Bayesian Model Averaging

Mark F.J. Steel
University of Warwick, UK

Summary

Bayesian Model averaging is a natural response to model uncertainty. It has become an important practical tool for dealing with model uncertainty, in particular in empirical settings with large numbers of potential models and relatively limited numbers of observations. Most of this paper focuses on the problem of variable selection in normal linear regression models, but also briefly considers a more general setting. The article surveys a number of prior structures proposed in the literature, and presents in some detail the most commonly used prior setup. Model selection consistency and practical implementation are also briefly discussed.

Keywords: Bayesian methods; Consistency; Covariate selection; Generalized linear model; Model uncertainty; Normal linear regression model; Prior specification; Robustness

1 Model uncertainty

The issue of model uncertainty <**stat07398**> is a very important one in statistical practice. It affects the vast majority of applied modelling and its consequences need to be taken into account whenever we are interested in quantities that are not model-specific (such as predictions or effects of regressors). Generally, one important and potentially dangerous consequence of neglecting model uncertainty is that we assign more precision to our inference than is warranted by the data, and this leads to overly confident decisions and predictions. In addition, our inference can be severely biased. See [6] and [10] for extensive discussions of model uncertainty.

The standard Bayesian response to dealing with uncertainty is to average. When dealing with parameter uncertainty, this involves averaging over parameter values with the posterior distribution in order to conduct inference. Analogously, model uncertainty is also resolved through averaging, but this time averaging over models with the (discrete) posterior model distribution. The latter procedure is usually called Bayesian Model Averaging (BMA) and was already described in [22] and later in e.g. [31]. Averaging immediately leads to predictive distributions, usually a key quantity of interest. In particular, assume we are interested in predicting the unobserved quantity y_f on the basis of the observations y. Let us denote the sampling model for y_f and y jointly by $p(y_f|y,\theta_j,M_j)p(y|\theta_j,M_j)$, where M_j is the model selected from a set of K possible models, and $\theta_j \in \Theta_j$ groups the (unknown) parameters of M_j. In a Bayesian framework, any uncertainty is reflected by a probability distribution (or measure) so we assign a (typically continuous) prior $p(\theta_j|M_j)$ for the parameters and a discrete prior $P(M_j)$ defined on the model space. We then compute the predictive distribution as

$$p(y_f|y) = \sum_{j=1}^{K} \left[\int_{\Theta_j} p(y_f|y,\theta_j,M_j)p(\theta_j|y,M_j)\mathrm{d}\theta_j \right] P(M_j|y), \tag{1}$$

where the quantity in square brackets is the predictive distribution given M_j obtained using the posterior of θ_j given M_j, which is computed as

$$p(\theta_j|y, M_j) = \frac{p(y|\theta_j, M_j)p(\theta_j|M_j)}{\int_{\Theta_j} p(y|\theta_j, M_j)p(\theta_j|M_j)\mathrm{d}\theta_j} \equiv \frac{p(y|\theta_j, M_j)p(\theta_j|M_j)}{p(y|M_j)}, \tag{2}$$

with the second equality defining $p(y|M_j)$, which is used in computing the posterior probability assigned to M_j as

$$P(M_j|y) = \frac{p(y|M_j)P(M_j)}{\sum_{i=1}^{K} p(y|M_i)P(M_i)} \equiv \frac{p(y|M_j)P(M_j)}{p(y)}. \tag{3}$$

Clearly, the predictive in (1) indeed involves averaging at two levels: over (continuous) parameter values, given each possible model, and discrete averaging over all possible models. The denominator (or integrating constant) in (2) is the so-called marginal likelihood <**stat05901**> of model j, $p(y|M_j)$ and is a key quantity for model comparison. In particular, the Bayes factor <**stat00224.pub2**> between two models is the ratio of their marginal likelihoods and the posterior odds are directly obtained as the product of the Bayes factor and the prior odds. The denominator in (3), $p(y)$, is defined as a sum and the challenge in its calculation often lies in the very large number of possible models, i.e. K.

Bayesian model averaging as described above is thus the formal probabilistic way of obtaining predictive inference or, more generally, inference on quantities of interest that are not model-specific.

The importance of Bayesian model averaging as a solution to model uncertainty is illustrated by the fact that Google Scholar returns 12,800 articles in a search for "Bayesian model averaging", about half of which date from 2011 and later (data from October 30, 2015).

1.1 Covariate selection in the normal linear regression model

Most of the relevant literature assumes the simple case of the normal linear regression model. This model is often used in empirical work and is fortunately also quite tractable. Model uncertainty then typically relates to the choice of which covariates should be included in the model, *i.e.* under model j the n observation in y are generated from the normal sampling model

$$y|\theta_j, M_j \sim N(\alpha + X_j\beta_j, \sigma^2), \tag{4}$$

where the $n \times p_j$ matrix X_j groups p_j of the possible p covariates and $\beta_j \in \Re^{p_j}$ are its corresponding regression coefficients. Furthermore, the models considered all contain an intercept $\alpha \in \Re$ and the scale $\sigma > 0$ has a common interpretation across all models. The covariates are standardized by subtracting their means, which makes them orthogonal to the intercept and renders the interpretation of the intercept common to all models. The model space then consists of all possible subsets of the covariates and thus contains $K = 2^p$ models in total.

This problem has received quite a lot of attention from the statistical community. Traditional methods (such as stepwise regression <**stat03252**>, LASSO <**stat07543.pub2**> and methods based on

information criteria <**stat04463**>) aim at uncovering a single "best" model. In other words, they are model selection methods (**see stat00228**), as opposed to model averaging methods. As it is often unlikely that reality can be adequately captured by a simple normal linear regression model, it is risky to rely on a single model for inference, forecasts and (policy) conclusions. An averaging method usually gives a better approximation to reality and improves our estimate of the uncertainty associated with our conclusions.

2 Bayesian Model averaging

2.1 Priors on model parameters

The priors for the model parameters, i.e. $p(\theta_j|M_j)$ needs to be proper on model-specific parameters. In our normal linear model in (4) the model-specific parameters are the β_js.

Here we focus on the prior structure proposed by [12], which is in line with the majority of the current literature. [12] adopt Jeffreys-style non-informative priors for α and σ^2. For the regression coefficients β_j, they propose the g-prior specification of [35] for the covariance structure. The prior density is then

$$p(\alpha, \beta_j, \sigma \mid M_j) \propto \sigma^{-1} f_N^{p_j}(\beta_j|0, \sigma^2 g(X_j'X_j)^{-1}), \tag{5}$$

where $f_N^q(\cdot|m, V)$ denotes the density function of a q-dimensional Normal distribution with mean m and covariance matrix V. The regression coefficients not appearing in M_j are exactly zero, corresponding to a prior point mass at zero. The amount of prior information requested from the user is limited to a single scalar $g > 0$. The g-prior is a convenient prior with nice properties, such as invariance to reparameterization under affine transformations. In addition, the marginal likelihood for each model (and thus the Bayes factor between each pair of models) can be calculated in closed form.

There are a number of suggestions in the literature for the choice of values for g. The unit information prior of [20] corresponds to the amount of information contained in one observation. This gives us $g = n$ and log Bayes factors that behave asymptotically like the BIC (see [12]). The RIC prior of [13] is based on the Risk inflation criterion which leads to $g = p^2$ using a minimax <**stat01687**> perspective. Finally, the benchmark prior of [12] corresponds to $g = \max(n, p^2)$.

When faced with a variety of possible prior choices for g, a natural Bayesian response is to formulate a hyperprior on g. This was already implicit in the Zellner-Siow prior of [36] who use a Cauchy <**stat05830**> prior on the regression coefficients, corresponding to an inverse gamma <**stat01030**> prior on g. This idea was investigated further in [26], where hyperpriors on g are shown to alleviate certain paradoxes that appear with fixed choices for g. Hierarchical priors on g were also discussed and compared in [24], whereas [33] proposed independent mixtures of g-priors on blocks of regression coefficients.

In the context of applications to genetic data, it is common to use priors where $X_j'X_j$ in the covariance in (5) is replaced by an identity matrix (since the covariates are typically comparable in scale). Similarly, [31] propose a (proper natural-conjugate <**stat00214**>) prior with a diagonal covariance structure (except for categorical predictors where a g-prior structure is used).

A different approach was proposed by [16, 17], who use a prior on the regression coefficients which does not include point masses at zero. In particular, they advocate a normal prior with mean zero on the entire p-dimensional vector of regression coefficients given the model M_j which assigns a small prior variance to the coefficients of the variables excluded in M_j and a larger variance to the remaining coefficients. In addition, their overall prior is proper and does not assume a common intercept.

Another class of priors are intrinsic priors, which underlie the intrinsic Bayes factors of [2], and are used in this context in [29].

2.2 Priors over models

The prior $P(M_j)$ on model space is typically constructed through the probability of inclusion of the covariates. If the latter is the same for each covariate, say w, and we assume covariate inclusions are prior independent, then

$$P(M_j) = w^{p_j}(1-w)^{p-p_j}. \tag{6}$$

This implies that prior odds will favour larger models if $w > 0.5$ and the opposite if $w < 0.5$. For $w = 0.5$ all model have equal prior probability $1/K$. Defining model size as the number of included covariates, it is easy to elicit w through the prior mean model size, which is pw. As the choice of w can have a large effect on the results, various authors [5, 8, 23] have suggested to put a Beta(a, b) <**stat04850**> hyperprior on w. This leads to

$$P(M_j) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(a+p_j)\Gamma(b+p-p_j)}{\Gamma(a+b+p)}, \tag{7}$$

which results in much less informative priors in terms of model size. [23] compare both approaches and suggest choosing $a = 1$ and $b = (p-m)/m$, where m is the chosen prior mean model size. They conclude that the choice of m is much less critical for the results in the hierarchical case (7) than for fixed w in (6).

[14] raises the issue of "dilution", which occurs when posterior probabilities are spread among many similar models, and suggest that prior model probabilities could have a built-in adjustment to compensate for dilution by down-weighting prior probabilities on sets of similar models. These so-called "dilution priors" were implemented by [11] through defining priors that are uniform on theories (i.e. neighbourhoods of similar models) rather than on individual models.

2.3 Empirical Bayes Priors

As explained in *e.g.* [3] and [12], it is virtually impossible to find a single default choice for g and w that performs well in all cases. Therefore, as stated above, hyperpriors on g and w have been proposed. An alternative approach to this issue is to use the data to estimate g and w, leading to empirical Bayes <**stat00257.pub2**> (EB) procedures. A local EB method was proposed by [19] which uses a different g for each model estimated by maximizing the marginal likelihood. [15] develop a global EB approach, which assumes one common g and w for all models and borrows strength from all models by estimating g and w by maximizing the marginal likelihood, averaged over all models.

2.4 Consistency

One of the desiderata in [1] for objective model selection priors is model selection consistency (introduced by [12]), which implies that if data have been generated by M_j, then the posterior probability of M_j should converge to unity with sample size. [12] present general conditions for the case with non-random g and show that consistency holds for e.g. the unit information and benchmark priors (but not for the RIC prior). When we consider hierarchical priors on g, model selection consistency is achieved by the Zellner-Siow prior in [36] but not by local and global EB priors nor by the hyper-g prior in [26], who therefore introduce a consistent modification (the hyper-g/n prior). [29] consider model selection consistency when the number of potential regressors p grows with sample size. Consistency is found to depend not only on the priors for the model parameters, but also on the priors in model space. They conclude that if $p = O^b(n)$, the unit information prior, the Zellner-Siow prior and the intrinsic prior lead to consistency for $0 \leq b < 1/2$ under the uniform prior over model space, while consistency holds for $0 \leq b \leq 1$ if we use a Beta(1,1) hyperprior on w in (7).

2.5 Practical implementation

The main challenge is typically the very large number of models, which makes complete enumeration impossible. Note that the marginal likelihood for each visited model can be computed analytically under the prior structure in (5), so all we need is to run a Markov chain Monte Carlo (MCMC) <**stat03616**> algorithm in model space, such as the MC^3 methods of [27], which is a random-walk Metropolis-Hastings sampler. If we use the prior by [16, 17] a Gibbs sampler (**see stat00211, stat07189**) can be used for inference: usually this is referred to as Stochastic search variable selection or SSVS <**stat07829**>. For mixtures of g-priors, we can either add a step for g (see [24]) or use Laplace approximations as in [26].

These methods are quite easy to implement. However, more sophisticated methods can be required for settings where covariates are highly correlated or for very large model spaces, such as adaptive MCMC (e.g. [30], [21] and [18]), evolutionary Monte Carlo [4] and Bayesian adaptive sampling proposed in [9], which samples models without replacement.

2.6 Covariate selection in generalized linear models

Generalized linear models (GLMs) describe a more general class of models (see [28], **stat06576, stat06942**) that covers the normal linear model but also regression models where the response variable is non-normal, such as binomial <**stat04852**> (e.g. logistic or logit regression <**stat06902**> models, probit models <**stat07555**>), Poisson <**stat01112**>, multinomial <**stat04877**> (e.g. ordered response models, proportional odds models <**stat06805**>) or gamma <**stat00989**> distributed. [32] consider the interpretation of the g-prior in linear models as the conditional posterior of the regression coefficients given a locally uniform prior and an imaginary sample of zeros with design matrix X_j and a scaled error variance, and extend this to the GLM context. Asymptotically, this leads to a prior which is very similar to the

standard g-prior. This idea was already used in the conjugate prior proposed by [7], although they do not treat the intercept separately. For priors on g, [32] consider a Zellner-Siow prior and a hyper-g/n prior. Both choices lead to consistent model selection, as shown in [34].

The priors on the model parameters designed for GLMs in [25] are based on a different type of "centering" (induced by a block diagonal observed information matrix at the MLE of the coefficients) and employ a wider class of (potentially truncated) hyper-priors for g. As a consequence, these priors are slightly more complicated, and, more importantly, are data-dependent (depending on y, not just the design matrix).

Acknowledgements: I am grateful to the Editors for giving me the opportunity to write this article. In July 2013, Eduardo Ley, who has made important contributions in this area, tragically passed away. He was a very dear friend and a much valued coauthor and this piece is dedicated to his memory.

Related Articles: stats07398; stat00243.pub2; stat05772; stat00219; stat00207; stat00224.pub2

References

[1] BAYARRI, M.-J., BERGER, J., FORTE, A., AND GARCÍA-DONATO, G. Criteria for Bayesian model choice with application to variable selection. *Annals of Statistics 40* (2012), 1550–77.

[2] BERGER, J., AND PERICCHI, L. The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association 91* (1995), 109–22.

[3] BERGER, J., AND PERICCHI, L. Objective Bayesian methods for model selection: Introduction and comparison. In *Model Selection* (Beachwood, OH: IMS, 2001), P. Lahiri, Ed., Institute of Mathematical Statistics Lecture Notes - Monograph Series 38, pp. 135–207.

[4] BOTTOLO, L., AND RICHARDSON, S. Evolutionary stochastic search for Bayesian model exploration. *Bayesian Analysis 5* (2010), 583–618.

[5] BROWN, P., VANNUCCI, M., AND FEARN, T. Bayesian wavelength selection in multicomponent analysis. *Journal of Chemometrics 12* (1998), 173–82.

[6] CHATFIELD, C. Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society, Series A 158* (1995), 419–66 (with discussion).

[7] CHEN, M., AND IBRAHIM, J. Conjugate priors for generalized linear models. *Statistica Sinica 13* (2003), 461–76.

[8] CLYDE, M., AND GEORGE, E. I. Model uncertainty. *Statistical Science 19* (2004), 81–94.

[9] CLYDE, M., GHOSH, J., AND LITTMAN, M. Bayesian adaptive sampling for variable selection and model averaging. *Journal of Computational and Graphical Statistics 20* (2011), 80–101.

[10] DRAPER, D. Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society, Series B 57* (1995), 45–97 (with discussion).

[11] DURLAUF, S., KOURTELLOS, A., AND TAN, C. Are any growth theories robust? *Economic Journal 118* (2008), 329–46.

[12] FERNÁNDEZ, C., LEY, E., AND STEEL, M. Benchmark priors for Bayesian model averaging. *Journal of Econometrics 100* (2001), 381–427.

[13] FOSTER, D., AND GEORGE, E. The risk inflation criterion for multiple regression. *Annals of Statistics 22* (1994), 1947–75.

[14] GEORGE, E. Discussion of "Bayesian model averaging and model search strategies" by M. Clyde. In *Bayesian Statistics 6* (Oxford: Oxford University Press, 1999), J. Bernardo, J. Berger, A. Dawid, and A. Smith, Eds., pp. 175–7.

[15] GEORGE, E., AND FOSTER, D. Calibration and empirical bayes variable selection. *Biometrika 87* (2000), 731–47.

[16] GEORGE, E., AND MCCULLOCH, R. Variable selection via Gibbs sampling. *Journal of the American Statistical Association 88* (1993), 881–89.

[17] GEORGE, E., AND MCCULLOCH, R. Approaches for Bayesian variable selection. *Statistica Sinica 7* (1997), 339–73.

[18] GRIFFIN, J., ŁATUSZYŃSKI, K., AND STEEL, M. Individual adaptation: An adaptive MCMC scheme for variable selection problems. CRiSM Working Paper 15-01, University of Warwick, 2015.

[19] HANSEN, M. H., AND YU, B. Model selection and the principle of minimum description length. *Journal of the American Statistical Association 96* (2001), 746–74.

[20] KASS, R., AND WASSERMAN, L. A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association 90* (1995), 928–34.

[21] LAMNISOS, D. S., GRIFFIN, J. E., AND STEEL, M. F. J. Adaptive Monte Carlo for Bayesian variable selection in regression models. *Journal of Computational and Graphical Statistics 22* (2013), 729–748.

[22] LEAMER, E. *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. New York: Wiley, 1978.

[23] LEY, E., AND STEEL, M. On the effect of prior assumptions in Bayesian model averaging with applications to growth regression. *Journal of Applied Econometrics 24* (2009), 651–74.

[24] LEY, E., AND STEEL, M. Mixtures of g-priors for Bayesian model averaging with economic applications. *Journal of Econometrics 171* (2012), 251–266.

[25] LI, Y., AND CLYDE, M. Mixtures of g-priors in generalized linear models. Working paper, Duke University, 2015.

[26] LIANG, F., PAULO, R., MOLINA, G., CLYDE, M., AND BERGER, J. Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association 103* (2008), 410–423.

[27] MADIGAN, D., AND YORK, J. Bayesian graphical models for discrete data. *International Statistical Review 63* (1995), 215–32.

[28] MCCULLAGH, P., AND NELDER, J. A. *Generalized Linear Models*. Chapman and Hall, 1989.

[29] MORENO, E., GIRÓN, J., AND CASELLA, G. Posterior model consistency in variable selection as the model dimension grows. *Statistical Science 30* (2015), 228–41.

[30] NOTT, D., AND KOHN, R. Adaptive sampling for Bayesian variable selection. *Biometrika 92* (2005), 747–63.

[31] RAFTERY, A., MADIGAN, D., AND HOETING, J. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association 92* (1997), 179–91.

[32] SABANÉS BOVÉ, D., AND HELD, L. Hyper-g priors for generalized linear models. *Bayesian Analysis 6* (2011), 387–410.

[33] SOM, A., HANS, C. M., AND MACEACHERN, S. N. Bayesian modeling with mixtures of block g priors. Technical report, Dept. of Statistics, Ohio State University, 2015.

[34] WU, H., FERREIRA, M., AND GOMPPER, M. Consistency of hyper-g-prior-based Bayesian variable selection for generalized linear models. *Brazilian Journal of Probability and Statistics 29* (2015), forthcoming.

[35] ZELLNER, A. On assessing prior distributions and bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti* (Amsterdam: North-Holland, 1986), P. K. Goel and A. Zellner, Eds., pp. 233–43.

[36] ZELLNER, A., AND SIOW, A. Posterior odds ratios for selected regression hypotheses (with discussion). In *Bayesian Statistics* (Valencia: University Press, 1980), J. Bernardo, M. DeGroot, D. Lindley, and A. Smith, Eds., pp. 585–603.