Bayesian Model Averaging and some applications

Mark Steel

Department of Statistics, University of Warwick

O'Bayes 2022, Santa Cruz, 6 September 2022

Introduction

- Bayesian Model Averaging (BMA) as a response to model uncertainty

Introduction

- Bayesian Model Averaging (BMA) as a response to model uncertainty
- Substantial field in statistics and many areas of application (economics, biology, ecology, sociology, meteorology, psychology and hydrology)

Introduction

- Bayesian Model Averaging (BMA) as a response to model uncertainty
- Substantial field in statistics and many areas of application (economics, biology, ecology, sociology, meteorology, psychology and hydrology)
- Methodology and key operational aspects

Introduction

- Bayesian Model Averaging (BMA) as a response to model uncertainty
- Substantial field in statistics and many areas of application (economics, biology, ecology, sociology, meteorology, psychology and hydrology)
- Methodology and key operational aspects
- Applications in economics (do institutions drive economic growth?) and political science (do proportional electoral rules generate higher turnout?)

Introduction

- Bayesian Model Averaging (BMA) as a response to model uncertainty
- Substantial field in statistics and many areas of application (economics, biology, ecology, sociology, meteorology, psychology and hydrology)
- Methodology and key operational aspects
- Applications in economics (do institutions drive economic growth?) and political science (do proportional electoral rules generate higher turnout?)
- Publicly available software

Introduction

- Bayesian Model Averaging (BMA) as a response to model uncertainty
- Substantial field in statistics and many areas of application (economics, biology, ecology, sociology, meteorology, psychology and hydrology)
- Methodology and key operational aspects
- Applications in economics (do institutions drive economic growth?) and political science (do proportional electoral rules generate higher turnout?)
- Publicly available software
- Some recommendations and open questions

Introduction

- Bayesian Model Averaging (BMA) as a response to model uncertainty
- Substantial field in statistics and many areas of application (economics, biology, ecology, sociology, meteorology, psychology and hydrology)
- Methodology and key operational aspects
- Applications in economics (do institutions drive economic growth?) and political science (do proportional electoral rules generate higher turnout?)
- Publicly available software
- Some recommendations and open questions
- Key references

Model uncertainty

Model uncertainty is an inherent part of modelling, especially in the social sciences. Ignore this at your peril!

Model uncertainty

Model uncertainty is an inherent part of modelling, especially in the social sciences. Ignore this at your peril!

Two main strategies for dealing with model uncertainty:

- Model selection: choose "best" model and then conduct inference conditionally upon the assumption that this model actually generated the data.

Model uncertainty

Model uncertainty is an inherent part of modelling, especially in the social sciences. Ignore this at your peril!

Two main strategies for dealing with model uncertainty:

- Model selection: choose "best" model and then conduct inference conditionally upon the assumption that this model actually generated the data. Only works well if model selected is (a really good approximation to) the data generating process. Otherwise it will miss important aspects of reality and inference will be systematically wrong or overly precise

Model uncertainty

Model uncertainty is an inherent part of modelling, especially in the social sciences. Ignore this at your peril!

Two main strategies for dealing with model uncertainty:

- Model selection: choose "best" model and then conduct inference conditionally upon the assumption that this model actually generated the data. Only works well if model selected is (a really good approximation to) the data generating process. Otherwise it will miss important aspects of reality and inference will be systematically wrong or overly precise
- Model averaging: our inference is averaged over all the models in the model space considered, using weights that are either derived from Bayes' theorem (BMA) or from sampling-theoretic optimality considerations (FMA). Here focus on BMA

BMA

In line with probability theory, the formal Bayesian response to dealing with uncertainty is to average

BMA

In line with probability theory, the formal Bayesian response to dealing with uncertainty is to average

E.g. wish to predict the unobserved y_f on the basis of the observed y. Sampling model for y_f and y jointly is $p(y_f|y,\theta_j,M_j)p(y|\theta_j,M_j)$, where M_j is the model selected from K possible models, and $\theta_j \in \Theta_j$ are the parameters of M_j.

BMA

In line with probability theory, the formal Bayesian response to dealing with uncertainty is to average

E.g. wish to predict the unobserved y_f on the basis of the observed y. Sampling model for y_f and y jointly is $p(y_f|y, \theta_j, M_j)p(y|\theta_j, M_j)$, where M_j is the model selected from K possible models, and $\theta_j \in \Theta_j$ are the parameters of M_j. Assign a (continuous) prior $p(\theta_j|M_j)$ for the parameters and a discrete prior $P(M_j)$ on the model space. Predictive distribution is

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

BMA

In line with probability theory, the formal Bayesian response to dealing with uncertainty is to average

E.g. wish to predict the unobserved y_f on the basis of the observed y. Sampling model for y_f and y jointly is $p(y_f|y,\theta_j,M_j)p(y|\theta_j,M_j)$, where M_j is the model selected from K possible models, and $\theta_j \in \Theta_j$ are the parameters of M_j. Assign a (continuous) prior $p(\theta_j|M_j)$ for the parameters and a discrete prior $P(M_j)$ on the model space. Predictive distribution is

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

Averaging at two levels: over parameter values, given each possible model, and discrete averaging over all possible models

BMA

Square brackets in (1): predictive given M_j obtained using the posterior of θ_j given M_j, which is

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

BMA

Square brackets in (1): predictive given M_j obtained using the posterior of θ_j given M_j, which is

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

with the second equality defining $p(y|M_j)$, used in computing the posterior probability of M_j:

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

BMA

Square brackets in (1): predictive given M_j obtained using the posterior of θ_j given M_j, which is

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

with the second equality defining $p(y|M_j)$, used in computing the posterior probability of M_j:

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

Denominators of both averaging operations are made explicit in (2) and (3). $p(y|M_j)$ in (2) is the marginal likelihood of M_j and is a key quantity: Bayes factor is the ratio of marginal likelihoods (posterior odds $=$ Bayes factor $*$ prior odds). $p(y)$ in (3) is a sum (challenge often lies in the number of models K).

BMA

More generally, the posterior distribution of any quantity of interest, say Δ, which has a common interpretation across models is a mixture of the model-specific posteriors with the posterior model probabilities as weights, *i.e.*

$$P_{\Delta|y} = \sum_{j=1}^{K} P_{\Delta \,|\, y, M_j} P(M_j \,|\, y). \tag{4}$$

Construction of the Model Space

Need to define $\mathcal{M}$, the space of all possible models

Construction of the Model Space

Need to define $\mathcal{M}$, the space of all possible models

Three key sources of uncertainty:

- Theory: which variables are important drivers? Often, theories regarding variable inclusion do not contradict each other ("open-endedness" of the theory)
- Specification: functional form, distributions, lag lengths, proxies for theoretical variables
- Heterogeneity : same model for all observations?

Construction of the Model Space

Need to define $\mathcal{M}$, the space of all possible models

Three key sources of uncertainty:

- Theory: which variables are important drivers? Often, theories regarding variable inclusion do not contradict each other ("open-endedness" of the theory)
- Specification: functional form, distributions, lag lengths, proxies for theoretical variables
- Heterogeneity : same model for all observations?

$\mathcal{M}$ should be as broad as possible: you can not learn about anything outside of it! Size is not really an issue.

Construction of the Model Space

Need to define $\mathcal{M}$, the space of all possible models

Three key sources of uncertainty:

- Theory: which variables are important drivers? Often, theories regarding variable inclusion do not contradict each other ("open-endedness" of the theory)
- Specification: functional form, distributions, lag lengths, proxies for theoretical variables
- Heterogeneity : same model for all observations?

$\mathcal{M}$ should be as broad as possible: you can not learn about anything outside of it! Size is not really an issue.

Usually theoretical results are derived under the assumption that $\mathcal{M}$ contains the "true" data-generating model ("$\mathcal{M}$-closed"), but most important results like model selection consistency extend to "$\mathcal{M}$-open" settings in an intuitive manner.

Covariate uncertainty in normal linear regression

Most common setting: model uncertainty about which covariates to include, *i.e.* under model j the n obs. in y are generated from

$$y|\theta_j, M_j \sim N(\alpha\iota + Z_j\beta_j, \sigma^2). \tag{5}$$

Here ι is a vector of ones, Z_j groups k_j of the possible k regressors and $\beta_j \in \Re^{k_j}$ are the regression coefficients. All models contain an intercept $\alpha \in \Re$ and a scale $\sigma > 0$ with a common interpretation.

Covariate uncertainty in normal linear regression

Most common setting: model uncertainty about which covariates to include, *i.e.* under model j the n obs. in y are generated from

$$y|\theta_j, M_j \sim N(\alpha\iota + Z_j\beta_j, \sigma^2). \quad (5)$$

Here ι is a vector of ones, Z_j groups k_j of the possible k regressors and $\beta_j \in \Re^{k_j}$ are the regression coefficients. All models contain an intercept $\alpha \in \Re$ and a scale $\sigma > 0$ with a common interpretation. We standardize the regressors by subtracting their means, which makes them orthogonal to the intercept and makes the interpretation of the intercept common to all models. Assume $n > k$ and design matrix has full column rank

Covariate uncertainty in normal linear regression

Most common setting: model uncertainty about which covariates to include, *i.e.* under model j the n obs. in y are generated from

$$y|\theta_j, M_j \sim N(\alpha\iota + Z_j\beta_j, \sigma^2). \quad (5)$$

Here ι is a vector of ones, Z_j groups k_j of the possible k regressors and $\beta_j \in \Re^{k_j}$ are the regression coefficients. All models contain an intercept $\alpha \in \Re$ and a scale $\sigma > 0$ with a common interpretation. We standardize the regressors by subtracting their means, which makes them orthogonal to the intercept and makes the interpretation of the intercept common to all models. Assume $n > k$ and design matrix has full column rank

$\mathcal{M}$: all subsets of the covariates and thus contains $K = 2^k$ models

Covariate uncertainty in normal linear regression

Most common setting: model uncertainty about which covariates to include, *i.e.* under model j the n obs. in y are generated from

$$y|\theta_j, M_j \sim N(\alpha\iota + Z_j\beta_j, \sigma^2). \quad (5)$$

Here ι is a vector of ones, Z_j groups k_j of the possible k regressors and $\beta_j \in \Re^{k_j}$ are the regression coefficients. All models contain an intercept $\alpha \in \Re$ and a scale $\sigma > 0$ with a common interpretation. We standardize the regressors by subtracting their means, which makes them orthogonal to the intercept and makes the interpretation of the intercept common to all models. Assume $n > k$ and design matrix has full column rank

$\mathcal{M}$: all subsets of the covariates and thus contains $K = 2^k$ models

Economics: k up to 100 (growth), so $K > 10^{30}$ and need efficient computational tools. Genetics (usually $n << k$): k could be up to 100,000, leading to $K > 10^{30,000}$!

BMA: Prior structures

Most commonly used structure is by Fernández et al. (2001)

BMA: Prior structures

Most commonly used structure is by Fernández et al. (2001)

Prior on model parameters:

$$p(\alpha, \beta_j, \sigma \mid M_j) \propto \sigma^{-1} f_N^{k_j}(\beta_j | 0, \sigma^2 g (Z_j' Z_j)^{-1}), \qquad (6)$$

which is a "g-prior" and leads to closed form for integral in (1) and the marginal likelihood (likelihood integrated out with the prior).

BMA: Prior structures

Most commonly used structure is by Fernández et al. (2001)

Prior on model parameters:

$$p(\alpha, \beta_j, \sigma \mid M_j) \propto \sigma^{-1} f_N^{k_j}(\beta_j | 0, \sigma^2 g(Z_j' Z_j)^{-1}), \tag{6}$$

which is a "g-prior" and leads to closed form for integral in (1) and the marginal likelihood (likelihood integrated out with the prior).

Prior over models:

$$P(M_j) = w^{k_j}(1 - w)^{k - k_j}, \tag{7}$$

so that covariates are a priori included independently and with probability w.

BMA: Prior structures

Most commonly used structure is by Fernández et al. (2001)

Prior on model parameters:

$$p(\alpha, \beta_j, \sigma \mid M_j) \propto \sigma^{-1} f_N^{k_j}(\beta_j | 0, \sigma^2 g (Z_j' Z_j)^{-1}), \qquad (6)$$

which is a "g-prior" and leads to closed form for integral in (1) and the marginal likelihood (likelihood integrated out with the prior).

Prior over models:

$$P(M_j) = w^{k_j}(1-w)^{k-k_j}, \qquad (7)$$

so that covariates are a priori included independently and with probability w.

Only requires choice of two scalars g and w: hyperpriors are recommended (adaptive and more robust)

BMA: Properties

BMA has a number of attractive properties for popular choices of (priors on) g, w:

BMA: Properties

BMA has a number of attractive properties for popular choices of (priors on) g, w:

- model selection consistency: if data have been generated by M_j, then the posterior probability of M_j converges to 1 with n (or to the "closest" model in $\mathcal{M}$ if we consider $\mathcal{M}$-open setting)

BMA: Properties

BMA has a number of attractive properties for popular choices of (priors on) g, w:

- model selection consistency: if data have been generated by M_j, then the posterior probability of M_j converges to 1 with n (or to the "closest" model in $\mathcal{M}$ if we consider $\mathcal{M}$-open setting)
- BMA predicts at least as well as any single model (assuming data is generated by (1)) and there is ample empirical evidence for clear superiority (probabilistic forecasts)

BMA: Properties

BMA has a number of attractive properties for popular choices of (priors on) g, w:

- model selection consistency: if data have been generated by M_j, then the posterior probability of M_j converges to 1 with n (or to the "closest" model in $\mathcal{M}$ if we consider $\mathcal{M}$-open setting)
- BMA predicts at least as well as any single model (assuming data is generated by (1)) and there is ample empirical evidence for clear superiority (probabilistic forecasts)
- BMA leads to point estimates that minimize MSE and BMA estimation intervals are calibrated in the sense that the average coverage probability of a BMA interval with posterior probability α is at least equal to α (assuming data is generated by (1)).

Numerical methods for large model spaces

Markov chain Monte Carlo methods on model space only (rest is integrated out with prior in (6))

Numerical methods for large model spaces

Markov chain Monte Carlo methods on model space only (rest is integrated out with prior in (6))

MC^3 is a random-walk Metropolis-Hastings algorithm which suggests proposals in a neighbourhood of the current model and accepts with a certain probability to ensure that draws converge to the correct posterior

Numerical methods for large model spaces

Markov chain Monte Carlo methods on model space only (rest is integrated out with prior in (6))

MC^3 is a random-walk Metropolis-Hastings algorithm which suggests proposals in a neighbourhood of the current model and accepts with a certain probability to ensure that draws converge to the correct posterior

MC^3 is implemented in freely available software and has been shown to work very well

Numerical methods for large model spaces

Markov chain Monte Carlo methods on model space only (rest is integrated out with prior in (6))

MC^3 is a random-walk Metropolis-Hastings algorithm which suggests proposals in a neighbourhood of the current model and accepts with a certain probability to ensure that draws converge to the correct posterior

MC^3 is implemented in freely available software and has been shown to work very well

More challenging cases such as with very correlated regressors and/or $k >> 100$ might need adaptive samplers (tuning as the chain is being generated): some have been shown to work well for k in order of $10,000$s.

Numerical methods for large model spaces

Markov chain Monte Carlo methods on model space only (rest is integrated out with prior in (6))

MC^3 is a random-walk Metropolis-Hastings algorithm which suggests proposals in a neighbourhood of the current model and accepts with a certain probability to ensure that draws converge to the correct posterior

MC^3 is implemented in freely available software and has been shown to work very well

More challenging cases such as with very correlated regressors and/or $k >> 100$ might need adaptive samplers (tuning as the chain is being generated): some have been shown to work well for k in order of $10,000$s.

So no need to avoid using large $\mathcal{M}$!

Role of the prior

Effect of the prior on posterior model probabilities can be much more pronounced than on posterior inference given a model

Role of the prior

Effect of the prior on posterior model probabilities can be much more pronounced than on posterior inference given a model

Posterior odds between models, given g and w:

$$\frac{P(M_i \mid y, w, g)}{P(M_j \mid y, w, g)} = \left(\frac{w}{1-w}\right)^{k_i-k_j} (1+g)^{\frac{k_j-k_i}{2}} \left(\frac{1+g(1-R_i^2)}{1+g(1-R_j^2)}\right)^{-\frac{n-1}{2}} \tag{8}$$

The three factors correspond to a model size penalty induced by the prior on the model space, a model size penalty resulting from the marginal likelihood and a lack-of-fit penalty from the marginal likelihood.

Role of the prior

Effect of the prior on posterior model probabilities can be much more pronounced than on posterior inference given a model

Posterior odds between models, given g and w:

$$\frac{P(M_i \mid y, w, g)}{P(M_j \mid y, w, g)} = \left(\frac{w}{1-w}\right)^{k_i - k_j} (1+g)^{\frac{k_j - k_i}{2}} \left(\frac{1+g(1-R_i^2)}{1+g(1-R_j^2)}\right)^{-\frac{n-1}{2}} \tag{8}$$

The three factors correspond to a model size penalty induced by the prior on the model space, a model size penalty resulting from the marginal likelihood and a lack-of-fit penalty from the marginal likelihood.

Hyperpriors on g and w can have a large effect on the induced penalties for model complexity but not on the impact of the relative fit of the models

Complexity penalties

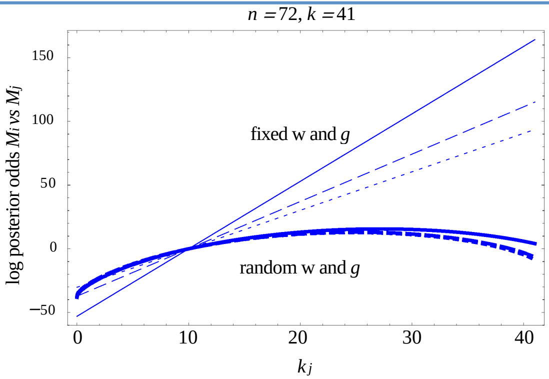

Figure: Posterior odds as a function of k_j when $k_i = 10$ with equal fit, using prior mean model size $m = 7$ (solid), $m = k/2$ (dashed), and $m = 2k/3$ (dotted). Bold lines correspond to random w and g

Complexity penalties

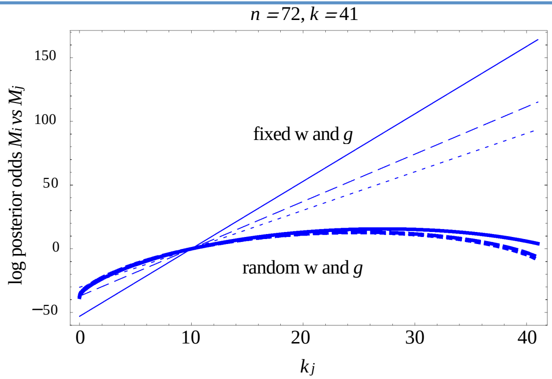

Figure: Posterior odds as a function of k_j when $k_i = 10$ with equal fit, using prior mean model size $m = 7$ (solid), $m = k/2$ (dashed), and $m = 2k/3$ (dotted). Bold lines correspond to random w and g

Hyperpriors more robust to m, less extreme and penalize models of size around $k/2$ (multiplicity penalty)

Approximations and hybrids

Closed-form marginal likelihood may not be available with other model structures (different sampling models, like Student-t, GLMs; different priors). Formally correct approach is to include model parameters in the MCMC, but this may be cumbersome

Approximations and hybrids

Closed-form marginal likelihood may not be available with other model structures (different sampling models, like Student-t, GLMs; different priors). Formally correct approach is to include model parameters in the MCMC, but this may be cumbersome

In regular models BIC tends to log Bayes factor with n; so BIC often used as (easy) approximation in more complex settings

Approximations and hybrids

Closed-form marginal likelihood may not be available with other model structures (different sampling models, like Student-t, GLMs; different priors). Formally correct approach is to include model parameters in the MCMC, but this may be cumbersome

In regular models BIC tends to log Bayes factor with n; so BIC often used as (easy) approximation in more complex settings

Hybrids of frequentist and Bayesian methods were used e.g. to deal with endogenous regressors: BIC approximations to posterior model probabilities for averaging over classical two-stage least squares (2SLS) estimates.

Other sampling models

BMA for many other models has been considered:

Other sampling models

BMA for many other models has been considered:

- Generalized linear models (GLMs), for example logistic, probit or ordered response models
- Generalized additive models (nonlinear effects)
- Models for outliers and non-normal models (e.g. Student-t)
- Dynamic models, e.g. AR(F)IMA and DLMs
- Models with endogenous regressors (IV models)
- Models for longitudinal data with individual effects
- Models for spatial data (Spatial AR models)
- Duration models

Endogeneity

Endogeneity occurs if one or more of the covariates is correlated with the error term in the equation corresponding to (5). In particular, consider the following extension of the model in (5):

$$
\begin{aligned}
y &= \alpha\iota + x\gamma + Z_j\beta_j + \varepsilon & (9)\\
x &= W\delta + \nu, & (10)
\end{aligned}
$$

where x is an endogenous regressor and W is a set of instruments, independent of ε. The error terms are iid:

$$
(\varepsilon_i, \nu_i)' \sim N(0, \Sigma), \tag{11}
$$

with $\Sigma = (\sigma_{ij})$ a 2×2 covariance matrix. Whenever $\sigma_{12} \neq 0$ this introduces a bias in the OLS estimator of γ and a standard classical approach is the use of 2SLS estimators instead.

BMA with Endogeneity

Posterior inference on coefficients and model probabilities is affected, even for large n.

BMA with Endogeneity

Posterior inference on coefficients and model probabilities is affected, even for large n.

Durlauf et al. (2008) focus on uncertainty in the selection of the endogenous and exogenous variables and propose to average over 2SLS model-specific estimates with BIC-based weights.

BMA with Endogeneity

Posterior inference on coefficients and model probabilities is affected, even for large *n*.

Durlauf et al. (2008) focus on uncertainty in the selection of the endogenous and exogenous variables and propose to average over 2SLS model-specific estimates with BIC-based weights.

Lenkoski et al. (2014) also account for model uncertainty in the selection of instruments. They propose a two-step procedure that first averages across the first-stage models (for the endogenous variables) and then, given the fitted endogenous regressors from the first stage, it again takes averages in the second stage. Both steps use BIC weights (approximations).

BMA with Endogeneity

Posterior inference on coefficients and model probabilities is affected, even for large *n*.

Durlauf et al. (2008) focus on uncertainty in the selection of the endogenous and exogenous variables and propose to average over 2SLS model-specific estimates with BIC-based weights.

Lenkoski et al. (2014) also account for model uncertainty in the selection of instruments. They propose a two-step procedure that first averages across the first-stage models (for the endogenous variables) and then, given the fitted endogenous regressors from the first stage, it again takes averages in the second stage. Both steps use BIC weights (approximations).

Karl and Lenkoski (2012) propose IVBMA, which uses conditional Bayes factors to account for model uncertainty within a Gibbs algorithm. Their algorithm hinges on certain restrictions (*e.g.* joint Normality and conditionally conjugate priors), but it is exact, efficient and is implemented in an R-package.

Application in Economics

Lenkoski et al. (2014). *Econometric Reviews*

Three main theories about what drives economic growth: geography (natural and human resources), international trade (linked with market integration) and institutions (property rights and the rule of law)

Application in Economics

Lenkoski et al. (2014). *Econometric Reviews*

Three main theories about what drives economic growth: geography (natural and human resources), international trade (linked with market integration) and institutions (property rights and the rule of law)

Many ways in which these theoretical determinants could be measured, so a large collection of possible models

Application in Economics

Lenkoski et al. (2014). *Econometric Reviews*

Three main theories about what drives economic growth: geography (natural and human resources), international trade (linked with market integration) and institutions (property rights and the rule of law)

Many ways in which these theoretical determinants could be measured, so a large collection of possible models

Only geography can be safely assumed to be exogenous.

Application in Economics

Lenkoski et al. (2014). *Econometric Reviews*

Three main theories about what drives economic growth: geography (natural and human resources), international trade (linked with market integration) and institutions (property rights and the rule of law)

Many ways in which these theoretical determinants could be measured, so a large collection of possible models

Only geography can be safely assumed to be exogenous.

In previous literature some (influential) studies found evidence that property rights are a strong driver for growth, but without considering many alternative models. Similarly, others concluded that trade variables were key drivers (without controlling for the effect of institutions). Rodrik et al. (2004) (RST) provide a "horse race" among alternative theories that propose candidate instruments and regressors, but don't use BMA (they just compare a limited set of models) and conclude only institutions matter

Application in Economics

Lenkoski et al. (2014). *Econometric Reviews*

Consider e.g. Rule of Law and Integration (Openness):

	Rule of Law			Integration		
Models	PIP	mean	sd	PIP	mean	sd
RST core	1.00	1.28	0.18	0.20	0.11	0.26
limited $\mathcal{M}$	1.00	0.95	0.13	0.07	0.07	0.14
full $\mathcal{M}$	0.96	0.80	0.32	0.85	0.93	0.38

Table: Some BMA (2nd stage) results with different sets of possible covariates (PIP is posterior inclusion probability)

Divergence of results (between 2SLS and BMA) grows as we allow for more uncertainty (bigger model spaces). Integration becomes important driver and all three theories are supported in the BMA results using all available variables.

Application in Political Science

Rainey, Electoral Studies (2016)

Do proportional electoral rules cause higher turnout?

Application in Political Science

Rainey, Electoral Studies (2016)

Do proportional electoral rules cause higher turnout?

Survey data from the 2001 Taiwan legislative election: useful since there is substantial variation in the proportionality of electoral rules across districts (magnitude varies with one to 13 seats per district)

Do proportional electoral rules cause higher turnout?

Survey data from the 2001 Taiwan legislative election: useful since there is substantial variation in the proportionality of electoral rules across districts (magnitude varies with one to 13 seats per district)

Hypotheses: larger district magnitude (more proportionality) could make potential voters more likely to

- feel represented
- feel close to a political party
- be contacted by a political party
- turn out to vote

Application in Political Science

Rainey, Electoral Studies (2016)

BMA explaining survey outcomes addressing the hypotheses through 16 potential regressors (logistic regression with BIC approximation to compute posterior model probabilities)

Application in Political Science

Rainey, Electoral Studies (2016)

BMA explaining survey outcomes addressing the hypotheses through 16 potential regressors (logistic regression with BIC approximation to compute posterior model probabilities)

Posterior probability that district magnitude has no effect on each of the four possible outcomes is at least 0.97

Application in Political Science

Rainey, Electoral Studies (2016)

- BMA explaining survey outcomes addressing the hypotheses through 16 potential regressors (logistic regression with BIC approximation to compute posterior model probabilities)
- Posterior probability that district magnitude has no effect on each of the four possible outcomes is at least 0.97
- In contrast with some of the literature, but this study takes into account model uncertainty in a reasonably large model space (with other potential explanatory variables) rather than focus on a very limited set of models

BMA explaining survey outcomes addressing the hypotheses through 16 potential regressors (logistic regression with BIC approximation to compute posterior model probabilities)

Posterior probability that district magnitude has no effect on each of the four possible outcomes is at least 0.97

In contrast with some of the literature, but this study takes into account model uncertainty in a reasonably large model space (with other potential explanatory variables) rather than focus on a very limited set of models

Important predictors for voting turnout are age and marital status

Software and resources

A number of free `R`-packages: BMS, BAS, BayesVarSel, ivbma (endogenous regressors)

Software and resources

A number of free `R`-packages: BMS, BAS, BayesVarSel, ivbma (endogenous regressors)

Packages for "standard" normal linear model all tend to be efficient and accurate, leading to reliable inference within 10 minutes on a simple PC for problems up to $k = 100$ or so covariates

Software and resources

A number of free `R`-packages: BMS, BAS, BayesVarSel, ivbma (endogenous regressors)

Packages for "standard" normal linear model all tend to be efficient and accurate, leading to reliable inference within 10 minutes on a simple PC for problems up to $k = 100$ or so covariates

Some useful resources available online, e.g. `http://bms.zeugner.eu/resources/` (also introductory material)

Conclusion

- Model uncertainty is a pervasive problem in applications

Conclusion

- Model uncertainty is a pervasive problem in applications
- Simply ignoring the problem is not a solution (biased and overconfident inference)

Conclusion

- Model uncertainty is a pervasive problem in applications
- Simply ignoring the problem is not a solution (biased and overconfident inference)
- BMA presents a formal solution, which optimally takes into account uncertainty, within model space

Conclusion

- Model uncertainty is a pervasive problem in applications
- Simply ignoring the problem is not a solution (biased and overconfident inference)
- BMA presents a formal solution, which optimally takes into account uncertainty, within model space
- Priors matter for BMA and it is crucial to be aware of this
  - this needs to be understood and properly communicated
  - "robustify" priors through hyperpriors on e.g w and g
  - elicited through intuitive quantities, e.g. prior mean model size

Conclusion

- Model uncertainty is a pervasive problem in applications
- Simply ignoring the problem is not a solution (biased and overconfident inference)
- BMA presents a formal solution, which optimally takes into account uncertainty, within model space
- Priors matter for BMA and it is crucial to be aware of this
  - this needs to be understood and properly communicated
  - "robustify" priors through hyperpriors on e.g w and g
  - elicited through intuitive quantities, e.g. prior mean model size
- Lots of unexplored areas especially outside of the normal linear model (prior structures, elicitation and effects; properties; computation)

Conclusion

- Model uncertainty is a pervasive problem in applications
- Simply ignoring the problem is not a solution (biased and overconfident inference)
- BMA presents a formal solution, which optimally takes into account uncertainty, within model space
- Priors matter for BMA and it is crucial to be aware of this
  - this needs to be understood and properly communicated
  - "robustify" priors through hyperpriors on e.g w and g
  - elicited through intuitive quantities, e.g. prior mean model size
- Lots of unexplored areas especially outside of the normal linear model (prior structures, elicitation and effects; properties; computation)
- I hope that BMA can become a key methodology in many areas of application (it already is in macroeconomics), and can contribute to constructive communication by better understanding the reasons for differences in empirical findings

Some useful references

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

Lenkoski, A., T. Eicher, and A. Raftery (2014). Two-stage Bayesian model averaging in endogenous variable models. *Econometric Reviews* 33, 122–51.

Liang, F., R. Paulo, G. Molina, M. Clyde, and J. Berger (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association* 103, 410–23.

Madigan, D. and J. York (1995). Bayesian graphical models for discrete data. *International Statistical Review* 63, 215–32.

Steel, M.F.J. (2020). Model Averaging and its Use in Economics. *Journal of Economic Literature*, 58, 644–719. https://arxiv.org/abs/1709.08221