The Book of Statistical Proofs

Joram Soch, BCCN Berlin
joram.soch@bccn-berlin.de

https://github.com/JoramSoch/TBSP

Contents

Chapter I

General Theorems

1 Probability theory: Bayes' theorem

Index:
▷ The Book of Statistical Proofs
  ▷ General Theorems
    ▷ Probability theory: 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

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

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:

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

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

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

Both equations can be rearranged for the joint probability

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

from which Bayes' theorem can be directly derived:

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

■

Dependencies:
- law of conditional probability, also called "product rule of probability"

Source:
- Koch, Karl-Rudolf (2007): *Introduction to Bayesian Statistics*, second edition, Springer, Berlin/Heidelberg, 2007, pp. 6/13, eqs. 2.12/2.38.

Index: number: A004; shortcut: prob-bayes; author: JoramSoch; date: 2019/05/03.

Chapter II

Probability Distributions

1 Multivariate normal distribution: Linear transformation theorem

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

Theorem: Let x follow a multivariate normal distribution:

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

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

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

Proof: The moment-generating function of a random vector x is

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

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

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

The joint moment-generating function of the multivariate normal distribution is

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

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

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

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

Dependencies:
- moment-generating function of a random vector
- joint moment-generating function of the multivariate normal distribution

Source:

- 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.

Index: number: A001; shortcut: mvn-ltt; author: JoramSoch; date: 2019/05/02.

Chapter III

Statistical Models

1 Mutiple linear regression: Ordinary least squares

Index:
▷ The Book of Statistical Proofs
  ▷ Statistical Models
    ▷ Mutiple 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:

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

■

Source:
- Stephen, 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

Index: number: A002; shortcut: mlr-ols; author: JoramSoch; date: 2019/05/02.

Chapter IV

Model Selection

1 Log model evidence: Partition into accuracy and complexity

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

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

$$\mathrm{LME}(m) = \mathrm{Acc}(m) - \mathrm{Com}(m) \tag{1}$$

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

$$\mathrm{Acc}(m) = \left\langle p(y|\theta,m) \right\rangle_{p(\theta|y,m)} \tag{2}$$

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

$$\mathrm{Com}(m) = \mathrm{KL}\left[p(\theta|y,m) \,||\, p(\theta|m)\right] \ . \tag{3}$$

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:

$$p(\theta|y,m) = \frac{p(y|\theta,m)\,p(\theta|m)}{p(y|m)} \ . \tag{4}$$

Rearranging this for the model evidence, we have:

$$p(y|m) = \frac{p(y|\theta,m)\,p(\theta|m)}{p(\theta|y,m)} \ . \tag{5}$$

Logarthmizing both sides of the equation, we obtain:

$$\log p(y|m) = \log p(y|\theta,m) - \log \frac{p(\theta|y,m)}{p(\theta|m)} \ . \tag{6}$$

Now taking the expectation over the posterior distribution yields:

$$\log p(y|m) = \int p(\theta|y,m) \log p(y|\theta,m) \,\mathrm{d}\theta - \int p(\theta|y,m) \log \frac{p(\theta|y,m)}{p(\theta|m)} \,\mathrm{d}\theta \ . \tag{7}$$

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

$$\mathrm{LME}(m) = \left\langle p(y|\theta,m) \right\rangle_{p(\theta|y,m)} - \mathrm{KL}\left[p(\theta|y,m) \,||\, p(\theta|m)\right] \tag{8}$$

which proofs the partition given by (1). ■

Dependencies:
- Bayes' theorem
- derivation of the log model evidence
- expectation with respect to a random variable
- Kullback-Leibler divergence of two random variables

Source:
- Penny et al. (2007): "Bayesian Comparison of Spatially Regularised General Linear Models". *Human Brain Mapping*, vol. 28, pp. 275–293.
- Soch et al. (2016): "How to avoid mismodelling in GLM-based fMRI data analysis: cross-validated Bayesian model selection". *NeuroImage*, vol. 141, pp. 469–489.

Index: number: A003; shortcut: lme-anc; author: JoramSoch; date: 2019/05/02.