Probability and Distribution Refresher

Korbinian Strimmer

18 December 2025

Table of contents

Welcome

The Probability and Distribution Refresher notes were written by Korbinian Strimmer from 2018–2025. This version is from 18 December 2025.

If you have any questions, comments, or corrections please get in touch! [1]

Updates

The lecture notes will be updated from time to time.

The most current version is found at the web page for the

- online version of the Probability and Distribution Refresher notes.

There you can also download the Probability and Distribution Refresher notes as

- PDF in A4 format for printing (double page layout), or as
- 6x9 inch PDF for use on tablets (single page layout).

License

[1] Email address: korbinian.strimmer@manchester.ac.uk

Preface

About the author

Hello! My name is Korbinian Strimmer and I am a Professor in Statistics. I am a member of the Statistics group at the Department of Mathematics of the University of Manchester. You can find more information about me on my home page.

About the notes

These supplementary notes provide a quick refresher on some essential combinatorics and probability concepts, and present an overview of selected univariate and multivariate distributions, along with an introduction to exponential families.

The notes are supporting information for a number of lecture courses in statistics that I teach or have taught at the Department of Mathematics of the University of Manchester.

This includes the currently offered modules:

- MATH27720 Statistics 2: Likelihood and Bayes and
- MATH38161 Multivariate Statistics

as well as the retired module (not offered any more):

- MATH20802 Statistical Methods.

1 Combinatorics

1.1 Some basic mathematical notation

Scalar quantity: plain font, typically lower case (x, θ, n), sometimes upper case (K, R^2, distribution functions F, P, Q).

Sets: plain font, upper case ($\Omega, \mathcal{F}$)

Vector quantity: bold font, lower case ($\boldsymbol{x}$, $\boldsymbol{\theta}$).

Matrix quantity: bold font, upper case ($\boldsymbol{X}$, $\boldsymbol{\Sigma}$).

Summation:

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n$$

Product:

$$\prod_{i=1}^{n} x_i = x_1 \times x_2 \times \ldots \times x_n$$
$$= x_1 x_2 \ldots x_n$$

The multiplication sign $\times$ between the factors is usually omitted unless it is needed for clarity.

Indicator function (in Iverson bracket notation):

$$[A] = \begin{cases} 1 & \text{if } A \text{ is true} \\ 0 & \text{if } A \text{ is not true} \end{cases}$$

1.2 Number of permutations

A **permutation** or **ordering** is a specific arrangement of items in a sequence, or equivalently, a specific assignment to labelled positions.

The **factorial**

$$n! = \prod_{i=1}^{n} i = 1 \times 2 \times \ldots \times n$$

is the number of permutations of n distinct items, where n is a positive integer.

For $n = 0$ the factorial is defined as

$$0! = 1$$

Thus, the factorial $n!$ equals the number of ways to place n distinct items into n labelled boxes so that each box contains exactly one item.

1.3 Multinomial and binomial coefficient

The **multinomial coefficient** for K groups

$$\begin{aligned} W_K &= \binom{n}{n_1, \ldots, n_K} \\ &= \frac{n!}{n_1!\, n_2! \ldots n_K!} \end{aligned}$$

is the number of permutations of n distinct items allocated to K groups, with n_k unordered items in group k and $\sum_{k=1}^{K} n_k = n$.

Thus, the multinomial coefficient W_K equals the number of ways to place n distinct items into K labelled boxes so that box k contains exactly n_k unordered items, with $\sum_{k=1}^{K} n_k = n$.

For $n_k = 1$ (and thus $K = n$) the multinomial coefficient reduces to the factorial.

For two groups ($K = 2$) the multinomial coefficient becomes the **binomial coefficient**

$$\begin{aligned} W_2 &= \binom{n}{n_1, n_2} = \binom{n}{n_1, n - n_1} \\ &= \frac{n!}{n_1!\,(n - n_1)!} \\ &= \binom{n}{n_1} \end{aligned}$$

1.4 De Moivre-Sterling approximation

The factorial is frequently approximated by the following formula derived by Abraham de Moivre (1667–1754) and James Stirling (1692-1770)

$$n! \approx \sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n}$$

or equivalently on logarithmic scale

$$\log n! \approx \left(n + \frac{1}{2}\right) \log n - n + \frac{1}{2} \log(2\pi)$$

The approximation is good for small n (but fails for $n = 0$) and becomes more and more accurate with increasing n. For large n the approximation can be simplified to

$$\log n! \approx n \log n - n$$

The de Moivre-Sterling approximation applied to the multinomial coefficient yields

$$\begin{aligned} \log W_K &\approx -n \sum_{k=1}^{K} \frac{n_k}{n} \log\left(\frac{n_k}{n}\right) \\ &= -n \sum_{k=1}^{K} q_k \log q_k = n H(\hat{Q}) \end{aligned}$$

Hence, for large n and large n_k the logarithm of the multinomial coefficient equals n times the information entropy $H(\hat{Q})$ of the empirical categorical distribution $\hat{Q}$ with class frequencies $\hat{q}_k = n_k / n$.

2 Probability

2.1 Random variables

A **random variable** describes a random experiment. The set of all possible outcomes is the **sample space** of the random variable and is denoted by Ω. If Ω is countable then the random variable is **discrete**, otherwise it is **continuous**. For a discrete random variable the sample space $\Omega = \{\omega_1, \omega_2, \ldots\}$ is composed of a finite or infinite number of **elementary outcomes** ω_i.

An event $A \subseteq \Omega$ is a subset of Ω. This includes as special cases the complete set Ω ("certain event") and the empty set $\emptyset$ ("impossible event"). The set of all possible events is denoted by $\mathcal{F}$. The complementary event $A^C = \Omega \setminus A$ is the complement of the set A in the sample space Ω. Two events A_1 and A_2 are mutually exclusive if the sets are disjoint with $A_1 \cap A_2 = \emptyset$.

For a discrete random variable, the elementary outcomes ω_i are referred to as **elementary events**, and they are all mutually exclusive. An event A consists of a number of elementary events $\omega_i \in A$ and the complementary event is given by $A^C = \{\omega_i \in \Omega : \omega_i \notin A\}$.

The **probability of an event** A is denoted by $\Pr(A)$. Broadly, $\Pr(A)$ provides a measure of the size of the set A relative to the set Ω. The probability measure $\Pr(A)$ satisfies the three axioms of probability:

1) $\Pr(A) \geq 0$, probabilities are non-negative,
2) $\Pr(\Omega) = 1$, the certain event has probability 1, and
3) $\Pr(A_1 \cup A_2 \cup \ldots) = \sum_i \Pr(A_i)$, the probability of countable mutually exclusive events A_i is additive.

This implies

- $\Pr(A) \leq 1$, probability values lie within the range $[0, 1]$,
- $\Pr(A^C) = 1 - \Pr(A)$, the probability of the complement, and
- $\Pr(\emptyset) = 0$, the impossible event has probability 0.

From the above it is evident that probability is closely linked to set theory, in particular to measure theory which serves as the theoretical foundations of probability and generalisations. For instance, if $\Pr(\emptyset) = 0$ is assumed instead of $\Pr(\Omega) = 1$, this leads to the axioms for a **positive measure** (of which probability is a special case).

2.2 Conditional probability

Consider two events A and B, which may not be be mutually exclusive. The probability of the event "A and B" is given by the probability of the set intersection $\Pr(A \cap B)$. The probability of the event "A or B" is given by the probability of the set union

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)\,.$$

This identity follows from the axioms.

The **conditional probability** of event A assuming event B has occurred is given by

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}$$

Essentially, now B acts as the new sample space relative to which A is measured, restricting it from Ω. Note that $\Pr(A|B)$ is generally not the same as $\Pr(B|A)$, see Bayes' theorem below.

Importantly, it can be seen that any probability may be viewed as conditional, namely relative to Ω as $\Pr(A) = \Pr(A|\Omega)$.

From the definition of conditional probability we derive the **product rule**

$$\begin{aligned}\Pr(A \cap B) &= \Pr(A|B)\ \Pr(B)\\ &= \Pr(B|A)\ \Pr(A)\end{aligned}$$

which in turn yields **Bayes' theorem**

$$\Pr(A|B) = \Pr(B|A)\frac{\Pr(A)}{\Pr(B)}$$

This theorem is useful for changing the order of conditioning and it plays a key role in Bayesian statistics.

If $\Pr(A \cap B) = \Pr(A)\ \Pr(B)$ then the two events A and B are **independent** with $\Pr(A|B) = \Pr(A)$ and $\Pr(B|A) = \Pr(B)$.

2.3 Probability mass and density function

The **distribution** (or **law**) of a random variable x with sample space Ω is the probability measure that assigns probabilities to values or ranges of x. This is done in practise by employing probability mass functions (pmf, for discrete random variables) or probability density functions (pdf, for continuous random variables).

The scalar random variable x is written in lowercase plain font. We use the same symbol x for both the random variable and its realisations.[1]

For a discrete random variable we define the event $A = \{x : x = a\} = \{a\}$ (corresponding to a single elementary event) and get the probability

$$\Pr(A) = \Pr(x = a) = f(a)$$

directly from the **probability mass function** (**pmf**). The pmf has the property that $\sum_{x\in\Omega} f(x) = 1$ and that $f(x) \in [0, 1]$.

For continuous random variables employ a **probability density function** (**pdf**) instead. We define the event $A = \{x : a < x \leq a + da\}$ (corresponding to an infinitesimal interval) and then assign the probability

$$\Pr(A) = \Pr(a < x \leq a + da) = f(a)da\,.$$

Similarly, the probability of the event $A = \{x : a_1 < x \leq a_2\}$ is given by

$$\Pr(A) = \Pr(a_1 < x \leq a_2) = \int_{a_1}^{a_2} f(a)da\,.$$

The pdf has the property that $\int_{x\in\Omega} f(x)dx = 1$ but in contrast to a pmf the density $f(x) \geq 0$ may take on values larger than 1.

It is sometimes convenient to refer to a pdf or a pmf collectively **as probability density mass function** (**pdmf**) without specifying whether x is continuous or discrete.

The set of all x for which $f(x)$ is positive is called the **support** of the pdmf.

[1] This notation is common in statistical machine learning and multivariate statistics, see for example Mardia, Kent, and Bibby (1979). An alternative convention uses uppercase letters for random variables and lowercase for outcomes, but that convention is problematic for multivariate objects (random vectors and random matrices) and is also ill-suited in Bayesian statistics where parameters are modelled as random variables.

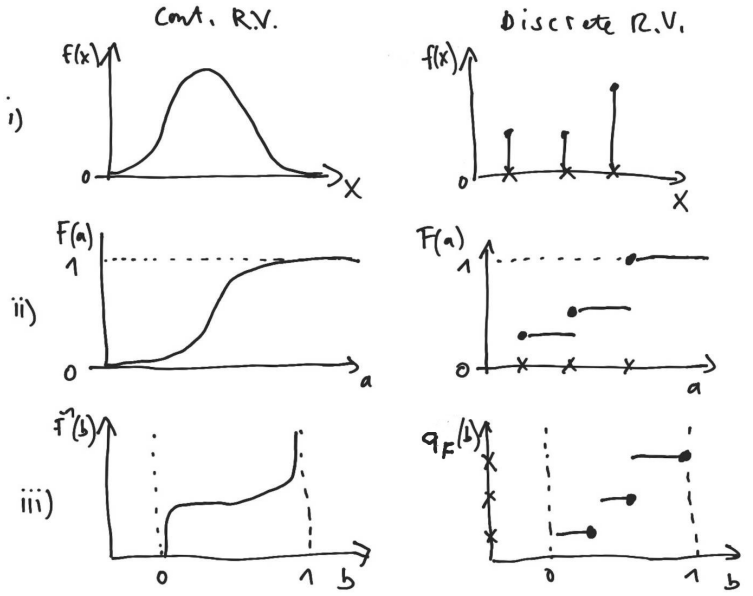

Figure 2.1: Illustration of i) pdmf, ii) distribution function and iii) quantile function for a continuous (first column) and a discrete random variable (second column).

Using the pdmf, the probability of general event $A \subseteq \Omega$ is given by

$$\Pr(A) = \begin{cases} \sum_{x\in A} f(x) & \text{discrete case} \\ \int_{x\in A} f(x)dx & \text{continuous case} \end{cases}$$

Figure 2.1 (first row) illustrates the pdmf for a continuous and discrete random variable.

In the above we denoted the pdmf by the lower case letter f though we also often use p or q.

2.4 Cumulative distribution function

As alternative to the pdmf we can describe the random variable using a **cumulative distribution function** (**cdf**). This requires an ordering so

that we can define the event $A = \{x : x \leq a\}$ and compute its probability as

$$F(a) = \Pr(A) = \Pr(x \leq a) = \begin{cases} \sum_{x \in A} f(x) & \text{discrete case} \\ \int_{x \in A} f(x)dx & \text{continuous case} \end{cases}$$

Th cdf is denoted by the same letter as the pdmf but in upper case (usually F, P and Q). By construction the cumulative distribution function is monotonically non-decreasing and its value ranges from 0 to 1. For a discrete random variable $F(a)$ is a step function with jumps of size $f(\omega_i)$ at the elementary outcomes ω_i.

With the help of the cdf we can compute the probability of the event $A = \{x : a_1 < x \leq a_2\}$ simply as

$$\Pr(A) = F(a_2) - F(a_1)\,.$$

This works both for discrete and continuous random variables.

Figure 2.1 (second row) illustrates the distribution function for a continuous and discrete random variable.

It is common to use the same upper case letter as the cdf to name the distribution. Thus, if a random variable x has distribution F we write $x \sim F$, and this implies it has a pdmf $f(x)$ and cdf $F(x)$.

2.5 Quantile function and quantiles

The **quantile function** is defined as $q_F(b) = \min\{x : F(x) \geq b\}$. For a continuous random variable the quantile function simplifies to $q_F(b) = F^{-1}(b)$, i.e. it is the ordinary inverse $F^{-1}(b)$ of the distribution function.

Figure 2.1 (third row) illustrates the quantile function for a continuous and discrete random variable.

The quantile x of order b of the distribution F is often denoted by $x_b = q_F(b)$.

The 25% quantile $x_{1/4} = x_{25\%} = q_F(1/4)$ is called the **first quartile** or **lower quartile**.

The 50% quantile $x_{1/2} = x_{50\%} = q_F(1/2)$ is called the **second quartile** or **median**.

The 75% quantile $x_{3/4} = x_{75\%} = q_F(3/4)$ is called the **third quartile** or **upper quartile**.

The interquartile range is the difference between the upper and lower quartiles and equals $\text{IQR}(F) = q_F(3/4) - q_F(1/4)$.

The quantile function is also useful for generating general random variates from uniform random variates. If $y \sim \text{Unif}(0, 1)$ then $x = q_F(y) \sim F$.

2.6 Expectation or mean

The expected value of a random variable $x \sim F$ is defined as the weighted average over all possible outcomes, with the weight given by the pdmf $f(x)$:

$$\begin{aligned} \text{E}(x) &= \text{E}_F(x) = \text{E}(F) \\ &= \begin{cases} \sum_{x \in \Omega} f(x)\, x & \text{discrete case} \\ \int_{x \in \Omega} f(x)\, x\, dx & \text{continuous case} \end{cases} \end{aligned}$$

The subscript F in $\text{E}_F(x)$ indicates that the expectation is taken with regard to the distribution F, but is usually left out if there are no ambiguities. The notation $\text{E}(F)$ emphasises that the mean is a functional of the distribution F.

Because the sum or integral may diverge, not all distributions have finite means so the mean does not always exist (in contrast to the median, or quantiles in general). For example, the location-scale t-distribution $t_\nu(\mu, \tau^2)$ does not have a mean for a degree of freedom in the range $0 < \nu \leq 1$ (see Section 5.6).

Expectation is a **linear operator**, meaning that

$$\text{E}(a_1 x_1 + a_2 x_2) = a_1\, \text{E}(x_1) + a_2\, \text{E}(x_2)$$

for random variables $x_1 \sim F_1$ and $x_2 \sim F_2$ and constants a_1 and a_2.

Consequently, expectation is **mixture preserving** so that

$$\text{E}(Q_\lambda) = (1 - \lambda)\, \text{E}(Q_0) + \lambda\, \text{E}(Q_1)$$

for the mixture $Q_\lambda = (1 - \lambda)Q_0 + \lambda Q_1$ with $0 < \lambda < 1$ and $Q_0 \neq Q_1$.

2.7 Variance

The variance of a random variable $x \sim F$ is the expected value of the squared deviation around the mean $\mu = \mathrm{E}(x)$:

$$\begin{aligned}\mathrm{Var}(x) &= \mathrm{Var}_F(x) = \mathrm{Var}(F) \\ &= \mathrm{E}\left((x-\mu))^2\right) \\ &= \mathrm{E}(x^2) - \mu^2\end{aligned}$$

By construction, $\mathrm{Var}(x) \geq 0$.

The notation $\mathrm{Var}(F)$ highlights that the variance is a functional of the distribution F. Occasionally, we write $\mathrm{Var}_F(x)$ indicate that the expectation is taken with regard to the distribution F.

Like the mean, the variance may diverge and hence not necessarily exists for all distribution. For example, the location-scale t-distribution $t_\nu(\mu, \tau^2)$ does not have a variance for the degree of freedom in the range $0 < \nu \leq 2$ (see Section 5.6).

2.8 Moments of a distribution

The n-th moment of a distribution F for a random variable x is defined as follows:

$$\mu_n(F) = \mathrm{E}(x^n)$$

Special important cases are the

- Zeroth moment: $\mu_0(F) = \mathrm{E}(x^0) = 1$ (since the pdmf integrates to one)
- First moment: $\mu_1(F) = \mathrm{E}(x^1) = \mathrm{E}(x) = \mu$ (=the mean)
- Second moment: $\mu_2(F) = \mathrm{E}(x^2)$

The n-th central moment centred around the mean $\mathrm{E}(x) = \mu$ is given by

$$m_n(F) = \mathrm{E}((x-\mu)^n)$$

The first few central moments are the

- Zeroth central moment: $m_0(F) = \mathrm{E}((x-\mu)^0) = 1$
- First central moment: $m_1(F) = \mathrm{E}((x-\mu)^1) = 0$
- Second central moment: $m_2(F) = \mathrm{E}\left((x-\mu)^2\right)$ (=the variance)

The moments of a distribution are not necessarily all finite, i.e. some moments may not exist. For example, the location-scale t-distribution $t_\nu(\mu, \tau^2)$ only has finite moments of degree smaller than the degree of freedom ν (see Section 5.6).

2.9 Expectation of a transformed random variable

Often, one needs to find the mean of a transformed random variable. If $x \sim F_x$ and $y = h(x)$ with $y \sim F_y$ then one can directly apply the above definition to obtain $\mathrm{E}(y) = \mathrm{E}(F_y)$. However, this requires knowledge of the transformed pdmf $f_y(y)$ (see Chapter 3 for more details about variable transformations).

As an alternative, the "law of the unconscious statistician"(LOTUS) provides a convenient shortcut to compute the mean of the transformed random variable $y = h(x)$ using only the pdmf of the original variable x:

$$\mathrm{E}(h(x)) = \begin{cases} \sum_{x\in\Omega} f(x)\, h(x) & \text{discrete case} \\ \int_{x\in\Omega} f(x)\, h(x)\, dx & \text{continuous case} \end{cases}$$

Note this is not an approximation but equivalent to obtaining the mean using the transformed pdmf.

2.10 Probability as expectation

Probability itself can also be understood as an expectation.

For an event $A \subseteq \Omega$ we define a corresponding indicator function $[x \in A]$. From LOTUS it then follows immediately that

$$\begin{aligned}\mathrm{E}([x \in A]) &= \begin{cases} \sum_{x\in A} f(x) & \text{discrete case} \\ \int_{x\in A} f(x)\, dx & \text{continuous case} \end{cases} \\ &= \Pr(A)\end{aligned}$$

This relation is called the "fundamental bridge" between probability and expectation. Interestingly, one can develop the whole theory of probability from this perspective (e.g., Whittle 2000).

2.11 Jensen's inequality for the expectation

If $h(\boldsymbol{x})$ is a *convex* function then the following inequality holds:

$$\mathrm{E}(h(\boldsymbol{x})) \geq h(\mathrm{E}(\boldsymbol{x}))$$

Recall: a con**ve**x function (such as x^2) has the shape of a "**va**lley".

An example of Jensen's inequality is $\mathrm{E}(x^2) \geq \mathrm{E}(x)^2$.

2.12 Random vectors and their mean and variance

In addition to scalar random variables we often make use of random vectors and random matrices.[2]

The mean of a random vector $\boldsymbol{x} = (x_1, x_2, ..., x_d)^T \sim F$ is given by

$$\begin{aligned}\mathrm{E}(\boldsymbol{x}) &= \mathrm{E}(F) \\ &= \underbrace{\boldsymbol{\mu}}_{d\times 1} = (\mu_1, \ldots, \mu_d)^T\end{aligned}$$

and thus is a vector of the same dimension as $\boldsymbol{x}$, where $\mu_i = \mathrm{E}(x_i)$ are the means of the individual components x_i.

The variance of a random vector $\boldsymbol{x}$ of length d, however, is not a vector but a matrix of size $d \times d$. This matrix is called the **covariance matrix**:

$$\begin{aligned}\mathrm{Var}(\boldsymbol{x}) &= \mathrm{Var}(F) \\ &= \underbrace{\boldsymbol{\Sigma}}_{d\times d} = (\sigma_{ij}) = \begin{pmatrix} \sigma_{11} & \ldots & \sigma_{1d} \\ \vdots & \ddots & \vdots \\ \sigma_{d1} & \ldots & \sigma_{dd} \end{pmatrix} \\ &= \mathrm{E}\left(\underbrace{(\boldsymbol{x} - \boldsymbol{\mu})}_{d\times 1} \underbrace{(\boldsymbol{x} - \boldsymbol{\mu})^T}_{1\times d} \right) \\ &= \mathrm{E}(\boldsymbol{x}\boldsymbol{x}^T) - \boldsymbol{\mu}\boldsymbol{\mu}^T\end{aligned}$$

[2] In our notational conventions, a scalar x is written in lower case plain font, a vector $\boldsymbol{x}$ is written in lower case bold font, a matrix $\boldsymbol{X}$ in upper case bold font.

The elements $\mathrm{Cov}(x_i, x_j) = \sigma_{ij}$ describe the covariance between the random variables x_i and x_j. The covariance matrix is symmetric, hence $\sigma_{ij} = \sigma_{ji}$. The diagonal elements $\mathrm{Cov}(x_i, x_i) = \sigma_{ii}$ correspond to the individual variances $\mathrm{Var}(x_i) = \sigma_i^2$. By construction, the covariance matrix $\boldsymbol{\Sigma}$ is **positive semi-definite**, i.e. the eigenvalues of $\boldsymbol{\Sigma}$ are all positive or equal to zero.

However, wherever possible one will aim to use models with non-singular covariance matrices, with all eigenvalues positive, so that the covariance matrix is invertible.

2.13 Correlation matrix

The **correlation matrix** $\boldsymbol{P}$ ("upper case rho", not "upper case p") is the variance standardised version of the covariance matrix $\boldsymbol{\Sigma}$.

Specifically, denote by $\boldsymbol{V}$ the diagonal matrix containing the variances

$$\boldsymbol{V} = \begin{pmatrix} \sigma_{11} & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \sigma_{dd} \end{pmatrix}$$

then the correlation matrix $\boldsymbol{P}$ is given by

$$\boldsymbol{P} = (\rho_{ij}) = \begin{pmatrix} 1 & \ldots & \rho_{1d} \\ \vdots & \ddots & \vdots \\ \rho_{d1} & \ldots & 1 \end{pmatrix} = \boldsymbol{V}^{-1/2}\,\boldsymbol{\Sigma}\,\boldsymbol{V}^{-1/2}$$

Like the covariance matrix the correlation matrix is symmetric. The elements of the diagonal of $\boldsymbol{P}$ are all set to 1.

Equivalently, in component notation the correlation between x_i and x_j is given by

$$\rho_{ij} = \mathrm{Cor}(x_i, x_j) = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$$

Following from the definition above, a covariance matrix $\boldsymbol{\Sigma}$ can be factorised into the product of standard deviations $\boldsymbol{V}^{1/2}$ and the correlation matrix $\boldsymbol{P}$ as follows:

$$\boldsymbol{\Sigma} = \boldsymbol{V}^{1/2}\,\boldsymbol{P}\,\boldsymbol{V}^{1/2}$$

2.14 Parameters and families of distributions

A **distribution family** $F(\theta)$ is a collection of distributions obtained by varying a parameter θ. Each specific value of the parameter θ indexes one distribution in that family.

Common distribution families are usually denoted by familiar abbreviation such as $N(\mu, \sigma^2)$ for the normal family. We also call these simply "distributions" with parameters and omit the word "family".

If a random variable x has distribution $F(\theta)$ we write $x \sim F(\theta)$. In case of named distributions, we use the corresponding abbreviation such $x \sim N(\mu, \sigma^2)$ (normal distribution) or $x \sim \text{Beta}(\alpha_1, \alpha_2)$.

The associated pdmf is written $f(x; \theta)$ or $f(x|\theta)$. The conditional notation is more general because it implies the parameter θ may have its own distribution, yielding a joint density $f(x, \theta) = f(x|\theta)f(\theta)$. Similarly, the corresponding cumulative distribution function is written $F(x; \theta)$ or $F(x|\theta)$.

Note that parametrisations are generally not unique, as any one-to-one transformation of θ yields an equivalent index of the same distribution family. For most commonly used distribution families there exist several standard parametrisations. We usually prefer those whose parameters that can be interpreted easily (e.g. in terms of moments) or that help to simplify calculations.

If distinct parameters correspond to distinct distributions they are called **identifiable**. This is an important property as it allows parameters to be estimated from data. Specifically, if parameters are identifiable then $P(\theta_1) = P(\theta_2)$ implies $\theta_1 = \theta_2$, and conversely if $P(\theta_1) \neq P(\theta_2)$ then $\theta_1 \neq \theta_2$. If parameters and distributions are unique within a neighbourhood $\theta_0 + \epsilon$ relative to a reference value θ_0, rather than globally, they are **locally identifiable** at θ_0.

An important class of distribution families are **exponential families** discussed in Chapter 7.

3 Transformations and convolution

3.1 Affine or location-scale transformation

Transformation rule

Suppose x is a scalar. The variable

$$y = a + bx$$

is a **location-scale transformation** or **affine transformation** of x, where a plays the role of the **location parameter** and b is the **scale parameter**. For $a = 0$ this is a **linear transformation**.

If $b \neq 0$ then the transformation is **invertible**, with back-transformation

$$x = (y - a)/b$$

Invertible transformations provide a one-to-one map between x and y.

For a vector $\boldsymbol{x}$ of dimension d the location-scale transformation is

$$\boldsymbol{y} = \boldsymbol{a} + \boldsymbol{B}\boldsymbol{x}$$

where $\boldsymbol{a}$ (a $m \times 1$ vector) is the **location parameter** and $\boldsymbol{B}$ (a $m \times d$ matrix) the **scale parameter**. For $\boldsymbol{a} = 0$ this is a **linear transformation**.

For $m = d$ (square $\boldsymbol{B}$) and $\det(\boldsymbol{B}) \neq 0$ the affine transformation is **invertible** with back-transformation

$$\boldsymbol{x} = \boldsymbol{B}^{-1}(\boldsymbol{y} - \boldsymbol{a})$$

Probability mass function

If $x \sim F_x$ is a discrete scalar random variable with pmf $f_x(x)$ and assuming an invertible transformation $y(x) = a + bx$ the pmf $f_y(y)$ for the discrete scalar random variable y is given by

$$f_y(y) = f_x\left(\frac{y-a}{b}\right)$$

Likewise, if $\boldsymbol{x} \sim F_{\boldsymbol{x}}$ is a discrete random vector with pmf $f_{\boldsymbol{x}}(\boldsymbol{x})$ and assuming an invertible transformation $\boldsymbol{y}(\boldsymbol{x}) = \boldsymbol{a} + \boldsymbol{B}\boldsymbol{x}$ the pmf $f_{\boldsymbol{y}}(\boldsymbol{y})$ for the discrete random vector $\boldsymbol{y}$ is given by

$$f_{\boldsymbol{y}}(\boldsymbol{y}) = f_{\boldsymbol{x}}\left(\boldsymbol{B}^{-1}(\boldsymbol{y}-\boldsymbol{a})\right)$$

Density

If $x \sim F_x$ is a continuous scalar random variable with pdf $f_x(x)$ and assuming an invertible transformation $y(x) = a + bx$ the pdf $f_y(y)$ for the continuous random scalar y is given by

$$f_y(y) = |b|^{-1} f_x\left(\frac{y-a}{b}\right)$$

where $|b|$ is the absolute value of b. The transformation of the corresponding differential element is

$$dy = |b|\, dx$$

Likewise, if $\boldsymbol{x} \sim F_{\boldsymbol{x}}$ is a continuous random vector with pdf $f_{\boldsymbol{x}}(\boldsymbol{x})$ and assuming an invertible transformation $\boldsymbol{y}(\boldsymbol{x}) = \boldsymbol{a} + \boldsymbol{B}\boldsymbol{x}$ the pdf $f_{\boldsymbol{y}}(\boldsymbol{y})$ for the continuous random vector $\boldsymbol{y}$ is given by

$$f_{\boldsymbol{y}}(\boldsymbol{y}) = |\det(\boldsymbol{B})|^{-1} f_{\boldsymbol{x}}\left(\boldsymbol{B}^{-1}(\boldsymbol{y}-\boldsymbol{a})\right)$$

where $|\det(\boldsymbol{B})|$ is the absolute value of the determinant $\det(\boldsymbol{B})$. The transformation of the corresponding infinitesimal volume element is

$$d\boldsymbol{y} = |\det(\boldsymbol{B})|\, d\boldsymbol{x}$$

Moments

The transformed random variable $y \sim F_y$ has mean

$$\mathrm{E}(y) = a + b\mu_x$$

and variance

$$\mathrm{Var}(y) = b^2\sigma_x^2$$

where $\mathrm{E}(x) = \mu_x$ and $\mathrm{Var}(x) = \sigma_x^2$ are the mean and variance of the original variable x.

The mean and variance of the transformed random vector $\boldsymbol{y} \sim F_{\boldsymbol{y}}$ is

$$\mathrm{E}(\boldsymbol{y}) = \boldsymbol{a} + \boldsymbol{B}\,\boldsymbol{\mu}_{\boldsymbol{x}}$$

and

$$\mathrm{Var}(\boldsymbol{y}) = \boldsymbol{B}\,\boldsymbol{\Sigma}_{\boldsymbol{x}}\,\boldsymbol{B}^T$$

where $\mathrm{E}(\boldsymbol{x}) = \boldsymbol{\mu}_{\boldsymbol{x}}$ and $\mathrm{Var}(\boldsymbol{x}) = \boldsymbol{\Sigma}_{\boldsymbol{x}}$ are the mean and variance of the original random vector $\boldsymbol{x}$.

Importance of affine transformations

The constants $\boldsymbol{a}$ and $\boldsymbol{B}$ (or a and b in the univariate case) are the parameters of the **location-scale family** $F_{\boldsymbol{y}}(\boldsymbol{a}, \boldsymbol{B})$ created from $F_{\boldsymbol{x}}$. Many important distributions are location-scale families such as the normal distribution (cf. Section 5.3 and Section 6.3) and the location-scale t-distribution (Section 5.6 and Section 6.6). Furthermore, key procedures in multivariate statistics such as orthogonal transformations (including PCA) or whitening transformations (e.g. the Mahalanobis transformation) are affine transformations.

3.2 General invertible transformation

Transformation rule

As above we assume x is a scalar and $\boldsymbol{x}$ is a vector and consider the general invertible transformation.

For a scalar variable the transformation is specified by $y(x) = h(x)$ and the back-transformation by $x(y) = h^{-1}(y)$. For a vector this becomes $\boldsymbol{y}(\boldsymbol{x}) = \boldsymbol{h}(\boldsymbol{x})$ with back-transformation $\boldsymbol{x}(\boldsymbol{y}) = \boldsymbol{h}^{-1}(\boldsymbol{y})$. The functions $h(x)$ and $\boldsymbol{h}(\boldsymbol{x})$ are assumed to be invertible.

Probability mass function

If $x \sim F_x$ is a discrete scalar random variable with pmf $f_x(x)$ then the pmf $f_y(y)$ of the transformed discrete scalar random variable $y(x)$ is given by

$$f_y(y) = f_x(x(y))$$

Likewise, for a discrete random vector $\boldsymbol{x} \sim F_{\boldsymbol{x}}$ with pmf $f_{\boldsymbol{x}}(\boldsymbol{x})$ the pmf $f_{\boldsymbol{y}}(\boldsymbol{y})$ for the discrete random vector $\boldsymbol{y}(\boldsymbol{x})$ is obtained by

$$f_{\boldsymbol{y}}(\boldsymbol{y}) = f_{\boldsymbol{x}}\left(\boldsymbol{x}(\boldsymbol{y})\right)$$

Density

If $x \sim F_x$ is a continuous scalar random variable with pdf $f_x(x)$ the pdf $f_y(y)$ of the transformed continuous scalar random variable $y(x)$ is given by

$$f_y(y) = \left|Dx(y)\right| \, f_x(x(y))$$

where $Dx(y)$ is the derivative of the inverse transformation $x(y)$. The transformation of the differential element is

$$dy = \left|Dy(x)\right| \, dx$$

Note that $|Dx(y)| = |Dy(x)|^{-1}|_{x=x(y)}$.

Likewise, for a continuous random vector $\boldsymbol{x} \sim F_{\boldsymbol{x}}$ with pdf $f_{\boldsymbol{x}}(\boldsymbol{x})$ the pdf $f_{\boldsymbol{y}}(\boldsymbol{y})$ for the continuous random vector $\boldsymbol{y}(\boldsymbol{x})$ is obtained by

$$f_{\boldsymbol{y}}(\boldsymbol{y}) = |\det\left(D\boldsymbol{x}(\boldsymbol{y})\right)| \, f_{\boldsymbol{x}}\left(\boldsymbol{x}(\boldsymbol{y})\right)$$

where $D\boldsymbol{x}(\boldsymbol{y})$ is the Jacobian matrix of the inverse transformation $\boldsymbol{x}(\boldsymbol{y})$. The transformation of the infinitesimal volume element is

$$d\boldsymbol{y} = |\det\left(D\boldsymbol{y}(\boldsymbol{x})\right)| \, d\boldsymbol{x}$$

Note that $|\det\left(D\boldsymbol{x}(\boldsymbol{y})\right)| = |\det\left(D\boldsymbol{y}(\boldsymbol{x})\right)|^{-1}|_{\boldsymbol{x}=\boldsymbol{x}(\boldsymbol{y})}$.

Moments

The mean and variance of the transformed random variable can typically only be approximated. Assume that $\mathrm{E}(x) = \mu_x$ and $\mathrm{Var}(x) = \sigma_x^2$ are the mean and variance of the original random variable x and $\mathrm{E}(\boldsymbol{x}) = \boldsymbol{\mu}_{\boldsymbol{x}}$ and $\mathrm{Var}(\boldsymbol{x}) = \boldsymbol{\Sigma}_{\boldsymbol{x}}$ are the mean and variance of the original random vector $\boldsymbol{x}$. In the **delta method** the transformation $y(x)$ resp. $\boldsymbol{y}(\boldsymbol{x})$ is linearised around the mean μ_x respectively $\boldsymbol{\mu}_{\boldsymbol{x}}$ and the mean and variance resulting from the linear transformation is reported.

Specifically, the linear approximation for the scalar-valued function is

$$y(x) \approx y\left(\mu_x\right) + Dy\left(\mu_x\right) \, \left(x - \mu_x\right)$$

where $Dy(x) = y'(x)$ is the first derivative of the transformation $y(x)$ and $Dy\left(\mu_x\right)$ is the first derivative evaluated at the mean μ_x, and for the vector-valued function

$$\boldsymbol{y}(\boldsymbol{x}) \approx \boldsymbol{y}\left(\boldsymbol{\mu}_{\boldsymbol{x}}\right) + D\boldsymbol{y}\left(\boldsymbol{\mu}_{\boldsymbol{x}}\right) \, \left(\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{x}}\right)$$

where $D\boldsymbol{y}(\boldsymbol{x})$ is the Jacobian matrix (vector derivative) for the transformation $\boldsymbol{y}(\boldsymbol{x})$ and $D\boldsymbol{y}\left(\boldsymbol{\mu}_{\boldsymbol{x}}\right)$ is the Jacobian matrix evaluated at the mean $\boldsymbol{\mu}_{\boldsymbol{x}}$.

In the univariate case the delta method yields as approximation for the mean and variance of the transformed random variable y

$$\mathrm{E}(y) \approx y\left(\mu_x\right)$$

and

$$\mathrm{Var}(y) \approx \left(Dy\left(\mu_x\right)\right)^2 \, \sigma_x^2$$

For the vector random variable $\boldsymbol{y}$ the delta method yields

$$\mathrm{E}(\boldsymbol{y}) \approx \boldsymbol{y}\left(\boldsymbol{\mu}_{\boldsymbol{x}}\right)$$

and

$$\mathrm{Var}(\boldsymbol{y}) \approx D\boldsymbol{y}\left(\boldsymbol{\mu}_{\boldsymbol{x}}\right) \, \boldsymbol{\Sigma}_{\boldsymbol{x}} \, D\boldsymbol{y}\left(\boldsymbol{\mu}_{\boldsymbol{x}}\right)^T$$

Invertible affine transformation as special case

The invertible affine transformation (Section 3.1) is special case of the general invertible transformation.

Assuming $y(x) = a + bx$, with $x(y) = (y - a)/b$, $Dy(x) = b$ and $Dx(y) = b^{-1}$, recovers the univariate location-scale transformation.

Likewise, assuming $\boldsymbol{y}(\boldsymbol{x}) = \boldsymbol{a} + \boldsymbol{B}\boldsymbol{x}$, with $\boldsymbol{x}(\boldsymbol{y}) = \boldsymbol{B}^{-1}(\boldsymbol{y} - \boldsymbol{a})$, $D\boldsymbol{y}(\boldsymbol{x}) = \boldsymbol{B}$ and $D\boldsymbol{x}(\boldsymbol{y}) = \boldsymbol{B}^{-1}$, recovers the multivariate location-scale transformation.

3.3 Convolution of random variables

Sum of independent random variables

Suppose we have a sum of n *independent* scalar random variables.

$$y = x_1 + x_2 + \ldots + x_n$$

where each $x_i \sim F_{x_i}$ has its own distribution and corresponding pdmf $f_{x_i}(x)$. The corresponding means are $\mathrm{E}(x_i) = \mu_i$ and the variances are $\mathrm{Var}(x_i) = \sigma_i^2$. As the x_i are independent, and therefore uncorrelated, the covariances $\mathrm{Cov}(x_i, x_j) = 0$ vanish for $i \neq j$.

With $\boldsymbol{x} = (x_1, \ldots, x_n)^T$ and $\mathbf{1}_n = (1, 1, \ldots, 1)^T$ the relationship between y and $\boldsymbol{x}$ can be written as the linear transformation

$$y = \mathbf{1}_n^T \boldsymbol{x}$$

As y is a scalar and $\boldsymbol{x}$ a vector the transformation from $\boldsymbol{x}$ to y is not invertible.

Moments

With $\mathrm{E}(\boldsymbol{x}) = \boldsymbol{\mu}$ and $\mathrm{Var}(\boldsymbol{x}) = \mathrm{Diag}(\sigma_1^2, \ldots, \sigma_n^2)$ the mean of the random variable y equals

$$\mathrm{E}(y) = \mathbf{1}_n^T \boldsymbol{\mu} = \sum_{i=1}^{n} \mu_i$$

and the variance of y is

$$\mathrm{Var}(y) = \mathbf{1}_n^T \, \mathrm{Var}(\boldsymbol{x}) \, \mathbf{1}_n = \sum_{i=1}^{n} \sigma_i^2$$

(cf. Section 3.1). Thus both the mean and variance of y are simply the sums of the individual means and variances (note that for the variance this only holds because the individual variables are uncorrelated).

Convolution

The pdmf $f_y(y)$ for y is obtained by repeatedly convolving (denoted by the asterisk $*$ operator) the pdmfs of the x_i:

$$f_y(y) = \left(f_{x_1} * f_{x_2} * \ldots f_{x_n}\right)(y)$$

The **convolution** of two functions is defined as (continuous case)

$$\left(f_{x_1} * f_{x_2}\right)(y) = \int_x f_{x_1}(x) \, f_{x_2}(y - x) dx$$

and (discrete case)

$$\left(f_{x_1} * f_{x_2}\right)(y) = \sum_x f_{x_1}(x) \, f_{x_2}(y - x)$$

Convolution is commutative and associative so so you may convolve multiple pdmfs in any order or grouping. Furthermore, the convolution of pdmfs yields another pdmf, i.e. the resulting function integrates to one.

Many commonly used random variables can be viewed as the outcome of convolutions. For example, the sum of Bernoulli variables yields a binomial random variable and the sum of normal variables yields another normal random variable.

See also (Wikipedia): list of convolutions of probability distributions.

Central limit theorem

The **central limit theorem**, first postulated by Abraham de Moivre (1667–1754) and later proved by Pierre-Simon Laplace (1749–1827) asserts that the distribution of the sum of n independent and identically distributed random variables with finite mean and finite variance converges in the limit of large n to a normal distribution (Section 5.3), even if the individual random variables are not themselves normal. In other words, it asserts that for large n the convolution of n identical distributions with finite first two moments converges to a normal distribution.

4 Evaluation

4.1 Loss functions

Loss function

A **loss or cost function** $L(x, a)$ evaluates a prediction a, for example a parameter or a probability distribution, on the basis of an observed outcome x, and returns a numerical score.

A loss function measures, informally, the error between x and a. During optimisation the prediction a is varied and the aim is minimisation of the error (hence a loss function has *negative orientation*, smaller is better).

A **utility or reward function** is a loss function with a reversed sign (hence it has *positive orientation*, larger is better).

Risk function

The **risk** of a under the distribution Q for x is defined as the expected loss

$$R(Q, a) = \mathrm{E}_Q(L(x, a))$$

The risk is **mixture preserving in Q** meaning that

$$R(Q_\lambda, a) = (1 - \lambda)R(Q_0, a) + \lambda R(Q_1, a)$$

for the mixture $Q_\lambda = (1 - \lambda)Q_0 + \lambda Q_1$ with $0 < \lambda < 1$ and $Q_0 \neq Q_1$. This follows from the linearity of expectation.

The risk of a under the empirical distribution $\hat{Q}_n$ obtained from observations $x_1, \ldots, x_n$ is the **empirical risk**

$$R(\hat{Q}_n, a) = \frac{1}{n} \sum_{i=1}^{n} L(x_i, a)$$

where the expectation is replaced by the sample average.

Minimising risk

Minimising $R(Q, a)$ with regard to a finds optimal predictions

$$a^* = \arg\min_a R(Q, a)$$

with associated minimum risk $R(Q, a^*)$.

Depending on the choice of underlying loss $L(x, a)$ minimising the risk provides a very general optimisation-based way to identify distributional features of the distribution Q and to obtain parameter estimates.

Equivalent loss functions

Adding a positive scaling factor $k > 0$ or an additive term $c(x)$ to a loss function generates a family of **equivalent loss functions**

$$L^{\text{equiv}}(x, a) = kL(x, a) + c(x)$$

with associated risk

$$R^{\text{equiv}}(Q, a) = kR(Q, a) + \mathrm{E}_Q(c(x))$$

Equivalent losses yield the **same risk minimiser** $\arg\min_a R(Q, a)$ and the **same loss minimiser** $\arg\min_a L(x, a)$ for fixed x.

4.2 Common loss functions

Squared loss

The **squared loss** or **squared error**

$$L_{\text{sq}}(x, a) = (x - a)^2$$

is one of the most commonly used loss functions. The corresponding risk is the **mean squared loss** or **mean squared error** (MSE)

$$R_{\text{sq}}(Q, a) = \mathrm{E}_Q((x - a)^2)$$

which is **minimised at the mean** $a^* = \mathrm{E}(Q)$. This follows from $R_{\text{sq}}(Q, a) = \mathrm{E}_Q(x^2) - 2a\,\mathrm{E}_Q(x) + a^2$ and $dR_{\text{sq}}(Q, a)/da = -2\,\mathrm{E}_Q(x) + 2a$. The **minimum risk** $R_{\text{sq}}(a^*) = \mathrm{Var}(Q)$ equals the **variance**.

0-1 loss

The **0-1 loss** function can be written as

$$L_{01}(x, a) = \begin{cases} -[x = a] & \text{discrete case} \\ -\delta(x - a) & \text{continuous case} \end{cases}$$

employing the indicator function and Dirac delta function, respectively. The corresponding risk assuming $x \sim Q$ and pdmf $q(x)$ is

$$R_{01}(Q, a) = -q(a)$$

which is **minimised at the mode** of the pdmf.

Asymmetric loss

The **asymmetric loss** can be defined as

$$L_{\text{asym}}(x, a; \tau) = \begin{cases} 2\tau(x - a) & \text{for } x \geq a \\ 2(1 - \tau)(a - x) & \text{for } x < a \end{cases}$$

and the corresponding risk is **minimised at the quantile** x_τ.

Absolute loss

For $\tau = 1/2$ it reduces to the **absolute loss**

$$L_{\text{abs}}(x, a) = |x - a|$$

whose corresponding risk is **minimised at the median** $x_{1/2}$.

4.3 Scoring rules

Proper scoring rules

A **scoring rule** $S(x, P)$ is special type of loss function[1] that assesses the probabilistic forecast P by assigning a numerical score based on P and the observed outcome x.

[1] Treating scoring rules as loss functions implies a negative orientation. However, some authors adopt the opposite convention and treat scoring rules as positively oriented utility functions.

The **risk** of P under Q is the expected score

$$R(Q, P) = \mathrm{E}_Q\left(S(x, P)\right)$$

For a **proper scoring rule**, the risk $R(Q, P)$ is smallest when the quoted model P matches the true model Q. The minimal risk, achieved for $P = Q$, leads to the **properness inequality**

$$R(Q, P) \geq R(Q, Q)$$

For a **strictly proper** scoring rule, the minimum risk is realised only for the true model, so **equality holds exclusively** for $P = Q$.

Proper scoring rules are useful because they enable identification and approximation of data-generating distributions and their parameters via risk minimisation or minimisation of the associated scoring-rule divergences. This allows to generalise conventional statistical approaches based on the logarithmic scoring rule (Section 4.4).

Scoring-rule entropy

The minimum risk associated with a proper scoring rule is called the **scoring-rule entropy** $R(Q) = R(Q, Q)$. With it the properness inequality becomes

$$R(Q, P) \geq R(Q)$$

For a **proper** scoring rule, the entropy $R(Q)$ is **concave** in Q. For a **strictly proper** scoring rule, the entropy $R(Q)$ is **strictly concave**. This means that

$$R(Q_\lambda) \geq (1 - \lambda)R(Q_0) + \lambda R(Q_1)$$

for the mixture $Q_\lambda = (1 - \lambda)Q_0 + \lambda Q_1$ with $0 < \lambda < 1$ and $Q_0 \neq Q_1$ (for strict concavity replace $\geq$ by $>$).

This follows from the fact that the risk $R(Q, P)$ is mixture-preserving in Q. Hence, $R(Q_\lambda) = R(Q_\lambda, Q_\lambda) = (1 - \lambda)R(Q_0, Q_\lambda) + \lambda R(Q_1, Q_\lambda)$. Applying properness $R(Q_i, Q_\lambda) \geq R(Q_i)$ with $i \in \{0, 1\}$ yields concavity.

Scoring-rule divergence

The **scoring-rule divergence** between the distributions Q and P equals the excess risk given by

$$D(Q,P) = R(Q,P) - R(Q)$$

For a **proper** scoring rule, the divergence $D(Q,P) \geq 0$ is always **non-negative** and with $D(Q,P) = 0$ if $P = Q$. For a **strictly proper** scoring rule $D(Q,P) = 0$ only when $P = Q$.

The scoring-rule divergence $D(Q,P)$ is **convex in** Q for fixed P for a **proper** scoring rule. It is **strictly convex in** Q for a **strictly proper** scoring rule. The convexity of $D(Q,P)$ in Q derives from the concavity of $R(Q)$ and the fact that $R(Q,P)$ is mixture-preserving in Q.

Divergences induced by proper scoring rules correspond to **Bregman divergences** applied to probability distributions, with the convex generator being the negative entropy.

Equivalent scoring rules

Equivalent scoring rules

$$S^{\text{equiv}}(x,P) = kS(x,P) + c(x)$$

have associated equivalent divergences

$$D^{\text{equiv}}(Q,P) = R^{\text{equiv}}(Q,P) - R^{\text{equiv}}(Q) = kD(Q,P)$$

Thus, (strictly) proper scoring rules remain (strictly) proper under equivalence transformations. Furthermore, for $k = 1$ equivalent scoring rules are **strongly equivalent** as their divergences are identical.

Further properties

Proper scoring rules also enjoy several additional properties not mentioned above. For example, various **decompositions** exist for their risk, and the scoring-rule divergence satisfies a **generalised Pythagorean theorem**.

4.4 Common scoring rules

Logarithmic scoring rule

The most important scoring rule is the **logarithmic scoring rule** or **log-loss**

$$S_{\log}(x,P) = -\log p(x)$$

The risk of P under Q based on the log-loss is the **mean log-loss**

$$R_{\log}(Q,P) = -\operatorname{E}_Q \log p(x) = H(Q,P)$$

which is uniquely minimised for $P = Q$. Thus, the log-loss is **strictly proper**. Moreover, the log-loss is noted as the only **local** strictly proper scoring rule, as it solely depends on the value of the pdmf at the observed outcome x, and not on any other features of the distribution P.

The mean log-loss is also known as **cross-entropy** denoted by $H(Q,P)$.

The minimum risk (scoring-rule entropy) equals the **information entropy** denoted by $H(Q)$:

$$R_{\log}(Q) = -\operatorname{E}_Q \log q(x) = H(Q)$$

The properness inequality $H(Q,P) \geq H(Q)$, with equality exclusively for $P = Q$ and relating cross-entropy and information entropy, is known as **Gibbs' inequality**.

The divergence induced by the log-loss is the Kullback-Leibler (KL) divergence

$$\begin{aligned} D_{\text{KL}}(Q,P) &= R_{\log}(Q,P) - R_{\log}(Q)) \\ &= H(Q,P) - H(Q) \\ &= \operatorname{E}_Q \log\left(\frac{q(x)}{p(x)}\right) \end{aligned}$$

The KL divergence is **invariant under a change of variables**. Specifically, for arbitrary one-to-one transformations between x and y, even for continuous random variables, $D_{\text{KL}}(Q_x, P_x) = D_{\text{KL}}(Q_y, P_y)$.

More generally, the KL divergence also satisfies the **data processing inequality** (DPI). Applying a transformation (stochastic or deterministic, possibly coarsening) that produces y from x, cannot increase the divergence, so that $D_{\text{KL}}(Q_x, P_x) \geq D_{\text{KL}}(Q_y, P_y)$. Note that a change of variables is a special case of data processing (with identity).

Divergences between distributions satisfying the DPI (and hence being invariant under a change of variables) form the class of f-divergences. The KL divergence is the *only* divergence induced by a proper scoring rule (i.e. the only Bregman divergence) that is also an f-divergence.

The empirical risk of a distribution family $P(\theta)$ based on the log-loss is proportional to the log-likelihood function

$$\begin{aligned} R_{\log}(\hat{Q}_n, P(\theta)) &= -\frac{1}{n}\sum_{i=1}^{n} \log p(x_i|\theta) \\ &= -\frac{1}{n}\ell_n(\theta) \end{aligned}$$

Minimising the empirical risk is thus equivalent to maximising the log-likelihood $\ell_n(\theta)$.

Similarly, minimising the KL divergence $D_{\mathrm{KL}}(\hat{Q}_n, P(\theta))$ with regard to θ is equivalent to minimising the empirical risk and hence to maximum likelihood.

Quadratic or Brier scoring rule

Assume the categorical distributions $Q = \mathrm{Cat}(\boldsymbol{q})$ with class probabilities $\boldsymbol{q} = (q_1, \ldots, q_K)^T$ and $P = \mathrm{Cat}(\boldsymbol{p})$ with corresponding class probabilities $\boldsymbol{p} = (p_1, \ldots, p_K)^T$ and an indicator vector $\boldsymbol{x} = (x_1, \ldots, x_K)^T = (0, 0, \ldots, 1, \ldots, 0)^T$ containing zeros everywhere except for a single element $x_k = 1$.

The **quadratic scoring rule**, also known as **Brier scoring rule**, evaluates the categorical forecast P given a realisation $\boldsymbol{x}$ from the categorical distribution Q using the **squared Euclidean distance** between $\boldsymbol{x}$ and $\boldsymbol{p}$:

$$\begin{aligned} S_{\mathrm{Brier}}(\boldsymbol{x}, P) &= ||\boldsymbol{x} - \boldsymbol{p}||^2 \\ &= (\boldsymbol{x} - \boldsymbol{p})^T(\boldsymbol{x} - \boldsymbol{p}) \\ &= 1 - 2\boldsymbol{x}^T\boldsymbol{p} + \boldsymbol{p}^T\boldsymbol{p} \\ &= 1 - 2p_k + \boldsymbol{p}^T\boldsymbol{p} \end{aligned}$$

Unlike the log-loss, the Brier score is *not local* as the pmf for P is evaluated across all K classes, not just at the realised class k.

The corresponding risk is

$$\begin{aligned} R_{\mathrm{Brier}}(Q, P) &= \mathrm{E}_Q(S(\boldsymbol{x}, P)) \\ &= 1 - 2\boldsymbol{q}^T\boldsymbol{p} + \boldsymbol{p}^T\boldsymbol{p} \end{aligned}$$

which is uniquely minimised for $P = Q$ with $\boldsymbol{p} = \boldsymbol{q}$. Thus, the Brier score is **strictly proper**.

The minimum risk (scoring-rule entropy) is

$$R_{\mathrm{Brier}}(Q) = 1 - \boldsymbol{q}^T\boldsymbol{q}$$

The divergence induced by the Brier score is the **squared Euclidean distance** between $\boldsymbol{q}$ and $\boldsymbol{p}$:

$$\begin{aligned} D_{\mathrm{Brier}}(Q, P) &= R_{\mathrm{Brier}}(Q, P) - R_{\mathrm{Brier}}(Q) \\ &= (\boldsymbol{q} - \boldsymbol{p})^T(\boldsymbol{q} - \boldsymbol{p}) \\ &= ||\boldsymbol{q} - \boldsymbol{p}||^2 \end{aligned}$$

Spherical scoring rule

The **cosine similarity** between two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ is the cosine of the angle ϕ between the two vectors:

$$\mathrm{cos\_sim}(\boldsymbol{a}, \boldsymbol{b}) = \cos\phi(\boldsymbol{a}, \boldsymbol{b}) = \frac{\boldsymbol{a} \cdot \boldsymbol{b}}{||a||\,||b||}$$

with $\boldsymbol{a} \cdot \boldsymbol{b} = \boldsymbol{a}^T\boldsymbol{b}$, $||a|| = (\boldsymbol{a}^T\boldsymbol{a})^{1/2}$ and $||b|| = (\boldsymbol{b}^T\boldsymbol{b})^{1/2}$.

The **cosine distance** is the complement

$$\mathrm{cos\_dist}(\boldsymbol{a}, \boldsymbol{b}) = 1 - \mathrm{cos\_sim}(\boldsymbol{a}, \boldsymbol{b})$$

The **spherical scoring rule** evaluates the categorical forecast $P = \mathrm{Cat}(\boldsymbol{p})$ given a realisation $\boldsymbol{x}$ from the categorical distribution $Q = \mathrm{Cat}(\boldsymbol{q})$ using the **negative cosine similarity** between the probability vectors $\boldsymbol{x}$ and $\boldsymbol{p}$:

$$\begin{aligned} S_{\mathrm{sph}}(\boldsymbol{x}, P) &= -\mathrm{cos\_sim}(\boldsymbol{x}, \boldsymbol{p}) \\ &= -\boldsymbol{x}^T\boldsymbol{p}/||\boldsymbol{p}|| \\ &= -p_k/||\boldsymbol{p}|| \end{aligned}$$

The spherical score is *not local* as the pmf for P is evaluated across all K classes, not just at the realised class k. Depending on $\boldsymbol{p}$, it may range from a minimum of -1 (angle $\phi = 0$, zero degrees) and to a maximum of 0 (angle, $\phi = \pi/2$, 90 degrees).

The corresponding risk is

$$\begin{aligned} R_{\text{sph}}(Q, P) &= \mathrm{E}_Q(S(\boldsymbol{x}, P)) \\ &= -\boldsymbol{q}^T \boldsymbol{p}/||\boldsymbol{p}|| \\ &= -||\boldsymbol{q}||\ \text{cos\_sim}(\boldsymbol{q}, \boldsymbol{p}) \end{aligned}$$

which is proportional to the negative cosine similarity between $\boldsymbol{q}$ and $\boldsymbol{p}$. It is uniquely minimised for $P = Q$ with $\boldsymbol{p} = \boldsymbol{q}$. Therefore, the spherical scoring rule is **strictly proper**.

The minimum risk (scoring-rule entropy) is

$$R_{\text{sph}}(Q) = -||\boldsymbol{q}||$$

The divergence induced by the spherical score is

$$\begin{aligned} D_{\text{sph}}(Q, P) &= R_{\text{sph}}(Q, P) - R_{\text{sph}}(Q) \\ &= -\boldsymbol{q}^T \boldsymbol{p}/||\boldsymbol{p}|| + ||\boldsymbol{q}|| \\ &= ||\boldsymbol{q}||\ \text{cos\_dist}(\boldsymbol{q}, \boldsymbol{p}) \end{aligned}$$

and hence is proportional to the cosine distance between $\boldsymbol{q}$ and $\boldsymbol{p}$.

Proper but not strictly proper scoring rules

An example of a proper, but not strictly proper, scoring rule is the **squared error relative to the mean** of the quoted model P:

$$S_{\text{sq}}(x, P) = (x - \mathrm{E}(P))^2$$

The corresponding risk is

$$\begin{aligned} R_{\text{sq}}(Q, P) &= \mathrm{E}_Q\left((x - \mathrm{E}(P))^2\right) \\ &= (\mathrm{E}(Q) - \mathrm{E}(P))^2 + \mathrm{Var}(Q) \end{aligned}$$

which is minimised at $P = Q$ but also at any distribution P with the same mean as Q.

The minimum risk (scoring-rule entropy) is the variance

$$R_{\text{sq}}(Q) = \mathrm{Var}(Q)$$

The scoring-rule divergence is the squared distance between the two means

$$\begin{aligned} D_{\text{sq}}(Q, P) &= R_{\text{sq}}(Q, P) - R_{\text{sq}}(Q)) \\ &= (\mathrm{E}(Q) - \mathrm{E}(P))^2 \end{aligned}$$

which vanishes at $P = Q$ but also at any P with $\mathrm{E}(P) = \mathrm{E}(Q)$.

The **Dawid-Sebastiani scoring rule** is a related scoring rule given by

$$S_{\text{DS}}(x, P) = \log \mathrm{Var}(P) + \frac{(x - \mathrm{E}(P))^2}{\mathrm{Var}(P)}$$

It is equivalent to the log-loss applied to a normal model P.

The corresponding risk is

$$R_{\text{DS}}(Q, P) = \log \mathrm{Var}(P) + \frac{(\mathrm{E}(Q) - \mathrm{E}(P))^2}{\mathrm{Var}(P)} + \frac{\mathrm{Var}(Q)}{\mathrm{Var}(P)}$$

which is minimised at $P = Q$ but also at any distribution P with $\mathrm{E}(P) = \mathrm{E}(Q)$ and $\mathrm{Var}(P) = \mathrm{Var}(Q)$.

The minimum risk (scoring-rule entropy) is

$$R_{\text{DS}}(Q) = \log \mathrm{Var}(Q) + 1$$

The scoring-rule divergence is

$$\begin{aligned} D_{\text{DS}}(Q, P) &= R_{\text{DS}}(Q, P) - R_{\text{DS}}(Q) \\ &= \frac{(\mathrm{E}(Q) - \mathrm{E}(P))^2}{\mathrm{Var}(P)} + \frac{\mathrm{Var}(Q)}{\mathrm{Var}(P)} - \log\left(\frac{\mathrm{Var}(Q}{\mathrm{Var}(P}\right) - 1 \end{aligned}$$

which vanishes at $P = Q$ but also at any P for which $\mathrm{E}(P) = \mathrm{E}(Q)$ and $\mathrm{Var}(P) = \mathrm{Var}(Q)$.

Other strictly proper scoring rules

Other useful strictly proper scoring rules include:

- the continuous ranked probability score (CRPS),
- the energy score (multivariate CRPS), and
- the Hyvärinen scoring rule.

See also (Wikipedia): scoring rule.

5 Univariate distributions

5.1 Binomial distribution

The **binomial distribution** $\text{Bin}(n, \theta)$ is a discrete distribution counting binary outcomes.

The **Bernoulli distribution** $\text{Ber}(\theta)$ is a special case of the binomial distribution.

Standard parametrisation

A binomial random variable x describes the number of successful outcomes in n identical and independent trials. We write

$$x \sim \text{Bin}(n, \theta)$$

where $\theta \in [0, 1]$ is the probability of a positive outcome ("success") in a single trial. Conversely, $1 - \theta \in [0, 1]$ is the complementary probability ("failure"). The support is $x \in \{0, 1, 2, \ldots, n\}$ which notably depends on n.

The binomial distribution is often motivated by a coin tossing experiment where θ is the probability of "head" when flipping the coin and x is the number of observed "heads" among n throws. Another common interpretation is that of an urn model where n items are distributed into two bins (Figure 5.1). Here θ is the probability to put an item into one urn (representing "success", "head") and $1 - \theta$ the probability to put it in the other urn (representing "failure", "tail").

The expected value is

$$\text{E}(x) = n\theta$$

and the variance is

$$\text{Var}(x) = n\theta(1 - \theta)$$

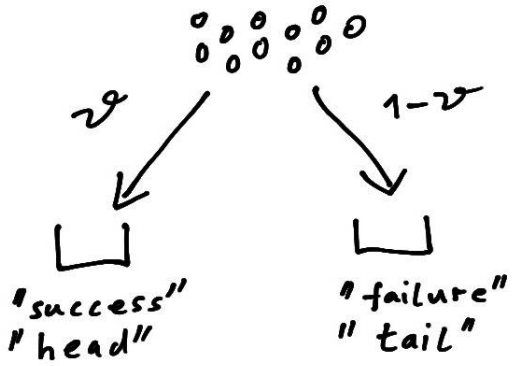

Figure 5.1: Binomial urn model.

The corresponding pmf is

$$p(x|n, \theta) = W_2 \, \theta^x (1 - \theta)^{n-x}$$

The binomial coefficient

$$W_2 = \binom{n}{x}$$

in the pmf accounts for the number of possible permutations of n items of two distinct types ("success" and "failure"). Note that the binomial coefficient W_2 does not depend on θ.

> R code
>
> The pmf of the binomial distribution is given by `dbinom()`, the distribution function is `pbinom()` and the quantile function is `qbinom()`. The corresponding random number generator is `rbinom()`.

Mean parametrisation

Instead of θ one may also use a mean parameter $\mu \in [0, n]$ so that

$$x \sim \text{Bin}\left(n, \theta = \frac{\mu}{n}\right)$$

The mean parameter μ can be obtained from θ and n by $\mu = n\theta$.

The mean and variance of the binomial distribution expressed in terms of μ and n are

$$\mathrm{E}(x) = \mu$$

and

$$\mathrm{Var}(x) = \mu - \frac{\mu^2}{n}$$

Special case: Bernoulli distribution

For $n = 1$ the binomial distribution reduces to the **Bernoulli distribution** $\mathrm{Ber}(\theta)$. This is the simplest of all distribution families and is named after Jacob Bernoulli (1655-1705) who also discovered the law of large numbers.

If a random variable x follows the Bernoulli distribution we write

$$x \sim \mathrm{Ber}(\theta)$$

with "success" probability $\theta \in [0, 1]$. Conversely, the complementary "failure" probability is $1 - \theta \in [0, 1]$. The support is $x \in \{0, 1\}$. The variable x acts as an indicator variable, with "success" indicated by $x = 1$ and "failure" indicated by $x = 0$.

Often the Bernoulli distribution is referred to as "coin flipping" model. Then θ is the probability of "head" and $1 - \theta$ the complementary probability of "tail" and $x = 1$ corresponds to the outcome "head" and $x = 0$ to the outcome "tail".

The expected value is

$$\mathrm{E}(x) = \theta$$

and the variance is

$$\mathrm{Var}(x) = \theta(1 - \theta)$$

The pmf of $\mathrm{Ber}(\theta)$ is

$$p(x|\theta) = \theta^x (1 - \theta)^{1-x} = \begin{cases} \theta & \text{if } x = 1 \\ 1 - \theta & \text{if } x = 0 \end{cases}$$

Convolution property and normal approximation

The convolution of n binomial distributions, each with identical success probability θ but possibly different number of trials n_i, yields another binomial distribution with the same parameter θ:

$$\sum_{i=1}^{n} \mathrm{Bin}(n_i, \theta) \sim \mathrm{Bin}\left(\sum_{i=1}^{n} n_i, \theta\right)$$

It follows that the binomial distribution with n trials is the result of the convolution of n Bernoulli distributions:

$$\sum_{i=1}^{n} \mathrm{Ber}(\theta) \sim \mathrm{Bin}(n, \theta)$$

Thus, repeating the same Bernoulli trial n times and counting the total number of successes yields a binomial random variable.

As a consequence, following the central limit theorem (Section 3.3), for large n the binomial distribution can be well approximated by a normal distribution (Section 5.3) with the same mean and variance. This is known as the De Moivre–Laplace theorem.

5.2 Beta distribution

The **beta distribution** $\mathrm{Beta}(\alpha_1, \alpha_2)$ is a continuous distribution that is useful to model proportions or probabilities for $K = 2$ classes.

It includes the **uniform distribution** over the unit interval as a special case.

Standard parametrisation

A beta-distributed random variable is denoted by

$$x \sim \mathrm{Beta}(\alpha_1, \alpha_2)$$

with shape parameters $\alpha_1 > 0$ and $\alpha_2 > 0$. Let $m = \alpha_1 + \alpha_2$. The support of x is the unit interval given by $x \in [0, 1]$. Thus, the beta distribution is defined over a one-dimensional space.

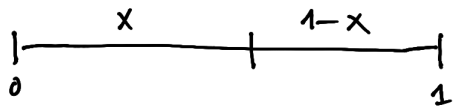

Figure 5.2: Stick breaking visualisation of a beta random variable.

A beta random variable can be visualised as breaking a unit stick of length one into two pieces of length $x_1 = x$ and $x_2 = 1 - x$ (Figure 5.2). Thus, the x_i may be used as the exclusive proportions or probabilities for $K = 2$ classes.

The mean is

$$\mathrm{E}(x) = \mathrm{E}(x_1) = \frac{\alpha_1}{m}$$

and hence

$$\mathrm{E}(1 - x) = \mathrm{E}(x_2) = \frac{\alpha_2}{m}$$

The variance is

$$\mathrm{Var}(x) = \mathrm{Var}(x_1) = \mathrm{Var}(x_2) = \frac{\alpha_1 \alpha_2}{m^2(m+1)}$$

The pdf of the beta distribution $\mathrm{Beta}(\alpha_1, \alpha_2)$ is

$$p(x|\alpha_1, \alpha_2) = \frac{1}{B(\alpha_1, \alpha_2)} x^{\alpha_1 - 1}(1 - x)^{\alpha_2 - 1}$$

In this density the **beta function**

$$B(\alpha_1, \alpha_1) = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1 + \alpha_2)}$$

serves as normalisation factor.

The beta distribution can assume a number of different shapes, depending on the values of α_1 and α_2 (see Figure 5.3).

> R code
>
> The pdf of the beta distribution is given by `dbeta()`, the distribution function is `pbeta()` and the quantile function is `qbeta()`. The corresponding random number generator is `rbeta()`.

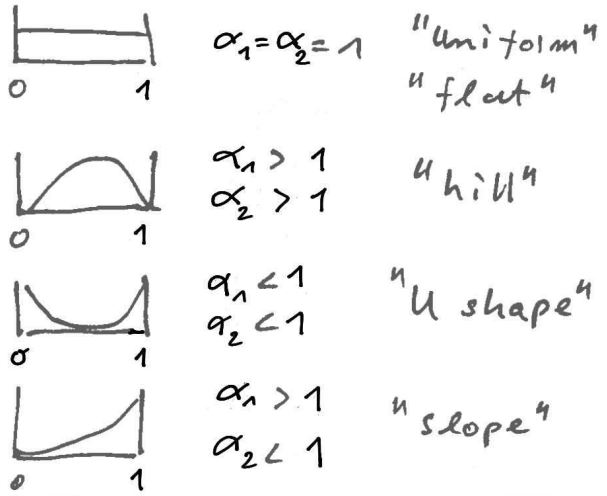

Figure 5.3: Shapes of the pdf of the beta distribution.

Mean parametrisation

Instead of employing α_1 and α_2 as parameters another useful reparametrisation of the beta distribution is in terms of a mean parameter $\mu \in [0, 1]$ and a concentration parameter $m > 0$ so that

$$x \sim \mathrm{Beta}(\alpha_1 = m\mu, \alpha_2 = m(1 - \mu))$$

The concentration and mean parameters can be obtained from α_1 and α_2 by $m = \alpha_1 + \alpha_2$ and $\mu = \alpha_1/m$.

The mean and variance of the beta distribution expressed in terms of μ and m are

$$\mathrm{E}(x) = \mu$$

and

$$\mathrm{Var}(x) = \frac{\mu(1 - \mu)}{m + 1}$$

With increasing concentration parameter m the variance decreases and thus the probability mass becomes more concentrated around the mean.

Special case: symmetric beta distribution

For $\alpha_1 = \alpha_2 = \alpha$ the beta distribution becomes the **symmetric beta distribution** with a single shape parameter $\alpha > 0$. In mean parametrisation the symmetric beta distribution corresponds to $\mu = 1/2$ and $m = 2\alpha$.

Special case: uniform distribution

For $\alpha_1 = \alpha_2 = 1$ the beta distribution becomes the **uniform distribution over the unit interval** with pdf $p(x) = 1$. In mean parametrisation the uniform distribution corresponds to $\mu = 1/2$ and $m = 2$.

5.3 Normal distribution

The **normal distribution** $N(\mu, \sigma^2)$ is the most important continuous probability distribution. It is also called **Gaussian distribution** named after Carl Friedrich Gauss (1777–1855).

Special cases are the **standard normal distribution** $N(0, 1)$ and the **delta distribution** $\delta(\mu)$.

Standard parametrisation

The univariate normal distribution $N(\mu, \sigma^2)$ has two parameters μ (location) and $\sigma^2 > 0$ (variance) and support $x \in \mathbb{R}$.

If a random variable x is normally distributed we write

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

with mean

$$\mathrm{E}(x) = \mu$$

and variance

$$\mathrm{Var}(x) = \sigma^2$$

The pdf is given by

$$\begin{aligned} p(x|\mu, \sigma^2) &= (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \\ &= (\sigma^2)^{-1/2}(2\pi)^{-1/2} e^{-\Delta^2/2} \end{aligned}$$

Here $\Delta^2 = (x-\mu)^2/\sigma^2$ is the squared distance between x and μ weighted by the variance σ^2, also known as **squared Mahalanobis distance**.

The normal distribution is sometimes also used by specifying the precision $1/\sigma^2$ instead of the variance σ^2.

> R code
>
> The normal pdf is given by `dnorm()`, the distribution function is `pnorm()` and the quantile function is `qnorm()`. The corresponding random number generator is `rnorm()`.

Scale parametrisation

Instead of the variance parameter σ^2 it is often also convenient to use the standard deviation $\sigma = \sqrt{\sigma^2} > 0$ as scale parameter. Similarly, instead of the precision $1/\sigma^2$ one may wish to use the inverse standard deviation $w = 1/\sigma$.

The scale parametrisation is central for location-scale transformations (see below).

Special case: standard normal distribution

The **standard normal distribution** $N(0, 1)$ has mean $\mu = 0$ and variance $\sigma^2 = 1$. The corresponding pdf is

$$p(x) = (2\pi)^{-1/2} e^{-x^2/2}$$

with the squared Mahalanobis distance reduced to $\Delta^2 = x^2$.

The cumulative distribution function (cdf) of the standard normal $N(0, 1)$ is

$$\Phi(x) = \int_{-\infty}^{x} p(x'|\mu = 0, \sigma^2 = 1)dx'$$

There is no analytic expression for $\Phi(x)$. The inverse $\Phi^{-1}(p)$ is called the quantile function of the standard normal distribution.

Figure 5.4 shows the pdf and cdf of the standard normal distribution.

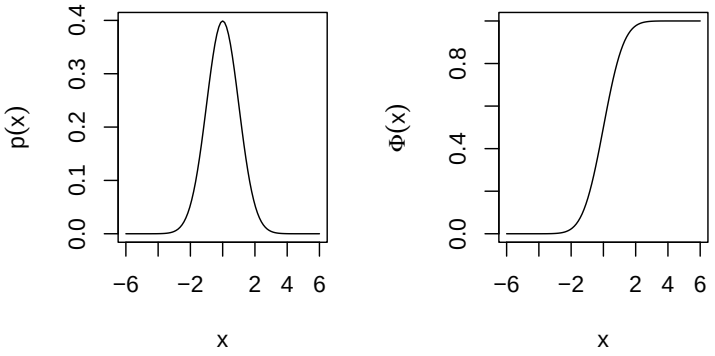

Figure 5.4: Probability density function (left) and cumulative density function (right) of the standard normal distribution.

Special case: delta distribution

The **delta distribution** $\delta(\mu)$ is obtained as the limit of $N(\mu, \varepsilon\sigma^2)$ for $\varepsilon \to 0$ and where σ^2 is a positive number (e.g. $\sigma^2 = 1$). Thus $\delta(\mu)$ is a continuous distribution representing a point mass at μ.

The corresponding pdf $\delta(x|\mu)$ is called the **Dirac delta function**, even though it is not an ordinary function. It satisfies $\delta(x|\mu) = 0$ for all $x \neq \mu$ with an infinite spike at μ but still integrates to one.

Location-scale transformation

Let $\sigma > 0$ be the positive square root of the variance σ^2 and $w = 1/\sigma$.

If $x \sim N(\mu, \sigma^2)$ then $y = w(x - \mu) \sim N(0, 1)$. This location-scale transformation corresponds to centring and standardisation of a normal random variable, reducing it to a standard normal random variable.

Conversely, if $y \sim N(0, 1)$ then $x = \mu + \sigma y \sim N(\mu, \sigma^2)$. This location-scale transformation generates the normal distribution from the standard normal distribution.

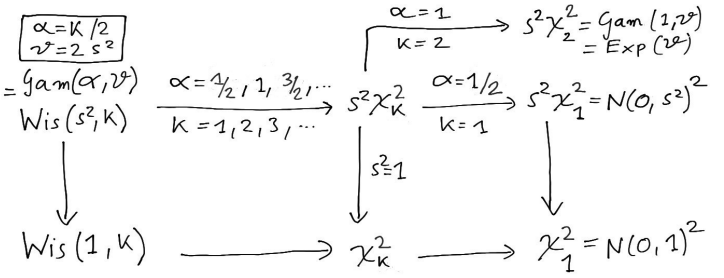

Figure 5.5: The gamma and the univariate Wishart distribution and their relatives.

Convolution property

The convolution of n independent, but not necessarily identical, normal distributions results in another normal distribution with corresponding mean and variance:

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

Hence, any normal random variable can be constructed as the sum of n suitable independent normal random variables.

Since n is an arbitrary positive integer the normal distribution is said to be **infinitely divisible**.

5.4 Gamma distribution

The **gamma distribution** $\text{Gam}(\alpha, \theta)$ is another widely used continuous distribution and is also known as **univariate Wishart distribution** $\text{Wis}\left(s^2, k\right)$ using a different parametrisation.

It contains as special cases the **scaled chi-squared distribution** $s^2\chi^2_k$ (two parameter restrictions) as well as the **univariate standard Wishart distribution** $\text{Wis}\,(1, k)$, the **chi-squared distribution** χ^2_k and the **exponential distribution** $\text{Exp}(\theta)$ (one parameter restrictions). Figure 5.5 illustrates the relationship of the gamma and the univariate Wishart distribution with these related distributions.

Standard parametrisation

The gamma distribution $\text{Gam}(\alpha, \theta)$ is a continuous distribution with two parameters $\alpha > 0$ (shape) and $\theta > 0$ (scale):

$$x \sim \text{Gam}(\alpha, \theta)$$

and support $x \in [0, \infty[$ with mean

$$\text{E}(x) = \alpha\theta$$

and variance

$$\text{Var}(x) = \alpha\theta^2$$

The gamma distribution is also often used with a rate parameter $\beta = 1/\theta$. Therefore one needs to pay attention which parametrisation is used.

The pdf is

$$p(x|\alpha, \theta) = \frac{1}{\Gamma(\alpha)\theta^\alpha} x^{\alpha-1} e^{-x/\theta}$$

In this density the **gamma function**

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$$

is part of the normalisation factor.

The gamma function can also be used as continuous version of the factorial as $\Gamma(x) = (x-1)!$ for any positive integer x.

R code

The pdf of the gamma distribution is available in the function `dgamma()`, the distribution function is `pgamma()` and the quantile function is `qgamma()`. The corresponding random number generator is `rgamma()`.

Wishart parametrisation

The gamma distribution is often used with a different set of parameters $s^2 = \theta/2 > 0$ (scale) and $k = 2\alpha > 0$ (shape or concentration). In this form it is known as **univariate or one-dimensional Wishart distribution**

$$x \sim \text{Wis}\left(s^2, k\right) = \text{Gam}\left(\alpha = \frac{k}{2}, \theta = 2s^2\right)$$

named after John Wishart (1898–1954).

In the above the scale parameter s^2 is a scalar and hence the resulting Wishart distribution is univariate. If instead a matrix-valued scale parameter $\boldsymbol{S}$ is used this yields the multivariate or d-dimensional Wishart distribution, see Section 6.4.

In the Wishart parametrisation the mean is

$$\text{E}(x) = ks^2$$

and the variance

$$\text{Var}(x) = 2ks^4$$

The pdf in terms of s^2 and k is

$$p(x|s^2, k) = \frac{1}{\Gamma(k/2)(2s^2)^{k/2}} x^{(k-2)/2} e^{-s^{-2}x/2}$$

Mean parametrisation

Finally, we also often employ the Wishart resp. gamma distribution in **mean parametrisation**

$$x \sim \text{Wis}\left(s^2 = \frac{\mu}{k}, k\right) = \text{Gam}\left(\alpha = \frac{k}{2}, \theta = \frac{2\mu}{k}\right)$$

with parameters $\mu = ks^2 > 0$ and $k > 0$. In this parametrisation the mean is

$$\text{E}(x) = \mu$$

and the variance

$$\text{Var}(x) = \frac{2\mu^2}{k}$$

Special case: univariate standard Wishart distribution

For $s^2 = 1$ the univariate Wishart distribution reduces to the **univariate standard Wishart distribution**

$$x \sim \text{Wis}\,(1, k) = \text{Gam}\left(\alpha = \frac{k}{2}, \theta = 2\right)$$

with mean

$$\text{E}(x) = k$$

and variance

$$\text{Var}(x) = 2k$$

The pdf is

$$p(x|k) = \frac{1}{\Gamma(k/2)2^{k/2}} x^{(k-2)/2} e^{-x/2}$$

Special case: scaled chi-squared distribution

If the shape parameter of the Wishart distribution Wis (s^2, k) is restricted to the positive integers $k \in \{1, 2, \ldots\}$ the Wishart distribution becomes the **scaled chi-squared distribution** $s^2\chi^2_k$ where k is called the *degree of freedom*.

This is equivalent to restricting the shape parameter α of the gamma distribution $\text{Gam}(\alpha = k/2, \theta = 2s^2)$ to $\alpha \in \{1/2, 1, 3/2, 2, \ldots\}$.

The scaled chi-squared distribution with $k = 1$ is the distribution of a squared normal random variable with mean zero. Specifically, if $z \sim N(0, s^2)$ then $z^2 \sim s^2\chi^2_1 = \text{Wis}(s^2, 1) = N(0, s^2)^2$.

Special case: chi-squared distribution

If $k \in \{1, 2, \ldots\}$ is restricted to the positive integers the univariate standard Wishart distribution reduces to the **chi-squared distribution**

$$x \sim \chi^2_k = \text{Wis}\left(s^2 = 1, k\right) = \text{Gam}\left(\alpha = \frac{k}{2}, \theta = 2\right)$$

where k is called the degree of freedom.

The chi-squared distribution has mean

$$\text{E}(x) = k$$

and variance

$$\text{Var}(x) = 2k$$

Figure 5.6 shows the pdf of the chi-squared distribution for degrees of freedom $k = 1$ and $k = 3$.

The chi-squared distribution with $k = 1$ is the distribution of a squared standard normal random variable. Specifically, if $z \sim N(0, 1)$ then $z^2 \sim \chi^2_1 = \text{Wis}(1, 1) = N(0, 1)^2$.

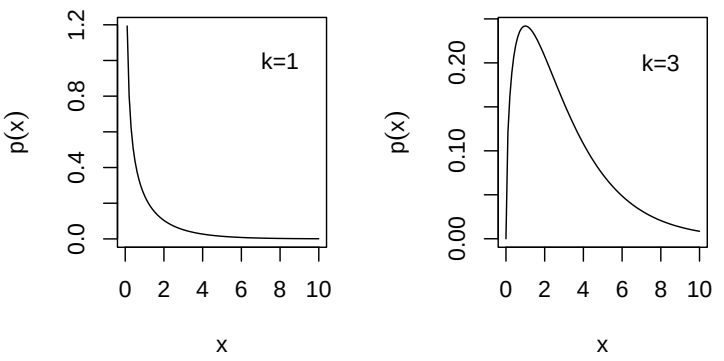

Figure 5.6: Probability density function of the chi-squared distribution.

R code

The pdf of the chi-squared distribution is given by `dchisq()`. The distribution function is `pchisq()` and the quantile function is `qchisq()`. The corresponding random number generator is `rchisq()`.

Special case: exponential distribution

If the shape parameter α of the gamma distribution $\text{Gam}(\alpha, \theta)$ is set to $\alpha = 1$, or if the shape parameter k of the Wishart distribution $\text{Wis}(s^2, k)$ is set to $k = 2$, we obtain the **exponential distribution**

$$x \sim \text{Exp}(\theta) = \text{Gam}(\alpha = 1, \theta) = \text{Wis}(s^2 = \theta/2, k = 2)$$

with scale parameter θ.

It has mean

$$\text{E}(x) = \theta$$

and variance

$$\text{Var}(x) = \theta^2$$

and the pdf is

$$p(x|\theta) = \theta^{-1} e^{-x/\theta}$$

Just like the gamma distribution the exponential distribution is also often specified using a rate parameter $\beta = 1/\theta$ instead of a scale parameter θ.

R code

The command `dexp()` returns the pdf of the exponential distribution, `pexp()` is the distribution function and `qexp()` is the quantile function. The corresponding random number generator is `rexp()`.

Scale transformation

If $x \sim \text{Gam}(\alpha, \theta)$ then the scaled random variable bx with $b > 0$ is also gamma distributed with $bx \sim \text{Gam}(\alpha, b\theta)$.

Hence,

- θ Gam$(\alpha, 1)$ = Gam(α, θ),
- θ Exp(1) = Exp(θ),
- (μ/k) Wis$(1, k)$ = Wis$(s^2 = \mu/k, k)$ and
- s^2 Wis$(1, k)$ = Wis(s^2, k).

As χ^2_k equals Wis$(1, k)$ the last example demonstrates that the **scaled chi-squared distribution** $s^2\chi^2_k$ equals the univariate Wishart distribution Wis(s^2, k).

Convolution property

The convolution of n gamma distributions with the same scale parameter θ but possible different shape parameters α_i yields another gamma distribution:

$$\sum_{i=1}^{n} \text{Gam}(\alpha_i, \theta) \sim \text{Gam}\left(\sum_{i=1}^{n} \alpha_i, \theta\right)$$

Thus, any gamma random variable can be obtained as the sum of n suitable independent gamma random variables.

In Wishart parametrisation this becomes

$$\sum_{i=1}^{n} \text{Wis}(s^2, k_i) \sim \text{Wis}\left(s^2, \sum_{i=1}^{n} k_i\right)$$

As a result, since n is an arbitrary positive integer, the gamma resp. univariate Wishart distribution is **infinitely divisible**.

The above includes the following two specific constructions:

a) If $x_1, \ldots, x_n \sim \text{Exp}(\theta)$ are independent samples from Exp(θ) then the sum $y = \sum_{i=1}^{n} x_i \sim \text{Gam}(\alpha = n, \theta)$ is gamma distributed with the same scale parameter.

b) The sum of k independent scaled chi-squared random variables $s^2\chi^2_1$ with one degree of freedom and identical scale parameter s^2 yields a scaled chi-squared random variable $s^2\chi^2_k$ with degree of freedom k and the same scale parameter. Thus, if $z_1, z_2, \ldots, z_k \sim N(0, 1)$ are k independent samples from $N(0, 1)$ then $\sum_{i=1}^{k} z_i^2 \sim \chi^2_k$.

5.5 Inverse gamma distribution

The **inverse gamma distribution** IGam(α, β) is a continuous distribution and is also known as **univariate inverse Wishart distribution** IWis(ψ, k) using a different parametrisation. It is linked to the gamma distribution Gam(α, θ) aka univariate Wishart distribution Wis $\left(s^2, k\right)$ (Section 5.4).

Special cases include the **inverse chi-squared distribution** χ^{-2}_k and the **scaled inverse chi-squared distribution** $s^2\chi^{-2}_k$.

Standard parametrisation

A random variable x following an **inverse gamma distribution** is denoted by

$$x \sim \text{IGam}(\alpha, \beta)$$

with two parameters $\alpha > 0$ (shape parameter) and $\beta > 0$ (scale parameter) and support $x > 0$.

The mean of the inverse gamma distribution is (for $\alpha > 1$)

$$\text{E}(x) = \frac{\beta}{\alpha - 1}$$

and the variance (for $\alpha > 2$)

$$\text{Var}(x) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$$

The inverse gamma distribution $\text{IGam}(\alpha, \beta)$ has pdf

$$p(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} e^{-\beta/x}$$

R code

The `extraDistr` package implements the inverse gamma distribution. The function `extraDistr::dinvgamma()` provides the pdf, `extraDistr::pinvgamma()` the distribution function and `extraDistr::qinvgamma()` is the quantile function. The corresponding random number generator is `extraDistr::rinvgamma()`.

Relation to the gamma distribution

The inverse gamma distribution is closely linked to the gamma distribution. Assume that the random variable y follows a gamma distribution with

$$y \sim \text{Gam}(\alpha, \theta)$$

then the inverse random variable $x = 1/y$ follows an inverse gamma distribution with inverted scale parameter

$$x = \frac{1}{y} \sim \text{IGam}\left(\alpha, \beta = \frac{1}{\theta}\right)$$

where α is the shared shape parameter, θ the scale parameter of the gamma distribution and β the scale parameter of the inverse gamma distribution.

Correspondingly, the density $p_x(x|\alpha,\beta)$ of the inverse gamma distribution is obtained from the density $p_y(y|\alpha,\theta)$ of the gamma distribution via

$$p_x(x|\alpha,\beta) = \frac{1}{x^2}\, p_y\left(\frac{1}{x}\middle|\, \alpha, \theta = \frac{1}{\beta}\right)$$

Wishart parametrisation

The inverse gamma distribution is frequently used with a different set of parameters $\psi = 2\beta$ (scale parameter) and $k = 2\alpha$ (shape parameter). In this form it is called **univariate inverse Wishart distribution**

$$x \sim \text{IWis}(\psi, k) = \text{IGam}\left(\alpha = \frac{k}{2}, \beta = \frac{\psi}{2}\right)$$

In the above the scale parameter ψ is scalar and hence the resulting inverse Wishart distribution is univariate. If instead a matrix-valued scale parameter $\mathbf{\Psi}$ is used this yields the multivariate or d-dimensional inverse Wishart distribution, see Section 6.5.

In the Wishart parametrisation the mean is (for $k > 2$)

$$\text{E}(x) = \frac{\psi}{k-2}$$

and the variance is (for $k > 4$)

$$\text{Var}(x) = \frac{2\psi^2}{(k-4)(k-2)^2}$$

The pdf in terms of ψ and k is

$$p(x|\psi,k) = \frac{(\psi/2)^{(k/2)}}{\Gamma(k/2)}\, x^{-(k+2)/2}\, e^{-\psi x^{-1}/2}$$

Relation to the univariate Wishart distribution

The univariate inverse Wishart distribution is closely linked to the univariate Wishart distribution. Assume that the random variable y follows an univariate Wishart distribution with

$$y \sim \text{Wis}(s^2, k)$$

then the inverse random variable $x = 1/y$ follows a univariate inverse Wishart distribution with inverted scale parameter

$$x = \frac{1}{y} \sim \text{IWis}\left(\psi = \frac{1}{s^2}, k\right)$$

where k is the shared shape parameter, s^2 the scale parameter of the univariate Wishart distribution and ψ the scale parameter of the univariate inverse Wishart distribution.

Correspondingly, the density $p_x(x|\psi, k)$ of the univariate inverse Wishart distribution is obtained from the density $p_y(y|s^2, k)$ of the univariate Wishart distribution via

$$p_x(x|\psi, k) = \frac{1}{x^2}\, p_y\left(\frac{1}{x}\middle| s^2 = \frac{1}{\psi}, k\right)$$

Mean parametrisation

Instead of ψ and k we may also equivalently use $\mu = \psi/(\nu - 2)$ and $\kappa = \nu - 2$ as parameters for the univariate inverse Wishart distribution, so that

$$x \sim \text{IWis}(\psi = \kappa\mu, k = \kappa + 2) = \text{IGam}\left(\alpha = \frac{\kappa + 2}{2}, \beta = \frac{\mu\kappa}{2}\right)$$

has mean (for $\kappa > 0$)

$$\text{E}(x) = \mu$$

and the variance (for $\kappa > 2$)

$$\text{Var}(x) = \frac{2\mu^2}{\kappa - 2}$$

The **mean parametrisation** is useful in Bayesian analysis when employing the inverse gamma aka univariate inverse Wishart distribution as prior and posterior distribution.

Biased mean parametrisation

Using $\tau^2 = \frac{\psi}{k}$ as biased mean parameter together with $\nu = k$ we arrive at the **biased mean parametrisation**

$$x \sim \text{IWis}(\psi = \nu\tau^2, k = \nu) = \text{IGam}\left(\alpha = \frac{\nu}{2}, \beta = \frac{\nu\tau^2}{2}\right)$$

with mean (for $\nu > 2$)

$$\text{E}(x) = \frac{\nu}{\nu - 2}\tau^2 = \mu$$

and variance ($\nu > 4$)

$$\text{Var}(x) = \left(\frac{\nu}{\nu - 2}\right)^2 \frac{2\tau^4}{\nu - 4}$$

As $\tau^2 = \mu(\nu - 2)/\nu$ for large ν the parameter τ^2 will become identical to the true mean μ.

This parametrisation is useful to derive the location-scale t-distribution with its matching parameters (see Section 5.6). It is also common in Bayesian analysis.

Special case: inverse chi-squared distribution

If the scale parameter in $\text{IWis}(\psi, k)$ is set to $\psi = 1$ and $k \in \{1, 2, \ldots\}$ is restricted to the positive integers the univariate inverse Wishart distribution reduces to the **inverse chi-squared distribution**

$$x \sim \chi_k^{-2} = \text{IWis}(\psi = 1, k) = \text{IGam}\left(\alpha = \frac{k}{2}, \beta = \frac{1}{2}\right)$$

where k is called the degree of freedom.

The inverse chi-squared distribution has mean (for $k > 2$)

$$\text{E}(x) = \frac{1}{k - 2}$$

and the variance is (for $k > 4$)

$$\text{Var}(x) = \frac{2}{(k - 2)^2(k - 4)}$$

Relation to the chi-squared distribution

The inverse chi-squared distribution is closely linked to the chi-squared distribution. Assume that the random variable y follows a chi-squared distribution with

$$y \sim \chi_k^2$$

then the inverse random variable $x = 1/y$ follows an inverse chi-squared distribution

$$x = \frac{1}{y} \sim \chi_k^{-2}$$

where k is the shared degree of freedom (shape parameter).

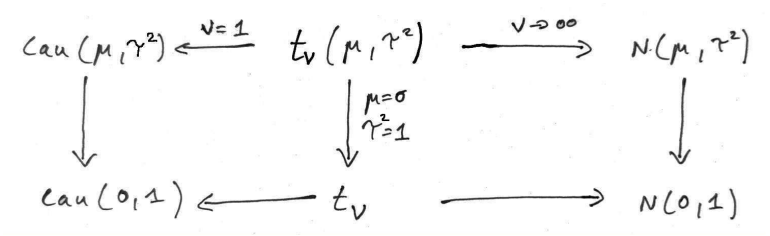

Figure 5.7: The location-scale t-distribution and its relatives.

Scale transformation

If $x \sim \text{IGam}(\alpha, \beta)$ then the scaled random variable bx with $b > 0$ is also inverse gamma distributed with $bx \sim \text{IGam}(\alpha, b\beta)$.

Hence,

- β IGam$(\alpha, 1) = \text{IGam}(\alpha, \beta)$,
- ψ IWis$(1, k) = \text{IWis}(\psi, k)$,
- $\kappa\mu$ IWis$(1, k = \kappa + 2) = \text{IWis}(\psi = \kappa\mu, k = \kappa + 2)$ and
- $\nu\tau^2$ IWis$(1, k = \nu) = \text{IWis}(\psi = \nu\tau^2, k = \nu)$

As χ_k^{-2} equals IWis$(\psi = 1, k)$ the **scaled inverse chi-squared distribution** $\psi\chi_k^{-2}$ is thus equivalent to the univariate inverse Wishart distribution IWis(ψ, k). If a random variable y follows the scaled chi-squared distribution its inverse $y = 1/y$ follows the corresponding scaled inverse chi-squared distribution. Specifically, if $y \sim s^2\,\chi_k^2$ then $x = 1/y \sim s^{-2}\,\chi_k^{-2}$.

The scaled inverse chi-squared distribution is frequently used in the biased mean parametrisation with $\tau = \psi/\nu$ and $\nu = k$. Then $\psi\,\chi_k^{-2}$ is equal to $\nu\tau^2\,\chi_\nu^{-2} = \text{IWis}(\psi = \nu\tau^2, k = \nu)$ which is sometimes also written as $\chi^{-2}(\nu, \tau^2)$.

5.6 Location-scale t-distribution

The **location-scale t-distribution** $t_\nu(\mu, \tau^2)$ is a continuous distribution and is a generalisation of the normal distribution $N(\mu, \tau^2)$ (Section 5.3) with an additional parameter $\nu > 0$ (degrees of freedom) controlling the probability mass in the tails.

Special cases include the **Student's t-distribution** t_ν, the **normal distribution** $N(\mu, \tau^2)$ and the **Cauchy distribution** Cau(μ, τ^2). Figure 5.7 illustrates the relationship of the location-scale t-distribution $t_\nu(\mu, \tau^2)$ with these related distributions.

Standard parametrisation

If a random variable $x \in \mathbb{R}$ follows the **location-scale t-distribution** we write

$$x \sim t_\nu(\mu, \tau^2)$$

where μ is the location and τ^2 the dispersion parameter. The parameter $\nu > 0$ prescribes the degrees of freedom. For small values of ν the distribution is heavy-tailed and as a result only moments of order smaller than ν are finite and defined.

The mean is (for $\nu > 1$)

$$\text{E}(x) = \mu$$

and the variance (for $\nu > 2$)

$$\text{Var}(x) = \frac{\nu}{\nu - 2}\tau^2$$

The pdf of $t_\nu(\mu, \tau^2)$ is

$$p(x|\mu, \tau^2, \nu) = (\tau^2)^{-1/2}\frac{\Gamma(\frac{\nu+1}{2})}{(\pi\nu)^{1/2}\,\Gamma(\frac{\nu}{2})}\left(1 + \frac{\Delta^2}{\nu}\right)^{-(\nu+1)/2}$$

with $\Delta^2 = (x - \mu)^2/\tau^2$ the squared Mahalanobis distance between x and μ.

> **R code**
>
> The package `extraDistr` implements the location-scale t-distribution. The function `extraDistr::dlst()` returns the pdf, `extraDistr::plst()` the distribution function function and `extraDistr::qlst()` is the quantile function. The corresponding random number generator is `extraDistr::rlst()`.

Scale parametrisation

Instead of the dispersion parameter τ^2 it is often also convenient to use the scale parameter $\tau = \sqrt{\tau^2} > 0$. Similarly, instead of the inverse dispersion $1/\tau^2$ one may wish to use the inverse scale $w = 1/\tau$.

The scale parametrisation is central for location-scale transformations (see below).

Special case: Student's t-distribution

With $\mu = 0$ and $\tau^2 = 1$ the location-scale t-distribution reduces to the **standard t-distribution** $t_\nu = t_\nu(0, 1)$. It is commonly known **Student's t-distribution** named after "Student" which was the pseudonym of William Sealy Gosset (1876–1937). It is a generalisation of the standard normal distribution $N(0, 1)$ to allow for heavy tails.

The distribution has mean $\mathrm{E}(x) = 0$ (for $\nu > 1$) and variance $\mathrm{Var}(x) = \frac{\nu}{\nu-2}$ (for $\nu > 2$).

The pdf of t_ν is

$$p(x|\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{(\pi\nu)^{1/2}\,\Gamma(\frac{\nu}{2})} \left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2}$$

with the squared Mahalanobis distance reducing to $\Delta^2 = x^2$.

> R code
>
> The command `dt()` returns the pdf of the t-distribution, `pt()` is distribution function and `qt()` the quantile function. The corresponding random number generator is `rt()`.

Special case: normal distribution

For $\nu \to \infty$ the location-scale t-distribution $t_\nu(\mu, \tau^2)$ reduces to the **normal distribution** $N(\mu, \tau^2)$ (Section 5.3). Correspondingly, for $\nu \to \infty$ the Student's t-distribution becomes equal to the **standard normal distribution** $N(0, 1)$.

See Section 6.6 for further details.

Special case: Cauchy distribution

For $\nu = 1$ the location-scale t-distribution becomes the **Cauchy distribution** $\mathrm{Cau}(\mu, \tau^2) = t_1(\mu, \tau^2)$ named after Augustin-Louis Cauchy (1789–1857).

Its mean, variance and other higher moments are all undefined.

It has pdf

$$\begin{aligned} p(x|\mu, \tau^2) &= (\tau^2)^{-1/2}(\pi(1 + \Delta^2))^{-1} \\ &= \frac{\tau}{\pi(\tau^2 + (x - \mu)^2)} \end{aligned}$$

with $\tau = \sqrt{\tau^2} > 0$.

Note that in the above we employ τ^2 as dispersion parameter as this parallels the location-scale t-distribution and the normal distribution but very often the Cauchy distribution is used with $\tau > 0$ as scale parameter.

> R code
>
> The command `dcauchy()` returns the pdf of the Cauchy distribution, `pcauchy()` is the distribution function and `qcauchy()` the quantile function. The corresponding random number generator is `rcauchy()`.

Special case: standard Cauchy distribution

The **standard Cauchy distribution** $\mathrm{Cau}(0, 1) = t_1(0, 1) = t_1$ is obtained by setting $\mu = 0$ and $\tau^2 = 1$ (Cauchy distribution) or, equivalently, by setting $\nu = 1$ (Student's t-distribution).

It has pdf

$$p(x) = \frac{1}{\pi(1 + x^2)}$$

Location-scale transformation

Let $\tau > 0$ be the positive square root of τ^2 and $w = 1/\tau$.

If $x \sim t_\nu(\mu, \tau^2)$ then $y = w(x - \mu) \sim t_\nu$. This location-scale transformation reduces a location-scale t-distributed random variable to a Student's t-distributed random variable.

Conversely, if $y \sim t_\nu$ then $x = \mu + \tau y \sim t_\nu(\mu, \tau^2)$. This location-scale transformation generates the location-scale t-distribution from the Student's t -distribution.

For the special case of the Cauchy distribution (corresponding to $\nu = 1$) similar relations hold between it and the standard Cauchy distribution. If $x \sim \text{Cau}(\mu, \tau^2)$ then $y = w(x - \mu) \sim \text{Cau}(0, 1)$. Conversely, if $y \sim \text{Cau}(0, 1)$ then $x = \mu + \tau y \sim \text{Cau}(\mu, \tau^2)$.

Convolution property

The location-scale t-distribution is not generally closed under convolution, with the exception of two special cases, the normal distribution ($\nu = \infty$), see Section 5.3, and the Cauchy distribution ($\nu = 1$).

For the Cauchy distribution with $\tau_i^2 = a_i^2\tau^2$, where $a_i > 0$ are positive scalars,

$$\sum_{i=1}^{n} \text{Cau}(\mu_i, a_i^2\tau^2) \sim \text{Cau}\left(\sum_{i=1}^{n} \mu_i, \left(\sum_{i=1}^{n} a_i\right)^2 \tau^2\right)$$

Location-scale t-distribution as compound distribution

The location-scale t-distribution can be obtained as mixture of normal distributions with identical mean and varying variance. Specifically, let z be a univariate inverse Wishart random variable

$$z \sim \text{IWis}(\psi = \nu, k = \nu) = \text{IGam}\left(\alpha = \frac{\nu}{2}, \beta = \frac{\nu}{2}\right)$$

and let $x|z$ be normal

$$x|z \sim N(\mu, \sigma^2 = z\tau^2)$$

Then the resulting marginal (scale mixture) distribution for x is the location-scale t-distribution

$$x \sim t_\nu\left(\mu, \tau^2\right)$$

An alternative way to arrive at $t_\nu\left(\mu, \tau^2\right)$ is to include τ^2 as parameter in the inverse Wishart distribution

$$z \sim \tau^2\,\text{IWis}(\psi = \nu, k = \nu) = \text{IWis}(\psi = \nu\tau^2, k = \nu)$$

and let

$$x|z \sim N(\mu, \sigma^2 = z)$$

Note that τ^2 is now the biased mean parameter of the univariate inverse Wishart distribution. This characterisation is useful in Bayesian analysis.

6 Multivariate distributions

6.1 Multinomial distribution

The **multinomial distribution** $\text{Mult}(n, \boldsymbol{\theta})$ is the multivariate generalisation of the binomial distribution $\text{Bin}(n, \theta)$ (Section 5.1) from two classes to K classes.

A special case is the **categorical distribution** $\text{Cat}(\boldsymbol{\theta})$ that generalises the Bernoulli distribution $\text{Ber}(\theta)$.

Standard parametrisation

A multinomial random variable $\boldsymbol{x}$ describes describes the allocation of n items to K classes. We write

$$\boldsymbol{x} \sim \text{Mult}(n, \boldsymbol{\theta})$$

where the parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)^T$ specifies the probability of each of K classes, with $\theta_k \in [0, 1]$ and $\boldsymbol{\theta}^T\mathbf{1}_K = \sum_{k=1}^{K} \theta_k = 1$. Thus there are $K - 1$ independent elements in $\boldsymbol{\theta}$. The number of classes K is implicitly given by the dimension of the vector $\boldsymbol{\theta}$. Each element of the vector $\boldsymbol{x} = (x_1, \ldots, x_K)^T$ is an integer $x_k \in \{0, 1, \ldots, n\}$ and $\boldsymbol{x}$ satisfies the constraint $\boldsymbol{x}^T\mathbf{1}_K = \sum_{k=1}^{K} x_k = n$. Therefore the support of $\boldsymbol{x}$ is a $K - 1$ dimensional space and it notably depends on n.

The multinomial distribution is best illustrated by an urn model distributing n items into K bins where $\boldsymbol{\theta}$ contains the corresponding bin probabilities (Figure 6.1).

The expected value is

$$\text{E}(\boldsymbol{x}) = n\boldsymbol{\theta}$$

The covariance matrix is

$$\text{Var}(\boldsymbol{x}) = n(\text{Diag}(\boldsymbol{\theta}) - \boldsymbol{\theta}\boldsymbol{\theta}^T)$$

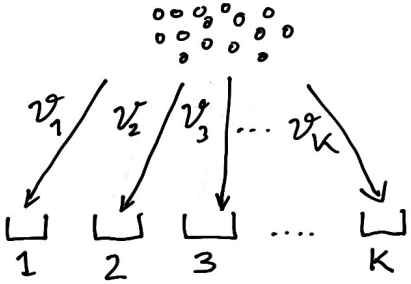

Figure 6.1: Multinomial urn model.

The covariance matrix is singular by construction because of the dependencies among the elements of $\boldsymbol{x}$.

The corresponding pmf is

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = W_K \prod_{k=1}^{K} \theta_k^{x_k}$$

The multinomial coefficient

$$W_K = \binom{n}{x_1, \ldots, x_K}$$

in the pmf accounts for the number of possible permutation of n items of K distinct types. Note that the multinomial coefficient W_K does not depend on $\boldsymbol{\theta}$.

While all K elements of $\boldsymbol{x}$ appear in the pmf recall that due the dependencies among the x_k the pmf is defined over a $K-1$ dimensional support.

For $K = 2$ the multinomial distribution reduces to the binomial distribution (Section 5.1).

R code

The pmf of the multinomial distribution is given by `dmultinom()`. The corresponding random number generator is `rmultinom()`.

Mean parametrisation

Instead of $\boldsymbol{\theta}$ one may also use a mean parameter $\boldsymbol{\mu}$, with elements $\mu_k \in [0, n]$ and $\boldsymbol{\mu}^T \mathbf{1}_K = \sum_{k=1}^{K} \mu_k = n$, so that

$$\boldsymbol{x} \sim \text{Mult}\left(n, \boldsymbol{\theta} = \frac{\boldsymbol{\mu}}{n}\right)$$

The mean parameter $\boldsymbol{\mu}$ can be obtained from $\boldsymbol{\theta}$ and n by $\boldsymbol{\mu} = n\boldsymbol{\theta}$. Note that the parameter space for $\boldsymbol{\mu}$ and the support of $\boldsymbol{x}$ are both of dimension $K-1$.

The mean and variance of the multinomial distribution expressed in terms of $\boldsymbol{\mu}$ and n are

$$\text{E}(x) = \boldsymbol{\mu}$$

and

$$\text{Var}(x) = \text{Diag}(\boldsymbol{\mu}) - \frac{\boldsymbol{\mu}\boldsymbol{\mu}^T}{n}$$

The covariance matrix is singular by construction because of the dependencies among the elements of $\boldsymbol{x}$.

Special case: categorical distribution

For $n = 1$ the multinomial distribution reduces to the **categorical distribution** $\text{Cat}(\boldsymbol{\theta})$ which in turn is the multivariate generalisation of the Bernoulli distribution $\text{Ber}(\theta)$ from two classes to K classes.

If a random variable $\boldsymbol{x}$ follows the categorical distribution we write

$$\boldsymbol{x} \sim \text{Cat}(\boldsymbol{\theta})$$

with class probabilities $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^T \mathbf{1}_K = 1$. The support is $x_k \in \{0, 1\}$ and $\boldsymbol{x}^T \mathbf{1}_K = 1$ and is a $K-1$ dimensional space.

The random vector $\boldsymbol{x}$ takes the form of an indicator vector $\boldsymbol{x} = (x_1, \ldots, x_K)^T = (0, 0, \ldots, 1, \ldots, 0)^T$ containing zeros everywhere except for a single element indicating the class to which the item has been allocated. This is called "one hot encoding", as opposed to "integer encoding" (i.e. stating the class number directly).

The expected value is

$$\text{E}(\boldsymbol{x}) = \boldsymbol{\theta}$$

The covariance matrix is

$$\text{Var}(\boldsymbol{x}) = \text{Diag}(\boldsymbol{\theta}) - \boldsymbol{\theta}\boldsymbol{\theta}^T$$

The covariance matrix is singular because of the dependencies among the elements of $\boldsymbol{x}$.

The corresponding pmf is

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_{k=1}^{K} \theta_k^{x_k} = \begin{cases} \theta_k & \text{if } x_k = 1 \end{cases}$$

Recall that the pmf is defined over the $K-1$ dimensional support of $\boldsymbol{x}$.

For $K = 2$ the categorical distribution reduces to the Bernoulli $\text{Ber}(\theta)$ distribution, with $\theta_1 = \theta$, $\theta_2 = 1 - \theta$ and $x_1 = x$ and $x_2 = 1 - x$.

Convolution property

The convolution of n multinomial distributions, each with identical bin probabilities $\boldsymbol{\theta}$ but possibly different number of items n_i, yields another multinomial distribution with the same parameter $\boldsymbol{\theta}$:

$$\sum_{i=1}^{n} \text{Mult}(n_i, \boldsymbol{\theta}) \sim \text{Mult}\left(\sum_{i=1}^{n} n_i, \boldsymbol{\theta}\right)$$

It follows that the multinomial distribution with n items is the result of the convolution of n categorical distributions:

$$\sum_{i=1}^{n} \text{Cat}(\boldsymbol{\theta}) \sim \text{Mult}(n, \boldsymbol{\theta})$$

Thus, repeating the same categorical trial n times and counting the total number of allocations in each bin yields a multinomial random variable.

6.2 Dirichlet distribution

The **Dirichlet distribution** $\text{Dir}(\boldsymbol{\alpha})$ is the multivariate generalisation of the beta distribution $\text{Beta}(\alpha_1, \alpha_2)$ (Section 5.2) that is useful to model proportions or probabilities for $K \geq 2$ classes. It is named after Peter Gustav Lejeune Dirichlet (1805–1859).

It includes the **uniform distribution** over the $K-1$ unit simplex as special case.

Standard parametrisation

A Dirichlet distributed random vector is denoted by

$$\boldsymbol{x} \sim \text{Dir}(\boldsymbol{\alpha})$$

with shape parameter $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_K)^T > 0$ and $K \geq 2$. Let $m = \boldsymbol{\alpha}^T \mathbf{1}_K = \sum_{k=1}^{K} \alpha_k$. The support of $\boldsymbol{x}$ is the $K-1$ dimensional unit simplex given by $x_k \in [0, 1]$ and $\boldsymbol{x}^T \mathbf{1}_K = \sum_{k=1}^{K} x_k = 1$. Thus, the Dirichlet distribution is defined over a $K-1$ dimensional space.

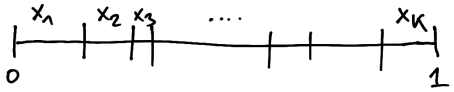

Figure 6.2: Stick breaking visualisation of a Dirichlet random variable.

A Dirichlet random variable can be visualised as breaking a unit stick into K individual pieces of lengths $x_1, x_2, \ldots, x_K$ adding up to one (Figure 6.2). Thus, the x_k may be used as the exclusive proportions or probabilities for K classes.

The mean is

$$\text{E}(\boldsymbol{x}) = \frac{\boldsymbol{\alpha}}{m}$$

and the variance is

$$\text{Var}(\boldsymbol{x}) = \frac{m\,\text{Diag}(\boldsymbol{\alpha}) - \boldsymbol{\alpha}\boldsymbol{\alpha}^T}{m^2(m+1)}$$

The covariance matrix is singular by construction because of the dependencies among the elements of $\boldsymbol{x}$. In component notation it is

$$\text{Cov}(x_i, x_j) = \frac{[i = j]m\alpha_i - \alpha_i\alpha_j}{m^2(m+1)}$$

where the indicator function $[i = j]$ equals 1 if $i = j$ and 0 otherwise.

The pdf of the Dirichlet distribution $\text{Dir}(\boldsymbol{\alpha})$ is

$$p(\boldsymbol{x}|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} x_k^{\alpha_k - 1}$$

In this density the normalisation factor is given by the **multivariate beta function**

$$B(\boldsymbol{\alpha}) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K} \alpha_k)}$$

For $K = 2$ it reduces to the conventional beta function $B(\alpha_1, \alpha_2)$.

While all K elements of $\boldsymbol{x}$ appear in the pdf recall that due the dependencies among the x_k the pdf is defined over a $K - 1$ dimensional support.

For $K = 2$ the Dirichlet distribution reduces to the beta distribution (Section 5.2).

R code

The `extraDistr` package implements the Dirichlet distribution. The pmf of the Dirichlet distribution is given by `extraDistr::ddirichlet()`. The corresponding random number generator is `extraDistr::rdirichlet()`.

Mean parametrisation

Instead of employing $\boldsymbol{\alpha}$ as parameter vector another useful reparametrisation of the Dirichlet distribution is in terms of a mean parameter $\boldsymbol{\mu}$, with elements $\mu_k \in [0, 1]$ and $\boldsymbol{\mu}^T \mathbf{1}_K = \sum_{k=1}^{K} \mu_k = 1$, and a concentration parameter $m > 0$ so that

$$\boldsymbol{x} \sim \text{Dir}(\boldsymbol{\alpha} = m\boldsymbol{\mu})$$

The concentration and mean parameters can be obtained from $\boldsymbol{\alpha}$ by $m = \boldsymbol{\alpha}^T \mathbf{1}_K$ and $\boldsymbol{\mu} = \boldsymbol{\alpha}/m$. The space of possible values for the mean parameter $\boldsymbol{\mu}$ and the support of $\boldsymbol{x}$ are both of dimension $K - 1$.

The mean and variance of the Dirichlet distribution expressed in terms of $\boldsymbol{\mu}$ and m are

$$\text{E}(\boldsymbol{x}) = \boldsymbol{\mu}$$

and

$$\text{Var}(\boldsymbol{x}) = \frac{\text{Diag}(\boldsymbol{\mu}) - \boldsymbol{\mu}\boldsymbol{\mu}^T}{m + 1}$$

The covariance matrix is singular by construction because of the dependencies among the elements of $\boldsymbol{x}$. In component notation it is

$$\begin{aligned}\text{Cov}(x_i, x_j) &= \frac{[i = j]\mu_i - \mu_i\mu_j}{m + 1} \\ &= \begin{cases} \mu_i(1 - \mu_i)/(m + 1) & \text{if } i = j \\ -\mu_i\mu_j/(m + 1) & \text{if } i \neq j \end{cases}\end{aligned}$$

Special case: symmetric Dirichlet distribution

For $\boldsymbol{\alpha} = \alpha \mathbf{1}_K$ the Dirichlet distribution becomes the **symmetric beta distribution** with a single shape parameters $\alpha > 0$. In mean parametrisation the symmetric Dirichlet distribution corresponds to $\boldsymbol{\mu} = \mathbf{1}_K/K$ and $m = \alpha K$.

Special case: uniform distribution

For $\boldsymbol{\alpha} = \mathbf{1}_K$ the Dirichlet distribution becomes the uniform distribution over the $K - 1$ unit simplex with pdf $p(\boldsymbol{x}) = 1/\Gamma(K)$. In mean parametrisation the uniform distribution corresponds to $\boldsymbol{\mu} = \mathbf{1}_K/K$ and $m = K$.

6.3 Multivariate normal distribution

The **multivariate normal distribution** $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ generalises the univariate normal distribution $N(\mu, \sigma^2)$ (Section 5.3) from one to d dimensions.

Special cases are the **multivariate standard normal distribution** $N(0, \boldsymbol{I})$ and the **multivariate delta distribution** $\delta(\boldsymbol{\mu})$.

Standard parametrisation

The multivariate normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ has a mean or location parameter $\boldsymbol{\mu}$ (a d dimensional vector), a variance parameter $\boldsymbol{\Sigma}$ (a $d \times d$ positive definite symmetric matrix) and support $\boldsymbol{x} \in \mathbb{R}^d$.

If a random vector $\boldsymbol{x} = (x_1, x_2, ..., x_d)^T$ follows a multivariate normal distribution we write

$$\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

with mean

$$\mathrm{E}(\boldsymbol{x}) = \boldsymbol{\mu}$$

and variance

$$\mathrm{Var}(\boldsymbol{x}) = \boldsymbol{\Sigma}$$

In the above notation the dimension d is implicitly given by the dimensions of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ but for clarity one often also writes $N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to explicitly indicate the dimension.

The pdf is given by

$$\begin{aligned} p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \det(2\pi\boldsymbol{\Sigma})^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \\ &= \det(\boldsymbol{\Sigma})^{-1/2}(2\pi)^{-d/2}e^{-\Delta^2/2} \end{aligned}$$

Here $\Delta^2 = (\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$ is the **squared Mahalanobis distance** between $\boldsymbol{x}$ and $\boldsymbol{\mu}$ taking into account the variance $\boldsymbol{\Sigma}$. Note that this pdf is a joint pdf over the d elements $x_1, \ldots, x_d$ of the random vector $\boldsymbol{x}$.

The multivariate normal distribution is sometimes also used by specifying the precision matrix $\boldsymbol{\Sigma}^{-1}$ instead of the variance $\boldsymbol{\Sigma}$.

For $d = 1$ the random vector $\boldsymbol{x} = x$ is a scalar, $\boldsymbol{\mu} = \mu$, $\boldsymbol{\Sigma} = \sigma^2$ and thus the multivariate normal distribution reduces to the univariate normal distribution (Section 5.3).

R code

The `mnormt` package implements the multivariate normal distribution. The function `mnormt::dmnorm()` provides the pdf and `mnormt::pmnorm()` returns the distribution function. The function `mnormt::rmnorm()` is the corresponding random number generator.
The `mniw` package also implements the multivariate normal distribution. The pdf of the Wishart distribution is given by `mniw::dmNorm()`. The corresponding random number generator is `mniw::rmNorm()`.

Scale parametrisation

In the univariate case it is straightforward to use the standard deviation σ as scale parameter instead of the variance σ^2, and similarly the inverse standard deviation $w = 1/\sigma$ instead of the precision σ^{-2}. However, in the multivariate setting with a matrix variance parameter $\boldsymbol{\Sigma}$ it is less obvious how to define a suitable matrix scale parameter.

Let $\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$ be the eigendecomposition of the positive definite matrix $\boldsymbol{\Sigma}$. Then $\boldsymbol{\Sigma}^{1/2} = \boldsymbol{U}\boldsymbol{\Lambda}^{1/2}\boldsymbol{U}^T$ is the principal matrix square root and $\boldsymbol{\Sigma}^{-1/2} = \boldsymbol{U}\boldsymbol{\Lambda}^{-1/2}\boldsymbol{U}^T$ the inverse principal matrix square root. Furthermore, let $\boldsymbol{Q}$ be an arbitrary orthogonal matrix with $\boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{Q}\boldsymbol{Q}^T = \boldsymbol{I}$.

Then $\boldsymbol{W} = \boldsymbol{Q}\boldsymbol{\Sigma}^{-1/2}$ is called a **whitening matrix** based on $\boldsymbol{\Sigma}$ and $\boldsymbol{L} = \boldsymbol{W}^{-1} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{Q}^T$ is the corresponding **inverse whitening matrix**. By construction, the matrix $\boldsymbol{L}$ provides a factorisation of the covariance matrix by $\boldsymbol{L}\boldsymbol{L}^T = \boldsymbol{\Sigma}$. Similarly, $\boldsymbol{W}$ factorises the precision matrix by $\boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{\Sigma}^{-1}$. The two matrices thus provide the basis for the scale parametrisation of the multivariate normal distribution.

Specifically, the matrix $\boldsymbol{L}$ is used in place of $\boldsymbol{\Sigma}$ and plays the role of the matrix scale parameter (corresponding to σ in the univariate setting) and $\boldsymbol{W}$ is used in place of the precision matrix $\boldsymbol{\Sigma}^{-1}$ and plays the role of the inverse matrix scale parameter (corresponding to $1/\sigma$ in the univariate case). The determinants occurring in the multivariate normal pdf can be rewritten in terms of $\boldsymbol{L}$ and $\boldsymbol{W}$ using the identities $|\det(\boldsymbol{W})| = \det(\boldsymbol{\Sigma})^{-1/2}$ and $|\det(\boldsymbol{L})| = \det(\boldsymbol{\Sigma})^{1/2}$ as $\det(\boldsymbol{Q}) = \pm 1$.

Since $\boldsymbol{Q}$ can be freely chosen the matrices $\boldsymbol{W}$ and $\boldsymbol{L}$ are not fully determined by $\boldsymbol{\Sigma}$ alone but there is rotational freedom due to $\boldsymbol{Q}$. Standard choices are

- $\boldsymbol{Q}^{\mathrm{ZCA}} = \boldsymbol{I}$ for ZCA-type factorisation with $\boldsymbol{W}^{\mathrm{ZCA}} = \boldsymbol{\Sigma}^{-1/2}$ and
- $\boldsymbol{Q}^{\mathrm{PCA}} = \boldsymbol{U}^T$ for PCA-type factorisation with $\boldsymbol{W}^{\mathrm{PCA}} = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{U}^T$. Note that the matrix $\boldsymbol{U}$ is not unique because its columns (eigenvectors) can have different signs (directions), hence $\boldsymbol{W}^{\mathrm{PCA}}$ and $\boldsymbol{L}^{\mathrm{PCA}}$ are also not unique without further constraints, such as positive diagonal elements of the (inverse) whitening matrix.
- A third common choice is to compute $\boldsymbol{L}$ directly by Cholesky decomposition of $\boldsymbol{\Sigma}$, which yields an $\boldsymbol{L}^{\mathrm{Chol}}$ (and also a $\boldsymbol{W}^{\mathrm{Chol}}$) in the form of a lower-triangular matrix with a positive diagonal, and a corresponding underlying $\boldsymbol{Q}^{\mathrm{Chol}} = (\boldsymbol{L}^{\mathrm{Chol}})^T\boldsymbol{\Sigma}^{-1/2}$.

Finally, the whitening matrix $\boldsymbol{W}$ and its inverse may also be constructed from the correlation matrix $\boldsymbol{P}$ and the diagonal matrix containing the variances $\boldsymbol{V}$ (with $\boldsymbol{\Sigma} = \boldsymbol{V}^{1/2}\boldsymbol{P}\boldsymbol{V}^{1/2}$) in the form $\boldsymbol{W} = \boldsymbol{Q}\boldsymbol{P}^{-1/2}\boldsymbol{V}^{-1/2}$ and $\boldsymbol{L} = \boldsymbol{V}^{1/2}\boldsymbol{P}^{1/2}\boldsymbol{Q}^T$.

Special case: multivariate standard normal distribution

The **multivariate standard normal distribution** $N(0, \boldsymbol{I})$ has mean $\boldsymbol{\mu} = 0$ and variance $\boldsymbol{\Sigma} = \boldsymbol{I}$. The corresponding pdf is

$$p(\boldsymbol{x}) = (2\pi)^{-d/2} e^{-\boldsymbol{x}^T \boldsymbol{x}/2}$$

with the squared Mahalanobis distance reduced to $\Delta^2 = \boldsymbol{x}^T \boldsymbol{x} = \sum_{i=1}^{d} x_i^2$.

The density of the multivariate standard normal distribution is the product of the corresponding univariate standard normal densities

$$p(\boldsymbol{x}) = \prod_{i=1}^{d} (2\pi)^{-1/2} e^{-x_i^2/2}$$

and therefore the elements x_i of $\boldsymbol{x} = (x_1, \ldots, x_d)^T$ are independent of each other.

Special case: multivariate delta distribution

The **multivariate delta distribution** $\delta(\boldsymbol{\mu})$ is obtained as the limit of $N(\boldsymbol{\mu}, \varepsilon \boldsymbol{A})$ for $\varepsilon \to 0$ and where $\boldsymbol{A}$ is a positive definite matrix (e.g. $\boldsymbol{A} = \boldsymbol{I}$). Thus $\delta(\boldsymbol{\mu})$ is a continuous distribution representing a point mass at $\boldsymbol{\mu}$.

The corresponding pdf $\delta(\boldsymbol{x}|\boldsymbol{\mu})$ is called the **multivariate Dirac delta function**, even though it is not an ordinary function. It satisfies $\delta(\boldsymbol{x}|\boldsymbol{\mu}) = 0$ for all $\boldsymbol{x} \neq \boldsymbol{\mu}$ with an infinite spike at $\boldsymbol{\mu}$ but still integrates to one.

Location-scale transformation

Let $\boldsymbol{W}$ be a whitening matrix for $\boldsymbol{\Sigma}$ and $\boldsymbol{L}$ the corresponding inverse whitening matrix.

If $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then $\boldsymbol{y} = \boldsymbol{W}(\boldsymbol{x} - \boldsymbol{\mu}) \sim N(0, \boldsymbol{I})$. This location-scale transformation corresponds to centring and whitening (i.e. standardisation and decorrelation) of a multivariate normal random variable.

Conversely, if $\boldsymbol{y} \sim N(0, \boldsymbol{I})$ then $\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{L}\boldsymbol{y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. This location-scale transformation generates the multivariate normal distribution from the multivariate standard normal distribution.

Note that under the location-scale transformation $\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{L}\boldsymbol{y}$ with $\mathrm{Var}(\boldsymbol{y}) = \boldsymbol{I}$ we get $\mathrm{Cov}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{L}$. This provides a means to choose

between different (inverse) whitening transformation and the corresponding factorisations of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}^{-1}$. For example, if positive correlation between corresponding elements in $\boldsymbol{x}$ and $\boldsymbol{y}$ is desired then the diagonal elements in $\boldsymbol{L}$ must be positive.

Convolution property

The convolution of n independent, but not necessarily identical, multivariate normal distributions of the same dimension d results in another d-dimensional multivariate normal distribution with corresponding mean and variance:

$$\sum_{i=1}^{n} N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \sim N\left(\sum_{i=1}^{n} \boldsymbol{\mu}_i, \sum_{i=1}^{n} \boldsymbol{\Sigma}_i\right)$$

Hence, any multivariate normal random variable can be constructed as the sum of n suitable independent multivariate normal random variables.

Since n is an arbitrary positive integer the multivariate normal distribution is said to be **infinitely divisible**.

6.4 Wishart distribution

The Wishart distribution $\mathrm{Wis}(\boldsymbol{S}, k)$ is a multivariate generalisation of the gamma distribution $\mathrm{Gam}(\alpha, \theta)$ (Section 5.4) from one to d dimensions.

Standard parametrisation

If the symmetric random matrix $\boldsymbol{X}$ of dimension $d \times d$ is Wishart distributed we write

$$\boldsymbol{X} \sim \mathrm{Wis}(\boldsymbol{S}, k)$$

where $\boldsymbol{S} = (s_{ij})$ is the scale parameter (a symmetric $d \times d$ positive definite matrix with elements s_{ij}). The dimension d is implicit in the scale parameter $\boldsymbol{S}$.

The shape parameter k takes on real values in the range $k > d - 1$ and integer values in the range $k \in 1, \ldots, d - 1$ for $d > 1$. For $k > d - 1$ the matrix $\boldsymbol{X}$ is positive definite and invertible (see also Section 6.5), otherwise $\boldsymbol{X}$ is singular and positive semi-definite.

The distribution has mean

$$\mathrm{E}(\boldsymbol{X}) = k\boldsymbol{S}$$

and variances of the elements of $\boldsymbol{X}$ are

$$\mathrm{Var}(x_{ij}) = k\left(s_{ij}^2 + s_{ii}s_{jj}\right)$$

The pdf is (for $k > d-1$)

$$p(\boldsymbol{X}|\boldsymbol{S}, k) = \frac{1}{\Gamma_d(k/2)\det(2\boldsymbol{S})^{k/2}}\det(\boldsymbol{X})^{(k-d-1)/2}\exp\left(-\mathrm{Tr}(\boldsymbol{S}^{-1}\boldsymbol{X})/2\right)$$

This pdf is a joint pdf over the d diagonal elements x_{ii} and the $d(d-1)/2$ off-diagonal elements x_{ij} of the symmetric random matrix $\boldsymbol{X}$.

Part of the normalisation factor in the density is the **multivariate gamma function**

$$\Gamma_d(x) = \pi^{d(d-1)/4}\prod_{j=1}^{d}\Gamma(x-(j-1)/2)$$

For $d = 1$ it reduces to the standard gamma function. The multivariate gamma function also occurs in densities of other matrix-variate distributions.

If $\boldsymbol{S}$ is a scalar rather than a matrix (and hence $d = 1$) then the multivariate Wishart distribution reduces to the univariate Wishart aka gamma distribution (Section 5.4).

The Wishart distribution is closely related to the multivariate normal distribution with mean zero. Specifically, if $\boldsymbol{z} \sim N(0, \boldsymbol{S})$ then $\boldsymbol{z}\boldsymbol{z}^T \sim \mathrm{Wis}(\boldsymbol{S}, 1)$.

R code

The `mniw` package implements the Wishart distribution. The pdf of the Wishart distribution is given by `mniw::dwish()`. The corresponding random number generator is `mniw::rwish()`.

Mean parametrisation

It is useful to employ the Wishart distribution in **mean parametrisation**

$$\mathrm{Wis}\left(\boldsymbol{S} = \frac{\boldsymbol{M}}{k}, k\right)$$

with parameters $\boldsymbol{M} = k\boldsymbol{S}$ and k. In this parametrisation the mean is

$$\mathrm{E}(\boldsymbol{X}) = \boldsymbol{M} = (\mu_{ij})$$

and variances of the elements of $\boldsymbol{X}$ are

$$\mathrm{Var}(x_{ij}) = \frac{\mu_{ij}^2 + \mu_{ii}\mu_{jj}}{k}$$

Special case: standard Wishart distribution

For $\boldsymbol{S} = \boldsymbol{I}$ the Wishart distribution reduces to the **standard Wishart distribution**

$$\boldsymbol{X} \sim \mathrm{Wis}(\boldsymbol{I}, k)$$

with a single shape parameter k. The mean is

$$\mathrm{E}(\boldsymbol{X}) = k\boldsymbol{I}$$

and variances of the elements of $\boldsymbol{X}$ are

$$\mathrm{Var}(x_{ij}) = \begin{cases} 2k & \text{if } i = j \\ k & \text{if } i \neq j \end{cases}$$

The pdf is (for $k > d-1$)

$$p(\boldsymbol{X}|k) = \frac{1}{\Gamma_d(k/2)2^{dk/2}}\det(\boldsymbol{X})^{(k-d-1)/2}\exp\left(-\mathrm{Tr}(\boldsymbol{X})/2\right)$$

The standard Wishart distribution is closely related to the standard multivariate normal distribution with mean zero. Specifically, if $\boldsymbol{z} \sim N(0, \boldsymbol{I})$ then $\boldsymbol{z}\boldsymbol{z}^T \sim \mathrm{Wis}(\boldsymbol{I}, 1)$.

The **Bartlett decomposition** of the standard multivariate Wishart $\mathrm{Wis}(\boldsymbol{I}, k)$ distribution for any real $k > d-1$ is obtained by Cholesky factorisation of the random matrix $\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{Z}^T$. By construction $\boldsymbol{Z}$ is a lower-triangular matrix with positive diagonal elements z_{ii} and lower off-diagonal elements z_{ij} with $i > j$ and $i, j \in \{1, \ldots, d\}$. The corresponding upper off-diagonal elements are set to zero ($z_{ji} = 0$).

The $d(d+1)/2$ elements of $\boldsymbol{Z}$ are independent and allow to generate a standard Wishart variate as follows:

1) the *squared* diagonal elements follow a univariate standard Wishart distribution $z_{ii}^2 \sim \mathrm{Wis}(1, k-i+1)$ and
2) the off-diagonal elements follow the univariate standard normal distribution $z_{ij} \sim N(0, 1)$.
3) Then $\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{Z}^T \sim \mathrm{Wis}(\boldsymbol{I}, k)$.

Scale transformation

If $\boldsymbol{X} \sim \text{Wis}(\boldsymbol{S}, k)$ then the scaled symmetric random matrix $\boldsymbol{AXA}^T$ is also Wishart distributed with $\boldsymbol{AXA}^T \sim \text{Wis}(\boldsymbol{ASA}^T, k)$ where the matrix $\boldsymbol{A}$ must be full rank and $\boldsymbol{ASA}^T$ remains positive definite. The matrix $\boldsymbol{A}$ may be rectangular, hence the size of $\boldsymbol{AXA}^T$ and $\boldsymbol{ASA}^T$ may be smaller compared to $\boldsymbol{X}$ and $\boldsymbol{S}$.

The transformations between the Wishart distribution and the standard Wishart distribution are two important special cases:

1) With $\boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{S}^{-1}$ and $\boldsymbol{X} \sim \text{Wis}(\boldsymbol{S}, k)$ then $\boldsymbol{Y} = \boldsymbol{WXW}^T \sim \text{Wis}(\boldsymbol{I}, k)$ as $\boldsymbol{WSW}^T = \boldsymbol{I}$. This transformation reduces the Wishart distribution to the standard Wishart distribution.

2) Conversely, with $\boldsymbol{LL}^T = \boldsymbol{S}$ and $\boldsymbol{Y} \sim \text{Wis}(\boldsymbol{I}, k)$ then $\boldsymbol{X} = \boldsymbol{LYL}^T \sim \text{Wis}(\boldsymbol{S}, k)$ as $\boldsymbol{LIL}^T = \boldsymbol{S}$. This transformation generates the Wishart distribution from the standard Wishart distribution.

Convolution property

The convolution of n Wishart distributions with the same scale parameter $\boldsymbol{S}$ but possible different shape parameters k_i yields another Wishart distribution:

$$\sum_{i=1}^{n} \text{Wis}(\boldsymbol{S}, k_i) \sim \text{Wis}\left(\boldsymbol{S}, \sum_{i=1}^{n} k_i\right)$$

Note that the shape parameter k is restricted to be an integer in the range $1, \ldots, d-1$ for $d > 1$ but is a real number in the range $k > d-1$. Thus, if the k_i are all valid shape parameters (for dimension d) then $\sum_{i=1}^{n} k_i$ is also a valid shape parameter.

Due the partial restriction of the shape parameter k to integer values the multivariate Wishart distribution is **not infinitely divisible** for $d > 1$.

The above includes the following construction of the multivariate Wishart distribution $\text{Wis}(S, k)$ for integer-valued k. The sum of k independent Wishart random variables $\text{Wis}(\boldsymbol{S}, 1)$ with one degree of freedom and identical scale parameter yields a Wishart random variable $\text{Wis}(\boldsymbol{S}, k)$ with degree of freedom k and the same scale parameter. Thus, if $\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_k \sim N(0, \boldsymbol{S})$ are k independent samples from $N(0, \boldsymbol{S})$ then $\sum_{i_1}^{k} \boldsymbol{z}_i \boldsymbol{z}_i^T \sim \text{Wis}(\boldsymbol{S}, k)$.

6.5 Inverse Wishart distribution

The **inverse Wishart distribution** $\text{IWis}(\boldsymbol{\Psi}, k)$ is a multivariate generalisation of the inverse gamma distribution $\text{IGam}(\alpha, \beta)$ (Section 5.5) from one to d dimensions. It is linked to the Wishart distribution $\text{Wis}(\boldsymbol{S}, k)$ (Section 6.4).

Standard parametrisation

A symmetric positive definite random matrix $\boldsymbol{X}$ of dimension $d \times d$ following an inverse Wishart distribution is denoted by

$$\boldsymbol{X} \sim \text{IWis}(\boldsymbol{\Psi}, k)$$

where $\boldsymbol{\Psi} = (\psi_{ij})$ is the scale parameter (a $d \times d$ positive definite symmetric matrix) and $k > d-1$ is the shape parameter. The dimension d is implicit in the scale parameter $\boldsymbol{\Psi}$.

The mean is (for $k > d+1$)

$$\text{E}(\boldsymbol{X}) = \frac{\boldsymbol{\Psi}}{k-d-1}$$

and the variances of elements of $\boldsymbol{X}$ are (for $k > d+3$)

$$\text{Var}(x_{ij}) = \frac{(k-d-1)\,\psi_{ii}\psi_{jj} + (k-d+1)\,\psi_{ij}^2}{(k-d)(k-d-3)(k-d-1)^2}$$

The inverse Wishart distribution $\text{IWis}(\boldsymbol{\Psi}, k)$ has pdf

$$p(\boldsymbol{X}|\boldsymbol{\Psi}, k) = \frac{\det(\boldsymbol{\Psi}/2)^{k/2}}{\Gamma_d(k/2)} \det(\boldsymbol{X})^{-(k+d+1)/2} \exp\left(-\text{Tr}(\boldsymbol{\Psi}\boldsymbol{X}^{-1})/2\right)$$

As with the Wishart distribution his pdf is a joint pdf over the d diagonal elements x_{ii} and the $d(d-1)/2$ off-diagonal elements x_{ij} of the symmetric random matrix $\boldsymbol{X}$.

If $\boldsymbol{\Psi}$ is a scalar ψ (and $d = 1$) then the multivariate inverse Wishart distribution reduces to the univariate inverse Wishart distribution (Section 5.5).

R code

The `mniw` package implements the inverse Wishart distribution. The pdf of the inverse Wishart distribution is given by `mniw::diwish()`. The corresponding random number generator is `mniw::riwish()`.

Relation to the Wishart distribution

The inverse Wishart distribution is closely linked to the Wishart distribution. Assume that the random variable $\boldsymbol{Y}$ (positive definite and symmetric) follows a Wishart distribution with

$$\boldsymbol{Y} \sim \text{Wis}\left(\boldsymbol{S}, k\right)$$

then the inverse random variable $\boldsymbol{X} = \boldsymbol{Y}^{-1}$ (positive definite and symmetric) follows an inverse Wishart distribution with inverted scale parameter

$$\boldsymbol{X} = \boldsymbol{Y}^{-1} \sim \text{IWis}\left(\boldsymbol{\Psi} = \boldsymbol{S}^{-1}, k\right)$$

where k is the shared shape parameter, $\boldsymbol{S}$ the scale parameter of the Wishart distribution and $\boldsymbol{\Psi}$ the scale parameter of the inverse Wishart distribution.

Correspondingly, the density $p_X(\boldsymbol{X}|\boldsymbol{\Psi}, k)$ of the inverse Wishart distribution is obtained from the density $p_Y(\boldsymbol{Y}|\boldsymbol{S}, k)$ of the Wishart distribution via

$$p_X(\boldsymbol{X}|\boldsymbol{\Psi}, k) = \det(\boldsymbol{X})^{-d-1} p_Y\left(\boldsymbol{X}^{-1}|\boldsymbol{S} = \boldsymbol{\Psi}^{-1}, k\right)$$

The exponent $-d-1$ in the Jacobian determinant for matrix inversion arises because $\boldsymbol{X}$ and $\boldsymbol{Y}$ are symmetric (for non-symmetric matrices the exponent is $-2d$).

Mean parametrisation

Instead of $\boldsymbol{\Psi}$ and k we may also equivalently use $\boldsymbol{M} = \boldsymbol{\Psi}/(k-d-1)$ and $\kappa = k-d-1$ as parameters for the inverse Wishart distribution, so that

$$\boldsymbol{X} \sim \text{IWis}\left(\boldsymbol{\Psi} = \kappa \boldsymbol{M}\,,\; k = \kappa + d + 1\right)$$

with mean (for $\kappa > 0$)

$$\text{E}(\boldsymbol{X}) = \boldsymbol{M}$$

and variances (for $\kappa > 2$)

$$\text{Var}(x_{ij}) = \frac{\kappa\, \mu_{ii}\mu_{jj} + (\kappa+2)\, \mu_{ij}^2}{(\kappa+1)(\kappa-2)}$$

For $\boldsymbol{M}$ equal to scalar μ with $d = 1$ the above reduces to the univariate inverse Wishart distribution in mean parametrisation.

Biased mean parametrisation

Using $\boldsymbol{T} = (t_{ij}) = \boldsymbol{\Psi}/(k-d+1) = \boldsymbol{\Psi}/\nu$ as biased mean parameter together with $\nu = k - d + 1$ we arrive at the **biased mean parametrisation**

$$\boldsymbol{X} \sim \text{IWis}\left(\boldsymbol{\Psi} = \nu \boldsymbol{T}\,,\; k = \nu + d - 1\right)$$

The corresponding mean is (for $\nu > 2$)

$$\text{E}(\boldsymbol{X}) = \frac{\nu}{\nu-2}\boldsymbol{T} = \boldsymbol{M}$$

and the variances of elements of $\boldsymbol{X}$ are (for $\nu > 4$)

$$\text{Var}(x_{ij}) = \left(\frac{\nu}{\nu-2}\right)^2 \frac{(\nu-2)\, t_{ii}t_{jj} + \nu\, t_{ij}^2}{(\nu-1)(\nu-4)}$$

As $\boldsymbol{T} = \boldsymbol{M}(\nu-2)/\nu$ for large ν the parameter $\boldsymbol{T}$ will become identical to the true mean $\boldsymbol{M}$.

For $\boldsymbol{T}$ equal to scalar τ^2 with $d = 1$ the above reduces to the univariate inverse Wishart distribution in biased mean parametrisation.

Scale transformation

If $\boldsymbol{X} \sim \text{IWis}(\boldsymbol{\Psi}, k)$ then the scaled symmetric random matrix $\boldsymbol{A}\boldsymbol{X}\boldsymbol{A}^T$ is also inverse Wishart distributed with $\boldsymbol{A}\boldsymbol{X}\boldsymbol{A}^T \sim \text{IWis}(\boldsymbol{A}\boldsymbol{\Psi}\boldsymbol{A}^T, k)$ where the matrix $\boldsymbol{A}$ has full rank and both $\boldsymbol{A}\boldsymbol{X}\boldsymbol{A}^T$ and $\boldsymbol{A}\boldsymbol{\Psi}\boldsymbol{A}^T$ remain positive definite. The matrix $\boldsymbol{A}$ may be rectangular, hence the size of $\boldsymbol{A}\boldsymbol{X}\boldsymbol{A}^T$ and $\boldsymbol{A}\boldsymbol{\Psi}\boldsymbol{A}^T$ may be smaller compared to $\boldsymbol{X}$ and $\boldsymbol{\Psi}$.

6.6 Multivariate t-distribution

The **multivariate t-distribution** $t_\nu(\boldsymbol{\mu}, \boldsymbol{T})$ is a multivariate generalisation of the location-scale t-distribution $t_\nu(\mu, \tau^2)$ (Section 5.6) from one to d dimensions. It is a generalisation of the multivariate normal distribution $N(\boldsymbol{\mu}, \boldsymbol{T})$ (Section 6.3) with an additional parameter $\nu > 0$ (degrees of freedom) controlling the probability mass in the tails.

Special cases include the **multivariate standard t-distribution** $t_\nu(0, \boldsymbol{I})$, the **multivariate normal distribution** $N(\boldsymbol{\mu}, \boldsymbol{T})$ and the **multivariate Cauchy** distribution $\text{Cau}(\boldsymbol{\mu}, \boldsymbol{T})$.

Standard parametrisation

If $\boldsymbol{x} \in \mathbb{R}^d$ is a multivariate t-distributed random variable we write

$$\boldsymbol{x} \sim t_\nu(\boldsymbol{\mu}, \boldsymbol{T})$$

where the vector $\boldsymbol{\mu}$ is the location parameter (a d dimensional vector) and the dispersion parameter $\boldsymbol{T}$ is a symmetric positive definite matrix of dimension $d \times d$. The dimension d is implicit in both parameters. The parameter $\nu > 0$ prescribes the degrees of freedom. For small values of ν the distribution is heavy-tailed and as a result only moments of order smaller than ν are finite and defined.

The mean is (for $\nu > 1$)

$$\text{E}(\boldsymbol{x}) = \boldsymbol{\mu}$$

and the variance (for $\nu > 2$)

$$\text{Var}(\boldsymbol{x}) = \frac{\nu}{\nu - 2}\boldsymbol{T}$$

The pdf of $t_\nu(\boldsymbol{\mu}, \boldsymbol{T})$ is

$$p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{T}, \nu) = \det(\boldsymbol{T})^{-1/2} \frac{\Gamma(\frac{\nu+d}{2})}{(\pi\nu)^{d/2}\,\Gamma(\frac{\nu}{2})} \left(1 + \frac{\Delta^2}{\nu}\right)^{-(\nu+d)/2}$$

with $\Delta^2 = (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{T}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})$ the squared Mahalanobis distance between $\boldsymbol{x}$ and $\boldsymbol{\mu}$. Note that this pdf is a joint pdf over the d elements $x_1, \ldots, x_d$ of the random vector $\boldsymbol{x}$.

For $d = 1$ the random vector $\boldsymbol{x} = x$ is a scalar, $\boldsymbol{\mu} = \mu$, $\boldsymbol{T} = \tau^2$ and thus the multivariate t-distribution reduces to the location-scale t-distribution (Section 5.6).

R code

The `mnormt` package implements the multivariate t-distribution. The function `mnormt::dmt()` provides the pdf and `mnormt::pmt()` returns the distribution function. The function `mnormt::rmt()` is the corresponding random number generator.

Scale parametrisation

The multivariate t-distribution, like the multivariate distribution, can also be represented with a matrix scale parameter $\boldsymbol{L}$ in place of a matrix dispersion parameter $\boldsymbol{T}$.

Let $\boldsymbol{L}$ be a matrix scale parameter such that $\boldsymbol{L}\boldsymbol{L}^T = \boldsymbol{T}$ and $\boldsymbol{W} = \boldsymbol{L}^{-1}$ be the corresponding inverse matrix scale parameter with $\boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{T}^{-1}$. By construction $|\det(\boldsymbol{W})| = \det(\boldsymbol{T})^{-1/2}$ and $|\det(\boldsymbol{L})| = \det(\boldsymbol{T})^{1/2}$.

Note that $\boldsymbol{T}$ alone does not fully determine $\boldsymbol{L}$ and $\boldsymbol{W}$ due to rotational freedom, see the discussion in Section 6.3 for details.

Special case: multivariate standard t-distribution

With $\boldsymbol{\mu} = 0$ and $\boldsymbol{T} = \boldsymbol{I}$ the multivariate t-distribution reduces to the **multivariate standard t-distribution** $t_\nu(0, \boldsymbol{I})$. It is a generalisation of the multivariate standard normal distribution $N(0, \boldsymbol{I})$ to allow for heavy tails.

The distribution has mean $\text{E}(\boldsymbol{x}) = 0$ (for $\nu > 1$) and variance $\text{Var}(\boldsymbol{x}) = \frac{\nu}{\nu-2}\boldsymbol{I}$ (for $\nu > 2$).

The pdf of $t_\nu(0, \boldsymbol{I})$ is

$$p(\boldsymbol{x}|\nu) = \frac{\Gamma(\frac{\nu+d}{2})}{(\pi\nu)^{d/2}\,\Gamma(\frac{\nu}{2})} \left(1 + \frac{\boldsymbol{x}^T\boldsymbol{x}}{\nu}\right)^{-(\nu+d)/2}$$

with the squared Mahalanobis distance reducing to $\Delta^2 = \boldsymbol{x}^T\boldsymbol{x}$.

For scalar x (and hence $d = 1$) the multivariate standard t-distribution reduces to the Student's t-distribution $t_\nu = t_\nu(0, 1)$.

Unlike the multivariate standard normal distribution, the density of the multivariate standard t-distribution cannot be written as product of corresponding univariate standard densities.

Special case: multivariate normal distribution

For $\nu \to \infty$ the multivariate t-distribution $t_\nu(\boldsymbol{\mu}, \boldsymbol{T})$ reduces to the **multivariate normal distribution** $N(\boldsymbol{\mu}, \boldsymbol{T})$ (Section 6.3). Correspondingly, for $\nu \to \infty$ the multivariate standard t-distribution $t_\nu(0, \boldsymbol{I})$ becomes equal to the **multivariate standard normal distribution** $N(0, \boldsymbol{I})$.

This can be seen from the corresponding limits of the two factors in the pdf of the multivariate t-distribution that depend on ν:

1) Following Sterling's approximation for large x we can approximate $\log\Gamma(x) \approx (x-1)\log(x-1)$. For large ν this implies that

$$\frac{\Gamma((\nu+d)/2)}{(\pi\nu)^{d/2}\,\Gamma(\nu/2)} \to (2\pi)^{-d/2}$$

2) For small x we can approximate $\log(1+x) \approx x$. Thus for large $\nu \gg d$ (and hence small Δ^2/ν) this yields $(\nu+d)\log(1+\Delta^2/\nu) \to \Delta^2$ and hence $\left(1+\Delta^2/\nu\right)^{-(\nu+d)/2} \to e^{-\Delta^2/2}$.

Hence, the pdf of $t_\infty(\boldsymbol{\mu}, \boldsymbol{T})$ is the multivariate normal pdf

$$p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{T}, \nu = \infty) = \det(\boldsymbol{T})^{-1/2}(2\pi)^{-d/2}e^{-\Delta^2/2}$$

Special case: multivariate Cauchy distribution

For $\nu = 1$ the multivariate t-distribution becomes the **multivariate Cauchy distribution** $\mathrm{Cau}(\boldsymbol{\mu}, \boldsymbol{T}) = t_1(\boldsymbol{\mu}, \boldsymbol{T})$.

Its mean, variance and other higher moments are all undefined.

It has pdf

$$p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{T}) = \det(\boldsymbol{T})^{-1/2}\Gamma\left(\frac{d+1}{2}\right)\left(\pi(1+\Delta^2)\right)^{-(d+1)/2}$$

For scalar x (and hence $d = 1$) the multivariate Cauchy distribution $\mathrm{Cau}(\boldsymbol{\mu}, \boldsymbol{T})$ reduces to the univariate Cauchy distribution $\mathrm{Cau}(\mu, \tau^2)$.

Special case: multivariate standard Cauchy distribution

The **multivariate standard Cauchy distribution** $\mathrm{Cau}(0, \boldsymbol{I}) = t_1(0, \boldsymbol{I})$ is obtained by setting $\boldsymbol{\mu} = 0$ and $\boldsymbol{T} = \boldsymbol{I}$ in the multivariate Cauchy distribution or, equivalently, by setting $\nu = 1$ in the multivariate standard t-distribution.

It has pdf

$$p(\boldsymbol{x}) = \Gamma\left(\frac{d+1}{2}\right)\left(\pi(1+\boldsymbol{x}^T\boldsymbol{x})\right)^{-(d+1)/2}$$

For scalar x (and hence $d = 1$) the multivariate standard Cauchy distribution $\mathrm{Cau}(0, \boldsymbol{I})$ reduces to the standard univariate Cauchy distribution $\mathrm{Cau}(0, 1)$.

Location-scale transformation

Let $\boldsymbol{L}$ be a scale matrix for $\boldsymbol{T}$ and $\boldsymbol{W}$ the corresponding inverse scale matrix.

If $\boldsymbol{x} \sim t_\nu(\boldsymbol{\mu}, \boldsymbol{T})$ then $\boldsymbol{y} = \boldsymbol{W}(\boldsymbol{x} - \boldsymbol{\mu}) \sim t_\nu(0, \boldsymbol{I})$. This location-scale transformation reduces a multivariate t-distributed random variable to a standard multivariate t-distributed random variable.

Conversely, if $\boldsymbol{y} \sim t_\nu(0, \boldsymbol{I})$ then $\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{L}\boldsymbol{y} \sim t_\nu(\boldsymbol{\mu}, \boldsymbol{T})$. This location-scale transformation generates the multivariate t-distribution from the multivariate standard t-distribution.

Note that for $\nu > 2$ under the location-scale transformation $\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{L}\boldsymbol{y}$ with $\mathrm{Var}(\boldsymbol{y}) = \nu/(\nu-2)\boldsymbol{I}$ we get $\mathrm{Cov}(\boldsymbol{x}, \boldsymbol{y}) = \nu/(\nu-2)\boldsymbol{L}$. This provides a means to choose between different factorisations of $\boldsymbol{T}$ and $\boldsymbol{T}^{-1}$. For example, if positive correlation between corresponding elements in $\boldsymbol{x}$ and $\boldsymbol{y}$ is desired then the diagonal elements in $\boldsymbol{L}$ must be positive.

For the special case of the multivariate Cauchy distribution (corresponding to $\nu = 1$) similar relations hold between it and the multivariate standard Cauchy distribution. If $\boldsymbol{x} \sim \mathrm{Cau}(\boldsymbol{\mu}, \boldsymbol{T})$ then $\boldsymbol{y} = \boldsymbol{W}(\boldsymbol{x} - \boldsymbol{\mu}) \sim \mathrm{Cau}(0, \boldsymbol{I})$. Conversely, if $\boldsymbol{y} \sim \mathrm{Cau}(0, \boldsymbol{I})$ then $\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{L}\boldsymbol{y} \sim \mathrm{Cau}(\boldsymbol{\mu}, \boldsymbol{T})$.

Convolution property

The multivariate t-distribution is not generally closed under convolution, with the exception of two special cases, the multivariate normal distribution ($\nu = \infty$), see Section 6.3, and the multivariate Cauchy distribution ($\nu = 1$) with the additional restriction that the dispersion parameters are proportional.

For the Cauchy distribution with $\boldsymbol{T}_i = a_i^2 \boldsymbol{T}$, where $a_i > 0$ are positive scalars,

$$\sum_{i=1}^{n} \text{Cau}(\boldsymbol{\mu}_i, a_i^2 \boldsymbol{T}) \sim \text{Cau}\left(\sum_{i=1}^{n} \boldsymbol{\mu}_i, \left(\sum_{i=1}^{n} a_i\right)^2 \boldsymbol{T}\right)$$

Multivariate t-distribution as compound distribution

The multivariate t-distribution can be obtained as mixture of multivariate normal distributions with identical mean and varying covariance matrix. Specifically, let z be a univariate inverse Wishart random variable

$$z \sim \text{IWis}(\psi = \nu, k = \nu) = \text{IGam}\left(\alpha = \frac{\nu}{2}, \beta = \frac{\nu}{2}\right)$$

and let $\boldsymbol{x}|z$ be multivariate normal

$$\boldsymbol{x}|z \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma} = z\boldsymbol{T})$$

The resulting marginal (scale mixture) distribution for $\boldsymbol{x}$ is the multivariate t-distribution

$$\boldsymbol{x} \sim t_\nu\left(\boldsymbol{\mu}, \boldsymbol{T}\right)$$

An alternative way to arrive at $t_\nu\left(\boldsymbol{\mu}, \boldsymbol{T}\right)$ is to include $\boldsymbol{T}$ as parameter in the inverse Wishart distribution

$$\boldsymbol{Z} \sim \text{IWis}(\boldsymbol{\Psi} = \nu\boldsymbol{T}, k = \nu + d - 1)$$

and let

$$\boldsymbol{x}|\boldsymbol{Z} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma} = \boldsymbol{Z})$$

Note that $\boldsymbol{T}$ is now the biased mean parameter of the multivariate inverse Wishart distribution. This characterisation is useful in Bayesian analysis.

7 Exponential families

7.1 Definition of an exponential family

Exponential tilting

A distribution family $P(\boldsymbol{\eta})$ for a random variable x is an **exponential family** if it is generated by **exponential tilting** of a base distribution B, resulting in a pdmf of the form

$$\begin{aligned} p(x|\boldsymbol{\eta}) &= \underbrace{h(x)}_{\text{base}} \underbrace{e^{\langle \boldsymbol{\eta}, \boldsymbol{t}(x)\rangle}}_{\text{exponential tilt}} / \underbrace{z(\boldsymbol{\eta})}_{\text{normaliser}} \\ &= h(x)\, e^{\langle \boldsymbol{\eta}, \boldsymbol{t}(x)\rangle - a(\boldsymbol{\eta})} \end{aligned}$$

where

- $\boldsymbol{t}(x)$ are the **canonical statistics**,
- $\boldsymbol{\eta}$ are the **canonical parameters**,
- $h(x)$ is a positive **base function** (typically unnormalised),
- $z(\boldsymbol{\eta})$ is the **partition function** and
- $a(\boldsymbol{\eta}) = \log z(\boldsymbol{\eta})$ the corresponding **log-partition function**.

The base pdmf is obtained at $\boldsymbol{\eta} = 0$ yielding $b(x) = p(x|\boldsymbol{\eta} = 0) = h(x)/z(0)$. If $h(x)$ is already a normalised pdmf then $z(0) = 1$ and $b(x) = h(x)$.

The above presentation of exponential families assumes a univariate random variable (scalar x) but also applies to multivariate random variables (vector $\boldsymbol{x}$ or matrix $\boldsymbol{X}$) .

Likewise, canonical statistics and parameters are written as vectors but these may also be scalars or matrices (or a combination of both). The use of inner product notation $\langle , \rangle$ includes all these cases, recall that for scalars $\langle a, b\rangle = ab$, for vectors $\langle \boldsymbol{a}, \boldsymbol{b}\rangle = \boldsymbol{a}^T\boldsymbol{a}$ and for matrices $\langle \boldsymbol{A}, \boldsymbol{B}\rangle = \text{Tr}(\boldsymbol{A}^T\boldsymbol{B}) = \text{Vec}(\boldsymbol{A})^T \text{Vec}(\boldsymbol{B})$.

Canonical statistics

The canonical statistics $\boldsymbol{t}(x)$ are transformations of x, usually simple functions such as the identity (x), the square (x^2), the inverse ($1/x$) or the logarithm ($\log x$).

Typically, the number of canonical statistics and hence the dimension of $\boldsymbol{t}$ is small.

The canonical statistics $\boldsymbol{t}(x)$ may be affinely dependent. If this is the case there is a vector $\boldsymbol{\eta}_0$ for which

$$\langle \boldsymbol{\eta}_0, \boldsymbol{t}(x) \rangle = \text{const.}$$

A common example is when $\boldsymbol{x}$ is a vector of counts $(n_1, \ldots, n_K)^T$ for K classes with a fixed total count $n = \sum_{i=k}^{K} n_k = \boldsymbol{x}^T \mathbf{1}_K$ and the canonical statistics are $\boldsymbol{t}(\boldsymbol{x}) = \boldsymbol{x}$ and thus include all K counts.

If the elements in $\boldsymbol{t}(x)$ are affinely independent the representation of the exponential family is **minimal** or **complete**, otherwise the representation is **non-minimal** or **overcomplete**.

Canonical parameters and identifiability

For each canonical statistic there is a corresponding canonical parameter so the dimensions and shape of $\boldsymbol{t}(x)$ and $\boldsymbol{\eta}$ match.

In a minimal representation the canonical parameters of the exponential family are **identifiable** and hence distinct parameter settings for $\boldsymbol{\eta}$ yield distinct distributions.

Conversely, in a non-minimal or overcomplete representation there are redundant elements in the canonical parameters $\boldsymbol{\eta}$ and the distributions within the exponential family are **not identifiable**. Specifically, there will be multiple $\boldsymbol{\eta}$ yielding the same underlying distribution.

In the example above, where $\boldsymbol{x}$ is a vector of counts with a fixed total count and $\boldsymbol{t}(\boldsymbol{x}) = \boldsymbol{x}$ and corresponding canonical parameters $\boldsymbol{\eta}$, the effective number of parameter is $K - 1$ rather than K, so there is one redundant parameter in $\boldsymbol{\eta}$.

Moment and cumulant generating functions

The moment generating function for the canonical statistics $\boldsymbol{t}(x)$ is

$$\begin{aligned} M(\boldsymbol{\tau}) &= \mathrm{E}\left(e^{\langle \boldsymbol{\tau}, \boldsymbol{t}(x) \rangle}\right) \\ &= \int_x e^{\langle \boldsymbol{\tau}, \boldsymbol{t}(x) \rangle} p(x|\boldsymbol{\eta}) dx \\ &= \int_x e^{\langle \boldsymbol{\tau}, \boldsymbol{t}(x) \rangle} e^{\langle \boldsymbol{\eta}, \boldsymbol{t}(x) \rangle} h(x) / z(\boldsymbol{\eta}) dx \\ &= \left(\int_x e^{\langle \boldsymbol{\tau}+\boldsymbol{\eta}, \boldsymbol{t}(x) \rangle} h(x)\, dx \right) / z(\boldsymbol{\eta}) \\ &= z(\boldsymbol{\tau} + \boldsymbol{\eta}) / z(\boldsymbol{\eta}) \end{aligned}$$

Correspondingly, the cumulant generating function is

$$\begin{aligned} K(\boldsymbol{\tau}) &= \log M(\boldsymbol{\tau}) \\ &= a(\boldsymbol{\tau} + \boldsymbol{\eta}) - a(\boldsymbol{\eta}) \end{aligned}$$

Thus, the moment and cumulant generating functions for the canonical statistics are closely linked to the partition and log-partition functions, respectively.

7.2 Roles of the partition function

Normalising factor

The pdmf $p(x|\boldsymbol{\eta})$ must integrate to one. Therefore, given $h(x)$ and $\boldsymbol{t}(x)$ the partition function $z(\boldsymbol{\eta})$ is obtained by

$$z(\boldsymbol{\eta}) = \int_x e^{\langle \boldsymbol{\eta}, \boldsymbol{t}(x) \rangle} h(x)\, dx$$

For discrete x replace the integral by a sum.

Consequently partition function $z(\boldsymbol{\eta})$ is also called the **normaliser** and and the log-partition function $a(\boldsymbol{\eta})$ is called the **log-normaliser**.

Constructed as weighted integral of exponentials of linear functions, the partition function $z(\boldsymbol{\eta})$ and the log-partition function $a(\boldsymbol{\eta})$ are **convex** with regard to $\boldsymbol{\eta}$.

For an exponential family in **minimal representation** the (log)-partition function(s) are **strictly convex**.

Definition of parameter space

The set of values of $\boldsymbol{\eta}$ for which $z(\boldsymbol{\eta}) < \infty$, and hence for which the pdmf $p(x|\boldsymbol{\eta})$ is well defined, comprises the parameter space of the exponential family. Some choices of $h(x)$ and $\boldsymbol{t}(x)$ do not yield a finite normalising factor for any $\boldsymbol{\eta}$ and hence these cannot be used to form an exponential family.

Moments of canonical statistics

The first cumulant (the mean) and second cumulant (the variance) are obtained as the first and second derivatives of the cumulant generating function $K(\boldsymbol{\tau}) = a(\boldsymbol{\tau} + \boldsymbol{\eta}) - a(\boldsymbol{\eta})$ evaluated at $\boldsymbol{\tau} = 0$, respectively. As a result, the log-partition function $a(\boldsymbol{\eta})$ provides a practical way to obtain the mean and variance of the canonical statistics $\boldsymbol{t}(x)$.

Specifically, computing its gradient yields the mean

$$\begin{aligned} \mathrm{E}(\boldsymbol{t}(x)) = \boldsymbol{\mu}_{\boldsymbol{t}} &= \nabla a(\boldsymbol{\eta}) \\ &= \frac{\nabla z(\boldsymbol{\eta})}{z(\boldsymbol{\eta})} \end{aligned}$$

and computing the Hessian matrix the covariance matrix

$$\begin{aligned} \mathrm{Var}(\boldsymbol{t}(x)) = \boldsymbol{\Sigma}_{\boldsymbol{t}} &= \nabla\nabla^T a(\boldsymbol{\eta}) \\ &= \frac{\nabla\nabla^T z(\boldsymbol{\eta})}{z(\boldsymbol{\eta})} - \left(\frac{\nabla z(\boldsymbol{\eta})}{z(\boldsymbol{\eta})}\right)\left(\frac{\nabla z(\boldsymbol{\eta})}{z(\boldsymbol{\eta})}\right)^T \end{aligned}$$

For an exponential family with minimal representation the log-partition function is strictly convex. Furthermore, the variance $\boldsymbol{\Sigma}_{\boldsymbol{t}}$ is a positive definite matrix and invertible.

For overcomplete representations the log-partition function is convex but not strictly convex. Furthermore, the covariance matrix is positive semi-definite and not invertible.

By construction, the log-partition function $a(\boldsymbol{\eta})$ is finite in the interior of its parameter space. Therefore *all* moments and cumulants of the canonical statistics $\boldsymbol{t}(x)$ exist (are finite) and are given by the derivatives of $a(\boldsymbol{\eta})$ at $\boldsymbol{\eta}$.

7.3 Further properties

Equivalent representations

An exponential family admits many equivalent representations such that different specifications of canonical statistics $\boldsymbol{t}(x)$ and base function $h(x)$ describe the same family.

First, any invertible linear transformation of the canonical statistic $\boldsymbol{t}(x)$ yields the same distribution family.

Second, any member of the family, say $P(\boldsymbol{\eta}_0)$, can serve as its base distribution. Specifically, with $p(x|\boldsymbol{\eta}_0)$ used as base the pdmf for the exponential family $P(\boldsymbol{\eta})$ is

$$p(x|\boldsymbol{\eta}) = e^{\langle \boldsymbol{\eta}-\boldsymbol{\eta}_0, \boldsymbol{t}(x)\rangle - (a(\boldsymbol{\eta}) - a(\boldsymbol{\eta}_0))}\, p(x|\boldsymbol{\eta}_0)$$

which is in exponential family form.

Third, the base function $h(x)$ can be left unnormalised, so there are infinitely many positive base functions $h(x)$ that yield the same base pdmf $b(x)$ after normalisation.

Fourth, any factors in $h(x)$ of the form $e^{\langle \boldsymbol{\eta}_0, \boldsymbol{t}(x)\rangle}$ for some constant $\boldsymbol{\eta}_0$ can be removed from $h(x)$ and absorbed into the exponential by replacing $\boldsymbol{\eta}$ with $\boldsymbol{\eta} + \boldsymbol{\eta}_0$ (thus leading to a different set of canonical parameters).

As a result, for many (but not all) commonly used exponential families the base function can be set to $h(x) = 1$ or some other constant value, so that all dependence on x enters through the canonical statistics $\boldsymbol{t}(x)$ via $\langle \boldsymbol{\eta}, \boldsymbol{t}(x)\rangle$.

Natural exponential families

If the canonical statistic is the identity, $t(x) = x$ or $\boldsymbol{t}(\boldsymbol{x}) = \boldsymbol{x}$, the family is a **natural exponential family** (**NEF**). A univariate NEF has a scalar canonical statistic $t(x)$ and a scalar canonical parameter η and is thus a one-parameter family.

Alternative parametrisations

An exponential family can be parametrised by three different sets of parameters:

1) **canonical parameters** $\boldsymbol{\eta}$,
2) **expectation parameters** $\boldsymbol{\mu_t} = \mathrm{E}(\boldsymbol{t}(x))$ (the mean of the canonical statistics $\boldsymbol{t}(x)$), as well as
3) **conventional parameters** $\boldsymbol{\theta}$ (such as mean and variance of x).

If the exponential family is minimal then there is a one-to-one map between the canonical parameters $\boldsymbol{\eta}$ and the expectation parameters $\boldsymbol{\mu_t}$.

The canonical and the expectation parameters can be expressed as a function of the conventional parameters $\boldsymbol{\theta}$.

Often, some expectation parameters $\boldsymbol{\mu_t}$ correspond to conventional parameters $\boldsymbol{\theta}$ (e.g., if one of the canonical statistics is x, then the corresponding expectation parameter is the mean of x).

Exponential families and change of variables

Assume that x is a random variable with an exponential-family distribution and $y(x)$ is an invertible transformation (see Section 3.2).

The resulting pdmf for y is

$$p(y|\boldsymbol{\eta}) = h_y(y)\, e^{\langle \boldsymbol{\eta}, \boldsymbol{t}_y(y) \rangle} / z(\boldsymbol{\eta})$$

with transformed canonical statistics

$$\boldsymbol{t}_y(y) = \boldsymbol{t}_x(x(y))$$

For a discrete random variable the base function changes to

$$h_y(y) = h_x(x(y))$$

whereas for a continuous random variable the base function becomes

$$h_y(y) = \left|Dx(y)\right| \, h_x(x(y))$$

where $Dx(y)$ is the Jacobian matrix of the inverse transformation $x(y)$.

Thus, for both discrete and continuous random variables the exponential-family form is preserved under a change of variables.

7.4 Univariate exponential families

Table 7.1 lists univariate exponential families, more details about these distributions are found in Chapter 5.

$\mathrm{Bin}(n, \theta)$ is a one-parameter exponential family with n assumed fixed.

For $\mathrm{Bin}(n, \theta)$ and $\mathrm{Ber}(\theta)$ the conventional parameter is

$$\theta = \mathrm{logit}^{-1}(\eta) = \frac{e^\eta}{1 + e^\eta}$$

(logistic function). Conversely, since this is a one-to-one map, the canonical parameter equals

$$\eta = \mathrm{logit}(\theta) = \log\left(\frac{\theta}{1-\theta}\right)$$

Apart from $\mathrm{Bin}(n, \theta)$ all families listed in Table 7.1 have constant base function ($h(x) = 1$).

Furthermore, $\mathrm{Bin}(n, \theta)$, $\mathrm{Ber}(\theta)$ and $\mathrm{Exp}(\theta)$ are NEFs. $N(\mu, \sigma^2)$ with fixed σ^2 (variance), $\mathrm{Gam}(\alpha, \theta)$ with fixed α (shape) and $\mathrm{Wis}(s^2, k)$ with fixed k (shape) are also NEFs.

Table 7.1: Common univariate exponential families

Distribution	$h(x)$	$z(\boldsymbol{\eta})$	$\boldsymbol{\eta}$	$\boldsymbol{t}(x)$	$\boldsymbol{\mu}_{\boldsymbol{t}}$
$\text{Bin}(n, \theta)$	W_2	$(1 + e^{\eta})^n$	$\text{logit}(\theta)$	x	θ
$\text{Ber}(\theta)$	1	$1 + e^{\eta}$	$\text{logit}(\theta)$	x	θ
$\text{Beta}(\alpha_1, \alpha_2)$	1	$B(\eta_1+1, \eta_2+1)$	$\begin{pmatrix}\alpha_1 - 1\\ \alpha_2 - 1\end{pmatrix}$	$\begin{pmatrix}\log x\\ \log(1-x)\end{pmatrix}$	$\begin{pmatrix}\psi^{(0)}(\alpha_1) - \psi^{(0)}(m)\\ \psi^{(0)}(\alpha_2) - \psi^{(0)}(m)\end{pmatrix}$
$N(\mu, \sigma^2)$	1	$(-\pi\eta_2^{-1})^{1/2}$ $\exp(-\frac{1}{4}\eta_1^2\eta_2^{-1})$	$\begin{pmatrix}\sigma^{-2}\mu\\ -\frac{1}{2}\sigma^{-2}\end{pmatrix}$	$\begin{pmatrix}x\\ x^2\end{pmatrix}$	$\begin{pmatrix}\mu\\ \sigma^2 + \mu^2\end{pmatrix}$
$\text{Gam}(\alpha, \theta)$	1	$(-\eta_1)^{-\eta_2-1}$ $\Gamma(\eta_2 + 1)$	$\begin{pmatrix}-1/\theta\\ \alpha - 1\end{pmatrix}$	$\begin{pmatrix}x\\ \log x\end{pmatrix}$	$\begin{pmatrix}\alpha\theta\\ \psi^{(0)}(\alpha) + \log\theta\end{pmatrix}$
$\text{Exp}(\theta)$	1	$-\eta^{-1}$	$-1/\theta$	x	θ
$\text{Wis}(s^2, k)$	1	$(-\eta_1)^{-\eta_2-1}$ $\Gamma(\eta_2 + 1)$	$\begin{pmatrix}-\frac{1}{2}s^{-2}\\ \frac{k}{2} - 1\end{pmatrix}$	$\begin{pmatrix}x\\ \log x\end{pmatrix}$	$\begin{pmatrix}ks^2\\ \psi^{(0)}(\frac{k}{2}) + \log(2s^2)\end{pmatrix}$
$\text{IGam}(\alpha, \beta)$	1	$(-\eta_1)^{\eta_2+1}$ $\Gamma(-\eta_2 - 1)$	$\begin{pmatrix}-\beta\\ -\alpha - 1\end{pmatrix}$	$\begin{pmatrix}x^{-1}\\ \log x\end{pmatrix}$	$\begin{pmatrix}\alpha/\beta\\ -\psi^{(0)}(\alpha) + \log\beta\end{pmatrix}$
$\text{IWis}(\psi, k)$	1	$(-\eta_1)^{\eta_2+1}$ $\Gamma(-\eta_2 - 1)$	$\begin{pmatrix}-\frac{\psi}{2}\\ -\frac{k}{2} - 1\end{pmatrix}$	$\begin{pmatrix}x^{-1}\\ \log x\end{pmatrix}$	$\begin{pmatrix}k/\psi\\ -\psi^{(0)}(\frac{k}{2}) + \log(\frac{\psi}{2})\end{pmatrix}$

Notes:

- $W_2 = \binom{n}{x}$ is the binomial coefficient.
- $B(\alpha_1, \alpha_2)$ is the beta function.
- $m = \alpha_1 + \alpha_2$.
- $\psi^{(0)}(x) = \frac{d}{dx}\log\Gamma(x)$ is the **digamma function**.

7.5 Multivariate exponential families

Table 7.2 lists multvariate exponential families, more details about these distributions are found in Chapter 6.

$\text{Mult}(n, \boldsymbol{\theta})$ is a $(K-1)$-parameter exponential family with n assumed fixed.

For $\text{Mult}(n, \boldsymbol{\theta})$ and $\text{Cat}(\boldsymbol{\theta})$ the conventional parameters are given by

$$\boldsymbol{\theta} = \text{softmax}(\boldsymbol{\eta}) = \frac{(\exp\eta_k)}{\sum_{i=1}^K \exp\eta_i}$$

As the softmax function is invariant against translation and is a many-to-one map, its inverse is not unique and the canonical parameters

$$\boldsymbol{\eta} = (c + \log\theta_k)$$

are determined by the conventional parameters $\boldsymbol{\theta}$ only up to a constant c. This representation using K canonical parameters $\boldsymbol{\eta}$ is non-minimal, hence $\boldsymbol{\eta}$ is not identifiable and different values of $\boldsymbol{\eta}$ can represent the same distribution.

A minimal representation with $K-1$ parameters $\eta_1, \ldots, \eta_{K-1}$ and $\eta_K = 0$ corresponds to $c = -\log\theta_K$ and $\eta_k = \log(\theta_k/\theta_K)$. For $K = 2$ this yields the minimal representations of $\text{Bin}(n, \theta)$ and $\text{Ber}(\theta)$ shown in Table 7.1.

Apart from $\text{Mult}(n, \boldsymbol{\theta})$ all families listed in Table 7.2 have constant base function ($h(\boldsymbol{x}) = 1$).

Furthermore, $\text{Mult}(n, \boldsymbol{\theta})$ and $\text{Cat}(\boldsymbol{\theta})$ are NEFs. $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with fixed $\boldsymbol{\Sigma}$ (variance) and $\text{Wis}(\boldsymbol{S}, k)$ with fixed k (shape) are NEFs as well.

Table 7.2: Common multivariate exponential families

Distribution	$h(\boldsymbol{x})$	$z(\boldsymbol{\eta})$	$\boldsymbol{\eta}$	$\boldsymbol{t}(\boldsymbol{x})$	$\boldsymbol{\mu}_{\boldsymbol{t}}$
$\text{Mult}(n, \boldsymbol{\theta})$	W_K	$(\sum_{k=1}^K \exp \eta_k)^n$	$(c + \log \theta_k)$	$\boldsymbol{x}$	$\boldsymbol{\theta}$
$\text{Cat}(\boldsymbol{\theta})$	1	$\sum_{k=1}^K \exp \eta_k$	$(c + \log \theta_k)$	$\boldsymbol{x}$	$\boldsymbol{\theta}$
$\text{Dir}(\boldsymbol{\alpha})$	1	$B(\boldsymbol{\eta} + 1)$	$(\alpha_k - 1)$	$(\log x_k)$	$(\psi^{(0)}(\alpha_k) - \psi^{(0)}(m))$
$N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	1	$\det(-\pi\boldsymbol{\eta}_2^{-1})^{1/2}$ $\exp(-\frac{1}{4}\boldsymbol{\eta}_1^T\boldsymbol{\eta}_2^{-1}\boldsymbol{\eta}_1)$	$\begin{pmatrix}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\ -\frac{1}{2}\boldsymbol{\Sigma}^{-1}\end{pmatrix}$	$\begin{pmatrix}\boldsymbol{x} \\ \boldsymbol{x}\boldsymbol{x}^T\end{pmatrix}$	$\begin{pmatrix}\boldsymbol{\mu} \\ \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^T\end{pmatrix}$
$\text{Wis}(\boldsymbol{S}, k)$	1	$\det(-\boldsymbol{\eta}_1)^{-\eta_2 - \frac{d+1}{2}}$ $\Gamma_d(\eta_2 + \frac{d+1}{2})$	$\begin{pmatrix}-\frac{1}{2}\boldsymbol{S}^{-1} \\ \frac{k}{2} - \frac{d+1}{2}\end{pmatrix}$	$\begin{pmatrix}\boldsymbol{X} \\ \log\det(\boldsymbol{X})\end{pmatrix}$	$\begin{pmatrix}k\boldsymbol{S} \\ \psi_d^{(0)}(\frac{k}{2}) + \log\det(2\boldsymbol{S})\end{pmatrix}$
$\text{IWis}(\boldsymbol{\Psi}, k)$	1	$\det(-\boldsymbol{\eta}_1)^{\eta_2 + \frac{d+1}{2}}$ $\Gamma_d(-\eta_2 - \frac{d+1}{2})$	$\begin{pmatrix}-\frac{1}{2}\boldsymbol{\Psi} \\ -\frac{k}{2} - \frac{d+1}{2}\end{pmatrix}$	$\begin{pmatrix}\boldsymbol{X}^{-1} \\ \log\det(\boldsymbol{X})\end{pmatrix}$	$\begin{pmatrix}k\boldsymbol{\Psi}^{-1} \\ -\psi_d^{(0)}(\frac{k}{2}) + \log\det(\frac{\boldsymbol{\Psi}}{2})\end{pmatrix}$

Notes:

- $W_K = \binom{n}{x_1,\ldots,x_K}$ is the multinomial coefficient for K groups.
- $B(\boldsymbol{\alpha})$ is the multivariate beta function.
- $m = \sum_{i=k}^K \alpha_k$.
- $\psi_d^{(0)} = \frac{d}{dx}\log\Gamma_d(x) = \sum_{i=1}^d \psi^{(0)}(x - (i-1)/2)$ is the **multivariate digamma function**.

See also: Exponential family (Wikipedia) and Efron (2022).

Bibliography

Efron, B. 2022. *Exponential Families in Theory and Practise*. Cambridge University Press. https://doi.org/10.1017/9781108773157.

Mardia, K. V., J. T. Kent, and J. M. Bibby. 1979. *Multivariate Analysis*. Academic Press.

Whittle, P. 2000. *Probability via Expectation*. 3rd ed. Springer. https://doi.org/10.1007/978-1-4612-0509-8.