The Multivariate Normal Distribution
Edps/Soc 584 and Psych 594
Applied Multivariate Statistics

Carolyn J. Anderson

Department of Educational Psychology

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

©Board of Trustees, University of Illinois

Outline

- Motivation
- The multivariate normal distribution
- The Bivariate Normal Distribution
- More properties of multivariate normal
- Estimation of μ and Σ
- Central Limit Theorem

Reading: Johnson & Wichern pages 149–176

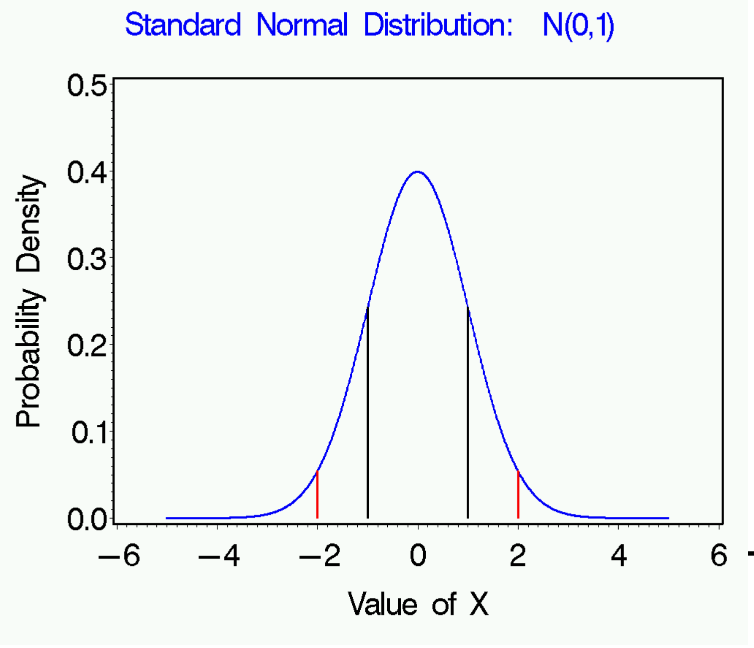
Motivation

- To be able to make inferences about populations, we need a model for the distribution of random variables $\longrightarrow$ We'll use the multivariate normal distribution, because. . .
- It's often a good population model. It's a reasonably good approximation of many phenomenon. A lot of variables are approximately normal (due to the central limit theorem for sums and averages).
- The sampling distribution of (test) statistics are often approximately multivariate or univariate normal due to the central limit theorem.
- Due to it's central importance, we need to thoroughly understand and know it's properties.

- The probability density function of the Univariate normal distribution ($p = 1$ variables):

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} \qquad \text{for } -\infty < x < \infty$$

- The parameters that completely characterize the distribution:
  - $\mu = E(X) =$ mean
  - $\sigma^2 =$ var$(X) =$ variance
- Area corresponds to probability:
  68% area between $\mu \pm \sigma$ and 95% between $\mu \pm 1.96\sigma$:



Generalization to Multivariate Normal

$$\left(\frac{x-\mu}{\sigma}\right)^2 = (x-\mu)(\sigma^2)^{-1}(x-\mu)$$

This is a squared statistical distance of x from μ in standard deviation units.

Generalization to $p > 1$ variables:

- We have $\boldsymbol{x}_{p\times 1}$ and parameters $\boldsymbol{\mu}_{p\times 1}$ and $\boldsymbol{\Sigma}_{p\times p}$.
- The exponent term for multivariate normal is

$$(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$$

  where $-\infty < x_i < \infty$ for $i = 1, \ldots, p$.
  - This is a scalar and reduces to what's at the top for $p = 1$.
  - It is a squared statistical distance of $\boldsymbol{x}$ to $\boldsymbol{\mu}$ (if $\boldsymbol{\Sigma}^{-1}$ exists). It takes into consideration both variability and covariability.
  - Integrating

$$\int_{x_1} \ldots \int_{x_p} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) = (2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}$$

Proper Distribution

Since the sum of probabilities over all possible values must add up to 1, we need to divide by $(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}$ to get a "proper" density function.

Multivariate Normal density function:

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

where $-\infty < x_i < \infty$ for $i = 1, \ldots, p$.

To denote this, we use

$$\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

For $p = 1$, this reduces to the univariate normal p.d.f.

Bivariate Normal: $p = 2$

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \qquad E(\boldsymbol{x}) = \begin{pmatrix} E(x_1) \\ E(x_2) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \boldsymbol{\mu}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}$$

and

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{\sigma_{11}\sigma_{22} - \sigma_{12}^2} \begin{pmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{12} & \sigma_{11} \end{pmatrix}$$

If we replace σ_{12} by $\rho_{12}\sqrt{\sigma_{11}\sigma_{22}}$, then we get

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{\sigma_{11}\sigma_{22}(1 - \rho_{12}^2)} \begin{pmatrix} \sigma_{22} & -\rho_{12}\sqrt{\sigma_{11}\sigma_{22}} \\ -\rho_{12}\sqrt{\sigma_{11}\sigma_{22}} & \sigma_{11} \end{pmatrix}$$

Using this, let's look at the statistical distance of $\boldsymbol{x}$ from $\boldsymbol{\mu}$. . .

Bivariate Normal & Statistical Distance

The quantity in the exponent of the bivariate normal is

$$(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$$

$$= ((x_1-\mu_1),(x_2-\mu_2))\left(\frac{1}{\sigma_{11}\sigma_{22}(1-\rho_{12}^2)}\right)$$

$$\times\begin{pmatrix}\sigma_{22} & -\rho_{12}\sqrt{\sigma_{11}\sigma_{22}}\\ -\rho_{12}\sqrt{\sigma_{11}\sigma_{22}} & \sigma_{11}\end{pmatrix}\begin{pmatrix}x_1-\mu_1\\ x_2-\mu_2\end{pmatrix}$$

$$= \frac{1}{1-\rho_{12}^2}\left\{\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)^2+\left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)^2-2\rho_{12}\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)\left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)\right\}$$

$$= \frac{1}{1-\rho_{12}^2}\left\{z_1^2+z_2^2-2\rho_{12}z_1z_2\right\}$$

Bivariate Normal & Independence

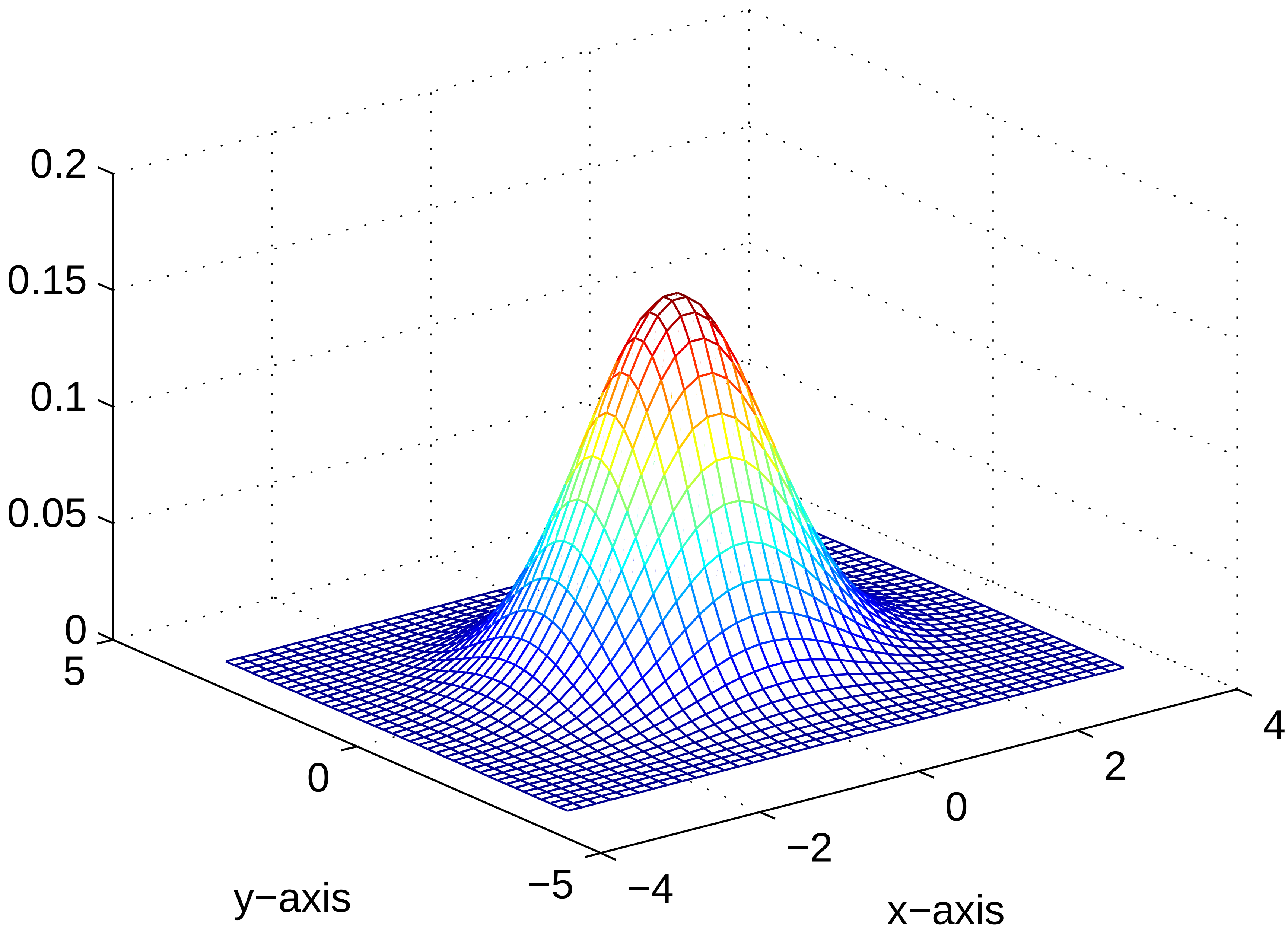
$$
\begin{aligned}
f(\boldsymbol{x}) &= \frac{1}{2\pi\sqrt{\sigma_{11}\sigma_{22}}} \exp\left[\frac{-1}{2(1-\rho_{12}^2)}\left\{\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)^2 + \left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)^2 \right.\right. \\
&\qquad \left.\left. -2\rho_{12}\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)\left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)\right\}\right]
\end{aligned}
$$

If $\sigma_{12} = 0$ or equivalently $\rho_{12} = 0$, then X_1 and X_2 are uncorrelated.

For bivariate normal, $\sigma_{12} = 0$ implies that X_1 and X_2 are statistically independent, because the density factors

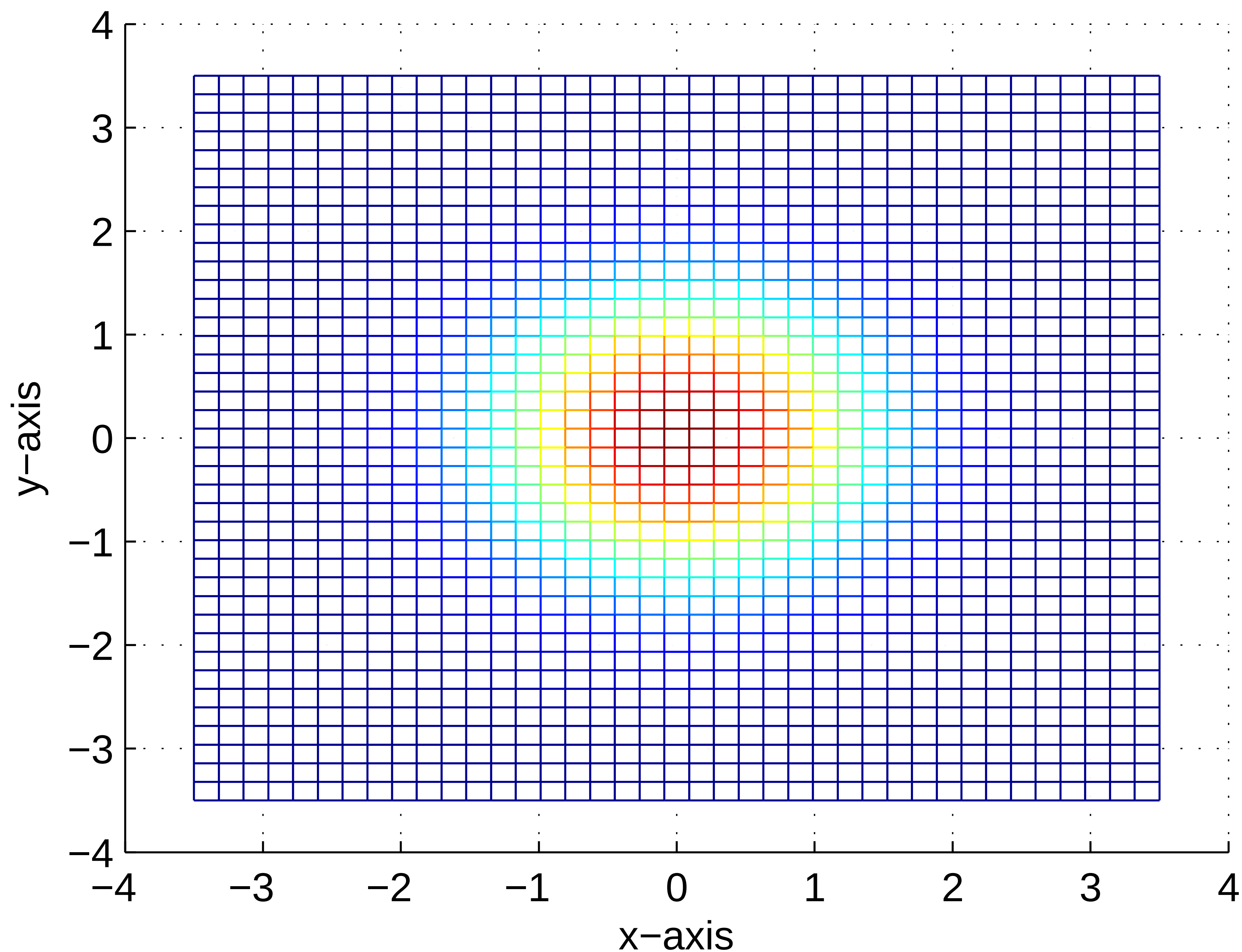
$$
\begin{aligned}
f(\boldsymbol{x}) &= \frac{1}{2\pi\sqrt{\sigma_{11}\sigma_{22}}} \exp\left[\frac{-1}{2}\left\{\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)^2 + \left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)^2\right\}\right] \\
&= \frac{1}{\sqrt{2\pi\sigma_{11}}} \exp\left[\frac{-1}{2}\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)^2\right] \frac{1}{\sqrt{2\pi\sigma_{22}}} \exp\left[\frac{-1}{2}\left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)^2\right] \\
&= f_1(x_1) \times f_2(x_2)
\end{aligned}
$$

Picture: $\mu_k = 0$, $\sigma_{kk} = 1$, $r = 0.0$



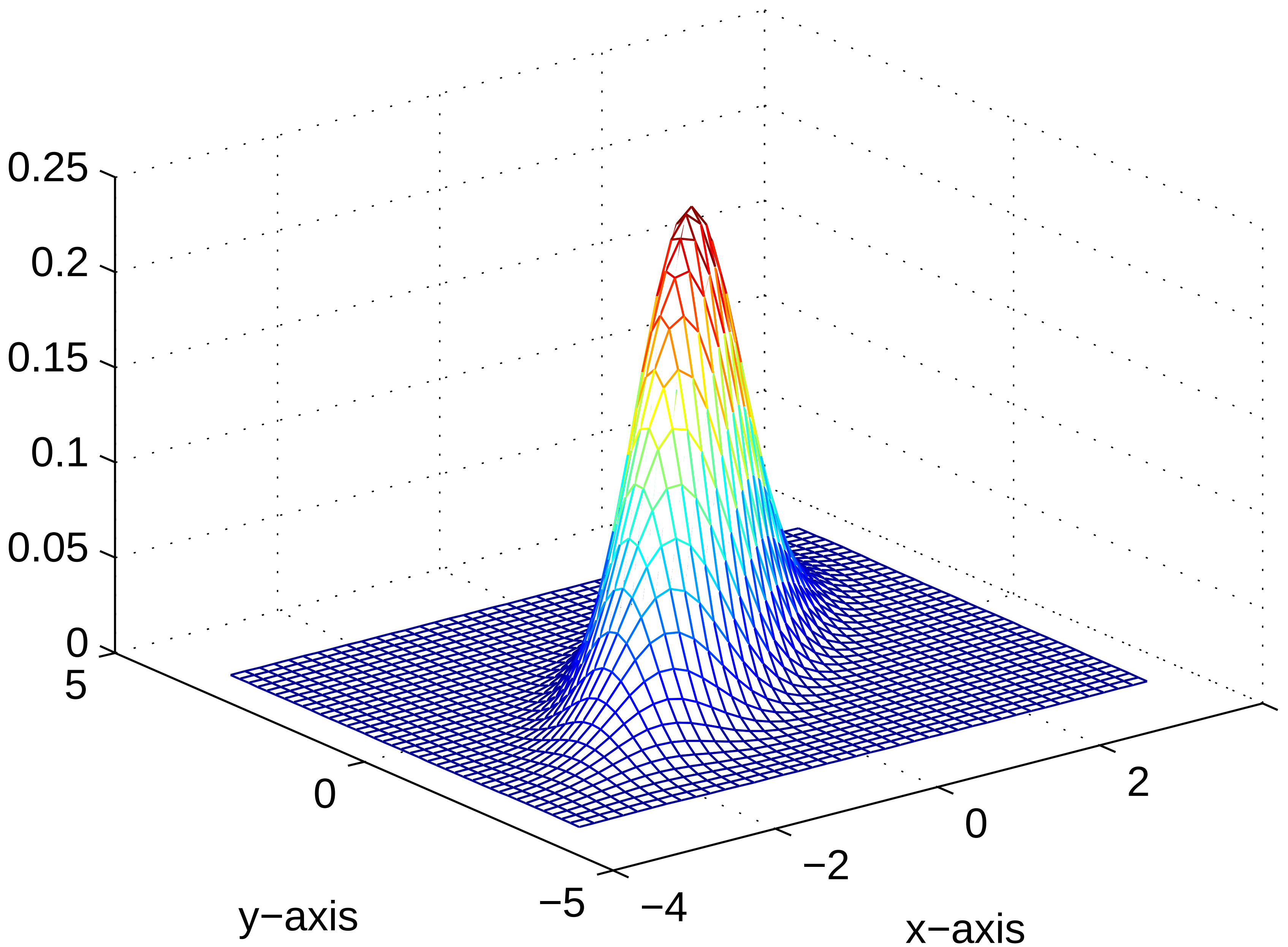
mu x | mu y | sigma x | sigma y

Overhead $\mu_k = 0$, $\sigma_{kk} = 1$, $r = 0.0$



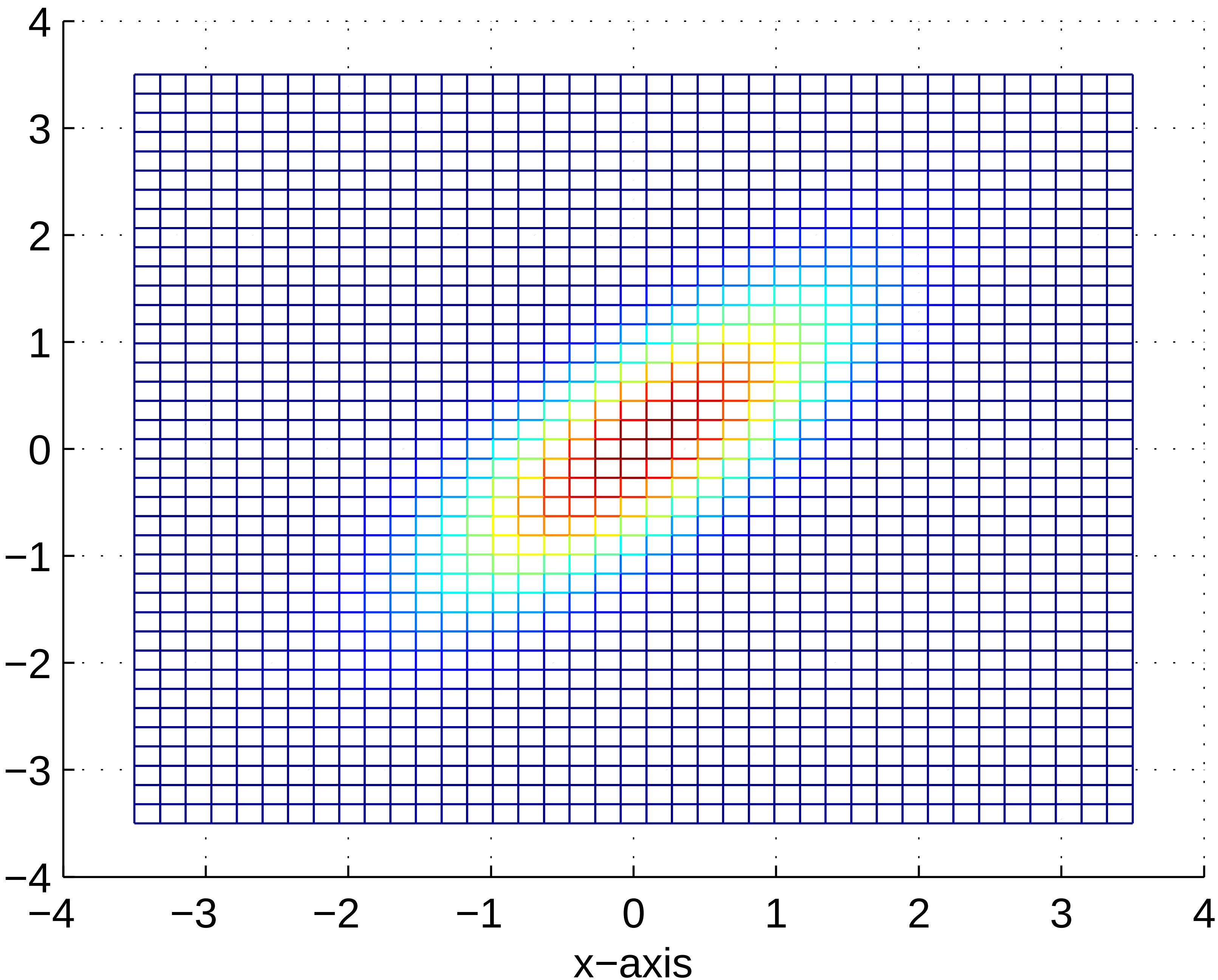
mu x | mu y | sigma x | sigma y

Picture: $\mu_k = 0$, $\sigma_{kk} = 1$, $r = 0.75$



mu x | mu y | sigma x | sigma y

Overhead: $\mu_k = 0$, $\sigma_{kk} = 1$, $r = 0.75$



mu x | mu y | sigma x | sigma y

Summary: Comparing $r = 0.0$ vs $r = 0.75$

For the figures shown, $\mu_1 = \mu_2 = 0$ and $\sigma_{11} = \sigma_{22} = 1$:

- With $r = 0.0$,
  - $\boldsymbol{\Sigma} = \text{diag}(\sigma_{11}, \sigma_{22})$, a diagonal matrix.
  - Density is “random” in the x-y plane.
  - When take a slice parallel to x-y, you get a circle.
- When $r = .75$,
  - $\boldsymbol{\Sigma}$ is not a diagonal .
  - Density is not random in x-y plane.
  - There is a linear tilt (ie., density is concentrated on a line).
  - When you take a slice you get an ellipse that’s tilted.
  - Tilt depends on relative values of σ_{11} and σ_{22} (and scale used in plotting).
- When $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$ (i.e., diagonal with equal variances), it’s “spherical normal”.

Real Time Software Demo

- binormal.m (Peter Dunn)
- Graph_Bivariate_.R (http://www.stat.ucl.ac.be/ISpersonnel/lecoutre/stats/fichiers/˜gallery.pdf)

Slices of Multivariate Normal Density

- For bi-variate normal, you get an ellipse whose equation is

$$(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = c^2$$

which gives all (x_1, x_2) pairs with constant probability.

- The ellipses are call contours and all are centered around $\boldsymbol{\mu}$.
- Definition:

A constant probability contour equals

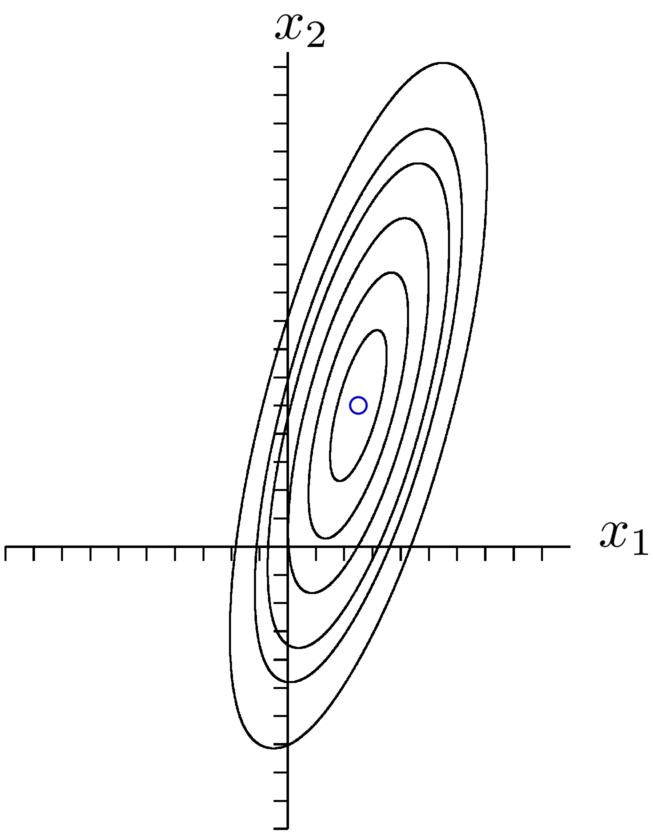
$$\begin{aligned} &= \{\text{all } \boldsymbol{x} \text{ such that } (\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = c^2\} \\ &= \{\text{surface of ellipsoid centered at } \boldsymbol{\mu}\} \end{aligned}$$

Example Probability Contours

Suppose that $\boldsymbol{x} \sim N_2$ with

$$\boldsymbol{\mu} = \begin{pmatrix} 5 \\ 10 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{pmatrix} 9 & 16 \\ 16 & 64 \end{pmatrix} \rightarrow \rho = .67$$

Probability contours for 99%, 95%, 90%, 75%, 50%, and 20%:



Eigenvectors & Values

To get the axes of the ellipsoids of probability contours. . . a little more on eigenvectors and eigenvalues . . .

- Recall that eigenvectors $\boldsymbol{e}_i$ and eigenvalues λ_i are solutions to

$$\boldsymbol{\Sigma}\boldsymbol{e}_i = \lambda_i \boldsymbol{e}_i \quad i = 1, 2, \ldots p$$

  where $\boldsymbol{\Sigma}_{p\times p}$ symmetric and $\boldsymbol{e}_i$ is a $(p \times 1)$ vector.

- If $\boldsymbol{\Sigma}$ is positive definite (so that $\boldsymbol{\Sigma}^{-1}$ exists), then

$$\boldsymbol{\Sigma}\boldsymbol{e} = \lambda\boldsymbol{e} \quad \text{implies} \quad \boldsymbol{\Sigma}^{-1}\boldsymbol{e} = \left(\frac{1}{\lambda}\right)\boldsymbol{e}$$

Probability Contours: Axes of ellipsoid

- The axes of the ellipsoid are
  - In the direction of the eigenvectors of $\mathbf{\Sigma}^{-1}$ (or $\mathbf{\Sigma}$)
  - Their lengths are proportional to the square roots of the eigenvalues of $\mathbf{\Sigma}$ (or reciprocals of square roots of eigenvalues of $\mathbf{\Sigma}^{-1}$).

- Specifically:

$$\text{axes are} \quad \boldsymbol{\mu} \pm c\sqrt{\lambda_i}\,\boldsymbol{e}_i$$

where $\boldsymbol{e}_i$ and λ_i are the i^{th} eigenvectors and eigenvalues of $\mathbf{\Sigma}$.

- What's c?

- The nature of $\mathbf{\Sigma}^{-1}$ determines the shape of the ellipses.

Probability Contours: Axes of ellipsoid

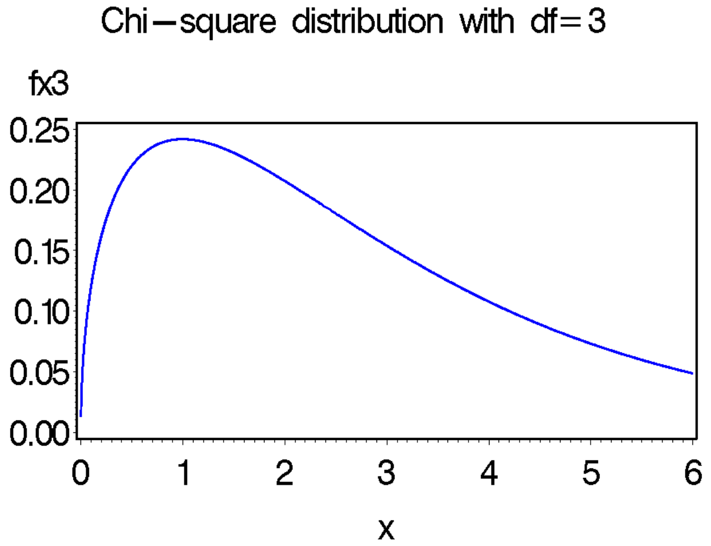
Important Points:

1. $(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \sim \chi^2_p$ (if $|\boldsymbol{\Sigma}| > 0$)

2. The solid ellipsoid of values $\boldsymbol{x}$ that satisfy

$$(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \leq c^2 = \chi^2_{p(\alpha)}$$

has probability $(1-\alpha)$ where $\chi^2_{p(\alpha)}$ is the $(1-\alpha)^{th}100\%$ point of the chi-square distribution with p degrees of freedom.



Example: Axses of Ellipses & Prob. Contours

Back to the example where $\boldsymbol{x} \sim N_2$ with

$$\boldsymbol{\mu} = \begin{pmatrix} 5 \\ 10 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{pmatrix} 9 & 16 \\ 16 & 64 \end{pmatrix} \rightarrow \rho = .667$$

and we want the "95% probability contour".

The upper 5% point of the chi-square distribution with 2 degrees of freedom is $\chi^2_{2(.05)} = 5.9915$, so $c = \sqrt{5.9915} = 2.4478$

Axes: $\boldsymbol{\mu} \pm c\sqrt{\lambda_i}\boldsymbol{e}_i$ where $(\lambda_i, \boldsymbol{e}_i)$ is the i^{th} ($i = 1, 2$) eigenvalue/eigenvector pair of $\boldsymbol{\Sigma}$.

$$\lambda_1 = 68.316 \qquad \boldsymbol{e}_1' = (.2604, .9655)$$
$$\lambda_2 = 4.684 \qquad \boldsymbol{e}_2' = (.9655, -.2604)$$

Major Axis

Using the largest eigenvalue and corresponding eigenvector:

$$\underbrace{\begin{pmatrix} 5 \\ 10 \end{pmatrix}}_{\boldsymbol{\mu}} \pm \underbrace{2.45}_{\sqrt{\chi^2_{2(.05)}}} \underbrace{\sqrt{68.316}}_{\lambda_1} \underbrace{\begin{pmatrix} .2604 \\ .9655 \end{pmatrix}}_{\boldsymbol{e}_1}$$

$$\begin{pmatrix} 5 \\ 10 \end{pmatrix} \pm 20.250 \begin{pmatrix} .2604 \\ .9655 \end{pmatrix}$$

$$\begin{pmatrix} 5 \\ 10 \end{pmatrix} \pm \begin{pmatrix} 5.273 \\ 19.551 \end{pmatrix}$$

$$\longrightarrow \begin{pmatrix} -.273 \\ -9.551 \end{pmatrix}, \quad \begin{pmatrix} 10.273 \\ 29.551 \end{pmatrix}$$

Minor Axis

Same process but now use λ_2 and e_2, the smallest eigenvalue and corresponding eigenvector:

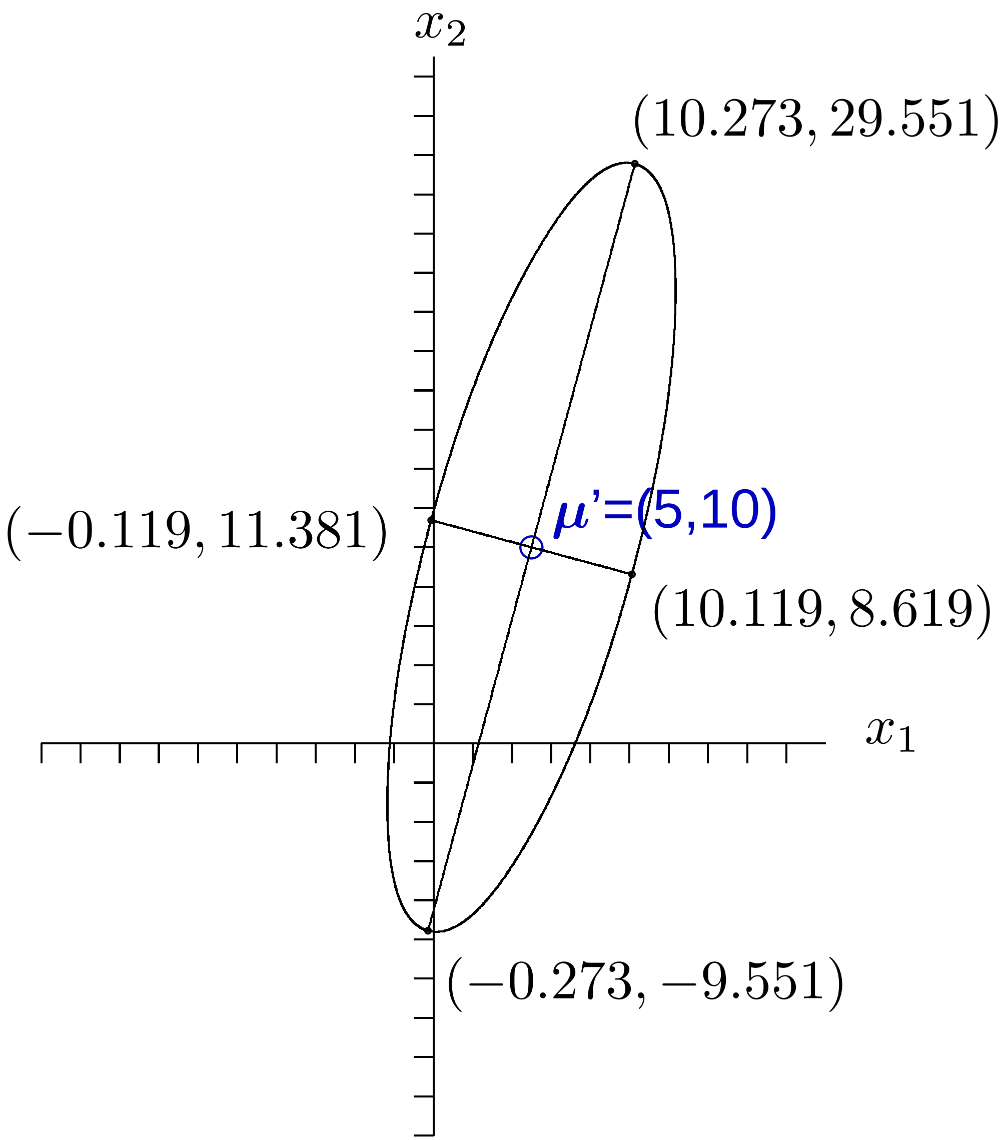
$$\begin{pmatrix} 5 \\ 10 \end{pmatrix} \pm 2.45\sqrt{4.684}\begin{pmatrix} .9655 \\ -.2604 \end{pmatrix}$$

$$\begin{pmatrix} 5 \\ 10 \end{pmatrix} \pm 5.30\begin{pmatrix} .9655 \\ -.2604 \end{pmatrix}$$

$$\begin{pmatrix} 5 \\ 10 \end{pmatrix} \pm \begin{pmatrix} 5.119 \\ -1.381 \end{pmatrix}$$

$$\longrightarrow \begin{pmatrix} -.119 \\ 11.381 \end{pmatrix}, \quad \begin{pmatrix} 10.119 \\ 8.619 \end{pmatrix}$$

Graph of 95% Probability Contour



Example: Equation for Contour

Equation for Contour:

$$(\boldsymbol{x}-\boldsymbol{\mu})' \quad \boldsymbol{\Sigma}^{-1} \quad (\boldsymbol{x}-\boldsymbol{\mu}) \quad \leq \quad 5.99$$

$$((x_1-5),(x_2-10))\begin{pmatrix} 9 & 16 \\ 16 & 64 \end{pmatrix}^{-1}\begin{pmatrix} (x_1-5) \\ (x_2-10) \end{pmatrix} \leq 5.99$$

$$((x_1-5),(x_2-10))\begin{pmatrix} .200 & -.050 \\ -.050 & .028 \end{pmatrix}\begin{pmatrix} (x_1-5) \\ (x_2-10) \end{pmatrix} \leq 5.99$$

$$.2(x_1-5)^2 + .028(x_2-10)^2 - .1(x_1-5)(x_2-10) \leq 5.99$$

$(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$ is a quadratic form, which is equation for a polynomial

Points inside or outside?

Are the following points inside or outside the 95% probability contour?

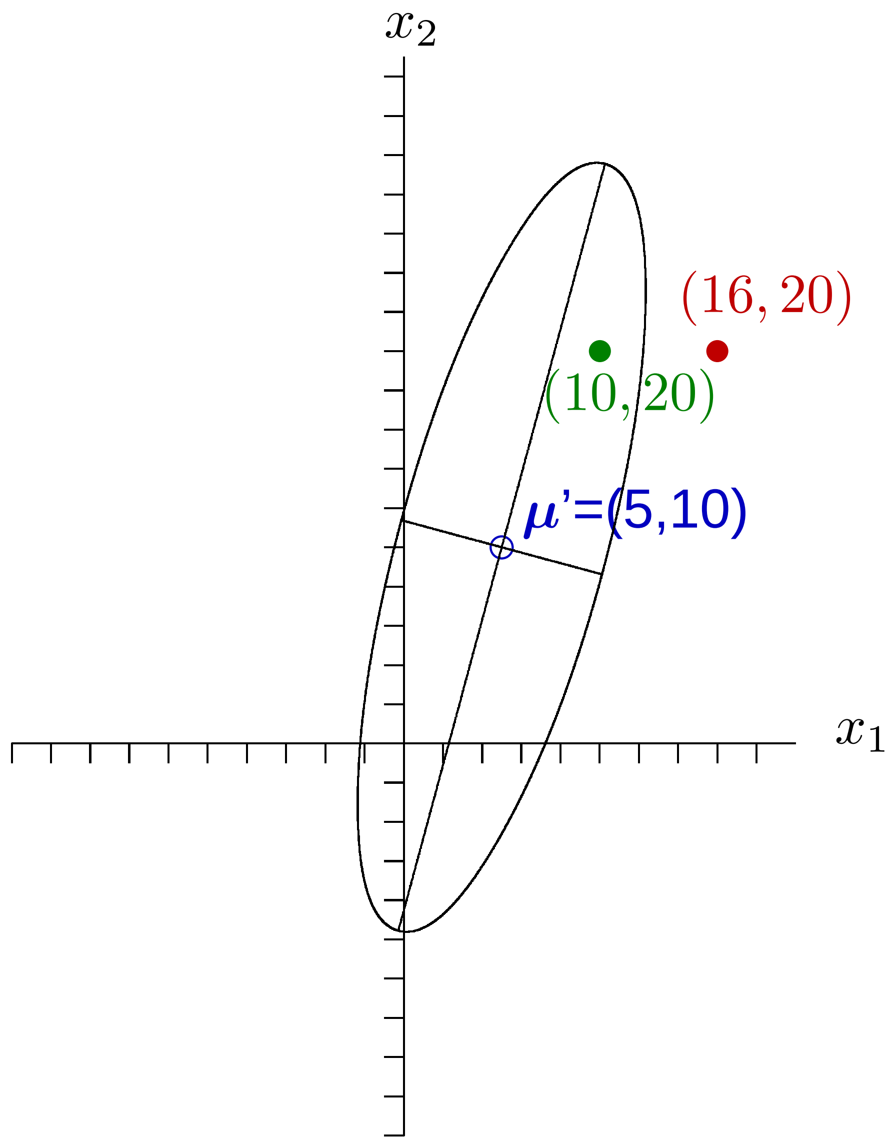
- Is the point (10,20) inside or outside the 95% probability contour?

$$
\begin{aligned}
(10, 20) \quad \longrightarrow \quad & .2(10-5)^2 + .028(20-10)^2 - .1(10-5)(20-10) \\
& = .2(25) + .028(100) - .1(50) \\
& = 2.8
\end{aligned}
$$

- Is the point (16,20) inside or outside the 95% probability contour?

$$
\begin{aligned}
(16, 20) \quad \longrightarrow \quad & .2(16-5)^2 + .028(20-10)^2 - .1(16-5)(20-10) \\
& .2(121) + .028(100) - .1(11)(10) \\
& = 16
\end{aligned}
$$

Points Inside and Outside



More Properties that we'll Expand on

If $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then

1. Linear combinations of components of $\boldsymbol{X}$ are (multivariate) normal.
2. All sub-sets of the components of $\boldsymbol{X}$ are (multivariate) normal.
3. Zero covariance implies that the corresponding components of $\boldsymbol{X}$ are statistical independent.
4. The conditional distributions of the components of $\boldsymbol{X}$ are (multivariate) normal.

1: Linear Combinations

If $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then any linear combination

$$\boldsymbol{a}'\boldsymbol{X} = a_1X_1 + a_2X_2 + \cdots + a_pX_p$$

is distributed as

$$\boldsymbol{a}'\boldsymbol{X} \sim \mathcal{N}_1(\boldsymbol{a}'\boldsymbol{\mu}, \boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a})$$

Also, If $\boldsymbol{a}'\boldsymbol{X}$ is normal $\mathcal{N}(\boldsymbol{a}'\boldsymbol{\mu}, \boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a})$ for all possible $\boldsymbol{a}$, then $\boldsymbol{X}$ must be $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

$$\boldsymbol{X} \sim \mathcal{N}\left(\begin{pmatrix} 5 \\ 10 \end{pmatrix}, \begin{pmatrix} 16 & 12 \\ 12 & 36 \end{pmatrix}\right) \qquad \begin{aligned} \boldsymbol{a}' &= (3, 2) \\ Y = \boldsymbol{a}'\boldsymbol{X} &= 3X_1 + 2X_2 \end{aligned}$$

$$\mu_Y = (3, 2)\begin{pmatrix} 5 \\ 10 \end{pmatrix} = 35 \quad \text{and } \sigma_Y^2 = (3, 2)\begin{pmatrix} 16 & 12 \\ 12 & 36 \end{pmatrix}\begin{pmatrix} 3 \\ 2 \end{pmatrix} = 432$$

$$Y \sim \mathcal{N}(35, 432)$$

More Linear Combinations

If $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then the q linear combinations

$$\boldsymbol{Y}_{q\times 1} = \boldsymbol{A}_{q\times p}\boldsymbol{X} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{q1} & a_{q2} & \cdots & a_{qp} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}$$

is distributed as $\mathcal{N}_q(\boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}')$.

Also, if

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{d},$$

where $\boldsymbol{d}_{q\times 1}$ is a vector constants, then

$$\boldsymbol{Y} = \mathcal{N}(\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{d}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}').$$

Numerical Example with Multiple Combinations

$$\boldsymbol{X} \sim \mathcal{N}_2\left(\begin{pmatrix} 5 \\ 10 \end{pmatrix} \begin{pmatrix} 16 & 12 \\ 12 & 36 \end{pmatrix}\right)$$

$$\begin{aligned} Y_1 &= X_1 + X_2 \\ Y_2 &= X_1 - X_2 \end{aligned} \quad \text{so} \quad \boldsymbol{A}_{2\times 2} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

$$\mu_Y = \boldsymbol{A}\boldsymbol{\mu} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 5 \\ 10 \end{pmatrix} = \begin{pmatrix} 15 \\ -5 \end{pmatrix}$$

$$\boldsymbol{\Sigma}_Y = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}' = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 16 & 12 \\ 12 & 36 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} = \begin{pmatrix} 76 & -20 \\ -20 & 28 \end{pmatrix}$$

So

$$\boldsymbol{Y} \sim \mathcal{N}_2\left(\begin{pmatrix} 15 \\ -5 \end{pmatrix}, \begin{pmatrix} 76 & -20 \\ -20 & 28 \end{pmatrix}\right)$$

Multiple Regression as an Example

This example will use what we know about linear combinations and now what we know about the distribution of linear combinations.

Linear Regression Model

- $Y =$ response variable.
- $Z_1, Z_2, \ldots, Z_r$ are predictor/explanatory variables, which are considered to be fixed.
- The model is

$$Y = \beta_o + \beta_1 Z_1 + \beta_2 Z_2 + \ldots + \beta_r Z_r + \epsilon$$

- The error of prediction ϵ is viewed as a random variable.

Multiple Regression as an Example

Suppose we have n observations on Y and have values of Z_i for all $i = 1, \ldots, n$; that is,

$$\begin{aligned}
Y_1 &= \beta_o + \beta_1 Z_{11} + \beta_2 Z_{12} + \ldots + \beta_r Z_{1r} + \epsilon_1 \\
Y_2 &= \beta_o + \beta_1 Z_{21} + \beta_2 Z_{22} + \ldots + \beta_r Z_{2r} + \epsilon_2 \\
\vdots & \quad \vdots \\
Y_n &= \beta_o + \beta_1 Z_{n1} + \beta_2 Z_{n2} + \ldots + \beta_r Z_{nr} + \epsilon_n
\end{aligned}$$

where $E(\epsilon_j) = 0$, var$(\epsilon_j) = \sigma^2$ (a constant), and cov$(\epsilon_j, \epsilon_k) = 0$ for $j \neq k$.
In terms of matrices,

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & Z_{11} & Z_{12} & \ldots & Z_{1r} \\ 1 & Z_{21} & Z_{22} & \ldots & Z_{2r} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & Z_{n1} & Z_{n2} & \ldots & Z_{nr} \end{pmatrix} \begin{pmatrix} \beta_o \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_r \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

$$\boldsymbol{Y} = \boldsymbol{Z\beta} + \boldsymbol{\epsilon} \quad \text{where } E(\boldsymbol{\epsilon}) = \boldsymbol{0} \text{ and cov}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}.$$

Distribution of Y

$$\boldsymbol{Y} = \underbrace{\boldsymbol{Z\beta}}_{\text{vector of constants}} + \underbrace{\boldsymbol{\epsilon}}_{\text{random}} \quad \text{where } E(\boldsymbol{\epsilon}) = \boldsymbol{0} \text{ and } \text{cov}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{I}.$$

So $\boldsymbol{Y}$ is a linear combination of a multivariate normally distributed variable, $\boldsymbol{\epsilon}$.

- Mean of $\boldsymbol{Y}$:

$$\boldsymbol{\mu_Y} = E(\boldsymbol{Y}) = E(\boldsymbol{Z\beta} + \boldsymbol{\epsilon}) = \boldsymbol{Z\beta} + E(\boldsymbol{\epsilon}) = \boldsymbol{Z\beta}$$

- Covariance of $\boldsymbol{Y}$:

$$\boldsymbol{\Sigma_Y} = \sigma^2\boldsymbol{I}$$

  (the same as $\boldsymbol{\epsilon}$).

- Distribution of $\boldsymbol{Y}$ is multivariate normal because $\boldsymbol{\epsilon}$ is multivariate normal:

$$\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{Z\beta}, \sigma^2\boldsymbol{I})$$

Least Square Estimation

$$\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad \text{where } E(\boldsymbol{\epsilon}) = \boldsymbol{0} \text{ and cov}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}$$

$\boldsymbol{\beta}$ and σ^2 are unknown parameters that need to be estimated from data.

Let $y_1, y_2, \ldots, y_n$ be a random sample with values $\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_r$ on the explanatory variables. The least squares estimate of $\boldsymbol{\beta}$ is the vector $\boldsymbol{b}$ that minimizes

$$\begin{aligned}\sum_{j=1}^{n}(y_j - \boldsymbol{z}_j'\boldsymbol{b})^2 &= \sum_{j=1}^{n}(y_j - b_o - b_1 z_{j1} - b_2 z_{j2} - \ldots - b_r z_{jr})^2 \\ &= (\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{b})'(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{b}) \\ &= \boldsymbol{\epsilon}'\boldsymbol{\epsilon}\end{aligned}$$

where $\boldsymbol{z}_j'$ is the j^{th} row of $\boldsymbol{Z}$, and $\boldsymbol{b} = (b_o, b_1, b_2, \ldots, b_r)'$.
If $\boldsymbol{Z}$ has full rank (i.e., the rank of $\boldsymbol{Z}$ is $r + 1 \leq n$), then the least squares estimate of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{y}$$

What's the distribution of $\hat{\beta}$?

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{y} = \boldsymbol{A}\boldsymbol{y}$$

We showed that $\boldsymbol{Y} \sim \mathcal{N}_n(\boldsymbol{Z}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$.

- Mean of $\hat{\boldsymbol{\beta}}$:

$$\begin{aligned}
\mu_{\hat{\boldsymbol{\beta}}} = E(\hat{\boldsymbol{\beta}}) &= E(\boldsymbol{A}\boldsymbol{Y}) \\
&= \boldsymbol{A}E(\boldsymbol{Y}) \\
&= \boldsymbol{A}\boldsymbol{Z}\boldsymbol{\beta} \\
&= \underbrace{(\boldsymbol{Z}'\boldsymbol{Z})^{-1}}\underbrace{\boldsymbol{Z}'\boldsymbol{Z}}\boldsymbol{\beta} = \boldsymbol{\beta}
\end{aligned}$$

- Covariance matrix for $\hat{\boldsymbol{\beta}}$

$$\begin{aligned}
\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}} &= \boldsymbol{A}\boldsymbol{\Sigma}_Y\boldsymbol{A}' \\
&= ((\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}')(\sigma^2\boldsymbol{I})(\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}) \\
&= \sigma^2(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1} \\
&= \sigma^2(\boldsymbol{Z}'\boldsymbol{Z})^{-1}
\end{aligned}$$

- The distribution of $\hat{\beta}$: $\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{Z}'\boldsymbol{Z})^{-1})$.

The distribution of $\hat{Y}$

The "fitted values" or predicted values are

$$\hat{\boldsymbol{y}} = \boldsymbol{Z}\hat{\boldsymbol{\beta}} = \boldsymbol{H}\boldsymbol{y}$$

where $\boldsymbol{H} = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'$. The matrix $\boldsymbol{H}$ is the "hat" matrix.

- We just showed that $\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{Z}'\boldsymbol{Z})^{-1})$, and so $\hat{\boldsymbol{y}}$ is a linear combination of a vector that's multivariate normal.
- Mean of $\hat{\boldsymbol{Y}}$:

$$\mu_{\hat{\mathbf{Y}}} = E(\boldsymbol{Z}\hat{\boldsymbol{\beta}}) = \boldsymbol{Z}E(\hat{\boldsymbol{\beta}}) = \boldsymbol{Z}\boldsymbol{\beta}$$

- Covariance matrix for $\hat{\boldsymbol{Y}}$

$$\boldsymbol{Z}\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}}\boldsymbol{Z}' = \boldsymbol{Z}(\sigma^2\underbrace{(\boldsymbol{Z}'\boldsymbol{Z})^{-1}})\underbrace{\boldsymbol{Z}'} = \sigma^2\boldsymbol{Z}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1} = \sigma^2\boldsymbol{I}$$

- Distribution of $\hat{\boldsymbol{Y}}$:

$$\hat{\boldsymbol{Y}} \sim \mathcal{N}(\boldsymbol{Z}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$$

The distribution of $\hat{\epsilon}$

The estimated residuals are

$$\hat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$$

and they contain the information necessary to estimate σ^2.

The least squares estimate of σ^2 is

$$s^2 = \frac{\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}}{n - (r+1)}$$

The estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\epsilon}}$ are uncorrelated.

Multivariate Normality Assumption $\boldsymbol{\epsilon} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ and what we know about linear combinations of random variables allowed us to derive the distribution of various random variables.

Last few comments on this example:

1. The least squares estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\epsilon}$ are also the maximum likelihood estimates.
2. The maximum likelihood estimate of σ^2 is $\hat{\sigma}^2 = \hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}/n$
3. $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\epsilon}}$ are statistically independent.

2: Sub-sets of Variables

If $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then all sub-sets of $\boldsymbol{X}$ are (multivariate) normally distributed.

For example, let's partition $\boldsymbol{X}$ into two sub-sets

$$\boldsymbol{X}_{p\times 1} = \begin{pmatrix} X_1 \\ \vdots \\ X_q \\ \hline X_{q+1} \\ \vdots \\ X_p \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_{1(q\times 1)} \\ \hline \boldsymbol{X}_{2((p-q)\times 1)} \end{pmatrix} \text{ and } \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_q \\ \hline \mu_{q+1} \\ \vdots \\ \mu_p \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu}_{1(q\times 1)} \\ \hline \boldsymbol{\mu}_{2((p-q)\times 1)} \end{pmatrix}$$

$$\boldsymbol{\Sigma}_{p\times p} = \left(\begin{array}{c|c} \boldsymbol{\Sigma}_{11(q\times q)} & \boldsymbol{\Sigma}_{12(q\times(p-q))} \\ \hline \boldsymbol{\Sigma}_{21((p-q)\times p)} & \boldsymbol{\Sigma}_{22((p-q)\times(p-q))} \end{array}\right) = \left(\begin{array}{c|c} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \hline \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{array}\right)$$

Sub-sets of Variables continued

Then for

$$\boldsymbol{X} = \left(\frac{\boldsymbol{X}_{1(q\times 1)}}{\boldsymbol{X}_{2((p-q)\times 1)}} \right)$$

The distributions of the sub-sets are

$$\boldsymbol{X}_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}) \qquad \text{and} \qquad \boldsymbol{X}_2 \sim \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$$

The result means that

- Each of the $\boldsymbol{X}_i$'s are univariate normals (next page)
- All possible sub-sets are multivariate normal.
- All marginal distributions are (multivariate) normal.

Little Example on Sub-sets

Suppose that

$$\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim \mathcal{N}_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Due to the result on sub-sets of multivariate normals,

$$X_1 \sim \mathcal{N}(\mu_1, \sigma_{11})$$
$$X_2 \sim \mathcal{N}(\mu_2, \sigma_{22})$$
$$X_3 \sim \mathcal{N}(\mu_3, \sigma_{33})$$

Also

$$\begin{pmatrix} X_2 \\ X_3 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_2 \\ \mu_3 \end{pmatrix} \begin{pmatrix} \sigma_{22} & \sigma_{23} \\ \sigma_{32} & \sigma_{33} \end{pmatrix}\right)$$

3: Zero Covariance & Statistical Independence

There are three parts to this one:

1. If $\boldsymbol{X}_1$ is $(q_1 \times 1)$ and $\boldsymbol{X}_2$ is $(q_2 \times 1)$ are statistically independent, then cov$(\boldsymbol{X}_1, \boldsymbol{X}_2) = \boldsymbol{\Sigma}_{12} = \boldsymbol{0}$.

2. If
$$\begin{pmatrix} \boldsymbol{X}_1 \\ \hline \boldsymbol{X}_2 \end{pmatrix} \sim \mathcal{N}_{q_1+q_2}\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \hline \boldsymbol{\mu}_2 \end{pmatrix}, \left(\begin{array}{c|c} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \hline \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{array}\right)\right),$$
Then $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are statistically independent if and only if $\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21}' = \boldsymbol{0}$.

3. If $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are statistically independent and distributed as $\mathcal{N}_{q_1}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ and $\mathcal{N}_{q_2}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, respectively, then
$$\begin{pmatrix} \boldsymbol{X}_1 \\ \hline \boldsymbol{X}_2 \end{pmatrix} \sim \mathcal{N}_{q_1+q_2}\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \hline \boldsymbol{\mu}_2 \end{pmatrix}, \left(\begin{array}{c|c} \boldsymbol{\Sigma}_{11} & \boldsymbol{0} \\ \hline \boldsymbol{0} & \boldsymbol{\Sigma}_{22} \end{array}\right)\right).$$

Example

$$\boldsymbol{Y}_{4\times 1} = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{pmatrix} \text{ and } \boldsymbol{\Sigma}_{\boldsymbol{Y}} = \begin{pmatrix} 2 & 1 & 0 & .5 \\ 1 & 3 & 0 & .5 \\ 0 & 0 & 4 & 0 \\ .5 & .5 & 0 & 1 \end{pmatrix}$$

and $\boldsymbol{Y} \sim \mathcal{N}_4(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
Let's take $\boldsymbol{X}_1' = (Y_1, Y_2, Y_4)$ and $\boldsymbol{X}_2' = (Y_3)$.

Then

$$\left(\frac{\boldsymbol{X}_1}{\boldsymbol{X}_2}\right) \sim \mathcal{N}_4\left(\begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_4 \\ \hline \mu_3 \end{pmatrix}, \left(\begin{array}{ccc|c} 2 & 1 & .5 & 0 \\ 1 & 3 & .5 & 0 \\ .5 & .5 & 1 & 0 \\ \hline 0 & 0 & 0 & 4 \end{array}\right)\right)$$

So set $\boldsymbol{X}_1$ is statistically independent of $\boldsymbol{X}_2$.

4: Conditional Distributions

Let $\boldsymbol{X}' = (\boldsymbol{X}_{1(q1\times 1)}, \boldsymbol{X}_{2(q_2\times 1)})$ be distributed at $\mathcal{N}_{q1+q2}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with

$$\boldsymbol{\mu} = \left(\begin{array}{c} \boldsymbol{\mu}_1 \\ \hline \boldsymbol{\mu}_2 \end{array}\right) \quad \text{and} \quad \boldsymbol{\Sigma} = \left(\begin{array}{c|c} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \hline \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{array}\right)$$

and $|\boldsymbol{\Sigma}| > 0$ (i.e., positive definite). Then the <u>conditional distribution</u> of $\boldsymbol{X}_1$ given $\boldsymbol{X}_2 = \boldsymbol{x}_2$ is (multivariate) normal with mean and covariance matrix

$$\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2) \quad \text{and} \quad \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$$

Let's look more closely at this for a simple case of $q_1 = q_2 = 1$.

Conditional Distribution for $q_1 = q_2 = 1$

Bivariate normal distribution

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}_2\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}\right)$$

$$f(x_1|x_2) \quad \text{is} \quad \mathcal{N}_1\left(\mu_1 + \frac{\sigma_{12}}{\sigma_{22}}(x_2 - \mu_2), \sigma_{11} - \sigma_{12}\left(\frac{\sigma_{12}}{\sigma_{22}}\right)\right)$$

Notes:

- $\sigma_{12} = \rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}$
- $\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1} = \sigma_{12}/\sigma_{22} = \rho_{12}(\sqrt{\sigma_{11}}/\sqrt{\sigma_{22}})$
- $\mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21} = \sigma_{11} - \sigma_{12}^2/\sigma_{22} = \sigma_{11}(1 - \rho_{12}^2)$

Alternative way to write $f(x_1|x_2)$:

$$f(x_1|x_2) \quad \text{is} \quad \mathcal{N}_1\left(\mu_1 + \rho_{12}\frac{\sqrt{\sigma_{11}}}{\sqrt{\sigma_{22}}}(x_2 - \mu_2), \sigma_{11}(1 - \rho_{12}^2)\right)$$

What is this?

Multiple Regression as a Conditional Dist.

Consider the case where $q_1 = 1$ and $q_2 > 1$.

- All conditional distributions are normal.
- The conditional covariance matrix $\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$ does not depend on the values of the conditioning variables.
- The conditional means have the following form:

Let $\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} = \boldsymbol{\beta}_{q_1 \times q_2}$

$$\begin{pmatrix} \beta_{1,q_1+1} & \beta_{1,q_1+2} & \cdots & \beta_{1,q_1+q_2} \\ \beta_{2,q_1+1} & \beta_{2,q_1+2} & \cdots & \beta_{2,q_1+q_2} \\ \cdots & \cdots & \ddots & \cdots \\ \beta_{q_1,q_1+1} & \beta_{q_1,q_1+2} & \cdots & \beta_{q_1,q_1+q_2} \end{pmatrix}$$

Condtional means

$$\begin{pmatrix} \mu_1 + \sum_{i=q_1+1}^{q_1+q_2} \beta_{1i}(x_i - \mu_i) \\ \mu_2 + \sum_{i=q_1+1}^{q_1+q_2} \beta_{2i}(x_i - \mu_i) \\ \vdots \\ \mu_{q_1} + \sum_{i=q_1+1}^{q_1+q_2} \beta_{q_1 i}(x_i - \mu_i) \end{pmatrix}$$

Estimation of μ and Σ

& sampling distribution of estimators.

Suppose we have a p dimensional normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

Take n observations $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ (these are each $(p \times 1)$ vectors).

$$\boldsymbol{X}_j \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad j = 1, 2, \ldots, n \text{ and independent}$$

For $p = 1$, we know that the MLEs are

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j \sim \mathcal{N}\left(\mu, \frac{1}{n}\sigma^2\right)$$

And $$n\hat{\sigma}^2 = \sum_{j=1}^{n}(x_j - \bar{x})^2 \text{ and } \frac{1}{\sigma^2}\sum_{j=1}^{n}(x_j - \bar{x})^2 \sim \chi^2_{(n-1)}$$

Or $$\hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})^2 \sim \sigma^2\chi^2_{(n-1)}$$

Estimation of μ and Σ: Multivariate Case

The maximum likelihood estimator of $\boldsymbol{\mu}$ is

$$\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{X}} = \frac{1}{n}\sum_{j=1}^{n} X_j$$

and the ML estimator of $\boldsymbol{\Sigma}$ is

$$\hat{\boldsymbol{\Sigma}} = \frac{n-1}{n}\boldsymbol{S}^2 = \boldsymbol{S}_n = \frac{1}{n}\sum_{j=1}^{n}(\boldsymbol{X}_j - \hat{\boldsymbol{\mu}})(\boldsymbol{X}_j - \hat{\boldsymbol{\mu}})'$$

Sampling Distribution of $\hat{\boldsymbol{\mu}}$:

The estimator is a linear combination of normal random vectors each from $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad i.i.d.$:

$$\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{X}} = \frac{1}{n}\boldsymbol{X}_1 + \frac{1}{n}\boldsymbol{X}_2 + \cdots + \frac{1}{n}\boldsymbol{X}_n$$

So $\hat{\boldsymbol{\mu}}$ also has a normal distribution,

$$\bar{\boldsymbol{X}} \sim \mathcal{N}_p(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma})$$

Sampling Distribution of $\hat{\Sigma}$

$$\hat{\boldsymbol{\Sigma}} = \frac{n-1}{n}\boldsymbol{S}$$

The matrix

$$(n-1)\boldsymbol{S} = \sum_{j=1}^{n}(\boldsymbol{x}_j - \bar{\boldsymbol{x}})(\boldsymbol{x}_j - \bar{\boldsymbol{x}})'$$

is distributed as a Wishart random matrix with $(n-1)$ degrees of freedom.

Whishart distribution:

- A multivariate analogue to the chi-square distribution.
- It's defined as

$$\begin{aligned} W_m(\cdot|\boldsymbol{\Sigma}) &= \text{Wishart distribution with } m \text{ degrees of freedom} \\ &= \text{The distribution of } \sum_{j=1}^{m}\boldsymbol{Z}_j\boldsymbol{Z}_j' \end{aligned}$$

where $\boldsymbol{Z}_j \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$ and independent.

Note : $\bar{\boldsymbol{X}}$ and $\boldsymbol{S}$ are independent.

Data are not always (multivariate) normal

The Law of Large Numbers (for multivariate data):

Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be independent observations from a population with mean $E(\boldsymbol{X}) = \boldsymbol{\mu}$.

Then $\bar{\boldsymbol{X}} = (1/n)\sum_{j=1}^{n} \boldsymbol{X}_j$ converges in probability to μ as n gets large; that is,

$$\bar{\boldsymbol{X}} \rightarrow \boldsymbol{\mu} \text{ for large samples}$$

And

$$\boldsymbol{S} (\text{or } \boldsymbol{S}_n) \text{ approach } \boldsymbol{\Sigma} \text{ for large samples}$$

These are true regardless of the true distribution of the $\boldsymbol{X}_j$'s.

Central Limit Theorem

Let $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_n$ be independent observations from a population with mean $E(\boldsymbol{Y}) = \boldsymbol{\mu}$ and finite (non-singular, full rank), covariance matrix $\boldsymbol{\Sigma}$.

Then $\sqrt{n}(\boldsymbol{X} - \boldsymbol{\mu})$ has an approximate $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ distribution if $n >> p$ (i.e., "much larger than").

So, for "large" n

$$\bar{\boldsymbol{X}} = \text{ Sample mean vector } \approx \mathcal{N}(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma}),$$

regardless of the underlying distribution of the $\boldsymbol{Y}_j$'s.

What if $\boldsymbol{\Sigma}$ is unknown? If n is large "enough", $\boldsymbol{S}$ will be close to $\boldsymbol{\Sigma}$, so

$$\sqrt{n}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}) \approx \mathcal{N}_p(\mathbf{0}, \boldsymbol{S}) \text{ or } \bar{\boldsymbol{X}} \approx \mathcal{N}_p(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{S}).$$

Since $n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}) \sim \chi^2_p$,

$$n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})'\boldsymbol{S}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}) \approx \chi^2_p$$

Few more comments

- Using $\boldsymbol{S}$ instead of $\boldsymbol{\Sigma}$ does not seriously effect approximation.

- n must be large relative to p; that is, $(n-p)$ is large.

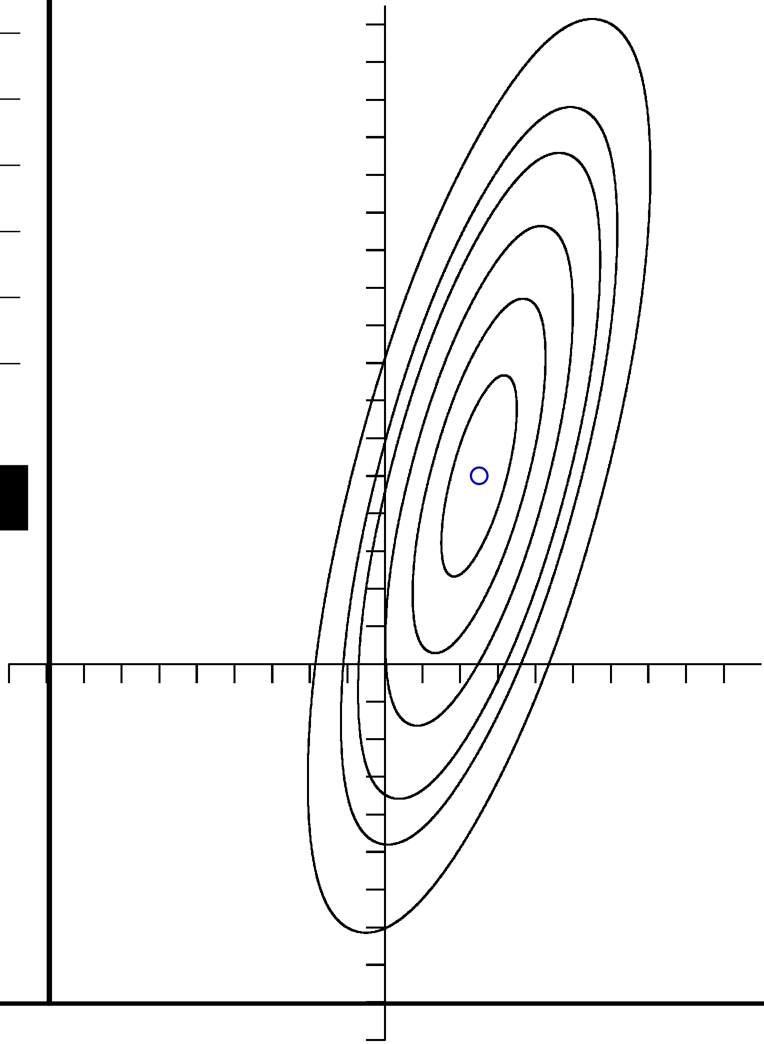
- The probability contours for $\bar{\boldsymbol{X}}$ are tighter than those for $\boldsymbol{X}$ since we have $(1/n)\boldsymbol{\Sigma}$ for $\bar{\boldsymbol{X}}$ rather than $\boldsymbol{\Sigma}$ for $\boldsymbol{X}$.

  See next slide for an example of the latter.

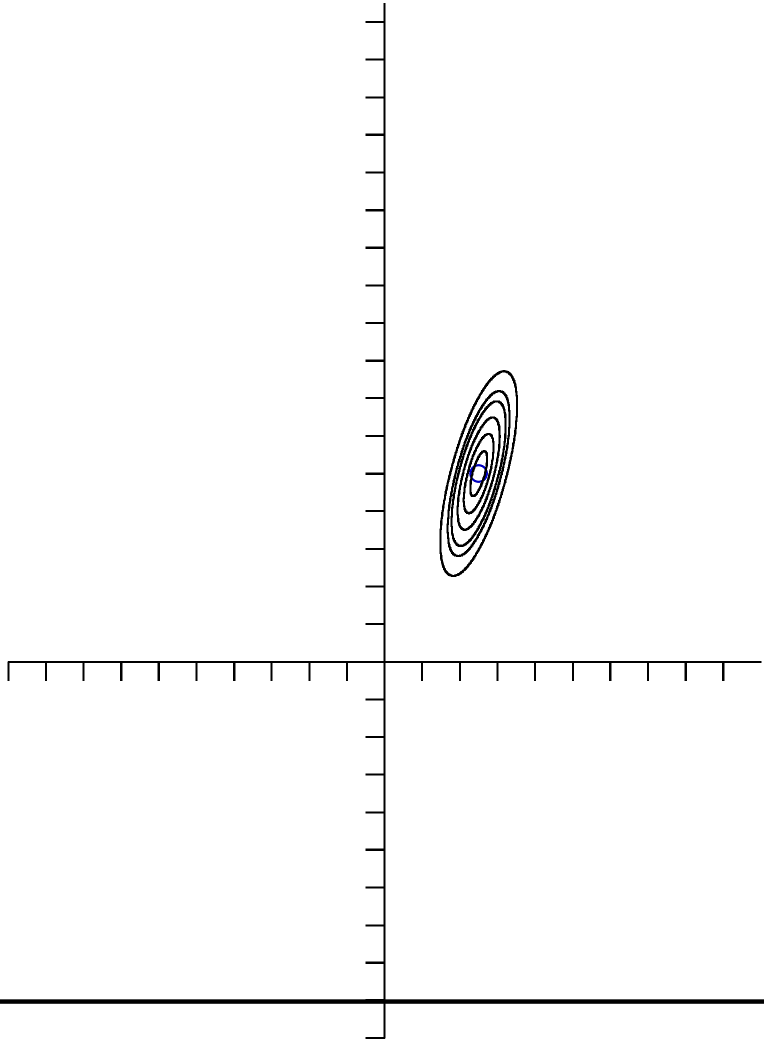
Comparison of Probability Contours

Returning to our example and pretending we have $n = 20$.
Below are contours for 99%, 95%, 90%, 75%, 50% and 20%:

Contours for $\boldsymbol{X}_j$



Contours for $\bar{\boldsymbol{X}}$



Why So Much a Difference with Only 20?

For $\boldsymbol{X}_j$

$$\boldsymbol{\Sigma} = \begin{pmatrix} 9 & 16 \\ 16 & 64 \end{pmatrix} \longrightarrow \lambda_1 = 68.316 \text{ and } \lambda_2 = 4.684$$

For $\bar{\boldsymbol{X}}$ with $n = 20$

$$\boldsymbol{\Sigma} = \frac{1}{20}\begin{pmatrix} 9 & 16 \\ 16 & 64 \end{pmatrix} = \begin{pmatrix} 0.45 & 0.80 \\ 0.80 & 3.20 \end{pmatrix} \longrightarrow \lambda_1 = 3.42 \text{ and } \lambda_2 = 0.23$$

Note that $68.316/20 = 3.42$ and $4.684/20 = 0.23$.