STAT 730 Chapter 2: Random vectors

Timothy Hanson

Department of Statistics, University of South Carolina

Stat 730: Multivariate Analysis

Random vectors

Let $\mathbf{x} \in \mathbb{R}^p$ be a random vector. The cumulative distribution function (cdf) of $\mathbf{x}$ is $F : \mathbb{R}^p \to [0, 1]$ such that

$$P(\mathbf{x} \leq \mathbf{x}^o) \overset{def}{=} P(x_1 \leq x_1^o, \ldots, x_p \leq x_p^o) = F(x_1^o, \ldots, x_p^o) = F(\mathbf{x}^o).$$

The probability of $\mathbf{x} \in A \subset \mathbb{R}^p$ is given by

$$P(\mathbf{x} \in A) = \int_A dF(\mathbf{u}),$$

the Riemann-Stieltjes integral over A.

The Riemann-Stieltjes intergral is essentially a weighted Riemann integral (but not quite). It's simpler than Lebesgue, yet powerful enough to handle continuous, discrete, and mixtures of continuous/discrete random vectors. Riemann-Stieltjes integrals reduce to the usual Riemann integrals or sums if $\mathbf{x}$ is absolutely continuous or discrete.

Density and mass function

If $\mathbf{x}$ is absolutely continuous, it has a joint density $f : \mathbb{R}^p \to [0, \infty)$ such that

$$P(\mathbf{x} \in A) = \int_A f(\mathbf{u}) d\mathbf{u};$$

if $\mathbf{x}$ is discrete then f is rather a probability mass function such that

$$P(\mathbf{x} \in A) = \sum_{\mathbf{u}_i \in A} f(\mathbf{u}_i),$$

where $\{\mathbf{u}_1, \mathbf{u}_2, \dots\}$ are those (countable) values in the support $S = \{\mathbf{u} : f(\mathbf{u}) > 0\}$.

For absolutely continuous $\mathbf{x}$ the density is given in terms of the cdf

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

Marginal and conditional distributions

Let $\mathbf{x}' = (\mathbf{x}_1', \mathbf{x}_2')$ where $\mathbf{x}_1 \in \mathbb{R}^k$ & $\mathbf{x}_2 \in \mathbb{R}^{p-k}$.

$$P(\mathbf{x}_1 \leq \mathbf{x}_1^o) = F(x_1^o, \ldots, x_k^o, \infty, \ldots, \infty),$$

is the marginal cdf of $\mathbf{x}_1$. If $\mathbf{x}$ is absolutely continuous then

$$f_1(\mathbf{x}_1) = \int_{\mathbb{R}^{p-k}} f(\mathbf{x}_1, \mathbf{x}_2) d\mathbf{x}_2,$$

is the marginal density of $\mathbf{x}_1$; $f_2(\mathbf{x}_2)$ is defined similarly. The conditional density of $\mathbf{x}_1$ given $\mathbf{x}_2 = \mathbf{x}_2^o$ is

$$f(\mathbf{x}_1 | \mathbf{x}_2 = \mathbf{x}_2^o) = \frac{f(\mathbf{x}_1, \mathbf{x}_2^o)}{f_2(\mathbf{x}_2^o)}.$$

Independence

Let $\mathbf{x}' = (\mathbf{x}_1', \mathbf{x}_2')$. $\mathbf{x}_1$ is independent of $\mathbf{x}_2$ when

- $f(\mathbf{x}_1 | \mathbf{x}_2 = \mathbf{x}_2^o) = f_1(\mathbf{x}_1)$ for all $\mathbf{x}_1$.
- $f(\mathbf{x}_2 | \mathbf{x}_1 = \mathbf{x}_1^o) = f_2(\mathbf{x}_2)$ for all $\mathbf{x}_2$.
- $f(\mathbf{x}_1, \mathbf{x}_2) = f_1(\mathbf{x}_1) f_2(\mathbf{x}_2)$.
- $F(\mathbf{x}_1, \mathbf{x}_2) = F_1(\mathbf{x}_1) F_2(\mathbf{x}_2)$.

First three are for absolutely continuous $\mathbf{x}$, last one for all $\mathbf{x}$.

Population moments

Let $\mathbf{x}$ be a random vector. The expectation of the random variable $g(\mathbf{x})$, where $g : \mathbb{R}^p \to \mathbb{R}$, is

$$E\{g(\mathbf{x})\} = \int_{\mathbb{R}^p} g(\mathbf{x}) dF(\mathbf{x}).$$

Properties:

- Linearity $E\{a_1 g_1(\mathbf{x}) + a_2 g_2(\mathbf{x})\} = a_1 E\{g_1(\mathbf{x})\} + a_2 E\{g_2(\mathbf{x})\}$.
- Partition $\mathbf{x}' = (\mathbf{x}_1', \mathbf{x}_2')$, then $E\{g(\mathbf{x}_1)\} = \int_{\mathbb{R}^k} g(\mathbf{x}_1) f_1(\mathbf{x}_1) d\mathbf{x}_1$.
- x_1 and x_2 independent
  $\Rightarrow E\{g_1(\mathbf{x}_1) g_2(\mathbf{x}_2)\} = E\{g_1(\mathbf{x}_1)\} E\{g_2(\mathbf{x}_2)\}$.

In general, let $\mathbf{G} : \mathbb{R}^p \to \mathbb{R}^{a \times b}$ have elements $[g_{ij}(\mathbf{x})]_{a \times b}$. Then $E\{\mathbf{G}(\mathbf{x})\}$ is the matrix with ijth element $E\{g_{ij}(\mathbf{x})\}$.

$\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$

Define

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_p \end{bmatrix} = \begin{bmatrix} E(x_1) \\ \vdots \\ E(x_p) \end{bmatrix} = E(\mathbf{x}),$$

and

$$\boldsymbol{\Sigma} = V(\mathbf{x}) = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix} = E\{(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})'\}.$$

These are the population mean vector and covariance matrix. Note that the *ij*th entry of $\boldsymbol{\Sigma}$ is $\sigma_{ij} = E\{x_i - \mu_i)(x_j - \mu_j)\} = C(x_i, x_j)$.

We write $\mathbf{x} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for short.

Property of expectation

Let $\mathbf{A} \in \mathbb{R}^{q \times p}$ and $\mathbf{b} \in \mathbb{R}^q$. Then

$$E\{\mathbf{A}\mathbf{x} + \mathbf{b}\} = \mathbf{A}E(\mathbf{x}) + \mathbf{b}.$$

To show this, let $\mathbf{G} : \mathbb{R}^p \to \mathbb{R}^q$ have ith element $g_i(\mathbf{x}) = (a_{i1} \cdots a_{ip})\mathbf{x} + b_i$ and use definition two slides earlier.

More covariance

Let $\mathbf{x} \in \mathbb{R}^p$ and $\mathbf{y} \in \mathbb{R}^q$ be random vectors. Then

$$C(\mathbf{x}, \mathbf{y}) = E\{(\mathbf{x} - E(\mathbf{x}))(\mathbf{y} - E(\mathbf{y}))'\}.$$

What is the *ij*th element of $C(\mathbf{x}, \mathbf{y})$? Properties:

- $\boldsymbol{\Sigma} = E(\mathbf{x}\mathbf{x}') - \boldsymbol{\mu}\boldsymbol{\mu}'$.
- $V(\mathbf{A}\mathbf{x} + \mathbf{b}) = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'$, $V(\mathbf{a}'\mathbf{x}) = \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}$.
- $\boldsymbol{\Sigma} \geq 0$ because $V(\cdot) \geq 0$ & def'n pos. def.
- $C(\mathbf{x}, \mathbf{x}) = V(\mathbf{x})$.
- $C(\mathbf{x}, \mathbf{y}) = C(\mathbf{y}, \mathbf{x})'$.
- $C(\mathbf{x}_1 + \mathbf{x}_2, \mathbf{y}) = C(\mathbf{x}_1, \mathbf{y}) + C(\mathbf{x}_2, \mathbf{y})$.
- $p = q \Rightarrow V(\mathbf{x} + \mathbf{y}) = V(\mathbf{x}) + C(\mathbf{x}, \mathbf{y}) + C(\mathbf{y}, \mathbf{x}) + V(\mathbf{y})$.
- $C(\mathbf{A}\mathbf{x}, \mathbf{B}\mathbf{y}) = \mathbf{A}C(\mathbf{x}, \mathbf{y})\mathbf{B}'$.
- $\mathbf{x}$ ind. $\mathbf{y} \Rightarrow C(\mathbf{x}, \mathbf{y}) = \mathbf{0}$.
- $V(\mathbf{x}_1 + \cdots + \mathbf{x}_n) = \sum_{i=1}^n V(\mathbf{x}_i) + \sum_{i \neq j} C(\mathbf{x}_i, \mathbf{x}_j)$.

Correlation matrix & generalized variance

Recall $\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$. Let $\mathbf{P} = [\rho_{ij}]$ be the $p \times p$ correlation matrix. Define $\boldsymbol{\Delta} = \text{diag}(\sigma_1, \ldots, \sigma_p)$. Then

$$\mathbf{P} = \boldsymbol{\Delta}^{-1} \boldsymbol{\Sigma} \boldsymbol{\Delta}^{-1} \text{ and } \boldsymbol{\Sigma} = \boldsymbol{\Delta} \mathbf{P} \boldsymbol{\Delta}.$$

Population generalized variance is $|\boldsymbol{\Sigma}|$ and population total variance is tr $\boldsymbol{\Sigma}$.

Mahalanobis space

Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$. M-distance between $\mathbf{a}$ and $\mathbf{b}$ based on $\mathbf{\Sigma}$ is $\Delta^2_{\mathbf{\Sigma}}(\mathbf{a}, \mathbf{b}) = (\mathbf{a} - \mathbf{b})'\mathbf{\Sigma}^{-1}(\mathbf{a} - \mathbf{b})$. Unitless distance that takes scale and correlation into account.

- Let $E(\mathbf{x}) = \mu_{\mathbf{x}}$, $E(\mathbf{y}) = \mu_{\mathbf{y}}$, and $V(\mathbf{x}) = V(\mathbf{y}) = \mathbf{\Sigma}$. Then $\Delta^2_{\mathbf{\Sigma}}(\boldsymbol{\mu}_{\mathbf{x}}, \boldsymbol{\mu}_{\mathbf{y}}) = (\boldsymbol{\mu}_{\mathbf{x}} - \boldsymbol{\mu}_{\mathbf{y}})'\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_{\mathbf{x}} - \boldsymbol{\mu}_{\mathbf{y}})$ is M-distance between two population means. Used, e.g., in anthropology to measure distance between groups based on bone measurements.
- $\Delta^2_{\mathbf{\Sigma}}(\mathbf{x}, \boldsymbol{\mu}) = (\mathbf{x} - \boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})$ used for outlier detection.

Multivariate skew and kurtosis (p. 31)

Here are two measures proposed by your book; there are others.

Multivariate skew (is mass "piled up" in the "tails" more in one direction that others) is measured by

$$\beta_{1,p} = E\{(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu})\}^3,$$

where $\mathbf{x}, \mathbf{y} \overset{iid}{\sim} (\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Multivariate kurtosis (how "peaked" the density is at the mode) is measured by

$$\beta_{2,p} = E\{(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu})\}^2.$$

There are natural sample analogues of these. These measures can be used to form tests for multivariate normality assessment.

Characteristic functions

The c.f. of $\mathbf{x}$ is the Fourier transform

$$\phi_{\mathbf{x}}(\mathbf{t}) = E(e^{i\mathbf{t}'\mathbf{x}}) = \int_{\mathbb{R}^p} e^{i\mathbf{t}'\mathbf{x}} dF(\mathbf{x}).$$

Properties:

- Always exists, $\phi_{\mathbf{x}}(\mathbf{0}) = 1$ and $|\phi_{\mathbf{x}}(\mathbf{t})| \leq 1$.
- $\mathbf{x}$ and $\mathbf{y}$ have same c.f. $\Leftrightarrow F_{\mathbf{x}}(\cdot) = F_{\mathbf{y}}(\cdot)$.
- If $\phi_{\mathbf{x}}(\mathbf{t})$ absolutely integrable then $\mathbf{x}$ has PDF $f(\mathbf{x}) = \frac{1}{(2\pi)^p} \int_{\mathbb{R}^p} e^{-i\mathbf{t}'\mathbf{x}} \phi_{\mathbf{x}}(\mathbf{t}) d\mathbf{t}$.
- For $\mathbf{x}' = (\mathbf{x}_1', \mathbf{x}_2')$, $\mathbf{x}_1$ ind. $\mathbf{x}_2 \Leftrightarrow \phi_{\mathbf{x}}(\mathbf{t}) = \phi_{\mathbf{x}_1}(\mathbf{t}_1)\phi_{\mathbf{x}_2}(\mathbf{t}_2)$. Also $\phi_{\mathbf{x}_1} = \phi_{\mathbf{x}}(\mathbf{t}_1, \mathbf{0})$.
- $E(x_1^{j_1} \cdots x_p^{j_p}) = \frac{1}{i^{j_1+\cdots+j_p}} \left[\frac{\partial^{j_1+\cdots+j_p}}{\partial t_1^{j_1} \cdots \partial t_p^{j_p}} \phi_{\mathbf{x}}(\mathbf{t}) \right]_{\mathbf{t}=\mathbf{0}}.$
- $\mathbf{x}$ and $\mathbf{y}$ independent p-vectors $\Rightarrow \phi_{\mathbf{x}+\mathbf{y}}(\mathbf{t}) = \phi_{\mathbf{x}}(\mathbf{t})\phi_{\mathbf{y}}(\mathbf{t})$.

Cramer-Wold theorem

First note that the distribution of $\mathbf{x}$ is completely determined by its c.f.

<u>thm</u>: The distribution of $\mathbf{x} \in \mathbb{R}^p$ is completely determined by the distributions of linear combinations in $\mathcal{D} = \{\mathbf{t}'\mathbf{x} : \mathbf{t} \in \mathbb{R}^p\}$.

$\boxed{\text{Proof}}$: Let $y_{\mathbf{t}} = \mathbf{t}'\mathbf{x} \in \mathcal{D}$. The c.f. is $\phi_{y_{\mathbf{t}}}(s) = E(e^{isy_{\mathbf{t}}}) = E(e^{is\mathbf{t}'\mathbf{x}})$. Now note $\phi_{y_{\mathbf{t}}}(1)$ is the c.f. of $\mathbf{x}$ evaluated at $\mathbf{t}$. $\square$.

In other words, for each $\mathbf{t} \in \mathbb{R}^p$, we can evaluate $\phi_{\mathbf{x}}(\mathbf{t})$ via the c.f. of the univariate $\phi_{y_{\mathbf{t}}}(1)$. We need all of the elements of $\mathcal{D}$ to completely specify the c.f. of $\mathbf{x}$ though.

This theorem will come in handy in specifying the multivariate normal distribution a bit later.

Transformations

Let $\mathbf{x} \in \mathbb{R}^p$ have a PDF $f(\cdot)$ and let the function $\mathbf{u} : \mathbb{R}^p \to \mathbb{R}^p$ be one-to-one. Then the PDF of $\mathbf{y} = \mathbf{u}(\mathbf{x})$ is

$$f_{\mathbf{y}}(\mathbf{y}) = f\{\mathbf{u}^{-1}(\mathbf{y})\}|\mathbf{J}|,$$

where $\mathbf{J}$ is the $p \times p$ matrix with ijth element $\frac{\partial x_i}{\partial y_j}$. Book makes note of sets of Lebesgue measure zero, but don't worry about this.

Pages 35–36 have Jacobians $|\mathbf{J}|$ for several transformations used in this class as well as some examples.

Multivariate normal distribution

Recall if $x \sim N(\mu, \sigma^2)$ then

$$f(x) = \{2\pi\sigma^2\}^{-1/2} \exp\{-\tfrac{1}{2}(x-\mu)(\sigma^2)^{-1}(x-\mu)\}.$$

<u>def'n</u>: $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} > 0$, has pdf

$$f(x) = |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\{-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\}.$$

A couple results

thm: If $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathbf{y} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$ then $y_1, \dots, y_p \overset{iid}{\sim} N(0,1)$.

$\boxed{\text{Proof}}$: The change-of-variables formula, two slides back, with $\mathbf{u}(\mathbf{x}) = \boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$ and $|\mathbf{J}| = |\boldsymbol{\Sigma}|^{1/2}$ yields

$$f_{\mathbf{y}}(\mathbf{y}) = \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y_i^2},$$

i.e. $\mathbf{y} \sim N_p(\mathbf{0}, \mathcal{I})$. $\square$.

Corollary: $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \Rightarrow E(\mathbf{x}) = \boldsymbol{\mu}, \quad V(\mathbf{x}) = \boldsymbol{\Sigma}$.

This is immediate by starting with $\mathbf{y} \sim N_p(\mathbf{0}, \mathcal{I})$, which implies $E(\mathbf{y}) = \mathbf{0}$ and $V(\mathbf{y}) = \mathcal{I}$, and taking the transformation $\mathbf{x} = \boldsymbol{\Sigma}^{1/2}\mathbf{y} + \boldsymbol{\mu}$ (one-to-one!). Clearly, $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ from above and properties of expectation and covariance yield the desired result.

Simulating multivariate normals

The inverse transformation $\mathbf{x} = \mathbf{\Sigma}^{1/2}\mathbf{y} + \boldsymbol{\mu}$ allows us to generate $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \mathbf{\Sigma})$ as long as well can generate *iid* $N(0,1)$ variables.

In fact, we can use any square root of $\mathbf{\Sigma}$, i.e. any $\mathbf{A}$ s.t. $\mathbf{A}\mathbf{A}' = \mathbf{\Sigma}$ by taking $\mathbf{x} = \mathbf{A}\mathbf{y} + \boldsymbol{\mu}$. In fact, $\mathbf{\Sigma}^{1/2} = \mathbf{M}\mathbf{\Lambda}^{1/2}\mathbf{M}'$ (spectral decomposition) gives the unique symmetric square root. All other square roots are given by $\mathbf{\Sigma}^{1/2}\mathbf{O}$ where $\mathbf{O}$ are unitary matrices.

One important non-symmetric square root is the Cholesky decomposition. Here we simulate

$$\mathbf{x}_1, \ldots, \mathbf{x}_{1000} \overset{iid}{\sim} N_2\left(\begin{bmatrix} 3 \\ 8 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 2 \end{bmatrix} \right):$$

```
m=matrix(c(1,0.5,0.5,2),2,2)
m # covariance matrix
t(chol(m))%*%chol(m) # same
X=t(t(chol(m))%*%matrix(rnorm(2*1000),2,1000)+c(3,8))
plot(X)
```

Geometry

If $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then the density $f(\mathbf{x})$ is constant on ellipsoids $(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = c^2$. These ellipsoids are called contours, and plots of them for different c yield a (topographic) contour plot.

```
library(mvtnorm)
mu=c(1,3)
sigma=matrix(c(1,0.7,0.7,1),2,2)
z=matrix(0,50,50)
x1=seq(qnorm(0.01,mu[1],sigma[1,1]),qnorm(0.99,mu[1],sigma[1,1]),length.out=50)
x2=seq(qnorm(0.01,mu[2],sigma[2,2]),qnorm(0.99,mu[2],sigma[2,2]),length.out=50)
for(i in 1:50){for(j in 1:50){z[i,j]=dmvnorm(c(x1[i],x2[j]),mu,sigma)}}
contour(x1,x2,z)
filled.contour(x1,x2,z)
```

The eigenvectors of $\boldsymbol{\Sigma}$ give the major and minor axes. The eigenvalues are how much the ellipse is "stretched" along its axis. More examples on pp. 39–40.

Some useful properties of $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

thm: $U = (\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \sim \chi^2_p$.

$\boxed{\text{Proof}}$: Again using $\mathbf{y} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$ we have $(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = \mathbf{y}'\mathbf{y}$ which is the sum of p independent squared standard normals. $\square$.

thm: The c.f. of $\mathbf{x}$ is $\phi_{\mathbf{x}}(\mathbf{t}) = \exp(i\mathbf{t}'\boldsymbol{\mu} - \frac{1}{2}\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t})$.

$\boxed{\text{Proof}}$: Use $\mathbf{x} = \boldsymbol{\Sigma}^{1/2}\mathbf{y} + \boldsymbol{\mu}$ and let $u_i = \sum_{j=1}^p (\boldsymbol{\Sigma}^{1/2})_{ij} t_j$ be the ith element of $\mathbf{u} = \boldsymbol{\Sigma}^{1/2}\mathbf{t} \in \mathbb{R}^p$. Then
$\phi_{\mathbf{x}}(\mathbf{t}) = E(e^{i\mathbf{t}'\mathbf{x}}) = e^{i\mathbf{t}'\boldsymbol{\mu}} E(e^{i\mathbf{t}'\boldsymbol{\Sigma}^{1/2}\mathbf{y}}) = e^{i\mathbf{t}'\boldsymbol{\mu}} \phi_{\mathbf{y}}(\boldsymbol{\Sigma}^{1/2}\mathbf{t}) =$
$e^{i\mathbf{t}'\boldsymbol{\mu}} \prod_{i=1}^p \phi_{y_i}(u_i) = e^{i\mathbf{t}'\boldsymbol{\mu}} \prod_{i=1}^p \exp\{-\frac{1}{2}\sum_{j=1}^p u_i^2\} =$
$e^{i\mathbf{t}'\boldsymbol{\mu}} \exp\{-\frac{1}{2}||\mathbf{u}||^2\} = e^{i\mathbf{t}'\boldsymbol{\mu}} \exp\{-\frac{1}{2}\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}\}$. $\square$

Some useful properties of $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

<u>thm</u>: Let $\mathbf{a} \in \mathbb{R}^p$. Then $\mathbf{a}'\mathbf{x} \sim N(\mathbf{a}'\boldsymbol{\mu}, \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a})$.

$\boxed{\text{Proof}}$: $\phi_{\mathbf{a}'\mathbf{x}}(t) = \phi_{\mathbf{x}}(t\mathbf{a}) = \exp\{it\mathbf{a}'\boldsymbol{\mu} - \frac{1}{2}t^2\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}\}$. $\square$.

<u>thm</u>: $\beta_{1,p} = 0$ and $\beta_{2,p} = p(p+2)$.

$\boxed{\text{Proof}}$: See book p. 41 if interested. $\square$

Singular normal distributions

First theorem on last slide doesn't make sense if $\mathbf{a} = \mathbf{0}$ unless we define particular normal distributions that have zero variance for some linear combinations $\mathbf{a}'\mathbf{x}$.

Another example: $x_1 \sim N(0,1)$ and $x_2 = x_1$. Want to have $\mathbf{x} \sim N_2(\mathbf{0}, \mathbf{1}_2\mathbf{1}_2')$, gives $\rho_{12} = 1$, but x_1 and x_2 both $N(0,1)$ marginally.

<u>def'n</u>: $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\text{rank}(\boldsymbol{\Sigma}) = k < p$ has density

$$f(\mathbf{x}) = \frac{(2\pi)^{-k/2}}{(\lambda_1 \cdots \lambda_k)^{1/2}} \exp\{(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-}(\mathbf{x} - \boldsymbol{\mu})\}$$

where $\mathbf{x}$ lives in hyperplane $\mathbf{N}'(\mathbf{x} - \boldsymbol{\mu}) = \mathbf{0}$ where $\mathbf{N}$ is $p \times (p - k)$ matrix w/ columns spanning null space of $\boldsymbol{\Sigma}$, i.e. $\mathbf{N}'\boldsymbol{\Sigma} = \mathbf{0}$ and $\mathbf{N}'\mathbf{N} = \mathcal{I}_{p-k}$. Here, $\boldsymbol{\Sigma}^{-}$ is a g-inverse of $\boldsymbol{\Sigma}$ and $\lambda_1, \ldots, \lambda_k$ are non-zero eigenvalues.

Singular normal distribution example

For $\mathbf{x} \sim N_2(\mathbf{0}, \mathbf{1}_2\mathbf{1}_2')$ we have $\text{rank}(\boldsymbol{\Sigma}) = 1$ and $\mathbf{N} = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix}$.

Note that $\boldsymbol{\Sigma}^- = \frac{1}{4}\mathbf{1}_2\mathbf{1}_2'$ yields $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^-\boldsymbol{\Sigma} = \boldsymbol{\Sigma}$ (SVD). Also, $\lambda_1 = 1$. Then

$$f(x_1, x_2) = (2\pi)^{-1/2} \exp\{-\tfrac{1}{8}(x_1 + x_2)^2)\}$$

on the hyperplane $x_1 = x_2$.

For singular normal, $\mathbf{a} \in \mathcal{C}(\mathbf{N})$ yields $\mathbf{a}'\mathbf{x} = \mathbf{a}\boldsymbol{\mu}$ with probability one.

Define $N_p(\boldsymbol{\mu}, \mathbf{0})$ to be point mass at $\boldsymbol{\mu}$, i.e. $\delta_{\boldsymbol{\mu}}$. Has c.f. $\phi_{\delta_{\boldsymbol{\mu}}}(\mathbf{t}) = e^{i\mathbf{t}'\boldsymbol{\mu}}$.

Multivariate generalizations of common distributions

- We will deal mainly w/ multivariate normal.
- $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $u_i = \exp(x_i)$, then $\mathbf{u}$ has multivariate log-normal distribution. Let $y \sim \chi^2_\nu$ and $u_i = x_i/\sqrt{y/\nu}$, then $\mathbf{u}$ has a multivariate t distribution.
- Wishart distribution generalizes χ^2.
- Multivariate Pareto dist'n, p. 44.
- Dirichlet and multinomial distributions generalize beta and binomial, respectively.
- Common components (Sec. 2.6.2) used in random effect models.

Exponential family

$$f(\mathbf{x};\boldsymbol{\theta}) = \exp\left[a_0(\boldsymbol{\theta}) + b_0(\mathbf{x}) + \sum_{i=1}^{q} a_i(\boldsymbol{\theta}) b_i(\mathbf{x})\right], \mathbf{x} \in S,$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_r)'$ is parameter vector, $e^{a_0(\boldsymbol{\theta})}$ is normalizing constant, and S is support.

If $r = q$ and $a_i(\boldsymbol{\theta}) = \theta_i$ for $i = 1, \ldots, r$ then $\mathbf{x}$ belongs to simple exponential family.

The multivariate normal is an exponential family.

Spherical family

If the pdf of $\mathbf{x}$ can be written $f(\mathbf{x}) = g(\mathbf{x}'\mathbf{x}) = g(||\mathbf{x}||^2)$ then $\mathbf{x}$ belongs to the spherical family because it is spherically symmetric: pdf contours $c^2 = g(\mathbf{x}'\mathbf{x})$ are equispheres.

Examples: $\mathbf{x} \sim N_p(\mathbf{0}, \mathcal{I})$ and $f(\mathbf{x}) = \pi^{-(p+1)/2}\Gamma(\frac{1}{2}(p+1))(1 + \mathbf{x}'\mathbf{x})^{-(p+1)/2}$, the multivariate Cauchy, belong to the spherical family.

Random samples

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \overset{iid}{\sim} (\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

thm: $E(\bar{\mathbf{x}}) = \boldsymbol{\mu}$ and $V(\bar{\mathbf{x}}) = \frac{1}{n}\boldsymbol{\Sigma}$.

$\boxed{\text{Proof}}$:

$$E(\mathbf{x}) = \tfrac{1}{n}\sum_{i=1}^{n} E(\mathbf{x}_i) = \boldsymbol{\mu}.$$

$$V(\bar{\mathbf{x}}) = \frac{1}{n^2} V(\mathbf{x}_1 + \cdots + \mathbf{x}_n) = \frac{1}{n^2}\left[\sum_{i=1}^{n} V(\mathbf{x}_i) + \sum_{i \neq j} C(\mathbf{x}_i, \mathbf{x}_j)\right] = \frac{1}{n^2}\sum_{i=1}^{n} V(\mathbf{x}_i) = \tfrac{1}{n}\boldsymbol{\Sigma}.\square$$

Note then $E\{(\bar{\mathbf{x}} - \boldsymbol{\mu})(\bar{\mathbf{x}} - \boldsymbol{\mu})'\} = V(\bar{\mathbf{x}}) = \frac{1}{n}\boldsymbol{\Sigma}$.

What is $E(\mathbf{S})$?

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})' &= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu}+\boldsymbol{\mu}-\bar{\mathbf{x}})(\mathbf{x}_i-\boldsymbol{\mu}+\boldsymbol{\mu}-\bar{\mathbf{x}})' \\
&= \frac{1}{n}\sum_{i=1}(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})' + \frac{1}{n}\sum_{i=1}^{n}(\bar{\mathbf{x}}-\boldsymbol{\mu})(\bar{\mathbf{x}}-\boldsymbol{\mu})' \\
&\quad +\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu})(\boldsymbol{\mu}-\bar{\mathbf{x}})' + \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\mu}-\bar{\mathbf{x}})(\mathbf{x}_i-\boldsymbol{\mu})' \\
&= \left[\frac{1}{n}\sum_{i=1}(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})'\right] - (\bar{\mathbf{x}}-\boldsymbol{\mu})(\bar{\mathbf{x}}-\boldsymbol{\mu})'
\end{aligned}$$

The expectation of the first term is $\boldsymbol{\Sigma}$; the expecation of the second term is $\frac{1}{n}\boldsymbol{\Sigma}$ from the previous slide. So $E(\mathbf{S}) = \frac{n-1}{n}\boldsymbol{\Sigma}$. Note then that for $\mathbf{S}_u = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})'$ that $E(\mathbf{S}_u) = \boldsymbol{\Sigma}$.

Simple results for later...

$$
\begin{aligned}
E(\mathbf{x}_i) &= \boldsymbol{\mu}, \\
E(\mathbf{x}_{(j)}) &= \boldsymbol{\mu}_j \mathbf{1}_n, \\
C(\mathbf{x}_i, \mathbf{x}_j) &= \delta_{ij} \boldsymbol{\Sigma}, \\
C(\mathbf{x}_{(i)}, \mathbf{x}_{(j)}) &= \sigma_{ij} \mathcal{I}_n.
\end{aligned}
$$

$E\{D_{ij}\} = E\{(\mathbf{x}_i - \bar{\mathbf{x}})\mathbf{S}^{-1}(\mathbf{x}_j - \bar{\mathbf{x}})\}$

Assume $\mathbf{X}$ is full rank and $n > p$. Recall from Chapter 1 that $n\mathbf{S} = \mathbf{X}'\mathbf{H}\mathbf{X}$. Then the $n \times n$ matrix $\mathbf{D} = [D_{ij}]$ can be written

$$\mathbf{D} = [\mathbf{H}\mathbf{X}][n(\mathbf{X}'\mathbf{H}\mathbf{X})^{-1}]\mathbf{X}'\mathbf{H}.$$

Since $\mathbf{H}$ is the orthogonal projection onto $\mathcal{C}(\mathbf{1}_n)^{\perp}$, $\mathbf{D}\mathbf{1}_n' = \mathbf{0}$, i.e. $\sum_{i=1}^n D_{ij} = 0$. You can also show this directly. Now, $\frac{1}{n}\mathbf{D}$ is an orthogonal projection onto $\mathbf{H}\mathbf{X}$, and the trace of an idempotent matrix is it's rank. The rank of $\mathbf{H}\mathbf{X}$ (just $\mathbf{X}$ with $\bar{\mathbf{x}}'$ subtracted from each row) is p a.s.

D_{ii} identically distributed, and D_{ij} identically distributed for $i \neq j$. The above implies

$$\sum_{i=1}^n D_{ij} = 0 \text{ and } \sum_{i=1}^n D_{ii} = np.$$

Taking expectations of both sides and solving gives $E\{D_{ii}\} = p$ and $E\{D_{ij}\} = -\frac{p}{n-1}$ for $i \neq j$.

Convergence in distribution

$\mathbf{x}_n \overset{D}{\to} \mathbf{x} \Leftrightarrow F_n(\mathbf{x}) \to F(\mathbf{x}) \Leftrightarrow P(\mathbf{x}_n \in A) \to P(\mathbf{x} \in A)$ for all measurable A.

Cramer-Wold implies $\mathbf{x}_n \overset{D}{\to} \mathbf{x} \Leftrightarrow \mathbf{t}'\mathbf{x}_n \overset{D}{\to} \mathbf{t}'\mathbf{x}$ for all $\mathbf{t}$. Again, multivariate achieved through all linear combinations.

This further implies for any $\mathbf{A} \in \mathbb{R}^{q \times p}$ that $\mathbf{A}\mathbf{x}_n \overset{D}{\to} \mathbf{A}\mathbf{x}$. Why? $\mathbf{t}'(\mathbf{A}\mathbf{x}_n) = (\mathbf{t}'\mathbf{A})\mathbf{x}_n \overset{D}{\to} \mathbf{t}'\mathbf{A}\mathbf{x}$. We need this for the Delta method, coming up after the CLT.

Central Limit Theorem

thm: Let $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3 \cdots \overset{iid}{\sim} (\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then

$$\sqrt{n}(\bar{\mathbf{x}} - \boldsymbol{\mu}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{\Sigma}).$$

General proof is in STAT 823, but uses Cramer-Wold & continuity theorem for c.f. Informally we write $\bar{x} \overset{\bullet}{\sim} N_p(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma})$.

The theorem follows from the univariate CLT by noting $\sqrt{n}\mathbf{t}'(\bar{\mathbf{x}} - \boldsymbol{\mu}) \xrightarrow{D} N(0, \mathbf{t}'\boldsymbol{\Sigma}\mathbf{t})$ holds for all $\mathbf{t}$. The result for convergence in distribution (previous slide) implies the multivariate CLT!

o_p and O_p

Convergence in probability:

$$x_n = o_p(a_n) \Leftrightarrow \frac{x_n}{a_n} = o_p(1) \Leftrightarrow \frac{x_n}{a_n} \overset{P}{\rightarrow} 0 \Leftrightarrow \lim_{n\rightarrow\infty} P(|x_n/a_n| \geq \epsilon) = 0 \ \forall \epsilon > 0.$$

Note: $\mathbf{x}_n \overset{P}{\rightarrow} \mathbf{a} \Leftrightarrow \mathbf{x}_{in} \overset{P}{\rightarrow} a_i$ for $i = 1, \ldots, p$. Matrix version similar.

Bounded in probability:

$$x_n = O_p(a_n) \Leftrightarrow \frac{x_n}{a_n} = O_p(1) \Leftrightarrow \forall \epsilon > 0 \ \exists M_\epsilon \ s.t. \ P(|x_n/a_n| \geq M_\epsilon) < \epsilon \ \forall n.$$

Also multivariate generalizations of Slutsky, etc. We'll discuss them if/when we need them. Two immediately useful results are $O_p(1)o_p(1) = o_p(1)$ and $O_p(1) + o_p(1) = O_p(1)$.

Multivariate Delta Method

<u>thm</u>: Let $\sqrt{n}(\mathbf{x}_n - \boldsymbol{\mu}) \overset{D}{\rightarrow} N_p(0, \boldsymbol{\Sigma})$ and $\mathbf{f} : \mathbb{R}^p \rightarrow \mathbb{R}^q$ is differentiable at $\boldsymbol{\mu}$. Then $\sqrt{n}[\mathbf{f}(\mathbf{x}_n) - \mathbf{f}(\boldsymbol{\mu})] \overset{D}{\rightarrow} N_q(\mathbf{0}, \mathbf{D}\boldsymbol{\Sigma}\mathbf{D}')$ where $\mathbf{D} = [\frac{\partial f_i}{\partial x_j}]_{\mathbf{x}=\boldsymbol{\mu}}$.

$\boxed{\text{Proof}}$: The multivariate Taylor's theorem for $\mathbf{x}_n$ expanded about $\boldsymbol{\mu}$ gives us

$$\mathbf{f}(\mathbf{x}_n) - \mathbf{f}(\boldsymbol{\mu}) = \mathbf{D}(\mathbf{x}_n - \boldsymbol{\mu}) + ||\mathbf{x}_n - \boldsymbol{\mu}||\delta(\mathbf{x}_n - \boldsymbol{\mu}),$$

where $\delta(\mathbf{a}_n - \boldsymbol{\mu}) \rightarrow \mathbf{0}$ as $\mathbf{a}_n \rightarrow \boldsymbol{\mu}$. Since $\sqrt{n}||\mathbf{x}_n - \boldsymbol{\mu}|| = O_p(1)$ and $\delta(\mathbf{x}_n - \boldsymbol{\mu}) = o_p(\mathbf{1})$ we have

$$\sqrt{n}[\mathbf{f}(\mathbf{x}_n) - \mathbf{f}(\boldsymbol{\mu})] = \sqrt{n}\mathbf{D}'(\mathbf{x}_n - \boldsymbol{\mu}) + O_p(1)o_p(\mathbf{1}) \overset{D}{\rightarrow} N_q(\mathbf{0}, \mathbf{D}'\boldsymbol{\Sigma}\mathbf{D}).\square$$

Two useful consequences...

In general $\mathbf{x}_n \overset{P}{\to} \mathbf{x} \Rightarrow \mathbf{x}_n \overset{D}{\to} \mathbf{x}$ but not the converse. However, the CLT implies two WLLN when all necessary expectations exist, if $\mathbf{x}_1, \mathbf{x}_2, \cdots \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then

$$\bar{\mathbf{x}} \overset{P}{\to} \boldsymbol{\mu},$$

and

$$\mathbf{S} \overset{P}{\to} \boldsymbol{\Sigma}.$$

Also, the univariate WLLN imply these results as well (convergence for each element).

Useful in MCMC approach to obtaining Bayesian inference. Also used in MCEM algorithm and elsewhere.

Try `mean(data.frame(X))` and `cov(X)` to obtain estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ from the simulated normal example.

Comments...

We are not going to be doing much asymptotics in this course, but some. STAT 823 beats this to death. Many results in STAT 730 are finite sample results based on normality, or else the asymptotic results are simply stated and cited.

The Delta method is very, very useful and used a lot due to the asymptotic normality of maximum likelihood estimators.