Available online at www.sciencedirect.com
ScienceDirect
Environmental Modelling & Software 23 (2008) 973–985
Environmental Modelling & Software
www.elsevier.com/locate/envsoft

Applying Bayesian Model Averaging to mechanistic models: An example and comparison of methods

J.M. Gibbons [a,*], G.M. Cox [a], A.T.A. Wood [b], J. Craigon [a], S.J. Ramsden [a], D. Tarsitano [a], N.M.J. Crout [a]

[a] Division of Agricultural and Environmental Sciences, School of Biosciences, University of Nottingham, Sutton Bonington LE12 5RD, UK
[b] Division of Statistics, School of Mathematical Sciences, University of Nottingham, University Park, Nottingham NG7 2RD, UK

Received 23 May 2007; received in revised form 29 November 2007; accepted 29 November 2007
Available online 18 January 2008

Abstract

Model averaging is a group of methods for combining predictions from several models which have the benefit of considering model uncertainty in addition to parameter uncertainty. The aim of this paper is to introduce these methods in the context of mechanistic model development. In model averaging predictions are combined, by weighting with factors related to model performance, resulting in ensemble predictions. Bayesian Model Averaging (BMA) is model averaging in a Bayesian framework where the model weights are Posterior model probabilities (PMPs). We describe three approximation methods (AIC, BIC and Laplace) for calculating PMPs and to compare with a full Bayesian approach implemented using a Markov Chain Monte Carlo (MCMC) method (Metropolis–Hastings). We also describe a simplified BMA approach which is readily implemented, as it only requires the maximum likelihood parameter estimates and Laplace approximation of the marginal likelihoods. We illustrate the application of BMA using a mechanistic model for predicting the plant uptake of radiocaesium from contaminated soils (the 'Absalom Model'). Ten models were selected for averaging, these comprised the full Absalom model and nine reduced models each derived from the full model. To assess performance model predictions and ensemble predictions were compared using an independent data set. The PMPs estimated using the MCMC approach and the Laplace approximation were similar and strongly weighted the models with fewer parameters. The AIC- and BIC-based estimates of the PMPs were correlated but differed considerably from the Laplace and MCMC-based PMP methods. For our example the simplified BMA approach was performed as well as the full approach. Individual predictions differed among models and the prediction ensembles resulting from all the approaches captured this uncertainty. We conclude that BMA is a valuable approach, relevant to mechanistic model development, and suggest a framework for incorporating BMA into model development.

Keywords: Model averaging; Bayesian Model Averaging; Model selection; Mechanistic models; Markov Chain Monte Carlo; Metropolis–Hastings algorithm; Posterior model probability

1. Introduction

It is routine to present model predictions with an estimate of prediction uncertainty. Typically, these estimates focus on the effect of uncertainty in the parameters of a single model, whose structure is taken as given and not subjected to uncertainty. In principle this might be reasonable as mechanistic, or process based, models represent scientific knowledge of a system. However, while this knowledge constrains model structure, understanding is rarely complete and consequently there are often several plausible candidate model structures which result in uncertainty in the model structure itself. This uncertainty can manifest itself both as uncertainty in how to model part of the system or in whether it is even necessary to include a component at all.

One rigorous approach to choice in model structure is to use model selection criteria (e.g. Myung, 2000; Burnham and Anderson, 2002; Cox et al., 2006) to select a single 'best'

* Corresponding author. Present address: School of the Environment and Natural Resources, College of Natural Sciences, Bangor University, Thoday Building, Deiniol Road, Bangor, Gwynedd LL57 2UW, UK. Tel.: +44 1248 382461.
E-mail address: j.gibbons@bangor.ac.uk (J.M. Gibbons).

1364-8152/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.envsoft.2007.11.008

model. However, this approach neglects uncertainty in the choice of models; this uncertainty may be important especially if there are several models that differ in predictions but have similar criterion scores. An alternative approach is to combine predictions from several models by a model averaging or weighting procedure. In this case, the final predictions are a weighted ensemble of the set of model predictions that takes account of uncertainty in model choice which is an area in which there is a paucity of methods available (Jakeman et al., 2006).

1.1. An overview of model averaging

We define model averaging as a method of combining results from several models into a single set of predictions. The simplest approach is an unweighted average across all models. However, there is an obvious advantage to weighting predictions by a measure of model performance. In this case, predictions from models that perform well (but possibly differ in predictions) would be weighted higher than predictions from poorly performing models. For ease of implementation, an attractive way of achieving this is to weight model predictions by model likelihood, normalized so the total model weights sum to unity. (Model likelihood is the probability of observing the data given the model and is maximised at the maximum likelihood estimates of the model parameter values.) For example, the GLUE (Generalized Likelihood Uncertainty Estimation) approach proposed by Beven and Binley (1992) weights model predictions by model maximum likelihood. However, model maximum likelihood is essentially a measure of goodness of fit. For a given data set it can always be increased by increasing the complexity of a model resulting in a model that fits existing data well but predicts new data poorly, so called overfitting or over-parameterisation. Jakeman et al. (2006) suggest that over-parameterisation is 'endemic in environmental modelling'; therefore, an alternative weighting procedure is preferable. Model selection criteria have been proposed as model weights (e.g. Burnham and Anderson, 2002). There also exist theoretically justifiable weighting criteria, which can be approximated by model selection criteria. A further approach, with some useful benefits, is to undertake model averaging within a Bayesian framework which we describe after first outlining the Bayesian approach to model inference.

1.2. A Bayesian approach to model inference

Bayesian model inference estimates the *posterior density*, a probability distribution of the parameter values given the data. This is calculated using the *model likelihood* and a *prior density* (see Appendix A for full details). This prior density represents previous knowledge about the parameter values before observing the current data. Controversy can arise because of the difficulties specifying prior belief mathematically and the resulting possibility that different conclusions can be reached from the same set of data. However, this effect rigorously encapsulates the common observations that (i) more evidence is often demanded to support surprising results and (ii) some scientists require more convincing than others. To avoid subjective difficulties uninformative priors can be used, allowing the application of Bayes' theorem, while ''letting the data speak for themselves'', uninfluenced by prior belief. More formally, with uninformative priors and large samples, the mode of the posterior density approximates the maximum likelihood parameter values. The choice between informative or uninformative priors may make little difference provided a sufficiently large sample of data is available so the posterior is dominated by the data. In the Bayesian framework every unknown is a random quantity, hence, model predictions also have posterior densities and this can be derived by evaluating the model, for a set of input data, at each point in the joint posterior density. Therefore, uncertainty of prediction is explicitly and automatically quantified given the model and data under consideration. New predictions can then be made by evaluating the likelihood of new input data over the posterior density. Except in very simple cases direct computation of the posterior is not being possible because of intractable integrals. This difficulty can be overcome by directly sampling from the posterior using Gibbs sampling or more generally a Markov Chain Monte Carlo (MCMC) approach as described in Appendix A.

1.3. A Bayesian approach to model averaging

In a model averaging framework, in addition to uncertainty in model parameter values (which can be regarded as within-model uncertainty), there is also uncertainty in model choice (uncertainty about which model is the best performing). Extension of Bayesian inference to models provides a natural treatment of this uncertainty (see Hoeting et al., 1999 for a full description). A clear advantage of Bayesian Model Averaging (BMA) is that no common structure is required for the set of models under consideration. They can be derived from entirely different principles and operate quite independently. The only requirement is that they predict the same quantities.

BMA is achieved by estimating *posterior model probabilities* (PMPs) for each model in addition to the posterior density of the parameters. PMPs are non-negative scalar values that sum to unity and are the relative probability of the model being true, given the data. The ratio of PMP values for any two models gives the relative support of those models and is known as a *Bayes factor*. Model averaged or *ensemble predictions* can be estimated by using the PMPs to weight predictions from individual models. Estimation of PMPs requires specification of prior model probabilities. These priors encapsulate the prior belief that the model is the true model. With no prior knowledge models can be given equal prior weight. Burnham and Anderson (2004) suggested that, following the parsimony principle, simpler models should be given higher prior weights. Alternatively, domain specialists might find models that are more detailed, more credible, and give them higher weight. See Link and Barker (2006) for an illustration of the consequences of different model priors.

While the parameter posterior density is estimated using the model likelihood, estimation of PMPs requires estimation of