Bayesian regression models in gretl

L. Pedini

Dipartimento di Scienze Economiche e Sociali (DiSES)
Univpm - Facoltà di Economia G.Fuà

8th gretl Conference
Gdańsk, Poland
Baytool introduces common Bayesian model estimations in gretl.

At the moment, the package provides:

- homoskedastic linear model;
- heteroskedastic linear model;
- Bayesian LASSO;
- Latent variable models (under development).
The package includes:

- easy and immediate syntax (script) and a GUI;
- flexibility in priors;
- parallelization via MPI;
- plots and diagnostics.

The idea behind and the final aim

Providing a Bayesian alternative to each frequentist estimation command (native or not)
Why going Bayesian?

Pros:
- “more elegant and rigorous” (Koop, 2003);
- parameters are random variables;
- priors add information and regularize;
- exploit Monte Carlo integration techniques.

Cons
- computationally more complex;
- sensitivity to prior choices;
- prior hyperparameters.
Most of the Bayesian interest lies in the posterior of a given parameter $\theta$, given data $\mathcal{D}$:

$$P(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)P(\theta)}{P(\mathcal{D})} \propto p(\mathcal{D}|\theta)P(\theta)$$

- $P(\theta)$, prior;
- $p(\mathcal{D}|\theta)$, likelihood;
- $P(\theta|\mathcal{D})$, posterior
An example: the linear model

\[ y_{n \times 1} = X_{n \times k} \beta_{k \times 1} + \varepsilon_{n \times 1}, \quad \varepsilon \sim N(0, \sigma^2 I_n) \]

The likelihood is given by,

\[
p(y|\beta, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left[-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)\right]
\]
Rewrite the likelihood as,

\[ p(y|\beta, \sigma^2) = \left( \frac{1}{2\pi} \right)^{n/2} \left\{ \frac{1}{\sigma} \exp \left[ -\frac{1}{2\sigma^2} (\beta - \hat{\beta})^T (X^T X)(\beta - \hat{\beta}) \right] \right\} \times \left\{ \left( \frac{1}{\sigma^2} \right)^{\hat{a}/2} \exp \left[ -\frac{\hat{a}s^2}{2\sigma^2} \right] \right\} \]

where:

\[ \hat{\beta} = (X^T X)^{-1}X^T y \]
\[ \hat{s}^2 = \frac{(y - X\hat{\beta})^T (y - X\hat{\beta})}{\hat{a}} \]
\[ \hat{a} = n - k \]
An obvious choice for the priors should follow a similar pattern: 
\( \beta | \sigma^2 \sim \text{MVN}; \sigma^2 \sim \text{IG} \)

**Conjugate**

\[
P(\beta, \sigma^2) = P(\beta | \sigma^2)P(\sigma^2)
\]

\[
\Downarrow
\]

\[
\beta, \sigma^2 | y \sim \text{NIG}
\]

\[
\beta | y \sim \text{MV} t
\]

\[
\sigma^2 | y \sim \text{IG}
\]

**Independent**

\[
P(\beta, \sigma^2) = P(\beta)P(\sigma^2)
\]

\[
\Downarrow
\]

\[
\beta, \sigma^2 | y \sim ???
\]

\[
\Downarrow
\]

**Gibbs sampler**

\[
\beta | \sigma^2, y \sim \text{MVN}
\]

\[
\sigma^2 | \beta, y \sim \text{IG}
\]
House price example (Koop, 2003)

Anglin and Gencay (1996) dataset on $n = 546$ house sold in Canada in 1987

- $y$, house sales price
- $x_1$, lot size
- $x_2$, number of bedrooms
- $x_3$, number of bathrooms
- $x_4$, number of storeys
bundle ret = BT_LinMod(series y, list X, bundle prior, bundle opt)

Dependent variable: y
Prior specification: Independent NIG prior
Estimation method: Gibbs Sampler - 10000 iter (1000 burn-in)

Summary statistics:

<table>
<thead>
<tr>
<th></th>
<th>Prior_m</th>
<th>Prior_se</th>
<th>NI-Post_m</th>
<th>NI-Post_se</th>
<th>I-Post_m</th>
<th>I-Post_se</th>
</tr>
</thead>
<tbody>
<tr>
<td>const</td>
<td>0.00</td>
<td>10000.00</td>
<td>-4009.55</td>
<td>3609.79</td>
<td>-4131.25</td>
<td>3269.36</td>
</tr>
<tr>
<td>x1</td>
<td>10.00</td>
<td>5.00</td>
<td>5.42</td>
<td>0.37</td>
<td>5.45</td>
<td>0.37</td>
</tr>
<tr>
<td>x2</td>
<td>5000.00</td>
<td>2500.00</td>
<td>2824.61</td>
<td>1217.06</td>
<td>3224.68</td>
<td>1068.05</td>
</tr>
<tr>
<td>x3</td>
<td>10000.00</td>
<td>5000.00</td>
<td>17105.17</td>
<td>1737.65</td>
<td>16123.06</td>
<td>1624.84</td>
</tr>
<tr>
<td>x4</td>
<td>10000.00</td>
<td>5000.00</td>
<td>7634.90</td>
<td>1009.84</td>
<td>7691.85</td>
<td>974.03</td>
</tr>
<tr>
<td>Sig2</td>
<td>4e+07</td>
<td>6e+07</td>
<td>3e+08</td>
<td>2e+07</td>
<td>3e+08</td>
<td>2e+07</td>
</tr>
</tbody>
</table>
Introduction

Some background: the linear model

Shrinkage

Diagnostics

Conclusion

References
Suppose to have a sample of $n = 300$ observations ($n_0 = 200$ train and $n_1 = 100$ test) of a given $y$ and $X$, with $k = 300$.

$$DGP: y = Z\beta + \varepsilon, \quad Z = \{x_1, \ldots, x_{75}\}$$

Under these circumstances common OLS fails when using $X$

- $\ell_1$ norm - LASSO (Tibshirani, 1996)
- $\ell_2$ norm - ridge
Ridge regression

**Frequentist framework:**

\[
\min_{\beta} \left( \frac{1}{n_0} \| y - X\beta \|_2^2 + \lambda \sum_{j=1}^{k} \beta_j^2 \right) \rightarrow \hat{\beta}_r = (X^TX + \lambda I)^{-1}X^Ty
\]

**Bayesian framework:** introducing regularization is quite natural; priors deal the job.

\[
\beta | \sigma^2 \sim \text{MVN}(0, \sigma^2 \gamma I) \rightarrow E(\beta | y) = \left( X^TX + \frac{1}{\gamma} I \right)^{-1} X^Ty
\]
LASSO

**Frequentist framework** (Tibshirani, 1996):

\[
\min_{\beta} \left( \frac{1}{n_0} \| y - X\beta \|_2^2 + \lambda \sum_{j=1}^k |\beta_j| \right)
\]

**Bayesian framework** (Park and Casella, 2008):

\[
\beta | \sigma^2, \tau_1^2, \ldots, \tau_k^2 \sim \text{MVN}(0, \sigma^2 D), \quad D = \text{diag}\{\tau_1^2, \ldots, \tau_k^2\}
\]

\[
P(\tau_1^2, \ldots, \tau_k^2) = \prod_{j=1}^k \frac{\lambda^2}{2} \exp \left( -\frac{\lambda^2 \tau_j^2}{2} \right)
\]
### Comparing predictions on test set

<table>
<thead>
<tr>
<th></th>
<th>MSE($\hat{y}$)</th>
<th>CPU (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ridge (cv)</td>
<td>10.45</td>
<td>0.42</td>
</tr>
<tr>
<td>Bay. ridge ($\gamma = 0.01$)</td>
<td>20.55</td>
<td>81.81</td>
</tr>
<tr>
<td>Bay. ridge ($\gamma = 0.05$)</td>
<td>11.17</td>
<td>79.19</td>
</tr>
<tr>
<td>Bay. ridge ($\gamma = 0.10$)</td>
<td>9.49</td>
<td>85.76</td>
</tr>
<tr>
<td>Bay. ridge ($\gamma = 1$)</td>
<td>9.50</td>
<td>78.86</td>
</tr>
<tr>
<td>Bay. ridge ($\gamma = 10$)</td>
<td>9.45</td>
<td>80.18</td>
</tr>
<tr>
<td>LASSO (cv)</td>
<td>3.37</td>
<td>5.56</td>
</tr>
<tr>
<td>Bay. LASSO (hyper)</td>
<td>4.80</td>
<td>77.74</td>
</tr>
</tbody>
</table>

Notes: cv = cross-validation with 10 folds, 50 $\lambda$ sequences; hyper = hyperprior set-up. Posterior mean of sampled predictions used in Bayesian experiments with 10000 Gibbs iterations with 1000 burn-in (MPI with 4 threads).
In the Bayesian context deriving inference tools is straightforward.
MCMC diagnostics

- autocorrelation/mixing “raw” indicators
  - time series plot, ACF;
  - effective sample size (Vats et al., 2019);
  - numerical standard errors.

- more “formal” testing
  - Heidelberger and Welch (1983);
  - Geweke (1992);
Autocorrelation plots

Diagnostic plots from previous Bayesian LASSO example
Conclusion

- Baytool includes common Bayesian alternatives to frequentist commands;

- this becomes very useful when the frequentist alternatives are complex or lack from proper inferential tools;

- computational burden is dealt with MPI;

- Diagnostic tool available for MCMC cases.
References


