Multivariate Stats, Week 4

Noah Silbert

September 15, 2017

To do

1. homework review
2. properties of multivariate normal random variables

Probability & Random Variables

Probability maps *events* to *numbers* ranging from 0 to 1.

We denote a possible event E, and we denote the set of all possible events Ω.

The probabilities of events E_i, $i \in \{1, 2, ...\}$ obey the following axioms of probability theory:

1. $0 \leq \Pr(E) \leq 1$
2. $\Pr(\Omega) = 1$
3. For mutually exclusive events E_i, $i \in \{1, 2, ...\}$:

$$\Pr\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \Pr(E_i)$$

An Example

A coin flip is a canonical example used to illustrate probability. We denote a possible event E. The set of possible events for a single coin flip is: $E \in \{H, T\}$. In this case $\Omega = \{H, T\}$.

The axioms for this case are:

1. $0 \leq \Pr(H), \Pr(T) \leq 1$
2. $\Pr(H \cup T) = 1$
3. Because H and T are mutually exclusive:

$$\Pr(H \cup T) = \Pr(H) + \Pr(T)$$

Joint Probabilities, Independence

Rolling a six-sided die is also useful for illustrating joint and conditional probabilities. Suppose we roll a fair die, i.e., $\Pr(X = k) = \frac{1}{6}$, $k \in \{1, 2, 3, 4, 5, 6\}$

The joint probability of two events E and F is denoted $\Pr(E, F)$ or $\Pr(E \cap F)$. With our fair die, we can estimate the joint probability of rolling an even number *and* a number greater than 2:

$$\begin{aligned}
\Pr\left(X \in \{2, 4, 6\} \cap X > 2\right) &= \Pr(X \in \{4, 6\}) \\
&= \Pr(X = 4 \cup X = 6) \\
&= \frac{1}{6} + \frac{1}{6} = \frac{1}{3}
\end{aligned}$$

Two events E, F are independent if $\Pr(E, F) = \Pr(E)\Pr(F)$.

Conditional Probabilities

The conditional probability of event E given event F is denoted $\Pr(E|F)$ and is defined as $\Pr(E|F) = \frac{\Pr(E,F)}{\Pr(F)}$. Rearranging terms, we see that $\Pr(E,F) = \Pr(E|F)\Pr(F)$.

For example, the probability of rolling an even number given that you've rolled a number greater than 2 is:

$$\begin{aligned}\Pr(X \in \{2,4,6\}|X > 2) &= \frac{\Pr(X \in \{2,4,6\} \cap X > 2)}{\Pr(X > 2)} \\ &= \frac{1/3}{2/3} = \frac{1}{2}\end{aligned}$$

Events E and F are conditionally independent given G if $\Pr(E,F|G) = \Pr(E|G)\Pr(F|G)$.

Random Variables

Random variables (RVs) are *numerical* events (or numbers corresponding to non-numerical events) that are mapped to probabilities.

For discrete random variables, a probability mass function relates the events and probabilities. For example, a coin flip is a Bernoulli RV. If we map H to $Y = 1$ and T to $Y = 0$, the following probability mass function models a Bernoulli random variable:

$$\Pr(Y = y|\theta) = \theta^y \left(1 - \theta\right)^{1-y}$$

Here θ is the probability of H, and $1 - \theta$ is the probability of T. We'll come back to this soon (when we talk about logistic regression).

Univariate Normal Random Variables

A continuous RV is governed by a probability *density* function (PDF), from which probabilities are obtained via integration.

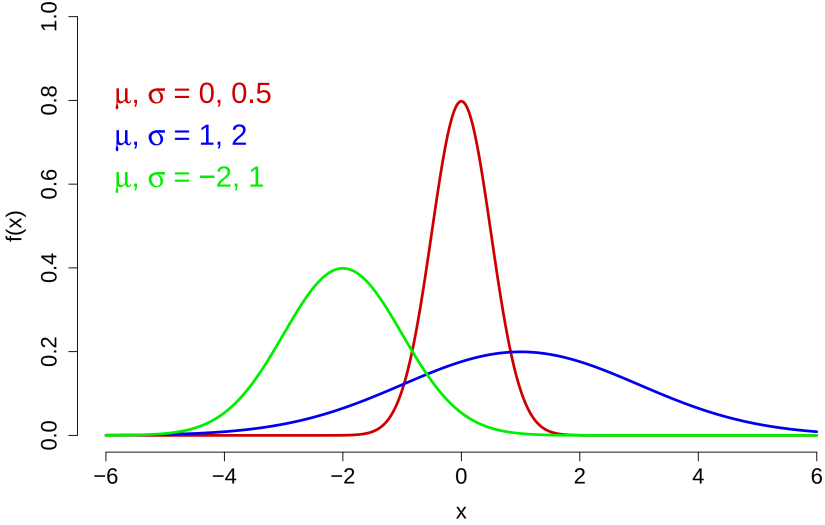
A univariate normal RV with mean μ and variance σ^2 is governed by this PDF:

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

The probability that X is in the interval $[a, b]$ is $\int_a^b f(x|\mu, \sigma)dx$.

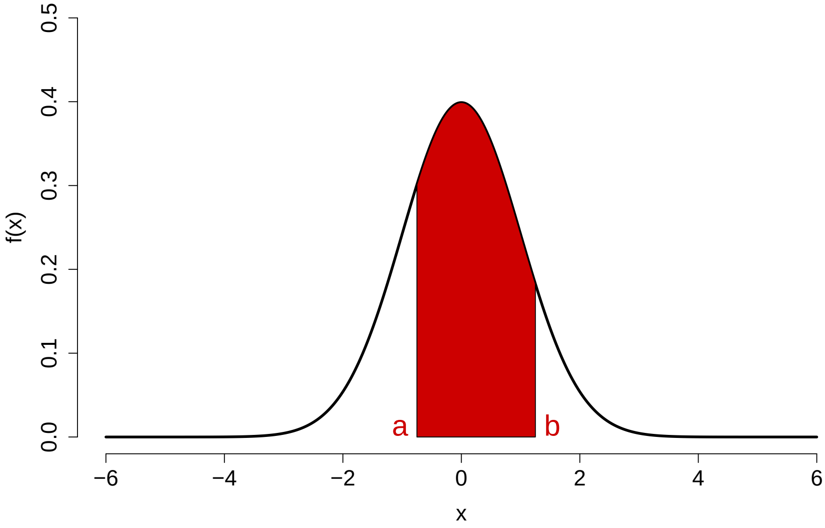
Thankfully, we can generally get estimates of this kind of integral using computers.

Univariate normal density function examples



Probability of a normal RV

$$\int_a^b f(x|\mu, \sigma)dx$$



Likelihood vs Probability

If we have a set of independent normal random variables $\mathbf{x} = \{x_1, x_2, ..., x_n\}$, and we let the parameters μ and σ vary, then we have the likelihood function:

$$
\begin{aligned}
f(\mathbf{x}|\mu,\sigma) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) \\
&= \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^n} \exp\left(-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}\right)
\end{aligned}
$$

The values of μ and σ that maximize this function are called *maximum likelihood estimates*. If $\mathbf{x}$ are normally distributed random variates, then $\hat{\mu} = \bar{x}$ (the sample mean) is the MLE for μ, and $\hat{\sigma} = s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \hat{\mu})^2}{n}}$ is the MLE for σ. **More here**.

Multivariate Normal Random Variables

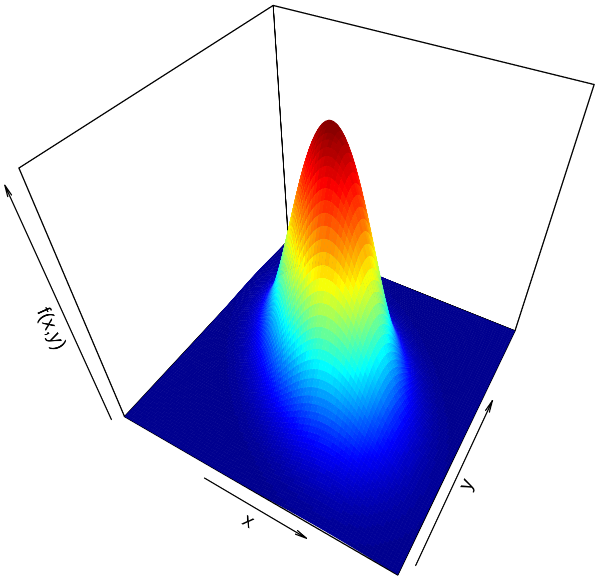
The multivariate normal density function with p dimensions is:

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

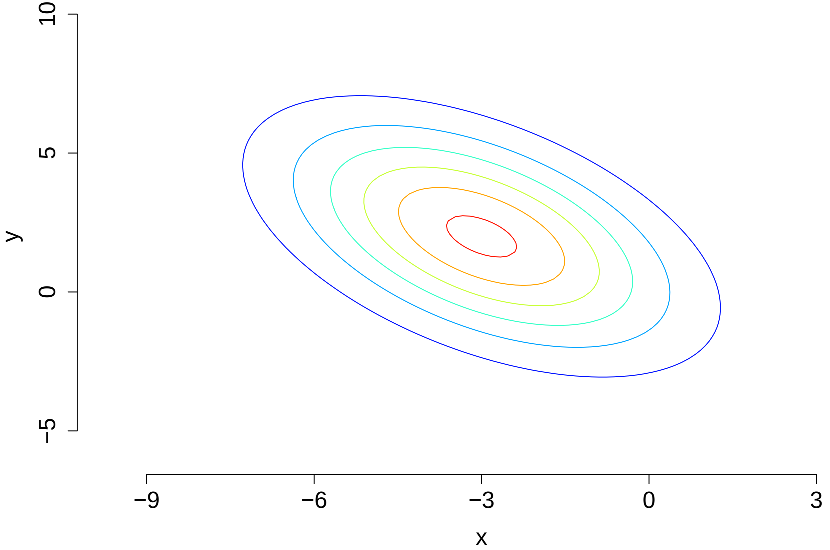
Note the square root of the determinant of Σ is analogous to $\sqrt{\sigma^2}$ in the denominator of the univariate normal PDF. Note, too, the power of p in the denominator.

Also note that the term $(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})$ is known as the Mahalanobis distance (between $\mathbf{x}$ and $\boldsymbol{\mu}$). We briefly discussed the square root of a similar formula as "statistical distance" last week.

Visualizing a 2D multivariate normal PDF



Bird's Eye View (Contour Plot)



Some properties of multivariate normal random variables

- Linear transformations of normal RVs are also normal RVs: if $\mathbf{x}$ is a p-dimensional normal RV with mean $\boldsymbol{\mu}$ and covariance Σ, and if $\mathbf{A}$ is a $q \times p$ constant matrix with rank q, then $\mathbf{Ax}$ is also p-dimensional normal with mean $\mathbf{A}\boldsymbol{\mu}$ and covariance $\mathbf{A}\Sigma\mathbf{A}^T$.

- Multivariate standard normals: if $\mathbf{z} = \left(\Sigma^{1/2}\right)^{-1}(\mathbf{x} - \boldsymbol{\mu})$, with $\mathbf{x}$ p-dimensional normal, then $\mathbf{z}$ is p-dimensional normal with mean $\mathbf{0}$ and covariance $\mathbf{I}$.

- Two components X and Y of a multivariate normal are independent if and only if $\rho_{XY} = \sigma_{XY} = 0$.

- $\mathbf{z}^T\mathbf{z} = (\mathbf{x} - \boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})$ is a χ^2 random variable with p degrees of freedom.

More properties

- Subsets of components of $\mathbf{x}$ are also multivariate normal. For example, if we split the p dimensions up into a two sets $\mathbf{x}_1$ of dimensionality $q < p$ and $\mathbf{x}_2$ of dimensionlity $r = p - q$, then $\mathbf{x}_1$ is $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ and $\mathbf{x}_2$ is $N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$, where

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$$

- Each individual component i of $\mathbf{x}$ is univariate normal with mean μ_i and variance σ_i^2.

Still more properties

If we partition a multivariate normal RV into two subsets $\mathbf{x}$ and $\mathbf{y}$:

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_y \\ \boldsymbol{\mu}_x \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yx} \\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{pmatrix}$$

then the conditional distribution of $\mathbf{y}$ given $\mathbf{x}$ is multivariate normal with:

$$\boldsymbol{\mu}_{y|x} = \boldsymbol{\mu}_y + \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}(\mathbf{x} - \boldsymbol{\mu}_x)$$

$$\boldsymbol{\Sigma}_{y|x} = \boldsymbol{\Sigma}_{yy} - \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}$$

Maximum likelihood estimation

- $\bar{\mathbf{x}}$ is the maximum likelihood estimate of $\boldsymbol{\mu}$, and $\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\mathbf{x}^T\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right)\mathbf{x}$ is the maximum likelihood estimate of $\boldsymbol{\Sigma}$. See pp. 99-100 of the textbook for more.

- If $\mathbf{x}$ is $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\bar{\mathbf{x}}$ is $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}/n)$. If $\mathbf{x}$ is non-normal, then $\bar{\mathbf{x}}$ is *approximately* $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}/n)$, as long as n is large enough.

- $\mathbf{W} = \mathbf{x}^T\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right)\mathbf{x}$ is a Wishart random variable (i.e., a multivariate generalization of χ^2) with $n-1$ degrees of freedom and scale matrix $\mathbf{V}$.

- $\bar{\mathbf{x}}$ and $\hat{\boldsymbol{\Sigma}}$ (or $\mathbf{S} = \frac{n}{n-1}\hat{\boldsymbol{\Sigma}}$) are statistically independent. (So are $\hat{\mu}$ and $\hat{\sigma}$ with univariate normals)

Simulating Multivariate Normal Data

The packages mnormt has a number of useful functions, including one for generating samples of multivariate normal RVs:

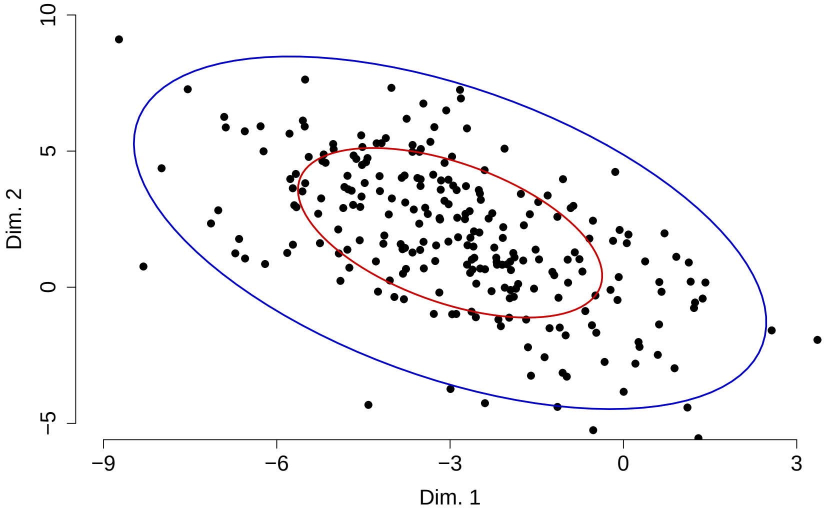
```
library(mnormt)
mu = c(-3,2)
Sigma = matrix(c(5,-3,-3,7),nrow=2)
N = 250
X = rmnorm(n=N,mean=mu,varcov=Sigma)
```

Visualing the Simulated and Expected Data

Here is code for a scatterplot and ellipses enclosing 50% and 95% of the probability mass for **x**:

```
library(ellipse)
plot(X[,1],X[,2],pch=16,xlim=c(-9,3),ylim=c(-5,10),
     cex=1.25,xlab="Dim. 1",ylab="Dim. 2",main="",
     axes=F,cex.lab=1.5)
lines(ellipse(S,centre=mu,level=0.5),lw=2,col="red3")
lines(ellipse(S,centre=mu,level=0.95),lw=2,col="blue3")
axis(1,at=seq(-9,3,by=3),cex.axis=1.5)
axis(2,at=seq(-5,10,by=5),cex.axis=1.5)
```

Visualing the Simulated and Expected Data

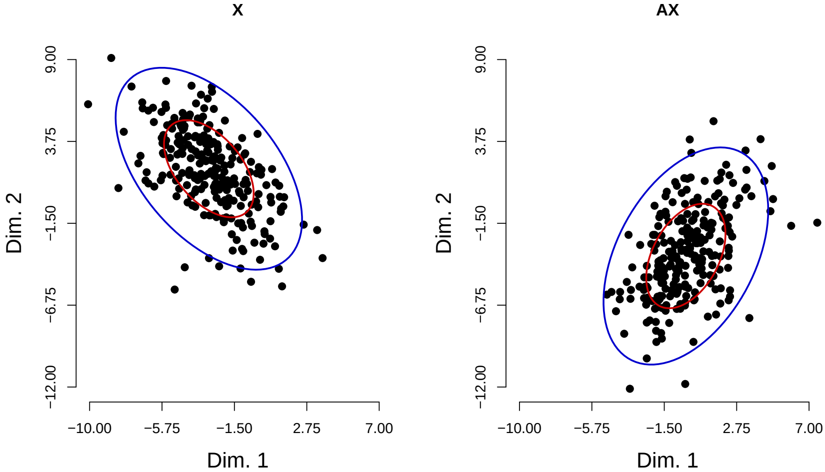


Linear Transformations in R

We can illustrate linear transformations of multivariate normal data to verify (or help build up our intuitions) about some of the facts we discussed earlier. In this case, we will rotate the data and distribution counter-clockwise $2\pi/3$ radians using a rotation matrix:

```
theta = 2*pi/3
A = matrix(c(cos(theta),sin(theta),-sin(theta),cos(theta)),
Amu = A %*% mu
ASAt = A %*% Sigma %*% t(A)
AX = X %*% t(A)
```

Visualization

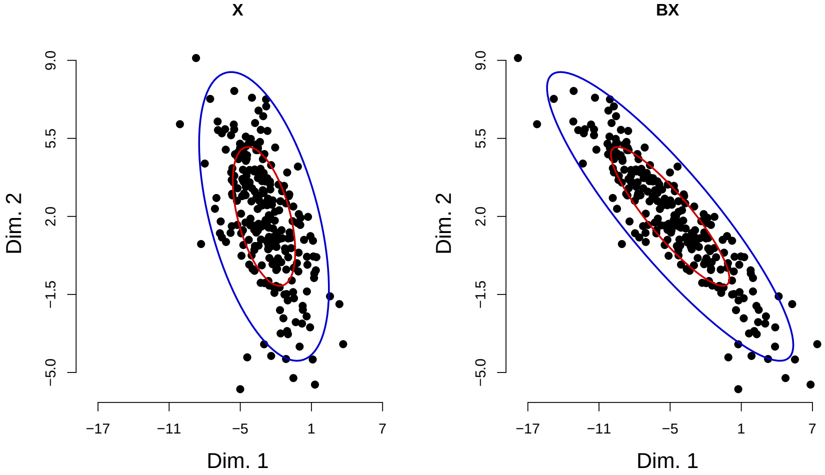


More Linear Transformations

Here is a shear transformation, wherein the values on the second dimension (y axis) are left alone and the values on the first dimension are shifted (leftward, in this case):

```
omega = pi/4
B = matrix(c(1,0,-1/tan(omega),1),nrow=2)
Bmu = B %*% mu
BSBt = B %*% Sigma %*% t(B)
BX = X %*% t(B)
```

Visualization



Standardization

Recall that we can standardize our multivariate normal data by subtracting the mean vector and (left-)multiplying by $\left(\boldsymbol{\Sigma}^{1/2}\right)^{-1}$.

```
eig = eigen(Sigma)
lam = eig$values
V = eig$vectors
Ssq = V %*% diag(sqrt(lam)) %*% t(V)
Si = solve(Ssq)
Z = t(Si %*% (t(X)-mu))
ZSZt = Si %*% Sigma %*% t(Si)
```

(This is closely related to Principal Components Analysis, which we will be talking about very soon.)

Visualization