

The Book of Statistical Proofs

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

2020-03-24, 07:25

Contents

I	General Theorems	1
1	Probability theory	2
1.1	Probability distributions	2
1.1.1	<i>Probability mass function</i>	2
1.1.2	<i>Probability density function</i>	2
1.1.3	<i>Cumulative distribution function</i>	2
1.1.4	<i>Quantile function</i>	3
1.1.5	<i>Moment-generating function</i>	3
1.2	Expected value	4
1.2.1	<i>Definition</i>	4
1.2.2	Non-negativity	4
1.2.3	Linearity	5
1.2.4	Monotonicity	7
1.2.5	(Non-)Multiplicativity	7
1.3	Variance	9
1.3.1	<i>Definition</i>	9
2	Bayesian statistics	10
2.1	Bayesian inference	10
2.1.1	Bayes' theorem	10
2.1.2	Bayes' rule	10
2.2	Probabilistic modeling	11
2.2.1	<i>Generative model</i>	11
2.2.2	<i>Likelihood function</i>	11
2.2.3	<i>Prior distribution</i>	12
2.2.4	<i>Full probability model</i>	12
2.2.5	<i>Joint likelihood</i>	12
2.2.6	<i>Posterior distribution</i>	13
2.2.7	<i>Marginal likelihood</i>	13
3	Estimation theory	14
3.1	Point estimates	14
3.1.1	Partition of the mean squared error into bias and variance	14
3.2	Interval estimates	15
3.2.1	Construction of confidence intervals using Wilks' theorem	15
4	Information theory	17
4.1	Shannon entropy	17
4.1.1	<i>Definition</i>	17
4.1.2	Non-negativity	17
4.1.3	<i>Conditional entropy</i>	18

4.1.4	<i>Joint entropy</i>	18
4.2	Differential entropy	18
4.2.1	<i>Definition</i>	18
4.2.2	Negativity	19
4.2.3	<i>Conditional differential entropy</i>	20
4.2.4	<i>Joint differential entropy</i>	20
4.3	Discrete mutual information	20
4.3.1	<i>Definition</i>	20
4.3.2	Relation to marginal and conditional entropy	21
4.3.3	Relation to marginal and joint entropy	22
4.3.4	Relation to joint and conditional entropy	23
4.4	Continuous mutual information	24
4.4.1	<i>Definition</i>	24
4.4.2	Relation to marginal and conditional differential entropy . .	25
4.4.3	Relation to marginal and joint differential entropy	26
4.4.4	Relation to joint and conditional differential entropy	27
II	Probability Distributions	29
1	Univariate discrete distributions	30
1.1	Bernoulli distribution	30
1.1.1	<i>Definition</i>	30
1.1.2	Mean	30
1.2	Binomial distribution	31
1.2.1	<i>Definition</i>	31
1.2.2	Mean	31
2	Multivariate discrete distributions	33
2.1	Categorical distribution	33
2.1.1	<i>Definition</i>	33
2.1.2	Mean	33
2.2	Multinomial distribution	34
2.2.1	<i>Definition</i>	34
2.2.2	Mean	34
3	Univariate continuous distributions	36
3.1	Continuous uniform distribution	36
3.1.1	<i>Definition</i>	36
3.1.2	Probability density function	36
3.1.3	Cumulative distribution function	37
3.1.4	Quantile function	38
3.1.5	Mean	39
3.1.6	Median	39
3.1.7	Mode	40
3.2	Normal distribution	41
3.2.1	<i>Definition</i>	41
3.2.2	Probability density function	41
3.2.3	Cumulative distribution function	42
3.2.4	Cumulative distribution function without error function . .	43
3.2.5	Quantile function	45
3.2.6	Mean	46

3.2.7	Median	47
3.2.8	Mode	48
3.2.9	Variance	49
3.2.10	Moment-generating function	51
3.3	Gamma distribution	52
3.3.1	<i>Definition</i>	52
3.3.2	Probability density function	53
3.4	Exponential distribution	53
3.4.1	<i>Definition</i>	53
3.4.2	Special case of gamma distribution	54
3.4.3	Probability density function	54
3.4.4	Cumulative distribution function	55
3.4.5	Quantile function	55
3.4.6	Mean	56
3.4.7	Median	57
3.4.8	Mode	58
4	Multivariate continuous distributions	60
4.1	Multivariate normal distribution	60
4.1.1	<i>Definition</i>	60
4.1.2	Probability density function	60
4.1.3	Linear transformation theorem	60
4.1.4	Marginal distributions	61
4.1.5	Conditional distributions	62
4.2	Normal-gamma distribution	65
4.2.1	<i>Definition</i>	65
4.2.2	Probability density function	66
4.2.3	Kullback-Leibler divergence	67
4.2.4	Marginal distributions	68
5	Matrix-variate continuous distributions	72
5.1	Matrix-normal distribution	72
5.1.1	<i>Definition</i>	72
5.1.2	Probability density function	72
5.1.3	Equivalence to multivariate normal distribution	73
5.2	Wishart distribution	74
5.2.1	<i>Definition</i>	74
III Statistical Models		75
1	Normal data	76
1.1	Multiple linear regression	76
1.1.1	<i>Definition</i>	76
1.1.2	Ordinary least squares (1)	77
1.1.3	Ordinary least squares (2)	77
1.1.4	<i>Total sum of squares</i>	78
1.1.5	<i>Explained sum of squares</i>	79
1.1.6	<i>Residual sum of squares</i>	79
1.1.7	Total, explained and residual sum of squares	79
1.1.8	Estimation, projection and residual-forming matrix	81
1.1.9	Weighted least squares	82

	1.1.10	Maximum likelihood estimation	83
1.2		Bayesian linear regression	85
	1.2.1	Conjugate prior distribution	85
	1.2.2	Posterior distribution	87
	1.2.3	Log model evidence	89
1.3		General linear model	91
	1.3.1	<i>Definition</i>	91
	1.3.2	Maximum likelihood estimation	92
2		Poisson data	95
2.1		Poisson-distributed data	95
	2.1.1	<i>Definition</i>	95
	2.1.2	Maximum likelihood estimation	95
2.2		Poisson distribution with exposure values	97
	2.2.1	<i>Definition</i>	97
	2.2.2	Conjugate prior distribution	97
	2.2.3	Posterior distribution	98
	2.2.4	Log model evidence	100
3		Probability data	103
3.1		Beta-distributed data	103
	3.1.1	Method of moments	103
3.2		Logistic regression	105
	3.2.1	Log-odds and probability	105
4		Categorical data	106
4.1		Binomial observations	106
	4.1.1	Conjugate prior distribution	106
	4.1.2	Posterior distribution	107
	4.1.3	Log model evidence	108
4.2		Multinomial observations	109
	4.2.1	Conjugate prior distribution	109
	4.2.2	Posterior distribution	110
	4.2.3	Log model evidence	112
IV Model Selection			115
1		Goodness-of-fit measures	116
1.1		Residual variance	116
	1.1.1	<i>Definition</i>	116
	1.1.2	Maximum likelihood estimator is biased	116
	1.1.3	Construction of unbiased estimator	118
1.2		R-squared	119
	1.2.1	<i>Definition</i>	119
	1.2.2	Derivation of R^2 and adjusted R^2	120
	1.2.3	Relationship to maximum log-likelihood	121
1.3		Signal-to-noise ratio	123
	1.3.1	<i>Definition</i>	123
2		Classical information criteria	124
2.1		Akaike information criterion	124
	2.1.1	<i>Definition</i>	124
2.2		Bayesian information criterion	124

	2.2.1	<i>Definition</i>	124
	2.2.2	Derivation	124
	2.3	Deviance information criterion	126
	2.3.1	<i>Definition</i>	126
3		Bayesian model selection	127
	3.1	Log model evidence	127
	3.1.1	<i>Definition</i>	127
	3.1.2	Derivation	127
	3.1.3	Partition into accuracy and complexity	128
	3.2	Log-evidence derivatives	129
	3.2.1	Log Bayes factor in terms of log model evidences	129
	3.2.2	Log family evidences in terms of log model evidences	130
	3.2.3	Posterior model probability in terms of log Bayes factor	131
	3.2.4	Posterior model probabilities in terms of Bayes factors	132
	3.2.5	Posterior model probabilities in terms of log model evidences	133
	3.2.6	Bayesian model averaging in terms of log model evidences	133

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 (\rightarrow Definition “rvar”) with possible outcomes \mathcal{X} . Then, $f_X(x) : \mathbb{R} \rightarrow [0, 1]$ is the probability mass function of X , if

$$\Pr(X = x) = f_X(x) \quad (1)$$

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

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

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 (\rightarrow Definition “rvar”) with possible outcomes \mathcal{X} . Then, $f_X(x) : \mathbb{R} \rightarrow \mathbb{R}$ is the probability density function of X , if

$$f_X(x) \geq 0 \quad (1)$$

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

$$\Pr(X \in A) = \int_A f_X(x) dx \quad (2)$$

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

$$\int_{\mathcal{X}} f_X(x) dx = 1. \quad (3)$$

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 (\rightarrow Definition “rvar”) with possible outcomes \mathcal{X} and the probability mass function (\rightarrow Definition I/1.1.1) $f_X(x)$. Then, the function $F_X(x) : \mathbb{R} \rightarrow [0, 1]$ with

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

is the cumulative distribution function of X .

2) Let X be a scalar continuous random variable (\rightarrow Definition “rvar”) with the probability density function (\rightarrow Definition I/1.1.2) $f_X(x)$. Then, the function $F_X(x) : \mathbb{R} \rightarrow [0, 1]$ with

$$F_X(x) = \int_{-\infty}^x f_X(z) \, dx \quad (2)$$

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 (\rightarrow Definition “rvar”) with the cumulative distribution function (\rightarrow Definition I/1.1.3) (CDF) $F_X(x)$. Then, the function $Q_X(p) : [0, 1] \rightarrow \mathbb{R}$ which is the inverse CDF

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

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 $F_X(x) = p$:

$$Q_X(p) = \min \{x \in \mathbb{R} \mid F_X(x) = p\} . \quad (2)$$

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 (\rightarrow Definition “rvar”) $X \in \mathbb{R}$ is

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

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

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

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 (\rightarrow Definition “rvar”) X with domain \mathcal{X} is

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

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

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

$$E(X) = \int_{\mathcal{X}} x \cdot f_X(x) dx \quad (2)$$

where $f_X(x)$ is the probability density function (\rightarrow 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 (\rightarrow Definition “rvar”) is strictly non-negative, its expected value (\rightarrow Definition I/1.2.1) is also non-negative, i.e.

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

Proof:

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

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

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 (\rightarrow Definition I/1.1.2) is always non-negative, the integrand in

$$E(X) = \int_{\mathcal{X}} x \cdot f_X(x) dx \quad (3)$$

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 (\rightarrow Definition I/1.2.1) is a linear operator, i.e.

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

for random variables (\rightarrow Definition “rvar”) X and Y and a constant a .

Proof:

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

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

and the law of marginal probability (\rightarrow Definition “prob-marg”) states that

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

Applying this, we have

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

as well as

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

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

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

and the law of marginal probability (\rightarrow Definition “prob-marg”) states that

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

Applying this, we have

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

as well as

$$\begin{aligned}
E(a X) &= \int_{\mathcal{X}} a x \cdot f_X(x) dx \\
&= a \int_{\mathcal{X}} x \cdot f_X(x) dx \\
&\stackrel{(6)}{=} a E(X) .
\end{aligned} \tag{9}$$

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 (\rightarrow Definition I/1.2.1) is monotonic, i.e.

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

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

$$E(Z) = E(Y - X) = E(Y) - E(X). \quad (2)$$

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

$$Z \geq 0 \quad \Rightarrow \quad E(Z) \geq 0. \quad (3)$$

Together with (2), this yields

$$E(Y) - E(X) \geq 0. \quad (4)$$

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-)Multiplicativity

Theorem:

1) If two random variables (\rightarrow Definition “rvar”) X and Y are independent (\rightarrow Definition “ind”), the expected value (\rightarrow Definition I/1.2.1) is multiplicative, i.e.

$$E(XY) = E(X)E(Y). \quad (1)$$

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

$$E(XY) \neq E(X)E(Y). \quad (2)$$

Proof:

1) If X and Y are independent (\rightarrow Definition “ind”), it holds that

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

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

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

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

$$\begin{aligned}
E(XY) &= \int_{\mathcal{X}} \int_{\mathcal{Y}} (x \cdot y) \cdot f_{X,Y}(x, y) \, dy \, dx \\
&\stackrel{(3)}{=} \int_{\mathcal{X}} \int_{\mathcal{Y}} (x \cdot y) \cdot (f_X(x) \cdot f_Y(y)) \, dy \, dx \\
&= \int_{\mathcal{X}} x \cdot f_X(x) \int_{\mathcal{Y}} y \cdot f_Y(y) \, dy \, dx \\
&= \int_{\mathcal{X}} x \cdot f_X(x) \cdot E(Y) \, dx \\
&= E(X) E(Y) .
\end{aligned} \tag{5}$$

2) Let X and Y be Bernoulli random variables (\rightarrow Definition II/1.1.1) with the following joint probability (\rightarrow Definition “prob-joint”) mass function (\rightarrow 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

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

and the expected value of their product is

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

while the product of their expected values is

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

and thus,

$$E(XY) \neq E(X) E(Y) . \tag{11}$$

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 (\rightarrow Definition “rvar”) X is defined as the expected value (\rightarrow Definition I/1.2.1) of the squared deviation from its expected value (\rightarrow Definition I/1.2.1):

$$\text{Var}(X) = E[(X - E(X))^2] . \tag{1}$$

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 (\rightarrow 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

$$p(A|B) = \frac{p(B|A) p(A)}{p(B)} . \quad (1)$$

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

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

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:

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

Both equations can be rearranged for the joint probability

$$p(A|B) p(B) \stackrel{(2)}{=} p(A, B) \stackrel{(3)}{=} p(B|A) p(A) \quad (4)$$

from which Bayes' theorem can be directly derived:

$$p(A|B) \stackrel{(4)}{=} \frac{p(B|A) p(A)}{p(B)} . \quad (5)$$

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 (\rightarrow Definition “rvar”) where A_1 and A_2 are mutually exclusive. Then, Bayes' rule states that the posterior odds (\rightarrow Definition “post-odd”) are equal to the Bayes factor (\rightarrow Definition “bf”) times the prior odds (\rightarrow Definition “prior-odd”), i.e.

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

Proof: Using Bayes' theorem (\rightarrow Proof I/2.1.1), the conditional probabilities (\rightarrow Definition “cp”) on the left are given by

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

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

Dividing the two conditional probabilities by each other

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

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 :

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

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 (\rightarrow Definition I/2.2.1) m describing measured data y using model parameters θ . Then, the probability density function (\rightarrow Definition I/1.1.2) of the distribution of y given θ is called the likelihood function of m :

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

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:

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

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

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

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 (\rightarrow Definition I/2.2.1) for y and a prior distribution (\rightarrow Definition I/2.2.3) on θ is called a full probability model m :

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

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 (\rightarrow Definition I/2.2.1) m describing measured data y using model parameters θ and a prior distribution (\rightarrow Definition I/2.2.3) on θ . Then, the joint probability (\rightarrow Definition “prob-joint”) density function (\rightarrow Definition I/1.1.2) of y and θ is called the joint likelihood:

$$p(y, \theta|m) = p(y|\theta, m) p(\theta|m) . \quad (1)$$

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:

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

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

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

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

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

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 (\rightarrow Definition “mse”) can be partitioned into variance and squared bias

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

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

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

and the bias (\rightarrow Definition “bias”) is given by

$$\text{Bias}(\hat{\theta}, \theta) = \left(\mathbb{E}_{\hat{\theta}}(\hat{\theta}) - \theta \right) . \quad (3)$$

Proof: The mean squared error (MSE) is defined as (\rightarrow Definition “mse”) the expected value (\rightarrow 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 $\hat{\theta}$:

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

This formula can be evaluated in the following way:

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

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

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

This proves 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 (\rightarrow Definition I/2.2.1) for measured data y with model parameters θ , consisting of a parameter of interest ϕ and nuisance parameters λ :

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

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

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

Then, an asymptotic confidence interval (\rightarrow Definition “ci”) for θ is given by

$$\text{CI}_{1-\alpha}(\hat{\phi}) = \left\{ \phi \mid \log p(y|\phi, \hat{\lambda}) \geq \log p(y|\hat{\phi}, \hat{\lambda}) - \frac{1}{2} \chi_{1,1-\alpha}^2 \right\} \quad (3)$$

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

Proof: The confidence interval (\rightarrow Definition “ci”) is defined as the interval that, under infinitely repeated random experiments (\rightarrow Definition “rexp”), contains the true parameter value with a certain probability.

Let us define the likelihood ratio (\rightarrow Definition “lr”)

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

and compute the log-likelihood ratio (\rightarrow Definition “llr”)

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

[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 (\rightarrow Definition “chi2”), if the null hypothesis is true:

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

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

$$-2 \log \Lambda(\phi) \sim \chi_1^2. \quad (7)$$

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

$$-2 \log \Lambda(\phi) \leq \chi_{1,1-\alpha}^2. \quad (8)$$

Applying (5) and rearranging, we can evaluate

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

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 (\rightarrow Definition “rvar”) with possible outcomes x_i , $i = 1, \dots, k$ and the (observed or assumed) probability mass function (\rightarrow Definition I/1.1.1) $p(x) = f_X(x)$. Then, the entropy (also referred to as “Shannon entropy”) of X is defined as

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

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 (\rightarrow Definition “rvar”) is a non-negative number:

$$H(X) \geq 0 . \quad (1)$$

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

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

The minus sign can be moved into the sum:

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

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

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

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

$$\lim_{p \rightarrow 0} [p \log_b(p)] = 0 . \quad (5)$$

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 (\rightarrow Definition “rvar”) with possible outcomes \mathcal{X} and \mathcal{Y} and probability mass functions (\rightarrow 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

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

where $H(Y|X = x)$ is the (marginal) entropy (\rightarrow 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 (\rightarrow Definition “rvar”) with possible outcomes \mathcal{X} and \mathcal{Y} and joint probability (\rightarrow Definition “prob-joint”) mass function (\rightarrow Definition I/1.1.1) $p(x, y)$. Then, the joint entropy of X and Y is defined as

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

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 (\rightarrow Definition “rvar”) with possible outcomes \mathcal{X} and the (estimated or assumed) probability density function (\rightarrow 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

$$h(X) = - \int_{\mathcal{X}} p(x) \log_b p(x) dx \quad (1)$$

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 (\rightarrow Proof I/4.1.2), the differential entropy (\rightarrow Definition I/4.2.1) can become negative.

Proof: Let X be a random variable (\rightarrow Definition “rvar”) following a continuous uniform distribution (\rightarrow Definition II/3.1.1) with minimum 0 and maximum $1/2$:

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

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

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

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

$$\begin{aligned} h(X) &= - \int_{\mathcal{X}} f_X(x) \log_b f_X(x) dx \\ &= - \int_0^{\frac{1}{2}} 2 \log_b(2) dx \\ &= - \log_b(2) \int_0^{\frac{1}{2}} 2 dx \\ &= - \log_b(2) [2x]_0^{\frac{1}{2}} \\ &= - \log_b(2) \end{aligned} \quad (3)$$

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.

4.2.3 Conditional differential entropy

Definition: Let X and Y be continuous random variables (\rightarrow Definition “rvar”) with possible outcomes \mathcal{X} and \mathcal{Y} and probability density functions (\rightarrow Definition I/1.1.2) $p(x)$ and $p(y)$. Then, the conditional differential entropy of Y given X or, differential entropy of Y conditioned on X , is defined as

$$h(Y|X) = \int_{x \in \mathcal{X}} p(x) \cdot h(Y|X = x) \quad (1)$$

where $h(Y|X = x)$ is the (marginal) differential entropy (\rightarrow Definition I/4.2.1) of Y , evaluated at x .

Sources:

- original work

Metadata: ID: D34 | shortcut: dent-cond | author: JoramSoch | date: 2020-03-21, 12:27.

4.2.4 Joint differential entropy

Definition: Let X and Y be continuous random variables (\rightarrow Definition “rvar”) with possible outcomes \mathcal{X} and \mathcal{Y} and joint probability (\rightarrow Definition “prob-joint”) density function (\rightarrow Definition I/1.1.2) $p(x, y)$. Then, the joint differential entropy of X and Y is defined as

$$h(X, Y) = - \int_{x \in \mathcal{X}} \int_{y \in \mathcal{Y}} p(x, y) \cdot \log_b p(x, y) \, dy \, dx \quad (1)$$

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

Sources:

- original work

Metadata: ID: D35 | shortcut: dent-joint | author: JoramSoch | date: 2020-03-21, 12:37.

4.3 Discrete mutual information

4.3.1 Definition

Definition:

1) The mutual information of two discrete random variables (\rightarrow Definition “rvar”) X and Y is defined as

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

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

2) The mutual information of two continuous random variables (\rightarrow Definition “rvar”) X and Y is defined as

$$I(X, Y) = - \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x, y) \cdot \log \frac{p(x, y)}{p(x) \cdot p(y)} \, dy \, dx \quad (2)$$

where $p(x)$ and $p(y)$ are the probability density functions (\rightarrow Definition I/1.1.1) of X and Y and $p(x, y)$ is the joint probability (\rightarrow 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 (\rightarrow Definition “rvar”) with the joint probability (\rightarrow Definition “prob-joint”) $p(x, y)$ for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Then, the mutual information (\rightarrow Definition I/4.4.1) of X and Y can be expressed as

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

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

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

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

Separating the logarithm, we have:

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

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

$$I(X, Y) = \sum_x \sum_y p(x|y) p(y) \log p(x|y) - \sum_x \sum_y p(x, y) \log p(x) . \quad (4)$$

Regrouping the variables, we have:

$$I(X, Y) = \sum_y p(y) \sum_x p(x|y) \log p(x|y) - \sum_x \left(\sum_y p(x, y) \right) \log p(x) . \quad (5)$$

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

$$I(X, Y) = \sum_y p(y) \sum_x p(x|y) \log p(x|y) - \sum_x p(x) \log p(x) . \quad (6)$$

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

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

we can finally show:

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

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 (\rightarrow Definition “rvar”) with the joint probability (\rightarrow Definition “prob-joint”) $p(x, y)$ for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Then, the mutual information (\rightarrow Definition I/4.4.1) of X and Y can be expressed as

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

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

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

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

Separating the logarithm, we have:

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

Regrouping the variables, this reads:

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

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

$$I(X, Y) = \sum_x \sum_y p(x, y) \log p(x, y) - \sum_x p(x) \log p(x) - \sum_y p(y) \log p(y) . \quad (5)$$

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

$$\begin{aligned} H(X) &= - \sum_{x \in \mathcal{X}} p(x) \log p(x) \\ H(X, Y) &= - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x, y) , \end{aligned} \quad (6)$$

we can finally show:

$$\begin{aligned} I(X, Y) &= -H(X, Y) + H(X) + H(Y) \\ &= H(X) + H(Y) - H(X, Y) . \end{aligned} \quad (7)$$

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 (\rightarrow Definition “rvar”) with the joint probability (\rightarrow Definition “prob-joint”) $p(x, y)$ for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Then, the mutual information (\rightarrow Definition I/4.4.1) of X and Y can be expressed as

$$I(X, Y) = H(X, Y) - H(X|Y) - H(Y|X) \quad (1)$$

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

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

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

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

$$I(X, Y) = H(X) - H(X|Y) \quad (3)$$

$$I(X, Y) = H(Y) - H(Y|X) \quad (4)$$

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

$$I(X, Y) = H(X) + H(Y) - H(X, Y) . \quad (5)$$

It is true that

$$I(X, Y) = I(X, Y) + I(X, Y) - I(X, Y) . \quad (6)$$

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

$$\begin{aligned} I(X, Y) &= H(X) - H(X|Y) + H(Y) - H(Y|X) - H(X) - H(Y) + H(X, Y) \\ &= H(X, Y) - H(X|Y) - H(Y|X) \end{aligned} \quad (7)$$

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 (\rightarrow Definition “rvar”) X and Y is defined as

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

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

2) The mutual information of two continuous random variables (\rightarrow Definition “rvar”) X and Y is defined as

$$I(X, Y) = - \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x, y) \cdot \log \frac{p(x, y)}{p(x) \cdot p(y)} dy dx \quad (2)$$

where $p(x)$ and $p(y)$ are the probability density functions (\rightarrow Definition I/1.1.1) of X and Y and $p(x, y)$ is the joint probability (\rightarrow 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 (\rightarrow Definition “rvar”) with the joint probability (\rightarrow Definition “prob-joint”) $p(x, y)$ for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Then, the mutual information (\rightarrow Definition I/4.4.1) of X and Y can be expressed as

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

where $h(X)$ and $h(Y)$ are the marginal differential entropies (\rightarrow Definition I/4.2.1) of X and Y and $h(X|Y)$ and $h(Y|X)$ are the conditional differential entropies (\rightarrow Definition I/4.2.3).

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

$$I(X, Y) = \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dy dx . \quad (2)$$

Separating the logarithm, we have:

$$I(X, Y) = \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p(y)} dy dx - \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x, y) \log p(x) dx dy . \quad (3)$$

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

$$I(X, Y) = \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x|y)p(y) \log p(x|y) dy dx - \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x, y) \log p(x) dy dx . \quad (4)$$

Regrouping the variables, we have:

$$I(X, Y) = \int_{\mathcal{Y}} p(y) \int_{\mathcal{X}} p(x|y) \log p(x|y) dx dy - \int_{\mathcal{X}} \left(\int_{\mathcal{Y}} p(x, y) dy \right) \log p(x) dx . \quad (5)$$

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

$$I(X, Y) = \int_{\mathcal{Y}} p(y) \int_{\mathcal{X}} p(x|y) \log p(x|y) dx dy - \int_{\mathcal{X}} p(x) \log p(x) dx . \quad (6)$$

Now considering the definitions of marginal (\rightarrow Definition I/4.2.1) and conditional (\rightarrow Definition I/4.2.3) differential entropy

$$\begin{aligned} h(X) &= - \int_{\mathcal{X}} p(x) \log p(x) dx \\ h(X|Y) &= \int_{\mathcal{Y}} p(y) h(X|Y = y) dy , \end{aligned} \quad (7)$$

we can finally show:

$$I(X, Y) = -h(X|Y) + h(X) = h(X) - h(X|Y) . \quad (8)$$

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 (\rightarrow Definition “rvar”) with the joint probability (\rightarrow Definition “prob-joint”) $p(x, y)$ for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Then, the mutual information (\rightarrow Definition I/4.4.1) of X and Y can be expressed as

$$I(X, Y) = h(X) + h(Y) - h(X, Y) \quad (1)$$

where $h(X)$ and $h(Y)$ are the marginal differential entropies (\rightarrow Definition I/4.2.1) of X and Y and $h(X, Y)$ is the joint differential entropy (\rightarrow Definition I/4.2.4).

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

$$I(X, Y) = \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p(x) p(y)} dy dx . \quad (2)$$

Separating the logarithm, we have:

$$I(X, Y) = \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x, y) \log p(x, y) dy dx - \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x, y) \log p(x) dy dx - \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x, y) \log p(y) dy dx . \quad (3)$$

Regrouping the variables, this reads:

$$I(X, Y) = \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x, y) \log p(x, y) dy dx - \int_{\mathcal{X}} \left(\int_{\mathcal{Y}} p(x, y) dy \right) \log p(x) dx - \int_{\mathcal{Y}} \left(\int_{\mathcal{X}} p(x, y) dx \right) \log p(y) dy . \quad (4)$$

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

$$I(X, Y) = \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x, y) \log p(x, y) dy dx - \int_{\mathcal{X}} p(x) \log p(x) dx - \int_{\mathcal{Y}} p(y) \log p(y) dy . \quad (5)$$

Now considering the definitions of marginal (\rightarrow Definition I/4.2.1) and joint (\rightarrow Definition I/4.2.4) differential entropy

$$\begin{aligned} h(X) &= - \int_{\mathcal{X}} p(x) \log p(x) dx \\ h(X, Y) &= - \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x, y) \log p(x, y) dy dx , \end{aligned} \quad (6)$$

we can finally show:

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

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 (\rightarrow Definition “rvar”) with the joint probability (\rightarrow Definition “prob-joint”) $p(x, y)$ for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Then, the mutual information (\rightarrow Definition I/4.4.1) of X and Y can be expressed as

$$I(X, Y) = h(X, Y) - h(X|Y) - h(Y|X) \quad (1)$$

where $h(X, Y)$ is the joint differential entropy (\rightarrow Definition I/4.2.4) of X and Y and $h(X|Y)$ and $h(Y|X)$ are the conditional differential entropies (\rightarrow Definition I/4.2.3).

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

$$I(X, Y) = \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dy dx . \quad (2)$$

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

$$I(X, Y) = h(X) - h(X|Y) \quad (3)$$

$$I(X, Y) = h(Y) - h(Y|X) \quad (4)$$

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

$$I(X, Y) = h(X) + h(Y) - h(X, Y) . \quad (5)$$

It is true that

$$I(X, Y) = I(X, Y) + I(X, Y) - I(X, Y) . \quad (6)$$

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

$$\begin{aligned} I(X, Y) &= h(X) - h(X|Y) + h(Y) - h(Y|X) - h(X) - h(Y) + h(X, Y) \\ &= h(X, Y) - h(X|Y) - h(Y|X) \end{aligned} \quad (7)$$

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 Definition

Definition: Let X be a random variable (\rightarrow Definition “rvar”). Then, X is said to follow a Bernoulli distribution with success probability p

$$X \sim \text{Bern}(p) , \quad (1)$$

if $X = 1$ with probability (\rightarrow Definition “prob”) p and $X = 0$ with probability (\rightarrow Definition “prob”) $q = 1 - p$.

Sources:

- Wikipedia (2020): “Bernoulli distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-22; URL: https://en.wikipedia.org/wiki/Bernoulli_distribution.

Metadata: ID: D44 | shortcut: bern | author: JoramSoch | date: 2020-03-22, 17:40.

1.1.2 Mean

Theorem: Let X be a random variable (\rightarrow Definition “rvar”) following a Bernoulli distribution (\rightarrow Definition II/1.1.1):

$$X \sim \text{Bern}(p) . \quad (1)$$

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

$$\text{E}(X) = p . \quad (2)$$

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

$$\text{E}(X) = \sum_{x \in \mathcal{X}} x \cdot \text{Pr}(X = x) . \quad (3)$$

Since there are only two possible outcomes for a Bernoulli random variable (\rightarrow Proof “bern-pmf”), we have:

$$\begin{aligned} \text{E}(X) &= 0 \cdot \text{Pr}(X = 0) + 1 \cdot \text{Pr}(X = 1) \\ &= 0 \cdot (1 - p) + 1 \cdot p \\ &= p . \end{aligned} \quad (4)$$

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 Definition

Definition: Let X be a random variable (\rightarrow Definition “rvar”). Then, X is said to follow a binomial distribution with number of trials n and success probability p

$$X \sim \text{Bin}(n, p) , \quad (1)$$

if X is the number of successes observed in n independent (\rightarrow Definition “ind”) trials, where each trial has two possible outcomes (\rightarrow Definition II/1.1.1) (success/failure) and the probability of success and failure are identical across trials ($p/q = 1 - p$).

Sources:

- Wikipedia (2020): “Binomial distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-22; URL: https://en.wikipedia.org/wiki/Binomial_distribution.

Metadata: ID: D45 | shortcut: bin | author: JoramSoch | date: 2020-03-22, 17:52.

1.2.2 Mean

Theorem: Let X be a random variable (\rightarrow Definition “rvar”) following a binomial distribution (\rightarrow Definition II/1.2.1):

$$X \sim \text{Bin}(n, p) . \quad (1)$$

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

$$E(X) = np . \quad (2)$$

Proof: By definition, a binomial random variable (\rightarrow Definition II/1.2.1) is the sum of n independent and identical Bernoulli trials (\rightarrow Definition II/1.1.1) with success probability p . Therefore, the expected value is

$$E(X) = E(X_1 + \dots + X_n) \quad (3)$$

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

$$\begin{aligned} E(X) &= E(X_1) + \dots + E(X_n) \\ &= \sum_{i=1}^n E(X_i) . \end{aligned} \quad (4)$$

With the expected value of the Bernoulli distribution (\rightarrow Proof II/1.1.2), we have:

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

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 Definition

Definition: Let X be a random vector (\rightarrow Definition “rvec”). Then, X is said to follow a categorical distribution with success probability p_1, \dots, p_k

$$X \sim \text{Cat}([p_1, \dots, p_k]) , \quad (1)$$

if $X = e_i$ with probability (\rightarrow Definition “prob”) p_i for all $i = 1, \dots, k$, where e_i is the i -th elementary row vector, i.e. a $1 \times k$ vector of zeros with a one in i -th position.

Sources:

- Wikipedia (2020): “Categorical distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-22; URL: https://en.wikipedia.org/wiki/Categorical_distribution.

Metadata: ID: D46 | shortcut: cat | author: JoramSoch | date: 2020-03-22, 18:09.

2.1.2 Mean

Theorem: Let X be a random vector (\rightarrow Definition “rvec”) following a categorical distribution (\rightarrow Definition II/2.1.1):

$$X \sim \text{Cat}([p_1, \dots, p_k]) . \quad (1)$$

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

$$\mathbb{E}(X) = [p_1, \dots, p_k] . \quad (2)$$

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

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

Sources:

- original work

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

2.2 Multinomial distribution

2.2.1 Definition

Definition: Let X be a random vector (\rightarrow Definition “rvec”). Then, X is said to follow a multinomial distribution with number of trials n and category probabilities p_1, \dots, p_k

$$X \sim \text{Mult}(n, [p_1, \dots, p_k]) , \quad (1)$$

if X are the numbers of observations belonging to k distinct categories in n independent (\rightarrow Definition “ind”) trials, where each trial has k possible outcomes (\rightarrow Definition II/2.1.1) and the category probabilities are identical across trials.

Sources:

- Wikipedia (2020): “Binomial distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-22; URL: https://en.wikipedia.org/wiki/Multinomial_distribution.

Metadata: ID: D47 | shortcut: mult | author: JoramSoch | date: 2020-03-22, 17:52.

2.2.2 Mean

Theorem: Let X be a random vector (\rightarrow Definition “rvec”) following a multinomial distribution (\rightarrow Definition II/2.2.1):

$$X \sim \text{Mult}(n, [p_1, \dots, p_k]) . \quad (1)$$

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

$$\mathbb{E}(X) = [np_1, \dots, np_k] . \quad (2)$$

Proof: By definition, a multinomial random variable (\rightarrow Definition II/2.2.1) is the sum of n independent and identical categorical trials (\rightarrow Definition II/2.1.1) with category probabilities p_1, \dots, p_k . Therefore, the expected value is

$$\mathbb{E}(X) = \mathbb{E}(X_1 + \dots + X_n) \quad (3)$$

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

$$\begin{aligned} \mathbb{E}(X) &= \mathbb{E}(X_1) + \dots + \mathbb{E}(X_n) \\ &= \sum_{i=1}^n \mathbb{E}(X_i) . \end{aligned} \quad (4)$$

With the expected value of the categorical distribution (\rightarrow Proof II/2.1.2), we have:

$$\mathbb{E}(X) = \sum_{i=1}^n [p_1, \dots, p_k] = n \cdot [p_1, \dots, p_k] = [np_1, \dots, np_k] . \quad (5)$$

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 (\rightarrow Definition “rvar”). Then, X is said to be uniformly distributed with minimum a and maximum b

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

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_\(continuous\)](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 (\rightarrow Definition “rvar”) following a continuous uniform distribution (\rightarrow Definition II/3.1.1):

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

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

$$f_X(x) = \begin{cases} \frac{1}{b-a} , & \text{if } a \leq x \leq b \\ 0 , & \text{otherwise .} \end{cases} \quad (2)$$

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

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

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

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

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

$$\frac{1}{c(a, b)} \int_a^b 1 \, dx = 1 . \quad (5)$$

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

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

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 (\rightarrow Definition “rvar”) following a continuous uniform distribution (\rightarrow Definition II/3.1.1):

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

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

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

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

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

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

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

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

$$F_X(x) = \int_{-\infty}^x 0 \, dz = 0 . \tag{5}$$

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

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

Finally, if $x > b$, we have

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

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 (\rightarrow Definition “rvar”) following a continuous uniform distribution (\rightarrow Definition II/3.1.1):

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

Then, the quantile function (\rightarrow 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 (\rightarrow Proof II/3.1.3) is:

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

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

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

This can be derived by rearranging equation (3):

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

Sources:

- original work

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

3.1.5 Mean

Theorem: Let X be a random variable (\rightarrow Definition “rvar”) following a continuous uniform distribution (\rightarrow Definition II/3.1.1):

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

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

$$\mathbb{E}(X) = \frac{1}{2}(a + b) . \quad (2)$$

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

$$\mathbb{E}(X) = \int_{\mathbb{R}} x \cdot f_X(x) \, dx . \quad (3)$$

With the probability density function of the continuous uniform distribution (\rightarrow Proof II/3.1.2), this becomes:

$$\begin{aligned} \mathbb{E}(X) &= \int_a^b x \cdot \frac{1}{b-a} \, dx \\ &= \left[\frac{1}{2} \frac{x^2}{b-a} \right]_a^b \\ &= \frac{1}{2} \frac{b^2 - a^2}{b-a} \\ &= \frac{1}{2} \frac{(b+a)(b-a)}{b-a} \\ &= \frac{1}{2}(a+b) . \end{aligned} \quad (4)$$

Sources:

- original work

Metadata: ID: P82 | shortcut: cuni-mean | author: JoramSoch | date: 2020-03-16, 16:12.

3.1.6 Median

Theorem: Let X be a random variable (\rightarrow Definition “rvar”) following a continuous uniform distribution (\rightarrow Definition II/3.1.1):

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

Then, the median (\rightarrow Definition “med”) of X is

$$\text{median}(X) = \frac{1}{2}(a + b) . \quad (2)$$

Proof: The median (\rightarrow Definition “med”) is the value at which the cumulative distribution function (\rightarrow Definition I/1.1.3) is $1/2$:

$$F_X(\text{median}(X)) = \frac{1}{2} . \quad (3)$$

The cumulative distribution function of the continuous uniform distribution (\rightarrow Proof II/3.1.3) is

$$F_X(x) = \begin{cases} 0 , & \text{if } x < a \\ \frac{x-a}{b-a} , & \text{if } a \leq x \leq b \\ 1 , & \text{if } x > b . \end{cases} \quad (4)$$

Thus, the inverse CDF (\rightarrow Proof II/3.1.4) is

$$x = bp + a(1 - p) . \quad (5)$$

Setting $p = 1/2$, we obtain:

$$\text{median}(X) = b \cdot \frac{1}{2} + a \cdot \left(1 - \frac{1}{2}\right) = \frac{1}{2}(a + b) . \quad (6)$$

Sources:

- original work

Metadata: ID: P83 | shortcut: cuni-med | author: JoramSoch | date: 2020-03-16, 16:19.

3.1.7 Mode

Theorem: Let X be a random variable (\rightarrow Definition “rvar”) following a continuous uniform distribution (\rightarrow Definition II/3.1.1):

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

Then, the mode (\rightarrow Definition “mode”) of X is

$$\text{mode}(X) \in [a, b] . \quad (2)$$

Proof: The mode (\rightarrow Definition “mode”) is the value which maximizes the probability density function (\rightarrow Definition I/1.1.2):

$$\text{mode}(X) = \arg \max_x f_X(x) . \quad (3)$$

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

$$f_X(x) = \begin{cases} \frac{1}{b-a} , & \text{if } a \leq x \leq b \\ 0 , & \text{otherwise} . \end{cases} \quad (4)$$

Since the PDF attains its only non-zero value whenever $a \leq x \leq b$,

$$\max_x f_X(x) = \frac{1}{b-a}, \quad (5)$$

any value in the interval $[a, b]$ may be considered the mode of X .

Sources:

- original work

Metadata: ID: P84 | shortcut: cuni-med | author: JoramSoch | date: 2020-03-16, 16:29.

3.2 Normal distribution

3.2.1 Definition

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

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

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

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

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 (\rightarrow Definition “rvar”) following a normal distribution (\rightarrow Definition II/3.2.1):

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

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

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

Proof: This follows directly from the definition of the normal distribution (\rightarrow 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 Cumulative distribution function

Theorem: Let X be a random variable (\rightarrow Definition “rvar”) following a normal distributions (\rightarrow Definition II/3.2.1):

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

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

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

where $\operatorname{erf}(x)$ is the error function defined as

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2) dt . \quad (3)$$

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

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

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

$$\begin{aligned} F_X(x) &= \int_{-\infty}^x \mathcal{N}(z; \mu, \sigma^2) dz \\ &= \int_{-\infty}^x \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp \left[-\frac{1}{2} \left(\frac{z - \mu}{\sigma} \right)^2 \right] dz \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^x \exp \left[-\left(\frac{z - \mu}{\sqrt{2}\sigma} \right)^2 \right] dz . \end{aligned} \quad (5)$$

Substituting $t = (z - \mu)/(\sqrt{2}\sigma)$, i.e. $z = \sqrt{2}\sigma t + \mu$, this becomes:

$$\begin{aligned} F_X(x) &= \frac{1}{\sqrt{2\pi}\sigma} \int_{(-\infty - \mu)/(\sqrt{2}\sigma)}^{(x - \mu)/(\sqrt{2}\sigma)} \exp(-t^2) d(\sqrt{2}\sigma t + \mu) \\ &= \frac{\sqrt{2}\sigma}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\frac{x - \mu}{\sqrt{2}\sigma}} \exp(-t^2) dt \\ &= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\frac{x - \mu}{\sqrt{2}\sigma}} \exp(-t^2) dt \\ &= \frac{1}{\sqrt{\pi}} \int_{-\infty}^0 \exp(-t^2) dt + \frac{1}{\sqrt{\pi}} \int_0^{\frac{x - \mu}{\sqrt{2}\sigma}} \exp(-t^2) dt \\ &= \frac{1}{\sqrt{\pi}} \int_0^{\infty} \exp(-t^2) dt + \frac{1}{\sqrt{\pi}} \int_0^{\frac{x - \mu}{\sqrt{2}\sigma}} \exp(-t^2) dt . \end{aligned} \quad (6)$$

Applying (3) to (6), we have:

$$\begin{aligned}
 F_X(x) &= \frac{1}{2} \lim_{x \rightarrow \infty} \operatorname{erf}(x) + \frac{1}{2} \operatorname{erf}\left(\frac{x - \mu}{\sqrt{2}\sigma}\right) \\
 &= \frac{1}{2} + \frac{1}{2} \operatorname{erf}\left(\frac{x - \mu}{\sqrt{2}\sigma}\right) \\
 &= \frac{1}{2} \left[1 + \operatorname{erf}\left(\frac{x - \mu}{\sqrt{2}\sigma}\right) \right].
 \end{aligned} \tag{7}$$

Sources:

- Wikipedia (2020): “Normal distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-20; URL: https://en.wikipedia.org/wiki/Normal_distribution#Cumulative_distribution_function.
- Wikipedia (2020): “Error function”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-20; URL: https://en.wikipedia.org/wiki/Error_function.

Metadata: ID: P85 | shortcut: norm-cdf | author: JoramSoch | date: 2020-03-20, 01:33.

3.2.4 Cumulative distribution function without error function

Theorem: Let X be a random variable (\rightarrow Definition “rvar”) following a normal distributions (\rightarrow Definition II/3.2.1):

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

Then, the cumulative distribution function (\rightarrow Definition I/1.1.3) of X can be expressed as

$$f_X(x) = \Phi_{\mu, \sigma}(x) = \varphi\left(\frac{x - \mu}{\sigma}\right) \cdot \sum_{i=1}^{\infty} \frac{\left(\frac{x - \mu}{\sigma}\right)^{2i-1}}{(2i-1)!!} + \frac{1}{2} \tag{2}$$

where $\varphi(x)$ is the probability density function (\rightarrow Definition I/1.1.2) of the standard normal distribution (\rightarrow Definition “snorm”) and $n!!$ is a double factorial.

Proof:

1) First, consider the standard normal distribution (\rightarrow Definition “snorm”) $\mathcal{N}(0, 1)$ which has the probability density function (\rightarrow Proof II/3.2.2)

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}x^2}. \tag{3}$$

Let $T(x)$ be the indefinite integral of this function. It can be obtained using infinitely repeated integration by parts as follows:

$$\begin{aligned}
T(x) &= \int \varphi(x) \, dx \\
&= \int \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}x^2} \, dx \\
&= \frac{1}{\sqrt{2\pi}} \int 1 \cdot e^{-\frac{1}{2}x^2} \, dx \\
&= \frac{1}{\sqrt{2\pi}} \cdot \left[x \cdot e^{-\frac{1}{2}x^2} + \int x^2 \cdot e^{-\frac{1}{2}x^2} \, dx \right] \\
&= \frac{1}{\sqrt{2\pi}} \cdot \left[x \cdot e^{-\frac{1}{2}x^2} + \left[\frac{1}{3}x^3 \cdot e^{-\frac{1}{2}x^2} + \int \frac{1}{3}x^4 \cdot e^{-\frac{1}{2}x^2} \, dx \right] \right] \\
&= \frac{1}{\sqrt{2\pi}} \cdot \left[x \cdot e^{-\frac{1}{2}x^2} + \left[\frac{1}{3}x^3 \cdot e^{-\frac{1}{2}x^2} + \left[\frac{1}{15}x^5 \cdot e^{-\frac{1}{2}x^2} + \int \frac{1}{15}x^6 \cdot e^{-\frac{1}{2}x^2} \, dx \right] \right] \right] \\
&= \dots \\
&= \frac{1}{\sqrt{2\pi}} \cdot \left[\sum_{i=1}^n \left(\frac{x^{2i-1}}{(2i-1)!!} \cdot e^{-\frac{1}{2}x^2} \right) + \int \left(\frac{x^{2n}}{(2n-1)!!} \cdot e^{-\frac{1}{2}x^2} \right) \, dx \right] \\
&= \frac{1}{\sqrt{2\pi}} \cdot \left[\sum_{i=1}^{\infty} \left(\frac{x^{2i-1}}{(2i-1)!!} \cdot e^{-\frac{1}{2}x^2} \right) + \lim_{n \rightarrow \infty} \int \left(\frac{x^{2n}}{(2n-1)!!} \cdot e^{-\frac{1}{2}x^2} \right) \, dx \right].
\end{aligned} \tag{4}$$

Since $(2n-1)!!$ grows faster than x^{2n} , it holds that

$$\frac{1}{\sqrt{2\pi}} \cdot \lim_{n \rightarrow \infty} \int \left(\frac{x^{2n}}{(2n-1)!!} \cdot e^{-\frac{1}{2}x^2} \right) \, dx = \int 0 \, dx = c \tag{5}$$

for constant c , such that the indefinite integral becomes

$$\begin{aligned}
T(x) &= \frac{1}{\sqrt{2\pi}} \cdot \sum_{i=1}^{\infty} \left(\frac{x^{2i-1}}{(2i-1)!!} \cdot e^{-\frac{1}{2}x^2} \right) + c \\
&= \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}x^2} \cdot \sum_{i=1}^{\infty} \frac{x^{2i-1}}{(2i-1)!!} + c \\
&\stackrel{(3)}{=} \varphi(x) \cdot \sum_{i=1}^{\infty} \frac{x^{2i-1}}{(2i-1)!!} + c.
\end{aligned} \tag{6}$$

2) Next, let $\Phi(x)$ be the cumulative distribution function (\rightarrow Definition I/1.1.3) of the standard normal distribution (\rightarrow Definition “snorm”):

$$\Phi(x) = \int_{-\infty}^x \varphi(x) \, dx. \tag{7}$$

It can be obtained by matching $T(0)$ to $\Phi(0)$ which is $1/2$, because the standard normal distribution is symmetric around zero:

$$\begin{aligned}
T(0) &= \varphi(0) \cdot \sum_{i=1}^{\infty} \frac{0^{2i-1}}{(2i-1)!!} + c = \frac{1}{2} = \Phi(0) \\
&\Leftrightarrow c = \frac{1}{2} \\
\Rightarrow \Phi(x) &= \varphi(x) \cdot \sum_{i=1}^{\infty} \frac{x^{2i-1}}{(2i-1)!!} + \frac{1}{2}.
\end{aligned} \tag{8}$$

3) Finally, the cumulative distribution functions of the standard normal distribution and the general normal distribution are related to each other (\rightarrow Proof “norm-snorm”) as

$$\Phi_{\mu,\sigma}(x) = \Phi\left(\frac{x-\mu}{\sigma}\right). \tag{9}$$

Combining (9) with (8), we have:

$$\Phi_{\mu,\sigma}(x) = \varphi\left(\frac{x-\mu}{\sigma}\right) \cdot \sum_{i=1}^{\infty} \frac{\left(\frac{x-\mu}{\sigma}\right)^{2i-1}}{(2i-1)!!} + \frac{1}{2}. \tag{10}$$

Sources:

- Soch J (2015): “Solution for the Indefinite Integral of the Standard Normal Probability Density Function”; in: *arXiv stat.OT*, arXiv:1512.04858; URL: <https://arxiv.org/abs/1512.04858>.
- Wikipedia (2020): “Normal distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-20; URL: https://en.wikipedia.org/wiki/Normal_distribution#Cumulative_distribution_function.

Metadata: ID: P86 | shortcut: norm-cdfwerf | author: JoramSoch | date: 2020-03-20, 04:26.

3.2.5 Quantile function

Theorem: Let X be a random variable (\rightarrow Definition “rvar”) following a normal distributions (\rightarrow Definition II/3.2.1):

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

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

$$Q_X(p) = \sqrt{2}\sigma \cdot \text{erf}^{-1}(2p-1) + \mu \tag{2}$$

where $\text{erf}^{-1}(x)$ is the inverse error function.

Proof: The cumulative distribution function of the normal distribution (\rightarrow Proof II/3.2.3) is:

$$F_X(x) = \frac{1}{2} \left[1 + \text{erf} \left(\frac{x-\mu}{\sqrt{2}\sigma} \right) \right]. \tag{3}$$

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

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

This can be derived by rearranging equation (3):

$$\begin{aligned}
p &= \frac{1}{2} \left[1 + \operatorname{erf} \left(\frac{x - \mu}{\sqrt{2}\sigma} \right) \right] \\
2p - 1 &= \operatorname{erf} \left(\frac{x - \mu}{\sqrt{2}\sigma} \right) \\
\operatorname{erf}^{-1}(2p - 1) &= \frac{x - \mu}{\sqrt{2}\sigma} \\
x &= \sqrt{2}\sigma \cdot \operatorname{erf}^{-1}(2p - 1) + \mu .
\end{aligned} \tag{5}$$

Sources:

- Wikipedia (2020): “Normal distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-20; URL: https://en.wikipedia.org/wiki/Normal_distribution#Quantile_function.

Metadata: ID: P87 | shortcut: norm-qf | author: JoramSoch | date: 2020-03-20, 04:47.

3.2.6 Mean

Theorem: Let X be a random variable (\rightarrow Definition “rvar”) following a normal distribution (\rightarrow Definition II/3.2.1):

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

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

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

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

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

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

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

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

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

The general antiderivatives are

$$\begin{aligned}
\int x \cdot \exp [-ax^2] dx &= -\frac{1}{2a} \cdot \exp [-ax^2] \\
\int \exp [-ax^2] dx &= \frac{1}{2} \sqrt{\frac{\pi}{a}} \cdot \operatorname{erf} [\sqrt{a}x]
\end{aligned} \tag{6}$$

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

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

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.

3.2.7 Median

Theorem: Let X be a random variable (\rightarrow Definition “rvar”) following a normal distribution (\rightarrow Definition II/3.2.1):

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

Then, the median (\rightarrow Definition “med”) of X is

$$\text{median}(X) = \mu . \quad (2)$$

Proof: The median (\rightarrow Definition “med”) is the value at which the cumulative distribution function (\rightarrow Definition I/1.1.3) is $1/2$:

$$F_X(\text{median}(X)) = \frac{1}{2} . \quad (3)$$

The cumulative distribution function of the normal distribution (\rightarrow Proof II/3.2.3) is

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

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

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

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

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

Sources:

- original work

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

3.2.8 Mode

Theorem: Let X be a random variable (\rightarrow Definition “rvar”) following a normal distribution (\rightarrow Definition II/3.2.1):

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

Then, the mode (\rightarrow Definition “mode”) of X is

$$\text{mode}(X) = \mu . \quad (2)$$

Proof: The mode (\rightarrow Definition “mode”) is the value which maximizes the probability density function (\rightarrow Definition I/1.1.2):

$$\text{mode}(X) = \arg \max_x f_X(x) . \quad (3)$$

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

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

The first two derivatives of this function are:

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

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

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

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

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

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

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

$$\text{mode}(X) = \mu . \quad (9)$$

Sources:

- original work

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

3.2.9 Variance

Theorem: Let X be a random variable (\rightarrow Definition “rvar”) following a normal distribution (\rightarrow Definition II/3.2.1):

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

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

$$\text{Var}(X) = \sigma^2 . \quad (2)$$

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

$$\text{Var}(X) = \int_{\mathbb{R}} (x - E(X))^2 \cdot f_X(x) dx . \quad (3)$$

With the expected value (\rightarrow Proof II/3.2.6) and probability density function (\rightarrow Proof II/3.2.2) of the normal distribution, this reads:

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

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

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

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

$$\begin{aligned} \text{Var}(X) &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} (\sqrt{2}\sigma x)^2 \cdot \exp \left[-\frac{1}{2} \left(\frac{\sqrt{2}\sigma x}{\sigma} \right)^2 \right] d(\sqrt{2}\sigma x) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \cdot 2\sigma^2 \cdot \sqrt{2}\sigma \int_{-\infty}^{+\infty} x^2 \cdot \exp [-x^2] dx \\ &= \frac{2\sigma^2}{\sqrt{\pi}} \int_{-\infty}^{+\infty} x^2 \cdot e^{-x^2} dx . \end{aligned} \quad (6)$$

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

$$\text{Var}(X) = \frac{4\sigma^2}{\sqrt{\pi}} \int_0^{\infty} x^2 \cdot e^{-x^2} dx . \quad (7)$$

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

$$\text{Var}(X) = \frac{4\sigma^2}{\sqrt{\pi}} \int_0^{\infty} z \cdot e^{-z} \cdot \frac{1}{2} z^{-1/2} dz = \frac{2\sigma^2}{\sqrt{\pi}} \int_0^{\infty} z^{\frac{3}{2}-1} \cdot e^{-z} dz \quad (8)$$

and using the definition of the gamma function

$$\Gamma(x) = \int_0^{\infty} z^{x-1} \cdot e^{-z} dz , \quad (9)$$

we can finally show that

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

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.

3.2.10 Moment-generating function

Theorem: Let X be a random variable (\rightarrow Definition “rvar”) following a normal distribution (\rightarrow Definition II/3.2.1):

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

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

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

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

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

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

$$M_X(t) = \mathbb{E} [e^{tX}] . \quad (4)$$

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

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

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

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

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

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

With the Gaussian integral (\rightarrow Proof “norm-gi”)

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

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 (\rightarrow Definition “rvar”). Then, X is said to follow a gamma distribution with shape a and rate b

$$X \sim \text{Gam}(a, b) , \quad (1)$$

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

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

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 (\rightarrow Definition “rvar”) following a gamma distribution (\rightarrow Definition II/3.3.1):

$$X \sim \text{Gam}(a, b) . \quad (1)$$

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

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

Proof: This follows directly from the definition of the gamma distribution (\rightarrow 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 (\rightarrow Definition “rvar”). Then, X is said to be exponentially distributed with rate (or, inverse scale) λ

$$X \sim \text{Exp}(\lambda) , \quad (1)$$

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

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

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

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

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

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

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

which is equivalent to the probability density function of the exponential distribution (\rightarrow 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 (\rightarrow Definition “rvar”) following an exponential distribution (\rightarrow Definition II/3.4.1):

$$X \sim \text{Exp}(\lambda) . \quad (1)$$

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

$$f_X(x) = \lambda \exp[-\lambda x] . \quad (2)$$

Proof: This follows directly from the definition of the exponential distribution (\rightarrow 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 (\rightarrow Definition “rvar”) following an exponential distribution (\rightarrow Definition II/3.4.1):

$$X \sim \text{Exp}(\lambda) . \quad (1)$$

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

$$F_X(x) = \begin{cases} 0 , & \text{if } x < 0 \\ 1 - \exp[-\lambda x] , & \text{if } x \geq 0 . \end{cases} \quad (2)$$

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

$$\text{Exp}(x; \lambda) = \begin{cases} 0 , & \text{if } x < 0 \\ \lambda \exp[-\lambda x] , & \text{if } x \geq 0 . \end{cases} \quad (3)$$

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

$$F_X(x) = \int_{-\infty}^x \text{Exp}(z; \lambda) \, dz . \quad (4)$$

If $x < 0$, we have:

$$F_X(x) = \int_{-\infty}^x 0 \, dz = 0 . \quad (5)$$

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

$$\begin{aligned} F_X(x) &= \int_{-\infty}^0 \text{Exp}(z; \lambda) \, dz + \int_0^x \text{Exp}(z; \lambda) \, dz \\ &= \int_{-\infty}^0 0 \, dz + \int_0^x \lambda \exp[-\lambda z] \, dz \\ &= 0 + \lambda \left[-\frac{1}{\lambda} \exp[-\lambda z] \right]_0^x \\ &= \lambda \left[\left(-\frac{1}{\lambda} \exp[-\lambda x] \right) - \left(-\frac{1}{\lambda} \exp[-\lambda \cdot 0] \right) \right] \\ &= 1 - \exp[-\lambda x] . \end{aligned} \quad (6)$$

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 (\rightarrow Definition “rvar”) following an exponential distribution (\rightarrow Definition II/3.4.1):

$$X \sim \text{Exp}(\lambda) . \quad (1)$$

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

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

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

$$F_X(x) = \begin{cases} 0 , & \text{if } x < 0 \\ 1 - \exp[-\lambda x] , & \text{if } x \geq 0 . \end{cases} \quad (3)$$

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

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

This can be derived by rearranging equation (3):

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

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 (\rightarrow Definition “rvar”) following an exponential distribution (\rightarrow Definition II/3.4.1):

$$X \sim \text{Exp}(\lambda) . \quad (1)$$

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

$$E(X) = \frac{1}{\lambda} . \quad (2)$$

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

$$E(X) = \int_{\mathbb{R}} x \cdot f_X(x) dx . \quad (3)$$

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

$$\begin{aligned}
E(X) &= \int_0^{+\infty} x \cdot \lambda \exp(-\lambda x) dx \\
&= \lambda \int_0^{+\infty} x \cdot \exp(-\lambda x) dx .
\end{aligned} \tag{4}$$

Using the following anti-derivative

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

the expected value becomes

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

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 (\rightarrow Definition “rvar”) following an exponential distribution (\rightarrow Definition II/3.4.1):

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

Then, the median (\rightarrow Definition “med”) of X is

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

Proof: The median (\rightarrow Definition “med”) is the value at which the cumulative distribution function (\rightarrow Definition I/1.1.3) is $1/2$:

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

The cumulative distribution function of the exponential distribution (\rightarrow Definition “exp-cdf”) is

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

Thus, the inverse CDF is

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

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

$$\text{median}(X) = -\frac{\ln(1 - \frac{1}{2})}{\lambda} = \frac{\ln 2}{\lambda}. \quad (6)$$

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 (\rightarrow Definition “rvar”) following an exponential distribution (\rightarrow Definition II/3.4.1):

$$X \sim \text{Exp}(\lambda). \quad (1)$$

Then, the mode (\rightarrow Definition “mode”) of X is

$$\text{mode}(X) = 0. \quad (2)$$

Proof: The mode (\rightarrow Definition “mode”) is the value which maximizes the probability density function (\rightarrow Definition I/1.1.2):

$$\text{mode}(X) = \arg \max_x f_X(x). \quad (3)$$

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

$$f_X(x) = \begin{cases} 0, & \text{if } x < 0 \\ \lambda \exp[-\lambda x], & \text{if } x \geq 0. \end{cases} \quad (4)$$

Since

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

and

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

it follows that

$$\text{mode}(X) = 0. \quad (7)$$

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 (\rightarrow Definition “rvec”). Then, X is said to be multivariate normally distributed with mean μ and covariance Σ

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

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

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

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 (\rightarrow Definition “rvec”) following a multivariate normal distribution (\rightarrow Definition II/4.1.1):

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

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

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

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

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

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

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

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

$$M_x(t) = \mathbb{E}(\exp[t^T x]) \quad (3)$$

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

$$\begin{aligned} M_y(t) &= \mathbb{E}(\exp[t^T(Ax + b)]) \\ &= \mathbb{E}(\exp[t^T Ax] \cdot \exp[t^T b]) \\ &= \exp[t^T b] \cdot \mathbb{E}(\exp[t^T Ax]) \\ &= \exp[t^T b] \cdot M_x(At) . \end{aligned} \quad (4)$$

The moment-generating function of the multivariate normal distribution (\rightarrow Proof “mvn-mgf”) is

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

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

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

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 $A\Sigma A^T$.

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-combinations>

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 (\rightarrow Definition II/4.1.1):

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

Then, the marginal distribution (\rightarrow Definition “dist-marg”) of any subset vector x_s is also a multivariate normal distribution

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

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,

$$x_s = Sx \quad (3)$$

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

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

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

Sources:

- original work

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

4.1.5 Conditional distributions

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

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

Then, the conditional distribution (\rightarrow Definition “dist-cond”) of any subset vector x_1 , given the complement vector x_2 , is also a multivariate normal distribution

$$x_1|x_2 \sim \mathcal{N}(\mu_{1|2}, \Sigma_{1|2}) \quad (2)$$

where the conditional mean and covariance are

$$\begin{aligned} \mu_{1|2} &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \\ \Sigma_{1|2} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} . \end{aligned} \quad (3)$$

with block-wise mean and covariance defined as

$$\begin{aligned} \mu &= \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \\ \Sigma &= \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} . \end{aligned} \quad (4)$$

Proof: Without loss of generality, we assume that, in parallel to (4),

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad (5)$$

where x_1 is an $n_1 \times 1$ vector, x_2 is an $n_2 \times 1$ vector and x is an $n_1 + n_2 = n \times 1$ vector.

By construction, the joint distribution (\rightarrow Definition “dist-joint”) of x_1 and x_2 is:

$$x_1, x_2 \sim \mathcal{N}(\mu, \Sigma) . \quad (6)$$

Moreover, the marginal distribution (\rightarrow Definition “dist-marg”) of x_2 follows from (\rightarrow Proof II/4.1.4) (1) and (4) as

$$x_2 \sim \mathcal{N}(\mu_2, \Sigma_{22}) . \quad (7)$$

According to the law of conditional probability (\rightarrow Definition “prob-cond”), it holds that

$$p(x_1|x_2) = \frac{p(x_1, x_2)}{p(x_2)} \quad (8)$$

Applying (6) and (7) to (8), we have:

$$p(x_1|x_2) = \frac{\mathcal{N}(x; \mu, \Sigma)}{\mathcal{N}(x_2; \mu_2, \Sigma_{22})} . \quad (9)$$

Using the probability density function of the multivariate normal distribution (\rightarrow Proof II/4.1.2), this becomes:

$$\begin{aligned} p(x_1|x_2) &= \frac{1/\sqrt{(2\pi)^n |\Sigma|} \cdot \exp \left[-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu) \right]}{1/\sqrt{(2\pi)^{n_2} |\Sigma_{22}|} \cdot \exp \left[-\frac{1}{2}(x_2 - \mu_2)^T \Sigma_{22}^{-1} (x_2 - \mu_2) \right]} \\ &= \frac{1}{\sqrt{(2\pi)^{n-n_2}}} \cdot \sqrt{\frac{|\Sigma_{22}|}{|\Sigma|}} \cdot \exp \left[-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu) + \frac{1}{2}(x_2 - \mu_2)^T \Sigma_{22}^{-1} (x_2 - \mu_2) \right] . \end{aligned} \quad (10)$$

Writing the inverse of Σ as

$$\Sigma^{-1} = \begin{bmatrix} \Sigma^{11} & \Sigma^{12} \\ \Sigma^{21} & \Sigma^{22} \end{bmatrix} \quad (11)$$

and applying (4) to (10), we get:

$$\begin{aligned} p(x_1|x_2) &= \frac{1}{\sqrt{(2\pi)^{n-n_2}}} \cdot \sqrt{\frac{|\Sigma_{22}|}{|\Sigma|}} \cdot \\ &\quad \exp \left[-\frac{1}{2} \left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \right)^T \begin{bmatrix} \Sigma^{11} & \Sigma^{12} \\ \Sigma^{21} & \Sigma^{22} \end{bmatrix} \left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \right) \right. \\ &\quad \left. + \frac{1}{2} (x_2 - \mu_2)^T \Sigma_{22}^{-1} (x_2 - \mu_2) \right] . \end{aligned} \quad (12)$$

Multiplying out within the exponent of (12), we have

$$\begin{aligned} p(x_1|x_2) &= \frac{1}{\sqrt{(2\pi)^{n-n_2}}} \cdot \sqrt{\frac{|\Sigma_{22}|}{|\Sigma|}} \cdot \\ &\quad \exp \left[-\frac{1}{2} \left((x_1 - \mu_1)^T \Sigma^{11} (x_1 - \mu_1) + 2(x_1 - \mu_1)^T \Sigma^{12} (x_2 - \mu_2) + (x_2 - \mu_2)^T \Sigma^{22} (x_2 - \mu_2) \right) \right. \\ &\quad \left. + \frac{1}{2} (x_2 - \mu_2)^T \Sigma_{22}^{-1} (x_2 - \mu_2) \right] \end{aligned} \quad (13)$$

where we have used the fact that $\Sigma^{21T} = \Sigma^{12}$, because Σ^{-1} is a symmetric matrix.

The inverse of a block matrix is

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} (A - BD^{-1}C)^{-1} & -(A - BD^{-1}C)^{-1}BD^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1} \end{bmatrix}, \quad (14)$$

thus the inverse of Σ in (11) is

$$\begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}^{-1} = \begin{bmatrix} (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} & -(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}\Sigma_{12}\Sigma_{22}^{-1} \\ -\Sigma_{22}^{-1}\Sigma_{21}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} & \Sigma_{22}^{-1} + \Sigma_{22}^{-1}\Sigma_{21}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}\Sigma_{12}\Sigma_{22}^{-1} \end{bmatrix}. \quad (15)$$

Plugging this into (13), we have:

$$\begin{aligned} p(x_1|x_2) &= \frac{1}{\sqrt{(2\pi)^{n-n_2}}} \cdot \sqrt{\frac{|\Sigma_{22}|}{|\Sigma|}} \cdot \\ &\exp \left[-\frac{1}{2} \left((x_1 - \mu_1)^T (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} (x_1 - \mu_1) - \right. \right. \\ &\quad 2(x_1 - \mu_1)^T (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} \Sigma_{12}\Sigma_{22}^{-1} (x_2 - \mu_2) + \\ &\quad (x_2 - \mu_2)^T \left[\Sigma_{22}^{-1} + \Sigma_{22}^{-1}\Sigma_{21}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}\Sigma_{12}\Sigma_{22}^{-1} \right] (x_2 - \mu_2) \left. \right) \\ &\quad \left. + \frac{1}{2} \left((x_2 - \mu_2)^T \Sigma_{22}^{-1} (x_2 - \mu_2) \right) \right]. \end{aligned} \quad (16)$$

Eliminating some terms, we have:

$$\begin{aligned} p(x_1|x_2) &= \frac{1}{\sqrt{(2\pi)^{n-n_2}}} \cdot \sqrt{\frac{|\Sigma_{22}|}{|\Sigma|}} \cdot \\ &\exp \left[-\frac{1}{2} \left((x_1 - \mu_1)^T (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} (x_1 - \mu_1) - \right. \right. \\ &\quad 2(x_1 - \mu_1)^T (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} \Sigma_{12}\Sigma_{22}^{-1} (x_2 - \mu_2) + \\ &\quad \left. \left. (x_2 - \mu_2)^T \Sigma_{22}^{-1} \Sigma_{21} (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} \Sigma_{12}\Sigma_{22}^{-1} (x_2 - \mu_2) \right) \right]. \end{aligned} \quad (17)$$

Rearranging the terms, we have

$$\begin{aligned} p(x_1|x_2) &= \frac{1}{\sqrt{(2\pi)^{n-n_2}}} \cdot \sqrt{\frac{|\Sigma_{22}|}{|\Sigma|}} \cdot \exp \left[-\frac{1}{2} \cdot \right. \\ &\quad \left. \left[(x_1 - \mu_1) - \Sigma_{12}^T \Sigma_{22}^{-1} (x_2 - \mu_2) \right]^T (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} \left[(x_1 - \mu_1) - \Sigma_{12}^T \Sigma_{22}^{-1} (x_2 - \mu_2) \right] \right] \\ &= \frac{1}{\sqrt{(2\pi)^{n-n_2}}} \cdot \sqrt{\frac{|\Sigma_{22}|}{|\Sigma|}} \cdot \exp \left[-\frac{1}{2} \cdot \right. \\ &\quad \left. \left[x_1 - (\mu_1 + \Sigma_{12}^T \Sigma_{22}^{-1} (x_2 - \mu_2)) \right]^T (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} \left[x_1 - (\mu_1 + \Sigma_{12}^T \Sigma_{22}^{-1} (x_2 - \mu_2)) \right] \right] \end{aligned} \quad (18)$$

where we have used the fact that $\Sigma_{21}^T = \Sigma_{12}$, because Σ is a covariance matrix.

The determinant of a block matrix is

$$\begin{vmatrix} A & B \\ C & D \end{vmatrix} = |D| \cdot |A - BD^{-1}C|, \quad (19)$$

such that we have for Σ that

$$\begin{vmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{vmatrix} = |\Sigma_{22}| \cdot |\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}|. \quad (20)$$

With this and $n - n_2 = n_1$, we finally arrive at

$$p(x_1|x_2) = \frac{1}{\sqrt{(2\pi)^{n_1} |\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}|}} \cdot \exp \left[-\frac{1}{2} \cdot \left[x_1 - (\mu_1 + \Sigma_{12}^T \Sigma_{22}^{-1} (x_2 - \mu_2)) \right]^T (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} \left[x_1 - (\mu_1 + \Sigma_{12}^T \Sigma_{22}^{-1} (x_2 - \mu_2)) \right] \right] \quad (21)$$

which is the probability density function of a multivariate normal distribution (\rightarrow Proof II/4.1.2)

$$p(x_1|x_2) = \mathcal{N}(x_1; \mu_{1|2}, \Sigma_{1|2}) \quad (22)$$

with the mean $\mu_{1|2}$ and variance $\Sigma_{1|2}$ given by (3).

Sources:

- Wang, Ruye (2006): “Marginal and conditional distributions of multivariate normal distribution”; in: *Computer Image Processing and Analysis*; URL: <http://fourier.eng.hmc.edu/e161/lectures/gaussianprocess/node7.html>.
- Wikipedia (2020): “Multivariate normal distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-20; URL: https://en.wikipedia.org/wiki/Multivariate_normal_distribution#Conditional_distributions.

Metadata: ID: P88 | shortcut: mvn-cond | author: JoramSoch | date: 2020-03-20, 08:44.

4.2 Normal-gamma distribution

4.2.1 Definition

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

$$X, Y \sim \text{NG}(\mu, \Lambda, a, b), \quad (1)$$

if and only if their joint probability (\rightarrow Definition “prob-joint”) density function (\rightarrow Definition I/1.1.2) is given by

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

where $\mathcal{N}(x; \mu, \Sigma)$ is the probability density function of the multivariate normal distribution (\rightarrow Proof II/4.1.2) with mean μ and covariance Σ and $\text{Gam}(x; a, b)$ is the probability density function of the

gamma distribution (\rightarrow 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 (\rightarrow Definition II/4.2.1):

$$x, y \sim \text{NG}(\mu, \Lambda, a, b) . \quad (1)$$

Then, the joint probability (\rightarrow Definition “prob-joint”) density function (\rightarrow Definition I/1.1.2) of x and y is

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

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

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

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

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

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

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

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 (\rightarrow Definition “rvec”) and $y > 0$ be a random variable (\rightarrow Definition “rvar”). Assume two normal-gamma distributions (\rightarrow Definition II/4.2.1) P and Q specifying the joint distribution of x and y as

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

Then, the Kullback-Leibler divergence (\rightarrow Definition “kl”) of P from Q is given by

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

Proof: The probability density function of the normal-gamma distribution (\rightarrow Proof II/4.2.2) is

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

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

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

and the Kullback-Leibler divergence of the univariate gamma distribution (\rightarrow Proof “gam-kl”) is

$$\text{KL}[P \parallel Q] = a_2 \ln \frac{b_1}{b_2} - \ln \frac{\Gamma(a_1)}{\Gamma(a_2)} + (a_1 - a_2) \psi(a_1) - (b_1 - b_2) \frac{a_1}{b_1} \quad (5)$$

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

The KL divergence for a continuous random variable (\rightarrow Definition “kl”) is given by

$$\text{KL}[P \parallel Q] = \int_{\mathcal{Z}} p(z) \ln \frac{p(z)}{q(z)} dz \quad (6)$$

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

$$\text{KL}[P \parallel Q] = \int_0^\infty \int_{\mathbb{R}^k} p(x, y) \ln \frac{p(x, y)}{q(x, y)} dx dy . \quad (7)$$

Using the law of conditional probability (\rightarrow Definition “prob-cond”), this can be evaluated as follows:

$$\begin{aligned}
\text{KL}[P \parallel Q] &= \int_0^\infty \int_{\mathbb{R}^k} p(x|y) p(y) \ln \frac{p(x|y) p(y)}{q(x|y) q(y)} dx dy \\
&= \int_0^\infty \int_{\mathbb{R}^k} p(x|y) p(y) \ln \frac{p(x|y)}{q(x|y)} dx dy + \int_0^\infty \int_{\mathbb{R}^k} p(x|y) p(y) \ln \frac{p(y)}{q(y)} dx dy \\
&= \int_0^\infty p(y) \int_{\mathbb{R}^k} p(x|y) \ln \frac{p(x|y)}{q(x|y)} dx dy + \int_0^\infty p(y) \ln \frac{p(y)}{q(y)} \int_{\mathbb{R}^k} p(x|y) dx dy \\
&= \langle \text{KL}[p(x|y) \parallel q(x|y)] \rangle_{p(y)} + \text{KL}[p(y) \parallel q(y)] .
\end{aligned} \tag{8}$$

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

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

and using the relation (\rightarrow Proof “gam-mean”) $y \sim \text{Gam}(a, b) \Rightarrow \langle y \rangle = a/b$, we have

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

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 (\rightarrow Definition II/4.2.1):

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

Then, the marginal distribution (\rightarrow Definition “dist-marg”) of y is a gamma distribution (\rightarrow Definition II/3.3.1)

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

and the marginal distribution (\rightarrow Definition “dist-marg”) of x is a multivariate t-distribution (\rightarrow Definition “mvt”)

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

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

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

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

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

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

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

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

which is the probability density function of a multivariate t-distribution (\rightarrow 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 (\rightarrow Definition “rmat”). Then, X is said to be matrix-normally distributed with mean M , covariance across rows U and covariance across columns V

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

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

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

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 (\rightarrow Definition “rmat”) following a matrix-normal distribution (\rightarrow Definition II/5.1.1):

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

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

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

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

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

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

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

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

Proof: The probability density function of the matrix-normal distribution (\rightarrow 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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

By showing that the probability density functions (\rightarrow Definition I/1.1.2) are identical, it is proven that the associated probability distributions (\rightarrow 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.

5.2 Wishart distribution

5.2.1 Definition

Definition: Let X be an $n \times p$ matrix following a matrix-normal distribution (\rightarrow Definition II/5.1.1) with mean zero, independence across rows and covariance across columns V :

$$X \sim \mathcal{MN}(0, I_n, V) . \quad (1)$$

Define the scatter matrix S as the product of the transpose of X with itself:

$$S = X^T X = \sum_{i=1}^n x_i^T x_i . \quad (2)$$

Then, the matrix S is said to follow a Wishart distribution with scale matrix V and degrees of freedom n

$$S \sim \mathcal{W}(V, n) \quad (3)$$

where $n > p - 1$ and V is a positive definite symmetric covariance matrix.

Sources:

- Wikipedia (2020): “Wishart distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-22; URL: https://en.wikipedia.org/wiki/Wishart_distribution#Definition.

Metadata: ID: D43 | shortcut: wish | author: JoramSoch | date: 2020-03-22, 17:15.

Chapter III

Statistical Models

1 Normal data

1.1 Multiple linear regression

1.1.1 Definition

Definition: Let y be an $n \times 1$ vector and let X be an $n \times p$ matrix. Then, a statement asserting a linear combination of X into y

$$y = X\beta + \varepsilon, \quad (1)$$

together with a statement asserting a normal distribution (\rightarrow Definition II/4.1.1) for ε

$$\varepsilon \sim \mathcal{N}(0, \sigma^2 V) \quad (2)$$

is called a univariate linear regression model or simply, “multiple linear regression”.

- y is called “measured data”, “dependent variable” or “measurements”;
- X is called “design matrix”, “set of independent variables” or “predictors”;
- V is called “covariance matrix” or “covariance structure”;
- β are called “regression coefficients” or “weights”;
- ε is called “noise”, “errors” or “error terms”;
- σ^2 is called “noise variance” or “error variance”;
- n is the number of observations;
- p is the number of predictors.

Alternatively, the linear combination may also be written as

$$y = \sum_{i=1}^p \beta_i x_i + \varepsilon \quad (3)$$

or, when the model includes an intercept term, as

$$y = \beta_0 + \sum_{i=1}^p \beta_i x_i + \varepsilon \quad (4)$$

which is equivalent to adding a constant regressor $x_0 = 1_n$ to the design matrix X .

When the covariance structure V is equal to the $n \times n$ identity matrix, this is called multiple linear regression with independent and identically distributed (i.i.d.) observations:

$$V = I_n \quad \Rightarrow \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n) \quad \Rightarrow \quad \varepsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2). \quad (5)$$

Otherwise, it is called multiple linear regression with correlated observations.

Sources:

- Wikipedia (2020): “Linear regression”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-21; URL: https://en.wikipedia.org/wiki/Linear_regression#Simple_and_multiple_linear_regression.

Metadata: ID: D36 | shortcut: mlr | author: JoramSoch | date: 2020-03-21, 20:09.

1.1.2 Ordinary least squares (1)

Theorem: Given a linear regression model (\rightarrow Definition III/1.1.1) with independent observations

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

the parameters minimizing the residual sum of squares (\rightarrow Definition III/1.1.6) are given by

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

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,

$$X^T \hat{\varepsilon} = 0, \quad (3)$$

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:

$$\begin{aligned} X^T \hat{\varepsilon} &= 0 \\ X^T (y - X\hat{\beta}) &= 0 \\ X^T y - X^T X\hat{\beta} &= 0 \\ X^T X\hat{\beta} &= X^T y \\ \hat{\beta} &= (X^T X)^{-1} X^T y. \end{aligned} \quad (4)$$

Sources:

- Stephan, Klaas Enno (2010): “The General Linear Model (GLM)”; in: *Methods and models for fMRI data analysis in neuroeconomics*, Lecture 3, Slides 10/11; URL: <http://www.socialbehavior.uzh.ch/teaching/methodspring10.html>.

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

1.1.3 Ordinary least squares (2)

Theorem: Given a linear regression model (\rightarrow Definition III/1.1.1) with independent observations

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

the parameters minimizing the residual sum of squares (\rightarrow Definition III/1.1.6) are given by

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

Proof: The residual sum of squares (\rightarrow Definition III/1.1.6) is defined as

$$\text{RSS}(\beta) = \sum_{i=1}^n \varepsilon_i^2 = \varepsilon^T \varepsilon = (y - X\beta)^T (y - X\beta) \quad (3)$$

which can be developed into

$$\begin{aligned} \text{RSS}(\beta) &= y^T y - y^T X \beta - \beta^T X^T y + \beta^T X^T X \beta \\ &= y^T y - 2\beta^T X^T y + \beta^T X^T X \beta . \end{aligned} \quad (4)$$

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

$$\frac{d\text{RSS}(\beta)}{d\beta} = -2X^T y + 2X^T X \beta \quad (5)$$

and setting this derivative to zero, we obtain:

$$\begin{aligned} \frac{d\text{RSS}(\hat{\beta})}{d\beta} &= 0 \\ 0 &= -2X^T y + 2X^T X \hat{\beta} \\ X^T X \hat{\beta} &= X^T y \\ \hat{\beta} &= (X^T X)^{-1} X^T y . \end{aligned} \quad (6)$$

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

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.

1.1.4 Total sum of squares

Definition: Let there be a multiple linear regression with independent observations (\rightarrow Definition III/1.1.1) using measured data y and design matrix X :

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

Then, the total sum of squares (TSS) is defined as the sum of squared deviations of the measured signal from the average signal:

$$\text{TSS} = \sum_{i=1}^n (y_i - \bar{y})^2 \quad \text{where} \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i . \quad (2)$$

Sources:

- Wikipedia (2020): “Total sum of squares”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-21; URL: https://en.wikipedia.org/wiki/Total_sum_of_squares.

Metadata: ID: D37 | shortcut: tss | author: JoramSoch | date: 2020-03-21, 21:44.

1.1.5 Explained sum of squares

Definition: Let there be a multiple linear regression with independent observations (\rightarrow Definition III/1.1.1) using measured data y and design matrix X :

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

Then, the explained sum of squares (ESS) is defined as the sum of squared deviations of the fitted signal from the average signal:

$$\text{ESS} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \quad \text{where} \quad \hat{y} = X\hat{\beta} \quad \text{and} \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \quad (2)$$

with estimated regression coefficients $\hat{\beta}$, e.g. obtained via ordinary least squares (\rightarrow Proof III/1.1.2).

Sources:

- Wikipedia (2020): “Explained sum of squares”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-21; URL: https://en.wikipedia.org/wiki/Explained_sum_of_squares.

Metadata: ID: D38 | shortcut: ess | author: JoramSoch | date: 2020-03-21, 21:57.

1.1.6 Residual sum of squares

Definition: Let there be a multiple linear regression with independent observations (\rightarrow Definition III/1.1.1) using measured data y and design matrix X :

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

Then, the residual sum of squares (RSS) is defined as the sum of squared deviations of the measured signal from the fitted signal:

$$\text{RSS} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad \text{where} \quad \hat{y} = X\hat{\beta} \quad (2)$$

with estimated regression coefficients $\hat{\beta}$, e.g. obtained via ordinary least squares (\rightarrow Proof III/1.1.2).

Sources:

- Wikipedia (2020): “Residual sum of squares”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-21; URL: https://en.wikipedia.org/wiki/Residual_sum_of_squares.

Metadata: ID: D39 | shortcut: rss | author: JoramSoch | date: 2020-03-21, 22:03.

1.1.7 Total, explained and residual sum of squares

Theorem: Assume a linear regression model (\rightarrow Definition III/1.1.1) with independent observations

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

and let X contain a constant regressor 1_n modelling the intercept term. Then, it holds that

$$\text{TSS} = \text{ESS} + \text{RSS} \quad (2)$$

where TSS is the total sum of squares (\rightarrow Definition III/1.1.4), ESS is the explained sum of squares (\rightarrow Definition III/1.1.5) and RSS is the residual sum of squares (\rightarrow Definition III/1.1.6).

Proof: The total sum of squares (\rightarrow Definition III/1.1.4) is given by

$$\text{TSS} = \sum_{i=1}^n (y_i - \bar{y})^2 \quad (3)$$

where \bar{y} is the mean across all y_i . The TSS can be rewritten as

$$\begin{aligned} \text{TSS} &= \sum_{i=1}^n (y_i - \bar{y} + \hat{y}_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^n ((\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i))^2 \\ &= \sum_{i=1}^n ((\hat{y}_i - \bar{y}) + \hat{\varepsilon}_i)^2 \\ &= \sum_{i=1}^n ((\hat{y}_i - \bar{y})^2 + 2\hat{\varepsilon}_i(\hat{y}_i - \bar{y}) + \hat{\varepsilon}_i^2) \\ &= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n \hat{\varepsilon}_i^2 + 2 \sum_{i=1}^n \hat{\varepsilon}_i(\hat{y}_i - \bar{y}) \\ &= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n \hat{\varepsilon}_i^2 + 2 \sum_{i=1}^n \hat{\varepsilon}_i(x_i\hat{\beta} - \bar{y}) \\ &= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n \hat{\varepsilon}_i^2 + 2 \sum_{i=1}^n \hat{\varepsilon}_i \left(\sum_{j=1}^p x_{ij}\hat{\beta}_j \right) - 2 \sum_{i=1}^n \hat{\varepsilon}_i \bar{y} \\ &= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n \hat{\varepsilon}_i^2 + 2 \sum_{j=1}^p \hat{\beta}_j \sum_{i=1}^n \hat{\varepsilon}_i x_{ij} - 2\bar{y} \sum_{i=1}^n \hat{\varepsilon}_i \end{aligned} \quad (4)$$

The fact that the design matrix includes a constant regressor ensures that

$$\sum_{i=1}^n \hat{\varepsilon}_i = \hat{\varepsilon}^T \mathbf{1}_n = 0 \quad (5)$$

and because the residuals are orthogonal to the design matrix (\rightarrow Proof III/1.1.2), we have

$$\sum_{i=1}^n \hat{\varepsilon}_i x_{ij} = \hat{\varepsilon}^T \mathbf{x}_j = 0. \quad (6)$$

Applying (5) and (6) to (4), this becomes

$$\text{TSS} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n \hat{\varepsilon}_i^2 \quad (7)$$

and, with the definitions of explained (\rightarrow Definition III/1.1.5) and residual sum of squares (\rightarrow Definition III/1.1.6), it is

$$\text{TSS} = \text{ESS} + \text{RSS} . \quad (8)$$

Sources:

- Wikipedia (2020): “Partition of sums of squares”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-09; URL: https://en.wikipedia.org/wiki/Partition_of_sums_of_squares#Partitioning_the_sum_of_squares_in_linear_regression.

Metadata: ID: P76 | shortcut: mlr-pss | author: JoramSoch | date: 2019-09-27, 07:18.

1.1.8 Estimation, projection and residual-forming matrix

Theorem: Assume a linear regression model (\rightarrow Definition III/1.1.1) with independent observations

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

and consider estimation using ordinary least squares (\rightarrow Proof III/1.1.2). Then, the estimated parameters, fitted signal and residuals are given by

$$\begin{aligned} \hat{\beta} &= Ey \\ \hat{y} &= Py \\ \hat{\varepsilon} &= Ry \end{aligned} \quad (2)$$

where

$$\begin{aligned} E &= (X^T X)^{-1} X^T \\ P &= X(X^T X)^{-1} X^T \\ R &= I_n - X(X^T X)^{-1} X^T \end{aligned} \quad (3)$$

are the estimation matrix, projection matrix and residual-forming matrix and n is the number of observations.

Proof:

1) Ordinary least squares parameter estimates of β are defined as minimizing the residual sum of squares (\rightarrow Definition III/1.1.6)

$$\hat{\beta} = \arg \min_{\beta} [(y - X\beta)^T (y - X\beta)] \quad (4)$$

and the solution to this (\rightarrow Proof III/1.1.2) is given by

$$\begin{aligned} \hat{\beta} &= (X^T X)^{-1} X^T y \\ &\stackrel{(3)}{=} Ey . \end{aligned} \quad (5)$$

2) The fitted signal is given by multiplying the design matrix with the estimated regression coefficients

$$\hat{y} = X\hat{\beta} \quad (6)$$

and using (5), this becomes

$$\begin{aligned} \hat{y} &= X(X^T X)^{-1} X^T y \\ &\stackrel{(3)}{=} P y . \end{aligned} \quad (7)$$

3) The residuals of the model are calculated by subtracting the fitted signal from the measured signal

$$\hat{\varepsilon} = y - \hat{y} \quad (8)$$

and using (7), this becomes

$$\begin{aligned} \hat{\varepsilon} &= y - X(X^T X)^{-1} X^T y \\ &= (I_n - X(X^T X)^{-1} X^T) y \\ &\stackrel{(3)}{=} R y . \end{aligned} \quad (9)$$

Sources:

- Stephan, Klaas Enno (2010): “The General Linear Model (GLM)” in: *Methods and models for fMRI data analysis in neuroeconomics*, Lecture 3, Slide 10; URL: <http://www.socialbehavior.uzh.ch/teaching/methodspring10.html>.

Metadata: ID: P75 | shortcut: mlr-mat | author: JoramSoch | date: 2020-03-09, 21:18.

1.1.9 Weighted least squares

Theorem: Given a linear regression model (\rightarrow Definition III/1.1.1) with correlated observations

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

the parameters minimizing the weighted residual sum of squares (\rightarrow Definition III/1.1.6) are given by

$$\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} y . \quad (2)$$

Proof: Let there be an $n \times n$ square matrix W , such that

$$W V W^T = I_n . \quad (3)$$

Since V is a covariance matrix and thus symmetric, W is also symmetric and can be expressed the matrix square root of the inverse of V :

$$W W = V^{-1} \quad \Leftrightarrow \quad W = V^{-1/2} . \quad (4)$$

Left-multiplying the linear regression equation (1) with W , the linear transformation theorem (\rightarrow Definition “mvn-ltt”) implies that

$$Wy = WX\beta + W\varepsilon, \quad W\varepsilon \sim \mathcal{N}(0, \sigma^2 W V W^T) . \quad (5)$$

Applying (3), we see that (5) is actually a linear regression model (\rightarrow Definition III/1.1.1) with independent observations

$$\tilde{y} = \tilde{X}\beta + \tilde{\varepsilon}, \quad \tilde{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n) \quad (6)$$

where $\tilde{y} = Wy$, $\tilde{X} = WX$ and $\tilde{\varepsilon} = W\varepsilon$, such that we can apply the ordinary least squares solution (\rightarrow Proof III/1.1.2) giving

$$\begin{aligned} \hat{\beta} &= (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \tilde{y} \\ &= ((WX)^T WX)^{-1} (WX)^T Wy \\ &= (X^T W^T W X)^{-1} X^T W^T W y \\ &= (X^T W W X)^{-1} X^T W W y \\ &\stackrel{(4)}{=} (X^T V^{-1} X)^{-1} X^T V^{-1} y . \end{aligned} \quad (7)$$

which corresponds to the weighted least squares solution (2).

Sources:

- Stephan, Klaas Enno (2010): “The General Linear Model (GLM)”;
in: *Methods and models for fMRI data analysis in neuroeconomics*, Lecture 3, Slides 20/23; URL: <http://www.socialbehavior.uzh.ch/teaching/methodspring10.html>.

Metadata: ID: P77 | shortcut: mlr-wls | author: JoramSoch | date: 2020-03-11, 11:22.

1.1.10 Maximum likelihood estimation

Theorem: Given a linear regression model (\rightarrow Definition III/1.1.1) with correlated observations

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

the maximum likelihood estimates (\rightarrow Definition “mle”) of β and σ^2 are given by

$$\begin{aligned} \hat{\beta} &= (X^T V^{-1} X)^{-1} X^T V^{-1} y \\ \hat{\sigma}^2 &= \frac{1}{n} (y - X\hat{\beta})^T (y - X\hat{\beta}) . \end{aligned} \quad (2)$$

Proof: With the probability density function of the multivariate normal distribution (\rightarrow Proof II/4.1.2), the linear regression equation (1) implies the following likelihood function (\rightarrow Definition I/2.2.2)

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

and, using $|\sigma^2 V| = (\sigma^2)^n |V|$, the log-likelihood function (\rightarrow Definition “llf”)

$$\begin{aligned} \text{LL}(\beta, \sigma^2) &= \log p(y|\beta, \sigma^2) \\ &= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2} \log |V| \\ &\quad - \frac{1}{2\sigma^2} (y - X\beta)^T V^{-1} (y - X\beta) . \end{aligned} \quad (4)$$

Substituting the precision matrix $P = V^{-1}$ into (4) to ease notation, we have:

$$\begin{aligned} \text{LL}(\beta, \sigma^2) &= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2} \log(|V|) \\ &\quad - \frac{1}{2\sigma^2} (y^T P y - 2\beta^T X^T P y + \beta^T X^T P X \beta) . \end{aligned} \quad (5)$$

The derivative of the log-likelihood function (5) with respect to β is

$$\begin{aligned} \frac{d\text{LL}(\beta, \sigma^2)}{d\beta} &= \frac{d}{d\beta} \left(-\frac{1}{2\sigma^2} (y^T P y - 2\beta^T X^T P y + \beta^T X^T P X \beta) \right) \\ &= \frac{1}{2\sigma^2} \frac{d}{d\beta} (2\beta^T X^T P y - \beta^T X^T P X \beta) \\ &= \frac{1}{2\sigma^2} (2X^T P y - 2X^T P X \beta) \\ &= \frac{1}{\sigma^2} (X^T P y - X^T P X \beta) \end{aligned} \quad (6)$$

and setting this derivative to zero gives the MLE for β :

$$\begin{aligned} \frac{d\text{LL}(\hat{\beta}, \sigma^2)}{d\beta} &= 0 \\ 0 &= \frac{1}{\sigma^2} (X^T P y - X^T P X \hat{\beta}) \\ 0 &= X^T P y - X^T P X \hat{\beta} \\ X^T P X \hat{\beta} &= X^T P y \\ \hat{\beta} &= (X^T P X)^{-1} X^T P y \end{aligned} \quad (7)$$

The derivative of the log-likelihood function (4) at $\hat{\beta}$ with respect to σ^2 is

$$\begin{aligned} \frac{d\text{LL}(\hat{\beta}, \sigma^2)}{d\sigma^2} &= \frac{d}{d\sigma^2} \left(-\frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} (y - X\hat{\beta})^T V^{-1} (y - X\hat{\beta}) \right) \\ &= -\frac{n}{2} \frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2} (y - X\hat{\beta})^T V^{-1} (y - X\hat{\beta}) \\ &= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} (y - X\hat{\beta})^T V^{-1} (y - X\hat{\beta}) \end{aligned} \quad (8)$$

and setting this derivative to zero gives the MLE for σ^2 :

$$\begin{aligned}
\frac{dLL(\hat{\beta}, \hat{\sigma}^2)}{d\sigma^2} &= 0 \\
0 &= -\frac{n}{2\hat{\sigma}^2} + \frac{1}{2(\hat{\sigma}^2)^2}(y - X\hat{\beta})^T V^{-1}(y - X\hat{\beta}) \\
\frac{n}{2\hat{\sigma}^2} &= \frac{1}{2(\hat{\sigma}^2)^2}(y - X\hat{\beta})^T V^{-1}(y - X\hat{\beta}) \\
\frac{2(\hat{\sigma}^2)^2}{n} \cdot \frac{n}{2\hat{\sigma}^2} &= \frac{2(\hat{\sigma}^2)^2}{n} \cdot \frac{1}{2(\hat{\sigma}^2)^2}(y - X\hat{\beta})^T V^{-1}(y - X\hat{\beta}) \\
\hat{\sigma}^2 &= \frac{1}{n}(y - X\hat{\beta})^T V^{-1}(y - X\hat{\beta})
\end{aligned} \tag{9}$$

Together, (7) and (9) constitute the MLE for multiple linear regression.

Sources:

- original work

Metadata: ID: P78 | shortcut: mlr-mle | author: JoramSoch | date: 2020-03-11, 12:27.

1.2 Bayesian linear regression

1.2.1 Conjugate prior distribution

Theorem: Let

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

be a linear regression model (\rightarrow Definition III/1.1.1) 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 (\rightarrow Definition “prior-conj”) for this model is a normal-gamma distribution (\rightarrow Definition II/4.2.1)

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

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

Proof: By definition, a conjugate prior (\rightarrow Definition “prior-conj”) is a prior distribution (\rightarrow Definition I/2.2.3) that, when combined with the likelihood function (\rightarrow Definition I/2.2.2), leads to a posterior distribution (\rightarrow Definition I/2.2.6) that belongs to the same family of probability distributions (\rightarrow 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 (\rightarrow Definition I/2.2.2)

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

which, for mathematical convenience, can also be parametrized as

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

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

Separating constant and variable terms, we have:

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

Expanding the product in the exponent, we have:

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

Completing the square over β , finally gives

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

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

In other words, the likelihood function (\rightarrow 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 :

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

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

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

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

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

exhibits the same proportionality

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

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.

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

Collecting identical variables gives:

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

Expanding the products in the exponent gives:

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

Completing the square over β , we finally have

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

with the posterior hyperparameters (\rightarrow Definition I/2.2.6)

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

Ergo, the joint likelihood is proportional to

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

with the posterior hyperparameters (\rightarrow Definition I/2.2.6)

$$\begin{aligned}
a_n &= a_0 + \frac{n}{2} \\
b_n &= b_0 + \frac{1}{2} (y^T P y + \mu_0^T \Lambda_0 \mu_0 - \mu_n^T \Lambda_n \mu_n) .
\end{aligned} \tag{15}$$

From the term in (14), we can isolate the posterior distribution over β given τ :

$$p(\beta|\tau, y) = \mathcal{N}(\beta; \mu_n, (\tau\Lambda_n)^{-1}) . \quad (16)$$

From the remaining term, we can isolate the posterior distribution over τ :

$$p(\tau|y) = \text{Gam}(\tau; a_n, b_n) . \quad (17)$$

Together, (16) and (17) constitute the joint (\rightarrow Definition “prob-joint”) posterior distribution (\rightarrow 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

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

be a linear regression model (\rightarrow Definition III/1.1.1) 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 (\rightarrow Proof III/1.2.1) over the model parameters β and $\tau = 1/\sigma^2$:

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

Then, the log model evidence (\rightarrow Definition IV/3.1.1) for this model is

$$\begin{aligned} \log p(y|m) = & \frac{1}{2} \log |P| - \frac{n}{2} \log(2\pi) + \frac{1}{2} \log |\Lambda_0| - \frac{1}{2} \log |\Lambda_n| + \\ & \log \Gamma(a_n) - \log \Gamma(a_0) + a_0 \log b_0 - a_n \log b_n \end{aligned} \quad (3)$$

where the posterior hyperparameters (\rightarrow Definition I/2.2.6) are given by

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

Proof: According to the law of marginal probability (\rightarrow Definition “prob-marg”), the model evidence (\rightarrow Definition I/2.2.7) for this model is:

$$p(y|m) = \iint p(y|\beta, \tau) p(\beta, \tau) d\beta d\tau . \quad (5)$$

According to the law of conditional probability (\rightarrow Definition “prob-cond”), the integrand is equivalent to the joint likelihood (\rightarrow Definition I/2.2.5):

$$p(y|m) = \iint p(y, \beta, \tau) d\beta d\tau . \quad (6)$$

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

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

which, for mathematical convenience, can also be parametrized as

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

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

When deriving the posterior distribution (\rightarrow Proof III/1.2.2) $p(\beta, \tau|y)$, the joint likelihood $p(y, \beta, \tau)$ is obtained as

$$p(y, \beta, \tau) = \sqrt{\frac{\tau^n |P|}{(2\pi)^n}} \sqrt{\frac{\tau^p |\Lambda_0|}{(2\pi)^p}} \frac{b_0^{a_0}}{\Gamma(a_0)} \tau^{a_0-1} \exp[-b_0 \tau] \cdot \exp \left[-\frac{\tau}{2} ((\beta - \mu_n)^T \Lambda_n (\beta - \mu_n) + (y^T P y + \mu_0^T \Lambda_0 \mu_0 - \mu_n^T \Lambda_n \mu_n)) \right] . \quad (9)$$

Using the probability density function of the multivariate normal distribution (\rightarrow Proof II/4.1.2), we can rewrite this as

$$p(y, \beta, \tau) = \sqrt{\frac{\tau^n |P|}{(2\pi)^n}} \sqrt{\frac{\tau^p |\Lambda_0|}{(2\pi)^p}} \sqrt{\frac{(2\pi)^p}{\tau^p |\Lambda_n|}} \frac{b_0^{a_0}}{\Gamma(a_0)} \tau^{a_0-1} \exp[-b_0 \tau] \cdot \mathcal{N}(\beta; \mu_n, (\tau \Lambda_n)^{-1}) \exp \left[-\frac{\tau}{2} (y^T P y + \mu_0^T \Lambda_0 \mu_0 - \mu_n^T \Lambda_n \mu_n) \right] . \quad (10)$$

Now, β can be integrated out easily:

$$\int p(y, \beta, \tau) d\beta = \sqrt{\frac{\tau^n |P|}{(2\pi)^n}} \sqrt{\frac{|\Lambda_0|}{|\Lambda_n|}} \frac{b_0^{a_0}}{\Gamma(a_0)} \tau^{a_0-1} \exp[-b_0 \tau] \cdot \exp \left[-\frac{\tau}{2} (y^T P y + \mu_0^T \Lambda_0 \mu_0 - \mu_n^T \Lambda_n \mu_n) \right] . \quad (11)$$

Using the probability density function of the gamma distribution (\rightarrow Proof II/3.3.2), we can rewrite this as

$$\int p(y, \beta, \tau) d\beta = \sqrt{\frac{|P|}{(2\pi)^n}} \sqrt{\frac{|\Lambda_0|}{|\Lambda_n|}} \frac{b_0^{a_0}}{\Gamma(a_0)} \frac{\Gamma(a_n)}{b_n^{a_n}} \text{Gam}(\tau; a_n, b_n) . \quad (12)$$

Finally, τ can also be integrated out:

$$\iint p(y, \beta, \tau) d\beta d\tau = \sqrt{\frac{|P|}{(2\pi)^n}} \sqrt{\frac{|\Lambda_0|}{|\Lambda_n|}} \frac{\Gamma(a_n)}{\Gamma(a_0)} \frac{b_0^{a_0}}{b_n^{a_n}} = p(y|m) . \quad (13)$$

Thus, the log model evidence (\rightarrow Definition IV/3.1.1) of this model is given by

$$\begin{aligned} \log p(y|m) = & \frac{1}{2} \log |P| - \frac{n}{2} \log(2\pi) + \frac{1}{2} \log |\Lambda_0| - \frac{1}{2} \log |\Lambda_n| + \\ & \log \Gamma(a_n) - \log \Gamma(a_0) + a_0 \log b_0 - a_n \log b_n . \end{aligned} \quad (14)$$

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 Definition

Definition: Let Y be an $n \times v$ matrix and let X be an $n \times p$ matrix. Then, a statement asserting a linear mapping from X to Y with parameters B and matrix-normally distributed (\rightarrow Definition II/5.1.1) errors E

$$Y = XB + E, \quad E \sim \mathcal{MN}(0, V, \Sigma) \quad (1)$$

is called a multivariate linear regression model or simply, “general linear model”.

- Y is called “data matrix”, “set of dependent variables” or “measurements”;
- X is called “design matrix”, “set of independent variables” or “predictors”;
- B are called “regression coefficients” or “weights”;
- E is called “noise matrix” or “error terms”;
- V is called “covariance across rows”;
- Σ is called “covariance across columns”;
- n is the number of observations;
- v is the number of measurements;
- p is the number of predictors.

When rows of Y correspond to units of time, e.g. subsequent measurements, V is called “temporal covariance”. When columns of Y correspond to units of space, e.g. measurement channels, Σ is called “spatial covariance”.

When the covariance matrix V is a scalar multiple of the $n \times n$ identity matrix, this is called a general linear model with independent and identically distributed (i.i.d.) observations:

$$V = \lambda I_n \quad \Rightarrow \quad E \sim \mathcal{MN}(0, \lambda I_n, \Sigma) \quad \Rightarrow \quad \varepsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \lambda \Sigma) . \quad (2)$$

Otherwise, it is called a general linear model with correlated observations.

Sources:

- Wikipedia (2020): “General linear model”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-21; URL: https://en.wikipedia.org/wiki/General_linear_model.

Metadata: ID: D40 | shortcut: glm | author: JoramSoch | date: 2020-03-21, 22:24.

1.3.2 Maximum likelihood estimation

Theorem: Given a general linear model (\rightarrow Definition III/1.3.1) with matrix-normally distributed (\rightarrow Definition II/5.1.1) errors

$$Y = XB + E, \quad E \sim \mathcal{MN}(0, V, \Sigma), \quad (1)$$

maximum likelihood estimates (\rightarrow Definition “mle”) for the unknown parameters B and Σ are given by

$$\begin{aligned} \hat{B} &= (X^T V^{-1} X)^{-1} X^T V^{-1} Y \\ \hat{\Sigma} &= \frac{1}{n} (Y - X \hat{B})^T V^{-1} (Y - X \hat{B}). \end{aligned} \quad (2)$$

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 (\rightarrow Definition I/2.2.2)

$$\begin{aligned} p(Y|B, \Sigma) &= \mathcal{MN}(Y; XB, V, \Sigma) \\ &= \sqrt{\frac{1}{(2\pi)^{nv} |\Sigma|^n |V|^v}} \cdot \exp \left[-\frac{1}{2} \text{tr} (\Sigma^{-1} (Y - XB)^T V^{-1} (Y - XB)) \right] \end{aligned} \quad (3)$$

and the log-likelihood function (\rightarrow Definition “llf”)

$$\begin{aligned} \text{LL}(B, \Sigma) &= \log p(Y|B, \Sigma) \\ &= -\frac{nv}{2} \log(2\pi) - \frac{n}{2} \log |\Sigma| - \frac{v}{2} \log |V| \\ &\quad - \frac{1}{2} \text{tr} [\Sigma^{-1} (Y - XB)^T V^{-1} (Y - XB)] . \end{aligned} \quad (4)$$

Substituting V^{-1} by the precision matrix P to ease notation, we have:

$$\begin{aligned} \text{LL}(B, \Sigma) &= -\frac{nv}{2} \log(2\pi) - \frac{n}{2} \log(|\Sigma|) - \frac{v}{2} \log(|V|) \\ &\quad - \frac{1}{2} \text{tr} [\Sigma^{-1} (Y^T P Y - Y^T P X B - B^T X^T P Y + B^T X^T P X B)] . \end{aligned} \quad (5)$$

The derivative of the log-likelihood function (5) with respect to B is

$$\begin{aligned}
\frac{dLL(B, \Sigma)}{dB} &= \frac{d}{dB} \left(-\frac{1}{2} \text{tr} [\Sigma^{-1} (Y^T PY - Y^T P X B - B^T X^T P Y + B^T X^T P X B)] \right) \\
&= \frac{d}{dB} \left(-\frac{1}{2} \text{tr} [-2 \Sigma^{-1} Y^T P X B] \right) + \frac{d}{dB} \left(-\frac{1}{2} \text{tr} [\Sigma^{-1} B^T X^T P X B] \right) \\
&= -\frac{1}{2} (-2 X^T P Y \Sigma^{-1}) - \frac{1}{2} (X^T P X B \Sigma^{-1} + (X^T P X)^T B (\Sigma^{-1})^T) \\
&= X^T P Y \Sigma^{-1} - X^T P X B \Sigma^{-1}
\end{aligned} \tag{6}$$

and setting this derivative to zero gives the MLE for B :

$$\begin{aligned}
\frac{dLL(\hat{B}, \Sigma)}{dB} &= 0 \\
0 &= X^T P Y \Sigma^{-1} - X^T P X \hat{B} \Sigma^{-1} \\
0 &= X^T P Y - X^T P X \hat{B} \\
X^T P X \hat{B} &= X^T P Y \\
\hat{B} &= (X^T P X)^{-1} X^T P Y
\end{aligned} \tag{7}$$

The derivative of the log-likelihood function (4) at \hat{B} with respect to Σ is

$$\begin{aligned}
\frac{dLL(\hat{B}, \Sigma)}{d\Sigma} &= \frac{d}{d\Sigma} \left(-\frac{n}{2} \log |\Sigma| - \frac{1}{2} \text{tr} [\Sigma^{-1} (Y - X \hat{B})^T V^{-1} (Y - X \hat{B})] \right) \\
&= -\frac{n}{2} (\Sigma^{-1})^T + \frac{1}{2} \left(\Sigma^{-1} (Y - X \hat{B})^T V^{-1} (Y - X \hat{B}) \Sigma^{-1} \right)^T \\
&= -\frac{n}{2} \Sigma^{-1} + \frac{1}{2} \Sigma^{-1} (Y - X \hat{B})^T V^{-1} (Y - X \hat{B}) \Sigma^{-1}
\end{aligned} \tag{8}$$

and setting this derivative to zero gives the MLE for Σ :

$$\begin{aligned}
\frac{dLL(\hat{B}, \hat{\Sigma})}{d\Sigma} &= 0 \\
0 &= -\frac{n}{2} \hat{\Sigma}^{-1} + \frac{1}{2} \hat{\Sigma}^{-1} (Y - X \hat{B})^T V^{-1} (Y - X \hat{B}) \hat{\Sigma}^{-1} \\
\frac{n}{2} \hat{\Sigma}^{-1} &= \frac{1}{2} \hat{\Sigma}^{-1} (Y - X \hat{B})^T V^{-1} (Y - X \hat{B}) \hat{\Sigma}^{-1} \\
\hat{\Sigma}^{-1} &= \frac{1}{n} \hat{\Sigma}^{-1} (Y - X \hat{B})^T V^{-1} (Y - X \hat{B}) \hat{\Sigma}^{-1} \\
I_v &= \frac{1}{n} (Y - X \hat{B})^T V^{-1} (Y - X \hat{B}) \hat{\Sigma}^{-1} \\
\hat{\Sigma} &= \frac{1}{n} (Y - X \hat{B})^T V^{-1} (Y - X \hat{B})
\end{aligned} \tag{9}$$

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 Definition

Definition: Poisson-distributed data are defined as a set of observed counts $y = \{y_1, \dots, y_n\}$, independent and identically distributed according to a Poisson distribution (\rightarrow Definition “poiss”) with rate λ :

$$y_i \sim \text{Poiss}(\lambda), \quad i = 1, \dots, n. \quad (1)$$

Sources:

- Wikipedia (2020): “Poisson distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-22; URL: https://en.wikipedia.org/wiki/Poisson_distribution#Parameter_estimation.

Metadata: ID: D41 | shortcut: poiss-data | author: JoramSoch | date: 2020-03-22, 22:50.

2.1.2 Maximum likelihood estimation

Theorem: Let there be a Poisson-distributed data (\rightarrow Definition III/2.1.1) set $y = \{y_1, \dots, y_n\}$:

$$y_i \sim \text{Poiss}(\lambda), \quad i = 1, \dots, n. \quad (1)$$

Then, the maximum likelihood estimate (\rightarrow Definition “mle”) for the rate parameter λ is given by

$$\hat{\lambda} = \bar{y} \quad (2)$$

where \bar{y} is the sample mean (\rightarrow Proof “mean-sample”)

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i. \quad (3)$$

Proof: The likelihood function (\rightarrow Definition I/2.2.2) for each observation is given by the probability mass function of the Poisson distribution (\rightarrow Proof “poiss-pdf”)

$$p(y_i|\lambda) = \text{Poiss}(y_i; \lambda) = \frac{\lambda^{y_i} \cdot \exp(-\lambda)}{y_i!} \quad (4)$$

and because observations are independent (\rightarrow Definition “ind”), the likelihood function for all observations is the product of the individual ones:

$$p(y|\lambda) = \prod_{i=1}^n p(y_i|\lambda) = \prod_{i=1}^n \frac{\lambda^{y_i} \cdot \exp(-\lambda)}{y_i!}. \quad (5)$$

Thus, the log-likelihood function (\rightarrow Definition “llf”) is

$$\text{LL}(\lambda) = \log p(y|\lambda) = \log \left[\prod_{i=1}^n \frac{\lambda^{y_i} \cdot \exp(-\lambda)}{y_i!} \right] \quad (6)$$

which can be developed into

$$\begin{aligned}
\text{LL}(\lambda) &= \sum_{i=1}^n \log \left[\frac{\lambda^{y_i} \cdot \exp(-\lambda)}{y_i!} \right] \\
&= \sum_{i=1}^n [y_i \cdot \log(\lambda) - \lambda - \log(y_i!)] \\
&= -\sum_{i=1}^n \lambda + \sum_{i=1}^n y_i \cdot \log(\lambda) - \sum_{i=1}^n \log(y_i!) \\
&= -n\lambda + \log(\lambda) \sum_{i=1}^n y_i - \sum_{i=1}^n \log(y_i!)
\end{aligned} \tag{7}$$

The derivatives of the log-likelihood with respect to λ are

$$\begin{aligned}
\frac{d\text{LL}(\lambda)}{d\lambda} &= \frac{1}{\lambda} \sum_{i=1}^n y_i - n \\
\frac{d^2\text{LL}(\lambda)}{d\lambda^2} &= -\frac{1}{\lambda^2} \sum_{i=1}^n y_i .
\end{aligned} \tag{8}$$

Setting the first derivative to zero, we obtain:

$$\begin{aligned}
\frac{d\text{LL}(\hat{\lambda})}{d\lambda} &= 0 \\
0 &= \frac{1}{\hat{\lambda}} \sum_{i=1}^n y_i - n \\
\hat{\lambda} &= \frac{1}{n} \sum_{i=1}^n y_i = \bar{y} .
\end{aligned} \tag{9}$$

Plugging this value into the second derivative, we confirm:

$$\begin{aligned}
\frac{d^2\text{LL}(\hat{\lambda})}{d\lambda^2} &= -\frac{1}{\bar{y}^2} \sum_{i=1}^n y_i \\
&= -\frac{n \cdot \bar{y}}{\bar{y}^2} \\
&= -\frac{n}{\bar{y}} < 0 .
\end{aligned} \tag{10}$$

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 Definition

Definition: A Poisson distribution with exposure values is defined as a set of observed counts $y = \{y_1, \dots, y_n\}$, independently distributed according to a Poisson distribution (\rightarrow Definition “poiss”) with common rate λ and a set of concurrent exposures $x = \{x_1, \dots, x_n\}$:

$$y_i \sim \text{Poiss}(\lambda x_i), \quad i = 1, \dots, n. \quad (1)$$

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. 2.14; URL: <http://www.stat.columbia.edu/~gelman/book/>.

Metadata: ID: D42 | shortcut: poissexp | author: JoramSoch | date: 2020-03-22, 22:57.

2.2.2 Conjugate prior distribution

Theorem: Consider data $y = \{y_1, \dots, y_n\}$ following a Poisson distribution with exposure values (\rightarrow Definition III/2.2.1):

$$y_i \sim \text{Poiss}(\lambda x_i), \quad i = 1, \dots, n. \quad (1)$$

Then, the conjugate prior (\rightarrow Definition “prior-conj”) for the model parameter λ is a gamma distribution (\rightarrow Definition II/3.3.1):

$$p(\lambda) = \text{Gam}(\lambda; a_0, b_0). \quad (2)$$

Proof: With the probability mass function of the Poisson distribution (\rightarrow Proof “poiss-pmf”), the likelihood function (\rightarrow Definition I/2.2.2) for each observation implied by (1) is given by

$$p(y_i|\lambda) = \text{Poiss}(y_i; \lambda x_i) = \frac{(\lambda x_i)^{y_i} \cdot \exp[-\lambda x_i]}{y_i!} \quad (3)$$

and because observations are independent (\rightarrow Definition “ind”), the likelihood function for all observations is the product of the individual ones:

$$p(y|\lambda) = \prod_{i=1}^n p(y_i|\lambda) = \prod_{i=1}^n \frac{(\lambda x_i)^{y_i} \cdot \exp[-\lambda x_i]}{y_i!}. \quad (4)$$

Resolving the product in the likelihood function, we have

$$\begin{aligned} p(y|\lambda) &= \prod_{i=1}^n \frac{x_i^{y_i}}{y_i!} \cdot \prod_{i=1}^n \lambda^{y_i} \cdot \prod_{i=1}^n \exp[-\lambda x_i] \\ &= \prod_{i=1}^n \left(\frac{x_i^{y_i}}{y_i!} \right) \cdot \lambda^{\sum_{i=1}^n y_i} \cdot \exp \left[-\lambda \sum_{i=1}^n x_i \right] \\ &= \prod_{i=1}^n \left(\frac{x_i^{y_i}}{y_i!} \right) \cdot \lambda^{n\bar{y}} \cdot \exp[-n\bar{x}\lambda] \end{aligned} \quad (5)$$

where \bar{y} and \bar{x} are the means (\rightarrow 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 :

$$p(y|\lambda) \propto \lambda^{n\bar{y}} \cdot \exp[-n\bar{x}\lambda] .\tag{7}$$

The same is true for a gamma distribution over λ

$$p(\lambda) = \text{Gam}(\lambda; a_0, b_0)\tag{8}$$

the probability density function of which (\rightarrow 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.

2.2.3 Posterior distribution

Theorem: Consider data $y = \{y_1, \dots, y_n\}$ following a Poisson distribution with exposure values (\rightarrow Definition III/2.2.1):

$$y_i \sim \text{Poiss}(\lambda x_i), \quad i = 1, \dots, n .\tag{1}$$

Moreover, assume a gamma prior distribution (\rightarrow Proof III/2.2.2) over the model parameter λ :

$$p(\lambda) = \text{Gam}(\lambda; a_0, b_0) .\tag{2}$$

Then, the posterior distribution (\rightarrow Definition I/2.2.6) is also a gamma distribution (\rightarrow Definition II/3.3.1)

$$p(\lambda|y) = \text{Gam}(\lambda; a_n, b_n)\tag{3}$$

and the posterior hyperparameters (\rightarrow Definition I/2.2.6) 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 (\rightarrow Proof “poiss-pmf”), the likelihood function (\rightarrow Definition I/2.2.2) for each observation implied by (1) is given by

$$p(y_i|\lambda) = \text{Poiss}(y_i; \lambda x_i) = \frac{(\lambda x_i)^{y_i} \cdot \exp[-\lambda x_i]}{y_i!} \tag{5}$$

and because observations are independent (\rightarrow Definition “ind”), the likelihood function for all observations is the product of the individual ones:

$$p(y|\lambda) = \prod_{i=1}^n p(y_i|\lambda) = \prod_{i=1}^n \frac{(\lambda x_i)^{y_i} \cdot \exp[-\lambda x_i]}{y_i!} . \tag{6}$$

Combining the likelihood function (6) with the prior distribution (2), the joint likelihood (\rightarrow Definition I/2.2.5) of the model is given by

$$\begin{aligned} p(y, \lambda) &= p(y|\lambda) p(\lambda) \\ &= \prod_{i=1}^n \frac{(\lambda x_i)^{y_i} \cdot \exp[-\lambda x_i]}{y_i!} \cdot \frac{b_0^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0 \lambda] . \end{aligned} \tag{7}$$

Resolving the product in the joint likelihood, we have

$$\begin{aligned} p(y, \lambda) &= \prod_{i=1}^n \frac{x_i^{y_i}}{y_i!} \prod_{i=1}^n \lambda^{y_i} \prod_{i=1}^n \exp[-\lambda x_i] \cdot \frac{b_0^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0 \lambda] \\ &= \prod_{i=1}^n \left(\frac{x_i^{y_i}}{y_i!} \right) \lambda^{\sum_{i=1}^n y_i} \exp \left[-\lambda \sum_{i=1}^n x_i \right] \cdot \frac{b_0^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0 \lambda] \\ &= \prod_{i=1}^n \left(\frac{x_i^{y_i}}{y_i!} \right) \lambda^{n\bar{y}} \exp[-n\bar{x}\lambda] \cdot \frac{b_0^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0 \lambda] \\ &= \prod_{i=1}^n \left(\frac{x_i^{y_i}}{y_i!} \right) \frac{b_0^{a_0}}{\Gamma(a_0)} \cdot \lambda^{a_0+n\bar{y}-1} \cdot \exp[-(b_0 + n\bar{x})\lambda] \end{aligned} \tag{8}$$

where \bar{y} and \bar{x} are the means (\rightarrow 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 (\rightarrow Proof “post-jl”):

$$p(\lambda|y) \propto p(y, \lambda) . \quad (10)$$

Setting $a_n = a_0 + n\bar{y}$ and $b_n = b_0 + n\bar{x}$, the posterior distribution is therefore proportional to

$$p(\lambda|y) \propto \lambda^{a_n-1} \cdot \exp[-b_n \lambda] \quad (11)$$

which, when normalized to one, results in the probability density function of the gamma distribution (\rightarrow Proof II/3.3.2):

$$p(\lambda|y) = \frac{b_n^{a_n}}{\Gamma(a_n)} \lambda^{a_n-1} \exp[-b_n \lambda] = \text{Gam}(\lambda; a_n, b_n) . \quad (12)$$

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.

2.2.4 Log model evidence

Theorem: Consider data $y = \{y_1, \dots, y_n\}$ following a Poisson distribution with exposure values (\rightarrow Definition III/2.2.1):

$$y_i \sim \text{Poiss}(\lambda x_i), \quad i = 1, \dots, n . \quad (1)$$

Moreover, assume a gamma prior distribution (\rightarrow Proof III/2.2.2) over the model parameter λ :

$$p(\lambda) = \text{Gam}(\lambda; a_0, b_0) . \quad (2)$$

Then, the log model evidence (\rightarrow Definition IV/3.1.1) for this model is

$$\begin{aligned} \log p(y|m) &= \sum_{i=1}^n y_i \log(x_i) - \sum_{i=1}^n \log y_i! + \\ &\quad \log \Gamma(a_n) - \log \Gamma(a_0) + a_0 \log b_0 - a_n \log b_n . \end{aligned} \quad (3)$$

where the posterior hyperparameters (\rightarrow Definition I/2.2.6) are given by

$$\begin{aligned} a_n &= a_0 + n\bar{y} \\ b_n &= b_0 + n\bar{x} . \end{aligned} \quad (4)$$

Proof: With the probability mass function of the Poisson distribution (\rightarrow Proof “poiss-pmf”), the likelihood function (\rightarrow Definition I/2.2.2) for each observation implied by (1) is given by

$$p(y_i|\lambda) = \text{Poiss}(y_i; \lambda x_i) = \frac{(\lambda x_i)^{y_i} \cdot \exp[-\lambda x_i]}{y_i!} \quad (5)$$

and because observations are independent (\rightarrow Definition “ind”), the likelihood function for all observations is the product of the individual ones:

$$p(y|\lambda) = \prod_{i=1}^n p(y_i|\lambda) = \prod_{i=1}^n \frac{(\lambda x_i)^{y_i} \cdot \exp[-\lambda x_i]}{y_i!} . \quad (6)$$

Combining the likelihood function (6) with the prior distribution (2), the joint likelihood (\rightarrow Definition I/2.2.5) of the model is given by

$$\begin{aligned} p(y, \lambda) &= p(y|\lambda) p(\lambda) \\ &= \prod_{i=1}^n \frac{(\lambda x_i)^{y_i} \cdot \exp[-\lambda x_i]}{y_i!} \cdot \frac{b_0^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0 \lambda] . \end{aligned} \quad (7)$$

Resolving the product in the joint likelihood, we have

$$\begin{aligned} p(y, \lambda) &= \prod_{i=1}^n \frac{x_i^{y_i}}{y_i!} \prod_{i=1}^n \lambda^{y_i} \prod_{i=1}^n \exp[-\lambda x_i] \cdot \frac{b_0^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0 \lambda] \\ &= \prod_{i=1}^n \left(\frac{x_i^{y_i}}{y_i!} \right) \lambda^{\sum_{i=1}^n y_i} \exp \left[-\lambda \sum_{i=1}^n x_i \right] \cdot \frac{b_0^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0 \lambda] \\ &= \prod_{i=1}^n \left(\frac{x_i^{y_i}}{y_i!} \right) \lambda^{n\bar{y}} \exp[-n\bar{x}\lambda] \cdot \frac{b_0^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0 \lambda] \\ &= \prod_{i=1}^n \left(\frac{x_i^{y_i}}{y_i!} \right) \frac{b_0^{a_0}}{\Gamma(a_0)} \cdot \lambda^{a_0+n\bar{y}-1} \cdot \exp[-(b_0 + n\bar{x})\lambda] \end{aligned} \quad (8)$$

where \bar{y} and \bar{x} are the means (\rightarrow Proof “mean-sample”) of y and x respectively:

$$\begin{aligned} \bar{y} &= \frac{1}{n} \sum_{i=1}^n y_i \\ \bar{x} &= \frac{1}{n} \sum_{i=1}^n x_i . \end{aligned} \quad (9)$$

Note that the model evidence is the marginal density of the joint likelihood (\rightarrow Definition I/2.2.7):

$$p(y) = \int p(y, \lambda) d\lambda . \quad (10)$$

Setting $a_n = a_0 + n\bar{y}$ and $b_n = b_0 + n\bar{x}$, the joint likelihood can also be written as

$$p(y, \lambda) = \prod_{i=1}^n \left(\frac{x_i^{y_i}}{y_i!} \right) \frac{b_0^{a_0}}{\Gamma(a_0)} \frac{\Gamma(a_n)}{b_n^{a_n}} \cdot \frac{b_n^{a_n}}{\Gamma(a_n)} \lambda^{a_n-1} \exp[-b_n \lambda] . \quad (11)$$

Using the probability density function of the gamma distribution (\rightarrow Proof II/3.3.2), λ can now be integrated out easily

$$\begin{aligned}
p(y) &= \int \prod_{i=1}^n \left(\frac{x_i^{y_i}}{y_i!} \right) \frac{b_0^{a_0}}{\Gamma(a_0)} \frac{\Gamma(a_n)}{b_n^{a_n}} \cdot \frac{b_n^{a_n}}{\Gamma(a_n)} \lambda^{a_n-1} \exp[-b_n \lambda] \, d\lambda \\
&= \prod_{i=1}^n \left(\frac{x_i^{y_i}}{y_i!} \right) \frac{\Gamma(a_n)}{\Gamma(a_0)} \frac{b_0^{a_0}}{b_n^{a_n}} \int \text{Gam}(\lambda; a_n, b_n) \, d\lambda \\
&= \prod_{i=1}^n \left(\frac{x_i^{y_i}}{y_i!} \right) \frac{\Gamma(a_n)}{\Gamma(a_0)} \frac{b_0^{a_0}}{b_n^{a_n}} ,
\end{aligned} \tag{12}$$

such that the log model evidence (\rightarrow Definition IV/3.1.1) is shown to be

$$\begin{aligned}
\log p(y|m) &= \sum_{i=1}^n y_i \log(x_i) - \sum_{i=1}^n \log y_i! + \\
&\quad \log \Gamma(a_n) - \log \Gamma(a_0) + a_0 \log b_0 - a_n \log b_n .
\end{aligned} \tag{13}$$

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 $y = \{y_1, \dots, y_n\}$ be a set of observed counts independent and identically distributed (\rightarrow Definition “iid”) according to a beta distribution (\rightarrow Definition “beta”) with shapes α and β :

$$y_i \sim \text{Bet}(\alpha, \beta), \quad i = 1, \dots, n. \quad (1)$$

Then, the method-of-moments estimates (\rightarrow Definition “mom”) for the shape parameters α and β are given by

$$\begin{aligned} \hat{\alpha} &= \bar{y} \left(\frac{\bar{y}(1 - \bar{y})}{\bar{v}} - 1 \right) \\ \hat{\beta} &= (1 - \bar{y}) \left(\frac{\bar{y}(1 - \bar{y})}{\bar{v}} - 1 \right) \end{aligned} \quad (2)$$

where \bar{y} is the sample mean (\rightarrow Proof “mean-sample”) and \bar{v} is the unbiased sample variance (\rightarrow Proof IV/1.1.3):

$$\begin{aligned} \bar{y} &= \frac{1}{n} \sum_{i=1}^n y_i \\ \bar{v} &= \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2. \end{aligned} \quad (3)$$

Proof: Mean (\rightarrow Proof “beta-mean”) and variance (\rightarrow Proof “beta-var”) of the beta distribution (\rightarrow Definition “beta”) in terms of the parameters α and β are given by

$$\begin{aligned} E(X) &= \frac{\alpha}{\alpha + \beta} \\ \text{Var}(X) &= \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}. \end{aligned} \quad (4)$$

Thus, matching the moments (\rightarrow Definition “mom”) requires us to solve the following equation system for α and β :

$$\begin{aligned} \bar{y} &= \frac{\alpha}{\alpha + \beta} \\ \bar{v} &= \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}. \end{aligned} \quad (5)$$

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:

$$\begin{aligned}
\bar{v} &= \frac{\alpha \cdot \alpha q}{(\alpha + \alpha q)^2 (\alpha + \alpha q + 1)} \\
&= \frac{\alpha^2 q}{(\alpha(1 + q))^2 (\alpha(1 + q) + 1)} \\
&= \frac{q}{(1 + q)^2 (\alpha(1 + q) + 1)} \\
&= \frac{q}{\alpha(1 + q)^3 + (1 + q)^2} \\
q &= \bar{v} [\alpha(1 + q)^3 + (1 + q)^2] .
\end{aligned} \tag{7}$$

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 (\rightarrow Definition “logreg”)

$$l_i = x_i\beta + \varepsilon_i, \quad i = 1, \dots, n \quad (1)$$

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

$$\Pr(y_i = 1) = \frac{1}{1 + b^{-(x_i\beta + \varepsilon_i)}} \quad (2)$$

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

$$l_i = \log_b \frac{p_i}{1 - p_i} \quad (3)$$

and using (1), we have

$$\begin{aligned} \log_b \frac{p_i}{1 - p_i} &= x_i\beta + \varepsilon_i \\ \frac{p_i}{1 - p_i} &= b^{x_i\beta + \varepsilon_i} \\ p_i &= (b^{x_i\beta + \varepsilon_i}) (1 - p_i) \\ p_i (1 + b^{x_i\beta + \varepsilon_i}) &= b^{x_i\beta + \varepsilon_i} \\ p_i &= \frac{b^{x_i\beta + \varepsilon_i}}{1 + b^{x_i\beta + \varepsilon_i}} \\ p_i &= \frac{b^{x_i\beta + \varepsilon_i}}{b^{x_i\beta + \varepsilon_i} (1 + b^{-(x_i\beta + \varepsilon_i)})} \\ p_i &= \frac{1}{1 + b^{-(x_i\beta + \varepsilon_i)}} \end{aligned} \quad (4)$$

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 (\rightarrow Definition II/1.2.1):

$$y \sim \text{Bin}(n, p) . \quad (1)$$

Then, the conjugate prior (\rightarrow Definition “prior-conj”) for the model parameter p is a beta distribution (\rightarrow Definition “beta”):

$$p(p) = \text{Bet}(p; \alpha_0, \beta_0) . \quad (2)$$

Proof: With the probability mass function of the binomial distribution (\rightarrow Proof “bin-pmf”), the likelihood function (\rightarrow Definition I/2.2.2) implied by (1) is given by

$$p(y|p) = \binom{n}{y} p^y (1-p)^{n-y} . \quad (3)$$

In other words, the likelihood function is proportional to a power of p times a power of $(1-p)$:

$$p(y|p) \propto p^y (1-p)^{n-y} . \quad (4)$$

The same is true for a beta distribution over p

$$p(p) = \text{Bet}(p; \alpha_0, \beta_0) \quad (5)$$

the probability density function of which (\rightarrow Proof “beta-pdf”)

$$p(p) = \frac{1}{B(\alpha_0, \beta_0)} p^{\alpha_0-1} (1-p)^{\beta_0-1} \quad (6)$$

exhibits the same proportionality

$$p(p) \propto p^{\alpha_0-1} (1-p)^{\beta_0-1} \quad (7)$$

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 (\rightarrow Definition II/1.2.1):

$$y \sim \text{Bin}(n, p) . \quad (1)$$

Moreover, assume a beta prior distribution (\rightarrow Proof III/4.1.1) over the model parameter p :

$$p(p) = \text{Bet}(p; \alpha_0, \beta_0) . \quad (2)$$

Then, the posterior distribution (\rightarrow Definition I/2.2.6) is also a beta distribution (\rightarrow Definition “beta”)

$$p(p|y) = \text{Bet}(p; \alpha_n, \beta_n) . \quad (3)$$

and the posterior hyperparameters (\rightarrow Definition I/2.2.6) are given by

$$\begin{aligned} \alpha_n &= \alpha_0 + y \\ \beta_n &= \beta_0 + (n - y) . \end{aligned} \quad (4)$$

Proof: With the probability mass function of the binomial distribution (\rightarrow Proof “bin-pmf”), the likelihood function (\rightarrow Definition I/2.2.2) implied by (1) is given by

$$p(y|p) = \binom{n}{y} p^y (1 - p)^{n-y} . \quad (5)$$

Combining the likelihood function (5) with the prior distribution (2), the joint likelihood (\rightarrow Definition I/2.2.5) of the model is given by

$$\begin{aligned} p(y, p) &= p(y|p) p(p) \\ &= \binom{n}{y} p^y (1 - p)^{n-y} \cdot \frac{1}{B(\alpha_0, \beta_0)} p^{\alpha_0-1} (1 - p)^{\beta_0-1} \\ &= \frac{1}{B(\alpha_0, \beta_0)} \binom{n}{y} p^{\alpha_0+y-1} (1 - p)^{\beta_0+(n-y)-1} . \end{aligned} \quad (6)$$

Note that the posterior distribution is proportional to the joint likelihood (\rightarrow Proof “post-jl”):

$$p(p|y) \propto p(y, p) . \quad (7)$$

Setting $\alpha_n = \alpha_0 + y$ and $\beta_n = \beta_0 + (n - y)$, the posterior distribution is therefore proportional to

$$p(p|y) \propto p^{\alpha_n-1} (1 - p)^{\beta_n-1} \quad (8)$$

which, when normalized to one, results in the probability density function of the beta distribution (\rightarrow Proof “beta-pdf”):

$$p(p|y) = \frac{1}{B(\alpha_n, \beta_n)} p^{\alpha_n-1} (1 - p)^{\beta_n-1} = \text{Bet}(p; \alpha_n, \beta_n) . \quad (9)$$

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 (\rightarrow Definition II/1.2.1):

$$y \sim \text{Bin}(n, p) . \quad (1)$$

Moreover, assume a beta prior distribution (\rightarrow Proof III/4.1.1) over the model parameter p :

$$p(p) = \text{Bet}(p; \alpha_0, \beta_0) . \quad (2)$$

Then, the log model evidence (\rightarrow Definition IV/3.1.1) for this model is

$$\begin{aligned} \log p(y|m) &= \log \Gamma(n+1) - \log \Gamma(k+1) - \log \Gamma(n-k+1) \\ &\quad + \log B(\alpha_n, \beta_n) - \log B(\alpha_0, \beta_0) . \end{aligned} \quad (3)$$

where the posterior hyperparameters (\rightarrow Definition I/2.2.6) are given by

$$\begin{aligned} \alpha_n &= \alpha_0 + y \\ \beta_n &= \beta_0 + (n - y) . \end{aligned} \quad (4)$$

Proof: With the probability mass function of the binomial distribution (\rightarrow Proof “bin-pmf”), the likelihood function (\rightarrow Definition I/2.2.2) implied by (1) is given by

$$p(y|p) = \binom{n}{y} p^y (1-p)^{n-y} . \quad (5)$$

Combining the likelihood function (5) with the prior distribution (2), the joint likelihood (\rightarrow Definition I/2.2.5) of the model is given by

$$\begin{aligned} p(y, p) &= p(y|p) p(p) \\ &= \binom{n}{y} p^y (1-p)^{n-y} \cdot \frac{1}{B(\alpha_0, \beta_0)} p^{\alpha_0-1} (1-p)^{\beta_0-1} \\ &= \binom{n}{y} \frac{1}{B(\alpha_0, \beta_0)} p^{\alpha_0+y-1} (1-p)^{\beta_0+(n-y)-1} . \end{aligned} \quad (6)$$

Note that the model evidence is the marginal density of the joint likelihood (\rightarrow Definition I/2.2.7):

$$p(y) = \int p(y, p) dp . \quad (7)$$

Setting $\alpha_n = \alpha_0 + y$ and $\beta_n = \beta_0 + (n - y)$, the joint likelihood can also be written as

$$p(y, p) = \binom{n}{y} \frac{1}{B(\alpha_0, \beta_0)} \frac{B(\alpha_n, \beta_n)}{1} \frac{1}{B(\alpha_n, \beta_n)} p^{\alpha_n-1} (1-p)^{\beta_n-1} . \quad (8)$$

Using the probability density function of the beta distribution (\rightarrow Proof “beta-pdf”), p can now be integrated out easily

$$\begin{aligned} p(y) &= \int \binom{n}{y} \frac{1}{B(\alpha_0, \beta_0)} \frac{B(\alpha_n, \beta_n)}{1} \frac{1}{B(\alpha_n, \beta_n)} p^{\alpha_n-1} (1-p)^{\beta_n-1} dp \\ &= \binom{n}{y} \frac{B(\alpha_n, \beta_n)}{B(\alpha_0, \beta_0)} \int \text{Bet}(p; \alpha_n, \beta_n) dp \\ &= \binom{n}{y} \frac{B(\alpha_n, \beta_n)}{B(\alpha_0, \beta_0)} , \end{aligned} \quad (9)$$

such that the log model evidence (\rightarrow Definition IV/3.1.1) (LME) is shown to be

$$\log p(y|m) = \log \binom{n}{y} + \log B(\alpha_n, \beta_n) - \log B(\alpha_0, \beta_0) . \quad (10)$$

With the definition of the binomial coefficient

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \quad (11)$$

and the definition of the gamma function

$$\Gamma(n) = (n-1)! , \quad (12)$$

the LME finally becomes

$$\begin{aligned} \log p(y|m) &= \log \Gamma(n+1) - \log \Gamma(k+1) - \log \Gamma(n-k+1) \\ &\quad + \log B(\alpha_n, \beta_n) - \log B(\alpha_0, \beta_0) . \end{aligned} \quad (13)$$

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.

4.2 Multinomial observations

4.2.1 Conjugate prior distribution

Theorem: Let $y = [y_1, \dots, y_k]$ be the number of observations in k categories resulting from n independent trials with unknown category probabilities $p = [p_1, \dots, p_k]$, such that y follows a multinomial distribution (\rightarrow Definition II/2.2.1):

$$y \sim \text{Mult}(n, p) . \quad (1)$$

Then, the conjugate prior (\rightarrow Definition “prior-conj”) for the model parameter p is a Dirichlet distribution (\rightarrow Definition “dir”):

$$p(p) = \text{Dir}(p; \alpha_0) . \quad (2)$$

Proof: With the probability mass function of the multinomial distribution (\rightarrow Proof “mult-pmf”), the likelihood function (\rightarrow Definition I/2.2.2) implied by (1) is given by

$$p(y|p) = \binom{n}{y_1, \dots, y_k} \prod_{j=1}^k p_j^{y_j} . \quad (3)$$

In other words, the likelihood function is proportional to a product of powers of the entries of the vector p :

$$p(y|p) \propto \prod_{j=1}^k p_j^{y_j} . \quad (4)$$

The same is true for a Dirichlet distribution over p

$$p(p) = \text{Dir}(p; \alpha_0) \quad (5)$$

the probability density function of which (\rightarrow Proof “dir-pdf”)

$$p(p) = \frac{\Gamma\left(\sum_{j=1}^k \alpha_{0j}\right)}{\prod_{j=1}^k \Gamma(\alpha_{0j})} \prod_{j=1}^k p_j^{\alpha_{0j}-1} \quad (6)$$

exhibits the same proportionality

$$p(p) \propto \prod_{j=1}^k p_j^{\alpha_{0j}-1} \quad (7)$$

and is therefore conjugate relative to the likelihood.

Sources:

- Wikipedia (2020): “Dirichlet distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-11; URL: https://en.wikipedia.org/wiki/Dirichlet_distribution#Conjugate_to_categorical/multinomial

Metadata: ID: P79 | shortcut: mult-prior | author: JoramSoch | date: 2020-11-03, 14:15.

4.2.2 Posterior distribution

Theorem: Let $y = [y_1, \dots, y_k]$ be the number of observations in k categories resulting from n independent trials with unknown category probabilities $p = [p_1, \dots, p_k]$, such that y follows a multinomial distribution (\rightarrow Definition II/2.2.1):

$$y \sim \text{Mult}(n, p) . \quad (1)$$

Moreover, assume a Dirichlet prior distribution (\rightarrow Proof III/4.2.1) over the model parameter p :

$$p(p) = \text{Dir}(p; \alpha_0) . \quad (2)$$

Then, the posterior distribution (\rightarrow Definition I/2.2.6) is also a Dirichlet distribution (\rightarrow Definition “dir”)

$$p(p|y) = \text{Dir}(p; \alpha_n) . \quad (3)$$

and the posterior hyperparameters (\rightarrow Definition I/2.2.6) are given by

$$\alpha_{nj} = \alpha_{0j} + y_j, \quad j = 1, \dots, k . \quad (4)$$

Proof: With the probability mass function of the multinomial distribution (\rightarrow Proof “mult-pmf”), the likelihood function (\rightarrow Definition I/2.2.2) implied by (1) is given by

$$p(y|p) = \binom{n}{y_1, \dots, y_k} \prod_{j=1}^k p_j^{y_j} . \quad (5)$$

Combining the likelihood function (5) with the prior distribution (2), the joint likelihood (\rightarrow Definition I/2.2.5) of the model is given by

$$\begin{aligned} p(y, p) &= p(y|p) p(p) \\ &= \binom{n}{y_1, \dots, y_k} \prod_{j=1}^k p_j^{y_j} \cdot \frac{\Gamma\left(\sum_{j=1}^k \alpha_{0j}\right)}{\prod_{j=1}^k \Gamma(\alpha_{0j})} \prod_{j=1}^k p_j^{\alpha_{0j}-1} \\ &= \frac{\Gamma\left(\sum_{j=1}^k \alpha_{0j}\right)}{\prod_{j=1}^k \Gamma(\alpha_{0j})} \binom{n}{y_1, \dots, y_k} \prod_{j=1}^k p_j^{\alpha_{0j}+y_j-1} . \end{aligned} \quad (6)$$

Note that the posterior distribution is proportional to the joint likelihood (\rightarrow Proof “post-jl”):

$$p(p|y) \propto p(y, p) . \quad (7)$$

Setting $\alpha_{nj} = \alpha_{0j} + y_j$, the posterior distribution is therefore proportional to

$$p(p|y) \propto \prod_{j=1}^k p_j^{\alpha_{nj}-1} \quad (8)$$

which, when normalized to one, results in the probability density function of the Dirichlet distribution (\rightarrow Proof “dir-pdf”):

$$p(p|y) = \frac{\Gamma\left(\sum_{j=1}^k \alpha_{nj}\right)}{\prod_{j=1}^k \Gamma(\alpha_{nj})} \prod_{j=1}^k p_j^{\alpha_{nj}-1} = \text{Dir}(p; \alpha_n) . \quad (9)$$

Sources:

- Wikipedia (2020): “Dirichlet distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-11; URL: https://en.wikipedia.org/wiki/Dirichlet_distribution#Conjugate_to_categorical/multinomial

Metadata: ID: P80 | shortcut: mult-post | author: JoramSoch | date: 2020-03-11, 14:40.

4.2.3 Log model evidence

Theorem: Let $y = [y_1, \dots, y_k]$ be the number of observations in k categories resulting from n independent trials with unknown category probabilities $p = [p_1, \dots, p_k]$, such that y follows a multinomial distribution (\rightarrow Definition II/2.2.1):

$$y \sim \text{Mult}(n, p) . \quad (1)$$

Moreover, assume a Dirichlet prior distribution (\rightarrow Proof III/4.2.1) over the model parameter p :

$$p(p) = \text{Dir}(p; \alpha_0) . \quad (2)$$

Then, the log model evidence (\rightarrow Definition IV/3.1.1) for this model is

$$\begin{aligned} \log p(y|m) &= \log \Gamma(n+1) - \sum_{j=1}^k \log \Gamma(k_j+1) \\ &\quad + \log \Gamma\left(\sum_{j=1}^k \alpha_{0j}\right) - \log \Gamma\left(\sum_{j=1}^k \alpha_{nj}\right) \\ &\quad + \sum_{j=1}^k \log \Gamma(\alpha_{nj}) - \sum_{j=1}^k \log \Gamma(\alpha_{0j}) . \end{aligned} \quad (3)$$

and the posterior hyperparameters (\rightarrow Definition I/2.2.6) are given by

$$\alpha_{nj} = \alpha_{0j} + y_j, \quad j = 1, \dots, k . \quad (4)$$

Proof: With the probability mass function of the multinomial distribution (\rightarrow Proof “mult-pmf”), the likelihood function (\rightarrow Definition I/2.2.2) implied by (1) is given by

$$p(y|p) = \binom{n}{y_1, \dots, y_k} \prod_{j=1}^k p_j^{y_j} . \quad (5)$$

Combining the likelihood function (5) with the prior distribution (2), the joint likelihood (\rightarrow Definition I/2.2.5) of the model is given by

$$\begin{aligned} p(y, p) &= p(y|p) p(p) \\ &= \binom{n}{y_1, \dots, y_k} \prod_{j=1}^k p_j^{y_j} \cdot \frac{\Gamma\left(\sum_{j=1}^k \alpha_{0j}\right)}{\prod_{j=1}^k \Gamma(\alpha_{0j})} \prod_{j=1}^k p_j^{\alpha_{0j}-1} \\ &= \binom{n}{y_1, \dots, y_k} \frac{\Gamma\left(\sum_{j=1}^k \alpha_{0j}\right)}{\prod_{j=1}^k \Gamma(\alpha_{0j})} \prod_{j=1}^k p_j^{\alpha_{0j}+y_j-1} . \end{aligned} \quad (6)$$

Note that the model evidence is the marginal density of the joint likelihood:

$$p(y) = \int p(y, p) dp . \quad (7)$$

Setting $\alpha_{nj} = \alpha_{0j} + y_j$, the joint likelihood can also be written as

$$p(y, p) = \binom{n}{y_1, \dots, y_k} \frac{\Gamma\left(\sum_{j=1}^k \alpha_{0j}\right)}{\prod_{j=1}^k \Gamma(\alpha_{0j})} \frac{\prod_{j=1}^k \Gamma(\alpha_{nj})}{\Gamma\left(\sum_{j=1}^k \alpha_{nj}\right)} \frac{\Gamma\left(\sum_{j=1}^k \alpha_{nj}\right)}{\prod_{j=1}^k \Gamma(\alpha_{nj})} \prod_{j=1}^k p_j^{\alpha_{nj}-1}. \quad (8)$$

Using the probability density function of the Dirichlet distribution (\rightarrow Proof “dir-pdf”), p can now be integrated out easily

$$\begin{aligned} p(y) &= \int \binom{n}{y_1, \dots, y_k} \frac{\Gamma\left(\sum_{j=1}^k \alpha_{0j}\right)}{\prod_{j=1}^k \Gamma(\alpha_{0j})} \frac{\prod_{j=1}^k \Gamma(\alpha_{nj})}{\Gamma\left(\sum_{j=1}^k \alpha_{nj}\right)} \frac{\Gamma\left(\sum_{j=1}^k \alpha_{nj}\right)}{\prod_{j=1}^k \Gamma(\alpha_{nj})} \prod_{j=1}^k p_j^{\alpha_{nj}-1} dp \\ &= \binom{n}{y_1, \dots, y_k} \frac{\Gamma\left(\sum_{j=1}^k \alpha_{0j}\right)}{\prod_{j=1}^k \Gamma(\alpha_{0j})} \frac{\prod_{j=1}^k \Gamma(\alpha_{nj})}{\Gamma\left(\sum_{j=1}^k \alpha_{nj}\right)} \int \text{Dir}(p; \alpha_n) dp \\ &= \binom{n}{y_1, \dots, y_k} \frac{\Gamma\left(\sum_{j=1}^k \alpha_{0j}\right)}{\Gamma\left(\sum_{j=1}^k \alpha_{nj}\right)} \frac{\prod_{j=1}^k \Gamma(\alpha_{nj})}{\prod_{j=1}^k \Gamma(\alpha_{0j})}, \end{aligned} \quad (9)$$

such that the log model evidence (\rightarrow Definition IV/3.1.1) (LME) is shown to be

$$\begin{aligned} \log p(y|m) &= \log \binom{n}{y_1, \dots, y_k} + \log \Gamma\left(\sum_{j=1}^k \alpha_{0j}\right) - \log \Gamma\left(\sum_{j=1}^k \alpha_{nj}\right) \\ &\quad + \sum_{j=1}^k \log \Gamma(\alpha_{nj}) - \sum_{j=1}^k \log \Gamma(\alpha_{0j}). \end{aligned} \quad (10)$$

With the definition of the multinomial coefficient

$$\binom{n}{k_1, \dots, k_m} = \frac{n!}{k_1! \cdot \dots \cdot k_m!} \quad (11)$$

and the definition of the gamma function

$$\Gamma(n) = (n-1)!, \quad (12)$$

the LME finally becomes

$$\begin{aligned} \log p(y|m) &= \log \Gamma(n+1) - \sum_{j=1}^k \log \Gamma(k_j+1) \\ &\quad + \log \Gamma\left(\sum_{j=1}^k \alpha_{0j}\right) - \log \Gamma\left(\sum_{j=1}^k \alpha_{nj}\right) \\ &\quad + \sum_{j=1}^k \log \Gamma(\alpha_{nj}) - \sum_{j=1}^k \log \Gamma(\alpha_{0j}). \end{aligned} \quad (13)$$

Sources:

- original work

Metadata: ID: P81 | shortcut: mult-lme | author: JoramSoch | date: 2020-03-11, 15:17.

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 (\rightarrow Definition III/1.1.1)

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

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 (\rightarrow 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 $x = \{x_1, \dots, x_n\}$ be a set of independent normally distributed (\rightarrow Definition II/3.2.1) observations with unknown mean (\rightarrow Definition I/1.2.1) μ and variance (\rightarrow Definition I/1.3.1) σ^2 :

$$x_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2), \quad i = 1, \dots, n. \quad (1)$$

Then,

1) the maximum likelihood estimator (\rightarrow Definition “mle”) of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (2)$$

where

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (3)$$

2) and $\hat{\sigma}^2$ is a biased estimator (\rightarrow Definition “est-unb”) of σ^2

$$\mathbb{E} [\hat{\sigma}^2] \neq \sigma^2, \quad (4)$$

more precisely:

$$\mathbb{E} [\hat{\sigma}^2] = \frac{n-1}{n} \sigma^2. \quad (5)$$

Proof:

1) This is equivalent to the maximum likelihood estimator for the univariate Gaussian with unknown variance (\rightarrow Proof “ug-mle”) and a special case of the maximum likelihood estimator for multiple linear regression (\rightarrow Proof III/1.1.10) in which $y = x$, $X = 1_n$ and $\hat{\beta} = \bar{x}$:

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n}(y - X\hat{\beta})^T(y - X\hat{\beta}) \\
&= \frac{1}{n}(x - 1_n\bar{x})^T(x - 1_n\bar{x}) \\
&= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 .
\end{aligned} \tag{6}$$

2) The expectation (\rightarrow Definition I/1.2.1) of the maximum likelihood estimator (\rightarrow Definition “mle”) can be developed as follows:

$$\begin{aligned}
\mathbb{E} [\hat{\sigma}^2] &= \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right] \\
&= \frac{1}{n} \mathbb{E} \left[\sum_{i=1}^n (x_i - \bar{x})^2 \right] \\
&= \frac{1}{n} \mathbb{E} \left[\sum_{i=1}^n (x_i^2 - 2x_i\bar{x} + \bar{x}^2) \right] \\
&= \frac{1}{n} \mathbb{E} \left[\sum_{i=1}^n x_i^2 - 2 \sum_{i=1}^n x_i\bar{x} + \sum_{i=1}^n \bar{x}^2 \right] \\
&= \frac{1}{n} \mathbb{E} \left[\sum_{i=1}^n x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 \right] \\
&= \frac{1}{n} \mathbb{E} \left[\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right] \\
&= \frac{1}{n} \left(\sum_{i=1}^n \mathbb{E} [x_i^2] - n\mathbb{E} [\bar{x}^2] \right) \\
&= \frac{1}{n} \sum_{i=1}^n \mathbb{E} [x_i^2] - \mathbb{E} [\bar{x}^2]
\end{aligned} \tag{7}$$

Due to the partition of variance into expected values (\rightarrow Proof “var-mean”)

$$\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 , \tag{8}$$

we have

$$\begin{aligned}
\text{Var}(x_i) &= \mathbb{E}(x_i^2) - \mathbb{E}(x_i)^2 \\
\text{Var}(\bar{x}) &= \mathbb{E}(\bar{x}^2) - \mathbb{E}(\bar{x})^2 ,
\end{aligned} \tag{9}$$

such that (7) becomes

$$\mathbb{E} [\hat{\sigma}^2] = \frac{1}{n} \sum_{i=1}^n (\text{Var}(x_i) + \mathbb{E}(x_i)^2) - (\text{Var}(\bar{x}) + \mathbb{E}(\bar{x})^2) . \tag{10}$$

From (1), it follows that

$$\mathbb{E}(x_i) = \mu \quad \text{and} \quad \text{Var}(x_i) = \sigma^2. \quad (11)$$

The expectation of (\rightarrow Proof “ug-unb”) \bar{x} given by (3) is

$$\begin{aligned} \mathbb{E}[\bar{x}] &= \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n x_i\right] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[x_i] \\ &\stackrel{(11)}{=} \frac{1}{n} \sum_{i=1}^n \mu = \frac{1}{n} \cdot n \cdot \mu \\ &= \mu. \end{aligned} \quad (12)$$

The variance of \bar{x} given by (3) is

$$\begin{aligned} \text{Var}[\bar{x}] &= \text{Var}\left[\frac{1}{n} \sum_{i=1}^n x_i\right] = \frac{1}{n^2} \sum_{i=1}^n \text{Var}[x_i] \\ &\stackrel{(11)}{=} \frac{1}{n^2} \sum_{i=1}^n \sigma^2 = \frac{1}{n^2} \cdot n \cdot \sigma^2 \\ &= \frac{1}{n} \sigma^2. \end{aligned} \quad (13)$$

Plugging (11), (12) and (13) into (10), we have

$$\begin{aligned} \mathbb{E}[\hat{\sigma}^2] &= \frac{1}{n} \sum_{i=1}^n (\sigma^2 + \mu^2) - \left(\frac{1}{n} \sigma^2 + \mu^2\right) \\ \mathbb{E}[\hat{\sigma}^2] &= \frac{1}{n} \cdot n \cdot (\sigma^2 + \mu^2) - \left(\frac{1}{n} \sigma^2 + \mu^2\right) \\ \mathbb{E}[\hat{\sigma}^2] &= \sigma^2 + \mu^2 - \frac{1}{n} \sigma^2 - \mu^2 \\ \mathbb{E}[\hat{\sigma}^2] &= \frac{n-1}{n} \sigma^2 \end{aligned} \quad (14)$$

which proves the bias (\rightarrow 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 $x = \{x_1, \dots, x_n\}$ be a set of independent normally distributed (\rightarrow Definition II/3.2.1) observations with unknown mean (\rightarrow Definition I/1.2.1) μ and variance (\rightarrow Definition I/1.3.1) σ^2 :

$$x_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2), \quad i = 1, \dots, n. \quad (1)$$

An unbiased estimator (\rightarrow Definition “est-unb”) of σ^2 is given by

$$\hat{\sigma}_{\text{unb}}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2. \quad (2)$$

Proof: It can be shown that (\rightarrow Proof IV/1.1.2) the maximum likelihood estimator (\rightarrow Definition “mle”) of σ^2

$$\hat{\sigma}_{\text{MLE}}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (3)$$

is a biased estimator (\rightarrow Definition “est-unb”) in the sense that

$$\mathbb{E} [\hat{\sigma}_{\text{MLE}}^2] = \frac{n-1}{n} \sigma^2. \quad (4)$$

From (4), it follows that

$$\begin{aligned} \mathbb{E} \left[\frac{n}{n-1} \hat{\sigma}_{\text{MLE}}^2 \right] &= \frac{n}{n-1} \mathbb{E} [\hat{\sigma}_{\text{MLE}}^2] \\ &\stackrel{(4)}{=} \frac{n}{n-1} \cdot \frac{n-1}{n} \sigma^2 \\ &= \sigma^2, \end{aligned} \quad (5)$$

such that an unbiased estimator (\rightarrow Definition “est-unb”) can be constructed as

$$\begin{aligned} \hat{\sigma}_{\text{unb}}^2 &= \frac{n}{n-1} \hat{\sigma}_{\text{MLE}}^2 \\ &\stackrel{(3)}{=} \frac{n}{n-1} \cdot \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \\ &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2. \end{aligned} \quad (6)$$

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 (\rightarrow Definition III/1.1.1) with independent (\rightarrow Definition “ind”) observations

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

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 (\rightarrow Definition III/1.1.4)”) that can be predicted from the independent variables X (“explained variance (\rightarrow Definition III/1.1.5)”) 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 R^2 and adjusted R^2

Theorem: Given a linear regression model (\rightarrow Definition III/1.1.1)

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

with n independent observations and p independent variables,

1) the coefficient of determination (\rightarrow Definition IV/1.2.1) is given by

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} \quad (2)$$

2) the adjusted coefficient of determination is

$$R_{\text{adj}}^2 = 1 - \frac{\text{RSS}/(n-p)}{\text{TSS}/(n-1)} \quad (3)$$

where the residual (\rightarrow Definition III/1.1.6) and total sum of squares (\rightarrow Definition III/1.1.4) are

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^n (y_i - \hat{y}_i)^2, \quad \hat{y} = X\hat{\beta} \\ \text{TSS} &= \sum_{i=1}^n (y_i - \bar{y})^2, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \end{aligned} \quad (4)$$

where X is the $n \times p$ design matrix and $\hat{\beta}$ are the ordinary least squares (\rightarrow Definition “mlr-ols”) estimates.

Proof: The coefficient of determination R^2 is defined as (\rightarrow 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 (\rightarrow Definition III/1.1.5) as

$$\text{ESS} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2, \quad (5)$$

then R^2 is given by

$$R^2 = \frac{\text{ESS}}{\text{TSS}} . \quad (6)$$

which is equal to

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}} , \quad (7)$$

because (\rightarrow Proof III/1.1.7) $\text{TSS} = \text{ESS} + \text{RSS}$.

2) Using (4), the coefficient of determination can be also written as:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2} . \quad (8)$$

If we replace the variance estimates by their unbiased estimators (\rightarrow Proof IV/1.1.3), we obtain

$$R_{\text{adj}}^2 = 1 - \frac{\frac{1}{n-p} \sum_{i=1}^n (y_i - \hat{y}_i)^2}{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\text{RSS}/\text{df}_r}{\text{TSS}/\text{df}_t} \quad (9)$$

where $\text{df}_r = n - p$ and $\text{df}_t = n - 1$ are the residual and total degrees of freedom (\rightarrow 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 (\rightarrow Definition III/1.1.1) with independent observations

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

the coefficient of determination (\rightarrow Definition IV/1.2.1) can be expressed in terms of the maximum log-likelihood (\rightarrow Definition “mll”) as

$$R^2 = 1 - (\exp[\Delta\text{MLL}])^{-2/n} \quad (2)$$

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 (\rightarrow Definition “mll”) (MLL) of a linear regression model in terms of its residual sum of squares (\rightarrow Definition III/1.1.6) (RSS). The model in (1) implies the following log-likelihood function (\rightarrow Definition “llf”)

$$\text{LL}(\beta, \sigma^2) = \log p(y|\beta, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} (y - X\beta)^T (y - X\beta) , \quad (3)$$

such that maximum likelihood estimates are (\rightarrow Proof III/1.1.10)

$$\hat{\beta} = (X^T X)^{-1} X^T y \quad (4)$$

$$\hat{\sigma}^2 = \frac{1}{n} (y - X\hat{\beta})^T (y - X\hat{\beta}) \quad (5)$$

and the residual sum of squares (\rightarrow Definition III/1.1.6) is

$$\text{RSS} = \sum_{i=1}^n \hat{\varepsilon}_i = \hat{\varepsilon}^T \hat{\varepsilon} = (y - X\hat{\beta})^T (y - X\hat{\beta}) = n \cdot \hat{\sigma}^2 . \quad (6)$$

Since $\hat{\beta}$ and $\hat{\sigma}^2$ are maximum likelihood estimates (\rightarrow Definition “mle”), plugging them into the log-likelihood function gives the maximum log-likelihood:

$$\text{MLL} = \text{LL}(\hat{\beta}, \hat{\sigma}^2) = -\frac{n}{2} \log(2\pi\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2} (y - X\hat{\beta})^T (y - X\hat{\beta}) . \quad (7)$$

With (6) for the first $\hat{\sigma}^2$ and (5) for the second $\hat{\sigma}^2$, the MLL becomes

$$\text{MLL} = -\frac{n}{2} \log(\text{RSS}) - \frac{n}{2} \log\left(\frac{2\pi}{n}\right) - \frac{n}{2} . \quad (8)$$

Second, we establish the relationship between maximum log-likelihood (MLL) and coefficient of determination (R^2). Consider the two models

$$\begin{aligned} m_0 : X_0 &= 1_n \\ m_1 : X_1 &= X \end{aligned} \quad (9)$$

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 (\rightarrow Definition III/1.1.4):

$$\text{TSS} = \sum_{i=1}^n (y_i - \bar{y})^2 . \quad (10)$$

Using (8), we can therefore write

$$\Delta\text{MLL} = \text{MLL}(m_1) - \text{MLL}(m_0) = -\frac{n}{2} \log(\text{RSS}) + \frac{n}{2} \log(\text{TSS}) . \quad (11)$$

Exponentiating both sides of the equation, we have:

$$\begin{aligned} \exp[\Delta\text{MLL}] &= \exp\left[-\frac{n}{2} \log(\text{RSS}) + \frac{n}{2} \log(\text{TSS})\right] \\ &= (\exp[\log(\text{RSS}) - \log(\text{TSS})])^{-n/2} \\ &= \left(\frac{\exp[\log(\text{RSS})]}{\exp[\log(\text{TSS})]}\right)^{-n/2} \\ &= \left(\frac{\text{RSS}}{\text{TSS}}\right)^{-n/2} . \end{aligned} \quad (12)$$

Taking both sides to the power of $-2/n$ and subtracting from 1, we have

$$\begin{aligned}
(\exp[\Delta\text{MLL}])^{-2/n} &= \frac{\text{RSS}}{\text{TSS}} \\
1 - (\exp[\Delta\text{MLL}])^{-2/n} &= 1 - \frac{\text{RSS}}{\text{TSS}} = R^2
\end{aligned} \tag{13}$$

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 (\rightarrow Definition III/1.1.1) with independent (\rightarrow Definition “ind”) observations

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

with measured data y , known design matrix X as well as unknown regression coefficients β and noise variance σ^2 .

Given estimated regression coefficients (\rightarrow Definition “mlr-beta”) $\hat{\beta}$ and residual variance (\rightarrow 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:

$$\text{SNR} = \frac{\text{Var}(X\hat{\beta})}{\hat{\sigma}^2}. \tag{2}$$

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 (\rightarrow Definition I/2.2.1) with likelihood function (\rightarrow Definition I/2.2.2) $p(y|\theta, m)$ and maximum likelihood estimates (\rightarrow Definition “mle”)

$$\hat{\theta} = \arg \max_{\theta} \log p(y|\theta, m) . \quad (1)$$

Then, the Akaike information criterion (AIC) of this model is defined as

$$\text{AIC}(m) = -2 \log p(y|\hat{\theta}, m) + 2p \quad (2)$$

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 (\rightarrow Definition I/2.2.1) with likelihood function (\rightarrow Definition I/2.2.2) $p(y|\theta, m)$ and maximum likelihood estimates (\rightarrow Definition “mle”)

$$\hat{\theta} = \arg \max_{\theta} \log p(y|\theta, m) . \quad (1)$$

Then, the Bayesian information criterion (BIC) of this model is defined as

$$\text{BIC}(m) = -2 \log p(y|\hat{\theta}, m) + p \log n \quad (2)$$

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 (\rightarrow Definition I/2.2.2) of a generative model (\rightarrow 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 (\rightarrow 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 (\rightarrow 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 (\rightarrow Definition IV/2.2.1) (BIC) as

$$-2 \log p(y|m) \approx \text{BIC}(m) = -2 \log p(y|\hat{\theta}, m) + p \log n \quad (1)$$

where $\hat{\theta}$ is the maximum likelihood estimator (\rightarrow Definition “mle”) (MLE) of θ , n is the number of data points and p is the number of model parameters.

Proof: Let $\text{LL}(\theta)$ be the log-likelihood function (\rightarrow Definition “llf”)

$$\text{LL}(\theta) = \log p(y|\theta, m) \quad (2)$$

and define the functions g and h as follows:

$$\begin{aligned} g(\theta) &= p(\theta|m) \\ h(\theta) &= \frac{1}{n} \text{LL}(\theta) . \end{aligned} \quad (3)$$

Then, the marginal likelihood (\rightarrow Definition I/2.2.7) can be written as follows:

$$\begin{aligned} p(y|m) &= \int_{\Theta} p(y|\theta, m) p(\theta|m) d\theta \\ &= \int_{\Theta} \exp[n h(\theta)] g(\theta) d\theta . \end{aligned} \quad (4)$$

This is an integral suitable for Laplace approximation which states that

$$\int_{\Theta} \exp[n h(\theta)] g(\theta) d\theta = \left(\sqrt{\frac{2\pi}{n}} \right)^p \exp[n h(\theta_0)] \left(g(\theta_0) |J(\theta_0)|^{-1/2} + O(1/n) \right) \quad (5)$$

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 $h(\theta) = 1/n \text{LL}(\theta)$ such that θ_0 is the maximum likelihood estimator $\hat{\theta}$:

$$\hat{\theta} = \arg \max_{\theta} \text{LL}(\theta) . \quad (6)$$

With this, (5) can be applied to (4) using (3) to give:

$$p(y|m) \approx \left(\sqrt{\frac{2\pi}{n}} \right)^p p(y|\hat{\theta}, m) p(\hat{\theta}|m) |J(\hat{\theta})|^{-1/2} . \quad (7)$$

Logarithmizing and multiplying with -2 , we have:

$$-2 \log p(y|m) \approx -2 \text{LL}(\hat{\theta}) + p \log n - p \log(2\pi) - 2 \log p(\hat{\theta}|m) + \log |J(\hat{\theta})| . \quad (8)$$

As $n \rightarrow \infty$, the last three terms are $O_p(1)$ and can therefore be ignored when comparing between models $\mathcal{M} = \{m_1, \dots, m_M\}$ and using $p(y|m_j)$ to compute posterior model probabilities (\rightarrow Definition “led-pmp”) $p(m_j|y)$. With that, the BIC is given as

$$\text{BIC}(m) = -2 \log p(y|\hat{\theta}, m) + p \log n . \quad (9)$$

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-averaging/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 (\rightarrow Definition I/2.2.1) with likelihood function (\rightarrow Definition I/2.2.2) $p(y|\theta, m)$ and prior distribution (\rightarrow Definition I/2.2.3) $p(\theta|m)$. Together, likelihood function and prior distribution imply a posterior distribution (\rightarrow Definition I/2.2.6) $p(\theta|y, m)$. Define the posterior expected log-likelihood (\rightarrow Definition “llf”) (PLL)

$$\text{PLL}(m) = \langle \log p(y|\theta, m) \rangle_{\theta|y} \quad (1)$$

and the log-likelihood (\rightarrow Definition “llf”) at the posterior expectation (LLP)

$$\text{LLP}(m) = \log p(y | \langle \theta \rangle_{\theta|y}, m) \quad (2)$$

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

$$\text{DIC}(m) = -2 \text{LLP}(m) + 2 p_D \quad \text{or} \quad \text{DIC}(m) = -2 \text{PLL}(m) + p_D \quad (3)$$

where the “effective number of parameters” p_D is given by

$$p_D = -2 \text{PLL}(m) + 2 \text{LLP}(m) . \quad (4)$$

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 (\rightarrow Definition I/2.2.1) with likelihood function (\rightarrow Definition I/2.2.2) $p(y|\theta, m)$ and prior distribution (\rightarrow 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 (\rightarrow Definition I/2.2.7):

$$\text{LME}(m) = \log p(y|m) . \quad (1)$$

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 (\rightarrow Definition I/2.2.2) of a generative model (\rightarrow 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 (\rightarrow Definition I/2.2.3) on model parameters θ . Then, the log model evidence (\rightarrow Definition IV/3.1.1) (LME), also called marginal log-likelihood,

$$\text{LME}(m) = \log p(y|m) , \quad (1)$$

can be expressed

1) as

$$\text{LME}(m) = \log \int p(y|\theta, m) p(\theta|m) d\theta \quad (2)$$

2) or

$$\text{LME}(m) = \log p(y|\theta, m) + \log p(\theta|m) - \log p(\theta|y, m) . \quad (3)$$

Proof:

1) The first expression is a simple consequence of the law of marginal probability (\rightarrow Definition “prob-marg”) for continuous variables according to which

$$p(y|m) = \int p(y|\theta, m) p(\theta|m) d\theta \quad (4)$$

which, when logarithmized, gives

$$\text{LME}(m) = \log p(y|m) = \log \int p(y|\theta, m) p(\theta|m) d\theta . \quad (5)$$

2) The second expression can be derived from Bayes' theorem (\rightarrow Proof I/2.1.1) which makes a statement about the posterior distribution (\rightarrow Definition I/2.2.6):

$$p(\theta|y, m) = \frac{p(y|\theta, m) p(\theta|m)}{p(y|m)} . \quad (6)$$

Rearranging for $p(y|m)$ and logarithmizing, we have:

$$\begin{aligned} \text{LME}(m) = \log p(y|m) &= \log \frac{p(y|\theta, m) p(\theta|m)}{p(\theta|y, m)} \\ &= \log p(y|\theta, m) + \log p(\theta|m) - \log p(\theta|y, m) . \end{aligned} \quad (7)$$

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 (\rightarrow Definition IV/3.1.1) can be partitioned into accuracy and complexity

$$\text{LME}(m) = \text{Acc}(m) - \text{Com}(m) \quad (1)$$

where the accuracy term is the posterior expectation of the log-likelihood function (\rightarrow Definition I/2.2.2)

$$\text{Acc}(m) = \langle p(y|\theta, m) \rangle_{p(\theta|y, m)} \quad (2)$$

and the complexity penalty is the Kullback-Leibler divergence (\rightarrow Definition “kl”) of posterior (\rightarrow Definition I/2.2.6) from prior (\rightarrow Definition I/2.2.3)

$$\text{Com}(m) = \text{KL} [p(\theta|y, m) || p(\theta|m)] . \quad (3)$$

Proof: We consider Bayesian inference on data y using model m with parameters θ . Then, Bayes' theorem (\rightarrow Proof I/2.1.1) makes a statement about the posterior distribution, i.e. the probability of parameters, given the data and the model:

$$p(\theta|y, m) = \frac{p(y|\theta, m) p(\theta|m)}{p(y|m)} . \quad (4)$$

Rearranging this for the model evidence (\rightarrow Proof IV/3.1.2), we have:

$$p(y|m) = \frac{p(y|\theta, m) p(\theta|m)}{p(\theta|y, m)} . \quad (5)$$

Logarithmizing both sides of the equation, we obtain:

$$\log p(y|m) = \log p(y|\theta, m) - \log \frac{p(\theta|y, m)}{p(\theta|m)} . \quad (6)$$

Now taking the expectation over the posterior distribution yields:

$$\log p(y|m) = \int p(\theta|y, m) \log p(y|\theta, m) d\theta - \int p(\theta|y, m) \log \frac{p(\theta|y, m)}{p(\theta|m)} d\theta . \quad (7)$$

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

$$\text{LME}(m) = \langle p(y|\theta, m) \rangle_{p(\theta|y, m)} - \text{KL} [p(\theta|y, m) || p(\theta|m)] \quad (8)$$

which proves 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 (\rightarrow Definition IV/3.1.1) $\text{LME}(m_1)$ and $\text{LME}(m_2)$. Then, the log Bayes factor (\rightarrow Definition “lbf”) in favor of model m_1 and against model m_2 is the difference of the log model evidences:

$$\text{LBF}_{1,2} = \text{LME}(m_1) - \text{LME}(m_2) . \quad (1)$$

Proof: The log Bayes factor (\rightarrow Definition “lbf”) (LBF) is defined as the logarithm of the Bayes factor (\rightarrow Definition “bf”) (BF) which is defined as the posterior odds ratio when both models are equally likely apriori:

$$\begin{aligned} \text{LBF}_{1,2} &= \log \text{BF}_{1,2} \\ &= \log \frac{p(m_1|y)}{p(m_2|y)} . \end{aligned} \quad (2)$$

Plugging in the posterior odds ratio according to Bayes’ rule (\rightarrow Proof I/2.1.2), we have

$$\text{LBF}_{1,2} = \log \left[\frac{p(y|m_1)}{p(y|m_2)} \cdot \frac{p(m_1)}{p(m_2)} \right] . \quad (3)$$

When both models are equally likely apriori, the prior odds ratio is one, such that

$$\text{LBF}_{1,2} = \log \frac{p(y|m_1)}{p(y|m_2)} . \quad (4)$$

Resolving the logarithm and applying the definition of the log model evidence (\rightarrow Definition IV/3.1.1), we finally have:

$$\begin{aligned} \text{LBF}_{1,2} &= \log p(y|m_1) - \log p(y|m_2) \\ &= \text{LME}(m_1) - \text{LME}(m_2) . \end{aligned} \quad (5)$$

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, \dots, m_M be M statistical models with log model evidences (\rightarrow Definition IV/3.1.1) $\text{LME}(m_1), \dots, \text{LME}(m_M)$ and belonging to F mutually exclusive model families f_1, \dots, f_F . Then, the log family evidences (\rightarrow Definition “lfe”) are given by:

$$\text{LFE}(f_j) = \log \sum_{m_i \in f_j} [\exp[\text{LME}(m_i)] \cdot p(m_i|f_j)], \quad j = 1, \dots, F, \quad (1)$$

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 (\rightarrow Definition “prob-marg”), this conditional probability is given by

$$p(y|f_j) = \sum_{m_i \in f_j} [p(y|m_i, f_j) \cdot p(m_i|f_j)] . \quad (2)$$

Because model families are mutually exclusive, it holds that $p(y|m_i, f_j) = p(y|m_i)$, such that

$$p(y|f_j) = \sum_{m_i \in f_j} [p(y|m_i) \cdot p(m_i|f_j)] . \quad (3)$$

Logarithmizing transforms the family evidence $p(y|f_j)$ into the log family evidence $\text{LFE}(f_j)$:

$$\text{LFE}(f_j) = \log \sum_{m_i \in f_j} [p(y|m_i) \cdot p(m_i|f_j)] . \quad (4)$$

The definition of the log model evidence (\rightarrow Definition IV/3.1.1)

$$\text{LME}(m) = \log p(y|m) \quad (5)$$

can be exponentiated to then read

$$\exp [\text{LME}(m)] = p(y|m) \quad (6)$$

and applying (6) to (4), we finally have:

$$\text{LFE}(f_j) = \log \sum_{m_i \in f_j} [\exp[\text{LME}(m_i)] \cdot p(m_i|f_j)] . \quad (7)$$

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 (\rightarrow Definition “lbf”) $\text{LBF}_{1,2}$ in favor of model m_1 and against model m_2 . Then, if both models are equally likely apriori, the posterior model probability (\rightarrow Definition “pmp”) of m_1 is

$$p(m_1|y) = \frac{\exp(\text{LBF}_{1,2})}{\exp(\text{LBF}_{1,2}) + 1} . \quad (1)$$

Proof: From Bayes’ rule (\rightarrow Proof I/2.1.2), the posterior odds ratio is

$$\frac{p(m_1|y)}{p(m_2|y)} = \frac{p(y|m_1)}{p(y|m_2)} \cdot \frac{p(m_1)}{p(m_2)} . \quad (2)$$

When both models are equally likely apriori, the prior odds ratio is one, such that

$$\frac{p(m_1|y)}{p(m_2|y)} = \frac{p(y|m_1)}{p(y|m_2)} . \quad (3)$$

Now the right-hand side corresponds to the Bayes factor (\rightarrow Definition “bf”), therefore

$$\frac{p(m_1|y)}{1 - p(m_1|y)} = \text{BF}_{1,2} . \quad (4)$$

Because the two models are collectively exhaustive, we have

$$\frac{p(m_1|y)}{1 - p(m_1|y)} = \text{BF}_{1,2} . \quad (5)$$

Now rearranging for the posterior probability (\rightarrow Definition “pmp”), this gives

$$p(m_1|y) = \frac{\text{BF}_{1,2}}{\text{BF}_{1,2} + 1} . \quad (6)$$

Because the log Bayes factor is the logarithm of the Bayes factor (\rightarrow Proof IV/3.2.1), we finally have

$$p(m_1|y) = \frac{\exp(\text{LBF}_{1,2})}{\exp(\text{LBF}_{1,2}) + 1} . \quad (7)$$

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, \dots, m_M be $M + 1$ statistical models with model evidences (\rightarrow Definition IV/3.1.1) $p(y|m_0), p(y|m_1), \dots, p(y|m_M)$. Then, the posterior model probabilities (\rightarrow Definition “pmp”) of the models m_1, \dots, m_M are given by:

$$p(m_i|y) = \frac{\text{BF}_{i,0} \cdot \alpha_i}{\sum_{j=1}^M \text{BF}_{j,0} \cdot \alpha_j}, \quad i = 1, \dots, M, \quad (1)$$

where $\text{BF}_{i,0}$ is the Bayes factor (\rightarrow 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

$$\text{BF}_{i,0} = \frac{p(y|m_i)}{p(y|m_0)} \quad (2)$$

and prior odds ratio of m_i against m_0

$$\alpha_i = \frac{p(m_i)}{p(m_0)}. \quad (3)$$

From Bayes’ theorem (\rightarrow Proof I/2.1.1), the posterior probability of m_i follows as

$$p(m_i|y) = \frac{p(y|m_i) \cdot p(m_i)}{\sum_{j=1}^M p(y|m_j) \cdot p(m_j)}. \quad (4)$$

Now applying (2) and (3) to (4), we have

$$\begin{aligned} p(m_i|y) &= \frac{\text{BF}_{i,0} p(y|m_0) \cdot \alpha_i p(m_0)}{\sum_{j=1}^M \text{BF}_{j,0} p(y|m_0) \cdot \alpha_j p(m_0)} \\ &= \frac{[p(y|m_0) p(m_0)] \text{BF}_{i,0} \cdot \alpha_i}{[p(y|m_0) p(m_0)] \sum_{j=1}^M \text{BF}_{j,0} \cdot \alpha_j}, \end{aligned} \quad (5)$$

such that

$$p(m_i|y) = \frac{\text{BF}_{i,0} \cdot \alpha_i}{\sum_{j=1}^M \text{BF}_{j,0} \cdot \alpha_j}. \quad (6)$$

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, \dots, m_M be M statistical models with log model evidences (\rightarrow Definition IV/3.1.1) $\text{LME}(m_1), \dots, \text{LME}(m_M)$. Then, the posterior model probabilities (\rightarrow Definition “pmp”) are given by:

$$p(m_i|y) = \frac{\exp[\text{LME}(m_i)] p(m_i)}{\sum_{j=1}^M \exp[\text{LME}(m_j)] p(m_j)}, \quad i = 1, \dots, M, \quad (1)$$

where $p(m_i)$ are prior model probabilities.

Proof: From Bayes’ theorem (\rightarrow Proof I/2.1.1), the posterior model probability (\rightarrow Definition “pmp”) of model m_i can be derived as

$$p(m_i|y) = \frac{p(y|m_i) p(m_i)}{p(y)}. \quad (2)$$

Using the law of marginal probability (\rightarrow Definition “prob-marg”), the denominator can be written as

$$p(m_i|y) = \frac{p(y|m_i) p(m_i)}{\sum_{j=1}^M p(y|m_j) p(m_j)}. \quad (3)$$

The definition of the log model evidence (\rightarrow Definition IV/3.1.1)

$$\text{LME}(m) = \log p(y|m) \quad (4)$$

can be exponentiated to then read

$$\exp[\text{LME}(m)] = p(y|m) \quad (5)$$

and applying (5) to (3), we finally have:

$$p(m_i|y) = \frac{\exp[\text{LME}(m_i)] p(m_i)}{\sum_{j=1}^M \exp[\text{LME}(m_j)] p(m_j)}. \quad (6)$$

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, \dots, m_M be M statistical models describing the same measured data y with log model evidences (\rightarrow Definition IV/3.1.1) $\text{LME}(m_1), \dots, \text{LME}(m_M)$ and shared model parameters θ . Then, Bayesian model averaging (BMA) determines the following posterior distribution over θ :

$$p(\theta|y) = \sum_{i=1}^M p(\theta|m_i, y) \cdot \frac{\exp[\text{LME}(m_i)] p(m_i)}{\sum_{j=1}^M \exp[\text{LME}(m_j)] p(m_j)}, \quad (1)$$

where $p(\theta|m_i, y)$ is the posterior distributions over θ obtained using m_i .

Proof: According to the law of marginal probability (\rightarrow 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 :

$$p(\theta|y) = \sum_{i=1}^M p(\theta|m_i, y) \cdot p(m_i|y) , \quad (2)$$

where $p(m_i|y)$ is the posterior probability (\rightarrow Definition “pmp”) of the i -th model. One can express posterior model probabilities in terms of log model evidences (\rightarrow Proof IV/3.2.5) as

$$p(m_i|y) = \frac{\exp[\text{LME}(m_i)] p(m_i)}{\sum_{j=1}^M \exp[\text{LME}(m_j)] p(m_j)} \quad (3)$$

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.