Statistics for Data Science

Wintersemester 2023/24

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

Third lecture, October 30, 2023

Joint distributions

Often, instead of dealing with one random variable only, we are interested in several random variables $X_1, \ldots, X_n$.

Let $(\Omega, \mathbb{P})$ be a probability space and let $X_j : \Omega \to E_j$ be random variables with target spaces E_j, $j = 1, \ldots, n$. One can view the map

$$X := (X_1, \ldots, X_n) \colon \Omega \mapsto E_1 \times \cdots \times E_n, \quad \omega \mapsto (X_1(\omega), \ldots, X_n(\omega))$$

as a single, multivariate random variable.

In analogy to the univariate case, the joint probability distribution of $X_1, \ldots, X_n$ is

$$P_{X_1,\ldots,X_n}(C) = \mathbb{P}((X_1, \ldots, X_n) \in C) \quad \text{for all } C \subset E_1 \times \cdots \times E_n.$$

Informally speaking, the marginal distribution of X_i is obtained by "integrating out" (continuous RVs) / "summation over" (discrete RVs) all variables except the i^{th} one. The precise definition is

$$\begin{aligned} P_{X_i}(A) &= P_{X_1,\ldots,X_n}(E_1 \times \cdots \times E_{i-1} \times A \times E_{i+1} \times \cdots \times E_n) \\ &= \mathbb{P}(X_1 \in E_1, \ldots, X_{i-1} \in E_{i-1}, X_i \in A, X_{i+1} \in E_{i+1}, \ldots, X_n \in E_n) \end{aligned}$$

for all events $A \subset E_i$.

Joint PMF (discrete RVs)

Assume that $X_j\colon \Omega \to E_j$ are discrete random variables (recall that this means each E_j is countable). This means that $E_1 \times \cdots \times E_n$ is also countable. The joint PMF $p_{X_1,\dots,X_n}\colon E_1 \times \cdots \times E_n \to [0,1]$ is defined as

$$p_{X_1,\dots,X_n}(x_1,\dots,x_n) = \mathbb{P}(X_1 = x_1,\dots,X_n = x_n), \quad (x_1,\dots,x_n) \in E_1 \times \cdots \times E_n.$$

The probability distribution can be expressed as follows in the discrete case.

Proposition

For all events $C \subset E_1 \times \cdots \times E_n$, there holds

$$P_{X_1,\dots,X_n}(C) = \sum_{(x_1,\dots,x_n)\in C} p_{X_1,\dots,X_n}(x_1,\dots,x_n).$$

Proof. The claim is an immediate consequence of σ-additivity of disjoint events

$$\{(X_1,\dots,X_n) \in C\} = \bigcup_{(x_1,\dots,x_n)\in C} \{X_1 = x_1,\dots,X_n = x_n\}. \quad \square$$

The marginal PMF of a discrete RV X_i can be obtained from the joint PMF by summation over all the other RVs:

$$p_{X_i}(x) = \sum_{\substack{x_1 \in E_1, \ldots, \\ x_{i-1} \in E_{i-1}, \\ x_{i+1} \in E_{i+1}, \ldots \\ x_n \in E_n}} p_{X_1,\ldots,X_n}(x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_n).$$

More generally, for any subset of indices $\mathcal{I} \subset \{1, \ldots, n\}$, we can recover the joint PMF of the random variables $(X_i)_{i \in \mathcal{I}}$ from the joint PMF of $X_1, \ldots, X_n$ by summing up $p_{X_1,\ldots,X_n}$ over all possible values in the coordinates $j \notin \mathcal{I}$.

For example, if $n = 4$, we can recover the joint PMF of (X_2, X_3) via

$$p_{X_2,X_3}(x, y) = \sum_{x_1 \in E_1,\ x_4 \in E_4} p_{X_1,X_2,X_3,X_4}(x_1, x, y, x_4).$$

Example (Bivariate case $n = 2$)

If (X, Y) is a bivariate discrete RV with PMF $p_{X,Y}$, then the PMFs of X and Y are respectively given by

$$p_X(x) = \sum_{y \in E_2} p_{X,Y}(x, y) \quad \text{and} \quad p_Y(y) = \sum_{x \in E_1} p_{X,Y}(x, y).$$

Example

Let (X, Y) be a bivariate RV taking values in $\{1, 2\} \times \{1, 2, 3\}$ and with joint PMF p given as below

$p(x, y)$	$y = 1$	$y = 2$	$y = 3$
$x = 1$	0.1	0.3	0.2
$x = 2$	0.2	0.2	0

The values of the marginal PMF $p_X(x)$, $x = 1, 2$, are obtained by summing up the probabilities in each of the corresponding rows

$$p_X(1) = 0.1 + 0.3 + 0.2 = 0.6$$
$$p_X(2) = 0.2 + 0.2 + 0 = 0.4.$$

Similarly, the values of the marginal PMF $p_Y(y)$, $y = 1, 2, 3$, are obtained by summing up the probabilities in each of the corresponding columns:

$$p_Y(1) = 0.1 + 0.2 = 0.3, \quad p_Y(2) = 0.3 + 0.2 = 0.5, \quad p_Y(3) = 0.2 + 0 = 0.2.$$

Joint PDF (continuous RVs)

Definition

A function $f\colon \mathbb{R}^n \to \mathbb{R}$ is called a probability density function (PDF) if the following conditions hold:

- $f(x_1, \dots, x_n) \geq 0$ for all $(x_1, \dots, x_n) \in \mathbb{R}^n$;
- $\int_{\mathbb{R}} \cdots \int_{\mathbb{R}} f(x_1, \dots, x_n) \, \mathrm{d}x_1 \cdots \mathrm{d}x_n = 1$.

The real-valued random variables $X_1, \dots, X_n$ admit a continuous joint distribution (resp. admit a joint density) if there exists a PDF $f_{X_1,\dots,X_n}\colon \mathbb{R}^n \to \mathbb{R}$ such that, for all subsets $A \subset \mathbb{R}^n$, there holds

$$\mathbb{P}((X_1, \dots, X_n) \in A) = \int_A f_{X_1,\dots,X_n}(x_1, \dots, x_n) \, \mathrm{d}x_1 \cdots \mathrm{d}x_n.$$

Then we call $f_{X_1,\dots,X_n}$ the probability density function (PDF) of X.

Lemma

If $X_1, \ldots, X_n$ admit a joint density $f_{X_1,\ldots,X_n}$, then $X_1, \ldots, X_n$ are continuous RVs with PDF given by

$$f_{X_i}(x) = \int_{\mathbb{R}^{n-1}} f_{X_1,\ldots,X_n}(x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_n) \, \mathrm{d}x_1 \cdots \mathrm{d}x_{i-1} \mathrm{d}x_{i+1} \cdots \mathrm{d}x_n$$

for $x \in \mathbb{R}$. We call f_{X_i} the marginal PDF of X_i.

More generally, for any subset of indices $\mathcal{I} \subset \{1, \ldots, n\}$ we can recover the joint PDF of the random variables $(X_i)_{i \in \mathcal{I}}$ from the joint PDF of $X_1, \ldots, X_n$ by integrating over all possible values in the coordinates $j \notin \mathcal{I}$.

For example, if $n = 4$, we can recover the joint PDF of (X_2, X_3) via

$$f_{X_2,X_3}(x, y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X_1,X_2,X_3,X_4}(x_1, x, y, x_4) \, \mathrm{d}x_1 \, \mathrm{d}x_4.$$

Example

Let $a, b, c, d \in \mathbb{R}$ be such that $a < b$ and $c < d$. Then the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$f(z) = \frac{1}{(b-a)(d-c)} \mathbf{1}_{[a,b]\times[c,d]}(z), \quad z \in \mathbb{R}^2,$$

is a PDF. It corresponds to the uniform distribution on the rectangle $[a, b] \times [c, d]$. The marginal distributions are univariate distributions on the $[a, b]$ and $[c, d]$, respectively:

$$X \sim \mathcal{U}(a, b), \quad Y \sim \mathcal{U}(c, d).$$

Example (Bivariate Gaussian distribution)

Let $\mu \in \mathbb{R}^2$ and let $C \in \mathbb{R}^{2\times 2}$ be a symmetric, positive definite 2×2 matrix. The function $f : \mathbb{R}^2 \to \mathbb{R}$ given by

$$f(z) = \frac{1}{2\pi\sqrt{\det C}} \exp\left(-\frac{1}{2}(z-\mu)^{\mathrm{T}} C^{-1}(z-\mu)\right), \quad z \in \mathbb{R}^2,$$

is a PDF. A random vector $Z = (X, Y)$ with PDF p is said to have Gaussian distribution with mean μ and covariance matrix C. Denoting

$$\mu = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \quad C = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix},$$

then the marginal PDFs are given by

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma_X^2}} \exp\left(-\frac{(x-\mu_X)^2}{2\sigma_X^2}\right),$$

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma_Y^2}} \exp\left(-\frac{(y-\mu_Y)^2}{2\sigma_Y^2}\right).$$

Thus $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$.

In the special case $\mu = 0$ and $C = I_2$, i.e., $\mu_X = \mu_Y = 0$, $\sigma_{XY} = 0$, and $\sigma_X^2 = \sigma_Y^2 = 1$:

$$f(z) = \frac{1}{2\pi} \exp\left(-\frac{1}{2}\|z\|^2\right), \quad z \in \mathbb{R}^2,$$

where $\|z\| = \sqrt{x^2 + y^2}$ denotes the Euclidean norm of $z = (x, y)$.

Independence of random variables

Definition

The random variables $X_1, \dots, X_n$ are said to be independent if, for any subsets $A_1 \subset E_1, \dots, A_n \subset E_n$, there holds

$$\mathbb{P}(X_1 \in A_1, \dots, X_n \in A_n) = \mathbb{P}(X_1 \in A_1) \cdots \mathbb{P}(X_n \in A_n).$$

Theorem (Independence of discrete RVs)

Assume that $X_1, \dots, X_n$ are discrete random variables with joint PMF $p_{X_1,\dots,X_n}$ and marginal PMFs $p_{X_1}, \dots, p_{X_n}$. Then $X_1, \dots, X_n$ are independent if and only if

$$p_{X_1,\dots,X_n}(x_1, \dots, x_n) = p_{X_1}(x_1) \cdots p_{X_n}(x_n), \quad (x_1, \dots, x_n) \in E_1 \times \cdots \times E_n.$$

Theorem (Independence of continuous RVs)

Assume that $X_1, \dots, X_n$ are continuous random variables with joint PDF $f_{X_1,\dots,X_n}$ and marginal PDFs $f_{X_1}, \dots, f_{X_n}$. Then $X_1, \dots, X_n$ are independent if and only if

$$f_{X_1,\dots,X_n}(x_1, \dots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n), \quad (x_1, \dots, x_n) \in \mathbb{R}^n.$$

Example, independence

Let X and Y have the joint PDF

$$f(x,y) = \begin{cases} x+y & \text{if } 0 \le x \le 1,\ 0 \le y \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Are the variables X and Y independent?

Now

$$f(x) = \int_0^1 (x+y)\,\mathrm{d}y = x + \frac{1}{2},\ 0 < x < 1$$

and

$$f(y) = \int_0^1 (x+y)\,\mathrm{d}x = y + \frac{1}{2},\ 0 < y < 1.$$

If the random variables are independent, then $f(x,y) = f(x) \cdot f(y)$. Let $x = 1/3$ and $y = 1/3$. Now

$$\begin{aligned} f(x,y) &= x + y = \tfrac{1}{3} + \tfrac{1}{3} = \tfrac{2}{3}, \\ f(x) \cdot f(y) &= (x + \tfrac{1}{2})(y + \tfrac{1}{2}) = \tfrac{5}{6} \cdot \tfrac{5}{6} = \tfrac{25}{36} \neq \tfrac{2}{3}. \end{aligned}$$

Thus X and Y are not independent.

Example, independence

Let X and Y have the joint PMF

$$p(x, y) = \begin{cases} \frac{1}{4} & \text{if } x \in \{1, 2\},\ y \in \{1, 2\}, \\ 0 & \text{otherwise.} \end{cases}$$

Now

$$p(x) = \sum_{y \in \{1,2\}} p(x, y) = \tfrac{1}{4} + \tfrac{1}{4} = \tfrac{1}{2},\ x \in \{1, 2\},$$

and otherwise $p(x) = 0$,

and

$$p(y) = \sum_{x \in \{1,2\}} p(x, y) = \tfrac{1}{4} + \tfrac{1}{4} = \tfrac{1}{2},\ y \in \{1, 2\},$$

and otherwise $p(y) = 0$.

Therefore $p(x, y) = p(x)p(y)$ for all x and y, meaning that X and Y are independent.

Conditional distribution

Definition

Let (X, Y) be a discrete random variable in $E_1 \times E_2$ with joint PMF $p_{X,Y}$ and marginal PMFs p_X and p_Y. The conditional PMF $p_{X|Y}$ of X given Y is defined by

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x, y)}{p_Y(y)},$$

for all $x \in E_1$ and $y \in E_2$ such that $p_Y(y) > 0$.

Definition

Let (X, Y) be a continuous random variable in $\mathbb{R}^n \times \mathbb{R}^k$ with joint PDF $f_{X,Y}$ and marginal PMFs f_X and f_Y. The conditional PDF $f_{X|Y}$ of X given Y is defined by

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)},$$

for all $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^k$ such that $f_Y(y) > 0$.

Transformations of random variables

When we perform arithmetic with random variables, it is natural to ask

- if X and Y are random variables, what is the distribution of $Z = X + Y$?
- if X is an $\mathbb{R}^k$-valued random variable with known distribution and $g\colon \mathbb{R}^k \to \mathbb{R}^k$ is a function, what is the distribution of the transformed random variable $Y = g(X)$?

Theorem

Let X be a continuous real-valued random variable with CDF F_X and quantile function F_X^{-1}.

1. *The random variable $U = F_X(X) \sim \mathcal{U}(0,1)$.*
2. *If $U \sim \mathcal{U}(0,1)$, then $F_X^{-1}(U)$ has the same distribution as X (they are equal in law).*

Proof. (1) Note that $\mathbb{P}(F_X(X) \leq t) = \mathbb{P}(X \leq F_X^{-1}(t))$.† We observe that for all $t \in (0,1)$,

$$\mathbb{P}(U \leq t) = \mathbb{P}(F_X(X) \leq t) = \mathbb{P}(X \leq F_X^{-1}(t)) = F_X(F_X^{-1}(t)) = t.$$

Therefore $\mathbb{P}(U \leq t) = t$, meaning that $U \sim \mathcal{U}(0,1)$.

(2) $\mathbb{P}(F_X^{-1}(U) \leq t) = \mathbb{P}(U \leq F_X(t)) = F_X(t)$. $\square$

---

†If $F_X(X) < t$, then $X < F_X^{-1}(t)$, which implies (since X is a continuous RV) that $\mathbb{P}(F_X(X) \leq t) = \mathbb{P}(F_X(X) < t) \leq \mathbb{P}(X < F_X^{-1}(t)) = \mathbb{P}(X \leq F_X^{-1}(t))$.
On the other hand, $X \leq F_X^{-1}(t)$ implies $F_X(X) \leq F_X(F_X^{-1}(t)) = t$, so $\mathbb{P}(X \leq F_X^{-1}(t)) \leq \mathbb{P}(F_X(X) \leq t)$. Therefore $\mathbb{P}(F_X(X) \leq t) = \mathbb{P}(X \leq F_X^{-1}(t))$.

The previous theorem is very useful for simulations: if we have a uniform random number generator, we can generate samples from any distribution provided that we have access to its quantile function.

Algorithm (Inverse transform sampling)

1. *Draw* $U \sim \mathcal{U}(0,1)$.
2. *Calculate* $X = F_X^{-1}(U)$.

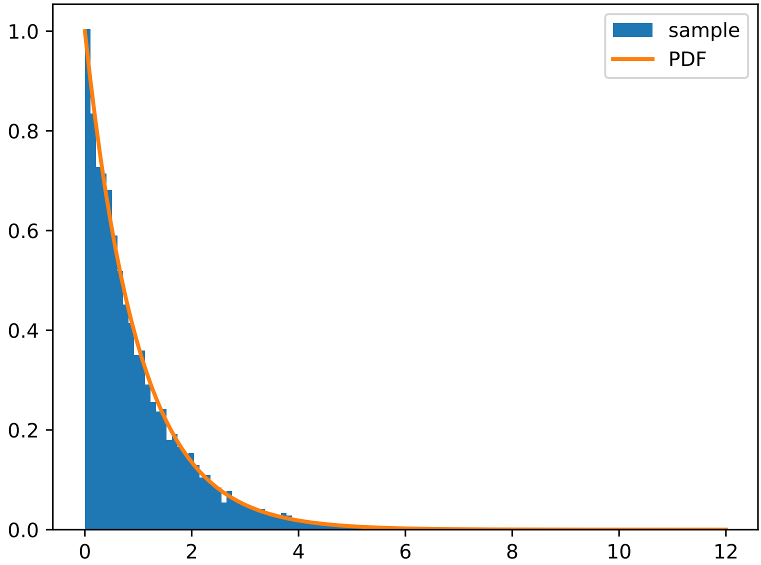
If a closed form expression for the inverse CDF is not available, then a computationally attractive formula for approximating the value $F_X^{-1}(U)$ is given by the generalized inverse:

$$F_X^{-1}(q) = \inf\{x \in \mathbb{R} \mid F_X(x) \geq q\}.$$

Example (Exponential distribution)

Let $X \sim \mathrm{Exp}(\lambda)$, $\lambda > 0$, with the PDF $f_X(x) = \lambda \mathrm{e}^{-\lambda x} \mathbf{1}_{[0,\infty)}(x)$. In this case, $F_X(a) = \mathbf{1}_{[0,\infty)}(a)(1 - \mathrm{e}^{-\lambda a})$ and $F_X^{-1}(q) = -\frac{1}{\lambda} \log(1-q)$, $q \in (0,1)$. We implement inverse transform sampling to draw a sample $X \sim \mathrm{Exp}(1)$.

```
import numpy as np
import matplotlib.pyplot as plt
n = int(1e5) # sample size
x = np.linspace(0,12,1000)
lam = 1 # lambda parameter of Exp distribution
p = lambda x: lam * np.exp(-lam*x) # PDF
invF = lambda q: -1/lam * np.log(1-q) # quantile function
u = np.random.uniform(size=n) # i.i.d. sample from U(0,1)
sample = invF(u) # inverse transform
plt.hist(sample,bins='auto',
         density=True,label='sample') # draw histogram
plt.plot(x,p(x),linewidth=2,label='PDF') # plot the PDF
plt.legend()
plt.show()
```

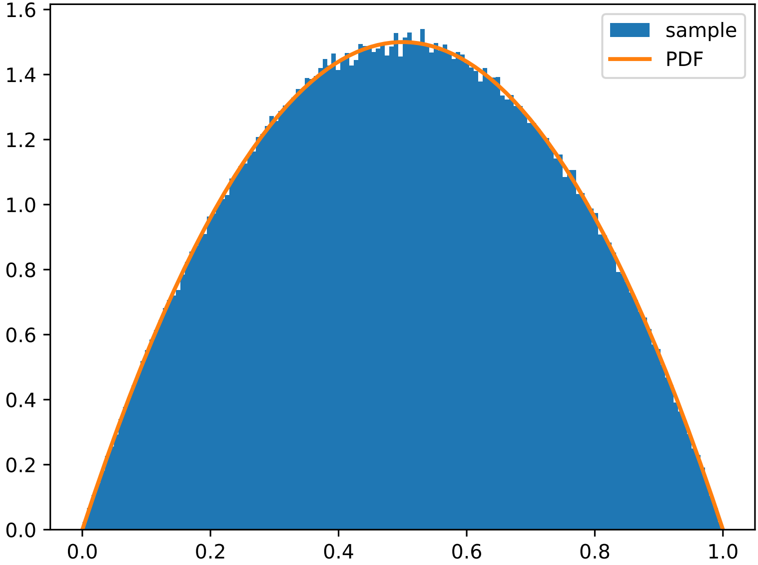


sample
PDF
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10
12

Example

Let the random variable X have the PDF $f_X(x) = (6x - 6x^2)\mathbf{1}_{[0,1]}(x)$. In this case, the quantile function is difficult to write down, but we can still implement inverse transform sampling numerically.

```
import numpy as np
import matplotlib.pyplot as plt
n = int(1e6) # sample size
x = np.linspace(0,1,10000)
p = lambda x: 6*x-6*x**2 # PDF
P = np.cumsum(p(x)); P = P/P[-1] # "empirical" CDF of p
sample = []
for _ in range(n):
    u = np.random.uniform() # realization of U(0,1)
    ind = np.where(u<=P)[0][0] # inverse transform
    sample.append(x[ind]) # store sample
plt.hist(sample,bins='auto',
        density=True,label='sample') # draw histogram
plt.plot(x,p(x),linewidth=2,label='PDF') # plot the PDF
plt.legend(); plt.show()
```


sample
PDF
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
0.0
0.2
0.4
0.6
0.8
1.0

Change of variables formula (discrete RVs)

Proposition

Let $X : \Omega \to E$ *and* $Y : \Omega \to F$ *be discrete random variables such that* $Y = g(X)$, *where* $g : E \to F$. *Then the PMF of* Y *is given by*

$$p_Y(y) = \sum_{x \in g^{-1}(\{y\})} p_X(x) = \sum_{\substack{x \in E \\ g(x) = y}} p_X(x).$$

In other words, the PMF of Y *at point* y *is obtained by summing up the PMF of* X *over the preimage* $g^{-1}(\{y\})$.

Proof. Recall that $g^{-1}(\{y\}) = \{x \in E \mid g(x) = y\}$. Thus

$$p_Y(y) = \mathbb{P}(Y = y) = \mathbb{P}(g(X) = y) = \mathbb{P}(X = g^{-1}(\{y\}))$$

$$= \mathbb{P}\Big(\bigcup_{x \in g^{-1}(\{y\})} \{X = x\} \Big) = \sum_{x \in g^{-1}(\{y\})} \mathbb{P}(X = x) = \sum_{x \in g^{-1}(\{y\})} p_X(x),$$

where we used the σ-additivity of the disjoint sets $(\{X = x\})_{x \in g^{-1}(y)}$. $\square$

Change of variables formula (continuous, univariate case)

Let X and Y be real-valued random variables such that $Y = g(X)$, where $g \colon \mathbb{R} \to \mathbb{R}$. By noting that the CDF of Y satisfies

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \leq y),$$

one can use the following method to obtain the PDF of Y given the PDF of X:

- Compute the CDF of Y using

$$F_Y(y) = \mathbb{P}(g(X) \leq y) \quad \text{for } y \in \mathbb{R}.$$

- If F_Y is differentiable, then Y has the PDF $f_Y = F_Y'$.

Example

Let $X \sim \mathcal{U}(0,1)$, $g(x) = x^2$, and define $Y = g(X)$. We wish to find $f_Y(y)$. We begin by noting that

$$F_Y(y) = \mathbb{P}(g(X) \leq y) = \mathbb{P}(X^2 \leq y) = \begin{cases} \mathbb{P}(\varnothing) & \text{if } y < 0, \\ \mathbb{P}(-\sqrt{y} \leq X \leq \sqrt{y}) & \text{if } y \geq 0. \end{cases}$$

Here, $\mathbb{P}(\varnothing) = 0$ and

$$\mathbb{P}(-\sqrt{y} \leq X \leq \sqrt{y}) = \int_{-\sqrt{y}}^{\sqrt{y}} \mathbf{1}_{[0,1]}(x)\,\mathrm{d}x = \begin{cases} \sqrt{y} & \text{if } y \in [0,1], \\ 1 & \text{if } y > 1. \end{cases}$$

Hence

$$F_Y(y) = \begin{cases} 0 & \text{if } y < 0 \\ \sqrt{y} & \text{if } y \in [0,1], \\ 1 & \text{if } y > 1 \end{cases} \quad \overset{\frac{\mathrm{d}}{\mathrm{d}y}}{\Rightarrow} \quad f_Y(y) = \frac{\mathbf{1}_{[0,1]}(y)}{2\sqrt{y}}, \; y \in \mathbb{R}.$$

In the special case where $g\colon \mathbb{R} \to \mathbb{R}$ is a strictly monotonic, continuously differentiable function, one has the following formula.

Theorem

Let $g\colon \mathbb{R} \to \mathbb{R}$ be a continuously differentiable and strictly monotonic function. Let X and Y be continuous, real-valued random variables satisfying $Y = g(X)$. Then we have the following:

$$f_X(x) = f_Y(g(x))|g'(x)|, \quad x \in \mathbb{R},$$

and

$$f_Y(y) = f_X(g^{-1}(y))|(g^{-1})'(y)| = f_X(g^{-1}(y))\frac{1}{|g'(g^{-1}(y))|}, \quad y \in \mathbb{R}.$$

Proof. For each (measurable) subset $B \subset \mathbb{R}$, there holds

$$\mathbb{P}(X \in B) = \mathbb{P}(Y \in g(B)) = \int_{g(B)} f_Y(y)\,\mathrm{d}y = \int_B f_Y(g(x))|g'(x)|\,\mathrm{d}x.$$

Since B is arbitrary, we conclude that $f_X(x) = f_Y(g(x))|g'(x)|$.

The second claim follows from the first one by writing $X = g^{-1}(Y)$. $\square$

Change of variables formula (continuous, multivariate case)

The change of variables formulae can be generalized to higher dimensions. For example, let $X_1, \ldots, X_k$ be real-valued random variables and let $g\colon \mathbb{R}^k \to \mathbb{R}$. We wish to derive the PDF of the real-valued random variable $Z = g(X_1, \ldots, X_k)$.

One can proceed as follows:

1. Compute the CDF F_Z of Z by

$$F_Z(z) = \mathbb{P}(g(X_1, \ldots, X_k) \leq z).$$

2. If F_Z is differentiable, then its PDF is given by $f_Z = F_Z'$.

Example

Let $X, Y \sim \mathcal{U}(0,1)$ be independent random variables and define $Z = \max(X, Y)$. Now[†]

$$F_Z(z) = \mathbb{P}(\max(X, Y) \leq z) = \mathbb{P}(X \leq z, Y \leq z).$$

Since X and Y were assumed to be independent, and both X and Y are uniformly distributed in $[0,1]$, we get

$$F_Z(z) = \mathbb{P}(X \leq z)\mathbb{P}(Y \leq z) = \left(\int_{-\infty}^{z} \mathbf{1}_{[0,1]}(t)\,\mathrm{d}t\right)^2 = \begin{cases} 0 & \text{if } z < 0, \\ z^2 & \text{if } z \in [0,1], \\ 1 & \text{if } z > 1. \end{cases}$$

Differentiating the above yields

$$f_Z(z) = 2z\,\mathbf{1}_{[0,1]}(z), \quad z \in \mathbb{R}.$$

[†] Note that $\max(X, Y) \leq z \Leftrightarrow X \leq z$ and $Y \leq z$. Recall also the notation $\mathbb{P}(X \leq z, Y \leq z) = \mathbb{P}(X \leq z \text{ and } Y \leq z)$.

The following change of variable formula works in the case where X, Y are $\mathbb{R}^n$-valued random variables and $g\colon \mathbb{R}^n \to \mathbb{R}^n$ is C^1-diffeomorphism (i.e., g is a bijection and both g and its inverse g^{-1} are continuously differentiable). The Jacobian matrix of a vector field $F(x) = [F_1(x), \ldots, F_n(x)]^{\mathrm{T}}$, where $F_j\colon \mathbb{R}^n \to \mathbb{R}$ for $j = 1, \ldots, n$, is

$$DF(x) = \begin{bmatrix} \frac{\partial}{\partial x_1} F_1(x) & \cdots & \frac{\partial}{\partial x_n} F_1(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial x_1} F_n(x) & \cdots & \frac{\partial}{\partial x_n} F_n(x) \end{bmatrix}.$$

Theorem

Let $g\colon \mathbb{R}^n \to \mathbb{R}^n$ be a C^1-diffeomorphism and let X and Y be $\mathbb{R}^n$-valued random variables such that $Y = g(X)$. Then

$$f_X(x) = f_Y(g(x))|\det Dg(x)|, \quad x \in \mathbb{R}^n,$$

and

$$f_Y(y) = f_X(g^{-1}(y))|\det Dg^{-1}(y)|, \quad y \in \mathbb{R}^n.$$

Proof. The argument is exactly the same as the univariate version (use the multivariate change of variables formula for integration). □

Example

Assume that g is an affine transformation

$$g(x) = Ax + b, \quad x \in \mathbb{R}^n,$$

for some fixed vector $b \in \mathbb{R}^n$ and invertible matrix $A \in \mathbb{R}^{n \times n}$. Suppose that X has the PDF f_X and $Y = g(X)$. We wish to find the PDF f_Y of Y.

The Jacobian matrix of g is given by

$$Dg(x) = A, \quad x \in \mathbb{R}^n,$$

and we have

$$g^{-1}(y) = A^{-1}(y - b).$$

Therefore the change of variables formula yields

$$f_Y(y) = f_X(A^{-1}(y-b))|\det A^{-1}| = f_X(A^{-1}(y-b))\frac{1}{|\det A|}, \quad y \in \mathbb{R}^n.$$

Sums of independent random variables

Theorem

Let X and Y be independent, real-valued discrete random variables with PMFs p_X and p_Y, respectively. Then the random variable $Z = X + Y$ has the PMF

$$p_Z(z) = \sum_{x \in E} p_X(x) p_Y(z - x).$$

Example

Let $X \sim \mathrm{Poisson}(\lambda)$ and $Y \sim \mathrm{Poisson}(\mu)$ be two independent Poisson random variables with parameters $\lambda, \mu > 0$. Then $X + Y \sim \mathrm{Poisson}(\lambda + \mu)$.

Theorem

Let X and Y be independent, real-valued continuous random variables with PDFs f_X and f_Y, respectively. Then the random variable $Z = X + Y$ has the PDF

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z - x) \,\mathrm{d}x, \quad z \in \mathbb{R}.$$

This is the convolution of f_X and f_Y and denoted $f_Z(z) = (f_X * f_Y)(z)$.

Positive definite matrices

Definition

Let $A \in \mathbb{R}^{d\times d}$ be a *symmetric matrix*. We call A a positive definite matrix if

$$x^{\mathrm{T}} A x > 0 \quad \text{for all } x \in \mathbb{R}^d \setminus \{0\}.$$

This implies that A is invertible and that A^{-1} is positive definite if A is.

Characterization

Let $A \in \mathbb{R}^{d\times d}$ be a symmetric matrix. Then the following are equivalent:

- *The matrix A is positive definite.*
- *The eigenvalues of A are positive.*
- *The matrix A has a Cholesky decomposition: there exists an upper triangular matrix $R \in \mathbb{R}^{d\times d}$ such that*
$$A = R^{\mathrm{T}} R.$$
- *The matrix A has a matrix square root, denoted by $A^{1/2}$, which satisfies*
$$A = A^{1/2} A^{1/2}.$$
Note that the matrix square root $A^{1/2}$ is also positive definite.

Multivariate Gaussian random variables

Definition

Let $\mu \in \mathbb{R}^d$ and let $C \in \mathbb{R}^{d\times d}$ be a positive definite matrix. We call a random variable X on $\mathbb{R}^d$ a multivariate Gaussian random variable with mean μ and covariance C if it has the PDF

$$f_X(x) = \left(\frac{1}{(2\pi)^d \det C}\right)^{1/2} \exp\left(-\frac{1}{2}(x-\mu)^{\mathrm{T}} C^{-1}(x-\mu)\right), \quad x \in \mathbb{R}^d.$$

In this case, we denote $X \sim \mathcal{N}(\mu, C)$.

Remark. There exists a concept of Gaussian random variable even in the case where the matrix C is positive semi-definite, i.e., at least one of its eigenvalues is 0, but such a random variable does not have a well-defined PDF (it is a "degenerate" random variable). The definition uses the so-called characteristic function. We omit the details.

The inverse of the covariance matrix is sometimes called a *precision matrix*. An often used notation is $\|x\|_C = \sqrt{x^{\mathrm{T}} C^{-1} x}$ for $x \in \mathbb{R}^d$.

Transformations of Gaussian random variables

Gaussian random variables behave predictably under affine transformations:

- Multiplying a Gaussian RV yields another Gaussian RV with an updated variance, but the same mean.
- Translating a Gaussian RV yields another Gaussian RV with an updated mean, but the same variance.
- An affine transformation of a Gaussian RV yields another Gaussian RVs with an updated mean and variance.
- Nonlinear transformations of Gaussian RVs are typically no longer Gaussian RVs!
  - For example, the Euclidean norm $Y = \|X\|$ of a Gaussian RV is not Gaussian (it follows a so-called "folded normal distribution").
  - The sum of squares of independent Gaussian RVs $Z = X_1^2 + \cdots + X_k^2$, where X_i are assumed to be independent Gaussian RVs, has the $\chi^2(k)$ distribution.

Proposition (ZCA transform, univariate version)

Let $\mu \in \mathbb{R}$ and $\sigma > 0$. The univariate Gaussian distribution satisfies the following properties:

1. *If $X \sim \mathcal{N}(0,1)$, then $Y := \mu + \sigma X \sim \mathcal{N}(\mu, \sigma^2)$.*
2. *If $Y \sim \mathcal{N}(\mu, \sigma^2)$, then $X := \frac{1}{\sigma}(Y - \mu) \sim \mathcal{N}(0,1)$.*

Proposition (ZCA transform, multivariate version)

Let $\mu \in \mathbb{R}^d$ and let $C \in \mathbb{R}^{d \times d}$ be a symmetric positive definite covariance matrix. The multivariate Gaussian distribution satisfies the following properties:

1. *If $X \sim \mathcal{N}(0, I_d)$, then $Y := \mu + C^{1/2} X \sim \mathcal{N}(\mu, C)$.*
2. *If $Y \sim \mathcal{N}(\mu, C)$, then $X := C^{-1/2}(Y - \mu) \sim \mathcal{N}(0, I_d)$.*

(Here, $C^{-1/2} := (C^{1/2})^{-1}$ is the inverse of the matrix square root of C.)

Remark. (1) is called a Mahalanobis or ZCA[†] coloring transform: it turns a *standard* Gaussian RV into a Gaussian RV with specified mean and covariance. (2) is called a Mahalanobis or ZCA[†] whitening transform: it turns a Gaussian RV with a specified mean and covariance into a *standard* Gaussian RV.

[†]Zero-phase component analysis

Proof. Let us prove claim (1) of the multivariate version. Let $X \sim \mathcal{N}(0, I_d)$ and define $Y = \mu + C^{1/2}x$. By defining $g(x) = \mu + C^{1/2}x$, we can write

$$Y = g(X) \quad \Rightarrow \quad f_Y(y) = f_X(g^{-1}(y))|\det Dg^{-1}(y)|.$$

In this case, we have

$$g^{-1}(y) = C^{-1/2}(y - \mu) \quad \text{and} \quad |\det Dg^{-1}(y)| = |\det C^{-1/2}| = \frac{1}{\sqrt{\det C}}.$$

Therefore

$$\begin{aligned} f_Y(y) &= \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}\|C^{-1/2}(y - \mu)\|^2 \right) \frac{1}{\sqrt{\det C}} \\ &= \left(\frac{1}{(2\pi)^d \det C} \right)^{1/2} \exp\left(-\frac{1}{2}(x - \mu)^{\mathrm{T}} C^{-1}(x - \mu) \right), \end{aligned}$$

which implies that $Y \sim \mathcal{N}(\mu, C)$.

The proof for (2) follows by writing $X = g^{-1}(Y)$ and using the change of variables formula $f_X(x) = f_Y(g(x))|\det Dg(x)|$. □

Different coloring transforms

Let $\mu \in \mathbb{R}^d$, let $C \in \mathbb{R}^{d \times d}$ be a symmetric positive covariance matrix, and let $X \sim \mathcal{N}(0, I_d)$.

- The Mahalanobis or ZCA coloring transform uses the matrix square root factorization $C = C^{1/2}C^{1/2}$ to write a standard Gaussian RV as

$$Y = \mu + C^{1/2}X \sim \mathcal{N}(\mu, C).$$

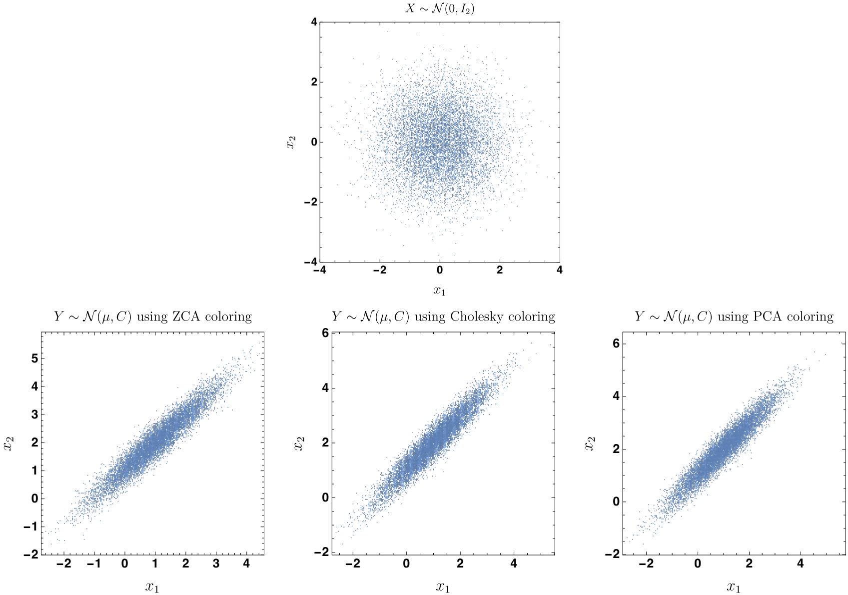
- One could alternatively use the Cholesky decomposition $C = R^{\mathrm{T}}R$ to obtain the Cholesky coloring transform

$$Y = \mu + R^{\mathrm{T}}X \sim \mathcal{N}(\mu, C).$$

- Finally, one could use the eigendecomposition $C = U\Lambda U^{\mathrm{T}} = (U\Lambda^{1/2})(U\Lambda^{1/2})^{\mathrm{T}}$, where $UU^{\mathrm{T}} = I = U^{\mathrm{T}}U$ and Λ is a diagonal matrix containing the eigenvalues of C, to obtain the PCA† coloring transform

$$Y = \mu + U\Lambda^{1/2}X \sim \mathcal{N}(\mu, C).$$

†Principal component analysis



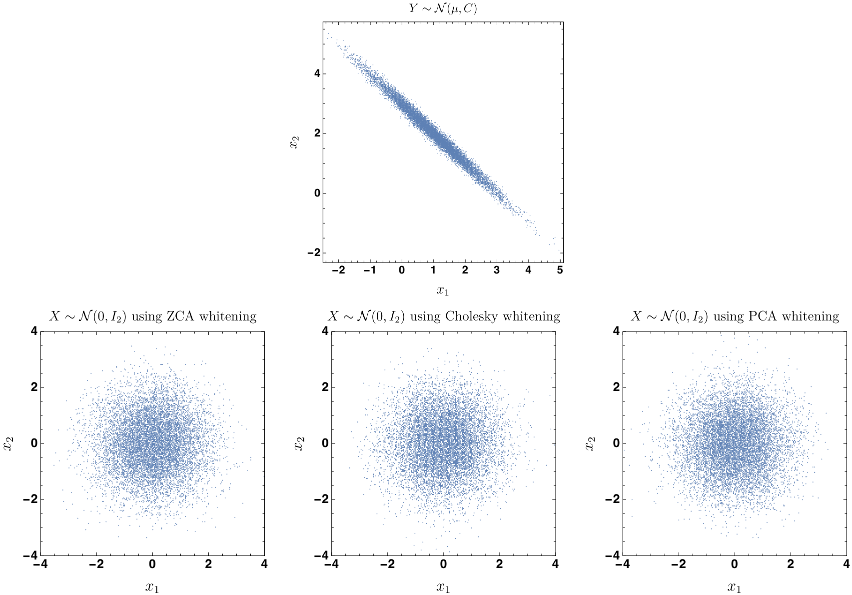
Coloring transforms with $\mu = [1, 2]^{\mathrm{T}}$ and $C = \begin{bmatrix} 1 & 0.95 \\ 0.95 & 1 \end{bmatrix}$.

Different whitening transforms

Let $\mu \in \mathbb{R}^d$, let $C \in \mathbb{R}^{d\times d}$ be a symmetric positive covariance matrix, and let $Y \sim \mathcal{N}(\mu, C)$.

- The Mahalanobis or ZCA whitening transform uses the matrix square root factorization $C = C^{1/2}C^{1/2}$ to write a standard Gaussian RV as
$$X = C^{-1/2}(Y - \mu) \sim \mathcal{N}(0, I_d).$$
- One could alternatively use the Cholesky decomposition $C = R^{\mathrm{T}}R$ to obtain the Cholesky whitening transform
$$X = R^{-\mathrm{T}}(Y - \mu) \sim \mathcal{N}(0, I_d).$$
- Finally, one could use the eigendecomposition $C = U\Lambda U^{\mathrm{T}} = (U\Lambda^{1/2})(U\Lambda^{1/2})^{\mathrm{T}}$, where $UU^{\mathrm{T}} = I = U^{\mathrm{T}}U$ and Λ is a diagonal matrix containing the eigenvalues of C, to obtain the PCA whitening transform
$$X = \Lambda^{-1/2}U^{\mathrm{T}}(Y - \mu) \sim \mathcal{N}(0, I_d).$$

Note: the three RVs obtained using different whitening transforms are rotations of one another.



Whitening transforms with $\mu = [1, 2]^{\mathrm{T}}$ and $C = \begin{bmatrix} 1 & -0.99 \\ -0.99 & 1 \end{bmatrix}$.

By inductive reasoning, one can deduce that any finite linear combination of Gaussian RVs is a Gaussian RV.

Proposition (Univariate version)

Let $X_j \sim \mathcal{N}(\mu_i, \sigma_i^2)$ *be independent Gaussian random variables with* $\mu_i \in \mathbb{R}$ *and* $\sigma_i > 0$ *for* $i = 1, \dots, n$*. Then*

$$X := \sum_{i=1}^{n} X_i \sim \mathcal{N}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right).$$

Proposition (Multivariate version)

Let $X_j \sim \mathcal{N}(\mu_i, C_i)$ *be independent Gaussian random variables with* $\mu_i \in \mathbb{R}^d$ *and symmetric, positive definite* $C_i \in \mathbb{R}^{d \times d}$ *for* $i = 1, \dots, n$*. Then*

$$X := \sum_{i=1}^{n} X_i \sim \mathcal{N}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} C_i\right).$$

Proposition

Let $\mu \in \mathbb{R}^d$ *and let* $C \in \mathbb{R}^{d\times d}$ *be a symmetric, positive definite matrix. Let* $X \sim \mathcal{N}(\mu, C)$*. If* $k \le d$ *and* $L \in \mathbb{R}^{k\times d}$ *is a matrix with full rank, then*

$$Y = LX \sim \mathcal{N}(L\mu, LCL^{\mathrm{T}}).$$

Proof. Omitted. □