ECOM168: Bayesian model averaging

Daniele Bianchi

School of Economics and Finance,
Queen Mary University of London

Outline

1. Introduction
2. Implementing Bayesian model averaging
3. Properties of Bayesian Model Averaging
4. Bayesian Model Averaging for regression models

Introduction

Introduction

A very general problem in statistics is where several models represent *plausible* descriptions for the target y.

Standard statistical analysis practice is based on the assumption that we can select a single *best* model and then proceed as if the model generated the data.

This approach inherently ignores a key component of uncertainty, that is *model uncertainty*, leading to possibly over-confident inferences.

N.B., obviously this perspective is not always applicable; sometimes the decision maker must select a specific model and/or eliminate redundant covariates.

Introduction

As highlighted by Madigan et al. (1996), a natural approach towards model uncertainty is to include all models $\mathcal{M}_k$ under consideration, thus by-passing the model choice step.

If model uncertainty is to be formally reflected into the inferential process, we will *average* model-specific estimates over the set of possible models under consideration.

Bayesian Model Averaging (BMA) provides a coherent mechanism for accounting for model uncertainty.

$\hookrightarrow$ posterior model probabilities are used as *weights* for the average.

Introduction

If Δ is the quantity of interest, posterior inference obtained via Bayesian model averaging (BMA) entails integrating over the model space $\mathcal{M}_\gamma \in \mathcal{M}$

$$p\left(\Delta|y\right) = \sum_\gamma p\left(\Delta|\mathcal{M}_\gamma, y\right) p\left(\mathcal{M}_\gamma|y\right), \tag{1}$$

where given a prior $p\left(\mathcal{M}_\gamma\right)$, the posterior model probability is defined by

$$p\left(\mathcal{M}_\gamma|y\right) = \frac{p\left(y|\mathcal{M}_\gamma\right) p\left(\mathcal{M}_\gamma\right)}{\sum_\gamma p\left(y|\mathcal{M}_\gamma\right) p\left(\mathcal{M}_\gamma\right)},$$

with $p\left(y|\mathcal{M}_\gamma\right) = \int p\left(y|\mathcal{M}_\gamma, \theta_\gamma\right) p\left(\theta_\gamma|\mathcal{M}_\gamma\right) d\theta_\gamma$ the marginal likelihood of model $\mathcal{M}_\gamma$, and θ_γ the corresponding vector of model parameters.

Introduction

For example, optimal prediction of future values $\widetilde{y}$ under a squared-error loss is defined as

$$E\left(\widetilde{y}|y\right) \equiv \sum_{\gamma} E\left(\widetilde{y}|\mathcal{M}_{\gamma}, y\right) p\left(\mathcal{M}_{\gamma}|y\right)$$

Madigan and Raftery (1994) note that averaging over *all* models in this fashion provides better predictive ability, compared to using a single model $\mathcal{M}_{\gamma}$.

Considerable empirical evidence now exist to support theoretical claims (see Bianchi and McAlinn, 2022 and the references therein).

Introduction

While BMA is an intuitively attractive solution to account for model uncertainty, it is not yet part of common statistical inference tools.

This is part due to the fact that BMA presents several difficulties:

$\hookrightarrow$ The number of terms in Eq.(1) could be very large.

$\hookrightarrow$ The integrals implicit in Eq.(1) could be computationally expensive.

$\hookrightarrow$ Specification of $p\left(\mathcal{M}_\gamma\right)$ is not obvious.

$\hookrightarrow$ Choosing the class of models over which to average becomes a fundamental modeling task.

We are going to discuss some possible solution within the context of linear regression models.

Implementing Bayesian model averaging

Managing the summation

The size of interesting model classes potentially renders the exhaustive summation of Eq.(1) impractical.

Two (not so) different approaches can be used to reduce the aggregate model space:

- $\hookrightarrow$ pre-screening models supported by the data.
- $\hookrightarrow$ stochastic search of most suitable models.

Occam's window method

Occam's window method of Madigan and Raftery (1994):

↪ average over a subset of models that are supported by the data.

Exclude models which predict the data far less well than the model with best predictions. Thus models not belonging to $\mathcal{M}'$,

$$\mathcal{M}' = \left\{ \mathcal{M}_\gamma : \frac{\max_l \{p(\mathcal{M}_l|y)\}}{p(\mathcal{M}_\gamma|y)} \leq C \right\}, \tag{2}$$

with C arbitrarily small, should be excluded from the model set.

Occam's window method

Exclude complex models which receive less support from the data than simpler counterparts.

$$\mathcal{M}^{''} = \left\{ \mathcal{M}_\gamma : \exists \, \mathcal{M}_l \in \mathcal{M}', \mathcal{M}_l \subset \mathcal{M}_k, \frac{p\left(\mathcal{M}_l | y\right)}{p\left(\mathcal{M}_k | y\right)} > 1 \right\}, \tag{3}$$

As a result, the set $\mathcal{M}$ in Eq.(1) can be replaced with $\mathcal{M}'$ or the even more restrictive $\mathcal{M}^{''}$.

This greatly reduces the number of models in the summation in Eq.(1);

↪ now all that is required is to define a strategy to identify the models in $\mathcal{M}', \mathcal{M}^{''}$.

Markov chain monte carlo

A second approach to pre-select a subset of possible models is to use Markov Chain Monte Carlo (MCMC) methods to explore the model space $\mathcal{M}$.

Specifically, one can construct a Markov chain $\mathcal{M}_k, \; k = 1, 2, \dots$ with state space $\mathcal{M}$. Consider a neighbourhood $ne\,(\mathcal{M}_k)$ for each $\mathcal{M}_k \in \mathcal{M}$.

↪ e.g., in regression models the neighbourhood could be the set of models with one regressor more or fewer than $\mathcal{M}_k$, plus model $\mathcal{M}_k$ itself.

Two key ingredients:

↪ the description of the possible moves from one model $\mathcal{M}_k$ to another $\mathcal{M}_l$.

↪ the determination of the acceptance probability for $\mathcal{M}_l$.

Markov chain monte carlo

The description of the possible moves takes the form of a proposal distribution

$$\mathcal{M}_l \sim q\left(\mathcal{M}_k \to \mathcal{M}_l\right)$$

Then the sampled model indicator is accepted with a probability

$$\alpha\left(\mathcal{M}_l | \mathcal{M}_k\right) = \min\left\{1, \frac{p\left(\mathcal{M}_l | y\right)}{p\left(\mathcal{M}_k | y\right)}\right\}$$

otherwise we keep model $\mathcal{M}_k$. We can then average only models with the highest posterior probability, i.e., $\mathcal{M}'$ in Eq.(2).

The SSVS of George and McCulloch (1993) could be used in the same spirit by averaging the forecasts for each $\gamma \sim p\left(\gamma | \ldots\right)$.

A word of caution

N.B., A major caveat of pre-selecting the models which are mostly supported by the data for the summation in Eq.(1), is that for p large exploring the model space could be a prohibitive task in and of itself.

$\hookrightarrow$ The posterior model probability $p\left(\mathcal{M}_\gamma | y\right)$ could also be highly computationally expensive to approximate for complex models.

Approximation of the marginal likelihood

Bayesian model averaging is based on the assumption that posterior model probabilities can be used as weights to integrate out model uncertainty.

For a given model prior $p\left(\mathcal{M}_\gamma\right)$ a posterior model probability is a direct function of the model marginal likelihood $p\left(y|\mathcal{M}_\gamma\right)$.

Unfortunately, the marginal likelihood is available in closed form only for simple models, such as Gaussian linear regression models.

Approximation of the marginal likelihood

Often we need to resort to monte carlo integration to approximate $p\left(y|\mathcal{M}_\gamma\right)$. Several methods have been proposed in the statistics literature.

For instance, Newton and Raftery (1994) showed that the marginal likelihood under model $\mathcal{M}_\gamma$ can be approximated as

$$\widehat{p}\left(y|\mathcal{M}_\gamma\right) = \left\{\frac{1}{G}\sum_{g=1}^{G}\left(\frac{1}{p\left(y|\theta_\gamma^{(g)},\mathcal{M}_\gamma\right)}\right)\right\}^{-1}, \tag{4}$$

where $\theta_\gamma^{(g)} \sim p\left(\theta_\gamma|\mathcal{M}_\gamma\right)$ are the draws from the posterior distribution obtained using a given sampling method.

Eq.(4) represents the harmonic mean of the likelihood values evaluated at the posterior draws.

Approximation of the marginal likelihood

Although Eq.(4) represents a consistent estimate of $p\left(\theta_\gamma|\mathcal{M}_\gamma\right)$, it is not numerically stable.

$\hookrightarrow$ the inverse of the likelihood does not have a finite variance.

Gelfand and Dey (1994) proposed a more stable approximation

$$\widehat{p}\left(y|\mathcal{M}_\gamma\right) = \left\{ \frac{1}{G} \sum_{g=1}^{G} \left(\frac{p\left(\theta_\gamma^{(g)}\right)}{p\left(y|\theta_\gamma^{(g)}, \mathcal{M}_\gamma\right) p\left(\theta_\gamma^{(g)}|\mathcal{M}_\gamma\right)} \right) \right\}^{-1}, \tag{5}$$

where $p\left(\theta_\gamma^{(g)}|\mathcal{M}_\gamma\right)$ is the prior for model $\mathcal{M}_\gamma$ and $p\left(\theta_\gamma^{(g)}\right)$ is a density with "thinner" tails than the product of the likelihood and the prior.

Gelfand and Dey (1994) showed that for G large, Eq.(5) converges to $p\left(y|\mathcal{M}_\gamma\right)$.

Approximation of the marginal likelihood

Nonetheless, Eq.(5) implies that a suitable tuning function $p\left(\theta_\gamma^{(g)}\right)$ should be found.

$\hookrightarrow$ Some of the obvious choices, such as Gaussian and Student-t, may not be “thin” enough.

Chib (1995) proposed one of the most used method to approximate the marginal likelihood based on Gibbs sampling.

The estimator takes the form

$$\log \widehat{p}\left(y|\mathcal{M}_\gamma\right) = \log p\left(y|\theta_\gamma^*\right) + \log p\left(\theta_\gamma^*\right) - \log \widehat{p}\left(\theta_\gamma^*|y\right),$$

The densities $p\left(y|\theta_\gamma^*\right)$ and $p\left(\theta_\gamma^*\right)$ are often ”easy” to evaluate, while $\widehat{p}\left(\theta_\gamma^*|y\right)$ gets more complicated.

Approximation of the marginal likelihood

Suppose we can decompose θ_γ into $(\theta_{1\gamma}, \theta_{2\gamma})$, such that $p\left(\theta_{1\gamma} | y, \theta_{2\gamma}\right)$ and $p\left(\theta_{2\gamma} | y, \theta_{1\gamma}\right)$ are available in closed form.

A Gibbs sampler draws from $p\left(\theta_\gamma | y\right)$ and hence marginally from $p\left(\theta_{2\gamma} | y\right)$, such that

$$p\left(\theta_{1\gamma}^* | y\right) = \int p\left(\theta_{1\gamma}^* | y, \theta_{2\gamma}\right) p\left(\theta_{2\gamma} | y\right) d\theta_{2\gamma} \approx \frac{1}{G} \sum_{g=1}^{G} p\left(\theta_{1\gamma}^* | y, \theta_{2\gamma}^{(g)}\right),$$

A classical example is a simple linear regression where $\theta_\gamma = \left(\beta_\gamma, \sigma^2\right)$.

Approximation of the marginal likelihood

If the marginal likelihood is too complex to obtain numerically, a viable solution is to rely on information criteria.

For example, the Bayesian Information Criterion (BIC) is a first-order approximation to the marginal likelihood,

$$\begin{aligned}\log p\left(y|\mathcal{M}_\gamma\right) = {} & \log p\left(y|\mathcal{M}_\gamma, \widetilde{\beta}, \widetilde{\sigma}^2\right) + \log p\left(\widetilde{\beta}, \widetilde{\sigma}^2|\mathcal{M}_\gamma\right) + \frac{p}{2}\log\left(2\pi\right) - \frac{p}{2}\log n \\ & - \frac{1}{2}\log\left|J_n\left(\widetilde{\beta}, \widetilde{\sigma}^2|\mathcal{M}_\gamma\right)\right| + \mathcal{O}\left(n^{-1}\right),\end{aligned}$$

where $J_n\left(\widetilde{\beta}, \widetilde{\sigma}^2|\mathcal{M}_\gamma\right)$ is the expected Fisher information evaluated at the mode $\left(\widetilde{\beta}, \widetilde{\sigma}^2\right)$ for a given model $\mathcal{M}_\gamma$.

Approximation of the marginal likelihood

Removing any terms of order $\mathcal{O}\left(1\right)$ or less we obtain:

$$\log p\left(y|\mathcal{M}_{\gamma}\right) = \log p\left(y|\mathcal{M}_{\gamma}, \widehat{\beta}, \widehat{\sigma}^{2}\right) - \frac{p}{2}\log n + \mathcal{O}\left(1\right), \tag{6}$$

The approximation in Eq.(6) gives the basics for the BIC

$$BIC = -2\log p\left(y|\mathcal{M}_{\gamma}, \widehat{\beta}, \widehat{\sigma}^{2}\right) + p\log n,$$

with $\left(\widehat{\beta}, \widehat{\sigma}^{2}\right)$ the parameters point estimates from model $\mathcal{M}_{\gamma}$.

Specifying prior model probabilities

Before implementing any of the BMA strategies as per Eq.(1), prior model probabilities must be assigned for each model $\mathcal{M}_\gamma \in \mathcal{M}$.

When there is little prior information about the relative plausibility of the models considered, the assumption that all models are equally likely a priori is a reasonable "neutral" choice.

However, Spiegelhalter et al. (1993) showed how beneficial could be to incorporate informative priors $p\left(\mathcal{M}_\gamma\right)$.

Specifying prior model probabilities

When prior information about the importance of a variable, e.g., in a linear regression model, a prior probability on model $\mathcal{M}_j$ can be specified as

$$p\left(\mathcal{M}_j\right) = \prod_{j=1}^{p} \pi_j^{\gamma_{ij}} \left(1 - \pi_j\right)^{1-\gamma_{ij}},$$

where

- $\hookrightarrow$ $\pi_j \in [0,1]$ is the prior probability that $\beta_j \neq 0$ in the regression.
- $\hookrightarrow$ γ_{ij} is an indicator variable that takes value 1 is the regressor j is included in the model i, and zero otherwise.

Specifying prior model probabilities

Among the most popular default choices for π_j are:

- ↪ $\pi_j = 0.5$, which assigns equal probability to each model, i.e., $p\left(\mathcal{M}_j\right) = 1/2^p$.
- ↪ $\pi_j \sim U\left(0, 1\right)$, giving equal probability to each possible number of covariates.
- ↪ $\pi_j < 0.5$ for all js impose a penalty for large models.

In general, fixed values of π_j have been shown to perform poorly in controlling for the occurrence of spurious explanatory variables.

- ↪ Ley and Steel (2009) consider $\pi_j \sim \text{Beta}\left(1, b\right)$, i.e. a binomial-beta prior.

Properties of Bayesian Model Averaging

Consistency

One of the desiderata when implementing BMA is model selection consistency.

$\hookrightarrow$ the posterior probability $p\left(\mathcal{M}_\gamma | y\right)$ should converge to 1 if the data are indeed generated from $\mathcal{M}_\gamma$.

In principle, consistency can depend not only on the priors for the model parameters, but also on the priors in the model space.

Consistency

Mukhopadhyay et al. (2015) suggest that in situations where the true model is not one of the candidate models ($\mathcal{M}$-open setting), BMA could be sub-optimal.

Put it differently, BMA is "optimal" only in an $\mathcal{M}$-closed setting, i.e., when the *true* model $\mathcal{M}^T$ is in the set of models considered, i.e., $\mathcal{M}^T \in \mathcal{M}$.

To see this consider the true model as

$$y = x_1\beta_1 + x_2\beta_2 + \epsilon,$$

and you estimate

$\hookrightarrow$ $\mathcal{M}_1 : y = x_1\beta_1 + \epsilon$

$\hookrightarrow$ $\mathcal{M}_2 : y = x_2\beta_2 + \epsilon$

Consistency

With BMA your prediction will be

$$\hat{y} = \omega_1 \cdot x_1 \hat{\beta}_1 + \omega_2 \cdot x_2 \hat{\beta}_2,$$

Under standard regularity conditions:

$\hookrightarrow$ $\omega_i \to 1$ for the model "closest" to the truth (in a Kullback-Leibler sense).

$\hookrightarrow$ BMA is not consistent if $\mathcal{M}^T \notin \mathcal{M}$.

N.B., Mukhopadhyay et al. (2015) showed that using a g-prior for the model parameters leads to select a model "as close as possible" to the true $\mathcal{M}^T$.

Shrinkage

BMA has also some important properties in terms of shrinkage in high-dimensional regression problems.

This has been highlighted by Castillo et al. (2015); they show that BMA in linear regressions leads to optimal rate of contraction of the posterior to a sparse “true” data-generating model.

↪ provided the prior sufficiently penalizes model complexity.

Similarly, Rossell and Telesca (2017) show that BMA leads to a rapid shrinkage of spurious coefficients.

↪ this is linked to the use of information criteria as proxy for the marginal likelihood

Bayesian Model Averaging for regression models

BMA for linear regression models

Let $y \in \mathbb{R}^n$ be the vector of response variable and $X \in \mathbb{R}^{n \times p}$ the design matrix.

The regression model we build upon has the form;

$$y = X\beta + \varepsilon, \qquad \varepsilon \sim N\left(0, \sigma^2 \mathbf{I}_n\right), \tag{7}$$

where $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$ is an $n \times p$ matrix of covariates and $\beta = (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p$ is the p-dimensional vector of regression parameters.

We have seen in the previous lectures that a typical approach to data analysis is to carry out model selection and make inference on the selected model.

BMA for linear regression models

This ignores a major component of uncertainty, that is uncertainty about the model itself.

As a consequence, uncertainty about quantities of interest – such as posterior parameter estimates – can be underestimated.

A complete Bayesian solution to this problem within the context of linear regressions is to average over all possible combinations of predictors when making inference about $\theta = \left(\beta, \sigma^2\right)$.

Indeed, this approach provides optimal predictive ability (see Madigan and Raftery, 1994).

BMA for linear regression models

Where possible, conjugate prior distributions should be used. Recall that a default prior for Eq.(7) is the natural conjugate prior:

$$\beta|\sigma^2 \sim N\left(\beta_0, \sigma^2\Sigma_0\right), \qquad \sigma^2 \sim \mathsf{IG}\left(v_0, s_0\right)$$

which leads to a marginal likelihood of the form

$$p\left(y|\mathcal{M}_\gamma\right) \propto |\widetilde{X}_\gamma^\top \widetilde{X}_\gamma|^{-1/2}|\Sigma_0|^{-1/2}\left(s_0 + \frac{S_\gamma^2}{2}\right)^{-(v_0+n/2)} \tag{8}$$

where $\widetilde{X}_\gamma = \left[\begin{array}{c} X_\gamma \\ \Sigma_0^{-1} \end{array}\right] \in \mathbb{R}^{(n+p_\gamma)\times p_\gamma}$, $S_\gamma^2 = y^\top H_\gamma y$, with $H_\gamma = I - X_\gamma \widehat{\Sigma}_\gamma X_\gamma^\top$ and $\widehat{\Sigma}_\gamma = \left(X_\gamma^\top X_\gamma + \Sigma_0^{-1}\right)^{-1}$.

BMA for linear regression models

We have seen in Lecture 1 how the g-prior addresses the issue of setting a prior on different regression models that might be nested and have correlated variables.

Given the analytical form of the posterior and the marginal likelihood, when doing BMA the Zellner's g-prior is a common benchmark (see, Fernandez et al., 2001).

For the regression coefficients the prior takes the form

$$\beta_\gamma|\sigma^2 \sim N\left(0, g\sigma^2\left(X_\gamma^\top X_\gamma\right)^{-1}\right),$$

Such that

$$p\left(y|\mathcal{M}_\gamma\right) \propto (g+1)^{-p_\gamma/2}\left[s_\gamma + \frac{\widehat{\beta}_\gamma^\top X_\gamma^\top X_\gamma \widehat{\beta}_\gamma}{g+1}\right]^{-\frac{n-1}{2}},$$

represents the marginal likelihood.

BMA for linear regression models

N.B., The conjugate prior structure outlined above gives the flexibility to modify the hyper-parameters $\beta_0, \Sigma_0, v_0, s_0$.

The posterior model probability for model $\mathcal{M}_\gamma$ can then be calculated from Eq.(8) as

$$p\left(\mathcal{M}_\gamma|y\right) = \frac{p\left(y|\mathcal{M}_\gamma\right)p\left(M_\gamma\right)}{\sum_\gamma p\left(y|\mathcal{M}_\gamma\right)p\left(M_\gamma\right)} \quad \longrightarrow \quad \frac{p\left(y|\mathcal{M}_\gamma\right)}{\sum_\gamma p\left(y|\mathcal{M}_\gamma\right)} \qquad \text{for} \qquad p\left(M_\gamma\right) = 1/2^p,$$

Parameter estimates are then carried out as based on $p\left(\mathcal{M}_\gamma|y\right)$, for instance,

$$\widehat{\beta} = \sum_\gamma \widehat{\beta}_\gamma p\left(\mathcal{M}_\gamma|y\right)$$

where $\widehat{\beta}_\gamma$ is the posterior mean estimate of the regression coefficients for model $\mathcal{M}_\gamma$.

BMA for linear regression models

More generally, the posterior mean and variance of a given quantity of interest Δ are:

$$E\left(\Delta|y\right)=\sum_{\gamma}E\left(\Delta|y,\mathcal{M}_{\gamma}\right)p\left(\mathcal{M}_{\gamma}|y\right),$$

and

$$Var\left(\Delta|y\right)=\sum_{\gamma}\left[Var\left(\Delta|y,\mathcal{M}_{\gamma}\right)+E^{2}\left(\Delta|y,\mathcal{M}_{k}\right)\right]p\left(\mathcal{M}_{\gamma}|y\right)-E^{2}\left(\Delta|y\right),$$

where $Var\left(\Delta|y,\mathcal{M}_{\gamma}\right)$ and $E\left(\Delta|y,\mathcal{M}_{k}\right)$ are the posterior variance and mean of Δ under $\mathcal{M}_{\gamma}$.

A frequentist view of Bayesian Model Averaging

Raftery (1995) introduced an approximation of the marginal likelihood based on the BIC information criterion,

$$BIC = -2 \log p\left(y|\widehat{\beta}_\gamma, \widehat{\sigma}^2, \mathcal{M}_\gamma\right) + p_\gamma \log n,$$

with $\left(\widehat{\beta}_\gamma, \widehat{\sigma}^2\right)$ the parameters point estimates, and p_γ the number of parameters for model $\mathcal{M}_\gamma$. This allows posterior probabilities to be calculated with minimal effort.

For instance, in the context of linear regression models, the BIC for each model further simplifies as

$$BIC_\gamma = n \log\left(\frac{RSS_\gamma}{n}\right) + p_\gamma \log n + C,$$

where RSS_γ is the residual sum of squares under $\mathcal{M}_\gamma$ and C is a constant that does not depend on p_γ.

A frequentist view of Bayesian Model Averaging

If we assume a uniform prior for the occurrence of each model, i.e., $p\left(\mathcal{M}_\gamma\right) = 1/2^p$, the posterior model probability can thus be approximated by:

$$\widehat{p}\left(\mathcal{M}_\gamma | y\right) \approx \frac{\exp\left(-\frac{1}{2}\Delta BIC_\gamma\right)}{\sum_\gamma \exp\left(-\frac{1}{2}\Delta BIC_\gamma\right)}, \tag{9}$$

where $\Delta BIC_\gamma = BIC_\gamma - BIC_{min}$, with $BIC_{min} = \min_\gamma BIC_\gamma$.[1]

Similar to BMA, we can then obtain weighted average estimates of a quantity of interest from different models.

[1] N.B., one can replace ΔBIC_γ with BIC_γ in Eq.(9) (see Raftery, 1995).

A frequentist view of Bayesian Model Averaging

For e.g., if $\mathbb{I}\left(X_i \in \mathcal{M}_\gamma\right)$ is an indicator of the presence of the variable i in model $\mathcal{M}_\gamma$, we can calculate

$$\widehat{\gamma}_i = \sum_\gamma \widehat{p}\left(\mathcal{M}_\gamma | y\right) \mathbb{I}\left(X_i \in \mathcal{M}_\gamma\right),$$

as the posterior inclusion probability that the ith variable. Larger $\widehat{\gamma}_i$ indicates higher "importance" of the variable X_i in the model.

Similarly, the estimated posterior mean for β_γ is

$$\widehat{\beta} = \sum_\gamma \widehat{\beta}_\gamma \widehat{p}\left(\mathcal{M}_\gamma | y\right),$$

where $\widehat{\beta}_\gamma$ is the estimate of β in model $\mathcal{M}_\gamma$.

References

Bianchi, D. and McAlinn, K. (2022). Divide and conquer: Financial ratios and industry returns predictability. *Available at SSRN*.

Castillo, I., Schmidt-Hieber, J., and Van der Vaart, A. (2015). Bayesian linear regression with sparse priors. *Annals of Statistics*.

Chib, S. (1995). Marginal likelihood from the gibbs output. *Journal of the american statistical association*, 90(432):1313–1321.

Fernandez, C., Ley, E., and Steel, M. F. (2001). Benchmark priors for bayesian model averaging. *Journal of Econometrics*, 100(2):381–427.

Gelfand, A. E. and Dey, D. K. (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(3):501–514.

George, E. I. and McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

Ley, E. and Steel, M. F. (2009). On the effect of prior assumptions in bayesian model averaging with applications to growth regression. *Journal of applied econometrics*, 24(4):651–674.

Madigan, D. and Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using occam's window. *Journal of the American Statistical Association*, 89(428):1535–1546.

Madigan, D., Raftery, A. E., Volinsky, C., and Hoeting, J. (1996). Bayesian model averaging. In *Proceedings of the AAAI Workshop on Integrating Multiple Learned Models, Portland, OR*, pages 77–83.

Mukhopadhyay, M., Samanta, T., and Chakrabarti, A. (2015). On consistency and optimality of bayesian variable selection based on g-prior in normal linear regression models. *Annals of the Institute of Statistical Mathematics*, 67(5):963–997.

Newton, M. A. and Raftery, A. E. (1994). Approximate bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(1):3–26.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological methodology*, pages 111–163.

Rossell, D. and Telesca, D. (2017). Nonlocal priors for high-dimensional estimation. *Journal of the American Statistical Association*, 112(517):254–265.

Spiegelhalter, D. J., Dawid, A. P., Lauritzen, S. L., and Cowell, R. G. (1993). Bayesian analysis in expert systems. *Statistical science*, pages 219–247.