L03: Parameters of conditional distributions

1. Parameters of conditional distributions

(1) Definitions

Suppose $\begin{pmatrix} X \\ Y \end{pmatrix}$ has a joint distribution where $X \in R^p$ and $Y \in R^q$.

Then $E(Y|x) = (E(Y_i|x))_{q\times 1} = \left(\iint_{R^q} y_i f_{Y|x}(y) dy_1, ..dy_q\right)_{q\times 1}$

$\text{Cov}(Y|X) = E(YY'|x) - E(Y|x)[E(Y|x)]'$.

Here $E(YY'|x) = (E(Y_iY_j|x))_{q\times q} = \left(\iint_{R^q} y_i y_j f_{Y|x}(y) dy_1, ..dy_q\right)_{q\times q}$

Comments: $Y \sim (E(Y),\, \text{Cov}(Y))$, $Y|X \sim (E(Y|X),\, \text{Cov}(Y|X))$ where $E(Y|X)$ and $\text{Cov}(Y|X)$ are vector-valued and matrix-valued functions of X, and hence are still random. So one can further consider $E(Y|X) \sim (E[E(Y|X)],\, \text{Cov}(E(Y|X))$

(2) Relations

(i) $E[E(Y|X)] = E(Y)$.

Proof. We show $E[E(Y_i|X)] = E(Y_i)$.

$$\begin{aligned} E[E(Y_i|X)] &= \iint_{R^p} E(Y_i|x)\, f_X(x_1, .., x_p)\, dx_1, ..., dx_p \\ &= \iint_{R^p} [\iint_{R^q} y_i f_{Y|x}(y_1, .., y_q)\, dy_1, ...dy_q]\, f_X(x_1, .., x_p)\, dx_1, \cdots, dx_p \\ &= \iint_{R^{p+q}} y_i f(x_1, ..., x_p, y_1, ..., y_q)\, dx_1, \cdots, dx_p, dy_1, \cdots, dy_q \\ &= E(Y_i) \end{aligned}.$$

(ii) $E[\text{Cov}(Y|X)] + \text{Cov}[E(Y|X)] = \text{Cov}(Y)$

$$\begin{aligned} \textbf{Proof.}\text{Cov}(Y) &= E(YY') - E(Y)[E(Y)]' \\ E[\text{Cov}(Y|X)] &= E\{E(YY'|X) - E(Y|X)[E(Y|X)]'\} \\ &= E(YY') - E\{E(Y|X)[E(Y|X)]'\} \\ \text{Cov}[E(Y|X)] &= E\{E(Y|X)[E(Y|X)]'\} - E[E(Y|X)]\{E[E(Y|X)]\}' \\ &= E\{E(Y|X)[E(Y|X)]'\} - E(Y)[E(Y)]'. \end{aligned}$$

So $E[\text{Cov}(Y|X)] + \text{Cov}[E(Y|X)] = E(YY') - E(Y)[E(Y)]' = \text{Cov}(Y)$.

Ex1: For $\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix}\right)$, $Y \sim N(\mu_y,\, \Sigma_{yy})$

$Y|X \sim N\left(\mu_y + \Sigma_{yx}\Sigma_{xx}^{-1}(X - \mu_x),\, \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}\right)$.

$E(Y|X) = \mu_y + \Sigma_{yx}\Sigma_{xx}^{-1}(X - \mu_x) \sim N\left(\mu_y,\, \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}\right)$.

So $E[E(Y|X)] = \mu_y = E(Y)$ and

$E[\text{Cov}(Y|X)] + \text{Cov}[E(Y|X)] = \left(\Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}\right) + \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy} = \Sigma_{yy} = \text{Cov}(Y)$

2. Independence

(1) Independence

$$\begin{aligned} X \in R^p \text{ and } Y \in R^q \text{ are independent } &\overset{def}{\Longleftrightarrow} f_X(x) = f_{X|y}(x) \Longleftrightarrow f_X(x) = \tfrac{f(x,y)}{f_Y(y)} \\ &\Longleftrightarrow f(x,y) = f_X(x)\, f_Y(y) \neq 0 \\ &\Longleftrightarrow f_Y(y) = f_{Y|x}(y). \end{aligned}$$

(2) Impact of independence on parameters

$X \in R^p$ and $Y \in R^q$ are independent $\Rightarrow$ X and Y are uncorrelated.

Proof. We show $\text{cov}(X_i,\, Y_j) = E(X_iY_j) - E(X_i)E(Y_j) = 0$.

$$\begin{aligned} E(X_iY_j) &= \iint_{R^{p+q}} x_i y_j f(x_1, .., x_p, y_1, .., y_q) dx_1, .., dx_p, dy_1, ..., dy_q \\ &= \iint_{R^{p+q}} x_i y_j\, f_X(x_1, .., x_p)\, f_Y(y_1, ..., y_q) dx_1, .., dx_p, dy_1, ..., dy_q \\ &= \iint_{R^p} x_i f_X(x_1, .., x_p)\, dx_1, .., dx_p \iint_{R^q} y_j f_Y(y_1, ..., y_q), dy_1, ..., dy_q \\ &= E(X_i)E(Y_j) \end{aligned}$$

(3) Independence for normal vectors

Suppose $\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix}\right)$. Then

$$X \text{ and } Y \text{ are independent} \iff X \text{ and } Y \text{ are uncorrelated}$$

Proof. Only give the sketch for $\Leftarrow$.

X and Y are uncorrelated $\Longrightarrow \Sigma_{xy} = \mathrm{Cov}(X,\, Y) = 0 \Longrightarrow \Sigma_{xy} = 0$ and $\Sigma_{yx} = \Sigma'_{xy} = 0$.

With $\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \Sigma_{xx} & 0 \\ 0 & \Sigma_{yy} \end{pmatrix}\right)$, $X \sim N(\mu_x,\, \Sigma_{xx})$ and $Y \sim N(\mu_y,\, \Sigma_{yy})$ one can check $f(x_1, ..., x_p, y_1, ..., y_p) = f_X(x_1, ..., x_p) f_Y(y_1, ..., y_q)$.

Ex2: In 4.3 on page 201 $X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim N(\mu,\, \Sigma)$ where $\Sigma = \begin{pmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{pmatrix}$.

(i) Are X_1 and X_2 independent?
$\mathrm{cov}(X_1,\, X_2) = -2 \neq 0$. So X_1 and X_2 are not independent.

(ii) Are X_1 and X_3 independent?
$\mathrm{cov}(X_1,\, X_3) = 0$. So X_1 and X_3 are independent.

(iii) Are X_1 and $2X_1 + X_2 + X_3$ independent?
$\mathrm{Cov}(X_1,\, 2X_1 + X_2 + X_3) = (1,\, 0,\, 0) \begin{pmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{pmatrix} \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix} = 0.$
So X_1 and $2X_1 + X_2 + X_3$ are independent.

3. Extended definitions for normality and independence

(1) Generalized normal vector

$X \sim N(\mu,\, \Sigma) \overset{def}{\iff} X = AZ_r + \mu$ where $Z_r \sim N(0,\, I_r)$ and $AA' = \Sigma$

So a vector is normal if it is transformed from a normal vector with pdf by function $Ax + b$.

Under this definition there is a convenient transformation for normal vectors,

$$X \sim N(\mu,\, \Sigma) \iff AX + \beta \sim N(A\mu + \beta,\, A\Sigma A') \text{ for all } A \text{ and } \beta$$

Caution: For a p-dimensional normal vector, its support may not be R^p.

Ex3: For $Z \sim N(0,\, 1^2)$, $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} Z \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}\right)$ has support $x_1 = x_2$.

(2) Concept of independence

$g(X)$ and $h(Y)$ are independent if X and Y are independent by pdfs or pmfs.

Under this extended definition, X and Y are independent $\Longrightarrow$ X and Y are uncorrelated.

For normal vectors $(X,\, Y)'$, X and Y are independent $\iff$ X and Y are uncorrelated.

(3) Independence from normal vectors

Suppose $X \sim N(\mu,\, \Sigma)$

(i) AX and BX: $A\Sigma B' = 0 \iff AX$ and BX are independent.
Pf: AX and BX are independent $\Longrightarrow \mathrm{Cov}(AX,\, BX) = 0 \iff A\Sigma B' = 0$.

(ii) AX and $X'BX$ where $B' = B$: $A\Sigma B = 0 \Longrightarrow AX$ and $X'BX$ are independent.
Pf: By the compact form of EVD, $B = P\Lambda P'$. So $A\Sigma B = A\Sigma P\Lambda P' = 0 \Longrightarrow A\Sigma P = 0$.
So AX and $P'X$ are independent.
Hence AX and $X'BX = (P'X)'\Lambda(P'X)$ are independent.

(iii) $X'AX$ and $X'BX$ where $A' = A$ and $B' = B$
$A\Sigma B = 0 \Longrightarrow X'AX$ and $X'BX$ are independent.
Pf: Skipped.

L04: Principal components

1. Principal components for X

   (1) Components of $X \sim (\mu, \Sigma)$
   The importance of the components of $X \in R^p$ is usually measured by their variances. The larger the variance, the more information provided by the component.
   The information provided by correlated components overlapped.
   So an ideal vector would have $\Sigma = V_x$ with non-increasing diagonal elements.

   (2) Principal components of X
   Among all linear combinations of the components of X with unit vector coefficients, select one with the largest variance and call it the first principal component of X.
   Once the first k principal components of X are selected, among all linear combinations of the components of X with unit vector coefficients and uncorrelated to the first k principal components already selected, select one with the largest variance and call it the $k+1$ the principal component of X.

2. Eigenvalue decomposition and principal components

   (1) Eigenvalue decomposition of Σ
   If $\Sigma v = \lambda v$ where $0 \neq v \in R^p$, then λ is an eigenvalue of Σ, v is an eigenvector of Σ belonging to the eigenvalue λ.
   $\Sigma > 0$ has p eigenvalues, all are positive numbers. Those eigenvalues are the roots of the characteristic polynomial $|\Sigma - \lambda I|$, i.e., the solutions to the characteristic equation $|\Sigma - \lambda I| = 0$.
   All vectors except 0 in the eigenspace $\mathcal{N}(\Sigma - \lambda I)$ are eigenvectors belonging to λ.
   For real symmetric Σ, the eigenvectors belonging to different eigenvalues are orthogonal.
   So one can find eigenvalue matrix $\Lambda = \text{diag}(\lambda_1, ..., \lambda_p)$ with $\lambda_1 \geq \cdots \geq \lambda_p$ such that all diagonal elements of Λ are eigenvalues of Σ; and eigenvector matrix $P \in R^{p\times p}$ such that the P_i, ith column of P, is the eigenvector for λ_i, $\|P_i\| = 1$, and all columns of P are orthogonal. Thus P is an orthogonal matrix, i.e., $P' = P^{-1}$.
   Clearly $\Sigma P = P\Lambda \Longleftrightarrow \Sigma = P\Lambda P'$, called the EVD for Σ.
   $|\Sigma - \lambda I|$ is a polynomial of λ called characteristic polynomial of Σ.
   Equation $|\Sigma - \lambda I| = 0$ is called the characteristic equation for Σ.
   The characteristic equation has p solutions $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ called the first, second,..., the pth largest eigenvalue of Σ.
   Solution $x \neq 0$ to $(\Sigma - \lambda_i I)x = 0$ is an eigenvector of Σ belonging to the eigenvalue λ_i.
   It can be shown that $\Sigma = P\Lambda P'$ where $P = (P_1, ..., P_p) \in R^{p\times p}$ is an orthogonal matrix such that $P'P = I$, and P_i is an eigenvector of Σ belonging to λ_i.
   $\Sigma = P\Lambda P'$ is called an eigenvalue decomposition of Σ.

   Ex1: Find eigenvalue decomposition for $\Sigma = \begin{pmatrix} 3 & -1 \\ -1 & 3 \end{pmatrix}$.

   $$|\Sigma - \lambda I| = \begin{vmatrix} 3-\lambda & -1 \\ -1 & 3-\lambda \end{vmatrix} = (\lambda - 3)^2 - 1 = (\lambda - 2)(\lambda - 4).$$

   $|\Sigma - \lambda I| = 0$, $\lambda_1 = 4$ and $\lambda_2 = 2$. So $\Lambda = \begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix}$.

   $$(\Sigma - \lambda_1)x = 0,\ \begin{pmatrix} -1 & -1 \\ -1 & -1 \end{pmatrix} x = 0,\ x_1 = -x_2,\ x = c\begin{pmatrix} 1 \\ -1 \end{pmatrix},\ P_1 = \tfrac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

   $$(\Sigma - \lambda_2)x = 0,\ \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} x = 0,\ x_1 = x_2,\ x = c\begin{pmatrix} 1 \\ 1 \end{pmatrix},\ P_2 = \tfrac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

   $$P = \tfrac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}.\ \Sigma = P\Lambda P' = \left[\tfrac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}\right]\begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix}\left[\tfrac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}\right]'.$$

(2) EVD and principal components

For $X \sim (\mu, \Sigma)$ with EVD $\Sigma = P\Lambda P'$ with non-decreasing eigenvalues, the compoents of $Y = P'X$ are the principal components of X.

Proof. First, note that $\mathrm{Cov}(Y) = P'\Sigma P = P'P\Lambda P'P = \Lambda$. So $Y_i = P_i'X$ with $\|P_i\| = 1$; $\mathrm{Cov}(Y_i, Y_j) = 0$ for $i \neq j$, and $\mathrm{var}(Y_i) = \lambda_i$ with $\lambda_1 \geq \cdots \geq \lambda_p$.
Secondly we show that it is the best choice. When creating Y_1, in $\{\alpha'X : \|\alpha\| = 1\}$, $\mathrm{var}(\alpha'X) = \alpha'P\Lambda P'\alpha = \beta'\Lambda\beta = \beta_1^2\lambda_1 + \cdots + \beta_p^2\lambda_p$. Here $\beta_1^2 + \cdots + \beta_p^2 = \alpha'PP'\alpha = 1$.
When creating Y_2, in $\{\alpha'X : \|\alpha\| = 1 \text{ and } \mathrm{cov}(\alpha'X, Y_1) = 0\}$, $\mathrm{var}(\alpha X) = \beta_1^2\lambda_1 + \cdots + \beta_p^2\lambda_p$ where $\beta_1^2 + \cdots + \beta_p^2 = 1$. But

$$0 = \mathrm{cov}(\alpha X, Y_1) = \mathrm{cov}(\alpha'X, P_1'X) = \alpha'P\Sigma P'P_1 = \beta'\Lambda e_i = \beta_1\lambda_1.$$

So $\beta_1 = 0$. Hence $\max[\mathrm{var}(\alpha'X) : \|\alpha\| = 1, \mathrm{cov}(\alpha'X, Y_1) = 0) = \lambda_2$.
Similarly one can show $\max[\mathrm{var}(\alpha'X) : \|\alpha\| = 1 \text{ and } \mathrm{cov}(\alpha'X, Y_i) = 0 \text{ for } i = 1, .., k\}$ is λ_{k+1}.

Ex2: For X in Ex1, $\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = P'X = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \frac{1}{\sqrt{2}}\begin{pmatrix} X_1 - X_2 \\ X_1 + X_2 \end{pmatrix}$. Here Y_1 and Y_2 are the first and the second principal components of X, and $\mathrm{Cov}(Y) = \begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix}$.

3. Properties

(1) Total variance
Total variance in X and total variance in Y, the principal component vector of X, are equal.

Proof. $\mathrm{Cov}(X) = \Sigma = P\Lambda P'$. So $\mathrm{tr}(\Sigma) = \mathrm{tr}(P\Lambda P') = \mathrm{tr}(P'P\Lambda) = \mathrm{tr}(\Lambda) = \lambda_1 + \cdots + \lambda_p$.
Here $\mathrm{tr}(\Sigma) = \mathrm{var}(X_1) + \cdots + \mathrm{var}(X_p)$, the total variance in X.
With principal component vector $Y = P'X$, the total variance in Y is
$\mathrm{var}(Y_1) + \cdots + \mathrm{var}(Y_p) = \lambda_1 + \cdots + \lambda_p$.

Ex3: In Ex1, $\mathrm{var}(X_1) + \mathrm{var}(X_2) = 3 + 3 = 6$. In Ex2, $\mathrm{var}(Y_1) + \mathrm{var}(Y_2) = 4 + 2 = 6$.

(2) Variations explained by principal components
The proportion of total variation in the original X explained by the ith principal component is

$$\frac{\mathrm{var}(Y_i)}{\text{Total varaince in } X} = \frac{\lambda_i}{\lambda_1 + \cdots + \lambda_p}.$$

The proportion of total variance explained by the first k principal components is

$$\frac{\mathrm{var}(Y_1) + \cdots + \mathrm{var}(Y_k)}{\text{Total variance in } X} = \frac{\lambda_1 + \cdots + \lambda_k}{\lambda_1 + \cdots + \lambda_p}.$$

Using principal components to achieve certain proportion of total variation explained, one can reduce the number of components in the vector.

Ex4: In Ex1, if we want to explain 60% of total variations in X, we have to use both X_1 and X_2. But if using principal components, from Ex2, we see we onely need the first principal component since $\frac{4}{6}$ =66.7%>60%.

Comment: The concept of principal component is a good mathematical work. But the new component might be $\frac{1}{\sqrt{3}}$(Age) + $\frac{1}{\sqrt{3}}$(Height) + $\frac{1}{\sqrt{3}}$(Number of hours in study) that does not have clear meaning.