The Book of Statistical Proofs

https://statproofbook.github.io/
StatProofBook@gmail.com

2020-03-07, 16:20

Contents

Chapter I

General Theorems

1 Probability theory

1.1 Probability distributions

1.1.1 Probability mass function

Definition: Let X be a discrete random variable (→ Definition "rvar") with possible outcomes $\mathcal{X}$. Then, $f_X(x) : \mathbb{R} \rightarrow [0,1]$ is the probability mass function of X, if

$$\mathrm{Pr}(X = x) = f_X(x) \tag{1}$$

for all $x \in \mathcal{X}$ and

$$\sum_{x \in \mathcal{X}} f_X(x) = 1 \; . \tag{2}$$

Sources:
- Wikipedia (2020): "Probability mass function"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-13; URL: https://en.wikipedia.org/wiki/Probability_mass_function.

Metadata: ID: D9 | shortcut: pmf | author: JoramSoch | date: 2020-02-13, 19:09.

1.1.2 Probability density function

Definition: Let X be a continuous random variable (→ Definition "rvar") with possible outcomes $\mathcal{X}$. Then, $f_X(x) : \mathbb{R} \rightarrow \mathbb{R}$ is the probability density function of X, if

$$f_X(x) \geq 0 \tag{1}$$

for all $x \in \mathbb{R}$,

$$\mathrm{Pr}(X \in A) = \int_A f_X(x) \, \mathrm{d}x \tag{2}$$

for any $A \subset \mathcal{X}$ and

$$\int_{\mathcal{X}} f_X(x) \, \mathrm{d}x = 1 \; . \tag{3}$$

Sources:
- Wikipedia (2020): "Probability density function"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-13; URL: https://en.wikipedia.org/wiki/Probability_density_function.

Metadata: ID: D10 | shortcut: pdf | author: JoramSoch | date: 2020-02-13, 19:26.

1.1.3 Cumulative distribution function

Definition:
1) Let X be a discrete random variable (→ Definition "rvar") with possible outcomes $\mathcal{X}$ and the probability mass function (→ Definition I/1.1.1) $f_X(x)$. Then, the function $F_X(x) : \mathbb{R} \rightarrow [0,1]$ with

$$F_X(x) = \sum_{\substack{z \in \mathcal{X} \\ z \leq x}} f_X(z) \tag{1}$$

is the cumulative distribution function of X.
2) Let X be a scalar continuous random variable (→ Definition "rvar") with the probability density function (→ Definition I/1.1.2) $f_X(x)$. Then, the function $F_X(x) : \mathbb{R} \to [0, 1]$ with

$$F_X(x) = \int_{-\infty}^{x} f_X(z) \, \mathrm{d}x \tag{2}$$

is the cumulative distribution function of X.

Sources:
- Wikipedia (2020): "Probability density function"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-17; URL: https://en.wikipedia.org/wiki/Cumulative_distribution_function#Definition.

Metadata: ID: D13 | shortcut: cdf | author: JoramSoch | date: 2020-02-17, 22:07.

1.1.4 Quantile function

Definition: Let X be a random variable (→ Definition "rvar") with the cumulative distribution function (→ Definition I/1.1.3) (CDF) $F_X(x)$. Then, the function $Q_X(p) : [0, 1] \to \mathbb{R}$ which is the inverse CDF

$$Q_X(p) = F_X^{-1}(x) \tag{1}$$

is the quantile function (QF) of X. More precisely, the QF is the function that, for a given quantile $p \in [0, 1]$, returns the smallest x for which $F_X(x) = p$:

$$Q_X(p) = \min \left\{ x \in \mathbb{R} \,|\, F_X(x) = p \right\} \; . \tag{2}$$

Sources:
- Wikipedia (2020): "Probability density function"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-17; URL: https://en.wikipedia.org/wiki/Quantile_function#Definition.

Metadata: ID: D14 | shortcut: qf | author: JoramSoch | date: 2020-02-17, 22:18.

1.1.5 Moment-generating function

Definition:
1) The moment-generating function of a random variable (→ Definition "rvar") $X \in \mathbb{R}$ is

$$M_X(t) = \mathrm{E}\left[e^{tX} \right], \quad t \in \mathbb{R} \; . \tag{1}$$

2) The moment-generating function of a random vector (→ Definition "rvec") $X \in \mathbb{R}^n$ is

$$M_X(t) = \mathrm{E}\left[e^{t^\mathrm{T} X} \right], \quad t \in \mathbb{R}^n \; . \tag{2}$$

Sources:

- Wikipedia (2020): “Moment-generating function”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-01-22; URL: https://en.wikipedia.org/wiki/Moment-generating_function#Definition.

Metadata: ID: D2 | shortcut: mgf | author: JoramSoch | date: 2020-01-22, 10:58.

1.2 Expected value

1.2.1 Definition

Definition:
1) The expected value (or, mean) of a discrete random variable (→ Definition “rvar”) X with domain $\mathcal{X}$ is

$$\mathrm{E}(X) = \sum_{x \in \mathcal{X}} x \cdot f_X(x) \tag{1}$$

where $f_X(x)$ is the probability mass function (→ Definition I/1.1.1) of X.

2) The expected value (or, mean) of a continuous random variable (→ Definition “rvar”) X with domain $\mathcal{X}$ is

$$\mathrm{E}(X) = \int_{\mathcal{X}} x \cdot f_X(x) \, \mathrm{d}x \tag{2}$$

where $f_X(x)$ is the probability density function (→ Definition I/1.1.2) of X.

Sources:
- Wikipedia (2020): “Expected value”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-13; URL: https://en.wikipedia.org/wiki/Expected_value#Definition.

Metadata: ID: D11 | shortcut: mean | author: JoramSoch | date: 2020-02-13, 19:38.

1.2.2 Non-negativity

Theorem: If a random variable (→ Definition “rvar”) is strictly non-negative, its expected value (→ Definition I/1.2.1) is also non-negative, i.e.

$$\mathrm{E}(X) \geq 0, \quad \text{if} \quad X \geq 0 \; . \tag{1}$$

Proof:
1) If $X \geq 0$ is a discrete random variable, then, because the probability mass function (→ Definition I/1.1.1) is always non-negative, all the addends in

$$\mathrm{E}(X) = \sum_{x \in \mathcal{X}} x \cdot f_X(x) \tag{2}$$

are non-negative, thus the entire sum must be non-negative.

2) If $X \geq 0$ is a continuous random variable, then, because the probability density function (→ Definition I/1.1.2) is always non-negative, the integrand in

$$\mathrm{E}(X) = \int_{\mathcal{X}} x \cdot f_X(x) \, \mathrm{d}x \tag{3}$$

is strictly non-negative, thus the term on the right-hand side is a Lebesgue integral, so that the result on the left-hand side must be non-negative.

Sources:

- Wikipedia (2020): "Expected value"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-13; URL: https://en.wikipedia.org/wiki/Expected_value#Basic_properties.

Metadata: ID: P52 | shortcut: mean-nonneg | author: JoramSoch | date: 2020-02-13, 20:14.

1.2.3 Linearity

Theorem: The expected value (→ Definition I/1.2.1) is a linear operator, i.e.

$$\begin{aligned} \mathrm{E}(X+Y) &= \mathrm{E}(X) + \mathrm{E}(Y) \\ \mathrm{E}(a\,X) &= a\,\mathrm{E}(X) \end{aligned} \tag{1}$$

for random variables (→ Definition "rvar") X and Y and a constant a.

Proof:
1) If X and Y are discrete random variables, the expected value (→ Definition I/1.2.1) is

$$\mathrm{E}(X) = \sum_{x \in \mathcal{X}} x \cdot f_X(x) \tag{2}$$

and the law of marginal probability (→ Definition "prob-marg") states that

$$p(x) = \sum_{y \in \mathcal{Y}} p(x,y) \; . \tag{3}$$

Applying this, we have

$$\begin{aligned} \mathrm{E}(X+Y) &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} (x+y) \cdot f_{X,Y}(x,y) \\ &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} x \cdot f_{X,Y}(x,y) + \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} y \cdot f_{X,Y}(x,y) \\ &= \sum_{x \in \mathcal{X}} x \sum_{y \in \mathcal{Y}} f_{X,Y}(x,y) + \sum_{y \in \mathcal{Y}} y \sum_{x \in \mathcal{X}} f_{X,Y}(x,y) \\ &\overset{(3)}{=} \sum_{x \in \mathcal{X}} x \cdot f_X(x) + \sum_{y \in \mathcal{Y}} y \cdot f_Y(y) \\ &\overset{(2)}{=} \mathrm{E}(X) + \mathrm{E}(Y) \end{aligned} \tag{4}$$

as well as

$$
\begin{aligned}
\mathrm{E}(a\,X) &= \sum_{x \in \mathcal{X}} a\,x \cdot f_X(x) \\
&= a \sum_{x \in \mathcal{X}} x \cdot f_X(x) \\
&\overset{(2)}{=} a\,\mathrm{E}(X) \; .
\end{aligned} \tag{5}
$$

2) If X and Y are continuous random variables, the expected value (→ Definition I/1.2.1) is

$$
\mathrm{E}(X) = \int_{\mathcal{X}} x \cdot f_X(x) \, \mathrm{d}x \tag{6}
$$

and the law of marginal probability (→ Definition "prob-marg") states that

$$
p(x) = \int_{\mathcal{Y}} p(x,y) \, \mathrm{d}y \; . \tag{7}
$$

Applying this, we have

$$
\begin{aligned}
\mathrm{E}(X+Y) &= \int_{\mathcal{X}} \int_{\mathcal{Y}} (x+y) \cdot f_{X,Y}(x,y) \, \mathrm{d}y \, \mathrm{d}x \\
&= \int_{\mathcal{X}} \int_{\mathcal{Y}} x \cdot f_{X,Y}(x,y) \, \mathrm{d}y \, \mathrm{d}x + \int_{\mathcal{X}} \int_{\mathcal{Y}} y \cdot f_{X,Y}(x,y) \, \mathrm{d}y \, \mathrm{d}x \\
&= \int_{\mathcal{X}} x \int_{\mathcal{Y}} f_{X,Y}(x,y) \, \mathrm{d}y \, \mathrm{d}x + \int_{\mathcal{Y}} y \int_{\mathcal{X}} f_{X,Y}(x,y) \, \mathrm{d}x \, \mathrm{d}y \\
&\overset{(7)}{=} \int_{\mathcal{X}} x \cdot f_X(x) \, \mathrm{d}x + \int_{\mathcal{Y}} y \cdot f_Y(y) \, \mathrm{d}y \\
&\overset{(6)}{=} \mathrm{E}(X) + \mathrm{E}(Y)
\end{aligned} \tag{8}
$$

as well as

$$
\begin{aligned}
\mathrm{E}(a\,X) &= \int_{\mathcal{X}} a\,x \cdot f_X(x) \, \mathrm{d}x \\
&= a \int_{\mathcal{X}} x \cdot f_X(x) \, \mathrm{d}x \\
&\overset{(6)}{=} a\,\mathrm{E}(X) \; .
\end{aligned} \tag{9}
$$

Collectively, this shows that both requirements for linearity are fulfilled for the expected value, for discrete as well as for continuous random variables.

Sources:

- Wikipedia (2020): "Expected value"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-13; URL: https://en.wikipedia.org/wiki/Expected_value#Basic_properties.
- Michael B, Kuldeep Guha Mazumder, Geoff Pilling et al. (2020): "Linearity of Expectation"; in: *brilliant.org*; URL: https://brilliant.org/wiki/linearity-of-expectation/.

Metadata: ID: P53 | shortcut: mean-lin | author: JoramSoch | date: 2020-02-13, 21:08.

1.2.4 Monotonicity

Theorem: The expected value (→ Definition I/1.2.1) is monotonic, i.e.

$$\mathrm{E}(X) \leq \mathrm{E}(Y), \quad \text{if} \quad X \leq Y \; . \tag{1}$$

Proof: Let $Z = Y - X$. Due to the linearity of the expected value (→ Proof I/1.2.3), we have

$$\mathrm{E}(Z) = \mathrm{E}(Y - X) = \mathrm{E}(Y) - \mathrm{E}(X) \; . \tag{2}$$

With the non-negativity property of the expected value (→ Proof I/1.2.2), it also holds that

$$Z \geq 0 \quad \Rightarrow \quad \mathrm{E}(Z) \geq 0 \; . \tag{3}$$

Together with (2), this yields

$$\mathrm{E}(Y) - \mathrm{E}(X) \geq 0 \; . \tag{4}$$

Sources:
- Wikipedia (2020): "Expected value"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-17; URL: https://en.wikipedia.org/wiki/Expected_value#Basic_properties.

Metadata: ID: P54 | shortcut: mean-mono | author: JoramSoch | date: 2020-02-17, 21:00.

1.2.5 (Non-)Multiplicitavity

Theorem:
1) If two random variables (→ Definition "rvar") X and Y are independent (→ Definition "ind"), the expected value (→ Definition I/1.2.1) is multiplicative, i.e.

$$\mathrm{E}(X\,Y) = \mathrm{E}(X)\,\mathrm{E}(Y) \; . \tag{1}$$

2) If two random variables (→ Definition "rvar") X and Y are dependent (→ Definition "ind"), the expected value (→ Definition I/1.2.1) is not necessarily multiplicative, i.e. there exist X and Y such that

$$\mathrm{E}(X\,Y) \neq \mathrm{E}(X)\,\mathrm{E}(Y) \; . \tag{2}$$

Proof:
1) If X and Y are independent (→ Definition "ind"), it holds that

$$p(x,y) = p(x)\,p(y) \quad \text{for all} \quad x \in \mathcal{X}, y \in \mathcal{Y} \; . \tag{3}$$

Applying this to the expected value for discrete random variables (→ Definition I/1.2.1), we have

$$
\begin{aligned}
\mathrm{E}(X\,Y) &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} (x \cdot y) \cdot f_{X,Y}(x,y) \\
&\overset{(3)}{=} \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} (x \cdot y) \cdot (f_X(x) \cdot f_Y(y)) \\
&= \sum_{x \in \mathcal{X}} x \cdot f_X(x) \sum_{y \in \mathcal{Y}} y \cdot f_Y(y) \\
&= \sum_{x \in \mathcal{X}} x \cdot f_X(x) \cdot \mathrm{E}(Y) \\
&= \mathrm{E}(X)\,\mathrm{E}(Y) \; .
\end{aligned} \tag{4}
$$

And applying it to the expected value for continuous random variables (→ Definition I/1.2.1), we have

$$
\begin{aligned}
\mathrm{E}(X\,Y) &= \int_{\mathcal{X}} \int_{\mathcal{Y}} (x \cdot y) \cdot f_{X,Y}(x,y) \,\mathrm{d}y \,\mathrm{d}x \\
&\overset{(3)}{=} \int_{\mathcal{X}} \int_{\mathcal{Y}} (x \cdot y) \cdot (f_X(x) \cdot f_Y(y)) \,\mathrm{d}y \,\mathrm{d}x \\
&= \int_{\mathcal{X}} x \cdot f_X(x) \int_{\mathcal{Y}} y \cdot f_Y(y) \,\mathrm{d}y \,\mathrm{d}x \\
&= \int_{\mathcal{X}} x \cdot f_X(x) \cdot \mathrm{E}(Y) \,\mathrm{d}x \\
&= \mathrm{E}(X)\,\mathrm{E}(Y) \; .
\end{aligned} \tag{5}
$$

2) Let X and Y be Bernoulli random variables (→ Definition "bern") with the following joint probability mass function (→ Definition I/1.1.1)

$$
\begin{aligned}
p(X=0, Y=0) &= 1/2 \\
p(X=0, Y=1) &= 0 \\
p(X=1, Y=0) &= 0 \\
p(X=1, Y=1) &= 1/2
\end{aligned} \tag{6}
$$

and thus, the following marginal probabilities:

$$
\begin{aligned}
p(X=0) = p(X=1) &= 1/2 \\
p(Y=0) = p(Y=1) &= 1/2 \; .
\end{aligned} \tag{7}
$$

Then, X and Y are dependent, because

$$
p(X=0, Y=1) \overset{(6)}{=} 0 \neq \frac{1}{2} \cdot \frac{1}{2} \overset{(7)}{=} p(X=0)\,p(Y=1) \; , \tag{8}
$$

and the expected value of their product is

$$
\begin{aligned}
\mathrm{E}(X\,Y) &= \sum_{x \in \{0,1\}} \sum_{y \in \{0,1\}} (x \cdot y) \cdot p(x,y) \\
&= (1 \cdot 1) \cdot p(X=1, Y=1) \\
&\overset{(6)}{=} \frac{1}{2}
\end{aligned} \tag{9}
$$

while the product of their expected values is

$$
\begin{aligned}
\mathrm{E}(X)\,\mathrm{E}(Y) &= \left(\sum_{x \in \{0,1\}} x \cdot p(x) \right) \cdot \left(\sum_{y \in \{0,1\}} y \cdot p(y) \right) \\
&= (1 \cdot p(X=1)) \cdot (1 \cdot p(Y=1)) \\
&\overset{(7)}{=} \frac{1}{4}
\end{aligned} \tag{10}
$$

and thus,

$$
\mathrm{E}(X\,Y) \neq \mathrm{E}(X)\,\mathrm{E}(Y) \; . \tag{11}
$$

Sources:

- Wikipedia (2020): "Expected value"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-17; URL: https://en.wikipedia.org/wiki/Expected_value#Basic_properties.

Metadata: ID: P55 | shortcut: mean-mult | author: JoramSoch | date: 2020-02-17, 21:51.

1.3 Variance

1.3.1 Definition

Definition: The variance of a random variable (→ Definition "rvar") X is defined as the expected value (→ Definition I/1.2.1) of the squared deviation from its expected value (→ Definition I/1.2.1):

$$
\mathrm{Var}(X) = \mathrm{E}\left[(X - \mathrm{E}(X))^2 \right] \; . \tag{1}
$$

Sources:

- Wikipedia (2020): "Variance"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-13; URL: https://en.wikipedia.org/wiki/Variance#Definition.

Metadata: ID: D12 | shortcut: var | author: JoramSoch | date: 2020-02-13, 19:55.

2 Bayesian statistics

2.1 Bayesian inference

2.1.1 Bayes' theorem

Theorem: Let A and B be two arbitrary statements about random variables (→ Definition "rvar"), such as statements about the presence or absence of an event or about the value of a scalar, vector or matrix. Then, the conditional probability that A is true, given that B is true, is equal to

$$p(A|B) = \frac{p(B|A)\, p(A)}{p(B)} \; . \tag{1}$$

Proof: The conditional probability (→ Definition "prob-cond") is defined as the ratio of joint probability (→ Definition "prob-joint"), i.e. the probability of both statements being true, and marginal probability (→ Definition "prob-marg"), i.e. the probability of only the second one being true:

$$p(A|B) = \frac{p(A,B)}{p(B)} \; . \tag{2}$$

It can also be written down for the reverse situation, i.e. to calculate the probability that B is true, given that A is true:

$$p(B|A) = \frac{p(A,B)}{p(A)} \; . \tag{3}$$

Both equations can be rearranged for the joint probability

$$p(A|B)\, p(B) \overset{(2)}{=} p(A,B) \overset{(3)}{=} p(B|A)\, p(A) \tag{4}$$

from which Bayes' theorem can be directly derived:

$$p(A|B) \overset{(4)}{=} \frac{p(B|A)\, p(A)}{p(B)} \; . \tag{5}$$

Sources:

- Koch, Karl-Rudolf (2007): "Rules of Probability"; in: *Introduction to Bayesian Statistics*, Springer, Berlin/Heidelberg, 2007, pp. 6/13, eqs. 2.12/2.38; URL: https://www.springer.com/de/book/9783540727231; DOI: 10.1007/978-3-540-72726-2.

Metadata: ID: P4 | shortcut: bayes-th | author: JoramSoch | date: 2019-09-27, 16:24.

2.1.2 Bayes' rule

Theorem: Let A_1, A_2 and B be arbitrary statements about random variables (→ Definition "rvar") where A_1 and A_2 are mutually exclusive. Then, Bayes' rule states that the posterior odds (→ Definition "post-odd") are equal to the Bayes factor (→ Definition "bf") times the prior odds (→ Definition "prior-odd"), i.e.

$$\frac{p(A_1|B)}{p(A_2|B)} = \frac{p(B|A_1)}{p(B|A_2)} \cdot \frac{p(A_1)}{p(A_2)} \; . \tag{1}$$

Proof: Using Bayes' theorem (→ Proof I/2.1.1), the conditional probabilities (→ Definition "cp") on the left are given by

$$p(A_1|B) = \frac{p(B|A_1) \cdot p(A_1)}{p(B)} \tag{2}$$

$$p(A_2|B) = \frac{p(B|A_2) \cdot p(A_2)}{p(B)} \; . \tag{3}$$

Dividing the two conditional probabilities by each other

$$\begin{split} \frac{p(A_1|B)}{p(A_2|B)} &= \frac{p(B|A_1) \cdot p(A_1)/p(B)}{p(B|A_2) \cdot p(A_2)/p(B)} \\ &= \frac{p(B|A_1)}{p(B|A_2)} \cdot \frac{p(A_1)}{p(A_2)} \; , \end{split} \tag{4}$$

one obtains the posterior odds ratio as given by the theorem.

Sources:
- Wikipedia (2019): "Bayes' theorem"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-01-06; URL: https://en.wikipedia.org/wiki/Bayes%27_theorem#Bayes%E2%80%99_rule.

Metadata: ID: P12 | shortcut: bayes-rule | author: JoramSoch | date: 2020-01-06, 20:55.

2.2 Probabilistic modeling

2.2.1 Generative model

Definition: Consider measured data y and some unknown latent parameters θ. A statement about the distribution of y given θ is called a generative model m:

$$m : \; y \sim \mathcal{D}(\theta) \; . \tag{1}$$

Sources:
- original work

Metadata: ID: D27 | shortcut: gm | author: JoramSoch | date: 2020-03-03, 15:50.

2.2.2 Likelihood function

Definition: Let there be a generative model (→ Definition I/2.2.1) m describing measured data y using model parameters θ. Then, the probability density function (→ Definition I/1.1.2) of the distribution of y given θ is called the likelihood function of m:

$$\mathcal{L}_m(\theta) = p(y|\theta, m) = \mathcal{D}(y; \theta) \; . \tag{1}$$

Sources:
- original work

Metadata: ID: D28 | shortcut: lf | author: JoramSoch | date: 2020-03-03, 15:50.

2.2.3 Prior distribution

Definition: Consider measured data y and some unknown latent parameters θ. A distribution of θ unconditional on y is called a prior distribution:

$$\theta \sim \mathcal{D}(\lambda) \; . \tag{1}$$

The parameters λ of this distribution are called the prior hyperparameters and the probability density function (→ Definition I/1.1.2) is called the prior density:

$$p(\theta|m) = \mathcal{D}(\theta; \lambda) \; . \tag{2}$$

Sources:
- original work

Metadata: ID: D29 | shortcut: prior | author: JoramSoch | date: 2020-03-03, 16:09.

2.2.4 Full probability model

Definition: Consider measured data y and some unknown latent parameters θ. The combination of a generative model (→ Definition I/2.2.1) for y and a prior distribution (→ Definition I/2.2.3) on θ is called a full probability model m:

$$m: \; y \sim \mathcal{D}(\theta), \, \theta \sim \mathcal{D}(\lambda) \; . \tag{1}$$

Sources:
- original work

Metadata: ID: D30 | shortcut: fpm | author: JoramSoch | date: 2020-03-03, 16:16.

2.2.5 Joint likelihood

Definition: Let there be a generative model (→ Definition I/2.2.1) m describing measured data y using model parameters θ and a prior distribution (→ Definition I/2.2.3) on θ. Then, the joint probability (→ Definition "prob-joint") density function (→ Definition I/1.1.2) of y and θ is called the joint likelihood:

$$p(y,\theta|m) = p(y|\theta,m) \, p(\theta|m) \; . \tag{1}$$

Sources:
- original work

Metadata: ID: D31 | shortcut: jl | author: JoramSoch | date: 2020-03-03, 16:36.

2.2.6 Posterior distribution

Definition: Consider measured data y and some unknown latent parameters θ. The distribution of θ conditional on y is called the posterior distribution:

$$\theta|y \sim \mathcal{D}(\phi) \; . \tag{1}$$

The parameters ϕ of this distribution are called the posterior hyperparameters and the probability density function (→ Definition I/1.1.2) is called the posterior density:

$$p(\theta|y,m) = \mathcal{D}(\theta;\phi) \; . \tag{2}$$

Sources:
- original work

Metadata: ID: D32 | shortcut: post | author: JoramSoch | date: 2020-03-03, 16:43.

2.2.7 Marginal likelihood

Definition: Let there be a generative model (→ Definition I/2.2.1) m describing measured data y using model parameters θ and a prior distribution (→ Definition I/2.2.3) on θ. Then, the marginal probability (→ Definition "mp") density function (→ Definition I/1.1.2) of y across the parameter space Θ is called the marginal likelihood:

$$p(y|m) = \int_{\Theta} p(y|\theta,m) \, p(\theta|m) \, \mathrm{d}\theta \; . \tag{1}$$

Sources:
- original work

Metadata: ID: D33 | shortcut: ml | author: JoramSoch | date: 2020-03-03, 16:49.

3 Estimation theory

3.1 Point estimates

3.1.1 Partition of the mean squared error into bias and variance

Theorem: The mean squared error (→ Definition "mse") can be partitioned into variance and squared bias

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) - \mathrm{Bias}(\hat{\theta}, \theta)^2 \tag{1}$$

where the variance (→ Definition I/1.3.1) is given by

$$\mathrm{Var}(\hat{\theta}) = \mathbb{E}_{\hat{\theta}}\left[\left(\hat{\theta} - \mathbb{E}_{\hat{\theta}}(\hat{\theta})\right)^2\right] \tag{2}$$

and the bias (→ Definition "bias") is given by

$$\mathrm{Bias}(\hat{\theta}, \theta) = \left(\mathbb{E}_{\hat{\theta}}(\hat{\theta}) - \theta\right) \; . \tag{3}$$

Proof: The mean squared error (MSE) is defined as (→ Definition "mse") the expected value (→ Definition I/1.2.1) of the squared deviation of the estimated value $\hat{\theta}$ from the true value θ of a parameter, over all values $\hat{\theta}$:

$$\mathrm{MSE}(\hat{\theta}) = \mathbb{E}_{\hat{\theta}}\left[\left(\hat{\theta} - \theta\right)^2\right] \; . \tag{4}$$

This formula can be evaluated in the following way:

$$\begin{aligned}
\mathrm{MSE}(\hat{\theta}) &= \mathbb{E}_{\hat{\theta}}\left[\left(\hat{\theta} - \theta\right)^2\right] \\
&= \mathbb{E}_{\hat{\theta}}\left[\left(\hat{\theta} - \mathbb{E}_{\hat{\theta}}(\hat{\theta}) + \mathbb{E}_{\hat{\theta}}(\hat{\theta}) - \theta\right)^2\right] \\
&= \mathbb{E}_{\hat{\theta}}\left[\left(\hat{\theta} - \mathbb{E}_{\hat{\theta}}(\hat{\theta})\right)^2 + 2\left(\hat{\theta} - \mathbb{E}_{\hat{\theta}}(\hat{\theta})\right)\left(\mathbb{E}_{\hat{\theta}}(\hat{\theta}) - \theta\right) + \left(\mathbb{E}_{\hat{\theta}}(\hat{\theta}) - \theta\right)^2\right] \\
&= \mathbb{E}_{\hat{\theta}}\left[\left(\hat{\theta} - \mathbb{E}_{\hat{\theta}}(\hat{\theta})\right)^2\right] + \mathbb{E}_{\hat{\theta}}\left[2\left(\hat{\theta} - \mathbb{E}_{\hat{\theta}}(\hat{\theta})\right)\left(\mathbb{E}_{\hat{\theta}}(\hat{\theta}) - \theta\right)\right] + \mathbb{E}_{\hat{\theta}}\left[\left(\mathbb{E}_{\hat{\theta}}(\hat{\theta}) - \theta\right)^2\right] \; .
\end{aligned} \tag{5}$$

Because $\mathbb{E}_{\hat{\theta}}(\hat{\theta}) - \theta$ is constant as a function of $\hat{\theta}$, we have:

$$\begin{aligned}
\mathrm{MSE}(\hat{\theta}) &= \mathbb{E}_{\hat{\theta}}\left[\left(\hat{\theta} - \mathbb{E}_{\hat{\theta}}(\hat{\theta})\right)^2\right] + 2\left(\mathbb{E}_{\hat{\theta}}(\hat{\theta}) - \theta\right)\mathbb{E}_{\hat{\theta}}\left[\hat{\theta} - \mathbb{E}_{\hat{\theta}}(\hat{\theta})\right] + \left(\mathbb{E}_{\hat{\theta}}(\hat{\theta}) - \theta\right)^2 \\
&= \mathbb{E}_{\hat{\theta}}\left[\left(\hat{\theta} - \mathbb{E}_{\hat{\theta}}(\hat{\theta})\right)^2\right] + 2\left(\mathbb{E}_{\hat{\theta}}(\hat{\theta}) - \theta\right)\left(\mathbb{E}_{\hat{\theta}}(\hat{\theta}) - \mathbb{E}_{\hat{\theta}}(\hat{\theta})\right) + \left(\mathbb{E}_{\hat{\theta}}(\hat{\theta}) - \theta\right)^2 \\
&= \mathbb{E}_{\hat{\theta}}\left[\left(\hat{\theta} - \mathbb{E}_{\hat{\theta}}(\hat{\theta})\right)^2\right] + \left(\mathbb{E}_{\hat{\theta}}(\hat{\theta}) - \theta\right)^2 \; .
\end{aligned} \tag{6}$$

This proofs the partition given by (1).

Sources:

- Wikipedia (2019): "Mean squared error"; in: *Wikipedia, the free encyclopedia*, retrieved on 2019-11-27; URL: https://en.wikipedia.org/wiki/Mean_squared_error#Proof_of_variance_and_bias_relationship.

Metadata: ID: P5 | shortcut: mse-bnv | author: JoramSoch | date: 2019-11-27, 14:26.

3.2 Interval estimates

3.2.1 Construction of confidence intervals using Wilks' theorem

Theorem: Let m be a generative model (→ Definition I/2.2.1) for measured data y with model parameters θ, consisting of a parameter of interest ϕ and nuisance parameters λ:

$$m : p(y|\theta) = \mathcal{D}(y;\theta), \quad \theta = \{\phi, \lambda\} \; . \tag{1}$$

Further, let $\hat{\theta}$ be an estimate of θ, obtained using maximum-likelihood-estimation (→ Definition "mle"):

$$\hat{\theta} = \arg\max_{\theta} \log p(y|\theta), \quad \hat{\theta} = \left\{ \hat{\phi}, \hat{\lambda} \right\} \; . \tag{2}$$

Then, an asymptotic confidence interval (→ Definition "ci") for θ is given by

$$\mathrm{CI}_{1-\alpha}(\hat{\phi}) = \left\{ \phi \, | \, \log p(y|\phi,\hat{\lambda}) \geq \log p(y|\hat{\phi},\hat{\lambda}) - \frac{1}{2}\chi^2_{1,1-\alpha} \right\} \tag{3}$$

where $1-\alpha$ is the confidence level and $\chi^2_{1,1-\alpha}$ is the $(1-\alpha)$-quantile of the chi-squared distribution (→ Definition "chi2") with 1 degree of freedom (→ Definition "dof").

Proof: The confidence interval (→ Definition "ci") is defined as the interval that, under infinitely repeated random experiments (→ Definition "rexp"), contains the true parameter value with a certain probability.
Let us define the likelihood ratio (→ Definition "lr")

$$\Lambda(\phi) = \frac{p(y|\phi,\hat{\lambda})}{p(y|\hat{\phi},\hat{\lambda})} \tag{4}$$

and compute the log-likelihood ratio (→ Definition "llr")

$$\log \Lambda(\phi) = \log p(y|\phi,\hat{\lambda}) - \log p(y|\hat{\phi},\hat{\lambda}) \; . \tag{5}$$

[Wilks' theorem](llr-wilks) states that, when comparing two statistical models with parameter spaces Θ_1 and $\Theta_0 \subset \Theta_1$, as the sample size approaches infinity, the quantity calculated as -2 times the log-ratio of maximum likelihoods follows a chi-squared distribution (→ Definition "chi2"), if the null hypothesis is true:

$$H_0 : \theta \in \Theta_0 \quad \Rightarrow \quad -2 \log \frac{\max_{\theta \in \Theta_0} p(y|\theta)}{\max_{\theta \in \Theta_1} p(y|\theta)} \sim \chi^2_{\Delta k} \tag{6}$$

where Δk is the difference in dimensionality between Θ_0 and Θ_1. Applied to our example in (5), we note that $\Theta_1 = \left\{ \phi, \hat{\phi} \right\}$ and $\Theta_0 = \{\phi\}$, such that $\Delta k = 1$ and Wilks' theorem implies:

$$-2 \log \Lambda(\phi) \sim \chi^2_1 \; . \tag{7}$$

Using the quantile function (→ Definition I/1.1.4) $\chi^2_{k,p}$ of the chi-squared distribution (→ Definition "chi2"), an $(1-\alpha)$-confidence interval is therefore given by all values ϕ that satisfy

$$-2 \log \Lambda(\phi) \leq \chi^2_{1,1-\alpha} \; . \tag{8}$$

Applying (5) and rearranging, we can evaluate

$$\begin{split}
-2 \left[\log p(y|\phi,\hat{\lambda}) - \log p(y|\hat{\phi},\hat{\lambda}) \right] &\leq \chi^2_{1,1-\alpha} \\
\log p(y|\phi,\hat{\lambda}) - \log p(y|\hat{\phi},\hat{\lambda}) &\geq -\frac{1}{2} \chi^2_{1,1-\alpha} \\
\log p(y|\phi,\hat{\lambda}) &\geq \log p(y|\hat{\phi},\hat{\lambda}) - \frac{1}{2} \chi^2_{1,1-\alpha}
\end{split} \tag{9}$$

which is equivalent to the confidence interval given by (3).

Sources:

- Wikipedia (2020): "Confidence interval"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-19; URL: https://en.wikipedia.org/wiki/Confidence_interval#Methods_of_derivation.
- Wikipedia (2020): "Likelihood-ratio test"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-19; URL: https://en.wikipedia.org/wiki/Likelihood-ratio_test#Definition.
- Wikipedia (2020): "Wilks' theorem"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-19; URL: https://en.wikipedia.org/wiki/Wilks%27_theorem.

Metadata: ID: P56 | shortcut: ci-wilks | author: JoramSoch | date: 2020-02-19, 17:15.

4 Information theory

4.1 Shannon entropy

4.1.1 Definition

Definition: Let X be a discrete random variable (→ Definition "rvar") with possible outcomes $x_i, \; i = 1, \ldots, k$ and the (observed or assumed) probability mass function (→ Definition I/1.1.1) $p(x) = f_X(x)$. Then, the entropy (also referred to as "Shannon entropy") of X is defined as

$$\mathrm{H}(X) = -\sum_{i=1}^{k} p(x_i) \cdot \log_b p(x_i) \tag{1}$$

where b is the base of the logarithm specifying in which unit the entropy is determined.

Sources:

- Shannon CE (1948): "A Mathematical Theory of Communication"; in: *Bell System Technical Journal*, vol. 27, iss. 3, pp. 379-423; URL: https://ieeexplore.ieee.org/document/6773024; DOI: 10.1002/j.1538-7305.1948.tb01338.x.

Metadata: ID: D15 | shortcut: ent | author: JoramSoch | date: 2020-02-19, 17:36.

4.1.2 Non-negativity

Theorem: The entropy of a discrete random variable (→ Definition "rvar") is a non-negative number:

$$\mathrm{H}(X) \geq 0 \; . \tag{1}$$

Proof: The entropy of a discrete random variable (→ Definition I/4.1.1) is defined as

$$\mathrm{H}(X) = -\sum_{i=1}^{k} p(x_i) \cdot \log_b p(x_i) \tag{2}$$

The minus sign can be moved into the sum:

$$\mathrm{H}(X) = \sum_{i=1}^{k} \left[p(x_i) \cdot (-\log_b p(x_i)) \right] \tag{3}$$

Because the co-domain of probability mass functions (→ Definition I/1.1.1) is $[0, 1]$, we can deduce:

$$\begin{array}{rcccl} 0 & \leq & p(x_i) & \leq & 1 \\ -\infty & \leq & \log_b p(x_i) & \leq & 0 \\ 0 & \leq & -\log_b p(x_i) & \leq & +\infty \\ 0 & \leq & p(x_i) \cdot (-\log_b p(x_i)) & \leq & +\infty \; . \end{array} \tag{4}$$

By convention, $0 \cdot \log_b(0)$ is taken to be 0 when calculating entropy, consistent with

$$\lim_{p \to 0} \left[p \log_b(p) \right] = 0 \; . \tag{5}$$

Taking this together, each addend in (3) is positive or zero and thus, the entire sum must also be non-negative.

Sources:
- Cover TM, Thomas JA (1991): "Elements of Information Theory", p. 15; URL: https://www.wiley.com/en-us/Elements+of+Information+Theory%2C+2nd+Edition-p-9780471241959.

Metadata: ID: P57 | shortcut: ent-nonneg | author: JoramSoch | date: 2020-02-19, 19:10.

4.1.3 Conditional entropy

Definition: Let X and Y be discrete random variables (→ Definition "rvar") with possible outcomes $\mathcal{X}$ and $\mathcal{Y}$ and probability mass functions (→ Definition I/1.1.1) $p(x)$ and $p(y)$. Then, the conditional entropy of Y given X or, entropy of Y conditioned on X, is defined as

$$\mathrm{H}(X) = \sum_{x \in \mathcal{X}} p(x) \cdot \mathrm{H}(Y|X = x) \tag{1}$$

where $\mathrm{H}(Y|X = x)$ is the (marginal) entropy (→ Definition I/4.1.1) of Y, evaluated at x.

Sources:
- Cover TM, Thomas JA (1991): "Joint Entropy and Conditional Entropy"; in: *Elements of Information Theory*, ch. 2.2, p. 15; URL: https://www.wiley.com/en-us/Elements+of+Information+Theory%2C+2nd+Edition-p-9780471241959.

Metadata: ID: D17 | shortcut: ent-cond | author: JoramSoch | date: 2020-02-19, 18:08.

4.1.4 Joint entropy

Definition: Let X and Y be discrete random variables (→ Definition "rvar") with possible outcomes $\mathcal{X}$ and $\mathcal{Y}$ and joint probability (→ Definition "prob-joint") mass function (→ Definition I/1.1.1) $p(x, y)$. Then, the joint entropy of X and Y is defined as

$$\mathrm{H}(X, Y) = -\sum_{x \in \mathcal{X}} \sum_{x \in \mathcal{Y}} p(x, y) \cdot \log_b p(x, y) \tag{1}$$

where b is the base of the logarithm specifying in which unit the entropy is determined.

Sources:
- Cover TM, Thomas JA (1991): "Joint Entropy and Conditional Entropy"; in: *Elements of Information Theory*, ch. 2.2, p. 16; URL: https://www.wiley.com/en-us/Elements+of+Information+Theory%2C+2nd+Edition-p-9780471241959.

Metadata: ID: D18 | shortcut: ent-joint | author: JoramSoch | date: 2020-02-19, 18:18.

4.2 Differential entropy

4.2.1 Definition

Definition: Let X be a continuous random variable (→ Definition "rvar") with possible outcomes $\mathcal{X}$ and the (estimated or assumed) probability density function (→ Definition I/1.1.2) $p(x) = f_X(x)$. Then, the differential entropy (also referred to as "continuous entropy") of X is defined as

$$\mathrm{h}(X) = -\int_{\mathcal{X}} p(x) \log_b p(x) \, \mathrm{d}x \tag{1}$$

where b is the base of the logarithm specifying in which unit the entropy is determined.

Sources:

- Cover TM, Thomas JA (1991): "Differential Entropy"; in: *Elements of Information Theory*, ch. 8.1, p. 243; URL: https://www.wiley.com/en-us/Elements+of+Information+Theory%2C+2nd+Edition-p-9780471241959.

Metadata: ID: D16 | shortcut: dent | author: JoramSoch | date: 2020-02-19, 17:53.

4.2.2 Negativity

Theorem: Unlike its discrete analogue (→ Proof I/4.1.2), the differential entropy (→ Definition I/4.2.1) can become negative.

Proof: Let X be a random variable (→ Definition "rvar") following a continuous uniform distribution (→ Definition II/3.1.1) with minimum 0 and maximum 1/2:

$$X \sim \mathcal{U}(0, 1/2) \; . \tag{1}$$

Then, its probability density function (→ Proof II/3.1.2) is:

$$f_X(x) = 2 \quad \text{for} \quad 0 \leq x \leq \frac{1}{2} \; . \tag{2}$$

Thus, the differential entropy (→ Definition I/4.2.1) follows as

$$\begin{split}
\mathrm{h}(X) &= -\int_{\mathcal{X}} f_X(x) \log_b f_X(x) \, \mathrm{d}x \\
&= -\int_0^{\frac{1}{2}} 2 \, \log_b(2) \, \mathrm{d}x \\
&= -\log_b(2) \int_0^{\frac{1}{2}} 2 \, \mathrm{d}x \\
&= -\log_b(2) \left[2x \right]_0^{\frac{1}{2}} \\
&= -\log_b(2)
\end{split} \tag{3}$$

which is negative for any base $b > 1$.

Sources:

- Wikipedia (2020): "Differential entropy"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-02; URL: https://en.wikipedia.org/wiki/Differential_entropy#Definition.

Metadata: ID: P68 | shortcut: dent-neg | author: JoramSoch | date: 2020-03-02, 20:32.

4.3 Discrete mutual information

4.3.1 Definition

Definition:
1) The mutual information of two discrete random variables (→ Definition "rvar") X and Y is defined as

$$\mathrm{I}(X,Y) = - \sum_{x \in \mathcal{X}} \sum_{x \in \mathcal{Y}} p(x,y) \cdot \log \frac{p(x,y)}{p(x) \cdot p(y)} \tag{1}$$

where $p(x)$ and $p(y)$ are the probability mass functions (→ Definition I/1.1.1) of X and Y and $p(x,y)$ is the joint probability (→ Definition "prob-joint") mass function of X and Y.
2) The mutual information of two continuous random variables (→ Definition "rvar") X and Y is defined as

$$\mathrm{I}(X,Y) = - \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x,y) \cdot \log \frac{p(x,y)}{p(x) \cdot p(y)} \, \mathrm{d}y \, \mathrm{d}x \tag{2}$$

where $p(x)$ and $p(y)$ are the probability density functions (→ Definition I/1.1.1) of X and Y and $p(x,y)$ is the joint probability (→ Definition "prob-joint") density function of X and Y.

Sources:
- Cover TM, Thomas JA (1991): "Relative Entropy and Mutual Information"; in: *Elements of Information Theory*, ch. 2.3/8.5, p. 20/251; URL: https://www.wiley.com/en-us/Elements+of+Information+Theory%2C+2nd+Edition-p-9780471241959.

Metadata: ID: D19 | shortcut: mi | author: JoramSoch | date: 2020-02-19, 18:35.

4.3.2 Relation to marginal and conditional entropy

Theorem: Let X and Y be discrete random variables (→ Definition "rvar") with the joint probability (→ Definition "prob-joint") $p(x,y)$ for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Then, the mutual information (→ Definition I/4.4.1) of X and Y can be expressed as

$$\begin{aligned} \mathrm{I}(X,Y) &= \mathrm{H}(X) - \mathrm{H}(X|Y) \\ &= \mathrm{H}(Y) - \mathrm{H}(Y|X) \end{aligned} \tag{1}$$

where $\mathrm{H}(X)$ and $\mathrm{H}(Y)$ are the marginal entropies (→ Definition I/4.1.1) of X and Y and $\mathrm{H}(X|Y)$ and $\mathrm{H}(Y|X)$ are the conditional entropies (→ Definition I/4.1.3).

Proof: The mutual information (→ Definition I/4.4.1) of X and Y is defined as

$$\mathrm{I}(X,Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)\,p(y)} \; . \tag{2}$$

Separating the logarithm, we have:

$$\mathrm{I}(X,Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(y)} - \sum_x \sum_y p(x,y) \log p(x) \; . \tag{3}$$

Applying the law of conditional probability (→ Definition "prob-cond"), i.e. $p(x,y) = p(x|y)\,p(y)$, we get:

$$\mathrm{I}(X,Y) = \sum_x \sum_y p(x|y)\,p(y) \log p(x|y) - \sum_x \sum_y p(x,y) \log p(x) \; . \tag{4}$$

Regrouping the variables, we have:

$$\mathrm{I}(X,Y) = \sum_y p(y) \sum_x p(x|y) \log p(x|y) - \sum_x \left(\sum_y p(x,y) \right) \log p(x) \; . \tag{5}$$

Applying the law of marginal probability (→ Definition "prob-marg"), i.e. $p(x) = \sum_y p(x,y)$, we get:

$$\mathrm{I}(X,Y) = \sum_y p(y) \sum_x p(x|y) \log p(x|y) - \sum_x p(x) \log p(x) \; . \tag{6}$$

Now considering the definitions of marginal (→ Definition I/4.1.1) and conditional (→ Definition I/4.1.3) entropy

$$\begin{aligned} \mathrm{H}(X) &= -\sum_{x \in \mathcal{X}} p(x) \log p(x) \\ \mathrm{H}(X|Y) &= \sum_{y \in \mathcal{Y}} p(y)\, \mathrm{H}(X|Y=y) \; , \end{aligned} \tag{7}$$

we can finally show:

$$\begin{aligned} \mathrm{I}(X,Y) &= -\mathrm{H}(X|Y) + \mathrm{H}(X) \\ &= \mathrm{H}(X) - \mathrm{H}(X|Y) \; . \end{aligned} \tag{8}$$

The conditioning of X on Y in this proof is without loss of generality. Thus, the proof for the expression using the reverse conditional entropy of Y given X is obtained by simply switching x and y in the derivation.

Sources:

- Wikipedia (2020): "Mutual information"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-01-13; URL: https://en.wikipedia.org/wiki/Mutual_information#Relation_to_conditional_and_joint_entropy.

Metadata: ID: P19 | shortcut: dmi-mce | author: JoramSoch | date: 2020-01-13, 18:20.

4.3.3 Relation to marginal and joint entropy

Theorem: Let X and Y be discrete random variables (→ Definition "rvar") with the joint probability (→ Definition "prob-joint") $p(x,y)$ for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Then, the mutual information (→ Definition I/4.4.1) of X and Y can be expressed as

$$\mathrm{I}(X,Y) = \mathrm{H}(X) + \mathrm{H}(Y) - \mathrm{H}(X,Y) \tag{1}$$

where $\mathrm{H}(X)$ and $\mathrm{H}(Y)$ are the marginal entropies (→ Definition I/4.1.1) of X and Y and $\mathrm{H}(X,Y)$ is the joint entropy (→ Definition I/4.1.4).

Proof: The mutual information (→ Definition I/4.4.1) of X and Y is defined as

$$\mathrm{I}(X,Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)\,p(y)} \; . \tag{2}$$

Separating the logarithm, we have:

$$\mathrm{I}(X,Y) = \sum_x \sum_y p(x,y) \log p(x,y) - \sum_x \sum_y p(x,y) \log p(x) - \sum_x \sum_y p(x,y) \log p(y) \; . \tag{3}$$

Regrouping the variables, this reads:

$$\mathrm{I}(X,Y) = \sum_x \sum_y p(x,y) \log p(x,y) - \sum_x \left(\sum_y p(x,y) \right) \log p(x) - \sum_y \left(\sum_x p(x,y) \right) \log p(y) \; . \tag{4}$$

Applying the law of marginal probability (→ Definition "prob-marg"), i.e. $p(x) = \sum_y p(x,y)$, we get:

$$\mathrm{I}(X,Y) = \sum_x \sum_y p(x,y) \log p(x,y) - \sum_x p(x) \log p(x) - \sum_y p(y) \log p(y) \; . \tag{5}$$

Now considering the definitions of marginal (→ Definition I/4.1.1) and joint (→ Definition I/4.1.4) entropy

$$\begin{split} \mathrm{H}(X) &= - \sum_{x \in \mathcal{X}} p(x) \log p(x) \\ \mathrm{H}(X,Y) &= - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log p(x,y) \; , \end{split} \tag{6}$$

we can finally show:

$$\begin{split} \mathrm{I}(X,Y) &= -\mathrm{H}(X,Y) + \mathrm{H}(X) + \mathrm{H}(Y) \\ &= \mathrm{H}(X) + \mathrm{H}(Y) - \mathrm{H}(X,Y) \; . \end{split} \tag{7}$$

Sources:

- Wikipedia (2020): "Mutual information"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-01-13; URL: https://en.wikipedia.org/wiki/Mutual_information#Relation_to_conditional_and_joint_entropy.

Metadata: ID: P20 | shortcut: dmi-mje | author: JoramSoch | date: 2020-01-13, 21:53.

4.3.4 Relation to joint and conditional entropy

Theorem: Let X and Y be discrete random variables (→ Definition "rvar") with the joint probability (→ Definition "prob-joint") $p(x,y)$ for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Then, the mutual information (→ Definition I/4.4.1) of X and Y can be expressed as

$$\mathrm{I}(X,Y) = \mathrm{H}(X,Y) - \mathrm{H}(X|Y) - \mathrm{H}(Y|X) \tag{1}$$

where $\mathrm{H}(X,Y)$ is the joint entropy (→ Definition I/4.1.4) of X and Y and $\mathrm{H}(X|Y)$ and $\mathrm{H}(Y|X)$ are the conditional entropies (→ Definition I/4.1.3).

Proof: The existence of the joint probability mass function (→ Definition I/1.1.1) ensures that the mutual information (→ Definition I/4.4.1) is defined:

$$\mathrm{I}(X,Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)\,p(y)} \; . \tag{2}$$

The relation of mutual information to conditional entropy (→ Proof I/4.3.2) is:

$$\mathrm{I}(X,Y) = \mathrm{H}(X) - \mathrm{H}(X|Y) \tag{3}$$

$$\mathrm{I}(X,Y) = \mathrm{H}(Y) - \mathrm{H}(Y|X) \tag{4}$$

The relation of mutual information to joint entropy (→ Proof I/4.3.3) is:

$$\mathrm{I}(X,Y) = \mathrm{H}(X) + \mathrm{H}(Y) - \mathrm{H}(X,Y) \; . \tag{5}$$

It is true that

$$\mathrm{I}(X,Y) = \mathrm{I}(X,Y) + \mathrm{I}(X,Y) - \mathrm{I}(X,Y) \; . \tag{6}$$

Plugging in (3), (4) and (5) on the right-hand side, we have

$$\begin{split} \mathrm{I}(X,Y) &= \mathrm{H}(X) - \mathrm{H}(X|Y) + \mathrm{H}(Y) - \mathrm{H}(Y|X) - \mathrm{H}(X) - \mathrm{H}(Y) + \mathrm{H}(X,Y) \\ &= \mathrm{H}(X,Y) - \mathrm{H}(X|Y) - \mathrm{H}(Y|X) \end{split} \tag{7}$$

which proves the identity given above.

Sources:
- Wikipedia (2020): "Mutual information"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-01-13; URL: https://en.wikipedia.org/wiki/Mutual_information#Relation_to_conditional_and_joint_entropy.

Metadata: ID: P21 | shortcut: dmi-jce | author: JoramSoch | date: 2020-01-13, 22:17.

4.4 Continuous mutual information

4.4.1 Definition

Definition:
1) The mutual information of two discrete random variables (→ Definition "rvar") X and Y is defined as

$$\mathrm{I}(X,Y) = -\sum_{x \in \mathcal{X}} \sum_{x \in \mathcal{Y}} p(x,y) \cdot \log \frac{p(x,y)}{p(x) \cdot p(y)} \tag{1}$$

where $p(x)$ and $p(y)$ are the probability mass functions (→ Definition I/1.1.1) of X and Y and $p(x,y)$ is the joint probability (→ Definition "prob-joint") mass function of X and Y.

2) The mutual information of two continuous random variables (→ Definition "rvar") X and Y is defined as

$$\mathrm{I}(X,Y) = -\int_{\mathcal{X}} \int_{\mathcal{Y}} p(x,y) \cdot \log \frac{p(x,y)}{p(x) \cdot p(y)} \, \mathrm{d}y \, \mathrm{d}x \tag{2}$$

where $p(x)$ and $p(y)$ are the probability density functions (→ Definition I/1.1.1) of X and Y and $p(x,y)$ is the joint probability (→ Definition "prob-joint") density function of X and Y.

Sources:

- Cover TM, Thomas JA (1991): "Relative Entropy and Mutual Information"; in: *Elements of Information Theory*, ch. 2.3/8.5, p. 20/251; URL: https://www.wiley.com/en-us/Elements+of+Information+Theory%2C+2nd+Edition-p-9780471241959.

Metadata: ID: D19 | shortcut: mi | author: JoramSoch | date: 2020-02-19, 18:35.

4.4.2 Relation to marginal and conditional differential entropy

Theorem: Let X and Y be continuous random variables (→ Definition "rvar") with the joint probability (→ Definition "prob-joint") $p(x,y)$ for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Then, the mutual information (→ Definition I/4.4.1) of X and Y can be expressed as

$$\begin{aligned} \mathrm{I}(X,Y) &= \mathrm{h}(X) - \mathrm{h}(X|Y) \\ &= \mathrm{h}(Y) - \mathrm{h}(Y|X) \end{aligned} \tag{1}$$

where $\mathrm{h}(X)$ and $\mathrm{h}(Y)$ are the marginal differential entropies (→ Definition I/4.2.1) of X and Y and $\mathrm{h}(X|Y)$ and $\mathrm{h}(Y|X)$ are the conditional differential entropies (→ Definition "dent-cond").

Proof: The mutual information (→ Definition I/4.4.1) of X and Y is defined as

$$\mathrm{I}(X,Y) = \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)\, p(y)} \, \mathrm{d}y \, \mathrm{d}x \; . \tag{2}$$

Separating the logarithm, we have:

$$\mathrm{I}(X,Y) = \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(y)} \, \mathrm{d}y \, \mathrm{d}x - \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x,y) \log p(x) \, \mathrm{d}x \, \mathrm{d}y \; . \tag{3}$$

Applying the law of conditional probability (→ Definition "prob-cond"), i.e. $p(x,y) = p(x|y)\, p(y)$, we get:

$$\mathrm{I}(X,Y) = \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x|y)\, p(y) \log p(x|y) \, \mathrm{d}y \, \mathrm{d}x - \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x,y) \log p(x) \, \mathrm{d}y \, \mathrm{d}x \; . \tag{4}$$

Regrouping the variables, we have:

$$\mathrm{I}(X,Y) = \int_{\mathcal{Y}} p(y) \int_{\mathcal{X}} p(x|y) \log p(x|y) \, \mathrm{d}x \, \mathrm{d}y - \int_{\mathcal{X}} \left(\int_{\mathcal{Y}} p(x,y) \, \mathrm{d}y \right) \log p(x) \, \mathrm{d}x \; . \tag{5}$$

Applying the law of marginal probability (→ Definition "prob-marg"), i.e. $p(x) = \int_{\mathcal{Y}} p(x,y) \, \mathrm{d}y$, we get:

$$\mathrm{I}(X,Y) = \int_{\mathcal{Y}} p(y) \int_{\mathcal{X}} p(x|y) \log p(x|y) \, \mathrm{d}x \, \mathrm{d}y - \int_{\mathcal{X}} p(x) \log p(x) \, \mathrm{d}x \; . \tag{6}$$

Now considering the definitions of marginal (→ Definition I/4.2.1) and conditional (→ Definition "dent-cond") differential entropy

$$\begin{aligned} \mathrm{h}(X) &= -\int_{\mathcal{X}} p(x) \log p(x) \, \mathrm{d}x \\ \mathrm{h}(X|Y) &= \sum_{\mathcal{Y}} p(y) \, \mathrm{h}(X|Y=y) \, \mathrm{d}y \; , \end{aligned} \tag{7}$$

we can finally show:

$$\mathrm{I}(X,Y) = -\mathrm{h}(X|Y) + \mathrm{h}(X) = \mathrm{h}(X) - \mathrm{h}(X|Y) \; . \tag{8}$$

The conditioning of X on Y in this proof is without loss of generality. Thus, the proof for the expression using the reverse conditional differential entropy of Y given X is obtained by simply switching x and y in the derivation.

Sources:

- Wikipedia (2020): "Mutual information"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-21; URL: https://en.wikipedia.org/wiki/Mutual_information#Relation_to_conditional_and_joint_entropy.

Metadata: ID: P58 | shortcut: cmi-mcde | author: JoramSoch | date: 2020-02-21, 16:53.

4.4.3 Relation to marginal and joint differential entropy

Theorem: Let X and Y be continuous random variables (→ Definition "rvar") with the joint probability (→ Definition "prob-joint") $p(x,y)$ for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Then, the mutual information (→ Definition I/4.4.1) of X and Y can be expressed as

$$\mathrm{I}(X,Y) = \mathrm{h}(X) + \mathrm{h}(Y) - \mathrm{h}(X,Y) \tag{1}$$

where $\mathrm{h}(X)$ and $\mathrm{h}(Y)$ are the marginal differential entropies (→ Definition I/4.2.1) of X and Y and $\mathrm{h}(X,Y)$ is the joint differential entropy (→ Definition "dent-joint").

Proof: The mutual information (→ Definition I/4.4.1) of X and Y is defined as

$$\mathrm{I}(X,Y) = \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)\,p(y)} \, \mathrm{d}y \, \mathrm{d}x \; . \tag{2}$$

Separating the logarithm, we have:

$$\mathrm{I}(X,Y) = \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x,y) \log p(x,y) \, \mathrm{d}y \, \mathrm{d}x - \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x,y) \log p(x) \, \mathrm{d}y \, \mathrm{d}x - \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x,y) \log p(y) \, \mathrm{d}y \, \mathrm{d}x \; . \tag{3}$$

Regrouping the variables, this reads:

$$\mathrm{I}(X,Y) = \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x,y) \log p(x,y) \, \mathrm{d}y \, \mathrm{d}x - \int_{\mathcal{X}} \left(\int_{\mathcal{Y}} p(x,y) \, \mathrm{d}y \right) \log p(x) \, \mathrm{d}x - \int_{\mathcal{Y}} \left(\int_{\mathcal{X}} p(x,y) \, \mathrm{d}x \right) \log p(y) \, \mathrm{d}y \; . \tag{4}$$

Applying the law of marginal probability (→ Definition "prob-marg"), i.e. $p(x) = \int_{\mathcal{Y}} p(x,y)$, we get:

$$\mathrm{I}(X,Y) = \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x,y) \log p(x,y) \, \mathrm{d}y \, \mathrm{d}x - \int_{\mathcal{X}} p(x) \log p(x) \, \mathrm{d}x - \int_{\mathcal{Y}} p(y) \log p(y) \, \mathrm{d}y \; . \tag{5}$$

Now considering the definitions of marginal (→ Definition I/4.2.1) and joint (→ Definition "dent-joint") differential entropy

$$\begin{aligned} \mathrm{h}(X) &= -\int_{\mathcal{X}} p(x) \log p(x) \, \mathrm{d}x \\ \mathrm{h}(X,Y) &= -\int_{\mathcal{X}} \int_{\mathcal{Y}} p(x,y) \log p(x,y) \, \mathrm{d}y \, \mathrm{d}x \; , \end{aligned} \tag{6}$$

we can finally show:

$$\begin{aligned} \mathrm{I}(X,Y) &= -\mathrm{h}(X,Y) + \mathrm{h}(X) + \mathrm{h}(Y) \\ &= \mathrm{h}(X) + \mathrm{h}(Y) - \mathrm{h}(X,Y) \; . \end{aligned} \tag{7}$$

Sources:

- Wikipedia (2020): "Mutual information"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-21; URL: https://en.wikipedia.org/wiki/Mutual_information#Relation_to_conditional_and_joint_entropy.

Metadata: ID: P59 | shortcut: cmi-mjde | author: JoramSoch | date: 2020-02-21, 17:13.

4.4.4 Relation to joint and conditional differential entropy

Theorem: Let X and Y be continuous random variables (→ Definition "rvar") with the joint probability (→ Definition "prob-joint") $p(x,y)$ for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Then, the mutual information (→ Definition I/4.4.1) of X and Y can be expressed as

$$\mathrm{I}(X,Y) = \mathrm{h}(X,Y) - \mathrm{h}(X|Y) - \mathrm{h}(Y|X) \tag{1}$$

where $\mathrm{h}(X,Y)$ is the joint differential entropy (→ Definition "dent-joint") of X and Y and $\mathrm{h}(X|Y)$ and $\mathrm{h}(Y|X)$ are the conditional differential entropies (→ Definition "dent-cond").

Proof: The existence of the joint probability density function (→ Definition I/1.1.2) ensures that the mutual information (→ Definition I/4.4.1) is defined:

$$\mathrm{I}(X,Y) = \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)\,p(y)} \, \mathrm{d}y \, \mathrm{d}x \; . \tag{2}$$

The relation of mutual information to conditional differential entropy (→ Proof I/4.4.2) is:

$$\mathrm{I}(X,Y) = \mathrm{h}(X) - \mathrm{h}(X|Y) \tag{3}$$

$$\mathrm{I}(X,Y) = \mathrm{h}(Y) - \mathrm{h}(Y|X) \tag{4}$$

The relation of mutual information to joint differential entropy (→ Proof I/4.4.3) is:

$$\mathrm{I}(X,Y) = \mathrm{h}(X) + \mathrm{h}(Y) - \mathrm{h}(X,Y) \; . \tag{5}$$

It is true that

$$\mathrm{I}(X,Y) = \mathrm{I}(X,Y) + \mathrm{I}(X,Y) - \mathrm{I}(X,Y) \; . \tag{6}$$

Plugging in (3), (4) and (5) on the right-hand side, we have

$$\begin{split} \mathrm{I}(X,Y) &= \mathrm{h}(X) - \mathrm{h}(X|Y) + \mathrm{h}(Y) - \mathrm{h}(Y|X) - \mathrm{h}(X) - \mathrm{h}(Y) + \mathrm{h}(X,Y) \\ &= \mathrm{h}(X,Y) - \mathrm{h}(X|Y) - \mathrm{h}(Y|X) \end{split} \tag{7}$$

which proves the identity given above.

Sources:

- Wikipedia (2020): "Mutual information"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-21; URL: https://en.wikipedia.org/wiki/Mutual_information#Relation_to_conditional_and_joint_entropy.

Metadata: ID: P60 | shortcut: cmi-jcde | author: JoramSoch | date: 2020-01-21, 17:23.

Chapter II

Probability Distributions

1 Univariate discrete distributions

1.1 Bernoulli distribution

1.1.1 Mean

Theorem: Let X be a random variable (→ Definition "rvar") following a Bernoulli distribution (→ Definition "bern"):

$$X \sim \mathrm{Bern}(p) \; . \tag{1}$$

Then, the mean or expected value (→ Definition I/1.2.1) of X is

$$\mathrm{E}(X) = p \; . \tag{2}$$

Proof: The expected value (→ Definition I/1.2.1) is the probability-weighted average of all possible values:

$$\mathrm{E}(X) = \sum_{x \in \mathcal{X}} x \cdot \mathrm{Pr}(X = x) \; . \tag{3}$$

Since there are only two possible outcomes for a Bernoulli random variable (→ Proof "bern-pmf"), we have:

$$\begin{split} \mathrm{E}(X) &= 0 \cdot \mathrm{Pr}(X = 0) + 1 \cdot \mathrm{Pr}(X = 1) \\ &= 0 \cdot (1-p) + 1 \cdot p \\ &= p \; . \end{split} \tag{4}$$

Sources:

- Wikipedia (2020): "Bernoulli distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-01-16; URL: https://en.wikipedia.org/wiki/Bernoulli_distribution#Mean.

Metadata: ID: P22 | shortcut: bern-mean | author: JoramSoch | date: 2020-01-16, 10:58.

1.2 Binomial distribution

1.2.1 Mean

Theorem: Let X be a random variable (→ Definition "rvar") following a binomial distribution (→ Definition "bin"):

$$X \sim \mathrm{Bin}(n,p) \; . \tag{1}$$

Then, the mean or expected value (→ Definition I/1.2.1) of X is

$$\mathrm{E}(X) = np \; . \tag{2}$$

Proof: By definition, a binomial random variable (→ Definition "bin") is the sum of n independent and identical Bernoulli trials (→ Definition "bern") with success probability p. Therefore, the expected value is

$$\mathrm{E}(X) = \mathrm{E}(X_1 + \ldots + X_n) \tag{3}$$

and because the expected value is a linear operator (→ Proof I/1.2.3), this is equal to

$$\begin{split} \mathrm{E}(X) &= \mathrm{E}(X_1) + \ldots + \mathrm{E}(X_n) \\ &= \sum_{i=1}^{n} \mathrm{E}(X_i) \; . \end{split} \tag{4}$$

With the expected value of the Bernoulli distribution (→ Proof II/1.1.1), we have:

$$\mathrm{E}(X) = \sum_{i=1}^{n} p = np \; . \tag{5}$$

Sources:

- Wikipedia (2020): "Binomial distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-01-16; URL: https://en.wikipedia.org/wiki/Binomial_distribution#Expected_value_and_variance.

Metadata: ID: P23 | shortcut: bin-mean | author: JoramSoch | date: 2020-01-16, 11:06.

2 Multivariate discrete distributions

2.1 Categorical distribution

2.1.1 Mean

Theorem: Let X be a random vector (→ Definition "rvec") following a categorical distribution (→ Definition "cat"):

$$X \sim \mathrm{Cat}([p_1, \ldots, p_k]) \; . \tag{1}$$

Then, the mean or expected value (→ Definition I/1.2.1) of X is

$$\mathrm{E}(X) = [p_1, \ldots, p_k] \; . \tag{2}$$

Proof: If we conceive the outcome of a categorical distribution (→ Definition "cat-pmf") to be a $1 \times k$ vector, then the elementary row vectors $e_1 = [1, 0, \ldots, 0]$, ..., $e_k = [0, \ldots, 0, 1]$ are all the possible outcomes and they occur with probabilities $\mathrm{Pr}(X = e_1) = p_1$, ..., $\mathrm{Pr}(X = e_k) = p_k$. Consequently, the expected value (→ Definition I/1.2.1) is

$$\begin{aligned} \mathrm{E}(X) &= \sum_{x \in \mathcal{X}} x \cdot \mathrm{Pr}(X = x) \\ &= \sum_{i=1}^{k} e_i \cdot \mathrm{Pr}(X = e_i) \\ &= \sum_{i=1}^{k} e_i \cdot p_i \\ &= [p_1, \ldots, p_k] \; . \end{aligned} \tag{3}$$

Sources:
- original work

Metadata: ID: P24 | shortcut: cat-mean | author: JoramSoch | date: 2020-01-16, 11:17.

2.2 Multinomial distribution

2.2.1 Mean

Theorem: Let X be a random vector (→ Definition "rvec") following a multinomial distribution (→ Definition "mult"):

$$X \sim \mathrm{Mult}(n, [p_1, \ldots, p_k]) \; . \tag{1}$$

Then, the mean or expected value (→ Definition I/1.2.1) of X is

$$\mathrm{E}(X) = [np_1, \ldots, np_k] \; . \tag{2}$$

Proof: By definition, a multinomial random variable (→ Definition "mult") is the sum of n independent and identical categorical trials (→ Definition "cat") with category probabilities $p_1, \ldots, p_k$. Therefore, the expected value is

$$\mathrm{E}(X) = \mathrm{E}(X_1 + \ldots + X_n) \tag{3}$$

and because the expected value is a linear operator (→ Proof I/1.2.3), this is equal to

$$\begin{split} \mathrm{E}(X) &= \mathrm{E}(X_1) + \ldots + \mathrm{E}(X_n) \\ &= \sum_{i=1}^{n} \mathrm{E}(X_i) \; . \end{split} \tag{4}$$

With the expected value of the categorical distribution (→ Proof II/2.1.1), we have:

$$\mathrm{E}(X) = \sum_{i=1}^{n} [p_1, \ldots, p_k] = n \cdot [p_1, \ldots, p_k] = [np_1, \ldots, np_k] \; . \tag{5}$$

Sources:
- original work

Metadata: ID: P25 | shortcut: mult-mean | author: JoramSoch | date: 2020-01-16, 11:26.

3 Univariate continuous distributions

3.1 Continuous uniform distribution

3.1.1 Definition

Definition: Let X be a continuous random variable (→ Definition "rvar"). Then, X is said to be uniformly distributed with minimum a and maximum b

$$X \sim \mathcal{U}(a, b) \; , \tag{1}$$

if and only if each value between and including a and b occurs with the same probability.

Sources:

- Wikipedia (2020): "Uniform distribution (continuous)"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-01-27; URL: https://en.wikipedia.org/wiki/Uniform_distribution_(continuous).

Metadata: ID: D3 | shortcut: cuni | author: JoramSoch | date: 2020-01-27, 14:05.

3.1.2 Probability density function

Theorem: Let X be a random variable (→ Definition "rvar") following a continuous uniform distribution (→ Definition II/3.1.1):

$$X \sim \mathcal{U}(a, b) \; . \tag{1}$$

Then, the probability density function (→ Definition I/1.1.2) of X is

$$f_X(x) = \left\{ \begin{array}{rl} \frac{1}{b-a} \; , & \text{if} \; a \leq x \leq b \\ 0 \; , & \text{otherwise} \; . \end{array} \right. \tag{2}$$

Proof: A continuous uniform variable is defined as (→ Definition II/3.1.1) having a constant probability density between minimum a and maximum b. Therefore,

$$\begin{split} & f_X(x) \propto 1 \quad \text{for all} \quad x \in [a,b] \quad \text{and} \\ & f_X(x) = 0, \quad \text{if} \quad x < a \quad \text{or} \quad x > b \; . \end{split} \tag{3}$$

To ensure that $f_X(x)$ is a proper probability density function (→ Definition I/1.1.2), the integral over all non-zero probabilities has to sum to 1. Therefore,

$$f_X(x) = \frac{1}{c(a,b)} \quad \text{for all} \quad x \in [a,b] \tag{4}$$

where the normalization factor $c(a,b)$ is specified, such that

$$\frac{1}{c(a,b)} \int_a^b 1 \, \mathrm{d}x = 1 \; . \tag{5}$$

Solving this for $c(a,b)$, we obtain:

$$
\begin{aligned}
\int_a^b 1 \, \mathrm{d}x &= c(a,b) \\
[x]_a^b &= c(a,b) \\
c(a,b) &= b-a \; .
\end{aligned} \tag{6}
$$

Sources:
- original work

Metadata: ID: P37 | shortcut: cuni-pdf | author: JoramSoch | date: 2020-01-31, 15:41.

3.1.3 Cumulative distribution function

Theorem: Let X be a random variable (→ Definition "rvar") following a continuous uniform distribution (→ Definition II/3.1.1):

$$
X \sim \mathcal{U}(a,b) \; . \tag{1}
$$

Then, the cumulative distribution function (→ Definition I/1.1.3) of X is

$$
F_X(x) = \left\{ \begin{array}{rl} 0 \; , & \text{if} \; x < a \\ \frac{x-a}{b-a} \; , & \text{if} \; a \leq x \leq b \\ 1 \; , & \text{if} \; x > b \; . \end{array} \right. \tag{2}
$$

Proof: The probability density function of the continuous uniform distribution (→ Proof II/3.1.2) is:

$$
\mathcal{U}(z;a,b) = \left\{ \begin{array}{rl} \frac{1}{b-a} \; , & \text{if} \; a \leq x \leq b \\ 0 \; , & \text{otherwise} \; . \end{array} \right. \tag{3}
$$

Thus, the cumulative distribution function (→ Definition I/1.1.3) is:

$$
F_X(x) = \int_{-\infty}^{x} \mathcal{U}(z;a,b) \, \mathrm{d}z \tag{4}
$$

First of all, if $x < a$, we have

$$
F_X(x) = \int_{-\infty}^{x} 0 \, \mathrm{d}z = 0 \; . \tag{5}
$$

Moreover, if $a \leq x \leq b$, we have using (3)

$$
\begin{aligned}
F_X(x) &= \int_{-\infty}^{a} \mathcal{U}(z;a,b) \, \mathrm{d}z + \int_{a}^{x} \mathcal{U}(z;a,b) \, \mathrm{d}z \\
&= \int_{-\infty}^{a} 0 \, \mathrm{d}z + \int_{a}^{x} \frac{1}{b-a} \, \mathrm{d}z \\
&= 0 + \frac{1}{b-a} [z]_a^x \\
&= \frac{x-a}{b-a} \; .
\end{aligned} \tag{6}
$$

Finally, if $x > b$, we have

$$
\begin{split}
F_X(x) &= \int_{-\infty}^{b} \mathcal{U}(z; a, b) \, \mathrm{d}z + \int_{b}^{x} \mathcal{U}(z; a, b) \, \mathrm{d}z \\
&= F_X(b) + \int_{b}^{x} 0 \, \mathrm{d}z \\
&= \frac{b-a}{b-a} + 0 \\
&= 1 \; .
\end{split}
\tag{7}
$$

This completes the proof.

Sources:
- original work

Metadata: ID: P38 | shortcut: cuni-cdf | author: JoramSoch | date: 2020-01-02, 18:05.

3.1.4 Quantile function

Theorem: Let X be a random variable (→ Definition "rvar") following a continuous uniform distribution (→ Definition II/3.1.1):

$$
X \sim \mathcal{U}(a, b) \; . \tag{1}
$$

Then, the quantile function (→ Definition I/1.1.4) of X is

$$
Q_X(p) = bp + a(1-p) \; . \tag{2}
$$

Proof: The cumulative distribution function of the continuous uniform distribution (→ Proof II/3.1.3) is:

$$
F_X(x) = \left\{
\begin{array}{rl}
0 \; , & \text{if} \; x < a \\
\frac{x-a}{b-a} \; , & \text{if} \; a \leq x \leq b \\
1 \; , & \text{if} \; x > b \; .
\end{array}
\right.
\tag{3}
$$

Thus, the quantile function (→ Definition I/1.1.4) is:

$$
Q_X(p) = F_X^{-1}(x) \; . \tag{4}
$$

This can be derived by rearranging equation (3):

$$
\begin{split}
p &= \frac{x-a}{b-a} \\
x &= p(b-a) + a \\
x &= bp + a(1-p) = Q_X(p) \; .
\end{split}
\tag{5}
$$

Sources:

- original work

Metadata: ID: P39 | shortcut: cuni-qf | author: JoramSoch | date: 2020-01-02, 18:27.

3.2 Normal distribution

3.2.1 Definition

Definition: Let X be a random variable (→ Definition "rvar"). Then, X is said to be normally distributed with mean μ and variance σ^2 (or, standard deviation σ)

$$X \sim \mathcal{N}(\mu, \sigma^2) \; , \tag{1}$$

if and only if its probability density function (→ Definition I/1.1.2) is given by

$$\mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right] \tag{2}$$

where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$.

Sources:
- Wikipedia (2020): "Normal distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-01-27; URL: https://en.wikipedia.org/wiki/Normal_distribution.

Metadata: ID: D4 | shortcut: norm | author: JoramSoch | date: 2020-01-27, 14:15.

3.2.2 Probability density function

Theorem: Let X be a random variable (→ Definition "rvar") following a normal distribution (→ Definition II/3.2.1):

$$X \sim \mathcal{N}(\mu, \sigma^2) \; . \tag{1}$$

Then, the probability density function (→ Definition I/1.1.2) of X is

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right] \; . \tag{2}$$

Proof: This follows directly from the definition of the normal distribution (→ Definition II/3.2.1).

Sources:
- original work

Metadata: ID: P33 | shortcut: norm-pdf | author: JoramSoch | date: 2020-01-27, 15:15.

3.2.3 Mean

Theorem: Let X be a random variable (→ Definition "rvar") following a normal distribution (→ Definition II/3.2.1):

$$X \sim \mathcal{N}(\mu, \sigma^2) \; . \tag{1}$$

Then, the mean or expected value (→ Definition I/1.2.1) of X is

$$\mathrm{E}(X) = \mu \; . \tag{2}$$

Proof: The expected value (→ Definition I/1.2.1) is the probability-weighted average over all possible values:

$$\mathrm{E}(X) = \int_{\mathbb{R}} x \cdot f_X(x) \, \mathrm{d}x \; . \tag{3}$$

With the probability density function of the normal distribution (→ Proof II/3.2.2), this reads:

$$\begin{split} \mathrm{E}(X) &= \int_{-\infty}^{+\infty} x \cdot \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} x \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \mathrm{d}x \; . \end{split} \tag{4}$$

Substituting $z = x - \mu$, we have:

$$\begin{split} \mathrm{E}(X) &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty-\mu}^{+\infty-\mu} (z+\mu) \cdot \exp\left[-\frac{1}{2}\left(\frac{z}{\sigma}\right)^2\right] \mathrm{d}(z+\mu) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} (z+\mu) \cdot \exp\left[-\frac{1}{2}\left(\frac{z}{\sigma}\right)^2\right] \mathrm{d}z \\ &= \frac{1}{\sqrt{2\pi}\sigma} \left(\int_{-\infty}^{+\infty} z \cdot \exp\left[-\frac{1}{2}\left(\frac{z}{\sigma}\right)^2\right] \mathrm{d}z + \mu \int_{-\infty}^{+\infty} \exp\left[-\frac{1}{2}\left(\frac{z}{\sigma}\right)^2\right] \mathrm{d}z \right) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \left(\int_{-\infty}^{+\infty} z \cdot \exp\left[-\frac{1}{2\sigma^2} \cdot z^2\right] \mathrm{d}z + \mu \int_{-\infty}^{+\infty} \exp\left[-\frac{1}{2\sigma^2} \cdot z^2\right] \mathrm{d}z \right) \; . \end{split} \tag{5}$$

The general antiderivatives are

$$\begin{split} \int x \cdot \exp\left[-ax^2\right] \mathrm{d}x &= -\frac{1}{2a} \cdot \exp\left[-ax^2\right] \\ \int \exp\left[-ax^2\right] \mathrm{d}x &= \frac{1}{2}\sqrt{\frac{\pi}{a}} \cdot \mathrm{erf}\left[\sqrt{a}x\right] \end{split} \tag{6}$$

where $\mathrm{erf}(x)$ is the error function. Using this, the integrals can be calculated as:

$$
\begin{split}
\mathrm{E}(X) &= \frac{1}{\sqrt{2\pi}\sigma} \left(\left[-\sigma^2 \cdot \exp\left[-\frac{1}{2\sigma^2} \cdot z^2 \right] \right]_{-\infty}^{+\infty} + \mu \left[\sqrt{\frac{\pi}{2}} \sigma \cdot \mathrm{erf}\left[\frac{1}{\sqrt{2}\sigma} z \right] \right]_{-\infty}^{+\infty} \right) \\
&= \frac{1}{\sqrt{2\pi}\sigma} \left(\left[\lim_{z \to \infty} \left(-\sigma^2 \cdot \exp\left[-\frac{1}{2\sigma^2} \cdot z^2 \right] \right) - \lim_{z \to -\infty} \left(-\sigma^2 \cdot \exp\left[-\frac{1}{2\sigma^2} \cdot z^2 \right] \right) \right] \right. \\
&\quad \left. + \mu \left[\lim_{z \to \infty} \left(\sqrt{\frac{\pi}{2}} \sigma \cdot \mathrm{erf}\left[\frac{1}{\sqrt{2}\sigma} z \right] \right) - \lim_{z \to -\infty} \left(\sqrt{\frac{\pi}{2}} \sigma \cdot \mathrm{erf}\left[\frac{1}{\sqrt{2}\sigma} z \right] \right) \right] \right) \\
&= \frac{1}{\sqrt{2\pi}\sigma} \left([0 - 0] + \mu \left[\sqrt{\frac{\pi}{2}} \sigma - \left(-\sqrt{\frac{\pi}{2}} \sigma \right) \right] \right) \\
&= \frac{1}{\sqrt{2\pi}\sigma} \cdot \mu \cdot 2 \sqrt{\frac{\pi}{2}} \sigma \\
&= \mu \; .
\end{split} \tag{7}
$$

Sources:

- Papadopoulos, Alecos (2013): "How to derive the mean and variance of Gaussian random variable?"; in: *StackExchange Mathematics*; URL: https://math.stackexchange.com/questions/518281/how-to-derive-the-mean-and-variance-of-a-gaussian-random-variable.

Metadata: ID: P15 | shortcut: norm-mean | author: JoramSoch | date: 2020-01-09, 15:04.

3.2.4 Median

Theorem: Let X be a random variable (→ Definition "rvar") following a normal distribution (→ Definition II/3.2.1):

$$
X \sim \mathcal{N}(\mu, \sigma^2) \; . \tag{1}
$$

Then, the median (→ Definition "med") of X is

$$
\mathrm{median}(X) = \mu \; . \tag{2}
$$

Proof: The median (→ Definition "med") is the value at which the cumulative distribution function (→ Definition I/1.1.3) is 1/2:

$$
F_X(\mathrm{median}(X)) = \frac{1}{2} \; . \tag{3}
$$

The cumulative distribution function of the normal distribution (→ Proof "norm-cdf") is

$$
F_X(x) = \frac{1}{2} \left[1 + \mathrm{erf}\left(\frac{x - \mu}{\sqrt{2}\sigma} \right) \right] \tag{4}
$$

where $\mathrm{erf}(x)$ is the error function. Thus, the inverse CDF is

$$
x = \sqrt{2}\sigma \cdot \mathrm{erf}^{-1}(2p - 1) + \mu \tag{5}
$$

where $\mathrm{erf}^{-1}(x)$ is the inverse error function. Setting $p = 1/2$, we obtain:

$$\mathrm{median}(X) = \sqrt{2}\sigma \cdot \mathrm{erf}^{-1}(0) + \mu = \mu \; . \tag{6}$$

Sources:
- original work

Metadata: ID: P16 | shortcut: norm-med | author: JoramSoch | date: 2020-01-09, 15:33.

3.2.5 Mode

Theorem: Let X be a random variable (→ Definition "rvar") following a normal distribution (→ Definition II/3.2.1):

$$X \sim \mathcal{N}(\mu, \sigma^2) \; . \tag{1}$$

Then, the mode (→ Definition "mode") of X is

$$\mathrm{mode}(X) = \mu \; . \tag{2}$$

Proof: The mode (→ Definition "mode") is the value which maximizes the probability density function (→ Definition I/1.1.2):

$$\mathrm{mode}(X) = \operatorname*{arg\,max}_x f_X(x) \; . \tag{3}$$

The probability density function of the normal distribution (→ Proof II/3.2.2) is:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right] \; . \tag{4}$$

The first two deriatives of this function are:

$$f'_X(x) = \frac{\mathrm{d}f_X(x)}{\mathrm{d}x} = \frac{1}{\sqrt{2\pi}\sigma^3} \cdot (-x+\mu) \cdot \exp\left[-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right] \tag{5}$$

$$f''_X(x) = \frac{\mathrm{d}^2 f_X(x)}{\mathrm{d}x^2} = -\frac{1}{\sqrt{2\pi}\sigma^3} \cdot \exp\left[-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right] + \frac{1}{\sqrt{2\pi}\sigma^5} \cdot (-x+\mu)^2 \cdot \exp\left[-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right] \; . \tag{6}$$

We now calculate the root of the first derivative (5):

$$\begin{aligned} f'_X(x) = 0 &= \frac{1}{\sqrt{2\pi}\sigma^3} \cdot (-x+\mu) \cdot \exp\left[-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right] \\ 0 &= -x + \mu \\ x &= \mu \; . \end{aligned} \tag{7}$$

By plugging this value into the second deriative (6),

$$\begin{aligned} f''_X(\mu) &= -\frac{1}{\sqrt{2\pi}\sigma^3} \cdot \exp(0) + \frac{1}{\sqrt{2\pi}\sigma^5} \cdot (0)^2 \cdot \exp(0) \\ &= -\frac{1}{\sqrt{2\pi}\sigma^3} < 0 \; , \end{aligned} \tag{8}$$

we confirm that it is in fact a maximum which shows that

$$\mathrm{mode}(X) = \mu \; . \tag{9}$$

Sources:
- original work

Metadata: ID: P17 | shortcut: norm-mode | author: JoramSoch | date: 2020-01-09, 15:58.

3.2.6 Variance

Theorem: Let X be a random variable (→ Definition "rvar") following a normal distribution (→ Definition II/3.2.1):

$$X \sim \mathcal{N}(\mu, \sigma^2) \; . \tag{1}$$

Then, the variance (→ Definition I/1.3.1) of X is

$$\mathrm{Var}(X) = \sigma^2 \; . \tag{2}$$

Proof: The variance (→ Definition I/1.3.1) is the probability-weighted average of the squared deviation from the mean (→ Definition I/1.2.1):

$$\mathrm{Var}(X) = \int_{\mathbb{R}} (x - \mathrm{E}(X))^2 \cdot f_\mathrm{X}(x) \, \mathrm{d}x \; . \tag{3}$$

With the expected value (→ Proof II/3.2.3) and probability density function (→ Proof II/3.2.2) of the normal distribution, this reads:

$$\begin{aligned} \mathrm{Var}(X) &= \int_{-\infty}^{+\infty} (x-\mu)^2 \cdot \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} (x-\mu)^2 \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \mathrm{d}x \; . \end{aligned} \tag{4}$$

Substituting $z = x - \mu$, we have:

$$\begin{aligned} \mathrm{Var}(X) &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty-\mu}^{+\infty-\mu} z^2 \cdot \exp\left[-\frac{1}{2}\left(\frac{z}{\sigma}\right)^2\right] \mathrm{d}(z+\mu) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} z^2 \cdot \exp\left[-\frac{1}{2}\left(\frac{z}{\sigma}\right)^2\right] \mathrm{d}z \; . \end{aligned} \tag{5}$$

Now substituting $z = \sqrt{2}\sigma x$, we have:

$$
\begin{split}
\mathrm{Var}(X) &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} (\sqrt{2}\sigma x)^2 \cdot \exp\left[-\frac{1}{2}\left(\frac{\sqrt{2}\sigma x}{\sigma}\right)^2\right] \, \mathrm{d}(\sqrt{2}\sigma x) \\
&= \frac{1}{\sqrt{2\pi}\sigma} \cdot 2\sigma^2 \cdot \sqrt{2}\sigma \int_{-\infty}^{+\infty} x^2 \cdot \exp\left[-x^2\right] \, \mathrm{d}x \\
&= \frac{2\sigma^2}{\sqrt{\pi}} \int_{-\infty}^{+\infty} x^2 \cdot e^{-x^2} \, \mathrm{d}x \; .
\end{split} \tag{6}
$$

Since the integrand is symmetric with respect to $x = 0$, we can write:

$$
\mathrm{Var}(X) = \frac{4\sigma^2}{\sqrt{\pi}} \int_{0}^{\infty} x^2 \cdot e^{-x^2} \, \mathrm{d}x \; . \tag{7}
$$

If we define $z = x^2$, then $x = \sqrt{z}$ and $\mathrm{d}x = 1/2 \, z^{-1/2} \, \mathrm{d}z$. Substituting this into the integral

$$
\mathrm{Var}(X) = \frac{4\sigma^2}{\sqrt{\pi}} \int_{0}^{\infty} z \cdot e^{-z} \cdot \frac{1}{2} z^{-\frac{1}{2}} \, \mathrm{d}z = \frac{2\sigma^2}{\sqrt{\pi}} \int_{0}^{\infty} z^{\frac{3}{2}-1} \cdot e^{-z} \, \mathrm{d}z \tag{8}
$$

and using the definition of the gamma function

$$
\Gamma(x) = \int_{0}^{\infty} z^{x-1} \cdot e^{-z} \, \mathrm{d}z \; , \tag{9}
$$

we can finally show that

$$
\mathrm{Var}(X) = \frac{2\sigma^2}{\sqrt{\pi}} \cdot \Gamma\left(\frac{3}{2}\right) = \frac{2\sigma^2}{\sqrt{\pi}} \cdot \frac{\sqrt{\pi}}{2} = \sigma^2 \; . \tag{10}
$$

Sources:

- Papadopoulos, Alecos (2013): "How to derive the mean and variance of Gaussian random variable?"; in: *StackExchange Mathematics*; URL: https://math.stackexchange.com/questions/518281/how-to-derive-the-mean-and-variance-of-a-gaussian-random-variable.

Metadata: ID: P18 | shortcut: norm-var | author: JoramSoch | date: 2020-01-09, 22:47.

3.2.7 Moment-generating function

Theorem: Let X be a random variable (→ Definition "rvar") following a normal distribution (→ Definition II/3.2.1):

$$
X \sim \mathcal{N}(\mu, \sigma^2) \; . \tag{1}
$$

Then, the moment-generating function (→ Definition I/1.1.5) of X is

$$
M_X(t) = \exp\left[\mu t + \frac{1}{2}\sigma^2 t^2\right] \; . \tag{2}
$$

Proof: The probability density function of the normal distribution (→ Proof II/3.2.2) is

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \tag{3}$$

and the moment-generating function (→ Definition I/1.1.5) is defined as

$$M_X(t) = \mathrm{E}\left[e^{tX}\right] \; . \tag{4}$$

Using the expected value for continuous random variables (→ Definition I/1.2.1), the moment-generating function of X therefore is

$$\begin{aligned} M_X(t) &= \int_{-\infty}^{+\infty} \exp[tx] \cdot \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \, \mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} \exp\left[tx - \frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \, \mathrm{d}x \; . \end{aligned} \tag{5}$$

Substituting $u = (x-\mu)/(\sqrt{2}\sigma)$, i.e. $x = \sqrt{2}\sigma u + \mu$, we have

$$\begin{aligned} M_X(t) &= \frac{1}{\sqrt{2\pi}\sigma} \int_{(-\infty-\mu)/(\sqrt{2}\sigma)}^{(+\infty-\mu)/(\sqrt{2}\sigma)} \exp\left[t\left(\sqrt{2}\sigma u + \mu\right) - \frac{1}{2}\left(\frac{\sqrt{2}\sigma u + \mu - \mu}{\sigma}\right)^2\right] \, \mathrm{d}\left(\sqrt{2}\sigma u + \mu\right) \\ &= \frac{\sqrt{2}\sigma}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} \exp\left[\left(\sqrt{2}\sigma u + \mu\right)t - u^2\right] \, \mathrm{d}u \\ &= \frac{\exp(\mu t)}{\sqrt{\pi}} \int_{-\infty}^{+\infty} \exp\left[\sqrt{2}\sigma u t - u^2\right] \, \mathrm{d}u \\ &= \frac{\exp(\mu t)}{\sqrt{\pi}} \int_{-\infty}^{+\infty} \exp\left[-\left(u^2 - \sqrt{2}\sigma u t\right)\right] \, \mathrm{d}u \\ &= \frac{\exp(\mu t)}{\sqrt{\pi}} \int_{-\infty}^{+\infty} \exp\left[-\left(u - \frac{\sqrt{2}}{2}\sigma t\right)^2 + \frac{1}{2}\sigma^2 t^2\right] \, \mathrm{d}u \\ &= \frac{\exp\left[\mu t + \frac{1}{2}\sigma^2 t^2\right]}{\sqrt{\pi}} \int_{-\infty}^{+\infty} \exp\left[-\left(u - \frac{\sqrt{2}}{2}\sigma t\right)^2\right] \, \mathrm{d}u \end{aligned} \tag{6}$$

Now substituting $v = u - \sqrt{2}/2\,\sigma t$, i.e. $u = v + \sqrt{2}/2\,\sigma t$, we have

$$\begin{aligned} M_X(t) &= \frac{\exp\left[\mu t + \frac{1}{2}\sigma^2 t^2\right]}{\sqrt{\pi}} \int_{-\infty-\sqrt{2}/2\,\sigma t}^{+\infty-\sqrt{2}/2\,\sigma t} \exp\left[-v^2\right] \, \mathrm{d}\left(v + \sqrt{2}/2\,\sigma t\right) \\ &= \frac{\exp\left[\mu t + \frac{1}{2}\sigma^2 t^2\right]}{\sqrt{\pi}} \int_{-\infty}^{+\infty} \exp\left[-v^2\right] \, \mathrm{d}v \; . \end{aligned} \tag{7}$$

With the Gaussian integral (→ Proof "norm-gi")

$$\int_{-\infty}^{+\infty} \exp\left[-x^2\right] \, \mathrm{d}x = \sqrt{\pi} \; , \tag{8}$$

this finally becomes

$$M_X(t) = \exp\left[\mu t + \frac{1}{2}\sigma^2 t^2\right] \; . \tag{9}$$

Sources:
- ProofWiki (2020): "Moment Generating Function of Gaussian Distribution"; in: *ProofWiki*, retrieved on 2020-03-03; URL: https://proofwiki.org/wiki/Moment_Generating_Function_of_Gaussian_Distribution.

Metadata: ID: P71 | shortcut: norm-mgf | author: JoramSoch | date: 2020-03-03, 11:29.

3.3 Gamma distribution

3.3.1 Definition

Definition: Let X be a random variable (→ Definition "rvar"). Then, X is said to follow a gamma distribution with shape a and rate b

$$X \sim \mathrm{Gam}(a,b) \; , \tag{1}$$

if and only if its probability density function (→ Definition I/1.1.2) is given by

$$\mathrm{Gam}(x; a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp[-bx], \quad x > 0 \tag{2}$$

where $a > 0$ and $b > 0$, and the density is zero, if $x \leq 0$.

Sources:
- Koch, Karl-Rudolf (2007): "Gamma Distribution"; in: *Introduction to Bayesian Statistics*, Springer, Berlin/Heidelberg, 2007, p. 47, eq. 2.172; URL: https://www.springer.com/de/book/9783540727231; DOI: 10.1007/978-3-540-72726-2.

Metadata: ID: D7 | shortcut: gam | author: JoramSoch | date: 2020-02-08, 23:29.

3.3.2 Probability density function

Theorem: Let X be a positive random variable (→ Definition "rvar") following a gamma distribution (→ Definition II/3.3.1):

$$X \sim \mathrm{Gam}(a,b) \; . \tag{1}$$

Then, the probability density function (→ Definition I/1.1.2) of X is

$$f_X(x) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp[-bx] \; . \tag{2}$$

Proof: This follows directly from the definition of the gamma distribution (→ Definition II/3.3.1).

Sources:

- original work

Metadata: ID: P45 | shortcut: gam-pdf | author: JoramSoch | date: 2020-02-08, 23:41.

3.4 Exponential distribution

3.4.1 Definition

Definition: Let X be a random variable (→ Definition "rvar"). Then, X is said to be exponentially distributed with rate (or, inverse scale) λ

$$X \sim \mathrm{Exp}(\lambda) \; , \tag{1}$$

if and only if its probability density function (→ Definition I/1.1.2) is given by

$$\mathrm{Exp}(x; \lambda) = \lambda \exp[-\lambda x], \quad x \geq 0 \tag{2}$$

where $\lambda > 0$, and the density is zero, if $x < 0$.

Sources:
- Wikipedia (2020): "Exponential distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-08; URL: https://en.wikipedia.org/wiki/Exponential_distribution#Definitions.

Metadata: ID: D8 | shortcut: exp | author: JoramSoch | date: 2020-02-08, 23:48.

3.4.2 Special case of gamma distribution

Theorem: The exponential distribution (→ Definition II/3.4.1) is a special case of the gamma distribution (→ Definition II/3.3.1) with shape $a = 1$ and rate $b = \lambda$.

Proof: The probability density function of the gamma distribution (→ Proof II/3.3.2) is

$$\mathrm{Gam}(x; a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp[-bx] \; . \tag{1}$$

Setting $a = 1$ and $b = \lambda$, we obtain

$$\begin{split} \mathrm{Gam}(x; 1, \lambda) &= \frac{\lambda^1}{\Gamma(1)} x^{1-1} \exp[-\lambda x] \\ &= \frac{x^0}{\Gamma(1)} \lambda \exp[-\lambda x] \\ &= \lambda \exp[-\lambda x] \end{split} \tag{2}$$

which is equivalent to the probability density function of the exponential distribution (→ Proof II/3.4.3).

Sources:
- original work

Metadata: ID: P69 | shortcut: exp-gam | author: JoramSoch | date: 2020-03-02, 20:49.

3.4.3 Probability density function

Theorem: Let X be a non-negative random variable (→ Definition "rvar") following an exponential distribution (→ Definition II/3.4.1):

$$X \sim \mathrm{Exp}(\lambda) \; . \tag{1}$$

Then, the probability density function (→ Definition I/1.1.2) of X is

$$f_X(x) = \lambda \exp[-\lambda x] \; . \tag{2}$$

Proof: This follows directly from the definition of the exponential distribution (→ Definition II/3.4.1).

Sources:
- original work

Metadata: ID: P46 | shortcut: exp-pdf | author: JoramSoch | date: 2020-02-08, 23:53.

3.4.4 Cumulative distribution function

Theorem: Let X be a random variable (→ Definition "rvar") following an exponential distribution (→ Definition II/3.4.1):

$$X \sim \mathrm{Exp}(\lambda) \; . \tag{1}$$

Then, the cumulative distribution function (→ Definition I/1.1.3) of X is

$$F_X(x) = \left\{ \begin{array}{rl} 0 \; , & \text{if} \; x < 0 \\ 1 - \exp[-\lambda x] \; , & \text{if} \; x \geq 0 \; . \end{array} \right. \tag{2}$$

Proof: The probability density function of the exponential distribution (→ Proof II/3.4.3) is:

$$\mathrm{Exp}(x; \lambda) = \left\{ \begin{array}{rl} 0 \; , & \text{if} \; x < 0 \\ \lambda \exp[-\lambda x] \; , & \text{if} \; x \geq 0 \; . \end{array} \right. \tag{3}$$

Thus, the cumulative distribution function (→ Definition I/1.1.3) is:

$$F_X(x) = \int_{-\infty}^{x} \mathrm{Exp}(z; \lambda) \, \mathrm{d}z \; . \tag{4}$$

If $x < 0$, we have:

$$F_X(x) = \int_{-\infty}^{x} 0 \, \mathrm{d}z = 0 \; . \tag{5}$$

If $x \geq 0$, we have using (3):

$$
\begin{aligned}
F_X(x) &= \int_{-\infty}^{0} \mathrm{Exp}(z;\lambda)\,\mathrm{d}z + \int_{0}^{x} \mathrm{Exp}(z;\lambda)\,\mathrm{d}z \\
&= \int_{-\infty}^{0} 0\,\mathrm{d}z + \int_{0}^{x} \lambda \exp[-\lambda z]\,\mathrm{d}z \\
&= 0 + \lambda \left[-\frac{1}{\lambda} \exp[-\lambda z] \right]_0^x \\
&= \lambda \left[\left(-\frac{1}{\lambda} \exp[-\lambda x] \right) - \left(-\frac{1}{\lambda} \exp[-\lambda \cdot 0] \right) \right] \\
&= 1 - \exp[-\lambda x] \; .
\end{aligned}
\tag{6}
$$

Sources:
- original work

Metadata: ID: P48 | shortcut: exp-cdf | author: JoramSoch | date: 2020-02-11, 14:48.

3.4.5 Quantile function

Theorem: Let X be a random variable (→ Definition "rvar") following an exponential distribution (→ Definition II/3.4.1):

$$
X \sim \mathrm{Exp}(\lambda) \; . \tag{1}
$$

Then, the quantile function (→ Definition I/1.1.4) of X is

$$
Q_X(p) = -\frac{\ln(1-p)}{\lambda} \; . \tag{2}
$$

Proof: The cumulative distribution function of the exponential distribution (→ Proof II/3.4.4) is:

$$
F_X(x) = \left\{ \begin{array}{rl} 0 \; , & \text{if } x < 0 \\ 1 - \exp[-\lambda x] \; , & \text{if } x \geq 0 \; . \end{array} \right. \tag{3}
$$

Thus, the quantile function (→ Definition I/1.1.4) is:

$$
Q_X(p) = F_X^{-1}(x) \; . \tag{4}
$$

This can be derived by rearranging equation (3):

$$
\begin{aligned}
p &= 1 - \exp[-\lambda x] \\
\exp[-\lambda x] &= 1 - p \\
-\lambda x &= \ln(1-p) \\
x &= -\frac{\ln(1-p)}{\lambda} \; .
\end{aligned}
\tag{5}
$$

Sources:

- original work

Metadata: ID: P50 | shortcut: exp-qf | author: JoramSoch | date: 2020-02-12, 15:48.

3.4.6 Mean

Theorem: Let X be a random variable (→ Definition "rvar") following an exponential distribution (→ Definition II/3.4.1):

$$X \sim \mathrm{Exp}(\lambda) \; . \tag{1}$$

Then, the mean or expected value (→ Definition I/1.2.1) of X is

$$\mathrm{E}(X) = \frac{1}{\lambda} \; . \tag{2}$$

Proof: The expected value (→ Definition I/1.2.1) is the probability-weighted average over all possible values:

$$\mathrm{E}(X) = \int_{\mathbb{R}} x \cdot f_{\mathrm{X}}(x) \, \mathrm{d}x \; . \tag{3}$$

With the probability density function of the exponential distribution (→ Proof II/3.4.3), this reads:

$$\begin{split} \mathrm{E}(X) &= \int_0^{+\infty} x \cdot \lambda \exp(-\lambda x) \, \mathrm{d}x \\ &= \lambda \int_0^{+\infty} x \cdot \exp(-\lambda x) \, \mathrm{d}x \; . \end{split} \tag{4}$$

Using the following anti-deriative

$$\int x \cdot \exp(-\lambda x) \, \mathrm{d}x = \left(-\frac{1}{\lambda} x - \frac{1}{\lambda^2} \right) \exp(-\lambda x) \; , \tag{5}$$

the expected value becomes

$$\begin{split} \mathrm{E}(X) &= \lambda \left[\left(-\frac{1}{\lambda} x - \frac{1}{\lambda^2} \right) \exp(-\lambda x) \right]_0^{+\infty} \\ &= \lambda \left[\lim_{x \to \infty} \left(-\frac{1}{\lambda} x - \frac{1}{\lambda^2} \right) \exp(-\lambda x) - \left(-\frac{1}{\lambda} \cdot 0 - \frac{1}{\lambda^2} \right) \exp(-\lambda \cdot 0) \right] \\ &= \lambda \left[0 + \frac{1}{\lambda^2} \right] \\ &= \frac{1}{\lambda} \; . \end{split} \tag{6}$$

Sources:

- Koch, Karl-Rudolf (2007): "Expected Value"; in: *Introduction to Bayesian Statistics*, Springer, Berlin/Heidelberg, 2007, p. 39, eq. 2.142a; URL: https://www.springer.com/de/book/9783540727231; DOI: 10.1007/978-3-540-72726-2.

Metadata: ID: P47 | shortcut: exp-mean | author: JoramSoch | date: 2020-02-10, 21:57.

3.4.7 Median

Theorem: Let X be a random variable (→ Definition "rvar") following an exponential distribution (→ Definition II/3.4.1):

$$X \sim \mathrm{Exp}(\lambda) \; . \tag{1}$$

Then, the median (→ Definition "med") of X is

$$\mathrm{median}(X) = \frac{\ln 2}{\lambda} \; . \tag{2}$$

Proof: The median (→ Definition "med") is the value at which the cumulative distribution function (→ Definition I/1.1.3) is 1/2:

$$F_X(\mathrm{median}(X)) = \frac{1}{2} \; . \tag{3}$$

The cumulative distribution function of the exponential distribution (→ Definition "exp-cdf") is

$$F_X(x) = 1 - \exp[-\lambda x], \quad x \geq 0 \; . \tag{4}$$

Thus, the inverse CDF is

$$x = -\frac{\ln(1-p)}{\lambda} \tag{5}$$

and setting $p = 1/2$, we obtain:

$$\mathrm{median}(X) = -\frac{\ln(1-\frac{1}{2})}{\lambda} = \frac{\ln 2}{\lambda} \; . \tag{6}$$

Sources:
- original work

Metadata: ID: P49 | shortcut: exp-med | author: JoramSoch | date: 2020-02-11, 15:03.

3.4.8 Mode

Theorem: Let X be a random variable (→ Definition "rvar") following an exponential distribution (→ Definition II/3.4.1):

$$X \sim \mathrm{Exp}(\lambda) \; . \tag{1}$$

Then, the mode (→ Definition "mode") of X is

$$\mathrm{mode}(X) = 0 \; . \tag{2}$$

Proof: The mode (→ Definition "mode") is the value which maximizes the probability density function (→ Definition I/1.1.2):

$$\mathrm{mode}(X) = \operatorname*{arg\,max}_x f_X(x) \; . \tag{3}$$

The probability density function of the exponential distribution (→ Proof II/3.4.3) is:

$$f_X(x) = \left\{ \begin{array}{rl} 0 \; , & \text{if} \; x < 0 \\ \lambda \exp[-\lambda x] \; , & \text{if} \; x \geq 0 \; . \end{array} \right. \tag{4}$$

Since

$$\lim_{x \rightarrow 0} f_X(x) = \infty \tag{5}$$

and

$$f_X(x) < \infty \quad \text{for any} \quad x \neq 0 \; , \tag{6}$$

it follows that

$$\mathrm{mode}(X) = 0 \; . \tag{7}$$

Sources:
- original work

Metadata: ID: P51 | shortcut: exp-mode | author: JoramSoch | date: 2020-02-12, 15:53.

4 Multivariate continuous distributions

4.1 Multivariate normal distribution

4.1.1 Definition

Definition: Let X be an $n \times 1$ random vector (→ Definition "rvec"). Then, X is said to be multivariate normally distributed with mean μ and covariance Σ

$$X \sim \mathcal{N}(\mu, \Sigma) \; , \tag{1}$$

if and only if its probability density function (→ Definition I/1.1.2) is given by

$$\mathcal{N}(x; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \cdot \exp\left[-\frac{1}{2} (x-\mu)^\mathrm{T} \Sigma^{-1} (x-\mu) \right] \tag{2}$$

where μ is an $n \times 1$ real vector and Σ is an $n \times n$ positive definite matrix.

Sources:
- Koch KR (2007): "Multivariate Normal Distribution"; in: *Introduction to Bayesian Statistics*, ch. 2.5.1, pp. 51-53, eq. 2.195; URL: https://www.springer.com/gp/book/9783540727231; DOI: 10.1007/978-3-540-72726-2.

Metadata: ID: D1 | shortcut: mvn | author: JoramSoch | date: 2020-01-22, 05:20.

4.1.2 Probability density function

Theorem: Let X be a random vector (→ Definition "rvec") following a multivariate normal distribution (→ Definition II/4.1.1):

$$X \sim \mathcal{N}(\mu, \Sigma) \; . \tag{1}$$

Then, the probability density function (→ Definition I/1.1.2) of X is

$$f_X(x) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \cdot \exp\left[-\frac{1}{2} (x-\mu)^\mathrm{T} \Sigma^{-1} (x-\mu) \right] \; . \tag{2}$$

Proof: This follows directly from the definition of the multivariate normal distribution (→ Definition II/4.1.1).

Sources:
- original work

Metadata: ID: P34 | shortcut: mvn-pdf | author: JoramSoch | date: 2020-01-27, 15:23.

4.1.3 Linear transformation theorem

Theorem: Let x follow a multivariate normal distribution (→ Definition II/4.1.1):

$$x \sim \mathcal{N}(\mu, \Sigma) \; . \tag{1}$$

Then, any linear transformation of x is also multivariate normally distributed:

$$y = Ax + b \sim \mathcal{N}(A\mu + b, A\Sigma A^\mathrm{T}) \; . \tag{2}$$

Proof: The moment-generating function of a random vector (→ Definition I/1.1.5) x is

$$M_x(t) = \mathbb{E}\left(\exp\left[t^\mathrm{T} x\right]\right) \tag{3}$$

and therefore the moment-generating function of the random vector y is given by

$$\begin{split} M_y(t) &= \mathbb{E}\left(\exp\left[t^\mathrm{T}(Ax+b)\right]\right) \\ &= \mathbb{E}\left(\exp\left[t^\mathrm{T}Ax\right] \cdot \exp\left[t^\mathrm{T}b\right]\right) \\ &= \exp\left[t^\mathrm{T}b\right] \cdot \mathbb{E}\left(\exp\left[t^\mathrm{T}Ax\right]\right) \\ &= \exp\left[t^\mathrm{T}b\right] \cdot M_x(At) \; . \end{split} \tag{4}$$

The moment-generating function of the multivariate normal distribution (→ Proof "mvn-mgf") is

$$M_x(t) = \exp\left[t^\mathrm{T}\mu + \frac{1}{2}t^\mathrm{T}\Sigma t\right] \tag{5}$$

and therefore the moment-generating function of the random vector y becomes

$$\begin{split} M_y(t) &= \exp\left[t^\mathrm{T}b\right] \cdot M_x(At) \\ &= \exp\left[t^\mathrm{T}b\right] \cdot \exp\left[t^\mathrm{T}A\mu + \frac{1}{2}t^\mathrm{T}A\Sigma A^\mathrm{T}t\right] \\ &= \exp\left[t^\mathrm{T}\left(A\mu + b\right) + \frac{1}{2}t^\mathrm{T}A\Sigma A^\mathrm{T}t\right] \; . \end{split} \tag{6}$$

Because moment-generating function and probability density function of a random variable are equivalent, this demonstrates that y is following a multivariate normal distribution with mean $A\mu + b$ and covariance $A\Sigma A^\mathrm{T}$.

Sources:

- Taboga, Marco (2010): "Linear combinations of normal random variables"; in: *Lectures on probability and statistics*; URL: https://www.statlect.com/probability-distributions/normal-distribution-linear-co

Metadata: ID: P1 | shortcut: mvn-ltt | author: JoramSoch | date: 2019-08-27, 12:14.

4.1.4 Marginal distributions

Theorem: Let x follow a multivariate normal distribution (→ Definition II/4.1.1):

$$x \sim \mathcal{N}(\mu, \Sigma) \; . \tag{1}$$

Then, the marginal distribution (→ Definition "md") of any subset vector x_s is also a multivariate normal distribution

$$x_s \sim \mathcal{N}(\mu_s, \Sigma_s) \tag{2}$$

where μ_s drops the irrelevant variables (the ones not in the subset, i.e. marginalized out) from the mean vector μ and Σ_s drops the corresponding rows and columns from the covariance matrix Σ.

Proof: Define an $m \times n$ subset matrix S such that $s_{ij} = 1$, if the j-th element in μ_s corresponds to the i-th element in x, and $s_{ij} = 0$ otherwise. Then,

$$x_s = Sx \tag{3}$$

and we can apply the linear transformation theorem (→ Proof II/4.1.3) to give

$$x_s \sim \mathcal{N}(S\mu, S\Sigma S^{\mathrm{T}}) \; . \tag{4}$$

Finally, we see that $S\mu = \mu_s$ and $S\Sigma S^{\mathrm{T}} = \Sigma_s$.

Sources:
- original work

Metadata: ID: P35 | shortcut: mvn-marg | author: JoramSoch | date: 2020-01-29, 15:12.

4.2 Normal-gamma distribution

4.2.1 Definition

Definition: Let X be an $n \times 1$ random vector (→ Definition "rvec") and let Y be a positive random variable (→ Definition "rvar"). Then, X and Y are said to follow a normal-gamma distribution

$$X, Y \sim \mathrm{NG}(\mu, \Lambda, a, b) \; , \tag{1}$$

if and only if their joint probability (→ Definition "prob-joint") density function (→ Definition I/1.1.2) is given by

$$f_{X,Y}(x, y) = \mathcal{N}(x; \mu, (y\Lambda)^{-1}) \cdot \mathrm{Gam}(y; a, b) \tag{2}$$

where $\mathcal{N}(x; \mu, \Sigma)$ is the probability density function of the multivariate normal distribution (→ Proof II/4.1.2) with mean μ and covariance Σ and $\mathrm{Gam}(x; a, b)$ is the probability density function of the gamma distribution (→ Proof II/3.3.2) with shape a and rate b. The $n \times n$ matrix Λ is referred to as the precision matrix of the normal-gamma distribution.

Sources:
- Koch KR (2007): "Normal-Gamma Distribution"; in: *Introduction to Bayesian Statistics*, ch. 2.5.3, pp. 55-56, eq. 2.212; URL: https://www.springer.com/gp/book/9783540727231; DOI: 10.1007/978-3-540-72726-2.

Metadata: ID: D5 | shortcut: ng | author: JoramSoch | date: 2020-01-27, 14:28.

4.2.2 Probability density function

Theorem: Let x and y follow a normal-gamma distribution (→ Definition II/4.2.1):

$$x, y \sim \mathrm{NG}(\mu, \Lambda, a, b) \; . \tag{1}$$

Then, the joint probability (→ Definition "prob-joint") density function (→ Definition I/1.1.2) of x and y is

$$p(x,y) = \sqrt{\frac{|\Lambda|}{(2\pi)^n}} \frac{b^a}{\Gamma(a)} \cdot y^{a+\frac{n}{2}-1} \exp\left[-\frac{y}{2}\left((x-\mu)^\mathrm{T}\Lambda(x-\mu)+2b\right)\right] \; . \tag{2}$$

Proof: The probability density of the normal-gamma distribution is defined as (→ Definition II/4.2.1) as the product of a multivariate normal distribution (→ Definition II/4.1.1) over x conditional on y and a univariate gamma distribution (→ Definition II/3.3.1) over y:

$$p(x,y) = \mathcal{N}(x;\mu,(y\Lambda)^{-1}) \cdot \mathrm{Gam}(y;a,b) \tag{3}$$

With the probability density function of the multivariate normal distribution (→ Proof II/4.1.2) and the probability density function of the gamma distribution (→ Proof II/3.3.2), this becomes:

$$p(x,y) = \sqrt{\frac{|y\Lambda|}{(2\pi)^n}} \exp\left[-\frac{1}{2}(x-\mu)^\mathrm{T}(y\Lambda)(x-\mu)\right] \cdot \frac{b^a}{\Gamma(a)} y^{a-1} \exp\left[-by\right] \; . \tag{4}$$

Using the relation $|yA| = y^n |A|$ for an $n \times n$ matrix A and rearranging the terms, we have:

$$p(x,y) = \sqrt{\frac{|\Lambda|}{(2\pi)^n}} \frac{b^a}{\Gamma(a)} \cdot y^{a+\frac{n}{2}-1} \exp\left[-\frac{y}{2}\left((x-\mu)^\mathrm{T}\Lambda(x-\mu)+2b\right)\right] \; . \tag{5}$$

Sources:

- Koch KR (2007): "Normal-Gamma Distribution"; in: *Introduction to Bayesian Statistics*, ch. 2.5.3, pp. 55-56, eq. 2.212; URL: https://www.springer.com/gp/book/9783540727231; DOI: 10.1007/978-3-540-72726-2.

Metadata: ID: P44 | shortcut: ng-pdf | author: JoramSoch | date: 2020-02-07, 20:44.

4.2.3 Kullback-Leibler divergence

Theorem: Let $x \in \mathbb{R}^k$ be a random vector (→ Definition "rvec") and $y > 0$ be a random variable (→ Definition "rvar"). Assume two normal-gamma distributions (→ Definition II/4.2.1) P and Q specifying the joint distribution of x and y as

$$\begin{aligned} P: \; (x,y) &\sim \mathrm{NG}(\mu_1, \Lambda_1^{-1}, a_1, b_1) \\ Q: \; (x,y) &\sim \mathrm{NG}(\mu_2, \Lambda_2^{-1}, a_2, b_2) \; . \end{aligned} \tag{1}$$

Then, the Kullback-Leibler divergence (→ Definition "kl") of P from Q is given by

$$\begin{aligned} \mathrm{KL}[P\,||\,Q] = &\frac{1}{2}\frac{a_1}{b_1}\left[(\mu_2-\mu_1)^T\Lambda_2(\mu_2-\mu_1)\right] + \frac{1}{2}\,\mathrm{tr}(\Lambda_2\Lambda_1^{-1}) - \frac{1}{2}\ln\frac{|\Lambda_2|}{|\Lambda_1|} - \frac{k}{2} \\ &+ a_2\,\ln\frac{b_1}{b_2} - \ln\frac{\Gamma(a_1)}{\Gamma(a_2)} + (a_1-a_2)\,\psi(a_1) - (b_1-b_2)\,\frac{a_1}{b_1} \; . \end{aligned} \tag{2}$$

Proof: The probabibility density function of the normal-gamma distribution (→ Proof II/4.2.2) is

$$p(x,y) = p(x|y) \cdot p(y) = \mathcal{N}(x; \mu, (y\Lambda)^{-1}) \cdot \mathrm{Gam}(y; a, b) \tag{3}$$

where $\mathcal{N}(x; \mu, \Sigma)$ is a multivariate normal density with mean μ and covariance Σ (hence, precision Λ) and $\mathrm{Gam}(y; a, b)$ is a univariate gamma density with shape a and rate b. The Kullback-Leibler divergence of the multivariate normal distribution (→ Proof "mvn-kl") is

$$\mathrm{KL}[P\,||\,Q] = \frac{1}{2}\left[(\mu_2-\mu_1)^T\Sigma_2^{-1}(\mu_2-\mu_1) + \mathrm{tr}(\Sigma_2^{-1}\Sigma_1) - \ln\frac{|\Sigma_1|}{|\Sigma_2|} - k\right] \tag{4}$$

and the Kullback-Leibler divergence of the univariate gamma distribution (→ Proof "gam-kl") is

$$\mathrm{KL}[P\,||\,Q] = a_2\,\ln\frac{b_1}{b_2} - \ln\frac{\Gamma(a_1)}{\Gamma(a_2)} + (a_1-a_2)\,\psi(a_1) - (b_1-b_2)\,\frac{a_1}{b_1} \tag{5}$$

where $\Gamma(x)$ is the gamma function and $\psi(x)$ is the digamma function.

The KL divergence for a continuous random variable (→ Definition "kl") is given by

$$\mathrm{KL}[P\,||\,Q] = \int_{\mathcal{Z}} p(z)\,\ln\frac{p(z)}{q(z)}\,\mathrm{d}z \tag{6}$$

which, applied to the normal-gamma distribution (→ Definition II/4.2.1) over x and y, yields

$$\mathrm{KL}[P\,||\,Q] = \int_0^\infty \int_{\mathbb{R}^k} p(x,y)\,\ln\frac{p(x,y)}{q(x,y)}\,\mathrm{d}x\,\mathrm{d}y\;. \tag{7}$$

Using the law of conditional probability (→ Definition "prob-cond"), this can be evaluated as follows:

$$\begin{split}
\mathrm{KL}[P\,||\,Q] &= \int_0^\infty \int_{\mathbb{R}^k} p(x|y)\,p(y)\,\ln\frac{p(x|y)\,p(y)}{q(x|y)\,q(y)}\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_0^\infty \int_{\mathbb{R}^k} p(x|y)\,p(y)\,\ln\frac{p(x|y)}{q(x|y)}\,\mathrm{d}x\,\mathrm{d}y + \int_0^\infty \int_{\mathbb{R}^k} p(x|y)\,p(y)\,\ln\frac{p(y)}{q(y)}\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_0^\infty p(y) \int_{\mathbb{R}^k} p(x|y)\,\ln\frac{p(x|y)}{q(x|y)}\,\mathrm{d}x\,\mathrm{d}y + \int_0^\infty p(y)\,\ln\frac{p(y)}{q(y)} \int_{\mathbb{R}^k} p(x|y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \left\langle \mathrm{KL}[p(x|y)\,||\,q(x|y)] \right\rangle_{p(y)} + \mathrm{KL}[p(y)\,||\,q(y)]\;.
\end{split} \tag{8}$$

In other words, the KL divergence between two normal-gamma distributions over x and y is equal to the sum of a multivariate normal KL divergence regarding x conditional on y, expected over y, and a univariate gamma KL divergence regarding y.

From equations (3) and (4), the first term becomes

$$\begin{split}
&\left\langle \mathrm{KL}[p(x|y)\,||\,q(x|y)] \right\rangle_{p(y)} \\
&= \left\langle \frac{1}{2}\left[(\mu_2-\mu_1)^T(y\Lambda_2)(\mu_2-\mu_1) + \mathrm{tr}\left((y\Lambda_2)(y\Lambda_1)^{-1}\right) - \ln\frac{|(y\Lambda_1)^{-1}|}{|(y\Lambda_2)^{-1}|} - k\right] \right\rangle_{p(y)} \\
&= \left\langle \frac{y}{2}(\mu_2-\mu_1)^T\Lambda_2(\mu_2-\mu_1) + \frac{1}{2}\,\mathrm{tr}(\Lambda_2\Lambda_1^{-1}) - \frac{1}{2}\ln\frac{|\Lambda_2|}{|\Lambda_1|} - \frac{k}{2} \right\rangle_{p(y)}
\end{split} \tag{9}$$

and using the relation (→ Proof "gam-mean") $y \sim \mathrm{Gam}(a,b) \Rightarrow \langle y \rangle = a/b$, we have

$$\left\langle \mathrm{KL}[p(x|y)\,||\,q(x|y)] \right\rangle_{p(y)} = \frac{1}{2}\frac{a_1}{b_1}(\mu_2-\mu_1)^T\Lambda_2(\mu_2-\mu_1) + \frac{1}{2}\,\mathrm{tr}(\Lambda_2\Lambda_1^{-1}) - \frac{1}{2}\ln\frac{|\Lambda_2|}{|\Lambda_1|} - \frac{k}{2} \; . \tag{10}$$

By plugging (10) and (5) into (8), one arrives at the KL divergence given by (2).

Sources:
- Soch & Allefeld (2016): "Kullback-Leibler Divergence for the Normal-Gamma Distribution"; in: *arXiv math.ST*, 1611.01437; URL: https://arxiv.org/abs/1611.01437.

Metadata: ID: P6 | shortcut: ng-kl | author: JoramSoch | date: 2019-12-06, 09:35.

4.2.4 Marginal distributions

Theorem: Let x and y follow a normal-gamma distribution (→ Definition II/4.2.1):

$$x, y \sim \mathrm{NG}(\mu, \Lambda, a, b) \; . \tag{1}$$

Then, the marginal distribution (→ Definition "md") of y is a gamma distribution (→ Definition II/3.3.1)

$$y \sim \mathrm{Gam}(a, b) \tag{2}$$

and the marginal distribution (→ Definition "md") of x is a multivariate t-distribution (→ Definition "mvt")

$$x \sim \mathrm{t}\left(\mu, \left(\frac{a}{b}\Lambda\right)^{-1}, 2a\right) \; . \tag{3}$$

Proof: The probability density function of the normal-gamma distribution (→ Proof II/4.2.2) is given by

$$\begin{aligned} p(x,y) &= p(x|y) \cdot p(y) \\ p(x|y) &= \mathcal{N}(x; \mu, (y\Lambda)^{-1}) \\ p(y) &= \mathrm{Gam}(y; a, b) \; . \end{aligned} \tag{4}$$

Using the law of marginal probability (→ Definition "prob-marg"), the marginal distribution of y can be derived as

$$\begin{aligned} p(y) &= \int p(x,y)\,\mathrm{d}x \\ &= \int \mathcal{N}(x; \mu, (y\Lambda)^{-1})\,\mathrm{Gam}(y; a, b)\,\mathrm{d}x \\ &= \mathrm{Gam}(y; a, b) \int \mathcal{N}(x; \mu, (y\Lambda)^{-1})\,\mathrm{d}x \\ &= \mathrm{Gam}(y; a, b) \end{aligned} \tag{5}$$

which is the probability density function of the gamma distribution (→ Proof II/3.3.2) with shape parameter a and rate parameter b.

Using the law of marginal probability (→ Definition “prob-marg”), the marginal distribution of x can be derived as

$$
\begin{split}
p(x) &= \int p(x,y)\,\mathrm{d}y \\
&= \int \mathcal{N}(x;\mu,(y\Lambda)^{-1})\,\mathrm{Gam}(y;a,b)\,\mathrm{d}y \\
&= \int \sqrt{\frac{|y\Lambda|}{\sqrt{(2\pi)^n}}}\,\exp\left[-\frac{1}{2}(x-\mu)^\mathrm{T}(y\Lambda)(x-\mu)\right]\cdot\frac{b^a}{\Gamma(a)}\,y^{a-1}\exp[-by]\,\mathrm{d}y \\
&= \int \sqrt{\frac{y^n|\Lambda|}{\sqrt{(2\pi)^n}}}\,\exp\left[-\frac{1}{2}(x-\mu)^\mathrm{T}(y\Lambda)(x-\mu)\right]\cdot\frac{b^a}{\Gamma(a)}\,y^{a-1}\exp[-by]\,\mathrm{d}y \\
&= \int \sqrt{\frac{|\Lambda|}{\sqrt{(2\pi)^n}}}\cdot\frac{b^a}{\Gamma(a)}\cdot y^{a+\frac{n}{2}-1}\cdot\exp\left[-\left(b+\frac{1}{2}(x-\mu)^\mathrm{T}\Lambda(x-\mu)\right)y\right]\mathrm{d}y \\
&= \int \sqrt{\frac{|\Lambda|}{\sqrt{(2\pi)^n}}}\cdot\frac{b^a}{\Gamma(a)}\cdot\frac{\Gamma\left(a+\frac{n}{2}\right)}{\left(b+\frac{1}{2}(x-\mu)^\mathrm{T}\Lambda(x-\mu)\right)^{a+\frac{n}{2}}}\cdot\mathrm{Gam}\left(y;a+\frac{n}{2},b+\frac{1}{2}(x-\mu)^\mathrm{T}\Lambda(x-\mu)\right)\mathrm{d}y \\
&= \sqrt{\frac{|\Lambda|}{\sqrt{(2\pi)^n}}}\cdot\frac{b^a}{\Gamma(a)}\cdot\frac{\Gamma\left(a+\frac{n}{2}\right)}{\left(b+\frac{1}{2}(x-\mu)^\mathrm{T}\Lambda(x-\mu)\right)^{a+\frac{n}{2}}}\int\mathrm{Gam}\left(y;a+\frac{n}{2},b+\frac{1}{2}(x-\mu)^\mathrm{T}\Lambda(x-\mu)\right)\mathrm{d}y \\
&= \sqrt{\frac{|\Lambda|}{\sqrt{(2\pi)^n}}}\cdot\frac{b^a}{\Gamma(a)}\cdot\frac{\Gamma\left(a+\frac{n}{2}\right)}{\left(b+\frac{1}{2}(x-\mu)^\mathrm{T}\Lambda(x-\mu)\right)^{a+\frac{n}{2}}} \\
&= \frac{\sqrt{|\Lambda|}}{(2\pi)^{\frac{n}{2}}}\cdot\frac{\Gamma\left(\frac{2a+n}{2}\right)}{\Gamma\left(\frac{2a}{2}\right)}\cdot b^a\cdot\left(b+\frac{1}{2}(x-\mu)^\mathrm{T}\Lambda(x-\mu)\right)^{-\left(a+\frac{n}{2}\right)} \\
&= \frac{\sqrt{|\Lambda|}}{\pi^{\frac{n}{2}}}\cdot\frac{\Gamma\left(\frac{2a+n}{2}\right)}{\Gamma\left(\frac{2a}{2}\right)}\cdot\left(\frac{1}{b}\right)^{-a}\cdot\left(b+\frac{1}{2}(x-\mu)^\mathrm{T}\Lambda(x-\mu)\right)^{-a}\cdot 2^{-\frac{n}{2}}\cdot\left(b+\frac{1}{2}(x-\mu)^\mathrm{T}\Lambda(x-\mu)\right)^{-\frac{n}{2}} \\
&= \frac{\sqrt{|\Lambda|}}{\pi^{\frac{n}{2}}}\cdot\frac{\Gamma\left(\frac{2a+n}{2}\right)}{\Gamma\left(\frac{2a}{2}\right)}\cdot\left(1+\frac{1}{2b}(x-\mu)^\mathrm{T}\Lambda(x-\mu)\right)^{-a}\cdot\left(2b+(x-\mu)^\mathrm{T}\Lambda(x-\mu)\right)^{-\frac{n}{2}} \\
&= \frac{\sqrt{|\Lambda|}}{\pi^{\frac{n}{2}}}\cdot\frac{\Gamma\left(\frac{2a+n}{2}\right)}{\Gamma\left(\frac{2a}{2}\right)}\cdot\left(\frac{1}{2a}\right)^{-a}\cdot\left(2a+(x-\mu)^\mathrm{T}\left(\frac{a}{b}\Lambda\right)(x-\mu)\right)^{-a}\cdot\left(\frac{b}{a}\right)^{-\frac{n}{2}}\cdot\bigg(2a+(x-\mu)^\mathrm{T}\left(\frac{a}{b}\Lambda\right)(x-\mu \\
&= \frac{\sqrt{\left(\frac{a}{b}\right)^n|\Lambda|}}{(2a)^{-a}\,\pi^{\frac{n}{2}}}\cdot\frac{\Gamma\left(\frac{2a+n}{2}\right)}{\Gamma\left(\frac{2a}{2}\right)}\cdot\left(2a+(x-\mu)^\mathrm{T}\left(\frac{a}{b}\Lambda\right)(x-\mu)\right)^{-a}\cdot\left(2a+(x-\mu)^\mathrm{T}\left(\frac{a}{b}\Lambda\right)(x-\mu)\right)^{-\frac{n}{2}} \\
&= \frac{\sqrt{\left(\frac{a}{b}\right)^n|\Lambda|}}{(2a)^{-a}\,\pi^{\frac{n}{2}}}\cdot\frac{\Gamma\left(\frac{2a+n}{2}\right)}{\Gamma\left(\frac{2a}{2}\right)}\cdot(2a)^{-a}\cdot\left(1+\frac{1}{2a}(x-\mu)^\mathrm{T}\left(\frac{a}{b}\Lambda\right)(x-\mu)\right)^{-a}\cdot(2a)^{-\frac{n}{2}}\cdot\bigg(1+\frac{1}{2a}(x-\mu)^\mathrm{T}\left(\frac{a}{b}\Lambda\right) \\
&= \frac{\sqrt{\left(\frac{a}{b}\right)^n|\Lambda|}}{(2a)^{\frac{n}{2}}\,\pi^{\frac{n}{2}}}\cdot\frac{\Gamma\left(\frac{2a+n}{2}\right)}{\Gamma\left(\frac{2a}{2}\right)}\cdot\left(1+\frac{1}{2a}(x-\mu)^\mathrm{T}\left(\frac{a}{b}\Lambda\right)(x-\mu)\right)^{-\frac{2a+n}{2}} \\
&= \sqrt{\frac{\left|\frac{a}{b}\,\Lambda\right|}{(2a\,\pi)^n}}\cdot\frac{\Gamma\left(\frac{2a+n}{2}\right)}{\Gamma\left(\frac{2a}{2}\right)}\cdot\left(1+\frac{1}{2a}(x-\mu)^\mathrm{T}\left(\frac{a}{b}\Lambda\right)(x-\mu)\right)^{-\frac{2a+n}{2}}
\end{split}
\tag{6}
$$

which is the probability density function of a multivariate t-distribution (→ Proof "mvt-pdf") with mean vector μ, shape matrix $\left(\frac{a}{b}\Lambda\right)^{-1}$ and $2a$ degrees of freedom.

Sources:
- original work

Metadata: ID: P36 | shortcut: ng-marg | author: JoramSoch | date: 2020-01-29, 21:42.

5 Matrix-variate continuous distributions

5.1 Matrix-normal distribution

5.1.1 Definition

Definition: Let X be an $n \times p$ random matrix (→ Definition "rmat"). Then, X is said to be matrix-normally distributed with mean M, covariance across rows U and covariance across columns V

$$X \sim \mathcal{MN}(M, U, V) \; , \tag{1}$$

if and only if its probability density function (→ Definition I/1.1.2) is given by

$$\mathcal{MN}(X; M, U, V) = \frac{1}{\sqrt{(2\pi)^{np} |V|^n |U|^p}} \cdot \exp\left[-\frac{1}{2} \mathrm{tr}\left(V^{-1} (X-M)^\mathrm{T} \, U^{-1} (X-M) \right) \right] \tag{2}$$

where μ is an $n \times p$ real matrix, U is an $n \times n$ positive definite matrix and V is a $p \times p$ positive definite matrix.

Sources:
- Wikipedia (2020): "Matrix normal distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-01-27; URL: https://en.wikipedia.org/wiki/Matrix_normal_distribution#Definition.

Metadata: ID: D6 | shortcut: matn | author: JoramSoch | date: 2020-01-27, 14:37.

5.1.2 Probability density function

Theorem: Let X be a random matrix (→ Definition "rmat") following a matrix-normal distribution (→ Definition II/5.1.1):

$$X \sim \mathcal{MN}(M, U, V) \; . \tag{1}$$

Then, the probability density function (→ Definition I/1.1.2) of X is

$$f_X(x) = \frac{1}{\sqrt{(2\pi)^{np} |V|^n |U|^p}} \cdot \exp\left[-\frac{1}{2} \mathrm{tr}\left(V^{-1} (X-M)^\mathrm{T} \, U^{-1} (X-M) \right) \right] \; . \tag{2}$$

Proof: This follows directly from the definition of the matrix-normal distribution (→ Definition II/5.1.1).

Sources:
- original work

Metadata: ID: P70 | shortcut: matn-pdf | author: JoramSoch | date: 2020-02-03, 21:03.

5.1.3 Equivalence to multivariate normal distribution

Theorem: The matrix X is matrix-normally distributed (→ Definition II/5.1.1)

$$X \sim \mathcal{MN}(M, U, V) \; , \tag{1}$$

if and only if $\mathrm{vec}(X)$ is multivariate normally distributed (→ Definition II/4.1.1)

$$\mathrm{vec}(X) \sim \mathcal{MN}(\mathrm{vec}(M), V \otimes U) \tag{2}$$

where $\mathrm{vec}(X)$ is the vectorization operator and $\otimes$ is the Kronecker product.

Proof: The probability density function of the matrix-normal distribution (→ Proof II/5.1.2) with $n \times p$ mean M, $n \times n$ covariance across rows U and $p \times p$ covariance across columns V is

$$\mathcal{MN}(X; M, U, V) = \frac{1}{\sqrt{(2\pi)^{np}|V|^n|U|^p}} \cdot \exp\left[-\frac{1}{2}\mathrm{tr}\left(V^{-1}(X-M)^\mathrm{T}\,U^{-1}(X-M)\right)\right] \; . \tag{3}$$

Using the trace property $\mathrm{tr}(ABC) = \mathrm{tr}(BCA)$, we have:

$$\mathcal{MN}(X; M, U, V) = \frac{1}{\sqrt{(2\pi)^{np}|V|^n|U|^p}} \cdot \exp\left[-\frac{1}{2}\mathrm{tr}\left((X-M)^\mathrm{T}\,U^{-1}(X-M)\,V^{-1}\right)\right] \; . \tag{4}$$

Using the trace-vectorization relation $\mathrm{tr}(A^\mathrm{T}B) = \mathrm{vec}(A)^\mathrm{T}\,\mathrm{vec}(B)$, we have:

$$\mathcal{MN}(X; M, U, V) = \frac{1}{\sqrt{(2\pi)^{np}|V|^n|U|^p}} \cdot \exp\left[-\frac{1}{2}\mathrm{vec}(X-M)^\mathrm{T}\,\mathrm{vec}\left(U^{-1}(X-M)\,V^{-1}\right)\right] \; . \tag{5}$$

Using the vectorization-Kronecker relation $\mathrm{vec}(ABC) = \left(C^\mathrm{T} \otimes A\right)\mathrm{vec}(B)$, we have:

$$\mathcal{MN}(X; M, U, V) = \frac{1}{\sqrt{(2\pi)^{np}|V|^n|U|^p}} \cdot \exp\left[-\frac{1}{2}\mathrm{vec}(X-M)^\mathrm{T}\,\left(V^{-1} \otimes U^{-1}\right)\mathrm{vec}(X-M)\right] \; . \tag{6}$$

Using the Kronecker product property $(A^{-1} \otimes B^{-1}) = (A \otimes B)^{-1}$, we have:

$$\mathcal{MN}(X; M, U, V) = \frac{1}{\sqrt{(2\pi)^{np}|V|^n|U|^p}} \cdot \exp\left[-\frac{1}{2}\mathrm{vec}(X-M)^\mathrm{T}\,(V \otimes U)^{-1}\,\mathrm{vec}(X-M)\right] \; . \tag{7}$$

Using the vectorization property $\mathrm{vec}(A+B) = \mathrm{vec}(A) + \mathrm{vec}(B)$, we have:

$$\mathcal{MN}(X; M, U, V) = \frac{1}{\sqrt{(2\pi)^{np}|V|^n|U|^p}} \cdot \exp\left[-\frac{1}{2}\left[\mathrm{vec}(X) - \mathrm{vec}(M)\right]^\mathrm{T}\,(V \otimes U)^{-1}\left[\mathrm{vec}(X) - \mathrm{vec}(M)\right]\right] \; . \tag{8}$$

Using the Kronecker-determinant relation $|A \otimes B| = |A|^m|B|^n$, we have:

$$\mathcal{MN}(X; M, U, V) = \frac{1}{\sqrt{(2\pi)^{np}|V \otimes U|}} \cdot \exp\left[-\frac{1}{2}\left[\mathrm{vec}(X) - \mathrm{vec}(M)\right]^\mathrm{T}\,(V \otimes U)^{-1}\left[\mathrm{vec}(X) - \mathrm{vec}(M)\right]\right] \; . \tag{9}$$

This is the probability density function of the multivariate normal distribution (→ Proof II/4.1.2) with the $np \times 1$ mean vector $\mathrm{vec}(M)$ and the $np \times np$ covariance matrix $V \otimes U$:

$$\mathcal{MN}(X; M, U, V) = \mathcal{N}(\mathrm{vec}(X); \mathrm{vec}(M), V \otimes U) \; . \tag{10}$$

By showing that the probability density functions (→ Definition I/1.1.2) are identical, it is proven that the associated probability distributions (→ Definition "pd") are equivalent.

Sources:

- Wikipedia (2020): "Matrix normal distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-01-20; URL: https://en.wikipedia.org/wiki/Matrix_normal_distribution#Proof.

Metadata: ID: P26 | shortcut: matn-mvn | author: JoramSoch | date: 2020-01-20, 21:09.

Chapter III

Statistical Models

1 Normal data

1.1 Multiple linear regression

1.1.1 Ordinary least squares (1)

Theorem: Given a linear regression model (→ Definition "mlr") with independent observations

$$y = X\beta + \varepsilon, \; \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \; , \tag{1}$$

the parameters minimizing the residual sum of squares (→ Definition "rss") are given by

$$\hat{\beta} = (X^\mathrm{T} X)^{-1} X^\mathrm{T} y \; . \tag{2}$$

Proof: Let $\hat{\beta}$ be the ordinary least squares (OLS) solution and let $\hat{\varepsilon} = y - X\hat{\beta}$ be the resulting vector of residuals. Then, this vector must be orthogonal to the design matrix,

$$X^\mathrm{T} \hat{\varepsilon} = 0 \; , \tag{3}$$

because if it wasn't, there would be another solution $\tilde{\beta}$ giving another vector $\tilde{\varepsilon}$ with a smaller residual sum of squares. From (3), the OLS formula can be directly derived:

$$\begin{split} X^\mathrm{T} \hat{\varepsilon} &= 0 \\ X^\mathrm{T} \left(y - X\hat{\beta} \right) &= 0 \\ X^\mathrm{T} y - X^\mathrm{T} X \hat{\beta} &= 0 \\ X^\mathrm{T} X \hat{\beta} &= X^\mathrm{T} y \\ \hat{\beta} &= (X^\mathrm{T} X)^{-1} X^\mathrm{T} y \; . \end{split} \tag{4}$$

Sources:

- Stephan, Klaas Enno (2010): "The General Linear Model (GLM)"; in: *Methods and models for fMRI data analysis in neuroeconomics*; URL: http://www.socialbehavior.uzh.ch/teaching/methodsspring10.html.

Metadata: ID: P2 | shortcut: mlr-ols | author: JoramSoch | date: 2019-09-27, 07:18.

1.1.2 Ordinary least squares (2)

Theorem: Given a linear regression model (→ Definition "mlr") with independent observations

$$y = X\beta + \varepsilon, \; \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \; , \tag{1}$$

the parameters minimizing the residual sum of squares (→ Definition "rss") are given by

$$\hat{\beta} = (X^\mathrm{T} X)^{-1} X^\mathrm{T} y \; . \tag{2}$$

Proof: The residual sum of squares (→ Definition "rss") is defined as

$$\mathrm{RSS}(\beta) = \sum_{i=1}^{n} \varepsilon_i = \varepsilon^\mathrm{T} \varepsilon = (y - X\beta)^\mathrm{T} (y - X\beta) \tag{3}$$

which can be developed into

$$\begin{split} \mathrm{RSS}(\beta) &= y^\mathrm{T} y - y^\mathrm{T} X \beta - \beta^\mathrm{T} X^\mathrm{T} y + \beta^\mathrm{T} X^\mathrm{T} X \beta \\ &= y^\mathrm{T} y - 2 \beta^\mathrm{T} X^\mathrm{T} y + \beta^\mathrm{T} X^\mathrm{T} X \beta \; . \end{split} \tag{4}$$

The derivative of $\mathrm{RSS}(\beta)$ with respect to β is

$$\frac{\mathrm{dRSS}(\beta)}{\mathrm{d}\beta} = -2 X^\mathrm{T} y + 2 X^\mathrm{T} X \beta \tag{5}$$

and setting this deriative to zero, we obtain:

$$\begin{split} \frac{\mathrm{dRSS}(\hat{\beta})}{\mathrm{d}\beta} &= 0 \\ 0 &= -2 X^\mathrm{T} y + 2 X^\mathrm{T} X \hat{\beta} \\ X^\mathrm{T} X \hat{\beta} &= X^\mathrm{T} y \\ \hat{\beta} &= (X^\mathrm{T} X)^{-1} X^\mathrm{T} y \; . \end{split} \tag{6}$$

Since the quadratic form $y^\mathrm{T} y$ in (4) is positive, $\hat{\beta}$ minimizes $\mathrm{RSS}(\beta)$.

Sources:

- Wikipedia (2020): "Proofs involving ordinary least squares"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-03; URL: https://en.wikipedia.org/wiki/Proofs_involving_ordinary_least_squares#Least_squares_estimator_for_%CE%B2.

Metadata: ID: P40 | shortcut: mlr-ols2 | author: JoramSoch | date: 2020-02-03, 18:43.

1.2 Bayesian linear regression

1.2.1 Conjugate prior distribution

Theorem: Let

$$y = X\beta + \varepsilon, \; \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \tag{1}$$

be a linear regression model (→ Definition "mlr") with measured $n \times 1$ data vector y, known $n \times p$ design matrix X, known $n \times n$ covariance structure V and unknown $p \times 1$ regression coefficients β and noise variance σ^2.
Then, the conjugate prior (→ Definition "prior-conj") for this model is a normal-gamma distribution (→ Definition II/4.2.1)

$$p(\beta, \tau) = \mathcal{N}(\beta; \mu_0, (\tau \Lambda_0)^{-1}) \cdot \mathrm{Gam}(\tau; a_0, b_0) \tag{2}$$

where $\tau = 1/\sigma^2$ is the inverse noise variance or noise precision.

Proof: By definition, a conjugate prior (→ Definition "prior-conj") is a prior distribution (→ Definition I/2.2.3) that, when combined with the likelihood function (→ Definition I/2.2.2), leads to a posterior distribution (→ Definition I/2.2.6) that belongs to the same family of probability distributions (→ Definition "pd"). This is fulfilled when the prior density and the likelihood function are proportional to the model parameters in the same way, i.e. the model parameters appear in the same functional form in both.
Equation (1) implies the following likelihood function (→ Definition I/2.2.2)

$$p(y|\beta,\sigma^2) = \mathcal{N}(y; X\beta, \sigma^2 V) = \sqrt{\frac{1}{(2\pi)^n |\sigma^2 V|}} \exp\left[-\frac{1}{2\sigma^2}(y-X\beta)^\mathrm{T} V^{-1} (y-X\beta)\right] \tag{3}$$

which, for mathematical convenience, can also be parametrized as

$$p(y|\beta,\tau) = \mathcal{N}(y; X\beta, (\tau P)^{-1}) = \sqrt{\frac{|\tau P|}{(2\pi)^n}} \exp\left[-\frac{\tau}{2}(y-X\beta)^\mathrm{T} P (y-X\beta)\right] \tag{4}$$

using the noise precision $\tau = 1/\sigma^2$ and the $n \times n$ precision matrix $P = V^{-1}$.

Seperating constant and variable terms, we have:

$$p(y|\beta,\tau) = \sqrt{\frac{|P|}{(2\pi)^n}} \cdot \tau^{n/2} \cdot \exp\left[-\frac{\tau}{2}(y-X\beta)^\mathrm{T} P (y-X\beta)\right] \; . \tag{5}$$

Expanding the product in the exponent, we have:

$$p(y|\beta,\tau) = \sqrt{\frac{|P|}{(2\pi)^n}} \cdot \tau^{n/2} \cdot \exp\left[-\frac{\tau}{2}\left(y^\mathrm{T} P y - y^\mathrm{T} P X \beta - \beta^\mathrm{T} X^\mathrm{T} P y + \beta^\mathrm{T} X^\mathrm{T} P X \beta\right)\right] \; . \tag{6}$$

Completing the square over β, finally gives

$$p(y|\beta,\tau) = \sqrt{\frac{|P|}{(2\pi)^n}} \cdot \tau^{n/2} \cdot \exp\left[-\frac{\tau}{2}\left((\beta - \tilde{X}y)^\mathrm{T} X^\mathrm{T} P X (\beta - \tilde{X}y) - y^\mathrm{T} Q y + y^\mathrm{T} P y\right)\right] \tag{7}$$

where $\tilde{X} = \left(X^\mathrm{T} P X\right)^{-1} X^\mathrm{T} P$ and $Q = \tilde{X}^\mathrm{T}\left(X^\mathrm{T} P X\right)\tilde{X}$.

In other words, the likelihood function (→ Definition I/2.2.2) is proportional to a power of τ times an exponential of τ and an exponential of a squared form of β, weighted by τ:

$$p(y|\beta,\tau) \propto \tau^{n/2} \cdot \exp\left[-\frac{\tau}{2}\left(y^\mathrm{T} P y - y^\mathrm{T} Q y\right)\right] \cdot \exp\left[-\frac{\tau}{2}(\beta - \tilde{X}y)^\mathrm{T} X^\mathrm{T} P X (\beta - \tilde{X}y)\right] \; . \tag{8}$$

The same is true for a normal gamma distribution over β and τ

$$p(\beta,\tau) = \mathcal{N}(\beta; \mu_0, (\tau\Lambda_0)^{-1}) \cdot \mathrm{Gam}(\tau; a_0, b_0) \tag{9}$$

the probability density function of which (→ Proof II/4.2.2)

$$p(\beta,\tau) = \sqrt{\frac{|\tau\Lambda_0|}{(2\pi)^p}} \exp\left[-\frac{\tau}{2}(\beta-\mu_0)^\mathrm{T}\Lambda_0(\beta-\mu_0)\right] \cdot \frac{{b_0}^{a_0}}{\Gamma(a_0)} \, \tau^{a_0-1} \exp[-b_0\tau] \tag{10}$$

exhibits the same proportionality

$$p(\beta, \tau) \propto \tau^{a_0+p/2-1} \cdot \exp[-\tau b_0] \cdot \exp\left[-\frac{\tau}{2}(\beta - \mu_0)^\mathrm{T} \Lambda_0 (\beta - \mu_0)\right] \tag{11}$$

and is therefore conjugate relative to the likelihood.

Sources:

- Bishop CM (2006): "Bayesian linear regression"; in: *Pattern Recognition for Machine Learning*, pp. 152-161, ex. 3.12, eq. 3.112; URL: https://www.springer.com/gp/book/9780387310732.

Metadata: ID: P9 | shortcut: blr-prior | author: JoramSoch | date: 2020-01-03, 15:26.

1.2.2 Posterior distribution

Theorem: Let

$$y = X\beta + \varepsilon, \; \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \tag{1}$$

be a linear regression model (→ Definition "mlr") with measured $n \times 1$ data vector y, known $n \times p$ design matrix X, known $n \times n$ covariance structure V and unknown $p \times 1$ regression coefficients β and noise variance σ^2. Moreover, assume a normal-gamma prior distribution (→ Proof III/1.2.1) over the model parameters β and $\tau = 1/\sigma^2$:

$$p(\beta, \tau) = \mathcal{N}(\beta; \mu_0, (\tau \Lambda_0)^{-1}) \cdot \mathrm{Gam}(\tau; a_0, b_0) \; . \tag{2}$$

Then, the posterior distribution (→ Definition I/2.2.6) is also a normal-gamma distribution (→ Definition II/4.2.1)

$$p(\beta, \tau|y) = \mathcal{N}(\beta; \mu_n, (\tau \Lambda_n)^{-1}) \cdot \mathrm{Gam}(\tau; a_n, b_n) \tag{3}$$

and the posterior hyperparameters (→ Definition "post-hyp") are given by

$$\begin{aligned}
\mu_n &= \Lambda_n^{-1}(X^\mathrm{T} P y + \Lambda_0 \mu_0) \\
\Lambda_n &= X^\mathrm{T} P X + \Lambda_0 \\
a_n &= a_0 + \frac{n}{2} \\
b_n &= b_0 + \frac{1}{2}(y^\mathrm{T} P y + \mu_0^\mathrm{T} \Lambda_0 \mu_0 - \mu_n^\mathrm{T} \Lambda_n \mu_n) \; .
\end{aligned} \tag{4}$$

Proof: According to Bayes' theorem (→ Proof I/2.1.1), the posterior distribution (→ Definition I/2.2.6) is given by

$$p(\beta, \tau|y) = \frac{p(y|\beta, \tau) \, p(\beta, \tau)}{p(y)} \; . \tag{5}$$

Since $p(y)$ is just a normalization factor, the posterior is proportional (→ Proof "post-jl") to the numerator:

$$p(\beta, \tau|y) \propto p(y|\beta, \tau) \, p(\beta, \tau) = p(y, \beta, \tau) \; . \tag{6}$$

Equation (1) implies the following likelihood function (→ Definition I/2.2.2)

$$p(y|\beta, \sigma^2) = \mathcal{N}(y; X\beta, \sigma^2 V) = \sqrt{\frac{1}{(2\pi)^n |\sigma^2 V|}} \exp\left[-\frac{1}{2\sigma^2}(y - X\beta)^\mathrm{T} V^{-1} (y - X\beta)\right] \tag{7}$$

which, for mathematical convenience, can also be parametrized as

$$p(y|\beta, \tau) = \mathcal{N}(y; X\beta, (\tau P)^{-1}) = \sqrt{\frac{|\tau P|}{(2\pi)^n}} \exp\left[-\frac{\tau}{2}(y - X\beta)^\mathrm{T} P (y - X\beta)\right] \tag{8}$$

using the noise precision $\tau = 1/\sigma^2$ and the $n \times n$ precision matrix $P = V^{-1}$.

Combining the likelihood function (8) with the prior distribution (2), the ssssssssssssssssssssssssssssssssssssssss of the model is given by

$$\begin{aligned} p(y, \beta, \tau) &= p(y|\beta, \tau)\, p(\beta, \tau) \\ &= \sqrt{\frac{|\tau P|}{(2\pi)^n}} \exp\left[-\frac{\tau}{2}(y - X\beta)^\mathrm{T} P (y - X\beta)\right] \cdot \\ &\quad \sqrt{\frac{|\tau \Lambda_0|}{(2\pi)^p}} \exp\left[-\frac{\tau}{2}(\beta - \mu_0)^\mathrm{T} \Lambda_0 (\beta - \mu_0)\right] \cdot \\ &\quad \frac{{b_0}^{a_0}}{\Gamma(a_0)}\, \tau^{a_0 - 1} \exp[-b_0 \tau] \; . \end{aligned} \tag{9}$$

Collecting identical variables gives:

$$\begin{aligned} p(y, \beta, \tau) = &\sqrt{\frac{\tau^{n+p}}{(2\pi)^{n+p}} |P| |\Lambda_0|}\, \frac{{b_0}^{a_0}}{\Gamma(a_0)}\, \tau^{a_0 - 1} \exp[-b_0 \tau] \cdot \\ &\exp\left[-\frac{\tau}{2}\left((y - X\beta)^\mathrm{T} P (y - X\beta) + (\beta - \mu_0)^\mathrm{T} \Lambda_0 (\beta - \mu_0)\right)\right] \; . \end{aligned} \tag{10}$$

Expanding the products in the exponent gives:

$$\begin{aligned} p(y, \beta, \tau) = &\sqrt{\frac{\tau^{n+p}}{(2\pi)^{n+p}} |P| |\Lambda_0|}\, \frac{{b_0}^{a_0}}{\Gamma(a_0)}\, \tau^{a_0 - 1} \exp[-b_0 \tau] \cdot \\ &\exp\left[-\frac{\tau}{2}\left(y^\mathrm{T} P y - y^\mathrm{T} P X \beta - \beta^\mathrm{T} X^\mathrm{T} P y + \beta^\mathrm{T} X^\mathrm{T} P X \beta + \right.\right. \\ &\qquad \left.\left. \beta^\mathrm{T} \Lambda_0 \beta - \beta^\mathrm{T} \Lambda_0 \mu_0 - \mu_0^\mathrm{T} \Lambda_0 \beta + \mu_0^\mathrm{T} \Lambda_0 \mu_0\right)\right] \; . \end{aligned} \tag{11}$$

Completing the square over β, we finally have

$$\begin{aligned} p(y, \beta, \tau) = &\sqrt{\frac{\tau^{n+p}}{(2\pi)^{n+p}} |P| |\Lambda_0|}\, \frac{{b_0}^{a_0}}{\Gamma(a_0)}\, \tau^{a_0 - 1} \exp[-b_0 \tau] \cdot \\ &\exp\left[-\frac{\tau}{2}\left((\beta - \mu_n)^\mathrm{T} \Lambda_n (\beta - \mu_n) + (y^\mathrm{T} P y + \mu_0^\mathrm{T} \Lambda_0 \mu_0 - \mu_n^\mathrm{T} \Lambda_n \mu_n)\right)\right] \end{aligned} \tag{12}$$

with the posterior hyperparameters (→ Definition “post-hyp”)

$$\begin{aligned} \mu_n &= \Lambda_n^{-1}(X^\mathrm{T} P y + \Lambda_0 \mu_0) \\ \Lambda_n &= X^\mathrm{T} P X + \Lambda_0 \; . \end{aligned} \tag{13}$$

Ergo, the joint likelihood is proportional to

$$p(y, \beta, \tau) \propto \tau^{p/2} \cdot \exp\left[-\frac{\tau}{2} (\beta - \mu_n)^\mathrm{T} \Lambda_n (\beta - \mu_n) \right] \cdot \tau^{a_n - 1} \cdot \exp\left[-b_n \tau \right] \tag{14}$$

with the posterior hyperparameters (→ Definition "post-hyp")

$$\begin{aligned} a_n &= a_0 + \frac{n}{2} \\ b_n &= b_0 + \frac{1}{2} (y^\mathrm{T} P y + \mu_0^\mathrm{T} \Lambda_0 \mu_0 - \mu_n^\mathrm{T} \Lambda_n \mu_n) \; . \end{aligned} \tag{15}$$

From the term in (14), we can isolate the posterior distribution over β given τ:

$$p(\beta|\tau, y) = \mathcal{N}(\beta; \mu_n, (\tau \Lambda_n)^{-1}) \; . \tag{16}$$

From the remaining term, we can isolate the posterior distribution over τ:

$$p(\tau|y) = \mathrm{Gam}(\tau; a_n, b_n) \; . \tag{17}$$

Together, (16) and (17) constitute the joint (→ Definition "prob-joint") posterior distribution (→ Definition I/2.2.6) of β and τ.

Sources:

- Bishop CM (2006): "Bayesian linear regression"; in: *Pattern Recognition for Machine Learning*, pp. 152-161, ex. 3.12, eq. 3.113; URL: https://www.springer.com/gp/book/9780387310732.

Metadata: ID: P10 | shortcut: blr-post | author: JoramSoch | date: 2020-01-03, 17:53.

1.2.3 Log model evidence

Theorem: Let

$$m: \; y = X\beta + \varepsilon, \; \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \tag{1}$$

be a linear regression model (→ Definition "mlr") with measured $n \times 1$ data vector y, known $n \times p$ design matrix X, known $n \times n$ covariance structure V and unknown $p \times 1$ regression coefficients β and noise variance σ^2. Moreover, assume a normal-gamma prior distribution (→ Proof III/1.2.1) over the model parameters β and $\tau = 1/\sigma^2$:

$$p(\beta, \tau) = \mathcal{N}(\beta; \mu_0, (\tau \Lambda_0)^{-1}) \cdot \mathrm{Gam}(\tau; a_0, b_0) \; . \tag{2}$$

Then, the log model evidence (→ Definition IV/3.1.1) for this model is

$$\begin{aligned} \log p(y|m) = \frac{1}{2} \log |P| - \frac{n}{2} \log(2\pi) + \frac{1}{2} \log |\Lambda_0| - \frac{1}{2} \log |\Lambda_n| + \\ \log \Gamma(a_n) - \log \Gamma(a_0) + a_0 \log b_0 - a_n \log b_n \end{aligned} \tag{3}$$

where the posterior hyperparameters (→ Definition "post-hyp") are given by

$$
\begin{aligned}
\mu_n &= \Lambda_n^{-1}(X^\mathrm{T} P y + \Lambda_0 \mu_0) \\
\Lambda_n &= X^\mathrm{T} P X + \Lambda_0 \\
a_n &= a_0 + \frac{n}{2} \\
b_n &= b_0 + \frac{1}{2}(y^\mathrm{T} P y + \mu_0^\mathrm{T} \Lambda_0 \mu_0 - \mu_n^\mathrm{T} \Lambda_n \mu_n) \; .
\end{aligned} \tag{4}
$$

Proof: According to the law of marginal probability (→ Definition "prob-marg"), the model evidence (→ Definition I/2.2.7) for this model is:

$$
p(y|m) = \iint p(y|\beta,\tau) \, p(\beta,\tau) \, \mathrm{d}\beta \, \mathrm{d}\tau \; . \tag{5}
$$

According to the law of conditional probability (→ Definition "prob-cond"), the integrand is equivalent to the joint likelihood (→ Definition I/2.2.5):

$$
p(y|m) = \iint p(y,\beta,\tau) \, \mathrm{d}\beta \, \mathrm{d}\tau \; . \tag{6}
$$

Equation (1) implies the following likelihood function (→ Definition I/2.2.2)

$$
p(y|\beta,\sigma^2) = \mathcal{N}(y; X\beta, \sigma^2 V) = \sqrt{\frac{1}{(2\pi)^n |\sigma^2 V|}} \, \exp\left[-\frac{1}{2\sigma^2}(y-X\beta)^\mathrm{T} V^{-1} (y-X\beta) \right] \tag{7}
$$

which, for mathematical convenience, can also be parametrized as

$$
p(y|\beta,\tau) = \mathcal{N}(y; X\beta, (\tau P)^{-1}) = \sqrt{\frac{|\tau P|}{(2\pi)^n}} \, \exp\left[-\frac{\tau}{2}(y-X\beta)^\mathrm{T} P (y-X\beta) \right] \tag{8}
$$

using the noise precision $\tau = 1/\sigma^2$ and the $n \times n$ precision matrix $P = V^{-1}$.

When deriving the posterior distribution (→ Proof III/1.2.2) $p(\beta,\tau|y)$, the joint likelihood $p(y,\beta,\tau)$ is obtained as

$$
\begin{split}
p(y,\beta,\tau) = \; & \sqrt{\frac{\tau^n |P|}{(2\pi)^n}} \sqrt{\frac{\tau^p |\Lambda_0|}{(2\pi)^p}} \, \frac{{b_0}^{a_0}}{\Gamma(a_0)} \, \tau^{a_0-1} \exp[-b_0 \tau] \cdot \\
& \exp\left[-\frac{\tau}{2} \left((\beta-\mu_n)^T \Lambda_n (\beta-\mu_n) + (y^T P y + \mu_0^T \Lambda_0 \mu_0 - \mu_n^T \Lambda_n \mu_n) \right) \right] \; .
\end{split} \tag{9}
$$

Using the probability density function of the multivariate normal distribution (→ Proof II/4.1.2), we can rewrite this as

$$
\begin{split}
p(y,\beta,\tau) = \; & \sqrt{\frac{\tau^n |P|}{(2\pi)^n}} \sqrt{\frac{\tau^p |\Lambda_0|}{(2\pi)^p}} \sqrt{\frac{(2\pi)^p}{\tau^p |\Lambda_n|}} \, \frac{{b_0}^{a_0}}{\Gamma(a_0)} \, \tau^{a_0-1} \exp[-b_0 \tau] \cdot \\
& \mathcal{N}(\beta; \mu_n, (\tau\Lambda_n)^{-1}) \, \exp\left[-\frac{\tau}{2} (y^T P y + \mu_0^T \Lambda_0 \mu_0 - \mu_n^T \Lambda_n \mu_n) \right] \; .
\end{split} \tag{10}
$$

Now, β can be integrated out easily:

$$\begin{aligned}\int p(y,\beta,\tau)\,\mathrm{d}\beta = \sqrt{\frac{\tau^n|P|}{(2\pi)^n}}\sqrt{\frac{|\Lambda_0|}{|\Lambda_n|}}\,\frac{{b_0}^{a_0}}{\Gamma(a_0)}\,\tau^{a_0-1}\exp[-b_0\tau]\cdot \\ \exp\left[-\frac{\tau}{2}(y^TPy+\mu_0^T\Lambda_0\mu_0-\mu_n^T\Lambda_n\mu_n)\right]\;.\end{aligned} \tag{11}$$

Using the probability density function of the gamma distribution (→ Proof II/3.3.2), we can rewrite this as

$$\int p(y,\beta,\tau)\,\mathrm{d}\beta = \sqrt{\frac{|P|}{(2\pi)^n}}\sqrt{\frac{|\Lambda_0|}{|\Lambda_n|}}\,\frac{{b_0}^{a_0}}{\Gamma(a_0)}\,\frac{\Gamma(a_n)}{{b_n}^{a_n}}\,\mathrm{Gam}(\tau;a_n,b_n)\;. \tag{12}$$

Finally, τ can also be integrated out:

$$\iint p(y,\beta,\tau)\,\mathrm{d}\beta\,\mathrm{d}\tau = \sqrt{\frac{|P|}{(2\pi)^n}}\sqrt{\frac{|\Lambda_0|}{|\Lambda_n|}}\,\frac{\Gamma(a_n)}{\Gamma(a_0)}\,\frac{{b_0}^{a_0}}{{b_n}^{a_n}} = p(y|m)\;. \tag{13}$$

Thus, the log model evidence (→ Definition IV/3.1.1) of this model is given by

$$\begin{aligned}\log p(y|m) = \frac{1}{2}\log|P| - \frac{n}{2}\log(2\pi) + \frac{1}{2}\log|\Lambda_0| - \frac{1}{2}\log|\Lambda_n| + \\ \log\Gamma(a_n) - \log\Gamma(a_0) + a_0\log b_0 - a_n\log b_n\;.\end{aligned} \tag{14}$$

Sources:
- Bishop CM (2006): "Bayesian linear regression"; in: *Pattern Recognition for Machine Learning*, pp. 152-161, ex. 3.23, eq. 3.118; URL: https://www.springer.com/gp/book/9780387310732.

Metadata: ID: P11 | shortcut: blr-lme | author: JoramSoch | date: 2020-01-03, 22:05.

1.3 General linear model

1.3.1 Maximum likelihood estimation

Theorem: Given a general linear model (→ Definition "glm") with matrix-normally distributed (→ Definition II/5.1.1) errors

$$Y = XB + E,\; E \sim \mathcal{MN}(0,V,\Sigma)\;, \tag{1}$$

maximum likelihood estimates (→ Definition "mle") for the unknown parameters B and Σ are given by

$$\begin{aligned}\hat{B} &= (X^\mathrm{T}V^{-1}X)^{-1}X^\mathrm{T}V^{-1}Y \\ \hat{\Sigma} &= \frac{1}{n}(Y-X\hat{B})^\mathrm{T}V^{-1}(Y-X\hat{B})\;.\end{aligned} \tag{2}$$

Proof: In (1), Y is an $n\times v$ matrix of measurements (n observations, v dependent variables), X is an $n\times p$ design matrix (n observations, p independent variables) and V is an $n\times n$ covariance matrix across observations. This multivariate GLM implies the following likelihood function (→ Definition I/2.2.2)

$$
\begin{split}
p(Y|B,\Sigma) &= \mathcal{MN}(Y; XB, V, \Sigma) \\
&= \sqrt{\frac{1}{(2\pi)^{nv} |\Sigma|^n |V|^v}} \cdot \exp\left[-\frac{1}{2} \mathrm{tr}\left(\Sigma^{-1} (Y-XB)^\mathrm{T} V^{-1} (Y-XB) \right) \right]
\end{split} \tag{3}
$$

and the log-likelihood function (→ Definition "llf")

$$
\begin{split}
\mathrm{LL}(B,\Sigma) = &\log p(Y|B,\Sigma) \\
= &-\frac{nv}{2} \log(2\pi) - \frac{n}{2} \log|\Sigma| - \frac{v}{2} \log|V| \\
&- \frac{1}{2} \mathrm{tr}\left[\Sigma^{-1} (Y-XB)^\mathrm{T} V^{-1} (Y-XB) \right] \; .
\end{split} \tag{4}
$$

Substituting V^{-1} by the precision matrix P to ease notation, we have:

$$
\begin{split}
\mathrm{LL}(B,\Sigma) = &-\frac{nv}{2} \log(2\pi) - \frac{n}{2} \log(|\Sigma|) - \frac{v}{2} \log(|V|) \\
&- \frac{1}{2} \mathrm{tr}\left[\Sigma^{-1} \left(Y^\mathrm{T} P Y - Y^\mathrm{T} P X B - B^\mathrm{T} X^\mathrm{T} P Y + B^\mathrm{T} X^\mathrm{T} P X B \right) \right] \; .
\end{split} \tag{5}
$$

The derivative of the log-likelihood function (5) with respect to B is

$$
\begin{split}
\frac{\mathrm{dLL}(B,\Sigma)}{\mathrm{d}B} &= \frac{\mathrm{d}}{\mathrm{d}B} \left(-\frac{1}{2} \mathrm{tr}\left[\Sigma^{-1} \left(Y^\mathrm{T} P Y - Y^\mathrm{T} P X B - B^\mathrm{T} X^\mathrm{T} P Y + B^\mathrm{T} X^\mathrm{T} P X B \right) \right] \right) \\
&= \frac{\mathrm{d}}{\mathrm{d}B} \left(-\frac{1}{2} \mathrm{tr}\left[-2 \Sigma^{-1} Y^\mathrm{T} P X B \right] \right) + \frac{\mathrm{d}}{\mathrm{d}B} \left(-\frac{1}{2} \mathrm{tr}\left[\Sigma^{-1} B^\mathrm{T} X^\mathrm{T} P X B \right] \right) \\
&= -\frac{1}{2} \left(-2 X^\mathrm{T} P Y \Sigma^{-1} \right) - \frac{1}{2} \left(X^\mathrm{T} P X B \Sigma^{-1} + (X^\mathrm{T} P X)^\mathrm{T} B (\Sigma^{-1})^\mathrm{T} \right) \\
&= X^\mathrm{T} P Y \Sigma^{-1} - X^\mathrm{T} P X B \Sigma^{-1}
\end{split} \tag{6}
$$

and setting this derivative to zero gives the MLE for B:

$$
\begin{split}
\frac{\mathrm{dLL}(\hat{B},\Sigma)}{\mathrm{d}B} &= 0 \\
0 &= X^\mathrm{T} P Y \Sigma^{-1} - X^\mathrm{T} P X \hat{B} \Sigma^{-1} \\
0 &= X^\mathrm{T} P Y - X^\mathrm{T} P X \hat{B} \\
X^\mathrm{T} P X \hat{B} &= X^\mathrm{T} P Y \\
\hat{B} &= \left(X^\mathrm{T} P X \right)^{-1} X^\mathrm{T} P Y
\end{split} \tag{7}
$$

The derivative of the log-likelihood function (4) at $\hat{B}$ with respect to Σ is

$$
\begin{aligned}
\frac{\mathrm{dLL}(\hat{B},\Sigma)}{\mathrm{d}\Sigma} &= \frac{\mathrm{d}}{\mathrm{d}\Sigma}\left(-\frac{n}{2}\log|\Sigma| - \frac{1}{2}\,\mathrm{tr}\left[\Sigma^{-1}(Y-X\hat{B})^\mathrm{T}V^{-1}(Y-X\hat{B})\right]\right) \\
&= -\frac{n}{2}\left(\Sigma^{-1}\right)^\mathrm{T} + \frac{1}{2}\left(\Sigma^{-1}(Y-X\hat{B})^\mathrm{T}V^{-1}(Y-X\hat{B})\,\Sigma^{-1}\right)^\mathrm{T} \\
&= -\frac{n}{2}\,\Sigma^{-1} + \frac{1}{2}\,\Sigma^{-1}(Y-X\hat{B})^\mathrm{T}V^{-1}(Y-X\hat{B})\,\Sigma^{-1}
\end{aligned}
\tag{8}
$$

and setting this derivative to zero gives the MLE for Σ:

$$
\begin{aligned}
\frac{\mathrm{dLL}(\hat{B},\hat{\Sigma})}{\mathrm{d}\Sigma} &= 0 \\
0 &= -\frac{n}{2}\,\hat{\Sigma}^{-1} + \frac{1}{2}\,\hat{\Sigma}^{-1}(Y-X\hat{B})^\mathrm{T}V^{-1}(Y-X\hat{B})\,\hat{\Sigma}^{-1} \\
\frac{n}{2}\,\hat{\Sigma}^{-1} &= \frac{1}{2}\,\hat{\Sigma}^{-1}(Y-X\hat{B})^\mathrm{T}V^{-1}(Y-X\hat{B})\,\hat{\Sigma}^{-1} \\
\hat{\Sigma}^{-1} &= \frac{1}{n}\,\hat{\Sigma}^{-1}(Y-X\hat{B})^\mathrm{T}V^{-1}(Y-X\hat{B})\,\hat{\Sigma}^{-1} \\
I_v &= \frac{1}{n}\,(Y-X\hat{B})^\mathrm{T}V^{-1}(Y-X\hat{B})\,\hat{\Sigma}^{-1} \\
\hat{\Sigma} &= \frac{1}{n}\,(Y-X\hat{B})^\mathrm{T}V^{-1}(Y-X\hat{B})
\end{aligned}
\tag{9}
$$

Together, (7) and (9) constitute the MLE for the GLM.

Sources:
- original work

Metadata: ID: P7 | shortcut: glm-mle | author: JoramSoch | date: 2019-12-06, 10:40.

2 Poisson data

2.1 Poisson-distributed data

2.1.1 Maximum likelihood estimation

Theorem: Let $y = \{y_1, \ldots, y_n\}$ be a set of observed counts independent and identically distributed according to a Poisson distribution (→ Definition "poiss") with rate λ:

$$y_i \sim \mathrm{Poiss}(\lambda), \quad i = 1, \ldots, n \; . \tag{1}$$

Then, the maximum likelihood estimate (→ Definition "mle") for the rate parameter λ is given by

$$\hat{\lambda} = \bar{y} \tag{2}$$

where $\bar{y}$ is the sample mean (→ Proof "mean-sample")

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \; . \tag{3}$$

Proof: The likelihood function (→ Definition I/2.2.2) for each observation is given by the probability mass function of the Poisson distribution (→ Proof "poiss-pdf")

$$p(y_i|\lambda) = \mathrm{Poiss}(y_i; \lambda) = \frac{\lambda^{y_i} \cdot \exp(-\lambda)}{y_i!} \tag{4}$$

and because observations are independent (→ Definition "ind"), the likelihood function for all observations is the product of the individual ones:

$$p(y|\lambda) = \prod_{i=1}^{n} p(y_i|\lambda) = \prod_{i=1}^{n} \frac{\lambda^{y_i} \cdot \exp(-\lambda)}{y_i!} \; . \tag{5}$$

Thus, the log-likelihood function (→ Definition "llf") is

$$\mathrm{LL}(\lambda) = \log p(y|\lambda) = \log \left[\prod_{i=1}^{n} \frac{\lambda^{y_i} \cdot \exp(-\lambda)}{y_i!} \right] \tag{6}$$

which can be developed into

$$\begin{split}
\mathrm{LL}(\lambda) &= \sum_{i=1}^{n} \log \left[\frac{\lambda^{y_i} \cdot \exp(-\lambda)}{y_i!} \right] \\
&= \sum_{i=1}^{n} \left[y_i \cdot \log(\lambda) - \lambda - \log(y_i!) \right] \\
&= - \sum_{i=1}^{n} \lambda + \sum_{i=1}^{n} y_i \cdot \log(\lambda) - \sum_{i=1}^{n} \log(y_i!) \\
&= -n\lambda + \log(\lambda) \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \log(y_i!)
\end{split} \tag{7}$$

The derivatives of the log-likelihood with respect to λ are

$$
\begin{aligned}
\frac{\mathrm{dLL}(\lambda)}{\mathrm{d}\lambda} &= \frac{1}{\lambda}\sum_{i=1}^{n} y_i - n \\
\frac{\mathrm{d}^2\mathrm{LL}(\lambda)}{\mathrm{d}\lambda^2} &= -\frac{1}{\lambda^2}\sum_{i=1}^{n} y_i \; .
\end{aligned} \tag{8}
$$

Setting the first derivative to zero, we obtain:

$$
\begin{aligned}
\frac{\mathrm{dLL}(\hat{\lambda})}{\mathrm{d}\lambda} &= 0 \\
0 &= \frac{1}{\hat{\lambda}}\sum_{i=1}^{n} y_i - n \\
\hat{\lambda} &= \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y} \; .
\end{aligned} \tag{9}
$$

Plugging this value into the second deriative, we confirm:

$$
\begin{aligned}
\frac{\mathrm{d}^2\mathrm{LL}(\hat{\lambda})}{\mathrm{d}\lambda^2} &= -\frac{1}{\bar{y}^2}\sum_{i=1}^{n} y_i \\
&= -\frac{n \cdot \bar{y}}{\bar{y}^2} \\
&= -\frac{n}{\bar{y}} < 0 \; .
\end{aligned} \tag{10}
$$

This demonstrates that the estimate $\hat{\lambda} = \bar{y}$ maximizes the likelihood $p(y|\lambda)$.

Sources:
- original work

Metadata: ID: P27 | shortcut: poiss-mle | author: JoramSoch | date: 2020-01-20, 21:53.

2.2 Poisson distribution with exposure values

2.2.1 Conjugate prior distribution

Theorem: Let $y = \{y_1, \ldots, y_n\}$ be a series of observed counts which are independently distributed according to a Poisson distribution (→ Definition "poiss") with common rate λ and concurrent exposures $\{x_1, \ldots, x_n\}$:

$$
y_i \sim \mathrm{Poiss}(\lambda x_i), \quad i = 1, \ldots, n \; . \tag{1}
$$

Then, the conjugate prior (→ Definition "prior-conj") for the model parameter λ is a gamma distribution (→ Definition II/3.3.1):

$$
p(\lambda) = \mathrm{Gam}(\lambda; a_0, b_0) \; . \tag{2}
$$

Proof: With the probability mass function of the Poisson distribution (→ Proof "poiss-pmf"), the likelihood function (→ Definition I/2.2.2) for each observation implied by (1) is given by

$$p(y_i|\lambda) = \mathrm{Poiss}(y_i; \lambda x_i) = \frac{(\lambda x_i)^{y_i} \cdot \exp\left[-\lambda x_i\right]}{y_i!} \tag{3}$$

and because observations are independent (→ Definition "ind"), the likelihood function for all observations is the product of the individual ones:

$$p(y|\lambda) = \prod_{i=1}^{n} p(y_i|\lambda) = \prod_{i=1}^{n} \frac{(\lambda x_i)^{y_i} \cdot \exp\left[-\lambda x_i\right]}{y_i!} \; . \tag{4}$$

Resolving the product in the likelihood function, we have

$$\begin{aligned} p(y|\lambda) &= \prod_{i=1}^{n} \frac{{x_i}^{y_i}}{y_i!} \cdot \prod_{i=1}^{n} \lambda^{y_i} \cdot \prod_{i=1}^{n} \exp\left[-\lambda x_i\right] \\ &= \prod_{i=1}^{n} \left(\frac{{x_i}^{y_i}}{y_i!} \right) \cdot \lambda^{\sum_{i=1}^{n} y_i} \cdot \exp\left[-\lambda \sum_{i=1}^{n} x_i \right] \\ &= \prod_{i=1}^{n} \left(\frac{{x_i}^{y_i}}{y_i!} \right) \cdot \lambda^{n\bar{y}} \cdot \exp\left[-n\bar{x}\lambda\right] \end{aligned} \tag{5}$$

where $\bar{y}$ and $\bar{x}$ are the means (→ Proof "mean-sample") of y and x respectively:

$$\begin{aligned} \bar{y} &= \frac{1}{n} \sum_{i=1}^{n} y_i \\ \bar{x} &= \frac{1}{n} \sum_{i=1}^{n} x_i \; . \end{aligned} \tag{6}$$

In other words, the likelihood function is proportional to a power of λ times an exponential of λ:

$$p(y|\lambda) \propto \lambda^{n\bar{y}} \cdot \exp\left[-n\bar{x}\lambda\right] \; . \tag{7}$$

The same is true for a gamma distribution over λ

$$p(\lambda) = \mathrm{Gam}(\lambda; a_0, b_0) \tag{8}$$

the probability density function of which (→ Proof II/3.3.2)

$$p(\lambda) = \frac{{b_0}^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0\lambda] \tag{9}$$

exhibits the same proportionality

$$p(\lambda) \propto \lambda^{a_0-1} \cdot \exp[-b_0\lambda] \tag{10}$$

and is therefore conjugate relative to the likelihood.

Sources:

- Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB (2014): "Other standard single-parameter models"; in: *Bayesian Data Analysis*, 3rd edition, ch. 2.6, p. 45, eq. 2.14ff.; URL: http://www.stat.columbia.edu/~gelman/book/.

Metadata: ID: P41 | shortcut: poissexp-prior | author: JoramSoch | date: 2020-02-04, 14:11.

2.2.2 Posterior distribution

Theorem: Let $y = \{y_1, \ldots, y_n\}$ be a series of observed counts which are independently distributed according to a Poisson distribution (→ Definition "poiss") with common rate λ and concurrent exposures $\{x_1, \ldots, x_n\}$:

$$y_i \sim \mathrm{Poiss}(\lambda x_i), \quad i = 1, \ldots, n \; . \tag{1}$$

Moreover, assume a gamma prior distribution (→ Proof III/2.2.1) over the model parameter λ:

$$p(\lambda) = \mathrm{Gam}(\lambda; a_0, b_0) \; . \tag{2}$$

Then, the posterior distribution (→ Definition I/2.2.6) is also a gamma distribution (→ Definition II/3.3.1)

$$p(\lambda|y) = \mathrm{Gam}(\lambda; a_n, b_n) \tag{3}$$

and the posterior hyperparameters (→ Definition "post-hyp") are given by

$$\begin{aligned} a_n &= a_0 + n\bar{y} \\ a_n &= a_0 + n\bar{x} \; . \end{aligned} \tag{4}$$

Proof: With the probability mass function of the Poisson distribution (→ Proof "poiss-pmf"), the likelihood function (→ Definition I/2.2.2) for each observation implied by (1) is given by

$$p(y_i|\lambda) = \mathrm{Poiss}(y_i; \lambda x_i) = \frac{(\lambda x_i)^{y_i} \cdot \exp\left[-\lambda x_i\right]}{y_i!} \tag{5}$$

and because observations are independent (→ Definition "ind"), the likelihood function for all observations is the product of the individual ones:

$$p(y|\lambda) = \prod_{i=1}^{n} p(y_i|\lambda) = \prod_{i=1}^{n} \frac{(\lambda x_i)^{y_i} \cdot \exp\left[-\lambda x_i\right]}{y_i!} \; . \tag{6}$$

Combining the likelihood function (6) with the prior distribution (2), the joint likelihood (→ Definition I/2.2.5) of the model is given by

$$\begin{aligned} p(y, \lambda) &= p(y|\lambda)\, p(\lambda) \\ &= \prod_{i=1}^{n} \frac{(\lambda x_i)^{y_i} \cdot \exp\left[-\lambda x_i\right]}{y_i!} \cdot \frac{{b_0}^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0 \lambda] \; . \end{aligned} \tag{7}$$

Resolving the product in the joint likelihood, we have

$$
\begin{aligned}
p(y,\lambda) &= \prod_{i=1}^n \frac{{x_i}^{y_i}}{y_i!} \prod_{i=1}^n \lambda^{y_i} \prod_{i=1}^n \exp\left[-\lambda x_i\right] \cdot \frac{{b_0}^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0\lambda] \\
&= \prod_{i=1}^n \left(\frac{{x_i}^{y_i}}{y_i!}\right) \lambda^{\sum_{i=1}^n y_i} \exp\left[-\lambda \sum_{i=1}^n x_i\right] \cdot \frac{{b_0}^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0\lambda] \\
&= \prod_{i=1}^n \left(\frac{{x_i}^{y_i}}{y_i!}\right) \lambda^{n\bar{y}} \exp\left[-n\bar{x}\lambda\right] \cdot \frac{{b_0}^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0\lambda] \\
&= \prod_{i=1}^n \left(\frac{{x_i}^{y_i}}{y_i!}\right) \frac{{b_0}^{a_0}}{\Gamma(a_0)} \cdot \lambda^{a_0+n\bar{y}-1} \cdot \exp\left[-(b_0+n\bar{x})\lambda\right]
\end{aligned} \tag{8}
$$

where $\bar{y}$ and $\bar{x}$ are the means (→ Proof "mean-sample") of y and x respectively:

$$
\begin{aligned}
\bar{y} &= \frac{1}{n}\sum_{i=1}^n y_i \\
\bar{x} &= \frac{1}{n}\sum_{i=1}^n x_i \; .
\end{aligned} \tag{9}
$$

Note that the posterior distribution is proportional to the joint likelihood (→ Proof "post-jl"):

$$
p(\lambda|y) \propto p(y,\lambda) \; . \tag{10}
$$

Setting $a_n = a_0 + n\bar{y}$ and $b_n = b_0 + n\bar{x}$, the posterior distribution is therefore proportional to

$$
p(\lambda|y) \propto \lambda^{a_n-1} \cdot \exp\left[-b_n\lambda\right] \tag{11}
$$

which, when normalized to one, results in the probability density function of the gamma distribution (→ Proof II/3.3.2):

$$
p(\lambda|y) = \frac{{b_n}^{a_n}}{\Gamma(a_0)} \lambda^{a_n-1} \exp\left[-b_n\lambda\right] = \mathrm{Gam}(\lambda; a_n, b_n) \; . \tag{12}
$$

Sources:

- Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB (2014): "Other standard single-parameter models"; in: *Bayesian Data Analysis*, 3rd edition, ch. 2.6, p. 45, eq. 2.15; URL: http://www.stat.columbia.edu/~gelman/book/.

Metadata: ID: P42 | shortcut: poissexp-post | author: JoramSoch | date: 2020-02-04, 14:42.

2.2.3 Log model evidence

Theorem: Let $y = \{y_1, \ldots, y_n\}$ be a series of observed counts which are independently distributed according to a Poisson distribution (→ Definition "poiss") with common rate λ and concurrent exposures $\{x_1, \ldots, x_n\}$:

$$
y_i \sim \mathrm{Poiss}(\lambda x_i), \quad i = 1, \ldots, n \; . \tag{1}
$$

Moreover, assume a gamma prior distribution (→ Proof III/2.2.1) over the model parameter λ:

$$p(\lambda) = \mathrm{Gam}(\lambda; a_0, b_0) \; . \tag{2}$$

Then, the log model evidence (→ Definition IV/3.1.1) for this model is

$$\begin{split} \log p(y|m) = &\sum_{i=1}^{n} y_i \log(x_i) - \sum_{i=1}^{n} \log y_i ! + \\ &\log \Gamma(a_n) - \log \Gamma(a_0) + a_0 \log b_0 - a_n \log b_n \; . \end{split} \tag{3}$$

where the posterior hyperparameters (→ Definition “post-hyp”) are given by

$$\begin{split} a_n &= a_0 + n\bar{y} \\ a_n &= a_0 + n\bar{x} \; . \end{split} \tag{4}$$

Proof: With the probability mass function of the Poisson distribution (→ Proof “poiss-pmf”), the likelihood function (→ Definition I/2.2.2) for each observation implied by (1) is given by

$$p(y_i|\lambda) = \mathrm{Poiss}(y_i; \lambda x_i) = \frac{(\lambda x_i)^{y_i} \cdot \exp\left[-\lambda x_i\right]}{y_i !} \tag{5}$$

and because observations are independent (→ Definition “ind”), the likelihood function for all observations is the product of the individual ones:

$$p(y|\lambda) = \prod_{i=1}^{n} p(y_i|\lambda) = \prod_{i=1}^{n} \frac{(\lambda x_i)^{y_i} \cdot \exp\left[-\lambda x_i\right]}{y_i !} \; . \tag{6}$$

Combining the likelihood function (6) with the prior distribution (2), the joint likelihood (→ Definition I/2.2.5) of the model is given by

$$\begin{split} p(y, \lambda) &= p(y|\lambda) \, p(\lambda) \\ &= \prod_{i=1}^{n} \frac{(\lambda x_i)^{y_i} \cdot \exp\left[-\lambda x_i\right]}{y_i !} \cdot \frac{{b_0}^{a_0}}{\Gamma(a_0)} \lambda^{a_0 - 1} \exp[-b_0 \lambda] \; . \end{split} \tag{7}$$

Resolving the product in the joint likelihood, we have

$$\begin{split} p(y, \lambda) &= \prod_{i=1}^{n} \frac{{x_i}^{y_i}}{y_i !} \prod_{i=1}^{n} \lambda^{y_i} \prod_{i=1}^{n} \exp\left[-\lambda x_i\right] \cdot \frac{{b_0}^{a_0}}{\Gamma(a_0)} \lambda^{a_0 - 1} \exp[-b_0 \lambda] \\ &= \prod_{i=1}^{n} \left(\frac{{x_i}^{y_i}}{y_i !} \right) \lambda^{\sum_{i=1}^{n} y_i} \exp\left[-\lambda \sum_{i=1}^{n} x_i\right] \cdot \frac{{b_0}^{a_0}}{\Gamma(a_0)} \lambda^{a_0 - 1} \exp[-b_0 \lambda] \\ &= \prod_{i=1}^{n} \left(\frac{{x_i}^{y_i}}{y_i !} \right) \lambda^{n\bar{y}} \exp\left[-n\bar{x}\lambda\right] \cdot \frac{{b_0}^{a_0}}{\Gamma(a_0)} \lambda^{a_0 - 1} \exp[-b_0 \lambda] \\ &= \prod_{i=1}^{n} \left(\frac{{x_i}^{y_i}}{y_i !} \right) \frac{{b_0}^{a_0}}{\Gamma(a_0)} \cdot \lambda^{a_0 + n\bar{y} - 1} \cdot \exp\left[-(b_0 + n\bar{x})\lambda\right] \end{split} \tag{8}$$

where $\bar{y}$ and $\bar{x}$ are the means (→ Proof "mean-sample") of y and x respectively:

$$
\begin{aligned}
\bar{y} &= \frac{1}{n} \sum_{i=1}^{n} y_i \\
\bar{x} &= \frac{1}{n} \sum_{i=1}^{n} x_i \; .
\end{aligned} \tag{9}
$$

Note that the model evidence is the marginal density of the joint likelihood (→ Definition I/2.2.7):

$$
p(y) = \int p(y, \lambda) \, \mathrm{d}\lambda \; . \tag{10}
$$

Setting $a_n = a_0 + n\bar{y}$ and $b_n = b_0 + n\bar{x}$, the joint likelihood can also be written as

$$
p(y, \lambda) = \prod_{i=1}^{n} \left(\frac{{x_i}^{y_i}}{y_i!} \right) \frac{{b_0}^{a_0}}{\Gamma(a_0)} \frac{\Gamma(a_n)}{{b_n}^{a_n}} \cdot \frac{{b_n}^{a_n}}{\Gamma(a_n)} \lambda^{a_n - 1} \exp\left[-b_n \lambda\right] \; . \tag{11}
$$

Using the probability density function of the gamma distribution (→ Proof II/3.3.2), λ can now be integrated out easily

$$
\begin{aligned}
\mathrm{p}(y) &= \int \prod_{i=1}^{n} \left(\frac{{x_i}^{y_i}}{y_i!} \right) \frac{{b_0}^{a_0}}{\Gamma(a_0)} \frac{\Gamma(a_n)}{{b_n}^{a_n}} \cdot \frac{{b_n}^{a_n}}{\Gamma(a_n)} \lambda^{a_n - 1} \exp\left[-b_n \lambda\right] \, \mathrm{d}\lambda \\
&= \prod_{i=1}^{n} \left(\frac{{x_i}^{y_i}}{y_i!} \right) \frac{\Gamma(a_n)}{\Gamma(a_0)} \frac{{b_0}^{a_0}}{{b_n}^{a_n}} \int \mathrm{Gam}(\lambda; a_n, b_n) \, \mathrm{d}\lambda \\
&= \prod_{i=1}^{n} \left(\frac{{x_i}^{y_i}}{y_i!} \right) \frac{\Gamma(a_n)}{\Gamma(a_0)} \frac{{b_0}^{a_0}}{{b_n}^{a_n}} \; ,
\end{aligned} \tag{12}
$$

such that the log model evidence (→ Definition IV/3.1.1) is shown to be

$$
\begin{aligned}
\log p(y|m) = &\sum_{i=1}^{n} y_i \log(x_i) - \sum_{i=1}^{n} \log y_i! + \\
&\log \Gamma(a_n) - \log \Gamma(a_0) + a_0 \log b_0 - a_n \log b_n \; .
\end{aligned} \tag{13}
$$

Sources:
- original work

Metadata: ID: P43 | shortcut: poissexp-lme | author: JoramSoch | date: 2020-02-04, 15:12.

3 Probability data

3.1 Beta-distributed data

3.1.1 Method of moments

Theorem: Let $y = \{y_1, \ldots, y_n\}$ be a set of observed counts independent and identically distributed (→ Definition "iid") according to a beta distribution (→ Definition "beta") with shapes α and β:

$$y_i \sim \mathrm{Bet}(\alpha, \beta), \quad i = 1, \ldots, n \; . \tag{1}$$

Then, the method-of-moments estimates (→ Definition "mom") for the shape parameters α and β are given by

$$\begin{aligned} \hat{\alpha} &= \bar{y} \left(\frac{\bar{y}(1-\bar{y})}{\bar{v}} - 1 \right) \\ \hat{\beta} &= (1-\bar{y}) \left(\frac{\bar{y}(1-\bar{y})}{\bar{v}} - 1 \right) \end{aligned} \tag{2}$$

where $\bar{y}$ is the sample mean (→ Proof "mean-sample") and $\bar{v}$ is the unbiased sample variance (→ Proof IV/1.1.3):

$$\begin{aligned} \bar{y} &= \frac{1}{n} \sum_{i=1}^{n} y_i \\ \bar{v} &= \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 \; . \end{aligned} \tag{3}$$

Proof: Mean (→ Proof "beta-mean") and variance (→ Proof "beta-var") of the beta distribution (→ Definition "beta") in terms of the parameters α and β are given by

$$\begin{aligned} \mathrm{E}(X) &= \frac{\alpha}{\alpha+\beta} \\ \mathrm{Var}(X) &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \; . \end{aligned} \tag{4}$$

Thus, matching the moments (→ Definition "mom") requires us to solve the following equation system for α and β:

$$\begin{aligned} \bar{y} &= \frac{\alpha}{\alpha+\beta} \\ \bar{v} &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \; . \end{aligned} \tag{5}$$

From the first equation, we can deduce:

$$
\begin{aligned}
\bar{y}(\alpha+\beta) &= \alpha \\
\alpha\bar{y} + \beta\bar{y} &= \alpha \\
\beta\bar{y} &= \alpha - \alpha\bar{y} \\
\beta &= \frac{\alpha}{\bar{y}} - \alpha \\
\beta &= \alpha \left(\frac{1}{\bar{y}} - 1 \right) \; .
\end{aligned} \tag{6}
$$

If we define $q = 1/\bar{y} - 1$ and plug (6) into the second equation, we have:

$$
\begin{aligned}
\bar{v} &= \frac{\alpha \cdot \alpha q}{(\alpha + \alpha q)^2 (\alpha + \alpha q + 1)} \\
&= \frac{\alpha^2 q}{(\alpha(1+q))^2 (\alpha(1+q)+1)} \\
&= \frac{q}{(1+q)^2 (\alpha(1+q)+1)} \\
&= \frac{q}{\alpha(1+q)^3 + (1+q)^2} \\
q &= \bar{v} \left[\alpha(1+q)^3 + (1+q)^2 \right] \; .
\end{aligned} \tag{7}
$$

Noting that $1 + q = 1/\bar{y}$ and $q = (1 - \bar{y})/\bar{y}$, one obtains for α:

$$
\begin{aligned}
\frac{1-\bar{y}}{\bar{y}} &= \bar{v} \left[\frac{\alpha}{\bar{y}^3} + \frac{1}{\bar{y}^2} \right] \\
\frac{1-\bar{y}}{\bar{y}\,\bar{v}} &= \frac{\alpha}{\bar{y}^3} + \frac{1}{\bar{y}^2} \\
\frac{\bar{y}^3(1-\bar{y})}{\bar{y}\,\bar{v}} &= \alpha + \bar{y} \\
\alpha &= \frac{\bar{y}^2(1-\bar{y})}{\bar{v}} - \bar{y} \\
&= \bar{y} \left(\frac{\bar{y}(1-\bar{y})}{\bar{v}} - 1 \right) \; .
\end{aligned} \tag{8}
$$

Plugging this into equation (6), one obtains for β:

$$
\begin{aligned}
\beta &= \bar{y} \left(\frac{\bar{y}(1-\bar{y})}{\bar{v}} - 1 \right) \cdot \left(\frac{1-\bar{y}}{\bar{y}} \right) \\
&= (1-\bar{y}) \left(\frac{\bar{y}(1-\bar{y})}{\bar{v}} - 1 \right) \; .
\end{aligned} \tag{9}
$$

Together, (8) and (9) constitute the method-of-moment estimates of α and β.

Sources:

- Wikipedia (2020): “Beta distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-01-20; URL: https://en.wikipedia.org/wiki/Beta_distribution#Method_of_moments.

Metadata: ID: P28 | shortcut: beta-mom | author: JoramSoch | date: 2020-01-22, 02:53.

3.2 Logistic regression

3.2.1 Log-odds and probability

Theorem: Assume a logistic regression model (→ Definition "logreg")

$$l_i = x_i\beta + \varepsilon_i, \; i = 1, \ldots, n \tag{1}$$

where x_i are the predictors corresponding to the i-th observation y_i and l_i are the log-odds that $y_i = 1$.
Then, the probability that $y_i = 1$ is given by

$$Pr(y_i = 1) = \frac{1}{1 + b^{-(x_i\beta+\varepsilon_i)}} \tag{2}$$

where b is the base used to form the log-odds l_i.

Proof: Let us denote $Pr(y_i = 1)$ as p_i. Then, the log-odds are

$$l_i = \log_b \frac{p_i}{1-p_i} \tag{3}$$

and using (1), we have

$$\begin{aligned}
\log_b \frac{p_i}{1-p_i} &= x_i\beta + \varepsilon_i \\
\frac{p_i}{1-p_i} &= b^{x_i\beta+\varepsilon_i} \\
p_i &= \left(b^{x_i\beta+\varepsilon_i}\right)(1-p_i) \\
p_i\left(1 + b^{x_i\beta+\varepsilon_i}\right) &= b^{x_i\beta+\varepsilon_i} \\
p_i &= \frac{b^{x_i\beta+\varepsilon_i}}{1 + b^{x_i\beta+\varepsilon_i}} \\
p_i &= \frac{b^{x_i\beta+\varepsilon_i}}{b^{x_i\beta+\varepsilon_i}\left(1 + b^{-(x_i\beta+\varepsilon_i)}\right)} \\
p_i &= \frac{1}{1 + b^{-(x_i\beta+\varepsilon_i)}}
\end{aligned} \tag{4}$$

which proves the identity given by (2).

Sources:

- Wikipedia (2020): "Logistic regression"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-03; URL: https://en.wikipedia.org/wiki/Logistic_regression#Logistic_model.

Metadata: ID: P72 | shortcut: logreg-lonp | author: JoramSoch | date: 2020-03-03, 12:01.

4 Categorical data

4.1 Binomial observations

4.1.1 Conjugate prior distribution

Theorem: Let y be the number of successes resulting from n independent trials with unknown success probability p, such that y follows a binomial distribution (→ Definition "bin"):

$$y \sim \mathrm{Bin}(n, p) \; . \tag{1}$$

Then, the conjugate prior (→ Definition "prior-conj") for the model parameter p is a beta distribution (→ Definition "beta"):

$$\mathrm{p}(p) = \mathrm{Bet}(p; \alpha_0, \beta_0) \; . \tag{2}$$

Proof: With the probability mass function of the binomial distribution (→ Proof "bin-pmf"), the likelihood function (→ Definition I/2.2.2) implied by (1) is given by

$$\mathrm{p}(y|p) = \binom{n}{y} \, p^y \, (1-p)^{n-y} \; . \tag{3}$$

In other words, the likelihood function is proportional to a power of p times a power of $(1-p)$:

$$\mathrm{p}(y|p) \propto p^y \, (1-p)^{n-y} \; . \tag{4}$$

The same is true for a beta distribution over p

$$\mathrm{p}(p) = \mathrm{Bet}(p; \alpha_0, \beta_0) \tag{5}$$

the probability density function of which (→ Proof "beta-pdf")

$$\mathrm{p}(p) = \frac{1}{B(\alpha_0, \beta_0)} \, p^{\alpha_0 - 1} \, (1-p)^{\beta_0 - 1} \tag{6}$$

exhibits the same proportionality

$$\mathrm{p}(p) \propto p^{\alpha_0 - 1} \, (1-p)^{\beta_0 - 1} \tag{7}$$

and is therefore conjugate relative to the likelihood.

Sources:

- Wikipedia (2020): "Binomial distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-01-23; URL: https://en.wikipedia.org/wiki/Binomial_distribution#Estimation_of_parameters.

Metadata: ID: P29 | shortcut: bin-prior | author: JoramSoch | date: 2020-01-23, 23:38.

4.1.2 Posterior distribution

Theorem: Let y be the number of successes resulting from n independent trials with unknown success probability p, such that y follows a binomial distribution (→ Definition "bin"):

$$y \sim \mathrm{Bin}(n, p) . \tag{1}$$

Moreover, assume a beta prior distribution (→ Proof III/4.1.1) over the model parameter p:

$$\mathrm{p}(p) = \mathrm{Bet}(p; \alpha_0, \beta_0) . \tag{2}$$

Then, the posterior distribution (→ Definition I/2.2.6) is also a beta distribution (→ Definition "beta")

$$\mathrm{p}(p|y) = \mathrm{Bet}(p; \alpha_n, \beta_n) . \tag{3}$$

and the posterior hyperparameters (→ Definition "post-hyp") are given by

$$\begin{aligned} \alpha_n &= \alpha_0 + y \\ \beta_n &= \beta_0 + (n - y) . \end{aligned} \tag{4}$$

Proof: With the probability mass function of the binomial distribution (→ Proof "bin-pmf"), the likelihood function (→ Definition I/2.2.2) implied by (1) is given by

$$\mathrm{p}(y|p) = \binom{n}{y} \, p^y \, (1-p)^{n-y} . \tag{5}$$

Combining the likelihood function (5) with the prior distribution (2), the joint likelihood (→ Definition I/2.2.5) of the model is given by

$$\begin{aligned} \mathrm{p}(y, p) &= \mathrm{p}(y|p) \, \mathrm{p}(p) \\ &= \binom{n}{y} \, p^y \, (1-p)^{n-y} \cdot frac1B(\alpha_0, \beta_0) \, p^{\alpha_0 - 1} \, (1-p)^{\beta_0 - 1} \\ &= \frac{1}{B(\alpha_0, \beta_0)} \binom{n}{y} \, p^{\alpha_0 + y - 1} \, (1-p)^{\beta_0 + (n-y) - 1} . \end{aligned} \tag{6}$$

Note that the posterior distribution is proportional to the joint likelihood (→ Proof "post-jl"):

$$\mathrm{p}(p|y) \propto \mathrm{p}(y, p) . \tag{7}$$

Setting $\alpha_n = \alpha_0 + y$ and $\beta_n = \beta_0 + (n - y)$, the posterior distribution is therefore proportional to

$$\mathrm{p}(p|y) \propto p^{\alpha_n - 1} \, (1-p)^{\beta_n - 1} \tag{8}$$

which, when normalized to one, results in the probability density function of the beta distribution (→ Proof "beta-pdf"):

$$\mathrm{p}(p|y) = \frac{1}{B(\alpha_n, \beta_n)} \, p^{\alpha_n - 1} \, (1-p)^{\beta_n - 1} = \mathrm{Bet}(p; \alpha_n, \beta_n) . \tag{9}$$

Sources:

• Wikipedia (2020): "Binomial distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-01-23; URL: https://en.wikipedia.org/wiki/Binomial_distribution#Estimation_of_parameters.

Metadata: ID: P30 | shortcut: bin-post | author: JoramSoch | date: 2020-01-24, 00:20.

4.1.3 Log model evidence

Theorem: Let y be the number of successes resulting from n independent trials with unknown success probability p, such that y follows a binomial distribution (→ Definition "bin"):

$$y \sim \mathrm{Bin}(n, p) \; . \tag{1}$$

Moreover, assume a beta prior distribution (→ Proof III/4.1.1) over the model parameter p:

$$\mathrm{p}(p) = \mathrm{Bet}(p; \alpha_0, \beta_0) \; . \tag{2}$$

Then, the log model evidence (→ Definition IV/3.1.1) for this model is

$$\log \mathrm{p}(y|m) = \log \binom{n}{y} + \log B(\alpha_n, \beta_n) - \log B(\alpha_0, \beta_0) \tag{3}$$

where the posterior hyperparameters (→ Definition "post-hyp") are given by

$$\begin{split} \alpha_n &= \alpha_0 + y \\ \beta_n &= \beta_0 + (n - y) \; . \end{split} \tag{4}$$

Proof: With the probability mass function of the binomial distribution (→ Proof "bin-pmf"), the likelihood function (→ Definition I/2.2.2) implied by (1) is given by

$$\mathrm{p}(y|p) = \binom{n}{y} p^y \, (1-p)^{n-y} \; . \tag{5}$$

Combining the likelihood function (5) with the prior distribution (2), the joint likelihood (→ Definition I/2.2.5) of the model is given by

$$\begin{split} \mathrm{p}(y, p) &= \mathrm{p}(y|p) \, \mathrm{p}(p) \\ &= \binom{n}{y} p^y \, (1-p)^{n-y} \cdot frac1B(\alpha_0, \beta_0) \, p^{\alpha_0 - 1} \, (1-p)^{\beta_0 - 1} \\ &= \binom{n}{y} \frac{1}{B(\alpha_0, \beta_0)} \, p^{\alpha_0 + y - 1} \, (1-p)^{\beta_0 + (n-y) - 1} \; . \end{split} \tag{6}$$

Note that the model evidence is the marginal density of the joint likelihood (→ Definition I/2.2.7):

$$\mathrm{p}(y) = \int \mathrm{p}(y, p) \, \mathrm{d}p \; . \tag{7}$$

Setting $\alpha_n = \alpha_0 + y$ and $\beta_n = \beta_0 + (n - y)$, the joint likelihood can also be written as

$$\mathrm{p}(y, p) = \binom{n}{y} \frac{1}{B(\alpha_0, \beta_0)} \frac{B(\alpha_n, \beta_n)}{1} \frac{1}{B(\alpha_n, \beta_n)} \, p^{\alpha_n - 1} \, (1-p)^{\beta_n - 1} \; . \tag{8}$$

Using the probability density function of the beta distribution (→ Proof "beta-pdf"), p can now be integrated out easily

$$
\begin{aligned}
\mathrm{p}(y) &= \int \binom{n}{y} \frac{1}{B(\alpha_0, \beta_0)} \frac{B(\alpha_n, \beta_n)}{1} \frac{1}{B(\alpha_n, \beta_n)} \, p^{\alpha_n - 1} \, (1-p)^{\beta_n - 1} \, \mathrm{d}p \\
&= \binom{n}{y} \frac{B(\alpha_n, \beta_n)}{B(\alpha_0, \beta_0)} \int \mathrm{Bet}(p; \alpha_n, \beta_n) \, \mathrm{d}p \\
&= \binom{n}{y} \frac{B(\alpha_n, \beta_n)}{B(\alpha_0, \beta_0)} \; ,
\end{aligned} \tag{9}
$$

such that the log model evidence (→ Definition IV/3.1.1) is shown to be

$$
\log \mathrm{p}(y|m) = \log \binom{n}{y} + \log B(\alpha_n, \beta_n) - \log B(\alpha_0, \beta_0) \; . \tag{10}
$$

Sources:

- Wikipedia (2020): "Beta-binomial distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-01-24; URL: https://en.wikipedia.org/wiki/Beta-binomial_distribution#Motivation_and_derivation.

Metadata: ID: P31 | shortcut: bin-lme | author: JoramSoch | date: 2020-01-24, 00:44.

Chapter IV

Model Selection

1 Goodness-of-fit measures

1.1 Residual variance

1.1.1 Definition

Definition: Let there be a linear regression model (→ Definition "mlr")

$$y = X\beta + \varepsilon, \; \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \tag{1}$$

with measured data y, known design matrix X and covariance structure V as well as unknown regression coefficients β and noise variance σ^2.
Then, an estimate of the noise variance σ^2 is called the "residual variance" $\hat{\sigma}^2$, e.g. obtained via maximum likelihood estimation (→ Definition "mle").

Sources:
- original work

Metadata: ID: D20 | shortcut: resvar | author: JoramSoch | date: 2020-02-25, 11:21.

1.1.2 Maximum likelihood estimator is biased

Theorem: Let $x = \{x_1, \ldots, x_n\}$ be a set of independent normally distributed (→ Definition II/3.2.1) observations with unknown mean (→ Definition I/1.2.1) μ and variance (→ Definition I/1.3.1) σ^2:

$$x_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2), \quad i = 1, \ldots, n \; . \tag{1}$$

Then,
1) the maximum likelihood estimator (→ Definition "mle") of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{2}$$

where

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{3}$$

2) and $\hat{\sigma}^2$ is a biased estimator (→ Definition "est-unb") of σ^2

$$\mathbb{E}\left[\hat{\sigma}^2\right] \neq \sigma^2 \; , \tag{4}$$

more precisely:

$$\mathbb{E}\left[\hat{\sigma}^2\right] = \frac{n-1}{n} \sigma^2 \; . \tag{5}$$

Proof:
1) This is equivalent to the maximum likelihood estimator for the univariate Gaussian with unknown variance (→ Proof "ug-mle") and a special case of the maximum likelihood estimator for multiple linear regression (→ Proof "mlr-mle") in which $y = x$, $X = 1_n$ and $\hat{\beta} = \bar{x}$:

$$
\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n}(y - X\hat{\beta})^\mathrm{T}(y - X\hat{\beta}) \\
&= \frac{1}{n}(x - 1_n \bar{x})^\mathrm{T}(x - 1_n \bar{x}) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \; .
\end{aligned} \tag{6}
$$

2) The expectation (→ Definition I/1.2.1) of the maximum likelihood estimator (→ Definition "mle") can be developed as follows:

$$
\begin{aligned}
\mathbb{E}\left[\hat{\sigma}^2\right] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2\right] \\
&= \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2\right] \\
&= \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n}\left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right)\right] \\
&= \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n}x_i^2 - 2\sum_{i=1}^{n}x_i\bar{x} + \sum_{i=1}^{n}\bar{x}^2\right] \\
&= \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n}x_i^2 - 2n\bar{x}^2 + n\bar{x}^2\right] \\
&= \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n}x_i^2 - n\bar{x}^2\right] \\
&= \frac{1}{n}\left(\sum_{i=1}^{n}\mathbb{E}\left[x_i^2\right] - n\mathbb{E}\left[\bar{x}^2\right]\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[x_i^2\right] - \mathbb{E}\left[\bar{x}^2\right]
\end{aligned} \tag{7}
$$

Due to the partition of variance into expected values (→ Proof "var-mean")

$$
\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 \; , \tag{8}
$$

we have

$$
\begin{aligned}
\mathrm{Var}(x_i) &= \mathbb{E}(x_i^2) - \mathbb{E}(x_i)^2 \\
\mathrm{Var}(\bar{x}) &= \mathbb{E}(\bar{x}^2) - \mathbb{E}(\bar{x})^2 \; ,
\end{aligned} \tag{9}
$$

such that (7) becomes

$$
\mathbb{E}\left[\hat{\sigma}^2\right] = \frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{Var}(x_i) + \mathbb{E}(x_i)^2\right) - \left(\mathrm{Var}(\bar{x}) + \mathbb{E}(\bar{x})^2\right) \; . \tag{10}
$$

From (1), it follows that

$$\mathbb{E}(x_i) = \mu \quad \text{and} \quad \mathrm{Var}(x_i) = \sigma^2 \; . \tag{11}$$

The expectation of (→ Proof "ug-unb") $\bar{x}$ given by (3) is

$$\begin{split} \mathbb{E}\left[\bar{x}\right] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[x_i\right] \\ &\overset{(11)}{=} \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n} \cdot n \cdot \mu \\ &= \mu \; . \end{split} \tag{12}$$

The variance of $\bar{x}$ given by (3) is

$$\begin{split} \mathrm{Var}\left[\bar{x}\right] &= \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}\left[x_i\right] \\ &\overset{(11)}{=} \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{1}{n^2} \cdot n \cdot \sigma^2 \\ &= \frac{1}{n}\sigma^2 \; . \end{split} \tag{13}$$

Plugging (11), (12) and (13) into (10), we have

$$\begin{split} \mathbb{E}\left[\hat{\sigma}^2\right] &= \frac{1}{n}\sum_{i=1}^{n} \left(\sigma^2 + \mu^2\right) - \left(\frac{1}{n}\sigma^2 + \mu^2\right) \\ \mathbb{E}\left[\hat{\sigma}^2\right] &= \frac{1}{n} \cdot n \cdot \left(\sigma^2 + \mu^2\right) - \left(\frac{1}{n}\sigma^2 + \mu^2\right) \\ \mathbb{E}\left[\hat{\sigma}^2\right] &= \sigma^2 + \mu^2 - \frac{1}{n}\sigma^2 - \mu^2 \\ \mathbb{E}\left[\hat{\sigma}^2\right] &= \frac{n-1}{n}\sigma^2 \end{split} \tag{14}$$

which proves the bias (→ Definition "est-unb") given by (5).

Sources:

- Liang, Dawen (????): "Maximum Likelihood Estimator for Variance is Biased: Proof", retrieved on 2020-02-24; URL: https://dawenl.github.io/files/mle_biased.pdf.

Metadata: ID: P61 | shortcut: resvar-bias | author: JoramSoch | date: 2020-02-24, 23:44.

1.1.3 Construction of unbiased estimator

Theorem: Let $x = \{x_1, \ldots, x_n\}$ be a set of independent normally distributed (→ Definition II/3.2.1) observations with unknown mean (→ Definition I/1.2.1) μ and variance (→ Definition I/1.3.1) σ^2:

$$x_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2), \quad i = 1, \ldots, n \; . \tag{1}$$

An unbiased estimator (→ Definition "est-unb") of σ^2 is given by

$$\hat{\sigma}^2_{\text{unb}} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \; . \tag{2}$$

Proof: It can be shown that (→ Proof IV/1.1.2) the maximum likelihood estimator (→ Definition "mle") of σ^2

$$\hat{\sigma}^2_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{3}$$

is a biased estimator (→ Definition "est-unb") in the sense that

$$\mathbb{E}\left[\hat{\sigma}^2_{\text{MLE}}\right] = \frac{n-1}{n} \sigma^2 \; . \tag{4}$$

From (4), it follows that

$$\begin{split} \mathbb{E}\left[\frac{n}{n-1} \hat{\sigma}^2_{\text{MLE}}\right] &= \frac{n}{n-1} \mathbb{E}\left[\hat{\sigma}^2_{\text{MLE}}\right] \\ &\overset{(4)}{=} \frac{n}{n-1} \cdot \frac{n-1}{n} \sigma^2 \\ &= \sigma^2 \; , \end{split} \tag{5}$$

such that an unbiased estimator (→ Definition "est-unb") can be constructed as

$$\begin{split} \hat{\sigma}^2_{\text{unb}} &= \frac{n}{n-1} \hat{\sigma}^2_{\text{MLE}} \\ &\overset{(3)}{=} \frac{n}{n-1} \cdot \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \\ &= \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \; . \end{split} \tag{6}$$

Sources:

- Liang, Dawen (????): "Maximum Likelihood Estimator for Variance is Biased: Proof", retrieved on 2020-02-25; URL: https://dawenl.github.io/files/mle_biased.pdf.

Metadata: ID: P62 | shortcut: resvar-unb | author: JoramSoch | date: 2020-02-25, 15:38.

1.2 R-squared

1.2.1 Definition

Definition: Let there be a linear regression model (→ Definition "mlr") with independent (→ Definition "ind") observations

$$y = X\beta + \varepsilon, \; \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \tag{1}$$

with measured data y, known design matrix X as well as unknown regression coefficients β and noise variance σ^2.
Then, the proportion of the variance of the dependent variable y ("total variance (→ Definition "tss")") that can be predicted from the independent variables X ("explained variance (→ Definition "ess")") is called "coefficient of determination", "R-squared" or R^2.

Sources:
- Wikipedia (2020): "Coefficient of determination"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-25; URL: https://en.wikipedia.org/wiki/Mean_squared_error#Proof_of_variance_and_bias_relationship.

Metadata: ID: D21 | shortcut: rsq | author: JoramSoch | date: 2020-02-25, 11:41.

1.2.2 Derivation of R² and adjusted R²

Theorem: Given a linear regression model (→ Definition "mlr")

$$y = X\beta + \varepsilon, \; \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \tag{1}$$

with n independent observations and p independent variables,
1) the coefficient of determination (→ Definition IV/1.2.1) is given by

$$R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} \tag{2}$$

2) the adjusted coefficient of determination is

$$R^2_{\mathrm{adj}} = 1 - \frac{\mathrm{RSS}/(n-p)}{\mathrm{TSS}/(n-1)} \tag{3}$$

where the residual (→ Definition "rss") and total sum of squares (→ Definition "tss") are

$$\begin{aligned} \mathrm{RSS} &= \sum_{i=1}^n (y_i - \hat{y}_i)^2, \quad \hat{y} = X\hat{\beta} \\ \mathrm{TSS} &= \sum_{i=1}^n (y_i - \bar{y})^2 \;, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^n y_i \end{aligned} \tag{4}$$

where X is the $n \times p$ design matrix and $\hat{\beta}$ are the ordinary least squares (→ Definition "mlr-ols") estimates.

Proof: The coefficient of determination R^2 is defined as (→ Definition IV/1.2.1) the proportion of the variance explained by the independent variables, relative to the total variance in the data.

1) If we define the explained sum of squares (→ Definition "ess") as

$$\mathrm{ESS} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \;, \tag{5}$$

then R^2 is given by

$$R^2 = \frac{\mathrm{ESS}}{\mathrm{TSS}} \; . \tag{6}$$

which is equal to

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} \; , \tag{7}$$

because (→ Proof "mlr-pss") TSS = ESS + RSS.

2) Using (4), the coefficient of determination can be also written as:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\frac{1}{n}\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2} \; . \tag{8}$$

If we replace the variance estimates by their unbiased estimators (→ Proof IV/1.1.3), we obtain

$$R^2_{\mathrm{adj}} = 1 - \frac{\frac{1}{n-p}\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\mathrm{RSS}/\mathrm{df}_r}{\mathrm{TSS}/\mathrm{df}_t} \tag{9}$$

where $\mathrm{df}_r = n-p$ and $\mathrm{df}_t = n-1$ are the residual and total degrees of freedom (→ Definition "dof").

This gives the adjusted R^2 which adjusts R^2 for the number of explanatory variables.

Sources:
- Wikipedia (2019): "Coefficient of determination"; in: *Wikipedia, the free encyclopedia*, retrieved on 2019-12-06; URL: https://en.wikipedia.org/wiki/Coefficient_of_determination#Adjusted_R2.

Metadata: ID: P8 | shortcut: rsq-der | author: JoramSoch | date: 2019-12-06, 11:19.

1.2.3 Relationship to maximum log-likelihood

Theorem: Given a linear regression model (→ Definition "mlr") with independent observations

$$y = X\beta + \varepsilon, \; \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \; , \tag{1}$$

the coefficient of determination (→ Definition IV/1.2.1) can be expressed in terms of the maximum log-likelihood (→ Definition "mll") as

$$R^2 = 1 - \left(\exp[\Delta\mathrm{MLL}]\right)^{-2/n} \tag{2}$$

where n is the number of observations and ΔMLL is the difference in maximum log-likelihood between the model given by (1) and a linear regression model with only a constant regressor.

Proof: First, we express the maximum log-likelihood (→ Definition "mll") (MLL) of a linear regression model in terms of its residual sum of squares (→ Definition "rss") (RSS). The model in (1) implies the following log-likelihood function (→ Definition "llf")

$$\mathrm{LL}(\beta, \sigma^2) = \log p(y|\beta, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y - X\beta)^\mathrm{T}(y - X\beta) \; , \tag{3}$$

such that maximum likelihood estimates are (→ Proof "mlr-mle")

$$\hat{\beta} = (X^{\mathrm{T}} X)^{-1} X^{\mathrm{T}} y \tag{4}$$

$$\hat{\sigma}^2 = \frac{1}{n} (y - X\hat{\beta})^{\mathrm{T}} (y - X\hat{\beta}) \tag{5}$$

and the residual sum of squares (→ Definition "rss") is

$$\mathrm{RSS} = \sum_{i=1}^{n} \hat{\varepsilon}_i = \hat{\varepsilon}^{\mathrm{T}} \hat{\varepsilon} = (y - X\hat{\beta})^{\mathrm{T}} (y - X\hat{\beta}) = n \cdot \hat{\sigma}^2 \; . \tag{6}$$

Since $\hat{\beta}$ and $\hat{\sigma}^2$ are maximum likelihood estimates (→ Definition "mle"), plugging them into the log-likelihood function gives the maximum log-likelihood:

$$\mathrm{MLL} = \mathrm{LL}(\hat{\beta}, \hat{\sigma}^2) = -\frac{n}{2} \log(2\pi\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2} (y - X\hat{\beta})^{\mathrm{T}} (y - X\hat{\beta}) \; . \tag{7}$$

With (6) for the first $\hat{\sigma}^2$ and (5) for the second $\hat{\sigma}^2$, the MLL becomes

$$\mathrm{MLL} = -\frac{n}{2} \log(\mathrm{RSS}) - \frac{n}{2} \log\left(\frac{2\pi}{n}\right) - \frac{n}{2} \; . \tag{8}$$

Second, we establish the relationship between maximum log-likelihood (MLL) and coefficient of determination (R^2). Consider the two models

$$\begin{aligned} m_0 &: \; X_0 = 1_n \\ m_1 &: \; X_1 = X \end{aligned} \tag{9}$$

For m_1, the residual sum of squares is given by (6); and for m_0, the residual sum of squares is equal to the total sum of squares (→ Definition "tss"):

$$\mathrm{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \; . \tag{10}$$

Using (8), we can therefore write

$$\Delta\mathrm{MLL} = \mathrm{MLL}(m_1) - \mathrm{MLL}(m_0) = -\frac{n}{2} \log(\mathrm{RSS}) + \frac{n}{2} \log(\mathrm{TSS}) \; . \tag{11}$$

Exponentiating both sides of the equation, we have:

$$\begin{aligned} \exp[\Delta\mathrm{MLL}] &= \exp\left[-\frac{n}{2} \log(\mathrm{RSS}) + \frac{n}{2} \log(\mathrm{TSS})\right] \\ &= \left(\exp\left[\log(\mathrm{RSS}) - \log(\mathrm{TSS})\right]\right)^{-n/2} \\ &= \left(\frac{\exp[\log(\mathrm{RSS})]}{\exp[\log(\mathrm{TSS})]}\right)^{-n/2} \\ &= \left(\frac{\mathrm{RSS}}{\mathrm{TSS}}\right)^{-n/2} \; . \end{aligned} \tag{12}$$

Taking both sides to the power of $-2/n$ and subtracting from 1, we have

$$
\begin{aligned}
(\exp[\Delta\mathrm{MLL}])^{-2/n} &= \frac{\mathrm{RSS}}{\mathrm{TSS}} \\
1 - (\exp[\Delta\mathrm{MLL}])^{-2/n} &= 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} = R^2
\end{aligned} \tag{13}
$$

which proves the identity given above.

Sources:
- original work

Metadata: ID: P14 | shortcut: rsq-mll | author: JoramSoch | date: 2020-01-08, 04:46.

1.3 Signal-to-noise ratio

1.3.1 Definition

Definition: Let there be a linear regression model (→ Definition "mlr") with independent (→ Definition "ind") observations

$$
y = X\beta + \varepsilon, \; \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \tag{1}
$$

with measured data y, known design matrix X as well as unknown regression coefficients β and noise variance σ^2.
Given estimated regression coefficients (→ Definition "mlr-beta") $\hat{\beta}$ and residual variance (→ Definition IV/1.1.1) $\hat{\sigma}^2$, the signal-to-noise ratio (SNR) is defined as the ratio of estimated signal variance to estimated noise variance:

$$
\mathrm{SNR} = \frac{\mathrm{Var}(X\hat{\beta})}{\hat{\sigma}^2} \; . \tag{2}
$$

Sources:
- Soch J, Allefeld C (2018): "MACS – a new SPM toolbox for model assessment, comparison and selection"; in: *Journal of Neuroscience Methods*, vol. 306, pp. 19-31, eq. 6; URL: https://www.sciencedirect.com/science/article/pii/S0165027018301468; DOI: 10.1016/j.jneumeth.2018.05.017.

Metadata: ID: D22 | shortcut: snr | author: JoramSoch | date: 2020-02-25, 12:01.

2 Classical information criteria

2.1 Akaike information criterion

2.1.1 Definition

Definition: Let m be a generative model (→ Definition I/2.2.1) with likelihood function (→ Definition I/2.2.2) $p(y|\theta, m)$ and maximum likelihood estimates (→ Definition "mle")

$$\hat{\theta} = \arg\max_{\theta} \log p(y|\theta, m) \; . \tag{1}$$

Then, the Akaike information criterion (AIC) of this model is defined as

$$\mathrm{AIC}(m) = -2 \log p(y|\hat{\theta}, m) + 2\, p \tag{2}$$

where p is the number of free parameters estimated via (1).

Sources:
- Akaike H (1974): "A New Look at the Statistical Model Identification"; in: *IEEE Transactions on Automatic Control*, vol. AC-19, no. 6, pp. 716-723; URL: https://ieeexplore.ieee.org/document/1100705; DOI: 10.1109/TAC.1974.1100705.

Metadata: ID: D23 | shortcut: aic | author: JoramSoch | date: 2020-02-25, 12:31.

2.2 Bayesian information criterion

2.2.1 Definition

Definition: Let m be a generative model (→ Definition I/2.2.1) with likelihood function (→ Definition I/2.2.2) $p(y|\theta, m)$ and maximum likelihood estimates (→ Definition "mle")

$$\hat{\theta} = \arg\max_{\theta} \log p(y|\theta, m) \; . \tag{1}$$

Then, the Bayesian information criterion (BIC) of this model is defined as

$$\mathrm{BIC}(m) = -2 \log p(y|\hat{\theta}, m) + p \log n \tag{2}$$

where n is the number of data points and p is the number of free parameters estimated via (1).

Sources:
- Schwarz G (1978): "Estimating the Dimension of a Model"; in: *The Annals of Statistics*, vol. 6, no. 2, pp. 461-464; URL: https://www.jstor.org/stable/2958889.

Metadata: ID: D24 | shortcut: bic | author: JoramSoch | date: 2020-02-25, 12:21.

2.2.2 Derivation

Theorem: Let $p(y|\theta, m)$ be the likelihood function (→ Definition I/2.2.2) of a generative model (→ Definition I/2.2.1) $m \in \mathcal{M}$ with model parameters $\theta \in \Theta$ describing measured data $y \in \mathbb{R}^n$.

Let $p(\theta|m)$ be a prior distribution (→ Definition I/2.2.3) on the model parameters. Assume that likelihood function and prior density are twice differentiable.
Then, as the number of data points goes to infinity, an approximation to the log-marginal likelihood (→ Definition I/2.2.7) $\log p(y|m)$, up to constant terms not depending on the model, is given by the Bayesian information criterion (→ Definition IV/2.2.1) (BIC) as

$$-2 \log p(y|m) \approx \mathrm{BIC}(m) = -2 \log p(y|\hat{\theta}, m) + p \log n \tag{1}$$

where $\hat{\theta}$ is the maximum likelihood estimator (→ Definition "mle") (MLE) of θ, n is the number of data points and p is the number of model parameters.

Proof: Let $\mathrm{LL}(\theta)$ be the log-likelihood function (→ Definition "llf")

$$\mathrm{LL}(\theta) = \log p(y|\theta, m) \tag{2}$$

and define the functions g and h as follows:

$$\begin{aligned} g(\theta) &= p(\theta|m) \\ h(\theta) &= \frac{1}{n} \mathrm{LL}(\theta) \; . \end{aligned} \tag{3}$$

Then, the marginal likelihood (→ Definition I/2.2.7) can be written as follows:

$$\begin{aligned} p(y|m) &= \int_{\Theta} p(y|\theta, m) \, p(\theta|m) \, \mathrm{d}\theta \\ &= \int_{\Theta} \exp\left[n \, h(\theta)\right] \, g(\theta) \, \mathrm{d}\theta \; . \end{aligned} \tag{4}$$

This is an integral suitable for Laplace approximation which states that

$$\int_{\Theta} \exp\left[n \, h(\theta)\right] \, g(\theta) \, \mathrm{d}\theta = \left(\sqrt{\frac{2\pi}{n}} \right)^{p} \exp\left[n \, h(\theta_0)\right] \left(g(\theta_0) \left| J(\theta_0) \right|^{-1/2} + O(1/n) \right) \tag{5}$$

where θ_0 is the value that maximizes $h(\theta)$ and $J(\theta_0)$ is the Hessian matrix evaluated at θ_0. In our case, we have $h(\theta) = 1/n \, \mathrm{LL}(\theta)$ such that θ_0 is the maximum likelihood estimator $\hat{\theta}$:

$$\hat{\theta} = \arg\max_{\theta} \mathrm{LL}(\theta) \; . \tag{6}$$

With this, (5) can be applied to (4) using (3) to give:

$$p(y|m) \approx \left(\sqrt{\frac{2\pi}{n}} \right)^{p} p(y|\hat{\theta}, m) \, p(\hat{\theta}|m) \, \left| J(\hat{\theta}) \right|^{-1/2} \; . \tag{7}$$

Logarithmizing and multiplying with -2, we have:

$$-2 \log p(y|m) \approx -2 \, \mathrm{LL}(\hat{\theta}) + p \log n - p \log(2\pi) - 2 \log p(\hat{\theta}|m) + \log \left| J(\hat{\theta}) \right| \; . \tag{8}$$

As $n \to \infty$, the last three terms are $O_p(1)$ and can therefore be ignored when comparing between models $\mathcal{M} = \{m_1, \ldots, m_M\}$ and using $p(y|m_j)$ to compute posterior model probabilies (→ Definition "led-pmp") $p(m_j|y)$. With that, the BIC is given as

$$\mathrm{BIC}(m) = -2 \log p(y|\hat{\theta}, m) + p \log n \; . \tag{9}$$

Sources:
- Claeskens G, Hjort NL (2008): "The Bayesian information criterion"; in: *Model Selection and Model Averaging*, ch. 3.2, pp. 78-81; URL: https://www.cambridge.org/core/books/model-selection-and-model-av E6F1EC77279D1223423BB64FC3A12C37; DOI: 10.1017/CBO9780511790485.

Metadata: ID: P32 | shortcut: bic-der | author: JoramSoch | date: 2020-01-26, 23:36.

2.3 Deviance information criterion

2.3.1 Definition

Definition: Let m be a generative model (→ Definition I/2.2.1) with likelihood function (→ Definition I/2.2.2) $p(y|\theta, m)$ and prior distribution (→ Definition I/2.2.3) $p(\theta|m)$. Together, likelihood function and prior distribution imply a posterior distribution (→ Definition I/2.2.6) $p(\theta|y, m)$. Define the posterior expected log-likelihood (→ Definition "llf") (PLL)

$$\mathrm{PLL}(m) = \left\langle \log p(y|\theta, m) \right\rangle_{\theta|y} \tag{1}$$

and the log-likelihood (→ Definition "llf") at the posterior expectation (LLP)

$$\mathrm{LLP}(m) = \log p(y| \left\langle \theta \right\rangle_{\theta|y}, m) \tag{2}$$

where $\langle \cdot \rangle_{\theta|y}$ denotes an expectation across the posterior distribution.
Then, the deviance information criterion (DIC) of the model is defined as

$$\mathrm{DIC}(m) = -2\,\mathrm{LLP}(m) + 2\,p_D \quad \text{or} \quad \mathrm{DIC}(m) = -2\,\mathrm{PLL}(m) + p_D \tag{3}$$

where the "effective number of parameters" p_D is given by

$$p_D = -2\,\mathrm{PLL}(m) + 2\,\mathrm{LLP}(m) \; . \tag{4}$$

Sources:
- Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A (2002): "Bayesian measures of model complexity and fit"; in: *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, vol. 64, iss. 4, pp. 583-639; URL: https://rss.onlinelibrary.wiley.com/doi/10.1111/1467-9868.00353; DOI: 10.1111/1467-9868.00353.
- Soch J, Allefeld C (2018): "MACS – a new SPM toolbox for model assessment, comparison and selection"; in: *Journal of Neuroscience Methods*, vol. 306, pp. 19-31, eqs. 10-12; URL: https://www.sciencedirect.com/science/article/pii/S0165027018301468; DOI: 10.1016/j.jneumeth.2018.05.017.

Metadata: ID: D25 | shortcut: dic | author: JoramSoch | date: 2020-02-25, 12:46.

3 Bayesian model selection

3.1 Log model evidence

3.1.1 Definition

Definition: Let m be a generative model (→ Definition I/2.2.1) with likelihood function (→ Definition I/2.2.2) $p(y|\theta, m)$ and prior distribution (→ Definition I/2.2.3) $p(\theta|m)$. Then, the log model evidence (LME) of this model is defined as the logarithm of the marginal likelihood (→ Definition I/2.2.7):

$$\mathrm{LME}(m) = \log p(y|m) \; . \tag{1}$$

Sources:
- Soch J, Allefeld C (2018): "MACS – a new SPM toolbox for model assessment, comparison and selection"; in: *Journal of Neuroscience Methods*, vol. 306, pp. 19-31, eq. 13; URL: https://www.sciencedirect.com/science/article/pii/S0165027018301468; DOI: 10.1016/j.jneumeth.2018.05.017.

Metadata: ID: D26 | shortcut: lme | author: JoramSoch | date: 2020-02-25, 12:56.

3.1.2 Derivation

Theorem: Let $p(y|\theta, m)$ be a likelihood function (→ Definition I/2.2.2) of a generative model (→ Definition I/2.2.1) m for making inferences on model parameters θ given measured data y. Moreover, let $p(\theta|m)$ be a prior distribution (→ Definition I/2.2.3) on model parameters θ. Then, the log model evidence (→ Definition IV/3.1.1) (LME), also called marginal log-likelihood,

$$\mathrm{LME}(m) = \log p(y|m) \; , \tag{1}$$

can be expressed
1) as

$$\mathrm{LME}(m) = \log \int p(y|\theta, m) \, p(\theta|m) \, \mathrm{d}\theta \tag{2}$$

2) or

$$\mathrm{LME}(m) = \log p(y|\theta, m) + \log p(\theta|m) - \log p(\theta|y, m) \; . \tag{3}$$

Proof:
1) The first expression is a simple consequence of the law of marginal probability (→ Definition "prob-marg") for continuous variables according to which

$$p(y|m) = \int p(y|\theta, m) \, p(\theta|m) \, \mathrm{d}\theta \tag{4}$$

which, when logarithmized, gives

$$\mathrm{LME}(m) = \log p(y|m) = \log \int p(y|\theta, m) \, p(\theta|m) \, \mathrm{d}\theta \; . \tag{5}$$

2) The second expression can be derived from Bayes' theorem (→ Proof I/2.1.1) which makes a statement about the posterior distribution (→ Definition I/2.2.6):

$$p(\theta|y,m) = \frac{p(y|\theta,m)\,p(\theta|m)}{p(y|m)} \; . \tag{6}$$

Rearranging for $p(y|m)$ and logarithmizing, we have:

$$\begin{split} \mathrm{LME}(m) = \log p(y|m) &= \log \frac{p(y|\theta,m)\,p(\theta|m)}{p(\theta|y,m)} \\ &= \log p(y|\theta,m) + \log p(\theta|m) - \log p(\theta|y,m) \; . \end{split} \tag{7}$$

Sources:
- original work

Metadata: ID: P13 | shortcut: lme-der | author: JoramSoch | date: 2020-01-06, 21:27.

3.1.3 Partition into accuracy and complexity

Theorem: The log model evidence (→ Definition IV/3.1.1) can be partitioned into accuracy and complexity

$$\mathrm{LME}(m) = \mathrm{Acc}(m) - \mathrm{Com}(m) \tag{1}$$

where the accuracy term is the posterior expectation of the log-likelihood function (→ Definition I/2.2.2)

$$\mathrm{Acc}(m) = \left\langle p(y|\theta,m) \right\rangle_{p(\theta|y,m)} \tag{2}$$

and the complexity penalty is the Kullback-Leibler divergence (→ Definition "kl") of posterior (→ Definition I/2.2.6) from prior (→ Definition I/2.2.3)

$$\mathrm{Com}(m) = \mathrm{KL}\left[p(\theta|y,m)\,||\,p(\theta|m)\right] \; . \tag{3}$$

Proof: We consider Bayesian inference on data y using model m with parameters θ. Then, Bayes' theorem (→ Proof I/2.1.1) makes a statement about the posterior distribution, i.e. the probability of parameters, given the data and the model:

$$p(\theta|y,m) = \frac{p(y|\theta,m)\,p(\theta|m)}{p(y|m)} \; . \tag{4}$$

Rearranging this for the model evidence (→ Proof IV/3.1.2), we have:

$$p(y|m) = \frac{p(y|\theta,m)\,p(\theta|m)}{p(\theta|y,m)} \; . \tag{5}$$

Logarthmizing both sides of the equation, we obtain:

$$\log p(y|m) = \log p(y|\theta,m) - \log \frac{p(\theta|y,m)}{p(\theta|m)} \; . \tag{6}$$

Now taking the expectation over the posterior distribution yields:

$$\log p(y|m) = \int p(\theta|y,m) \log p(y|\theta,m) \, \mathrm{d}\theta - \int p(\theta|y,m) \log \frac{p(\theta|y,m)}{p(\theta|m)} \, \mathrm{d}\theta \; . \tag{7}$$

By definition, the left-hand side is the log model evidence and the terms on the right-hand side correspond to the posterior expectation of the log-likelihood function and the Kullback-Leibler divergence of posterior from prior

$$\mathrm{LME}(m) = \left\langle p(y|\theta,m) \right\rangle_{p(\theta|y,m)} - \mathrm{KL}\left[p(\theta|y,m) \, || \, p(\theta|m) \right] \tag{8}$$

which proofs the partition given by (1).

Sources:

- Penny et al. (2007): "Bayesian Comparison of Spatially Regularised General Linear Models"; in: *Human Brain Mapping*, vol. 28, pp. 275–293; URL: https://onlinelibrary.wiley.com/doi/full/10.1002/hbm.20327; DOI: 10.1002/hbm.20327.
- Soch et al. (2016): "How to avoid mismodelling in GLM-based fMRI data analysis: cross-validated Bayesian model selection"; in: *NeuroImage*, vol. 141, pp. 469–489; URL: https://www.sciencedirect.com/science/article/pii/S1053811916303615; DOI: 10.1016/j.neuroimage.2016.07.047.

Metadata: ID: P3 | shortcut: lme-anc | author: JoramSoch | date: 2019-09-27, 16:13.

3.2 Log-evidence derivatives

3.2.1 Log Bayes factor in terms of log model evidences

Theorem: Let m_1 and m_2 be two statistical models with log model evidences (→ Definition IV/3.1.1) $\mathrm{LME}(m_1)$ and $\mathrm{LME}(m_2)$. Then, the log Bayes factor (→ Definition "lbf") in favor of model m_1 and against model m_2 is the difference of the log model evidences:

$$\mathrm{LBF}_{1,2} = \mathrm{LME}(m_1) - \mathrm{LME}(m_2) \; . \tag{1}$$

Proof: The log Bayes factor (→ Definition "lbf") (LBF) is defined as the logarithm of the Bayes factor (→ Definition "bf") (BF) which is defined as the posterior odds ratio when both models are equally likely apriori:

$$\begin{split} \mathrm{LBF}_{1,2} &= \log \mathrm{BF}_{1,2} \\ &= \log \frac{p(m_1|y)}{p(m_2|y)} \; . \end{split} \tag{2}$$

Plugging in the posterior odds ratio according to Bayes' rule (→ Proof I/2.1.2), we have

$$\mathrm{LBF}_{1,2} = \log \left[\frac{p(y|m_1)}{p(y|m_2)} \cdot \frac{p(m_1)}{p(m_2)} \right] \; . \tag{3}$$

When both models are equally likely apriori, the prior odds ratio is one, such that

$$\mathrm{LBF}_{1,2} = \log \frac{p(y|m_1)}{p(y|m_2)} \; . \tag{4}$$

Resolving the logarithm and applying the definition of the log model evidence (→ Definition IV/3.1.1), we finally have:

$$
\begin{aligned}
\mathrm{LBF}_{1,2} &= \log p(y|m_1) - \log p(y|m_2) \\
&= \mathrm{LME}(m_1) - \mathrm{LME}(m_2) \; .
\end{aligned} \tag{5}
$$

Sources:

- Soch J, Allefeld C (2018): "MACS – a new SPM toolbox for model assessment, comparison and selection"; in: *Journal of Neuroscience Methods*, vol. 306, pp. 19-31, eq. 18; URL: https://www.sciencedirect.com/science/article/pii/S0165027018301468; DOI: 10.1016/j.jneumeth.2018.05.017.

Metadata: ID: P64 | shortcut: lbf-lme | author: JoramSoch | date: 2020-02-27, 20:51.

3.2.2 Log family evidences in terms of log model evidences

Theorem: Let $m_1, \ldots, m_M$ be M statistical models with log model evidences (→ Definition IV/3.1.1) $\mathrm{LME}(m_1), \ldots, \mathrm{LME}(m_M)$ and belonging to F mutually exclusive model families $f_1, \ldots, f_F$. Then, the log family evidences (→ Definition "lfe") are given by:

$$
\mathrm{LFE}(f_j) = \log \sum_{m_i \in f_j} \left[\exp[\mathrm{LME}(m_i)] \cdot p(m_i|f_j) \right] , \quad j = 1, \ldots, F, \tag{1}
$$

where $p(m_i|f_j)$ are within-family prior model probabilities.

Proof: Let us consider the (unlogarithmized) family evidence $p(y|f_j)$. According to the law of marginal probability (→ Definition "prob-marg"), this conditional probability is given by

$$
p(y|f_j) = \sum_{m_i \in f_j} \left[p(y|m_i, f_j) \cdot p(m_i|f_j) \right] \; . \tag{2}
$$

Because model families are mutually exclusive, it holds that $p(y|m_i, f_j) = p(y|m_i)$, such that

$$
p(y|f_j) = \sum_{m_i \in f_j} \left[p(y|m_i) \cdot p(m_i|f_j) \right] \; . \tag{3}
$$

Logarithmizing transforms the family evidence $p(y|f_j)$ into the log family evidence $\mathrm{LFE}(f_j)$:

$$
\mathrm{LFE}(f_j) = \log \sum_{m_i \in f_j} \left[p(y|m_i) \cdot p(m_i|f_j) \right] \; . \tag{4}
$$

The definition of the log model evidence (→ Definition IV/3.1.1)

$$
\mathrm{LME}(m) = \log p(y|m) \tag{5}
$$

can be exponentiated to then read

$$
\exp\left[\mathrm{LME}(m)\right] = p(y|m) \tag{6}
$$

and applying (6) to (4), we finally have:

$$\mathrm{LFE}(f_j) = \log \sum_{m_i \in f_j} \left[\exp[\mathrm{LME}(m_i)] \cdot p(m_i|f_j) \right] \; . \tag{7}$$

Sources:

- Soch J, Allefeld C (2018): "MACS – a new SPM toolbox for model assessment, comparison and selection"; in: *Journal of Neuroscience Methods*, vol. 306, pp. 19-31, eq. 16; URL: https://www.sciencedirect.com/science/article/pii/S0165027018301468; DOI: 10.1016/j.jneumeth.2018.05.017.

Metadata: ID: P65 | shortcut: lfe-lme | author: JoramSoch | date: 2020-02-27, 21:16.

3.2.3 Posterior model probability in terms of log Bayes factor

Theorem: Let m_1 and m_2 be two statistical models log Bayes factor (→ Definition "lbf") $\mathrm{LBF}_{1,2}$ in favor of model m_1 and against model m_2. Then, if both models are equally likely apriori, the posterior model probability (→ Definition "pmp") of m_1 is

$$p(m_1|y) = \frac{\exp(\mathrm{LBF}_{1,2})}{\exp(\mathrm{LBF}_{1,2}) + 1} \; . \tag{1}$$

Proof: From Bayes' rule (→ Proof I/2.1.2), the posterior odds ratio is

$$\frac{p(m_1|y)}{p(m_2|y)} = \frac{p(y|m_1)}{p(y|m_2)} \cdot \frac{p(m_1)}{p(m_2)} \; . \tag{2}$$

When both models are equally likely apriori, the prior odds ratio is one, such that

$$\frac{p(m_1|y)}{p(m_2|y)} = \frac{p(y|m_1)}{p(y|m_2)} \; . \tag{3}$$

Now the right-hand side corresponds to the Bayes factor (→ Definition "bf"), therefore

$$\frac{p(m_1|y)}{1 - p(m_1|y)} = \mathrm{BF}_{1,2} \; . \tag{4}$$

Because the two models are collectively exhaustive, we have

$$\frac{p(m_1|y)}{1 - p(m_1|y)} = \mathrm{BF}_{1,2} \; . \tag{5}$$

Now rearranging for the posterior probability (→ Definition "pmp"), this gives

$$p(m_1|y) = \frac{\mathrm{BF}_{1,2}}{\mathrm{BF}_{1,2} + 1} \; . \tag{6}$$

Because the log Bayes factor is the logarithm of the Bayes factor (→ Proof IV/3.2.1), we finally have

$$p(m_1|y) = \frac{\exp(\mathrm{LBF}_{1,2})}{\exp(\mathrm{LBF}_{1,2}) + 1} \; . \tag{7}$$

Sources:

- Soch J, Allefeld C (2018): "MACS – a new SPM toolbox for model assessment, comparison and selection"; in: *Journal of Neuroscience Methods*, vol. 306, pp. 19-31, eq. 21; URL: https://www.sciencedirect.com/science/article/pii/S0165027018301468; DOI: 10.1016/j.jneumeth.2018.05.017.

Metadata: ID: P73 | shortcut: pmp-lbf | author: JoramSoch | date: 2020-03-03, 12:27.

3.2.4 Posterior model probabilities in terms of Bayes factors

Theorem: Let $m_0, m_1, \ldots, m_M$ be $M+1$ statistical models with model evidences (→ Definition IV/3.1.1) $p(y|m_0), p(y|m_1), \ldots, p(y|m_M)$. Then, the posterior model probabilities (→ Definition "pmp") of the models $m_1, \ldots, m_M$ are given by:

$$p(m_i|y) = \frac{\mathrm{BF}_{i,0} \cdot \alpha_i}{\sum_{j=1}^M \mathrm{BF}_{j,0} \cdot \alpha_j}, \quad i = 1, \ldots, M \; , \tag{1}$$

where $\mathrm{BF}_{i,0}$ is the Bayes factor (→ Definition "bf") comparing model m_i with m_0 and α_i is the prior odds ratio of model m_i against m_0.

Proof: Define the Bayes factor

$$\mathrm{BF}_{i,0} = \frac{p(y|m_i)}{p(y|m_0)} \tag{2}$$

and prior odds ratio of m_i against m_0

$$\alpha_i = \frac{p(m_i)}{p(m_0)} \; . \tag{3}$$

From Bayes' theorem (→ Proof I/2.1.1), the posterior probability of m_i follows as

$$p(m_i|y) = \frac{p(y|m_i) \cdot p(m_i)}{\sum_{j=1}^M p(y|m_j) \cdot p(m_j)} \; . \tag{4}$$

Now applying (2) and (3) to (4), we have

$$\begin{split} p(m_i|y) &= \frac{\mathrm{BF}_{i,0} \, p(y|m_0) \cdot \alpha_i \, p(m_0)}{\sum_{j=1}^M \mathrm{BF}_{j,0} \, p(y|m_0) \cdot \alpha_j \, p(m_0)} \\ &= \frac{[p(y|m_0) \, p(m_0)] \, \mathrm{BF}_{i,0} \cdot \alpha_i}{[p(y|m_0) \, p(m_0)] \sum_{j=1}^M \mathrm{BF}_{j,0} \cdot \alpha_j} \; , \end{split} \tag{5}$$

such that

$$p(m_i|y) = \frac{\mathrm{BF}_{i,0} \cdot \alpha_i}{\sum_{j=1}^M \mathrm{BF}_{j,0} \cdot \alpha_j} \; . \tag{6}$$

Sources:

- Hoeting JA, Madigan D, Raftery AE, Volinsky CT (1999): "Bayesian Model Averaging: A Tutorial"; in: *Statistical Science*, vol. 14, no. 4, pp. 382–417, eq. 9; URL: https://projecteuclid.org/euclid.ss/1009212519; DOI: 10.1214/ss/1009212519.

Metadata: ID: P74 | shortcut: pmp-bf | author: JoramSoch | date: 2020-03-03, 13:13.

3.2.5 Posterior model probabilities in terms of log model evidences

Theorem: Let $m_1, \ldots, m_M$ be M statistical models with log model evidences (→ Definition IV/3.1.1) $\mathrm{LME}(m_1), \ldots, \mathrm{LME}(m_M)$. Then, the posterior model probabilities (→ Definition "pmp") are given by:

$$p(m_i|y) = \frac{\exp[\mathrm{LME}(m_i)]\, p(m_i)}{\sum_{j=1}^{M} \exp[\mathrm{LME}(m_j)]\, p(m_j)}, \quad i = 1, \ldots, M \; , \tag{1}$$

where $p(m_i)$ are prior model probabilities.

Proof: From Bayes' theorem (→ Proof I/2.1.1), the posterior model probability (→ Definition "pmp") of model m_i can be derived as

$$p(m_i|y) = \frac{p(y|m_i)\, p(m_i)}{p(y)} \; . \tag{2}$$

Using the law of marginal probability (→ Definition "prob-marg"), the denominator can be written as

$$p(m_i|y) = \frac{p(y|m_i)\, p(m_i)}{\sum_{j=1}^{M} p(y|m_j)\, p(m_j)} \; . \tag{3}$$

The definition of the log model evidence (→ Definition IV/3.1.1)

$$\mathrm{LME}(m) = \log p(y|m) \tag{4}$$

can be exponentiated to then read

$$\exp\left[\mathrm{LME}(m)\right] = p(y|m) \tag{5}$$

and applying (5) to (3), we finally have:

$$p(m_i|y) = \frac{\exp[\mathrm{LME}(m_i)]\, p(m_i)}{\sum_{j=1}^{M} \exp[\mathrm{LME}(m_j)]\, p(m_j)} \; . \tag{6}$$

Sources:

- Soch J, Allefeld C (2018): "MACS – a new SPM toolbox for model assessment, comparison and selection"; in: *Journal of Neuroscience Methods*, vol. 306, pp. 19-31, eq. 23; URL: https://www.sciencedirect.com/science/article/pii/S0165027018301468; DOI: 10.1016/j.jneumeth.2018.05.017.

Metadata: ID: P66 | shortcut: pmp-lme | author: JoramSoch | date: 2020-02-27, 21:33.

3.2.6 Bayesian model averaging in terms of log model evidences

Theorem: Let $m_1, \ldots, m_M$ be M statistical models describing the same measured data y with log model evidences (→ Definition IV/3.1.1) $\mathrm{LME}(m_1), \ldots, \mathrm{LME}(m_M)$ and shared model parameters θ. Then, Bayesian model averaging (BMA) determines the following posterior distribution over θ:

$$p(\theta|y) = \sum_{i=1}^{M} p(\theta|m_i, y) \cdot \frac{\exp[\mathrm{LME}(m_i)]\, p(m_i)}{\sum_{j=1}^{M} \exp[\mathrm{LME}(m_j)]\, p(m_j)} \; , \tag{1}$$

where $p(\theta|m_i, y)$ is the posterior distributions over θ obtained using m_i.

Proof: According to the law of marginal probability (→ Definition "prob-marg"), the probability of the shared parameters θ conditional on the measured data y can be obtained by marginalizing over the discrete variable model m:

$$p(\theta|y) = \sum_{i=1}^{M} p(\theta|m_i, y) \cdot p(m_i|y) \; , \tag{2}$$

where $p(m_i|y)$ is the posterior probability (→ Definition "pmp") of the i-th model. One can express posterior model probabilities in terms of log model evidences (→ Proof IV/3.2.5) as

$$p(m_i|y) = \frac{\exp[\mathrm{LME}(m_i)]\, p(m_i)}{\sum_{j=1}^{M} \exp[\mathrm{LME}(m_j)]\, p(m_j)} \tag{3}$$

and by plugging (3) into (2), one arrives at (1).

Sources:

- Soch J, Allefeld C (2018): "MACS – a new SPM toolbox for model assessment, comparison and selection"; in: *Journal of Neuroscience Methods*, vol. 306, pp. 19-31, eq. 25; URL: https://www.sciencedirect.com/science/article/pii/S0165027018301468; DOI: 10.1016/j.jneumeth.2018.05.017.

Metadata: ID: P67 | shortcut: bma-lme | author: JoramSoch | date: 2020-02-27, 21:58.