The Book of Statistical Proofs

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

2020-02-06, 12:01

Contents

Chapter I

General Theorems

1 Probability theory

1.1 Probability distributions

1.1.1 Moment-generating function

Definition: Moment-generating function

Index: The Book of Statistical Proofs ▷ General Theorems ▷ Probability theory ▷ Probability distributions ▷ Moment-generating function

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

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

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

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

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

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

1.2 Bayesian inference

1.2.1 Bayes' theorem

Proof: Bayes' theorem

Index: The Book of Statistical Proofs ▷ General Theorems ▷ Probability theory ▷ Bayesian inference ▷ Bayes' theorem

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

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

Proof: The conditional probability is defined as the ratio of joint probability, i.e. the probability of both statements being true, and marginal probability, i.e. the probability of only the second one being true:

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

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

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

Both equations can be rearranged for the joint probability

$$p(A|B) \, p(B) \overset{(4)}{=} p(A,B) \overset{(5)}{=} p(B|A) \, p(A) \tag{6}$$

from which Bayes' theorem can be directly derived:

$$p(A|B) \overset{(6)}{=} \frac{p(B|A) \, p(A)}{p(B)} \; . \tag{7}$$

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.

1.2.2 Bayes' rule

Proof: Bayes' rule

Index: The Book of Statistical Proofs ▷ General Theorems ▷ Probability theory ▷ Bayesian inference ▷ Bayes' theorem

Theorem: Let A_1, A_2 and B be arbitrary statements about random variables where A_1 and A_2 are mutually exclusive. Then, Bayes' rule states that the posterior odds are equal to the Bayes factor times the prior odds, i.e.

$$\frac{p(A_1|B)}{p(A_2|B)} = \frac{p(B|A_1)}{p(B|A_2)} \cdot \frac{p(A_1)}{p(A_2)} \; . \tag{8}$$

Proof: Using Bayes' theorem, the conditional probabilities on the left are given by

$$p(A_1|B) = \frac{p(B|A_1) \cdot p(A_1)}{p(B)} \tag{9}$$

$$p(A_2|B) = \frac{p(B|A_2) \cdot p(A_2)}{p(B)} \; . \tag{10}$$

Dividing the two condition probabilities by each other

$$\begin{aligned}\frac{p(A_1|B)}{p(A_2|B)} &= \frac{p(B|A_1) \cdot p(A_1)/p(B)}{p(B|A_2) \cdot p(A_2)/p(B)} \\ &= \frac{p(B|A_1)}{p(B|A_2)} \cdot \frac{p(A_1)}{p(A_2)} \; ,\end{aligned} \tag{11}$$

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

Sources:

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

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

2 Estimation theory

2.1 Point estimates

2.1.1 Partition of the mean squared error into bias and variance

Proof: Partition of the mean squared error into bias and variance

Index: The Book of Statistical Proofs ▷ General Theorems ▷ Estimation theory ▷ Point estimates ▷ Partition of the mean squared error into bias and variance

Theorem: The mean squared error can be partitioned into variance and squared bias

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) - \mathrm{Bias}(\hat{\theta}, \theta)^2 \tag{12}$$

where the variance is given by

$$\mathrm{Var}(\hat{\theta}) = \mathbb{E}_{\hat{\theta}}\left[\left(\hat{\theta} - \mathbb{E}_{\hat{\theta}}(\hat{\theta})\right)^2\right] \tag{13}$$

and the bias is given by

$$\mathrm{Bias}(\hat{\theta}, \theta) = \left(\mathbb{E}_{\hat{\theta}}(\hat{\theta}) - \theta\right) \; . \tag{14}$$

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

$$\mathrm{MSE}(\hat{\theta}) = \mathbb{E}_{\hat{\theta}}\left[\left(\hat{\theta} - \theta\right)^2\right] \; . \tag{15}$$

This formula can be evaluated in the following way:

$$\begin{aligned}
\mathrm{MSE}(\hat{\theta}) &= \mathbb{E}_{\hat{\theta}}\left[\left(\hat{\theta} - \theta\right)^2\right] \\
&= \mathbb{E}_{\hat{\theta}}\left[\left(\hat{\theta} - \mathbb{E}_{\hat{\theta}}(\hat{\theta}) + \mathbb{E}_{\hat{\theta}}(\hat{\theta}) - \theta\right)^2\right] \\
&= \mathbb{E}_{\hat{\theta}}\left[\left(\hat{\theta} - \mathbb{E}_{\hat{\theta}}(\hat{\theta})\right)^2 + 2\left(\hat{\theta} - \mathbb{E}_{\hat{\theta}}(\hat{\theta})\right)\left(\mathbb{E}_{\hat{\theta}}(\hat{\theta}) - \theta\right) + \left(\mathbb{E}_{\hat{\theta}}(\hat{\theta}) - \theta\right)^2\right] \\
&= \mathbb{E}_{\hat{\theta}}\left[\left(\hat{\theta} - \mathbb{E}_{\hat{\theta}}(\hat{\theta})\right)^2\right] + \mathbb{E}_{\hat{\theta}}\left[2\left(\hat{\theta} - \mathbb{E}_{\hat{\theta}}(\hat{\theta})\right)\left(\mathbb{E}_{\hat{\theta}}(\hat{\theta}) - \theta\right)\right] + \mathbb{E}_{\hat{\theta}}\left[\left(\mathbb{E}_{\hat{\theta}}(\hat{\theta}) - \theta\right)^2\right] \; .
\end{aligned} \tag{16}$$

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

$$\begin{aligned}
\mathrm{MSE}(\hat{\theta}) &= \mathbb{E}_{\hat{\theta}}\left[\left(\hat{\theta} - \mathbb{E}_{\hat{\theta}}(\hat{\theta})\right)^2\right] + 2\left(\mathbb{E}_{\hat{\theta}}(\hat{\theta}) - \theta\right)\mathbb{E}_{\hat{\theta}}\left[\hat{\theta} - \mathbb{E}_{\hat{\theta}}(\hat{\theta})\right] + \left(\mathbb{E}_{\hat{\theta}}(\hat{\theta}) - \theta\right)^2 \\
&= \mathbb{E}_{\hat{\theta}}\left[\left(\hat{\theta} - \mathbb{E}_{\hat{\theta}}(\hat{\theta})\right)^2\right] + 2\left(\mathbb{E}_{\hat{\theta}}(\hat{\theta}) - \theta\right)\left(\mathbb{E}_{\hat{\theta}}(\hat{\theta}) - \mathbb{E}_{\hat{\theta}}(\hat{\theta})\right) + \left(\mathbb{E}_{\hat{\theta}}(\hat{\theta}) - \theta\right)^2 \\
&= \mathbb{E}_{\hat{\theta}}\left[\left(\hat{\theta} - \mathbb{E}_{\hat{\theta}}(\hat{\theta})\right)^2\right] + \left(\mathbb{E}_{\hat{\theta}}(\hat{\theta}) - \theta\right)^2 \; .
\end{aligned} \tag{17}$$

This proofs the partition given by (12).

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 Information theory

3.1 Discrete mutual information

3.1.1 Relation to marginal and conditional entropy

Proof: Relation of mutual information to marginal and conditional entropy

Index: The Book of Statistical Proofs ▷ General Theorems ▷ Information theory ▷ Discrete mutual information ▷ Relation to marginal and conditional entropy

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

$$\begin{split} \mathrm{I}(X,Y) &= \mathrm{H}(X) - \mathrm{H}(X|Y) \\ &= \mathrm{H}(Y) - \mathrm{H}(Y|X) \end{split} \tag{18}$$

where $\mathrm{H}(X)$ and $\mathrm{H}(Y)$ are the marginal entropies of X and Y and $\mathrm{H}(X \mid Y)$ and $\mathrm{H}(Y \mid X)$ are the conditional entropies.

Proof: The mutual information of X and Y is defined as

$$\mathrm{I}(X,Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)\,p(y)} \; . \tag{19}$$

Separating the logarithm, we have:

$$\mathrm{I}(X,Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(y)} - \sum_x \sum_y p(x,y) \log p(x) \; . \tag{20}$$

Applying the law of conditional probability, i.e. $p(x,y) = p(x \mid y)\,p(y)$, we get:

$$\mathrm{I}(X,Y) = \sum_x \sum_y p(x|y)\,p(y) \log p(x|y) - \sum_x \sum_y p(x,y) \log p(x) \; . \tag{21}$$

Regrouping the variables, we have:

$$\mathrm{I}(X,Y) = \sum_y p(y) \sum_x p(x|y) \log p(x|y) - \sum_x \left(\sum_y p(x,y) \right) \log p(x) \; . \tag{22}$$

Applying the law of marginal probability, i.e. $p(x) = \sum_y p(x,y)$, we get:

$$\mathrm{I}(X,Y) = \sum_y p(y) \sum_x p(x|y) \log p(x|y) - \sum_x p(x) \log p(x) \; . \tag{23}$$

Now considering the definitions of marginal and conditional entropy

$$\begin{split} \mathrm{H}(X) &= -\sum_{x \in \mathcal{X}} p(x) \log p(x) \\ \mathrm{H}(X|Y) &= \sum_{y \in \mathcal{Y}} p(y)\, \mathrm{H}(X|Y=y) \; , \end{split} \tag{24}$$

we can finally show:

$$\begin{aligned} \mathrm{I}(X,Y) &= -\mathrm{H}(X|Y) + \mathrm{H}(X) \\ &= \mathrm{H}(X) - \mathrm{H}(X|Y) \; . \end{aligned} \tag{25}$$

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.

3.1.2 Relation to marginal and joint entropy

Proof: Relation of mutual information to marginal and joint entropy

Index: The Book of Statistical Proofs ▷ General Theorems ▷ Information theory ▷ Discrete mutual information ▷ Relation to marginal and joint entropy

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

$$\mathrm{I}(X,Y) = \mathrm{H}(X) + \mathrm{H}(Y) - \mathrm{H}(X,Y) \tag{26}$$

where $\mathrm{H}(X)$ and $\mathrm{H}(Y)$ are the marginal entropies of X and Y and $\mathrm{H}(X,Y)$ is the joint entropy.

Proof: The mutual information of X and Y is defined as

$$\mathrm{I}(X,Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)\, p(y)} \; . \tag{27}$$

Separating the logarithm, we have:

$$\mathrm{I}(X,Y) = \sum_x \sum_y p(x,y) \log p(x,y) - \sum_x \sum_y p(x,y) \log p(x) - \sum_x \sum_y p(x,y) \log p(y) \; . \tag{28}$$

Regrouping the variables, this reads:

$$\mathrm{I}(X,Y) = \sum_x \sum_y p(x,y) \log p(x,y) - \sum_x \left(\sum_y p(x,y) \right) \log p(x) - \sum_y \left(\sum_x p(x,y) \right) \log p(y) \; . \tag{29}$$

Applying the law of marginal probability, i.e. $p(x) = \sum_y p(x,y)$, we get:

$$\mathrm{I}(X,Y) = \sum_x \sum_y p(x,y) \log p(x,y) - \sum_x p(x) \log p(x) - \sum_x p(y) \log p(y) \; . \tag{30}$$

Now considering the definitions of marginal and joint entropy

$$\begin{aligned} \mathrm{H}(X) &= - \sum_{x \in \mathcal{X}} p(x) \log p(x) \\ \mathrm{H}(X,Y) &= - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log p(x,y) \; , \end{aligned} \tag{31}$$

we can finally show:

$$\begin{aligned} \mathrm{I}(X,Y) &= -\mathrm{H}(X,Y) + \mathrm{H}(X) + \mathrm{H}(Y) \\ &= \mathrm{H}(X) + \mathrm{H}(Y) - \mathrm{H}(X,Y) \; . \end{aligned} \tag{32}$$

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.

3.1.3 Relation to joint and conditional entropy

Proof: Relation of mutual information to joint and conditional entropy

Index: The Book of Statistical Proofs ▷ General Theorems ▷ Information theory ▷ Discrete mutual information ▷ Relation to joint and conditional entropy

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

$$\mathrm{I}(X,Y) = \mathrm{H}(X,Y) - \mathrm{H}(X|Y) - \mathrm{H}(Y|X) \tag{33}$$

where $\mathrm{H}(X,Y)$ is the joint entropy of X and Y and $\mathrm{H}(X \mid Y)$ and $\mathrm{H}(Y \mid X)$ are the conditional entropies.

Proof: The existence of the joint probability function ensures that the mutual information is defined:

$$\mathrm{I}(X,Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)\, p(y)} \; . \tag{34}$$

The relation of mutual information to conditional entropy is:

$$\mathrm{I}(X,Y) = \mathrm{H}(X) - \mathrm{H}(X|Y) \tag{35}$$

$$\mathrm{I}(X,Y) = \mathrm{H}(Y) - \mathrm{H}(Y|X) \tag{36}$$

The relation of mutual information to joint entropy is:

$$\mathrm{I}(X,Y) = \mathrm{H}(X) + \mathrm{H}(Y) - \mathrm{H}(X,Y) \; . \tag{37}$$

It is true that

$$\mathrm{I}(X,Y) = \mathrm{I}(X,Y) + \mathrm{I}(X,Y) - \mathrm{I}(X,Y) \; . \tag{38}$$

Plugging in (35), (36) and (37) on the right-hand side, we have

$$\begin{split} \mathrm{I}(X,Y) &= \mathrm{H}(X) - \mathrm{H}(X|Y) + \mathrm{H}(Y) - \mathrm{H}(Y|X) - \mathrm{H}(X) - \mathrm{H}(Y) + \mathrm{H}(X,Y) \\ &= \mathrm{H}(X,Y) - \mathrm{H}(X|Y) - \mathrm{H}(Y|X) \end{split} \tag{39}$$

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.

Chapter II

Probability Distributions

1 Univariate discrete distributions

1.1 Bernoulli distribution

1.1.1 Mean

Proof: Mean of the Bernoulli distribution

Index: The Book of Statistical Proofs ▷ Probability Distributions ▷ Univariate discrete distributions ▷ Bernoulli distribution ▷ Mean

Theorem: Let X be a random variable following a Bernoulli distribution:

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

Then, the mean or expected value of X is

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

Proof: The expected value is the probability-weighted average of all possible values:

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

Since there are only two possible outcomes for a Bernoulli random variable, we have:

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

Sources:

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

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

1.2 Binomial distribution

1.2.1 Mean

Proof: Mean of the binomial distribution

Index: The Book of Statistical Proofs ▷ Probability Distributions ▷ Univariate discrete distributions ▷ Binomial distribution ▷ Mean

Theorem: Let X be a random variable following a binomial distribution:

$$X \sim \mathrm{Bin}(n,p) \; . \tag{5}$$

Then, the mean or expected value of X is

$$\mathrm{E}(X) = np \; . \tag{6}$$

Proof: By definition, a binomial random variable is the sum of n independent and identical Bernoulli trials with success probability p. Therefore, the expected value is

$$\mathrm{E}(X) = \mathrm{E}(X_1 + \ldots + X_n) \tag{7}$$

and because the expected value is a linear operator, this is equal to

$$\begin{split} \mathrm{E}(X) &= \mathrm{E}(X_1) + \ldots + \mathrm{E}(X_n) \\ &= \sum_{i=1}^{n} \mathrm{E}(X_i) \; . \end{split} \tag{8}$$

With the expected value of the Bernoulli distribution, we have:

$$\mathrm{E}(X) = \sum_{i=1}^{n} p = np \; . \tag{9}$$

Sources:

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

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

2 Multivariate discrete distributions

2.1 Categorical distribution

2.1.1 Mean

Proof: Mean of the categorical distribution

Index: The Book of Statistical Proofs ▷ Probability Distributions ▷ Multivariate discrete distributions ▷ Categorical distribution ▷ Mean

Theorem: Let X be a random variable following a categorical distribution:

$$X \sim \mathrm{Cat}([p_1, \ldots, p_k]) \; . \tag{10}$$

Then, the mean or expected value of X is

$$\mathrm{E}(X) = [p_1, \ldots, p_k] \; . \tag{11}$$

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

$$\begin{aligned}
\mathrm{E}(X) &= \sum_{x \in \mathcal{X}} x \cdot \mathrm{Pr}(X = x) \\
&= \sum_{i=1}^{k} e_i \cdot \mathrm{Pr}(X = e_i) \\
&= \sum_{i=1}^{k} e_i \cdot p_i \\
&= [p_1, \ldots, p_k] \; .
\end{aligned} \tag{12}$$

Sources:
- own work

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

2.2 Multinomial distribution

2.2.1 Mean

Proof: Mean of the multinomial distribution

Index: The Book of Statistical Proofs ▷ Probability Distributions ▷ Multivariate discrete distributions ▷ Multinomial distribution ▷ Mean

Theorem: Let X be a random variable following a multinomial distribution:

$$X \sim \mathrm{Mult}(n, [p_1, \ldots, p_k]) \; . \tag{13}$$

Then, the mean or expected value of X is

$$\mathrm{E}(X) = [np_1, \ldots, np_k] \; . \tag{14}$$

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

$$\mathrm{E}(X) = \mathrm{E}(X_1 + \ldots + X_n) \tag{15}$$

and because the expected value is a linear operator, this is equal to

$$\begin{split} \mathrm{E}(X) &= \mathrm{E}(X_1) + \ldots + \mathrm{E}(X_n) \\ &= \sum_{i=1}^{n} \mathrm{E}(X_i) \; . \end{split} \tag{16}$$

With the expected value of the categorical distribution, we have:

$$\mathrm{E}(X) = \sum_{i=1}^{n} [p_1, \ldots, p_k] = n \cdot [p_1, \ldots, p_k] = [np_1, \ldots, np_k] \; . \tag{17}$$

Sources:
- own 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: Continuous uniform distribution

Index: The Book of Statistical Proofs ▷ Probability Distributions ▷ Univariate continuous distributions ▷ Continuous uniform distribution ▷ Definition

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

$$X \sim \mathcal{U}(a, b) \; , \tag{18}$$

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

Sources:

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

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

3.1.2 Probability density function

Proof: Probability density function of the continuous uniform distribution

Index: The Book of Statistical Proofs ▷ Probability Distributions ▷ Univariate continuous distributions ▷ Continuous uniform distribution ▷ Probability density function

Theorem: Let X be a random variable following a continuous uniform distribution:

$$X \sim \mathcal{U}(a, b) \; . \tag{19}$$

Then, the probability density function of X is

$$f_X(x) = \left\{ \begin{array}{rl} \frac{1}{b-a} \; , & \text{if} \; a \leq x \leq b \\ 0 \; , & \text{otherwise} \; . \end{array} \right. \tag{20}$$

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

$$\begin{split} f_X(x) \propto 1 \quad & \text{for all} \quad x \in [a, b] \quad \text{and} \\ f_X(x) = 0, \quad & \text{if} \quad x < a \quad \text{or} \quad x > b \; . \end{split} \tag{21}$$

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

$$f_X(x) = \frac{1}{c(a,b)} \quad \text{for all} \quad x \in [a,b] \tag{22}$$

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

$$\frac{1}{c(a,b)} \int_a^b 1 \,\mathrm{d}x = 1 \; . \tag{23}$$

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

$$\begin{split} \int_a^b 1 \,\mathrm{d}x &= c(a,b) \\ [x]_a^b &= c(a,b) \\ c(a,b) &= b-a \; . \end{split} \tag{24}$$

Sources:

- own work

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

3.1.3 Cumulative distribution function

Proof: Cumulative distribution function of the continuous uniform distribution

Index: The Book of Statistical Proofs ▷ Probability Distributions ▷ Univariate continuous distributions ▷ Continuous uniform distribution ▷ Cumulative distribution function

Theorem: Let X be a random variable following a continuous uniform distribution:

$$X \sim \mathcal{U}(a,b) \; . \tag{25}$$

Then, the cumulative distribution function of X is

$$F_X(x) = \left\{ \begin{array}{rl} 0 \; , & \text{if } x < a \\ \frac{x-a}{b-a} \; , & \text{if } a \leq x \leq b \\ 1 \; , & \text{if } x > b \; . \end{array} \right. \tag{26}$$

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

$$f_X(x) = \left\{ \begin{array}{rl} \frac{1}{b-a} \; , & \text{if } a \leq x \leq b \\ 0 \; , & \text{otherwise} \; . \end{array} \right. \tag{27}$$

Thus, the cumulative distribution function (→ Definition "cdf") is:

$$F_X(x) = \int_{-\infty}^{x} \mathcal{U}(z;a,b) \,\mathrm{d}z \tag{28}$$

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

$$F_X(x) = \int_{-\infty}^{x} 0 \, \mathrm{d}z = 0 \; . \tag{29}$$

Moreover, if $a \leq x \leq b$, we have using (27)

$$\begin{split} F_X(x) &= \int_{-\infty}^{a} \mathcal{U}(z; a, b) \, \mathrm{d}z + \int_{a}^{x} \mathcal{U}(z; a, b) \, \mathrm{d}z \\ &= \int_{-\infty}^{a} 0 \, \mathrm{d}z + \int_{a}^{x} \frac{1}{b-a} \, \mathrm{d}z \\ &= 0 + \frac{1}{b-a} [z]_a^x \\ &= \frac{x-a}{b-a} \; . \end{split} \tag{30}$$

Finally, if $x > b$, we have

$$\begin{split} F_X(x) &= \int_{-\infty}^{b} \mathcal{U}(z; a, b) \, \mathrm{d}z + \int_{b}^{x} \mathcal{U}(z; a, b) \, \mathrm{d}z \\ &= F_X(b) + \int_{b}^{x} 0 \, \mathrm{d}z \\ &= \frac{b-a}{b-a} + 0 \\ &= 1 \; . \end{split} \tag{31}$$

This completes the proof.

Sources:
- own work

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

3.1.4 Quantile function

Proof: Quantile function of the continuous uniform distribution

Index: The Book of Statistical Proofs ▷ Probability Distributions ▷ Univariate continuous distributions ▷ Continuous uniform distribution ▷ Quantile function

Theorem: Let X be a random variable following a continuous uniform distribution:

$$X \sim \mathcal{U}(a, b) \; . \tag{32}$$

Then, the quantile function of X is

$$Q_X(p) = bp + a(1-p) \; . \tag{33}$$

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

$$F_X(x) = \begin{cases} 0 \; , & \text{if } x < a \\ \frac{x-a}{b-a} \; , & \text{if } a \leq x \leq b \\ 1 \; , & \text{if } x > b \; . \end{cases} \tag{34}$$

Thus, the quantile function (→ Definition "qf") is:

$$Q_X(p) = F_X^{-1}(x) \; . \tag{35}$$

This can be derived by rearranging equation (34):

$$\begin{aligned} p &= \frac{x-a}{b-a} \\ x &= p(b-a) + a \\ x &= bp + a(1-p) = Q_X(p) \; . \end{aligned} \tag{36}$$

Sources:

- own work

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

3.2 Normal distribution

3.2.1 Definition

Definition: Normal distribution

Index: The Book of Statistical Proofs ▷ Probability Distributions ▷ Univariate continuous distributions ▷ Normal distribution ▷ Definition

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

$$X \sim \mathcal{N}(\mu, \sigma^2) \; , \tag{37}$$

if and only if its probability density function is given by

$$f_X(x) = \mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right] \tag{38}$$

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

Sources:

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

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

3.2.2 Probability density function

Proof: Probability density function of the normal distribution

Index: The Book of Statistical Proofs ▷ Probability Distributions ▷ Univariate continuous distributions ▷ Normal distribution ▷ Probability density function

Theorem: Let X be a random variable following a normal distribution:

$$X \sim \mathcal{N}(\mu, \sigma^2) \; . \tag{39}$$

Then, the probability density function of X is

$$f_X(x) = \mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right] \; . \tag{40}$$

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

Sources:
- own work

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

3.2.3 Mean

Proof: Mean of the normal distribution

Index: The Book of Statistical Proofs ▷ Probability Distributions ▷ Univariate continuous distributions ▷ Normal distribution ▷ Mean

Theorem: Let X be a random variable following a normal distribution:

$$X \sim \mathcal{N}(\mu, \sigma^2) \; . \tag{41}$$

Then, the mean or expected value of X is

$$\mathrm{E}(X) = \mu \; . \tag{42}$$

Proof: The expected value is the probability-weighted average over all possible values:

$$\mathrm{E}(X) = \int_{\mathbb{R}} x \cdot f_X(x) \, \mathrm{d}x \; . \tag{43}$$

With the probability density function of the normal distribution, this reads:

$$\begin{aligned}
\mathrm{E}(X) &= \int_{-\infty}^{+\infty} x \cdot \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} x \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \mathrm{d}x \; .
\end{aligned} \tag{44}$$

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

$$\begin{aligned}
\mathrm{E}(X) &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty-\mu}^{+\infty-\mu} (z+\mu) \cdot \exp\left[-\frac{1}{2}\left(\frac{z}{\sigma}\right)^2\right] \mathrm{d}(z+\mu) \\
&= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} (z+\mu) \cdot \exp\left[-\frac{1}{2}\left(\frac{z}{\sigma}\right)^2\right] \mathrm{d}z \\
&= \frac{1}{\sqrt{2\pi}\sigma} \left(\int_{-\infty}^{+\infty} z \cdot \exp\left[-\frac{1}{2}\left(\frac{z}{\sigma}\right)^2\right] \mathrm{d}z + \mu \int_{-\infty}^{+\infty} \exp\left[-\frac{1}{2}\left(\frac{z}{\sigma}\right)^2\right] \mathrm{d}z\right) \\
&= \frac{1}{\sqrt{2\pi}\sigma} \left(\int_{-\infty}^{+\infty} z \cdot \exp\left[-\frac{1}{2\sigma^2} \cdot z^2\right] \mathrm{d}z + \mu \int_{-\infty}^{+\infty} \exp\left[-\frac{1}{2\sigma^2} \cdot z^2\right] \mathrm{d}z\right) \; .
\end{aligned} \tag{45}$$

The general antiderivatives are

$$\begin{aligned}
\int x \cdot \exp\left[-ax^2\right] \mathrm{d}x &= -\frac{1}{2a} \cdot \exp\left[-ax^2\right] \\
\int \exp\left[-ax^2\right] \mathrm{d}x &= \frac{1}{2}\sqrt{\frac{\pi}{a}} \cdot \operatorname{erf}\left[\sqrt{a}x\right]
\end{aligned} \tag{46}$$

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

$$\begin{aligned}
\mathrm{E}(X) &= \frac{1}{\sqrt{2\pi}\sigma} \left(\left[-\sigma^2 \cdot \exp\left[-\frac{1}{2\sigma^2} \cdot z^2\right]\right]_{-\infty}^{+\infty} + \mu \left[\sqrt{\frac{\pi}{2}}\sigma \cdot \operatorname{erf}\left[\frac{1}{\sqrt{2}\sigma}z\right]\right]_{-\infty}^{+\infty}\right) \\
&= \frac{1}{\sqrt{2\pi}\sigma} \left(\left[\lim_{z\to\infty}\left(-\sigma^2 \cdot \exp\left[-\frac{1}{2\sigma^2} \cdot z^2\right]\right) - \lim_{z\to-\infty}\left(-\sigma^2 \cdot \exp\left[-\frac{1}{2\sigma^2} \cdot z^2\right]\right)\right]\right. \\
&\qquad \left. + \mu \left[\lim_{z\to\infty}\left(\sqrt{\frac{\pi}{2}}\sigma \cdot \operatorname{erf}\left[\frac{1}{\sqrt{2}\sigma}z\right]\right) - \lim_{z\to-\infty}\left(\sqrt{\frac{\pi}{2}}\sigma \cdot \operatorname{erf}\left[\frac{1}{\sqrt{2}\sigma}z\right]\right)\right]\right) \\
&= \frac{1}{\sqrt{2\pi}\sigma} \left([0-0] + \mu \left[\sqrt{\frac{\pi}{2}}\sigma - \left(-\sqrt{\frac{\pi}{2}}\sigma\right)\right]\right) \\
&= \frac{1}{\sqrt{2\pi}\sigma} \cdot \mu \cdot 2\sqrt{\frac{\pi}{2}}\sigma \\
&= \mu \; .
\end{aligned} \tag{47}$$

Sources:

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

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

3.2.4 Median

Proof: Median of the normal distribution

Index: The Book of Statistical Proofs ▷ Probability Distributions ▷ Univariate continuous distributions ▷ Normal distribution ▷ Median

Theorem: Let X be a random variable following a normal distribution:

$$X \sim \mathcal{N}(\mu, \sigma^2) \; . \tag{48}$$

Then, the median of X is

$$\mathrm{median}(X) = \mu \; . \tag{49}$$

Proof: The median is the value at which the cumulative distribution function is 1/2:

$$F_X(\mathrm{median}(X)) = \frac{1}{2} \; . \tag{50}$$

The cumulative distribution function (CDF) of the normal distribution is

$$F_X(x) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)\right] \tag{51}$$

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

$$x = \sqrt{2}\sigma \cdot \mathrm{erf}^{-1}(2p-1) + \mu \tag{52}$$

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

$$\mathrm{median}(X) = \sqrt{2}\sigma \cdot \mathrm{erf}^{-1}(0) + \mu = \mu \; . \tag{53}$$

Sources:
- own work

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

3.2.5 Mode

Proof: Mode of the normal distribution

Index: The Book of Statistical Proofs ▷ Probability Distributions ▷ Univariate continuous distributions ▷ Normal distribution ▷ Mode

Theorem: Let X be a random variable following a normal distribution:

$$X \sim \mathcal{N}(\mu, \sigma^2) \; . \tag{54}$$

Then, the mode of X is

$$\mathrm{mode}(X) = \mu \; . \tag{55}$$

Proof: The mode is the value which maximizes the probability density function:

$$\mathrm{mode}(X) = \operatorname*{arg\,max}_x f_X(x) \; . \tag{56}$$

The probability density function of the normal distribution is:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right] \; . \tag{57}$$

The first two deriatives of this function are:

$$f'_X(x) = \frac{\mathrm{d}f_X(x)}{\mathrm{d}x} = \frac{1}{\sqrt{2\pi}\sigma^3} \cdot (-x+\mu) \cdot \exp\left[-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right] \tag{58}$$

$$f''_X(x) = \frac{\mathrm{d}^2 f_X(x)}{\mathrm{d}x^2} = -\frac{1}{\sqrt{2\pi}\sigma^3} \cdot \exp\left[-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right] + \frac{1}{\sqrt{2\pi}\sigma^5} \cdot (-x+\mu)^2 \cdot \exp\left[-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right] \; . \tag{59}$$

We now calculate the root of the first derivative (58):

$$\begin{aligned} f'_X(x) = 0 &= \frac{1}{\sqrt{2\pi}\sigma^3} \cdot (-x+\mu) \cdot \exp\left[-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right] \\ 0 &= -x + \mu \\ x &= \mu \; . \end{aligned} \tag{60}$$

By plugging this value into the second deriative (59),

$$\begin{aligned} f''_X(\mu) &= -\frac{1}{\sqrt{2\pi}\sigma^3} \cdot \exp(0) + \frac{1}{\sqrt{2\pi}\sigma^5} \cdot (0)^2 \cdot \exp(0) \\ &= -\frac{1}{\sqrt{2\pi}\sigma^3} < 0 \; , \end{aligned} \tag{61}$$

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

$$\mathrm{mode}(X) = \mu \; . \tag{62}$$

Sources:

- own work

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

3.2.6 Variance

Proof: Variance of the normal distribution

Index: The Book of Statistical Proofs ▷ Probability Distributions ▷ Univariate continuous distributions ▷ Normal distribution ▷ Variance

Theorem: Let X be a random variable following a normal distribution:

$$X \sim \mathcal{N}(\mu, \sigma^2) \; . \tag{63}$$

Then, the variance of X is

$$\mathrm{Var}(X) = \sigma^2 \; . \tag{64}$$

Proof: The variance is the probability-weighted average of the squared deviation from the mean:

$$\mathrm{Var}(X) = \int_{\mathbb{R}} (x - \mathrm{E}(X))^2 \cdot f_{\mathrm{X}}(x) \, \mathrm{d}x \; . \tag{65}$$

With the expeted value and probability density function of the normal distribution, this reads:

$$\begin{split} \mathrm{Var}(X) &= \int_{-\infty}^{+\infty} (x-\mu)^2 \cdot \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right] \mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} (x-\mu)^2 \cdot \exp\left[-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right] \mathrm{d}x \; . \end{split} \tag{66}$$

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

$$\begin{split} \mathrm{Var}(X) &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty-\mu}^{+\infty-\mu} z^2 \cdot \exp\left[-\frac{1}{2} \left(\frac{z}{\sigma} \right)^2 \right] \mathrm{d}(z+\mu) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} z^2 \cdot \exp\left[-\frac{1}{2} \left(\frac{z}{\sigma} \right)^2 \right] \mathrm{d}z \; . \end{split} \tag{67}$$

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

$$\begin{split} \mathrm{Var}(X) &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} (\sqrt{2}\sigma x)^2 \cdot \exp\left[-\frac{1}{2} \left(\frac{\sqrt{2}\sigma x}{\sigma} \right)^2 \right] \mathrm{d}(\sqrt{2}\sigma x) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \cdot 2\sigma^2 \cdot \sqrt{2}\sigma \int_{-\infty}^{+\infty} x^2 \cdot \exp\left[-x^2 \right] \mathrm{d}x \\ &= \frac{2\sigma^2}{\sqrt{\pi}} \int_{-\infty}^{+\infty} x^2 \cdot e^{-x^2} \, \mathrm{d}x \; . \end{split} \tag{68}$$

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

$$\mathrm{Var}(X) = \frac{4\sigma^2}{\sqrt{\pi}} \int_0^\infty x^2 \cdot e^{-x^2} \, \mathrm{d}x \; . \tag{69}$$

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

$$\mathrm{Var}(X) = \frac{4\sigma^2}{\sqrt{\pi}} \int_0^\infty z \cdot e^{-z} \cdot \frac{1}{2} z^{-\frac{1}{2}} \, \mathrm{d}z = \frac{2\sigma^2}{\sqrt{\pi}} \int_0^\infty z^{\frac{3}{2}-1} \cdot e^{-z} \, \mathrm{d}z \tag{70}$$

and using the definition of the gamma function

$$\Gamma(x) = \int_0^\infty z^{x-1} \cdot e^{-z} \, \mathrm{d}z \; , \tag{71}$$

we can finally show that

$$\mathrm{Var}(X) = \frac{2\sigma^2}{\sqrt{\pi}} \cdot \Gamma\left(\frac{3}{2}\right) = \frac{2\sigma^2}{\sqrt{\pi}} \cdot \frac{\sqrt{\pi}}{2} = \sigma^2 \; . \tag{72}$$

Sources:

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

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

4 Multivariate continuous distributions

4.1 Multivariate normal distribution

4.1.1 Definition

Definition: Multivariate normal distribution

Index: The Book of Statistical Proofs ▷ Probability Distributions ▷ Multivariate continuous distributions ▷ Multivariate normal distribution ▷ Definition

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

$$X \sim \mathcal{N}(\mu, \Sigma) \; , \tag{73}$$

if and only if its probability density function is given by

$$\mathcal{N}(x; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \cdot \exp\left[-\frac{1}{2} (x-\mu)^\mathrm{T} \Sigma^{-1} (x-\mu) \right] \tag{74}$$

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

Proof: Probability density function of the multivariate normal distribution

Index: The Book of Statistical Proofs ▷ Probability Distributions ▷ Multivariate continuous distributions ▷ Multivariate normal distribution ▷ Probability density function

Theorem: Let X be a random vector following a multivariate normal distribution:

$$X \sim \mathcal{N}(\mu, \Sigma) \; . \tag{75}$$

Then, the probability density function of X is

$$f_X(x) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \cdot \exp\left[-\frac{1}{2} (x-\mu)^\mathrm{T} \Sigma^{-1} (x-\mu) \right] \; . \tag{76}$$

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

Sources:

- own work

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

4.1.3 Linear transformation theorem

Proof: Linear transformation theorem for the multivariate normal distribution

Index: The Book of Statistical Proofs ▷ Probability Distributions ▷ Multivariate continuous distributions ▷ Multivariate normal distribution ▷ Linear transformation theorem

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

$$x \sim \mathcal{N}(\mu, \Sigma) \; . \tag{77}$$

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

$$y = Ax + b \sim \mathcal{N}(A\mu + b, A \Sigma A^\mathrm{T}) \; . \tag{78}$$

Proof: The moment-generating function of a random vector (→ Definition I/1.1.1) x is

$$M_x(t) = \mathbb{E}\left(\exp\left[t^\mathrm{T} x \right] \right) \tag{79}$$

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

$$\begin{split} M_y(t) &= \mathbb{E}\left(\exp\left[t^\mathrm{T} (Ax+b) \right] \right) \\ &= \mathbb{E}\left(\exp\left[t^\mathrm{T} Ax \right] \cdot \exp\left[t^\mathrm{T} b \right] \right) \\ &= \exp\left[t^\mathrm{T} b \right] \cdot \mathbb{E}\left(\exp\left[t^\mathrm{T} Ax \right] \right) \\ &= \exp\left[t^\mathrm{T} b \right] \cdot M_x(At) \; . \end{split} \tag{80}$$

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

$$M_x(t) = \exp\left[t^\mathrm{T} \mu + \frac{1}{2} t^\mathrm{T} \Sigma t \right] \tag{81}$$

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

$$\begin{split} M_y(t) &= \exp\left[t^\mathrm{T} b \right] \cdot M_x(At) \\ &= \exp\left[t^\mathrm{T} b \right] \cdot \exp\left[t^\mathrm{T} A \mu + \frac{1}{2} t^\mathrm{T} A \Sigma A^\mathrm{T} t \right] \\ &= \exp\left[t^\mathrm{T} \left(A\mu + b \right) + \frac{1}{2} t^\mathrm{T} A \Sigma A^\mathrm{T} t \right] \; . \end{split} \tag{82}$$

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

Sources:

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

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

4.1.4 Marginal distributions

Proof: Marginal distributions of the multivariate normal distribution

Index: The Book of Statistical Proofs ▷ Probability Distributions ▷ Multivariate continuous distributions ▷ Multivariate normal distribution ▷ Marginal distributions

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

$$x \sim \mathcal{N}(\mu, \Sigma) \; . \tag{83}$$

Then, the marginal distribution of any subset vector x_s is also a multivariate normal distribution

$$x_s \sim \mathcal{N}(\mu_s, \Sigma_s) \tag{84}$$

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

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

$$x_s = Sx \tag{85}$$

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

$$x_s \sim \mathcal{N}(S\mu, S\Sigma S^{\mathrm{T}}) \; . \tag{86}$$

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

Sources:
- own work

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

4.2 Normal-gamma distribution

4.2.1 Definition

Definition: Normal-gamma distribution

Index: The Book of Statistical Proofs ▷ Probability Distributions ▷ Multivariate continuous distributions ▷ Normal-gamma distribution ▷ Definition

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

$$X, Y \sim \mathrm{NG}(\mu, \Lambda, a, b) \; , \tag{87}$$

if and only if their joint probability density function is given by

$$f_{X,Y}(x, y) = \mathcal{N}(x; \mu, (y\Lambda)^{-1}) \cdot \mathrm{Gam}(y; a, b) \tag{88}$$

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

Proof: Marginal distributions of the normal-gamma distribution

Index: The Book of Statistical Proofs ▷ Probability Distributions ▷ Multivariate continuous distributions ▷ Normal-gamma distribution ▷ Marginal distributions

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

$$x, y \sim \mathrm{NG}(\mu, \Lambda, a, b) \; . \tag{89}$$

Then, the marginal distribution of y is a gamma distribution

$$y \sim \mathrm{Gam}(a, b) \tag{90}$$

and the marginal distribution of x is a multivariate t-distribution

$$x \sim \mathrm{t}\left(\mu, \left(\frac{a}{b}\Lambda\right)^{-1}, 2a\right) \; . \tag{91}$$

Proof: The probability density function of the normal-gamma distribution (→ Proof "ng-pdf") is given by

$$\begin{split} p(x, y) &= p(x|y) \cdot p(y) \\ p(x|y) &= \mathcal{N}(x; \mu, (y\Lambda)^{-1}) \\ p(y) &= \mathrm{Gam}(y; a, b) \; . \end{split} \tag{92}$$

Using the law of marginal probability, the marginal distribution of y can be derived as

$$
\begin{aligned}
p(y) &= \int p(x,y) \, \mathrm{d}x \\
&= \int \mathcal{N}(x;\mu,(y\Lambda)^{-1}) \, \mathrm{Gam}(y;a,b) \, \mathrm{d}x \\
&= \mathrm{Gam}(y;a,b) \int \mathcal{N}(x;\mu,(y\Lambda)^{-1}) \, \mathrm{d}x \\
&= \mathrm{Gam}(y;a,b)
\end{aligned} \tag{93}
$$

which is the probability density function of the gamma distribution (→ Proof “ng-pdf”) with shape parameter a and rate parameter b.

Using the law of marginal probability, the marginal distribution of x can be derived as

$$
\begin{aligned}
p(x) &= \int p(x,y)\,\mathrm{d}y \\
&= \int \mathcal{N}(x;\mu,(y\Lambda)^{-1})\,\mathrm{Gam}(y;a,b)\,\mathrm{d}y \\
&= \int \sqrt{\frac{|y\Lambda|}{\sqrt{(2\pi)^n}}}\,\exp\left[-\frac{1}{2}(x-\mu)^\mathrm{T}(y\Lambda)(x-\mu)\right]\cdot\frac{b^a}{\Gamma(a)}\,y^{a-1}\exp[-by]\,\mathrm{d}y \\
&= \int \sqrt{\frac{y^n|\Lambda|}{\sqrt{(2\pi)^n}}}\,\exp\left[-\frac{1}{2}(x-\mu)^\mathrm{T}(y\Lambda)(x-\mu)\right]\cdot\frac{b^a}{\Gamma(a)}\,y^{a-1}\exp[-by]\,\mathrm{d}y \\
&= \int \sqrt{\frac{|\Lambda|}{\sqrt{(2\pi)^n}}}\cdot\frac{b^a}{\Gamma(a)}\cdot y^{a+\frac{n}{2}-1}\cdot\exp\left[-\left(b+\frac{1}{2}(x-\mu)^\mathrm{T}\Lambda(x-\mu)\right)y\right]\mathrm{d}y \\
&= \int \sqrt{\frac{|\Lambda|}{\sqrt{(2\pi)^n}}}\cdot\frac{b^a}{\Gamma(a)}\cdot\frac{\Gamma\left(a+\frac{n}{2}\right)}{\left(b+\frac{1}{2}(x-\mu)^\mathrm{T}\Lambda(x-\mu)\right)^{a+\frac{n}{2}}}\cdot\mathrm{Gam}\left(y;a+\frac{n}{2},b+\frac{1}{2}(x-\mu)^\mathrm{T}\Lambda(x-\mu)\right)\mathrm{d} \\
&= \sqrt{\frac{|\Lambda|}{\sqrt{(2\pi)^n}}}\cdot\frac{b^a}{\Gamma(a)}\cdot\frac{\Gamma\left(a+\frac{n}{2}\right)}{\left(b+\frac{1}{2}(x-\mu)^\mathrm{T}\Lambda(x-\mu)\right)^{a+\frac{n}{2}}}\int\mathrm{Gam}\left(y;a+\frac{n}{2},b+\frac{1}{2}(x-\mu)^\mathrm{T}\Lambda(x-\mu)\right)\mathrm{d}y \\
&= \sqrt{\frac{|\Lambda|}{\sqrt{(2\pi)^n}}}\cdot\frac{b^a}{\Gamma(a)}\cdot\frac{\Gamma\left(a+\frac{n}{2}\right)}{\left(b+\frac{1}{2}(x-\mu)^\mathrm{T}\Lambda(x-\mu)\right)^{a+\frac{n}{2}}} \\
&= \frac{\sqrt{|\Lambda|}}{(2\pi)^{\frac{n}{2}}}\cdot\frac{\Gamma\left(\frac{2a+n}{2}\right)}{\Gamma\left(\frac{2a}{2}\right)}\cdot b^a\cdot\left(b+\frac{1}{2}(x-\mu)^\mathrm{T}\Lambda(x-\mu)\right)^{-\left(a+\frac{n}{2}\right)} \\
&= \frac{\sqrt{|\Lambda|}}{\pi^{\frac{n}{2}}}\cdot\frac{\Gamma\left(\frac{2a+n}{2}\right)}{\Gamma\left(\frac{2a}{2}\right)}\cdot\left(\frac{1}{b}\right)^{-a}\cdot\left(b+\frac{1}{2}(x-\mu)^\mathrm{T}\Lambda(x-\mu)\right)^{-a}\cdot 2^{-\frac{n}{2}}\cdot\left(b+\frac{1}{2}(x-\mu)^\mathrm{T}\Lambda(x-\mu)\right)^{-\frac{n}{2}} \\
&= \frac{\sqrt{|\Lambda|}}{\pi^{\frac{n}{2}}}\cdot\frac{\Gamma\left(\frac{2a+n}{2}\right)}{\Gamma\left(\frac{2a}{2}\right)}\cdot\left(1+\frac{1}{2b}(x-\mu)^\mathrm{T}\Lambda(x-\mu)\right)^{-a}\cdot\left(2b+(x-\mu)^\mathrm{T}\Lambda(x-\mu)\right)^{-\frac{n}{2}} \\
&= \frac{\sqrt{|\Lambda|}}{\pi^{\frac{n}{2}}}\cdot\frac{\Gamma\left(\frac{2a+n}{2}\right)}{\Gamma\left(\frac{2a}{2}\right)}\cdot\left(\frac{1}{2a}\right)^{-a}\cdot\left(2a+(x-\mu)^\mathrm{T}\left(\frac{a}{b}\Lambda\right)(x-\mu)\right)^{-a}\cdot\left(\frac{b}{a}\right)^{-\frac{n}{2}}\cdot\Big(2a+(x-\mu)^\mathrm{T}\left(\frac{a}{b}\Lambda\right)\big(\\
&= \frac{\sqrt{\left(\frac{a}{b}\right)^n|\Lambda|}}{(2a)^{-a}\,\pi^{\frac{n}{2}}}\cdot\frac{\Gamma\left(\frac{2a+n}{2}\right)}{\Gamma\left(\frac{2a}{2}\right)}\cdot\left(2a+(x-\mu)^\mathrm{T}\left(\frac{a}{b}\Lambda\right)(x-\mu)\right)^{-a}\cdot\left(2a+(x-\mu)^\mathrm{T}\left(\frac{a}{b}\Lambda\right)(x-\mu)\right)^{-\frac{n}{2}} \\
&= \frac{\sqrt{\left(\frac{a}{b}\right)^n|\Lambda|}}{(2a)^{-a}\,\pi^{\frac{n}{2}}}\cdot\frac{\Gamma\left(\frac{2a+n}{2}\right)}{\Gamma\left(\frac{2a}{2}\right)}\cdot(2a)^{-a}\cdot\left(1+\frac{1}{2a}(x-\mu)^\mathrm{T}\left(\frac{a}{b}\Lambda\right)(x-\mu)\right)^{-a}\cdot(2a)^{-\frac{n}{2}}\cdot\Bigg(1+\frac{1}{2a}(x-\mu)^\mathrm{T}\Big(\\
&= \frac{\sqrt{\left(\frac{a}{b}\right)^n|\Lambda|}}{(2a)^{\frac{n}{2}}\,\pi^{\frac{n}{2}}}\cdot\frac{\Gamma\left(\frac{2a+n}{2}\right)}{\Gamma\left(\frac{2a}{2}\right)}\cdot\left(1+\frac{1}{2a}(x-\mu)^\mathrm{T}\left(\frac{a}{b}\Lambda\right)(x-\mu)\right)^{-\frac{2a+n}{2}} \\
&= \sqrt{\frac{\left|\frac{a}{b}\,\Lambda\right|}{(2a\,\pi)^n}}\cdot\frac{\Gamma\left(\frac{2a+n}{2}\right)}{\Gamma\left(\frac{2a}{2}\right)}\cdot\left(1+\frac{1}{2a}(x-\mu)^\mathrm{T}\left(\frac{a}{b}\Lambda\right)(x-\mu)\right)^{-\frac{2a+n}{2}}
\end{aligned}
\tag{94}
$$

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

Sources:

- own work

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

4.2.3 Kullback-Leibler divergence

Proof: Kullback-Leibler divergence for the normal-gamma distribution

Index: The Book of Statistical Proofs ▷ Probability Distributions ▷ Multivariate continuous distributions ▷ Normal-gamma distribution ▷ Kullback-Leibler divergence

Theorem: Let $x \in \mathbb{R}^k$ be a random vector and $y > 0$ be a random variable. Assume two normal-gamma distributions P and Q specifying the joint distribution of x and y as

$$\begin{aligned} P &: \; (x, y) \sim \mathrm{NG}(\mu_1, \Lambda_1^{-1}, a_1, b_1) \\ Q &: \; (x, y) \sim \mathrm{NG}(\mu_2, \Lambda_2^{-1}, a_2, b_2) \; . \end{aligned} \tag{95}$$

Then, the Kullback-Leibler divergence of P from Q is given by

$$\begin{aligned} \mathrm{KL}[P \,||\, Q] = & \frac{1}{2} \frac{a_1}{b_1} \left[(\mu_2 - \mu_1)^T \Lambda_2 (\mu_2 - \mu_1) \right] + \frac{1}{2} \mathrm{tr}(\Lambda_2 \Lambda_1^{-1}) - \frac{1}{2} \ln \frac{|\Lambda_2|}{|\Lambda_1|} - \frac{k}{2} \\ & + a_2 \, \ln \frac{b_1}{b_2} - \ln \frac{\Gamma(a_1)}{\Gamma(a_2)} + (a_1 - a_2) \, \psi(a_1) - (b_1 - b_2) \, \frac{a_1}{b_1} \; . \end{aligned} \tag{96}$$

Proof: The probability density function of the normal-gamma (NG) distribution is

$$p(x, y) = p(x|y) \cdot p(y) = \mathcal{N}(x; \mu, (y\Lambda)^{-1}) \cdot \mathrm{Gam}(y; a, b) \tag{97}$$

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

$$\mathrm{KL}[P \,||\, Q] = \frac{1}{2} \left[(\mu_2 - \mu_1)^T \Sigma_2^{-1} (\mu_2 - \mu_1) + \mathrm{tr}(\Sigma_2^{-1} \Sigma_1) - \ln \frac{|\Sigma_1|}{|\Sigma_2|} - k \right] \tag{98}$$

and the Kullback-Leibler divergence of the univariate gamma distribution is

$$\mathrm{KL}[P \,||\, Q] = a_2 \, \ln \frac{b_1}{b_2} - \ln \frac{\Gamma(a_1)}{\Gamma(a_2)} + (a_1 - a_2) \, \psi(a_1) - (b_1 - b_2) \, \frac{a_1}{b_1} \tag{99}$$

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

The KL divergence for a continuous random variable is given by

$$\mathrm{KL}[P\,||\,Q] = \int_{\mathcal{Z}} p(z)\,\ln\frac{p(z)}{q(z)}\,\mathrm{d}z \tag{100}$$

which, applied to the normal-gamma distribution over x and y, yields

$$\mathrm{KL}[P\,||\,Q] = \int_0^\infty \int_{\mathbb{R}^k} p(x,y)\,\ln\frac{p(x,y)}{q(x,y)}\,\mathrm{d}x\,\mathrm{d}y\;. \tag{101}$$

Using the law of conditional probability, this can be evaluated as follows:

$$\begin{aligned}
\mathrm{KL}[P\,||\,Q] &= \int_0^\infty \int_{\mathbb{R}^k} p(x|y)\,p(y)\,\ln\frac{p(x|y)\,p(y)}{q(x|y)\,q(y)}\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_0^\infty \int_{\mathbb{R}^k} p(x|y)\,p(y)\,\ln\frac{p(x|y)}{q(x|y)}\,\mathrm{d}x\,\mathrm{d}y + \int_0^\infty \int_{\mathbb{R}^k} p(x|y)\,p(y)\,\ln\frac{p(y)}{q(y)}\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_0^\infty p(y) \int_{\mathbb{R}^k} p(x|y)\,\ln\frac{p(x|y)}{q(x|y)}\,\mathrm{d}x\,\mathrm{d}y + \int_0^\infty p(y)\,\ln\frac{p(y)}{q(y)} \int_{\mathbb{R}^k} p(x|y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \left\langle \mathrm{KL}[p(x|y)\,||\,q(x|y)] \right\rangle_{p(y)} + \mathrm{KL}[p(y)\,||\,q(y)]\;.
\end{aligned} \tag{102}$$

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 (97) and (98), the first term becomes

$$\begin{aligned}
&\left\langle \mathrm{KL}[p(x|y)\,||\,q(x|y)] \right\rangle_{p(y)} \\
&= \left\langle \frac{1}{2}\left[(\mu_2-\mu_1)^T (y\Lambda_2)(\mu_2-\mu_1) + \mathrm{tr}\left((y\Lambda_2)(y\Lambda_1)^{-1} \right) - \ln\frac{|(y\Lambda_1)^{-1}|}{|(y\Lambda_2)^{-1}|} - k \right] \right\rangle_{p(y)} \\
&= \left\langle \frac{y}{2}(\mu_2-\mu_1)^T \Lambda_2 (\mu_2-\mu_1) + \frac{1}{2}\,\mathrm{tr}(\Lambda_2\Lambda_1^{-1}) - \frac{1}{2}\ln\frac{|\Lambda_2|}{|\Lambda_1|} - \frac{k}{2} \right\rangle_{p(y)}
\end{aligned} \tag{103}$$

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

$$\left\langle \mathrm{KL}[p(x|y)\,||\,q(x|y)] \right\rangle_{p(y)} = \frac{1}{2}\frac{a_1}{b_1}(\mu_2-\mu_1)^T \Lambda_2 (\mu_2-\mu_1) + \frac{1}{2}\,\mathrm{tr}(\Lambda_2\Lambda_1^{-1}) - \frac{1}{2}\ln\frac{|\Lambda_2|}{|\Lambda_1|} - \frac{k}{2}\;. \tag{104}$$

By plugging (104) and (99) into (102), one arrives at the KL divergence given by (96).

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.

5 Matrix-variate continuous distributions

5.1 Matrix-normal distribution

5.1.1 Definition

Definition: Matrix-normal distribution

Index: The Book of Statistical Proofs ▷ Probability Distributions ▷ Matrix-variate continuous distributions ▷ Matrix-normal distribution ▷ Definition
Definition: Let X be an $n \times p$ random matrix. Then, X is said to be matrix-normally distributed with mean M, covariance across rows U and covariance across columns V

$$X \sim \mathcal{MN}(M, U, V) \; , \tag{105}$$

if and only if its probability density function is given by

$$f_X(X) = \mathcal{MN}(X; M, U, V) = \frac{1}{\sqrt{(2\pi)^{np} |V|^n |U|^p}} \cdot \exp\left[-\frac{1}{2} \mathrm{tr}\left(V^{-1} (X-M)^\mathrm{T} \, U^{-1} (X-M) \right) \right] \tag{106}$$

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

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

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

5.1.2 Equivalence to multivariate normal distribution

Proof: Equivalence of matrix-normal distribution and multivariate normal distribution

Index: The Book of Statistical Proofs ▷ Probability Distributions ▷ Matrix-variate continuous distributions ▷ Matrix-normal distribution ▷ Equivalence to multivariate normal distribution

Theorem: The matrix X is matrix-normally distributed

$$X \sim \mathcal{MN}(M, U, V) \; , \tag{107}$$

if and only if $\mathrm{vec}(X)$ is multivariate normally distributed

$$\mathrm{vec}(X) \sim \mathcal{MN}(\mathrm{vec}(M), V \otimes U) \tag{108}$$

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

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

$$\mathcal{MN}(X; M, U, V) = \frac{1}{\sqrt{(2\pi)^{np} |V|^n |U|^p}} \cdot \exp\left[-\frac{1}{2} \mathrm{tr}\left(V^{-1} (X-M)^\mathrm{T} \, U^{-1} (X-M) \right) \right] \; . \tag{109}$$

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

$$\mathcal{MN}(X; M, U, V) = \frac{1}{\sqrt{(2\pi)^{np} |V|^n |U|^p}} \cdot \exp\left[-\frac{1}{2} \mathrm{tr}\left((X-M)^\mathrm{T} \, U^{-1} (X-M) \, V^{-1} \right) \right] \; . \tag{110}$$

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

$$\mathcal{MN}(X; M, U, V) = \frac{1}{\sqrt{(2\pi)^{np} |V|^n |U|^p}} \cdot \exp\left[-\frac{1}{2} \mathrm{vec}(X-M)^\mathrm{T} \, \mathrm{vec}\left(U^{-1} (X-M) \, V^{-1} \right) \right] \; . \tag{111}$$

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

$$\mathcal{MN}(X; M, U, V) = \frac{1}{\sqrt{(2\pi)^{np} |V|^n |U|^p}} \cdot \exp\left[-\frac{1}{2} \mathrm{vec}(X-M)^\mathrm{T} \, \left(V^{-1} \otimes U^{-1} \right) \mathrm{vec}(X-M) \right] \; . \tag{112}$$

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

$$\mathcal{MN}(X; M, U, V) = \frac{1}{\sqrt{(2\pi)^{np} |V|^n |U|^p}} \cdot \exp\left[-\frac{1}{2} \mathrm{vec}(X-M)^\mathrm{T} \, (V \otimes U)^{-1} \, \mathrm{vec}(X-M) \right] \; . \tag{113}$$

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

$$\mathcal{MN}(X; M, U, V) = \frac{1}{\sqrt{(2\pi)^{np} |V|^n |U|^p}} \cdot \exp\left[-\frac{1}{2} \left[\mathrm{vec}(X) - \mathrm{vec}(M) \right]^\mathrm{T} \, (V \otimes U)^{-1} \left[\mathrm{vec}(X) - \mathrm{vec}(M) \right] \right] \; . \tag{114}$$

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

$$\mathcal{MN}(X; M, U, V) = \frac{1}{\sqrt{(2\pi)^{np} |V \otimes U|}} \cdot \exp\left[-\frac{1}{2} \left[\mathrm{vec}(X) - \mathrm{vec}(M) \right]^\mathrm{T} \, (V \otimes U)^{-1} \left[\mathrm{vec}(X) - \mathrm{vec}(M) \right] \right] \; . \tag{115}$$

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

$$\mathcal{MN}(X; M, U, V) = \mathcal{N}(\mathrm{vec}(X); \mathrm{vec}(M), V \otimes U) \; . \tag{116}$$

By showing that the probability density functions are identical, it is proven that the associated probability distributions are equivalent.

Sources:

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

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

Chapter III

Statistical Models

1 Normal data

1.1 Multiple linear regression

1.1.1 Ordinary least squares (1)

Proof: Ordinary least squares for multiple linear regression

Index: The Book of Statistical Proofs ▷ Statistical Models ▷ Normal data ▷ Multiple linear regression ▷ Ordinary least squares

Theorem: Given a linear regression model with independent observations

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

the parameters minimizing the residual sum of squares are given by

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

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

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

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

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

Sources:

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

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

1.1.2 Ordinary least squares (2)

Proof: Ordinary least squares for multiple linear regression

Index: The Book of Statistical Proofs ▷ Statistical Models ▷ Normal data ▷ Multiple linear regression ▷ Ordinary least squares

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

$$y = X\beta + \varepsilon, \; \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \; , \tag{5}$$

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

$$\hat{\beta} = (X^\mathrm{T} X)^{-1} X^\mathrm{T} y \; . \tag{6}$$

Proof: The residual sum of squares is defined as

$$\mathrm{RSS}(\beta) = \sum_{i=1}^{n} \varepsilon_i = \varepsilon^\mathrm{T} \varepsilon = (y - X\beta)^\mathrm{T} (y - X\beta) \tag{7}$$

which can be developed into

$$\begin{split} \mathrm{RSS}(\beta) &= y^\mathrm{T} y - y^\mathrm{T} X\beta - \beta^\mathrm{T} X^\mathrm{T} y + \beta^\mathrm{T} X^\mathrm{T} X\beta \\ &= y^\mathrm{T} y - 2\beta^\mathrm{T} X^\mathrm{T} y + \beta^\mathrm{T} X^\mathrm{T} X\beta \; . \end{split} \tag{8}$$

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

$$\frac{\mathrm{dRSS}(\beta)}{\mathrm{d}\beta} = -2X^\mathrm{T} y + 2X^\mathrm{T} X\beta \tag{9}$$

and setting this deriative to zero, we obtain:

$$\begin{split} \frac{\mathrm{dRSS}(\hat{\beta})}{\mathrm{d}\beta} &= 0 \\ 0 &= -2X^\mathrm{T} y + 2X^\mathrm{T} X\hat{\beta} \\ X^\mathrm{T} X\hat{\beta} &= X^\mathrm{T} y \\ \hat{\beta} &= (X^\mathrm{T} X)^{-1} X^\mathrm{T} y \; . \end{split} \tag{10}$$

Since the quadratic form $y^\mathrm{T} y$ in (8) is positive, $\hat{\beta}$ minimizes $\mathrm{RSS}(\beta)$.

Sources:

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

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

1.2 Bayesian linear regression

1.2.1 Conjugate prior distribution

Proof: Conjugate prior distribution for Bayesian linear regression

Index: The Book of Statistical Proofs ▷ Statistical Models ▷ Normal data ▷ Bayesian linear regression ▷ Conjugate prior distribution

Theorem: Let

$$y = X\beta + \varepsilon, \; \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \tag{11}$$

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

$$p(\beta, \tau) = \mathcal{N}(\beta; \mu_0, (\tau \Lambda_0)^{-1}) \cdot \mathrm{Gam}(\tau; a_0, b_0) \tag{12}$$

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

Proof: By definition, a conjugate prior is a prior distribution that, when combined with the likelihood function, leads to a posterior distribution that belongs to the same family of probability distributions. 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 (11) implies the following likelihood function

$$p(y|\beta, \sigma^2) = \mathcal{N}(y; X\beta, \sigma^2 V) = \sqrt{\frac{1}{(2\pi)^n |\sigma^2 V|}} \, \exp\left[-\frac{1}{2\sigma^2} (y - X\beta)^\mathrm{T} V^{-1} (y - X\beta) \right] \tag{13}$$

which, for mathematical convenience, can also be parametrized as

$$p(y|\beta, \tau) = \mathcal{N}(y; X\beta, (\tau P)^{-1}) = \sqrt{\frac{|\tau P|}{(2\pi)^n}} \, \exp\left[-\frac{\tau}{2} (y - X\beta)^\mathrm{T} P (y - X\beta) \right] \tag{14}$$

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

Seperating constant and variable terms, we have:

$$p(y|\beta, \tau) = \sqrt{\frac{|P|}{(2\pi)^n}} \cdot \tau^{n/2} \cdot \exp\left[-\frac{\tau}{2} (y - X\beta)^\mathrm{T} P (y - X\beta) \right] \; . \tag{15}$$

Expanding the product in the exponent, we have:

$$p(y|\beta, \tau) = \sqrt{\frac{|P|}{(2\pi)^n}} \cdot \tau^{n/2} \cdot \exp\left[-\frac{\tau}{2} \left(y^\mathrm{T} P y - y^\mathrm{T} P X \beta - \beta^\mathrm{T} X^\mathrm{T} P y + \beta^\mathrm{T} X^\mathrm{T} P X \beta \right) \right] \; . \tag{16}$$

Completing the square over β, finally gives

$$p(y|\beta, \tau) = \sqrt{\frac{|P|}{(2\pi)^n}} \cdot \tau^{n/2} \cdot \exp\left[-\frac{\tau}{2} \left((\beta - \tilde{X} y)^\mathrm{T} X^\mathrm{T} P X (\beta - \tilde{X} y) - y^\mathrm{T} Q y + y^\mathrm{T} P y \right) \right] \tag{17}$$

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

In other words, the likelihood function is proportional to a power of τ times an exponential of τ and an exponential of a squared form of β, weighted by τ:

$$p(y|\beta,\tau) \propto \tau^{n/2} \cdot \exp\left[-\frac{\tau}{2}\left(y^\mathrm{T} P y - y^\mathrm{T} Q y\right)\right] \cdot \exp\left[-\frac{\tau}{2}(\beta - \tilde{X}y)^\mathrm{T} X^\mathrm{T} P X (\beta - \tilde{X}y)\right] \; . \tag{18}$$

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

$$p(\beta,\tau) = \mathcal{N}(\beta; \mu_0, (\tau \Lambda_0)^{-1}) \cdot \mathrm{Gam}(\tau; a_0, b_0) \tag{19}$$

whose probability density function

$$p(\beta,\tau) = \sqrt{\frac{|\tau \Lambda_0|}{(2\pi)^p}} \exp\left[-\frac{\tau}{2}(\beta - \mu_0)^\mathrm{T} \Lambda_0 (\beta - \mu_0)\right] \cdot \frac{{b_0}^{a_0}}{\Gamma(a_0)} \tau^{a_0-1} \exp[-b_0 \tau] \tag{20}$$

exhibits the same proportionality

$$p(\beta,\tau) \propto \tau^{a_0+p/2-1} \cdot \exp[-\tau b_0] \cdot \exp\left[-\frac{\tau}{2}(\beta - \mu_0)^\mathrm{T} \Lambda_0 (\beta - \mu_0)\right] \tag{21}$$

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

Proof: Posterior distribution for Bayesian linear regression

Index: The Book of Statistical Proofs ▷ Statistical Models ▷ Normal data ▷ Bayesian linear regression ▷ Posterior distribution

Theorem: Let

$$y = X\beta + \varepsilon, \; \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \tag{22}$$

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

$$p(\beta,\tau) = \mathcal{N}(\beta; \mu_0, (\tau \Lambda_0)^{-1}) \cdot \mathrm{Gam}(\tau; a_0, b_0) \; . \tag{23}$$

Then, the posterior distribution is also a normal-gamma distribution

$$p(\beta, \tau|y) = \mathcal{N}(\beta; \mu_n, (\tau \Lambda_n)^{-1}) \cdot \mathrm{Gam}(\tau; a_n, b_n) \tag{24}$$

and the posterior hyperparameters are given by

$$\begin{aligned}
\mu_n &= \Lambda_n^{-1}(X^\mathrm{T} P y + \Lambda_0 \mu_0) \\
\Lambda_n &= X^\mathrm{T} P X + \Lambda_0 \\
a_n &= a_0 + \frac{n}{2} \\
b_n &= b_0 + \frac{1}{2}(y^\mathrm{T} P y + \mu_0^\mathrm{T} \Lambda_0 \mu_0 - \mu_n^\mathrm{T} \Lambda_n \mu_n) \; .
\end{aligned} \tag{25}$$

Proof: According to Bayes' theorem, the posterior distribution is given by

$$p(\beta, \tau|y) = \frac{p(y|\beta, \tau) \, p(\beta, \tau)}{p(y)} \; . \tag{26}$$

Since $p(y)$ is just a normalization factor, the posterior is proportional to the numerator:

$$p(\beta, \tau|y) \propto p(y|\beta, \tau) \, p(\beta, \tau) = p(y, \beta, \tau) \; . \tag{27}$$

Equation (22) implies the following likelihood function

$$p(y|\beta, \sigma^2) = \mathcal{N}(y; X\beta, \sigma^2 V) = \sqrt{\frac{1}{(2\pi)^n |\sigma^2 V|}} \, \exp\left[-\frac{1}{2\sigma^2} (y - X\beta)^\mathrm{T} V^{-1} (y - X\beta) \right] \tag{28}$$

which, for mathematical convenience, can also be parametrized as

$$p(y|\beta, \tau) = \mathcal{N}(y; X\beta, (\tau P)^{-1}) = \sqrt{\frac{|\tau P|}{(2\pi)^n}} \, \exp\left[-\frac{\tau}{2} (y - X\beta)^\mathrm{T} P (y - X\beta) \right] \tag{29}$$

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

Combining the likelihood function (29) with the prior distribution (23), the joint likelihood of the model is given by

$$\begin{aligned}
p(y, \beta, \tau) &= p(y|\beta, \tau) \, p(\beta, \tau) \\
&= \sqrt{\frac{|\tau P|}{(2\pi)^n}} \, \exp\left[-\frac{\tau}{2} (y - X\beta)^\mathrm{T} P (y - X\beta) \right] \cdot \\
&\quad \sqrt{\frac{|\tau \Lambda_0|}{(2\pi)^p}} \, \exp\left[-\frac{\tau}{2} (\beta - \mu_0)^\mathrm{T} \Lambda_0 (\beta - \mu_0) \right] \cdot \\
&\quad \frac{{b_0}^{a_0}}{\Gamma(a_0)} \, \tau^{a_0 - 1} \exp[-b_0 \tau] \; .
\end{aligned} \tag{30}$$

Collecting identical variables gives:

$$
\begin{aligned}
p(y, \beta, \tau) = \sqrt{\frac{\tau^{n+p}}{(2\pi)^{n+p}} |P||\Lambda_0|} \, \frac{{b_0}^{a_0}}{\Gamma(a_0)} \, \tau^{a_0-1} \exp[-b_0 \tau] \cdot \\
\exp\left[-\frac{\tau}{2} \left((y - X\beta)^\mathrm{T} P (y - X\beta) + (\beta - \mu_0)^\mathrm{T} \Lambda_0 (\beta - \mu_0) \right) \right] \; .
\end{aligned} \tag{31}
$$

Expanding the products in the exponent gives:

$$
\begin{aligned}
p(y, \beta, \tau) = \sqrt{\frac{\tau^{n+p}}{(2\pi)^{n+p}} |P||\Lambda_0|} \, \frac{{b_0}^{a_0}}{\Gamma(a_0)} \, \tau^{a_0-1} \exp[-b_0 \tau] \cdot \\
\exp\left[-\frac{\tau}{2} \left(y^\mathrm{T} P y - y^\mathrm{T} P X \beta - \beta^\mathrm{T} X^\mathrm{T} P y + \beta^\mathrm{T} X^\mathrm{T} P X \beta + \right. \right. \\
\left. \left. \beta^\mathrm{T} \Lambda_0 \beta - \beta^\mathrm{T} \Lambda_0 \mu_0 - \mu_0^\mathrm{T} \Lambda_0 \beta + \mu_0^\mathrm{T} \Lambda_0 \mu_0 \right) \right] \; .
\end{aligned} \tag{32}
$$

Completing the square over β, we finally have

$$
\begin{aligned}
p(y, \beta, \tau) = \sqrt{\frac{\tau^{n+p}}{(2\pi)^{n+p}} |P||\Lambda_0|} \, \frac{{b_0}^{a_0}}{\Gamma(a_0)} \, \tau^{a_0-1} \exp[-b_0 \tau] \cdot \\
\exp\left[-\frac{\tau}{2} \left((\beta - \mu_n)^\mathrm{T} \Lambda_n (\beta - \mu_n) + (y^\mathrm{T} P y + \mu_0^\mathrm{T} \Lambda_0 \mu_0 - \mu_n^\mathrm{T} \Lambda_n \mu_n) \right) \right]
\end{aligned} \tag{33}
$$

with the posterior hyperparameters

$$
\begin{aligned}
\mu_n &= \Lambda_n^{-1} (X^\mathrm{T} P y + \Lambda_0 \mu_0) \\
\Lambda_n &= X^\mathrm{T} P X + \Lambda_0 \; .
\end{aligned} \tag{34}
$$

Ergo, the joint likelihood is proportional to

$$
p(y, \beta, \tau) \propto \tau^{p/2} \cdot \exp\left[-\frac{\tau}{2} (\beta - \mu_n)^\mathrm{T} \Lambda_n (\beta - \mu_n) \right] \cdot \tau^{a_n - 1} \cdot \exp\left[-b_n \tau \right] \tag{35}
$$

with the posterior hyperparameters

$$
\begin{aligned}
a_n &= a_0 + \frac{n}{2} \\
b_n &= b_0 + \frac{1}{2} (y^\mathrm{T} P y + \mu_0^\mathrm{T} \Lambda_0 \mu_0 - \mu_n^\mathrm{T} \Lambda_n \mu_n) \; .
\end{aligned} \tag{36}
$$

From the term in (35), we can isolate the posterior distribution over β given τ:

$$
p(\beta | \tau, y) = \mathcal{N}(\beta; \mu_n, (\tau \Lambda_n)^{-1}) \; . \tag{37}
$$

From the remaining term, we can isolate the posterior distribution over τ:

$$
p(\tau | y) = \mathrm{Gam}(\tau; a_n, b_n) \; . \tag{38}
$$

Together, (37) and (38) constitute the joint posterior distribution 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

Proof: Log model evidence for Bayesian linear regression

Index: The Book of Statistical Proofs ▷ Statistical Models ▷ Normal data ▷ Bayesian linear regression ▷ Log model evidence

Theorem: Let

$$m:\; y = X\beta + \varepsilon,\; \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \tag{39}$$

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

$$p(\beta, \tau) = \mathcal{N}(\beta; \mu_0, (\tau\Lambda_0)^{-1}) \cdot \mathrm{Gam}(\tau; a_0, b_0) \; . \tag{40}$$

Then, the log model evidence for this model is

$$\begin{split} \log p(y|m) = \frac{1}{2}\log|P| - \frac{n}{2}\log(2\pi) + \frac{1}{2}\log|\Lambda_0| - \frac{1}{2}\log|\Lambda_n| + \\ \log\Gamma(a_n) - \log\Gamma(a_0) + a_0\log b_0 - a_n\log b_n \end{split} \tag{41}$$

where the posterior hyperparameters are given by

$$\begin{split} \mu_n &= \Lambda_n^{-1}(X^\mathrm{T} P y + \Lambda_0 \mu_0) \\ \Lambda_n &= X^\mathrm{T} P X + \Lambda_0 \\ a_n &= a_0 + \frac{n}{2} \\ b_n &= b_0 + \frac{1}{2}(y^\mathrm{T} P y + \mu_0^\mathrm{T} \Lambda_0 \mu_0 - \mu_n^\mathrm{T} \Lambda_n \mu_n) \; . \end{split} \tag{42}$$

Proof: According to the law of marginal probability, the model evidence for this model is:

$$p(y|m) = \iint p(y|\beta,\tau)\, p(\beta,\tau)\, \mathrm{d}\beta\, \mathrm{d}\tau \; . \tag{43}$$

According to the law of conditional probability, the integrand is equivalent to the joint likelihood:

$$p(y|m) = \iint p(y,\beta,\tau) \, \mathrm{d}\beta \, \mathrm{d}\tau \; . \tag{44}$$

Equation (39) implies the following likelihood function

$$p(y|\beta,\sigma^2) = \mathcal{N}(y; X\beta, \sigma^2 V) = \sqrt{\frac{1}{(2\pi)^n |\sigma^2 V|}} \, \exp\left[-\frac{1}{2\sigma^2}(y - X\beta)^{\mathrm{T}} V^{-1} (y - X\beta)\right] \tag{45}$$

which, for mathematical convenience, can also be parametrized as

$$p(y|\beta,\tau) = \mathcal{N}(y; X\beta, (\tau P)^{-1}) = \sqrt{\frac{|\tau P|}{(2\pi)^n}} \, \exp\left[-\frac{\tau}{2}(y - X\beta)^{\mathrm{T}} P (y - X\beta)\right] \tag{46}$$

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

When deriving the posterior distribution $p(\beta,\tau|y)$, the joint likelihood $p(y,\beta,\tau)$ is obtained as

$$\begin{split} p(y,\beta,\tau) = \sqrt{\frac{\tau^n |P|}{(2\pi)^n}} \sqrt{\frac{\tau^p |\Lambda_0|}{(2\pi)^p}} \, \frac{{b_0}^{a_0}}{\Gamma(a_0)} \, \tau^{a_0 - 1} \exp[-b_0 \tau] \cdot \\ \exp\left[-\frac{\tau}{2}\left((\beta - \mu_n)^T \Lambda_n (\beta - \mu_n) + (y^T P y + \mu_0^T \Lambda_0 \mu_0 - \mu_n^T \Lambda_n \mu_n)\right)\right] \; . \end{split} \tag{47}$$

Using the probability density function of the multivariate normal distribution, we can rewrite this as

$$\begin{split} p(y,\beta,\tau) = \sqrt{\frac{\tau^n |P|}{(2\pi)^n}} \sqrt{\frac{\tau^p |\Lambda_0|}{(2\pi)^p}} \sqrt{\frac{(2\pi)^p}{\tau^p |\Lambda_n|}} \, \frac{{b_0}^{a_0}}{\Gamma(a_0)} \, \tau^{a_0 - 1} \exp[-b_0 \tau] \cdot \\ \mathcal{N}(\beta; \mu_n, (\tau \Lambda_n)^{-1}) \, \exp\left[-\frac{\tau}{2}(y^T P y + \mu_0^T \Lambda_0 \mu_0 - \mu_n^T \Lambda_n \mu_n)\right] \; . \end{split} \tag{48}$$

Now, β can be integrated out easily:

$$\begin{split} \int p(y,\beta,\tau) \, \mathrm{d}\beta = \sqrt{\frac{\tau^n |P|}{(2\pi)^n}} \sqrt{\frac{|\Lambda_0|}{|\Lambda_n|}} \, \frac{{b_0}^{a_0}}{\Gamma(a_0)} \, \tau^{a_0 - 1} \exp[-b_0 \tau] \cdot \\ \exp\left[-\frac{\tau}{2}(y^T P y + \mu_0^T \Lambda_0 \mu_0 - \mu_n^T \Lambda_n \mu_n)\right] \; . \end{split} \tag{49}$$

Using the probability density function of the gamma distribution, we can rewrite this as

$$\int p(y,\beta,\tau) \, \mathrm{d}\beta = \sqrt{\frac{|P|}{(2\pi)^n}} \sqrt{\frac{|\Lambda_0|}{|\Lambda_n|}} \, \frac{{b_0}^{a_0}}{\Gamma(a_0)} \, \frac{\Gamma(a_n)}{{b_n}^{a_n}} \, \mathrm{Gam}(\tau; a_n, b_n) \; . \tag{50}$$

Finally, τ can also be integrated out:

$$\iint p(y,\beta,\tau)\,\mathrm{d}\beta\,\mathrm{d}\tau = \sqrt{\frac{|P|}{(2\pi)^n}} \sqrt{\frac{|\Lambda_0|}{|\Lambda_n|}} \, \frac{\Gamma(a_n)}{\Gamma(a_0)} \, \frac{{b_0}^{a_0}}{{b_n}^{a_n}} = p(y|m) \; . \tag{51}$$

Thus, the log model evidence of this model is given by

$$\begin{split} \log p(y|m) = \frac{1}{2} \log |P| - \frac{n}{2} \log(2\pi) + \frac{1}{2} \log |\Lambda_0| - \frac{1}{2} \log |\Lambda_n| + \\ \log \Gamma(a_n) - \log \Gamma(a_0) + a_0 \log b_0 - a_n \log b_n \; . \end{split} \tag{52}$$

Sources:

- Bishop CM (2006): “Bayesian linear regression”; in: *Pattern Recognition for Machine Learning*, pp. 152-161, ex. 3.23, eq. 3.118; URL: https://www.springer.com/gp/book/9780387310732.

Metadata: ID: P11 | shortcut: blr-lme | author: JoramSoch | date: 2020-01-03, 22:05.

1.3 General linear model

1.3.1 Maximum likelihood estimation

Proof: Maximum likelihood estimation for the general linear model

Index: The Book of Statistical Proofs ▷ Statistical Models ▷ Normal data ▷ General linear model ▷ Maximum likelihood estimation

Theorem: Given a general linear model with matrix-normally distributed errors

$$Y = XB + E, \; E \sim \mathcal{MN}(0, V, \Sigma) \; , \tag{53}$$

maximum likelihood estimates for the unknown parameters B and Σ are given by

$$\begin{split} \hat{B} &= (X^\mathrm{T} V^{-1} X)^{-1} X^\mathrm{T} V^{-1} Y \\ \hat{\Sigma} &= \frac{1}{n} (Y - X\hat{B})^\mathrm{T} V^{-1} (Y - X\hat{B}) \; . \end{split} \tag{54}$$

Proof: In (53), 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

$$\begin{split} p(Y|B,\Sigma) &= \mathcal{MN}(Y; XB, V, \Sigma) \\ &= \sqrt{\frac{1}{(2\pi)^{nv} |\Sigma|^n |V|^v}} \cdot \exp\left[-\frac{1}{2} \mathrm{tr}\left(\Sigma^{-1} (Y - XB)^\mathrm{T} V^{-1} (Y - XB) \right) \right] \end{split} \tag{55}$$

and the log-likelihood function

$$
\begin{aligned}
\mathrm{LL}(B, \Sigma) =& \log p(Y|B, \Sigma) \\
=& -\frac{nv}{2}\log(2\pi) - \frac{n}{2}\log|\Sigma| - \frac{v}{2}\log|V| \\
& - \frac{1}{2}\,\mathrm{tr}\left[\Sigma^{-1}(Y - XB)^{\mathrm{T}}V^{-1}(Y - XB)\right] \; .
\end{aligned} \tag{56}
$$

Substituting V^{-1} by the precision matrix P to ease notation, we have:

$$
\begin{aligned}
\mathrm{LL}(B, \Sigma) =& -\frac{nv}{2}\log(2\pi) - \frac{n}{2}\log(|\Sigma|) - \frac{v}{2}\log(|V|) \\
& - \frac{1}{2}\,\mathrm{tr}\left[\Sigma^{-1}\left(Y^{\mathrm{T}}PY - Y^{\mathrm{T}}PXB - B^{\mathrm{T}}X^{\mathrm{T}}PY + B^{\mathrm{T}}X^{\mathrm{T}}PXB\right)\right] \; .
\end{aligned} \tag{57}
$$

The derivative of the log-likelihood function (57) with respect to B is

$$
\begin{aligned}
\frac{\mathrm{dLL}(B, \Sigma)}{\mathrm{d}B} &= \frac{\mathrm{d}}{\mathrm{d}B}\left(-\frac{1}{2}\,\mathrm{tr}\left[\Sigma^{-1}\left(Y^{\mathrm{T}}PY - Y^{\mathrm{T}}PXB - B^{\mathrm{T}}X^{\mathrm{T}}PY + B^{\mathrm{T}}X^{\mathrm{T}}PXB\right)\right]\right) \\
&= \frac{\mathrm{d}}{\mathrm{d}B}\left(-\frac{1}{2}\,\mathrm{tr}\left[-2\Sigma^{-1}Y^{\mathrm{T}}PXB\right]\right) + \frac{\mathrm{d}}{\mathrm{d}B}\left(-\frac{1}{2}\,\mathrm{tr}\left[\Sigma^{-1}B^{\mathrm{T}}X^{\mathrm{T}}PXB\right]\right) \\
&= -\frac{1}{2}\left(-2X^{\mathrm{T}}PY\Sigma^{-1}\right) - \frac{1}{2}\left(X^{\mathrm{T}}PXB\Sigma^{-1} + (X^{\mathrm{T}}PX)^{\mathrm{T}}B(\Sigma^{-1})^{\mathrm{T}}\right) \\
&= X^{\mathrm{T}}PY\Sigma^{-1} - X^{\mathrm{T}}PXB\Sigma^{-1}
\end{aligned} \tag{58}
$$

and setting this derivative to zero gives the MLE for B:

$$
\begin{aligned}
\frac{\mathrm{dLL}(\hat{B}, \Sigma)}{\mathrm{d}B} &= 0 \\
0 &= X^{\mathrm{T}}PY\Sigma^{-1} - X^{\mathrm{T}}PX\hat{B}\Sigma^{-1} \\
0 &= X^{\mathrm{T}}PY - X^{\mathrm{T}}PX\hat{B} \\
X^{\mathrm{T}}PX\hat{B} &= X^{\mathrm{T}}PY \\
\hat{B} &= \left(X^{\mathrm{T}}PX\right)^{-1}X^{\mathrm{T}}PY
\end{aligned} \tag{59}
$$

The derivative of the log-likelihood function (56) at $\hat{B}$ with respect to Σ is

$$
\begin{aligned}
\frac{\mathrm{dLL}(\hat{B}, \Sigma)}{\mathrm{d}\Sigma} &= \frac{\mathrm{d}}{\mathrm{d}\Sigma}\left(-\frac{n}{2}\log|\Sigma| - \frac{1}{2}\,\mathrm{tr}\left[\Sigma^{-1}(Y - X\hat{B})^{\mathrm{T}}V^{-1}(Y - X\hat{B})\right]\right) \\
&= -\frac{n}{2}\left(\Sigma^{-1}\right)^{\mathrm{T}} + \frac{1}{2}\left(\Sigma^{-1}(Y - X\hat{B})^{\mathrm{T}}V^{-1}(Y - X\hat{B})\,\Sigma^{-1}\right)^{\mathrm{T}} \\
&= -\frac{n}{2}\,\Sigma^{-1} + \frac{1}{2}\,\Sigma^{-1}(Y - X\hat{B})^{\mathrm{T}}V^{-1}(Y - X\hat{B})\,\Sigma^{-1}
\end{aligned} \tag{60}
$$

and setting this derivative to zero gives the MLE for Σ:

$$
\begin{aligned}
\frac{\mathrm{dLL}(\hat{B},\hat{\Sigma})}{\mathrm{d}\Sigma} &= 0 \\
0 &= -\frac{n}{2}\,\hat{\Sigma}^{-1} + \frac{1}{2}\,\hat{\Sigma}^{-1}(Y - X\hat{B})^\mathrm{T} V^{-1} (Y - X\hat{B})\,\hat{\Sigma}^{-1} \\
\frac{n}{2}\,\hat{\Sigma}^{-1} &= \frac{1}{2}\,\hat{\Sigma}^{-1}(Y - X\hat{B})^\mathrm{T} V^{-1} (Y - X\hat{B})\,\hat{\Sigma}^{-1} \\
\hat{\Sigma}^{-1} &= \frac{1}{n}\,\hat{\Sigma}^{-1}(Y - X\hat{B})^\mathrm{T} V^{-1} (Y - X\hat{B})\,\hat{\Sigma}^{-1} \\
I_v &= \frac{1}{n}\,(Y - X\hat{B})^\mathrm{T} V^{-1} (Y - X\hat{B})\,\hat{\Sigma}^{-1} \\
\hat{\Sigma} &= \frac{1}{n}\,(Y - X\hat{B})^\mathrm{T} V^{-1} (Y - X\hat{B})
\end{aligned}
\tag{61}
$$

Together, (59) and (61) constitute the MLE for the GLM.

Sources:

- own work

Metadata: ID: P7 | shortcut: glm-mle | author: JoramSoch | date: 2019-12-06, 10:40.

2 Poisson data

2.1 Poisson-distributed data

2.1.1 Maximum likelihood estimation

Proof: Maximum likelihood estimation for Poisson-distributed data

Index: The Book of Statistical Proofs ▷ Statistical Models ▷ Poisson data ▷ Poisson-distributed data ▷ Maximum likelihood estimation

Theorem: Let $y = \{y_1, \ldots, y_n\}$ be a set of observed counts independent and identically distributed according to a Poisson distribution with rate λ:

$$y_i \sim \mathrm{Poiss}(\lambda), \quad i = 1, \ldots, n \; . \tag{62}$$

Then, the maximum likelihood estimate for the rate parameter λ is given by

$$\hat{\lambda} = \bar{y} \tag{63}$$

where $\bar{y}$ is the sample mean

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \; . \tag{64}$$

Proof: The likelihood function for each observation is given by the probability mass function of the Poisson distribution

$$p(y_i|\lambda) = \mathrm{Poiss}(y_i; \lambda) = \frac{\lambda^{y_i} \cdot \exp(-\lambda)}{y_i!} \tag{65}$$

and because observations are independent, the likelihood function for all observations is the product of the individual ones:

$$p(y|\lambda) = \prod_{i=1}^{n} p(y_i|\lambda) = \prod_{i=1}^{n} \frac{\lambda^{y_i} \cdot \exp(-\lambda)}{y_i!} \; . \tag{66}$$

Thus, the log-likelihood function is

$$\mathrm{LL}(\lambda) = \log p(y|\lambda) = \log \left[\prod_{i=1}^{n} \frac{\lambda^{y_i} \cdot \exp(-\lambda)}{y_i!} \right] \tag{67}$$

which can be developed into

$$\begin{aligned}
\mathrm{LL}(\lambda) &= \sum_{i=1}^{n} \log\left[\frac{\lambda^{y_i}\cdot\exp(-\lambda)}{y_i!}\right] \\
&= \sum_{i=1}^{n} \left[y_i\cdot\log(\lambda) - \lambda - \log(y_i!)\right] \\
&= -\sum_{i=1}^{n} \lambda + \sum_{i=1}^{n} y_i\cdot\log(\lambda) - \sum_{i=1}^{n}\log(y_i!) \\
&= -n\lambda + \log(\lambda)\sum_{i=1}^{n} y_i - \sum_{i=1}^{n}\log(y_i!)
\end{aligned} \tag{68}$$

The derivatives of the log-likelihood with respect to λ are

$$\begin{aligned}
\frac{\mathrm{dLL}(\lambda)}{\mathrm{d}\lambda} &= \frac{1}{\lambda}\sum_{i=1}^{n} y_i - n \\
\frac{\mathrm{d}^2\mathrm{LL}(\lambda)}{\mathrm{d}\lambda^2} &= -\frac{1}{\lambda^2}\sum_{i=1}^{n} y_i \; .
\end{aligned} \tag{69}$$

Setting the first derivative to zero, we obtain:

$$\begin{aligned}
\frac{\mathrm{dLL}(\hat{\lambda})}{\mathrm{d}\lambda} &= 0 \\
0 &= \frac{1}{\hat{\lambda}}\sum_{i=1}^{n} y_i - n \\
\hat{\lambda} &= \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y} \; .
\end{aligned} \tag{70}$$

Plugging this value into the second deriative, we confirm:

$$\begin{aligned}
\frac{\mathrm{d}^2\mathrm{LL}(\hat{\lambda})}{\mathrm{d}\lambda^2} &= -\frac{1}{\bar{y}^2}\sum_{i=1}^{n} y_i \\
&= -\frac{n\cdot\bar{y}}{\bar{y}^2} \\
&= -\frac{n}{\bar{y}} < 0 \; .
\end{aligned} \tag{71}$$

This demonstrates that the estimate $\hat{\lambda} = \bar{y}$ maximizes the likelihood $p(y \mid \lambda)$.

Sources:

- own work

Metadata: ID: P27 | shortcut: poiss-mle | author: JoramSoch | date: 2020-01-20, 21:53.

2.2 Poisson distribution with exposure values

2.2.1 Conjugate prior distribution

Proof: Conjugate prior distribution for the Poisson distribution with exposure values

Index: The Book of Statistical Proofs ▷ Statistical Models ▷ Poisson data ▷ Poisson distribution with exposure values ▷ Conjugate prior distribution

Theorem: Let $y = \{y_1, \ldots, y_n\}$ be a series of observed counts which are independently distributed according to a Poisson distribution (→ Definition "poiss") with common rate λ and concurrent exposures $\{x_1, \ldots, x_n\}$:

$$y_i \sim \mathrm{Poiss}(\lambda x_i), \quad i = 1, \ldots, n \; . \tag{72}$$

Then, the conjugate prior (→ Definition "conj-prior") for the model parameter λ is a gamma distribution:

$$p(\lambda) = \mathrm{Gam}(\lambda; a_0, b_0) \; . \tag{73}$$

Proof: With the probability mass function of the Poisson distribution (→ Proof "poiss-pmf"), the likelihood function for each observation implied by (72) is given by

$$p(y_i|\lambda) = \mathrm{Poiss}(y_i; \lambda x_i) = \frac{(\lambda x_i)^{y_i} \cdot \exp\left[-\lambda x_i\right]}{y_i!} \tag{74}$$

and because observations are independent, the likelihood function (→ Definition "lf") for all observations is the product of the individual ones:

$$p(y|\lambda) = \prod_{i=1}^{n} p(y_i|\lambda) = \prod_{i=1}^{n} \frac{(\lambda x_i)^{y_i} \cdot \exp\left[-\lambda x_i\right]}{y_i!} \; . \tag{75}$$

Resolving the product in the likelihood function, we have

$$\begin{aligned} p(y|\lambda) &= \prod_{i=1}^{n} \frac{{x_i}^{y_i}}{y_i!} \cdot \prod_{i=1}^{n} \lambda^{y_i} \cdot \prod_{i=1}^{n} \exp\left[-\lambda x_i\right] \\ &= \prod_{i=1}^{n} \left(\frac{{x_i}^{y_i}}{y_i!}\right) \cdot \lambda^{\sum_{i=1}^{n} y_i} \cdot \exp\left[-\lambda \sum_{i=1}^{n} x_i\right] \\ &= \prod_{i=1}^{n} \left(\frac{{x_i}^{y_i}}{y_i!}\right) \cdot \lambda^{n\bar{y}} \cdot \exp\left[-n\bar{x}\lambda\right] \end{aligned} \tag{76}$$

where $\bar{y}$ and $\bar{x}$ are the means of y and x respectively:

$$\begin{aligned} \bar{y} &= \frac{1}{n} \sum_{i=1}^{n} y_i \\ \bar{x} &= \frac{1}{n} \sum_{i=1}^{n} x_i \; . \end{aligned} \tag{77}$$

In other words, the likelihood function is proportional to a power of λ times an exponential of λ:

$$p(y|\lambda) \propto \lambda^{n\bar{y}} \cdot \exp\left[-n\bar{x}\lambda\right] \; . \tag{78}$$

The same is true for a gamma distribution over λ

$$p(\lambda) = \mathrm{Gam}(\lambda; a_0, b_0) \tag{79}$$

the probability density function of which (→ Proof "gamma-pdf")

$$p(\lambda) = \frac{{b_0}^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0\lambda] \tag{80}$$

exhibits the same proportionality

$$p(\lambda) \propto \lambda^{a_0-1} \cdot \exp[-b_0\lambda] \tag{81}$$

and is therefore conjugate relative to the likelihood.

Sources:

- Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB (2014): "Other standard single-parameter models"; in: *Bayesian Data Analysis*, 3rd edition, ch. 2.6, p. 45, eq. 2.14ff.; URL: http://www.stat.columbia.edu/~gelman/book/.

Metadata: ID: P41 | shortcut: poissexp-prior | author: JoramSoch | date: 2020-02-04, 14:11.

2.2.2 Posterior distribution

Proof: Posterior distribution for the Poisson distribution with exposure values

Index: The Book of Statistical Proofs ▷ Statistical Models ▷ Poisson data ▷ Poisson distribution with exposure values ▷ Posterior distribution

Theorem: Let $y = \{y_1, \ldots, y_n\}$ be a series of observed counts which are independently distributed according to a Poisson distribution (→ Definition "poiss") with common rate λ and concurrent exposures $\{x_1, \ldots, x_n\}$:

$$y_i \sim \mathrm{Poiss}(\lambda x_i), \quad i = 1, \ldots, n \; . \tag{82}$$

Moreover, assume a gamma prior distribution over the model parameter λ:

$$p(\lambda) = \mathrm{Gam}(\lambda; a_0, b_0) \; . \tag{83}$$

Then, the posterior distribution is also a gamma distribution

$$p(\lambda|y) = \mathrm{Gam}(\lambda; a_n, b_n) \tag{84}$$

and the posterior hyperparameters are given by

$$
\begin{aligned}
a_n &= a_0 + n\bar{y} \\
a_n &= a_0 + n\bar{x} \; .
\end{aligned} \tag{85}
$$

Proof: With the probability mass function of the Poisson distribution (→ Proof "poiss-pmf"), the likelihood function for each observation implied by (82) is given by

$$
p(y_i|\lambda) = \mathrm{Poiss}(y_i; \lambda x_i) = \frac{(\lambda x_i)^{y_i} \cdot \exp\left[-\lambda x_i\right]}{y_i!} \tag{86}
$$

and because observations are independent, the likelihood function (→ Definition "lf") for all observations is the product of the individual ones:

$$
p(y|\lambda) = \prod_{i=1}^{n} p(y_i|\lambda) = \prod_{i=1}^{n} \frac{(\lambda x_i)^{y_i} \cdot \exp\left[-\lambda x_i\right]}{y_i!} \; . \tag{87}
$$

Combining the likelihood function (87) with the prior distribution (83), the joint likelihood (→ Definition "jl") of the model is given by

$$
\begin{aligned}
p(y,\lambda) &= p(y|\lambda)\, p(\lambda) \\
&= \prod_{i=1}^{n} \frac{(\lambda x_i)^{y_i} \cdot \exp\left[-\lambda x_i\right]}{y_i!} \cdot \frac{{b_0}^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0 \lambda] \; .
\end{aligned} \tag{88}
$$

Resolving the product in the joint likelihood, we have

$$
\begin{aligned}
p(y,\lambda) &= \prod_{i=1}^{n} \frac{{x_i}^{y_i}}{y_i!} \prod_{i=1}^{n} \lambda^{y_i} \prod_{i=1}^{n} \exp\left[-\lambda x_i\right] \cdot \frac{{b_0}^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0 \lambda] \\
&= \prod_{i=1}^{n} \left(\frac{{x_i}^{y_i}}{y_i!} \right) \lambda^{\sum_{i=1}^{n} y_i} \exp\left[-\lambda \sum_{i=1}^{n} x_i \right] \cdot \frac{{b_0}^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0 \lambda] \\
&= \prod_{i=1}^{n} \left(\frac{{x_i}^{y_i}}{y_i!} \right) \lambda^{n\bar{y}} \exp\left[-n\bar{x}\lambda\right] \cdot \frac{{b_0}^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0 \lambda] \\
&= \prod_{i=1}^{n} \left(\frac{{x_i}^{y_i}}{y_i!} \right) \frac{{b_0}^{a_0}}{\Gamma(a_0)} \cdot \lambda^{a_0+n\bar{y}-1} \cdot \exp\left[-(b_0 + n\bar{x})\lambda\right]
\end{aligned} \tag{89}
$$

where $\bar{y}$ and $\bar{x}$ are the means of y and x respectively:

$$
\begin{aligned}
\bar{y} &= \frac{1}{n} \sum_{i=1}^{n} y_i \\
\bar{x} &= \frac{1}{n} \sum_{i=1}^{n} x_i \; .
\end{aligned} \tag{90}
$$

Note that the posterior distribution is proportional to the joint likelihood:

$$p(\lambda|y) \propto p(y, \lambda) \; . \tag{91}$$

Setting $a_n = a_0 + n\bar{y}$ and $b_n = b_0 + n\bar{x}$, the posterior distribution is therefore proportional to

$$p(\lambda|y) \propto \lambda^{a_n - 1} \cdot \exp\left[-b_n \lambda\right] \tag{92}$$

which, when normalized to one, results in the probability density function of the gamma distribution (→ Proof "gam-pdf"):

$$p(\lambda|y) = \frac{{b_n}^{a_n}}{\Gamma(a_0)} \lambda^{a_n - 1} \exp\left[-b_n \lambda\right] = \mathrm{Gam}(\lambda; a_n, b_n) \; . \tag{93}$$

Sources:

- Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB (2014): "Other standard single-parameter models"; in: *Bayesian Data Analysis*, 3rd edition, ch. 2.6, p. 45, eq. 2.15; URL: http://www.stat.columbia.edu/~gelman/book/.

Metadata: ID: P42 | shortcut: poissexp-post | author: JoramSoch | date: 2020-02-04, 14:42.

2.2.3 Log model evidence

Proof: Log model evidence for the Poisson distribution with exposure values

Index: The Book of Statistical Proofs ▷ Statistical Models ▷ Poisson data ▷ Poisson distribution with exposure values ▷ Log model evidence

Theorem: Let $y = \{y_1, \ldots, y_n\}$ be a series of observed counts which are independently distributed according to a Poisson distribution (→ Definition "poiss") with common rate λ and concurrent exposures $\{x_1, \ldots, x_n\}$:

$$y_i \sim \mathrm{Poiss}(\lambda x_i), \quad i = 1, \ldots, n \; . \tag{94}$$

Moreover, assume a gamma prior distribution over the model parameter λ:

$$p(\lambda) = \mathrm{Gam}(\lambda; a_0, b_0) \; . \tag{95}$$

Then, the log model evidence for this model is

$$\begin{split} \log p(y|m) = &\sum_{i=1}^{n} y_i \log(x_i) - \sum_{i=1}^{n} \log y_i ! + \\ &\log\Gamma(a_n) - \log\Gamma(a_0) + a_0 \log b_0 - a_n \log b_n \; . \end{split} \tag{96}$$

where the posterior hyperparameters are given by

$$\begin{split} a_n &= a_0 + n\bar{y} \\ a_n &= a_0 + n\bar{x} \; . \end{split} \tag{97}$$

Proof: With the probability mass function of the Poisson distribution (→ Proof "poiss-pmf"), the likelihood function for each observation implied by (94) is given by

$$p(y_i|\lambda) = \mathrm{Poiss}(y_i; \lambda x_i) = \frac{(\lambda x_i)^{y_i} \cdot \exp\left[-\lambda x_i\right]}{y_i!} \tag{98}$$

and because observations are independent, the likelihood function (→ Definition "lf") for all observations is the product of the individual ones:

$$p(y|\lambda) = \prod_{i=1}^{n} p(y_i|\lambda) = \prod_{i=1}^{n} \frac{(\lambda x_i)^{y_i} \cdot \exp\left[-\lambda x_i\right]}{y_i!} \; . \tag{99}$$

Combining the likelihood function (99) with the prior distribution (95), the joint likelihood (→ Definition "jl") of the model is given by

$$\begin{split} p(y,\lambda) &= p(y|\lambda) \, p(\lambda) \\ &= \prod_{i=1}^{n} \frac{(\lambda x_i)^{y_i} \cdot \exp\left[-\lambda x_i\right]}{y_i!} \cdot \frac{{b_0}^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0 \lambda] \; . \end{split} \tag{100}$$

Resolving the product in the joint likelihood, we have

$$\begin{split} p(y,\lambda) &= \prod_{i=1}^{n} \frac{{x_i}^{y_i}}{y_i!} \prod_{i=1}^{n} \lambda^{y_i} \prod_{i=1}^{n} \exp\left[-\lambda x_i\right] \cdot \frac{{b_0}^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0 \lambda] \\ &= \prod_{i=1}^{n} \left(\frac{{x_i}^{y_i}}{y_i!} \right) \lambda^{\sum_{i=1}^{n} y_i} \exp\left[-\lambda \sum_{i=1}^{n} x_i\right] \cdot \frac{{b_0}^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0 \lambda] \\ &= \prod_{i=1}^{n} \left(\frac{{x_i}^{y_i}}{y_i!} \right) \lambda^{n\bar{y}} \exp\left[-n\bar{x}\lambda\right] \cdot \frac{{b_0}^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0 \lambda] \\ &= \prod_{i=1}^{n} \left(\frac{{x_i}^{y_i}}{y_i!} \right) \frac{{b_0}^{a_0}}{\Gamma(a_0)} \cdot \lambda^{a_0+n\bar{y}-1} \cdot \exp\left[-(b_0 + n\bar{x})\lambda\right] \end{split} \tag{101}$$

where $\bar{y}$ and $\bar{x}$ are the means of y and x respectively:

$$\begin{split} \bar{y} &= \frac{1}{n} \sum_{i=1}^{n} y_i \\ \bar{x} &= \frac{1}{n} \sum_{i=1}^{n} x_i \; . \end{split} \tag{102}$$

Note that the model evidence is the marginal density of the joint likelihood:

$$p(y) = \int p(y,\lambda) \, \mathrm{d}\lambda \; . \tag{103}$$

Setting $a_n = a_0 + n\bar{y}$ and $b_n = b_0 + n\bar{x}$, the joint likelihood can also be written as

$$p(y, \lambda) = \prod_{i=1}^{n} \left(\frac{{x_i}^{y_i}}{y_i!} \right) \frac{{b_0}^{a_0}}{\Gamma(a_0)} \frac{\Gamma(a_n)}{{b_n}^{a_n}} \cdot \frac{{b_n}^{a_n}}{\Gamma(a_n)} \lambda^{a_n-1} \exp\left[-b_n \lambda\right] \; . \tag{104}$$

Using the probability density function of the gamma distribution (→ Proof "gam-pdf"), λ can now be integrated out easily

$$\begin{split} \mathrm{p}(y) &= \int \prod_{i=1}^{n} \left(\frac{{x_i}^{y_i}}{y_i!} \right) \frac{{b_0}^{a_0}}{\Gamma(a_0)} \frac{\Gamma(a_n)}{{b_n}^{a_n}} \cdot \frac{{b_n}^{a_n}}{\Gamma(a_n)} \lambda^{a_n-1} \exp\left[-b_n \lambda\right] \, \mathrm{d}\lambda \\ &= \prod_{i=1}^{n} \left(\frac{{x_i}^{y_i}}{y_i!} \right) \frac{\Gamma(a_n)}{\Gamma(a_0)} \frac{{b_0}^{a_0}}{{b_n}^{a_n}} \int \mathrm{Gam}(\lambda; a_n, b_n) \, \mathrm{d}\lambda \\ &= \prod_{i=1}^{n} \left(\frac{{x_i}^{y_i}}{y_i!} \right) \frac{\Gamma(a_n)}{\Gamma(a_0)} \frac{{b_0}^{a_0}}{{b_n}^{a_n}} \; , \end{split} \tag{105}$$

such that the log model evidence is shown to be

$$\begin{split} \log p(y|m) = &\sum_{i=1}^{n} y_i \log(x_i) - \sum_{i=1}^{n} \log y_i! + \\ &\log \Gamma(a_n) - \log \Gamma(a_0) + a_0 \log b_0 - a_n \log b_n \; . \end{split} \tag{106}$$

Sources:

- own work

Metadata: ID: P43 | shortcut: poissexp-lme | author: JoramSoch | date: 2020-02-04, 15:12.

3 Probability data

3.1 Beta-distributed data

3.1.1 Method of moments

Proof: Method of moments for beta-distributed data

Index: The Book of Statistical Proofs ▷ Statistical Models ▷ Probability data ▷ Beta-distributed data ▷ Method of moments

Theorem: Let $y = \{y_1, \ldots, y_n\}$ be a set of observed counts independent and identically distributed according to a beta distribution with shapes α and β:

$$y_i \sim \mathrm{Bet}(\alpha, \beta), \quad i = 1, \ldots, n \; . \tag{107}$$

Then, the method-of-moments estimates for the shape parameters α and β are given by

$$\begin{split} \hat{\alpha} &= \bar{y} \left(\frac{\bar{y}(1-\bar{y})}{\bar{v}} - 1 \right) \\ \hat{\beta} &= (1-\bar{y}) \left(\frac{\bar{y}(1-\bar{y})}{\bar{v}} - 1 \right) \end{split} \tag{108}$$

where $\bar{y}$ is the sample mean and $\bar{v}$ is the sample variance:

$$\begin{split} \bar{y} &= \frac{1}{n} \sum_{i=1}^n y_i \\ \bar{v} &= \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \; . \end{split} \tag{109}$$

Proof: Mean and variance of the beta distribution in terms of the parameters α and β are given by

$$\begin{split} \mathrm{E}(X) &= \frac{\alpha}{\alpha+\beta} \\ \mathrm{Var}(X) &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \; . \end{split} \tag{110}$$

Thus, matching the moments requires us to solve the following equation system for α and β:

$$\begin{split} \bar{y} &= \frac{\alpha}{\alpha+\beta} \\ \bar{v} &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \; . \end{split} \tag{111}$$

From the first equation, we can deduce:

$$\begin{aligned}
\bar{y}(\alpha+\beta) &= \alpha \\
\alpha\bar{y} + \beta\bar{y} &= \alpha \\
\beta\bar{y} &= \alpha - \alpha\bar{y} \\
\beta &= \frac{\alpha}{\bar{y}} - \alpha \\
\beta &= \alpha\left(\frac{1}{\bar{y}} - 1\right) \; .
\end{aligned} \tag{112}$$

If we define $q = 1/\bar{y} - 1$ and plug (112) into the second equation, we have:

$$\begin{aligned}
\bar{v} &= \frac{\alpha \cdot \alpha q}{(\alpha + \alpha q)^2 (\alpha + \alpha q + 1)} \\
&= \frac{\alpha^2 q}{(\alpha(1+q))^2 (\alpha(1+q) + 1)} \\
&= \frac{q}{(1+q)^2 (\alpha(1+q) + 1)} \\
&= \frac{q}{\alpha(1+q)^3 + (1+q)^2} \\
q &= \bar{v}\left[\alpha(1+q)^3 + (1+q)^2\right] \; .
\end{aligned} \tag{113}$$

Noting that $1 + q = 1/\bar{y}$ and $q = (1-\bar{y})/\bar{y}$, one obtains for α:

$$\begin{aligned}
\frac{1-\bar{y}}{\bar{y}} &= \bar{v}\left[\frac{\alpha}{\bar{y}^3} + \frac{1}{\bar{y}^2}\right] \\
\frac{1-\bar{y}}{\bar{y}\,\bar{v}} &= \frac{\alpha}{\bar{y}^3} + \frac{1}{\bar{y}^2} \\
\frac{\bar{y}^3(1-\bar{y})}{\bar{y}\,\bar{v}} &= \alpha + \bar{y} \\
\alpha &= \frac{\bar{y}^2(1-\bar{y})}{\bar{v}} - \bar{y} \\
&= \bar{y}\left(\frac{\bar{y}(1-\bar{y})}{\bar{v}} - 1\right) \; .
\end{aligned} \tag{114}$$

Plugging this into equation (112), one obtains for β:

$$\begin{aligned}
\beta &= \bar{y}\left(\frac{\bar{y}(1-\bar{y})}{\bar{v}} - 1\right) \cdot \left(\frac{1-\bar{y}}{\bar{y}}\right) \\
&= (1-\bar{y})\left(\frac{\bar{y}(1-\bar{y})}{\bar{v}} - 1\right) \; .
\end{aligned} \tag{115}$$

Together, (114) and (115) 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 Categorical data

4.1 Binomial observations

4.1.1 Conjugate prior distribution

Proof: Conjugate prior distribution for binomial observations

Index: The Book of Statistical Proofs ▷ Statistical Models ▷ Categorical data ▷ Binomial observations ▷ 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:

$$y \sim \mathrm{Bin}(n,p) \; . \tag{116}$$

Then, the conjugate prior for the model parameter p is a beta distribution:

$$\mathrm{p}(p) = \mathrm{Bet}(p; \alpha_0, \beta_0) \; . \tag{117}$$

Proof: With the probability mass function of the binomial distribution (→ Proof "bin-pmf"), the likelihood function implied by (116) is given by

$$\mathrm{p}(y|p) = \binom{n}{y} \, p^y \, (1-p)^{n-y} \; . \tag{118}$$

In other words, the likelihood function is proportional to a power of p times a power of $(1-p)$:

$$\mathrm{p}(y|p) \propto p^y \, (1-p)^{n-y} \; . \tag{119}$$

The same is true for a beta distribution over p

$$\mathrm{p}(p) = \mathrm{Bet}(p; \alpha_0, \beta_0) \tag{120}$$

the probability density function of which (→ Proof "beta-pdf")

$$\mathrm{p}(p) = \frac{1}{B(\alpha_0, \beta_0)} \, p^{\alpha_0-1} \, (1-p)^{\beta_0-1} \tag{121}$$

exhibits the same proportionality

$$\mathrm{p}(p) \propto p^{\alpha_0-1} \, (1-p)^{\beta_0-1} \tag{122}$$

and is therefore conjugate relative to the likelihood.

Sources:

- Wikipedia (2020): "Binomial distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-01-23; URL: https://en.wikipedia.org/wiki/Binomial_distribution#Estimation_of_parameters.

Metadata: ID: P29 | shortcut: bin-prior | author: JoramSoch | date: 2020-01-23, 23:38.

4.1.2 Posterior distribution

Proof: Posterior distribution for binomial observations

Index: The Book of Statistical Proofs ▷ Statistical Models ▷ Categorical data ▷ Binomial observations ▷ 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:

$$y \sim \mathrm{Bin}(n, p) \; . \tag{123}$$

Moreover, assume a beta prior distribution over the model parameter p:

$$\mathrm{p}(p) = \mathrm{Bet}(p; \alpha_0, \beta_0) \; . \tag{124}$$

Then, the posterior distribution is also a beta distribution

$$\mathrm{p}(p|y) = \mathrm{Bet}(p; \alpha_n, \beta_n) \; . \tag{125}$$

and the posterior hyperparameters are given by

$$\begin{split} \alpha_n &= \alpha_0 + y \\ \beta_n &= \beta_0 + (n - y) \; . \end{split} \tag{126}$$

Proof: With the probability mass function of the binomial distribution (→ Proof "bin-pmf"), the likelihood function implied by (123) is given by

$$\mathrm{p}(y|p) = \binom{n}{y} \, p^y \, (1-p)^{n-y} \; . \tag{127}$$

Combining the likelihood function (127) with the prior distribution (124), the joint likelihood of the model is given by

$$\begin{split} \mathrm{p}(y, p) &= \mathrm{p}(y|p) \, \mathrm{p}(p) \\ &= \binom{n}{y} \, p^y \, (1-p)^{n-y} \cdot frac1B(\alpha_0, \beta_0) \, p^{\alpha_0 - 1} \, (1-p)^{\beta_0 - 1} \\ &= \frac{1}{B(\alpha_0, \beta_0)} \binom{n}{y} \, p^{\alpha_0 + y - 1} \, (1-p)^{\beta_0 + (n-y) - 1} \; . \end{split} \tag{128}$$

Note that the posterior distribution is proportional to the joint likelihood:

$$\mathrm{p}(p|y) \propto \mathrm{p}(y, p) \; . \tag{129}$$

Setting $\alpha_n = \alpha_0 + y$ and $\beta_n = \beta_0 + (n - y)$, the posterior distribution is therefore proportional to

$$\mathrm{p}(p|y) \propto p^{\alpha_n - 1} \, (1-p)^{\beta_n - 1} \tag{130}$$

which, when normalized to one, results in the probability density function of the beta distribution (→ Proof "beta-pdf"):

$$\mathrm{p}(p|y) = \frac{1}{B(\alpha_n, \beta_n)} \, p^{\alpha_n - 1} \, (1-p)^{\beta_n - 1} = \mathrm{Bet}(p; \alpha_n, \beta_n) \; . \tag{131}$$

Sources:

- Wikipedia (2020): "Binomial distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-01-23; URL: https://en.wikipedia.org/wiki/Binomial_distribution#Estimation_of_parameters.

Metadata: ID: P30 | shortcut: bin-post | author: JoramSoch | date: 2020-01-24, 00:20.

4.1.3 Log model evidence

Proof: Log model evidence for binomial observations

Index: The Book of Statistical Proofs ▷ Statistical Models ▷ Categorical data ▷ Binomial observations ▷ 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:

$$y \sim \mathrm{Bin}(n, p) \; . \tag{132}$$

Moreover, assume a beta prior distribution over the model parameter p:

$$\mathrm{p}(p) = \mathrm{Bet}(p; \alpha_0, \beta_0) \; . \tag{133}$$

Then, the log model evidence for this model is

$$\log \mathrm{p}(y|m) = \log \binom{n}{y} + \log B(\alpha_n, \beta_n) - \log B(\alpha_0, \beta_0) \tag{134}$$

where the posterior hyperparameters are given by

$$\begin{aligned} \alpha_n &= \alpha_0 + y \\ \beta_n &= \beta_0 + (n - y) \; . \end{aligned} \tag{135}$$

Proof: With the probability mass function of the binomial distribution (→ Proof "bin-pmf"), the likelihood function implied by (132) is given by

$$\mathrm{p}(y|p) = \binom{n}{y} \, p^y \, (1-p)^{n-y} \; . \tag{136}$$

Combining the likelihood function (136) with the prior distribution (133), the joint likelihood of the model is given by

$$
\begin{aligned}
\mathrm{p}(y,p) &= \mathrm{p}(y|p)\,\mathrm{p}(p) \\
&= \binom{n}{y}\, p^y\, (1-p)^{n-y} \cdot frac1B(\alpha_0,\beta_0)\, p^{\alpha_0-1}\, (1-p)^{\beta_0-1} \\
&= \binom{n}{y}\, \frac{1}{B(\alpha_0,\beta_0)}\, p^{\alpha_0+y-1}\, (1-p)^{\beta_0+(n-y)-1} \; .
\end{aligned} \tag{137}
$$

Note that the model evidence is the marginal density of the joint likelihood:

$$
\mathrm{p}(y) = \int \mathrm{p}(y,p)\, \mathrm{d}p \; . \tag{138}
$$

Setting $\alpha_n = \alpha_0 + y$ and $\beta_n = \beta_0 + (n-y)$, the joint likelihood can also be written as

$$
\mathrm{p}(y,p) = \binom{n}{y}\, \frac{1}{B(\alpha_0,\beta_0)}\, \frac{B(\alpha_n,\beta_n)}{1}\, \frac{1}{B(\alpha_n,\beta_n)}\, p^{\alpha_n-1}\, (1-p)^{\beta_n-1} \; . \tag{139}
$$

Using the probability density function of the beta distribution (→ Proof “beta-pdf”), p can now be integrated out easily

$$
\begin{aligned}
\mathrm{p}(y) &= \int \binom{n}{y}\, \frac{1}{B(\alpha_0,\beta_0)}\, \frac{B(\alpha_n,\beta_n)}{1}\, \frac{1}{B(\alpha_n,\beta_n)}\, p^{\alpha_n-1}\, (1-p)^{\beta_n-1}\, \mathrm{d}p \\
&= \binom{n}{y}\, \frac{B(\alpha_n,\beta_n)}{B(\alpha_0,\beta_0)} \int \mathrm{Bet}(p;\alpha_n,\beta_n)\, \mathrm{d}p \\
&= \binom{n}{y}\, \frac{B(\alpha_n,\beta_n)}{B(\alpha_0,\beta_0)} \; ,
\end{aligned} \tag{140}
$$

such that the log model evidence is shown to be

$$
\log \mathrm{p}(y|m) = \log \binom{n}{y} + \log B(\alpha_n,\beta_n) - \log B(\alpha_0,\beta_0) \; . \tag{141}
$$

Sources:

- Wikipedia (2020): “Beta-binomial distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-01-24; URL: https://en.wikipedia.org/wiki/Beta-binomial_distribution#Motivation_and_derivation.

Metadata: ID: P31 | shortcut: bin-lme | author: JoramSoch | date: 2020-01-24, 00:44.

Chapter IV

Model Selection

1 Goodness-of-fit measures

1.1 R-squared

1.1.1 Derivation of R² and adjusted R²

Proof: Derivation of R^2 and adjusted R^2

Index: The Book of Statistical Proofs ▷ Model Selection ▷ Goodness-of-fit measures ▷ R-squared ▷ Derivation of R^2 and adjusted R^2

Theorem: Given a linear regression model

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

with n independent observations and p independent variables,
1) the coefficient of determination is given by

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} \tag{2}$$

2) the adjusted coefficient of determination is

$$R^2_{\text{adj}} = 1 - \frac{\text{RSS}/(n-p)}{\text{TSS}/(n-1)} \tag{3}$$

where the residual and total sum of squares are

$$\begin{split} \text{RSS} &= \sum_{i=1}^n (y_i - \hat{y}_i)^2, \quad \hat{y} = X\hat{\beta} \\ \text{TSS} &= \sum_{i=1}^n (y_i - \bar{y})^2 \; , \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \end{split} \tag{4}$$

where X is the $n \times p$ design matrix and $\hat{\beta}$ are the ordinary least squares estimates.

Proof: The coefficient of determination R^2 is defined as 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 as

$$\text{ESS} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \; , \tag{5}$$

then R^2 is given by

$$R^2 = \frac{\text{ESS}}{\text{TSS}} \; . \tag{6}$$

which is equal to

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} \; , \tag{7}$$

because TSS = ESS + RSS.

2) Using (4), the coefficient of determination can be also written as:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\frac{1}{n}\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2} \; . \tag{8}$$

If we replace the variance estimates by their unbiased estimators, we obtain

$$R^2_{\mathrm{adj}} = 1 - \frac{\frac{1}{n-p}\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\mathrm{RSS}/\mathrm{df}_r}{\mathrm{TSS}/\mathrm{df}_t} \tag{9}$$

where $\mathrm{df}_r = n - p$ and $\mathrm{df}_t = n - 1$ are the residual and total degrees of freedom.

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.

1.1.2 Relationship to maximum log-likelihood

Proof: Relationship between R^2 and maximum log-likelihood

Index: The Book of Statistical Proofs ▷ Model Selection ▷ Goodness-of-fit measures ▷ R-squared ▷ Relationship to maximum log-likelihood

Theorem: Given a linear regression model with independent observations

$$y = X\beta + \varepsilon, \; \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \; , \tag{10}$$

the coefficient of determination can be expressed in terms of the maximum log-likelihood as

$$R^2 = 1 - \left(\exp[\Delta\mathrm{MLL}]\right)^{-2/n} \tag{11}$$

where n is the number of observations and ΔMLL is the difference in maximum log-likelihood between the model given by (10) and a linear regression model with only a constant regressor.

Proof: First, we express the maximum log-likelihood (MLL) of a linear regression model in terms of its residual sum of squares (RSS). The model in (10) implies the following log-likelihood function

$$\mathrm{LL}(\beta, \sigma^2) = \log p(y|\beta, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y - X\beta)^\mathrm{T}(y - X\beta) \; , \tag{12}$$

such that maximum likelihood estimates are

$$\hat{\beta} = (X^\mathrm{T}X)^{-1}X^\mathrm{T}y \tag{13}$$

$$\hat{\sigma}^2 = \frac{1}{n}(y - X\hat{\beta})^\mathrm{T}(y - X\hat{\beta}) \tag{14}$$

and the residual sum of squares is

$$\mathrm{RSS} = \sum_{i=1}^{n} \hat{\varepsilon}_i = \hat{\varepsilon}^\mathrm{T}\hat{\varepsilon} = (y - X\hat{\beta})^\mathrm{T}(y - X\hat{\beta}) = n \cdot \hat{\sigma}^2 \; . \tag{15}$$

Since $\hat{\beta}$ and $\hat{\sigma}^2$ are maximum likelihood estimates, plugging them into the log-likelihood function gives the maximum log-likelihood:

$$\mathrm{MLL} = \mathrm{LL}(\hat{\beta}, \hat{\sigma}^2) = -\frac{n}{2}\log(2\pi\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2}(y - X\hat{\beta})^\mathrm{T}(y - X\hat{\beta}) \; . \tag{16}$$

With (15) for the first $\hat{\sigma}^2$ and (14) for the second $\hat{\sigma}^2$, the MLL becomes

$$\mathrm{MLL} = -\frac{n}{2}\log(\mathrm{RSS}) - \frac{n}{2}\log\left(\frac{2\pi}{n}\right) - \frac{n}{2} \; . \tag{17}$$

Second, we establish the relationship between maximum log-likelihood (MLL) and coefficient of determination (R^2). Consider the two models

$$\begin{aligned} m_0 &: \; X_0 = 1_n \\ m_1 &: \; X_1 = X \end{aligned} \tag{18}$$

For m_1, the residual sum of squares is given by (15); and for m_0, the residual sum of squares is equal to the total sum of squares:

$$\mathrm{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \; . \tag{19}$$

Using (17), we can therefore write

$$\Delta\mathrm{MLL} = \mathrm{MLL}(m_1) - \mathrm{MLL}(m_0) = -\frac{n}{2}\log(\mathrm{RSS}) + \frac{n}{2}\log(\mathrm{TSS}) \; . \tag{20}$$

Exponentiating both sides of the equation, we have:

$$\begin{aligned} \exp[\Delta\mathrm{MLL}] &= \exp\left[-\frac{n}{2}\log(\mathrm{RSS}) + \frac{n}{2}\log(\mathrm{TSS})\right] \\ &= \left(\exp\left[\log(\mathrm{RSS}) - \log(\mathrm{TSS})\right]\right)^{-n/2} \\ &= \left(\frac{\exp[\log(\mathrm{RSS})]}{\exp[\log(\mathrm{TSS})]}\right)^{-n/2} \\ &= \left(\frac{\mathrm{RSS}}{\mathrm{TSS}}\right)^{-n/2} \; . \end{aligned} \tag{21}$$

Taking both sides to the power of $-2/n$ and subtracting from 1, we have

$$\begin{aligned}
(\exp[\Delta\mathrm{MLL}])^{-2/n} &= \frac{\mathrm{RSS}}{\mathrm{TSS}} \\
1 - (\exp[\Delta\mathrm{MLL}])^{-2/n} &= 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} = R^2
\end{aligned} \tag{22}$$

which proves the identity given above.

Sources:

- own work

Metadata: ID: P14 | shortcut: rsq-mll | author: JoramSoch | date: 2020-01-08, 04:46.

2 Classical information criteria

2.1 Bayesian information criterion

2.1.1 Derivation

Proof: Derivation of the Bayesian information criterion

Index: The Book of Statistical Proofs ▷ Model Selection ▷ Classical information criteria ▷ Bayesian information criterion ▷ Derivation

Theorem: Let $p(y \mid \theta, m)$ be the likelihood function (→ Definition "lf") of a generative model $m \in \mathcal{M}$ with model parameters $\theta \in \Theta$ describing measured data $y \in \mathbb{R}^n$. Let $p(\theta \mid m)$ be a prior distribution (→ Definition "prior") on the model parameters. Assume that likelihood function and prior density are twice differentiable.
Then, as the number of data points goes to infinity, an approximation to the log-marginal likelihood (→ Definition "ml") $\log p(y \mid m)$, up to constant terms not depending on the model, is given by the Bayesian information criterion (BIC) as

$$-2 \log p(y \mid m) \approx \mathrm{BIC}(m) = -2 \log p(y \mid \hat{\theta}, m) + p \log n \tag{23}$$

where $\hat{\theta}$ is the maximum likelihood estimator (→ Definition "mle") (MLE) of θ, n is the number of data points and p is the number of model parameters.

Proof: Let $\mathrm{LL}(\theta)$ be the log-likelihood function

$$\mathrm{LL}(\theta) = \log p(y|\theta, m) \tag{24}$$

and define the functions g and h as follows:

$$\begin{aligned} g(\theta) &= p(\theta|m) \\ h(\theta) &= \frac{1}{n} \, \mathrm{LL}(\theta) \; . \end{aligned} \tag{25}$$

Then, the marginal likelihood can be written as follows:

$$\begin{aligned} p(y|m) &= \int_{\Theta} p(y|\theta, m) \, p(\theta|m) \, \mathrm{d}\theta \\ &= \int_{\Theta} \exp\left[n \, h(\theta)\right] \, g(\theta) \, \mathrm{d}\theta \; . \end{aligned} \tag{26}$$

This is an integral suitable for Laplace approximation which states that

$$\int_{\Theta} \exp\left[n \, h(\theta)\right] \, g(\theta) \, \mathrm{d}\theta = \left(\sqrt{\frac{2\pi}{n}}\right)^p \exp\left[n \, h(\theta_0)\right] \left(g(\theta_0) \left|J(\theta_0)\right|^{-1/2} + O(1/n)\right) \tag{27}$$

where θ_0 is the value that maximizes $h(\theta)$ and $J(\theta_0)$ is the Hessian matrix evaluated at θ_0. In our case, we have $h(\theta) = 1/n \, \mathrm{LL}(\theta)$ such that θ_0 is the maximum likelihood estimator $\hat{\theta}$:

$$\hat{\theta} = \operatorname*{arg\,max}_{\theta} \mathrm{LL}(\theta) \; . \tag{28}$$

With this, (27) can be applied to (26) using (25) to give:

$$p(y|m) \approx \left(\sqrt{\frac{2\pi}{n}} \right)^p p(y|\hat{\theta}, m) \, p(\hat{\theta}|m) \, \left| J(\hat{\theta}) \right|^{-1/2} \; . \tag{29}$$

Logarithmizing and multiplying with -2, we have:

$$-2 \log p(y|m) \approx -2 \, \mathrm{LL}(\hat{\theta}) + p \log n - p \log(2\pi) - 2 \log p(\hat{\theta}|m) + \log \left| J(\hat{\theta}) \right| \; . \tag{30}$$

As $n \to \infty$, the last three terms are $O_p(1)$ and can therefore be ignored when comparing between models $\mathcal{M} = \{m_1, \ldots, m_M\}$ and using $p(y \mid m_j)$ to compute posterior model probabilies $p(m_j \mid y)$. With that, the BIC is given as

$$\mathrm{BIC}(m) = -2 \log p(y|\hat{\theta}, m) + p \log n \; . \tag{31}$$

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.

3 Bayesian model selection

3.1 Log model evidence

3.1.1 Derivation

Proof: Derivation of the log model evidence

Index: The Book of Statistical Proofs ▷ Model Selection ▷ Bayesian model selection ▷ Log model evidence ▷ Derivation

Theorem: Let $p(y \mid \theta, m)$ be a likelihood function of a generative model m for making inferences on model parameters θ given measured data y. Moreover, let $p(\theta \mid m)$ be a prior distribution on model parameters θ. Then, the log model evidence (LME), also called marginal log-likelihood,

$$\mathrm{LME}(m) = \log p(y|m) \; , \tag{32}$$

can be expressed
1) as

$$\mathrm{LME}(m) = \log \int p(y|\theta,m) \, p(\theta|m) \, \mathrm{d}\theta \tag{33}$$

2) or

$$\mathrm{LME}(m) = \log p(y|\theta,m) + \log p(\theta|m) - \log p(\theta|y,m) \; . \tag{34}$$

Proof:
1) The first expression is a simple consequence of the law of marginal probability for continuous variables according to which

$$p(y|m) = \int p(y|\theta,m) \, p(\theta|m) \, \mathrm{d}\theta \tag{35}$$

which, when logarithmized, gives

$$\mathrm{LME}(m) = \log p(y|m) = \log \int p(y|\theta,m) \, p(\theta|m) \, \mathrm{d}\theta \; . \tag{36}$$

2) The second expression can be derived from Bayes' theorem which makes a statement about the posterior distribution:

$$p(\theta|y,m) = \frac{p(y|\theta,m) \, p(\theta|m)}{p(y|m)} \; . \tag{37}$$

Rearranging for $p(y \mid m)$ and logarithmizing, we have:

$$\begin{split} \mathrm{LME}(m) = \log p(y|m) &= \log \frac{p(y|\theta,m) \, p(\theta|m)}{p(\theta|y,m)} \\ &= \log p(y|\theta,m) + \log p(\theta|m) - \log p(\theta|y,m) \; . \end{split} \tag{38}$$

Sources:

- own work

Metadata: ID: P13 | shortcut: lme-der | author: JoramSoch | date: 2020-01-06, 21:27.

3.1.2 Partition into accuracy and complexity

Proof: Partition of the log model evidence into accuracy and complexity

Index: The Book of Statistical Proofs ▷ Model Selection ▷ Bayesian model selection ▷ Log model evidence ▷ Partition into accuracy and complexity

Theorem: The log model evidence can be partitioned into accuracy and complexity

$$\mathrm{LME}(m) = \mathrm{Acc}(m) - \mathrm{Com}(m) \tag{39}$$

where the accuracy term is the posterior expectation of the log-likelihood function

$$\mathrm{Acc}(m) = \left\langle p(y|\theta,m) \right\rangle_{p(\theta|y,m)} \tag{40}$$

and the complexity penalty is the Kullback-Leibler divergence of posterior from prior

$$\mathrm{Com}(m) = \mathrm{KL}\left[p(\theta|y,m) \, || \, p(\theta|m) \right] \; . \tag{41}$$

Proof: We consider Bayesian inference on data y using model m with parameters θ. Then, Bayes' theorem makes a statement about the posterior distribution, i.e. the probability of parameters, given the data and the model:

$$p(\theta|y,m) = \frac{p(y|\theta,m) \, p(\theta|m)}{p(y|m)} \; . \tag{42}$$

Rearranging this for the model evidence, we have:

$$p(y|m) = \frac{p(y|\theta,m) \, p(\theta|m)}{p(\theta|y,m)} \; . \tag{43}$$

Logarthmizing both sides of the equation, we obtain:

$$\log p(y|m) = \log p(y|\theta,m) - \log \frac{p(\theta|y,m)}{p(\theta|m)} \; . \tag{44}$$

Now taking the expectation over the posterior distribution yields:

$$\log p(y|m) = \int p(\theta|y,m) \log p(y|\theta,m) \, \mathrm{d}\theta - \int p(\theta|y,m) \log \frac{p(\theta|y,m)}{p(\theta|m)} \, \mathrm{d}\theta \; . \tag{45}$$

By definition, the left-hand side is the log model evidence and the terms on the right-hand side correspond to the posterior expectation of the log-likelihood function and the Kullback-Leibler divergence of posterior from prior

$$\mathrm{LME}(m) = \left\langle p(y|\theta,m) \right\rangle_{p(\theta|y,m)} - \mathrm{KL}\left[p(\theta|y,m) \, || \, p(\theta|m) \right] \tag{46}$$

which proofs the partition given by (39).

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.

Metadata: ID: P3 | shortcut: lme-anc | author: JoramSoch | date: 2019-09-27, 16:13.