Methodology for Bayesian Model Averaging: An Update

Jennifer A. Hoeting*
Colorado State University

Abstract

The standard practice of selecting a single model from some class of models and then making inferences based on this model ignores model uncertainty. Ignoring model uncertainty can impair predictive performance and lead to overstatement of the strength of evidence via p-values that are too small. Bayesian model averaging provides a coherent approach for accounting for model uncertainty. A variety of methods for implementing Bayesian model averaging have been developed. A brief overview of Bayesian model averaging is provided and recently developed methodology to perform Bayesian model averaging in specific model classes is described. Literature references as well as software descriptions and relevant webpage addresses are provided.

1 Introduction

Chronic wasting disease of the deer family has recently garnered daily headlines in newspapers in the state of Colorado. The disease is a member of a group of infectious diseases known as prion diseases which affect animals and humans. "Mad-cow" disease is one of the most widely known prion diseases and scientists are hopeful that the study of chronic wasting disease will increase understanding of this and other prion diseases.

Scientists would like to determine the prevalence of the disease in Colorado and are considering predictors of disease prevalence such as population density, habitat and other potentially related covariates. One concern about the study of prevalence is that the currently available data are generally from hunter returns where hunters bring in samples of harvested deer for testing. While this allows scientists to map disease prevalence around the state, the data are clearly subject to biases related to this non-random sampling.

Classical statistical theory relies on the notion of repeatability of an experiment. This is not always possible with ecological studies such as chronic wasting disease. Indeed, scientists are often pressed for early answers to questions, such as the prevalence of chronic wasting disease, without time or funding to perform experimental studies. In many situations, experimental manipulation of relevant factors is impossible.

For many of these problems, the typical data analysis approach is to select a set of predictors or risk factors and make inferences using this single model. A serious shortcoming this approach is the dependence of the inferences on the set of predictors selected for inclusion in the model. For example, it is quite possible that two different subsets of predictors of chronic wasting disease prevalence will fit the data well and provide sensible inferences. Without experimental modification of specific factors, it is impossible to specify the correct model. Similar arguments may be reasonable for certain experimental studies. In either case, choosing one model and basing inference on this single model ignores the uncertainty in the model selection.

*Postal address: Department of Statistics, Colorado State University, Fort Collins, CO 80526, USA; email address: jah@stat.colostate.edu, homepage: www.stat.colostate.edu/∼jah

Bayesian model averaging allows for the incorporation of model uncertainty into inference. The basic idea of Bayesian model averaging is to make inferences based on a weighted average over model space. This approach accounts for model uncertainty in both predictions and parameter estimates. The resulting estimates of uncertainty incorporate model uncertainty and thus may be better reflect the true uncertainty in the estimates. This paper provides a brief overview of Bayesian model averaging and some of the recently developed methodology to implement Bayesian model averaging for specific model classes.

2 Bayesian Model Averaging

Let $\mathcal{M} = (M_1, \ldots, M_K)$ be the set of models under consideration. A model may be defined by a variety of attributes such as the subset of explanatory variables in the model or the form of the error variance. If Δ is the quantity of interest, such as a future observable or a model parameter, then the posterior distribution of Δ given data Z is:

$$p\left(\Delta \,|\boldsymbol{Z}\right) = \sum_{k=1}^{K} p\left(\Delta \,|\boldsymbol{Z}, M_k\right) p\left(M_k \,|\boldsymbol{Z}\right), \tag{1}$$

This is an average of the posterior predictive distribution for Δ under each of the models considered, weighted by the corresponding posterior model probability. The posterior probability for model M_k is given by

$$p(M_k|\boldsymbol{Z}) = \frac{p(\boldsymbol{Z}|M_k)p(M_k)}{\sum_{l=1}^{K} p(\boldsymbol{Z}|M_l)p(M_l)}, \tag{2}$$

where

$$p\left(\boldsymbol{Z} \,|M_k\right) = \int \cdots \int p\left(\boldsymbol{Z} \,|\boldsymbol{\theta}_k, M_k\right) p\left(\boldsymbol{\theta}_k \,|M_k\right) d\boldsymbol{\theta}_k \tag{3}$$

is the integrated likelihood of model M_k, $\boldsymbol{\theta}_k$ is the vector of parameters of model M_k, $p\left(\boldsymbol{\theta}_k \,|M_k\right)$ is the prior density of the parameters under model M_k, $p\left(\boldsymbol{Z} \,|\boldsymbol{\theta}_k, M_k\right)$ is the likelihood, and $p(M_k)$ is the prior probability that M_k is the true model. All probabilities are implicitly conditional on $\mathcal{M}$, the set of all models being considered.

Parameter estimates and other quantities of interest are provided via straightforward application of the principles described above. For example, the Bayesian model averaging (BMA) estimate of a parameter θ is

$$\hat{\theta}_{\text{BMA}} = \sum_{k=1}^{K} \hat{\theta}_k p\left(M_k \,|\boldsymbol{Z}\right)$$

where $\hat{\theta}_k$ denotes the posterior mean for model k. Variances of these estimates and other quantities are also available (e.g., Hoeting *et al.* 1999 and Viallefont *et al.* 2001).

There are many challenges involved in the implementation of Bayesian model averaging (BMA), including the computation of (1) for a very large number of models, the evaluation of the integrals implicit in (3) which do not typically exist in closed form, and the specification of the prior model probabilities $p(M_k)$.

A number of researchers have considered the problem of managing the summation in equation (1) for a large number of models, some of which are described below for specific areas of application. A popular approach is to explore the space of models stochastically via a Markov chain Monte Carlo approach (e.g., George and McCulloch, 1997 and Raftery, Madigan, and Hoeting, 1997). Clyde (1999a; 1999b) shows that many of these approaches are a special case of reversible jump MCMC algorithms (Green, 1995). Clyde (1999a) and Dellaportas, Forster and Ntzoufras (2002) review model search and averaging algorithms. Godsill (2001) proposes a composite representation for model uncertainty problems which includes many of these Markov chain Monte Carlo approaches as special cases. This and other recent advances, such as implementation of perfect sampling, show promise in this area.

Hoeting, Madigan, Raftery, Volinsky (1999) discuss the historical development of BMA, provide additional description of the challenges of carrying out BMA, and describe some solutions to these problems for a variety of model classes. The next section provides a brief overview of some of the methodology that was described in that paper and describes more recent work in this area.

3 Bayesian Model Averaging for Specific Model Classes

The development of methodology to carry out model averaging is a rapidly growing area. In the last two years more than 60 papers related to model averaging have appeared in the peer reviewed literature. As this body of literature continues to increase, it is a daunting challenge to present a review of methodology to implement Bayesian model averaging. Below is a annotated bibliography of some of the available methodology for specific classes of models. Software to implement these procedures, when available, is described. All of the programs listed below are free of charge and internet addresses are provided. Many of these programs are available at the "Bayesian model averaging home page" (`www.research.att.com/∼volinsky/bma.html`).

3.1 Linear Regression Models

The selection of subsets of predictor variables is a fundamental issue in linear regression modeling. Several approaches have been developed to average over all possible sets of predictors. Raftery, Madigan, and Hoeting (1997) provide a closed form expression for the likelihood in this context (equation 3) and an extensive discussion of hyperparameter choice in the situation where little prior information is available. Fernández, Ley, and Steele (1997; 2001a) and Liang, Troung and Wong (2001) offer alternative prior structures aiming at a more automatic choice of hyperparameters.

Raftery *et al.* (1997) also develop approaches to overcoming the challenges of averaging over a large number of possible models in the context of linear regression models. They develop a Markov chain Monte Carlo (MCMC) approach and a more heuristic procedure. Other researchers have also considered this problem. See Clyde (1999a) for a review.

Hoeting, Raftery and Madigan (1996; 2002a) extend the BMA framework of averaging over subsets of predictors to account for uncertainty in the selection of transformations and identification of outliers in regression models.

Software: The programs listed below are written in S-Plus© and are available at `www.research.att.com/∼volinsky/bma.html`.

1. *bicreg* uses a Bayesian information criterion (BIC) approximation to perform Bayesian model selection and accounting for model uncertainty in linear regression models [Written by A.E. Raftery].

2. *BMA* implements BMA for linear regression models via the Markov chain Monte Carlo model model composition (MC^3) algorithm [Written by J.A. Hoeting].

3.2 Nonparametric Regression Models

An active area of research has been the development of methodology for nonparameteric regression models, including model averaging for nonparametric models. Shively, Kohn, and Wood (1999) develop methodology for additive nonparametric Gaussian and binary regression models. Posterior means of the regression functions are averaged across all possible models. The authors demonstrate that this approach can produce better estimates of the regression functions as compared to estimates that assume that all functions should be included in the model. A related approach is described in Lamon and Clyde (2000). Gustafson (2000) offers an alternative spline-based approach and Liang *et al.* (2001) develop an automatric prior set-up for curve ftting with regression splines.

Holmes and Mallick (2001) develop methodology for model averaged predictive distributions for a piecewise linear model constructed using basis functions. Their approach accounts for uncertainty in both the number and the locations of the splines.

3.3 Spatial Models

Stochastic models for spatial prediction are becoming increasingly important as the cost of collecting spatially referenced data continues to decrease. When the goal is prediction and the data are collected from spatially irregular locations, a typical approach is to assume that the response of interest is a realization of a Gaussian random field which can be described by a function of distance, direction between two sample locations, and, possibly, regression covariates. In this context, geostatistical methods such as universal kriging (Cressie, 1993, Chapter 3) are often used to produce predictions for a Gaussian random field.

An attractive alternative to standard kriging methods is a Bayesian approach for spatial prediction (see Gaudard *et al.*, 1999, Handcock and Stein, 1993, and the references therein). The Bayesian approach has the advantage that inferences and predictions fully include parameter uncertainty. De Oliverira *et al.* (1997) show that ignoring the uncertainty associated with specification of the parameters in a normalizing transformation for Gaussian random fields can lead to over-confident prediction intervals. Sun (1998) also demonstrates the perils of ignoring parameter uncertainty in a comparison of cokriging with a Bayesian alternative.

In addition to parameter uncertainty, there are other components to model uncertainty in spatial prediction. It is common practice in both standard and Bayesian kriging to first select the form of the model (including selecting the regression covariates, the form of the regression function and the autocorrelation function) and then make inferences assuming that the selected model is the "true" model. This ignores uncertainty due to model selection.

Thompson (2001) and Hoeting, Thompson and Davis (2002b) adopt a Bayesian approach for spatial prediction and develop methodology for accounting for model uncertainty. Spatial prediction via Bayesian model averaging accounts for the uncertainty associated with model selection including the selection of explanatory variables, the form of the regression function, and the form of the autocorrelation function. While Bayesian methods are often appealing, their solutions can be computationally complex, particularly in the context of spatial models. The proposed methods reduce computational complexity while preserving many of the advantages of Bayesian methodology. Thompson (2001) shows that averaging over all possible subsets of explanatory variables in spatial models can lead to more accurate predictive coverage and smaller predictive errors as compared to the standard practice of basing predictions on a single model.

3.4 Other Model Classes and Areas of Application

Generalized linear models: Uncertainty about the choice of the link function, the variance function as well as the explanatory variables all contribute to model uncertainty in the construction of generalized linear models (McCullagh and Nelder, 1989). Raftery (1996) proposes methodology to approximate the quantities necessary to compute the posterior model probabilities (equation 2) and carry out BMA in this context.

Clyde (2000) develops objective prior distributions for BMA in generalized linear models, allowing for direct comparison of Bayesian model selection with standard methods such as Akaike Information Criterion (AIC) and Bayes Information Criterion (BIC). This approach is used to investigate the impact of model uncertainty on inferences about the effect of particulate matter on mortality of the elderly.

Viallefont, Raftery and Richardson (2001) apply BMA for logistic regression models in the context of case-control studies and the determination of significant risk factors. Via a simulation study they show that p-values computed after traditional variable selection can greatly overstate the strength of conclusions while BMA appears to alleviate this problem. They demonstrate the approach on a case-control study investigating risk factors for cervical cancer and include careful description of the inferences available via BMA in this context.

Landrum and Becker (2001) develop a multiple imputation strategy for incomplete longitudinal data. They use model averaging for pooling predictions across different statistical models.

Software: The programs listed below are written in S-Plus© and are available at `www.research.att.com/~volinsky/bma.html`.

1. *bic.glm* performs BMA for generalized linear models using the Leaps and Bounds algorithm and the BIC approximation. [Written by C.T. Volinsky].
2. *bic.logit* performs Bayesian model selection and accounting for model uncertainty using the BIC approximation for logistic regression models [Written by A.E. Raftery].
3. *glib* carries out Bayesian estimation, model comparison and accounting for model uncertainty in generalized linear models, allowing user-specified prior distributions [Written by A.E. Raftery].

Graphical Models: Graphical models can be used to summarize a set of conditional independence statements. Madigan and York (1995) and York *et al.* (1995) describe methodology to implement BMA for missing data and latent variable problems in the context of Bayesian graphical models for discrete data. Andersson *et al.* (1997) show that a family of graphical models can be partitioned into Markov-equivalence classes, where each class is associated with a unique statistical model. Accounting for these relationships can reduce inefficiencies in computational procedures for statistical inference. Madigan *et al.* (1996) apply these principles to develop two stochastic algorithms to perform BMA for the analysis of discrete variable data.

Classification Models: Denison, Adams, Homes, and Hand (2002) compare several approaches to Bayesian classification modeling via partitioning and describe BMA methodology in this context. Model averaging and model uncertainty are also discussed in Chipman, George and McCulloch (1998) in the development of a Bayesian algorithm for classification and regression trees (CART).

Software: C++ software to perform Bayesian CART (Chipman *et al.*, 1998) is available at `gsbwww.uchicago.edu/fac/robert.mcculloch/research/code`.

Capture-recapture Models: A series of papers on Bayesian analysis of capture-recapture data have recently appeared. Brooks, Catchpole and Morgan (2000) develop a Bayesian approach to estimating parameters associated with animal survival. These same authors propose a similar approach for analyzing ringing data in a more recent paper (Brooks *et al.*, 2002).

King and Brooks (2002) discuss model discrimination as well as model averaging for multiple strata capture-recapture data.

Software: Brooks *et al.* (2000; 2002) include computer code to carry out the analyses. The code is written in winBUGS, software for the Bayesian analysis of statistical models using Markov chain Monte Carlo methods, which is available at no charge at `www.mrc-bsu.cam.ac.uk/bugs`.

Survival Analysis: Volinsky *et al.* (1997) use BIC and other approximations to enable BMA and the selection of explanatory variables for Cox proportional hazards models (Cox, 1972). To manage the summation in equation (1), they develop a leaps and bounds algorithm related to the algorithm developed by Furnival and Wilson (1974). The authors demonstrate the methodology using data from the Cardiovascular Health Study to investigate the risk factors for stroke.

Software: The S-plus© program *bic.surv* performs BMA for proportional hazard models using the BIC approximation. It is available at `www.research.att.com/~volinsky/bma.html`.

Economic Applications: Early contributors on model averaging of forecasts includes Palm and Zellner (1992) and Min and Zellner (1993). More recently, BMA is applied to economic data in the analyses of consumer demand systems (Chua *et al.*, 2001), cross-country growth regressions (Fernández *et al.*, 2001b), and option pricing (Bunnin *et al.*, 2002).

4 Discussion

There are many other issues in the application of and philosophy behind Bayesian model averaging that are not considered here such as specification of the prior model probabilities ($p(M_k)$ in equation 2), averaging over more than one class of models, interpretation of the results, and model checking. Some of these issues receive further treatment in Hoeting *et al.* (1999) and the comments therein by D. Draper, E.I. George and M. Clyde. However, many of these are open areas for further research.

Bayesian model averaging offers a coherent approach to accounting for model uncertainty. BMA has also been demonstrated to improve predictive performance (e.g., Hoeting *et al.* 1999) and to avoid the problem of overstatement of the strength of evidence, a problem when p-values are computed after traditional variable selection (Volinsky *et al.*, 1997; Viallefont *et al.*, 2001). It should be emphasized, however, that BMA should not be used as an excuse for poor science or to argue for the use of observational studies over experimental approaches. BMA is useful after careful scientific analysis of the problem at hand. Indeed, BMA offers one more tool in the toolbox of applied statisticians for improved data analysis and interpretation.

Acknowledgments

This work was partially funded by STAR Research Assistance Agreement CR-829095 awarded to Colorado State University by the U.S. Environmental Protection Agency (EPA). It has not been formally reviewed by EPA; the views expressed here are solely those of authors. EPA does not endorse any products or commercial services mentioned here.

References

Andersson, S. A., Madigan, D., and Perlman, M. D. (1997). A characterization of Markov equivalence classes for acyclic digraphs. *Annals of Statistics*, 25(2):505–541.

Brooks, S., Catchpole, E., and Morgan, B. (2000). Bayesian animal survival estimation. *Statistical Science*, 15:357–376.

Brooks, S., Catchpole, E., Morgan, B., and Harris, M. (2002). Bayesian methods for analysing ringing data. *Journal of Applied Statistics*, 29(1–4):187–206.

Bunnin, F., Guo, Y., and Ren, Y. (2002). Option pricing under model and parameter uncertainty using predictive densities. *Statistics and Computing*, 12(1):37–44.

Chipman, H., George, E., and McCulloch, R. (1998). Bayesian CART estimation (with discussion and rejoinder). *Journal of the American Statistical Association*, 93:935–960.

Chua, C., Griffiths, W., and O'Donnell, C. (2001). Bayesian model averaging in consumer demand systems with inequality constraints. *Canadian Journal of Agricultural Economics*, 49(3):269–291.

Clyde, M. (1999a). Bayesian model averaging and model search strategies (with discussion). In Bernardo, J., Berger, J., Dawid, A., and Smith, A., editors, *Bayesian Statistics 6*, pages 157–185. Oxford University Press.

Clyde, M. (1999b). Discussion of Bayesian model averaging: A tutorial with discussion by J. A. Hoeting, D. Madigan and A. E. Raftery and C. T. Volinsky. *Statistical Science*, 14:401–404.

Clyde, M. (2000). Model uncertainty and health effect studies for particulate matter. *Environmetrics*, 11(6):745–763.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society Ser. B*, 34:187–220.

Cressie, N. A. C. (1993). *Statistics for Spatial Data*. Wiley: New York.

De Oliveira, V., Kedem, B., and Short, D. (1997). Bayesian prediction of transformed Gaussian random fields. *Journal of the American Statistical Association*, 92(440):1422–1433.

Dellaportas, P., Forster, J., and Ntzoufras, I. (2002). Bayesian model and variable selection using MCMC. *Statistics and Computing*, 12(1):27–36.

Denison, D., Adams, N., Holmes, C., and Hand, D. (2002). Bayesian partition modelling. *Computational Statistics & Data Analysis*, 4(4):475–485.

Fernández, C., Ley, E., and Steel, M. F. (1997). Statistical modeling of fishing activities in the north atlantic. Technical report, Department of Econometrics, Tilburg Univeriesty, the Netherlands.

Fernández, C., Ley, E., and Steel, M. F. (2001a). Benchmark priors for Bayesian model averaging. *Journal of Econometrics*, 100:381–427.

Fernández, C., Ley, E., and Steel, M. F. (2001b). Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics*, 16(5):563–576.

Furnival, G. M. and Wilson, R. W. (1974). Regression by leaps and bounds. *Technometrics*, 16:499–511.

Gaudard, M., Karson, M., Linder, E., and Sinha, D. (1999). Bayesian spatial prediction. *Environmental and Ecological Statistics*, 6:147–171.

George, E. I. and Foster, D. P. (1997). Calibration and empirical bayes variable selection. Technical report, Dept of MSIS, University of Texas at Austin.

Godsill, S. (2001). On the relationship between Markov chain Monte Carlo for model uncertainty. *Journal of Computational and Graphical Statistics*, 10(2):230–248.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732.

Gustafson, P. (2000). Bayesian regression modeling with interactions and smooth effects. *Journal* *of the American Statistical Association*, 95:745–763.

Handcock, M. S. and Stein, M. L. (1993). A Bayesian analysis of kriging. *Technometrics*, 35:403–410.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: A tutorial with discussion. *Statistical Science*, 14:382–417. Printing errors corrected in version available at http://www.stat.washington.edu/www/research/online/hoeting1999.pdf.

Hoeting, J. A., Raftery, A. E., and Madigan, D. (1996). A method for simultaneous variable selection and outlier identification in linear regression. *Journal of Computational Statistics*, 22:251–271.

Hoeting, J. A., Raftery, A. E., and Madigan, D. (2002a). Bayesian variable and transformation selection in linear regression. *Statistical Science*, to appear.

Hoeting, J. A., Thompson, S. E., and Davis, R. A. (2002b). Spatial prediction via Bayesian model averaging. Technical report, Department of Statistics, Colorado State Univesity.

Holmes, C. and Mallick, B. (2001). Bayesian regression with multivariate linear splines. *Journal of the Royal Statistical Society Ser. B*, 63(1):3–17.

King, R. and Brooks, S. (2002). Bayesian model discrimination for multiple strata capture-recapture data. *Biometrika*, to appear.

Lamon, E. and Clyde, M. (2000). Accounting for model uncertainty in prediction of chlorophyll a in lake okeechobee. *Journal of Agricultural Biological and Environmental Statistics*, 5(3):297–322.

Landrum, M. and Becker, M. (2001). A multiple imputation strategy for incomplete longitudinal data. *Statistics in Medicine*, 20(17-18):2741–2760.

Liang, F. M., Troung, Y., and Wong, W. H. (2001). Automatic Bayesian model averaging for linear regression and applications in Bayesian curve fitting. *Statistica Sinica*, 11(4):1005–1029.

Madigan, D., Raftery, A. E., Volinsky, C. T., and Hoeting, J. A. (1996). Aaai workshop on integrating multiple learned models. pages 77–83.

Madigan, D. and York, J. (1995). Bayesian graphical models for discrete data. *International Statistical Review*, 63:215–232.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman & Hall, London, 2nd edition.

Min, C. and Zellner, A. (1993). Bayesian and non-Bayesian methods for combining models and forecasts with applications to forecasting international growth rates. *Journal of Econometrics*, 56(1-2):89–118.

Palm, F. and Zellner, A. (1992). To combine or not to combine - issues of combining forecasts. *Journal of Forecasting*, 11(8):687–701.

Raftery, A. E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika*, 83:251–266.

Raftery, A. E., Madigan, D., and Hoeting, J. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92:179–191.

Shively, T., Kohn, R., and Wood, S. (1999). Variable selection and function estimation in additive nonparametric regression using a data-based prior (with discussion). *Journal of the American Statistical Association*, 94(447):777–806.

Sun, W. (1998). Comparison of a cokriging method with a Bayesian alternative. *Environmetrics*, 9:445–457.

Thompson, S. E. (2001). *Bayesian Model Averaging and Spatial Prediction*. PhD thesis, Colorado State University.

Viallefont, V., Raftery, A., and Richardson, S. (2001). Variable selection and Bayesian model averaing in case-control studies. *Statistics in Medicine*, 20(21):3215–3230.

Volinsky, C. T., Madigan, D., Raftery, A. E., and Kronmal, R. A. (1997). Bayesian Model Averaging in proportional hazard models: Assessing the risk of a stroke. *Applied Statistics*, 46(3).

York, J., Madigan, D., Heuch, I., and Lie, R. T. (1995). Estimating a proportion of birth defects by double sampling: a Bayesian approach incorporating covariates and model uncertainty. *Applied Statistics*, 44:227–242.