STAT 802: Multivariate Analysis

Course outline:

- Multivariate Distributions.
- The Multivariate Normal Distribution.
- The 1 sample problem.
- Paired comparisons.
- Repeated measures: 1 sample.
- One way MANOVA.
- Two way MANOVA.
- Profile Analysis.

- Multivariate Multiple Regression.
- Discriminant Analysis.
- Clustering.
- Principal Components.
- Factor analysis.
- Canonical Correlations.

Basic structure of typical multivariate data set:

Case by variables: data in matrix. Each row is a case, each column is a variable.

Example: Fisher's iris data: 5 rows of 150 by 5 matrix:

Case #	Variety	Sepal Length	Sepal Width	Petal Length	Petal Width
1	Setosa	5.1	3.5	1.4	0.2
2	Setosa	4.9	3.0	1.4	0.2
⋮	⋮	⋮	⋮	⋮	⋮
51	Versicolor	7.0	3.2	4.7	1.4
⋮	⋮	⋮	⋮	⋮	⋮

Usual model: rows of data matrix are independent random variables.

Vector valued random variable: function $\mathbf{X} : \Omega \mapsto \mathbb{R}^p$ such that, writing $\mathbf{X} = (X_1, \ldots, X_p)^T$,

$$P(X_1 \le x_1, \ldots, X_p \le x_p)$$

defined for any const's $(x_1, \ldots, x_p)$.

Cumulative Distribution Function (CDF) of $\mathbf{X}$: function $F_{\mathbf{X}}$ on $\mathbb{R}^p$ defined by

$$F_{\mathbf{X}}(x_1, \ldots, x_p) = P(X_1 \le x_1, \ldots, X_p \le x_p)\,.$$

Defn: Distribution of rv $\mathbf{X}$ is **absolutely continuous** if there is a function f such that

$$P(\mathbf{X} \in A) = \int_A f(x)dx \qquad (1)$$

for any (Borel) set A. This is a p dimensional integral in general. Equivalently

$$F(x_1, \ldots, x_p) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_p} f(y_1, \ldots, y_p)\, dy_p, \ldots, dy_1 \,.$$

Defn: Any f satisfying (**??**) is a **density** of $\mathbf{X}$.

For most x F is differentiable at x and

$$\frac{\partial^p F(x)}{\partial x_1 \cdots \partial x_p} = f(x)\,.$$

Building Multivariate Models

Basic tactic: specify density of

$$\mathbf{X} = (X_1, \ldots, X_p)^T.$$

Tools: marginal densities, conditional densities, independence, transformation.

Marginalization: Simplest multivariate problem

$$\mathbf{X} = (X_1, \ldots, X_p), \qquad Y = X_1$$

(or in general Y is any X_j).

Theorem 1 *If* $\mathbf{X}$ *has density* $f(x_1, \ldots, x_p)$ *and* $q < p$ *then* $\mathbf{Y} = (X_1, \ldots, X_q)$ *has density*

$$f_{\mathbf{Y}}(x_1, \ldots, x_q) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_p)\, dx_{q+1} \ldots dx_p$$

$f_{X_1,\ldots,X_q}$ is the **marginal** density of $X_1, \ldots, X_q$ and $f_{\mathbf{X}}$ the **joint** density of $\mathbf{X}$ but they are both just densities. "Marginal" just to distinguish from the joint density of $\mathbf{X}$.

Independence, conditional distributions

Def'n: Events A and B are independent if

$$P(AB) = P(A)P(B)\,.$$

(Notation: AB is the event that both A and B happen, also written $A \cap B$.)

Def'n: A_i, $i = 1, \ldots, p$ are **independent** if

$$P(A_{i_1} \cdots A_{i_r}) = \prod_{j=1}^{r} P(A_{i_j})$$

for any $1 \le i_1 < \cdots < i_r \le p$.

Def'n: $\mathbf{X}$ and $\mathbf{Y}$ are **independent** if

$$P(\mathbf{X} \in A; \mathbf{Y} \in B) = P(\mathbf{X} \in A)P(\mathbf{Y} \in B)$$

for all A and B.

Def'n: Rvs $\mathbf{X}_1, \ldots, \mathbf{X}_p$ **independent**:

$$P(\mathbf{X}_1 \in A_1, \cdots, \mathbf{X}_p \in A_p) = \prod P(\mathbf{X}_i \in A_i)$$

for any $A_1, \ldots, A_p$.

Theorem:

1. If $\mathbf{X}$ and $\mathbf{Y}$ are independent with joint density $f_{\mathbf{X},\mathbf{Y}}(x,y)$ then $\mathbf{X}$ and $\mathbf{Y}$ have densities $f_{\mathbf{X}}$ and $f_{\mathbf{Y}}$, and
$$f_{\mathbf{X},\mathbf{Y}}(x,y) = f_{\mathbf{X}}(x) f_{\mathbf{Y}}(y)\,.$$

2. If $\mathbf{X}$ and $\mathbf{Y}$ independent with marginal densities $f_{\mathbf{X}}$ and $f_{\mathbf{Y}}$ then $(\mathbf{X},\mathbf{Y})$ has joint density
$$f_{\mathbf{X},\mathbf{Y}}(x,y) = f_{\mathbf{X}}(x) f_{\mathbf{Y}}(y)\,.$$

3. If $(\mathbf{X},\mathbf{Y})$ has density $f(x,y)$ and there exist $g(x)$ and $h(y)$ st $f(x,y) = g(x)h(y)$ for (almost) **all** (x,y) then $\mathbf{X}$ and $\mathbf{Y}$ are independent with densities given by
$$f_{\mathbf{X}}(x) = g(x) / \int_{-\infty}^{\infty} g(u)du$$
$$f_{\mathbf{Y}}(y) = h(y) / \int_{-\infty}^{\infty} h(u)du\,.$$

Theorem: If $\mathbf{X}_1, \ldots, \mathbf{X}_p$ are independent and $\mathbf{Y}_i = g_i(\mathbf{X}_i)$ then $\mathbf{Y}_1, \ldots, \mathbf{Y}_p$ are independent. Moreover, $(\mathbf{X}_1, \ldots, \mathbf{X}_q)$ and $(\mathbf{X}_{q+1}, \ldots, \mathbf{X}_p)$ are independent.

Conditional densities

Conditional density of $\mathbf{Y}$ given $\mathbf{X} = x$:

$$f_{\mathbf{Y}|\mathbf{X}}(y|x) = f_{\mathbf{X},\mathbf{Y}}(x, y)/f_{\mathbf{X}}(x)\,;$$

in words "conditional = joint/marginal".

Change of Variables

Suppose $\mathbf{Y} = g(\mathbf{X}) \in \mathbb{R}^p$ with $\mathbf{X} \in \mathbb{R}^p$ having density $f_{\mathbf{X}}$. **Assume g is a one to one ("injective") map,** i.e., $g(x_1) = g(x_2)$ if and only if $x_1 = x_2$. Find $f_{\mathbf{Y}}$:

Step 1: Solve for x in terms of y: $x = g^{-1}(y)$.

Step 2: Use basic equation:

$$f_{\mathbf{Y}}(y)dy = f_{\mathbf{X}}(x)dx$$

and rewrite it in the form

$$f_{\mathbf{Y}}(y) = f_{\mathbf{X}}(g^{-1}(y))\frac{dx}{dy}$$

Interpretation of derivative $\frac{dx}{dy}$ when $p > 1$:

$$\frac{dx}{dy} = \left|\det\left(\frac{\partial x_i}{\partial y_j}\right)\right|$$

which is the so called **Jacobian**.

Equivalent formula inverts the matrix:

$$f_{\mathbf{Y}}(y) = \frac{f_{\mathbf{X}}(g^{-1}(y))}{\left|\frac{dy}{dx}\right|}.$$

This notation means

$$\left|\frac{dy}{dx}\right| = \left|\det \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_p} \\ & \vdots & & \\ \frac{\partial y_p}{\partial x_1} & \frac{\partial y_p}{\partial x_2} & \cdots & \frac{\partial y_p}{\partial x_p} \end{bmatrix}\right|$$

but with x replaced by the corresponding value of y, that is, replace x by $g^{-1}(y)$.

Example: The density

$$f_{\mathbf{X}}(x_1, x_2) = \frac{1}{2\pi} \exp\left\{-\frac{x_1^2 + x_2^2}{2}\right\}$$

is the **standard bivariate normal density**. Let $\mathbf{Y} = (Y_1, Y_2)$ where $Y_1 = \sqrt{X_1^2 + X_2^2}$ and $0 \le Y_2 < 2\pi$ is angle from the positive x axis to the ray from the origin to the point (X_1, X_2). I.e., $\mathbf{Y}$ is $\mathbf{X}$ in polar co-ordinates.

Solve for x in terms of y:

$$\begin{aligned} X_1 &= Y_1 \cos(Y_2) \\ X_2 &= Y_1 \sin(Y_2) \end{aligned}$$

so that

$$\begin{aligned} g(x_1, x_2) &= (g_1(x_1, x_2), g_2(x_1, x_2)) \\ &= (\sqrt{x_1^2 + x_2^2}, \text{argument}(x_1, x_2)) \\ g^{-1}(y_1, y_2) &= (g_1^{-1}(y_1, y_2), g_2^{-1}(y_1, y_2)) \\ &= (y_1 \cos(y_2), y_1 \sin(y_2)) \\ \left|\frac{dx}{dy}\right| &= \left|\det \begin{pmatrix} \cos(y_2) & -y_1 \sin(y_2) \\ \sin(y_2) & y_1 \cos(y_2) \end{pmatrix}\right| \\ &= y_1 \,. \end{aligned}$$

It follows that

$$\begin{aligned} f_{\mathbf{Y}}(y_1, y_2) &= \frac{1}{2\pi} \exp\left\{-\frac{y_1^2}{2}\right\} y_1 \times \\ & \quad 1(0 \le y_1 < \infty) 1(0 \le y_2 < 2\pi) \,. \end{aligned}$$

Next: marginal densities of Y_1, Y_2?

Factor $f_{\mathbf{Y}}$ as $f_{\mathbf{Y}}(y_1, y_2) = h_1(y_1)h_2(y_2)$ where

$$h_1(y_1) = y_1 e^{-y_1^2/2} 1(0 \le y_1 < \infty)$$

and

$$h_2(y_2) = 1(0 \le y_2 < 2\pi)/(2\pi)\,.$$

Then

$$\begin{aligned} f_{Y_1}(y_1) &= \int_{-\infty}^{\infty} h_1(y_1)h_2(y_2)\,dy_2 \\ &= h_1(y_1)\int_{-\infty}^{\infty} h_2(y_2)\,dy_2 \end{aligned}$$

so marginal density of Y_1 is a multiple of h_1. Multiplier makes $\int f_{Y_1} = 1$ but in this case

$$\int_{-\infty}^{\infty} h_2(y_2)\,dy_2 = \int_0^{2\pi} (2\pi)^{-1} dy_2 = 1$$

so that

$$f_{Y_1}(y_1) = y_1 e^{-y_1^2/2} 1(0 \le y_1 < \infty)\,.$$

(Special Weibull or Rayleigh distribution.)

Similarly

$$f_{Y_2}(y_2) = 1(0 \le y_2 < 2\pi)/(2\pi)$$

which is the **Uniform**$(0, 2\pi)$ density. Exercise: $W = Y_1^2/2$ has standard exponential distribution. Recall: by definition $U = Y_1^2$ has a χ^2 distribution on 2 degrees of freedom. Exercise: find χ^2_2 density.

Remark: easy to check $\int_0^\infty ye^{-y^2/2}dy = 1$.

Thus: have proved original bivariate normal density integrates to 1.

Put $I = \int_{-\infty}^\infty e^{-x^2/2}dx$. Get

$$\begin{aligned} I^2 &= \int_{-\infty}^\infty e^{-x^2/2}dx \int_{-\infty}^\infty e^{-y^2/2}dy \\ &= \int_{-\infty}^\infty \int_{-\infty}^\infty e^{-(x^2+y^2)/2}dydx \\ &= 2\pi. \end{aligned}$$

So $I = \sqrt{2\pi}$.