Bayesian Statistics : Final Project

Instructor : Anna Simoni

December 11, 2023

The Final Project is the last piece you have to do to validate the course. Failure in returning the Final Project implies failure of the class.

Instructions :

1). Students have to work in **groups of 3 students** (same groups as for the Assignment 1 and Assignment 2). Only one document per group has to be returned.

2). The due date is **January 22, 2024** (assignment returned after this day will not be considered and the whole project will not be validated).

3). Each group has to upload the report on `https://app.compilatio.net/v5/document-submission/515-6EA-6D7` in the pdf format (no other format will be considered).

4). The pdf has to be named in the following way : `FAMILYNAME1_FAMILYNAME2_FAMILYNAME3_Ass_1`.

5). The report has to be typed (not handwritten).

6). Please provide the code that you have used. Since the code for the Gibbs sampling is the same as the one you have written for Assignment 2, you do not have to provide this part of the code.

You have to find a dataset with many covariates (large dimension) or you can choose among the 4 datasets described below (three of them are available on Pamplemousse but can be download as additional material from the indicated papers).

You consider the same model that you have used for Assignments 1 and 2, that is : for $t = 1, \dots, T$,

$$y_t = u_t'\phi + x_t'\beta + \varepsilon_t, \qquad \varepsilon_t \sim^{i.i.d.} \mathcal{N}(0, \sigma^2), \tag{0.1}$$

where x_t is a vector of regressors of dimension k and u_t is a vector of regressors of dimension ℓ. For simplicity, the variance of each regressor is normalized to 1.

You use the same prior specification for $\theta := (\sigma^2, \phi', \beta')' \in \mathbb{R}^{1+\ell+k}$ as in the Assignments 1 and 2 : $\pi(\theta) = \pi(\sigma^2) \times \pi(\phi) \times \pi(\beta|\sigma^2, \gamma^2, q)$ with

$$\begin{aligned}
\pi(\sigma^2) &\propto \frac{1}{\sigma^2} \\
\pi(\phi) &= \text{flat} \\
\pi(\beta_i|\sigma^2, \gamma^2, q) &\sim^{i.i.d.} \begin{cases} \mathcal{N}(0, \sigma^2\gamma^2) & \text{with probability } q \\ 0 & \text{with probability } 1-q \end{cases} \qquad \text{for } i = 1, \dots, k
\end{aligned}$$

and γ^2, q are hyperparameters taking respectively nonnegative values and values in $[0, 1]$.

In addition, we specify the prior for the hyperparameters q and γ^2 in the following way. First, we specify the mapping $(\gamma^2, q) \mapsto R^2(\gamma^2, q) := \frac{qk\gamma^2\overline{v}_x}{qk\gamma^2\overline{v}_x+1}$, where $\overline{v}_x := \frac{1}{k}\sum_{j=1}^k \widehat{Var(x_{t,j})}$ is the average sample variance of the predictors. Then, the prior for q and R^2 is

$$\begin{aligned}
\pi(q) &\sim Beta(a, b) \\
\pi(R^2) &\sim Beta(A, B).
\end{aligned}$$

You have to analyse the dataset by using the regression model (0.1). You are free to analyse the questions that you prefer. Examples of analysis are the following : (i) you can explore the posterior distribution in relation to the sparsity structure ; (ii) you can focus on inference (test and credible set) ; (iii) you can consider other aspects related to what we have seen in class.

For these analysis, you will use the prior specification above and the Gibbs sampling you have implemented in Assignment 2. A good project is a project that :

- clearly states the question you wants to analyse ;
- clearly states how you analyse it ;

— correctly implements and describes the methodology used

— provides interpretation and comments of results.

A good project does not have to be long but it has to be written in a clear way with a focus on the main points and main message of your analysis. It is a short project. So, **you are not expected to return a long report**. A suitable length is about 3-5 pages with fontsize not smaller than 12.

Description of the available datasets

Dataset 1 (Macro1) : This is a popular large data set used in the literature for macroeconomic predictions originally developed by Stock and Watson [2002a,b]. With this dataset, in your project you can for instance predict the *monthly growth rate of U.S. industrial production.*

The data set consists of 122 possible predictors, including various monthly macroeconomic indicators, such as measures of output, income, consumption, orders, surveys, labor market variables, house prices, consumer and producer prices, money, credit, and asset prices. The sample ranges from February 1961 to February 2019, and all the data have been transformed to obtain stationarity, as in the work of Stock and Watson, the outliers and missing values have been eliminated or imputed.

Dataset 2 (Macro2) : this dataset has been used for instance in Belloni et al. [2011, Section 7.2], to explain the average growth rate of GDP between 1960 and 1985 across countries. It comes from the data set constructed by Barro and Lee (1994) to find an answer to the debate on the cross-country determinants of long-term economic growth initiated by Barro (1991). The database includes data for 138 countries and 60 potential predictors, corresponding to the pre-1960 value of several measures of socio-economic, institutional, and geographical characteristics. The data can also be found on `https://rdrr.io/cran/hdm/man/Growth-Data.html`.

Dataset 3 (Micro1) : This dataset is used in the article Belloni et al. [2013] to study the impact of abortion on crime rates. It is an extension of the original dataset by Donohue and Levitt [2001]. Explanations about this dataset as well as the way the original dataset was enriched can be found in Belloni et al. [2013, Section 6]. The original dataset is available as supplementary data of the paper. You can either use the original dataset and enrich it as described in Belloni et al. [2013] or use the dataset provided (which contains 298 variables).

Dataset 4 (Micro2) : This dataset allows us to investigate the economic impact of eminent domain, that is, the right of the U.S. government to expropriate private property for public use. This dataset has been used in Belloni et al. [2012, Section 7]. You can look at this paper for additional explanations about the dataset. These data can be obtained with the R command `data(EminentDomain)` of the library ”library(hdm)”.

Références

A. Belloni, V. Chernozhukov, and C. Hansen. Inference for high-dimensional sparse econometric models. *arXiv preprint arXiv :1201.0220*, 2011.

A. Belloni, D. Chen, V. Chernozhukov, and C. Hansen. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6) :2369–2429, 2012.

A. Belloni, V. Chernozhukov, and C. Hansen. Inference on Treatment Effects after Selection among High-Dimensional Controls†. *The Review of Economic Studies*, 81(2) :608–650, 11 2013.

I. Donohue, John J. and S. D. Levitt. The Impact of Legalized Abortion on Crime*. *The Quarterly Journal of Economics*, 116(2) :379–420, 2001.

J. H. Stock and M. W. Watson. Forecasting Using Principal Components From a Large Number of Predictors. *Journal of the American Statistical Association*, 97(460) :1167–1179, 2002a.

J. H. Stock and M. W. Watson. Macroeconomic Forecasting Using Diffusion Indexes. *Journal of Business & Economic Statistics*, 20(2) :147–162, 2002b.