Maastricht University *Leading in Learning!*

Course EBC 2043

Introduction to Software in Econometrics and Operations Research

Econometrics part Assignments

Academic year 2020-21

Contents

1 **Out-of-Sample Model Comparison** 3

2 **Bayesian LASSO Application** 3

3 **Independence Analysis in a Markov Chain Model** 4

Introduction

For your assignment, choose one of the topics listed in the following sections. Within each topic, different tasks are specified.

Deadlines

- *Preliminary report:* January 21st, 12:00
- *Final report:* January 27th, 12:00

Late submissions will be penalized.

Deliverables

- a PDF file `XX-Report.pdf` (where XX is your student number) of a short report of 5-10 pages. Be concise and to the point.
  This report should include the econometric model definition, formulation of the research question, simulation procedure used for the analysis and a summary of the data used for the analysis. In addition, the report should include the main results of your R implementation and interpretation of these results.
  You are expected to prepare this report *individually*. Cooperation with the other students is not allowed.
- a zip file `XX-sourcefiles.zip` (no tar, rar or other compressing formats), where XX is your student number. This zip file should contain the data and all R source code that is needed to run and reproduce your analysis.
  You can work on the R code with your other team member. However, different data sets need to be chosen and simulations should be carried out individually with different seeds for initializing random number generation.
- (optional) If you need to submit any extra files, include them in a separate zip file named `XX-Extras.zip` (where XX is your student number).

Guidelines for the report

The following should be included in your report.

- Research question. Example research questions are:
  - Are the loan payment failures of two borrowers correlated? (research question for real data)
  - Does the Metropolis Sampling algorithm perform better than the Importance Sampling algorithm in a random effects model? (research question that can be addressed using simulated data)
- Data definition. The data can be simulated or real data which are appropriate for the model and the analysis. Your data definition should include a plot of the data and/or summary of data properties (e.g. number of observations, data source, interesting properties of data that affect your econometric model choice etc)
- The econometric model.
- Definition of the selected prior / priors and explanation of why you choose a specific prior.
- Likelihood and the posterior density of your model.
- The simulation algorithm(s) you use for the analysis. Explanation of why you choose this (these) simulation algorithms.
  For this part you may need to include the likelihood of the model and provide the equation for the posterior density of parameters.
- Important computational aspects of your simulation algorithm. E.g. did you use burn-in draws? Why/why not? Did you trim draws? Why/why not? Do you need to report weights of draws or acceptance rates?
- Results of the simulation algorithm. Posterior mean, standard deviation, 95% intervals for the parameters of interest (this depends on your research question).
- General conclusion about your findings. I.e. what your conclusion is in relation the research question you started with.

Data

The assignments do not specify an explicit data for the model. You can select any data that you want to analyze. In addition, you may prefer to simulate data and analyze simulated data. Make sure that a brief data definition is included in your report and the dataset is also included in your zip file.

Some useful resources for freely economic data are:

- *R* has an extensive list of available datasets. You can check the description of these datasets, and choose one dataset which you find interesting:

https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/00Index.html

- A detailed website for several economic data sets and sources:
  https://www.economicsnetwork.ac.uk/links/data_free

1 Out-of-Sample Model Comparison

For this topic, consider the univariate linear regression model with k explanatory variables, for observations $i = 1, \ldots, N$,

$$y_i = X_i\beta + \varepsilon_i,$$

where y_i is (scalar) dependent variable, X_i are $1 \times k$ vectors of explanatory variables, $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ are the errors, and β is the $k \times 1$ vector of coefficients.
The purpose is to consider different priors for the model parameters β and/or σ^2, and assess which prior leads to better results according to out-of-sample (forecasting) performance.
Note: Team members need to choose different datasets.

- Choose a dataset and apply a linear regression with multiple explanatory variables.
- Choose two different priors for the model. Explain the intuition behind choosing these priors.
- Split the data to an estimation and hold-out sample.
- Estimate the linear regression model with two priors. Notice that you may need to adjust the sampling algorithm for different priors. Clearly explain which sampling algorithms you use and how they are implemented.
- Obtain predictions of the hold-out sample data points.
- Compare the out-of-sample results with different priors. Which prior leads to better out-of-sample results? Explain the reasoning for this specific prior to work better than the other ones. Also explain which measure you use to compare out-of-sample results.

2 Bayesian LASSO Application

Many economic and statistical applications suffer from high dimensionality, namely, having too many parameters to estimate compared to the number of data points. There is a large literature (frequentist and Bayesian) aiming to reduce the dimension of the parameter space. The most common method for this is to use a LASSO estimator.
For this exercise, consider a linear regression model. The purpose is to report the parameter estimates under an uninformative prior and under the LASSO prior.

- Choose a data appropriate for a linear regression model, and at least 3 explanatory variables. (*Note: Group members need to choose different data.*)
- The linear regression model is given by:

$$y = X\,\beta + \varepsilon,$$

where y is the $n \times 1$ vector of dependent variable, X is an $n \times K$ matrix of explanatory variables and $\varepsilon \sim \mathcal{N}(0, \sigma^2\, I)$, where I indicates the identity matrix. Write down the log-likelihood of this model.
Perform the remaining analysis for two priors:
  (i) $p_1(\beta, \sigma^2) \propto \frac{1}{\sigma^2}$, corresponding to the flat prior.
  (ii) $p_2(\beta, \sigma^2) \propto \frac{1}{\sigma^2} \times \prod_{k=1}^{K} \frac{\lambda}{2\sqrt{\sigma^2}} e^{-\lambda|\beta_k|/\sqrt{\sigma^2}}$
- Create two functions to calculate the (log-)prior density for the above priors.
- Apply the Importance Sampling or Metropolis Hastings algorithm to obtain posterior results for the two priors separately. Clearly explain the candidate distribution you choose.
- Compare the posterior results of β, σ^2 obtained under the two priors. Consider reporting the posterior means, variances or quantiles of interest.
- What do you conclude in terms of the relation between X and y variables? Does the LASSO prior indeed shrink the parameter estimates?
- Plot the parameter draws for the elements of β. Comment on the convergence of your simulation algorithm in both cases. Do you need to trim these draws? Do you need to 'burn-in' part of the draws?
- How does the choice of λ effect your results?

References

Park, Trevor, and Casella, George (2008). The Bayesian LASSO. *Journal of the American Statistical Association* **103**, 681-686.

3 Independence Analysis in a Markov Chain Model

Markov Chain models are often used in practice to model discrete events that occur consecutively. Examples include systematic failures in loan payments and occurrence of economic recessions.
In this exercise, we consider simple first order Markov Chain models with two states. Such a Markov chain model is defined by its transition probabilities

$$\mathbb{P}(y_t = j \mid y_{t-1} = i) = p_{ij}, \tag{0.1}$$

for states $i \in \{1, 2\}$ and probabilities $p_{ij} \in (0, 1)$, Note that the model has two free parameters since $p_{12} = 1 - p_{11}$ and $p_{22} = 1 - p_{21}$. You will simulate data from such a model, estimate this model using the Metropolis Hastings algorithm, and assess the performance of the Metropolis Hastings algorithm that you define.

- Simulate $T = 1000$ observations from a first order Markov Chain model. (*Note: Group members need to set different random seeds and different* p_{ij}.)
- Write down the log-likelihood for the first order Markov chain in (0.1).
- Define a flat prior over the probabilities. Again, recall that the model has two free parameters since $p_{12} = 1 - p_{11}$ and $p_{22} = 1 - p_{21}$.
- Apply a Metropolis Hastings algorithm to obtain posterior draws from p_{ij}. Clearly explain the candidate distribution you choose. Do you need to trim these draws? Do you need to 'burn-in' part of the draws? Report the posterior results.
- Apply an Importance Sampling algorithm to obtain posterior draws from p_{ij}. Clearly explain the candidate distribution you choose. Do you need to trim these draws? Do you need to 'burn-in' part of the draws? Report the posterior results.
- Compare the results in the last two parts. Are the two algorithms leading to different results? Did you expect the results to be similar?