ISYE 7405: Multivariate Data Analysis **Georgia Tech**

Chapter 1: Review of Matrix Theory

Lecturer: Shihao Yang

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

1.1 Vector Space (Linear Space)

1. **Euclidean Space**

   In Euclidean space $\mathbb{R}^p$, given a vector $x = (x_1, x_2, \ldots, x_p) \in \mathbb{R}^p$, $y = (y_1, y_2, \ldots, y_p) \in \mathbb{R}^p$, the vector operations are defined as

$$x + y = (x_1 + y_1, x_2 + y_2, \ldots, x_p + y_p)$$

$$c \cdot x = (cx_1, cx_2, \ldots, cx_p)$$

2. **Inner Product**

   The inner product of two vectors is defined as

$$< x, y >_A = x^T Ay,$$

   where A is symmetric and positive definite. The norm of vectors

$$||x||_A = \sqrt{< x, x >_A} = \sqrt{x^T Ax}$$

   Examples: 1. $A = I$; 2. $A = \Sigma$ (covariance matrix); 3. $A = \Sigma^{-1}$

3. **Angle between vectors**:

$$\cos(\alpha_{xy}) = \frac{< x, y >}{||x||||y||}$$

   Cauchy-Schwartz inequality:

$$| < x, y > | \leq ||x|| \cdot ||y||$$

4. **Column space of a matrix**:

   Given a matrix $V_{p\times q} = (v_1, v_2, \ldots, v_q)$, the column space is defined as

$$\mathcal{L}_{col}(\mathrm{V}) = \{w : \ w = \sum_{i=1}^{q} v_i c_i\} = \{w = \mathrm{V}c\}$$

5. **Orthogonal basis**:

   Assume $v_1, v_2, \ldots, v_q$ are linearly independent, then we can always find an orthogonal set $(\gamma_1, \gamma_2, \ldots, \gamma_q)$, $\gamma_i \in \mathbb{R}^p$, s.t. $\mathrm{V} = \Gamma C_{q\times q}$, where C is invertible.

6. **Orthogonal space (null space)**:

$$\mathcal{L}_{col}^{\perp}(\mathrm{V}) = \{w :< w, v_i >= 0,\ i = 1, 2, \dots, q\} = \{w :< w, v >= 0,\ \forall v \in \mathcal{L}_{col}(\mathrm{V})\}$$

For Euclidean space, $\mathcal{L}_{col}^{\perp}(\mathrm{V}) = \{w : w^T\mathrm{V} = 0\}$.

Theorem 1.1 *Let $X_{p\times r} = (x_1, \dots, x_r)$, then $\mathcal{L}_{col}(\mathrm{X}) = \mathcal{L}_{col}(\mathrm{X}X^T)$.*

Proof:

(1) $\mathcal{L}_{col}(\mathrm{X}X^T) \subseteq \mathcal{L}_{col}(\mathrm{X})$

(2) $\mathcal{L}_{col}(\mathrm{X}) \subseteq \mathcal{L}_{col}(\mathrm{X}X^T) \Leftrightarrow \mathcal{L}_{col}^{\perp}(\mathrm{X}X^T) \subseteq \mathcal{L}_{col}^{\perp}(\mathrm{X})$

For $\forall w \in \mathcal{L}_{col}^{\perp}(\mathrm{X}X^T)$, $w^T xx^T = 0 \Rightarrow w^T xx^T w = 0 \Rightarrow w^T x = 0$. ■

1.2 Rank of a Matrix

Definition 1.2 *rank*$(X_{p\times r}) = \dim \mathcal{L}_{col}(\mathrm{X})$

Facts:

(1) rank(X)=rank(X^T)

(2) Suppose $B_{p\times p},\ C_{q\times q}$ are both nonsingular, then rank(BXC)=rank(X).

(3) rank(XX^T)=rank(X)=rank(X^TX)

Remark: Let $X = \begin{pmatrix} v_1^T \\ \vdots \\ v_p^T \end{pmatrix}$, then $G = XX^T = (v_i^T v_j)_{ij}$ is called "Gram Matrix".

1.3 Random Vectors

Given probability space $\mathcal{L}^2(\Omega, \mathcal{F}, \mathrm{P})$, $z_i, z_j \in \mathcal{L}^2$ (square integrable), $E[|z_i z_j|] < \infty$, $z = \begin{pmatrix} z_1 \\ \vdots \\ z_p \end{pmatrix}$.

- **Covariance matrix**: $\Sigma = Cov(z, z) \Leftrightarrow \Sigma_{ij} = \sigma_{ij} = Cov(z_i, z_j)$.

- **Linear functions**: $x \in \mathbb{R}^p$, $\mathcal{X} = x^T z = \sum_{i=1}^p x_i z_i$.

- **Inner product**: $< \mathcal{X}, \mathcal{Y} >= Cov(\mathcal{X}, \mathcal{Y}) = x^T \Sigma y$, where $\mathcal{X} = x^T z, \mathcal{Y} = y^T z$, and x, y are constant vectors.

  Recall that: 1. $Cov(az_1, z_2) = a\ Cov(z_1, z_2)$; 2. $Cov(z_1 + z_2, z_3) = Cov(z_1, z_3) + Cov(z_2, z_3)$

- **Length**: $||\mathcal{X}|| = \sqrt{< \mathcal{X}, \mathcal{X} >} = \sqrt{x^T \Sigma x}$

- **Angle**: $\alpha_{\mathcal{X}\mathcal{Y}} = \cos^{-1}(\frac{<\mathcal{X},\mathcal{Y}>}{||\mathcal{X}||\cdot||\mathcal{Y}||})$

  Remark: z takes value in $\mathbb{R}^p$, it is possible that z takes value in a subspace of $\mathbb{R}^p$.

Example: Suppose $z \sim$ multinomial$(n, \pi)/n$, $z_i \in [0,1]$, and $\pi = (\pi_1, \ldots, \pi_p)$ is the probability, i.e. $\pi_1 + \ldots + \pi_p = 1$.

1. What is the Σ of z?

2. $\mathbf{1} \in \mathcal{L}^{\perp}_{col}(\Sigma)$.

ISYE 7405: Multivariate Data Analysis **Georgia Tech**

Chapter 2: Projections

Lecturer: Shihao Yang

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

2.4 Projection

Suppose $V_{n\times p} = (v_1, v_2, \ldots, v_p)$ is of rank $p < n$. For $\forall y \in \mathbb{R}^n$, we want to decompose y into $y = \hat{y} + y^{\perp}$, where $\hat{y} \in \mathcal{L}_{col}(V), y^{\perp} \in \mathcal{L}_{col}^{\perp}(V)$.

$$\hat{y} = V\hat{\beta} = \sum_{i=1}^{p} \hat{\beta}_i v_i, \hat{\beta} \in \mathbb{R}^p$$

$$V^T y = V^T \hat{y} + V^T y^{\perp} = V^T \hat{y} = V^T V \hat{\beta} \Rightarrow \hat{\beta} = (V^T V)^{-1} V^T y$$

which is the linear regression estimator.

$$\hat{y} = V\hat{\beta} = V(V^T V)^{-1} V^T y = \hat{P} y$$

we call $\hat{P} = V(V^T V)^{-1} V^T$ projection matrix.

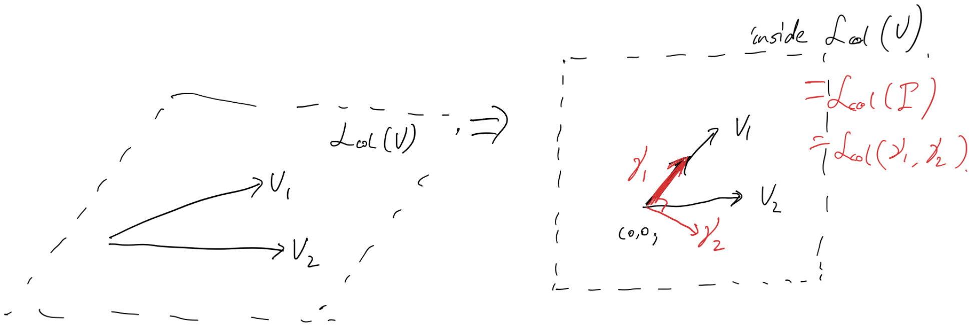
$$y^{\perp} = y - \hat{y} = (I - \hat{P})y \Rightarrow P^{\perp} := I - \hat{P}$$

Hence $\hat{P}$ is the projection matrix into $\mathcal{L}_{col}(V)$, $P^{\perp}$ is the projection matrix into $\mathcal{L}_{col}(V^{\perp})$.

Properties of projection matrix:
(i) $\hat{P}$ and $P^{\perp}$ are symmetric;
(ii) $\hat{P}^2 = \hat{P} \cdot \hat{P} = \hat{P}$;
(iii) $\mathcal{L}_{col}(\hat{P}) = \mathcal{L}_{col}(V)$;
(iv) rank($\hat{P}$)=p, rank($P^{\perp}$)=n-p;
(v) $\hat{y}^T y^{\perp} = 0 \Leftrightarrow < \hat{y}, y^{\perp} >= 0$;
(vi) $||\hat{y}||^2 = y^T \hat{P} y,\ ||y^{\perp}||^2 = y^T P^{\perp} y \Rightarrow ||y||^2 = ||\hat{y}||^2 + ||y^{\perp}||^2$;
(vii) $\cos^2(\alpha_{y,\hat{y}}) = \frac{||\hat{y}||^2}{||y||^2} = \frac{y^T \hat{P} y}{y^T y}$, $\alpha_{y,\hat{y}} = \min\{\alpha_{y,w} : \forall w \in \mathcal{L}_{col}(V)\}$.
(viii) Eigenvalues of $\hat{P}$ are either 0 or 1.
–Geometrical intuition: An eigenvector of projection matrix, with nonzero eigenvalue, is in the column space of this projection matrix both before and after the transformation. The projection matrix $\hat{P}$ project every vector into $\mathcal{L}_{col}(\hat{P})$. The only vectors that are still in their original column space after being projected into $\mathcal{L}_{col}(\hat{P})$ are those which are already in $\mathcal{L}_{col}(\hat{P})$, with eigenvalue = 1 since their lengths do not change. For vectors in $\mathcal{L}_{col}(\hat{P}^{\perp})$ are also the eigenvectors of $\hat{P}$ since after projection they become a dot in $\mathcal{L}_{col}(\hat{P}^{\perp})$, with eigenvalue = 0.
In other words, if $w \in \mathcal{L}_{col}(\hat{P})$ then $\hat{P}w = w$. If $w \in \mathcal{L}_{col}(\hat{P}^{\perp})$ then $\hat{P}w = 0$.
–Mathematical proof: Note that $\hat{P}^2 = \hat{P}$ and therefore $\forall v$:

$$\lambda^2 v = \hat{P}(\lambda v) = \hat{P}(\hat{P} v) = \hat{P} v = \lambda v \Rightarrow \lambda^2 = \lambda \Rightarrow \lambda = 0 \quad or \quad 1$$

(ix) $V_{n\times p} = (v_1, v_2, ..., v_n), \Gamma = (\gamma_1, \gamma_2, ..., \gamma_n)$ where γ_i are all orthogonal to each other with $\mathcal{L}_{col}(V) = \mathcal{L}_{col}(\Gamma)$ Porjection matrix is $\hat{P} = V(V^\top V)^{-1} V^\top = \Gamma\Gamma^\top$. Also when $p < n$ we have $\Gamma\Gamma^\top \neq I$.



2.5 Projection matrix in a general metric

We define $\langle y, v\rangle_A = y^\top Av$. Vector y can be written as $y = \hat{y} + y^\perp$. Then $\hat{y}$ can be written as $\hat{y} = V\hat{\beta}$.

$$\langle y, v\rangle = \langle \hat{y}, v\rangle + \langle y^\perp, v\rangle = \langle \hat{y}, v\rangle \Rightarrow V^\top Ay = V^\top AV\hat{\beta} \Rightarrow \hat{\beta} = (V^\top AV)^{-1}V^\top Ay$$

Thus we get,

$$\hat{y} = V\hat{\beta} = V(V^\top AV)^{-1}V^\top Ay = V\langle V, V\rangle^{-1}\langle V, y\rangle$$

Exercise: Verify the Pythagorean theorem:

$$||y||^2 = ||\hat{y}||^2 + ||y^\perp||^2$$

2.6 Linear Prediction

Assume Y to be a random variable and $V = (V_1, V_2, ..., V_p)$ where v_i is a random variable. Assume that $E[Y] = E[V_i] = 0$. Define $\langle X, Y\rangle = cov(X, Y)$.
How to find a linear prediction of of $(V_1, V_2, ..., V_p)$ such that best predict V?
"Best": norm of residual is smallest $\Rightarrow$ variance of residual is smallest, i.e., $\hat{Y}$ should be the projection of Y on V.
According to the conclusion in Section 1.7,

$$\hat{Y} = V\langle V, V\rangle^{-1}\langle V, Y\rangle = V\sigma_{VV}^{-1}\sigma_{VY}$$

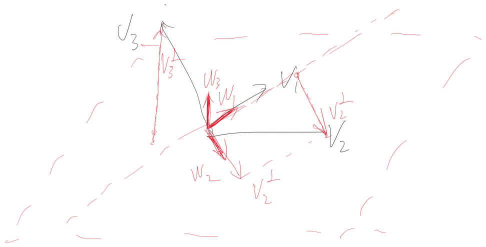
Properties of Linear Prediction:

1. $\hat{Y}, Y^\perp$ is uncorrelated since $\langle \hat{Y}, Y^\perp\rangle = 0$ (The reverse direction is true as well).
2. $\text{Var}(\hat{Y}) = \langle \hat{Y}, \hat{Y}\rangle = \sigma_{YV}\sigma_{VV}^{-1}\sigma_{VY}$.
   Proof: $\hat{Y} = \hat{\beta}V$ where $\hat{\beta} = \sigma_{YV}\sigma_{VV}^{-1}$.
   Then, $\text{Var}(\hat{Y}) = \text{Var}(\hat{\beta}V) = \hat{\beta}\text{Var}(V)\hat{\beta}^T = \sigma_{YV}\sigma_{VV}^{-1}\text{Var}(V)\sigma_{VV}^{-1}\sigma_{VY} = \sigma_{YV}\sigma_{VV}^{-1}\sigma_{VY}$
   Subsequently, $\text{Var}(Y^\perp) = \text{Var}(Y) - \text{Var}(\hat{Y}) = \sigma_{YY} - \sigma_{YV}\sigma_{VV}^{-1}\sigma_{VY}$
3. $\text{Var}(Y^\perp) = \min_{\beta\in\mathbb{R}^p} \text{Var}(Y - V\beta)$
4. $\frac{\text{Var}(\hat{Y})}{\text{Var}(Y)} = \frac{\sigma_{YV}\sigma_{VV}^{-1}\sigma_{VY}}{\sigma_{YY}} = \rho^2_{Y|V_1,V_2,...,V_p} = R^2$. R^2 is the multiple correlation coefficient between Y and $(v_1, v_2, ..., v_p)$.

2.7 Gram-Schimdt Orthogonalization

$V_{n\times p} = (v_1, v_2, ..., v_n)$, denote: $V_j = (v_1, v_2, ..., v_j)$ and $P_j = V_j(V_j^\top V_j)^{-1}V_j^\top$
Algorithm:

- Let $W_1 = V_1/||V_1||$
- $V_2^\perp = (I - \hat{P}_1)V_2$, $W_2 = V_2^\perp/||V_2^\perp||$
  ......
- $V_j^\perp = (I - \hat{P}_{j-1})V_j$, $W_j = V_j^\perp/||V_j^\perp||$
  ......
- $V_p^\perp = (I - \hat{P}_{p-1})V_p$, $W_p = V_p^\perp/||V_p^\perp||$



Let's change the direction of expression

- $V_1 = U_{11}W_1$
- $V_2 = U_{12}W_1 + U_{22}W_2$
  ......
- $V_j = U_{1j}W_1 + ... + U_{jj}W_j$
  ......
- $V_p = U_{1p}W_1 + ... + U_{pp}W_p$

$$V = WU \Rightarrow \text{QR Decomposition, where } W^TW = I_p,\ U_{p\times p} \text{ upper triangle}$$

2.8 Gram Matrix and Projection

Gram matrix: $G_{ij} = \langle x_i, x_j\rangle$
Assume

$$V_{n\times(p_1+p_2)} = (V_1, V_2), \qquad G = V^TV = \begin{pmatrix} V_1^TV_1 & V_1^TV_2 \\ V_2^TV_1 & V_2^TV_2 \end{pmatrix} = \begin{pmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{pmatrix}$$

Let's consider the projection of V_2 to V_1

$$\hat{P}_1 = V_1(V_1^TV_1)^{-1}V_1^T = V_1G_{11}^{-1}V_1^T \qquad P_1^\perp = I_1 - \hat{P}_1$$
$$\hat{V}_2 = \hat{P}_1V_2 = V_1G_{11}^{-1}V_1^TV_2 = V_1G_{11}^{-1}G_{12} \qquad V_2^\perp = P_1^\perp V_2$$

Then,

$$G_{22} = V_2^TV_2 = (\hat{V}_2 + V_2^\perp)^T(\hat{V}_2 + V_2^\perp) = \hat{V}_2^T\hat{V}_2 + (V_2^\perp)^TV_2^\perp = \langle\hat{V}_2, \hat{V}_2\rangle + \langle V_2^\perp, V_2^\perp\rangle$$
$$\hat{V}_2^T\hat{V}_2 = (V_2 - V_2^\perp)^T\hat{V}_2 = V_2^T\hat{V}_2 = V_2^T\hat{P}_1V_2 = V_2^TV_1G_{11}^{-1}V_1^TV_2 = G_{21}G_{11}^{-1}G_{12}$$
$$(V_2^\perp)^TV_2^\perp = G_{22} - G_{21}G_{11}^{-1}G_{12}$$

2.8.1 Application to Linear Predictions

Assume $E[Y_1] = E[Y_2] = 0$, $\dim(Y_1) = p_1$ and $\dim(Y_2) = p_2$, and we want to use Y_1 (a group of random variables) to predict Y_2 (another group of random variables):
$\Sigma = \left\langle \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}, \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \right\rangle = Cov\left(\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}, \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \right) = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}_{(p_1+p_2)\times(p_1+p_2)}$
$\hat{Y_2} = \Sigma_{21}\Sigma_{11}^{-1}Y_1 \quad Y_2^{\perp} = Y_2 - \hat{Y_2}$
$\Sigma_{22}^{\perp} = Cov(Y_2^{\perp}) = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$, partial covariance matrix of Y_2 after linear regression on Y_1

ISYE 7405: Multivariate Data Analysis **Georgia Tech**

Chapter 3: Determinants & Jacobians

Lecturer: Shihao Yang

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

3.9 Determinants

Definition 3.3 *Given a square matrix* $A_{p\times p} = \begin{pmatrix} \vdots & & \vdots \\ a_1 & \cdots & a_p \\ \vdots & & \vdots \end{pmatrix}$*, the determinant of* A *is*

$$det(A) = |A| = \sum_{\pi} sgn(\pi) A_{1\pi(1)} A_{2\pi(2)} \cdots A_{p\pi(p)}$$

where π *denotes all possible permutations.*

3.9.1 Properties of determinants

1. For any upper triangular matrix $U_{p\times p}$, $|U_{p\times p}| = \prod_{i=1}^{n} u_{ii}$
2. $|A^T| = |A|$
3. $\begin{vmatrix} \vdots & \vdots & & \vdots & & \vdots \\ a_1 & a_2 & \cdots & c\cdot a_j & \cdots & a_p \\ \vdots & \vdots & & \vdots & & \vdots \end{vmatrix} = c|A|$, $|cA| = c^p|A|$
4. $\begin{vmatrix} \vdots & \vdots & & \vdots & & \vdots & & \vdots \\ a_1 & a_2 & \cdots & a_j & \cdots & a_i & \cdots & a_p \\ \vdots & \vdots & & \vdots & & \vdots & & \vdots \end{vmatrix} = -|A|$, where $i < j$, swap a_i and a_j
5. $\begin{vmatrix} \vdots & & \vdots & & \vdots & & \vdots \\ a_1 & \cdots & a_i + a_j & \cdots & a_j & \cdots & a_p \\ \vdots & & \vdots & & \vdots & & \vdots \end{vmatrix} = |A|$
6. $|A_{p\times p}B_{p\times p}| = |A||B| \Rightarrow |A^{-1}| = \frac{1}{|A|}$; if A is an orthogonal matrix, $A^TA = I \Rightarrow |A| = \pm 1$
7. $|A_{p\times p}| = 0$ if and only if $rank(A) < p$
8. $\begin{vmatrix} A_{p\times p} & 0 \\ C_{q\times p} & D_{q\times q} \end{vmatrix} = |A| \cdot |D|$

9. $\begin{vmatrix} A_{p\times p} & B_{p\times q} \\ C_{q\times p} & D_{q\times q} \end{vmatrix} = \begin{cases} |D|\cdot|A-BD^{-1}C|, & \text{if } rank(D)=q \\ |A|\cdot|D-CA^{-1}B|, & \text{if } rank(A)=p \end{cases}$

   Proof:

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} A-BD^{-1}C & B \\ 0 & D \end{pmatrix} \begin{pmatrix} I_p & 0 \\ D^{-1}C & I_q \end{pmatrix}$$

■

10. $|I_p + A_{p\times q}B_{q\times p}| = |I_q + B_{q\times p}A_{p\times q}|$

    Proof: Apply property 9 to

$$\begin{vmatrix} I_p & -A_{p\times q} \\ B_{q\times p} & I_q \end{vmatrix}$$

■

(Exercise)

$$\begin{vmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{vmatrix} = (1-\rho)^{n-1}\left[1+(n-1)\rho\right]$$

3.9.2 Geometric Interpretation of Determinant

Recall

$$det(A) = |A|, \quad A = \begin{pmatrix} \vdots & \vdots & & \vdots \\ a_1 & a_2 & \cdots & a_p \\ \vdots & \vdots & & \vdots \end{pmatrix}_{p\times p}$$

Geometric meaning:

$|det(A)| = p$-dimensional volume of the parallelogram consisting of $\{\vec{a_1}, \vec{a_2}, \cdots, \vec{a_p}\}$

Example: Consider

$$A = \begin{pmatrix} \vdots & \vdots \\ a_1 & a_2 \\ \vdots & \vdots \end{pmatrix} = \begin{pmatrix} 2 & -1 \\ 1 & 1 \end{pmatrix}$$

$$|A| = 3$$

The change in volume is shown in Figure 3.1.

QR Decomposition:

$A = WU$, where W is orthogonal basis, U is upper triangular.

$$|A| = |W||U| = \pm|U| = \pm\prod_{i=1}^{p} u_{ii}$$

Figure 3.1: $det(A) = 3$

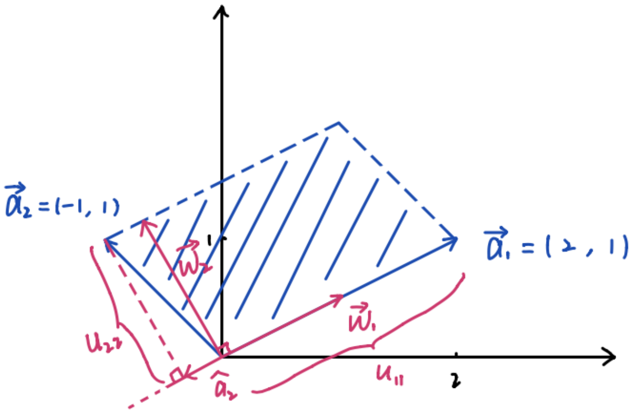
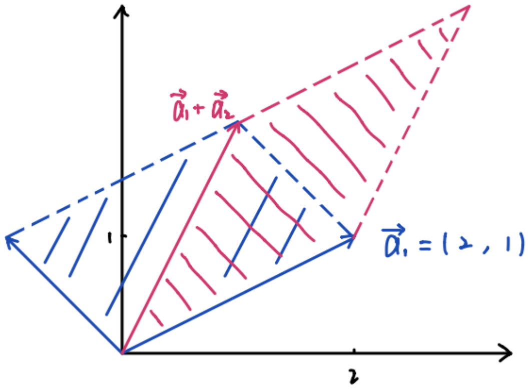


Figure 3.2: $|a_1, a_2| = |a_1, a_1 + a_2|$



Example. $\forall a_1, a_2 \in \mathbb{R}^2$, $|a_1, a_2| = |a_1, a_1 + a_2|$. Geometric interpretation as shown in Figure 3.2: Red Area = Blue Area.

3.10 Jacobian

Motivation for Jacobian: Suppose random variable X has density function $f(x)$. $\tilde{X} = m(X)$, where m is a 1-1 map from $\mathbb{R}^n \to \mathbb{R}^n$. What's the density function $\tilde{f}(\tilde{\mathbf{x}})$ for $\tilde{X}$?

Definition 3.4 *Suppose $\tilde{X} = m(X)$, where m is a 1-1 map from $\mathbb{R}^n \to \mathbb{R}^n$.*

$$M(X) = \left(\frac{\partial \tilde{x_i}}{\partial x_j}\right)_{ij}$$
$$M^{-1}(X) = \left(\frac{\partial x_i}{\partial \tilde{x_j}}\right)_{ij}$$

Jacobian

$$J(X \to \tilde{X}) = \left|det(M^{-1}(X))\right| = \left|det\Big(\frac{\partial x_i}{\partial \tilde{x_j}}\Big)\right|$$

Then, for our problem in random variable

$$\tilde{f}(\tilde{\mathbf{x}}) = f(\mathbf{x}) \cdot J(X \to \tilde{X})$$

3.10.1 Intuitive Sense: why use Jacobian (determinant)?

Example. Consider $n = 2$ as shown in Figure 3.3, where

$$\vec{e_1} = (1,0) \quad \vec{e_2} = (0,1)$$
$$d\mathbf{x} = (dx_1, dx_2)$$
$$M = (\vec{M_1}, \vec{M_2}) = \begin{pmatrix} \frac{\partial \tilde{x_1}}{\partial x_1} & \frac{\partial \tilde{x_1}}{\partial x_2} \\ \frac{\partial \tilde{x_2}}{\partial x_1} & \frac{\partial \tilde{x_2}}{\partial x_2} \end{pmatrix}$$
$$M(\mathbf{x} + \vec{e_1} dx_1) \approx \tilde{\mathbf{x}} + \vec{M_1} dx_1 + o(dx_1) \quad \text{by Taylor Expansion}$$

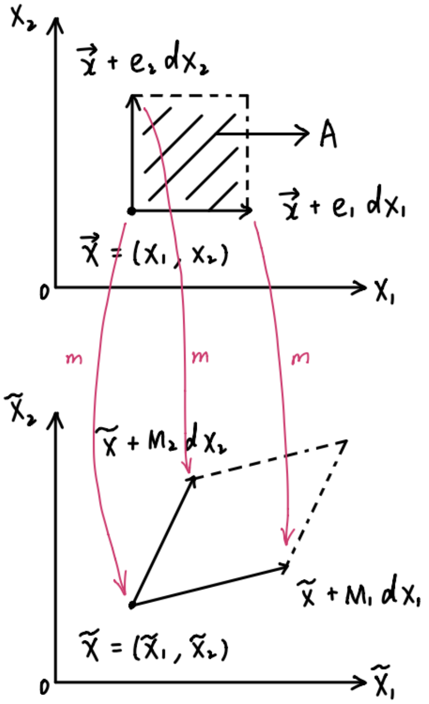
$$Vol(\tilde{A}) = det\begin{pmatrix} \vdots & \vdots \\ M_1 dx_1 & M_2 dx_2 \\ \vdots & \vdots \end{pmatrix} = \left(\prod_{i=1}^{n} dx_i\right) \cdot |det(M)|$$

Probability mass:

$$f(\mathbf{x})Vol(A) = \tilde{f}(\tilde{\mathbf{x}})Vol(\tilde{A})$$
$$\Rightarrow \tilde{f}(\tilde{\mathbf{x}}) = f(\mathbf{x}) \cdot \frac{Vol(A)}{Vol(\tilde{A})}$$

$$Vol(\tilde{A}) = det(M_1 dx_1, M_2 dx_2) = (\Pi_{i=1}^{n} dx_i) \cdot |det(M)| = Vol(A) \cdot |det(M)|$$
$$\Rightarrow \tilde{f}(\tilde{\mathbf{x}}) = f(\mathbf{x}) \cdot \frac{1}{|det(M)|} = f(\mathbf{x}) \cdot det(M^{-1})$$

Figure 3.3: n=2


x_2
$\vec{x} + e_2 dx_2$
A
$\vec{x} + e_1 dx_1$
$\vec{x} = (x_1, x_2)$
0
x_1
m
m
m
$\tilde{x}_2$
$\tilde{x} + M_2 dx_2$
$\tilde{x} + M_1 dx_1$
$\tilde{x} = (\tilde{x}_1, \tilde{x}_2)$
0
$\tilde{x}_1$

Example. 1-d space (n-d space):

$$\begin{aligned}\int \tilde{f}(\tilde{x})|d\tilde{x}| &= \int f(x)|dx| \\ \Leftrightarrow \tilde{f}(\tilde{x})|d\tilde{x}| &= f(x)|dx| \\ \Leftrightarrow \tilde{f}(\tilde{x}) &= f(x)\frac{|dx|}{|d\tilde{x}|}\end{aligned}$$

Classical Example. Polar system.

Figure 3.4: Polar coordinates

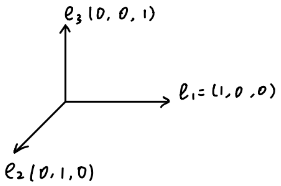
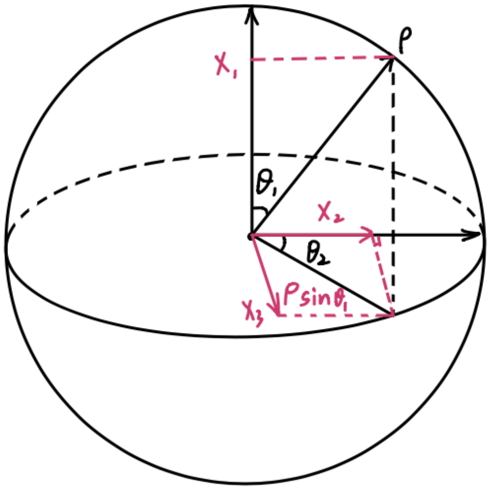


Figure 3.5: Polar to Cartesian



$$\begin{aligned}
x_1 &= \rho\cos\theta_1\\
x_2 &= \rho\sin\theta_1 cos\theta_2\\
&\vdots\\
x_{n-1} &= \rho\sin\theta_1\cos\theta_2\cdots\sin\theta_{n-1}\cos\theta_{n-1}\\
x_n &= \rho\sin\theta_1\cdots\sin\theta_{n-1}
\end{aligned}$$

where $\rho \geq 0,\ 0 \leq \theta_i \leq \pi,\ \theta_{n-1} \in [0, 2\pi]$.

Calculate: $J(X \to (\rho, \theta_1, \theta_2, ..., \theta_{n-1})) = \rho^{n-1}\sin^{n-2}\theta_1\sin^{n-3}\theta_2\cdots\sin\theta_{n-2}$
Proof: Math induction on $J_n = \rho\sin^{n-2}\theta_1 J_{n-1}$ ■

3.11 Integral Jacobian: $X \xrightarrow{m} \tilde{X}$

Example (Motivation) $X_1, X_2, ..., X_n \overset{i.i.d}{\sim} f(x)$, what is the density of :

$$\begin{gathered}
(1)\ x_1 + x_2 + ... + x_n\\
(2)\ x_1^2 + x_2^2 + ... + x_n^2\\
\text{(3) Other many to one mapping}
\end{gathered}$$

Definition 3.5 *$X \overset{m}{\to} Y_1$. We can find Y_2 , such that the mapping $X \to (Y_1, Y_2)$ is one-to-one. The integral Jacobian of $X \to Y_1$ is defined as*

$$J(X \to Y_1) = \int_{y_2} J(X \to (Y_1, y_2))dy_2$$

Lemma 3.6

$$J(X \to Y_1) = \lim_{dy_1 \to 0} \frac{Vol(m^{-1}([y_1, y_1 + dy_1]))}{Vol([y_1, y_1 + dy_1])}$$

Example. Suppose $Y_1 = m(X)$, X is 2-dimensional ($n = 2$), and Y_1 is scalar. (See Figure 3.6)

$$\begin{aligned}
\tilde{A} &= \{(\tilde{y_1}, \tilde{y_2}) : \tilde{y_1} \in [y_1, y_1 + dy_1]\}\\
A &= \{x : m(x) \in [y_1, y_1 + dy_1]\} = m^{-1}(\tilde{A}) \Leftrightarrow \tilde{A} = m(A)\\
\text{Jacobian} &= \lim_{dy_1 \to 0} \frac{Vol(A)}{Vol(\tilde{A})} = \lim_{dy_1 \to 0} \frac{Vol(m^{-1}([y_1, y_1 + dy_1]))}{Vol([y_1, y_1 + dy_1])}
\end{aligned}$$

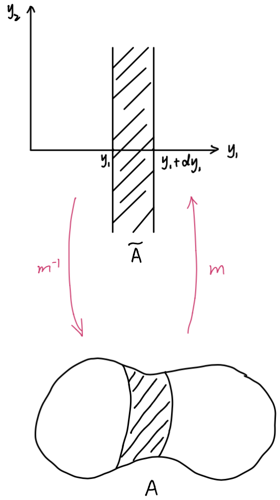
Lemma 3.7 *Suppose $X \overset{m}{\to} Y_1$, X has density $f^X(x) = g \circ m(x) = g(m(x)) = g(y_1)$. Then the density function of Y_1 is*

$$f^{Y_1}(y_1) = g(y_1)J(X \to y_1)$$

Example. $X_1, X_2, ..., X_n \overset{i.i.d}{\sim} \mathcal{N}(0, 1), \rho = \sqrt{X_1^2 + ... + X_n^2}$, what's density of ρ?

Solution.

Figure 3.6:



$$
\begin{aligned}
f^X(x) &= \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x_i^2} \right) \\
&= \left(\frac{1}{\sqrt{2\pi}} \right)^n e^{-\frac{1}{2}\sum x_i^2} \\
&= (2\pi)^{-\frac{n}{2}} \, e^{-\frac{1}{2}\rho^2}
\end{aligned}
$$

From lemma, we have

$$f^\rho(\rho) = (2\pi)^{-\frac{n}{2}} e^{-\frac{1}{2}\rho^2} J(X \to \rho)$$

Using polar system, we can show that

$$
\begin{aligned}
J(X \to \rho) &= \int J(X \to (\rho, \theta_1, \cdots, \theta_{n-1})) d\theta_1 \cdots d\theta_{n-1} \\
&= \rho^{n-1} \frac{2\pi^{n/2}}{\Gamma(n/2)}
\end{aligned}
$$

So,

$$f^\rho(\rho) = \frac{\rho^{n-1} e^{-\frac{1}{2}\rho^2}}{2^{\frac{n}{2}-1}\Gamma(n/2)} \sim \chi_n$$

Proof: (of Lemma 3.7)

Consider $X \to \binom{Y_1}{Y_2}$ such that it is one-to-one. By definition,

$$\begin{aligned}
f^{Y_1}(y_1) &= \int f^{Y_1,Y_2}(y_1,y_2)dy_2 \\
&= \int f^X(x)J(X \to (y_1,y_2))dy_2 \\
&= \int g(m(x))J(X \to (y_1,y_2))dy_2 \\
&= \int g(y_1)J(X \to (y_1,y_2))dy_2 \\
&= g(y_1)\int J(X \to (y_1,y_2))dy_2 \\
&= g(y_1)J(X \to (y_1)
\end{aligned}$$

■

Remark: In one-to-one mapping of m,

$$f^{Y_1}(y_1) = f^X(m^{-1}(y_1))J(X \to y_1)$$

Trick: Reverse Lemma 3.7 to find Jacobian.

$$J(X \to Y_1) = \frac{f^{Y_1}(y_1)}{f^X(x)}$$

provided that $f^X(x) = g(y_1)$.

Example. $X = (X_1, X_2, \ldots, X_n)$, $m : (\mathbb{R}^+)^n \to \mathbb{R}^+$ maps X to $S = \sum_{i=1}^n X_i$, . What is $J(X \to S)$?

Solution. Reverse Lemma 3.7. Consider $X_1, X_2, \ldots, X_n \overset{iid}{\sim} exp(1)$, $f(x) = e^{-x}$. Then, $S = X_1 + X_2 + \cdots + X_n \sim \Gamma(n, 1)$.

$$\begin{aligned}
f^S(s) &= \frac{1}{\Gamma(n)} s^{n-1} e^{-s} \\
f^X(x) &= \Pi_{i=1}^n e^{-x_i} = e^{-\sum_{i=1}^n x_i} = e^{-s} \\
J(X \to S) &= \frac{f^S(s)}{f^X(x)} = \frac{s^{n-1}}{\Gamma(n)}
\end{aligned}$$

Lemma 3.8 ***Chain rule of Integral Jacobian.*** *If $X \overset{1-1}{\to} (Y_1, Y_2)$, $Y \overset{1-1}{\to} (Z_1, Z_2)$, then $J(X \to Z_1) = J(X \to Y_1) \cdot J(Y_1 \to Z_1)$.*

Proof: Consider $X \overset{1-1}{\to} (Z_1, Z_2, Y_2)$. Using definition,

$$\begin{aligned}
J(X \to Z_1) &= \int\int J(X \to (Z_1, z_2, y_2)) dz_2 dy_2 \\
J(X \to Y_1) &= \int J(X \to (Y_1, y_2)) dy_2 \\
J(Y_1 \to Z_1) &= \int J(Y_1 \to (Z_1, z_2)) dz_2 \\
J(X \to Y_1) \cdot J(Y_1 \to Z_1) &= \int\int J(X \to (Y_1, y_2) J(Y_1 \to (Z_1, z_2)) dz_2 dy_2 \\
&= \int\int J(X \to (Z_1, z_2, y_2)) dz_2 dy_2 \\
&= J(X \to Z_1)
\end{aligned}$$

where the second last equality is using standard chain rule for Jacobian.

Remark: We can also use Lemma 3.6 with the intuition that as $dz_1 \to 0$, $dy_1 \to 0$.

$$J(X \to Z_1) = \lim_{z_1 \to 0} \frac{Vol(B)}{Vol(\tilde{A})} = \lim_{y_1 \to 0} \frac{Vol(A)}{Vol(\tilde{A})} \lim_{z_1 \to 0} \frac{Vol(B)}{Vol(A)}$$

■

Lemma 3.9 ***(Hsu's Lemma)*** $X_{n\times p} \overset{m}{\to} S_{p\times p} = X^T X$.

$$J(X \to S) = \frac{\pi^{\frac{np}{2} - \frac{p^2}{4} + \frac{p}{4}}}{\Pi_{j=1}^p \Gamma(\frac{n-j+1}{2})} |det(S)|^{\frac{n-p+1}{2}}$$

Proof: Use the trick of inverting Lemma 3.7. ■

3.12 Spectral Decomposition (Eigendecomposition)

Suppose $A_{p\times p}$ is a real-valued, symmetric with rank K. Then A can always be decomposed as

$$A_{p\times p} = \Gamma_{p\times k} \Lambda \Gamma_{k\times p}^T$$

where $\Gamma = \begin{pmatrix} \vdots & \vdots & & \vdots \\ r_1 & r_2 & \dots & r_k \\ \vdots & \vdots & & \vdots \end{pmatrix}$ is orthogonal matrix $\Gamma^T \Gamma = I_k$

$\Lambda = diag(\lambda_1, \lambda_2, \cdots \lambda_k)$, where $\lambda_i \neq 0 (i = 1, 2, \cdots k)$

We can find Eigenvalues (λ_i) and Eigenvectors (ρ_i) as follows.

$$A\Gamma = (\Gamma\Lambda\Gamma^T)\Gamma = \Gamma\Lambda \Leftrightarrow A\rho_i = \lambda_i \rho_i$$

, where $i = 1, 2, \cdots k$. Also, we can know $\mathcal{L}_{col}(A) = \mathcal{L}_{col}(\Gamma)$

3.13 Trace of matrix

Trace of $p \times p$ matrix B is defined as

$$tr(B) = \sum_{i=1}^{p} B_{ii}$$

Based on the definition of the trace matrix, we can swap the order of matrix multiplication as below.

$$tr(A_{p\times q}B_{q\times p}) = tr(B_{q\times p}A_{p\times q}) = \sum_i \sum_j A_{ij}B_{ji}$$

Suppose A is symmetric, we can calculate the spectral sum by using the trace property

$$tr(A) = tr(\Gamma\Lambda\Gamma^T) = tr(\Gamma^T\Gamma\Lambda) = tr(\Lambda) = \sum_i^k \lambda_i$$

Special case: Suppose $P_{n\times n}$ is a projection matrix into $k - dim$ linear space. $(tr(P) = k)$

1. k-dimensional linear space $= \mathcal{L}_{col}(X)$

$$P = X(X^TX)^{-1}X^T$$

QR decomposition on X, where $X = \Gamma_{p\times k}U$

$$P = \Gamma\Gamma^T \rightarrow tr(P) = tr(\Gamma\Gamma^T) = tr(\Gamma^T\Gamma) = tr(I_k) = k$$

2. P must have eigenvalues 0 or 1

3.14 Matrix square root

Let $S^2 = A$, where A is positive semi-definite.

(1) **Symmetric square root**: Suppose $A_{p\times p}$ is symmetric positive semi-definite and $rank(A) = k$, then the solution to $S^2 = A$ is

1. $S_{p\times p} = \Gamma\Lambda^{\frac{1}{2}}\Gamma^T$, where $\Lambda^{\frac{1}{2}} = diag(\sqrt{\lambda}_1, \sqrt{\lambda}_2, \cdots \sqrt{\lambda}_k)$ or
2. $S_{k\times p} = \Lambda^{\frac{1}{2}}\Gamma^T$

We have $S^TS = \Gamma\Lambda\Gamma^T = A$ in both case. We can see S and A share eigenvectors and also have close relationship as for eigenvalues $\frac{1}{2}$

(2) **Upper triangular square root**: $U^TU = A_{p\times p}$

- $S = \Lambda^{\frac{1}{2}}\Gamma^T = WU$, where W is orthogonal
- $A = S^TS = U^TW^TWU = U^TU$

3.15 Singular Value Decomposition (SVD)

While spectral decomposition can be used only for square matrix, SVD can be used for any matrix.

For any $A_{n\times p}$ with rank k, $A_{n\times p} = U_{n\times k}D_{k\times k}V^T_{k\times p}$, where $U_{n\times k}$, $V_{p\times k}$ are orthogonal ($\Leftrightarrow U^TU = I_k \quad V^TV = I_k$) and $D = diag(d_1, d_2 \cdots d_k)$, where $d_i > 0$. Furthermore, we can arrange them such that $d_1 \geq d_2 \geq \cdots d_k$. Here, $d_1 \cdots d_k$ are singular values, $U = (u_1 \cdot u_k)_{n\times k}$ components are left singular vectors, $V = (v_1 \cdot v_k)_{p\times k}$ components are right singular vectors.

U and V are not unique $(-U)$, $(-V)$. So you can put a sign in any column of U and V

$(AA^T)_{n\times n} = (UD^2U^T)_{n\times n}$ and $(A^TA)_{p\times p} = (VD^2V^T)_{p\times p}$. Then we have d_i^2 to be the eigenvalues of A^TA or AA^T, which also explains that eigenvectors of A^TA or AA^T should be columns of U and V.

Proof: Let AA^T is symmetric positive semi-definite. Spectral decomposition gives $AA^T = U\Lambda U^T$. Let $D = \Lambda^{\frac{1}{2}}$ and $V^T = D^{-1}U^TA$. Now we show U, D, V satisfy the requirements.

1. U is orthogonal
2. D is orthogonal
3. We have to check V is orthogonal.

$$V^TV = D^{-1}U^TAA^TUD^{-1} = I_k$$

4. $A = UDV^T = UDD^TU^TA = UU^TA = A$

UU^T is the projection matrix onto $\mathcal{L}_{col}(U) \Leftrightarrow \mathcal{L}_{col}(U) = \mathcal{L}_{col}(AA^T) = \mathcal{L}_{col}(A)$. Then, $UU^TA = A$.

■

ISYE 7405: Multivariate Data Analysis **Georgia Tech**

Chapter 5: Matrix Theory Review & Multivariate Normal Distribution

Lecturer: Shihao Yang

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

5.16 Matrix Theory Review (continued)

5.16.1 Pseudo-inverse X^- (continued)

Pseudo-inverse can be applied to solving linear systems

$$\eta = X\beta$$

where it assumes $\eta \in \mathcal{L}_{col}(X)$.
Claim: $\beta_0 = X^-\eta$ is a solution.
Verify: $X\beta_0 = XX^-\eta = \eta$ because XX^- is the projection matrix into $\mathcal{L}_{col}(X)$.
* Exercise (in the case $n < p$): in the case of multiple solutions, β_0 is the shortest solution in terms of the length $||\beta||$.
* What if $X\beta \approx y, y \notin \mathcal{L}_{col}(X), n \gg p$. (consider linear regression)

$$\widehat{\beta} = X^-y = (X^\top X)^{-1}X^\top y,$$

then $X\widehat{\beta} = XX^-y = \widehat{P}y = \widehat{y}$ is projection of y into $\mathcal{L}_{col}(X)$.

5.16.2 Orthogonal Representation of X

Suppose X has the following SVD decomposition

$$X = UDV^\top = (u_1, \cdots, u_k)\begin{pmatrix} d_1 & & \\ & \ddots & \\ & & d_k \end{pmatrix}\begin{pmatrix} v_1^\top \\ \vdots \\ v_k^\top \end{pmatrix},$$

where

$$U_{n\times k} = (u_1, u_2, \cdots, u_k), u_i \in \mathbb{R}^n,$$
$$V_{p\times k} = (v_1, v_2, \cdots, v_k), v_i \in \mathbb{R}^p,$$
$$D = \text{diag}(d_1, d_2, \cdots, d_k), d_1 \geq d_2 \geq \cdots \geq d_k > 0,$$
$$k = \text{rank}(X).$$

Then

$$X = \sum_{j=1}^{k} u_j d_j v_j^\top.$$

Let $B_j = u_j v_j^\top$. B_j is $(n \times p)$ matrix, $\text{rank}(B_j) = 1$. We have

$$X = \sum_{j=1}^{k} d_j B_j.$$

- B_j are orthogonal to each other, and $||B_j|| = 1$.
  Proof: $\forall i \neq j, < B_i, B_j >= 0$

$$\Leftrightarrow \sum_{l,m} (B_i)_{lm}(B_j)_{lm} = \mathrm{tr}(B_i^\top B_j) = \mathrm{tr}(v_i u_i^\top u_j v_j^\top) = \mathrm{tr}(u_i^\top u_j v_j^\top v_i) = \begin{cases} 0, \text{if } i \neq j, \\ 1, \text{if } i = j. \end{cases}$$

■

Define $\widehat{X}(J) = \sum_{j=1}^{J} d_j B_j, J \leq k$, $\mathrm{rank}(\widehat{X}(J)) = J$. $\widehat{X}(J)$ is an approximation to X with the following properties:

1. $||X - \widehat{X}(J)||^2 = \sum_{j=J+1}^{k} d_j^2$;
2. $\frac{||X-\widehat{X}(J)||^2}{||X||^2} = \frac{\sum_{j=J+1}^{k} d_j^2}{\sum_{j=1}^{k} d_j^2}$ (percentage).

Example: $X_{n\times p}, k = p$ but $\widehat{X}(p-1)$ approximates X very well, then X has multi-collinearity.

5.16.3 Block matrix inversion

Consider block matrix

$$A_{(p+q)\times(p+q)} = \begin{pmatrix} A_{11_{p\times p}} & A_{12_{p\times q}} \\ A_{21_{q\times p}} & A_{22_{q\times q}} \end{pmatrix},$$

where A_{11}, A_{22} are invertible, then we have

$$A^{-1} = \begin{pmatrix} A_{11}^{-1} + A_{11}^{-1}A_{12}(A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1}A_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}(A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1} \\ -A_{22}^{-1}A_{21}(A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1} & (A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1} \end{pmatrix}.$$

5.16.3.1 Woodbury's formula

Suppose $A_{p\times p}, B_{q\times q}$ are both non-singular, then

$$(A_{p\times p} + U_{p\times q}B_{q\times q}V_{q\times p})^{-1} = A^{-1} - A^{-1}U(B^{-1} + VA^{-1}U)^{-1}VA^{-1}.$$

Proof: Assume we are trying compute

$$\begin{pmatrix} A & U \\ V & B \end{pmatrix}^{-1}.$$

We can do a column elimination first and then a row elimination to make the matrix diagonal. Specifically, by column operation:

$$\begin{pmatrix} I & 0 \\ -VA^{-1} & I \end{pmatrix}\begin{pmatrix} A & U \\ V & B \end{pmatrix} = \begin{pmatrix} A & U \\ 0 & B - VA^{-1}U \end{pmatrix}.$$

By row operation:

$$\begin{pmatrix} A & U \\ V & B \end{pmatrix}\begin{pmatrix} I & -A^{-1}U \\ 0 & I \end{pmatrix} = \begin{pmatrix} A & 0 \\ V & B - VA^{-1}U \end{pmatrix}.$$

Then it follows

$$\begin{pmatrix} I & 0 \\ -VA^{-1} & I \end{pmatrix}\begin{pmatrix} A & U \\ V & B \end{pmatrix}\begin{pmatrix} I & -A^{-1}U \\ 0 & I \end{pmatrix} = \begin{pmatrix} A & 0 \\ 0 & B - VA^{-1}U \end{pmatrix}$$

$$\Longrightarrow \begin{pmatrix} A & U \\ V & B \end{pmatrix}^{-1} = \begin{pmatrix} A^{-1} + A^{-1}U(B - VA^{-1}U)^{-1}VA^{-1} & -A^{-1}U(B - VA^{-1}U)^{-1} \\ -(B - VA^{-1}U)^{-1}VA^{-1} & (B - VA^{-1}U)^{-1} \end{pmatrix}. \quad (1)$$

On the other hand, we can also do row elimination first and then column elimination. It follows that

$$\begin{pmatrix} A & U \\ V & B \end{pmatrix} = \begin{pmatrix} I & UB^{-1} \\ 0 & I \end{pmatrix}\begin{pmatrix} A - UB^{-1}V & 0 \\ 0 & B \end{pmatrix}\begin{pmatrix} I & 0 \\ B^{-1}V & 0 \end{pmatrix}$$

$$\Longrightarrow \begin{pmatrix} A & U \\ V & B \end{pmatrix}^{-1} = \begin{pmatrix} (A-UB^{-1}V)^{-1} & -(A-UB^{-1}V)^{-1}UB^{-1} \\ -B^{-1}V(A-UB^{-1}V)^{-1} & B^{-1}+B^{-1}V(A-UB^{-1}V)^{-1}VB^{-1} \end{pmatrix}. \tag{2}$$

Note that the right-hand-side of (1) and (2) should be equal. The equality of left-upper block gives Woodbury's formula. ■

For special case when $q = 1$:

$$(A + bUV)^{-1} = A^{-1} - \frac{A^{-1}UV^\top A^{-1}}{b + V^\top A^{-1}U}.$$

5.16.3.2 Statistical application of block matrix inversion

Consider

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}, \; \Sigma = \mathrm{Var}(X) = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

Using block matrix inversion, we have

$$(\Sigma)^{-1}_{22} = (\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1} = (\Sigma_{22}^{\perp})^{-1},$$

where the right-hand-side represents the conditional variance of X_2 conditioning on X_1.
* Exercise: it is known that $\Sigma_{ij} = 0 \Leftrightarrow \mathrm{Cov}(X_i, X_j) = 0$. Show that $(\Sigma^{-1})_{ij} = 0 \Leftrightarrow X_i, X_j$ have partial correlation of 0 after projecting on $X_1, ..., X_{i-1}, X_{i+1}, ..., X_{j-1}, X_{j+1}, ..., X_p$.

5.17 Multivariate Normal (Gaussian) Distribution

5.17.1 Standard normal

Random vector Z following standard normal distribution is denoted as:

$$Z \sim N_p(0, I_p) \Leftrightarrow Z = (z_1, .., z_p)^\top, z_i \overset{iid}{\sim} N(0, 1).$$

Density function is:

$$f_Z(\underline{x}) = (2\pi)^{-\frac{p}{2}} \exp\{-\frac{1}{2}||\underline{x}||^2\}.$$

Characteristic function is:

$$\Psi_Z(t) = E(e^{it^\top Z}) = e^{-\frac{1}{2}||t||^2} = e^{-\frac{1}{2}t^\top t}.$$

5.17.2 General normal (Gaussian)

For any $\mu \in \mathbb{R}^p$, and symmetric positive semi-definite matrix $\Sigma_{p\times p}$, a random vector following general normal distribution can be constructed as

$$X = \mu + \Sigma^{\frac{1}{2}} Z \sim N_p(\mu, \Sigma).$$

The following are some properties of general normal distribution:

1. $E(X) = \mu, \mathrm{Cov}(X, X) = \Sigma$.

2. If $\mathrm{rank}(\Sigma) = p$ (full rank), then Σ is positive definite, and density exists:

$$f_X(x) = f_Z(x)J(Z \to X) = (2\pi)^{-\frac{p}{2}} \exp\{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\}|\det(\Sigma)|^{-\frac{1}{2}}.$$

3. Characteristic function:

$$\Psi_X(x) = E[e^{it^\top X}] = \exp\{it^\top \mu - \frac{1}{2}t^\top \Sigma t\}.$$

4. For any $\mu \in \mathbb{R}^p, \Sigma_{p\times p} \geq 0$ being symmetric, there is a unique $N_p(\mu, \Sigma)$ distribution.

5. $X \in \mu + \mathcal{L}_{col}(\Sigma)$ with probability 1, since $\mathcal{L}_{col}(\Sigma^{-\frac{1}{2}}) = \mathcal{L}_{col}(\Sigma)$.

6. Let $Y = r_q + A_{q\times p}X_{p\times 1}$. Then $Y \sim N_q(r + A\mu, A\Sigma A^\top)$.

7. Skewness is always 0. For $X = (x_1, ..., x_p) \sim N_p(\mu, \Sigma)$,

$$E((x_i - \mu_i)(x_j - \mu_j)(x_k - \mu_k)) = 0, \forall i, j, k.$$

   Kurtosis is

$$E((x_1 - \mu_1)(x_2 - \mu_2)(x_3 - \mu_3)(x_4 - \mu_4)) = \sigma_{12}\sigma_{34} + \sigma_{13}\sigma_{24} + \sigma_{14}\sigma_{23}.$$

8. Take $\mu = 0, \Sigma^{\frac{1}{2}} = \Gamma$ as a $p \times p$ orthogonal matrix, then $X = \Gamma Z \sim N_p(0, I_p)$.
   $\Longrightarrow$ Standard normal is rotationally (spherically) invariant.

9. Let $U = Z/||Z||$. U is uniformly distributed on the surface of a p-dimensional sphere.

ISYE 7405: Multivariate Data Analysis **Georgia Tech**

Chapter 6: Multivariate Gaussian and Matrix Manipulation

Lecturer: Shihao Yang

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

6.18 Multivariate Normal Distribution

Properties (cont.)

9. $U = Z/\|Z\|$ is uniform on a sphere. This is a way to simulate a uniform distribution on a sphere.
10. $\begin{pmatrix} X_p \\ Y_q \end{pmatrix} \sim N_{p+q}\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{pmatrix}\right) \implies X \sim N_p(\mu_X, \Sigma_{XX})$
    A marginal distribution of a sub-vector is still multivariate Gaussian.
11. $\Sigma_{XY} = 0 \iff X \perp\!\!\!\perp Y$
    Proof: $\psi_{(X,Y)}(t,s) = \psi_X(t)\psi_Y(s)$
12. Conditional distribution, $Y|X \sim N_q(\mu_Y + \Sigma_{YX}\Sigma_{XX}^{-1}(X-\mu_X), \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY})$
    Proof: $Y = \hat{Y} + Y^{\perp}$ (projection on $\mathcal{L}(X)$)
    $\psi_{Y|X}(t) = E[e^{it^TY}|X] = E[e^{it^T(\hat{Y}+Y^{\perp})}|X] = e^{it^T\hat{Y}}E[e^{it^TY^{\perp}}|X]$
    $\text{Cov}(Y^{\perp},X) = 0 \iff Y^{\perp} \perp\!\!\!\perp X \implies e^{it^T\hat{Y}}E[e^{it^TY^{\perp}}|X] = e^{it^T\hat{Y}}E[e^{it^TY^{\perp}}]$
13. $E[Y|X]$ is the best linear predictor of Y in terms of X and the best predictor of Y in terms of X. This holds for any conditional expectation. In the Gaussian case, the best linear prediction is the best prediction.
    Proof: Recall $Y = \hat{Y} + Y^{\perp}$. The best linear predictor is $\hat{Y} = AX$ such that $\text{Var}(Y^{\perp})$ is minimized. The best predictor is $\hat{Y} = f(X)$ such that $\text{Var}(Y^{\perp})$ is minimized.

    $$E[Y|X] = E[\hat{Y}+Y^{\perp}|X] = f(X)+E[Y^{\perp}|X] \wedge Var(Y^{\perp}) = E[Var(Y^{\perp}|X)]+Var(E[Y^{\perp}|X]) \implies f(X) = E[Y|X]$$

Repeated Sampling

Now, instead of looking at just one sample, lets look at n independent and identically distributed p-dimensional random vectors:

$$X_1, ..., X_n \sim^{i.i.d} N(\mu, \Sigma).$$

We assume Σ is full rank, i.e. it is positive definite. we can collect all the random vectors as columns, and we get a random matrix:

$$\mathbb{X}_{p\times n} = \begin{pmatrix} \vdots & \vdots & & \vdots \\ X_1 & X_2 & \dots & X_n \\ \vdots & \vdots & & \vdots \end{pmatrix}.$$

We can find the joint distribution using independence:

$$f_{(\mu,\Sigma)}(\mathbb{X}) = \frac{1}{((2\pi)^p|\Sigma|)^{\frac{n}{2}}}\exp(-\frac{1}{2}\sum_{i=1}^{n}(X_i-\mu)^T\Sigma^{-1}(X_i-\mu)).$$

Let us simplify the exponent of e using sample mean $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$,

$$\sum_{i=1}^{n}(X_i-\mu)^T\Sigma^{-1}(X_i-\mu)$$

$$= \sum_{i=1}^{n}(X_i-\bar{X}+\bar{X}-\mu)^T\Sigma^{-1}(X_i-\bar{X}+\bar{X}-\mu)$$

$$= \sum_{i=1}^{n}(X_i-\bar{X})^T\Sigma^{-1}(X_i-\bar{X}) + n(\bar{X}-\mu)^T\Sigma^{-1}(\bar{X}-\mu)$$

$$= \mathrm{Tr}(\sum_{i=1}^{n}(X_i-\bar{X})^T\Sigma^{-1}(X_i-\bar{X})) + n(\bar{X}-\mu)^T\Sigma^{-1}(\bar{X}-\mu)$$

$$= \mathrm{Tr}(\Sigma^{-1}\sum_{i=1}^{n}(X_i-\bar{X})(X_i-\bar{X})^T) + n(\bar{X}-\mu)^T\Sigma^{-1}(\bar{X}-\mu)$$

$$= \mathrm{Tr}(\Sigma^{-1}S) + n(\bar{X}-\mu)^T\Sigma^{-1}(\bar{X}-\mu),$$

where $S = \sum_{i=1}^{n}(X_i-\bar{X})(X_i-\bar{X})^T$ is the sample covariance matrix. Thus we can write the joint distribution in terms of the sample mean $\bar{X}$ and sample covariance matrix S,

$$f_{(\mu,\Sigma)}(\mathbb{X}) = \frac{1}{((2\pi)^p|\Sigma|)^{\frac{n}{2}}}\exp(\mathrm{Tr}(\Sigma^{-1}S) + n(\bar{X}-\mu)^T\Sigma^{-1}(\bar{X}-\mu)).$$

$(\bar{X}, S)$ is a sufficient statistic (complete sufficient).
Multivariate Gaussian distribution is a special case of the exponential family, whose distribution is of the form

$$f(\tilde{x},\theta) = \exp(\sum_{i=1}^{k}\eta_i(\theta)T_i(\tilde{x}) - \psi(\theta))h(\tilde{x})$$

where $\tilde{x}$ can be either a vector or a matrix, and $\{T_i(k)\}$ are complete sufficient.

6.19 Matrix Manipulation

Vectorization

$$A = \begin{pmatrix} \vdots & \vdots & & \vdots \\ a_1 & a_2 & \dots & a_p \\ \vdots & \vdots & & \vdots \end{pmatrix}_{n\times p} \implies vec(A) = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix}_{np\times 1}$$

With this definition, we can define other things like matrix norm, inner product, and orthogonality between matrices.

$$\|A\|^2 = \|vec(A)\|^2 = tr(A^TA)$$

$$\langle A, B\rangle = \langle vec(A), vec(B)\rangle = tr(A^TB)$$

This leads us to a natural way to manipulate random matrix X_{pxn}, where $X_i \overset{iid}{\sim} N_p(\mu,\Sigma)$ for all $i \in \{1,...,n\}$.

$$X_{p\times n} = \begin{pmatrix} \vdots & \vdots & & \vdots \\ X_1 & X_2 & \dots & X_n \\ \vdots & \vdots & & \vdots \end{pmatrix} \implies vec(X) = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}_{pn\times 1}$$

$vec(X) \sim N_{pn}(\mu^*, \Sigma^*)$ such that:

$$\mu^* = \begin{pmatrix} \mu \\ \mu \\ \vdots \\ \mu \end{pmatrix}_{(pn)\times 1}$$

$$\Sigma^* = \begin{pmatrix} \Sigma & Cov(X_1, X_2) = 0 & \dots & Cov(X_1, X_3) = 0 & Cov(X_1, X_n) = 0 \\ Cov(X_2, X_1) = 0 & \Sigma & \ddots & Cov(X_2, X_3) = 0 & Cov(X_2, X_n) = 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ Cov(X_{n-1}, X_1) = 0 & Cov(X_{n-1}, X_2) = 0 & \ddots & \Sigma & Cov(X_{n-1}, X_n) = 0 \\ Cov(X_n, X_1) = 0 & Cov(X_n, X_2) = 0 & \dots & Cov(X_n, X_{n-1}) = 0 & \Sigma \end{pmatrix}$$

$$= diag(\Sigma, \Sigma, ..., \Sigma) = I_n \otimes \Sigma$$

Kronecker Product

The definition of a Kronecker product is as follows:

$$A_{pxq} \otimes B_{rxs} = \begin{pmatrix} a_{11}B & a_{12}B & \dots & a_{1q}B \\ a_{21}B & a_{22}B & \dots & a_{2q}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1}B & a_{p2}B & \dots & a_{pq}B \end{pmatrix}_{(pr)\times(qs)}$$

The Kronecker product has the following properties:

1. $(A \otimes B) \otimes C = A \otimes (B \otimes C)$
2. $(A \otimes B)^T = (A^T \otimes B^T)$
3. $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$ if A, B invertible
4. $(A \otimes B)(C \otimes D) = (AC \otimes BD)$
5. $tr(A \otimes B) = tr(A)tr(B)$
6. $vec(AXB) = (B^T \otimes A)vec(X)$

Returning to the repeated sampling setup, we see when

$$X_1, ..., X_n \sim^{i.i.d} N(\mu, \Sigma),$$

and

$$\mathbb{X}_{p\times n} = \begin{pmatrix} \vdots & \vdots & & \vdots \\ X_1 & X_2 & \dots & X_n \\ \vdots & \vdots & & \vdots \end{pmatrix};$$

and $\mu \cdot \mathbf{1}^T = (\mu \;\; \mu \cdots \mu)$, we have

$$\mathbb{X}_{p\times n} \sim N_{p\times n}(\mu \cdot \mathbf{1}^T, I_n \otimes \Sigma).$$

Generalization of (i.i.d.) Normal Data Matrix

We now consider

$$\mathbb{X}_{p\times n} = \begin{pmatrix} \vdots & \vdots & & \vdots \\ X_1 & X_2 & \dots & X_n \\ \vdots & \vdots & & \vdots \end{pmatrix};$$

where $X_1, ..., X_n \sim N(\mu, \Sigma)$, i.e. they are identically distributed but **not** independent. Under the independence assumption, we have

$$\text{Cov}(\mathbb{X}_{i_1 j_1}, \mathbb{X}_{i_2 j_2}) = \Sigma_{i_1 i_2} I_{j_1 j_2}.$$

In the dependent case we will have

$$\text{Cov}(\mathbb{X}_{i_1 j_1}, \mathbb{X}_{i_2 j_2}) = \Sigma_{i_1 i_2} \Delta_{j_1 j_2},$$

for some symmetric positive definite matrix Δ.

- $Cov(\mathbb{X}_{i_1,j_1}, \mathbb{X}_{i_2,j_2}) = \Sigma_{i_1,i_2} \triangle_{j_1,j_2}$ for $\Sigma, \triangle$ symmetric and $\Sigma > 0, \triangle > 0$.
- $Cov(\mathbb{X}_{\bullet,j}) = Cov((\mathbb{X}_{1,j}, \cdots, \mathbb{X}_{p,j})^T) = \Sigma \cdot \triangle_{j,j}$
- $Cov(\mathbb{X}_{i,\bullet}) = Cov((\mathbb{X}_{i,1}, \cdots, \mathbb{X}_{i,n})^T) = \Sigma_{i,i} \cdot \triangle$.

ISYE 7405: Multivariate Data Analysis **Georgia Tech**

Chapter 7: Multivariate Gaussian Distribution

Lecturer: Shihao Yang

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

7.20 Repeated Sampling (continued)

Now we consider $\mathbb{X} = (X_1, X_2, \cdots, X_n)$ with $X_1, X_2, \cdots, X_n$ are identically distributed but not independent. E.g.: stationary time series $X_1, X_2, \cdots, X_t$ with each vector marginally following same distribution, vector Auto Regressive (VAR) Model.

Then assume $\mathbb{X}_{p\times n} \sim N_{p\times n}(\mu, \triangle \otimes \Sigma)$, where $\mu_{i,j} = E(\mathbb{X}_{i,j})$ is a matrix, $\text{Cov}(\text{vec}(\mathbb{X})) = \triangle \otimes \Sigma$

7.20.1 Properties of Generalised Gaussian Matrix

- $\mathbb{X}^T \sim N_{n\times p}(\mu^T, \Sigma \otimes \triangle)$.
- If $\mathbb{X}_{p\times n} \sim N_{p\times n}(\mu, \triangle \otimes \Sigma)$, then $A_{q\times p}\mathbb{X}_{p\times n}B_{n\times m} \sim N(A\mu B, (B^T \triangle B) \otimes (A\Sigma A^T))$.

  Proof: $vec(A\mathbb{X}B) = (B^T \otimes A) \cdot vec(\mathbb{X})$, then

$$\begin{aligned} Cov(vec(A\mathbb{X}B)) &= Cov((B^T \otimes A) \cdot vec(\mathbb{X})) \\ &= (B^T \otimes A)Cov(vec(\mathbb{X}))(B^T \otimes A)^T \\ &= (B^T \otimes A)(\triangle \otimes \Sigma)(B \otimes A^T) \\ &= (B^T \triangle B) \otimes (A\Sigma A^T). \end{aligned} \tag{7.1}$$

  to remember intuition, think special case with $n = p = 1$. ■

- Suppose $\Sigma, \triangle$ have spectral decomposition

$$\Sigma = \Gamma_1 D_1 \Gamma_1^T, \triangle = \Gamma_2 D_2 \Gamma_2^T \tag{7.2}$$

  with $\text{rank}(\Sigma) = k_1, \text{rank}(\triangle) = k_2$. Then we can take

$$\mathbb{X} = \mu + \Gamma_1 D_1^{1/2} \mathbb{Z} D_2^{1/2} \Gamma_2^T \tag{7.3}$$

  then $\mathbb{Z}_{k_1\times k_2} \sim N(0, I_{k_2} \otimes I_{k_1})$(iid standard normal in all entries of matrix), which is equivalent to $\mathbb{X} \sim N_{p\times n}(\mu, \triangle \otimes \Sigma)$.

Consider $\mathbb{X}_{p\times n} \sim N_{p\times n}(0, I_n \otimes \Sigma) \iff \mathbb{X} = (X_1, X_2, \cdots, X_n)$ with $X_1, X_2, \cdots, X_n$ $N_p(0, \Sigma)$ follows iid.

$$\mathbb{X} = \begin{pmatrix} \vdots & \vdots & & \vdots \\ X_1 & X_2 & \cdots & X_n \\ \vdots & \vdots & & \vdots \end{pmatrix}$$

Let $V = \mathbb{X}_{p\times n}^T = (v_1, v_2, \cdots, v_p)$ and we have $v_i \sim N_n(0, \sigma_{ii} I_n)$.

$$V = \mathbb{X}^{\mathbb{T}} = \begin{pmatrix} \vdots & \vdots & & \vdots \\ v_1 & v_2 & \cdots & v_p \\ \vdots & \vdots & & \vdots \end{pmatrix}$$

$$\mathbb{X} = \begin{pmatrix} \cdots v_1^T \cdots \\ \cdots v_2^T \cdots \\ \vdots \\ \cdots v_p^T \cdots \end{pmatrix}$$

where v_p^T can be taken as individual features like height, weight, age, etc.
Rotational in-variance of mean 0, Normal data matrix. I_n is a full orthogonal matrix Moreover, let $\Gamma_{n\times n}$ be a full orthogonal basis, then

$$\Gamma v_i \sim N_n(0, \sigma_{ii}\Gamma I_n \Gamma^T) = N_n(0, \sigma_{ii} I_n) \sim v_i \tag{7.4}$$

and

$$X\Gamma^T = (\Gamma X^T)^T = (\Gamma V)^T \sim N(0, \Gamma\Gamma^T \otimes \Sigma) = N(0, I_n \otimes \Sigma) \tag{7.5}$$

7.21 Spherical Symmetry and t-test

Recall the univariate t-test -

$$X_1, X_2, \cdots, X_n \sim N(\mu, \sigma^2) \text{ follows iid}, H_0 : \mu = 0; H_1 : \mu \neq 0 \tag{7.6}$$

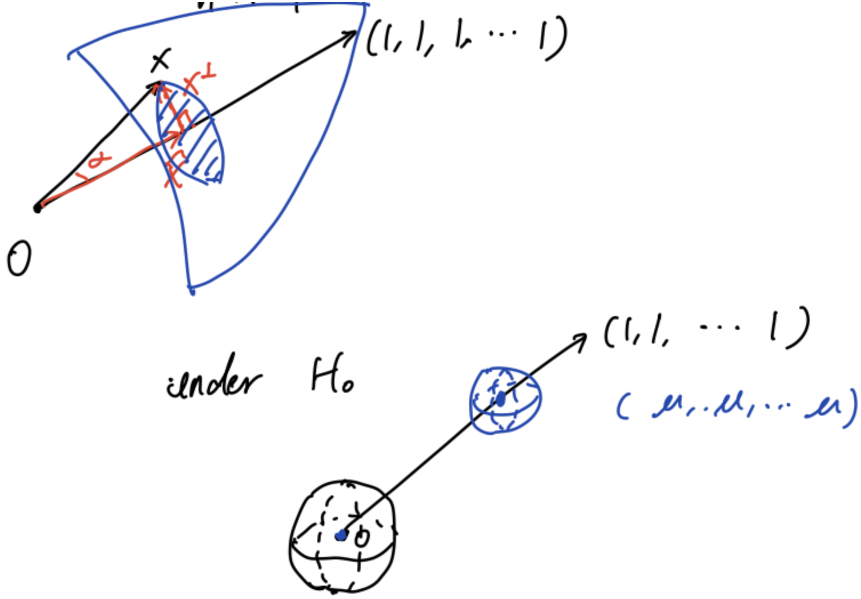
$$t = \frac{\sqrt{n}\overline{X}}{\sqrt{\sum_i (X_i - \overline{X})^2/(n-1)}} \tag{7.7}$$

In geometry, $1_{n\times 1} = (1, 1, \cdots, 1)^T$. Using projection onto $1_{n\times 1}$.
We have $\hat{X} = P_1 X = (\overline{X}, \overline{X}, \cdots, \overline{X})_{n\times 1}$.
Moreover, $X^\perp = X - \hat{X} = (X_1 - \hat{X}, X_2 - \hat{X}, \cdots, X_n - \hat{X}) = (I - P_1)X$, and

$$t = sgn(\hat{X}) \cdot \sqrt{n-1} \cdot \frac{\left\|\hat{X}\right\|}{\|X^\perp\|} \tag{7.8}$$

where $\left\|\hat{X}\right\| = \|(\overline{X}, \overline{X}, \cdots, \overline{X})\| = \sqrt{n(\overline{X}^2)} = \sqrt{n}|\overline{X}|$
P_1 (projection matrix onto 1^T) $= 1(1^T 1)^{-1} 1^T = \frac{1}{n} 1 \cdot 1^T$

Figure 7.7: Geometric intuition behind t-test



ISYE 7405: Multivariate Data Analysis **Georgia Tech**

Chapter 8: Enter the title

Lecturer: Shihao Yang

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

8.22 Centering matrix

$$\mathbb{X} = \begin{pmatrix} \vdots & \vdots & & \vdots \\ X_1 & X_2 & \cdots & X_n \\ \vdots & \vdots & & \vdots \end{pmatrix}$$

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \bar{X} = \frac{1}{n}\mathbb{X} I_n, I_n = \begin{pmatrix} I \\ I \\ I \\ \vdots \\ I \end{pmatrix}_{n\times 1}$$

$$\begin{aligned} Y =& \left(X_1 - \bar{X}, X_2 - \bar{X}, \cdots, X_n - \bar{X}\right) \\ =& \mathbb{X} - \left(\bar{X}, \bar{X}, \bar{X}, \cdots, \bar{X}\right) = \mathbb{X} - \bar{X}\ I_n^T = \mathbb{X} - \frac{1}{n}\left(\mathbb{X} I_n\right)\ I_n^T \\ =& \mathbb{X}\left(I_n - \frac{1}{n} I_n I_n^T\right) \end{aligned}$$

$$\begin{aligned}\hat{P}_I =& \frac{1}{n} I_n I_{_n}^T : \quad \text{projection matrix onto } \mathcal{L}_{col}^{\perp}\left(I_n\right) \\ =& I_n \left(I_{_n}^T I_n\right)^{-1} \ I_{_n}^T\end{aligned}$$

$$H_n = I_n - \frac{1}{n} I_n I_{_n}^T = \hat{P}_{_I}^{\perp} : \text{projection matrix onto } \ \mathcal{L}_{col}^{\perp}\left(I_n\right)$$

$$J_n = I_n \ I_{_n}^T = \begin{pmatrix} 1\ 1\ 1 \cdots 1 \\ 1\ 1\ 1 \cdots 1 \\ 1\ 1\ 1 \cdots 1 \\ 1\ 1\ 1 \cdots 1 \end{pmatrix}$$

Therefore ,

$$Y = XH_n$$

$$\mathbb{X}_{p\times n} \sim N\left(\mu \ I_{_n}^T, I_n \otimes \sum\right) \quad \mu \ I_{_n}^T = \left(\overset{1}{\underset{1}{\mu}}, \overset{1}{\underset{1}{\mu}}, \cdots \overset{1}{\underset{1}{\mu}}\right)$$

$$Y_{p\times n} \sim \mathbb{X}H_n \qquad \Updownarrow x_i \overset{iid}{\sim} N_p\left(\mu, \sum\right)$$

8.23 Properties of $Y_{p\times n}$

8.23.1 Distribution

$$Y_{p\times n} \sim N_{p\times n}\left(0, \underline{H_n} \otimes \sum\right)$$

$$\updownarrow$$

$$H_{_n}^T I_n \, H_n = H_n$$

$$\overline{X} \sim N_p\left(\mu, \frac{1}{\underline{n}} \sum\right)$$

$$\frac{1}{n^2}\left(I_n^T I_n I_n\right) \overset{\updownarrow}{\otimes} \sum$$

$$\updownarrow$$

$$\frac{1}{n} \sum$$

8.23.2 $Y \coprod \overline{X}$ only need uncorrelated

$$\begin{aligned}\left(Y, \overline{X}\right) =& \mathbb{X}\left(H_n, \frac{1}{n} I_n\right) \\ \sim& N\left[\quad , \left(H_n, \frac{1}{n} I_n\right)^T I_n \left(H_n, \frac{1}{n} I_n\right) \otimes \sum\right]\end{aligned}$$

$$Y \coprod \overline{X} \Leftrightarrow H_n \cdot \frac{1}{n} I_n = 0$$

$$I_n \overset{\updownarrow}{\cdot} \frac{1}{n} I_n I_{_n}^T$$

8.23.3 YY^T is independent of $\overline{X}$

$$\begin{aligned}S =& \sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(X_i - \overline{X}\right)^T \\ =& YY^T \coprod \overline{X}\end{aligned}$$

8.23.4 Distribution of YY^T

$I_n/\sqrt{n}$ vector, wren of I

$\exists I_{1,n\times(n-1)}$ is orthogonal basic of $\mathcal{L}_{ol}^{\perp}\left(I_n\right)$

Then $\Gamma = \left(\Gamma_1, I_n/\sqrt{n}\ \right)$ is full orthogonal basis

H_n projection matrix onto $\mathcal{L}_{ol}^{\perp}\left(I_n\right) = \Gamma_1\Gamma_1^T$

$Z = \mathbb{X}\Gamma_1 \sim N_{p\times(n-1)}\left(0, I_{n-1}\otimes\sum\right)$

$S = YY^T = YH_nY^T = Y\Gamma_1\Gamma_1^TY^T = ZZ^T$

$ZZ^T = \mathbb{X}\Gamma_1\Gamma^T\mathbb{X}^T = \mathbb{X}H_n\left(H_n^T\mathbb{X}^T\right) = YY^T$

$\Rightarrow YY^T$ has the same distribution of ZZ^T where $Z\sim N_{px(n-1)}\left(0, I_{n-1}\otimes\sum\right)$

$\Leftrightarrow YY^T$ has the same distribution of $\sum\limits_{i=1}^{n-1} Z_iZ_i^T$ where $Z_1, Z_2, \cdots Z_{n-1}\overset{idd}{\sim}N_p\left(\sum\right)$

8.24 Wishart distribution

Consider $\mathbb{X}_{p\times n}\sim N\left(0, I_n\otimes\sum\right)$

Def. For $\mathbb{X}_{p\times n}\sim N\left(0, I_n\otimes\sum\right)$,$S = XX^T$ is said to have wishart distribution with scale matrix $\sum$, and degree of freedom n

$S\sim W_p\left(\sum, n\right)$

Pd.f : $f^S\left(S\right) = f^{\mathbb{X}}\left(X\right)J\left(\mathbb{X}\rightarrow S\right)$

*Jacobian to Triangular coordinate $\Rightarrow$ Barlett decomposition

$J\left(\mathbb{X}\rightarrow S\right) = J\left(\mathbb{X}\rightarrow T\right)J\left(T\rightarrow S\right)$

$$\left.\begin{aligned}&T:\left(\nabla\right)V_{n\times p} = X^T, V = W_{n\times p}T_{p\times p}\ \left(QR\ decomp\right)\\&S = XX^T = V^TV = T^TT\\&Then\ \ J\left(\mathbb{X}\rightarrow S\right) = J\left(\mathbb{X}\rightarrow T\right)J\left(J\rightarrow S\right)\end{aligned}\right\} idea\ of\ proof$$

Lemma 1 The integral Jacobian of $\mathbb{X}_{p\times n}\rightarrow T_{p\times p}$

$J\left(\mathbb{X}\rightarrow T\right) = C_1\cdot\prod\limits_{i=1}^{p} t_{ii}^{n-i}$

where $C_1 = 2^p\pi^{\frac{np}{2}-\frac{p^2}{4}+\frac{p}{4}}/\prod\limits_{j=1}^{p}\Gamma\left(\frac{n-j+1}{2}\right)$

Corollary : $f^T\left(t\right) = f^{\mathbb{X}}\left(x\right)\cdot J\left(\mathbb{X}\rightarrow T\right)$

Hint: $\mathbb{X}_{p\times n}\sim N_{p\times n}\left(0, I_n\otimes I_p\right)$,$x : j\overset{iid}{\sim}N\left(0,1\right)$

ISYE 7405: Multivariate Data Analysis **Georgia Tech**

Chapter 9: Properties of Wishart Distribution

Lecturer: Shihao Yang

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

9.25 Integral Jacobian

$$X_{p\times n} \sim N(0, I_n \otimes \Sigma), \quad S = XX^T, \quad V = X^T = (v_1, v_2, ..., v_p)$$

Definition: $S\ W_p(\Sigma, \Lambda,\ V = WT$ (QR decomposition). $S = T^T T$

Lemma 1: Integral Jacobian of $X_{p\times n} \to T_{p\times p}$ is:

$$J(X \to T) = c_1 \prod_{i=1}^{p} t_{ii}^{n-i}, \quad \text{where} \quad c_1 = 2^p \frac{\pi^{\frac{np}{2} - \frac{p^2}{4} + \frac{p}{4}}}{\prod_{j=1}^{p} \Gamma(\frac{n-j+1}{2})} \tag{9.9}$$

Proof: Idea $X_{p\times n} \sim N(0, I_n \otimes I_p)$. Then if I find out the distribution of T, where $X^T = U = WT$. Then the $J(X \to T) = \frac{f^X(x)}{f^T(t)}$, where $f^X(x) = (2\pi)^{-\frac{np}{2}} \exp(-\frac{1}{2}\sum_{ij} x_{ij}^2)$.

$$\begin{aligned}\sum_{ij} x_{ij}^2 &= tr(XX^T) \\ &= tr(T^T W^T W T) \\ &= tr(T^T T) \\ &= \sum_{ij} t_{ij}^2\end{aligned}$$

Thus, we can write:

$$f^X(x) = (2\pi)^{-\frac{np}{2}} \exp(-\frac{1}{2}\sum_{ij} t_{ij}^2)$$

Next find out distribution of t_{ij}

$$X^T = V = (v_1, v_2, ..., v_p) = WT$$

$$v_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{in} \end{bmatrix}, \quad \text{each } N(0,1) \quad \text{iid}$$

By Gram-Schmidt orthogonalization:

$$\begin{aligned} v_1 &= t_{11} w_1 \\ v_2 &= t_{12} w_1 + t_{22} w_2 \\ v_3 &= t_{13} w_1 + t_{23} w_2 + t_{33} w_3 \\ &\vdots \\ v_p &= t_{1p} w_1 + t_{2p} w_2 + \cdots + t_{pp} w_p \end{aligned}$$

where $t_{11} = ||v_1||$, t_{12} =Projection of V_2 to W_1,..., etc.

Construction of W and T:

$$\begin{aligned}
&t_{11}^2 = ||v_1||^2 \sim \chi_n^2, \qquad w_1 = \frac{v_1}{||v_1||}\\
&t_{12} =< v_2, w_1 >\sim N(0,1), \quad w_2 = \text{Projection of } V_2 \text{ onto } L^{\perp}(V_1) = \frac{v_2^{\perp}}{||v_2^{\perp}||}\\
&t_{22}^2 = ||v_2^{\perp}||^2 \sim \chi_{n-1}^2,\\
&t_{13} =< v_3, w_1 >\sim N(0,1), \quad w_3 = \text{Projection of } V_3 \text{ onto } L^{\perp}(V_1, V_2) = \frac{v_3^{\perp}}{||v_3^{\perp}||}\\
&t_{23} =< v_3, w_2 >\sim N(0,1),\\
&t_{33}^2 = ||v_3^{\perp}||^2 \sim \chi_{n-2}^2,\\
&\quad\vdots\\
&t_{ij} \sim N(0,1)\\
&t_{ii} \sim \chi_{n-i+1}\\
&t_{ii}^2 \sim \chi_{n-i+1}^2
\end{aligned}$$

We have that:

$$\begin{aligned}
&t_{22}^2 = ||v_2||^2 = ||P_2^{\perp} V_2||^2 = V_2^T P_2^{\perp} V_2 = ||(\Gamma_2^{\perp})^T V_2||_2^2\\
&P_2^{\perp} = (\Gamma_2^{\perp})(\Gamma^{\perp})^T, \quad \Gamma_2^{\perp} = \text{is the orthogonal basis for } L_{col}^{\perp}(V_1, V_2)\\
&(\Gamma_2^{\perp})^T V_2 \sim N_{n-1}(0, I_{n-1}).
\end{aligned}$$

As,

$$\begin{aligned}
t_{12} =< W_1, V_2 > &= W_1^T V_2\\
V_2 &\sim N(0, I_n)\\
W_1^T V_2 &\sim N(0, W_1^T W_2) = N(0,1)\\
\text{Similarly, } t_{23} =< W_2, V_3 > &= W_2^T V_3\\
V_3 &\sim N(0, I_n)\\
W_2^T V_3 &\sim N(0, W_2^T W_3) = N(0,1)\\
cov(W_2^T V_3, W_1^T V_2) &= E(W_2^T I_n W_1) = 0
\end{aligned}$$

So, all of t_{ij}'s are mutually independent, $i < j$, t_{ij} is inner product onto orthogonal space $\Longrightarrow$ correlation zero, $\Longrightarrow$ independence. t_{ii} is the norm (HW: $||z|| \perp \frac{z}{||z||}$)

$$f^T(t) = \prod_{i=1}^{p} \frac{t_{ii}^{n-i} e^{-\frac{1}{2}t_{ii}^2}}{2^{\frac{n-i+1}{2}} \Gamma(\frac{n-i+2}{2})} \prod_{j>i} \frac{1}{\sqrt{2\pi}} e^{-\frac{t_{ij}^2}{2}}$$

$$J(X \to T) = \frac{f^T(t)}{f^X(x)}$$

$$T = \begin{bmatrix} \chi_n & N(0,1) & N(0,1) & \dots & N(0,1)\\ & \chi_{n-1} & N(0,1) & \dots & N(0,1)\\ & & \vdots & & \\ & & & \dots & \chi_{n-p+1} \end{bmatrix}_{p\times p}$$

9.26 Bartlett Decomposition for General Wishart Matrix

$$\begin{aligned}
S &\sim W_p(\Sigma, n) \\
X_{p\times n} &\sim N_{p\times n}(0, I_n \otimes \Sigma) \ \text{ s.t. } S = XX^T \\
X_{p\times n} &= \Sigma^{\frac{1}{2}} Z \quad Z \sim N(0, I_n \times I_p) \\
S &= \Sigma^{\frac{1}{2}} Z Z^T (\Sigma^{\frac{1}{2}})^T
\end{aligned}$$

If we use $\Sigma^{1/2} =$ LT Cholesky Decomposition

$$S = (LT)^{\Sigma^{1/2}} (LT)^{T^T} (UT)^T (UT)^{(\Sigma^{1/2})^T}$$

Lemma 9.10 *$J(T \to S)$*

Let T be upper triangular, $S = T^T T$. Then we have an injective mapping

$$J(T \to S) = (2^p \prod_{i=1}^{n} t_{ii}^{p-i+1})^{-1}$$

Intuition: Degrees of Freedom $T, S : \frac{p(p+1)}{2}$ Proof: HW3

9.26.1 Hsu's Lemma

$$\begin{aligned}
J(X \to S) = J(X \to T) J(T \to S) \\
= \frac{\pi^{\frac{np}{2} - \frac{p^2}{4} + \frac{p}{4}}}{\prod_{i=1}^{p} \Gamma(\frac{n-i+1}{2})} |S|^{\frac{1}{2}(n-p+1)}
\end{aligned}$$

Theorem 9.11 *If $\Sigma > 0$, then $S > 0$ w.p. 1 and has density*

$$\begin{aligned}
f^S(s) = C_2 |\Sigma|^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2} tr(\Sigma^{-1} S) \right\} |S|^{\frac{n-p-1}{2}} \\
\textit{where } C_2 = \left[2^{\frac{np}{2}} \pi^{\frac{p^2}{4} - \frac{p}{4}} \prod_{i=1}^{p} \Gamma(\frac{n-i+1}{2}) \right]^{-1}
\end{aligned}$$

9.26.2 Properties of Wishart Distribution

(1) If $S \sim W_p(\Sigma, n)$, let $\tilde{S}_{qq} = A_{q\times p} S_{p\times p} A^T_{p\times q}$. Then:

$$\tilde{S} \sim W_q(A\Sigma A^T, n)$$

Proof:

$$\begin{aligned}
S &= XX^T \\
X &\sim N(0, I_n \otimes \Sigma) \\
AX &\sim N(0, I_n \otimes A\Sigma A^T) \\
\Rightarrow ASA^T &= AXX^T A^T = (AX)(AX)^T \sim W_q(A\Sigma A^T, n)
\end{aligned}$$

(2) Let

$$\begin{aligned} &S_1 \sim W_p(\Sigma, n_1) \\ &S_2 \sim W_p(\Sigma, n_2) \\ &S_1 \perp S_2 \Rightarrow S_1 + S_2 \sim W_p(n_1 + n_2) \end{aligned}$$

Proof:

$$\begin{aligned} S_1 &= X_1 X_1^T \text{ where } X_1 \sim N(0, I_{n_1} \otimes \Sigma) \\ S_2 &= X_2 X_2^T \text{ where } X_2 \sim N(0, I_{n_2} \otimes \Sigma) \\ S_1 \perp S_2 &\iff X_1 \perp X_2 \\ \because \text{Block matrix } \begin{bmatrix} X_1 & X_2 \end{bmatrix} &\sim N(0, I_{n_1+n_2} \otimes \Sigma) \\ \therefore S_1 + S_2 = \begin{bmatrix} X_1 & X_2 \end{bmatrix} &\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim W_p(\Sigma, n_1 + n_2) \end{aligned}$$

(3) Let

$$X_{p\times n} \sim N(\mu_{p\times n}, I_n \otimes \Sigma)$$

Where $\Gamma_{n\times m}$ is an orthogonal matrix in $\mathcal{L}^{\perp}_{row}(\mu)$ i.e. $\mu\Gamma = 0$. Then:

$$\begin{aligned} Y_{p\times m} = X\Gamma &\sim N_{p\times m}(\underbrace{\mu\Gamma}_{0}, \underbrace{\Gamma^T\Gamma}_{I_m} \otimes \Sigma) \\ S = YY^T &\sim W_p(\Sigma, m) \end{aligned}$$

(4) Let

$$X \sim N(\mu, I_n \otimes \Sigma)$$

Where P is a projection matrix into m dimensional space that is a subspace of $\mathcal{L}^{\perp}_{row}(\mu)$ i.e. $\mu P = 0$ Then

$$Y := XP \Rightarrow YY^T \sim W_p(\Sigma, m)$$

Proof: Let Γ be an orthogonal basis $\Gamma_{n\times m}$

$$\begin{aligned} P &= \Gamma\Gamma^T \\ YY^T &= XPPX^T \\ &= XPX^T \\ &= X\Gamma\Gamma^T X^T \\ &= (X\Gamma)(X\Gamma)^T \\ \text{where } X\Gamma &\sim N(0, I_m \otimes \Sigma) \end{aligned}$$

(5) Application to Centering
Consider the projection matrix onto $\mathcal{L}^{\perp}_{col}(1_n)$

$$H_n = I_n - \frac{1}{n} 1_n 1_n^T$$

and the matrix

$$X \sim N(\mu 1_n^T, I_n \otimes \Sigma) \iff X_i \overset{iid}{\sim} N_p(\mu, \Sigma),\ i = 1, 2, \ldots, n$$

Then

$$Y = XH_n = (X_1 - \bar{X}, \ldots, X_n - \bar{X})$$
$$S = YY^T = \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})^T \sim W_p(\Sigma, n-1)$$

9.27 Wishart distribution

Definition 9.12 *Suppose X is a $p \times n$ matrix,each column of which is independently drawn from a p-variate normal distribution with zero mean:*
$S = XX^T = \sum_{i=1}^{n} X_i X_i^T$ *known as the scatter matrix.*
One indicates that S has that probability distribution by writing:
$S \sim W_p(V, n)$
The positive integer n is the number of degrees of freedom.

Theorem 9.13 *Correlation coefficient:* $\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$
Sample Correlation coefficient: $R_{ij} = \frac{S_{ij}}{\sqrt{S_{ii}S_{jj}}}$
$D_{p\times p} = diag(\sqrt{S_{ii}}), i = 1, \ldots, p$
So, $R = D^{-1}SD^{-1}$
$df : R : \frac{p(p-1)}{2} \quad D : p \quad S : \frac{p(p+1)}{2}$
Therefore, $S \to (D, R)$ *is an one-to-one mapping.*

Lemma 9.14 $J(S \to (D, R)) = 2^p |D|^p$ *proof:* $S = DRD$ *then calculate the derivative* $S_{ij} = d_i R_{ij} d_j$

Theorem 9.15 *The joint distribution of* (D, R) *is* $f_S(s) \cdot 2^p |D|^p$

Corollary 9.16 *If* $S \sim W_p(\Sigma, n-1)$ *then D is orthogonal to R with* $f^R(R) = C_4 \cdot |R|^{\frac{n-p-1}{2}}$ *and* $D_{ii} \overset{i.i.d}{\sim} \chi_n$
Proof: $f_S(s) \cdot 2^p |D|^p = C_3 \cdot e^{-\frac{1}{2}tr(S)} |S|^{\frac{n-p-1}{2}} \cdot 2^p$
Because we can have this separate expression, D is orthogonal to R

ISYE 7405: Multivariate Data Analysis **Georgia Tech**

Chapter 10: Intro to Hotelling's T^2 statistics

Lecturer: Shihao Yang

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

This notes was done by Hongqian Sun

10.28 Hotelling's T^2 statistics

Definition 10.17 *Univariate t-test: $X_1, X_2, \dots, X_n \overset{i.i.d}{\sim} N(\mu, \sigma^2)$. $H_0 : \mu = 0, H_1 : \mu \neq 0$*

$$t = \frac{\sqrt{n}\bar{X}}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2/n-1}} = sgn(\bar{X})\frac{||\hat{X}||}{||X^{\perp}||}\sqrt{n-1}$$

Definition 10.18 *Hotelling's T^2 test: $X_1, X_2, \dots, X_n \overset{i.i.d}{\sim} N_p(\mu, \Sigma)$. $H_0 : \mu = 0, H_1 : \mu \neq 0$*

$$S = \sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T = XX^T - n\bar{X}\bar{X}^T$$

The T^2 test statistics:

$$T^2 = \bar{X}^T(\frac{S}{n(n-1)})^{-1}\bar{X}$$

The Hotelling's T^2 statistics if a quadratic from w.r.t. shape S^{-1}.

10.28.1 Geometric interpretation

Claim 10.19 *look for the ”smallest angle” between 1_n and any linear combination of $V_1, V_2, \dots, V_p$, where $V_{n\times p} = X^T = (V_1, V_2, \dots, V_p)$. Looking at the angle between $L_{row}(X)$ & 1_n. Let's all this angle A.*

Proof:
Projection matrix onto $L_{row}(X)$ is $X^T(XX^T)^{-1}X$.
the project of 1_n on $L_row(X)$ is $u = X^T(XX^T)^{-1}X1_n$.

$$cos(A)||1_n||||u|| =< 1_n, u >$$

$$cos^2(A) = \frac{< 1_n, u >^2}{< 1_n, 1_n >< u, u >} = \frac{1_n^T X^T(XX^T)^{-}1X1_n}{n} = n\bar{X}^T(XX^T)^{-1}\bar{X} = n\bar{X}^T(S + n\bar{X}\bar{X}^T)^{-1}\bar{X}$$

Woodbury's Formula:

$$cos^2(A) = n\bar{X}^T[S^{-1} - \frac{S^{-1}\bar{X}\bar{X}^T S^{-1}}{1/n + \bar{X}S^{-1}\bar{X}}]\bar{X} = \frac{n\bar{X}^T S^{-1}\bar{X}}{1 + n\bar{X}^T S^{-1}\bar{X}}$$

Therefore,

$$cot^2 A = \frac{cos^2 A}{1 - cos^2 A} = n\bar{X}^T S^{-1}\bar{X} = \frac{T^2}{n-1}$$

■

10.28.2 Connection with linear regression

$1_n \overset{regress}{=} \beta_1 V_1 + \beta_2 V_2 + \cdots + \beta_p V_p \quad H_0 : \mu = 0 \iff H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$, H_1:at least 1 of β are not equal to 0

ISYE 7405: Multivariate Data Analysis **Georgia Tech**

Chapter 11:

Lecturer: Shihao Yang

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

11.29 Null Distribution of T^2

If H_0 is true, Under H_0: $X \sim N_{p\times n}(0, I_n \otimes \Sigma)$, assume the angle between 1_n and $\mathcal{L}_{row}(X)$ is A.
Change the frame as if we are sitting down on a fixed p-dimensional hyperplane, where the direction of 1_n is "uniformly distributed" on all possible directions. Therefore, as long as the projection is concerned, we can take the plane to be $\mathcal{L}(e_1, e_2, ..., e_p)$ where $e_i = (0, 0, ..., 0, 1(\text{i-th}), 0, ..., 0)$ and replace 1_n by $(y_1, y_2, ..., y_n)$ where $y_i \sim N(0, 1)$, iid. Then

$$\widehat{y} = (y_1, y_2, ..., y_p, 0, ..., 0)$$
$$y^{\perp} = (0, ..., 0, y_{p+1}, ..., y_n)$$
$$cot^2 A = \frac{||\widehat{y}||^2}{||y^{\perp}||^2} = \frac{\sum_{i=1}^{p} y_i^2}{\sum_{i=p+1}^{n} y_i^2} \sim \frac{p}{n-p} F_{p,n-p}$$

11.29.1 Linear Regression Interpretation

H_0: $\mu = 0 \Leftrightarrow H_0$: $\beta = 0$ in regression of $1_n \sim X^T \Leftrightarrow$ F-test.

$$X \sim N_{p\times n}(\mu 1_n^T, I_n \otimes \Sigma)$$
$$S = \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})^T = XX^T - n\bar{X}\bar{X}^T$$
$$\bar{X} \perp\!\!\!\perp S$$
$$test\ H_0 :\ \mu = 0$$
$$T^2 = \bar{X}^T \{\frac{S}{n(n-1)}\}^{-1} \bar{X}$$

Theorem 11.20 *Under the null hypothesis, the hotelling T^2 statistics has $\frac{(n-1)p}{n-p} F_{p,n-p}$ distribution.*

Ex: Let a_{p*1} be a fixed non-zero vector, $y = a^T x = (y_1, ..., y_n)$, $y_i \sim N(a^T\mu, a^T\Sigma a)$, iid. Let $t^2(a)$ be the sample t-test statistics

$$t^2(a) = \frac{n\bar{y}^2}{\frac{\sum (y_i - \bar{y})^2}{n-1}}$$

Show that Hotelling T^2 statistics. $T^2 = \max_a t^2(a)$ (Homework).

Lemma 11.21 *Suppose $S \sim W_p(\Sigma, m)$ $m > p$, then we have*

1) $\frac{a^T S a}{a^T \Sigma a} \sim \chi^2_m$ *where a is a p-vector*

2) $\frac{a^T\Sigma^{-1}a}{a^TS^{-1}a} \sim \chi^2_{m-p+1}$

Proof: 1):

$$\begin{aligned} S = XX^T,\ X &\sim N_{p*m}(0, I_m \otimes \Sigma) \\ then\ a^TX &\sim N(0, I_m \otimes a^T\Sigma) \\ then\ a^TSa &= (a^TX)(a^TX)^T \end{aligned}$$

The rest is trivial.

2): We can simply take $\Sigma = I_p$, WLOG, otherwise we can take $\tilde{S} = \Sigma^{-\frac{1}{2}}S(\Sigma^{-\frac{1}{2}})^T$, $b = \Sigma^{-\frac{1}{2}}a$, then $\frac{a^T\Sigma^{-1}a}{a^TS^{-1}a} = \frac{b^Tb}{b^T\tilde{S}^{-1}b}$ where $\tilde{S} \sim W_p(I_p, m)$. Also, we can assume $\|b\| = 1$.

So, we only need to show for $S \sim W_p(I_p, m)$ u unit length vector $\frac{1}{u^TS^{-1}u} \sim \chi^2_{m-p+1}$. Let $\Gamma_{p*p} = (\Gamma_1, u)$ be a full orthogonal matrix.

$$\begin{aligned} R = \Gamma^TS\Gamma &\sim W_p(I_p, m) \\ R^{-1} = \Gamma^TS^{-1}\Gamma &\sim \begin{bmatrix} \Gamma_1 \\ u^T \end{bmatrix} S^{-1}[\Gamma_1\ u] \end{aligned}$$

Then $u^TS^{-1}u = (R^{-1})_{pp}$, $R = \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix}$. Therefore, $u^TS^{-1}u = (R_{22} - R_{21}R_{11}^{-1}R_{12})^{-1} \sim \frac{1}{\chi^2_{m-p+1}}$ ■

Proof:[Proof of Theorem 11.20] Now we prove under H_0: $T^2 \sim (n-1)\frac{p}{n-p}F_{p,n-p}$.

$$\begin{aligned} T^2 &= \bar{X}^T\left[\frac{S}{n(n-1)}\right]^{-1}\bar{X} \\ &= (n-1)\frac{\bar{X}^T(\frac{\Sigma}{n})^{-1}\bar{X}}{\bar{X}^T(\frac{\Sigma}{n})^{-1}\bar{X}/\bar{X}^T(\frac{S}{n})^{-1}\bar{X}} \\ \bar{X} &\sim N_p(0, \frac{\Sigma}{n}) \\ \bar{X} &\perp\!\!\!\perp S \\ \frac{S}{n} &\sim W_p(\frac{\Sigma}{n}, n-1) \end{aligned}$$

Conditioning on $\bar{X}$, use 2), we have $\frac{\bar{X}^T(\frac{\Sigma}{n})^{-1}\bar{X}}{\bar{X}^T(\frac{S}{n})^{-1}\bar{X}} \sim \chi^2_{n-p}$ which implies $\frac{\bar{X}^T(\frac{\Sigma}{n})^{-1}\bar{X}}{\bar{X}^T(\frac{S}{n})^{-1}\bar{X}}$ is independent of $\bar{X}$. $\Rightarrow \bar{X}^T(\frac{\Sigma}{n})^{-1}\bar{X} \sim \chi^2_p$. Using 1) and also independent to $\frac{\bar{X}^T(\frac{\Sigma}{n})^{-1}\bar{X}}{\bar{X}^T(\frac{S}{n})^{-1}\bar{X}}$. $\Rightarrow T^2 \sim (n-1)\frac{p}{n-p}F_{p,n-p}$ ■

11.30 Non Null Distribution of T^2

What happens to T^2 if $\underset{\sim}{\mu} \neq \underset{\sim}{0}$?

Lemma 11.22 (Linear Invariance of T^2) *$X_{p*n} \sim N_{p*n}(\mu 1_n, I_n \otimes \Sigma)$, $\tilde{X} = A_{p*p}X$ when A_{p*p} is non-singular, similarly we can get $\tilde{\bar{\bar{X}}} = \frac{1}{n}\sum_{i=1}^n \widetilde{X_i}$, $\tilde{S} = \sum_{i=1}^n(\widetilde{X_i} - \tilde{\bar{\bar{X}}})(\widetilde{X_i} - \tilde{\bar{\bar{X}}})^T$ and $\widetilde{T}^2 = \tilde{\bar{\bar{X}}}\left[\frac{\widetilde{S}}{n(n-1)}\right]^{-1}\tilde{\bar{\bar{X}}}$.*

Claim 11.23 $T^2 = \widetilde{T}^2$

Proof:

$$\begin{aligned}
\tilde{\bar{X}} &= A\bar{X} \\
\widetilde{S} &= ASA^T \\
\widetilde{T}^2 &= \tilde{\bar{X}} \left[\frac{\widetilde{S}}{n(n-1)} \right]^{-1} \tilde{\bar{X}} = T^2
\end{aligned}$$

■

This means we can even assume $\Sigma = I_p$ in our proof for the distribution of T^2.

Theorem 11.24 *(Hotelling) If $X_{p*n} \sim N(\mu 1_n^T, I_n \otimes \Sigma)$, $\Sigma > 0$, then $T^2 = \bar{X} \left[\frac{\widetilde{S}}{n(n-1)} \right]^{-1} \bar{X}$ has distribution $T^2 \sim (n-1)\frac{p}{n-p} F_{p,n-p}(n\mu^T \Sigma^{-1} \mu)$ where $F_{n_1,n_2}(\delta^2)$ is the non-central F distribution. $F_{n_1,n_2}(\delta^2) = \frac{\chi^2_{n_1}(\delta^2)/n_1}{\chi^2_{n_2}/n_2}$.*

ISYE 7405: Multivariate Data Analysis **Georgia Tech**

Chapter 12: Two-sample T Test

Lecturer: Shihao Yang

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

12.31 Review

Theorem 12.25 (Hotelling's Theorem) *If $X_{p\times n} \sim N(\mu 1_n^T, I_n \otimes \Sigma)$, then $T^2 = \bar{X}^T(\frac{S}{n(n-1)})^{-1}\bar{X}$ has distribution:*

$$T^2 \sim (n-1)\tfrac{p}{n-p}F_{p,n-p}(n\mu^T\Sigma^{-1}\mu)$$

where $F_{n_1,n_2}(\delta^2)$ is the non-normal F - distribution:

$$F_{n_1,n_2}(\delta^2) = \tfrac{\chi^2_{n_1}(\delta^2)/n_1}{\chi^2_{n_2}/n_2}$$

Proof: Take $A = \Sigma^{-1}$, so $\widetilde{X} = \Sigma^{-\frac{1}{2}}X$, $\widetilde{\mu} = \Sigma^{-\frac{1}{2}}\mu$. Then:

$$T^2 = \widetilde{T}^2 = \tfrac{n\bar{\tilde{X}}^T\bar{\tilde{X}}}{n\bar{\tilde{X}}^T\bar{\tilde{X}}/\bar{\tilde{X}}^T(\frac{S}{n})^{-1}\bar{\tilde{X}}} \cdot (n-1)$$

$$Denominator \sim \chi^2_{n-p}$$

$$Nominator \sim N_p(\hat{\mu}, I_p/n) = \tfrac{1}{\sqrt{n}}N_p(\sqrt{n}\tilde{\mu}, I_p)$$

By definition of non-central Chi-square distribution with non-central parameter:

$$\tilde{\mu}_1^2 + \tilde{\mu}_2^2 + ... + \tilde{\mu}_p^2 = \tilde{\mu}^T\tilde{\mu} = \mu^T\Sigma^{-1}\mu$$

More generally, $y \sim N_p(\mu, c\Sigma)$ and is independent from $S \sim W_p(\Sigma, \mu)$, then:

$$y^T(\tfrac{\Delta S}{m})^{-1}y \sim \tfrac{mp}{m-p+1}F_{p,m-p+1}(\delta^2)$$

$$\delta^2 = \mu^T\Sigma^{-1}\mu/c$$

■

12.32 Two Sample T-Test

1. Assume $X_1, X_2, ..., X_{n_1} \sim N(\mu_1, \sigma^2), Y_1, ..., Y_{n_1} \sim N(\mu_2, \sigma^2)$, and test the hypothesis:

$$H_0 : \mu_1 = \mu_2 \quad H_1 : \mu_1 \neq \mu_2$$

T-statistics:

$$t = \frac{\bar{y}-\bar{x}}{\sqrt{(\frac{1}{n_1}+\frac{1}{n_2})\frac{S_1+S_2}{n-2}}}$$

$$S_1 = \sum_{j=1}^{n_1}(X_j - \bar{X})^2$$

$$S_2 = \sum_{j=1}^{n_2}(Y_j - \bar{Y})^2$$

2. Geometry of Two-Sample T-test:

$$d = (-\frac{1}{n_1}^{(1)}, -\frac{1}{n_1}^{(2)}, ..., -\frac{1}{n_1}^{(n_1)}, \frac{1}{n_2}^{(1)}, ..., \frac{1}{n_2}^{(n_2)})$$

$$Z = (X_1, X_2, ..., X_{n_1}, Y_1, Y_2, ..., Y_{n_2})$$

$$\bar{Y} - \bar{X} = d^T Z$$

3. Linear regression interpretation:
Regress variable Z onto vectors $(1, 1, ..., 1)$ and vector d:

$$Z \sim \alpha 1_n + \beta d$$

$$\mu_1 = \mu_2 \Leftrightarrow \beta = 0$$

12.33 Two Sample T-Test in Multivariate Situations

1. Assume $X_1, X_2, ..., X_{n_1} \sim N(\mu_1, \Sigma), Y_1, Y_2, ..., Y_{n_1} \sim N(\mu_2, \Sigma)$, and test the hypothesis:

$$H_0 : \mu_1 = \mu_2 \quad H_1 : \mu_1 \neq \mu_2$$

$$T^2 = (\bar{y} - \bar{x})^T((\frac{1}{n_1} + \frac{1}{n_2})\frac{S_1+S_2}{n-2})^{-1}(\bar{y} - \bar{x})$$

$$S_1 = \sum_{j=1}^{n_1}(X_j - \bar{X})(X_j - \bar{X})^T$$

$$S_2 = \sum_{j=1}^{n_2}(Y_j - \bar{Y})(Y_j - \bar{Y})^T$$

2. Geometric interpretation:

$$Z_{p\times n} = (X, Y) = (Z_1^T, Z_2^T, ..., Z_p^T)$$

3. Linear regression interpretation:

$$d = (-\frac{1}{n_1}^{(1)}, -\frac{1}{n_1}^{(2)}, ..., -\frac{1}{n_1}^{(n_1)}, \frac{1}{n_2}^{(1)}, ..., \frac{1}{n_2}^{(n_2)})$$

$$d \sim \beta_1 Z_1^{\perp} + \beta_2 Z_2^{\perp} + ... + \beta_n Z_p^{\perp}$$

$$H_0 : \mu 1 = \mu 2 \quad \Leftrightarrow \beta_1 = \beta_2 = ... = \beta_n = 0$$

4. Null distribution of T^2:
Theorem: Under null hypothesis,

$$T^2 \sim (n-2)\frac{p}{n-p-1}F_{p,n-p-1}$$

Under non-null hypothesis,

$$T^2 \sim (n-2)\tfrac{p}{n-p-1}F_{p,n-p-1}(\delta^2)$$

$$\delta^2 = \tfrac{n_1 n_2}{n}(\mu_2-\mu_1)^T\Sigma^{-1}(\mu_2-\mu_1)$$

12.34 Mahalanobis Distance

If $X_{p\times 1} \sim (\mu_X, \Sigma)$, which does not have to be Gaussian, and $Y_{p\times 1} \sim (\mu_Y, \Sigma)$ (share the same Σ). Then

$$\Delta = [(\mu_y - \mu_X)^T\Sigma^{-1}(\mu_Y - \mu_X)]^{\frac{1}{2}}$$

Δ is called Mahalanobis Distance between X and Y. Properties:
1. Linearly invariant: $A_{p\times p}$ is non-singular, $\tilde{\mu}_X = AX + B, \tilde{\mu}_Y = AY + B$, then $\tilde{\Delta} = \Delta$.
2. Connection with K-L divengence: $KL(N_p(\mu_x, \Sigma), N_p(\mu_Y, \Sigma)) = 1/2\Delta$;
3. Decomposition of Δ^2: Suppose $X_{p\times 1} = (X_{1(p_1\times 1)}, X_{2(p_2\times 1)})$, then:

$$\mu_2^{\perp} = \mu_2 - \Sigma_{21}\Sigma_{11}^{-1}\mu_1$$

$$\Sigma_{22}^{\perp} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$$

$$\Delta^2(X) = \mu_1^T\Sigma_{11}\mu_1 + (\mu_2^{\perp})^T(\Sigma_{22}^{\perp})^{-1}\mu_2^{\perp}$$

ISYE 7405: Multivariate Data Analysis **Georgia Tech**

Chapter 13: Principal Component Analysis

Lecturer: Shihao Yang

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

13.35 Fundamental Lemma for PCA

Quadratic form: $||g||_A^2 = g^T A g$, $A_{p\times p} \geq 0$, A is symmetric, $rank(A) = k$. Then

$$A = \Gamma_1 \Lambda \Gamma_1^T,$$

where $\Gamma_1 = (\gamma_1, ..., \gamma_k), \Lambda = diag\{\lambda_1, ..., \lambda_k\}$.

Enrich the orthogonal basis: $\Gamma = (\Gamma_1, \gamma_{k+1}, \ldots, \gamma_p)$ and $\Lambda = diag(\lambda_1, \ldots, \lambda_k, 0, \ldots, 0)$, Thus,

$$A = \sum_{i=1}^{p} \lambda_i \gamma_i \gamma_i^T$$

$\forall g \in \mathbb{R}^p$, $||g||_A^2 = g^T A g = \sum_{i=1}^{p} \lambda_i g^T \gamma_i \gamma_i^T g = \sum_{i=1}^{p} \lambda_i h_i^2$, where $h_i = \gamma_i^T g = g^T \gamma_i =< g, \gamma_i >$ is the coordinate under basis $\gamma_i, \ldots, \gamma_p$.

Claim 13.26 *Assume $rank(A) = p$ and i.e. $\lambda_i > 0$ for $i = 1, 2, \ldots, p$*

$$\begin{aligned} \Gamma &:= [\gamma_1, \gamma_2, \ldots, \gamma_p] \\ \Lambda &= diag(\lambda_1, \lambda_2, \ldots, \lambda_p) \\ A &= \Gamma \Lambda \Gamma^T \end{aligned}$$

Then

1. *$g^* := \gamma_1$ maximizes*
$$\max_g \left\{ \|g\|_A^2 : \|g\| = 1 \right\}$$
with objective value $\|g^*\|_A^2 = \lambda_1$

2. *$g^* := \gamma_2$ maximizes*
$$\max_g \left\{ \|g\|_A^2 : \|g\| = 1,\ < g, \gamma_1 >= 0 \right\}$$
with objective value $\|g^*\|_A^2 = \lambda_2$

$\vdots$

k. *$g^* := \gamma_k$ maximizes*
$$\max_g \left\{ \|g\|_A^2 : \|g\| = 1,\ < g, \gamma_1 >= 0, < g, \gamma_2 >= 0, \ldots, < g, \gamma_{k-1} >= 0 \right\}$$
with objective value $\|g^*\|_A^2 = \lambda_k$

$\vdots$

p. $g^* := \gamma_p$ *minimizes* $\|g\|_A^2$ *on* $\|g\|^2 = 1$ *with objective value* $\|g^*\|_A^2 = \lambda_p$

Note: If $\|g\|^2 = 1$ then what is the range of $\|g\|_A^2$

$$R(g) = \frac{\|g\|_A^2}{\|g\|^2}$$

Proof: We see that

$$\begin{aligned} h := \Gamma^T g = \begin{pmatrix} \gamma_1^T g \\ \gamma_2^T g \\ \vdots \\ \gamma_p^T g \end{pmatrix} &= \begin{pmatrix} h1 \\ h2 \\ \vdots \\ h_p \end{pmatrix} \\ \|g\|_A^2 := g^T A g = g^T \Gamma \Lambda \Gamma^T g \\ = \sum_{i=1}^p \lambda_i h_i^2 \\ \therefore \|h\|_2^2 = h^T h = g^T \Gamma \Gamma^T g = g^T g = 1 \end{aligned}$$

So given $\|h\|_2^2 = 1$, what is the range of $\|g\|_A^2 = \sum_{i=1}^p \lambda_i h_i^2$?

1. $h = [1; 0; 0; \ldots; 0]$ maximizes $\sum_{i=1}^p \lambda_i h_i^2$ to be λ_1

2. $g^T \gamma_1 = 0 \iff h_1 = 0 \iff h = [0; \Box; \Box; \ldots; \Box]$
   Given $h_1 = 0, h = [0; 1; 0; \ldots; 0]$ maximizes $\sum_{i=1}^p \lambda_i h_i^2$ to be λ_2

   $\vdots$

k. $g^T \gamma_1 = 0, g^T \gamma_2 = 0, \ldots, g^T \gamma_{k-1} = 0 \iff h_1 = h_2 = \cdots = h_{k-1} = 0$
   Given $h_1 = h_2 = \cdots = h_{k-1} = 0, h = [0; 0; 0; \ldots; \underbrace{1}_{k\text{-th}}; \ldots; 0]$ maximizes $\sum_{i=1}^p \lambda_i h_i^2$ to be λ_k

   $\vdots$

p. $h = [0; 0; \ldots; 0; 1]$ minimizes $\sum_{i=1}^p \lambda_i h_i^2$ to be λ_p

■

13.35.1 Simultaneous Orthogonality

Since

$$g^T A \gamma_i = g^T (\sum_{j=1}^p \lambda_j \gamma_j \gamma_j^T) \gamma_i = \lambda_i (g^T \gamma_i)$$

Then we have

$$g^T A \gamma_i = 0 \iff g^T \gamma_i = 0,$$

which means

$$< g, \gamma_i >= 0 \iff < g, \gamma_i >_A = 0.$$

Thus, $g^T \gamma_i = 0$ can be replaced by $< g, \gamma_i >_A = 0$

13.36 Principal Components in Sample Space

Given sample

$$\begin{aligned} X_{p\times n} &= [\vec{X}_1, \vec{X}_2, \ldots, \vec{X}_n] \\ &= \begin{pmatrix} \vec{V}_1^T \\ \vec{V}_2^T \\ \vdots \\ \vec{V}_p^T \end{pmatrix} \end{aligned}$$

Assume $\mathbb{E}[\vec{X}_i] = 0, S = XX^T$, with spectral decomposition

$$\begin{aligned} &S = \Gamma D \Gamma^T \\ &D = diag(d_1, \ldots, d_p),\ d_1 \geq d_2 \geq \ldots \geq d_p > 0 \\ &\Gamma = [\gamma_1, \gamma_2, \ldots, \gamma_p] \end{aligned}$$

Definition 13.27

1. *$\vec{Y}_j = \gamma^T X$ is known as the j-th principal component of X*
2. *γ_j is known as the j-th principal factor*
3. *Let $\Gamma(j) := [\gamma_1, \gamma_2, \ldots, \gamma_j]$*
   Then
   $$\vec{Y}(j) = [\Gamma(j)]^T X = \begin{pmatrix} \gamma_1^T X \\ \vdots \\ \gamma_j^T X \end{pmatrix} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_j \end{pmatrix}$$
   is called the j-th principal component representation
4. *The i-th component of the vector $\vec{Y}_j$ is called the loading of $\vec{X}_i$ on γ_j*
   $$(\vec{Y}_j)_i = \gamma_j^T \vec{X}_i$$

Theorem 13.28 *Among all p-dimensional **unit vectors** $\vec{g}$, the first principal factor γ_1 maximizes*

$$\sum_{i=1}^{n} (\vec{g}^T \vec{X}_i)^2 = \|\vec{g}^T X\|^2 \text{ (Sample Variance)}$$

where the maximum is d_1
Then among all unit vectors $\vec{g}$ satisfying $\vec{g}^T \gamma_1 = 0$ or equivalently

$$\vec{g}^T S \vec{g} = 0 \iff \vec{g}^T X X^T \vec{g} = 0 \iff Corr(\vec{g}^T X, \gamma_1^T X) = 0$$

The second principal factor γ_2 maximizes $\underbrace{\|g^T X\|_2^2}_{sample\ variance}$ *and the maximum is d_2*

$\vdots$

ISYE 7405: Multivariate Data Analysis **Georgia Tech**

Chapter 14: Principal Component Analysis

Lecturer: Shihao Yang

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

14.37 Basic notions

$$Y(j) = \Gamma(j)^T X = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \gamma_1^T X \\ \gamma_2^T X \\ \vdots \\ \gamma_n^T X \end{pmatrix}$$

where $Y(j)$ represents $X_{p\times n}$ in lower dimension $j \times n$.

$$Tr(Y(j)Y(j)^T) = \sum_{k=1}^{j}\sum_{i=1}^{n} Y_{ki}^2, \ Y = Y(p)$$

$$YY^T = (\Gamma^T X)(\Gamma^T X)^T = \Gamma^T X X^T \Gamma = \Gamma^T S \Gamma = \Gamma^T \Gamma D \Gamma^T \Gamma = D$$

$$\sum_{k=1}^{p}\sum_{i=1}^{n} Y_{ki}^2 = \sum_{k=1}^{p} d_k$$

$$\sum_{k=1}^{j}\sum_{i=1}^{n} Y_{ki}^2 = Tr(Y(j)Y(j)^T) = tr(\Gamma(j)^T X X^T \Gamma(j)) = tr(\Gamma(j)^T \Gamma D \Gamma^T \Gamma(j)) = \sum_{k=1}^{j} d_k$$

as

$$\Gamma^T\Gamma(j) = \begin{pmatrix} \gamma_1^T \\ \vdots \\ \gamma_p^T \end{pmatrix} \begin{pmatrix} \gamma_1 & \dots & \gamma_j \end{pmatrix} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \dots & 1 \\ 0 & 0 & \dots & 0 \end{pmatrix}_{p\times j}$$

The proportion of variation explained by $Y(j)$ is $\frac{\sum_{k=1}^{j} d_k}{\sum_{k=1}^{p} d_k}$.

14.38 Population principle components

Suppose X is a random vector in $\mathbb{R}^p$, $X \sim (\mu, \Sigma)$, subtracting $\mu \to X \sim (0, \Sigma)$ (does not have to be Gaussian). Spectral decomposition $\Sigma = \Gamma\Lambda\Gamma^T = \sum_{i=1}^{p} \lambda_i \gamma_i \gamma_i^T$, where $\Gamma = (\gamma_1, \gamma_2, \dots, \gamma_n)$ and $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$.

Theorem 14.29 *(1) $g = \gamma_1$ maximize $Var(g^T X)$ subject to $||g|| = 1$, and the maximum is λ_1.*
(2) Among all unit vectors satisfying $g^T\gamma_1 = 0$ (equivalently, $Cov(g^T X, \gamma_1^T X) = 0 \Leftrightarrow g^T \Sigma \gamma_1 = 0$), $g = \gamma_2$ maximize $Var(g^T X)$, with maximum being λ_2.

Proof:

$$Cov(g^T X, h^T X) = g^T \Sigma h$$

Then use lemma. ■

Definition 14.30 *$y_j = \gamma_j^T X \rightarrow j$-th principal component of X.*

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_p \end{pmatrix} = \begin{pmatrix} \gamma_1^T \\ \vdots \\ \gamma_p^T \end{pmatrix} X = \Gamma^T X$$

$Cov(y) = Cov(\Gamma^T X) = \Lambda$

14.39 Best linear prediction

Theorem 14.31 *Suppose $X \sim (0, \Sigma)$, $\Sigma = \Gamma\Lambda\Gamma^T$, $\lambda_i > 0, \forall i$. Let $\Gamma(j) = (\Gamma_1, \dots, \Gamma_j)$, then*
(a) The best linear prediction of X in terms of $Y(j) = \Gamma(j)^T X$ is $\hat{X} = \sum_{i=1}^{j} \gamma_j y_j$.
(b) The residual $X^\perp = X - \hat{X}$ has covariance matrix

$$\Sigma_{(j)}^\perp = \sum_{i=j+1}^{p} \lambda_i \gamma_i \gamma_i^T \text{ with } tr(\Sigma_{(j)}^\perp) = \sum_{i=j+1}^{p} \lambda_i$$

(c) For any matrix $A_{j\times p}$, Let $Z = AX$ and $X_Z^\perp = X - \Sigma_{XZ}\Sigma_{ZZ}^{-1} Z$, we have

$$tr(\Sigma_Z^\perp) = tr(Cov(X_Z^\perp)) \geq \sum_{i=j+1}^{p} \lambda_i$$

The equality holds if and only if $A = \Gamma(j)^T$.

14.40 PCA and SVD

$X_{p\times n} = (X_1, \dots, X_n) = \begin{pmatrix} \dots & V_1^T & \dots \\ \dots & \vdots & \dots \\ \dots & V_p^T & \dots \end{pmatrix}$ Singular value decomposition:

$$X = L_{p\times r} C_{r\times r} R_{r\times n}^T$$

Where

$$L = (l_1, \dots, l_r),\ C = diag(c_1, \dots, c_r),\ R = (r_1, \dots, r_r)$$
$$c_1 \geq c_2 \geq \dots \geq c_r > 0,\ r = rank(X)$$

Then

$$L_{col}(X) = L_{col}(L),\ L_{row}(X) = L_{row}(R^T) = L_{col}(R)$$
$$S = XX^T = LC^2L^T, T = X^T X = RC^2R^T, \text{ spectral decomposition}$$

So L is the same as the principal factor.

$$L(j) = (l_1, \dots, l_j), C(j) = diag(c_1, \dots, c_j), R(j) = (r_1, \dots, r_j)$$

$$X = \sum_{i=1}^{r} l_i c_i r_i^T, \hat{X}(j) = \sum_{i=1}^{j} l_i c_i r_i^T$$

Lemma 14.32

$$\hat{X}(j) = L(j)C(j)R(j)^T = L(j)L(j)^T X = XR(j)R(j)^T$$

where $L(j)L(j)^T$ is the projection matrix onto $L_{col}(l_1, \ldots, l_j)$, first j principal factor.

Define $\hat{X}_{(j)}^{\perp} = X - \hat{X}(j)$ the residual of the approximation.
Define matrix norm: $< A, B >= tr(AB^T) = \sum_{i=1}^{p} \sum_{j=1}^{q} a_{ij} b_{ij}$ and $||A|| =< A, A >^{\frac{1}{2}}$

Theorem 14.33 *Among all j-dimensional subspace of $\mathbb{R}^p$, $L(l_1, \ldots, l_j)$ maximized the projected square length, and minimized the total orthogonal square residuals.*

$$\hat{X}(j) = L(j)C(j)R(j)^T, (\hat{X}(j))_i = \sum_{i=1}^{j} l_k c_k r_{ik}$$

$$X_i = \sum_{i=1}^{j} l_k c_k r_{ik}$$

$$Y = L^T X = CR^T$$

14.41 Metric eigenvalues

$$Q(g) = \frac{||g||_A^2}{||g||_B^2} = \frac{g^T A g}{g^T B g}, A \geq 0, B > 0$$

Define $\tilde{g} = B^{\frac{1}{2}} g,\ g = B^{-\frac{1}{2}} \tilde{g}$. Then

$$Q(g) = \frac{\tilde{g}^T B^{-\frac{1}{2}} A B^{-\frac{1}{2}} \tilde{g}}{\tilde{g}^T \tilde{g}} = \frac{\tilde{g}^T \tilde{A} \tilde{g}}{\tilde{g}^T \tilde{g}}, \text{ where } \tilde{A} = B^{-\frac{1}{2}} A B^{-\frac{1}{2}}$$

Consider spectral decomposition of $\tilde{A} = B^{-\frac{1}{2}} A B^{-\frac{1}{2}} = \Gamma \lambda \Gamma^T$, where

$$\Lambda = diag(\lambda_1, \ldots, \lambda_r, 0, \ldots, 0), r = rank(\tilde{A}) = rank(A)$$

and $\Gamma = (\gamma_1, \ldots, \gamma_p)$

$$\xi_i = B^{-\frac{1}{2}} \gamma_i,\ \Xi = B^{-\frac{1}{2}} \Gamma$$

Simultaneously diagonalization:

$$\Xi A \Xi = \Lambda, < \xi_i, \xi_j >_A = \lambda_i \delta_{ij}$$

where $\delta_{ij} = 1$ for $i = j$ and $\delta_{ij} = 0$ for $i \neq j$. Similarly,

$$\Xi B \Xi = I_p, < \xi_i, \xi_j >_B = \delta_{ij}$$

$$A\Xi = (\Xi^T)^{-1} \Lambda, B\Xi = (\Xi^T)^{-1}, A\Xi = B\Xi\Lambda, A\xi_i = \lambda_i B \xi_i$$

Definition 14.34 *The values $\lambda_1, \ldots, \lambda_p$ are called the eigenvalues of A in the B metric, $\xi_1, \ldots, \xi_p$ are the corresponding eigenvectors $\Leftrightarrow det(A - \lambda B) = 0$.*

Notions: $L_s^{\perp}(v_1, \ldots, v_j) = \{v :\ v^T s v_i = 0, i = 1, \ldots, j\}$. Then we have the corollary of the fundamental lemma:
For $Q(g) = \frac{||g||_A^2}{||g||_B^2}, g = \xi_1$ maximizes $Q(g)$. $\cdots\cdots$

ISYE 7405: Multivariate Data Analysis **Georgia Tech**

Chapter 15: LDA and Critical Angles

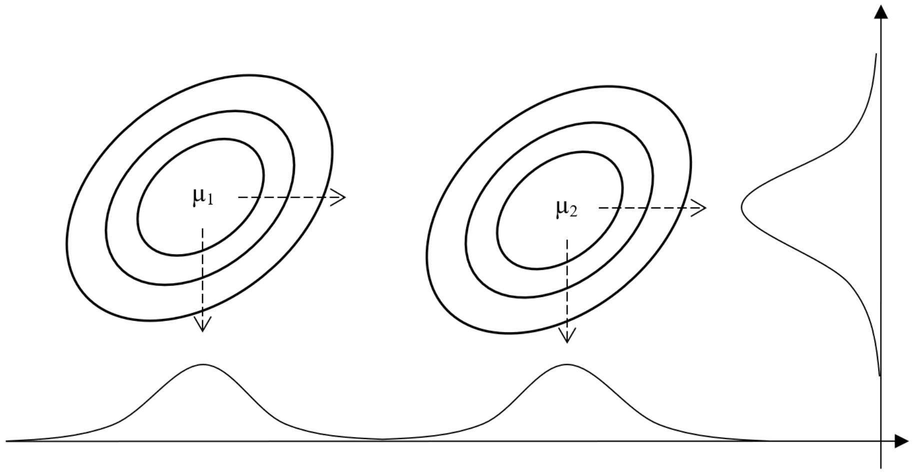
Lecturer: Shihao Yang

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

15.42 Fisher's Linear Discriminant Analysis

Suppose we have $X_1 \sim N_p(\mu_1, \Sigma)$ and $X_2 \sim N_p(\mu_2, \Sigma)$. We are interested in a 1-dimensional scalar that provides the maximum separation of X_1 and X_2. For example, see Figure 15.42.

Figure 15.8: Projecting the data in the downward direction provides more variable separation than projecting in the rightward direction.



In order to analyze this problem mathematically, we let $Y_1 = g^T X_1$ and $Y_2 = g^T X_2$ for some vector g. Then, $Y_1 \sim N_p(g^T\mu_1, g^T\Sigma g)$ and $Y_2 \sim N_p(g^T\mu_2, g^T\Sigma g)$. We want to maximize $\frac{(g^T\mu_1 - g^T\mu_2)^2}{g^T\Sigma g}$. To simplify, we let $\delta = \mu_1 - \mu_2$, $A = \delta\delta^T$, and $B = \Sigma$. Then, we consider the maximization of $Q(g) = \frac{g^T A g}{g^T B g}$.

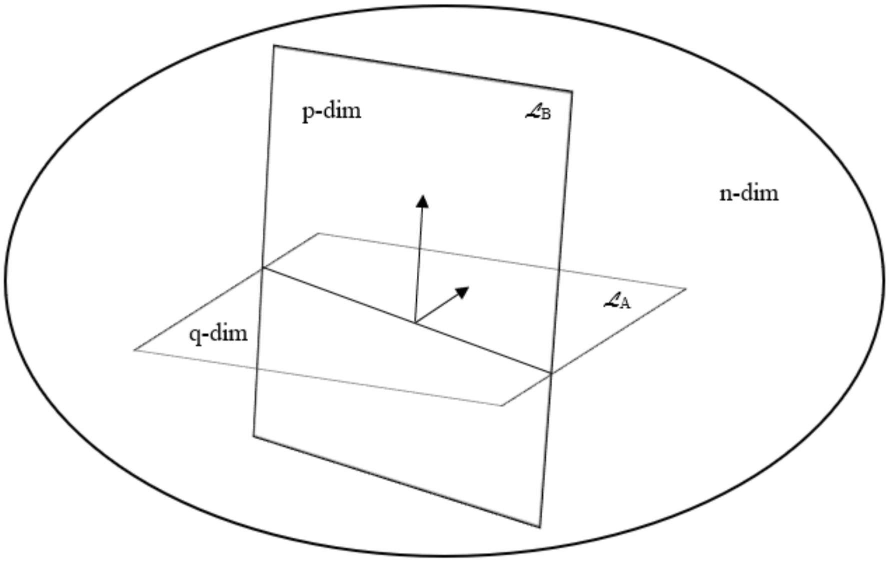
From the lemma in Lecture 14, we let $\tilde{A} = B^{-1/2} A B^{-1/2} = \Sigma^{-1/2}\delta\delta^T\Sigma^{-1/2}$ and $\gamma_1 = \Sigma^{-1/2}\delta$. We now have $\xi_1 = B^{-1/2}\gamma_1 = \Sigma^{-1}\delta$ and $\tilde{A}\gamma_1 = \Sigma^{-1/2}\delta\delta^T\Sigma^{-1/2}\Sigma^{-1/2}\delta = \lambda_1\gamma_1$ where $\lambda_1 = \delta^T\Sigma^{-1}\delta$. Because rank($\tilde{A}$) = 1, we also know $\lambda_2 = \lambda_3 = ... = 0$. The max value of $Q(g)$ is $\lambda_1 = \delta^T\Sigma^{-1}\delta = (\mu_1 - \mu_2)^T\Sigma^{-1}(\mu_1 - \mu_2)$ which is exactly the Mahalanobis Distance.

15.43 Critical Angles

Suppose there are two subspaces $\mathcal{L}_B$ and $\mathcal{L}_A$, of dimension p and q respectively, located in an n-dimensional space, as seen in Figure 15.43. We are interested in the smallest angles between subspaces, or critical angles.

To represent $\mathcal{L}_B$, we let $X_{p\times n} = (X_1, X_2, ..., X_n) = (v_1, v_2, ..., v_p)^T$. Then, $\mathcal{L}_B = \mathcal{L}_{row}(X)$. $\mathcal{L}_A$ is a q-dimensional subspace of $\mathbb{R}^n$, represented by projection matrix $(P_A)_{n\times n}$. For all $u \in \mathcal{L}_B$, we can write the row vector $u = g^T X$ for some vector g. Given u, the projection of u onto $\mathcal{L}_A$ $\hat{u}$, has the smallest angle

Figure 15.9: The two subspaces shown here share a line in n-dimensional space. The smallest angle between these two spaces is 0.



between u and $\mathcal{L}_A$ (i.e. smallest $cos^2(\theta(g))$ where $\theta(g)$ is the angle between u and $\hat{u}$). Mathematically, we represent this as follows.

$$\hat{u} = uP_A = g^T XP_A = g^T \hat{X} \text{ where } \hat{X} = XP_A$$

$$cos^2(\theta(g)) = \frac{||\hat{u}||^2}{||u||^2} = \frac{g^T \hat{X}\hat{X}^T g}{g^T XX^T g} = \frac{g^T XP_A X^T g}{g^T XX^T g}$$

We now define $A = XP_A X^T$, $B = XX^T$, and $\tilde{A} = B^{-1/2}AB^{-1/2}$. Then, $B - A = X(I - P_A)X^T = XP_A^{\perp}X^T \geq 0$ and $B^{-1/2}(B - A)B^{-1/2} = I - \tilde{A} \geq 0$. The latter inequality implies the eigenvalues of $\tilde{A}$ are less than or equal to 1. We can organize these eigenvalues such that $1 \geq \lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$. The results are as follows.

1. $g = \xi_1$ and $u_1 = \xi_1^T X$ maximizes $cos^2(\theta(g)) = \lambda_1$

2. $g = \xi_2$ and $u_1 = \xi_2^T X$ maximizes $cos^2(\theta(g)) = \lambda_2$ subject to $g^T A\xi_1 = 0$, $g^T B\xi_1 = 0$

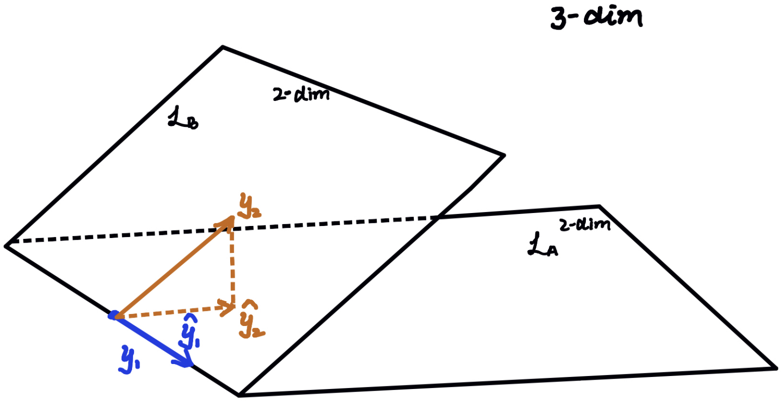
...

k. $g = \xi_k$ and $u_1 = \xi_k^T X$ maximizes $cos^2(\theta(g)) = \lambda_k$ subject to $g^T A\xi_i = 0$, $g^T B\xi_i = 0 \; \forall \; i \in \{1, 2, ..., k-1\}$

...

p. $g = \xi_p$ and $u_1 = \xi_p^T X$ maximizes $cos^2(\theta(g)) = \lambda_p$ subject to $g^T A\xi_i = 0$, $g^T B\xi_i = 0 \; \forall \; i \in \{1, 2, ..., p-1\}$

Geometry ($p = g = 2$, $n = 3$)



$$y_2 \perp y_1, \quad \hat{y}_2 \perp \hat{y}_1$$
$$Y = \Xi^T X = \begin{bmatrix} \xi_1^T X \\ \vdots \\ \xi_p^T X \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_p \end{bmatrix}$$
$$\hat{Y} = \Xi^T \hat{X} = \begin{bmatrix} \xi_1^T \hat{X} \\ \vdots \\ \xi_p^T \hat{X} \end{bmatrix} = \begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_p \end{bmatrix} = \Xi^T X P_A$$
$$YY^T = \Xi^T X X^T \Xi = \Xi^T B = I_p$$
$$\hat{Y}\hat{Y}^T = \Xi^T X P_A X^T \Xi = \Xi^T A = \Lambda$$

For pairs $(y_k, \hat{y}_k)$, $k = 1, 2, ..., p$

(1) The $y_k's$ are mutually orthogonal $||y_k||^2 = 1$.
(2) The $\hat{y}_k's$ are mutually orthogonal $||\hat{y}_k||^2 = \lambda_k$.
(3) All $2p$ vectors are mutually orthogonal $< y_i, \hat{y}_j >= 0$ if $i \neq j$.
(4) The smallest possible angle between $\mathcal{L}_A$ and $\mathcal{L}_B$ is between $y_1, \hat{y}_1$.
The next smallest angle is achieved by $y_2, \hat{y}_2$.
From (1)-(3) we get that:

$$\begin{bmatrix} Y \\ \hat{Y} \end{bmatrix} \begin{bmatrix} Y^T \hat{Y}^T \end{bmatrix} = \begin{bmatrix} I_p & \Lambda \\ \Lambda & \Lambda \end{bmatrix}$$

Comments:

(1) The y and $\hat{y}$ are intrinsic to $\mathcal{L}_A$ and $\mathcal{L}_B$. They do not depend on the choices of the base.

$$X_{p \times n} = \begin{bmatrix} v_1 \\ \vdots \\ v_p \end{bmatrix}$$

(2) You can start from $\mathcal{L}_A$, then project onto $\mathcal{L}_B$ and get the same answer.

Definition 15.35 *$\theta_1, \theta_2, ..., \theta_p$ are called critical angles between $\mathcal{L}_A$ and $\mathcal{L}_B$.*

ISYE 7405: Multivariate Data Analysis **Georgia Tech**

Chapter 16: Enter the title

Lecturer: Shihao Yang

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

16.44 Canonical Correlations

$R_1, R_2, ..., R_p$ and $S_1, S_2, ..., S_q$ are random variables, $R = \left[R_1, \cdots, R_p\right]^\intercal$, $S = \left[S_1, \cdots, S_q\right]^\intercal$. Assume that $\mathbb{E}(R) = \mathbb{E}(S) = 0$, $\Sigma_{RR(p\times p)} > 0, \Sigma_{SS(q\times q)} > 0$.

Recall:

$$\langle u, v\rangle = Cov(u, v), \quad \langle R, S\rangle = \Sigma_{RS}$$

$\hat{R} = \Sigma_{RS}\Sigma_{SS}^{-1}S$ is the linear combination of S that has the highest correlation with $R \equiv R$ projected to space span by S.

$$\begin{aligned}\rho^2(g) &= \frac{g^T Ag}{g^T Bg}\\ g^T Bg &= Cov(g^T R, g^T R)\\ g^T Ag &= Cov(g^T \hat{R}, g^T \hat{R})\end{aligned}$$

where $A = \Sigma_{\hat{R}\hat{R}} = \Sigma_{RS}\Sigma_{SS}^{-1}\Sigma_{SR}$ and $B = \Sigma_{RR}$. We want to find g such that $\rho^2(g)$ is maximized.

$$\begin{aligned}\tilde{A} &= B^{-1/2}AB^{-1/2} = \Sigma_{RR}^{-1/2}\Sigma_{RS}\Sigma_{SS}^{-1}\Sigma_{SR}\Sigma_{RR}^{-1/2}\\ \tilde{A} &= \Gamma\Lambda\Gamma^T\\ \Xi &= B^{-1/2}\Gamma = \Sigma_{RR}^{-1/2}\Gamma\end{aligned}$$

Theorem 16.36

1. *The greatest correlation^2 is achieved by*

$$Y_1 = \xi_1^T R \quad \textit{and} \quad \hat{Y}_1 = \xi_1^T \hat{R} = \xi_1^T \Sigma_{RS}\Sigma_{SS}^{-1}S,$$

the value is λ_1.

2. *The second greatest correlation^2 between linear combination of R and linear combination of S is*

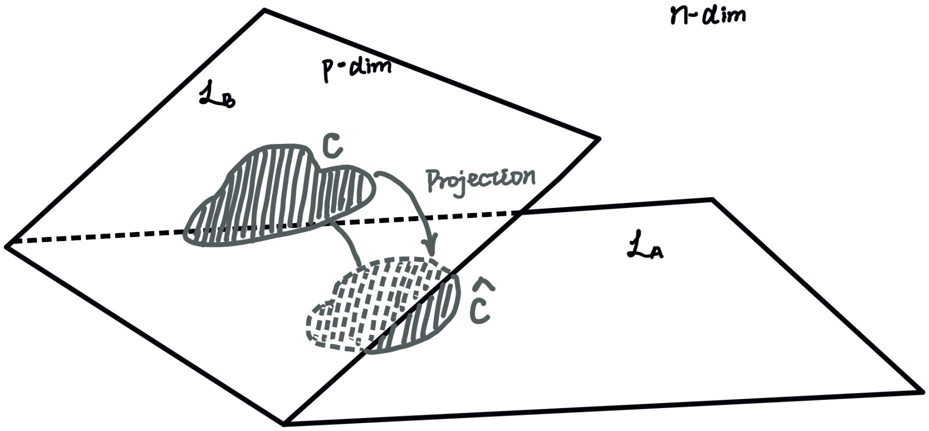
$$Y_2 = \xi_2^T R \quad \textit{and} \quad \hat{Y}_2 = \xi_2^T \hat{R},$$

subject to

$$\begin{aligned}Cov(Y_2, Y_1) = 0, &\quad Cov(\hat{Y}_2, \hat{Y}_1) = 0,\\ Cov(Y_2, \hat{Y}_1) = 0, &\quad Cov(\hat{Y}_2, Y_1) = 0.\end{aligned}$$

3. $\cdots\cdots$

Definition 16.37 $\sqrt{\lambda_1}, \sqrt{\lambda_2}, ..., \sqrt{\lambda_p}$ *is called the canonical correlation between R and S, when $p = 1$. This is multiple correlation.*



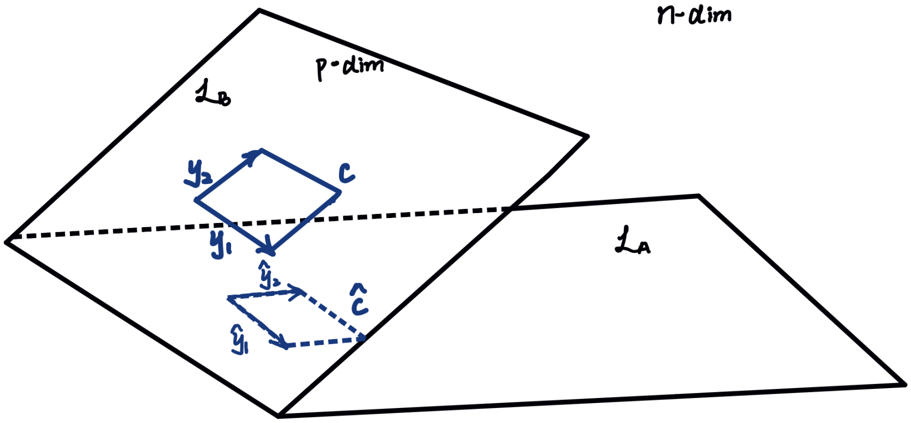
16.45 Projection ratio and critical angles

$p-$dim subspace $\mathcal{L}_B$
$p-$dim volume of C: $Vol_p(C)$
$q-$dim subspace $(p \leq q)$
$p-$dim volume of $\hat{C}$: $Vol_p(\hat{C})$

Theorem 16.38 *The $p-$dim volume of C and $\hat{C}$ are related by $Vol_p(\hat{C}) = Vol_p(C)\prod_{k=1}^{p}\cos\theta_k$, where $\theta_1, \theta_2, ..., \theta_k$ are critical angles.*

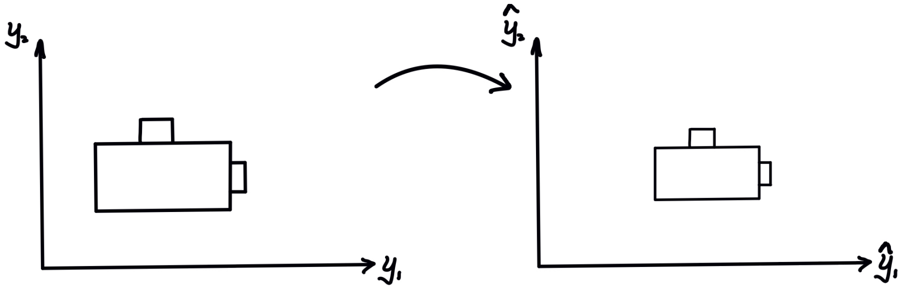
Proof:

1. Rectangles:

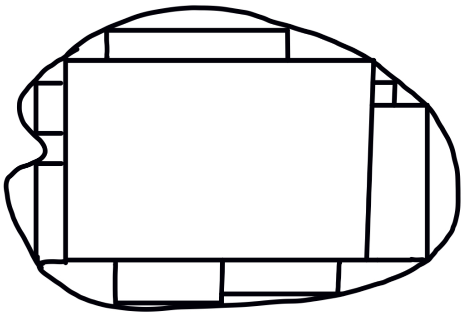


$$
\begin{aligned}
Vol(\hat{C}) &= ||\hat{y}_1|| \cdot ||\hat{y}_2|| \\
Vol(C) &= ||y_1|| \cdot ||y_2|| \\
\frac{Vol(\hat{C})}{Vol(C)} &= \frac{|\hat{y}_1||}{||y_1||} \cdot \frac{|\hat{y}_2||}{||y_2||} = \cos\theta_1 \cdot \cos\theta_2
\end{aligned}
$$

2. Project what we have onto axis $\hat{y}_1, \hat{y}_2$.



3. What we get can be approximated by rectangles, since $\hat{C}$ consists still only of rectangles.



■

ISYE 7405: Multivariate Data Analysis **Georgia Tech**

Chapter 17: Classification

Lecturer: Shihao Yang

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

There are k population group $\pi_1, \pi_2, ..., \pi_k$, each has probability density function $f_j(x)$, $j = 1, 2, ..., k$. The goal: given a future observation x, allocate x into one of the groups.

Classification rule: A division of R^p into disjoint regions $R_1, R_2, ..., R_k$ such that $\bigcup_{i=1}^{k} R_i = R^p$ (or the full space). Therefore, allocate x into π_j if $x \in R_j$.

17.46 Maximum Likelihood Classification Rule

Allocate x to the group that gives the largest likelihood to x, $L(x) = argmax\{f_j(x)\}$

17.46.1 Example: Multinomial distribution ($k = 2$)

$$\begin{aligned}
&\Pi_1 : multi(n, \alpha_1, \alpha_2, ..., \alpha_g)\\
&\Pi_2 : multi(n, \beta_1, \beta_2, ..., \beta_g)\\
&\frac{n!}{x_1!x_2!..x_g!}\prod_{i=1}^{g}\alpha_i^{x_i} \rightarrow \text{Group 1 likelihood}\\
&\frac{n!}{x_1!x_2!..x_g!}\prod_{i=1}^{g}\beta_i^{x_i} \rightarrow \text{Group 2 likelihood}
\end{aligned}$$

$$\Longrightarrow \text{log likelihood ratio} = \sum_{i=1}^{g} x_i \log(\frac{\alpha_i}{\beta_i}) \begin{cases} > 0 & x \text{ to } \pi_1 \\ < 0 & x \text{ to } \pi_2 \end{cases},$$

where $\sum_{i=1}^{g} x_i \log(\frac{\alpha_i}{\beta_i})$ is linear boundary.

17.46.2 Example: Fisher's Linear Discriminant Analysis

Group Π_i: $X \sim N_p(\mu_i, \Sigma)$.
Maximum likelihood classifier (general likelihood rule):

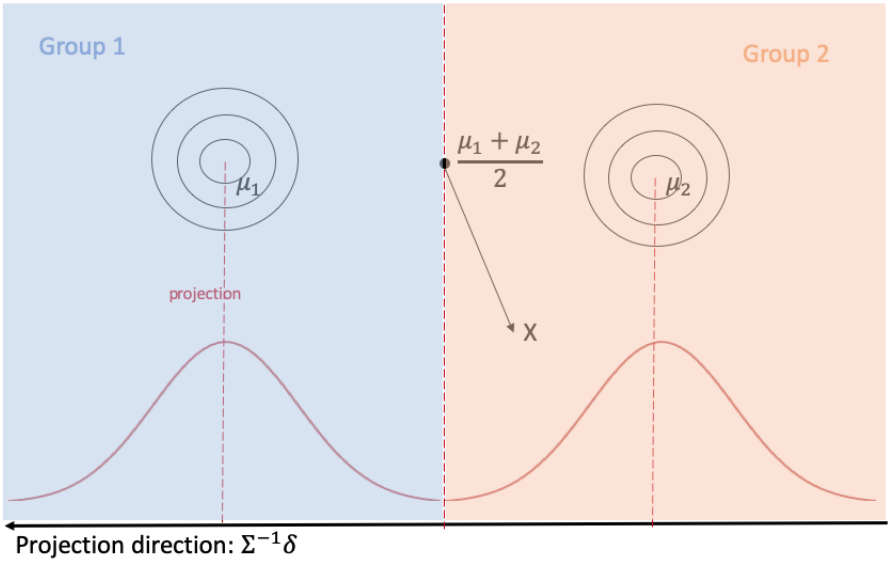
$$L(X) = \arg\min_j (X - \mu_j)^\top \Sigma^{-1} (X - \mu_j).$$

For specific case: $K = 2$,

$$L(X) : (\mu_1 - \mu_2)^\top \Sigma^{-1} (X - \frac{\mu_1 + \mu_2}{2}) \begin{cases} > 0 & \rightarrow X \in \Pi_1 \\ < 0 & \rightarrow X \in \Pi_2 \end{cases}.$$

In other words, if define $\delta := \mu_1 - \mu_2$, the classifier is checking the inner product of $\Sigma^{-1}\delta$ and $(X - \frac{\mu_1+\mu_2}{2})$ to classify X. The following is a pictorial demonstration.

Figure 17.10: Linear Discriminant Analysis Example


Group 1
Group 2
μ_1
$\frac{\mu_1 + \mu_2}{2}$
μ_2
projection
X
Projection direction: $\Sigma^{-1}\delta$

17.47 Bayesian Perspective

Joint model (X, Y), where Y is the class label. Denote
Likelihood: $f_j(X) = f(X|Y = j)$;
Prior: $\pi_j = \pi(Y = j)$;
Posterior: $P(Y = j|X)$.
For specific case: $K = 2$, the log posterior ratio is

$$\begin{aligned}\log \frac{P(Y=0|X)}{P(Y=1|X)} &= \log \frac{\pi_0 \cdot f(X|Y=0)}{\pi_1 \cdot f(X|Y=1)} = \log \frac{\pi_0}{\pi_1} + \log \frac{f(X|Y=0)}{f(X|Y=1)} \\ &(\text{in Normal case} \rightarrow) = \log \frac{\pi_0}{\pi_1} + X^\top \Sigma^{-1}\delta + \text{constant}.\end{aligned}$$

$$\Rightarrow \; P(Y=j|X) \propto \pi(Y=j) \cdot f(X|Y=j). \quad (\text{i.e. posterior} \propto \text{prior} \cdot \text{likelihood})$$

17.48 Sample Version

In practice, we do not know θ in the $f_j(X|\theta)$. Simple solution is to get an estimation $\hat{\theta}$ and plug in $\theta \leftarrow \hat{\theta}$. (For example, in linear discriminant analysis analysis, $\theta = (\mu_1, \mu_2, \Sigma)$ and $\hat{\theta} = (\hat{\mu_1}, \hat{\mu_2}, \hat{\Sigma})$.)
Issue: need to consider the variation in $\hat{\theta}$ caused by estimation uncertainty. This issue is hard to address using frequentist approach but **can be well addressed under Bayesian perspective**.
$\hat{\theta}$ in Bayesian comes from $P_j(\theta|X)$ where X is in the jth group. Then the classification rule

$$L_j(X_{new}|X) = \int f_j(X_{new}|\theta)P_j(\theta|X)d\theta$$

integrates variations of θ. It is called *posterior predictive distribution.*
The advantage of Bayesian setting over frequentist is that Bayesian setting incorporates the variation of θ in the posterior predictive distribution, which frequentist cannot do.

17.49 Logistic Regression

With $Y \in \{1, 2, ..., K\}$ and X, logistic regression has the following form:

$$P(Y = k|X = x) = \frac{\exp\{\beta_{k_0} + \beta_k^\top x\}}{1 + \sum_{l=1}^{K-1} \exp\{\beta_{k_0} + \beta_l^\top x\}},\ k = 1, 2, ..., K - 1,$$

$$P(Y = K|X = x) = \frac{1}{1 + \sum_{l=1}^{K-1} \exp\{\beta_{k_0} + \beta_l^\top x\}}.$$

Equivalently (relative form):

$$\log \frac{P(Y = k|X = x)}{P(Y = K|X = x)} = \beta_{k_0} + \beta_k^\top x,\ k = 1, 2, ..., K - 1.$$

Let's assume $K = 2$ from now on. Then

$$\log \frac{P(Y = 1|X = x)}{P(Y = 2|X = x)} = \beta_0 + \beta_1^\top x.$$

Recall Linear Discriminant Analysis (from previous notes), the decision boundary is

$$(\mu_1 - \mu_2)^\top \Sigma^{-1}(X - \frac{\mu_1 + \mu_2}{2}) \begin{cases} > 0 & \rightarrow X \in \Pi_1 \\ < 0 & \rightarrow X \in \Pi_2 \end{cases}.$$

In the posterior form:

$$\begin{aligned} \log \frac{P(Y = 1|X = x)}{P(Y = 2|X = x)} &= \log \frac{\pi_0}{\pi_1} + \log \frac{f(x|Y = 0)}{f(x|Y = 1)} = \text{constant} + x^\top \Sigma^{-1} \delta \ \ (\text{note:}\delta = \mu_1 - \mu_2) \\ &= \alpha_0 + \alpha_1^\top x. \end{aligned}$$

Therefore, Logistic Regression and Linear Discriminant Analysis both use hyperplane as decision boundary. However, they estimate the coefficients differently.

- Linear Discriminant Analysis: $(X|Y = k) \sim N(\mu_k, \Sigma)$. The full likelihood is

  $$P(X, Y) = P(X|Y) \cdot P(Y).$$

  In estimation, the method gets $\hat{\mu}_k, \hat{\Sigma}$ for boundary $\hat{\Sigma}^{-1}\hat{\delta}$ to maximize full likelihood.

- Logistic Regression: The partial likelihood is

  $$\prod_i P(Y = y_i|X = x_i).$$

  In estimation, the method gets β_0, β_1 directly to maximize partial likelihood.

Comments on the comparison of these two methods:

1. Linear Discriminant Analysis: model bottom up (model the distribution of each group $P(X|Y = k)$, more statistical).

   $$P(X, Y) = P(X|Y) \cdot P(Y),\ \ P(X) = \int P(X, Y)dY.$$

   $P(X)$ is normal mixture.
   If your underlying data is close to normal, then Linear Discriminant Analysis will be more efficient.

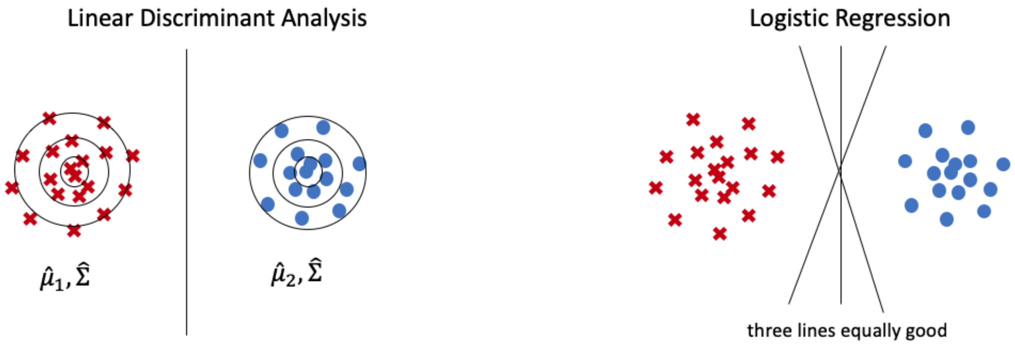
2. Logistic Regression: model top down (model partial data $P(Y=k|X=x)$, more machine learning). $P(X)$ is not specified. Since not assuming normality, it is “robust”. Being “robust” means if the real data is indeed not normal, the method still holds; but if the data is close to normal, it loses efficiency (roughly 30%).

   Also, if data is perfectly separable, it could cause issue for logistic regression, not for linear discriminant analysis.

The following are some pictorial comparisons of these two methods in 2-dimensional cases.

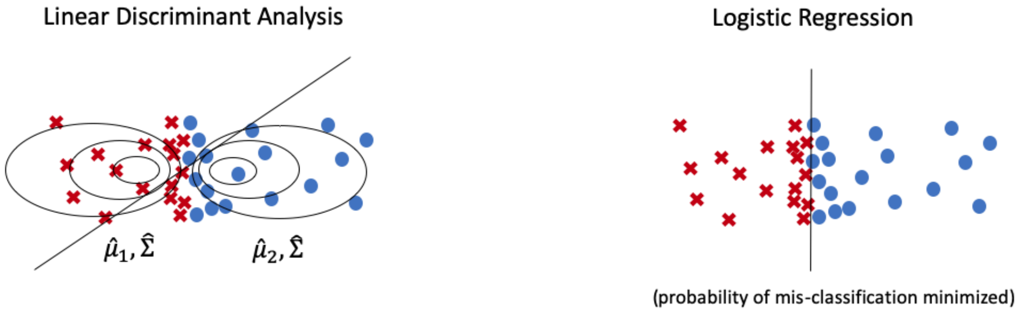
1. When the data is normal (and separable), Linear Discriminant Analysis is more efficient. The three lines for Logistic Regression are equally good, since the likelihood is at global maximum at all of them.

Figure 17.11: Data is normal


Linear Discriminant Analysis
Logistic Regression
$\hat{\mu}_1, \hat{\Sigma}$
$\hat{\mu}_2, \hat{\Sigma}$
three lines equally good

2. When the data is not normal, Logistic Regression is more efficient.

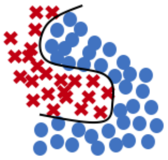
Figure 17.12: Data is not normal


Linear Discriminant Analysis
Logistic Regression
$\hat{\mu}_1, \hat{\Sigma}$
$\hat{\mu}_2, \hat{\Sigma}$
(probability of mis-classification minimized)

3. When the data is not linearly separable, two methods would both fail.

The common feature of these two methods is they both use separating hyperplane. In the next lecture, we will introduce support vector machine which uses data close to the margin of groups to find separating hyperplane.

Figure 17.13: Data is not linearly separable



ISYE 7405: Multivariate Data Analysis **Georgia Tech**

Chapter 18: Hyperplane

Lecturer: Shihao Yang

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

18.50 Separating hyperplace

Let $f(x) = \beta_0 + \beta_1^T x$ and define $L = \{x : f(x) = 0\}$. Then

- For any two points x_1, x_2 on L, we have $\beta_1^T(x_1 - x_2) = 0$.
- If we define the unit normal vector $\beta^* = \frac{\beta_1}{\|\beta_1\|}$, then for any x, the signal distance of x to L satisfies

$$d(x, L) = (\beta^*)^T(x - x_0) = \frac{\beta_1^T}{\|\beta_1\|}(x - x_0) = \frac{\beta_1^T x}{\|\beta_1\|} - \frac{\beta_1^T x_0}{\|\beta_1\|} = \frac{\beta_0 + \beta_1^T x}{\|\beta_1\|},$$

where x_0 is an arbitary point on L. Therefore, $G(x) = \operatorname{sgn}(f(x))$ gives the classification.

18.51 Rosenblatt's Perception learning algorithm

We try to find a hyperplace by minimizing the distance of misclassified points to the boundary.

Define M as the set of misclassified points, then we aim to minimize $D(\beta_0, \beta_1)$ subject to $\|\beta_1\| = 1$, where

$$D(\beta_0, \beta_1) = \sum_{i \in M} |\beta_0 + \beta_1^T x_i|.$$

Moreover, if we label $y_i \in \{-1, 1\}$, we have

$$\begin{aligned} D(\beta_0, \beta_1) = \sum_{i \in M} |\beta_0 + \beta_1^T x_i| &= \sum_{i \in M} -y_i(\beta_0 + \beta_1^T x_i) \\ &= \sum_{i} (-y_i(\beta_0 + \beta_1^T x_i))_+. \end{aligned}$$

Assuming that M is fixed, we take the derivatives and obtain

$$\begin{cases} \dfrac{\partial D}{\partial \beta_1} = -\displaystyle\sum_{i \in M} y_i x_i \\ \dfrac{\partial D}{\partial \beta_0} = -\displaystyle\sum_{i \in M} y_i \end{cases} \tag{18.10}$$

18.51.1 Algorithm: stochastic gradient decent

Instead of computing the sum in (18.10), the gradient decent step is taken after each observation is visited. For each i in M, we update $(\beta_1, \beta_0)^T$ by $(\beta_1, \beta_0)^T = (\beta_1, \beta_0)^T + \eta(y_i x_i, y_i)^T$.

Gradient decent possesses two nice properties: 1. the simple algorithm is easy to code; 2. if the classes are separable, then the algorithm will stop in finite iterations. However, there are some issues about the algorithms. It converges very slowly and never converge in non-separable case. Also, it produces infinite solutions in separable case.

18.52 Optimal separating hyperplane

In separable case, we want to find the hyperplane that maximizes the minimal distance from either class. Recall we have proved for any x, the signal distance to the hyperplane is $\frac{\beta_0+\beta_1^T x}{\|\beta_1\|}$, then we aim to solve the following optimization problem

$$\max_{\beta_0,\beta_1}\{\min_i \frac{y_i(\beta_0+\beta_1^T x)}{\|\beta_1\|}\},$$

or equivalently

$$\max_{\beta_0,\beta_1} C, s.t. \ \frac{y_i(\beta_0+\beta_1^T x)}{\|\beta_1\|}\} \geq C, \ \forall i,$$

or equivalently

$$\max_{\beta_0,\|\beta_1\|=1} C, s.t. \ y_i(\beta_0+\beta_1^T x) \geq C, \ \forall i,$$

or equivalently, take $\|\beta_1\| = \frac{1}{C}$,

$$\max_{\beta_0,\|\beta_1\|=1} \|\beta_1\|, s.t. \ y_i(\beta_0+\beta_1^T x) \geq 1, \ \forall i,$$

We introduce the Lagrange multiplier L_p

$$L_p = \frac{1}{2}\|\beta_1\|^2 - \sum_i \alpha_i(y_i(\beta_0+\beta_1^T x) - 1), \tag{18.11}$$

where α_i are non-negative coefficients. Taking derivatives with β_0 and β_1, we obtain

$$\begin{cases} \beta_1 = \sum_i \alpha_i y_i x_i \\ 0 = \sum_i \alpha_i y_i \end{cases} \tag{18.12}$$

Substitute (18.12) back into (18.11), we have

$$L_p = \sum_i \alpha_i - \frac{1}{2}\sum_i\sum_j \alpha_i\alpha_j y_i y_j x_i^T x_j \tag{18.13}$$

From convex optimization with convex constraint, we also have KKT conditions

$$\alpha_i(y_i(\beta_0+\beta_1^T x_i - 1) = 0, \forall i, \tag{18.14}$$

which means either $\alpha_i = 0$, or $\alpha_i{>}0$, $y_i(\beta_0+\beta_1^T x_i) = 1$.

We can get optimized β_0, β_1 from (18.12),(18.13),(18.14) and we can find only support vectors determine the β_0, β_1. Because (18.12) means β_1 depends only on points with $\alpha_i{>}0$. In other word, β_1 only depends on support points.

18.53 Comparison of Optimal Separating Hyperplane(OSH) with Linear Discriminant Analysis(LDA)

- LDA bottom up. OSH top down.
- OSH more robust to outliers.
- OSH sensitive to support vectors(support points).
- Only partially true that β_0, β_1 only depends on support points("which ones are support points" depends on the entire set.)
- If data is indeed normal, LDA is more efficient.
- If data is not linearly separable, OSH will fail but LDA will still work.

ISYE 7405: Multivariate Data Analysis **Georgia Tech**

Chapter 19: Convex Optimization

Lecturer: Shihao Yang

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

19.54 Convex Optimization

Assume f, g, h are all convex (h must be affine as well to have a convex program).

$$\begin{aligned} \text{argmin} \quad & f(x) && \text{(Primal Objective)} \\ \text{s.t} \quad & g_i(x) \leq 0, \forall i = 1, 2, ..., n_I && \text{(Constraints)} \\ & h_j(x) = 0, \forall j = 1, 2, ..., n_E \\ & x \in \mathbb{R}^p \end{aligned}$$

The set $\mathbb{X} = \{x \in \mathbb{R}^p : g_i(x) \leq 0, \forall i = 1, 2, ..., n_I, h_j(x) = 0, \forall j = 1, 2, ..., n_E\}$ is called the primal feasible set. If a solution exists to the problem is denoted by x^* and it's called primal solution. Respectively, $p^* = f(x^*)$ is the primal optimum.

19.54.1 Lagrangian Function

We define $\forall x \in \mathbb{R}^p, \forall \lambda \in \mathbb{R}^{n_I}$ and $\nu \in \mathbb{R}^E$.

$$L(x, \lambda, \nu) = f(x) + \sum_{i=1}^{n_I} \lambda_i g_i(x) + \sum_{j=1}^{n_E} \nu_j h_j(x)$$

The Lagrange dual function for $\lambda \in \mathbb{R}_+^{n_I}$

$$\Lambda(\lambda, \nu) = \inf_{x \in \mathbb{R}^p} L(x, \lambda, \nu)$$

It holds that the dual function is always less or equal to primal optimum p^*. Indeed, $\forall \lambda \in \mathbb{R}_+^{n_I}, \nu \in \mathbb{R}^E$:

$$\Lambda(\lambda, \nu) = \inf_{x \in \mathbb{R}^p} L(x, \lambda, \nu) \leq \inf_{\tilde{x} \in \mathbb{X}} L(\tilde{x}, \lambda, \nu) = inf_{\tilde{x} \in \mathbb{X}}(f(\tilde{x}) + \sum_i \underbrace{\lambda_i g_i(\tilde{x})}_{\leq 0} + \sum_j \underbrace{\nu_j h_j(\tilde{x}))}_{=0} \leq inf_{\tilde{x} \in \mathbb{X}}(f(\tilde{x})) = p^*$$

Thus $\forall \lambda \geq 0, \Lambda(\lambda, \nu) \leq p^*$

19.54.2 Dual problem

$$\begin{aligned} \text{argmax} \quad & \Lambda(\lambda, \nu) \\ \text{s.t} \quad & \lambda \geq 0, \nu \end{aligned}$$

Dual feasible set: $\mathbb{Z} = \{\lambda, \nu : \lambda \in \mathbb{R}_+^{n_I}, \nu \in \mathbb{R}^{n_E}\}$. Dual solution is denoted by λ^*, ν^* and the dual optimum as $d^* \leq p^*$.

19.54.3 KKT conditions

Necessary and sufficient for $p^* = d^*$. Necessary conditions:

$$p^* = f(x^*) = d^* = g(\lambda^*, \nu^*) = \inf_{x \in \mathbb{R}^p} (f(x) + \sum_i \lambda_i^* g_i(x) + \sum_j v_j^* h_j(x)) \leq f(x^*) + \sum_i \lambda_i^* g_i(x^*) + \sum_j v_j^* h_j(x^*) \leq f(x^*)$$

x^* minimizes $L(x, \lambda^*, \nu^*)$ over all $x \in \mathbb{R}^p$. If L is convex differentiable, then

$$\nabla_x L(x^*, \lambda^*, \nu^*) = 0 \Longleftrightarrow \nabla_x f(x^*) + \sum_i \lambda_i^* \nabla_x g_i(x^*) + \sum_j v_J^* \nabla_x h_j(x^*) = 0 \tag{19.15}$$

We also get

$$\lambda_i^* g_i(x^*) = 0, \forall i = 1, 2, ..., n_I \Longleftrightarrow \lambda_i^* = 0 \text{ or } g_i(x^*) = 0, \forall i = 1, 2, ..., n_I \tag{19.16}$$

Finally, for $x^* \in \mathbb{X}, (\lambda^*, \nu^*) \in \mathbb{Z}$:

$$g_i(x^*) \leq 0, \lambda_i^* \geq 0, h_j(x^*) = 0 \tag{19.17}$$

(19.1), (19.2), (19.3) constitute the KKT conditions.

19.54.4 Sufficient conditions

Suppose $\tilde{x}, \tilde{\lambda}, \tilde{\nu}$ satisfy (19.1), (19.2), (19.3). Then

$$x \rightarrow L(x, \tilde{\lambda}, \tilde{\nu})$$

has gradient zero at $\tilde{x}$. Now, if $x \in L(x, \tilde{\lambda}, \tilde{\nu}$ is convex then $\tilde{x}$ must be the minimizer.

$$g(\tilde{\lambda}, \tilde{\nu}) = \inf_{x \in \mathbb{R}} L(x, \tilde{\lambda}, \tilde{\nu}) = L(\tilde{x}, \tilde{\lambda}, \tilde{\nu}) = f(\tilde{x}) + \sum_i \tilde{\lambda}_i g_i(\tilde{x}) + \sum_j \tilde{\nu}_j h_j(\tilde{x}) = f(\tilde{x})$$

However, we also have $\forall \lambda, \nu \quad g(\lambda, \nu), \leq p^* = \inf_{x \in \mathbb{X}} f(x) \Rightarrow \tilde{\lambda}, \tilde{\nu}, \tilde{x}$ makes equality hold $\Rightarrow \tilde{x}$ is minimizer for $\inf_{x \in \mathbb{X}} f(x)$ and $\tilde{\lambda}, \tilde{\nu}$ minimize $g(\lambda, \nu)$.

19.55 Classification

Classes:

$$Y_1, ..., Y_N \in \{-1, 1\}$$

Features:

$$X_1, ..., X_N$$

Discriminant function $f(x)$ and choosen class $G(x) = sgn\{f(x)\}$.

19.55.1 Model based

Linear Discriminant Analysis. Full likelihood (X, Y)

$$P(Y = k) = \pi_k, \quad P(X|Y = k) \sim \mathcal{N}(\mu_k, \Sigma)$$

$$f(x) = \beta^\top x + \beta_0 = (\hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_2))^\top (x - \frac{\hat{\mu}_2 - \hat{\mu}_1}{2})$$

Logistic Regression:

$$P(Y|X) = \frac{exp(f(x))}{1 + exp(f(x))}, \quad f(x) = \beta^\top x + \beta_0$$

19.55.2 Separating Hyperplanes

minimize misclassification.

- Optimal SH. problem: Fails when data not linearly separable.
- Support vector classification

Idea for Support vector classification: introduce slack variables:

$$\xi_1, \xi_2, ..., \xi_N \quad s.t \quad \xi_i \geq 0, \quad \sum_i \xi_i \leq Constant$$

Optimize margin but with slack:

$$y_i(\beta_0 + \beta_1^\top x_i) \geq C(1 - \xi_i) \quad \forall i$$

$(\xi = 0)$ outside margin, $(0 < \xi_i < 1)$ inside margin, $(\xi > 1)$ wrong classification. Problem:

$$\begin{aligned} \max \quad & C \\ \text{s.t} \quad & \beta_0, ||\beta_1|| = 1 \end{aligned}$$

Let $\tilde{\beta} = \frac{\beta}{c}, \beta_0 = \frac{\beta_0}{c}, ||\tilde{\beta}|| = \frac{1}{c}$. The dual of the above problem is:

$$\begin{aligned} \min \quad & \frac{1}{2}||\beta_1||^2 + r\sum_i \xi_i \\ \text{s.t} \quad & y_i(\beta_0 + \beta_1^\top \xi_i) \geq 1 - \xi_i, \xi \geq 0 \end{aligned}$$

Using KKT conditions: $L_p = \frac{1}{2}||\beta_1||^2 + r\sum_i \xi_i + \alpha_i[(1 - \xi_i) - y_i(\xi_i^\top \beta_1 + \beta_0)] - \sum_i \xi_i \mu_i$.

$$0 = \frac{\partial L_p}{\partial \beta_1} = \beta_1 - \sum_{i=1}^N \alpha_i y_i x_i$$

$$0 = \frac{\partial L_p}{\partial \beta_0} = \sum \alpha_i y_i$$

$$0 = \frac{\partial L_p}{\partial \xi_i} = r - \mu_i - \alpha_i \Rightarrow \alpha_i = r - \mu_i \Rightarrow \alpha_i \in [0, r]$$

Substituting back, the dual objective function becomes:

$$L_D = \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j x_i^\top x_j$$

The problem becomes:

$$\begin{aligned} \max \quad & L_D \\ \text{s.t} \quad & 0 \leq \alpha_i \leq r, \quad \sum \alpha_i y_i = 0 \end{aligned}$$

The rest of the KKT:

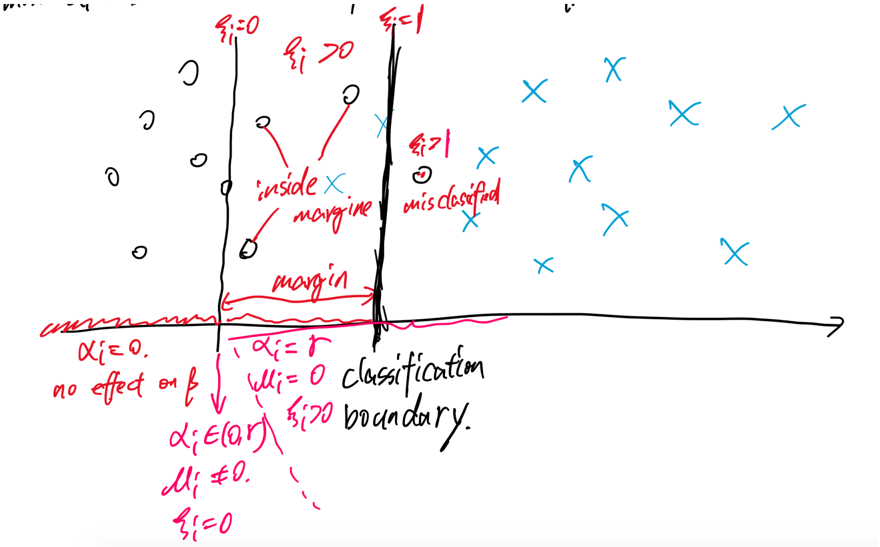
$$\begin{cases} \alpha_i[y_i(x_i^\top \beta_1 + \beta_0) - (1 - \xi_i)] = 0 \\ \mu_i \xi_i = 0 \end{cases}$$

$y_i(x_i^\top \beta_1 + \beta_0) - (1 - \xi_i) \geq 0, \xi_i \geq 0$. $\beta_1 = \sum_i \alpha_i y_i x_i$. β_1 depends only on i for which $\alpha_i \neq 0$. For support vectors $\alpha_i \neq 0 \Rightarrow y_i(x_i^\top \beta_1 + \beta_0) - (1 - \xi_i) = 0$.

For $\alpha_i \neq 0:$

$$\begin{cases} \xi_i = 0 & \text{on the edge} \\ 0 < \xi_i < 1 & \text{in the margin} \\ \xi_i > 1 & \text{misclassified} \end{cases}$$

For those $\alpha_i = 0$, ”inner points” no effect on β.



ISYE 7405: Multivariate Data Analysis **Georgia Tech**

Chapter 20: Support Vector Machine

Lecturer: Shihao Yang

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

20.56 Support Vector Machine

Enlarge feature x_i using basis expansions (e.g. x_i, x_i^2, x_i^3):

$$h(x_i) = (h_1(x_i), h_2(x_i), ...h_m(x_i))$$

Fit Support Vector linear classifier on $h_i(x)$ with discriminant function

$$f(x) = \beta^\top h(x) + \beta_0$$

and decision

$$G(x) = \text{sign}(f(x))$$

Lagrangian dual with the <u>enlarged</u> basis function:

$$\begin{aligned}
L_0 &= \sum_{i=1}^N \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j h(x_i)^\top h(x_j) \\
&= \sum_{i=1}^N \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \underline{\langle h(x_i), h(x_j)\rangle} \\
\beta &= \sum_{i=1}^N \alpha_i y_i h(x_i) \\
f(x) &= \beta^\top h(x) + \beta_0 \\
&= \sum_{i=1}^N \alpha_i y_i h(x_i)^\top h(x) + beta_0 \\
&= \sum_{i=1}^N \alpha_i y_i \underline{\langle h(x_i), h(x)\rangle} + \beta_0
\end{aligned}$$

KKT conditions:

$$\begin{aligned}
\alpha_i(y_i f(x_i) - (1-\xi_i)) &= 0 \\
y_i f(x_i) - (1-\xi_i) &\geq 0
\end{aligned}$$

<u>Observation</u>: h is related to prediction and optimization only through $\langle h(x_i), h(x_j)\rangle$ so we do not need to specify $h(x)$ at all. It is sufficient to specify $\langle h(x), h(x')\rangle$ for all x, x'.

Kernel function:

$$K(x, x') := \langle h(x), h(x')\rangle.$$

As long as I know $K(x, x')$, I don't need to know h.

$K(x, x')$ is symmetric:

$$K(x, x') = K(x', x);$$

and positive definite

$$\forall n,\ \forall x_1, ..., x_n,\ K\left(\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}\right) = \left(\begin{pmatrix} K(x_1,x_1) & K(x_1,x_2) & \cdots & K(x_1,x_n) \\ K(x_2,x_1) & K(x_2,x_2) & \cdots & K(x_2,x_n) \\ \vdots & \vdots & \ddots & \vdots \\ K(x_n,x_1) & K(x_n,x_2) & \cdots & K(x_n,x_n) \end{pmatrix}\right)$$

is positive definite matrix.

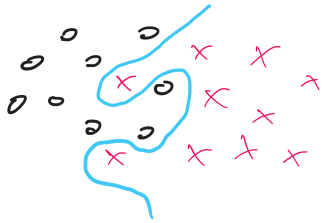
$$\begin{aligned} f(x) &= \sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + \beta_0 \\ L_0 &= \sum_{i=1}^{N} \alpha_i - \sum_{i,j} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \end{aligned}$$

With a kernel function, the support vector classifier is referred to as support vector machine. Some choices of the kernel $K(x_i, x_j)$

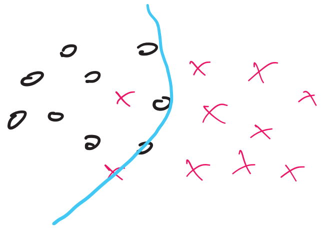
1. d-th polynomial: $K(x, x') = (1 + \langle x, x' \rangle)^d$
2. Gaussian radial kernel: $K(x, x') = \exp\left\{-\frac{\|x-x'\|^2}{c}\right\}$
3. Neural network: $K(x, x') = \tanh(K_1(x, x') + K_2)$ $\quad (tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}})$

20.56.1 Tuning parameter r

large $r \Rightarrow$ fewer possible $\xi \Rightarrow$ less mis-classification $\Rightarrow$ boundary more "wiggly"



small $r \Rightarrow \|\beta\|^2$ small, $\sum \xi_i$ to be large $\Rightarrow$ more positive $\xi \Rightarrow$ more tolerant on mis-classification $\Rightarrow$ boundary will be smooth



Cross-validation to tune r.

20.56.2 Loss function

SVM:

$$\begin{aligned} \min \quad & D(\beta) = \frac{1}{2}\|\beta\|^2 + r\sum_{i=1}^{N}\xi_i \\ \text{s.t.} \quad & \xi_i \geq 0 \\ & y_i(x_i^\top \beta + \beta_0) \geq 1 - \xi_i \\ \Leftrightarrow \quad \min \quad & \sum_{i=1}^{N}(1 - y_i(f(x_i)))_+ + \lambda\|\beta\|^2 \qquad \lambda = \frac{1}{2r} \end{aligned}$$

$$L(y, f) = (1 - y_i f)_+$$

where

$$f = \beta x + \beta_0$$

if linear.

LDA: $L(y, f) = (Y - f)^2 = (1 - Yf)^2$. Linear regression as if $y \in \{-1, 1\}$ are continuous response variable.

Hint: $Y = (-1, +1, -1, +1)^\top$, $X = (x_1, ..., x_n)$, $\bar{X} = 0$, and n_1, n_2 with $n_1 + n_2 = N$, $n_1 = n_2$.

$$\begin{aligned} (\hat{\beta}_0, \hat{\beta}) &= (\tilde{X}\tilde{X}^\top)^{-1}\tilde{X}Y, \qquad \tilde{X} = \begin{pmatrix} \mathbf{1}_n^\top \\ X \end{pmatrix} \\ \tilde{X}\tilde{X}^\top &= \begin{pmatrix} n & 0 \\ 0 & n\hat{\Sigma} \end{pmatrix} \qquad \hat{\Sigma} = \frac{XX^\top}{n} \\ \tilde{X}Y &= \begin{pmatrix} n_1 - n_2 \\ n_1\bar{X}_1 + n_2\bar{X}_2 \end{pmatrix} \\ (\hat{\beta}_0, \hat{\beta}) &= \left(0, \frac{1}{n}\hat{\Sigma}^{-1}(\bar{X}_1 - \bar{X}_2)\right) \\ f(x) &= \hat{\Sigma}^{-1}(\hat{\mu}_2 - \hat{\mu}_1)(x - \frac{\hat{\mu}_2 + \hat{\mu}_1}{2}) \end{aligned}$$

which is the same as LDA.

Logistic regression:

$$\begin{aligned} P(Y = 1|X) &= \frac{e^f}{1 + e^f} = \frac{1}{e^{-f} + 1} \\ P(Y = -1|X) &= \frac{1}{1 + e^f} \\ \Rightarrow P(Y|X) &= \frac{1}{1 + e^{-Yf}} \end{aligned}$$

$$\text{minimize}: \ -\log P(Y|X) = -\log\frac{1}{1 + e^{-Yf}} = L(y, f)$$

Support vector classifier, LDA, logistic regression are all linear classifier trained with different loss function.

20.56.3 Kernel and linear classifier

$$\begin{array}{rcl}
\text{Support Vector Classifier} & & \\
L(y,f) = (1-yf)_+ & \xrightarrow[\text{kernel}]{\text{non-linear}} & \text{Support Vector Machine} \\
\text{Linear Discriminant Analysis} & & \\
L(y,f) = (y-f)^2 = (1-yf)^2 & \xrightarrow[\text{kernel}]{\text{non-linear}} & \text{Gaussian Process Classification} \\
\text{Logistic Regression} & & \\
L(y,f) = \log\left(1+e^{-yf}\right) & \xrightarrow[\text{kernel}]{\text{non-linear}} & ? \\
\text{Linear Regression} & & \\
(y-f)^2 & \longrightarrow & \text{Gaussian Process Classification} \\
X(X^\top X)^{-1}X^\top y \quad (X^\top X \to \langle x, x\rangle) & &
\end{array}$$

20.57 Clustering Analysis

(X_i, Y_i) - classification. $Y_i = \pm 1$ have it at least for training data.
$(X_i)_{i=1}^n$ - clustering. X_i: p dimensions.

Goal: given n observations which are believed to be heterogenous. Want to group them into K homogenous subpopulations, where K is also unknown.

Distances and dissimulating measures - X, X', $d(X, X')$: real valued

d is said to be a dissimulating measures if

1. symmetry: $d(X, X') = d(X', X)$
2. non-negativity: $d(X, X') \geq 0$
3. identification: $d(X, X) = 0$

ISYE 7405: Multivariate Data Analysis **Georgia Tech**

Chapter 21: Distance & Clustering algorithms

Lecturer: Shihao Yang

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

21.58 Distance & dissimulating measures

x, x' $d(x, x')$: real valued. d is said to be a dissimulating measure if:

(1) symmetry: $d(x, x') = d(x', x)$

(2) non-negativity: $d(x, x') \geq 0$

(3) identification: $d(x, x) = 0$

If furthermore, d satisfies

(4) definiteness: $d(x, x') = 0$ iff $x = x'$

(5) triangular inequality: $d(x, x') \leq d(x, y) + d(y, x')$. Then d is called a distance metric.

21.58.1 Quantitative variables

$(x_i)_j \in R$, $x_i \in R^p$

- Euclidean distance: $d(x, x') = \|x - x'\|$

  e.g. L_2 distance: $d(x, x') = \sqrt{\sum_{j=1}^p (x_j - x_j')^2}$
  L_1 distance: $d(x, x') = \sum_{j=1}^p |x_j - x_j'|$

- Pearson distance: $d^2(x, x') = \sum_{j=1}^p \frac{(x_j - x_j')^2}{s_j^2}$

  s_j^2: the variance of $j_t h$ feature
  s_j can be replaced by some robust measure of "spread"
  e.g. $s_j =$ interquantile range

- Mahalanobis distance: $d^2(x, x') = (x - x')\hat{\sum}^{-1}(x - x')$

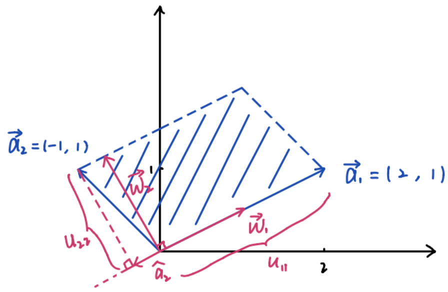
  $\hat{\sum} = I \rightarrow L_2$ distance
  $\hat{\sum} = diag \rightarrow$ Pearson distance
  "How to estimate precision matrix $\sum^{-1}$ (sparse)"

21.58.2 Ordinal variables

e.g. rank of preference
often coded by contiguous integers. such as $1, 2, ..., M$
often treated by replacing i by $\frac{i - \frac{1}{2}}{M}$ & pretend as if they are quantitative in nature

21.58.3 Categorical variables

"look up" table dissimulating matrix



21.59 Clustering algorithms

- Model based

Assume x_i are independent, each comes from any one of g possible sub-populations with density function $f(x_i, \theta_k), k = 1, 2, ..., g$.

likelihood: Let $\gamma = \begin{pmatrix} \gamma_1 \\ \vdots \\ \gamma_n \end{pmatrix}$ to be assignment of x_i. $\gamma_i \in \{1, 2, ..., g\}$. One way is to view γ as parameters.

$$C_k = \{i : \gamma_i = k\}$$

$$L(\gamma, \theta_1, ..., \theta_k) = \prod_{k=1}^{g} \prod_{i \in C_k} f(x_i, \theta_k)$$

The MLE:

$$(\hat{\gamma}, \hat{\theta_1}, ..., \hat{\theta_k}) = \arg\max_{\gamma, \theta_1, ..., \theta_k} L$$

Issue:

(1) combinetorial optimization

(2) treating γ as parameters is arguable

Better approach: view γ as missing data
special case: treat $f(x, \theta_k) \overset{i.i.d}{\sim} \mathcal{N}(\mu_k, \Sigma_k)$ complete data likelihood:

$$\begin{aligned} L(\theta_1, ..., \theta_k; x, \gamma) &= L(\mu_1, ..., \mu_g, \Sigma_1, ..., \Sigma_g, \tau_1, ..., \tau_g) \\ &= \prod_{i=1}^{g} \sum_{k=1}^{g} 1_{\{\gamma_i = k\}} \tau_k \phi(x_i; \mu_k, \Sigma_k)) \end{aligned}$$

log_ likelihood:

$$log\ L = \sum_{i=1}^{n} \sum_{k=1}^{g} 1_{\{\gamma_i = k\}} (log\ \tau_k + log\ \phi(x_i; \mu_k, \Sigma_k))$$

21.59.1 Expectation-Maximization Algorithm

E-step:

$$\begin{aligned} Q(\theta|\theta^{(t)}) &= E_{\gamma|x,\theta^{(t)}}(logL(\theta; x, \gamma)) \\ &= E_{\gamma|x,\theta^{(t)}} \left(\sum_{i=1}^{n} \sum_{k=1}^{g} 1_{\{\gamma_i = k\}} (log\ \tau_k + log\ \phi(x_i; \mu_k, \Sigma_k)) \right) \\ &= \sum_{i=1}^{n} \sum_{k=1}^{g} (E_{\gamma|x,\theta^{(t)}}(1_{\{\gamma_i = k\}}))(log\ \tau_k + log\ \phi(x_i; \mu_k, \Sigma_k)) \\ &= \sum_{i=1}^{n} \sum_{k=1}^{g} (T_{ik}^{(k)})(log\ \tau_k + log\ \phi(x_i; \mu_k, \Sigma_k)) \end{aligned}$$

$$\begin{aligned} T_{ik}^{(t)} &= P(\gamma_i = k | x_i = x_i, \theta^{(t)}) \\ &= \frac{\tau_k^{(t)} \phi(x_i; \mu_k^{(t)}, \Sigma_k^{(t)})}{\sum_{j=1}^{g} \tau_j^{(t)} \phi(x_i; \mu_j^{(t)}, \Sigma_j^{(t)})} \end{aligned}$$

M-step: maximizing

$$\theta^{(t+1)} = \arg\max_{\theta} Q(\theta|\theta^{(t)})$$

$$\arg\max_{\tau_k, \mu_k, \Sigma_k} \sum_{i=1}^{n} \sum_{k=1}^{g} T_{ik}^{(k)} (log\ \tau_k + log\ \phi(x_i; \mu_k, \Sigma_k))$$

where

$$\sum_{i=1}^{n} \sum_{k=1}^{g} T_{ik}^{(k)} (log\ \tau_k + log\ \phi(x_i; \mu_k, \Sigma_k)) = \sum_{i} \sum_{k} T_{ik}^{(k)} log\ \tau_k + \sum_{i} \sum_{k} T_{ik}^{(k)} \left(-\frac{1}{2} \left[(x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k) + det(\Sigma_k) \right] \right)$$

$$\begin{aligned} \tau_k^{(t+1)} &= \frac{\sum_{i=1}^{n} T_{ik}^{(t)}}{\sum_{j=1}^{g} \sum_{i=1}^{n} T_{ij}^{(t)}} = \frac{1}{n} \sum_{i=1}^{n} T_{ik}^{(t)} \\ \mu_k^{(t+1)} &= \frac{\sum_{i=1}^{n} T_{ik}^{(t)} x_i}{\sum_{i=1}^{n} T_{ik}^{(t)}} \\ \Sigma_k &= \frac{\sum_{i=1}^{n} T_{ik}^{(t)} (x_i - \mu_k)^{(t+1)})(x_i - \mu_k^{(t+1)})^T}{\sum_{i=1}^{n} T_{ik}^{(t)}} \end{aligned}$$

local maximizer

21.59.2 Combinatorial Algorithm

Seek the cluster assignment that minimizes some loss function based on dissimulating measures.

A natural choice of loss function

$$W(\gamma) = \frac{1}{2}\sum_{k=1}^{g}\sum_{i,j,\gamma(i)=\gamma(j)=k} d(x_i, x_j), \quad \gamma \to assignment$$

total within-class distance

Equivalent description:

$$B(\gamma) = \frac{1}{2}\sum_{k=1}^{g}\sum_{i,j,\gamma(i)=\gamma(j)\neq k} d(x_i, x_j)$$

total between-class distance

21.59.3 K-means Algorithm

K-means algorithm assumes all variables are quantitative and use Euclidean L_2 distance^2 as dissimulating metric.

$$\begin{aligned} W(\gamma) &= \frac{1}{2}\sum_{k=1}^{g}\sum_{i,j,\gamma(i)=\gamma(j)=k} \|x_i - x_j\|^2 \\ &= \frac{1}{2}\sum_{k=1}^{g} n_k \sum_{i,\gamma(i)=k} \|x_i - \bar{x}_k\|^2 \end{aligned}$$

where

$$\begin{aligned} \bar{x}_k &= \text{mean vector from the k-th cluster} \\ n_k &= \text{cluster size of the k-th cluster} \end{aligned}$$

An iterative desent algorithm: minimizing the following within-class distance

$$\min_{\gamma,m}\sum_{k=1}^{g} n_k \sum_{\gamma(i)=k} \|x_i - m_k\|^2$$

t

ISYE 7405: Multivariate Data Analysis **Georgia Tech**

Chapter 23: Factor Analysis

Lecturer: Shihao Yang

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

23.60 Factor Analysis

A mathematical model that attempts to explain the correlation between a large set of variables in terms of a small number of underlying factors (assuming that the underlying factors are not observed).
The factor model:

$$\begin{array}{ccccccc} Y_p & = & \Lambda f_k & + & u & + & \mu_p \\ \downarrow & & \downarrow & & & & \downarrow \\ \text{observation} & & \text{factors} & & & & \text{mean of } Y_p\text{-constant} \end{array}$$

$\Lambda_{p \times k}$: factor loading matrix, constant across observations.
u: random vector, unique factor (specific).
Assumptions:

$$\begin{aligned} & \mathrm{E}(f) = 0, \quad \mathrm{E}(u) = 0. \\ & \mathrm{Cov}(f) = I, \ \mathrm{Cov}(u) = \mathrm{diag}(\psi_1, ..., \psi_p) = \psi \\ & \mathrm{Cov}(f, u) = 0 \\ \Longrightarrow & \mathrm{E}(Y) = \mu_p, \ \ \mathrm{Cov}(Y) = \Lambda\Lambda^T + \psi \triangleq \Sigma \end{aligned}$$

23.60.1 Factor models are scale-invariant

$$\begin{aligned} \text{suppose } Z = C \cdot Y, & \quad C = \mathrm{diag}(c_1, ..., c_p) \\ = (C\Lambda)f + Cu + C\mu & \end{aligned}$$

23.60.2 Issue: Rotation Invariant

$$\begin{aligned} Y = (\Lambda\Gamma)(\Gamma^T f) + u + \mu, \ \Gamma : \text{orthogonal matrix} \\ f' = \Gamma^T f, \ \Lambda' = \Lambda f, \ Y = \Lambda' f' + u + \mu \end{aligned}$$

We need further constraints to make model identifiable.
Common constraints:

(1) $\Lambda^T \psi^{-1} \Lambda$ is diagonal, otherwise (if not diagonal) we do spectral decomposition on $\Lambda^T \psi^{-1} \Lambda = \Gamma \Lambda^* \Gamma^T$, then we take this Γ. Let $\tilde{\Lambda} \triangleq \Lambda\Gamma$, then $\tilde{\Lambda}^T \psi^{-1} \tilde{\Lambda}$ is to be diagonal.

or (2) $\Lambda^T D^{-1} \Lambda$ is diagonal, where $D = \mathrm{diag}(\Sigma) = \mathrm{diag}(\sigma_{11}, \sigma_{22}, ..., \sigma_{pp})$; If data is standardized, i.e., $\mathrm{diag}(\Sigma) = (1, 1, ..., 1)$, (2) then will be constraint: $\Lambda^T \Lambda = $ diagonal.

23.60.3 Count # of Free Parameters

$$\Sigma : \frac{1}{2}\, p\,(p+1)\,\text{for free}$$
$$\Lambda \,\&\, \psi : p\,k + p\,\text{for free}$$

Constraint (1) or (2) requires a $k \times k$ matrix $((\Lambda^T \psi^{-1} \Lambda)_{k\times k}, (\Lambda^T D^{-1} \Lambda)_{k\times k})$ to be diagonal.

$$\frac{1}{2}k(k-1)\text{ constraints}$$

$$\begin{aligned}\text{Total freedom : } & \frac{1}{2}p(p+1) - [pk + p - \frac{1}{2}k(k-1)] \\ &= \frac{1}{2}(p-k)^2 - \frac{1}{2}(p+k).\end{aligned}$$

No guarantee this is > 0. For the usual usecase, $k \ll p$, then the model is “fine”.

23.60.3.1 Example

$p = 3, k = 1$, degree of freedom $= 0$, the solution is unique.
$\psi \geq 0$, otherwise cannot be covariance matrix.

23.60.4 How to estimate

$$\hat{\mu} = \bar{Y}$$

“MLE” by assuming

$$Y \sim \mathcal{N}(\mu, \Sigma)$$

where

$$\Sigma = \Lambda\Lambda^T + \psi$$

subject to constraints

$$\begin{aligned}\Lambda^T \psi^{-1} \Lambda &= \text{diag} \\ \text{or} \quad \Lambda^T D^{-1} \Lambda &= \text{diag}\end{aligned}$$

$$f_Y(y) = \frac{1}{(\sqrt{2\pi})^p \det(\Lambda\Lambda^T + \psi)^{\frac{1}{2}}} \exp(-\frac{1}{2}(y-\mu)(\Lambda\Lambda^T + \psi)^{-1}(y-\mu))$$

$$\ell(\Lambda, \psi; y_1, ..., y_n) = \sum_{i=1}^{n} \log_Y(y_i)$$

$$\hat{\Lambda}, \hat{\psi} = \arg\max_{\Lambda, \psi} \ell(\Lambda, \psi)$$

Let $\hat{\Sigma}$ be sample variance-covariance matrix. $\hat{\Sigma} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})(y_i - \bar{y})^T$.
We know $\hat{\Sigma}$ is MLE without $\Sigma = \Lambda\Lambda^T + \psi$ constraints.

$$\max_{\Lambda} \ell(\Lambda, \psi) \;\Leftrightarrow \min_{\Lambda} \text{loss}(\hat{\Sigma}, \Lambda),$$

where loss function is sum of square of all off-diagonal elements.

$$\Leftrightarrow \min_{\Lambda} \sum_{i<j} (\hat{\Sigma} - \Lambda\Lambda^T)^2_{ij}$$

If not normal, this estimate can be thought as Method of Moments estimation. Or, can be thought as quasi-MLE if Gaussian is mis-specified.
"MLE "(or quasi-MLE) helps for hypothesis testing to decide k.

$$\text{Generalized likelihood testing: } -2\log LR \sim \chi^2_{df_1 - df_2}$$

In the special case where data are standardized, $\text{diag}(\hat{\Sigma}) = (1, 1, ..., 1)$

- then $\Lambda\Lambda^T$ is also the model correlation matrix (on off-diag elements).
- and $\text{diag}(\Lambda\Lambda^T)$ is the proportion of variance explained by factors.

Comparison:	F A	PCA
(1)	model based	model free
(2)	degree of freedom when choosing k	choose as many as you like
(3)	uniqueness(rotation)	no such issue
(4)	don't interpret the factor, but can be used for comparison, prediction.	interpretation: direction with max variability across data.
(5)	$\min_{\Lambda} \sum_{i<j} (\hat{\Sigma} - \Lambda\Lambda^T)^2_{ij}$ minimize off-diag difference $\Sigma = \Lambda\Lambda^T + \psi$ model for correlation	$\min \sum_{i,j} (\hat{\Sigma} - \Gamma(k)\Lambda(k)\Gamma(k)^T)^2_{ij}$ for the first k principal components. all elements difference $\Sigma = \Gamma(k)\Lambda(k)\Gamma(k)^T$ model for entire variance covariance

23.61 Independent Component Analysis (ICA) [Signal Processing]

Model $Y = \Lambda f + u + \mu$
Assumption: f and u are independent.

1. $f_i \amalg f_j, f_i \amalg u_j, u_i \amalg u_j$
2. f_i's are non-Gaussian.

ISYE 7405: Multivariate Data Analysis **Georgia Tech**

Chapter 24: Independent Component Analysis & Gaussian Process

Lecturer: Shihao Yang

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

24.62 Independent Component Analysis (ICA). [signal processing]

- Model

$$Y_P = \Lambda_{p\times k} f_k + u + \mu$$

- Assumption

  f and u are independent.

  - $f_i \perp f_j . f_i \perp u_j . u_i \perp u_j$
  - f_i' s are non-Gaussian

- Simple case

  Cocktail Party Problem:k person in the room with P microphone

$$Y = \Lambda f , p = k.$$

g: joint distribution of f. g_i: distribution of f_i under in dependence:

$$g(f) = \prod_{i=1}^{k} g_i(f_i)$$

Use KL divergence to measure distance between g and $\prod_{i=1}^{k} g_i(f_i)$

$$KL(g,h) = -Eg\left(\log \frac{h}{g}\right)$$

Then

$$\begin{aligned} KL\left(g, \prod_{i=1}^{k} g_i\right) &= -\int \left(\sum_{i=1}^{k} \log g_i(f_i) - \log g\left(\underset{\sim}{f}\right)\right) g\left(\underset{\sim}{f}\right) d\underset{\sim}{f} \\ &= \sum_{i=1}^{k} \int -\log g_i(f_i)\, g(f_i)\, df_i - \int -\log g\left(\underset{\sim}{f}\right) g\left(\underset{\sim}{f}\right) d\underset{\sim}{f} \\ &= \sum_{i=1}^{k} \text{Entropy}(g_i) - \text{Entropy}(g) \end{aligned}$$

Since $P = K$

$$Y = \Lambda f . f = \Lambda^{-1} Y.$$

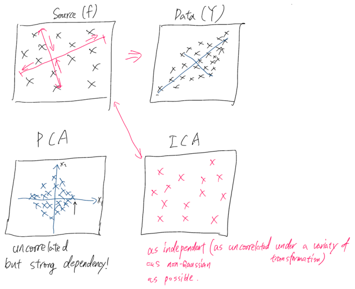
Entropy is in variant under linear transformation. So,

$$\text{Entropy } (g)\text{=Entropy}(Y) \text{ and is observed.}$$

$$\min \; KL\left(g, \prod_{i=1}^{k} g_i\right) \Leftrightarrow \min \sum_i \text{Entropy}(f_i) \quad (*)$$

Given mean and variance, Gaussian maximizes the entropy.

$$(*) \Leftrightarrow \text{Given mean and variance, we are moving away from Gaussian.}$$



24.63 Gaussian Process

- Linear Regression
- Ridge Regression
- Bayesian Linear Regression
- Gaussian Process Regression

24.63.1 Linear Regression

- Response

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

- Predictor (Feature Matrix)

$$X_{n \times p} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

where $x_1 \cdots x_n$ are row vectors of length p (p features)

- Model

$$Y = X\beta + \varepsilon$$
$$\varepsilon \sim \left(0, \sigma^2\right) \text{ doesn't have to be Gaussian}$$

- Training (Estimation of β)

  Minimize loss function in MSE (mean square error)

$$L_{MSE}\left(\beta\right) = \tfrac{1}{n}\sum_{i=1}^{n}\left(y_i - x_i\beta\right)^2 = \tfrac{1}{n}\left(Y - X\beta\right)^T\left(Y - X\beta\right)$$

  In the case where $\varepsilon \sim N\left(0, \sigma^2\right)$, the log likelihood can be loss function

$$\text{loglike}\left(\beta\right) = \sum_{i=1}^{n}\left[-\tfrac{1}{2}\log\left(2\pi\sigma^2\right) - \tfrac{1}{2}\tfrac{\left(y_i - x_i\beta\right)^2}{\sigma^2}\right]$$

  which is closely related to L_{MSE}

  Then,

$$\arg\min_{\beta} L_{MSE}\left(\beta\right) \Leftrightarrow \arg\max_{\beta} loglike\left(\beta\right)$$

  Check:

$$O = \tfrac{\partial L_{MSE}}{\partial\beta} = \tfrac{2}{n}\left(-X^T\right)\left(Y - X\beta\right) \Rightarrow \hat{\beta}^{OLS} = \left(X^TX\right)^{-1}X^TY$$

$$\hat{Y} = X\hat{\beta} = X\left(X^TX\right)^{-1}X^TY$$

  $\hat{Y}$ is the projection of Y into $\mathcal{L}_{col}\left(X\right)$

- *Questions*

  What if X^TX is (nearly) singular?

  Undesirable: small change in X or Y $\Rightarrow$big change in $\hat{\beta}$ and $\hat{Y}$

24.63.2 Ridge Regression

- Idea: Add constant on diag $\left(X^TX\right)$

$$\hat{\beta}^{ridge} = \left(X^TX + \lambda I\right)^{-1}X^TY$$
$$\hat{Y}^{ridge} = X\hat{\beta}^{ridge} = X\left(X^TX + \lambda I\right)^{-1}X^TY$$

  where λ is tuning parameter

- Bias -Variance trade-off (HW4)

  $\hat{\beta}^{ridge}$ is biasing towards 0, but has less variance.

$$\text{MSE} = \text{bias}^2 + \text{variance}$$

  $\hat{\beta}^{ridge}$ can have smaller MSE to the true β^* for some λ in the sweetspot

- Choice of λ: cross$-$validation
- Connection to noise$-$injected data

  What if feature data has nolse injected ?(noise in X)

  - data contamination
  - survey data inaccuracy
  - privacy (researcher add noise intentionally)

Define,

$$\hat{\beta}^{OLS} = \left(\tilde{X}^T\tilde{X}\right)^{-1}\tilde{X}^TY$$

$$= \left(\tfrac{1}{n}(X+Z)^T(X+Z)\right)^{-1}\tfrac{1}{n}(X+Z)^TY = \left(\tfrac{1}{n}X^TX+\tfrac{1}{n}Z^TZ+\tfrac{1}{n}X^TZ+\tfrac{1}{n}Z^TX\right)^{-1}\left(\tfrac{1}{n}X^TY+\tfrac{1}{n}Z^TY\right)$$

$$\approx \left(\tfrac{1}{n}X^TX+\sigma_Z^2 I\right)\left(\tfrac{1}{n}X^TY\right) = \left(X^TX+n\sigma_Z^2 I\right)\left(X^TY\right)$$

In real-world big data, X almost always have noise ⇒ implicit ridge penalty is applied.

Many ML tricks involving noise injection in training (stochastic GD dropout layer) which the model robust can be though of as implicit regularization.

ISYE 7405: Multivariate Data Analysis **Georgia Tech**

Chapter 26: LASSO Regression

Lecturer: Shihao Yang

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

26.64 (From previous lecture) Summary of Gaussian Process

Gaussian Process is appealing because it:

1. Quantify uncertainty, which includes
   - Intrinsic noise
   - Errors in parameter estimation
2. Non-parametric regression: can model any arbitrary functions
3. Introduce kernels into regression:
   - GP = Ridge Regression + kernel base
4. Simple and straightforward linear algebra implementations.

Downside: computational complexity (for K^{-1})

26.65 LASSO (Least Absolute Shrinkage and Selection Operator)

26.65.1 Key Feature

LASSO scales well with number of parameters p:

- statistical error
- computational cost

26.65.2 Setup

- Feature matrix: X is standardized so that column mean is 0 and variance is 1, i.e., $\sum_{i=1}^n X_{ij} = 0$, $\frac{1}{n}\sum_{i=1}^n X_{ij}^2 = 1$, for $j = 1, \ldots, p$
- Response: Y, $\sum_{i=1}^n Y_i = 0$

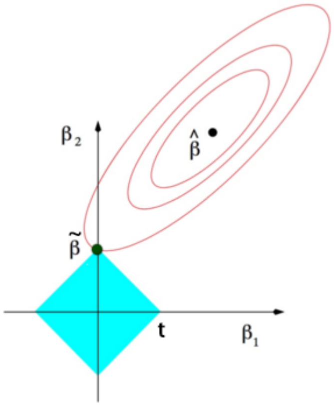
26.65.3 Loss Function

$$\hat{\beta}^L_\lambda = \text{argmin}_\beta Q_\lambda(\beta), \quad \text{where } Q_\lambda(\beta) = \frac{1}{2n}||Y - X\beta||_2^2 + \lambda||\beta||_1$$

Dual form:

$$\begin{aligned} &\text{argmin} \quad \frac{1}{2n}||Y - X\beta||_2^2 \\ &\text{s.t.} \qquad ||\beta||_1 \leq t, \text{for some } t \end{aligned}$$

The L_2 loss function is more like a ellipse, where the L_1 ball has corners, where most components are exactly zero, which means LASSO will give sparse solutions.



The loss functions is in red curve and the constraint is in blue. The constraint in each quadrant is a linear function, and formulate a diamond shape. The area of blue diamond is $\{\beta : ||\beta||_1 \leq t\}$. $\hat{\beta}$ is the unconstrained optimum, where $\tilde{\beta}$ is the constrained optimum which gives $\beta_1 = 0$.

26.65.3.1 Sparse Solutions

Any form $\{\beta : \sum_{j=1}^p |\beta_j|^q \leq t\}$ for $q \leq 1$ will have corners. When $q \geq 1$, then corner points will become smooth. $q = 1$ is the only convex constraint set.

26.65.4 Convex Optimization (Revisited)

Given a optimization problem

$$\begin{aligned} &\text{argmin} \quad f(x) \\ &\text{s.t.} \qquad g_i(x) \leq 0, h_j(x) = 0 \end{aligned}$$

The KKT conditions for the Lagrange function $L(x, \lambda, v) = f(x) + \sum_i \lambda_i g_i(x) + \sum_j v_j h_j(x)$ is

1. $0 = \nabla_x L(x, \lambda, v)$
2. $\lambda_i g_i(x) = 0$
3. $g_i(x) \leq 0$, $\lambda_i \geq 0$, $h_j(x) = 0$

KKT conditions cannot directly be applied to LASSO, we need to introduce the subgradient

$$f(y) \geq f(x) + \nabla f(x)^T (y - x)$$

26.65.4.1 Subgradient

$$\partial f(x) = \{v : \forall y, f(y) \geq f(x) + v^T(y-x)\}$$

if f is differentiable at x, then $\partial f(x) = \{\nabla f(x)\}$.
$x^* = \text{argmin} f(x) \iff \forall y, f(y) \geq f(x^*) \iff f(y) \geq f(x^*) + 0(y-x) \iff 0 \in \partial f(x^*)$. Then KKT(1) now becomes $0 \in \partial_x L(x, \lambda, v)$.
Back to the loss fucntion of LASSO:

$$Q_\lambda : Q_\lambda(\beta) = \frac{1}{2n}||Y - X\beta||_2^2 + \lambda||\beta||_1$$

The subdifferential of the L_1 norm is

$$\partial||x||_1 = \{v \in \mathbb{R}^p : ||v||_\infty \leq 1, v_{s(x)} = sgn(x_{s(x)})\}$$

where $s(x) = \{j \in \{1, \dots, p\}; x_j \neq 0\}$ The subdifferential of the Q_λ at some vector $\beta \in \mathbb{R}^p$ is

$$\partial Q_\lambda = \{\frac{1}{n}X^T(Y - X^T\beta) + \lambda v : v \in \partial||\beta||_1\}$$

i.e. for $j = 1, \dots, p$, $v_j = sgn(\beta_j)$ if $\beta_j \neq 0$ otherwise $v_j \in [-1, 1]$.
Then KKT(1) becomes:

$$0 \in \partial Q_\lambda(\hat{\beta}^L_\lambda) \iff \exists \hat{v} \text{ s.t. } \hat{v}_j = sgn(\hat{\beta}^L_{\lambda,j}) \text{ if } \hat{\beta}^L_{\lambda,j} \neq 0, \text{ and } \hat{v}_j \in [-1,1], \text{ otherwise}$$
$$\text{s.t. } \frac{1}{n}X^T(Y - X^T\hat{\beta}^L_\lambda) = \lambda\hat{v}$$

This is referred as the KKT condition for the LASSO.