WIRTSCHAFTS
UNIVERSITÄT
WIEN VIENNA
UNIVERSITY OF
ECONOMICS
AND BUSINESS

Data Science and Machine Learning

Bayesian Model Averaging

Katharina Fenz

November 29, 2023

Outline

The theory behind Bayesian model averaging

- Bayesian vs. frequentist statistics
- Bayes theorem
- Bayesian econometrics
- Bayesian model averaging

Bayesian model averaging in R and special cases

- Bayesian model averaging with the package `BMS`
- Bayesian model averaging with interaction terms
- Spatial Bayesian model averaging

The Intuition behind Frequentist vs. Bayesian Statistics

If I toss a fair coin, what is the probability that it lands heads up?

Now imagine that I actually toss the coin and it lands in my hand, but I cover it, so you cannot see which side is up. What is the probability that it is heads up?

- Frequentist: I don't know the answer, but the probability that it is heads up is either 100% if it landed heads up or 0% if it landed tails up.
  $\rightarrow$ Parameters are unknown, but fixed.
- Bayesian: For me the probability that the coin is heads up is 50%.
  $\rightarrow$ Parameters are random variables that can be described with a probability distribution.

Frequentist vs. Bayesian Statistics I

Probability

- Frequentist: Probability represents the likelihood of an event occurring based on its observed relative frequency.
- Bayesian: Probability represents the degree of belief in the occurrence of an event given prior knowledge and current evidence.

Parameters

- Frequentist: Parameters are considered fixed, unknown values. The goal is to estimate these fixed values based on sample data.
- Bayesian: Parameters are treated as random variables with probability distributions. Prior beliefs about the parameters are combined with the observed data to update the beliefs and obtain a posterior distribution of the parameters.

Prior information

- Frequentist: Frequentist approaches typically do not incorporate prior beliefs or information about parameters. Inference is based solely on the observed data.
- Bayesian: Bayesian methods include the use of prior distributions, which represent beliefs about the parameters before observing the data. This prior information is then updated using the observed data to obtain a posterior distribution.

Inference

- Frequentist: Inference is based on point estimates and confidence intervals.
- Bayesian: Inference is based on posterior distributions.

Bayes' Theorem I

Bayes' theorem can be written as

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

with the following key components:

- **Posterior Probability P(A|B):** the probability of event A occurring given that event B has occurred, represents our updated belief about A after considering new information B
- **Likelihood P(B|A):** the probability of observing event B given that event A has occured, describes the likelihood of the observed data B under the hypothesis A
- **Prior Probability P(A):** the initial belief or probability of A occurring before considering the new evidence B, represents our prior knowledge or assumptions about A
- **Marginal Likelihood P(B):** the probability of observing B irrespective of A, serves as a normalization factor

Bayes' Theorem II

Bayes' theorem

- allows us to update our beliefs (posterior probability)
- based on new evidence (likelihood),
- taking into account our initial beliefs (prior probability)
- and the overall likelihood of the new evidence (marginal likelihood).

Bayes' Theorem - Example I

Assume someone tests whether they have a rare disease, where

- A: the person has the disease
- B: the test is positive

and based on that

- $P(A|B)$: probability of having the disease given a positive test result
- $P(B|A)$: probability of testing positive if the person has the disease (sensitivity of the test)
- $P(A)$: prior probability of having the disease
- $P(B) = P(A)P(B|A) + P(-A)P(B| - A)$: probability of testing positive
- $P(B| - A)$: probability of testing positive if the person does not have the disease (specificity of the test)

Bayes' Theorem - Example II

Assume that $P(B|A) = 0.99$, $P(A) = 0.001$ and $P(B|-A) = 0.01$ and use Bayes' theorem to compute $P(A|B)$:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(A)P(B|A) + P(-A)P(B|-A)}$$

$$P(A|B) = \frac{0.99 * 0.001}{0.001 * 0.99 + 0.999 * 0.01}$$

$$P(A|B) = 0.09016 \approx 9\%$$

Assume we want to use a data set $\mathcal{D}$ to estimate β in the following model:

$$y = x\beta + \sigma\epsilon$$

We can use Bayes' theorem to predict

$$p(\beta|\mathcal{D}) = \frac{p(\mathcal{D}|\beta)p(\beta)}{p(\mathcal{D})}$$

where

- the posterior density $p(\beta|\mathcal{D})$ describes the probability of β taking a specific value given $\mathcal{D}$
- the likelihood function $p(\mathcal{D}|\beta)$ describes the probability of observing $\mathcal{D}$ given a specific value of β
- the prior density $p(\beta)$ represents our beliefs about β before observing $\mathcal{D}$
- the marginal likelihood $p(\mathcal{D})$ represents the probability of observing $\mathcal{D}$ irrespective of β and ensures that $p(\beta|\mathcal{D})$ integrates to 1

Priors I

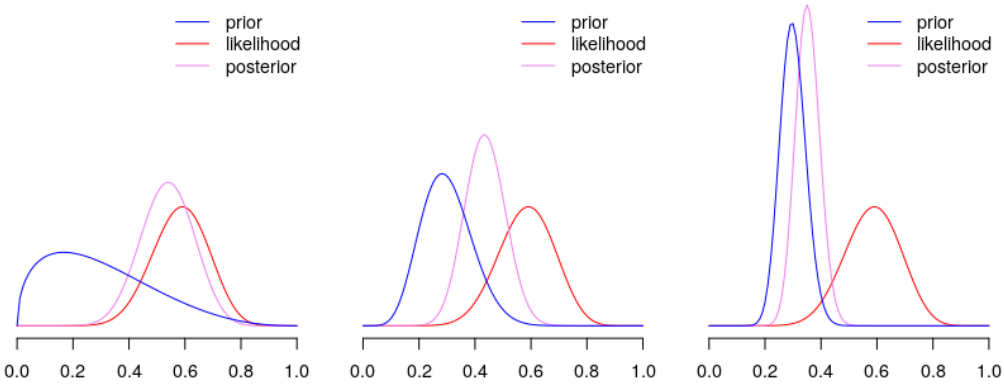
The prior density $p(\beta)$ represents our beliefs or information about the parameter β before observing new data. It encapsulates any prior knowledge, assumptions, or expectations we have about the likely values of $p(\beta)$ based on external information or expertise. As the prior is central to Bayesian econometrics, we should consider the following points:

- **Influence on the posterior:** Through Bayes' Theorem, the prior influences the posterior distribution. The posterior is a combination of the likelihood (how likely the data are given the model) and the prior (our initial beliefs about the parameter). The more informative or restrictive the prior, the stronger its influence on the posterior.

Priors II

- **Subjectivity:** The choice of the prior is subjective and may vary among different analysts or researchers. It reflects their beliefs or assumptions about the parameter.
- **Sensitivity:** As the prior is subjective and affects our results, it can make sense to perform a sensitivity analysis on the prior. This involves examining how changes in the prior distribution affect the posterior results.
- **Regularization:** In cases where the sample size is small or the data is insufficient to precisely estimate the parameters, the prior can act as a form of regularization. It helps prevent overfitting by constraining the parameter space and discouraging extreme parameter values.

How Priors can affect our Results



Bayesian Model Averaging I

- Bayesian model averaging (BMA) is based on Bayesian statistics and explicitly takes into account model uncertainty when making inferences or predictions.
- Considering model uncertainty helps to prevent issues like overfitting and multicollinearity, which particularly tend to come up when working with large sets of potential explanatory variables.
- BMA addresses the challenge of model selection by combining the parameter estimates of multiple models, which are weighted by their ability to explain the data.

Bayesian Model Averaging II

Assume

$$y = X_k \beta_k + \sigma \epsilon$$

where X_k is a matrix of k explanatory variables

- BMA creates a weighted average of all models that can be set up as combinations of the variables in X_k
- With k potential explanatory variables, we have $m = 2^k$ models, which can be denoted as M_i for $i = 1, ..., m$
- 10 covariates would already lead to over 1,000 models, 20 covariates to over 1,000,000 models
- The cardinality of such a big model space makes the estimation of each model infeasible $\rightarrow$ Markov chain Monte Carlo model composition (MC^3)

Bayesian Model Averaging III

- Weights are assigned to individual models based on the posterior model probabilities resulting from Bayes' theorem
- Posterior model probability of a model M_i can be written as

$$p(M_i|\mathcal{D}) = \frac{p(\mathcal{D}|M_i)p(M_i)}{p(\mathcal{D})} = \frac{p(\mathcal{D}|M_i)p(M_i)}{\sum_{i=1}^{m} p(\mathcal{D}|M_i)p(M_i)}$$

where

- $p(M_i)$ is the prior on model M_i
- $p(\mathcal{D}|M_i)$ is the likelihood of model M_i
- $p(\mathcal{D})$ is the probability of $\mathcal{D}$ integrated over the model space

Bayesian Model Averaging IV

- Given the posterior model probabilities, we can write the posterior distribution of coefficient β_k as

$$p(\beta_k|\mathcal{D}) = \sum_{i=1}^{m} p(\beta_k|M_i, \mathcal{D})p(M_i|\mathcal{D})$$

- The posterior distribution of estimator of β_k (given $\mathcal{D}$) is the average of its posterior distribution under each model, weighted by its posterior model probability

Parameter Priors for BMA I

- To perform BMA, we have to select priors for the parameters and the models.
- The most common choice for parameter priors is Zellner's g-prior, which can be written as

$$\beta_k | g \sim N(0, g\sigma^2 (X_k{}^\top X_k)^{-1})$$

- Zellner's g-prior implements the assumption that $\beta_i = 0$ for all β_i with $i = 1, ..., k$.
- This corresponds to assuming that non of the explanatory variables in X_k has an effect on y before seeing the data.

Parameter Priors for BMA II

- To use Zellner's g-prior, we have to set g, which expresses how sure we are about $\beta_i = 0$.
- A common choice for g was introduced by Fernandez, Ley and Steel in 2001 and considers the Bayesian information criterion (with $g = n$) and the risk inflation criterion (with $g = k^2$).
- This so-called BRIC prior sets $g = max(n, k^2)$, where n is the number of observations and k is the number of covariates.

Model Priors for BMA

- A common choice for model priors is the uniform model prior, which assigns the same prior probability to each possible model.
- As most models consider a number of covariates close to $\frac{k}{2}$, in this case the aggregate prior probabilities of all models considering $\frac{k}{2}$ covariates is much higher than that of all models with only one covariate or the model including all covariates k.
- Another option is to assign equal prior probability to each possible model size.
- In this case, the prior probability of the model considering all covariates k equals the sum of the prior probabilities of all models with $\frac{k}{2}$ covariates.

Posterior Odds Ratio

- The posterior odds ratio is a way of comparing the posterior probabilities of two models M_i and M_j. It can be written as

$$PO_{ij} = \underbrace{\frac{p(M_i|\mathcal{D})}{p(M_j|\mathcal{D})}}_{\text{posterior odds}} = \underbrace{\frac{p(\mathcal{D}|M_i)}{p(\mathcal{D}|M_j)}}_{\text{Bayes factor}} \cdot \underbrace{\frac{p(M_i)}{p(M_j)}}_{\text{prior odds}}$$

- This approach of comparing different models is part of a strategy that helps us to identify the most relevant models in our model space: Markov chain Monte Carlo model composition (MC^3)

Markov Chain Monte Carlo Model Composition (MC^3) I

- With an increasing number of covariates, our model space in BMA quickly becomes huge.
- 10 covariates lead to over 1,000 models, 20 covariates to over 1,000,000.
- The cardinality of such a big model space makes the estimation of each individual model infeasible.
- Markov chain Monte Carlo model composition allows us to identify the most important models and make reasonable predictions without checking every individual model in our model space.

Markov Chain Monte Carlo Model Composition (MC^3) II

- Markov Chain Monte Carlo model composition can be based on a birth-death sampler.
- This sampler starts at model M_i with a posterior model probability of $p(M_i|\mathcal{D})$.
- One of the k potential covariates in X_k is randomly chosen.
- If the covariate is part of M_i, it is dropped ("dies"), otherwise it is added ("born") to create another model M_j.
- The sampler now switches from M_i to M_j with the probability $p_{i,j}$:

$$p_{i,j} = min(1, \underbrace{\frac{p(M_j|\mathcal{D})}{p(M_i|\mathcal{D})}}_{\text{posterior odds}})$$

- The number of times each model is kept converges to the distribution of the posterior model probabilities.

BMA using MC^3

- Start with a random model M_i.
- Compute and save the corresponding parameter estimates β_i and the posterior model probability $p(M_i|\mathcal{D})$.
- Create a neighboring model M_j by adding or dropping one covariate.
- Switch to the new model with probability $p_{i,j}$.
- Compute and save the corresponding parameter estimates and repeat the whole process for a predefined number of iterations.
- Discard the results of the first iterations.
- Compute parameter estimates that are based on the parameter estimates of the remaining models, weighted by their posterior model probabilities.

BMA in R I

- Package for BMA: `BMS`
- Estimation:

```
bma_1 = bms(DATA, burn=1000, iter=3000, g="BRIC", mprior="uniform", mcmc="bd")
```

- Prediction:

```
pred_bma_1 = predict(bma_1, newdata = NEW_DATA])
```

To perform BMA in R, we set the following parameters:

- `burn`: number of models that are sampled and discarded at the beginning of the MC^3 sampling process
- `iter`: number of models that are sampled and used for the analysis after the burn-in draws
- `g`: hyperparameter on Zellner's g-prior for the regression coefficients, "BRIC" corresponds to $g = max(n, k^2)$
- `mprior`: model prior, "uniform" assigns the same prior probability to each possible model
- `mcmc`: model sampler, "bd" corresponds to a birth-death sampler

BMA with Interaction Terms I

BMA also allows to estimate models with interaction terms, like

$$y = X_k \beta_{1,k} + \sum_{i,j=1}^{k} x_i \times x_j \beta_{2,i,j} + \sigma\epsilon$$

where X_k is a matrix of k explanatory variables, $x_i, x_j \in X_k$ and $i \neq j$

- Estimating these models with BMA, we have to decide whether we want to consider models with interaction terms that do not include the corresponding parent variables.
- Uniform model priors consider all models, strong heredity priors only allow interaction terms in models including both parent variables.
- Strong heredity priors can be implemented in R using `mcmc="bd.int"`

BMA with Interaction Terms II

- Models with interaction terms that do not include both parent variables are harder to interpret. It is unclear whether effects come from the interaction terms or their missing parent variables.

Optional reading on the topic:

- Masanjala and Papageorgiou (2008) analyzed growth in Africa with BMA, interaction terms and a uniform model prior.
- Crespo Cuaresma (2011) and Moser and Hofmarcher (2014) criticized this approach and showed that strong heredity priors can improve interpretability and model performance.

Spatial BMA I

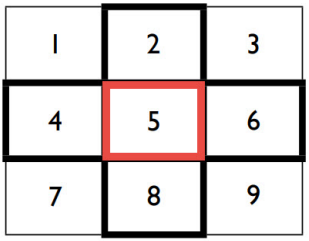
BMA also allows us to estimate spatial autoregressive models, like

$$y = \rho W y + X\beta + \epsilon$$

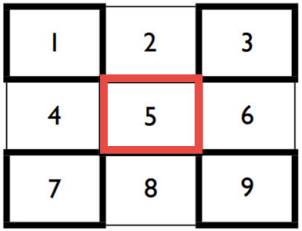
where W is a spatial weight matrix and ρ reflects the degree of spatial autocorrelation.

- Spatial autoregressive models can help to model spatial data referring to, for example, neighboring countries or districts.
- In this case, the observations y_i and y_j of two neighboring units i and j might not be independent from each other.
- The results of such a spatial autoregressive model depend on the structure of the spatial weight matrix that is taken into account. In other words, the results depend on which units we consider as being neighbors.

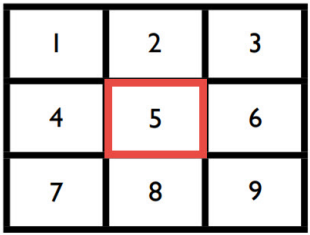
Spatial BMA II



rook contiguity - edges only
2, 4, 6, 8 are neighbors of 5



bishop contiguity - corners only
1, 3, 7, 9 are neighbors of 5



queen contiguity - edges and corners
5 has eight neighbors

Spatial BMA III

- Different types of spatial weight matrices can capture different types of spatial autocorrelation and it is not necessarily clear which one to use.
- Spatial BMA addresses this uncertainty by considering a weighted average of different spatial weight matrices.
- BMA with spatial effects can be implemented in R with the package `spatBMS`.

Optional reading on the topic:

- Crespo Cuaresma and Feldkircher (2013) used BMA with spatial effects to analyze the speed of income convergence in Europe.

References

- Crespo Cuaresma, Jesus (2011). "How different is Africa? A comment on Masanjala and Papageorgiou". In: Journal of Applied Econometrics 26.6, pp. 1041–1047.
- Crespo Cuaresma, Jesús and Martin Feldkircher (2013). "Spatial filtering, model uncertainty and the speed of income convergence in Europe". In: Journal of Applied Econometrics 28.4, pp. 720–741.
- Fernandez, Carmen, Eduardo Ley, and Mark FJ Steel (2001). "Benchmark priors for Bayesian model averaging". In: Journal of Econometrics 100.2, pp. 381–427.
- Masanjala, Winford H and Chris Papageorgiou (2008). "Rough and lonely road to prosperity: a reexamination of the sources of growth in Africa using Bayesian model averaging". In: Journal of Applied Econometrics 23.5, pp. 671–682.
- Moser, Mathias and Paul Hofmarcher (2014). "Model priors revisited: Interaction terms in BMA growth applications". In: Journal of Applied Econometrics 29.2, pp. 344–347.