

Bayesian model comparison

Suppose we have two models, model M_0 and model M_1 , not necessarily with the same number of parameters. Our prior information on these models can be used to assign each of them a prior probability. Usually one takes $P[M_1] = P[M_0] = \frac{1}{2}$, if there is no preference for a particular model. The prior is then called non-informative (see below). We can use a data-set D to update these prior probabilities using Bayes' formula which is

$$P[M_i|D] = \frac{P[D|M_i] P[M_i]}{P[D]} \quad (1)$$

Here, $P[M_i|D]$ is the posterior probability of model M_i (i takes the values 0 or 1 in this example), $P[D|M_i]$ is the probability of the data given model M_i and $P[D]$ is the so-called unconditional marginal likelihood of the data.

Bayesian model comparison and selection can be done in various ways (Berger 1985, Berger & Pericchi 1996, Carlin & Louis 2000; Wasserman 2000). We use a variant of the so-called Bayes factor (Kass & Raftery 1995).

Bayes factor

The Bayes factor B is the ratio of the posterior probabilities of two models (here M_0 and M_1 ; Kass & Raftery 1995):

$$B = \frac{P[M_1|D]}{P[M_0|D]} = \frac{P[D|M_1] P[M_1]}{P[D|M_0] P[M_0]} \quad (2)$$

Because $P[M_i]$ is assumed to be known, we only need to determine $P[D|M_i]$, the marginal likelihood of the data conditional on the model M_i . If there are no model parameters, the marginal likelihood immediately follows from the model. For example, the MacArthur broken-stick model (MacArthur 1957, 1960; Tokeshi 1990; see main text) has no free parameters. However, usually there are free parameters in the model, as in the two other models discussed in this paper. Let us denote the whole set of these model parameters in model i by Θ_i . The marginal likelihood is then a weighted average of the likelihoods $P[D|\Theta_i, M_i]$. When the weights are taken from the prior distribution of the model parameters (given the model), that is, from $P[\Theta_i|M_i]$, then we obtain the prior marginal likelihood,

$$P_{\text{prior}}[D|M_i] = \int P[D|\Theta_i, M_i] P[\Theta_i|M_i] d\Theta_i \quad (3)$$

Hence the prior marginal likelihood can be interpreted as an average likelihood weighted by the prior distribution.

When the weights are taken from the posterior distribution of the model parameters, that is, from $P[\Theta_i|D, M_i]$,

we obtain the posterior marginal likelihood,

$$P_{\text{posterior}}[D|M_i] = \int P[D|\Theta_i, M_i]P[\Theta_i|D, M_i]d\Theta_i \quad (4)$$

which is the likelihood averaged with the posterior distribution. We will use the posterior marginal likelihood to compare different models, because it is insensitive to the choice of the prior distribution.

The Bayes factor (2) with posterior marginal likelihoods (4) is known as the posterior Bayes factor (Aitkin 1991, Laud & Ibrahim 1995, De Santis & Spezzaferri 1997, Upadhyay & Peshwani 2003, Vlachos & Gelfand 2003). It is the ratio of the posterior marginal likelihoods of two models M_1 and M_0 :

$$B_{\text{posterior}, M_1 M_0} = \frac{P_{\text{posterior}}[D|M_1]P[M_1]}{P_{\text{posterior}}[D|M_0]P[M_0]} \quad (5)$$

Table 1 shows how a particular value of the Bayes factor should be interpreted.

Table I. Interpretation of the Bayes factor $B_{M_1 M_0}$ in comparing model M_1 to model M_0 (Kass & Raftery 1995) where M_1 is the model with the largest marginal likelihood. The categories are arbitrary in the same sense as a significance level of $\alpha = 0.05$ is arbitrary in classical statistics, but, just like this value of α , these categories seem to furnish appropriate guidelines (Kass & Raftery 1995).

$2 \ln B_{M_1 M_0}$	$B_{M_1 M_0}$	Evidence against model M_0
0 to 2	1 to 3	Not worth more than a bare mention
2 to 6	3 to 20	Positive
6 to 10	20 to 150	Strong
> 10	>150	Very strong

Below we will present a procedure to obtain the posterior probability distribution of the parameter set Θ_i . Crucial in this procedure (as in all Bayesian approaches) is again Bayes' formula, which, when applied to the model parameters Θ_i , reads:

$$P[\Theta_i|D, M_i] = \frac{P[D|\Theta_i, M_i]P[\Theta_i|M_i]}{P[D|M_i]} \quad (6)$$

As before, in equation (6), $P[\Theta|D, M_i]$ is the posterior distribution conditional on the model, $P[\Theta_i|M_i]$ is the (possibly multidimensional) prior distribution of the parameters Θ_i given the model and $P[D|M_i]$ is the prior marginal likelihood of the data given the model. Because all probabilities are conditioned on the model, we drop M_i and the subscript i for notational convenience whenever general statements are made that do not explicitly depend on a particular model.

Prior probability distribution - Jeffreys' prior

Although the prior probability $P[\Theta]$ is ideal to incorporate any expert knowledge about model parameters, sometimes this knowledge is so vague that a non-informative prior is desired. A simple prior which contains little information is the uniform probability distribution between realistic limits, but the probability distribution is not invariant under transformation of the parameter (for example, if one logtransforms the parameter). A natural choice of non-informative prior that is invariant under transformation is the Jeffreys priors (Jeffreys 1961; Kass & Wasserman 1996, 1998), defined as the square root of the determinant of the Fisher information matrix,

$$P[\Theta] \propto \sqrt{|I(\Theta)|} \quad (7)$$

where

$$I_{ij}(\Theta) = E_{D|\Theta} \left[-\frac{\partial^2}{\partial \Theta_i \partial \Theta_j} \ln P[D|\Theta] \right] \quad (8)$$

We use (variations of) the Jeffreys prior in our approach, but we note that there are several other ways to specify non-informative priors (Novick & Hall 1965, Zellner 1971, Box & Tiao 1973, Akaike 1978, Bernardo 1979). If sufficient data is available, it does not matter much which prior is chosen. We remark that non-informative priors are often improper, *i.e.* they do not integrate to 1, but this is no problem as long as the posterior is proper.

Posterior probability distribution - Markov Chain Monte Carlo simulation

(MCMC)

There are various ways to obtain the posterior distribution (Gelman *et al.* 2003, Carlin & Louis 2000), but we find the Markov Chain Monte Carlo (MCMC) approach (Chen *et al.* 2000) very convenient and suitable for our problem (see Ter Braak & Etienne 2003, Etienne *et al.* 2004, Link *et al.* 2002 for different ecological applications). The idea of MCMC simulation is to let the parameters perform a random walk in parameter space according to a Markov chain, set up in such a way that its stationary distribution is the posterior distribution. A useful algorithm for setting up the Markov chain is the Metropolis-Hastings (MH) algorithm (Metropolis and Ulam 1949; Metropolis *et al.* 1953; Hastings 1970). The MH-algorithm reads (Gelman *et al.* 2003):

1. Choose a starting value Θ^0 for parameters Θ .

For $u = 1 \dots$ repeat steps 2-4:

2. Choose a candidate point Θ^* (proposal) from a jumping distribution $J_u[\Theta^*|\Theta^{u-1}]$.

3. Calculate the acceptance ratio

$$\begin{aligned}
 r &= \frac{P[\Theta^*|D] J_u[\Theta^{u-1}|\Theta^*]}{P[\Theta^{u-1}|D] J_u[\Theta^*|\Theta^{u-1}]} = \\
 &= \frac{P[D|\Theta^*]P[\Theta^*]}{P[D|\Theta^{u-1}]P[\Theta^{u-1}]} \frac{J_u[\Theta^{u-1}|\Theta^*]}{J_u[\Theta^*|\Theta^{u-1}]} .
 \end{aligned} \tag{9}$$

4. Take

$$\Theta^u = \begin{cases} \Theta^* & \text{with probability } \min(r, 1) \\ \Theta^{u-1} & \text{otherwise} \end{cases} . \tag{10}$$

In this way a list of Θ^u is generated and the Θ^u with $u > u_{\text{burn-in}}$ constitute (a sample of) the posterior distribution for Θ , $u_{\text{burn-in}}$ being the point where the process is considered to have converged to its stationary distribution; the period up to this point is called the burn-in period. The efficiency of the algorithm depends largely on the choice of the jumping distribution. All parameters can be sampled simultaneously from a joint jumping distribution after which the entire set of parameters is either accepted or rejected, but they can also be sampled and updated (accepted/rejected) one by one by alternate sampling where the remaining parameters are fixed. Of course, any intermediate combination is also possible. Efficiency guides our choice.

The starting parameter value(s) Θ^0 can in principle be chosen arbitrarily, as the stationary distribution should be independent of Θ^0 . However, for rapid convergence of the Markov chain (*i.e.* a low $u_{\text{burn-in}}$), one should choose likely value(s) of Θ^0 , that are obtained, for example, by classical likelihood maximization.

Literature cited

- Aitkin, M. (1991). Posterior Bayes factors. *Journal of the Royal Statistical Society B* 53: 111-142.
- Akaike, H. (1978). A new look at the Bayes procedure. *Biometrika* 65: 53-59.
- Berger, J.O. (1985). *Statistical decision theory: foundations, concepts and methods*. Berlin, Germany: Springer.
- Berger, J.O. & L.R. Pericchi (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association* 91: 109-122.
- Bernardo, J.M. (1979). Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society B* 41:113-147.
- Box, G.E.P. & G.C. Tiao (1973). *Bayesian inference in statistical analysis*. Reading, MA: Addison-Wesley.

- Carlin, B.P. & T.A. Louis (2000). *Bayes and empirical Bayesian methods for data analysis*. London, U.K.: Chapman & Hall/CRC.
- Chen, M.-H., Q.-M. Shao & J.G. Ibrahim (2000). *Monte Carlo methods in Bayesian computation*. New York, NY: Springer.
- De Santis, F. & F. Spezzaferri (1997). Alternative Bayes factors for model selection. *Canadian Journal of Statistics* 25: 503-515.
- Etienne, R.S., C.J.F. ter Braak & C.C. Vos (2004). Application of stochastic patch occupancy models to real metapopulations. Pages 105-132 *in Ecology, Genetics, and Evolution of Metapopulations* (I. Hanski & O.E. Gaggiotti, eds.). London, U.K.: Elsevier Academic Press.
- Gelman, A., J.B. Carlin, H.S. Stern & D.B. Rubin (2003). *Bayesian data analysis*. London, U.K.: Chapman & Hall/CRC.
- Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57: 97-109.
- Jeffreys, H. 1961. *Theory of probability*. Oxford University Press, Oxford, U.K.
- Kass, R.E. & A.E. Raftery (1995). Bayes factors. *Journal of the American Statistical Association* 90: 773-795.
- Kass, R.E. & L. Wasserman (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association* 91: 1343-1370.
- Kass, R.E. & L. Wasserman (1998). The selection of prior distributions by formal rules (vol 91, pg 1343, 1996). *Journal of the American Statistical Association* 93: 412.
- Laud, P.W. & J.G. Ibrahim (1995). Predictive model selection. *Journal of the Royal Statistical Society B* 57: 247-262.
- Link, W.A., E. Cam, J.D. Nichols & E.G. Cooch (2002). Of BUGS and birds: Markov Chain Monte Carlo for hierarchical modeling wildlife research. *Journal of Wildlife Management* 66: 277-291.
- MacArthur, R.H. (1957). On the relative abundance of bird species. *Proceedings of the National Academy of Science, USA* 43: 293-295.
- MacArthur, R.H. (1960). On the relative abundance of species. *American Naturalist* 94: 25-36.
- Metropolis, N. & S. Ulam (1949). The Monte-Carlo method. *Journal of the American Statistical Association* 44:335-341.
- Metropolis, N., A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller & E. Teller (1953). Equation of state calculations

- by fast computing machines. *Journal of Chemical Physics* 21: 1087-1092.
- Novick, M.R. & W.J. Hall (1965). A Bayesian indifference procedure. *Journal of the American Statistical Association* 60: 1104-1117.
- Ter Braak, C.J.F. & R.S. Etienne (2003). Improved Bayesian analysis of metapopulation data with an application to a tree frog metapopulation. *Ecology* 84: 231-241.
- Tokeshi, M (1990). Niche apportionment or random assortment - species abundance patterns revisited. *Journal of Animal Ecology* 59: 1129-1146.
- Upadhyay, S.K. & M. Peshwani (2003). Choice between Weibull and lognormal models: a simulation based Bayesian study. *Communications in Statistics: Theory and Methods* 32: 381-405.
- Vlachos, P.K. & A.E. Gelfand (2003). On the calibration of Bayesian model choice criteria. *Journal of Statistical Planning and Inference* 111: 223-234.
- Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology* 44: 92-107.
- Zellner, A. (1971). An introduction to Bayesian inference in econometrics. New York, NY: John Wiley & Sons, Inc.