbf{0} \preceq \mathbf{a}, \mathbf{b} \preceq C\mathbf{1}_n
\end{aligned}
$$

If $\mathbf{a}^*$ and $\mathbf{b}^*$ solve this optimization problem. then

$$\boldsymbol{\beta}^* = \sum_{i=1}^{n} (a_i^* - b_i^*)\mathbf{x}_i$$

From KKT complementarity conditions:

$$
\begin{aligned}
a_i[(y_i - (\beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta}) - \varepsilon - \xi_i'] &= 0 \\
b_i[(\beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta}) - y_i - \varepsilon - \xi_i] &= 0 \\
c_i \xi_i' = (C - a_i)\xi_i' &= 0
\end{aligned}
$$

$$d_i\xi_i = (C - b_i)\xi_i = 0$$

we have when $a_i < C(b_i < C)$, $\xi_i' = 0(\xi_i = 0)$; when $a_i = C(b_i = C)$, $\xi_i' \geq 0(\xi_i \geq 0)$.

$$\beta_0^* = \frac{1}{|\{i|0 < a_i < C\}| + |\{i|0 < b_i < C\}|}\left[\sum_{\{i|0<a_i<C\}}(y_i - \mathbf{x}_i^\top\boldsymbol{\beta} - \varepsilon) + \sum_{\{i|0<b_i<C\}}(y_i - \mathbf{x}_i^\top\boldsymbol{\beta} + \varepsilon)\right]$$

□

23.2 Quadratic ε-Insensitive SVR

The primal optimization problem is givem by

$$\begin{aligned}
\min_{\beta_0,\boldsymbol{\beta},\boldsymbol{\xi},\boldsymbol{\xi}'} \quad & \frac{1}{2}\|\boldsymbol{\beta}\|^2 + C\sum_{i=1}^{n}(\xi_i^2 + \xi_i'^2) \\
s.t. \quad & y_i - (\beta_0 + \mathbf{x}_i^\top\boldsymbol{\beta}) \leq \varepsilon + \xi_i' \\
& (\beta_0 + \mathbf{x}_i^\top\boldsymbol{\beta}) - y_i \leq \varepsilon + \xi_i \\
& \boldsymbol{\xi} \succeq \mathbf{0} \\
& \boldsymbol{\xi}' \succeq \mathbf{0}
\end{aligned}$$

Proof. Noet that if $\xi_i < 0$ or $\xi_i' < 0$ violates the constraint, then there exists $\xi_i = 0$ or $\xi_i' = 0$ satisfies the constraint with a even smaller value of the objective function. Therefore, $\xi_i < 0$ or $\xi_i' < 0$ is never a solution and $\boldsymbol{\xi} \succeq \mathbf{0}$ and $\boldsymbol{\xi}' \succeq \mathbf{0}$ is redundant.

The primal function is given by

$$\begin{aligned}
F_P(\beta_0, \boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\xi}', \mathbf{a}, \mathbf{b}) = & \frac{1}{2}\|\beta\|^2 + \frac{C}{2}\sum_{i=1}^{n}(\xi_i^2 + \xi_i'^2) \\
& + \sum_{i=1}^{n} a_i[y_i - (\beta_0 + \mathbf{x}_i^\top\boldsymbol{\beta}) - \varepsilon - \xi_i'] \\
& + \sum_{i=1}^{n} b_i[(\beta_0 + \mathbf{x}_i^\top\boldsymbol{\beta}) - y_i - \varepsilon - \xi_i]
\end{aligned}$$

Setting the derivatives to 0,

$$\begin{aligned}
\frac{\partial F_P}{\partial \beta_0} &= \sum_{i=1}^{n}(a_i - b_i) = 0 \\
\frac{\partial F_P}{\partial \boldsymbol{\beta}} &= \boldsymbol{\beta} - \sum_{i=1}^{n} a_i\mathbf{x}_i + \sum_{i=1}^{n} b_i\mathbf{x}_i = 0 \\
\frac{\partial F_P}{\partial \boldsymbol{\xi}'} &= C\boldsymbol{\xi}' - \mathbf{a} = 0 \\
\frac{\partial F_P}{\partial \boldsymbol{\xi}} &= C\boldsymbol{\xi} - \mathbf{b} = 0
\end{aligned}$$

A stationary solution yields,

$$\begin{aligned}
\boldsymbol{\beta}^* &= \sum_{i=1}^{n}(a_i - b_i)\mathbf{x}_i \\
\sum_{i=1}^{n}(a_i - b_i) &= 0 \\
\boldsymbol{\xi}' &= \frac{1}{C}\mathbf{a} \\
\boldsymbol{\xi} &= \frac{1}{C}\mathbf{b}
\end{aligned}$$

Substituting the solution into the primal function gives us the dual function

$$\begin{aligned}
F_D(\mathbf{a},\mathbf{b}) &= \frac{1}{2}\|\boldsymbol{\beta}\|^2 + \frac{1}{2C}(\mathbf{a}^\top\mathbf{a} + \mathbf{b}^\top\mathbf{b}) \\
&\quad + (\mathbf{a}-\mathbf{b})\mathbf{y} - \sum_{i=1}^{n}(a_i - b_i)\mathbf{x}_i^\top\boldsymbol{\beta} - \sum_{i=1}^{n}(a_i + b_i)\varepsilon - \frac{1}{C}[\mathbf{a}^\top\mathbf{a} + \mathbf{b}^\top\mathbf{b}] \\
&= \frac{1}{2}\|\boldsymbol{\beta}\|^2 - \|\boldsymbol{\beta}\|^2 \\
&\quad - \frac{1}{2C}[\mathbf{a}^\top\mathbf{a} + \mathbf{b}^\top\mathbf{b}] + (\mathbf{a}-\mathbf{b})\mathbf{y} - \varepsilon(\mathbf{a}+\mathbf{b})^\top\mathbf{1} \\
&= -\frac{1}{2}\|\boldsymbol{\beta}\|^2 - \frac{1}{2C}[\mathbf{a}^\top\mathbf{a} + \mathbf{b}^\top\mathbf{b}] + (\mathbf{a}-\mathbf{b})\mathbf{y} - \varepsilon(\mathbf{a}+\mathbf{b})^\top\mathbf{1} \\
&= -\frac{1}{2}(\mathbf{a}-\mathbf{b})^\top\mathbf{K}(\mathbf{a}-\mathbf{b}) - \frac{1}{2C}[\mathbf{a}^\top\mathbf{a} + \mathbf{b}^\top\mathbf{b}] + (\mathbf{a}-\mathbf{b})\mathbf{y} - \varepsilon(\mathbf{a}+\mathbf{b})^\top\mathbf{1}
\end{aligned}$$

where

$$\mathbf{K} = \Big(\langle \mathbf{x}_i, \mathbf{x}_j\rangle\Big)_{1\le i,j\le n}$$

Therefore, the dual problem is given by

$$\begin{aligned}
\max \quad & P_D \\
s.t. \quad & \mathbf{a},\mathbf{b} \succeq \mathbf{0} \\
& (\mathbf{a}-\mathbf{b})^\top\mathbf{1} = 0
\end{aligned}$$

From the KKT conditions, for $i = 1, 2, \cdots, n$,

$$\begin{aligned}
a_i\left[y_i - (\beta_0 + \mathbf{x}_i^\top\boldsymbol{\beta}) - \varepsilon - \frac{a_i}{C}\right] &= 0 \\
b_i\left[(\beta_0 + \mathbf{x}_i^\top\boldsymbol{\beta}) - y_i - \varepsilon - \frac{b_i}{C}\right] &= 0 \\
a_i b_i &= 0
\end{aligned}$$

solve them for $\mathbf{a}$ and $\mathbf{b}$. If $\mathbf{a}^*$ and $\mathbf{b}^*$ are the solution, then

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \sum_{i=1}^{n}(\hat{a}_i^* - b_i^*)\mathbf{x}_i \\
\hat{\beta}_0 &= \frac{1}{|\{i|a_i^* > 0\}| + |\{i|b_i^* > 0\}|}\left[\sum_{\{i|a_i^*>0\}}\left(y_i - \mathbf{x}_i^\top\hat{\boldsymbol{\beta}} - \varepsilon - \frac{a_i^*}{C}\right) + \sum_{\{i|\hat{b}_i^*>0\}}\left(y_i - \mathbf{x}_i^\top\hat{\boldsymbol{\beta}} + \varepsilon + \frac{b_i^*}{C}\right)\right]
\end{aligned}$$

$$
\begin{aligned}
\|\boldsymbol{\beta}\|^2 &= \sum_{i=1}^{n}(\alpha_i^* - \beta_i^*)\mathbf{x}_i^\top \boldsymbol{\beta} \\
&= \sum_{\{i|\alpha_i^*>0\}} \alpha_i^* \mathbf{x}_i^\top \boldsymbol{\beta} - \sum_{\{i|\beta_i^*>0\}} \beta_i^* \mathbf{x}_i^\top \boldsymbol{\beta} \\
&= \sum_{\{i|\alpha_i^*>0\}} \alpha_i^* \left(y_i - \beta_0 - \varepsilon - \frac{\alpha_i^*}{C} \right) - \sum_{\{i|\beta_i^*>0\}} \beta_i^* \left(y_i - \beta_0 + \varepsilon + \frac{\beta_i^*}{C} \right) \\
&= \sum_{i=1}^{n} \alpha_i^* \left(y_i - \beta_0 - \varepsilon - \frac{\alpha_i^*}{C} \right) - \sum_{i=1}^{n} \beta_i^* \left(y_i - \beta_0 + \varepsilon + \frac{\beta_i^*}{C} \right) \\
&= \sum_{i=1}^{n}(\alpha_i^* - \beta_i^*) y_i - \sum_{i=1}^{n}(\alpha_i^* + \beta_i^*)\varepsilon - \frac{1}{C}\sum_{i=1}^{n}(\alpha_i^{*2} - \beta_i^{*2})
\end{aligned}
$$

□

Chapter 10 Nonparametric Density Estimation

24 Definition

Suppose we wish to estimate a bona fide density p, which is a continuous probability density function of a random r-vector variate X satisfying

$$p(\mathbf{x}) \geq 0, \quad \int_{\mathbb{R}^r} p(\mathbf{x}) \mathrm{d}\mathbf{x} = 1$$

24.1 Statistical Properties

1. Asymptotically Unbiasedness

   $\forall \mathbf{x} \in \mathbb{R}^r$, $\mathbb{E}_p[\hat{p}_n(\mathbf{x})] \to p(\mathbf{x})$ as the sample size $n \to \infty$.

2. Consistency

25 Histogram

26 Maximum Penalized Likelihood

Let Φ be a given nonnegative (roughness) penalty functional defined on H. The Φ-penalized likelihood of p is defined to be

$$\tilde{L}(p) = \left[\prod_{i=1}^{n} p(X_i) \right] e^{-\Phi(p)}$$

The optimization problem calls for $L(p)$, or its logarithm

$$L(p) = \log_e \tilde{L}(p) = \sum_{i=1}^{n} \log_e p(X_i) - \Phi(p)$$

to be maximized subject to

$$p \in H(\Omega), \quad \int_{\Omega} p(u) \mathrm{d}u, \quad p(u) \geq 0 \qquad \forall\, u \in \Omega$$

If it exists, a solution, $\hat{p}$, of that problem is called a maximum penalized likelihood (MPL) estimate of p corresponding to the penalty function Φ and class of functions H.

27 Kernel Density Estimation

Given i.i.d. univariate observations, $X_1, X_2, \cdots, X_n \sim p$, the kernel density estimator,

$$\hat{p}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h} \right), \quad x \in \mathbb{R}, h > 0$$

of $p(x)$, $x \in \mathbb{R}$ is used to obtain a smoother density estimate than the histogram. K is a kernel function, and the window width h determines the smoothness of the density estimate.

Given the r-vectors $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n$, the multivariate kernel density estimator of p is defined to have the general form,

$$\hat{p}_{\mathbf{H}}(\mathbf{x}) = \frac{1}{n|\mathbf{H}|} \sum_{i=1}^{n} K\left[\mathbf{H}^{-1}(x - \mathbf{X}_i)\right], \quad x \in \mathbb{R}^r$$

where $\mathbf{H}$ is an $(r \times r)$ nonsingular matrix that generalizes the window width h, and K is a multivariate function with mean 0 and integrates to 1.

If the variance of the projection, $Var(\mathbf{Y}) = \mathbf{w}^\top \mathbf{\Sigma}_{\mathbf{XX}} \mathbf{w}$ where $\|\mathbf{w}\| = 1$ is taken as the objective function, then maximizing that function with respect to $\mathbf{w}$ yields the first principal component of $\mathbf{W}$. PCA is, therefore, a special case of ICA.

28 Projection Pursuit Density Estimation

Algorithm 12 Projection Pursuit Density Estimation Algorithm

Require: $L = \{X_i, i = 1, 2, \cdots, n\}$

Ensure: $\hat{f}(\mathbf{X})$

1: Sphere the data to have mean 0 and covariance matrix I_r.

2: Initialize: Choose $\hat{p}^{(0)}$ to be an initial multivariate density estimate of p, usually taken to be the standard multivariate Gaussian.

3: for $j = 1, 2, \cdots$ do

4: Find the direction $\mathbf{a}_j \in \mathbb{R}^r$ for which the (model) marginal $p_{\mathbf{a}_j}$ along $\mathbf{a}_j$ differs most from the current estimated (data) marginal $\hat{p_{\mathbf{a}_j}}$ along $\mathbf{a}_j$. Choice of direction $\mathbf{a}_j$ will not generally be unique.

5: Given $\mathbf{a}_j$, define a univariate "augmenting function"

$$g_j(\mathbf{a}_j^\top \mathbf{x}) = \frac{p_{\mathbf{a}_j}(\mathbf{a}_j^\top \mathbf{x})}{\hat{p_{\mathbf{a}_j}}(\mathbf{a}_j^\top \mathbf{x})}$$

6: Update the previous estimate so that

$$\hat{p}^{(j)} = \hat{p}^{(j-1)}(\mathbf{x}) g_j(\mathbf{a}_j^\top \mathbf{x})$$

7: end for

Chapter 11 Cluster Analysis

Chapter 12 Multidimensional Scaling and Distance Geometry

32 Classical Scaling and Distance Geometry

32.1 Proximity Matrix

Consider a particular collection of n entities. Let δ_{ij} represent the dissimilarity of the ith entity to the jth entity. We arrange the m dissimilarities, $\{\delta_{ij}\}$, into an $(m \times m)$ square matrix, a proximity matrix is given by $\Delta = (\delta_{ij})$ with

$$\delta_{ij} \geq 0, \quad \delta_{ii} = 0, \quad \delta_{ij} = \delta_{ji}$$

In order for a dissimilarity measure to be regarded as a metric distance, we also require that δ_{ij} satisfy the triangle inequality,

$$\delta_{ij} \leq \delta_{ik} + \delta_{kj}, \qquad \forall k$$

When the matrix is asymmetrical, adjustments (e.g., setting $\delta_{ij} \leftarrow \frac{\delta_{ij}+\delta_{ji}}{2}$ to form a symmetrized version of Δ) can be made.

32.2 Definition

Let $\mathbf{X}_1, \cdots, \mathbf{X}_n \in \mathbb{R}_r$, $\delta_{ij} = \|\mathbf{X}_i - \mathbf{X}_j\|$, then

$$\delta_{ij}^2 = \|\mathbf{X}_i\|^2 + \|\mathbf{X}_j\|^2 - 2\langle \mathbf{X}_i, \mathbf{X}_j \rangle$$

Let

$$b_{ij} = \mathbf{X}_i^\top \mathbf{X}_j = -\frac{1}{2}\left(\delta_{ij}^2 - \delta_{i0}^2 - \delta_{j0}^2\right)$$

where $\delta_{i0}^2 = \|\mathbf{X}_i\|^2$.

Suppose that $\overline{\mathbf{X}} = \sum_{i=1}^{n} \mathbf{X}_i = 0$, summing over i and over j yields the following identities:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} \delta_{ij}^2 &= \frac{1}{n}\sum_{i=1}^{n} \delta_{i0}^2 + \delta_{j0}^2 \\
\frac{1}{n}\sum_{j=1}^{n} \delta_{ij}^2 &= \delta_{i0}^2 + \frac{1}{n}\sum_{j=1}^{n} \delta_{j0}^2 \\
\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \delta_{ij}^2 &= \frac{2}{n}\sum_{i=1}^{n} \delta_{i0}^2
\end{aligned}$$

Solve for δ_{i0} and δ_{j0} and then

$$\begin{aligned}
b_{ij} &= -\frac{1}{2}\delta_{ij}^2 - \frac{1}{2n}\sum_{j=1}^{n} \delta_{ij}^2 - \frac{1}{2n}\sum_{i=1}^{n} \delta_{ij}^2 + \frac{1}{4n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \delta_{ij}^2 \\
&= a_{ij} - a_{i\cdot} - a_{\cdot j} + a_{\cdot\cdot}
\end{aligned}$$

If we set $\mathbf{A} = -\frac{1}{2}\Delta = (a_{ij})$ and $\mathbf{B} = (b_{ij})$, then $\mathbf{B} = \mathbf{HAH}$, where $\mathbf{H} = \mathbf{I}_n - \frac{1}{n}\mathbf{J}_n$ is a centering matrix and $\mathbf{J}_n$ is an $(n \times n)$-matrix of ones.

The classical scaling algorithm is based upon an eigendecomposition of the matrix

$$\mathbf{B} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top = (\mathbf{V}_t\mathbf{\Lambda}_t^{\frac{1}{2}})(\mathbf{V}_t\mathbf{\Lambda}_t^{\frac{1}{2}})^\top = \mathbf{Y}\mathbf{Y}^\top$$

This eigendecomposition produces $\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 & \cdots & \mathbf{Y}_n \end{pmatrix} \in \mathbb{R}^{t\times n}$, $t = rank(\mathbf{B}) < r$, a configuration whose Euclidean interpoint distances

$$d_{ij}^2 = \|\mathbf{Y}_i - \mathbf{Y}_j\|_2^2$$

match those given in the matrix Δ.

The solution of the classical scaling problem is not unique, e.g. an orthogonal transformation of $\mathbf{X}_i$.

33 Metric Distance Scaling

Metric MDS refers to when the function f is usually taken to be a parametric monotonic function.

A given configuration of points $\{Y_{ij}\} \subset \mathbb{R}^t$ can be evaluated by computing the pairwise distances $\{d_{ij}\}$ and then, for an unknown monotone function f, using the weighted loss function,

$$L_f(\mathbf{Y}_1, \cdots, \mathbf{Y}_n; \mathbf{W}) = \sum_{i<j} w_{ij}[d_{ij} - f(\delta_{ij})]^2$$

as a goodness-of-fit criterion, where $\mathbf{W} = (w_{ij})$ is a given matrix of weights. For a specific dimensionality t, the square-root of L_f,

$$stress = \sqrt{L_f(\mathbf{Y}_1, \cdots, \mathbf{Y}_n; \mathbf{W})}$$

is known as the metric stress function. Minimizing stress over all t-dimensional configurations $\{\mathbf{Y}_{ij}\}$ and monotone f yields an optimal metric distance scaling solution.

Weighting systems include $w_{ij} = \left(\sum_{k<l} \delta_{kl}^2\right)^{-1}$ and $w_{ij} = \delta_{ij}^{-2}$. Sammon nonlinear mapping: $w_{ij} = \delta^{-1}\left(\sum_{k<l} \delta_{kl}\right)^{-1}$ and f is the identity function.

34 Non-Metric Distance Scaling

Non-metric MDS only makes use of the rank order of the dissimilarities. Such an f need only preserve rank order.

We assume that $m = \frac{1}{2}n(n-1)$ dissimilarities in Δ can be strictly ordered from smallest to largest:

$$\delta_{i_1 j_1} < \delta_{i_2 j_2} < \cdots < \delta_{i_m j_m}$$

The objective is to represent these r-dimensional entities as a configuration of n points in the lower-dimensional space $\mathbb{R}^t$. Denote the points in this configuration by $\mathbf{Y}_1, \cdots, \mathbf{Y}_n$ and let

$$d_{ij} = \|\mathbf{Y}_i - \mathbf{Y}_j\|_2$$

Nonmetric distance scaling finds a configuration such that the ordering of the distances

$$d_{i_1 j_1} < d_{i_2 j_2} < \cdots < d_{i_m j_m}$$

matches exactly the ordering of the dissimilarities of δ_{ij}.

To overcome this difficulty of non-monotonicity, we approximate the $\{d_{ij}\}$ by $\hat{d}_{ij}$, say (usually called disparities), which are monotonically related to the $\{d_{ij}\}$ and where

$$\hat{d}_{i_1 j_1} < \hat{d}_{i_2 j_2} < \cdots < \hat{d}_{i_m j_m}$$

Chapter 13 Latent Variable Models for Blind Source Separation

35 Independent Component Analysis

35.1 Definitionn

Given a random r-vector $\mathbf{X} = \begin{pmatrix} X_1 & \cdots & X_r \end{pmatrix}^\top$, $\mathbb{E}\mathbf{X} = \boldsymbol{\mu}$ and $Cov\mathbf{X} = \boldsymbol{\Sigma}_{\mathbf{XX}}$, we first center and shpere $\mathbf{X}$:

$$\mathbf{X} \leftarrow \boldsymbol{\Lambda}^{-\frac{1}{2}}\mathbf{U}^\top(\mathbf{X} - \boldsymbol{\mu})$$

The general ICA model assumes that $\mathbf{X}$ is generated by

$$\mathbf{X} = f(\mathbf{S}) + \mathbf{e}$$

where $\mathbf{S} = \begin{pmatrix} S_1 & \cdots & S_m \end{pmatrix}^\top$ is an (unobservable) random m-vector variate of sources whose components $\{S_j\}$ are independent latent variables each having zero mean, $f : \mathbb{R}^m \to \mathbb{R}^r$ is an unknown mixing function, and $\mathbf{e}$ is a zero-mean, additive, r-vector-valued component that represents measurement noise and any other type of variability that cannot be directly attributed to the sources.

We assume that $\mathbb{E}\mathbf{S} = \mathbf{0}$ and $Cov(\mathbf{S}) = \mathbf{I}_m$, but that the distribution of $\mathbf{S}$ is otherwise unknown.

The problem is to invert f and estimate $\mathbf{S}$. It is ill-posed and needs some additional constraints or regularization on $\mathbf{S}$, f, and $\mathbf{e}$. If f is a linear function, $f(\mathbf{S}) = \mathbf{AS}$, where A is a "mixing" matrix, then this is described as a linear ICA model, nonlinear ICA otherwise. A model is referred to as noiseless ICA if there is no additive noise $\mathbf{e}$, or nosiy ICA otherwise.

The noiseless ICA model with linear mixing, $\mathbf{X} = \mathbf{AS}$, can only be solved if the vector $\mathbf{S}$ with independent components is not Gaussian.

For a given $\mathbf{A} \in \mathbb{R}^{r \times m}$ $(m \leq r)$ with full-rank, there exists a separating (or unmixing matrix) $\mathbf{W}$ such that the sources can be recovered exactly from the observed $\mathbf{X}$ by $\mathbf{S} = \mathbf{WX}$, where

$$\mathbf{W} = (\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top$$

If $r = m$, then $\mathbf{W} = \mathbf{A}^{-1}$. If $\mathbf{X}$ has been centered and sphered, then the resulting square mixing matrix $\mathbf{A}$ is orthogonal, and so $\mathbf{W} = \mathbf{A}^\top$.

Given an estimate $\hat{\mathbf{W}} = \begin{pmatrix} \hat{\mathbf{w}}_1 & \cdots & \hat{\mathbf{w}}_m \end{pmatrix}^\top$ of the separating matrix $\mathbf{W}$, the source component vector $\mathbf{S}$ is approximated by $\mathbf{Y} = \hat{\mathbf{W}}\mathbf{X}$ where the elements, $Y_1 = \mathbf{w}_1^\top\mathbf{X}, \cdots, Y_m = \mathbf{w}_m^\top\mathbf{X}$, of $\mathbf{Y}$ are taken to be statistically independent and as non-Gaussian as possible.

35.2 The FastICA Algorithm

Let Y be a projection, $Y = \mathbf{w}^\top\mathbf{X}$, of $\mathbf{X}$. The idea is to find that direction $\mathbf{w}$ that optimizes a given objective function.