Lecture 5: Conditioning. Multivariate Normal Distribution.

Math 586

Conditioning

Conditional distribution of Y given $X = x$ describes probabilistic behavior of Y when a value of X is known. If X and Y are not independent it means that X contains some information about Y. Example: given the reflectance value x from a satellite measurement, we can guess roughly what the soil moisture Y is.

Let $f(x, y)$= joint density, $f_X(x)$, $f_Y(y)$ - marginals. We have $f_X(x) = \int_{-\infty}^{\infty} f(x, y)dy$ and similarly for $f_Y(y)$.

Define **conditional density** $f(y \,|\, x) = \dfrac{f(x, y)}{f_X(x)}$ then
$f(x, y) = f(y \,|\, x)f_X(x)$

Discrete case: conditional PMF $p(y \,|\, x) = \dfrac{P(X = x, Y = y)}{P(X = x)}$ $\quad\ldots(*)$

Conditional expectation: $\mathbb{E}\,(Y \,|\, X = x)$ is the integral

$$\mathbb{E}\,[Y \,|\, X = x] = \int_{-\infty}^{\infty} y\, f(y|x)\, dy = H(x) \quad \text{is some function of } x$$

Discrete case: sums are used.

Note: if X and Y are independent, then $f(y \,|\, x) = f_Y(y)$ and $\mathbb{E}\,(Y \,|\, X = x) = \mathbb{E}\,(Y)$

Example 1.

Given the probability table

Y	3	4	5	marginal of X
X = 0	.2	.1	0	0.3
X = 1	.1	.2	.1	0.4
X = 2	0	.2	.1	0.3
marginal of Y	0.3	0.5	0.2	1

Find $p(y \mid X = 1)$, $\mathbb{E}(Y \mid X = 1)$.

Solution: Applying formula (*), we get

Y	3	4	5	Total
$p(y \mid X = 1)$	$.1/.4 = 0.25$	$.2/.4 = 0.5$	$.1/.4 = 0.25$	1

Then, $\mathbb{E}(Y \mid X = 1) = 3 * 0.25 + 4 * 0.5 + 5 * 0.25 = 3$
□

Example 2.

Suppose that, instead of specifying the joint density $f(x, y)$, we define Y in a conditional way:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

where the error ε is independent of X and $\mathbb{E}(\varepsilon) = 0$. Then

$$\mathbb{E}(Y \mid X = x) = \beta_0 + \beta_1 x + \mathbb{E}(\varepsilon \mid X = x) = \beta_0 + \beta_1 x$$

□

Prediction

For two r.v. X, Y, what is the "best" prediction of Y given X?
Depends on what one means by "best". One possibility: minimum MSE (mean square error). That is, if $\mathcal{F}(X)$ is the predictor of Y then find $\mathcal{F}$ that minimizes MSE $= \mathbb{E}[(Y - \mathcal{F}(X))^2]$.

First, consider a simpler question: for a single r.v. Y, what is the best predictor that minimizes MSE, that is, find constant a such that $\mathbb{E}[(Y - a)^2] \mapsto \min$. Answer:

$$\mathbb{E}[(Y - a)^2] = \mathbb{E}[(Y - \mu + \mu - a)^2] = \mathbb{E}[(Y - \mu)^2] + 2\mathbb{E}[(Y - \mu)(\mu - a)] + (\mu - a)^2 =$$

$$= \mathbb{E}[(Y - \mu)^2] + (\mu - a)^2,$$

where $\mu = \mathbb{E}[Y]$. The minimum is reached when $a = \mu$.

A similar argument shows that **the best predictor of Y given $X = x$ is conditional expectation**:

$$\mathcal{F}(x) = \mathbb{E}[Y \mid X = x]$$

Also, the same extends to vectors (when X is replaced by $\mathbf{X}$).

Note: $\hat{Y} = \mathbb{E}[Y \mid X = x]$ is automatically an **unbiased** predictor of Y, that is

$\mathbb{E}(\hat{Y} \mid X = x) = \hat{Y} = \mathbb{E}(Y \mid X = x)$, for every possible x.

The prediction MSE is also called **conditional variance** $Var[Y \mid X = x]$.

BLUEs and BLUPs

BLUE = Best Linear Unbiased Estimator, BLUP = B.L.U. Predictor. We are interested in the predictor $\hat{Y}$ of unknown quantity Y.

Unbiased means that $\mathbb{E}\,(\hat{Y}\,|\,\texttt{data}) = \mathbb{E}\,(Y\,|\,\texttt{data})$.

Spatial prediction: consider the case when $Y = Y_1$ and $\mathbf{X} = (Y_2, Y_3, ..., Y_n)'$ where Y_i is the observation of some random quantity ("random field") at the geographical location $\boldsymbol{\xi}_i$.

The BLUP of Y_1 can be obtained from the covariance matrix $\mathbf{\Sigma}$ of vector $\mathbf{Y} = (Y_1, Y_2, Y_3, ..., Y_n)$ (see below).

Question: when is the best linear predictor i.e. $\hat{Y}_1 = a_1 + a_2Y_2 + a_3Y_3 + ... + a_nY_n$ also the **best** predictor i.e. $\mathbb{E}\,(Y_1\,|\,Y_2, Y_3, ..., Y_n)$?

Important case: $\mathbf{Y}$ is Multivariate Normal.

Multivariate Normal Distribution (MVN)

Univariate:

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}}\ \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\} = \frac{1}{\sigma\sqrt{2\pi}}\ \exp\left\{-\frac{1}{2}(y-\mu)\sigma^{-2}(y-\mu)\right\}$$

Generalize: $\mathbf{Y} \in \mathbb{R}^n$, $\mathbb{E}\,(\mathbf{Y}) = \boldsymbol{\mu} = (\mu_1, ..., \mu_n)'$.

Let $Var(\mathbf{Y}) = \mathbf{\Sigma} = (\sigma_{ij})$ be $n \times n$ matrix called *variance* or *variance-covariance* matrix of vector $\mathbf{Y}$, so that $\sigma_{ij} = Cov(Y_i, Y_j)$. Let also $det(\mathbf{\Sigma}) = |\mathbf{\Sigma}|$ be the determinant.

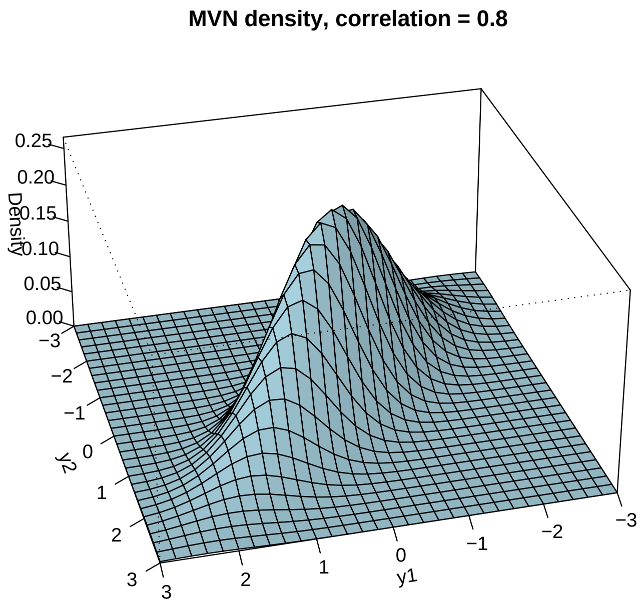
Then

$$f(\mathbf{y}) = \frac{1}{\sqrt{|\mathbf{\Sigma}|}(2\pi)^{n/2}}\ \exp\left\{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})'\,\mathbf{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right\}$$

where $'$ is the transposition and $\mathbf{\Sigma}^{-1}$ = inverse matrix.

Example: when $Y_1, Y_2, ..., Y_n$ are independent and $Var(Y_j) = \sigma_j^2$ then $\mathbf{\Sigma} = diag\{\sigma_1^2, ..., \sigma_n^2\}$.
In this case, can prove that the MVN density is the product of marginals.
Vice versa, if $\mathbf{Y}$ is MVN and its variance matrix is diagonal, then all $Y_1, Y_2, ..., Y_n$ are mutually independent.[1]

[1] Of course, generally uncorrelated RV's are not necessarily independent.



Alternative Definition: $\mathbf{Y}$ is MVN iff $\sum_{i=1}^{n} b_i Y_i$ is a univariate Normal for every set of $\{b_i\}_{i=1}^{n}$ (not all 0's). Often useful.

Properties of MVN

a. Each component Y_i is univariate Normal with mean μ_i and variance σ_{ii}.

b. Any subset of vector $\mathbf{Y}$ is also MVN, with variance matrix being a sub-matrix of $\mathbf{\Sigma}$.

c. If $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2)$ is MVN then conditional $f(\mathbf{Y_1}|\mathbf{Y_2} = \mathbf{y}_2)$ is also MVN.

d. $\mathbb{E}\,(\mathbf{Y_1}\,|\,\mathbf{Y_2})$ is linear in $\mathbf{Y}_2$. [See (1) below.]

e. Any linear combination of *independent* MVN is also MVN, i.e. for n-vector $\boldsymbol{\mu}_0$ and constants $\beta_1, ..., \beta_m$ (not all 0),

$$\mathbf{Y} := \boldsymbol{\mu}_0 + \sum_{i=1}^{m} \beta_i \mathbf{Y}_i \qquad \text{is } n\text{-dimensional MVN.}$$

Lemma. For any random vector $\mathbf{Y}$ its variance matrix $\mathbf{\Sigma}$ is symmetric and positive semidefinite, i.e. for each n-vector $\mathbf{b}$, $\mathbf{b}'\,\mathbf{\Sigma}\,\mathbf{b} \geq 0$.

Proof: Let r.v. $X = \mathbf{b}'\mathbf{Y} = \sum_i b_i Y_i$. Then

$$0 \leq Var(X) = \sum_i b_i^2 Var(Y_i) + \sum_i \sum_{j \neq i} b_i b_j Cov(Y_i, Y_j) = \mathbf{b}'\,\mathbf{\Sigma}\,\mathbf{b}.$$

Examples:

3) Let $\mathbf{Y} = (Y_1, Y_2, Y_3, Y_4)'$

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} \end{bmatrix}$$

then $(Y_1, Y_3)'$ is MVN with variance matrix

$$\mathbf{\Sigma}_{13} = \begin{bmatrix} \sigma_{11} & \sigma_{13} \\ \sigma_{31} & \sigma_{33} \end{bmatrix}$$

and $Y_2 - Y_3$ is univariate Normal with variance ... (see Lemma above, use $\mathbf{b} = (0, 1, -1, 0)'$).

4) Let $\mathbf{Y} = (Y_1, Y_2)$ with variance matrix

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

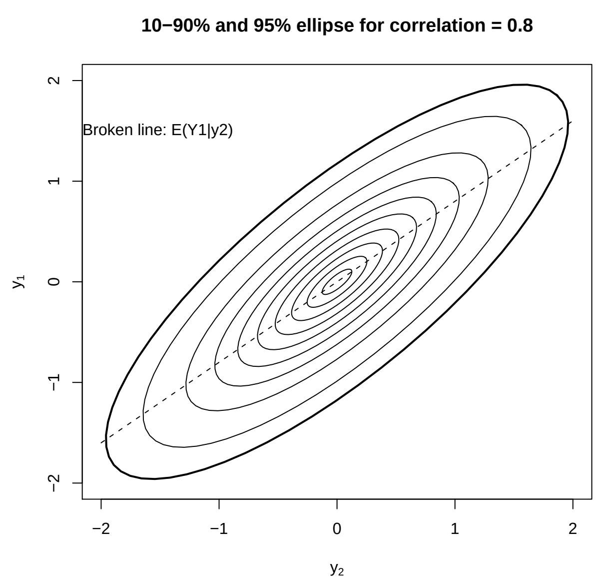
Then

$$\mathbb{E}\,(Y_1 \,|\, Y_2 = y_2) = \mu_1 + \rho\frac{\sigma_1}{\sigma_2}(y_2 - \mu_2), \tag{1}$$

$$Var(Y_1 \,|\, Y_2 = y_2) = \sigma_1^2(1 - \rho^2) \tag{2}$$

Surprisingly, the conditional variance does not depend on y_2!

Exercise. Prove (1), (2) by using the ratio formula for conditional density.



General case Let $\mathbf{Y} = (\mathbf{Y}_1', \mathbf{Y}_2')'$ with mean vector $\boldsymbol{\mu} = (\boldsymbol{\mu}_1', \boldsymbol{\mu}_2')'$ and variance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

Then $f(\mathbf{Y_1} \,|\, \mathbf{Y_2} = \mathbf{y}_2)$ is MVN with mean

$$\mathbb{E}\,(\mathbf{Y}_1 \,|\, \mathbf{Y}_2 = \mathbf{y}_2) = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{y}_2 - \boldsymbol{\mu}_2) \qquad \ldots (**)$$

and conditional var/covar

$$\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$$

does **not** depend on $\mathbf{y}_2$.

Note: if $\boldsymbol{\mu} = \mathbf{0}$ then $\mathbb{E}\,(\mathbf{Y}_1 \,|\, \mathbf{Y}_2 = \mathbf{y}_2) = \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{y}_2$. If we seek $\mathbf{B}$ such that

$$\mathbb{E}\,(\mathbf{Y}_1 \,|\, \mathbf{Y}_2 = \mathbf{y}_2) = \mathbf{B}\mathbf{y}_2$$

then $\mathbf{B} = \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}$. Thus, the bext unbiased estimator of $\mathbf{Y}_1$ given $\mathbf{y}_2$ is a *linear* function of $\mathbf{y}_2$.

Note: be careful. Sometimes the distribution can degenerate, for example normal distribution on a line $y_2 = a + b\,y_1$ is not MVN. Make sure that the variance matrix $\boldsymbol{\Sigma}$ is positive-definite, in particular $|\boldsymbol{\Sigma}| \neq 0$.

Note: The formula (**) can also be used in non-Normal case, as a formula for finding BLUP. If we set $\mathbf{Y}_1 = Y_1$ and $\mathbf{Y}_2 = (Y_2, Y_3, ..., Y_n)'$ then the BLUP of Y_1 given $\mathbf{Y}_2 = \mathbf{y}_2$ is

$$\hat{Y}_1 = \mu_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{y}_2 - \boldsymbol{\mu}_2)$$