MULTIVARIATE NORMAL DISTRIBUTION

Jerzy Szulga

Department of Mathematics and Statistics
Auburn University

MATH 5670-6670 FALL 2019

PROBABILITY I

November 15, 2019

Outline

Standard normal

A standard normal random variable Z has the density

$$\phi(z) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{z^2}{2}}, \quad -\infty < z < \infty.$$

The symbol $Z \sim N(0,1)$ indicates $\mathrm{E}\, Z = 0$ n and $\mathrm{Var}(Z) = 1$.

Its mgf is $M(t) = \mathrm{E}\, e^{tZ} = e^{t^2/2}$, the odd moments vanish, and the even moments are

$$\mathrm{E}\, Z^{2n} = \frac{(2n)!}{2^n n!}, \quad \text{in particular,} \quad \mathrm{E}\, Z^4 = 3.$$

General normal

Let μ and $\sigma > 0$ be any numbers. The transformed random variable

$$X = \sigma Z + \mu \qquad \leftrightarrow \qquad Z = \frac{X - \mu}{\sigma} \tag{1}$$

where μ is a translation parameter and σ is a scale parameter, has the density

$$\phi_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

We use then the symbol $X \sim N(\mu, \sigma^2)$.

Standard bivariate normal

The simplest way to create a 2D joint pdf is to multiply two pdfs:

$$f(x, y) \stackrel{\text{def}}{=} g(x)h(y)$$

Corresponding random variables X and Y will be independent.

For example, the multiplication of two standard normal pdf's,

$$\varphi(x, y) = \phi(x)\phi(y) = \frac{1}{2\pi} e^{-\frac{x^2+y^2}{2}},$$

yields the **standard bivariate normal** pdf.

Gauss 2D bell

The graph is a surface of revolution, a perfect bell.

Its horizontal sections are circles.

Its vertical sections are 1D Gauss bells.

It's invariant under rotation or symmetry.

Scaling and shifting

To independent (X, Y) apply

$$X \mapsto X' = aX + \mu, \qquad Y \mapsto Y' = bY + \nu$$

The new pdf is

$$\frac{1}{2\pi ab} \exp\left\{ -\frac{1}{2}\left(\frac{(x-\mu)^2}{a^2} + \frac{(y-\nu)^2}{b^2}\right)\right\} .$$

with elliptical sections centered at (μ, ν) and semiaxes $|a|$ and $|b|$.

Apparently the shift mars the formulas and concepts. So, for time being we assume no shift (it can be applied any time at need).

Rotation

Assume no scaling. Rotate by an angle θ. Then the point (x, y) rotates to the point (x', y'). Denote $c = \cos\theta$, $s = \sin\theta$. So,

$$x' = ux - sy, \quad y' = sx + cy.$$

(Clearly, $c^2 + s^2 = 1$.) Then, a surprise:

The seemingly dependent random variables

$$X' = cX - sY,\ Y' = sX + cY$$

are still independent because their pdf is still the same bell!

General bivariate normal

It is obtained from independent standard normal (Z_1, Z_2) by three transformations,

- scaling of x and scaling of y (two parameters)
- rotation (one parameter),
- shift (two parameters).

There are 5 parameters. We'd rather use:

$$\sigma_X^2,\ \sigma_Y^2, \quad \rho \text{ (or covariance } \sigma_{XY} = \rho\sigma_X\sigma_Y), \quad \text{and } \mu_X,\ \mu_Y.$$

General bivariate normal pdf

Assume no shift, $\mu_X = \mu_Y = 0$. Then (cf. Slide 27) the pdf $f(x,y) =$

$$\frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{x^2}{\sigma_X^2} - \frac{2\rho xy}{\sigma_X\sigma_Y} + \frac{y^2}{\sigma_Y^2}\right]\right\}$$

Independence implies zero covariance (or $\rho = 0$).

The inverse implication doesn't hold, in general. However

Theorem

Normal random variables X and Y are independent if and only if they are uncorrelated.

(Bernoulli random variables also have this property. It's rare.)

Vectors

A sequence $(x_1, \ldots, x_n)$ is a vector with a default vertical form.

The transpose toggles between the vertical and the horizontal.

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \qquad \mathbf{x}^{\top} = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}.$$

Matrices

The transpose of a quadratic matrix is a symmetry about the diagonal, and switches row vectors and column vectors.

$$A = \left[\begin{array}{c|c|c} a_{11} & \cdots & a_{1n} \\ \cdots & \cdots & \cdots \\ a_{n1} & \cdots & a_{nn} \end{array}\right] = \left[\begin{array}{c|c|c} \mathbf{a}_1 & \cdots & \mathbf{a}_n \end{array}\right],$$

$$A^{\intercal} = \left[\begin{array}{ccc} a_{11} & \cdots & a_{n1} \\ \hline \cdots & \cdots & \cdots \\ \hline a_{1n} & \cdots & a_{nn} \end{array}\right] = \left[\begin{array}{c} \mathbf{a}_1^{\intercal} \\ \hline \vdots \\ \hline \mathbf{a}_n^{\intercal} \end{array}\right].$$

Dot product and matrix product

The dot product

$$\mathbf{x}^\top \mathbf{y} = \sum_{i=1}^{n} x_i y_i \quad \Rightarrow \quad ||\mathbf{x}||^2 = \mathbf{x}^\top \mathbf{x}.$$

The matrix product $AB = \left[\mathbf{a}^i \mathbf{b}_j\right]$ (here $\mathbf{a}^i$ are rows of A).

Hence $\mathbf{y} = A\mathbf{x}$ means

$$\begin{aligned} y_1 &= \mathbf{a}_1^\top \mathbf{x} &&= a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\ &\dots && \dots \\ y_n &= \mathbf{a}_n^\top \mathbf{x} &&= a_{n1}x_1 + a_{n2}x_2 + + \cdots + a_{nn}x_n, \end{aligned}$$

Quadratic forms

A real symmetric matrix $C = [c_{ij}]$ yields the **quadratic form**:

$$\mathbf{x}^\top C\mathbf{x} = \sum_i \sum_j c_{ij}\, x_i x_j = \sum_i c_{ii} x_i^2 + 2 \sum_{i<j} c_{ij}\, x_i x_j, \quad \mathbf{x} \in \mathbb{R}^n.$$

The matrix C is **positive definite** if $\mathbf{x}^\top C\mathbf{x} > 0$ for every $\mathbf{x} \neq \mathbf{0}$ (**semi-positive definite** if $\mathbf{x}^\top C\mathbf{x} \geq 0$ for every $\mathbf{x}$).

If $C\mathbf{x} = \lambda\mathbf{x}$, $\mathbf{x} \neq \mathbf{0}$, then $\mathbf{x}$ is an **eigenvector** and λ is an **eigenvalue**.

C is positive definite iff all eigenvalues of C are strictly positive.

Covariance matrix

A RV $\mathbf{X}$ with components of finite variance yields the matrix of n^2 covariances

$$\Sigma = \mathrm{Cov}(\mathbf{X}) = \Big[\mathrm{Cov}(X_i, X_j)\Big] = [\sigma_{ij}]\,,$$

called the **covariance matrix**. The variances show on the diagonal. That is,

$$\Sigma = \mathrm{Cov}(\mathbf{X}) = \mathrm{E}\left(\mathbf{X} - \mathrm{E}\,\mathbf{X}\right)\left(\mathbf{X} - \mathrm{E}\,\mathbf{X}\right)^{\intercal} = \mathrm{E}\,\mathbf{X}\mathbf{X}^{\intercal} - \boldsymbol{\mu}_X \boldsymbol{\mu}_X^{\intercal}.$$

Again, the presence of shifts unnecessarily mars the transparency.

Mgf

Definition $M_X(\mathbf{v}) = \mathrm{E}\, e^{\mathbf{v}^\top \mathbf{X}} = \mathrm{E}\, \exp\left\{ \sum_i v_i X_i \right\}$.

In particular, in the case of iid components with a mgf $m(t) = \mathrm{E}\, e^{tX_i}$, we have $M_X(\mathbf{v}) = \prod_i m(v_i)$.

Then, for a linear transformation $\mathbf{Y} = A\mathbf{X}$,

$$M_Y(\mathbf{v}) = \mathrm{E}\, e^{\mathbf{v}^\top A\mathbf{X}} = \mathrm{E}\, e^{(A^\top \mathbf{v})^\top \mathbf{X}} = M_X(A^\top \mathbf{v}).$$

Properties

Theorem

Let C be an $n \times n$ matrix. TFAE:

1. *C is semi-positive definite;*
2. *C is a covariance matrix, $C = \Sigma$;*
3. *C admits a decomposition $C = AA^\intercal$.*

The latter decomposition is not unique. When A is lower triangular, then we call it **Cholesky decomposition**.

E.g., in Matlab, `A=chol(C)`.

Clearly, $|\Sigma| = |A|^2$.

Multivariate standard normal in algebraic notation

Let $\mathbf{Z}$ be a standard normal vector in $\mathbb{R}^n$. The pdf:

$$\varphi(\mathbf{z}) = \frac{1}{(2\pi)^{n/2}} \exp\left\{ -\frac{1}{2} \sum_i z_i^2 \right\} = \frac{1}{(2\pi)^{n/2}} \exp\left\{ -\frac{\mathbf{z}^\top \mathbf{z}}{2} \right\},$$

whose graph still may be called the **standard Gauss *n*-bell**.

Its mgf and covariance:

$$M(\mathbf{v}) = \mathrm{E}\, e^{\mathbf{v}^\top \mathbf{Z}} = \exp\left\{ \frac{\mathbf{v}^\top \mathbf{v}}{2} \right\}, \quad \mathrm{Cov}(\mathbf{Z}) = \mathrm{E}\, \mathbf{Z}\mathbf{Z}^\top = I.$$

An affine transformation of standard normal

A general normal RV in $\mathbb{R}^n$ is given by an affine transformation,

$$\mathbf{X} = A\mathbf{Z} + \boldsymbol{\mu}.$$

where A is an $n \times n$ matrix and $\boldsymbol{\mu}$ is a deterministic vector in $\mathbb{R}^n$.

From now on until further notice we assume that $\boldsymbol{\mu} = \mathbf{0}$, so $\mathbf{X} = A\mathbf{Z}$.
Then the covariance matrix is simpler :

$$\Sigma = \mathrm{E}\,\mathbf{X}\mathbf{X}^{\intercal} = \mathrm{E}\,A\mathbf{Z}\mathbf{Z}^{\intercal}A^{\intercal} = AA^{\intercal}.$$

Orthogonal transformation

Definition A matrix A or linear transformation $\mathbf{y} = A\mathbf{x}$ is called **orthogonal** if either of the following conditions is satisfied:

- *Columns (or rows) are orthonormal (orthogonal and unit)*;
- $A^{-1} = A^{\intercal}$ *(i.e.,* $AA^{\intercal} = A^{\intercal}A = I$*)*;
- *eigenvalues (may be complex) have modulus 1*;
- *the transformation is rigid, i.e.,* $||A\mathbf{x}|| = ||\mathbf{x}||$;

And one more (generalizing the observation in Slide 8):

- The transformation preserves the standard Gauss n-bell, i.e., $A\mathbf{Z}$ *and* $\mathbf{Z}$ *are equidistributed*;

General multivariate normal pdf

Theorem

Let A be nonsingular and $\mathbf{X} = A\mathbf{Z}$, yielding nonsingular $\Sigma = AA^{\intercal}$. Denote $C = \Sigma^{-1}$. Then

$$\varphi_X(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}\mathrm{abs}(|A|)} \exp\left\{ -\frac{1}{2}\mathbf{x}^{\intercal} C\mathbf{x} \right\} .$$

Proof. Let E be a region in $\mathbb{R}^n$ (e.g., an orthant). By definition

$$\begin{aligned} \mathrm{P}(\mathbf{X} \in E) &= \int \cdots \int_{E} \varphi_X(\mathbf{x})\, d\mathbf{x}, \\ \mathrm{P}(\mathbf{Z} \in A^{-1}E) &= \int \cdots \int_{A^{-1}E} \varphi(\mathbf{z})\, d\mathbf{z} \end{aligned}$$

Proof continued

Changing the variable $\mathbf{x} = A\mathbf{z}$ and using the Jacobian $\dfrac{\partial \mathbf{x}}{\partial \mathbf{z}} = |A|$, the former integral equals

$$\int \cdots \int_{A^{-1}E} f_X(A\mathbf{z}) \, \mathrm{abs}(|A|) \, d\mathbf{z}$$

Comparing the integrands, and returning to the variable $\mathbf{x}$, since $\mathbf{z}^\intercal \mathbf{z} = (A^{-1}\mathbf{x})^\intercal (A^{-1}\mathbf{x}) = \mathbf{x}^\intercal (A^\intercal A)^{-1} \mathbf{x} = \mathbf{x}^\intercal C \mathbf{x}$,

$$\varphi_X(\mathbf{x}) \, \mathrm{abs}(|A|) = \varphi(\mathbf{z}),$$

which is exactly our formula. ∎

Semantics - **orthant**

quadrant -

a region in $\mathbb{R}^2$, e.g., $\{\,(x_1, x_2) : x_i \le a_i,\ i = 1, 2\,\}$

octant -

a region in $\mathbb{R}^3$, e.g., $\{\,(x_1, x_2, x_3) : x_i \le a_i,\ i = 1, 2, 3\,\}$

hexant -

a region in $\mathbb{R}^4$, e.g., $\{\,(x_1, x_2, x_3, x_4) : x_i \le a_i,\ i = 1, 2, 3, 4\,\}$

triacontadiant -

a region in $\mathbb{R}^5$, e.g., $\{\,(x_1, x_2, x_3, x_4, x_5) : x_i \le a_i,\ i = 1, 2, 3, 4, 5\,\}$

hexacontatetrant ... (it'd be easy should you count in Old Greek)

Mgf

The formula for the mgf is derived directly (cf. Slide 16).

For the zero mean normal vector $\mathbf{X} = A\mathbf{Z}$ with the covariance matrix $\Sigma = AA^\intercal$,

$$M_X(\mathbf{v}) = \mathsf{E}\ \exp\left\{\mathbf{v}^\intercal\mathbf{X}\right\} \ = \ \exp\left\{\frac{1}{2}\mathbf{v}^\intercal\Sigma\mathbf{v}\right\} .$$

What if A is singular?

If $|A| = 0$, then the density formula on Slide 21 makes no sense.

Example. Let $n = 2$ and $A = [1, 0; 0, 0]$. That is, $\mathbf{X} = A\mathbf{Z}$ is the projection of a standard normal RV onto the horizontal axis, $X_1 = Z_1$, $X_2 = 0$. The implication

$$\text{area of a region } E \subset \mathbb{R}^2 \text{ is zero} \quad \Rightarrow \quad \mathrm{P}(\mathbf{X} \in E) = 0$$

fails (e.g., when E is the horizontal axis). This implication, if true for every region, would ensure the existence of the pdf (it is subject of the highly nontrivial **Radon-Nikodym Theorem**).

The distribution of $\mathbf{X}$ is neither discrete nor continuous.

Orthogonal split

Yet, in that "smaller universe", the first component of $\mathbf{X}$ is standard normal and has the reduced "flat" density $\phi(x_1)$.

The range (or column space) is the orthogonal complement of the nullspace, $R(A) = (N(A^\intercal))^\perp$ (cf. MATH 2660). That is

$$\mathbb{R}^n = R(A) \oplus N(A^\intercal),$$

where both subspaces are orthogonal to each other. Thus the transformation $\mathbf{X} = A\mathbf{Z}$ of less than full rank $m < n$ "squeezes" the standard normal n-vector to the proper subspace $R(A)$, where $\mathbf{X}$ has an m-variate density but no n-variate density exists.

Inverse of a 2×2 matrix

It's quick: $\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \dfrac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$.

Consider the covariance matrix Σ of a normal $[X \ \ Y]^\intercal$.

Then $|\Sigma| = \sigma_X^2 \sigma_Y^2 (1 - \rho^2)$ and its inverse

$$\Sigma^{-1} = \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix}^{-1} = \frac{1}{1-\rho^2} \begin{bmatrix} \frac{1}{\sigma_X^2} & -\frac{1}{\sigma_X\sigma_Y} \\ -\frac{1}{\sigma_X\sigma_Y} & \frac{1}{\sigma_Y^2} \end{bmatrix}.$$

Whence the shift-free bivariate normal density (Slide 10) follows.

Conditional density

Let a normal RV $[X \;\; Y]^\intercal$ have mean $\mathbf{0}$.

$$\phi(y|x) = \frac{\varphi(x,y)}{\varphi_X(x)} = \frac{1}{\sigma_Y\sqrt{1-\rho^2}\sqrt{2\pi}} \exp\left\{-\frac{Q}{2}\right\},$$

where

$$\begin{aligned} Q &= \frac{1}{1-\rho^2}\left(\frac{x^2}{\sigma_X^2} - \frac{2\rho xy}{\sigma_X\sigma_Y} + \frac{y^2}{\sigma_Y^2}\right) - \frac{x^2}{\sigma_X^2} \\ &= \frac{1}{1-\rho^2}\left(\frac{\rho^2 x^2}{\sigma_X^2} - \frac{2\rho xy}{\sigma_X\sigma_Y} + \frac{y^2}{\sigma_Y^2}\right) \\ &= \frac{1}{\sigma_Y^2(1-\rho^2)}\left(y - \rho\frac{\sigma_Y}{\sigma_X}x\right)^2 \end{aligned}$$

We recognize the univariate normal density with mean $\frac{\sigma_Y}{\sigma_X}x$ and variance $\sigma_Y^2(1-\rho^2)$.

Introducing now possibly nonzero shifts μ_X and μ_Y,

$$\mathbf{[Y|X = x] \sim \phi(y|x) \sim N\Big(r(x), \boldsymbol{\sigma}_Y^2(1-\boldsymbol{\rho}^2)\Big)},$$

where the mean is given by the linear regression function

$$r(x) = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X).$$

Observe that the variance does not depend on x.

Basic lemma

Let $X_1, \dots, X_n)$ be iid $\mathrm{N}(\mu, \sigma^2)$ and $u_1, \dots, u_n$ be arbitrary numbers.

Then $\sum_i u_i(X_i - \overline{X})$ *and* $\overline{X}$ *are independent.*

Indeed, w.l.o.g. we may and do assume that $\mu = 0$ and $\sigma = 1$. Let us compute the covariance (now, just the expectation of the product):

$$
\begin{aligned}
\mathrm{E}\left(\overline{X}\sum_i u_i(X_i - \overline{X})\right) &= \sum_i u_i \mathrm{E}\left(\overline{X}X_i\right) - \sum_i u_i \mathrm{E}\left(\overline{X}\right)^2 \\
&= \sum_i u_i/n - \sum_i u_i/n = 0
\end{aligned}
$$

Uncorrelated normal RVs are independent.

Sample variance

In virtue of Algebra of Expectations, motivated by the LLN, the sample average $\overline{X}$ is an **unbiased estimator** of the ideal mean (the expectation is equal to the estimated parameter, $\mathrm{E}\,\overline{X} = \mu$).

By the same token yet with a small modification we obtain an unbiased estimator of the ideal variance σ^2:

$$S^2 = \frac{1}{n-1}\sum_i (X_i - \overline{X})^2$$

The normalizing factor $1/(n-1)$ ensures unbiasedness, $\mathrm{E}\,S^2 = \sigma^2$ (in lieu of $1/n$).

The Fundamental Theorem of Statistics

Theorem

The sample variance and the sample mean from normal population are independent.

Proof. Indeed, putting $a_i = X_i - \overline{X}$,

$$\max\left\{ \left(\sum_i u_i a_i\right)^2 : \sum_i u_i^2 = 1 \right\} = \sum_i a_i^2$$

(a simple exercise in Lagrange Multipliers in Calculus 3). Then the property follows from our Basic Lemma since any transformations of independent random variables preserve independence. ∎

Exercises

- (doable) Let $p \geq 1$.
  Show that $\sum\limits_i |X_i - \overline{X}|^p$ and $\overline{X}$ are independent.
- (doable) Show that $\max\limits_i |X_i - \overline{X}|$ and $\overline{X}$ are independent.
- (advanced) Let $0 < p < 1$.
  Show that $\sum\limits_i |X_i - \overline{X}|^p$ and $\overline{X}$ are independent.
- (from #3, tricky) Let $q_i \geq 0, \sum_i q_i = 1$.
  Show that $\prod\limits_i |X_i - \overline{X}|^{q_i}$ and $\overline{X}$ are independent.

The easiest exercise

All of above, including the Basic Lemma, follow from it.

- Let $h(\mathbf{x})$ be an arbitrary real valued function.
  Put $Y = h(\mathbf{X} - \overline{X})$.

  Show that Y and $\overline{X}$ are independent.