

The Book of Statistical Proofs

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

2020-09-09, 08:49

Contents

I	General Theorems	1
1	Probability theory	2
1.1	Random variables	2
1.1.1	<i>Random variable</i>	2
1.1.2	<i>Random vector</i>	2
1.1.3	<i>Random matrix</i>	2
1.1.4	<i>Constant</i>	2
1.2	Probability	3
1.2.1	<i>Probability</i>	3
1.2.2	<i>Joint probability</i>	3
1.2.3	<i>Marginal probability</i>	4
1.2.4	<i>Conditional probability</i>	4
1.2.5	<i>Statistical independence</i>	4
1.3	Probability distributions	5
1.3.1	<i>Probability distribution</i>	5
1.3.2	<i>Joint distribution</i>	5
1.3.3	<i>Marginal distribution</i>	6
1.3.4	<i>Conditional distribution</i>	6
1.4	Probability functions	6
1.4.1	<i>Probability mass function</i>	6
1.4.2	<i>Probability density function</i>	7
1.4.3	<i>Cumulative distribution function</i>	7
1.4.4	<i>Quantile function</i>	8
1.4.5	<i>Moment-generating function</i>	8
1.4.6	Moment-generating function of linear transformation	9
1.4.7	Moment-generating function of linear combination	9
1.4.8	<i>Cumulant-generating function</i>	10
1.4.9	<i>Probability-generating function</i>	11
1.5	Expected value	11
1.5.1	<i>Definition</i>	11
1.5.2	Non-negative random variable	12
1.5.3	Non-negativity	13
1.5.4	Linearity	13
1.5.5	Monotonicity	15
1.5.6	(Non-)Multiplicativity	16
1.5.7	Expectation of a quadratic form	18
1.5.8	Law of the unconscious statistician	19
1.6	Variance	20

1.6.1	<i>Definition</i>	20
1.6.2	Partition into expected values	21
1.6.3	Non-negativity	21
1.6.4	Variance of a constant	22
1.6.5	Variance equals zero	22
1.6.6	Invariance under addition	23
1.6.7	Scaling upon multiplication	24
1.6.8	Variance of a sum	24
1.6.9	Variance of linear combination	25
1.6.10	Additivity under independence	26
1.7	Covariance	26
1.7.1	<i>Definition</i>	26
1.7.2	Partition into expected values	26
1.7.3	Covariance under independence	27
1.7.4	Relationship to correlation	28
1.7.5	<i>Covariance matrix</i>	28
1.7.6	Covariance matrix and expected values	28
1.7.7	Covariance matrix and correlation matrix	29
1.7.8	<i>Precision matrix</i>	30
1.7.9	Precision matrix and correlation matrix	30
1.8	Correlation	32
1.8.1	<i>Definition</i>	32
1.8.2	<i>Correlation matrix</i>	32
1.9	Measures of statistical dispersion	32
1.9.1	<i>Standard deviation</i>	32
1.9.2	<i>Full width at half maximum</i>	33
1.10	Further moments	33
1.10.1	<i>Moment</i>	33
1.10.2	Moment in terms of moment-generating function	34
1.10.3	First central moment is zero	35
2	Information theory	37
2.1	Shannon entropy	37
2.1.1	<i>Definition</i>	37
2.1.2	Non-negativity	37
2.1.3	Concavity	38
2.1.4	<i>Conditional entropy</i>	39
2.1.5	<i>Joint entropy</i>	39
2.1.6	<i>Cross-entropy</i>	40
2.1.7	Convexity of cross-entropy	40
2.1.8	Gibbs' inequality	41
2.1.9	Log sum inequality	42
2.2	Differential entropy	43
2.2.1	<i>Definition</i>	43
2.2.2	Negativity	43
2.2.3	<i>Conditional differential entropy</i>	44
2.2.4	<i>Joint differential entropy</i>	44
2.2.5	<i>Differential cross-entropy</i>	45
2.3	Discrete mutual information	45

2.3.1	<i>Definition</i>	45
2.3.2	Relation to marginal and conditional entropy	46
2.3.3	Relation to marginal and joint entropy	47
2.3.4	Relation to joint and conditional entropy	48
2.4	Continuous mutual information	49
2.4.1	<i>Definition</i>	49
2.4.2	Relation to marginal and conditional differential entropy	49
2.4.3	Relation to marginal and joint differential entropy	51
2.4.4	Relation to joint and conditional differential entropy	52
2.5	Kullback-Leibler divergence	53
2.5.1	<i>Definition</i>	53
2.5.2	Non-negativity	53
2.5.3	Non-negativity	54
2.5.4	Non-symmetry	55
2.5.5	Convexity	56
2.5.6	Additivity for independent distributions	57
2.5.7	Invariance under parameter transformation	58
2.5.8	Relation to discrete entropy	59
2.5.9	Relation to differential entropy	60
3	Estimation theory	61
3.1	Point estimates	61
3.1.1	Partition of the mean squared error into bias and variance	61
3.2	Interval estimates	62
3.2.1	Construction of confidence intervals using Wilks' theorem	62
4	Frequentist statistics	64
4.1	Likelihood theory	64
4.1.1	<i>Likelihood function</i>	64
4.1.2	<i>Log-likelihood function</i>	64
4.1.3	<i>Maximum likelihood estimation</i>	64
4.1.4	<i>Maximum log-likelihood</i>	65
5	Bayesian statistics	66
5.1	Probabilistic modeling	66
5.1.1	<i>Generative model</i>	66
5.1.2	<i>Likelihood function</i>	66
5.1.3	<i>Prior distribution</i>	66
5.1.4	<i>Full probability model</i>	67
5.1.5	<i>Joint likelihood</i>	67
5.1.6	Joint likelihood is product of likelihood and prior	67
5.1.7	<i>Posterior distribution</i>	68
5.1.8	Posterior density is proportional to joint likelihood	68
5.1.9	<i>Marginal likelihood</i>	69
5.1.10	Marginal likelihood is integral of joint likelihood	69
5.2	Bayesian inference	70
5.2.1	Bayes' theorem	70
5.2.2	Bayes' rule	70

II	Probability Distributions	73
1	Univariate discrete distributions	74
1.1	Discrete uniform distribution	74
1.1.1	<i>Definition</i>	74
1.1.2	Probability mass function	74
1.1.3	Cumulative distribution function	75
1.1.4	Quantile function	76
1.2	Bernoulli distribution	76
1.2.1	<i>Definition</i>	76
1.2.2	Probability mass function	77
1.2.3	Mean	77
1.3	Binomial distribution	78
1.3.1	<i>Definition</i>	78
1.3.2	Probability mass function	78
1.3.3	Mean	79
1.4	Poisson distribution	80
1.4.1	<i>Definition</i>	80
1.4.2	Probability mass function	80
1.4.3	Mean	81
2	Multivariate discrete distributions	82
2.1	Categorical distribution	82
2.1.1	<i>Definition</i>	82
2.1.2	Probability mass function	82
2.1.3	Mean	82
2.2	Multinomial distribution	83
2.2.1	<i>Definition</i>	83
2.2.2	Probability mass function	83
2.2.3	Mean	84
3	Univariate continuous distributions	86
3.1	Continuous uniform distribution	86
3.1.1	<i>Definition</i>	86
3.1.2	Probability density function	86
3.1.3	Cumulative distribution function	87
3.1.4	Quantile function	88
3.1.5	Mean	89
3.1.6	Median	89
3.1.7	Mode	90
3.2	Normal distribution	91
3.2.1	<i>Definition</i>	91
3.2.2	<i>Standard normal distribution</i>	91
3.2.3	Relation to standard normal distribution	92
3.2.4	Probability density function	92
3.2.5	Cumulative distribution function	93
3.2.6	Cumulative distribution function without error function	94
3.2.7	Quantile function	96
3.2.8	Mean	97
3.2.9	Median	98
3.2.10	Mode	99

	3.2.11	Variance	100
	3.2.12	Full width at half maximum	102
	3.2.13	Differential entropy	103
	3.2.14	Moment-generating function	104
3.3		Gamma distribution	106
	3.3.1	<i>Definition</i>	106
	3.3.2	<i>Standard gamma distribution</i>	106
	3.3.3	Relation to standard gamma distribution	107
	3.3.4	Probability density function	107
	3.3.5	Mean	108
	3.3.6	Variance	109
	3.3.7	Logarithmic expectation	110
	3.3.8	Kullback-Leibler divergence	112
3.4		Exponential distribution	113
	3.4.1	<i>Definition</i>	113
	3.4.2	Special case of gamma distribution	113
	3.4.3	Probability density function	114
	3.4.4	Cumulative distribution function	114
	3.4.5	Quantile function	115
	3.4.6	Mean	116
	3.4.7	Median	117
	3.4.8	Mode	117
3.5		Beta distribution	118
	3.5.1	<i>Definition</i>	118
	3.5.2	Probability density function	119
3.6		Wald distribution	119
	3.6.1	<i>Definition</i>	119
	3.6.2	Probability density function	119
4		Multivariate continuous distributions	121
	4.1	Multivariate normal distribution	121
		4.1.1 <i>Definition</i>	121
		4.1.2 Probability density function	121
		4.1.3 Differential entropy	121
		4.1.4 Kullback-Leibler divergence	123
		4.1.5 Linear transformation	124
		4.1.6 Marginal distributions	125
		4.1.7 Conditional distributions	126
	4.2	Normal-gamma distribution	129
		4.2.1 <i>Definition</i>	129
		4.2.2 Probability density function	130
		4.2.3 Kullback-Leibler divergence	131
		4.2.4 Marginal distributions	132
		4.2.5 Conditional distributions	135
	4.3	Dirichlet distribution	137
		4.3.1 <i>Definition</i>	137
		4.3.2 Probability density function	137
5		Matrix-variate continuous distributions	138
	5.1	Matrix-normal distribution	138

5.1.1	<i>Definition</i>	138
5.1.2	Probability density function	138
5.1.3	Equivalence to multivariate normal distribution	139
5.1.4	Transposition	140
5.1.5	Linear transformation	141
5.2	Wishart distribution	142
5.2.1	<i>Definition</i>	142
III Statistical Models		143
1	Univariate normal data	144
1.1	Multiple linear regression	144
1.1.1	<i>Definition</i>	144
1.1.2	Ordinary least squares	145
1.1.3	Ordinary least squares	145
1.1.4	<i>Total sum of squares</i>	146
1.1.5	<i>Explained sum of squares</i>	147
1.1.6	<i>Residual sum of squares</i>	147
1.1.7	Total, explained and residual sum of squares	147
1.1.8	<i>Estimation matrix</i>	149
1.1.9	<i>Projection matrix</i>	149
1.1.10	<i>Residual-forming matrix</i>	149
1.1.11	Estimation, projection and residual-forming matrix	150
1.1.12	Idempotence of projection and residual-forming matrix	151
1.1.13	Weighted least squares	152
1.1.14	Weighted least squares	153
1.1.15	Maximum likelihood estimation	154
1.2	Bayesian linear regression	156
1.2.1	Conjugate prior distribution	156
1.2.2	Posterior distribution	158
1.2.3	Log model evidence	160
1.2.4	Posterior probability of alternative hypothesis	162
1.2.5	Posterior credibility region excluding null hypothesis	164
2	Multivariate normal data	166
2.1	General linear model	166
2.1.1	<i>Definition</i>	166
2.1.2	Ordinary least squares	166
2.1.3	Weighted least squares	167
2.1.4	Maximum likelihood estimation	168
2.2	Multivariate Bayesian linear regression	170
2.2.1	Conjugate prior distribution	170
2.2.2	Posterior distribution	172
2.2.3	Log model evidence	174
3	Poisson data	177
3.1	Poisson-distributed data	177
3.1.1	<i>Definition</i>	177
3.1.2	Maximum likelihood estimation	177
3.2	Poisson distribution with exposure values	179
3.2.1	<i>Definition</i>	179

	3.2.2	Conjugate prior distribution	179
	3.2.3	Posterior distribution	180
	3.2.4	Log model evidence	182
4	Probability data		185
	4.1	Beta-distributed data	185
	4.1.1	<i>Definition</i>	185
	4.1.2	Method of moments	185
	4.2	Logistic regression	187
	4.2.1	<i>Definition</i>	187
	4.2.2	Probability and log-odds	188
	4.2.3	Log-odds and probability	188
5	Categorical data		190
	5.1	Binomial observations	190
	5.1.1	<i>Definition</i>	190
	5.1.2	Conjugate prior distribution	190
	5.1.3	Posterior distribution	191
	5.1.4	Log model evidence	192
	5.2	Multinomial observations	194
	5.2.1	<i>Definition</i>	194
	5.2.2	Conjugate prior distribution	194
	5.2.3	Posterior distribution	195
	5.2.4	Log model evidence	196
IV Model Selection			199
1	Goodness-of-fit measures		200
	1.1	Residual variance	200
	1.1.1	<i>Definition</i>	200
	1.1.2	Maximum likelihood estimator is biased	200
	1.1.3	Construction of unbiased estimator	202
	1.2	R-squared	203
	1.2.1	<i>Definition</i>	203
	1.2.2	Derivation of R^2 and adjusted R^2	204
	1.2.3	Relationship to maximum log-likelihood	205
	1.3	Signal-to-noise ratio	207
	1.3.1	<i>Definition</i>	207
	1.3.2	Relationship with R^2	207
2	Classical information criteria		209
	2.1	Akaike information criterion	209
	2.1.1	<i>Definition</i>	209
	2.2	Bayesian information criterion	209
	2.2.1	<i>Definition</i>	209
	2.2.2	Derivation	209
	2.3	Deviance information criterion	211
	2.3.1	<i>Definition</i>	211
3	Bayesian model selection		212
	3.1	Log model evidence	212
	3.1.1	<i>Definition</i>	212
	3.1.2	Derivation	212

	3.1.3	Partition into accuracy and complexity	213
3.2		Log family evidence	214
	3.2.1	<i>Definition</i>	214
	3.2.2	Derivation	215
	3.2.3	Calculation from log model evidences	216
3.3		Bayes factor	217
	3.3.1	<i>Definition</i>	217
	3.3.2	Transitivity	218
	3.3.3	Computation using Savage-Dickey Density Ratio	218
	3.3.4	Computation using Encompassing Prior Method	219
	3.3.5	<i>Encompassing model</i>	221
3.4		Log Bayes factor	221
	3.4.1	<i>Definition</i>	221
	3.4.2	Derivation	222
	3.4.3	Calculation from log model evidences	223
3.5		Posterior model probability	223
	3.5.1	<i>Definition</i>	223
	3.5.2	Derivation	224
	3.5.3	Calculation from Bayes factors	224
	3.5.4	Calculation from log Bayes factor	225
	3.5.5	Calculation from log model evidences	226
3.6		Bayesian model averaging	227
	3.6.1	<i>Definition</i>	227
	3.6.2	Derivation	227
	3.6.3	Calculation from log model evidences	228

Chapter I

General Theorems

1 Probability theory

1.1 Random variables

1.1.1 Random variable

Definition: A random variable may be understood

- informally, as a real number $X \in \mathbb{R}$ whose value is the outcome of a random experiment (\rightarrow Definition “rexp”);
- formally, as a measurable function (\rightarrow Definition “meas-fct”) X defined on a probability space (\rightarrow Definition “prob-spc”) (Ω, \mathcal{F}, P) that maps from a sample space Ω to the real numbers \mathbb{R} using an event space \mathcal{F} and a probability function (\rightarrow Definition I/1.4.1) P ;
- more broadly, as any random quantity X such as a random scalar (\rightarrow Definition I/1.1.1), a random vector (\rightarrow Definition I/1.1.2) or a random matrix (\rightarrow Definition I/1.1.3).

Sources:

- Wikipedia (2020): “Random variable”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-27; URL: https://en.wikipedia.org/wiki/Random_variable#Definition.

Metadata: ID: D65 | shortcut: rvar | author: JoramSoch | date: 2020-05-27, 22:36.

1.1.2 Random vector

Definition: A random vector, also called “multivariate random variable”, is an n -dimensional column vector $X \in \mathbb{R}^{n \times 1}$ whose entries are random variables (\rightarrow Definition I/1.1.1).

Sources:

- Wikipedia (2020): “Multivariate random variable”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-27; URL: https://en.wikipedia.org/wiki/Multivariate_random_variable.

Metadata: ID: D66 | shortcut: rvec | author: JoramSoch | date: 2020-05-27, 22:44.

1.1.3 Random matrix

Definition: A random matrix, also called “matrix-valued random variable”, is an $n \times p$ matrix $X \in \mathbb{R}^{n \times p}$ whose entries are random variables (\rightarrow Definition I/1.1.1). Equivalently, a random matrix is an $n \times p$ matrix whose columns are n -dimensional random vectors (\rightarrow Definition I/1.1.2).

Sources:

- Wikipedia (2020): “Random matrix”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-27; URL: https://en.wikipedia.org/wiki/Random_matrix.

Metadata: ID: D67 | shortcut: rmat | author: JoramSoch | date: 2020-05-27, 22:48.

1.1.4 Constant

Definition: A constant is a quantity which does not change and thus always has the same value. From a statistical perspective, a constant is a random variable (\rightarrow Definition I/1.1.1) which is equal

to its expected value (\rightarrow Definition I/1.5.1)

$$X = E(X) \quad (1)$$

or equivalently, whose variance (\rightarrow Definition I/1.6.1) is zero

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

Sources:

- ProofWiki (2020): “Definition: Constant”; in: *ProofWiki*, retrieved on 2020-09-09; URL: <https://proofwiki.org/wiki/Definition:Constant#Definition>.

Metadata: ID: D96 | shortcut: const | author: JoramSoch | date: 2020-09-09, 01:30.

1.2 Probability

1.2.1 Probability

Definition: Let E be a statement about an arbitrary event such as the outcome of a random experiment (\rightarrow Definition “rexp”). Then, $p(E)$ is called the probability of E and may be interpreted as

- (objectivist interpretation of probability:) some physical state of affairs, e.g. the relative frequency of occurrence of E , when repeating the experiment (“Frequentist probability”); or
- (subjectivist interpretation of probability:) a degree of belief in E , e.g. the price at which someone would buy or sell a bet that pays 1 unit of utility if E and 0 if not E (“Bayesian probability”).

Sources:

- Wikipedia (2020): “Probability”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-10; URL: <https://en.wikipedia.org/wiki/Probability#Interpretations>.

Metadata: ID: D48 | shortcut: prob | author: JoramSoch | date: 2020-05-10, 19:41.

1.2.2 Joint probability

Definition: Let A and B be two arbitrary statements about random variables (\rightarrow Definition I/1.1.1), such as statements about the presence or absence of an event or about the value of a scalar, vector or matrix. Then, $p(A, B)$ is called the joint probability of A and B and is defined as the probability (\rightarrow Definition I/1.2.1) that A and B are both true.

Sources:

- Wikipedia (2020): “Joint probability distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-10; URL: https://en.wikipedia.org/wiki/Joint_probability_distribution.

Metadata: ID: D49 | shortcut: prob-joint | author: JoramSoch | date: 2020-05-10, 19:49.

1.2.3 Marginal probability

Definition: Let A and X be two arbitrary statements about random variables (\rightarrow Definition I/1.1.1), such as statements about the presence or absence of an event or about the value of a scalar, vector or matrix. Furthermore, assume a joint probability (\rightarrow Definition I/1.2.2) distribution $p(A, X)$. Then, $p(A)$ is called the marginal probability of A and,

1) if X is a discrete random variable (\rightarrow Definition I/1.1.1) with domain \mathcal{X} , is given by

$$p(A) = \sum_{x \in \mathcal{X}} p(A, x) ; \quad (1)$$

2) if X is a continuous random variable (\rightarrow Definition I/1.1.1) with domain \mathcal{X} , is given by

$$p(A) = \int_{\mathcal{X}} p(A, x) dx . \quad (2)$$

Sources:

- Wikipedia (2020): “Marginal distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-10; URL: https://en.wikipedia.org/wiki/Marginal_distribution#Definition.

Metadata: ID: D50 | shortcut: prob-marg | author: JoramSoch | date: 2020-05-10, 20:01.

1.2.4 Conditional probability

Definition: Let A and B be two arbitrary statements about random variables (\rightarrow Definition I/1.1.1), such as statements about the presence or absence of an event or about the value of a scalar, vector or matrix. Furthermore, assume a joint probability (\rightarrow Definition I/1.2.2) distribution $p(A, B)$. Then, $p(A|B)$ is called the conditional probability that A is true, given that B is true, and is given by

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

where $p(B)$ is the marginal probability (\rightarrow Definition I/1.2.3) of B .

Sources:

- Wikipedia (2020): “Conditional probability”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-10; URL: https://en.wikipedia.org/wiki/Conditional_probability#Definition.

Metadata: ID: D51 | shortcut: prob-cond | author: JoramSoch | date: 2020-05-10, 20:06.

1.2.5 Statistical independence

Definition: Generally speaking, random variables (\rightarrow Definition I/1.1.1) are statistically independent, if their joint probability (\rightarrow Definition I/1.2.2) can be expressed in terms of their marginal probability (\rightarrow Definition I/1.2.3).

1) A set of discrete random variables (\rightarrow Definition I/1.1.1) X_1, \dots, X_n with possible values $\mathcal{X}_1, \dots, \mathcal{X}_n$ is called statistically independent, if

$$p(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^n p(X_i = x_i) \quad \text{for all } x_i \in \mathcal{X}_i, i = 1, \dots, n \quad (1)$$

where $p(x_1, \dots, x_n)$ are the joint probabilities (\rightarrow Definition I/1.2.2) of X_1, \dots, X_n and $p(x_i)$ are the marginal probabilities (\rightarrow Definition I/1.2.3) of X_i .

2) A set of continuous random variables (\rightarrow Definition I/1.1.1) X_1, \dots, X_n defined on the domains $\mathcal{X}_1, \dots, \mathcal{X}_n$ is called statistically independent, if

$$F_{X_1, \dots, X_n}(x_1, \dots, x_n) = \prod_{i=1}^n F_{X_i}(x_i) \quad \text{for all } x_i \in \mathcal{X}_i, i = 1, \dots, n \quad (2)$$

or equivalently, if the probability densities (\rightarrow Definition I/1.4.2) exist,

$$f_{X_1, \dots, X_n}(x_1, \dots, x_n) = \prod_{i=1}^n f_{X_i}(x_i) \quad \text{for all } x_i \in \mathcal{X}_i, i = 1, \dots, n \quad (3)$$

where F are the joint (\rightarrow Definition I/1.3.2) or marginal (\rightarrow Definition I/1.3.3) cumulative distribution functions (\rightarrow Definition I/1.4.3) and f are the respective probability density functions (\rightarrow Definition I/1.4.2).

Sources:

- Wikipedia (2020): “Independence (probability theory)”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-06-06; URL: [https://en.wikipedia.org/wiki/Independence_\(probability_theory\)#Definition](https://en.wikipedia.org/wiki/Independence_(probability_theory)#Definition).

Metadata: ID: D75 | shortcut: ind | author: JoramSoch | date: 2020-06-06, 07:16.

1.3 Probability distributions

1.3.1 Probability distribution

****Definition**:** Let X be a random variable (\rightarrow Definition I/1.1.1) with the set of possible outcomes \mathcal{X} . Then, a probability distribution of X is a mathematical function that gives the probabilities (\rightarrow Definition I/1.2.1) of occurrence of all possible outcomes $x \in \mathcal{X}$ of this random variable.

Sources:

- Wikipedia (2020): “Probability distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-17; URL: https://en.wikipedia.org/wiki/Probability_distribution.

Metadata: ID: D55 | shortcut: dist | author: JoramSoch | date: 2020-05-17, 20:23.

1.3.2 Joint distribution

****Definition**:** Let X and Y be random variables (\rightarrow Definition I/1.1.1) with sets of possible outcomes \mathcal{X} and \mathcal{Y} . Then, a joint distribution of X and Y is a probability distribution (\rightarrow Definition I/1.3.1) that specifies the probability of the event that $X = x$ and $Y = y$ for each possible combination of $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.

- The joint distribution of two scalar random variables (\rightarrow Definition I/1.1.1) is called a bivariate distribution.
- The joint distribution of a random vector (\rightarrow Definition I/1.1.2) is called a multivariate distribution.
- The joint distribution of a random matrix (\rightarrow Definition I/1.1.3) is called a matrix-variate distribution.

Sources:

- Wikipedia (2020): “Joint probability distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-17; URL: https://en.wikipedia.org/wiki/Joint_probability_distribution.

Metadata: ID: D56 | shortcut: dist-joint | author: JoramSoch | date: 2020-05-17, 20:43.

1.3.3 Marginal distribution

****Definition**:** Let X and Y be random variables (\rightarrow Definition I/1.1.1) with sets of possible outcomes \mathcal{X} and \mathcal{Y} . Then, the marginal distribution of X is a probability distribution (\rightarrow Definition I/1.3.1) that specifies the probability of the event that $X = x$ irrespective of the value of Y for each possible value $x \in \mathcal{X}$. The marginal distribution can be obtained from the joint distribution (\rightarrow Definition I/1.3.2) of X and Y using the law of marginal probability (\rightarrow Definition I/1.2.3).

Sources:

- Wikipedia (2020): “Marginal distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-17; URL: https://en.wikipedia.org/wiki/Marginal_distribution.

Metadata: ID: D57 | shortcut: dist-marg | author: JoramSoch | date: 2020-05-17, 21:02.

1.3.4 Conditional distribution

****Definition**:** Let X and Y be random variables (\rightarrow Definition I/1.1.1) with sets of possible outcomes \mathcal{X} and \mathcal{Y} . Then, the conditional distribution of X given that Y is a probability distribution (\rightarrow Definition I/1.3.1) that specifies the probability of the event that $X = x$ given that $Y = y$ for each possible combination of $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. The conditional distribution of X can be obtained from the joint distribution (\rightarrow Definition I/1.3.2) of X and Y and the marginal distribution (\rightarrow Definition I/1.3.3) of Y using the law of conditional probability (\rightarrow Definition I/1.2.4).

Sources:

- Wikipedia (2020): “Conditional probability distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-17; URL: https://en.wikipedia.org/wiki/Conditional_probability_distribution.

Metadata: ID: D58 | shortcut: dist-cond | author: JoramSoch | date: 2020-05-17, 21:25.

1.4 Probability functions

1.4.1 Probability mass function

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

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

for all $x \notin \mathcal{X}$,

$$\Pr(X = x) = f_X(x) \quad (2)$$

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

$$\sum_{x \in \mathcal{X}} f_X(x) = 1 . \quad (3)$$

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.

1.4.2 Probability density function

Definition: Let X be a continuous random variable (\rightarrow Definition I/1.1.1) 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.

1.4.3 Cumulative distribution function

Definition: The cumulative distribution function (CDF) of a random variable (\rightarrow Definition I/1.1.1) X at a given value x is defined as the probability (\rightarrow Definition I/1.2.1) that X is smaller than x :

$$F_X(x) = \Pr(X \leq x) . \quad (1)$$

1) Let X be a discrete random variable (\rightarrow Definition I/1.1.1) with possible outcomes \mathcal{X} and the probability mass function (\rightarrow Definition I/1.4.1) $f_X(x)$. Then, the function $F_X(x) : \mathbb{R} \rightarrow [0, 1]$ with

$$F_X(x) = \sum_{\substack{z \in \mathcal{X} \\ z \leq x}} f_X(z) \quad (2)$$

is the cumulative distribution function of X .

2) Let X be a scalar continuous random variable (\rightarrow Definition I/1.1.1) with the probability density function (\rightarrow Definition I/1.4.2) $f_X(x)$. Then, the function $F_X(x) : \mathbb{R} \rightarrow [0, 1]$ with

$$F_X(x) = \int_{-\infty}^x f_X(z) dz \quad (3)$$

is the cumulative distribution function of X .

Sources:

- Wikipedia (2020): “Cumulative distribution 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.

1.4.4 Quantile function

Definition: Let X be a random variable (\rightarrow Definition I/1.1.1) with the cumulative distribution function (\rightarrow Definition I/1.4.3) (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.

1.4.5 Moment-generating function

Definition:

1) The moment-generating function of a random variable (\rightarrow Definition I/1.1.1) $X \in \mathbb{R}$ 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 I/1.1.2) $X \in \mathbb{R}^n$ 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.

1.4.6 Moment-generating function of linear transformation

Theorem: Let X be an $n \times 1$ random vector (\rightarrow Definition I/1.1.2) with the moment-generating function (\rightarrow Definition I/1.4.5) $M_X(t)$. Then, the moment-generating function of the linear transformation $Y = AX + b$ is given by

$$M_Y(t) = \exp [t^T b] \cdot M_X(At) \quad (1)$$

where A is an $m \times n$ matrix and b is an $m \times 1$ vector.

Proof: The moment-generating function of a random vector (\rightarrow Definition I/1.4.5) X is

$$M_X(t) = E \left(\exp [t^T X] \right) \quad (2)$$

and therefore the moment-generating function of the random vector (\rightarrow Definition I/1.1.2) Y is given by

$$\begin{aligned} M_Y(t) &= E \left(\exp [t^T (AX + b)] \right) \\ &= E \left(\exp [t^T AX] \cdot \exp [t^T b] \right) \\ &= \exp [t^T b] \cdot E \left(\exp [(At)^T X] \right) \\ &= \exp [t^T b] \cdot M_X(At) . \end{aligned} \quad (3)$$

Sources:

- ProofWiki (2020): “Moment Generating Function of Linear Transformation of Random Variable”; in: *ProofWiki*, retrieved on 2020-08-19; URL: https://proofwiki.org/wiki/Moment_Generating_Function_of_Linear_Transformation_of_Random_Variable.

Metadata: ID: P154 | shortcut: mgf-ltt | author: JoramSoch | date: 2020-08-19, 08:09.

1.4.7 Moment-generating function of linear combination

Theorem: Let X_1, \dots, X_n be n independent (\rightarrow Definition I/1.2.5) random variables (\rightarrow Definition I/1.1.1) with moment-generating functions (\rightarrow Definition I/1.4.5) $M_{X_i}(t)$. Then, the moment-generating function of the linear combination $X = \sum_{i=1}^n a_i X_i$ is given by

$$M_X(t) = \prod_{i=1}^n M_{X_i}(a_i t) \quad (1)$$

where a_1, \dots, a_n are n real numbers.

Proof: The moment-generating function of a random variable (\rightarrow Definition I/1.4.5) X_i is

$$M_{X_i}(t) = \mathbb{E}(\exp[tX_i]) \quad (2)$$

and therefore the moment-generating function of the linear combination X is given by

$$\begin{aligned} M_X(t) &= \mathbb{E}(\exp[tX]) \\ &= \mathbb{E}\left(\exp\left[t \sum_{i=1}^n a_i X_i\right]\right) \\ &= \mathbb{E}\left(\prod_{i=1}^n \exp[t a_i X_i]\right). \end{aligned} \quad (3)$$

Because the expected value is multiplicative for independent random variables (\rightarrow Proof I/1.5.6), we have

$$\begin{aligned} M_X(t) &= \prod_{i=1}^n \mathbb{E}(\exp[(a_i t)X_i]) \\ &= \prod_{i=1}^n M_{X_i}(a_i t). \end{aligned} \quad (4)$$

Sources:

- ProofWiki (2020): “Moment Generating Function of Linear Combination of Independent Random Variables”; in: *ProofWiki*, retrieved on 2020-08-19; URL: https://proofwiki.org/wiki/Moment_Generating_Function_of_Linear_Combination_of_Independent_Random_Variables.

Metadata: ID: P155 | shortcut: mgf-lincomb | author: JoramSoch | date: 2020-08-19, 08:36.

1.4.8 Cumulant-generating function

Definition:

1) The cumulant-generating function of a random variable (\rightarrow Definition I/1.1.1) $X \in \mathbb{R}$ is

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

2) The cumulant-generating function of a random vector (\rightarrow Definition I/1.1.2) $X \in \mathbb{R}^n$ is

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

Sources:

- Wikipedia (2020): “Cumulant”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-31; URL: <https://en.wikipedia.org/wiki/Cumulant#Definition>.

Metadata: ID: D68 | shortcut: cgf | author: JoramSoch | date: 2020-05-31, 23:46.

1.4.9 Probability-generating function

Definition:

1) If X is a discrete random variable (\rightarrow Definition I/1.1.1) taking values in the non-negative integers $\{0, 1, \dots\}$, then the probability-generating function of X is defined as

$$G_X(z) = E[z^X] = \sum_{x=0}^{\infty} p(x) z^x \quad (1)$$

where $z \in \mathbb{C}$ and $p(x)$ is the probability mass function (\rightarrow Definition I/1.4.1) of X .

2) If X is a discrete random vector (\rightarrow Definition I/1.1.2) taking values in the n -dimensional integer lattice $x \in \{0, 1, \dots\}^n$, then the probability-generating function of X is defined as

$$G_X(z) = E[z_1^{X_1} \cdot \dots \cdot z_n^{X_n}] = \sum_{x_1=0}^{\infty} \dots \sum_{x_n=0}^{\infty} p(x_1, \dots, x_n) z_1^{x_1} \cdot \dots \cdot z_n^{x_n} \quad (2)$$

where $z \in \mathbb{C}^n$ and $p(x)$ is the probability mass function (\rightarrow Definition I/1.4.1) of X .

Sources:

- Wikipedia (2020): “Probability-generating function”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-31; URL: https://en.wikipedia.org/wiki/Probability-generating_function#Definition.

Metadata: ID: D69 | shortcut: pgf | author: JoramSoch | date: 2020-05-31, 23:59.

1.5 Expected value

1.5.1 Definition

Definition:

1) The expected value (or, mean) of a discrete random variable (\rightarrow Definition I/1.1.1) 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/1.4.1) of X .

2) The expected value (or, mean) of a continuous random variable (\rightarrow Definition I/1.1.1) 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/1.4.2) 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.

1.5.2 Non-negative random variable

Theorem: Let X be a non-negative random variable (\rightarrow Definition I/1.1.1). Then, the expected value (\rightarrow Definition I/1.5.1) of X is

$$E(X) = \int_0^{\infty} (1 - F_X(x)) dx \quad (1)$$

where $F_X(x)$ is the cumulative distribution function (\rightarrow Definition I/1.4.3) of X .

Proof: Because the cumulative distribution function gives the probability of a random variable being smaller than a given value (\rightarrow Definition I/1.4.3),

$$F_X(x) = \Pr(X \leq x), \quad (2)$$

we have

$$1 - F_X(x) = \Pr(X > x), \quad (3)$$

such that

$$\int_0^{\infty} (1 - F_X(x)) dx = \int_0^{\infty} \Pr(X > x) dx \quad (4)$$

which, using the probability density function (\rightarrow Definition I/1.4.2) of X , can be rewritten as

$$\begin{aligned} \int_0^{\infty} (1 - F_X(x)) dx &= \int_0^{\infty} \int_x^{\infty} f_X(z) dz dx \\ &= \int_0^{\infty} \int_0^z f_X(z) dx dz \\ &= \int_0^{\infty} f_X(z) \int_0^z 1 dx dz \\ &= \int_0^{\infty} [x]_0^z \cdot f_X(z) dz \\ &= \int_0^{\infty} z \cdot f_X(z) dz \end{aligned} \quad (5)$$

and by applying the definition of the expected value (\rightarrow Definition I/1.5.1), we see that

$$\int_0^{\infty} (1 - F_X(x)) dx = \int_0^{\infty} z \cdot f_X(z) dz = E(X) \quad (6)$$

which proves the identity given above.

Sources:

- Kemp, Graham (2014): “Expected value of a non-negative random variable”; in: *StackExchange Mathematics*, retrieved on 2020-05-18; URL: <https://math.stackexchange.com/questions/958472/expected-value-of-a-non-negative-random-variable>.

Metadata: ID: P103 | shortcut: mean-nnrvar | author: JoramSoch | date: 2020-05-18, 23:54.

1.5.3 Non-negativity

Theorem: If a random variable (\rightarrow Definition I/1.1.1) is strictly non-negative, its expected value (\rightarrow Definition I/1.5.1) 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/1.4.1) 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/1.4.2) 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.

1.5.4 Linearity

Theorem: The expected value (\rightarrow Definition I/1.5.1) 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 I/1.1.1) X and Y and a constant a .

Proof:

1) If X and Y are discrete random variables, the expected value (\rightarrow Definition I/1.5.1) is

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

and the law of marginal probability (\rightarrow Definition I/1.2.3) states that

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

Applying this, we have

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

as well as

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

2) If X and Y are continuous random variables, the expected value (\rightarrow Definition I/1.5.1) is

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

and the law of marginal probability (\rightarrow Definition I/1.2.3) 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

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

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*, retrieved on 2020-02-13; URL: <https://brilliant.org/wiki/linearity-of-expectation/>.

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

1.5.5 Monotonicity

Theorem: The expected value (\rightarrow Definition I/1.5.1) is monotonic, i.e.

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

Proof: Let $Z = Y - X$. Due to the linearity of the expected value (\rightarrow Proof I/1.5.4), we have

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

With the non-negativity property of the expected value (\rightarrow Proof I/1.5.3), it also holds that

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

Together with (2), this yields

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

Sources:

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

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

1.5.6 (Non-)Multiplicativity

Theorem:

1) If two random variables (\rightarrow Definition I/1.1.1) X and Y are independent (\rightarrow Definition I/1.2.5), the expected value (\rightarrow Definition I/1.5.1) is multiplicative, i.e.

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

2) If two random variables (\rightarrow Definition I/1.1.1) X and Y are dependent (\rightarrow Definition I/1.2.5), the expected value (\rightarrow Definition I/1.5.1) 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 I/1.2.5), 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/1.5.1), we have

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

And applying it to the expected value for continuous random variables (\rightarrow Definition I/1.5.1), we have

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

2) Let X and Y be Bernoulli random variables (\rightarrow Definition II/1.2.1) with the following joint probability (\rightarrow Definition I/1.2.2) mass function (\rightarrow Definition I/1.4.1)

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

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.

1.5.7 Expectation of a quadratic form

Theorem: Let X be an $n \times 1$ random vector (\rightarrow Definition I/1.1.2) with mean (\rightarrow Definition I/1.5.1) μ and covariance (\rightarrow Definition I/1.7.1) Σ and let A be a symmetric $n \times n$ matrix. Then, the expectation of the quadratic form $X^T A X$ is

$$E [X^T A X] = \mu^T A \mu + \text{tr}(A \Sigma) . \quad (1)$$

Proof: Note that $X^T A X$ is a 1×1 matrix. We can therefore write

$$E [X^T A X] = E [\text{tr} (X^T A X)] . \quad (2)$$

Using the trace property $\text{tr}(ABC) = \text{tr}(BCA)$, this becomes

$$E [X^T A X] = E [\text{tr} (A X X^T)] . \quad (3)$$

Because mean and trace are linear operators (\rightarrow Proof I/1.5.4), we have

$$E [X^T A X] = \text{tr} (A E [X X^T]) . \quad (4)$$

Note that the covariance matrix can be partitioned into expected values (\rightarrow Proof I/1.7.6)

$$\text{Cov}(X, X) = E(X X^T) - E(X)E(X)^T , \quad (5)$$

such that the expected value of the quadratic form becomes

$$E [X^T A X] = \text{tr} (A [\text{Cov}(X, X) + E(X)E(X)^T]) . \quad (6)$$

Finally, applying mean and covariance of X , we have

$$\begin{aligned} E [X^T A X] &= \text{tr} (A [\Sigma + \mu \mu^T]) \\ &= \text{tr} (A \Sigma + A \mu \mu^T) \\ &= \text{tr}(A \Sigma) + \text{tr}(A \mu \mu^T) \\ &= \text{tr}(A \Sigma) + \text{tr}(\mu^T A \mu) \\ &= \mu^T A \mu + \text{tr}(A \Sigma) . \end{aligned} \quad (7)$$

Sources:

- Kendrick, David (1981): “Expectation of a quadratic form”; in: *Stochastic Control for Economic Models*, pp. 170-171.
- Wikipedia (2020): “Multivariate random variable”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-07-13; URL: https://en.wikipedia.org/wiki/Multivariate_random_variable#Expectation_of_a_quadratic_form.
- Halvorsen, Kjetil B. (2012): “Expected value and variance of trace function”; in: *StackExchange Cross Validated*, retrieved on 2020-07-13; URL: <https://stats.stackexchange.com/questions/34477/expected-value-and-variance-of-trace-function>.
- Sarwate, Dilip (2013): “Expected Value of Quadratic Form”; in: *StackExchange Cross Validated*, retrieved on 2020-07-13; URL: <https://stats.stackexchange.com/questions/48066/expected-value-of-quadratic-form>.

Metadata: ID: P131 | shortcut: mean-qf | author: JoramSoch | date: 2020-07-13, 21:59.

1.5.8 Law of the unconscious statistician

Theorem: Let X be a random variable (\rightarrow Definition I/1.1.1) and let $Y = g(X)$ be a function of this random variable.

1) If X is a discrete random variable with possible outcomes \mathcal{X} and probability mass function (\rightarrow Definition I/1.4.1) $f_X(x)$, the expected value (\rightarrow Definition I/1.5.1) of $g(X)$ is

$$E[g(X)] = \sum_{x \in \mathcal{X}} g(x) f_X(x) . \quad (1)$$

2) If X is a continuous random variable with possible outcomes \mathcal{X} and probability density function (\rightarrow Definition I/1.4.2) $f_X(x)$, the expected value (\rightarrow Definition I/1.5.1) of $g(X)$ is

$$E[g(X)] = \int_{\mathcal{X}} g(x) f_X(x) dx . \quad (2)$$

Proof: Suppose that g is differentiable and that its inverse g^{-1} is monotonic.

1) The expected value (\rightarrow Definition I/1.5.1) of $Y = g(X)$ is defined as

$$E[Y] = \sum_{y \in \mathcal{Y}} y f_Y(y) . \quad (3)$$

Writing the probability mass function $f_Y(y)$ in terms of $y = g(x)$, we have:

$$E[g(X)] = \sum_{y \in \mathcal{Y}} y \Pr(g(x) = y) . \quad (4)$$

Replacing the probability that $g(x) = y$ in terms of $f_X(x)$, this becomes:

$$E[g(X)] = \sum_{y \in \mathcal{Y}} y \sum_{x: g(x)=y} f_X(x) . \quad (5)$$

Observe that $y = g(x)$ can be moved into the inner sum. Finally, noting that “for all y , then for all x , such that $g(x) = y$ ” is equivalent to “for all x ” if g^{-1} is a monotonic function, we can conclude that

$$E[g(X)] = \sum_{x \in \mathcal{X}} g(x) f_X(x) . \quad (6)$$

2) Let $y = g(x)$. The derivative of an inverse function is

$$\frac{d}{dy}(g^{-1}(y)) = \frac{1}{g'(g^{-1}(y))} \quad (7)$$

Because $x = g^{-1}(y)$, this can be rearranged into

$$dx = \frac{1}{g'(g^{-1}(y))} dy \quad (8)$$

and substituting (8) into (2), we get

$$\int_{\mathcal{X}} g(x) f_X(x) dx = \int_{\mathcal{Y}} y f_X(g^{-1}(y)) \frac{1}{g'(g^{-1}(y))} dy . \quad (9)$$

Considering the cumulative distribution function (\rightarrow Definition I/1.4.3) of Y , one can deduce:

$$\begin{aligned} F_Y(y) &= \Pr(Y \leq y) \\ &= \Pr(g(X) \leq y) \\ &= \Pr(X \leq g^{-1}(y)) \\ &= F_X(g^{-1}(y)) . \end{aligned} \tag{10}$$

Differentiating to get the probability density function (\rightarrow Definition I/1.4.2) of Y , the result is:

$$\begin{aligned} f_Y(y) &= \frac{d}{dy} F_Y(y) \\ &\stackrel{(10)}{=} \frac{d}{dy} F_X(g^{-1}(y)) \\ &= f_X(g^{-1}(y)) \frac{d}{dy} (g^{-1}(y)) \\ &\stackrel{(7)}{=} f_X(g^{-1}(y)) \frac{1}{g'(g^{-1}(y))} . \end{aligned} \tag{11}$$

Finally, substituting (11) into (9), we have:

$$\int_{\mathcal{X}} g(x) f_X(x) dx = \int_{\mathcal{Y}} y f_Y(y) dy = E[Y] = E[g(X)] . \tag{12}$$

Sources:

- Wikipedia (2020): “Law of the unconscious statistician”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-07-22; URL: https://en.wikipedia.org/wiki/Law_of_the_unconscious_statistician#Proof.

Metadata: ID: P138 | shortcut: mean-lotus | author: JoramSoch | date: 2020-07-22, 08:30.

1.6 Variance

1.6.1 Definition

Definition: The variance of a random variable (\rightarrow Definition I/1.1.1) X is defined as the expected value (\rightarrow Definition I/1.5.1) of the squared deviation from its expected value (\rightarrow Definition I/1.5.1):

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

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.

1.6.2 Partition into expected values

Theorem: Let X be a random variable (\rightarrow Definition I/1.1.1). Then, the variance (\rightarrow Definition I/1.6.1) of X is equal to the mean (\rightarrow Definition I/1.5.1) of the square of X minus the square of the mean (\rightarrow Definition I/1.5.1) of X :

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

Proof: The variance (\rightarrow Definition I/1.6.1) of X is defined as

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

which, due to the linearity of the expected value (\rightarrow Proof I/1.5.4), can be rewritten as

$$\begin{aligned} \text{Var}(X) &= E[(X - E[X])^2] \\ &= E[X^2 - 2X E(X) + E(X)^2] \\ &= E(X^2) - 2E(X)E(X) + E(X)^2 \\ &= E(X^2) - E(X)^2 . \end{aligned} \quad (3)$$

Sources:

- Wikipedia (2020): “Variance”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-19; URL: <https://en.wikipedia.org/wiki/Variance#Definition>.

Metadata: ID: P104 | shortcut: var-mean | author: JoramSoch | date: 2020-05-19, 00:17.

1.6.3 Non-negativity

Theorem: The variance (\rightarrow Definition I/1.6.1) is always non-negative, i.e.

$$\text{Var}(X) \geq 0 . \quad (1)$$

Proof: The variance (\rightarrow Definition I/1.6.1) of a random variable (\rightarrow Definition I/1.1.1) is defined as

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

1) If X is a discrete random variable (\rightarrow Definition I/1.1.1), then, because squares and probabilities are strictly non-negative, all the addends in

$$\text{Var}(X) = \sum_{x \in \mathcal{X}} (x - E(X))^2 \cdot f_X(x) \quad (3)$$

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

2) If X is a continuous random variable (\rightarrow Definition I/1.1.1), then, because squares and probability densities are strictly non-negative, the integrand in

$$\text{Var}(X) = \int_{\mathcal{X}} (x - E(X))^2 \cdot f_X(x) dx \quad (4)$$

is always 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): “Variance”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-06-06; URL: https://en.wikipedia.org/wiki/Variance#Basic_properties.

Metadata: ID: P123 | shortcut: var-nonneg | author: JoramSoch | date: 2020-06-06, 07:29.

1.6.4 Variance of a constant

Theorem: The variance (\rightarrow Definition I/1.6.1) of a constant (\rightarrow Definition I/1.1.4) is zero:

$$a = \text{const.} \quad \Rightarrow \quad \text{Var}(a) = 0 . \quad (1)$$

Proof: A constant (\rightarrow Definition I/1.1.4) is a quantity that always has the same value. Thus, if understood as a random variable (\rightarrow Definition I/1.1.1), the expected value (\rightarrow Definition I/1.5.1) of a constant is equal to itself:

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

Plugged into the formula of the variance (\rightarrow Definition I/1.6.1), we have

$$\begin{aligned} \text{Var}(a) &= \text{E} [(a - \text{E}(a))^2] \\ &= \text{E} [(a - a)^2] \\ &= \text{E}(0) . \end{aligned} \quad (3)$$

Applied to the formula of the expected value (\rightarrow Definition I/1.5.1), this gives

$$\text{E}(0) = \sum_{x=0} x \cdot f_X(x) = 0 \cdot 1 = 0 . \quad (4)$$

Together, (3) and (4) imply (1).

Sources:

- Wikipedia (2020): “Variance”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-06-27; URL: https://en.wikipedia.org/wiki/Variance#Basic_properties.

Metadata: ID: P124 | shortcut: var-const | author: JoramSoch | date: 2020-06-27, 06:44.

1.6.5 Variance equals zero

Theorem: If the variance (\rightarrow Definition I/1.6.1) of X is zero, then X is a constant (\rightarrow Definition I/1.1.4):

$$\text{Var}(X) = 0 \quad \Rightarrow \quad X = \text{const.} \quad (1)$$

Proof: The variance (\rightarrow Definition I/1.6.1) is defined as

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

Because $(X - \text{E}(X))^2$ is strictly non-negative (\rightarrow Proof I/1.5.3), the only way for the variance to become zero is, if the squared deviation is always zero:

$$(X - \text{E}(X))^2 = 0 . \quad (3)$$

Thus, in turn, requires that X is equal to its expected value (\rightarrow Definition I/1.5.1)

$$X = \text{E}(X) \quad (4)$$

which can only be the case, if X always has the same value (\rightarrow Definition I/1.1.4):

$$X = \text{const.} \quad (5)$$

Sources:

- Wikipedia (2020): “Variance”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-06-27; URL: https://en.wikipedia.org/wiki/Variance#Basic_properties.

Metadata: ID: P125 | shortcut: var-zero | author: JoramSoch | date: 2020-06-27, 07:05.

1.6.6 Invariance under addition

Theorem: The variance (\rightarrow Definition I/1.6.1) is invariant under addition of a constant (\rightarrow Definition I/1.1.4):

$$\text{Var}(X + a) = \text{Var}(X) \quad (1)$$

Proof: The variance (\rightarrow Definition I/1.6.1) is defined in terms of the expected value (\rightarrow Definition I/1.5.1) as

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

Using this and the linearity of the expected value (\rightarrow Proof I/1.5.4), we can derive (1) as follows:

$$\begin{aligned} \text{Var}(X + a) &\stackrel{(2)}{=} \text{E} [((X + a) - \text{E}(X + a))^2] \\ &= \text{E} [(X + a - \text{E}(X) - a)^2] \\ &= \text{E} [(X - \text{E}(X))^2] \\ &\stackrel{(2)}{=} \text{Var}(X) . \end{aligned} \quad (3)$$

Sources:

- Wikipedia (2020): “Variance”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-07-07; URL: https://en.wikipedia.org/wiki/Variance#Basic_properties.

Metadata: ID: P126 | shortcut: var-inv | author: JoramSoch | date: 2020-07-07, 05:23.

1.6.7 Scaling upon multiplication

Theorem: The variance (\rightarrow Definition I/1.6.1) scales upon multiplication with a constant (\rightarrow Definition I/1.1.4):

$$\text{Var}(aX) = a^2 \text{Var}(X) \quad (1)$$

Proof: The variance (\rightarrow Definition I/1.6.1) is defined in terms of the expected value (\rightarrow Definition I/1.5.1) as

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

Using this and the linearity of the expected value (\rightarrow Proof I/1.5.4), we can derive (1) as follows:

$$\begin{aligned} \text{Var}(aX) &\stackrel{(2)}{=} E[((aX) - E(aX))^2] \\ &= E[(aX - aE(X))^2] \\ &= E[(a(X - E(X)))^2] \\ &= E[a^2(X - E(X))^2] \\ &= a^2 E[(X - E(X))^2] \\ &\stackrel{(2)}{=} a^2 \text{Var}(X) . \end{aligned} \quad (3)$$

Sources:

- Wikipedia (2020): “Variance”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-07-07; URL: https://en.wikipedia.org/wiki/Variance#Basic_properties.

Metadata: ID: P127 | shortcut: var-scal | author: JoramSoch | date: 2020-07-07, 05:38.

1.6.8 Variance of a sum

Theorem: The variance (\rightarrow Definition I/1.6.1) of the sum of two random variables (\rightarrow Definition I/1.1.1) equals the sum of the variances of those random variables, plus two times their covariance (\rightarrow Definition I/1.7.1):

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2 \text{Cov}(X, Y) . \quad (1)$$

Proof: The variance (\rightarrow Definition I/1.6.1) is defined in terms of the expected value (\rightarrow Definition I/1.5.1) as

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

Using this and the linearity of the expected value (\rightarrow Proof I/1.5.4), we can derive (1) as follows:

$$\begin{aligned}
\text{Var}(X + Y) &\stackrel{(2)}{=} \mathbb{E} [((X + Y) - \mathbb{E}(X + Y))^2] \\
&= \mathbb{E} [(X - \mathbb{E}(X)) + (Y - \mathbb{E}(Y))]^2 \\
&= \mathbb{E} [(X - \mathbb{E}(X))^2 + (Y - \mathbb{E}(Y))^2 + 2(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))] \\
&= \mathbb{E} [(X - \mathbb{E}(X))^2] + \mathbb{E} [(Y - \mathbb{E}(Y))^2] + \mathbb{E} [2(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))] \\
&\stackrel{(2)}{=} \text{Var}(X) + \text{Var}(Y) + 2 \text{Cov}(X, Y) .
\end{aligned} \tag{3}$$

Sources:

- Wikipedia (2020): “Variance”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-07-07; URL: https://en.wikipedia.org/wiki/Variance#Basic_properties.

Metadata: ID: P128 | shortcut: var-sum | author: JoramSoch | date: 2020-07-07, 06:10.

1.6.9 Variance of linear combination

Theorem: The variance (\rightarrow Definition I/1.6.1) of the linear combination of two random variables (\rightarrow Definition I/1.1.1) is a function of the variances as well as the covariance (\rightarrow Definition I/1.7.1) of those random variables:

$$\text{Var}(aX + bY) = a^2 \text{Var}(X) + b^2 \text{Var}(Y) + 2ab \text{Cov}(X, Y) . \tag{1}$$

Proof: The variance (\rightarrow Definition I/1.6.1) is defined in terms of the expected value (\rightarrow Definition I/1.5.1) as

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

Using this and the linearity of the expected value (\rightarrow Proof I/1.5.4), we can derive (1) as follows:

$$\begin{aligned}
\text{Var}(aX + bY) &\stackrel{(2)}{=} \mathbb{E} [((aX + bY) - \mathbb{E}(aX + bY))^2] \\
&= \mathbb{E} [(a(X - \mathbb{E}(X)) + b(Y - \mathbb{E}(Y)))^2] \\
&= \mathbb{E} [a^2 (X - \mathbb{E}(X))^2 + b^2 (Y - \mathbb{E}(Y))^2 + 2ab (X - \mathbb{E}(X))(Y - \mathbb{E}(Y))] \\
&= \mathbb{E} [a^2 (X - \mathbb{E}(X))^2] + \mathbb{E} [b^2 (Y - \mathbb{E}(Y))^2] + \mathbb{E} [2ab (X - \mathbb{E}(X))(Y - \mathbb{E}(Y))] \\
&\stackrel{(2)}{=} a^2 \text{Var}(X) + b^2 \text{Var}(Y) + 2ab \text{Cov}(X, Y) .
\end{aligned} \tag{3}$$

Sources:

- Wikipedia (2020): “Variance”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-07-07; URL: https://en.wikipedia.org/wiki/Variance#Basic_properties.

Metadata: ID: P129 | shortcut: var-lincomb | author: JoramSoch | date: 2020-07-07, 06:21.

1.6.10 Additivity under independence

Theorem: The variance (\rightarrow Definition I/1.6.1) is additive for independent (\rightarrow Definition I/1.2.5) random variables (\rightarrow Definition I/1.1.1):

$$p(X, Y) = p(X)p(Y) \quad \Rightarrow \quad \text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) . \quad (1)$$

Proof: The variance of the sum of two random variables (\rightarrow Proof I/1.6.8) is given by

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2 \text{Cov}(X, Y) . \quad (2)$$

The covariance of independent random variables (\rightarrow Proof I/1.7.3) is zero:

$$p(X, Y) = p(X)p(Y) \quad \Rightarrow \quad \text{Cov}(X, Y) = 0 . \quad (3)$$

Combining (2) and (3), we have:

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) . \quad (4)$$

Sources:

- Wikipedia (2020): “Variance”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-07-07; URL: https://en.wikipedia.org/wiki/Variance#Basic_properties.

Metadata: ID: P130 | shortcut: var-add | author: JoramSoch | date: 2020-07-07, 06:52.

1.7 Covariance

1.7.1 Definition

Definition: The covariance of two random variables (\rightarrow Definition I/1.1.1) X and Y is defined as the expected value (\rightarrow Definition I/1.5.1) of the product of their deviations from their individual expected values (\rightarrow Definition I/1.5.1):

$$\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])] . \quad (1)$$

Sources:

- Wikipedia (2020): “Covariance”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-06; URL: <https://en.wikipedia.org/wiki/Covariance#Definition>.

Metadata: ID: D70 | shortcut: cov | author: JoramSoch | date: 2020-06-02, 20:20.

1.7.2 Partition into expected values

Theorem: Let X and Y be random variables (\rightarrow Definition I/1.1.1). Then, the covariance (\rightarrow Definition I/1.7.1) of X and Y is equal to the mean (\rightarrow Definition I/1.5.1) of the product of X and Y minus the product of the means (\rightarrow Definition I/1.5.1) of X and Y :

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y) . \quad (1)$$

Proof: The covariance (\rightarrow Definition I/1.7.1) of X and Y is defined as

$$\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])] . \quad (2)$$

which, due to the linearity of the expected value (\rightarrow Proof I/1.5.4), can be rewritten as

$$\begin{aligned} \text{Cov}(X, Y) &= E[(X - E[X])(Y - E[Y])] \\ &= E[XY - X E(Y) - E(X) Y + E(X)E(Y)] \\ &= E(XY) - E(X)E(Y) - E(X)E(Y) + E(X)E(Y) \\ &= E(XY) - E(X)E(Y) . \end{aligned} \quad (3)$$

Sources:

- Wikipedia (2020): “Covariance”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-06-02; URL: <https://en.wikipedia.org/wiki/Covariance#Definition>.

Metadata: ID: P118 | shortcut: cov-mean | author: JoramSoch | date: 2020-06-02, 20:50.

1.7.3 Covariance under independence

Theorem: Let X and Y be independent (\rightarrow Definition I/1.2.5) random variables (\rightarrow Definition I/1.1.1). Then, the covariance (\rightarrow Definition I/1.7.1) of X and Y is zero:

$$X, Y \text{ independent} \quad \Rightarrow \quad \text{Cov}(X, Y) = 0 . \quad (1)$$

Proof: The covariance can be expressed in terms of expected values (\rightarrow Proof I/1.7.2) as

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y) . \quad (2)$$

For independent random variables, the expected value of the product is equal to the product of the expected values (\rightarrow Proof I/1.5.6):

$$E(XY) = E(X)E(Y) . \quad (3)$$

Taking (2) and (3) together, we have

$$\begin{aligned} \text{Cov}(X, Y) &\stackrel{(2)}{=} E(XY) - E(X)E(Y) \\ &\stackrel{(3)}{=} E(X)E(Y) - E(X)E(Y) \\ &= 0 . \end{aligned} \quad (4)$$

Sources:

- Wikipedia (2020): “Covariance”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-09-03; URL: https://en.wikipedia.org/wiki/Covariance#Uncorrelatedness_and_independence.

Metadata: ID: P158 | shortcut: cov-ind | author: JoramSoch | date: 2020-09-03, 06:05.

1.7.4 Relationship to correlation

Theorem: Let X and Y be random variables (\rightarrow Definition I/1.1.1). Then, the covariance (\rightarrow Definition I/1.7.1) of X and Y is equal to the product of their correlation (\rightarrow Definition I/1.8.1) and the standard deviations (\rightarrow Definition I/1.9.1) of X and Y :

$$\text{Cov}(X, Y) = \sigma_X \text{Corr}(X, Y) \sigma_Y . \quad (1)$$

Proof: The correlation (\rightarrow Definition I/1.8.1) of X and Y is defined as

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} . \quad (2)$$

which can be rearranged for the covariance (\rightarrow Definition I/1.7.1) to give

$$\text{Cov}(X, Y) = \sigma_X \text{Corr}(X, Y) \sigma_Y \quad (3)$$

Sources:

- original work

Metadata: ID: P119 | shortcut: cov-corr | author: JoramSoch | date: 2020-06-02, 21:00.

1.7.5 Covariance matrix

Definition: Let $X = [X_1, \dots, X_n]^T$ be a random vector (\rightarrow Definition I/1.1.2). Then, the covariance matrix of X is defined as the $n \times n$ matrix in which the entry (i, j) is the covariance (\rightarrow Definition I/1.7.1) of X_i and X_j :

$$\Sigma_{XX} = \begin{bmatrix} \text{Cov}(X_1, X_1) & \dots & \text{Cov}(X_1, X_n) \\ \vdots & \ddots & \vdots \\ \text{Cov}(X_n, X_1) & \dots & \text{Cov}(X_n, X_n) \end{bmatrix} = \begin{bmatrix} \text{E}[(X_1 - \text{E}[X_1])(X_1 - \text{E}[X_1])] & \dots & \text{E}[(X_1 - \text{E}[X_1])(X_n - \text{E}[X_n])] \\ \vdots & \ddots & \vdots \\ \text{E}[(X_n - \text{E}[X_n])(X_1 - \text{E}[X_1])] & \dots & \text{E}[(X_n - \text{E}[X_n])(X_n - \text{E}[X_n])] \end{bmatrix} \quad (1)$$

Sources:

- Wikipedia (2020): “Covariance matrix”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-06-06; URL: https://en.wikipedia.org/wiki/Covariance_matrix#Definition.

Metadata: ID: D72 | shortcut: covmat | author: JoramSoch | date: 2020-06-06, 04:24.

1.7.6 Covariance matrix and expected values

Theorem: Let X be a random vector (\rightarrow Definition I/1.1.2). Then, the covariance matrix (\rightarrow Definition I/1.7.5) of X is equal to the mean (\rightarrow Definition I/1.5.1) of the outer product of X with itself minus the outer product of the mean (\rightarrow Definition I/1.5.1) of X with itself:

$$\Sigma_{XX} = \text{E}(XX^T) - \text{E}(X)\text{E}(X)^T . \quad (1)$$

Proof: The covariance matrix (\rightarrow Definition I/1.7.5) of X is defined as

$$\Sigma_{XX} = \begin{bmatrix} E[(X_1 - E[X_1])(X_1 - E[X_1])] & \dots & E[(X_1 - E[X_1])(X_n - E[X_n])] \\ \vdots & \ddots & \vdots \\ E[(X_n - E[X_n])(X_1 - E[X_1])] & \dots & E[(X_n - E[X_n])(X_n - E[X_n])] \end{bmatrix} \quad (2)$$

which can also be expressed using matrix multiplication as

$$\Sigma_{XX} = E[(X - E[X])(X - E[X])^T] \quad (3)$$

Due to the linearity of the expected value (\rightarrow Proof I/1.5.4), this can be rewritten as

$$\begin{aligned} \Sigma_{XX} &= E[(X - E[X])(X - E[X])^T] \\ &= E[XX^T - XE(X)^T - E(X)X^T + E(X)E(X)^T] \\ &= E(XX^T) - E(X)E(X)^T - E(X)E(X)^T + E(X)E(X)^T \\ &= E(XX^T) - E(X)E(X)^T. \end{aligned} \quad (4)$$

Sources:

- Taboga, Marco (2010): “Covariance matrix”; in: *Lectures on probability and statistics*, retrieved on 2020-06-06; URL: <https://www.statlect.com/fundamentals-of-probability/covariance-matrix>.

Metadata: ID: P120 | shortcut: covmat-mean | author: JoramSoch | date: 2020-06-06, 05:31.

1.7.7 Covariance matrix and correlation matrix

Theorem: Let X be a random vector (\rightarrow Definition I/1.1.2). Then, the covariance matrix (\rightarrow Definition I/1.7.5) of X can be expressed in terms of its correlation matrix (\rightarrow Definition I/1.8.2) as follows

$$\Sigma_{XX} = D_X \cdot C_{XX} \cdot D_X, \quad (1)$$

where D_X is a diagonal matrix with the standard deviations (\rightarrow Definition I/1.9.1) of X_1, \dots, X_n as entries on the diagonal:

$$D_X = \text{diag}(\sigma_{X_1}, \dots, \sigma_{X_n}) = \begin{bmatrix} \sigma_{X_1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_{X_n} \end{bmatrix}. \quad (2)$$

Proof: Reiterating (1) and applying (2), we have:

$$\Sigma_{XX} = \begin{bmatrix} \sigma_{X_1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_{X_n} \end{bmatrix} \cdot C_{XX} \cdot \begin{bmatrix} \sigma_{X_1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_{X_n} \end{bmatrix}. \quad (3)$$

Together with the definition of the correlation matrix (\rightarrow Definition I/1.8.2), this gives

$$\begin{aligned}
\Sigma_{XX} &= \begin{bmatrix} \sigma_{X_1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_{X_n} \end{bmatrix} \cdot \begin{bmatrix} \frac{E[(X_1 - E[X_1])(X_1 - E[X_1])]}{\sigma_{X_1} \sigma_{X_1}} & \dots & \frac{E[(X_1 - E[X_1])(X_n - E[X_n])]}{\sigma_{X_1} \sigma_{X_n}} \\ \vdots & \ddots & \vdots \\ \frac{E[(X_n - E[X_n])(X_1 - E[X_1])]}{\sigma_{X_n} \sigma_{X_1}} & \dots & \frac{E[(X_n - E[X_n])(X_n - E[X_n])]}{\sigma_{X_n} \sigma_{X_n}} \end{bmatrix} \cdot \begin{bmatrix} \sigma_{X_1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_{X_n} \end{bmatrix} \\
&= \begin{bmatrix} \frac{\sigma_{X_1} \cdot E[(X_1 - E[X_1])(X_1 - E[X_1])]}{\sigma_{X_1} \sigma_{X_1}} & \dots & \frac{\sigma_{X_1} \cdot E[(X_1 - E[X_1])(X_n - E[X_n])]}{\sigma_{X_1} \sigma_{X_n}} \\ \vdots & \ddots & \vdots \\ \frac{\sigma_{X_n} \cdot E[(X_n - E[X_n])(X_1 - E[X_1])]}{\sigma_{X_n} \sigma_{X_1}} & \dots & \frac{\sigma_{X_n} \cdot E[(X_n - E[X_n])(X_n - E[X_n])]}{\sigma_{X_n} \sigma_{X_n}} \end{bmatrix} \cdot \begin{bmatrix} \sigma_{X_1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_{X_n} \end{bmatrix} \\
&= \begin{bmatrix} \frac{\sigma_{X_1} \cdot E[(X_1 - E[X_1])(X_1 - E[X_1]) \cdot \sigma_{X_1}]}{\sigma_{X_1} \sigma_{X_1}} & \dots & \frac{\sigma_{X_1} \cdot E[(X_1 - E[X_1])(X_n - E[X_n]) \cdot \sigma_{X_n}]}{\sigma_{X_1} \sigma_{X_n}} \\ \vdots & \ddots & \vdots \\ \frac{\sigma_{X_n} \cdot E[(X_n - E[X_n])(X_1 - E[X_1]) \cdot \sigma_{X_1}]}{\sigma_{X_n} \sigma_{X_1}} & \dots & \frac{\sigma_{X_n} \cdot E[(X_n - E[X_n])(X_n - E[X_n]) \cdot \sigma_{X_n}]}{\sigma_{X_n} \sigma_{X_n}} \end{bmatrix} \\
&= \begin{bmatrix} E[(X_1 - E[X_1])(X_1 - E[X_1])] & \dots & E[(X_1 - E[X_1])(X_n - E[X_n])] \\ \vdots & \ddots & \vdots \\ E[(X_n - E[X_n])(X_1 - E[X_1])] & \dots & E[(X_n - E[X_n])(X_n - E[X_n])] \end{bmatrix}
\end{aligned} \tag{4}$$

which is nothing else than the definition of the covariance matrix (\rightarrow Definition I/1.7.5).

Sources:

- Penny, William (2006): “The correlation matrix”; in: *Mathematics for Brain Imaging*, ch. 1.4.5, p. 28, eq. 1.60; URL: https://ueapsylabs.co.uk/sites/wpenny/mbi/mbi_course.pdf.

Metadata: ID: P121 | shortcut: covmat-covmat | author: JoramSoch | date: 2020-06-06, 06:02.

1.7.8 Precision matrix

Definition: Let $X = [X_1, \dots, X_n]^T$ be a random vector (\rightarrow Definition I/1.1.2). Then, the precision matrix of X is defined as the inverse of the covariance matrix (\rightarrow Definition I/1.7.5) of X :

$$\Lambda_{XX} = \Sigma_{XX}^{-1} = \begin{bmatrix} \text{Cov}(X_1, X_1) & \dots & \text{Cov}(X_1, X_n) \\ \vdots & \ddots & \vdots \\ \text{Cov}(X_n, X_1) & \dots & \text{Cov}(X_n, X_n) \end{bmatrix}^{-1}. \tag{1}$$

Sources:

- Wikipedia (2020): “Precision (statistics)”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-06-06; URL: [https://en.wikipedia.org/wiki/Precision_\(statistics\)](https://en.wikipedia.org/wiki/Precision_(statistics)).

Metadata: ID: D74 | shortcut: precmat | author: JoramSoch | date: 2020-06-06, 05:08.

1.7.9 Precision matrix and correlation matrix

Theorem: Let X be a random vector (\rightarrow Definition I/1.1.2). Then, the precision matrix (\rightarrow Definition I/1.7.8) of X can be expressed in terms of its correlation matrix (\rightarrow Definition I/1.8.2) as follows

$$\Lambda_{XX} = D_X^{-1} \cdot C_{XX}^{-1} \cdot D_X^{-1}, \quad (1)$$

where D_X^{-1} is a diagonal matrix with the inverse standard deviations (\rightarrow Definition I/1.9.1) of X_1, \dots, X_n as entries on the diagonal:

$$D_X^{-1} = \text{diag}(1/\sigma_{X_1}, \dots, 1/\sigma_{X_n}) = \begin{bmatrix} \frac{1}{\sigma_{X_1}} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \frac{1}{\sigma_{X_n}} \end{bmatrix}. \quad (2)$$

Proof: The precision matrix (\rightarrow Definition I/1.7.8) is defined as the inverse of the covariance matrix (\rightarrow Definition I/1.7.5)

$$\Lambda_{XX} = \Sigma_{XX}^{-1} \quad (3)$$

and the relation between covariance matrix and correlation matrix (\rightarrow Proof I/1.7.7) is given by

$$\Sigma_{XX} = D_X \cdot C_{XX} \cdot D_X \quad (4)$$

where

$$D_X = \text{diag}(\sigma_{X_1}, \dots, \sigma_{X_n}) = \begin{bmatrix} \sigma_{X_1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_{X_n} \end{bmatrix}. \quad (5)$$

Using the matrix product property

$$(A \cdot B \cdot C)^{-1} = C^{-1} \cdot B^{-1} \cdot A^{-1} \quad (6)$$

and the diagonal matrix property

$$\text{diag}(a_1, \dots, a_n)^{-1} = \begin{bmatrix} a_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & a_n \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{a_1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \frac{1}{a_n} \end{bmatrix} = \text{diag}(1/a_1, \dots, 1/a_n), \quad (7)$$

we obtain

$$\begin{aligned} \Lambda_{XX} &\stackrel{(3)}{=} \Sigma_{XX}^{-1} \\ &\stackrel{(4)}{=} (D_X \cdot C_{XX} \cdot D_X)^{-1} \\ &\stackrel{(6)}{=} D_X^{-1} \cdot C_{XX}^{-1} \cdot D_X^{-1} \\ &\stackrel{(7)}{=} \begin{bmatrix} \frac{1}{\sigma_{X_1}} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \frac{1}{\sigma_{X_n}} \end{bmatrix} \cdot C_{XX}^{-1} \cdot \begin{bmatrix} \frac{1}{\sigma_{X_1}} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \frac{1}{\sigma_{X_n}} \end{bmatrix} \end{aligned} \quad (8)$$

which conforms to equation (1).

Sources:

- original work

Metadata: ID: P122 | shortcut: precmat-corrmat | author: JoramSoch | date: 2020-06-06, 06:28.

1.8 Correlation

1.8.1 Definition

Definition: The correlation of two random variables (\rightarrow Definition I/1.1.1) X and Y , also called Pearson product-moment correlation coefficient (PPMCC), is defined as the ratio of the covariance (\rightarrow Definition I/1.7.1) of X and Y relative to the product of their standard deviations (\rightarrow Definition I/1.9.1):

$$\text{Corr}(X, Y) = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)} \sqrt{\text{Var}(Y)}} = \frac{\text{E}[(X - \text{E}[X])(Y - \text{E}[Y])]}{\sqrt{\text{E}[(X - \text{E}[X])^2]} \sqrt{\text{E}[(Y - \text{E}[Y])^2]}}. \quad (1)$$

Sources:

- Wikipedia (2020): “Correlation and dependence”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-06; URL: https://en.wikipedia.org/wiki/Correlation_and_dependence#Pearson's_product-moment_coefficient.

Metadata: ID: D71 | shortcut: corr | author: JoramSoch | date: 2020-06-02, 20:34.

1.8.2 Correlation matrix

Definition: Let $X = [X_1, \dots, X_n]^T$ be a random vector (\rightarrow Definition I/1.1.2). Then, the correlation matrix of X is defined as the $n \times n$ matrix in which the entry (i, j) is the correlation (\rightarrow Definition I/1.8.1) of X_i and X_j :

$$C_{XX} = \begin{bmatrix} \text{Corr}(X_1, X_1) & \dots & \text{Corr}(X_1, X_n) \\ \vdots & \ddots & \vdots \\ \text{Corr}(X_n, X_1) & \dots & \text{Corr}(X_n, X_n) \end{bmatrix} = \begin{bmatrix} \frac{\text{E}[(X_1 - \text{E}[X_1])(X_1 - \text{E}[X_1])]}{\sigma_{X_1} \sigma_{X_1}} & \dots & \frac{\text{E}[(X_1 - \text{E}[X_1])(X_n - \text{E}[X_n])]}{\sigma_{X_1} \sigma_{X_n}} \\ \vdots & \ddots & \vdots \\ \frac{\text{E}[(X_n - \text{E}[X_n])(X_1 - \text{E}[X_1])]}{\sigma_{X_n} \sigma_{X_1}} & \dots & \frac{\text{E}[(X_n - \text{E}[X_n])(X_n - \text{E}[X_n])]}{\sigma_{X_n} \sigma_{X_n}} \end{bmatrix}. \quad (1)$$

Sources:

- Wikipedia (2020): “Correlation and dependence”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-06-06; URL: https://en.wikipedia.org/wiki/Correlation_and_dependence#Correlation_matrices.

Metadata: ID: D73 | shortcut: corrmatrix | author: JoramSoch | date: 2020-06-06, 04:56.

1.9 Measures of statistical dispersion

1.9.1 Standard deviation

Definition: The standard deviation σ of a random variable (\rightarrow Definition I/1.1.1) X with expected value (\rightarrow Definition I/1.5.1) μ is defined as the square root of the variance (\rightarrow Definition I/1.6.1), i.e.

$$\sigma(X) = \sqrt{\mathbb{E}[(X - \mu)^2]} . \quad (1)$$

Sources:

- Wikipedia (2020): “Standard deviation”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-09-03; URL: https://en.wikipedia.org/wiki/Standard_deviation#Definition_of_population_values.

Metadata: ID: D94 | shortcut: std | author: JoramSoch | date: 2020-09-03, 05:43.

1.9.2 Full width at half maximum

Definition: Let X be a continuous random variable (\rightarrow Definition I/1.1.1) with a unimodal probability density function (\rightarrow Definition I/1.4.2) $f_X(x)$ and mode (\rightarrow Definition “mode”) x_M . Then, the full width at half maximum of X is defined as

$$\text{FWHM}(X) = \Delta x = x_2 - x_1 \quad (1)$$

where x_1 and x_2 are specified, such that

$$f_X(x_1) = f_X(x_2) = \frac{1}{2}f_X(x_M) \quad \text{and} \quad x_1 < x_M < x_2 \quad (2)$$

Sources:

- Wikipedia (2020): “Full width at half maximum”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-08-19; URL: https://en.wikipedia.org/wiki/Full_width_at_half_maximum.

Metadata: ID: D91 | shortcut: fwhm | author: JoramSoch | date: 2020-08-19, 05:40.

1.10 Further moments**1.10.1 Moment**

Definition: Let X be a random variable (\rightarrow Definition I/1.1.1) and let n be a positive integer. Then, the n -th moment of X , also called (n -th) “raw moment” or “crude moment”, is defined as the expected value (\rightarrow Definition I/1.5.1) of the n -th power of X :

$$\mu'_n = \mathbb{E}[X^n] . \quad (1)$$

Sources:

- Wikipedia (2020): “Moment (mathematics)”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-08-19; URL: [https://en.wikipedia.org/wiki/Moment_\(mathematics\)#Significance_of_the_moments](https://en.wikipedia.org/wiki/Moment_(mathematics)#Significance_of_the_moments).

Metadata: ID: D90 | shortcut: mom | author: JoramSoch | date: 2020-08-19, 05:24.

1.10.2 Moment in terms of moment-generating function

Theorem: Let X be a scalar random variable (\rightarrow Definition I/1.1.1) with the moment-generating function (\rightarrow Definition I/1.4.5) $M_X(t)$. Then, the n -th moment (\rightarrow Definition I/1.10.1) of X can be calculated from the moment-generating function via

$$E(X^n) = M_X^{(n)}(0) \quad (1)$$

where n is a positive integer and $M_X^{(n)}(t)$ is the n -th derivative of $M_X(t)$.

Proof: Using the definition of the moment-generating function (\rightarrow Definition I/1.4.5), we can write:

$$M_X^{(n)}(t) = \frac{d^n}{dt^n} E(e^{tX}) . \quad (2)$$

Using the power series expansion of the exponential function

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} , \quad (3)$$

equation (2) becomes

$$M_X^{(n)}(t) = \frac{d^n}{dt^n} E \left(\sum_{m=0}^{\infty} \frac{t^m X^m}{m!} \right) . \quad (4)$$

Because the expected value is a linear operator (\rightarrow Proof I/1.5.4), we have:

$$\begin{aligned} M_X^{(n)}(t) &= \frac{d^n}{dt^n} \sum_{m=0}^{\infty} E \left(\frac{t^m X^m}{m!} \right) \\ &= \sum_{m=0}^{\infty} \frac{d^n}{dt^n} \frac{t^m}{m!} E(X^m) . \end{aligned} \quad (5)$$

Using the n -th derivative of the m -th power

$$\frac{d^n}{dx^n} x^m = \begin{cases} m^n x^{m-n} , & \text{if } n \leq m \\ 0 , & \text{if } n > m . \end{cases} \quad (6)$$

with the falling factorial

$$m^n = \prod_{i=0}^{n-1} (m - i) = \frac{m!}{(m - n)!} , \quad (7)$$

equation (5) becomes

$$\begin{aligned}
M_X^{(n)}(t) &= \sum_{m=n}^{\infty} \frac{m^n t^{m-n}}{m!} E(X^m) \\
&\stackrel{(7)}{=} \sum_{m=n}^{\infty} \frac{m! t^{m-n}}{(m-n)! m!} E(X^m) \\
&= \sum_{m=n}^{\infty} \frac{t^{m-n}}{(m-n)!} E(X^m) \\
&= \frac{t^{n-n}}{(n-n)!} E(X^n) + \sum_{m=n+1}^{\infty} \frac{t^{m-n}}{(m-n)!} E(X^m) \\
&= \frac{t^0}{0!} E(X^n) + \sum_{m=n+1}^{\infty} \frac{t^{m-n}}{(m-n)!} E(X^m) \\
&= E(X^n) + \sum_{m=n+1}^{\infty} \frac{t^{m-n}}{(m-n)!} E(X^m) .
\end{aligned} \tag{8}$$

Setting $t = 0$ in (8) yields

$$\begin{aligned}
M_X^{(n)}(0) &= E(X^n) + \sum_{m=n+1}^{\infty} \frac{0^{m-n}}{(m-n)!} E(X^m) \\
&= E(X^n)
\end{aligned} \tag{9}$$

which conforms to equation (1).

Sources:

- ProofWiki (2020): “Moment in terms of Moment Generating Function”; in: *ProofWiki*, retrieved on 2020-08-19; URL: https://proofwiki.org/wiki/Moment_in_terms_of_Moment_Generating_Function.

Metadata: ID: P153 | shortcut: mom-mgf | author: JoramSoch | date: 2020-08-19, 07:51.

1.10.3 First central moment is zero

Theorem: The first central moment (\rightarrow Definition “mom-cent”) is zero, i.e.

$$\mu_1 = 0 . \tag{1}$$

Proof: The first central moment (\rightarrow Definition “mom-cent”) of a random variable (\rightarrow Definition I/1.1.1) X with mean (\rightarrow Definition I/1.5.1) μ is defined as

$$\mu_1 = E[(X - \mu)^1] . \tag{2}$$

Due to the linearity of the expected value (\rightarrow Proof I/1.5.4) and by plugging in $\mu = E(X)$, we have

$$\begin{aligned}\mu_1 &= \mathbb{E}[X - \mu] \\ &= \mathbb{E}(X) - \mu \\ &= \mathbb{E}(X) - \mathbb{E}(X) \\ &= 0 .\end{aligned}\tag{3}$$

Sources:

- ProofWiki (2020): “First Central Moment is Zero”; in: *ProofWiki*, retrieved on 2020-09-09; URL: https://proofwiki.org/wiki/First_Central_Moment_is_Zero.

Metadata: ID: P167 | shortcut: momcent-1st | author: JoramSoch | date: 2020-08-19, 07:51.

2 Information theory

2.1 Shannon entropy

2.1.1 Definition

Definition: Let X be a discrete random variable (\rightarrow Definition I/1.1.1) with possible outcomes \mathcal{X} and the (observed or assumed) probability mass function (\rightarrow Definition I/1.4.1) $p(x) = f_X(x)$. Then, the entropy (also referred to as “Shannon entropy”) of X is defined as

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

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.

2.1.2 Non-negativity

Theorem: The entropy of a discrete random variable (\rightarrow Definition I/1.1.1) is a non-negative number:

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

Proof: The entropy of a discrete random variable (\rightarrow Definition I/2.1.1) is defined as

$$H(X) = - \sum_{x \in \mathcal{X}} p(x) \cdot \log_b p(x) \quad (2)$$

The minus sign can be moved into the sum:

$$H(X) = \sum_{x \in \mathcal{X}} [p(x) \cdot (-\log_b p(x))] \quad (3)$$

Because the co-domain of probability mass functions (\rightarrow Definition I/1.4.1) is $[0, 1]$, we can deduce:

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

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.

2.1.3 Concavity

Theorem: The entropy (\rightarrow Definition I/2.1.1) is concave in the probability mass function (\rightarrow Definition I/1.4.1) p , i.e.

$$H[\lambda p_1 + (1 - \lambda)p_2] \geq \lambda H[p_1] + (1 - \lambda)H[p_2] \quad (1)$$

where p_1 and p_2 are probability mass functions and $0 \leq \lambda \leq 1$.

Proof: Let X be a discrete random variable (\rightarrow Definition I/1.1.1) with possible outcomes \mathcal{X} and let $u(x)$ be the probability mass function (\rightarrow Definition I/1.4.1) of a discrete uniform distribution (\rightarrow Definition II/1.1.1) on $X \in \mathcal{X}$. Then, the entropy (\rightarrow Definition I/2.1.1) of an arbitrary probability mass function (\rightarrow Definition I/1.4.1) $p(x)$ can be rewritten as

$$\begin{aligned} H[p] &= - \sum_{x \in \mathcal{X}} p(x) \cdot \log p(x) \\ &= - \sum_{x \in \mathcal{X}} p(x) \cdot \log \frac{p(x)}{u(x)} u(x) \\ &= - \sum_{x \in \mathcal{X}} p(x) \cdot \log \frac{p(x)}{u(x)} - \sum_{x \in \mathcal{X}} p(x) \cdot \log u(x) \\ &= -\text{KL}[p||u] - \log \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} p(x) \\ &= \log |\mathcal{X}| - \text{KL}[p||u] \\ \log |\mathcal{X}| - H[p] &= \text{KL}[p||u] \end{aligned} \quad (2)$$

where we have applied the definition of the Kullback-Leibler divergence (\rightarrow Definition I/2.5.1), the probability mass function of the discrete uniform distribution (\rightarrow Proof II/1.1.2) and the total sum over the probability mass function (\rightarrow Definition I/1.4.1).

Note that the KL divergence is convex (\rightarrow Proof I/2.5.5) in the pair of probability distributions (\rightarrow Definition I/1.3.1) (p, q) :

$$\text{KL}[\lambda p_1 + (1 - \lambda)p_2 || \lambda q_1 + (1 - \lambda)q_2] \leq \lambda \text{KL}[p_1 || q_1] + (1 - \lambda) \text{KL}[p_2 || q_2] \quad (3)$$

A special case of this is given by

$$\begin{aligned} \text{KL}[\lambda p_1 + (1 - \lambda)p_2 || \lambda u + (1 - \lambda)u] &\leq \lambda \text{KL}[p_1 || u] + (1 - \lambda) \text{KL}[p_2 || u] \\ \text{KL}[\lambda p_1 + (1 - \lambda)p_2 || u] &\leq \lambda \text{KL}[p_1 || u] + (1 - \lambda) \text{KL}[p_2 || u] \end{aligned} \quad (4)$$

and applying equation (2), we have

$$\begin{aligned}
\log |\mathcal{X}| - H[\lambda p_1 + (1 - \lambda)p_2] &\leq \lambda (\log |\mathcal{X}| - H[p_1]) + (1 - \lambda) (\log |\mathcal{X}| - H[p_2]) \\
\log |\mathcal{X}| - H[\lambda p_1 + (1 - \lambda)p_2] &\leq \log |\mathcal{X}| - \lambda H[p_1] - (1 - \lambda)H[p_2] \\
-H[\lambda p_1 + (1 - \lambda)p_2] &\leq -\lambda H[p_1] - (1 - \lambda)H[p_2] \\
H[\lambda p_1 + (1 - \lambda)p_2] &\geq \lambda H[p_1] + (1 - \lambda)H[p_2]
\end{aligned} \tag{5}$$

which is equivalent to (1).

Sources:

- Wikipedia (2020): “Entropy (information theory)”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-08-11; URL: [https://en.wikipedia.org/wiki/Entropy_\(information_theory\)#Further_properties](https://en.wikipedia.org/wiki/Entropy_(information_theory)#Further_properties).
- Cover TM, Thomas JA (1991): “Elements of Information Theory”, p. 30; URL: <https://www.wiley.com/en-us/Elements+of+Information+Theory%2C+2nd+Edition-p-9780471241959>.
- Xie, Yao (2012): “Chain Rules and Inequalities”; in: *ECE587: Information Theory*, Lecture 3, Slide 25; URL: <https://www2.isye.gatech.edu/~yxie77/ece587/Lecture3.pdf>.
- Goh, Siong Thye (2016): “Understanding the proof of the concavity of entropy”; in: *StackExchange Mathematics*, retrieved on 2020-11-08; URL: <https://math.stackexchange.com/questions/2000194/understanding-the-proof-of-the-concavity-of-entropy>.

Metadata: ID: P149 | shortcut: ent-conc | author: JoramSoch | date: 2020-08-11, 08:29.

2.1.4 Conditional entropy

Definition: Let X and Y be discrete random variables (\rightarrow Definition I/1.1.1) with possible outcomes \mathcal{X} and \mathcal{Y} and probability mass functions (\rightarrow Definition I/1.4.1) $p(x)$ and $p(y)$. Then, the conditional entropy of Y given X or, entropy of Y conditioned on X , is defined as

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

where $H(Y|X = x)$ is the (marginal) entropy (\rightarrow Definition I/2.1.1) 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.

2.1.5 Joint entropy

Definition: Let X and Y be discrete random variables (\rightarrow Definition I/1.1.1) with possible outcomes \mathcal{X} and \mathcal{Y} and joint probability (\rightarrow Definition I/1.2.2) mass function (\rightarrow Definition I/1.4.1) $p(x, y)$. Then, the joint entropy of X and Y is defined as

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

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

Sources:

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

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

2.1.6 Cross-entropy

Definition: Let X be a discrete random variable (\rightarrow Definition I/1.1.1) with possible outcomes \mathcal{X} and let P and Q be two probability distributions (\rightarrow Definition I/1.3.1) on X with the probability mass functions (\rightarrow Definition I/1.4.1) $p(x)$ and $q(x)$. Then, the cross-entropy of Q relative to P is defined as

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

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

Sources:

- Wikipedia (2020): “Cross entropy”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-07-28; URL: https://en.wikipedia.org/wiki/Cross_entropy#Definition.

Metadata: ID: D85 | shortcut: ent-cross | author: JoramSoch | date: 2020-07-28, 02:51.

2.1.7 Convexity of cross-entropy

Theorem: The cross-entropy (\rightarrow Definition I/2.1.6) is convex in the probability distribution (\rightarrow Definition I/1.3.1) q , i.e.

$$H[p, \lambda q_1 + (1 - \lambda)q_2] \leq \lambda H[p, q_1] + (1 - \lambda)H[p, q_2] \quad (1)$$

where p is a fixed and q_1 and q_2 are any two probability distributions and $0 \leq \lambda \leq 1$.

Proof: The relationship between Kullback-Leibler divergence, entropy and cross-entropy (\rightarrow Proof I/2.5.8) is:

$$\text{KL}[P||Q] = H(P, Q) - H(P) . \quad (2)$$

Note that the KL divergence is convex (\rightarrow Proof I/2.5.5) in the pair of probability distributions (\rightarrow Definition I/1.3.1) (p, q) :

$$\text{KL}[\lambda p_1 + (1 - \lambda)p_2 || \lambda q_1 + (1 - \lambda)q_2] \leq \lambda \text{KL}[p_1 || q_1] + (1 - \lambda)\text{KL}[p_2 || q_2] \quad (3)$$

A special case of this is given by

$$\begin{aligned} \text{KL}[\lambda p + (1 - \lambda)p || \lambda q_1 + (1 - \lambda)q_2] &\leq \lambda \text{KL}[p || q_1] + (1 - \lambda)\text{KL}[p || q_2] \\ \text{KL}[p || \lambda q_1 + (1 - \lambda)q_2] &\leq \lambda \text{KL}[p || q_1] + (1 - \lambda)\text{KL}[p || q_2] \end{aligned} \quad (4)$$

and applying equation (2), we have

$$\begin{aligned} H[p, \lambda q_1 + (1 - \lambda)q_2] - H[p] &\leq \lambda (H[p, q_1] - H[p]) + (1 - \lambda) (H[p, q_2] - H[p]) \\ H[p, \lambda q_1 + (1 - \lambda)q_2] - H[p] &\leq \lambda H[p, q_1] + (1 - \lambda)H[p, q_2] - H[p] \\ H[p, \lambda q_1 + (1 - \lambda)q_2] &\leq \lambda H[p, q_1] + (1 - \lambda)H[p, q_2] \end{aligned} \quad (5)$$

which is equivalent to (1).

Sources:

- Wikipedia (2020): “Cross entropy”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-08-11; URL: https://en.wikipedia.org/wiki/Cross_entropy#Definition.
- gunes (2019): “Convexity of cross entropy”; in: *StackExchange CrossValidated*, retrieved on 2020-11-08; URL: <https://stats.stackexchange.com/questions/394463/convexity-of-cross-entropy>.

Metadata: ID: P150 | shortcut: entcross-conv | author: JoramSoch | date: 2020-08-11, 09:16.

2.1.8 Gibbs’ inequality

Theorem: Let X be a discrete random variable (\rightarrow Definition I/1.1.1) and consider two probability distributions (\rightarrow Definition I/1.3.1) with probability mass functions (\rightarrow Definition I/1.4.1) $p(x)$ and $q(x)$. Then, Gibbs’ inequality states that the entropy (\rightarrow Definition I/2.1.1) of X according to P is smaller than or equal to the cross-entropy (\rightarrow Definition I/2.1.6) of P and Q :

$$-\sum_{x \in \mathcal{X}} p(x) \log_b p(x) \leq -\sum_{x \in \mathcal{X}} p(x) \log_b q(x) . \quad (1)$$

Proof: Without loss of generality, we will use the natural logarithm, because a change in the base of the logarithm only implies multiplication by a constant:

$$\log_b a = \frac{\ln a}{\ln b} . \quad (2)$$

Let I be the set of all x for which $p(x)$ is non-zero. Then, proving (1) requires to show that

$$\sum_{x \in I} p(x) \ln \frac{p(x)}{q(x)} \geq 0 . \quad (3)$$

Because $\ln x \leq x - 1$, i.e. $-\ln x \geq 1 - x$, for all $x > 0$, with equality only if $x = 1$, we can say about the left-hand side that

$$\begin{aligned} \sum_{x \in I} p(x) \ln \frac{p(x)}{q(x)} &\geq \sum_{x \in I} p(x) \left(1 - \frac{p(x)}{q(x)} \right) \\ &= \sum_{x \in I} p(x) - \sum_{x \in I} q(x) . \end{aligned} \quad (4)$$

Finally, since $p(x)$ and $q(x)$ are probability mass functions (\rightarrow Definition I/1.4.1), we have

$$\begin{aligned}
0 \leq p(x) \leq 1, \quad \sum_{x \in I} p(x) &= 1 \quad \text{and} \\
0 \leq q(x) \leq 1, \quad \sum_{x \in I} q(x) &\leq 1,
\end{aligned} \tag{5}$$

such that it follows from (4) that

$$\begin{aligned}
\sum_{x \in I} p(x) \ln \frac{p(x)}{q(x)} &\geq \sum_{x \in I} p(x) - \sum_{x \in I} q(x) \\
&= 1 - \sum_{x \in I} q(x) \geq 0.
\end{aligned} \tag{6}$$

Sources:

- Wikipedia (2020): “Gibbs’ inequality”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-09-09; URL: https://en.wikipedia.org/wiki/Gibbs%27_inequality#Proof.

Metadata: ID: P164 | shortcut: gibbs-ineq | author: JoramSoch | date: 2020-09-09, 02:18.

2.1.9 Log sum inequality

Theorem: Let a_1, \dots, a_n and b_1, \dots, b_n be non-negative real numbers and define $a = \sum_{i=1}^n a_i$ and $b = \sum_{i=1}^n b_i$. Then, the log sum inequality states that

$$\sum_{i=1}^n a_i \log_c \frac{a_i}{b_i} \geq a \log_c \frac{a}{b}. \tag{1}$$

Proof: Without loss of generality, we will use the natural logarithm, because a change in the base of the logarithm only implies multiplication by a constant:

$$\log_c a = \frac{\ln a}{\ln c}. \tag{2}$$

Let $f(x) = x \ln x$. Then, the left-hand side of (1) can be rewritten as

$$\begin{aligned}
\sum_{i=1}^n a_i \ln \frac{a_i}{b_i} &= \sum_{i=1}^n b_i f\left(\frac{a_i}{b_i}\right) \\
&= b \sum_{i=1}^n \frac{b_i}{b} f\left(\frac{a_i}{b_i}\right).
\end{aligned} \tag{3}$$

Because $f(x)$ is a convex function and

$$\begin{aligned}
\frac{b_i}{b} &\geq 0 \\
\sum_{i=1}^n \frac{b_i}{b} &= 1,
\end{aligned} \tag{4}$$

applying Jensen's inequality yields

$$\begin{aligned}
 b \sum_{i=1}^n \frac{b_i}{b} f\left(\frac{a_i}{b_i}\right) &\geq b f\left(\sum_{i=1}^n \frac{b_i}{b} \frac{a_i}{b_i}\right) \\
 &= b f\left(\frac{1}{b} \sum_{i=1}^n a_i\right) \\
 &= b f\left(\frac{a}{b}\right) \\
 &= a \ln \frac{a}{b}.
 \end{aligned} \tag{5}$$

Finally, combining (3) and (5), this demonstrates (1).

Sources:

- Wikipedia (2020): “Log sum inequality”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-09-09; URL: https://en.wikipedia.org/wiki/Log_sum_inequality#Proof.
- Wikipedia (2020): “Jensen's inequality”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-09-09; URL: https://en.wikipedia.org/wiki/Jensen%27s_inequality#Statements.

Metadata: ID: P165 | shortcut: logsum-ineq | author: JoramSoch | date: 2020-09-09, 02:46.

2.2 Differential entropy

2.2.1 Definition

Definition: Let X be a continuous random variable (\rightarrow Definition I/1.1.1) with possible outcomes \mathcal{X} and the (estimated or assumed) probability density function (\rightarrow Definition I/1.4.2) $p(x) = f_X(x)$. Then, the differential entropy (also referred to as “continuous entropy”) of X is defined as

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

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

Sources:

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

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

2.2.2 Negativity

Theorem: Unlike its discrete analogue (\rightarrow Proof I/2.1.2), the differential entropy (\rightarrow Definition I/2.2.1) can become negative.

Proof: Let X be a random variable (\rightarrow Definition I/1.1.1) 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/2.2.1) 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.

2.2.3 Conditional differential entropy

Definition: Let X and Y be continuous random variables (\rightarrow Definition I/1.1.1) with possible outcomes \mathcal{X} and \mathcal{Y} and probability density functions (\rightarrow Definition I/1.4.2) $p(x)$ 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/2.2.1) of Y , evaluated at x .

Sources:

- original work

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

2.2.4 Joint differential entropy

Definition: Let X and Y be continuous random variables (\rightarrow Definition I/1.1.1) with possible outcomes \mathcal{X} and \mathcal{Y} and joint probability (\rightarrow Definition I/1.2.2) density function (\rightarrow Definition I/1.4.2) $p(x, y)$. 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.

2.2.5 Differential cross-entropy

Definition: Let X be a continuous random variable (\rightarrow Definition I/1.1.1) with possible outcomes \mathcal{X} and let P and Q be two probability distributions (\rightarrow Definition I/1.3.1) on X with the probability density functions (\rightarrow Definition I/1.4.2) $p(x)$ and $q(x)$. Then, the differential cross-entropy of Q relative to P is defined as

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

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

Sources:

- Wikipedia (2020): “Cross entropy”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-07-28; URL: https://en.wikipedia.org/wiki/Cross_entropy#Definition.

Metadata: ID: D86 | shortcut: dent-cross | author: JoramSoch | date: 2020-07-28, 03:03.

2.3 Discrete mutual information

2.3.1 Definition

Definition:

1) The mutual information of two discrete random variables (\rightarrow Definition I/1.1.1) 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/1.4.1) of X and Y and $p(x, y)$ is the joint probability (\rightarrow Definition I/1.2.2) mass function of X and Y .

2) The mutual information of two continuous random variables (\rightarrow Definition I/1.1.1) 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/1.4.1) of X and Y and $p(x, y)$ is the joint probability (\rightarrow Definition I/1.2.2) 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.

2.3.2 Relation to marginal and conditional entropy

Theorem: Let X and Y be discrete random variables (\rightarrow Definition I/1.1.1) with the joint probability (\rightarrow Definition I/1.2.2) $p(x, y)$ for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Then, the mutual information (\rightarrow Definition I/2.4.1) 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/2.1.1) of X and Y and $H(X|Y)$ and $H(Y|X)$ are the conditional entropies (\rightarrow Definition I/2.1.4).

Proof: The mutual information (\rightarrow Definition I/2.4.1) 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 I/1.2.4), 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 I/1.2.3), 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/2.1.1) and conditional (\rightarrow Definition I/2.1.4) entropy

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

we can finally show:

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

The conditioning of X on Y in this proof is without loss of generality. Thus, the proof for the expression using the reverse conditional 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.

2.3.3 Relation to marginal and joint entropy

Theorem: Let X and Y be discrete random variables (\rightarrow Definition I/1.1.1) with the joint probability (\rightarrow Definition I/1.2.2) $p(x, y)$ for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Then, the mutual information (\rightarrow Definition I/2.4.1) 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 entropies (\rightarrow Definition I/2.1.1) of X and Y and $H(X, Y)$ is the joint entropy (\rightarrow Definition I/2.1.5).

Proof: The mutual information (\rightarrow Definition I/2.4.1) 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:

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

Regrouping the variables, this reads:

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

Applying the law of marginal probability (\rightarrow Definition I/1.2.3), 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/2.1.1) and joint (\rightarrow Definition I/2.1.5) entropy

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

we can finally show:

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

Sources:

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

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

2.3.4 Relation to joint and conditional entropy

Theorem: Let X and Y be discrete random variables (\rightarrow Definition I/1.1.1) with the joint probability (\rightarrow Definition I/1.2.2) $p(x, y)$ for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Then, the mutual information (\rightarrow Definition I/2.4.1) of X and Y can be expressed as

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

where $H(X, Y)$ is the joint entropy (\rightarrow Definition I/2.1.5) of X and Y and $H(X|Y)$ and $H(Y|X)$ are the conditional entropies (\rightarrow Definition I/2.1.4).

Proof: The existence of the joint probability mass function (\rightarrow Definition I/1.4.1) ensures that the mutual information (\rightarrow Definition I/2.4.1) is defined:

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

The relation of mutual information to conditional entropy (\rightarrow Proof I/2.3.2) is:

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

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

The relation of mutual information to joint entropy (\rightarrow Proof I/2.3.3) is:

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

It is true that

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

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.

2.4 Continuous mutual information

2.4.1 Definition

Definition:

1) The mutual information of two discrete random variables (\rightarrow Definition I/1.1.1) 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/1.4.1) of X and Y and $p(x, y)$ is the joint probability (\rightarrow Definition I/1.2.2) mass function of X and Y .

2) The mutual information of two continuous random variables (\rightarrow Definition I/1.1.1) 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/1.4.1) of X and Y and $p(x, y)$ is the joint probability (\rightarrow Definition I/1.2.2) 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.

2.4.2 Relation to marginal and conditional differential entropy

Theorem: Let X and Y be continuous random variables (\rightarrow Definition I/1.1.1) with the joint probability (\rightarrow Definition I/1.2.2) $p(x, y)$ for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Then, the mutual information (\rightarrow Definition I/2.4.1) 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/2.2.1) of X and Y and $h(X|Y)$ and $h(Y|X)$ are the conditional differential entropies (\rightarrow Definition I/2.2.3).

Proof: The mutual information (\rightarrow Definition I/2.4.1) 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 I/1.2.4), 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 I/1.2.3), 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/2.2.1) and conditional (\rightarrow Definition I/2.2.3) 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.

2.4.3 Relation to marginal and joint differential entropy

Theorem: Let X and Y be continuous random variables (\rightarrow Definition I/1.1.1) with the joint probability (\rightarrow Definition I/1.2.2) $p(x, y)$ for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Then, the mutual information (\rightarrow Definition I/2.4.1) 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/2.2.1) of X and Y and $h(X, Y)$ is the joint differential entropy (\rightarrow Definition I/2.2.4).

Proof: The mutual information (\rightarrow Definition I/2.4.1) 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 I/1.2.3), 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/2.2.1) and joint (\rightarrow Definition I/2.2.4) 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.

2.4.4 Relation to joint and conditional differential entropy

Theorem: Let X and Y be continuous random variables (\rightarrow Definition I/1.1.1) with the joint probability (\rightarrow Definition I/1.2.2) $p(x, y)$ for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Then, the mutual information (\rightarrow Definition I/2.4.1) 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/2.2.4) of X and Y and $h(X|Y)$ and $h(Y|X)$ are the conditional differential entropies (\rightarrow Definition I/2.2.3).

Proof: The existence of the joint probability density function (\rightarrow Definition I/1.4.2) ensures that the mutual information (\rightarrow Definition I/2.4.1) 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/2.4.2) 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/2.4.3) 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: 2020-02-21, 17:23.

2.5 Kullback-Leibler divergence

2.5.1 Definition

Definition: Let X be a random variable (\rightarrow Definition I/1.1.1) with possible outcomes \mathcal{X} and let P and Q be two probability distributions (\rightarrow Definition I/1.3.1) on X .

1) The Kullback-Leibler divergence of P from Q for a discrete random variable X is defined as

$$\text{KL}[P||Q] = \sum_{x \in \mathcal{X}} p(x) \cdot \log \frac{p(x)}{q(x)} \quad (1)$$

where $p(x)$ and $q(x)$ are the probability mass functions (\rightarrow Definition I/1.4.1) of P and Q .

2) The Kullback-Leibler divergence of P from Q for a continuous random variable X is defined as

$$\text{KL}[P||Q] = \int_{\mathcal{X}} p(x) \cdot \log \frac{p(x)}{q(x)} dx \quad (2)$$

where $p(x)$ and $q(x)$ are the probability density functions (\rightarrow Definition I/1.4.2) of P and Q .

Sources:

- MacKay, David J.C. (2003): “Probability, Entropy, and Inference”; in: *Information Theory, Inference, and Learning Algorithms*, ch. 2.6, eq. 2.45, p. 34; URL: <https://www.inference.org.uk/itprnn/book.pdf>.

Metadata: ID: D52 | shortcut: kl | author: JoramSoch | date: 2020-05-10, 20:20.

2.5.2 Non-negativity

Theorem: The Kullback-Leibler divergence (\rightarrow Definition I/2.5.1) is always non-negative

$$\text{KL}[P||Q] \geq 0 \quad (1)$$

with $\text{KL}[P||Q] = 0$, if and only if $P = Q$.

Proof: The discrete Kullback-Leibler divergence (\rightarrow Definition I/2.5.1) is defined as

$$\text{KL}[P||Q] = \sum_{x \in \mathcal{X}} p(x) \cdot \log \frac{p(x)}{q(x)} \quad (2)$$

which can be reformulated into

$$\text{KL}[P||Q] = \sum_{x \in \mathcal{X}} p(x) \cdot \log p(x) - \sum_{x \in \mathcal{X}} p(x) \cdot \log q(x) . \quad (3)$$

Gibbs’ inequality (\rightarrow Proof I/2.1.8) states that the entropy (\rightarrow Definition I/2.1.1) of a probability distribution is always less than or equal to the cross-entropy (\rightarrow Definition I/2.1.6) with another probability distribution – with equality only if the distributions are identical –,

$$-\sum_{i=1}^n p(x_i) \log p(x_i) \leq -\sum_{i=1}^n p(x_i) \log q(x_i) \quad (4)$$

which can be reformulated into

$$\sum_{i=1}^n p(x_i) \log p(x_i) - \sum_{i=1}^n p(x_i) \log q(x_i) \geq 0 . \quad (5)$$

Applying (5) to (3), this proves equation (1).

Sources:

- Wikipedia (2020): “Kullback-Leibler divergence”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-31; URL: https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence#Properties.

Metadata: ID: P117 | shortcut: kl-nonneg | author: JoramSoch | date: 2020-05-31, 23:43.

2.5.3 Non-negativity

Theorem: The Kullback-Leibler divergence (\rightarrow Definition I/2.5.1) is always non-negative

$$\text{KL}[P||Q] \geq 0 \quad (1)$$

with $\text{KL}[P||Q] = 0$, if and only if $P = Q$.

Proof: The discrete Kullback-Leibler divergence (\rightarrow Definition I/2.5.1) is defined as

$$\text{KL}[P||Q] = \sum_{x \in \mathcal{X}} p(x) \cdot \log \frac{p(x)}{q(x)} . \quad (2)$$

The log sum inequality (\rightarrow Proof I/2.1.9) states that

$$\sum_{i=1}^n a_i \log_c \frac{a_i}{b_i} \geq a \log_c \frac{a}{b} . \quad (3)$$

where a_1, \dots, a_n and b_1, \dots, b_n be non-negative real numbers and $a = \sum_{i=1}^n a_i$ and $b = \sum_{i=1}^n b_i$. Because $p(x)$ and $q(x)$ are probability mass functions (\rightarrow Definition I/1.4.1), such that

$$\begin{aligned} p(x) &\geq 0, & \sum_{x \in \mathcal{X}} p(x) &= 1 & \text{ and} \\ q(x) &\geq 0, & \sum_{x \in \mathcal{X}} q(x) &= 1 , \end{aligned} \quad (4)$$

theorem (1) is simply a special case of (3), i.e.

$$\text{KL}[P||Q] \stackrel{(2)}{=} \sum_{x \in \mathcal{X}} p(x) \cdot \log \frac{p(x)}{q(x)} \stackrel{(3)}{\geq} 1 \log \frac{1}{1} = 0 . \quad (5)$$

Sources:

- Wikipedia (2020): “Log sum inequality”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-09-09; URL: https://en.wikipedia.org/wiki/Log_sum_inequality#Applications.

Metadata: ID: P166 | shortcut: kl-nonneg2 | author: JoramSoch | date: 2020-09-09, 07:02.

2.5.4 Non-symmetry

Theorem: The Kullback-Leibler divergence (\rightarrow Definition I/2.5.1) is non-symmetric, i.e.

$$\text{KL}[P||Q] \neq \text{KL}[Q||P] \quad (1)$$

for some [probability distributions](dist) P and Q .

Proof: Let $X \in \mathcal{X} = \{0, 1, 2\}$ be a discrete random variable (\rightarrow Definition I/1.1.1) and consider the two [probability distributions](dist)

$$\begin{aligned} P : X &\sim \text{Bin}(2, 0.5) \\ Q : X &\sim \mathcal{U}(0, 2) \end{aligned} \quad (2)$$

where $\text{Bin}(n, p)$ indicates a binomial distribution (\rightarrow Definition II/1.3.1) and $\mathcal{U}(a, b)$ indicates a discrete uniform distribution (\rightarrow Definition II/1.1.1).

Then, the probability mass function of the binomial distribution (\rightarrow Proof II/1.3.2) entails that

$$p(x) = \begin{cases} 1/4, & \text{if } x = 0 \\ 1/2, & \text{if } x = 1 \\ 1/4, & \text{if } x = 2 \end{cases} \quad (3)$$

and the probability mass function of the discrete uniform distribution (\rightarrow Proof II/1.1.2) entails that

$$q(x) = \frac{1}{3}, \quad (4)$$

such that the Kullback-Leibler divergence (\rightarrow Definition I/2.5.1) of P from Q is

$$\begin{aligned} \text{KL}[P||Q] &= \sum_{x \in \mathcal{X}} p(x) \cdot \log \frac{p(x)}{q(x)} \\ &= \frac{1}{4} \log \frac{3}{4} + \frac{1}{2} \log \frac{3}{2} + \frac{1}{4} \log \frac{3}{4} \\ &= \frac{1}{2} \log \frac{3}{4} + \frac{1}{2} \log \frac{3}{2} \\ &= \frac{1}{2} \left(\log \frac{3}{4} + \log \frac{3}{2} \right) \\ &= \frac{1}{2} \log \left(\frac{3}{4} \cdot \frac{3}{2} \right) \\ &= \frac{1}{2} \log \frac{9}{8} = 0.0589 \end{aligned} \quad (5)$$

and the Kullback-Leibler divergence (\rightarrow Definition I/2.5.1) of Q from P is

$$\begin{aligned}
\text{KL}[Q||P] &= \sum_{x \in \mathcal{X}} q(x) \cdot \log \frac{q(x)}{p(x)} \\
&= \frac{1}{3} \log \frac{4}{3} + \frac{1}{3} \log \frac{2}{3} + \frac{1}{3} \log \frac{4}{3} \\
&= \frac{1}{3} \left(\log \frac{4}{3} + \log \frac{2}{3} + \log \frac{4}{3} \right) \\
&= \frac{1}{3} \log \left(\frac{4}{3} \cdot \frac{2}{3} \cdot \frac{4}{3} \right) \\
&= \frac{1}{3} \log \frac{32}{27} = 0.0566
\end{aligned} \tag{6}$$

which provides an example for

$$\text{KL}[P||Q] \neq \text{KL}[Q||P] \tag{7}$$

and thus proves the theorem.

Sources:

- Kullback, Solomon (1959): “Divergence”; in: *Information Theory and Statistics*, ch. 1.3, pp. 6ff.; URL: <http://index-of.co.uk/Information-Theory/Information%20theory%20and%20statistics%20-%20Solomon%20Kullback.pdf>.
- Wikipedia (2020): “Kullback-Leibler divergence”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-08-11; URL: https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence#Basic_example.

Metadata: ID: P147 | shortcut: kl-nonsymm | author: JoramSoch | date: 2020-08-11, 06:57.

2.5.5 Convexity

Theorem: The Kullback-Leibler divergence (\rightarrow Definition I/2.5.1) is convex in the pair of probability distributions (\rightarrow Definition I/1.3.1) (p, q) , i.e.

$$\text{KL}[\lambda p_1 + (1 - \lambda)p_2 || \lambda q_1 + (1 - \lambda)q_2] \leq \lambda \text{KL}[p_1 || q_1] + (1 - \lambda) \text{KL}[p_2 || q_2] \tag{1}$$

where (p_1, q_1) and (p_2, q_2) are two pairs of probability distributions and $0 \leq \lambda \leq 1$.

Proof: The Kullback-Leibler divergence (\rightarrow Definition I/2.5.1) of P from Q is defined as

$$\text{KL}[P||Q] = \sum_{x \in \mathcal{X}} p(x) \cdot \log \frac{p(x)}{q(x)} \tag{2}$$

and the log sum inequality (\rightarrow Proof I/2.1.9) states that

$$\sum_{i=1}^n a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^n a_i \right) \log \frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i} \tag{3}$$

where a_1, \dots, a_n and b_1, \dots, b_n are non-negative real numbers.

Thus, we can rewrite the KL divergence of the mixture distribution as

$$\begin{aligned}
& \text{KL}[\lambda p_1 + (1 - \lambda)p_2 || \lambda q_1 + (1 - \lambda)q_2] \\
& \stackrel{(2)}{=} \sum_{x \in \mathcal{X}} \left[[\lambda p_1(x) + (1 - \lambda)p_2(x)] \cdot \log \frac{\lambda p_1(x) + (1 - \lambda)p_2(x)}{\lambda q_1(x) + (1 - \lambda)q_2(x)} \right] \\
& \stackrel{(3)}{\leq} \sum_{x \in \mathcal{X}} \left[\lambda p_1(x) \cdot \log \frac{\lambda p_1(x)}{\lambda q_1(x)} + (1 - \lambda)p_2(x) \cdot \log \frac{(1 - \lambda)p_2(x)}{(1 - \lambda)q_2(x)} \right] \\
& = \lambda \sum_{x \in \mathcal{X}} p_1(x) \cdot \log \frac{p_1(x)}{q_1(x)} + (1 - \lambda) \sum_{x \in \mathcal{X}} p_2(x) \cdot \log \frac{p_2(x)}{q_2(x)} \\
& \stackrel{(2)}{=} \lambda \text{KL}[p_1 || q_1] + (1 - \lambda) \text{KL}[p_2 || q_2]
\end{aligned} \tag{4}$$

which is equivalent to (1).

Sources:

- Wikipedia (2020): “Kullback-Leibler divergence”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-08-11; URL: https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence#Properties.
- Xie, Yao (2012): “Chain Rules and Inequalities”; in: *ECE587: Information Theory*, Lecture 3, Slides 22/24; URL: <https://www2.isye.gatech.edu/~yxie77/ece587/Lecture3.pdf>.

Metadata: ID: P148 | shortcut: kl-conv | author: JoramSoch | date: 2020-08-11, 07:30.

2.5.6 Additivity for independent distributions

Theorem: The Kullback-Leibler divergence (\rightarrow Definition I/2.5.1) is additive for independent distributions, i.e.

$$\text{KL}[P || Q] = \text{KL}[P_1 || Q_1] + \text{KL}[P_2 || Q_2] \tag{1}$$

where P_1 and P_2 are independent (\rightarrow Definition I/1.2.5) distributions (\rightarrow Definition I/1.3.1) with the joint distribution (\rightarrow Definition I/1.3.2) P , such that $p(x, y) = p_1(x) p_2(y)$, and equivalently for Q_1 , Q_2 and Q .

Proof: The continuous Kullback-Leibler divergence (\rightarrow Definition I/2.5.1) is defined as

$$\text{KL}[P || Q] = \int_{\mathcal{X}} p(x) \cdot \log \frac{p(x)}{q(x)} dx \tag{2}$$

which, applied to the joint distributions P and Q , yields

$$\text{KL}[P || Q] = \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x, y) \cdot \log \frac{p(x, y)}{q(x, y)} dy dx . \tag{3}$$

Applying $p(x, y) = p_1(x) p_2(y)$ and $q(x, y) = q_1(x) q_2(y)$, we have

$$\text{KL}[P || Q] = \int_{\mathcal{X}} \int_{\mathcal{Y}} p_1(x) p_2(y) \cdot \log \frac{p_1(x) p_2(y)}{q_1(x) q_2(y)} dy dx . \tag{4}$$

Now we can separate the logarithm and evaluate the integrals:

$$\begin{aligned}
\text{KL}[P||Q] &= \int_{\mathcal{X}} \int_{\mathcal{Y}} p_1(x) p_2(y) \cdot \left(\log \frac{p_1(x)}{q_1(x)} + \log \frac{p_2(y)}{q_2(y)} \right) dy dx \\
&= \int_{\mathcal{X}} \int_{\mathcal{Y}} p_1(x) p_2(y) \cdot \log \frac{p_1(x)}{q_1(x)} dy dx + \int_{\mathcal{X}} \int_{\mathcal{Y}} p_1(x) p_2(y) \cdot \log \frac{p_2(y)}{q_2(y)} dy dx \\
&= \int_{\mathcal{X}} p_1(x) \cdot \log \frac{p_1(x)}{q_1(x)} \int_{\mathcal{Y}} p_2(y) dy dx + \int_{\mathcal{Y}} p_2(y) \cdot \log \frac{p_2(y)}{q_2(y)} \int_{\mathcal{X}} p_1(x) dx dy \\
&= \int_{\mathcal{X}} p_1(x) \cdot \log \frac{p_1(x)}{q_1(x)} dx + \int_{\mathcal{Y}} p_2(y) \cdot \log \frac{p_2(y)}{q_2(y)} dy \\
&\stackrel{(2)}{=} \text{KL}[P_1||Q_1] + \text{KL}[P_2||Q_2] .
\end{aligned} \tag{5}$$

Sources:

- Wikipedia (2020): “Kullback-Leibler divergence”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-31; URL: https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence#Properties.

Metadata: ID: P116 | shortcut: kl-add | author: JoramSoch | date: 2020-05-31, 23:26.

2.5.7 Invariance under parameter transformation

Theorem: The Kullback-Leibler divergence (\rightarrow Definition I/2.5.1) is invariant under parameter transformation, i.e.

$$\text{KL}[p(x)||q(x)] = \text{KL}[p(y)||q(y)] \tag{1}$$

where $y(x) = mx + n$ is an affine transformation of x and $p(x)$ and $q(x)$ are the probability density functions (\rightarrow Definition I/1.4.2) of the probability distributions (\rightarrow Definition I/1.3.1) P and Q on the continuous random variable (\rightarrow Definition I/1.1.1) X .

Proof: The continuous Kullback-Leibler divergence (\rightarrow Definition I/2.5.1) (KL divergence) is defined as

$$\text{KL}[p(x)||q(x)] = \int_a^b p(x) \cdot \log \frac{p(x)}{q(x)} dx \tag{2}$$

where $a = \min(\mathcal{X})$ and $b = \max(\mathcal{X})$ are the lower and upper bound of the possible outcomes \mathcal{X} of X .

Due to the identity of the differentials

$$\begin{aligned}
p(x) dx &= p(y) dy \\
q(x) dx &= q(y) dy
\end{aligned} \tag{3}$$

which can be rearranged into

$$\begin{aligned}
p(x) &= p(y) \frac{dy}{dx} \\
q(x) &= q(y) \frac{dy}{dx} ,
\end{aligned} \tag{4}$$

the KL divergence can be evaluated as follows:

$$\begin{aligned}
 \text{KL}[p(x)||q(x)] &= \int_a^b p(x) \cdot \log \frac{p(x)}{q(x)} dx \\
 &= \int_{y(a)}^{y(b)} p(y) \frac{dy}{dx} \cdot \log \left(\frac{p(y) \frac{dy}{dx}}{q(y) \frac{dy}{dx}} \right) dx \\
 &= \int_{y(a)}^{y(b)} p(y) \cdot \log \frac{p(y)}{q(y)} dy \\
 &= \text{KL}[p(y)||q(y)] .
 \end{aligned} \tag{5}$$

Sources:

- Wikipedia (2020): “Kullback-Leibler divergence”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-27; URL: https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence#Properties.
- shimao (2018): “KL divergence invariant to affine transformation?”; in: *StackExchange CrossValidated*, retrieved on 2020-05-28; URL: <https://stats.stackexchange.com/questions/341922/kl-divergence-invariant-to-affine-transformation>.

Metadata: ID: P115 | shortcut: kl-inv | author: JoramSoch | date: 2020-05-28, 00:18.

2.5.8 Relation to discrete entropy

Theorem: Let X be a discrete random variable (\rightarrow Definition I/1.1.1) with possible outcomes \mathcal{X} and let P and Q be two probability distributions (\rightarrow Definition I/1.3.1) on X . Then, the Kullback-Leibler divergence (\rightarrow Definition I/2.5.1) of P from Q can be expressed as

$$\text{KL}[P||Q] = H(P, Q) - H(P) \tag{1}$$

where $H(P, Q)$ is the cross-entropy (\rightarrow Definition I/2.1.6) of P and Q and $H(P)$ is the marginal entropy (\rightarrow Definition I/2.1.1) of P .

Proof: The discrete Kullback-Leibler divergence (\rightarrow Definition I/2.5.1) is defined as

$$\text{KL}[P||Q] = \sum_{x \in \mathcal{X}} p(x) \cdot \log \frac{p(x)}{q(x)} \tag{2}$$

where $p(x)$ and $q(x)$ are the probability mass functions (\rightarrow Definition I/1.4.1) of P and Q . Separating the logarithm, we have:

$$\text{KL}[P||Q] = - \sum_{x \in \mathcal{X}} p(x) \log q(x) + \sum_{x \in \mathcal{X}} p(x) \log p(x) . \tag{3}$$

Now considering the definitions of marginal entropy (\rightarrow Definition I/2.1.1) and cross-entropy (\rightarrow Definition I/2.1.6)

$$\begin{aligned}
 H(P) &= - \sum_{x \in \mathcal{X}} p(x) \log p(x) \\
 H(P, Q) &= - \sum_{x \in \mathcal{X}} p(x) \log q(x) ,
 \end{aligned} \tag{4}$$

we can finally show:

$$\text{KL}[P||Q] = H(P, Q) - H(P) . \quad (5)$$

Sources:

- Wikipedia (2020): “Kullback-Leibler divergence”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-27; URL: https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence#Motivation.

Metadata: ID: P113 | shortcut: kl-ent | author: JoramSoch | date: 2020-05-27, 23:20.

2.5.9 Relation to differential entropy

Theorem: Let X be a continuous random variable (\rightarrow Definition I/1.1.1) with possible outcomes \mathcal{X} and let P and Q be two probability distributions (\rightarrow Definition I/1.3.1) on X . Then, the Kullback-Leibler divergence (\rightarrow Definition I/2.5.1) of P from Q can be expressed as

$$\text{KL}[P||Q] = h(P, Q) - h(P) \quad (1)$$

where $h(P, Q)$ is the differential cross-entropy (\rightarrow Definition I/2.2.5) of P and Q and $h(P)$ is the marginal differential entropy (\rightarrow Definition I/2.2.1) of P .

Proof: The continuous Kullback-Leibler divergence (\rightarrow Definition I/2.5.1) is defined as

$$\text{KL}[P||Q] = \int_{\mathcal{X}} p(x) \cdot \log \frac{p(x)}{q(x)} dx \quad (2)$$

where $p(x)$ and $q(x)$ are the probability density functions (\rightarrow Definition I/1.4.2) of P and Q . Separating the logarithm, we have:

$$\text{KL}[P||Q] = - \int_{\mathcal{X}} p(x) \log q(x) dx + \int_{\mathcal{X}} p(x) \log p(x) dx . \quad (3)$$

Now considering the definitions of marginal differential entropy (\rightarrow Definition I/2.2.1) and differential cross-entropy (\rightarrow Definition I/2.2.5)

$$\begin{aligned} h(P) &= - \int_{\mathcal{X}} p(x) \log p(x) dx \\ h(P, Q) &= - \int_{\mathcal{X}} p(x) \log q(x) dx , \end{aligned} \quad (4)$$

we can finally show:

$$\text{KL}[P||Q] = h(P, Q) - h(P) . \quad (5)$$

Sources:

- Wikipedia (2020): “Kullback-Leibler divergence”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-27; URL: https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence#Motivation.

Metadata: ID: P114 | shortcut: kl-dent | author: JoramSoch | date: 2020-05-27, 23:32.

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/1.6.1) 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/1.5.1) 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/5.1.1) 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 I/4.1.3):

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

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/1.4.4) $\chi_{k,p}^2$ 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 Frequentist statistics

4.1 Likelihood theory

4.1.1 Likelihood function

Definition: Let there be a generative model (\rightarrow Definition I/5.1.1) m describing measured data y using model parameters θ . Then, the probability density function (\rightarrow Definition I/1.4.2) 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.

4.1.2 Log-likelihood function

Definition: Let there be a generative model (\rightarrow Definition I/5.1.1) m describing measured data y using model parameters θ . Then, the logarithm of the probability density function (\rightarrow Definition I/1.4.2) of the distribution of y given θ is called the log-likelihood function (\rightarrow Definition I/5.1.2) of m :

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

Sources:

- original work

Metadata: ID: D59 | shortcut: llf | author: JoramSoch | date: 2020-05-17, 22:52.

4.1.3 Maximum likelihood estimation

Definition: Let there be a generative model (\rightarrow Definition I/5.1.1) m describing measured data y using model parameters θ . Then, the parameter values maximizing the likelihood function (\rightarrow Definition I/5.1.2) or log-likelihood function (\rightarrow Definition I/4.1.2) are called maximum likelihood estimates of θ :

$$\hat{\theta} = \arg \max_{\theta} \mathcal{L}_m(\theta) = \arg \max_{\theta} \text{LL}_m(\theta) . \quad (1)$$

The process of calculating $\hat{\theta}$ is called “maximum likelihood estimation” and the functional form leading from y to $\hat{\theta}$ given m is called “maximum likelihood estimator”. Maximum likelihood estimation, estimator and estimates may all be abbreviated as “MLE”.

Sources:

- original work

Metadata: ID: D60 | shortcut: mle | author: JoramSoch | date: 2020-05-15, 23:05.

4.1.4 Maximum log-likelihood

Definition: Let there be a generative model (\rightarrow Definition I/5.1.1) m describing measured data y using model parameters θ . Then, the maximum log-likelihood (MLL) of m is the maximal value of the log-likelihood function (\rightarrow Definition I/4.1.2) of this model:

$$\text{MLL}(m) = \max_{\theta} \text{LL}_m(\theta) . \quad (1)$$

The maximum log-likelihood can be obtained by plugging the maximum likelihood estimates (\rightarrow Definition I/4.1.3) into the log-likelihood function (\rightarrow Definition I/4.1.2).

Sources:

- original work

Metadata: ID: D61 | shortcut: mll | author: JoramSoch | date: 2020-05-15, 23:13.

5 Bayesian statistics

5.1 Probabilistic modeling

5.1.1 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.

5.1.2 Likelihood function

Definition: Let there be a generative model (\rightarrow Definition I/5.1.1) m describing measured data y using model parameters θ . Then, the probability density function (\rightarrow Definition I/1.4.2) 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.

5.1.3 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/1.4.2) 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.

5.1.4 Full probability model

Definition: Consider measured data y and some unknown latent parameters θ . The combination of a generative model (\rightarrow Definition I/5.1.1) for y and a prior distribution (\rightarrow Definition I/5.1.3) 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.

5.1.5 Joint likelihood

Definition: Let there be a generative model (\rightarrow Definition I/5.1.1) m describing measured data y using model parameters θ and a prior distribution (\rightarrow Definition I/5.1.3) on θ . Then, the joint probability (\rightarrow Definition I/1.2.2) density function (\rightarrow Definition I/1.4.2) 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.

5.1.6 Joint likelihood is product of likelihood and prior

Theorem: Let there be a generative model (\rightarrow Definition I/5.1.1) m describing measured data y using model parameters θ and a prior distribution (\rightarrow Definition I/5.1.3) on θ . Then, the joint likelihood (\rightarrow Definition I/5.1.5) is equal to the product of likelihood function (\rightarrow Definition I/5.1.2) and prior density (\rightarrow Definition I/5.1.3):

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

Proof: The joint likelihood (\rightarrow Definition I/5.1.5) is defined as the joint probability (\rightarrow Definition I/1.2.2) density function (\rightarrow Definition I/1.4.2) of data y and parameters θ :

$$p(y, \theta|m) . \quad (2)$$

Applying the law of conditional probability (\rightarrow Definition I/1.2.4), we have:

$$\begin{aligned} p(y|\theta, m) &= \frac{p(y, \theta|m)}{p(\theta|m)} \\ &\Leftrightarrow \\ p(y, \theta|m) &= p(y|\theta, m) p(\theta|m) . \end{aligned} \quad (3)$$

Sources:

- original work

Metadata: ID: P89 | shortcut: jl-lfnprior | author: JoramSoch | date: 2020-05-05, 04:21.

5.1.7 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/1.4.2) 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.

5.1.8 Posterior density is proportional to joint likelihood

Theorem: In a full probability model (\rightarrow Definition I/5.1.4) m describing measured data y using model parameters θ , the posterior density (\rightarrow Definition I/5.1.7) over the model parameters is proportional to the joint likelihood (\rightarrow Definition I/5.1.5):

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

Proof: In a full probability model (\rightarrow Definition I/5.1.4), the posterior distribution (\rightarrow Definition I/5.1.7) can be expressed using Bayes' theorem (\rightarrow Proof I/5.2.1):

$$p(\theta|y, m) = \frac{p(y|\theta, m) p(\theta|m)}{p(y|m)} . \quad (2)$$

Applying the law of conditional probability (\rightarrow Definition I/1.2.4) to the numerator, we have:

$$p(\theta|y, m) = \frac{p(y, \theta|m)}{p(y|m)} . \quad (3)$$

Because the denominator does not depend on θ , it is constant in θ and thus acts a proportionality factor between the posterior distribution and the joint likelihood:

$$p(\theta|y, m) \propto p(y, \theta|m) . \quad (4)$$

Sources:

- original work

Metadata: ID: P90 | shortcut: post-jl | author: JoramSoch | date: 2020-05-05, 04:46.

5.1.9 Marginal likelihood

Definition: Let there be a generative model (\rightarrow Definition I/5.1.1) m describing measured data y using model parameters θ and a prior distribution (\rightarrow Definition I/5.1.3) on θ . Then, the marginal probability (\rightarrow Definition I/1.2.3) density function (\rightarrow Definition I/1.4.2) 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.

5.1.10 Marginal likelihood is integral of joint likelihood

Theorem: In a full probability model (\rightarrow Definition I/5.1.4) m describing measured data y using model parameters θ , the marginal likelihood (\rightarrow Definition I/5.1.9) is the integral of the joint likelihood (\rightarrow Definition I/5.1.5) across the parameter space Θ :

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

Proof: In a full probability model (\rightarrow Definition I/5.1.4), the marginal likelihood (\rightarrow Definition I/5.1.9) is defined as the marginal probability (\rightarrow Definition I/1.2.3) of the data y , given only the model m :

$$p(y|m) . \quad (2)$$

Using the law of marginal probability (\rightarrow Definition I/1.2.3), this can be obtained by integrating over the product of likelihood function (\rightarrow Definition I/5.1.2) and prior density (\rightarrow Definition I/5.1.3):

$$p(y|m) = \int_{\Theta} p(y|\theta, m) p(\theta|m) d\theta . \quad (3)$$

Applying the law of conditional probability (\rightarrow Definition I/1.2.4) to the integrand, we have:

$$p(y|m) = \int_{\Theta} p(y, \theta|m) d\theta . \quad (4)$$

Sources:

- original work

Metadata: ID: P91 | shortcut: ml-jl | author: JoramSoch | date: 2020-05-05, 04:59.

5.2 Bayesian inference

5.2.1 Bayes' theorem

Theorem: Let A and B be two arbitrary statements about random variables (\rightarrow Definition I/1.1.1), 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 I/1.2.4) is defined as the ratio of joint probability (\rightarrow Definition I/1.2.2), i.e. the probability of both statements being true, and marginal probability (\rightarrow Definition I/1.2.3), 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.

5.2.2 Bayes' rule

Theorem: Let A_1 , A_2 and B be arbitrary statements about random variables (\rightarrow Definition I/1.1.1) 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 IV/3.3.1) 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/5.2.1), the conditional probabilities (\rightarrow Definition I/1.2.4) 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.

Chapter II

Probability Distributions

1 Univariate discrete distributions

1.1 Discrete uniform distribution

1.1.1 Definition

Definition: Let X be a discrete random variable (\rightarrow Definition I/1.1.1). 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 integer between and including a and b occurs with the same probability.

Sources:

- Wikipedia (2020): “Discrete uniform distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-07-28; URL: https://en.wikipedia.org/wiki/Discrete_uniform_distribution.

Metadata: ID: D88 | shortcut: duni | author: JoramSoch | date: 2020-07-28, 04:05.

1.1.2 Probability mass function

Theorem: Let X be a random variable (\rightarrow Definition I/1.1.1) following a discrete uniform distribution (\rightarrow Definition II/1.1.1):

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

Then, the probability mass function (\rightarrow Definition I/1.4.1) of X is

$$f_X(x) = \frac{1}{b - a + 1} \quad \text{where} \quad x \in \{a, a + 1, \dots, b - 1, b\} . \quad (2)$$

Proof: A discrete uniform variable is defined as (\rightarrow Definition II/1.1.1) having the same probability for each integer between and including a and b . The number of integers between and including a and b is

$$n = b - a + 1 \quad (3)$$

and because the sum across all probabilities (\rightarrow Definition I/1.4.1) is

$$\sum_{x=a}^b f_X(x) = 1 , \quad (4)$$

we have

$$f_X(x) = \frac{1}{n} = \frac{1}{b - a + 1} . \quad (5)$$

Sources:

- original work

Metadata: ID: P140 | shortcut: duni-pmf | author: JoramSoch | date: 2020-07-28, 04:57.

1.1.3 Cumulative distribution function

Theorem: Let X be a random variable (\rightarrow Definition I/1.1.1) following a discrete uniform distribution (\rightarrow Definition II/1.1.1):

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

Then, the cumulative distribution function (\rightarrow Definition I/1.4.3) of X is

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

Proof: The probability mass function of the discrete uniform distribution (\rightarrow Proof II/1.1.2) is

$$\mathcal{U}(x; a, b) = \frac{1}{b - a + 1} \quad \text{where } x \in \{a, a + 1, \dots, b - 1, b\} . \quad (3)$$

Thus, the cumulative distribution function (\rightarrow Definition I/1.4.3) is:

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

From (3), it follows that the cumulative probability increases step-wise by $1/n$ at each integer between and including a and b where

$$n = b - a + 1 \quad (5)$$

is the number of integers between and including a and b . This can be expressed by noting that

$$F_X(x) \stackrel{(3)}{=} \frac{\lfloor x \rfloor - a + 1}{n}, \quad \text{if } a \leq x \leq b . \quad (6)$$

Also, because $\Pr(X < a) = 0$, we have

$$F_X(x) \stackrel{(4)}{=} \int_{-\infty}^x 0 \, dz = 0, \quad \text{if } x < a \quad (7)$$

and because $\Pr(X > b) = 0$, we have

$$\begin{aligned} F_X(x) &\stackrel{(4)}{=} \int_{-\infty}^x \mathcal{U}(z; a, b) \, dz \\ &= \int_{-\infty}^b \mathcal{U}(z; a, b) \, dz + \int_b^x \mathcal{U}(z; a, b) \, dz \\ &= F_X(b) + \int_b^x 0 \, dz \stackrel{(6)}{=} 1 + 0 \\ &= 1, \quad \text{if } x > b . \end{aligned} \quad (8)$$

This completes the proof.

Sources:

- original work

Metadata: ID: P141 | shortcut: duni-cdf | author: JoramSoch | date: 2020-07-28, 05:34.

1.1.4 Quantile function

Theorem: Let X be a random variable (\rightarrow Definition I/1.1.1) following a discrete uniform distribution (\rightarrow Definition II/1.1.1):

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

Then, the quantile function (\rightarrow Definition I/1.4.4) of X is

$$Q_X(p) = a(1 - p) + (b + 1)p - 1 \quad \text{where} \quad p \in \left\{ \frac{1}{n}, \frac{2}{n}, \dots, \frac{b-a}{n}, 1 \right\} \quad (2)$$

with $n = b - a + 1$.

Proof: The cumulative distribution function of the discrete uniform distribution (\rightarrow Proof II/1.1.3) is:

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

Because the CDF only returns (\rightarrow Proof II/1.1.3) multiples of $1/n$ with $n = b - a + 1$, the quantile function (\rightarrow Definition I/1.4.4) is only defined for such values. Since the cumulative probability increases step-wise (\rightarrow Proof II/1.1.3) by $1/n$ at each integer between and including a and b , the minimum x at which

$$F_X(x) = \frac{c}{n} \quad \text{where} \quad c \in \{1, \dots, n\} \quad (4)$$

is given by

$$Q_X\left(\frac{c}{n}\right) = a + \frac{c}{n} \cdot n - 1 . \quad (5)$$

Substituting $p = c/n$ and $n = b - a + 1$, we can finally show:

$$\begin{aligned} Q_X(p) &= a + p \cdot (b - a + 1) - 1 \\ &= a + pb - pa + p - 1 \\ &= a(1 - p) + (b + 1)p - 1 . \end{aligned} \quad (6)$$

Sources:

- original work

Metadata: ID: P142 | shortcut: duni-qf | author: JoramSoch | date: 2020-07-28, 06:17.

1.2 Bernoulli distribution

1.2.1 Definition

Definition: Let X be a random variable (\rightarrow Definition I/1.1.1). 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 I/1.2.1) p and $X = 0$ with probability (\rightarrow Definition I/1.2.1) $q = 1 - p$.

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.

1.2.2 Probability mass function

Theorem: Let X be a random variable (\rightarrow Definition I/1.1.1) following a Bernoulli distribution (\rightarrow Definition II/1.2.1):

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

Then, the probability mass function (\rightarrow Definition I/1.4.1) of X is

$$f_X(x) = \begin{cases} p , & \text{if } x = 1 \\ 1 - p , & \text{if } x = 0 . \end{cases} \quad (2)$$

Proof: This follows directly from the definition of the Bernoulli distribution (\rightarrow Definition II/1.2.1).

Sources:

- original work

Metadata: ID: P96 | shortcut: bern-pmf | author: JoramSoch | date: 2020-05-11, 22:10.

1.2.3 Mean

Theorem: Let X be a random variable (\rightarrow Definition I/1.1.1) following a Bernoulli distribution (\rightarrow Definition II/1.2.1):

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

Then, the mean or expected value (\rightarrow Definition I/1.5.1) of X is

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

Proof: The expected value (\rightarrow Definition I/1.5.1) is the probability-weighted average of all possible values:

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

Since there are only two possible outcomes for a Bernoulli random variable (\rightarrow Proof II/1.2.2), we have:

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

Sources:

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

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

1.3 Binomial distribution

1.3.1 Definition

Definition: Let X be a random variable (\rightarrow Definition I/1.1.1). Then, X is said to follow a binomial distribution with number of trials n and success probability p

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

if X is the number of successes observed in n independent (\rightarrow Definition I/1.2.5) trials, where each trial has two possible outcomes (\rightarrow Definition II/1.2.1) (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.

1.3.2 Probability mass function

Theorem: Let X be a random variable (\rightarrow Definition I/1.1.1) following a binomial distribution (\rightarrow Definition II/1.3.1):

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

Then, the probability mass function (\rightarrow Definition I/1.4.1) of X is

$$f_X(x) = \binom{n}{x} p^x (1 - p)^{n-x} . \tag{2}$$

Proof: A binomial variable is defined as (\rightarrow Definition II/1.3.1) the number of successes observed in n independent (\rightarrow Definition I/1.2.5) trials, where each trial has two possible outcomes (\rightarrow Definition II/1.2.1) (success/failure) and the probability (\rightarrow Definition I/1.2.1) of success and failure are identical across trials ($p/q = 1 - p$).

If one has obtained x successes in n trials, one has also obtained $(n - x)$ failures. The probability of a particular series of x successes and $(n - x)$ failures, when order does matter, is

$$p^x (1 - p)^{n-x} . \quad (3)$$

When order does not matter, there is a number of series consisting of x successes and $(n - x)$ failures. This number is equal to the number of possibilities in which x objects can be chosen from n objects which is given by the binomial coefficient:

$$\binom{n}{x} . \quad (4)$$

In order to obtain the probability of x successes and $(n - x)$ failures, when order does not matter, the probability in (3) has to be multiplied with the number of possibilities in (4) which gives

$$p(X = x|n, p) = \binom{n}{x} p^x (1 - p)^{n-x} \quad (5)$$

which is equivalent to the expression above.

Sources:

- original work

Metadata: ID: P97 | shortcut: bin-pmf | author: JoramSoch | date: 2020-05-11, 22:35.

1.3.3 Mean

Theorem: Let X be a random variable (\rightarrow Definition I/1.1.1) following a binomial distribution (\rightarrow Definition II/1.3.1):

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

Then, the mean or expected value (\rightarrow Definition I/1.5.1) of X is

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

Proof: By definition, a binomial random variable (\rightarrow Definition II/1.3.1) is the sum of n independent and identical Bernoulli trials (\rightarrow Definition II/1.2.1) 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/1.5.4), 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/1.2.3), 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.

1.4 Poisson distribution

1.4.1 Definition

Definition: Let X be a random variable (\rightarrow Definition I/1.1.1). Then, X is said to follow a Poisson distribution with rate λ

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

if and only if its probability mass function (\rightarrow Definition I/1.4.1) is given by

$$\text{Poiss}(x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!} \quad (2)$$

where $x \in \mathbb{N}_0$ and $\lambda > 0$.

Sources:

- Wikipedia (2020): “Poisson distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-25; URL: https://en.wikipedia.org/wiki/Poisson_distribution#Definitions.

Metadata: ID: D62 | shortcut: poiss | author: JoramSoch | date: 2020-05-25, 23:34.

1.4.2 Probability mass function

Theorem: Let X be a random variable (\rightarrow Definition I/1.1.1) following a Poisson distribution (\rightarrow Definition II/1.4.1):

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

Then, the probability mass function (\rightarrow Definition I/1.4.1) of X is

$$f_X(x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x \in \mathbb{N}_0 . \quad (2)$$

Proof: This follows directly from the definition of the Poisson distribution (\rightarrow Definition II/1.4.1).

Sources:

- original work

Metadata: ID: P102 | shortcut: poiss-pmf | author: JoramSoch | date: 2020-05-14, 20:39.

1.4.3 Mean

Theorem: Let X be a random variable (\rightarrow Definition I/1.1.1) following a Poisson distribution (\rightarrow Definition II/1.4.1):

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

Then, the mean or expected value (\rightarrow Definition I/1.5.1) of X is

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

Proof: The expected value of a discrete random variable (\rightarrow Definition I/1.5.1) is defined as

$$E(X) = \sum_{x \in \mathcal{X}} x \cdot f_X(x) , \quad (3)$$

such that, with the probability mass function of the Poisson distribution (\rightarrow Proof II/1.4.2), we have:

$$\begin{aligned} E(X) &= \sum_{x=0}^{\infty} x \cdot \frac{\lambda^x e^{-\lambda}}{x!} \\ &= \sum_{x=1}^{\infty} x \cdot \frac{\lambda^x e^{-\lambda}}{x!} \\ &= e^{-\lambda} \cdot \sum_{x=1}^{\infty} \frac{x}{x!} \lambda^x \\ &= \lambda e^{-\lambda} \cdot \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} . \end{aligned} \quad (4)$$

Substituting $z = x - 1$, such that $x = z + 1$, we get:

$$E(X) = \lambda e^{-\lambda} \cdot \sum_{z=0}^{\infty} \frac{\lambda^z}{z!} . \quad (5)$$

Using the power series expansion of the exponential function

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} , \quad (6)$$

the expected value of X finally becomes

$$\begin{aligned} E(X) &= \lambda e^{-\lambda} \cdot e^{\lambda} \\ &= \lambda . \end{aligned} \quad (7)$$

Sources:

- ProofWiki (2020): “Expectation of Poisson Distribution”; in: *ProofWiki*, retrieved on 2020-08-19; URL: https://proofwiki.org/wiki/Expectation_of_Poisson_Distribution.

Metadata: ID: P151 | shortcut: poiss-mean | author: JoramSoch | date: 2020-08-19, 06:09.

2 Multivariate discrete distributions

2.1 Categorical distribution

2.1.1 Definition

Definition: Let X be a random vector (\rightarrow Definition I/1.1.2). 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 I/1.2.1) 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 Probability mass function

Theorem: Let X be a random vector (\rightarrow Definition I/1.1.2) following a categorical distribution (\rightarrow Definition II/2.1.1):

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

Then, the probability mass function (\rightarrow Definition I/1.4.1) of X is

$$f_X(x) = \begin{cases} p_1 , & \text{if } x = e_1 \\ \vdots & \vdots \\ p_k , & \text{if } x = e_k . \end{cases} \quad (2)$$

where e_1, \dots, e_k are the $1 \times k$ elementary row vectors.

Proof: This follows directly from the definition of the categorical distribution (\rightarrow Definition II/2.1.1).

Sources:

- original work

Metadata: ID: P98 | shortcut: cat-pmf | author: JoramSoch | date: 2020-05-11, 22:58.

2.1.3 Mean

Theorem: Let X be a random vector (\rightarrow Definition I/1.1.2) 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/1.5.1) of X is

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

Proof: If we conceive the outcome of a categorical distribution (\rightarrow Definition II/2.1.1) 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/1.5.1) is

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

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 I/1.1.2). 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 I/1.2.5) 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 Probability mass function

Theorem: Let X be a random vector (\rightarrow Definition I/1.1.2) following a multinomial distribution (\rightarrow Definition II/2.2.1):

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

Then, the probability mass function (\rightarrow Definition I/1.4.1) of X is

$$f_X(x) = \binom{n}{x_1, \dots, x_k} \prod_{i=1}^k p_i^{x_i} . \quad (2)$$

Proof: A multinomial variable is defined as (\rightarrow Definition II/2.2.1) a vector of the numbers of observations belonging to k distinct categories in n independent (\rightarrow Definition I/1.2.5) trials, where each trial has k possible outcomes (\rightarrow Definition II/2.1.1) and the category probabilities (\rightarrow Definition I/1.2.1) are identical across trials.

The probability of a particular series of x_1 observations for category 1, x_2 observations for category 2 etc., when order does matter, is

$$\prod_{i=1}^k p_i^{x_i} . \quad (3)$$

When order does not matter, there is a number of series consisting of x_1 observations for category 1, ..., x_k observations for category k . This number is equal to the number of possibilities in which x_1 category 1 objects, ..., x_k category k objects can be distributed in a sequence of n objects which is given by the multinomial coefficient that can be expressed in terms of factorials:

$$\binom{n}{x_1, \dots, x_k} = \frac{n!}{x_1! \cdot \dots \cdot x_k!} . \quad (4)$$

In order to obtain the probability of x_1 observations for category 1, ..., x_k observations for category k , when order does not matter, the probability in (3) has to be multiplied with the number of possibilities in (4) which gives

$$p(X = x | n, [p_1, \dots, p_k]) = \binom{n}{x_1, \dots, x_k} \prod_{i=1}^k p_i^{x_i} \quad (5)$$

which is equivalent to the expression above.

Sources:

- original work

Metadata: ID: P99 | shortcut: mult-pmf | author: JoramSoch | date: 2020-05-11, 23:30.

2.2.3 Mean

Theorem: Let X be a random vector (\rightarrow Definition I/1.1.2) 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/1.5.1) of X is

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

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

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

and because the expected value is a linear operator (\rightarrow Proof I/1.5.4), 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 categorical distribution (\rightarrow Proof II/2.1.3), we have:

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

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 I/1.1.1). 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 I/1.1.1) 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/1.4.2) 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/1.4.2), 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 I/1.1.1) 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/1.4.3) 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:

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

Thus, the cumulative distribution function (\rightarrow Definition I/1.4.3) 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 I/1.1.1) 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/1.4.4) 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/1.4.4) 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 I/1.1.1) 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/1.5.1) of X is

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

Proof: The expected value (\rightarrow Definition I/1.5.1) 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 I/1.1.1) 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/1.4.3) 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 I/1.1.1) 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/1.4.2):

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

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 I/1.1.1). 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/1.4.2) 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 Standard normal distribution

Definition: Let X be a random variable (\rightarrow Definition I/1.1.1). Then, X is said to be standard normally distributed, if X follows a normal distribution (\rightarrow Definition II/3.2.1) with mean $\mu = 0$ and variance $\sigma^2 = 1$:

$$X \sim \mathcal{N}(0, 1). \quad (1)$$

Sources:

- Wikipedia (2020): “Normal distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-26; URL: https://en.wikipedia.org/wiki/Normal_distribution#Standard_normal_distribution.

Metadata: ID: D63 | shortcut: snorm | author: JoramSoch | date: 2020-05-26, 23:32.

3.2.3 Relation to standard normal distribution

Theorem: Let X be a random variable (\rightarrow Definition I/1.1.1) following a normal distribution (\rightarrow Definition II/3.2.1) with mean μ and variance σ^2 :

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

Then, the quantity $Z = (X - \mu)/\sigma$ will have a standard normal distribution (\rightarrow Definition II/3.2.2) with mean 0 and variance 1:

$$Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1) . \quad (2)$$

Proof: Rearranging to get X in terms of Z , we have

$$X = \sigma Z + \mu . \quad (3)$$

The cumulative distribution function of the normally distributed (\rightarrow Proof II/3.2.5) X is

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

Substituting (3) into (4), we obtain

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

which is the cumulative distribution function (\rightarrow Definition I/1.4.3) of the standard normal distribution (\rightarrow Definition II/3.2.2).

Sources:

- original work

Metadata: ID: P111 | shortcut: norm-snorm | author: JoramSoch | date: 2020-05-26, 23:01.

3.2.4 Probability density function

Theorem: Let X be a random variable (\rightarrow Definition I/1.1.1) 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/1.4.2) 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.

3.2.5 Cumulative distribution function

Theorem: Let X be a random variable (\rightarrow Definition I/1.1.1) 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/1.4.3) 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/3.2.4) 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/1.4.3) 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:

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

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.

3.2.6 Cumulative distribution function without error function

Theorem: Let X be a random variable (\rightarrow Definition I/1.1.1) 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/1.4.3) 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/1.4.2) of the standard normal distribution (\rightarrow Definition II/3.2.2) and $n!!$ is a double factorial.

Proof:

1) First, consider the standard normal distribution (\rightarrow Definition II/3.2.2) $\mathcal{N}(0, 1)$ which has the probability density function (\rightarrow Proof II/3.2.4)

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

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

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

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

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

for constant c , such that the indefinite integral becomes

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

2) Next, let $\Phi(x)$ be the cumulative distribution function (\rightarrow Definition I/1.4.3) of the standard normal distribution (\rightarrow Definition II/3.2.2):

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

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

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

3) Finally, the cumulative distribution functions of the standard normal distribution and the general normal distribution are related to each other (\rightarrow Proof II/3.2.3) as

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

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

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

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.

3.2.7 Quantile function

Theorem: Let X be a random variable (\rightarrow Definition I/1.1.1) following a normal distributions (\rightarrow Definition II/3.2.1):

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

Then, the quantile function (\rightarrow Definition I/1.4.4) of X is

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

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

Proof: The cumulative distribution function of the normal distribution (\rightarrow Proof II/3.2.5) is:

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

Thus, the quantile function (\rightarrow Definition I/1.4.4) is:

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

This can be derived by rearranging equation (3):

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

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.

3.2.8 Mean

Theorem: Let X be a random variable (\rightarrow Definition I/1.1.1) following a normal distribution (\rightarrow Definition II/3.2.1):

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

Then, the mean or expected value (\rightarrow Definition I/1.5.1) of X is

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

Proof: The expected value (\rightarrow Definition I/1.5.1) 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 normal distribution (\rightarrow Proof II/3.2.4), this reads:

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

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*, retrieved on 2020-01-09; 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.

3.2.9 Median

Theorem: Let X be a random variable (\rightarrow Definition I/1.1.1) 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/1.4.3) is $1/2$:

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

The cumulative distribution function of the normal distribution (\rightarrow Proof II/3.2.5) 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.

3.2.10 Mode

Theorem: Let X be a random variable (\rightarrow Definition I/1.1.1) 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/1.4.2):

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

The probability density function of the normal distribution (\rightarrow Proof II/3.2.4) 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.

3.2.11 Variance

Theorem: Let X be a random variable (\rightarrow Definition I/1.1.1) 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/1.6.1) of X is

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

Proof: The variance (\rightarrow Definition I/1.6.1) is the probability-weighted average of the squared deviation from the mean (\rightarrow Definition I/1.5.1):

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

With the expected value (\rightarrow Proof II/3.2.8) and probability density function (\rightarrow Proof II/3.2.4) 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*, retrieved on 2020-01-09; 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.

3.2.12 Full width at half maximum

Theorem: Let X be a random variable (\rightarrow Definition I/1.1.1) following a normal distribution (\rightarrow Definition II/3.2.1):

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

Then, the full width at half maximum (\rightarrow Definition I/1.9.2) (FWHM) of X is

$$\text{FWHM}(X) = 2\sqrt{2 \ln 2} \sigma . \quad (2)$$

Proof: The probability density function of the normal distribution (\rightarrow Proof II/3.2.4) 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 mode of the normal distribution (\rightarrow Proof II/3.2.10) is

$$\text{mode}(X) = \mu , \quad (4)$$

such that

$$f_{\max} = f_X(\text{mode}(X)) \stackrel{(4)}{=} f_X(\mu) \stackrel{(3)}{=} \frac{1}{\sqrt{2\pi}\sigma} . \quad (5)$$

The FWHM bounds satisfy the equation (\rightarrow Definition I/1.9.2)

$$f_X(x_{\text{FWHM}}) = \frac{1}{2} f_{\max} \stackrel{(5)}{=} \frac{1}{2\sqrt{2\pi}\sigma} . \quad (6)$$

Using (3), we can develop this equation as follows:

$$\begin{aligned}
\frac{1}{\sqrt{2\pi}\sigma} \cdot \exp \left[-\frac{1}{2} \left(\frac{x_{\text{FWHM}} - \mu}{\sigma} \right)^2 \right] &= \frac{1}{2\sqrt{2\pi}\sigma} \\
\exp \left[-\frac{1}{2} \left(\frac{x_{\text{FWHM}} - \mu}{\sigma} \right)^2 \right] &= \frac{1}{2} \\
-\frac{1}{2} \left(\frac{x_{\text{FWHM}} - \mu}{\sigma} \right)^2 &= \ln \frac{1}{2} \\
\left(\frac{x_{\text{FWHM}} - \mu}{\sigma} \right)^2 &= -2 \ln \frac{1}{2} \\
\frac{x_{\text{FWHM}} - \mu}{\sigma} &= \pm \sqrt{2 \ln 2} \\
x_{\text{FWHM}} - \mu &= \pm \sqrt{2 \ln 2} \sigma \\
x_{\text{FWHM}} &= \pm \sqrt{2 \ln 2} \sigma + \mu .
\end{aligned} \tag{7}$$

This implies the following two solutions for x_{FWHM}

$$\begin{aligned}
x_1 &= \mu - \sqrt{2 \ln 2} \sigma \\
x_2 &= \mu + \sqrt{2 \ln 2} \sigma ,
\end{aligned} \tag{8}$$

such that the full width at half maximum (\rightarrow Definition I/1.9.2) of X is

$$\begin{aligned}
\text{FWHM}(X) &= \Delta x = x_2 - x_1 \\
&\stackrel{(8)}{=} \left(\mu + \sqrt{2 \ln 2} \sigma \right) - \left(\mu - \sqrt{2 \ln 2} \sigma \right) \\
&= 2\sqrt{2 \ln 2} \sigma .
\end{aligned} \tag{9}$$

Sources:

- Wikipedia (2020): “Full width at half maximum”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-08-19; URL: https://en.wikipedia.org/wiki/Full_width_at_half_maximum.

Metadata: ID: P152 | shortcut: norm-fwhm | author: JoramSoch | date: 2020-08-19, 06:39.

3.2.13 Differential entropy

Theorem: Let X be a random variable (\rightarrow Definition I/1.1.1) following a normal distribution (\rightarrow Definition II/3.2.1):

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

Then, the differential entropy (\rightarrow Definition I/2.2.1) of X is

$$h(X) = \frac{1}{2} \ln (2\pi\sigma^2 e) . \tag{2}$$

Proof: The differential entropy (\rightarrow Definition I/2.2.1) of a random variable is defined as

$$h(X) = - \int_{\mathcal{X}} p(x) \log_b p(x) dx . \quad (3)$$

To measure $h(X)$ in nats, we set $b = e$, such that (\rightarrow Definition I/1.5.1)

$$h(X) = -E [\ln p(x)] . \quad (4)$$

With the probability density function of the normal distribution (\rightarrow Proof II/3.2.4), the differential entropy of X is:

$$\begin{aligned} h(X) &= -E \left[\ln \left(\frac{1}{\sqrt{2\pi}\sigma} \cdot \exp \left[-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right] \right) \right] \\ &= -E \left[-\frac{1}{2} \ln(2\pi\sigma^2) - \frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right] \\ &= \frac{1}{2} \ln(2\pi\sigma^2) + \frac{1}{2} E \left[\left(\frac{x-\mu}{\sigma} \right)^2 \right] \\ &= \frac{1}{2} \ln(2\pi\sigma^2) + \frac{1}{2} \cdot \frac{1}{\sigma^2} \cdot E [(x-\mu)^2] \\ &= \frac{1}{2} \ln(2\pi\sigma^2) + \frac{1}{2} \cdot \frac{1}{\sigma^2} \cdot \sigma^2 \\ &= \frac{1}{2} \ln(2\pi\sigma^2) + \frac{1}{2} \\ &= \frac{1}{2} \ln(2\pi\sigma^2 e) . \end{aligned} \quad (5)$$

Sources:

- Wang, Peng-Hua (2012): “Differential Entropy”; in: *National Taipei University*; URL: <https://web.ntpu.edu.tw/~phwang/teaching/2012s/IT/slides/chap08.pdf>.

Metadata: ID: P101 | shortcut: norm-dent | author: JoramSoch | date: 2020-05-14, 20:09.

3.2.14 Moment-generating function

Theorem: Let X be a random variable (\rightarrow Definition I/1.1.1) 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/1.4.5) 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/3.2.4) 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/1.4.5) is defined as

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

Using the expected value for continuous random variables (\rightarrow Definition I/1.5.1), 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

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

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

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

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

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

this finally becomes

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

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 I/1.1.1). 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/1.4.2) 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 Standard gamma distribution

Definition: Let X be a random variable (\rightarrow Definition I/1.1.1). Then, X is said to have a standard gamma distribution, if X follows a gamma distribution (\rightarrow Definition II/3.3.1) with shape $a > 0$ and rate $b = 1$:

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

Sources:

- JoramSoch (2017): “Gamma-distributed random numbers”; in: *MACS – a new SPM toolbox for model assessment, comparison and selection*, retrieved on 2020-05-26; URL: https://github.com/JoramSoch/MACS/blob/master/MD_gamrnd.m; DOI: 10.5281/zenodo.845404.

Metadata: ID: D64 | shortcut: sgam | author: JoramSoch | date: 2020-05-26, 23:36.

3.3.3 Relation to standard gamma distribution

Theorem: Let X be a random variable (\rightarrow Definition I/1.1.1) following a gamma distribution (\rightarrow Definition II/3.3.1) with shape a and rate b :

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

Then, the quantity $Z = bX$ will have a standard gamma distribution (\rightarrow Definition II/3.3.2) with shape a and rate 1:

$$Z = bX \sim \text{Gam}(a, 1) . \quad (2)$$

Proof: Rearranging to get X in terms of Z , we have

$$X = \frac{1}{b}Z . \quad (3)$$

The cumulative distribution function of the gamma-distributed (\rightarrow Proof “gam-cdf”) X is

$$F_X(t) = \int_{-\infty}^t \frac{b^a}{\Gamma(a)} x^{a-1} \exp[-bx] dx . \quad (4)$$

Substituting (3) into (4), we obtain

$$\begin{aligned} F_Z(t) &= \int_{-\infty}^t \frac{b^a}{\Gamma(a)} \left(\frac{1}{b}z\right)^{a-1} \exp\left[-b\left(\frac{1}{b}z\right)\right] d\left(\frac{1}{b}z\right) \\ &= \int_{-\infty}^t \frac{b^a}{b} \left(\frac{1}{b}\right)^{a-1} \cdot \frac{1}{\Gamma(a)} z^{a-1} \exp[-z] dz \\ &= \int_{-\infty}^t \frac{1}{\Gamma(a)} z^{a-1} \exp[-z] dz \end{aligned} \quad (5)$$

which is the cumulative distribution function (\rightarrow Definition I/1.4.3) of the standard gamma distribution (\rightarrow Definition II/3.3.2).

Sources:

- original work

Metadata: ID: P112 | shortcut: gam-sgam | author: JoramSoch | date: 2020-05-26, 23:14.

3.3.4 Probability density function

Theorem: Let X be a positive random variable (\rightarrow Definition I/1.1.1) 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/1.4.2) 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.

3.3.5 Mean

Theorem: Let X be a random variable (\rightarrow Definition I/1.1.1) following a gamma distribution (\rightarrow Definition II/3.3.1):

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

Then, the mean or expected value (\rightarrow Definition I/1.5.1) of X is

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

Proof: The expected value (\rightarrow Definition I/1.5.1) 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 gamma distribution (\rightarrow Proof II/3.3.4), this reads:

$$\begin{aligned} \mathbb{E}(X) &= \int_0^\infty x \cdot \frac{b^a}{\Gamma(a)} x^{a-1} \exp[-bx] \, dx \\ &= \int_0^\infty \frac{b^a}{\Gamma(a)} x^{(a+1)-1} \exp[-bx] \, dx \\ &= \int_0^\infty \frac{1}{b} \cdot \frac{b^{a+1}}{\Gamma(a)} x^{(a+1)-1} \exp[-bx] \, dx . \end{aligned} \quad (4)$$

Employing the relation $\Gamma(x+1) = \Gamma(x) \cdot x$, we have

$$\mathbb{E}(X) = \int_0^\infty \frac{a}{b} \cdot \frac{b^{a+1}}{\Gamma(a+1)} x^{(a+1)-1} \exp[-bx] \, dx \quad (5)$$

and again using the density of the gamma distribution (\rightarrow Proof II/3.3.4), we get

$$\begin{aligned} \mathbb{E}(X) &= \frac{a}{b} \int_0^\infty \text{Gam}(x; a+1, b) \, dx \\ &= \frac{a}{b} . \end{aligned} \quad (6)$$

Sources:

- Turlapaty, Anish (2013): “Gamma random variable: mean & variance”; in: *YouTube*, retrieved on 2020-05-19; URL: <https://www.youtube.com/watch?v=Sy4wP-Y2dmA>.

Metadata: ID: P108 | shortcut: gam-mean | author: JoramSoch | date: 2020-05-19, 06:54.

3.3.6 Variance

Theorem: Let X be a random variable (\rightarrow Definition I/1.1.1) following a gamma distribution (\rightarrow Definition II/3.3.1):

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

Then, the variance (\rightarrow Definition I/1.6.1) of X is

$$\text{Var}(X) = \frac{a}{b^2} . \quad (2)$$

Proof: The variance (\rightarrow Definition I/1.6.1) can be expressed in terms of expected values (\rightarrow Proof I/1.6.2) as

$$\text{Var}(X) = E(X^2) - E(X)^2 . \quad (3)$$

The expected value of a gamma random variable (\rightarrow Proof II/3.3.5) is

$$E(X) = \frac{a}{b} . \quad (4)$$

With the probability density function of the gamma distribution (\rightarrow Proof II/3.3.4), the expected value of a squared gamma random variable is

$$\begin{aligned} E(X^2) &= \int_0^\infty x^2 \cdot \frac{b^a}{\Gamma(a)} x^{a-1} \exp[-bx] dx \\ &= \int_0^\infty \frac{b^a}{\Gamma(a)} x^{(a+2)-1} \exp[-bx] dx \\ &= \int_0^\infty \frac{1}{b^2} \cdot \frac{b^{a+2}}{\Gamma(a)} x^{(a+2)-1} \exp[-bx] dx . \end{aligned} \quad (5)$$

Twice-applying the relation $\Gamma(x+1) = \Gamma(x) \cdot x$, we have

$$E(X^2) = \int_0^\infty \frac{a(a+1)}{b^2} \cdot \frac{b^{a+2}}{\Gamma(a+2)} x^{(a+2)-1} \exp[-bx] dx \quad (6)$$

and again using the density of the gamma distribution (\rightarrow Proof II/3.3.4), we get

$$\begin{aligned} E(X^2) &= \frac{a(a+1)}{b^2} \int_0^\infty \text{Gam}(x; a+2, b) dx \\ &= \frac{a^2 + a}{b^2} . \end{aligned} \quad (7)$$

Plugging (7) and (4) into (3), the variance of a gamma random variable finally becomes

$$\begin{aligned} \text{Var}(X) &= \frac{a^2 + a}{b^2} - \left(\frac{a}{b}\right)^2 \\ &= \frac{a}{b^2} . \end{aligned} \quad (8)$$

Sources:

- Turlapaty, Anish (2013): “Gamma random variable: mean & variance”; in: *YouTube*, retrieved on 2020-05-19; URL: <https://www.youtube.com/watch?v=Sy4wP-Y2dmA>.

Metadata: ID: P109 | shortcut: gam-var | author: JoramSoch | date: 2020-05-19, 07:20.

3.3.7 Logarithmic expectation

Theorem: Let X be a random variable (\rightarrow Definition I/1.1.1) following a gamma distribution (\rightarrow Definition II/3.3.1):

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

Then, the expectation (\rightarrow Definition I/1.5.1) of the natural logarithm of X is

$$\mathbb{E}(\ln X) = \psi(a) - \ln(b) . \quad (2)$$

Proof: Let $Y = \ln(X)$, such that $\mathbb{E}(Y) = \mathbb{E}(\ln X)$ and consider the special case that $b = 1$. In this case, the probability density function of the gamma distribution (\rightarrow Proof II/3.3.4) is

$$f_X(x) = \frac{1}{\Gamma(a)} x^{a-1} \exp[-x] . \quad (3)$$

Multiplying this function with dx , we obtain

$$f_X(x) dx = \frac{1}{\Gamma(a)} x^a \exp[-x] \frac{dx}{x} . \quad (4)$$

Substituting $y = \ln x$, i.e. $x = e^y$, such that $dx/dy = x$, i.e. $dx/x = dy$, we get

$$\begin{aligned} f_Y(y) dy &= \frac{1}{\Gamma(a)} (e^y)^a \exp[-e^y] dy \\ &= \frac{1}{\Gamma(a)} \exp[ay - e^y] dy . \end{aligned} \quad (5)$$

Because $f_Y(y)$ integrates to one, we have

$$\begin{aligned} 1 &= \int_{\mathbb{R}} f_Y(y) dy \\ 1 &= \int_{\mathbb{R}} \frac{1}{\Gamma(a)} \exp[ay - e^y] dy \\ \Gamma(a) &= \int_{\mathbb{R}} \exp[ay - e^y] dy . \end{aligned} \quad (6)$$

Note that the integrand in (6) is differentiable with respect to a :

$$\begin{aligned} \frac{d}{da} \exp[ay - e^y] dy &= y \exp[ay - e^y] dy \\ &\stackrel{(5)}{=} \Gamma(a) y f_Y(y) dy . \end{aligned} \quad (7)$$

Now we can calculate the expected value of $Y = \ln(X)$:

$$\begin{aligned}
 E(Y) &= \int_{\mathbb{R}} y f_Y(y) dy \\
 &\stackrel{(7)}{=} \frac{1}{\Gamma(a)} \int_{\mathbb{R}} \frac{d}{da} \exp[ay - e^y] dy \\
 &= \frac{1}{\Gamma(a)} \frac{d}{da} \int_{\mathbb{R}} \exp[ay - e^y] dy \\
 &\stackrel{(6)}{=} \frac{1}{\Gamma(a)} \frac{d}{da} \Gamma(a) \\
 &= \frac{\Gamma'(a)}{\Gamma(a)} .
 \end{aligned} \tag{8}$$

Using the derivative of a logarithmized function

$$\frac{d}{dx} \ln f(x) = \frac{f'(x)}{f(x)} \tag{9}$$

and the definition of the digamma function

$$\psi(x) = \frac{d}{dx} \ln \Gamma(x) , \tag{10}$$

we have

$$E(Y) = \psi(a) . \tag{11}$$

Finally, noting that $1/b$ acts as a scaling parameter (\rightarrow Proof II/3.3.3) on a gamma-distributed (\rightarrow Definition II/3.3.1) random variable (\rightarrow Definition I/1.1.1),

$$X \sim \text{Gam}(a, 1) \quad \Rightarrow \quad \frac{1}{b}X \sim \text{Gam}(a, b) , \tag{12}$$

and that a scaling parameter acts additively on the logarithmic expectation of a random variable,

$$E[\ln(cX)] = E[\ln(X) + \ln(c)] = E[\ln(X)] + \ln(c) , \tag{13}$$

it follows that

$$X \sim \text{Gam}(a, b) \quad \Rightarrow \quad E(\ln X) = \psi(a) - \ln(b) . \tag{14}$$

Sources:

- whuber (2018): “What is the expected value of the logarithm of Gamma distribution?”; in: *StackExchange CrossValidated*, retrieved on 2020-05-25; URL: <https://stats.stackexchange.com/questions/370880/what-is-the-expected-value-of-the-logarithm-of-gamma-distribution>.

Metadata: ID: P110 | shortcut: gam-logmean | author: JoramSoch | date: 2020-05-25, 21:28.

3.3.8 Kullback-Leibler divergence

Theorem: Let x be a random variable (\rightarrow Definition I/1.1.1). Assume two gamma distributions (\rightarrow Definition II/3.3.1) P and Q specifying the probability distribution of x as

$$\begin{aligned} P : x &\sim \text{Gam}(a_1, b_1) \\ Q : x &\sim \text{Gam}(a_2, b_2) . \end{aligned} \quad (1)$$

Then, the Kullback-Leibler divergence (\rightarrow Definition I/2.5.1) of P from Q is given by

$$\text{KL}[P \parallel Q] = a_2 \ln \frac{b_1}{b_2} - \ln \frac{\Gamma(a_1)}{\Gamma(a_2)} + (a_1 - a_2) \psi(a_1) - (b_1 - b_2) \frac{a_1}{b_1} . \quad (2)$$

Proof: The KL divergence for a continuous random variable (\rightarrow Definition I/2.5.1) is given by

$$\text{KL}[P \parallel Q] = \int_{\mathcal{X}} p(x) \ln \frac{p(x)}{q(x)} dx \quad (3)$$

which, applied to the gamma distributions (\rightarrow Definition II/3.3.1) in (1), yields

$$\begin{aligned} \text{KL}[P \parallel Q] &= \int_{-\infty}^{+\infty} \text{Gam}(x; a_1, b_1) \ln \frac{\text{Gam}(x; a_1, b_1)}{\text{Gam}(x; a_2, b_2)} dx \\ &= \left\langle \ln \frac{\text{Gam}(x; a_1, b_1)}{\text{Gam}(x; a_2, b_2)} \right\rangle_{p(x)} . \end{aligned} \quad (4)$$

Using the probability density function of the gamma distribution (\rightarrow Proof II/3.3.4), this becomes:

$$\begin{aligned} \text{KL}[P \parallel Q] &= \left\langle \ln \frac{\frac{b_1^{a_1}}{\Gamma(a_1)} x^{a_1-1} \exp[-b_1 x]}{\frac{b_2^{a_2}}{\Gamma(a_2)} x^{a_2-1} \exp[-b_2 x]} \right\rangle_{p(x)} \\ &= \left\langle \ln \left(\frac{b_1^{a_1}}{b_2^{a_2}} \cdot \frac{\Gamma(a_2)}{\Gamma(a_1)} \cdot x^{a_1-a_2} \cdot \exp[-(b_1 - b_2)x] \right) \right\rangle_{p(x)} \\ &= \langle a_1 \cdot \ln b_1 - a_2 \cdot \ln b_2 - \ln \Gamma(a_1) + \ln \Gamma(a_2) + (a_1 - a_2) \cdot \ln x - (b_1 - b_2) \cdot x \rangle_{p(x)} . \end{aligned} \quad (5)$$

Using the mean of the gamma distribution (\rightarrow Proof II/3.3.5) and the expected value of a logarithmized gamma variate (\rightarrow Proof II/3.3.7)

$$\begin{aligned} x \sim \text{Gam}(a, b) \quad \Rightarrow \quad \langle x \rangle &= \frac{a}{b} \quad \text{and} \\ \langle \ln x \rangle &= \psi(a) - \ln(b) , \end{aligned} \quad (6)$$

the Kullback-Leibler divergence from (5) becomes:

$$\begin{aligned} \text{KL}[P \parallel Q] &= a_1 \cdot \ln b_1 - a_2 \cdot \ln b_2 - \ln \Gamma(a_1) + \ln \Gamma(a_2) + (a_1 - a_2) \cdot (\psi(a_1) - \ln(b_1)) - (b_1 - b_2) \cdot \frac{a_1}{b_1} \\ &= a_2 \cdot \ln b_1 - a_2 \cdot \ln b_2 - \ln \Gamma(a_1) + \ln \Gamma(a_2) + (a_1 - a_2) \cdot \psi(a_1) - (b_1 - b_2) \cdot \frac{a_1}{b_1} . \end{aligned} \quad (7)$$

Finally, combining the logarithms, we get:

$$\text{KL}[P || Q] = a_2 \ln \frac{b_1}{b_2} - \ln \frac{\Gamma(a_1)}{\Gamma(a_2)} + (a_1 - a_2) \psi(a_1) - (b_1 - b_2) \frac{a_1}{b_1} . \quad (8)$$

Sources:

- Penny, William D. (2001): “KL-Divergences of Normal, Gamma, Dirichlet and Wishart densities”; in: *University College, London*; URL: <https://www.fil.ion.ucl.ac.uk/~wpenny/publications/densities.ps>.

Metadata: ID: P93 | shortcut: gam-kl | author: JoramSoch | date: 2020-05-05, 08:41.

3.4 Exponential distribution

3.4.1 Definition

****Definition**:** Let X be a random variable (\rightarrow Definition I/1.1.1). 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/1.4.2) 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/3.3.4) 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 I/1.1.1) 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/1.4.2) 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 I/1.1.1) 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/1.4.3) 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/1.4.3) 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 I/1.1.1) following an exponential distribution (\rightarrow Definition II/3.4.1):

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

Then, the quantile function (\rightarrow Definition I/1.4.4) 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/1.4.4) 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 I/1.1.1) 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/1.5.1) of X is

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

Proof: The expected value (\rightarrow Definition I/1.5.1) 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 exponential distribution (\rightarrow Proof II/3.4.3), this reads:

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

Using the following anti-derivative

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

the expected value becomes

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

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 I/1.1.1) following an exponential distribution (\rightarrow Definition II/3.4.1):

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

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

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

Proof: The median (\rightarrow Definition “med”) is the value at which the cumulative distribution function (\rightarrow Definition I/1.4.3) is $1/2$:

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

The cumulative distribution function of the exponential distribution (\rightarrow Proof II/3.4.4) 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 I/1.1.1) 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/1.4.2):

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

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.

3.5 Beta distribution

3.5.1 Definition

****Definition**:** Let X be a random variable (\rightarrow Definition I/1.1.1). Then, X is said to follow a beta distribution with shape parameters α and β

$$X \sim \text{Bet}(\alpha, \beta) , \quad (1)$$

if and only if its probability density function (\rightarrow Definition I/1.4.2) is given by

$$\text{Bet}(x; \alpha, \beta) = \frac{1}{\text{B}(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1} \quad (2)$$

where $\alpha > 0$ and $\beta > 0$, and the density is zero, if $x \notin [0, 1]$.

Sources:

- Wikipedia (2020): “Beta distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-10; URL: https://en.wikipedia.org/wiki/Beta_distribution#Definitions.

Metadata: ID: D53 | shortcut: beta | author: JoramSoch | date: 2020-05-10, 20:29.

3.5.2 Probability density function

Theorem: Let X be a random variable (\rightarrow Definition I/1.1.1) following a beta distribution (\rightarrow Definition II/3.5.1):

$$X \sim \text{Bet}(\alpha, \beta) . \quad (1)$$

Then, the probability density function (\rightarrow Definition I/1.4.2) of X is

$$f_X(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1} . \quad (2)$$

Proof: This follows directly from the definition of the beta distribution (\rightarrow Definition II/3.5.1).

Sources:

- original work

Metadata: ID: P94 | shortcut: beta-pdf | author: JoramSoch | date: 2020-05-05, 21:03.

3.6 Wald distribution

3.6.1 Definition

****Definition**:** Let X be a random variable (\rightarrow Definition I/1.1.1). Then, X is said to follow a Wald distribution with drift rate γ and threshold α

$$X \sim \text{Wald}(\gamma, \alpha) , \quad (1)$$

if and only if its probability density function (\rightarrow Definition I/1.4.2) is given by

$$\text{Wald}(x; \gamma, \alpha) = \frac{\alpha}{\sqrt{2\pi x^3}} \exp\left(-\frac{(\alpha - \gamma x)^2}{2x}\right) \quad (2)$$

where $\gamma > 0$, $\alpha > 0$, and the density is zero if $x \leq 0$.

Sources:

- Anders, R., Alario, F.-X., and van Maanen, L. (2016): “The Shifted Wald Distribution for Response Time Data Analysis”; in: *Psychological Methods*, vol. 21, no. 3, pp. 309-327; URL: <https://dx.doi.org/10.1037/met0000066>; DOI: 10.1037/met0000066.

Metadata: ID: D95 | shortcut: wald | author: tomfaulkenberry | date: 2020-09-04, 12:00.

3.6.2 Probability density function

Theorem: Let X be a positive random variable (\rightarrow Definition I/1.1.1) following a Wald distribution (\rightarrow Definition II/3.6.1):

$$X \sim \text{Wald}(\gamma, \alpha) . \quad (1)$$

Then, the probability density function (\rightarrow Definition I/1.4.2) of X is

$$f_X(x) = \frac{\alpha}{\sqrt{2\pi x^3}} \exp\left(-\frac{(\alpha - \gamma x)^2}{2x}\right) . \quad (2)$$

Proof: This follows directly from the definition of the Wald distribution (\rightarrow Definition II/3.6.1).

Sources:

- original work

Metadata: ID: P162 | shortcut: wald-pdf | author: tomfaulkenberry | date: 2020-09-04, 12:00.

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 I/1.1.2). 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/1.4.2) 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 I/1.1.2) 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/1.4.2) 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 Differential entropy

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

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

Then, the differential entropy (\rightarrow Definition I/2.2.1) of x in nats is

$$h(x) = \frac{n}{2} \ln(2\pi) + \frac{1}{2} \ln |\Sigma| + \frac{1}{2} n . \quad (2)$$

Proof: The differential entropy (\rightarrow Definition I/2.2.1) of a random variable is defined as

$$h(X) = - \int_{\mathcal{X}} p(x) \log_b p(x) dx . \quad (3)$$

To measure $h(X)$ in nats, we set $b = e$, such that (\rightarrow Definition I/1.5.1)

$$h(X) = -E [\ln p(x)] . \quad (4)$$

With the probability density function of the multivariate normal distribution (\rightarrow Proof II/4.1.2), the differential entropy of x is:

$$\begin{aligned} h(x) &= -E \left[\ln \left(\frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \cdot \exp \left[-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right] \right) \right] \\ &= -E \left[-\frac{n}{2} \ln(2\pi) - \frac{1}{2} \ln |\Sigma| - \frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right] \\ &= \frac{n}{2} \ln(2\pi) + \frac{1}{2} \ln |\Sigma| + \frac{1}{2} E [(x - \mu)^T \Sigma^{-1} (x - \mu)] . \end{aligned} \quad (5)$$

The last term can be evaluated as

$$\begin{aligned} E [(x - \mu)^T \Sigma^{-1} (x - \mu)] &= E [\text{tr} ((x - \mu)^T \Sigma^{-1} (x - \mu))] \\ &= E [\text{tr} (\Sigma^{-1} (x - \mu) (x - \mu)^T)] \\ &= \text{tr} (\Sigma^{-1} E [(x - \mu) (x - \mu)^T]) \\ &= \text{tr} (\Sigma^{-1} \Sigma) \\ &= \text{tr} (I_n) \\ &= n , \end{aligned} \quad (6)$$

such that the differential entropy is

$$h(x) = \frac{n}{2} \ln(2\pi) + \frac{1}{2} \ln |\Sigma| + \frac{1}{2} n . \quad (7)$$

Sources:

- Kiuahn (2018): “Entropy of the multivariate Gaussian”; in: *StackExchange Mathematics*, retrieved on 2020-05-14; URL: <https://math.stackexchange.com/questions/2029707/entropy-of-the-multivariate-ga>

Metadata: ID: P100 | shortcut: mvn-dent | author: JoramSoch | date: 2020-05-14, 19:49.

4.1.4 Kullback-Leibler divergence

Theorem: Let x be an $n \times 1$ random vector (\rightarrow Definition I/1.1.2). Assume two multivariate normal distributions (\rightarrow Definition II/4.1.1) P and Q specifying the probability distribution of x as

$$\begin{aligned} P : x &\sim \mathcal{N}(\mu_1, \Sigma_1) \\ Q : x &\sim \mathcal{N}(\mu_2, \Sigma_2) . \end{aligned} \quad (1)$$

Then, the Kullback-Leibler divergence (\rightarrow Definition I/2.5.1) of P from Q is given by

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

Proof: The KL divergence for a continuous random variable (\rightarrow Definition I/2.5.1) is given by

$$\text{KL}[P \parallel Q] = \int_{\mathcal{X}} p(x) \ln \frac{p(x)}{q(x)} dx \quad (3)$$

which, applied to the multivariate normal distributions (\rightarrow Definition II/4.1.1) in (1), yields

$$\begin{aligned} \text{KL}[P \parallel Q] &= \int_{\mathbb{R}^n} \mathcal{N}(x; \mu_1, \Sigma_1) \ln \frac{\mathcal{N}(x; \mu_1, \Sigma_1)}{\mathcal{N}(x; \mu_2, \Sigma_2)} dx \\ &= \left\langle \ln \frac{\mathcal{N}(x; \mu_1, \Sigma_1)}{\mathcal{N}(x; \mu_2, \Sigma_2)} \right\rangle_{p(x)} . \end{aligned} \quad (4)$$

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

$$\begin{aligned} \text{KL}[P \parallel Q] &= \left\langle \ln \frac{\frac{1}{\sqrt{(2\pi)^n |\Sigma_1|}} \cdot \exp \left[-\frac{1}{2} (x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1) \right]}{\frac{1}{\sqrt{(2\pi)^n |\Sigma_2|}} \cdot \exp \left[-\frac{1}{2} (x - \mu_2)^T \Sigma_2^{-1} (x - \mu_2) \right]} \right\rangle_{p(x)} \\ &= \left\langle \frac{1}{2} \ln \frac{|\Sigma_2|}{|\Sigma_1|} - \frac{1}{2} (x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1) + \frac{1}{2} (x - \mu_2)^T \Sigma_2^{-1} (x - \mu_2) \right\rangle_{p(x)} \\ &= \frac{1}{2} \left\langle \ln \frac{|\Sigma_2|}{|\Sigma_1|} - (x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1) + (x - \mu_2)^T \Sigma_2^{-1} (x - \mu_2) \right\rangle_{p(x)} . \end{aligned} \quad (5)$$

Now, using the fact that $x = \text{tr}(x)$, if a is scalar, and the trace property $\text{tr}(ABC) = \text{tr}(BCA)$, we have:

$$\begin{aligned} \text{KL}[P \parallel Q] &= \frac{1}{2} \left\langle \ln \frac{|\Sigma_2|}{|\Sigma_1|} - \text{tr} [\Sigma_1^{-1} (x - \mu_1)(x - \mu_1)^T] + \text{tr} [\Sigma_2^{-1} (x - \mu_2)(x - \mu_2)^T] \right\rangle_{p(x)} \\ &= \frac{1}{2} \left\langle \ln \frac{|\Sigma_2|}{|\Sigma_1|} - \text{tr} [\Sigma_1^{-1} (x - \mu_1)(x - \mu_1)^T] + \text{tr} [\Sigma_2^{-1} (xx^T - 2\mu_2 x^T + \mu_2 \mu_2^T)] \right\rangle_{p(x)} . \end{aligned} \quad (6)$$

Because trace function and expected value (\rightarrow Definition I/1.5.1) are both linear operators, the expectation can be moved inside the trace:

$$\begin{aligned}
\text{KL}[P || Q] &= \frac{1}{2} \left(\ln \frac{|\Sigma_2|}{|\Sigma_1|} - \text{tr} \left[\Sigma_1^{-1} \langle (x - \mu_1)(x - \mu_1)^T \rangle_{p(x)} \right] + \text{tr} \left[\Sigma_2^{-1} \langle xx^T - 2\mu_2 x^T + \mu_2 \mu_2^T \rangle_{p(x)} \right] \right) \\
&= \frac{1}{2} \left(\ln \frac{|\Sigma_2|}{|\Sigma_1|} - \text{tr} \left[\Sigma_1^{-1} \langle (x - \mu_1)(x - \mu_1)^T \rangle_{p(x)} \right] + \text{tr} \left[\Sigma_2^{-1} \left(\langle xx^T \rangle_{p(x)} - \langle 2\mu_2 x^T \rangle_{p(x)} + \langle \mu_2 \mu_2^T \rangle_{p(x)} \right) \right] \right)
\end{aligned} \tag{7}$$

Using the expectation of a linear form for the multivariate normal distribution (\rightarrow Proof II/4.1.5)

$$x \sim \mathcal{N}(\mu, \Sigma) \quad \Rightarrow \quad \langle Ax \rangle = A\mu \tag{8}$$

and the expectation of a quadratic form for the multivariate normal distribution (\rightarrow Proof I/1.5.7)

$$x \sim \mathcal{N}(\mu, \Sigma) \quad \Rightarrow \quad \langle x^T A x \rangle = \mu^T A \mu + \text{tr}(A \Sigma), \tag{9}$$

the Kullback-Leibler divergence from (7) becomes:

$$\begin{aligned}
\text{KL}[P || Q] &= \frac{1}{2} \left(\ln \frac{|\Sigma_2|}{|\Sigma_1|} - \text{tr} [\Sigma_1^{-1} \Sigma_1] + \text{tr} [\Sigma_2^{-1} (\Sigma_1 + \mu_1 \mu_1^T - 2\mu_2 \mu_1^T + \mu_2 \mu_2^T)] \right) \\
&= \frac{1}{2} \left(\ln \frac{|\Sigma_2|}{|\Sigma_1|} - \text{tr} [I_n] + \text{tr} [\Sigma_2^{-1} \Sigma_1] + \text{tr} [\Sigma_2^{-1} (\mu_1 \mu_1^T - 2\mu_2 \mu_1^T + \mu_2 \mu_2^T)] \right) \\
&= \frac{1}{2} \left(\ln \frac{|\Sigma_2|}{|\Sigma_1|} - n + \text{tr} [\Sigma_2^{-1} \Sigma_1] + \text{tr} [\mu_1^T \Sigma_2^{-1} \mu_1 - 2\mu_1^T \Sigma_2^{-1} \mu_2 + \mu_2^T \Sigma_2^{-1} \mu_2] \right) \\
&= \frac{1}{2} \left[\ln \frac{|\Sigma_2|}{|\Sigma_1|} - n + \text{tr} [\Sigma_2^{-1} \Sigma_1] + (\mu_2 - \mu_1)^T \Sigma_2^{-1} (\mu_2 - \mu_1) \right].
\end{aligned} \tag{10}$$

Finally, rearranging the terms, we get:

$$\text{KL}[P || Q] = \frac{1}{2} \left[(\mu_2 - \mu_1)^T \Sigma_2^{-1} (\mu_2 - \mu_1) + \text{tr}(\Sigma_2^{-1} \Sigma_1) - \ln \frac{|\Sigma_1|}{|\Sigma_2|} - n \right]. \tag{11}$$

Sources:

- Duchi, John (2014): “Derivations for Linear Algebra and Optimization”; in: *University of California, Berkeley*; URL: http://www.eecs.berkeley.edu/~jduchi/projects/general_notes.pdf.

Metadata: ID: P92 | shortcut: mvn-kl | author: JoramSoch | date: 2020-05-05, 06:57.

4.1.5 Linear transformation

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

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

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

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

Proof: The moment-generating function of a random vector (\rightarrow Definition I/1.4.5) 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

$$\begin{aligned} M_y(t) &\stackrel{(2)}{=} \mathbb{E}(\exp[t^T(Ax + b)]) \\ &= \mathbb{E}(\exp[t^T Ax] \cdot \exp[t^T b]) \\ &= \exp[t^T b] \cdot \mathbb{E}(\exp[t^T Ax]) \\ &\stackrel{(3)}{=} \exp[t^T b] \cdot M_x(At) . \end{aligned} \quad (4)$$

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

$$\begin{aligned} M_y(t) &\stackrel{(4)}{=} \exp[t^T b] \cdot M_x(At) \\ &\stackrel{(5)}{=} \exp[t^T b] \cdot \exp\left[t^T A \mu + \frac{1}{2} t^T A \Sigma A^T t\right] \\ &= \exp\left[t^T (A \mu + b) + \frac{1}{2} t^T A \Sigma A^T t\right] . \end{aligned} \quad (6)$$

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*, retrieved on 2019-08-27; 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.

4.1.6 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 I/1.3.3) 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/4.1.5) 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.

4.1.7 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 I/1.3.4) 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 (\rightarrow Definition I/1.5.1) and covariance (\rightarrow Definition I/1.7.1) 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 I/1.3.2) of x_1 and x_2 is:

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

Moreover, the marginal distribution (\rightarrow Definition I/1.3.3) of x_2 follows from (\rightarrow Proof II/4.1.6) (1) and (4) as

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

According to the law of conditional probability (\rightarrow Definition I/1.2.4), 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:

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

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

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

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 I/1.1.2) and let Y be a positive random variable (\rightarrow Definition I/1.1.1). 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 I/1.2.2) density function (\rightarrow Definition I/1.4.2) 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/3.3.4) with shape a and rate b . The $n \times n$ matrix Λ is referred to as the precision matrix (\rightarrow Definition I/1.7.8) 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 I/1.2.2) density function (\rightarrow Definition I/1.4.2) 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/3.3.4), 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 be an $n \times 1$ random vector (\rightarrow Definition I/1.1.2) and let y be a positive random variable (\rightarrow Definition I/1.1.1). 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 I/2.5.1) of P from Q is given by

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

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

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

The Kullback-Leibler divergence of the multivariate normal distribution (\rightarrow Proof II/4.1.4) is

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

and the Kullback-Leibler divergence of the univariate gamma distribution (\rightarrow Proof II/3.3.8) 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 I/2.5.1) 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

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

Using the law of conditional probability (\rightarrow Definition I/1.2.4), this can be evaluated as follows:

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

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

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

and using the relation (\rightarrow Proof II/3.3.5) $y \sim \text{Gam}(a, b) \Rightarrow \langle y \rangle = a/b$, we have

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

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):

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

Then, the marginal distribution (\rightarrow Definition I/1.3.3) of y is a gamma distribution (\rightarrow Definition II/3.3.1)

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

and the marginal distribution (\rightarrow Definition I/1.3.3) of x is a multivariate t-distribution (\rightarrow Definition “mvt”)

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

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 I/1.2.3), the marginal distribution of y can be derived as

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

which is the probability density function of the gamma distribution (\rightarrow Proof II/3.3.4) with shape parameter a and rate parameter b .

Using the law of marginal probability (\rightarrow Definition I/1.2.3), 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|}{(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|}{(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|}{(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|}{(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|}{(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|}{(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}$$

(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.

4.2.5 Conditional distributions

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

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

Then,

1) the conditional distribution (\rightarrow Definition I/1.3.4) of x given y is a multivariate normal distribution (\rightarrow Definition II/4.1.1)

$$x|y \sim \mathcal{N}(\mu, (y\Lambda)^{-1}) ; \quad (2)$$

2) the conditional distribution (\rightarrow Definition I/1.3.4) of a subset vector x_1 , given the complement vector x_2 and y , is also a multivariate normal distribution (\rightarrow Definition II/4.1.1)

$$x_1|x_2, y \sim \mathcal{N}(\mu_{1|2}(y), \Sigma_{1|2}(y)) \quad (3)$$

with the conditional mean (\rightarrow Definition I/1.5.1) and covariance (\rightarrow Definition I/1.7.1)

$$\begin{aligned} \mu_{1|2}(y) &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \\ \Sigma_{1|2}(y) &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12} \end{aligned} \quad (4)$$

where μ_1, μ_2 and $\Sigma_{11}, \Sigma_{12}, \Sigma_{22}, \Sigma_{21}$ are block-wise components (\rightarrow Proof II/4.1.7) of μ and $\Sigma(y) = (y\Lambda)^{-1}$;

3) the conditional distribution (\rightarrow Definition I/1.3.4) of y given x is a gamma distribution (\rightarrow Definition II/3.3.1)

$$y|x \sim \text{Gam}\left(a + \frac{n}{2}, b + \frac{1}{2}(x - \mu)^T \Lambda (x - \mu)\right) \quad (5)$$

where n is the dimensionality of x .

Proof:

1) This follows from the definition of the normal-gamma distribution (\rightarrow Definition II/4.2.1):

$$\begin{aligned} p(x, y) &= p(x|y) \cdot p(y) \\ &= \mathcal{N}(x; \mu, (y\Lambda)^{-1}) \cdot \text{Gam}(y; a, b) . \end{aligned} \quad (6)$$

2) This follows from (2) and the conditional distributions of the multivariate normal distribution (\rightarrow Proof II/4.1.7):

$$\begin{aligned} x &\sim \mathcal{N}(\mu, \Sigma) \\ \Rightarrow x_1|x_2 &\sim \mathcal{N}(\mu_{1|2}, \Sigma_{1|2}) \\ \mu_{1|2} &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \\ \Sigma_{1|2} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} . \end{aligned} \quad (7)$$

3) The conditional density of y given x follows from Bayes' theorem (\rightarrow Proof I/5.2.1) as

$$p(y|x) = \frac{p(x|y) \cdot p(y)}{p(x)}. \quad (8)$$

The conditional distribution (\rightarrow Definition I/1.3.4) of x given y is a multivariate normal distribution (\rightarrow Proof II/4.2.2)

$$p(x|y) = \mathcal{N}(x; \mu, (y\Lambda)^{-1}) = \sqrt{\frac{|y\Lambda|}{(2\pi)^n}} \exp \left[-\frac{1}{2}(x - \mu)^T (y\Lambda)(x - \mu) \right], \quad (9)$$

the marginal distribution (\rightarrow Definition I/1.3.3) of y is a gamma distribution (\rightarrow Proof II/4.2.4)

$$p(y) = \text{Gam}(y; a, b) = \frac{b^a}{\Gamma(a)} y^{a-1} \exp[-by] \quad (10)$$

and the marginal distribution (\rightarrow Definition I/1.3.3) of x is a multivariate t-distribution (\rightarrow Proof II/4.2.4)

$$\begin{aligned} p(x) &= t \left(x; \mu, \left(\frac{a}{b} \Lambda \right)^{-1}, 2a \right) \\ &= \sqrt{\frac{\left| \frac{a}{b} \Lambda \right|}{(2a\pi)^n}} \cdot \frac{\Gamma\left(\frac{2a+n}{2}\right)}{\Gamma\left(\frac{2a}{2}\right)} \cdot \left(1 + \frac{1}{2a}(x - \mu)^T \left(\frac{a}{b} \Lambda \right) (x - \mu) \right)^{-\frac{2a+n}{2}} \\ &= \sqrt{\frac{|\Lambda|}{(2\pi)^n}} \cdot \frac{\Gamma\left(a + \frac{n}{2}\right)}{\Gamma(a)} \cdot b^a \cdot \left(b + \frac{1}{2}(x - \mu)^T \Lambda (x - \mu) \right)^{-(a + \frac{n}{2})}. \end{aligned} \quad (11)$$

Plugging (9), (10) and (11) into (8), we obtain

$$\begin{aligned} p(y|x) &= \frac{\sqrt{\frac{|y\Lambda|}{(2\pi)^n}} \exp \left[-\frac{1}{2}(x - \mu)^T (y\Lambda)(x - \mu) \right] \cdot \frac{b^a}{\Gamma(a)} y^{a-1} \exp[-by]}{\sqrt{\frac{|\Lambda|}{(2\pi)^n}} \cdot \frac{\Gamma\left(a + \frac{n}{2}\right)}{\Gamma(a)} \cdot b^a \cdot \left(b + \frac{1}{2}(x - \mu)^T \Lambda (x - \mu) \right)^{-(a + \frac{n}{2})}} \\ &= y^{\frac{n}{2}} \cdot \exp \left[-\frac{1}{2}(x - \mu)^T (y\Lambda)(x - \mu) \right] \cdot y^{a-1} \cdot \exp[-by] \cdot \frac{1}{\Gamma\left(a + \frac{n}{2}\right)} \cdot \left(b + \frac{1}{2}(x - \mu)^T \Lambda (x - \mu) \right)^{a + \frac{n}{2}} \\ &= \frac{\left(b + \frac{1}{2}(x - \mu)^T \Lambda (x - \mu) \right)^{a + \frac{n}{2}}}{\Gamma\left(a + \frac{n}{2}\right)} \cdot y^{a + \frac{n}{2} - 1} \cdot \exp \left[- \left(b + \frac{1}{2}(x - \mu)^T \Lambda (x - \mu) \right) y \right] \end{aligned} \quad (12)$$

which is the probability density function of a gamma distribution (\rightarrow Proof II/3.3.4) with shape and rate parameters

$$a + \frac{n}{2} \quad \text{and} \quad b + \frac{1}{2}(x - \mu)^T \Lambda (x - \mu), \quad (13)$$

such that

$$p(y|x) = \text{Gam} \left(y; a + \frac{n}{2}, b + \frac{1}{2}(x - \mu)^T \Lambda (x - \mu) \right). \quad (14)$$

Sources:

- original work

Metadata: ID: P146 | shortcut: ng-cond | author: JoramSoch | date: 2020-08-05, 06:54.

4.3 Dirichlet distribution

4.3.1 Definition

****Definition**:** Let X be a random vector (\rightarrow Definition I/1.1.2). Then, X is said to follow a Dirichlet distribution with concentration parameters $\alpha = [\alpha_1, \dots, \alpha_k]$

$$X \sim \text{Dir}(\alpha) , \quad (1)$$

if and only if its probability density function (\rightarrow Definition I/1.4.2) is given by

$$\text{Dir}(x; \alpha) = \frac{\Gamma\left(\sum_{i=1}^k \alpha_i\right)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{i=1}^k x_i^{\alpha_i-1} \quad (2)$$

where $\alpha_i > 0$ for all $i = 1, \dots, k$, and the density is zero, if $x_i \notin [0, 1]$ for any $i = 1, \dots, k$ or $\sum_{i=1}^k x_i \neq 1$.

Sources:

- Wikipedia (2020): “Dirichlet distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-10; URL: https://en.wikipedia.org/wiki/Dirichlet_distribution#Probability_density_function.

Metadata: ID: D54 | shortcut: dir | author: JoramSoch | date: 2020-05-10, 20:36.

4.3.2 Probability density function

Theorem: Let X be a random vector (\rightarrow Definition I/1.1.2) following a Dirichlet distribution (\rightarrow Definition II/4.3.1):

$$X \sim \text{Dir}(\alpha) . \quad (1)$$

Then, the probability density function (\rightarrow Definition I/1.4.2) of X is

$$f_X(x) = \frac{\Gamma\left(\sum_{i=1}^k \alpha_i\right)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{i=1}^k x_i^{\alpha_i-1} . \quad (2)$$

Proof: This follows directly from the definition of the Dirichlet distribution (\rightarrow Definition II/4.3.1).

Sources:

- original work

Metadata: ID: P95 | shortcut: dir-pdf | author: JoramSoch | date: 2020-05-05, 21:22.

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 I/1.1.3). Then, X is said to be matrix-normally distributed with mean M , covariance (\rightarrow Definition I/1.7.5) across rows U and covariance (\rightarrow Definition I/1.7.5) across columns V

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

if and only if its probability density function (\rightarrow Definition I/1.4.2) 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 M 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 I/1.1.3) 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/1.4.2) of X is

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

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: 2020-03-02, 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/1.4.2) are identical, it is proven that the associated probability distributions (\rightarrow Definition I/1.3.1) 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.

5.1.4 Transposition

Theorem: Let X be a random matrix (\rightarrow Definition I/1.1.3) following a matrix-normal distribution (\rightarrow Definition II/5.1.1):

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

Then, the transpose of X also has a matrix-normal distribution:

$$X^T \sim \mathcal{MN}(M^T, V, U) . \quad (2)$$

Proof: The probability density function of the matrix-normal distribution (\rightarrow Proof II/5.1.2) is:

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

Define $Y = X^T$. Then, $X = Y^T$ and we can substitute:

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

Using $(A + B)^T = (A^T + B^T)$, we have:

$$f(Y) = \frac{1}{\sqrt{(2\pi)^{np}|V|^n|U|^p}} \cdot \exp \left[-\frac{1}{2} \text{tr} (V^{-1}(Y - M^T) U^{-1}(Y - M^T)^T) \right] . \quad (5)$$

Using $\text{tr}(ABC) = \text{tr}(CAB)$, we obtain

$$f(Y) = \frac{1}{\sqrt{(2\pi)^{np}|V|^n|U|^p}} \cdot \exp \left[-\frac{1}{2} \text{tr} (U^{-1}(Y - M^T)^T V^{-1}(Y - M^T)) \right] \quad (6)$$

which is the probability density function of a matrix-normal distribution (\rightarrow Proof II/5.1.2) with mean M^T , covariance across rows V and covariance across columns U .

Sources:

- original work

Metadata: ID: P144 | shortcut: matn-trans | author: JoramSoch | date: 2020-08-03, 22:21.

5.1.5 Linear transformation

Theorem: Let X be an $n \times p$ random matrix (\rightarrow Definition I/1.1.3) following a matrix-normal distribution (\rightarrow Definition II/5.1.1):

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

Then, a linear transformation of X is also matrix-normally distributed

$$Y = AXB + C \sim \mathcal{MN}(AMB + C, AUA^T, B^TVB) \quad (2)$$

where A is an $r \times n$ matrix of full rank $r \leq n$ and B is a $p \times s$ matrix of full rank $s \leq p$ and C is an $r \times s$ matrix.

Proof: The matrix-normal distribution is equivalent to the multivariate normal distribution (\rightarrow Proof II/5.1.3),

$$X \sim \mathcal{MN}(M, U, V) \Leftrightarrow \text{vec}(X) \sim \mathcal{N}(\text{vec}(M), V \otimes U) , \quad (3)$$

and the linear transformation theorem for the multivariate normal distribution (\rightarrow Proof II/4.1.5) states:

$$x \sim \mathcal{N}(\mu, \Sigma) \Rightarrow y = Ax + b \sim \mathcal{N}(A\mu + b, A\Sigma A^T) . \quad (4)$$

The vectorization of $Y = AXB + C$ is

$$\begin{aligned} \text{vec}(Y) &= \text{vec}(AXB + C) \\ &= \text{vec}(AXB) + \text{vec}(C) \\ &= (B^T \otimes A)\text{vec}(X) + \text{vec}(C) . \end{aligned} \quad (5)$$

Using (3) and (4), we have

$$\begin{aligned} \text{vec}(Y) &\sim \mathcal{N}((B^T \otimes A)\text{vec}(M) + \text{vec}(C), (B^T \otimes A)(V \otimes U)(B^T \otimes A)^T) \\ &= \mathcal{N}(\text{vec}(AMB) + \text{vec}(C), (B^TV \otimes AU)(B^T \otimes A)^T) \\ &= \mathcal{N}(\text{vec}(AMB + C), B^TVB \otimes AUA^T) . \end{aligned} \quad (6)$$

Using (3), we finally have:

$$Y \sim \mathcal{MN}(AMB + C, AUA^T, B^TVB) . \quad (7)$$

Sources:

- original work

Metadata: ID: P145 | shortcut: matn-ltt | author: JoramSoch | date: 2020-08-03, 22:24.

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 Univariate 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

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

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

$$\sum_{i=1}^n \hat{\varepsilon}_i x_{ij} = \hat{\varepsilon}^T x_j = 0. \quad (6)$$

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: 2020-03-09, 22:18.

1.1.8 Estimation matrix

Definition: In multiple linear regression (\rightarrow Definition III/1.1.1), the estimation matrix is the matrix E that results in ordinary least squares (\rightarrow Proof III/1.1.2) or weighted least squares (\rightarrow Proof III/1.1.13) parameter estimates when right-multiplied with the measured data:

$$Ey = \hat{\beta} . \quad (1)$$

Sources:

- original work

Metadata: ID: D81 | shortcut: emat | author: JoramSoch | date: 2020-07-22, 05:17.

1.1.9 Projection matrix

Definition: In multiple linear regression (\rightarrow Definition III/1.1.1), the projection matrix is the matrix P that results in the fitted signal explained by estimated parameters (\rightarrow Definition III/1.1.8) when right-multiplied with the measured data:

$$Py = \hat{y} = X\hat{\beta} . \quad (1)$$

Sources:

- Wikipedia (2020): “Projection matrix”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-07-22; URL: https://en.wikipedia.org/wiki/Projection_matrix#Overview.

Metadata: ID: D82 | shortcut: pmat | author: JoramSoch | date: 2020-07-22, 05:25.

1.1.10 Residual-forming matrix

Definition: In multiple linear regression (\rightarrow Definition III/1.1.1), the residual-forming matrix is the matrix R that results in the vector of residuals left over by estimated parameters (\rightarrow Definition III/1.1.8) when right-multiplied with the measured data:

$$Ry = \hat{\varepsilon} = y - \hat{y} . \quad (1)$$

Sources:

- Wikipedia (2020): “Projection matrix”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-07-22; URL: https://en.wikipedia.org/wiki/Projection_matrix#Properties.

Metadata: ID: D83 | shortcut: rformat | author: JoramSoch | date: 2020-07-22, 05:35.

1.1.11 Estimation, projection and residual-forming matrix

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

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

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 (\rightarrow Definition III/1.1.8), projection matrix (\rightarrow Definition III/1.1.9) and residual-forming matrix (\rightarrow Definition III/1.1.10) 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

$$\begin{aligned}\hat{y} &= X(X^T X)^{-1} X^T y \\ &\stackrel{(3)}{=} Py .\end{aligned}\tag{7}$$

3) The residuals of the model are calculated by subtracting the fitted signal from the measured signal

$$\hat{\varepsilon} = y - \hat{y}\tag{8}$$

and using (7), this becomes

$$\begin{aligned}\hat{\varepsilon} &= y - X(X^T X)^{-1} X^T y \\ &= (I_n - X(X^T X)^{-1} X^T) y \\ &\stackrel{(3)}{=} Ry .\end{aligned}\tag{9}$$

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.

1.1.12 Idempotence of projection and residual-forming matrix

Theorem: The projection matrix (\rightarrow Definition III/1.1.9) and the residual-forming matrix (\rightarrow Definition III/1.1.10) are idempotent:

$$\begin{aligned}P^2 &= P \\ R^2 &= R .\end{aligned}\tag{1}$$

Proof:

1) The projection matrix for ordinary least squares is given by (\rightarrow Proof III/1.1.11)

$$P = X(X^T X)^{-1} X^T ,\tag{2}$$

such that

$$\begin{aligned}P^2 &= X(X^T X)^{-1} X^T X(X^T X)^{-1} X^T \\ &= X(X^T X)^{-1} X^T \\ &\stackrel{(2)}{=} P .\end{aligned}\tag{3}$$

2) The residual-forming matrix for ordinary least squares is given by (\rightarrow Proof III/1.1.11)

$$R = I_n - X(X^T X)^{-1} X^T = I_n - P, \quad (4)$$

such that

$$\begin{aligned} R^2 &= (I_n - P)(I_n - P) \\ &= I_n - P - P + P^2 \\ &\stackrel{(3)}{=} I_n - 2P + P \\ &= I_n - P \\ &\stackrel{(4)}{=} R. \end{aligned} \quad (5)$$

Sources:

- Wikipedia (2020): “Projection matrix”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-07-22; URL: https://en.wikipedia.org/wiki/Projection_matrix#Properties.

Metadata: ID: P135 | shortcut: mlr-idem | author: JoramSoch | date: 2020-07-22, 06:28.

1.1.13 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

$$WVW^T = I_n. \quad (3)$$

Since V is a covariance matrix and thus symmetric, W is also symmetric and can be expressed as the matrix square root of the inverse of V :

$$WVW = I_n \quad \Leftrightarrow \quad V = W^{-1}W^{-1} \quad \Leftrightarrow \quad V^{-1} = WW \quad \Leftrightarrow \quad W = V^{-1/2}. \quad (4)$$

Left-multiplying the linear regression equation (1) with W , the linear transformation theorem (\rightarrow Proof II/4.1.5) implies that

$$Wy = WX\beta + W\varepsilon, \quad W\varepsilon \sim \mathcal{N}(0, \sigma^2 WVW^T). \quad (5)$$

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

$$\begin{aligned}\hat{\beta} &= (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \tilde{y} \\ &= ((WX)^T WX)^{-1} (WX)^T Wy \\ &= (X^T W^T WX)^{-1} X^T W^T Wy \\ &= (X^T WWX)^{-1} X^T WWy \\ &\stackrel{(4)}{=} (X^T V^{-1} X)^{-1} X^T V^{-1} y\end{aligned}\tag{7}$$

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.

1.1.14 Weighted least squares

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

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

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

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

Proof: Let there be an $n \times n$ square matrix W , such that

$$WVW^T = I_n. \tag{3}$$

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 :

$$WVW = I_n \quad \Leftrightarrow \quad V = W^{-1}W^{-1} \quad \Leftrightarrow \quad V^{-1} = WW \quad \Leftrightarrow \quad W = V^{-1/2}. \tag{4}$$

Left-multiplying the linear regression equation (1) with W , the linear transformation theorem (\rightarrow Proof II/4.1.5) implies that

$$Wy = WX\beta + W\varepsilon, \quad W\varepsilon \sim \mathcal{N}(0, \sigma^2 WVW^T). \tag{5}$$

Applying (3), we see that (5) is actually a linear regression model (\rightarrow Definition III/1.1.1) with independent observations

$$Wy = WX\beta + W\varepsilon, \quad W\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n). \tag{6}$$

With this, we can express the weighted residual sum of squares (\rightarrow Definition III/1.1.6) as

$$\text{wRSS}(\beta) = \sum_{i=1}^n (W\varepsilon)_i = (W\varepsilon)^T(W\varepsilon) = (Wy - WX\beta)^T(Wy - WX\beta) \quad (7)$$

which can be developed into

$$\begin{aligned} \text{wRSS}(\beta) &= y^T W^T W y - y^T W^T W X \beta - \beta^T X^T W^T W y + \beta^T X^T W^T W X \beta \\ &= y^T W W y - 2\beta^T X^T W W y + \beta^T X^T W W X \beta \\ &\stackrel{(4)}{=} y^T V^{-1} y - 2\beta^T X^T V^{-1} y + \beta^T X^T V^{-1} X \beta . \end{aligned} \quad (8)$$

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

$$\frac{d\text{wRSS}(\beta)}{d\beta} = -2X^T V^{-1} y + 2X^T V^{-1} X \beta \quad (9)$$

and setting this derivative to zero, we obtain:

$$\begin{aligned} \frac{d\text{wRSS}(\hat{\beta})}{d\beta} &= 0 \\ 0 &= -2X^T V^{-1} y + 2X^T V^{-1} X \hat{\beta} \\ X^T V^{-1} X \hat{\beta} &= X^T V^{-1} y \\ \hat{\beta} &= (X^T V^{-1} X)^{-1} X^T V^{-1} y . \end{aligned} \quad (10)$$

Since the quadratic form $y^T V^{-1} y$ in (8) is positive, $\hat{\beta}$ minimizes $\text{wRSS}(\beta)$.

Sources:

- original work

Metadata: ID: P136 | shortcut: mlr-wls2 | author: JoramSoch | date: 2020-07-22, 06:48.

1.1.15 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 I/4.1.3) 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/5.1.2)

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

and, using $|\sigma^2 V| = (\sigma^2)^n |V|$, the log-likelihood function (\rightarrow Definition I/4.1.2)

$$\begin{aligned}
\text{LL}(\beta, \sigma^2) &= \log p(y|\beta, \sigma^2) \\
&= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2} \log |V| \\
&\quad - \frac{1}{2\sigma^2} (y - X\beta)^T V^{-1} (y - X\beta) .
\end{aligned} \tag{4}$$

Substituting the precision matrix $P = V^{-1}$ into (4) to ease notation, we have:

$$\begin{aligned}
\text{LL}(\beta, \sigma^2) &= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2} \log(|V|) \\
&\quad - \frac{1}{2\sigma^2} (y^T P y - 2\beta^T X^T P y + \beta^T X^T P X \beta) .
\end{aligned} \tag{5}$$

The derivative of the log-likelihood function (5) with respect to β is

$$\begin{aligned}
\frac{d\text{LL}(\beta, \sigma^2)}{d\beta} &= \frac{d}{d\beta} \left(-\frac{1}{2\sigma^2} (y^T P y - 2\beta^T X^T P y + \beta^T X^T P X \beta) \right) \\
&= \frac{1}{2\sigma^2} \frac{d}{d\beta} (2\beta^T X^T P y - \beta^T X^T P X \beta) \\
&= \frac{1}{2\sigma^2} (2X^T P y - 2X^T P X \beta) \\
&= \frac{1}{\sigma^2} (X^T P y - X^T P X \beta)
\end{aligned} \tag{6}$$

and setting this derivative to zero gives the MLE for β :

$$\begin{aligned}
\frac{d\text{LL}(\hat{\beta}, \sigma^2)}{d\beta} &= 0 \\
0 &= \frac{1}{\sigma^2} (X^T P y - X^T P X \hat{\beta}) \\
0 &= X^T P y - X^T P X \hat{\beta} \\
X^T P X \hat{\beta} &= X^T P y \\
\hat{\beta} &= (X^T P X)^{-1} X^T P y
\end{aligned} \tag{7}$$

The derivative of the log-likelihood function (4) at $\hat{\beta}$ with respect to σ^2 is

$$\begin{aligned}
\frac{dLL(\hat{\beta}, \sigma^2)}{d\sigma^2} &= \frac{d}{d\sigma^2} \left(-\frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} (y - X\hat{\beta})^T V^{-1} (y - X\hat{\beta}) \right) \\
&= -\frac{n}{2} \frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2} (y - X\hat{\beta})^T V^{-1} (y - X\hat{\beta}) \\
&= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} (y - X\hat{\beta})^T V^{-1} (y - X\hat{\beta})
\end{aligned} \tag{8}$$

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 as well as unknown $p \times 1$ regression coefficients β and unknown 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/5.1.3) that, when combined with the likelihood function (\rightarrow Definition I/5.1.2), leads to a

posterior distribution (\rightarrow Definition I/5.1.7) that belongs to the same family of probability distributions (\rightarrow Definition I/1.3.1). 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/5.1.2)

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

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/5.1.2) 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 (\rightarrow Definition II/4.2.1) 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.

1.2.2 Posterior distribution

Theorem: Let

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

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 as well as unknown $p \times 1$ regression coefficients β and unknown 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 posterior distribution (\rightarrow Definition I/5.1.7) is also a normal-gamma distribution (\rightarrow Definition II/4.2.1)

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

and the posterior hyperparameters (\rightarrow Definition I/5.1.7) 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 Bayes' theorem (\rightarrow Proof I/5.2.1), the posterior distribution (\rightarrow Definition I/5.1.7) is given by

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

Since $p(y)$ is just a normalization factor, the posterior is proportional (\rightarrow Proof I/5.1.8) to the numerator:

$$p(\beta, \tau | y) \propto p(y | \beta, \tau) p(\beta, \tau) = p(y, \beta, \tau) . \quad (6)$$

Equation (1) implies the following likelihood function (\rightarrow Definition I/5.1.2)

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

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 (\rightarrow Definition I/1.7.8) $P = V^{-1}$.

Combining the likelihood function (\rightarrow Definition I/5.1.2) (8) with the prior distribution (\rightarrow Definition I/5.1.3) (2), the joint likelihood (\rightarrow Definition I/5.1.5) of the model is given by

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

Collecting identical variables gives:

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

Expanding the products in the exponent gives:

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

Completing the square over β , we finally have

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

with the posterior hyperparameters (\rightarrow Definition I/5.1.7)

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

Ergo, the joint likelihood is proportional to

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

with the posterior hyperparameters (\rightarrow Definition I/5.1.7)

$$\begin{aligned} a_n &= a_0 + \frac{n}{2} \\ b_n &= b_0 + \frac{1}{2} (y^T P y + \mu_0^T \Lambda_0 \mu_0 - \mu_n^T \Lambda_n \mu_n) . \end{aligned} \quad (15)$$

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 I/1.2.2) posterior distribution (\rightarrow Definition I/5.1.7) 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 as well as unknown $p \times 1$ regression coefficients β and unknown 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

$$\begin{aligned} \log p(y|m) &= \frac{1}{2} \log |P| - \frac{n}{2} \log(2\pi) + \frac{1}{2} \log |\Lambda_0| - \frac{1}{2} \log |\Lambda_n| + \\ &\quad \log \Gamma(a_n) - \log \Gamma(a_0) + a_0 \log b_0 - a_n \log b_n \end{aligned} \quad (3)$$

where the posterior hyperparameters (\rightarrow Definition I/5.1.7) are given by

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

Proof: According to the law of marginal probability (\rightarrow Definition I/1.2.3), the model evidence (\rightarrow Definition I/5.1.9) for this model is:

$$p(y|m) = \iint p(y|\beta, \tau) p(\beta, \tau) d\beta d\tau . \tag{5}$$

According to the law of conditional probability (\rightarrow Definition I/1.2.4), the integrand is equivalent to the joint likelihood (\rightarrow Definition I/5.1.5):

$$p(y|m) = \iint p(y, \beta, \tau) d\beta d\tau . \tag{6}$$

Equation (1) implies the following likelihood function (\rightarrow Definition I/5.1.2)

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

which, for mathematical convenience, can also be parametrized as

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

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

When deriving the posterior distribution (\rightarrow Proof III/1.2.2) $p(\beta, \tau|y)$, the joint likelihood $p(y, \beta, \tau)$ is obtained as

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

Using the probability density function of the multivariate normal distribution (\rightarrow Proof II/4.1.2), we can rewrite this as

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

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/3.3.4), 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

$$\log p(y|m) = \frac{1}{2} \log |P| - \frac{n}{2} \log(2\pi) + \frac{1}{2} \log |\Lambda_0| - \frac{1}{2} \log |\Lambda_n| + \log \Gamma(a_n) - \log \Gamma(a_0) + a_0 \log b_0 - a_n \log b_n. \quad (14)$$

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.

1.2.4 Posterior probability of alternative hypothesis

Theorem: Let there be a linear regression model (\rightarrow Definition III/1.1.1) with normally distributed (\rightarrow Definition II/4.1.1) errors:

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

and assume a normal-gamma (\rightarrow Definition II/4.2.1) prior distribution (\rightarrow Definition I/5.1.3) 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 posterior (\rightarrow Definition I/5.1.7) probability (\rightarrow Definition I/1.2.1) of the alternative hypothesis (\rightarrow Definition “h1”)

$$H_1 : c^T \beta > 0 \quad (3)$$

is given by

$$\Pr(H_1|y) = 1 - T\left(-\frac{c^T \mu}{\sqrt{c^T \Sigma c}}; \nu\right) \quad (4)$$

where c is a $p \times 1$ contrast vector (\rightarrow Definition “con”), $T(x; \nu)$ is the cumulative distribution function (\rightarrow Definition I/1.4.3) of the t-distribution (\rightarrow Definition “t”) with ν degrees of freedom (\rightarrow Definition “dof”) and μ , Σ and ν can be obtained from the posterior hyperparameters (\rightarrow Definition I/5.1.7) of Bayesian linear regression.

Proof: The posterior distribution for Bayesian linear regression (\rightarrow Proof III/1.2.2) is given by a normal-gamma distribution (\rightarrow Definition II/4.2.1) over β and $\tau = 1/\sigma^2$

$$p(\beta, \tau|y) = \mathcal{N}(\beta; \mu_n, (\tau \Lambda_n)^{-1}) \cdot \text{Gam}(\tau; a_n, b_n) \quad (5)$$

with the posterior hyperparameters (\rightarrow Definition I/5.1.7)

$$\begin{aligned} \mu_n &= \Lambda_n^{-1}(X^T P y + \Lambda_0 \mu_0) \\ \Lambda_n &= X^T P X + \Lambda_0 \\ a_n &= a_0 + \frac{n}{2} \\ b_n &= b_0 + \frac{1}{2}(y^T P y + \mu_0^T \Lambda_0 \mu_0 - \mu_n^T \Lambda_n \mu_n) . \end{aligned} \quad (6)$$

The marginal distribution of a normal-gamma distribution is a multivariate t-distribution (\rightarrow Proof II/4.2.4), such that the marginal (\rightarrow Definition I/1.3.3) posterior (\rightarrow Definition I/5.1.7) distribution of β is

$$p(\beta|y) = t(\beta; \mu, \Sigma, \nu) \quad (7)$$

with the posterior hyperparameters (\rightarrow Definition I/5.1.7)

$$\begin{aligned} \mu &= \mu_n \\ \Sigma &= \left(\frac{a_n}{b_n} \Lambda_n\right)^{-1} \\ \nu &= 2 a_n . \end{aligned} \quad (8)$$

Define the quantity $\gamma = c^T \beta$. According to the linear transformation theorem for the multivariate t-distribution (\rightarrow Proof “mvt-ltt”), γ also follows a multivariate t-distribution (\rightarrow Definition “mvt”):

$$p(\gamma|y) = t(\gamma; c^T \mu, c^T \Sigma c, \nu) . \quad (9)$$

Because c^T is a $1 \times p$ vector, γ is a scalar and actually has a non-central scaled t-distribution (\rightarrow Definition “ncst”). Therefore, the posterior probability of H_1 can be calculated using a one-dimensional integral:

$$\begin{aligned} \Pr(H_1|y) &= p(\gamma > 0|y) \\ &= \int_0^{+\infty} p(\gamma|y) d\gamma \\ &= 1 - \int_{-\infty}^0 p(\gamma|y) d\gamma \\ &= 1 - T_{\text{ncst}}(0; c^T \mu, c^T \Sigma c, \nu) . \end{aligned} \quad (10)$$

Using the relation between non-central scaled t-distribution and standard t-distribution (\rightarrow Proof “ncst-t”), we can finally write:

$$\begin{aligned}\Pr(H_1|y) &= 1 - T\left(\frac{(0 - c^T\mu)}{\sqrt{c^T\Sigma c}}; \nu\right) \\ &= 1 - T\left(-\frac{c^T\mu}{\sqrt{c^T\Sigma c}}; \nu\right).\end{aligned}\tag{11}$$

Sources:

- Koch, Karl-Rudolf (2007): “Multivariate t-distribution”; in: *Introduction to Bayesian Statistics*, Springer, Berlin/Heidelberg, 2007, eqs. 2.235, 2.236, 2.213, 2.210, 2.188; URL: <https://www.springer.com/de/book/9783540727231>; DOI: 10.1007/978-3-540-72726-2.

Metadata: ID: P133 | shortcut: blr-pp | author: JoramSoch | date: 2020-07-17, 17:03.

1.2.5 Posterior credibility region excluding null hypothesis

Theorem: Let there be a linear regression model (\rightarrow Definition III/1.1.1) with normally distributed (\rightarrow Definition II/4.1.1) errors:

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

and assume a normal-gamma (\rightarrow Definition II/4.2.1) prior distribution (\rightarrow Definition I/5.1.3) over the model parameters β and $\tau = 1/\sigma^2$:

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

Then, the largest posterior (\rightarrow Definition I/5.1.7) credibility region (\rightarrow Definition “cr”) that does not contain the omnibus null hypothesis (\rightarrow Definition “h0”)

$$H_0 : C^T\beta = 0\tag{3}$$

is given by the credibility level (\rightarrow Definition “cr”)

$$(1 - \alpha) = F\left([\mu^T C (C^T \Sigma C)^{-1} C^T \mu] / q; q, \nu\right)\tag{4}$$

where C is a $p \times q$ contrast matrix (\rightarrow Definition “con”), $F(x; v, w)$ is the cumulative distribution function (\rightarrow Definition I/1.4.3) of the F-distribution (\rightarrow Definition “f”) with v numerator degrees of freedom (\rightarrow Definition “dof”) w denominator degrees of freedom (\rightarrow Definition “dof”) and μ , Σ and ν can be obtained from the posterior hyperparameters (\rightarrow Definition I/5.1.7) of Bayesian linear regression.

Proof: The posterior distribution for Bayesian linear regression (\rightarrow Proof III/1.2.2) is given by a normal-gamma distribution (\rightarrow Definition II/4.2.1) over β and $\tau = 1/\sigma^2$

$$p(\beta, \tau|y) = \mathcal{N}(\beta; \mu_n, (\tau\Lambda_n)^{-1}) \cdot \text{Gam}(\tau; a_n, b_n)\tag{5}$$

with the posterior hyperparameters (\rightarrow Definition I/5.1.7)

$$\begin{aligned}
\mu_n &= \Lambda_n^{-1}(X^T P y + \Lambda_0 \mu_0) \\
\Lambda_n &= X^T P X + \Lambda_0 \\
a_n &= a_0 + \frac{n}{2} \\
b_n &= b_0 + \frac{1}{2}(y^T P y + \mu_0^T \Lambda_0 \mu_0 - \mu_n^T \Lambda_n \mu_n) .
\end{aligned} \tag{6}$$

The marginal distribution of a normal-gamma distribution is a multivariate t-distribution (\rightarrow Proof II/4.2.4), such that the marginal (\rightarrow Definition I/1.3.3) posterior (\rightarrow Definition I/5.1.7) distribution of β is

$$p(\beta|y) = t(\beta; \mu, \Sigma, \nu) \tag{7}$$

with the posterior hyperparameters (\rightarrow Definition I/5.1.7)

$$\begin{aligned}
\mu &= \mu_n \\
\Sigma &= \left(\frac{a_n}{b_n} \Lambda_n \right)^{-1} \\
\nu &= 2 a_n .
\end{aligned} \tag{8}$$

Define the quantity $\gamma = C^T \beta$. According to the linear transformation theorem for the multivariate t-distribution (\rightarrow Proof “mvt-ltt”), γ also follows a multivariate t-distribution (\rightarrow Definition “mvt”):

$$p(\gamma|y) = t(\gamma; C^T \mu, C^T \Sigma C, \nu) . \tag{9}$$

Because C^T is a $q \times p$ matrix, γ is a $q \times 1$ vector. The quadratic form of a multivariate t-distributed random variable has an F-distribution (\rightarrow Proof “mvt-f”), such that we can write:

$$QF(\gamma) = (\gamma - C^T \mu)^T (C^T \Sigma C)^{-1} (\gamma - C^T \mu) / q \sim F(q, \nu) . \tag{10}$$

Therefore, the largest posterior credibility region for γ which does not contain $\gamma = 0_q$ (i.e. only touches this origin point) can be obtained by plugging $QF(0)$ into the cumulative distribution function of the F-distribution:

$$\begin{aligned}
(1 - \alpha) &= F(QF(0); q, \nu) \\
&= F([\mu^T C (C^T \Sigma C)^{-1} C^T \mu] / q; q, \nu) .
\end{aligned} \tag{11}$$

Sources:

- Koch, Karl-Rudolf (2007): “Multivariate t-distribution”; in: *Introduction to Bayesian Statistics*, Springer, Berlin/Heidelberg, 2007, eqs. 2.235, 2.236, 2.213, 2.210, 2.211, 2.183; URL: <https://www.springer.com/de/book/9783540727231>; DOI: 10.1007/978-3-540-72726-2.

Metadata: ID: P134 | shortcut: blr-pcr | author: JoramSoch | date: 2020-07-17, 17:41.

2 Multivariate normal data

2.1 General linear model

2.1.1 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.

2.1.2 Ordinary least squares

Theorem: Given a general linear model (\rightarrow Definition III/2.1.1) with independent observations

$$Y = XB + E, \quad E \sim \mathcal{MN}(0, \sigma^2 I_n, \Sigma) , \quad (1)$$

the ordinary least squares (\rightarrow Definition “ols”) parameters estimates are given by

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

Proof: Let \hat{B} be the ordinary least squares (\rightarrow Definition “ols”) (OLS) solution and let $\hat{E} = Y - X\hat{B}$ be the resulting matrix of residuals. According to the exogeneity assumption of OLS, the errors have conditional mean (\rightarrow Definition I/1.5.1) zero

$$E(E|X) = 0 , \quad (3)$$

a direct consequence of which is that the regressors are uncorrelated with the errors

$$E(X^T E) = 0 , \quad (4)$$

which, in the finite sample, means that the residual matrix must be orthogonal to the design matrix:

$$X^T \hat{E} = 0 . \quad (5)$$

From (5), the OLS formula can be directly derived:

$$\begin{aligned} X^T \hat{E} &= 0 \\ X^T (Y - X\hat{B}) &= 0 \\ X^T Y - X^T X\hat{B} &= 0 \\ X^T X\hat{B} &= X^T Y \\ \hat{B} &= (X^T X)^{-1} X^T Y . \end{aligned} \quad (6)$$

Sources:

- original work

Metadata: ID: P106 | shortcut: glm-ols | author: JoramSoch | date: 2020-05-19, 06:02.

2.1.3 Weighted least squares

Theorem: Given a general linear model (\rightarrow Definition III/2.1.1) with correlated observations

$$Y = XB + E, \quad E \sim \mathcal{MN}(0, V, \Sigma) , \quad (1)$$

the weighted least squares (\rightarrow Definition “wls”) parameter estimates are given by

$$\hat{B} = (X^T V^{-1} X)^{-1} X^T V^{-1} Y . \quad (2)$$

Proof: Let there be an $n \times n$ square matrix W , such that

$$WVW^T = I_n . \quad (3)$$

Since V is a covariance matrix and thus symmetric, W is also symmetric and can be expressed as the matrix square root of the inverse of V :

$$WW = V^{-1} \quad \Leftrightarrow \quad W = V^{-1/2} . \quad (4)$$

Left-multiplying the linear regression equation (1) with W , the linear transformation theorem (\rightarrow Proof II/5.1.5) implies that

$$WY = WXB + WE, \quad WE \sim \mathcal{MN}(0, WVW^T, \Sigma) . \quad (5)$$

Applying (3), we see that (5) is actually a general linear model (\rightarrow Definition III/2.1.1) with independent observations

$$\tilde{Y} = \tilde{X}B + \tilde{E}, \quad \tilde{E} \sim \mathcal{N}(0, I_n, \Sigma) \quad (6)$$

where $\tilde{Y} = WY$, $\tilde{X} = WX$ and $\tilde{E} = WE$, such that we can apply the ordinary least squares solution (\rightarrow Proof III/2.1.2) giving

$$\begin{aligned} \hat{B} &= (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \tilde{Y} \\ &= ((WX)^T WX)^{-1} (WX)^T WY \\ &= (X^T W^T W X)^{-1} X^T W^T WY \\ &= (X^T W W X)^{-1} X^T W WY \\ &\stackrel{(4)}{=} (X^T V^{-1} X)^{-1} X^T V^{-1} Y \end{aligned} \quad (7)$$

which corresponds to the weighted least squares solution (2).

Sources:

- original work

Metadata: ID: P107 | shortcut: glm-wls | author: JoramSoch | date: 2020-05-19, 06:27.

2.1.4 Maximum likelihood estimation

Theorem: Given a general linear model (\rightarrow Definition III/2.1.1) 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 I/4.1.3) 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/5.1.2)

$$\begin{aligned} p(Y|B, \Sigma) &= \mathcal{MN}(Y; XB, V, \Sigma) \\ &= \sqrt{\frac{1}{(2\pi)^{nv} |\Sigma|^n |V|^v}} \cdot \exp \left[-\frac{1}{2} \text{tr} (\Sigma^{-1} (Y - XB)^T V^{-1} (Y - XB)) \right] \end{aligned} \quad (3)$$

and the log-likelihood function (\rightarrow Definition I/4.1.2)

$$\begin{aligned}
\text{LL}(B, \Sigma) &= \log p(Y|B, \Sigma) \\
&= -\frac{nv}{2} \log(2\pi) - \frac{n}{2} \log |\Sigma| - \frac{v}{2} \log |V| \\
&\quad - \frac{1}{2} \text{tr} [\Sigma^{-1}(Y - XB)^T V^{-1}(Y - XB)] .
\end{aligned} \tag{4}$$

Substituting V^{-1} by the precision matrix P to ease notation, we have:

$$\begin{aligned}
\text{LL}(B, \Sigma) &= -\frac{nv}{2} \log(2\pi) - \frac{n}{2} \log(|\Sigma|) - \frac{v}{2} \log(|V|) \\
&\quad - \frac{1}{2} \text{tr} [\Sigma^{-1} (Y^T P Y - Y^T P X B - B^T X^T P Y + B^T X^T P X B)] .
\end{aligned} \tag{5}$$

The derivative of the log-likelihood function (5) with respect to B is

$$\begin{aligned}
\frac{d\text{LL}(B, \Sigma)}{dB} &= \frac{d}{dB} \left(-\frac{1}{2} \text{tr} [\Sigma^{-1} (Y^T P Y - Y^T P X B - B^T X^T P Y + B^T X^T P X B)] \right) \\
&= \frac{d}{dB} \left(-\frac{1}{2} \text{tr} [-2\Sigma^{-1} Y^T P X B] \right) + \frac{d}{dB} \left(-\frac{1}{2} \text{tr} [\Sigma^{-1} B^T X^T P X B] \right) \\
&= -\frac{1}{2} (-2X^T P Y \Sigma^{-1}) - \frac{1}{2} (X^T P X B \Sigma^{-1} + (X^T P X)^T B (\Sigma^{-1})^T) \\
&= X^T P Y \Sigma^{-1} - X^T P X B \Sigma^{-1}
\end{aligned} \tag{6}$$

and setting this derivative to zero gives the MLE for B :

$$\begin{aligned}
\frac{d\text{LL}(\hat{B}, \Sigma)}{dB} &= 0 \\
0 &= X^T P Y \Sigma^{-1} - X^T P X \hat{B} \Sigma^{-1} \\
0 &= X^T P Y - X^T P X \hat{B} \\
X^T P X \hat{B} &= X^T P Y \\
\hat{B} &= (X^T P X)^{-1} X^T P Y
\end{aligned} \tag{7}$$

The derivative of the log-likelihood function (4) at \hat{B} with respect to Σ is

$$\begin{aligned}
\frac{d\text{LL}(\hat{B}, \Sigma)}{d\Sigma} &= \frac{d}{d\Sigma} \left(-\frac{n}{2} \log |\Sigma| - \frac{1}{2} \text{tr} [\Sigma^{-1} (Y - X\hat{B})^T V^{-1} (Y - X\hat{B})] \right) \\
&= -\frac{n}{2} (\Sigma^{-1})^T + \frac{1}{2} \left(\Sigma^{-1} (Y - X\hat{B})^T V^{-1} (Y - X\hat{B}) \Sigma^{-1} \right)^T \\
&= -\frac{n}{2} \Sigma^{-1} + \frac{1}{2} \Sigma^{-1} (Y - X\hat{B})^T V^{-1} (Y - X\hat{B}) \Sigma^{-1}
\end{aligned} \tag{8}$$

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.

2.2 Multivariate Bayesian linear regression

2.2.1 Conjugate prior distribution

Theorem: Let

$$Y = XB + E, \quad E \sim \mathcal{MN}(0, V, \Sigma) \tag{1}$$

be a general linear model (\rightarrow Definition III/2.1.1) with measured $n \times v$ data matrix Y , known $n \times p$ design matrix X , known $n \times n$ covariance structure (\rightarrow Definition II/5.1.1) V as well as unknown $p \times v$ regression coefficients B and unknown $v \times v$ noise covariance (\rightarrow Definition II/5.1.1) Σ .

Then, the conjugate prior (\rightarrow Definition “prior-conj”) for this model is a normal-Wishart distribution (\rightarrow Definition “nw”)

$$p(B, T) = \mathcal{MN}(B; M_0, \Lambda_0^{-1}, T^{-1}) \cdot \mathcal{W}(T; \Omega_0^{-1}, \nu_0) \tag{2}$$

where $T = \Sigma^{-1}$ is the inverse noise covariance (\rightarrow Definition I/1.7.5) or noise precision matrix (\rightarrow Definition I/1.7.8).

Proof: By definition, a conjugate prior (\rightarrow Definition “prior-conj”) is a prior distribution (\rightarrow Definition I/5.1.3) that, when combined with the likelihood function (\rightarrow Definition I/5.1.2), leads to a posterior distribution (\rightarrow Definition I/5.1.7) that belongs to the same family of probability distributions (\rightarrow Definition I/1.3.1). 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/5.1.2)

$$p(Y|B, \Sigma) = \mathcal{MN}(Y; XB, V, \Sigma) = \sqrt{\frac{1}{(2\pi)^{nv} |\Sigma|^n |V|^v}} \exp \left[-\frac{1}{2} \text{tr} (\Sigma^{-1} (Y - XB)^T V^{-1} (Y - XB)) \right] \quad (3)$$

which, for mathematical convenience, can also be parametrized as

$$p(Y|B, T) = \mathcal{MN}(Y; XB, P, T^{-1}) = \sqrt{\frac{|T|^n |P|^v}{(2\pi)^{nv}}} \exp \left[-\frac{1}{2} \text{tr} (T(Y - XB)^T P(Y - XB)) \right] \quad (4)$$

using the $v \times v$ precision matrix (\rightarrow Definition I/1.7.8) $T = \Sigma^{-1}$ and the $n \times n$ precision matrix (\rightarrow Definition I/1.7.8) $P = V^{-1}$.

Separating constant and variable terms, we have:

$$p(Y|B, T) = \sqrt{\frac{|P|^v}{(2\pi)^{nv}}} \cdot |T|^{n/2} \cdot \exp \left[-\frac{1}{2} \text{tr} (T(Y - XB)^T P(Y - XB)) \right]. \quad (5)$$

Expanding the product in the exponent, we have:

$$p(Y|B, T) = \sqrt{\frac{|P|^v}{(2\pi)^{nv}}} \cdot |T|^{n/2} \cdot \exp \left[-\frac{1}{2} \text{tr} (T [Y^T P Y - Y^T P X B - B^T X^T P Y + B^T X^T P X B]) \right]. \quad (6)$$

Completing the square over β , finally gives

$$p(Y|B, T) = \sqrt{\frac{|P|^v}{(2\pi)^{nv}}} \cdot |T|^{n/2} \cdot \exp \left[-\frac{1}{2} \text{tr} \left(T \left[(B - \tilde{X} Y)^T X^T P X (B - \tilde{X} Y) - Y^T Q Y + Y^T P Y \right] \right) \right] \quad (7)$$

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/5.1.2) is proportional to a power of the determinant of T , times an exponential of the trace of T and an exponential of the trace of a squared form of B , weighted by T :

$$p(Y|B, T) \propto |T|^{n/2} \cdot \exp \left[-\frac{1}{2} \text{tr} (T [Y^T P Y - Y^T Q Y]) \right] \cdot \exp \left[-\frac{1}{2} \text{tr} \left(T \left[(B - \tilde{X} Y)^T X^T P X (B - \tilde{X} Y) \right] \right) \right]. \quad (8)$$

The same is true for a normal-Wishart distribution (\rightarrow Definition “nw”) over B and T

$$p(B, T) = \mathcal{MN}(B; M_0, \Lambda_0^{-1}, T^{-1}) \cdot \mathcal{W}(T; \Omega_0^{-1}, \nu_0) \quad (9)$$

the probability density function of which (\rightarrow Proof “nw-pdf”)

$$p(B, T) = \sqrt{\frac{|T|^p |\Lambda_0|^v}{(2\pi)^{pv}}} \exp \left[-\frac{1}{2} \text{tr} (T(B - M_0)^T \Lambda_0 (B - M_0)) \right] \cdot \frac{1}{\Gamma_v \left(\frac{\nu_0}{2} \right)} \sqrt{\frac{|\Omega_0|^{\nu_0}}{2^{\nu_0 v}}} |T|^{(\nu_0 - v - 1)/2} \exp \left[-\frac{1}{2} \text{tr} (\Omega_0 T) \right] \quad (10)$$

exhibits the same proportionality

$$p(B, T) \propto |T|^{(\nu_0 + p - v - 1)/2} \cdot \exp \left[-\frac{1}{2} \text{tr}(T \Omega_0) \right] \cdot \exp \left[-\frac{1}{2} \text{tr} \left(T [(B - M_0)^T \Lambda_0 (B - M_0)] \right) \right] \quad (11)$$

and is therefore conjugate relative to the likelihood.

Sources:

- Wikipedia (2020): “Bayesian multivariate linear regression”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-09-03; URL: https://en.wikipedia.org/wiki/Bayesian_multivariate_linear_regression#Conjugate_prior_distribution.

Metadata: ID: P159 | shortcut: mblr-prior | author: JoramSoch | date: 2020-09-03, 07:33.

2.2.2 Posterior distribution

Theorem: Let

$$Y = XB + E, \quad E \sim \mathcal{MN}(0, V, \Sigma) \quad (1)$$

be a general linear model (\rightarrow Definition III/2.1.1) with measured $n \times v$ data matrix Y , known $n \times p$ design matrix X , known $n \times n$ covariance structure (\rightarrow Definition II/5.1.1) V as well as unknown $p \times v$ regression coefficients B and unknown $v \times v$ noise covariance (\rightarrow Definition II/5.1.1) Σ . Moreover, assume a normal-Wishart prior distribution (\rightarrow Proof III/2.2.1) over the model parameters B and $T = \Sigma^{-1}$:

$$p(B, T) = \mathcal{MN}(B; M_0, \Lambda_0^{-1}, T^{-1}) \cdot \mathcal{W}(T; \Omega_0^{-1}, \nu_0) . \quad (2)$$

Then, the posterior distribution (\rightarrow Definition I/5.1.7) is also a normal-Wishart distribution (\rightarrow Definition “nw”)

$$p(B, T|Y) = \mathcal{MN}(B; M_n, \Lambda_n^{-1}, T^{-1}) \cdot \mathcal{W}(T; \Omega_n^{-1}, \nu_n) \quad (3)$$

and the posterior hyperparameters (\rightarrow Definition I/5.1.7) are given by

$$\begin{aligned} M_n &= \Lambda_n^{-1} (X^T P Y + \Lambda_0 M_0) \\ \Lambda_n &= X^T P X + \Lambda_0 \\ \Omega_n &= \Omega_0 + Y^T P Y + M_0^T \Lambda_0 M_0 - M_n^T \Lambda_n M_n \\ \nu_n &= \nu_0 + n . \end{aligned} \quad (4)$$

Proof: According to Bayes’ theorem (\rightarrow Proof I/5.2.1), the posterior distribution (\rightarrow Definition I/5.1.7) is given by

$$p(B, T|Y) = \frac{p(Y|B, T) p(B, T)}{p(Y)} . \quad (5)$$

Since $p(Y)$ is just a normalization factor, the posterior is proportional (\rightarrow Proof I/5.1.8) to the numerator:

$$p(B, T|Y) \propto p(Y|B, T) p(B, T) = p(Y, B, T) . \quad (6)$$

Equation (1) implies the following likelihood function (\rightarrow Definition I/5.1.2)

$$p(Y|B, \Sigma) = \mathcal{MN}(Y; XB, V, \Sigma) = \sqrt{\frac{1}{(2\pi)^{nv} |\Sigma|^n |V|^v}} \exp \left[-\frac{1}{2} \text{tr} (\Sigma^{-1} (Y - XB)^T V^{-1} (Y - XB)) \right] \quad (7)$$

which, for mathematical convenience, can also be parametrized as

$$p(Y|B, T) = \mathcal{MN}(Y; XB, P, T^{-1}) = \sqrt{\frac{|T|^n |P|^v}{(2\pi)^{nv}}} \exp \left[-\frac{1}{2} \text{tr} (T(Y - XB)^T P(Y - XB)) \right] \quad (8)$$

using the $v \times v$ precision matrix (\rightarrow Definition I/1.7.8) $T = \Sigma^{-1}$ and the $n \times n$ precision matrix (\rightarrow Definition I/1.7.8) $P = V^{-1}$.

Combining the likelihood function (\rightarrow Definition I/5.1.2) (8) with the prior distribution (\rightarrow Definition I/5.1.3) (2), the joint likelihood (\rightarrow Definition I/5.1.5) of the model is given by

$$\begin{aligned} p(Y, B, T) &= p(Y|B, T) p(B, T) \\ &= \sqrt{\frac{|T|^n |P|^v}{(2\pi)^{nv}}} \exp \left[-\frac{1}{2} \text{tr} (T(Y - XB)^T P(Y - XB)) \right] \cdot \\ &\quad \sqrt{\frac{|T|^p |\Lambda_0|^v}{(2\pi)^{pv}}} \exp \left[-\frac{1}{2} \text{tr} (T(B - M_0)^T \Lambda_0 (B - M_0)) \right] \cdot \\ &\quad \frac{1}{\Gamma_v \left(\frac{\nu_0}{2} \right)} \sqrt{\frac{|\Omega_0|^{\nu_0}}{2^{\nu_0 v}}} |T|^{(\nu_0 - v - 1)/2} \exp \left[-\frac{1}{2} \text{tr} (\Omega_0 T) \right] . \end{aligned} \quad (9)$$

Collecting identical variables gives:

$$\begin{aligned} p(Y, B, T) &= \sqrt{\frac{|T|^n |P|^v}{(2\pi)^{nv}}} \sqrt{\frac{|T|^p |\Lambda_0|^v}{(2\pi)^{pv}}} \sqrt{\frac{|\Omega_0|^{\nu_0}}{2^{\nu_0 v}}} \frac{1}{\Gamma_v \left(\frac{\nu_0}{2} \right)} \cdot |T|^{(\nu_0 - v - 1)/2} \exp \left[-\frac{1}{2} \text{tr} (\Omega_0 T) \right] \cdot \\ &\quad \exp \left[-\frac{1}{2} \text{tr} (T [(Y - XB)^T P(Y - XB) + (B - M_0)^T \Lambda_0 (B - M_0)]) \right] . \end{aligned} \quad (10)$$

Expanding the products in the exponent gives:

$$\begin{aligned} p(Y, B, T) &= \sqrt{\frac{|T|^n |P|^v}{(2\pi)^{nv}}} \sqrt{\frac{|T|^p |\Lambda_0|^v}{(2\pi)^{pv}}} \sqrt{\frac{|\Omega_0|^{\nu_0}}{2^{\nu_0 v}}} \frac{1}{\Gamma_v \left(\frac{\nu_0}{2} \right)} \cdot |T|^{(\nu_0 - v - 1)/2} \exp \left[-\frac{1}{2} \text{tr} (\Omega_0 T) \right] \cdot \\ &\quad \exp \left[-\frac{1}{2} \text{tr} (T [Y^T P Y - Y^T P X B - B^T X^T P Y + B^T X^T P X B + \right. \\ &\quad \left. B^T \Lambda_0 B - B^T \Lambda_0 M_0 - M_0^T \Lambda_0 B + M_0^T \Lambda_0 M_0]) \right] . \end{aligned} \quad (11)$$

Completing the square over B , we finally have

$$p(Y, B, T) = \sqrt{\frac{|T|^n |P|^v}{(2\pi)^{nv}}} \sqrt{\frac{|T|^p |\Lambda_0|^v}{(2\pi)^{pv}}} \sqrt{\frac{|\Omega_0|^{\nu_0}}{2^{\nu_0 v}} \frac{1}{\Gamma_v(\frac{\nu_0}{2})}} \cdot |T|^{(\nu_0 - v - 1)/2} \exp \left[-\frac{1}{2} \text{tr}(\Omega_0 T) \right] \cdot \exp \left[-\frac{1}{2} \text{tr} \left(T \left[(B - M_n)^T \Lambda_n (B - M_n) + (Y^T P Y + M_0^T \Lambda_0 M_0 - M_n^T \Lambda_n M_n) \right] \right) \right]. \quad (12)$$

with the posterior hyperparameters (\rightarrow Definition I/5.1.7)

$$\begin{aligned} M_n &= \Lambda_n^{-1} (X^T P Y + \Lambda_0 M_0) \\ \Lambda_n &= X^T P X + \Lambda_0. \end{aligned} \quad (13)$$

Ergo, the joint likelihood is proportional to

$$p(Y, B, T) \propto |T|^{p/2} \cdot \exp \left[-\frac{1}{2} \text{tr} \left(T \left[(B - M_n)^T \Lambda_n (B - M_n) \right] \right) \right] \cdot |T|^{(\nu_n - v - 1)/2} \cdot \exp \left[-\frac{1}{2} \text{tr}(\Omega_n T) \right] \quad (14)$$

with the posterior hyperparameters (\rightarrow Definition I/5.1.7)

$$\begin{aligned} \Omega_n &= \Omega_0 + Y^T P Y + M_0^T \Lambda_0 M_0 - M_n^T \Lambda_n M_n \\ \nu_n &= \nu_0 + n. \end{aligned} \quad (15)$$

From the term in (14), we can isolate the posterior distribution over B given T :

$$p(B|T, Y) = \mathcal{MN}(B; M_n, \Lambda_n^{-1}, T^{-1}). \quad (16)$$

From the remaining term, we can isolate the posterior distribution over T :

$$p(T|Y) = \mathcal{W}(T; \Omega_n^{-1}, \nu_n). \quad (17)$$

Together, (16) and (17) constitute the joint (\rightarrow Definition I/1.2.2) posterior distribution (\rightarrow Definition I/5.1.7) of B and T .

Sources:

- Wikipedia (2020): “Bayesian multivariate linear regression”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-09-03; URL: https://en.wikipedia.org/wiki/Bayesian_multivariate_linear_regression#Posterior_distribution.

Metadata: ID: P160 | shortcut: mblr-post | author: JoramSoch | date: 2020-09-03, 08:37.

2.2.3 Log model evidence

Theorem: Let

$$Y = X B + E, \quad E \sim \mathcal{MN}(0, V, \Sigma) \quad (1)$$

be a general linear model (\rightarrow Definition III/2.1.1) with measured $n \times v$ data matrix Y , known $n \times p$ design matrix X , known $n \times n$ covariance structure (\rightarrow Definition II/5.1.1) V as well as unknown $p \times v$

regression coefficients B and unknown $v \times v$ noise covariance (\rightarrow Definition II/5.1.1) Σ . Moreover, assume a normal-Wishart prior distribution (\rightarrow Proof III/2.2.1) over the model parameters B and $T = \Sigma^{-1}$:

$$p(B, T) = \mathcal{MN}(B; M_0, \Lambda_0^{-1}, T^{-1}) \cdot \mathcal{W}(T; \Omega_0^{-1}, \nu_0) . \quad (2)$$

Then, the log model evidence (\rightarrow Definition IV/3.1.1) for this model is

$$\begin{aligned} \log p(y|m) = & \frac{v}{2} \log |P| - \frac{nv}{2} \log(2\pi) + \frac{v}{2} \log |\Lambda_0| - \frac{v}{2} \log |\Lambda_n| + \\ & \frac{\nu_0}{2} \log \left| \frac{1}{2} \Omega_0 \right| - \frac{\nu_n}{2} \log \left| \frac{1}{2} \Omega_n \right| + \log \Gamma_v \left(\frac{\nu_n}{2} \right) - \log \Gamma_v \left(\frac{\nu_0}{2} \right) \end{aligned} \quad (3)$$

where the posterior hyperparameters (\rightarrow Definition I/5.1.7) are given by

$$\begin{aligned} M_n &= \Lambda_n^{-1} (X^T P Y + \Lambda_0 M_0) \\ \Lambda_n &= X^T P X + \Lambda_0 \\ \Omega_n &= \Omega_0 + Y^T P Y + M_0^T \Lambda_0 M_0 - M_n^T \Lambda_n M_n \\ \nu_n &= \nu_0 + n . \end{aligned} \quad (4)$$

Proof: According to the law of marginal probability (\rightarrow Definition I/1.2.3), the model evidence (\rightarrow Definition I/5.1.9) for this model is:

$$p(Y|m) = \iint p(Y|B, T) p(B, T) dB dT . \quad (5)$$

According to the law of conditional probability (\rightarrow Definition I/1.2.4), the integrand is equivalent to the joint likelihood (\rightarrow Definition I/5.1.5):

$$p(Y|m) = \iint p(Y, B, T) dB dT . \quad (6)$$

Equation (1) implies the following likelihood function (\rightarrow Definition I/5.1.2)

$$p(Y|B, \Sigma) = \mathcal{MN}(Y; XB, V, \Sigma) = \sqrt{\frac{1}{(2\pi)^{nv} |\Sigma| |V|^v}} \exp \left[-\frac{1}{2} \text{tr} (\Sigma^{-1} (Y - XB)^T V^{-1} (Y - XB)) \right] \quad (7)$$

which, for mathematical convenience, can also be parametrized as

$$p(Y|B, T) = \mathcal{MN}(Y; XB, P, T^{-1}) = \sqrt{\frac{|T|^n |P|^v}{(2\pi)^{nv}}} \exp \left[-\frac{1}{2} \text{tr} (T (Y - XB)^T P (Y - XB)) \right] \quad (8)$$

using the $v \times v$ precision matrix (\rightarrow Definition I/1.7.8) $T = \Sigma^{-1}$ and the $n \times n$ precision matrix (\rightarrow Definition I/1.7.8) $P = V^{-1}$.

When deriving the posterior distribution (\rightarrow Proof III/2.2.2) $p(B, T|Y)$, the joint likelihood $p(Y, B, T)$ is obtained as

$$p(Y, B, T) = \sqrt{\frac{|T|^n |P|^v}{(2\pi)^{nv}}} \sqrt{\frac{|T|^p |\Lambda_0|^v}{(2\pi)^{pv}}} \sqrt{\frac{|\Omega_0|^{\nu_0}}{2^{\nu_0 v}} \frac{1}{\Gamma_v\left(\frac{\nu_0}{2}\right)}} \cdot |T|^{(\nu_0 - v - 1)/2} \exp\left[-\frac{1}{2} \text{tr}(\Omega_0 T)\right] \cdot \exp\left[-\frac{1}{2} \text{tr}\left(T[(B - M_n)^T \Lambda_n (B - M_n) + (Y^T P Y + M_0^T \Lambda_0 M_0 - M_n^T \Lambda_n M_n)]\right)\right]. \quad (9)$$

Using the probability density function of the matrix-normal distribution (\rightarrow Proof II/5.1.2), we can rewrite this as

$$p(Y, B, T) = \sqrt{\frac{|T|^n |P|^v}{(2\pi)^{nv}}} \sqrt{\frac{|T|^p |\Lambda_0|^v}{(2\pi)^{pv}}} \sqrt{\frac{(2\pi)^{pv}}{|T|^p |\Lambda_n|^v}} \sqrt{\frac{|\Omega_0|^{\nu_0}}{2^{\nu_0 v}} \frac{1}{\Gamma_v\left(\frac{\nu_0}{2}\right)}} \cdot |T|^{(\nu_0 - v - 1)/2} \exp\left[-\frac{1}{2} \text{tr}(\Omega_0 T)\right] \cdot \mathcal{MN}(B; M_n, \Lambda_n^{-1}, T^{-1}) \cdot \exp\left[-\frac{1}{2} \text{tr}\left(T[Y^T P Y + M_0^T \Lambda_0 M_0 - M_n^T \Lambda_n M_n]\right)\right]. \quad (10)$$

Now, B can be integrated out easily:

$$\int p(Y, B, T) dB = \sqrt{\frac{|T|^n |P|^v}{(2\pi)^{nv}}} \sqrt{\frac{|\Lambda_0|^v}{|\Lambda_n|^v}} \sqrt{\frac{|\Omega_0|^{\nu_0}}{2^{\nu_0 v}} \frac{1}{\Gamma_v\left(\frac{\nu_0}{2}\right)}} \cdot |T|^{(\nu_0 - v - 1)/2} \cdot \exp\left[-\frac{1}{2} \text{tr}\left(T[\Omega_0 + Y^T P Y + M_0^T \Lambda_0 M_0 - M_n^T \Lambda_n M_n]\right)\right]. \quad (11)$$

Using the probability density function of the Wishart distribution (\rightarrow Proof “wish-pdf”), we can rewrite this as

$$\int p(Y, B, T) dB = \sqrt{\frac{|P|^v}{(2\pi)^{nv}}} \sqrt{\frac{|\Lambda_0|^v}{|\Lambda_n|^v}} \sqrt{\frac{|\Omega_0|^{\nu_0}}{2^{\nu_0 v}}} \sqrt{\frac{2^{\nu_n v}}{|\Omega_n|^{\nu_n}} \frac{\Gamma_v\left(\frac{\nu_n}{2}\right)}{\Gamma_v\left(\frac{\nu_0}{2}\right)}} \cdot \mathcal{W}(T; \Omega_n^{-1}, \nu_n). \quad (12)$$

Finally, T can also be integrated out:

$$\iint p(Y, B, T) dB dT = \sqrt{\frac{|P|^v}{(2\pi)^{nv}}} \sqrt{\frac{|\Lambda_0|^v}{|\Lambda_n|^v}} \sqrt{\frac{|\frac{1}{2}\Omega_0|^{\nu_0}}{|\frac{1}{2}\Omega_n|^{\nu_n}} \frac{\Gamma_v\left(\frac{\nu_n}{2}\right)}{\Gamma_v\left(\frac{\nu_0}{2}\right)}} = p(y|m). \quad (13)$$

Thus, the log model evidence (\rightarrow Definition IV/3.1.1) of this model is given by

$$\begin{aligned} \log p(y|m) &= \frac{v}{2} \log |P| - \frac{nv}{2} \log(2\pi) + \frac{v}{2} \log |\Lambda_0| - \frac{v}{2} \log |\Lambda_n| + \\ &\quad \frac{\nu_0}{2} \log \left| \frac{1}{2} \Omega_0 \right| - \frac{\nu_n}{2} \log \left| \frac{1}{2} \Omega_n \right| + \log \Gamma_v\left(\frac{\nu_n}{2}\right) - \log \Gamma_v\left(\frac{\nu_0}{2}\right). \end{aligned} \quad (14)$$

Sources:

- original work

Metadata: ID: P161 | shortcut: mblr-lme | author: JoramSoch | date: 2020-09-03, 09:23.

3 Poisson data

3.1 Poisson-distributed data

3.1.1 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 II/1.4.1) 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.

3.1.2 Maximum likelihood estimation

Theorem: Let there be a Poisson-distributed data (\rightarrow Definition III/3.1.1) 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 I/4.1.3) 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/5.1.2) for each observation is given by the probability mass function of the Poisson distribution (\rightarrow Proof II/1.4.2)

$$p(y_i|\lambda) = \text{Poiss}(y_i; \lambda) = \frac{\lambda^{y_i} \cdot \exp(-\lambda)}{y_i!} \quad (4)$$

and because observations are independent (\rightarrow Definition I/1.2.5), 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 I/4.1.2) 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.

3.2 Poisson distribution with exposure values

3.2.1 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 II/1.4.1) 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.

3.2.2 Conjugate prior distribution

Theorem: Consider data $y = \{y_1, \dots, y_n\}$ following a Poisson distribution with exposure values (\rightarrow Definition III/3.2.1):

$$y_i \sim \text{Poiss}(\lambda x_i), \quad i = 1, \dots, n. \quad (1)$$

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 II/1.4.2), the likelihood function (\rightarrow Definition I/5.1.2) 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 I/1.2.5), 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/3.3.4)

$$p(\lambda) = \frac{b_0^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0\lambda]\tag{9}$$

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.

3.2.3 Posterior distribution

Theorem: Consider data $y = \{y_1, \dots, y_n\}$ following a Poisson distribution with exposure values (\rightarrow Definition III/3.2.1):

$$y_i \sim \text{Poiss}(\lambda x_i), \quad i = 1, \dots, n .\tag{1}$$

Moreover, assume a gamma prior distribution (\rightarrow Proof III/3.2.2) over the model parameter λ :

$$p(\lambda) = \text{Gam}(\lambda; a_0, b_0) .\tag{2}$$

Then, the posterior distribution (\rightarrow Definition I/5.1.7) 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/5.1.7) are given by

$$\begin{aligned} a_n &= a_0 + n\bar{y} \\ a_n &= a_0 + n\bar{x} . \end{aligned} \quad (4)$$

Proof: With the probability mass function of the Poisson distribution (\rightarrow Proof II/1.4.2), the likelihood function (\rightarrow Definition I/5.1.2) 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 I/1.2.5), 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/5.1.5) 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 posterior distribution is proportional to the joint likelihood (\rightarrow Proof I/5.1.8):

$$p(\lambda|y) \propto p(y, \lambda) . \quad (10)$$

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/3.3.4):

$$p(\lambda|y) = \frac{b_n^{a_n}}{\Gamma(a_n)} \lambda^{a_n-1} \exp[-b_n\lambda] = \text{Gam}(\lambda; a_n, b_n) . \quad (12)$$

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.

3.2.4 Log model evidence

Theorem: Consider data $y = \{y_1, \dots, y_n\}$ following a Poisson distribution with exposure values (\rightarrow Definition III/3.2.1):

$$y_i \sim \text{Poiss}(\lambda x_i), \quad i = 1, \dots, n . \quad (1)$$

Moreover, assume a gamma prior distribution (\rightarrow Proof III/3.2.2) 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/5.1.7) 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 II/1.4.2), the likelihood function (\rightarrow Definition I/5.1.2) 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 I/1.2.5), 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/5.1.5) 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/5.1.9):

$$p(y) = \int p(y, \lambda) d\lambda . \quad (10)$$

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/3.3.4), 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.

4 Probability data

4.1 Beta-distributed data

4.1.1 Definition

Definition: Beta-distributed data are defined as a set of proportions $y = \{y_1, \dots, y_n\}$ with $y_i \in [0, 1]$, $i = 1, \dots, n$, independent and identically distributed according to a Beta distribution (\rightarrow Definition II/3.5.1) with shapes α and β :

$$y_i \sim \text{Bet}(\alpha, \beta), \quad i = 1, \dots, n. \quad (1)$$

Sources:

- original work

Metadata: ID: D77 | shortcut: beta-data | author: JoramSoch | date: 2020-06-28, 21:16.

4.1.2 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 II/3.5.1) 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 “mome”) 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 II/3.5.1) in terms of the parameters α and β are given by

$$\begin{aligned} \text{E}(X) &= \frac{\alpha}{\alpha + \beta} \\ \text{Var}(X) &= \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}. \end{aligned} \quad (4)$$

Thus, matching the moments (\rightarrow Definition “mome”) requires us to solve the following equation system for α and β :

$$\begin{aligned}\bar{y} &= \frac{\alpha}{\alpha + \beta} \\ \bar{v} &= \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.\end{aligned}\tag{5}$$

From the first equation, we can deduce:

$$\begin{aligned}\bar{y}(\alpha + \beta) &= \alpha \\ \alpha\bar{y} + \beta\bar{y} &= \alpha \\ \beta\bar{y} &= \alpha - \alpha\bar{y} \\ \beta &= \frac{\alpha}{\bar{y}} - \alpha \\ \beta &= \alpha \left(\frac{1}{\bar{y}} - 1 \right).\end{aligned}\tag{6}$$

If we define $q = 1/\bar{y} - 1$ and plug (6) into the second equation, we have:

$$\begin{aligned}\bar{v} &= \frac{\alpha \cdot \alpha q}{(\alpha + \alpha q)^2(\alpha + \alpha q + 1)} \\ &= \frac{\alpha^2 q}{(\alpha(1 + q))^2(\alpha(1 + q) + 1)} \\ &= \frac{q}{(1 + q)^2(\alpha(1 + q) + 1)} \\ &= \frac{q}{\alpha(1 + q)^3 + (1 + q)^2} \\ q &= \bar{v} [\alpha(1 + q)^3 + (1 + q)^2].\end{aligned}\tag{7}$$

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.

4.2 Logistic regression

4.2.1 Definition

Definition: A logistic regression model is given by a set of binary observations $y_i \in \{0, 1\}, i = 1, \dots, n$, a set of predictors $x_j \in \mathbb{R}^n, j = 1, \dots, p$, a base b and the assumption that the log-odds are a linear combination of the predictors:

$$l_i = x_i \beta + \varepsilon_i, \quad i = 1, \dots, n \tag{1}$$

where l_i are the log-odds that $y_i = 1$

$$l_i = \log_b \frac{\Pr(y_i = 1)}{\Pr(y_i = 0)} \tag{2}$$

and x_i is the i -th row of the $n \times p$ matrix

$$X = [x_1, \dots, x_p] . \tag{3}$$

Within this model,

- y are called “categorical observations” or “dependent variable”;
- X is called “design matrix” or “set of independent variables”;
- β are called “regression coefficients” or “weights”;
- ε_i is called “noise” or “error term”;
- n is the number of observations;
- p is the number of predictors.

Sources:

- Wikipedia (2020): “Logistic regression”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-06-28; URL: https://en.wikipedia.org/wiki/Logistic_regression#Logistic_model.

Metadata: ID: D76 | shortcut: logreg | author: JoramSoch | date: 2020-06-28, 20:51.

4.2.2 Probability and log-odds

Theorem: Assume a logistic regression model (\rightarrow Definition III/4.2.1)

$$l_i = x_i\beta + \varepsilon_i, \quad i = 1, \dots, n \quad (1)$$

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 log-odds in favor of $y_i = 1$ against $y_i = 0$ can also be expressed as

$$l_i = \log_b \frac{p(x_i|y_i = 1) p(y_i = 1)}{p(x_i|y_i = 0) p(y_i = 0)} \quad (2)$$

where $p(x_i|y_i)$ is a likelihood function (\rightarrow Definition I/5.1.2) consistent with (1), $p(y_i)$ are prior probabilities (\rightarrow Definition I/5.1.3) for $y_i = 1$ and $y_i = 0$ and where b is the base used to form the log-odds l_i .

Proof: Using Bayes' theorem (\rightarrow Proof I/5.2.1) and the law of marginal probability (\rightarrow Definition I/1.2.3), the posterior probabilities (\rightarrow Definition I/5.1.7) for $y_i = 1$ and $y_i = 0$ are given by

$$\begin{aligned} p(y_i = 1|x_i) &= \frac{p(x_i|y_i = 1) p(y_i = 1)}{p(x_i|y_i = 1) p(y_i = 1) + p(x_i|y_i = 0) p(y_i = 0)} \\ p(y_i = 0|x_i) &= \frac{p(x_i|y_i = 0) p(y_i = 0)}{p(x_i|y_i = 1) p(y_i = 1) + p(x_i|y_i = 0) p(y_i = 0)} . \end{aligned} \quad (3)$$

Calculating the log-odds from the posterior probabilities, we have

$$\begin{aligned} l_i &= \log_b \frac{p(y_i = 1|x_i)}{p(y_i = 0|x_i)} \\ &= \log_b \frac{p(x_i|y_i = 1) p(y_i = 1)}{p(x_i|y_i = 0) p(y_i = 0)} . \end{aligned} \quad (4)$$

Sources:

- Bishop, Christopher M. (2006): “Linear Models for Classification”; in: *Pattern Recognition for Machine Learning*, ch. 4, p. 197, eq. 4.58; URL: <http://users.isr.ist.utl.pt/~wurmd/Livros/school/Bishop%20-%20Pattern%20Recognition%20And%20Machine%20Learning%20-%20Springer%20%202006.pdf>.

Metadata: ID: P105 | shortcut: logreg-pnlo | author: JoramSoch | date: 2020-05-19, 05:08.

4.2.3 Log-odds and probability

Theorem: Assume a logistic regression model (\rightarrow Definition III/4.2.1)

$$l_i = x_i\beta + \varepsilon_i, \quad i = 1, \dots, n \quad (1)$$

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.

5 Categorical data

5.1 Binomial observations

5.1.1 Definition

Definition: An ordered pair (n, y) with $n \in \mathbb{N}$ and $y \in \mathbb{N}_0$, where y is the number of successes in n trials, constitutes a set of binomial observations.

Sources:

- original work

Metadata: ID: D78 | shortcut: bin-data | author: JoramSoch | date: 2020-07-07, 07:04.

5.1.2 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/1.3.1):

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

Then, the conjugate prior (\rightarrow Definition “prior-conj”) for the model parameter p is a beta distribution (\rightarrow Definition II/3.5.1):

$$p(p) = \text{Bet}(p; \alpha_0, \beta_0) . \quad (2)$$

Proof: With the probability mass function of the binomial distribution (\rightarrow Proof II/1.3.2), the likelihood function (\rightarrow Definition I/5.1.2) 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 II/3.5.2)

$$p(p) = \frac{1}{B(\alpha_0, \beta_0)} p^{\alpha_0-1} (1-p)^{\beta_0-1} \quad (6)$$

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.

5.1.3 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/1.3.1):

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

Moreover, assume a beta prior distribution (\rightarrow Proof III/5.1.2) over the model parameter p :

$$p(p) = \text{Bet}(p; \alpha_0, \beta_0) . \quad (2)$$

Then, the posterior distribution (\rightarrow Definition I/5.1.7) is also a beta distribution (\rightarrow Definition II/3.5.1)

$$p(p|y) = \text{Bet}(p; \alpha_n, \beta_n) . \quad (3)$$

and the posterior hyperparameters (\rightarrow Definition I/5.1.7) 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 II/1.3.2), the likelihood function (\rightarrow Definition I/5.1.2) 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/5.1.5) 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 I/5.1.8):

$$p(p|y) \propto p(y, p) . \quad (7)$$

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 II/3.5.2):

$$p(p|y) = \frac{1}{B(\alpha_n, \beta_n)} p^{\alpha_n-1} (1-p)^{\beta_n-1} = \text{Bet}(p; \alpha_n, \beta_n) . \quad (9)$$

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.

5.1.4 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/1.3.1):

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

Moreover, assume a beta prior distribution (\rightarrow Proof III/5.1.2) 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/5.1.7) 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 II/1.3.2), the likelihood function (\rightarrow Definition I/5.1.2) 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/5.1.5) 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/5.1.9):

$$p(y) = \int p(y, p) \, dp . \quad (7)$$

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 II/3.5.2), 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.

5.2 Multinomial observations

5.2.1 Definition

Definition: An ordered pair (n, y) with $n \in \mathbb{N}$ and $y = [y_1, \dots, y_k] \in \mathbb{N}_0^{1 \times k}$, where y_i is the number of observations for the i -th out of k categories obtained in n trials, $i = 1, \dots, k$, constitutes a set of multinomial observations.

Sources:

- original work

Metadata: ID: D79 | shortcut: mult-data | author: JoramSoch | date: 2020-07-07, 07:12.

5.2.2 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 II/4.3.1):

$$p(p) = \text{Dir}(p; \alpha_0) . \quad (2)$$

Proof: With the probability mass function of the multinomial distribution (\rightarrow Proof II/2.2.2), the likelihood function (\rightarrow Definition I/5.1.2) 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 II/4.3.2)

$$p(p) = \frac{\Gamma\left(\sum_{j=1}^k \alpha_{0j}\right)}{\prod_{j=1}^k \Gamma(\alpha_{0j})} \prod_{j=1}^k p_j^{\alpha_{0j}-1} \quad (6)$$

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: 2020-03-11, 14:15.

5.2.3 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/5.2.2) over the model parameter p :

$$p(p) = \text{Dir}(p; \alpha_0) . \quad (2)$$

Then, the posterior distribution (\rightarrow Definition I/5.1.7) is also a Dirichlet distribution (\rightarrow Definition II/4.3.1)

$$p(p|y) = \text{Dir}(p; \alpha_n) . \quad (3)$$

and the posterior hyperparameters (\rightarrow Definition I/5.1.7) 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 II/2.2.2), the likelihood function (\rightarrow Definition I/5.1.2) 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/5.1.5) of the model is given by

$$\begin{aligned} p(y, p) &= p(y|p) p(p) \\ &= \binom{n}{y_1, \dots, y_k} \prod_{j=1}^k p_j^{y_j} \cdot \frac{\Gamma(\sum_{j=1}^k \alpha_{0j})}{\prod_{j=1}^k \Gamma(\alpha_{0j})} \prod_{j=1}^k p_j^{\alpha_{0j}-1} \\ &= \frac{\Gamma(\sum_{j=1}^k \alpha_{0j})}{\prod_{j=1}^k \Gamma(\alpha_{0j})} \binom{n}{y_1, \dots, y_k} \prod_{j=1}^k p_j^{\alpha_{0j}+y_j-1} . \end{aligned} \quad (6)$$

Note that the posterior distribution is proportional to the joint likelihood (\rightarrow Proof I/5.1.8):

$$p(p|y) \propto p(y, p) . \quad (7)$$

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 II/4.3.2):

$$p(p|y) = \frac{\Gamma\left(\sum_{j=1}^k \alpha_{nj}\right)}{\prod_{j=1}^k \Gamma(\alpha_{nj})} \prod_{j=1}^k p_j^{\alpha_{nj}-1} = \text{Dir}(p; \alpha_n) . \quad (9)$$

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.

5.2.4 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/5.2.2) 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/5.1.7) 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 II/2.2.2), the likelihood function (\rightarrow Definition I/5.1.2) 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/5.1.5) 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 II/4.3.2), 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 I/4.1.3).

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/1.5.1) μ and variance (\rightarrow Definition I/1.6.1) 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 I/4.1.3) 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/1.1.15) 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/1.5.1) of the maximum likelihood estimator (\rightarrow Definition I/4.1.3) 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 I/1.6.2)

$$\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 , \tag{8}$$

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 (\rightarrow Definition I/1.5.1) of \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/1.5.1) μ and variance (\rightarrow Definition I/1.6.1) 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 I/4.1.3) 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 I/1.2.5) 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

$$R^2_{\text{adj}} = 1 - \frac{\text{RSS}/(n-p)}{\text{TSS}/(n-1)} \quad (3)$$

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 Proof III/1.1.2) 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 I/4.1.4) 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 I/4.1.4) (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 I/4.1.2)

$$\text{LL}(\beta, \sigma^2) = \log p(y|\beta, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} (y - X\beta)^T (y - X\beta) , \quad (3)$$

such that maximum likelihood estimates are (\rightarrow Proof III/1.1.15)

$$\hat{\beta} = (X^T X)^{-1} X^T y \quad (4)$$

$$\hat{\sigma}^2 = \frac{1}{n} (y - X\hat{\beta})^T (y - X\hat{\beta}) \quad (5)$$

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 I/4.1.3), 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-mlr | 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 I/1.2.5) 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 Proof III/1.1.15) $\hat{\beta}$ 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.

1.3.2 Relationship with R^2

Theorem: Let there be a linear regression model (\rightarrow Definition III/1.1.1) with independent (\rightarrow Definition I/1.2.5) observations

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

and parameter estimates (\rightarrow Definition “est”) obtained with ordinary least squares (\rightarrow Proof III/1.1.2)

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

Then, the signal-to noise ratio (\rightarrow Definition IV/1.3.1) can be expressed in terms of the coefficient of determination (\rightarrow Definition IV/1.2.1)

$$\text{SNR} = \frac{R^2}{1 - R^2} \quad (3)$$

and vice versa

$$R^2 = \frac{\text{SNR}}{1 + \text{SNR}} , \quad (4)$$

if the predicted signal mean is equal to the actual signal mean.

Proof: The signal-to-noise ratio (SNR) is defined as (\rightarrow Definition IV/1.3.1)

$$\text{SNR} = \frac{\text{Var}(X\hat{\beta})}{\hat{\sigma}^2} = \frac{\text{Var}(\hat{y})}{\hat{\sigma}^2} . \quad (5)$$

Writing out the variances, we have

$$\text{SNR} = \frac{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2}{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2}{\sum_{i=1}^n (y_i - \hat{y}_i)^2} . \quad (6)$$

Note that it is irrelevant whether we use the biased estimator of the variance (\rightarrow Proof IV/1.1.2) (dividing by n) or the unbiased estimator for the variance (\rightarrow Proof IV/1.1.3) (dividing by $n - 1$), because the relevant terms cancel out.

If the predicted signal mean is equal to the actual signal mean – which is the case when variable regressors in X have mean zero, such that they are orthogonal to a constant regressor in X –, this means that $\bar{\hat{y}} = \bar{y}$, such that

$$\text{SNR} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \hat{y}_i)^2} . \quad (7)$$

Then, the SNR can be written in terms of the explained (\rightarrow Definition III/1.1.5), residual (\rightarrow Definition III/1.1.6) and total sum of squares (\rightarrow Definition III/1.1.4):

$$\text{SNR} = \frac{\text{ESS}}{\text{RSS}} = \frac{\text{ESS}/\text{TSS}}{\text{RSS}/\text{TSS}} . \quad (8)$$

With the derivation of the coefficient of determination (\rightarrow Proof IV/1.2.2), this becomes

$$\text{SNR} = \frac{R^2}{1 - R^2} . \quad (9)$$

Rearranging this equation for the coefficient of determination (\rightarrow Definition IV/1.2.1), we have

$$R^2 = \frac{\text{SNR}}{1 + \text{SNR}} , \quad (10)$$

Sources:

- original work

Metadata: ID: P63 | shortcut: snr-rsq | author: JoramSoch | date: 2020-02-26, 10:37.

2 Classical information criteria

2.1 Akaike information criterion

2.1.1 Definition

Definition: Let m be a generative model (\rightarrow Definition I/5.1.1) with likelihood function (\rightarrow Definition I/5.1.2) $p(y|\theta, m)$ and maximum likelihood estimates (\rightarrow Definition I/4.1.3)

$$\hat{\theta} = \arg \max_{\theta} \log p(y|\theta, m) . \quad (1)$$

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/5.1.1) with likelihood function (\rightarrow Definition I/5.1.2) $p(y|\theta, m)$ and maximum likelihood estimates (\rightarrow Definition I/4.1.3)

$$\hat{\theta} = \arg \max_{\theta} \log p(y|\theta, m) . \quad (1)$$

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/5.1.2) of a generative model (\rightarrow Definition I/5.1.1) $m \in \mathcal{M}$ 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/5.1.3) 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/5.1.9) $\log p(y|m)$, 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 I/4.1.3) (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 I/4.1.2)

$$\text{LL}(\theta) = \log p(y|\theta, m) \quad (2)$$

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/5.1.9) 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 IV/3.5.1) $p(m_j|y)$. 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/5.1.1) with likelihood function (\rightarrow Definition I/5.1.2) $p(y|\theta, m)$ and prior distribution (\rightarrow Definition I/5.1.3) $p(\theta|m)$. Together, likelihood function and prior distribution imply a posterior distribution (\rightarrow Definition I/5.1.7) $p(\theta|y, m)$. Define the posterior expected log-likelihood (\rightarrow Definition I/4.1.2) (PLL)

$$\text{PLL}(m) = \langle \log p(y|\theta, m) \rangle_{\theta|y} \quad (1)$$

and the log-likelihood (\rightarrow Definition I/4.1.2) 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/5.1.1) with likelihood function (\rightarrow Definition I/5.1.2) $p(y|\theta, m)$ and prior distribution (\rightarrow Definition I/5.1.3) $p(\theta|m)$. Then, the log model evidence (LME) of this model is defined as the logarithm of the marginal likelihood (\rightarrow Definition I/5.1.9):

$$\text{LME}(m) = \log p(y|m) . \quad (1)$$

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/5.1.2) of a generative model (\rightarrow Definition I/5.1.1) m for making inferences on model parameters θ given measured data y . Moreover, let $p(\theta|m)$ be a prior distribution (\rightarrow Definition I/5.1.3) 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 I/1.2.3) 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/5.2.1) which makes a statement about the posterior distribution (\rightarrow Definition I/5.1.7):

$$p(\theta|y, m) = \frac{p(y|\theta, m) p(\theta|m)}{p(y|m)} . \quad (6)$$

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/5.1.2)

$$\text{Acc}(m) = \langle p(y|\theta, m) \rangle_{p(\theta|y, m)} \quad (2)$$

and the complexity penalty is the Kullback-Leibler divergence (\rightarrow Definition I/2.5.1) of posterior (\rightarrow Definition I/5.1.7) from prior (\rightarrow Definition I/5.1.3)

$$\text{Com}(m) = \text{KL} [p(\theta|y, m) || p(\theta|m)] . \quad (3)$$

Proof: We consider Bayesian inference on data y using model m with parameters θ . Then, Bayes' theorem (\rightarrow Proof I/5.2.1) 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 proofs the partition given by (1).

Sources:

- Penny et al. (2007): “Bayesian Comparison of Spatially Regularised General Linear Models”; in: *Human Brain Mapping*, vol. 28, pp. 275–293; URL: <https://onlinelibrary.wiley.com/doi/full/10.1002/hbm.20327>; DOI: 10.1002/hbm.20327.
- Soch et al. (2016): “How to avoid mismodelling in GLM-based fMRI data analysis: cross-validated Bayesian model selection”; in: *NeuroImage*, vol. 141, pp. 469–489; URL: <https://www.sciencedirect.com/science/article/pii/S1053811916303615>; DOI: 10.1016/j.neuroimage.2016.07.047.

Metadata: ID: P3 | shortcut: lme-anc | author: JoramSoch | date: 2019-09-27, 16:13.

3.2 Log family evidence

3.2.1 Definition

Definition: Let f be a family of M generative models (\rightarrow Definition I/5.1.1) m_1, \dots, m_M , such that the following statement holds true:

$$f \Leftrightarrow m_1 \vee \dots \vee m_M . \quad (1)$$

Then, the family evidence of f is the weighted average of the model evidences (\rightarrow Definition I/5.1.9) of m_1, \dots, m_M where the weights are the within-family prior model probabilities (\rightarrow Definition I/5.1.3)

$$p(y|f) = \sum_{i=1}^M p(y|m_i) p(m_i|f) . \quad (2)$$

The log family evidence is given by the logarithm of the family evidence:

$$\text{LFE}(f) = \log p(y|f) = \log \sum_{i=1}^M p(y|m_i) p(m_i|f) . \quad (3)$$

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: D80 | shortcut: lfe | author: JoramSoch | date: 2020-07-13, 22:31.

3.2.2 Derivation

Theorem: Let f be a family of M generative models (\rightarrow Definition I/5.1.1) m_1, \dots, m_M with model evidences (\rightarrow Definition I/5.1.9) $p(y|m_1), \dots, p(y|m_M)$. Then, the log family evidence (\rightarrow Definition IV/3.2.1)

$$\text{LFE}(f) = \log p(y|f) \quad (1)$$

can be expressed as

$$\text{LFE}(f) = \log \sum_{i=1}^M p(y|m_i) p(m_i|f) \quad (2)$$

where $p(m_i|f)$ are the within-family (\rightarrow Definition IV/3.2.1) prior (\rightarrow Definition I/5.1.3) model (\rightarrow Definition I/5.1.1) probabilities (\rightarrow Definition I/1.2.1).

Proof: We will assume “prior additivity”

$$p(f) = \sum_{i=1}^M p(m_i) \quad (3)$$

and “posterior additivity” for family probabilities:

$$p(f|y) = \sum_{i=1}^M p(m_i|y) \quad (4)$$

Bayes’ theorem (\rightarrow Proof I/5.2.1) for the family evidence (\rightarrow Definition IV/3.2.1) gives

$$p(y|f) = \frac{p(f|y) p(y)}{p(f)} . \quad (5)$$

Applying (3) and (4), we have

$$p(y|f) = \frac{\sum_{i=1}^M p(m_i|y) p(y)}{\sum_{i=1}^M p(m_i)} . \quad (6)$$

Bayes’ theorem (\rightarrow Proof I/5.2.1) for the model evidence (\rightarrow Definition IV/3.2.1) gives

$$p(y|m_i) = \frac{p(m_i|y) p(y)}{p(m_i)} \quad (7)$$

which can be rearranged into

$$p(m_i|y) p(y) = p(y|m_i) p(m_i) . \quad (8)$$

Plugging (8) into (6), we have

$$\begin{aligned}
p(y|f) &= \frac{\sum_{i=1}^M p(y|m_i) p(m_i)}{\sum_{i=1}^M p(m_i)} \\
&= \sum_{i=1}^M p(y|m_i) \cdot \frac{p(m_i)}{\sum_{i=1}^M p(m_i)} \\
&= \sum_{i=1}^M p(y|m_i) \cdot \frac{p(m_i, f)}{p(f)} \\
&= \sum_{i=1}^M p(y|m_i) \cdot p(m_i|f) .
\end{aligned} \tag{9}$$

Equation (2) follows by logarithmizing both sides of (9).

Sources:

- original work

Metadata: ID: P132 | shortcut: lfe-der | author: JoramSoch | date: 2020-07-13, 22:58.

3.2.3 Calculation from 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 IV/3.2.1) are given by:

$$\text{LFE}(f_j) = \log \sum_{m_i \in f_j} [\exp[\text{LME}(m_i)] \cdot p(m_i|f_j)], \quad j = 1, \dots, F, \tag{1}$$

where $p(m_i|f_j)$ are within-family (\rightarrow Definition IV/3.2.1) prior (\rightarrow Definition I/5.1.3) model (\rightarrow Definition I/5.1.1) probabilities (\rightarrow Definition I/1.2.1).

Proof: Let us consider the (unlogarithmized) family evidence $p(y|f_j)$. According to the law of marginal probability (\rightarrow Definition I/1.2.3), this conditional probability is given by

$$p(y|f_j) = \sum_{m_i \in f_j} [p(y|m_i, f_j) \cdot p(m_i|f_j)] . \tag{2}$$

Because model families are mutually exclusive, it holds that $p(y|m_i, f_j) = p(y|m_i)$, such that

$$p(y|f_j) = \sum_{m_i \in f_j} [p(y|m_i) \cdot p(m_i|f_j)] . \tag{3}$$

Logarithmizing transforms the family evidence $p(y|f_j)$ into the log family evidence $\text{LFE}(f_j)$:

$$\text{LFE}(f_j) = \log \sum_{m_i \in f_j} [p(y|m_i) \cdot p(m_i|f_j)] . \tag{4}$$

The definition of the log model evidence (\rightarrow Definition IV/3.1.1)

$$\text{LME}(m) = \log p(y|m) \tag{5}$$

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.

3.3 Bayes factor

3.3.1 Definition

Definition: Consider two competing generative models (\rightarrow Definition I/5.1.1) m_1 and m_2 for observed data y . Then the Bayes factor in favor m_1 over m_2 is the ratio of marginal likelihoods (\rightarrow Definition I/5.1.9) of m_1 and m_2 :

$$\text{BF}_{12} = \frac{p(y | m_1)}{p(y | m_2)}. \quad (1)$$

Note that by Bayes’ theorem (\rightarrow Proof I/5.2.1), the ratio of posterior model probabilities (\rightarrow Definition IV/3.5.1) (i.e., the posterior model odds) can be written as

$$\frac{p(m_1 | y)}{p(m_2 | y)} = \frac{p(m_1)}{p(m_2)} \cdot \frac{p(y | m_1)}{p(y | m_2)}, \quad (2)$$

or equivalently by (1),

$$\frac{p(m_1 | y)}{p(m_2 | y)} = \frac{p(m_1)}{p(m_2)} \cdot \text{BF}_{12}. \quad (3)$$

In other words, the Bayes factor can be viewed as the factor by which the prior model odds are updated (after observing data y) to posterior model odds – which is also expressed by Bayes’ rule (\rightarrow Proof I/5.2.2).

Sources:

- Kass, Robert E. and Raftery, Adrian E. (1995): “Bayes Factors”; in: *Journal of the American Statistical Association*, vol. 90, no. 430, pp. 773-795; URL: <https://dx.doi.org/10.1080/01621459.1995.10476572>; DOI: 10.1080/01621459.1995.10476572.

Metadata: ID: D92 | shortcut: bf | author: tomfaulkenberry | date: 2020-08-26, 12:00.

3.3.2 Transitivity

Theorem: Consider three competing models (\rightarrow Definition I/5.1.1) m_1 , m_2 , and m_3 for observed data y . Then the Bayes factor (\rightarrow Definition IV/3.3.1) for m_1 over m_3 can be written as:

$$\text{BF}_{13} = \text{BF}_{12} \cdot \text{BF}_{23}. \quad (1)$$

Proof: By definition (\rightarrow Definition IV/3.3.1), the Bayes factor BF_{13} is the ratio of marginal likelihoods of data y over m_1 and m_3 , respectively. That is,

$$\text{BF}_{13} = \frac{p(y \mid m_1)}{p(y \mid m_3)}. \quad (2)$$

We can equivalently write

$$\begin{aligned} \text{BF}_{13} &\stackrel{(2)}{=} \frac{p(y \mid m_1)}{p(y \mid m_3)} \\ &= \frac{p(y \mid m_1)}{p(y \mid m_3)} \cdot \frac{p(y \mid m_2)}{p(y \mid m_2)} \\ &= \frac{p(y \mid m_1)}{p(y \mid m_2)} \cdot \frac{p(y \mid m_2)}{p(y \mid m_3)} \\ &\stackrel{(2)}{=} \text{BF}_{12} \cdot \text{BF}_{23}, \end{aligned} \quad (3)$$

which completes the proof of (1).

Sources:

- original work

Metadata: ID: P163 | shortcut: bf-trans | author: tomfaulkenberry | date: 2020-09-07, 12:00.

3.3.3 Computation using Savage-Dickey Density Ratio

Theorem: Consider two competing models (\rightarrow Definition I/5.1.1) on data y containing parameters δ and φ , namely $m_0 : \delta = \delta_0, \varphi$ and $m_1 : \delta, \varphi$. In this context, we say that δ is a parameter of interest, φ is a nuisance parameter (i.e., common to both models), and m_0 is a sharp point hypothesis nested within m_1 . Suppose further that the prior for the nuisance parameter φ in m_0 is equal to the prior for φ in m_1 after conditioning on the restriction – that is, $p(\varphi \mid m_0) = p(\varphi \mid \delta = \delta_0, m_1)$. Then the Bayes factor (\rightarrow Definition IV/3.3.1) for m_0 over m_1 can be computed as:

$$\text{BF}_{01} = \frac{p(\delta = \delta_0 \mid y, m_1)}{p(\delta = \delta_0 \mid m_1)}. \quad (1)$$

Proof: By definition (\rightarrow Definition IV/3.3.1), the Bayes factor BF_{01} is the ratio of marginal likelihoods of data y over m_0 and m_1 , respectively. That is,

$$\text{BF}_{01} = \frac{p(y \mid m_0)}{p(y \mid m_1)}. \quad (2)$$

The key idea in the proof is that we can use a “change of variables” technique to express BF_{01} entirely in terms of the “encompassing” model m_1 . This proceeds by first unpacking the marginal likelihood (\rightarrow Definition I/5.1.9) for m_0 over the nuisance parameter φ and then using the fact that m_0 is a sharp hypothesis nested within m_1 to rewrite everything in terms of m_1 . Specifically,

$$\begin{aligned} p(y \mid m_0) &= \int p(y \mid \varphi, m_0) p(\varphi \mid m_0) d\varphi \\ &= \int p(y \mid \varphi, \delta = \delta_0, m_1) p(\varphi \mid \delta = \delta_0, m_1) d\varphi \\ &= p(y \mid \delta = \delta_0, m_1). \end{aligned} \tag{3}$$

By Bayes’ theorem (\rightarrow Proof I/5.2.1), we can rewrite this last line as

$$p(y \mid \delta = \delta_0, m_1) = \frac{p(\delta = \delta_0 \mid y, m_1) p(y \mid m_1)}{p(\delta = \delta_0 \mid m_1)}. \tag{4}$$

Thus we have

$$\begin{aligned} \text{BF}_{01} &\stackrel{(2)}{=} \frac{p(y \mid m_0)}{p(y \mid m_1)} \\ &= p(y \mid m_0) \cdot \frac{1}{p(y \mid m_1)} \\ &\stackrel{(3)}{=} p(y \mid \delta = \delta_0, m_1) \cdot \frac{1}{p(y \mid m_1)} \\ &\stackrel{(4)}{=} \frac{p(\delta = \delta_0 \mid y, m_1) p(y \mid m_1)}{p(\delta = \delta_0 \mid m_1)} \cdot \frac{1}{p(y \mid m_1)} \\ &= \frac{p(\delta = \delta_0 \mid y, m_1)}{p(\delta = \delta_0 \mid m_1)}, \end{aligned} \tag{5}$$

which completes the proof of (1).

Sources:

- Faulkenberry, Thomas J. (2019): “A tutorial on generalizing the default Bayesian t-test via posterior sampling and encompassing priors”; in: *Communications for Statistical Applications and Methods*, vol. 26, no. 2, pp. 217-238; URL: <https://dx.doi.org/10.29220/CSAM.2019.26.2.217>; DOI: 10.29220/CSAM.2019.26.2.217.
- Penny, W.D. and Ridgway, G.R. (2013): “Efficient Posterior Probability Mapping Using Savage-Dickey Ratios”; in: *PLoS ONE*, vol. 8, iss. 3, art. e59655, eq. 16; URL: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0059655>; DOI: 10.1371/journal.pone.0059655.

Metadata: ID: P156 | shortcut: bf-sddr | author: tomfaulkenberry | date: 2020-08-26, 12:00.

3.3.4 Computation using Encompassing Prior Method

Theorem: Consider two models m_1 and m_e , where m_1 is nested within an encompassing model (\rightarrow Definition IV/3.3.5) m_e via an inequality constraint on some parameter θ , and θ is unconstrained under m_e . Then, the Bayes factor (\rightarrow Definition IV/3.3.1) is

$$\text{BF}_{1e} = \frac{c}{d} = \frac{1/d}{1/c} \quad (1)$$

where $1/d$ and $1/c$ represent the proportions of the posterior and prior of the encompassing model, respectively, that are in agreement with the inequality constraint imposed by the nested model m_1 .

Proof: Consider first that for any model m_1 on data y with parameter θ , Bayes' theorem (\rightarrow Proof I/5.2.1) implies

$$p(\theta \mid y, m_1) = \frac{p(y \mid \theta, m_1) \cdot p(\theta \mid m_1)}{p(y \mid m_1)}. \quad (2)$$

Rearranging equation (2) allows us to write the marginal likelihood (\rightarrow Definition I/5.1.9) for y under m_1 as

$$p(y \mid m_1) = \frac{p(y \mid \theta, m_1) \cdot p(\theta \mid m_1)}{p(\theta \mid y, m_1)}. \quad (3)$$

Taking the ratio of the marginal likelihoods for m_1 and the encompassing model (\rightarrow Definition IV/3.3.5) m_e yields the following Bayes factor (\rightarrow Definition IV/3.3.1):

$$\text{BF}_{1e} = \frac{p(y \mid \theta, m_1) \cdot p(\theta \mid m_1) / p(\theta \mid y, m_1)}{p(y \mid \theta, m_e) \cdot p(\theta \mid m_e) / p(\theta \mid y, m_e)}. \quad (4)$$

Now, both the constrained model m_1 and the encompassing model (\rightarrow Definition IV/3.3.5) m_e contain the same parameter vector θ . Choose a specific value of θ , say θ' , that exists in the support of both models m_1 and m_e (we can do this, because m_1 is nested within m_e). Then, for this parameter value θ' , we have $p(y \mid \theta', m_1) = p(y \mid \theta', m_e)$, so the expression for the Bayes factor in equation (4) reduces to an expression involving only the priors and posteriors for θ' under m_1 and m_e :

$$\text{BF}_{1e} = \frac{p(\theta' \mid m_1) / p(\theta' \mid y, m_1)}{p(\theta' \mid m_e) / p(\theta' \mid y, m_e)}. \quad (5)$$

Because m_1 is nested within m_e via an inequality constraint, the prior $p(\theta' \mid m_1)$ is simply a truncation of the encompassing prior $p(\theta' \mid m_e)$. Thus, we can express $p(\theta' \mid m_1)$ in terms of the encompassing prior $p(\theta' \mid m_e)$ by multiplying the encompassing prior by an indicator function over m_1 and then normalizing the resulting product. That is,

$$\begin{aligned} p(\theta' \mid m_1) &= \frac{p(\theta' \mid m_e) \cdot I_{\theta' \in m_1}}{\int p(\theta' \mid m_e) \cdot I_{\theta' \in m_1} d\theta'} \\ &= \left(\frac{I_{\theta' \in m_1}}{\int p(\theta' \mid m_e) \cdot I_{\theta' \in m_1} d\theta'} \right) \cdot p(\theta' \mid m_e), \end{aligned} \quad (6)$$

where $I_{\theta' \in m_1}$ is an indicator function. For parameters $\theta' \in m_1$, this indicator function is identically equal to 1, so the expression in parentheses reduces to a constant, say c , allowing us to write the prior as

$$p(\theta' \mid m_1) = c \cdot p(\theta' \mid m_e). \quad (7)$$

By similar reasoning, we can write the posterior as

$$p(\theta' | y, m_1) = \left(\frac{I_{\theta' \in m_1}}{\int p(\theta' | y, m_e) \cdot I_{\theta' \in m_1} d\theta'} \right) \cdot p(\theta' | y, m_e) = d \cdot p(\theta' | y, m_e). \quad (8)$$

Plugging (7) and (8) into (5), this gives us

$$\text{BF}_{1e} = \frac{c \cdot p(\theta' | m_e) / d \cdot p(\theta' | y, m_e)}{p(\theta' | m_e) / p(\theta' | y, m_e)} = \frac{c}{d} = \frac{1/d}{1/c}, \quad (9)$$

which completes the proof. Note that by definition, $1/d$ represents the proportion of the posterior distribution for θ under the encompassing model (\rightarrow Definition IV/3.3.5) m_e that agrees with the constraints imposed by m_1 . Similarly, $1/c$ represents the proportion of the prior distribution for θ under the encompassing model (\rightarrow Definition IV/3.3.5) m_e that agrees with the constraints imposed by m_1 .

Sources:

- Klugkist, I., Kato, B., and Hoijtink, H. (2005): “Bayesian model selection using encompassing priors”; in: *Statistica Neerlandica*, vol. 59, no. 1., pp. 57-69; URL: <https://dx.doi.org/10.1111/j.1467-9574.2005.00279.x>; DOI: 10.1111/j.1467-9574.2005.00279.x.
- Faulkenberry, Thomas J. (2019): “A tutorial on generalizing the default Bayesian t-test via posterior sampling and encompassing priors”; in: *Communications for Statistical Applications and Methods*, vol. 26, no. 2, pp. 217-238; URL: <https://dx.doi.org/10.29220/CSAM.2019.26.2.217>; DOI: 10.29220/CSAM.2019.26.2.217.

Metadata: ID: P157 | shortcut: bf-ep | author: tomfaulkenberry | date: 2020-09-02, 12:00.

3.3.5 Encompassing model

Definition: Consider a family f of generative models (\rightarrow Definition I/5.1.1) m on data y , where each $m \in f$ is defined by placing an inequality constraint on model parameter(s) θ (e.g., $m : \theta > 0$). Then the encompassing model m_e is constructed such that each m is nested within m_e and all inequality constraints on the parameter(s) θ are removed.

Sources:

- Klugkist, I., Kato, B., and Hoijtink, H. (2005): “Bayesian model selection using encompassing priors”; in: *Statistica Neerlandica*, vol. 59, no. 1, pp. 57-69; URL: <https://dx.doi.org/10.1111/j.1467-9574.2005.00279.x>; DOI: 10.1111/j.1467-9574.2005.00279.x.

Metadata: ID: D93 | shortcut: encm | author: tomfaulkenberry | date: 2020-09-02, 12:00.

3.4 Log Bayes factor

3.4.1 Definition

Definition: Let there be two generative models (\rightarrow Definition I/5.1.1) m_1 and m_2 which are mutually exclusive, but not necessarily collectively exhaustive:

$$\neg(m_1 \wedge m_2) \quad (1)$$

Then, the Bayes factor in favor of m_1 and against m_2 is the ratio of the model evidences (\rightarrow Definition I/5.1.9) of m_1 and m_2 :

$$\text{BF}_{12} = \frac{p(y|m_1)}{p(y|m_2)}. \quad (2)$$

The log Bayes factor is given by the logarithm of the Bayes factor:

$$\text{LBF}_{12} = \log \text{BF}_{12} = \log \frac{p(y|m_1)}{p(y|m_2)}. \quad (3)$$

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: D84 | shortcut: lbf | author: JoramSoch | date: 2020-07-22, 07:02.

3.4.2 Derivation

Theorem: Let there be two generative models (\rightarrow Definition I/5.1.1) m_1 and m_2 with model evidences (\rightarrow Definition I/5.1.9) $p(y|m_1)$ and $p(y|m_2)$. Then, the log Bayes factor (\rightarrow Definition IV/3.4.1)

$$\text{LBF}_{12} = \log \text{BF}_{12} \quad (1)$$

can be expressed as

$$\text{LBF}_{12} = \log \frac{p(y|m_1)}{p(y|m_2)}. \quad (2)$$

Proof: The Bayes factor (\rightarrow Definition IV/3.3.1) is defined as the posterior (\rightarrow Definition I/5.1.7) odds ratio (\rightarrow Definition “odds”) when both models (\rightarrow Definition I/5.1.1) are equally likely apriori (\rightarrow Definition I/5.1.3):

$$\text{BF}_{12} = \frac{p(m_1|y)}{p(m_2|y)} \quad (3)$$

Plugging in the posterior odds ratio according to Bayes’ rule (\rightarrow Proof I/5.2.2), we have

$$\text{BF}_{12} = \frac{p(y|m_1)}{p(y|m_2)} \cdot \frac{p(m_1)}{p(m_2)}. \quad (4)$$

When both models are equally likely apriori, the prior (\rightarrow Definition I/5.1.3) odds ratio (\rightarrow Definition “odds”) is one, such that

$$\text{BF}_{12} = \frac{p(y|m_1)}{p(y|m_2)}. \quad (5)$$

Equation (2) follows by logarithmizing both sides of (5).

Sources:

- original work

Metadata: ID: P137 | shortcut: lbf-der | author: JoramSoch | date: 2020-07-22, 07:27.

3.4.3 Calculation from 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 IV/3.4.1) in favor of model m_1 and against model m_2 is the difference of the log model evidences:

$$\text{LBF}_{12} = \text{LME}(m_1) - \text{LME}(m_2) . \quad (1)$$

Proof: The Bayes factor (\rightarrow Definition IV/3.3.1) is defined as the ratio of the model evidences (\rightarrow Definition I/5.1.9) of m_1 and m_2

$$\text{BF}_{12} = \frac{p(y|m_1)}{p(y|m_2)} \quad (2)$$

and the log Bayes factor (\rightarrow Definition IV/3.4.1) is defined as the logarithm of the Bayes factor

$$\text{LBF}_{12} = \log \text{BF}_{12} = \log \frac{p(y|m_1)}{p(y|m_2)} . \quad (3)$$

With the definition of the log model evidence (\rightarrow Definition IV/3.1.1)

$$\text{LME}(m) = \log p(y|m) \quad (4)$$

the log Bayes factor can be expressed as:

Resolving the logarithm and applying the definition of the log model evidence (\rightarrow Definition IV/3.1.1), we finally have:

$$\begin{aligned} \text{LBF}_{12} &= \log p(y|m_1) - \log p(y|m_2) \\ &= \text{LME}(m_1) - \text{LME}(m_2) . \end{aligned} \quad (5)$$

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.

3.5 Posterior model probability

3.5.1 Definition

Definition: Let m_1, \dots, m_M be M statistical models (\rightarrow Definition I/5.1.4) with model evidences (\rightarrow Definition I/5.1.9) $p(y|m_1), \dots, p(y|m_M)$ and prior probabilities (\rightarrow Definition I/5.1.3) $p(m_1), \dots, p(m_M)$. Then, the conditional probability (\rightarrow Definition I/1.2.4) of model m_i , given the data y , is called the posterior probability (\rightarrow Definition I/5.1.7) of model m_i :

$$\text{PP}(m_i) = p(m_i|y) . \quad (1)$$

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. 23; URL: <https://www.sciencedirect.com/science/article/pii/S0165027018301468>; DOI: 10.1016/j.jneumeth.2018.05.017.

Metadata: ID: D87 | shortcut: pmp | author: JoramSoch | date: 2020-07-28, 03:30.

3.5.2 Derivation

Theorem: Let there be a set of generative models (\rightarrow Definition I/5.1.1) m_1, \dots, m_M with model evidences (\rightarrow Definition I/5.1.9) $p(y|m_1), \dots, p(y|m_M)$ and prior probabilities (\rightarrow Definition I/5.1.3) $p(m_1), \dots, p(m_M)$. Then, the posterior probability (\rightarrow Definition IV/3.5.1) of model m_i is given by

$$p(m_i|y) = \frac{p(y|m_i) p(m_i)}{\sum_{j=1}^M p(y|m_j) p(m_j)}, \quad i = 1, \dots, M . \quad (1)$$

Proof: From Bayes’ theorem (\rightarrow Proof I/5.2.1), the posterior model probability (\rightarrow Definition IV/3.5.1) of the i -th model 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 I/1.2.3), the denominator can be rewritten, such that

$$p(m_i|y) = \frac{p(y|m_i) p(m_i)}{\sum_{j=1}^M p(y, m_j)} . \quad (3)$$

Finally, using the law of conditional probability (\rightarrow Definition I/1.2.4), we have

$$p(m_i|y) = \frac{p(y|m_i) p(m_i)}{\sum_{j=1}^M p(y|m_j) p(m_j)} . \quad (4)$$

Sources:

- original work

Metadata: ID: P139 | shortcut: pmp-der | author: JoramSoch | date: 2020-07-28, 03:58.

3.5.3 Calculation from 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 IV/3.5.1) 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 IV/3.3.1) comparing model m_i with m_0 and α_i is the prior (\rightarrow Definition I/5.1.3) odds ratio (\rightarrow Definition “odds”) of model m_i against m_0 .

Proof: Define the Bayes factor (\rightarrow Definition IV/3.3.1) for m_i

$$\text{BF}_{i,0} = \frac{p(y|m_i)}{p(y|m_0)} \quad (2)$$

and prior odds ratio of m_i against m_0

$$\alpha_i = \frac{p(m_i)}{p(m_0)} . \quad (3)$$

The posterior model probability (\rightarrow Proof IV/3.5.2) of m_i is given by

$$p(m_i|y) = \frac{p(y|m_i) \cdot p(m_i)}{\sum_{j=1}^M p(y|m_j) \cdot p(m_j)} . \quad (4)$$

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.

3.5.4 Calculation from log Bayes factor

Theorem: Let m_1 and m_2 be two statistical models with the log Bayes factor (\rightarrow Definition IV/3.4.1) LBF_{12} in favor of model m_1 and against model m_2 . Then, if both models are equally likely apriori (\rightarrow Definition I/5.1.3), the posterior model probability (\rightarrow Definition IV/3.5.1) of m_1 is

$$p(m_1|y) = \frac{\exp(\text{LBF}_{12})}{\exp(\text{LBF}_{12}) + 1} . \quad (1)$$

Proof: From Bayes’ rule (\rightarrow Proof I/5.2.2), the posterior odds ratio (\rightarrow Definition “odds”) 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 (\rightarrow Definition I/5.1.3), the prior odds ratio (\rightarrow Definition “odds”) 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 IV/3.3.1), therefore

$$\frac{p(m_1|y)}{p(m_2|y)} = \text{BF}_{12} . \quad (4)$$

Because the two posterior model probabilities (\rightarrow Definition IV/3.5.1) add up to 1, we have

$$\frac{p(m_1|y)}{1 - p(m_1|y)} = \text{BF}_{12} . \quad (5)$$

Now rearranging for the posterior probability (\rightarrow Definition IV/3.5.1), this gives

$$p(m_1|y) = \frac{\text{BF}_{12}}{\text{BF}_{12} + 1} . \quad (6)$$

Because the log Bayes factor is the logarithm of the Bayes factor (\rightarrow Definition IV/3.4.1), we finally have

$$p(m_1|y) = \frac{\exp(\text{LBF}_{12})}{\exp(\text{LBF}_{12}) + 1} . \quad (7)$$

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.

3.5.5 Calculation from 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 IV/3.5.1) 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 (\rightarrow Definition I/5.1.3) model probabilities.

Proof: The posterior model probability (\rightarrow Proof IV/3.5.2) can be derived as

$$p(m_i|y) = \frac{p(y|m_i) p(m_i)}{\sum_{j=1}^M p(y|m_j) p(m_j)} . \quad (2)$$

The definition of the log model evidence (\rightarrow Definition IV/3.1.1)

$$\text{LME}(m) = \log p(y|m) \quad (3)$$

can be exponentiated to then read

$$\exp[\text{LME}(m)] = p(y|m) \quad (4)$$

and applying (4) to (2), we finally have:

$$p(m_i|y) = \frac{\exp[\text{LME}(m_i)] p(m_i)}{\sum_{j=1}^M \exp[\text{LME}(m_j)] p(m_j)} . \quad (5)$$

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. 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.

3.6 Bayesian model averaging

3.6.1 Definition

Definition: Let m_1, \dots, m_M be M statistical models (\rightarrow Definition I/5.1.4) with posterior model probabilities (\rightarrow Definition IV/3.5.1) $p(m_1|y), \dots, p(m_M|y)$ and posterior distributions (\rightarrow Definition I/5.1.7) $p(\theta|y, m_1), \dots, p(\theta|y, m_M)$. Then, Bayesian model averaging (BMA) consists in finding the marginal (\rightarrow Definition I/1.3.3) posterior (\rightarrow Definition I/5.1.7) density (\rightarrow Definition I/1.4.2), conditional (\rightarrow Definition I/1.2.4) on the measured data y , but unconditional (\rightarrow Definition I/1.2.3) on the modelling approach m :

$$p(\theta|y) = \sum_{i=1}^M p(\theta|y, m_i) \cdot p(m_i|y) . \quad (1)$$

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. 1; URL: <https://projecteuclid.org/euclid.ss/1009212519>; DOI: 10.1214/ss/1009212519.

Metadata: ID: D89 | shortcut: bma | author: JoramSoch | date: 2020-08-03, 21:34.

3.6.2 Derivation

Theorem: Let m_1, \dots, m_M be M statistical models (\rightarrow Definition I/5.1.4) with posterior model probabilities (\rightarrow Definition IV/3.5.1) $p(m_1|y), \dots, p(m_M|y)$ and posterior distributions (\rightarrow Definition I/5.1.7) $p(\theta|y, m_1), \dots, p(\theta|y, m_M)$. Then, the marginal (\rightarrow Definition I/1.3.3) posterior (\rightarrow Definition I/5.1.7) density (\rightarrow Definition I/1.4.2), conditional (\rightarrow Definition I/1.2.4) on the measured data y , but unconditional (\rightarrow Definition I/1.2.3) on the modelling approach m , is given by:

$$p(\theta|y) = \sum_{i=1}^M p(\theta|y, m_i) \cdot p(m_i|y) . \quad (1)$$

Proof: Using the law of marginal probability (\rightarrow Definition I/1.2.3), the probability distribution of the shared parameters θ conditional (\rightarrow Definition I/1.2.4) on the measured data y can be obtained by marginalizing (\rightarrow Definition I/1.2.3) over the discrete random variable (\rightarrow Definition I/1.1.1) model m :

$$p(\theta|y) = \sum_{i=1}^M p(\theta, m_i|y) . \quad (2)$$

Using the law of the conditional probability (\rightarrow Definition I/1.2.4), the summand can be expanded to give

$$p(\theta|y) = \sum_{i=1}^M p(\theta|y, m_i) \cdot p(m_i|y) \quad (3)$$

where $p(\theta|y, m_i)$ is the posterior distribution (\rightarrow Definition I/5.1.7) of the i -th model and $p(m_i|y)$ happens to be the posterior probability (\rightarrow Definition IV/3.5.1) of the i -th model.

Sources:

- original work

Metadata: ID: P143 | shortcut: bma-der | author: JoramSoch | date: 2020-08-03, 22:05.

3.6.3 Calculation from log model evidences

Theorem: Let m_1, \dots, m_M be M statistical models (\rightarrow Definition I/5.1.4) 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 (\rightarrow Definition IV/3.6.1) 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 I/1.2.3), the probability of the shared parameters θ conditional on the measured data y can be obtained (\rightarrow Proof IV/3.6.2) 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 IV/3.5.1) of the i -th model. One can express posterior model probabilities in terms of log model evidences (\rightarrow Proof IV/3.5.5) 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.