# Bayesian regression models in `gretl`

L. Pedini

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

$8^{th}$ `gretl` Conference
Gdańsk, Poland

# The package `BayTool`

`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

# Posterior

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

$$\underset{n\times 1}{y} = \underset{n\times k}{X}\beta + \underset{n\times 1}{\varepsilon}, \qquad \varepsilon \sim \mathrm{N}(0, \sigma^2 \mathrm{I}_n)$$

The likelihood is given by,

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

# Priors

Rewrite the likelihood as,

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

where:

$$\hat{\beta} = (X^TX)^{-1}X^Ty$$
$$\hat{s}^2 = \frac{(y-X\hat{\beta})^T(y-X\hat{\beta})}{\hat{a}}, \quad \hat{a} = n-k$$

An obvious choice for the priors should follow a similar pattern: $\beta|\sigma^2 \sim \mathrm{MVN}; \sigma^2 \sim \mathrm{IG}$

**Conjugate**

$$P(\beta, \sigma^2) = P(\beta|\sigma^2)P(\sigma^2)$$

$$\Downarrow$$

$$\beta, \sigma^2|y \sim \mathrm{NIG}$$

$$\beta|y \sim \mathrm{MV}t$$

$$\sigma^2|y \sim \mathrm{IG}$$

**Independent**

$$P(\beta, \sigma^2) = P(\beta)P(\sigma^2)$$

$$\Downarrow$$

$$\beta, \sigma^2|y \sim ???$$

$$\Downarrow$$

**Gibbs sampler**

$$\beta|\sigma^2, y \sim \mathrm{MVN}$$

$$\sigma^2|\beta, y \sim \mathrm{IG}$$

# House price example (Koop, 2003)

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

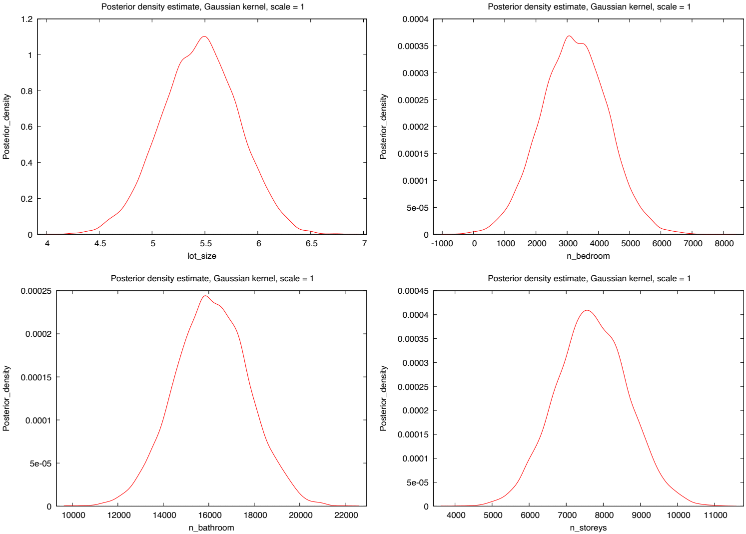
- $y$, house sales price
- $x_1$, lot size
- $x_2$, number of bedrooms
- $x_3$, number of bathrooms
- $x_4$, number of storeys

```
-------------------------------------------------------------------
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:
        Prior_m     Prior_se   NI-Post_m  NI-Post_se    I-Post_m    I-Post_se
const      0.00     10000.00    -4009.55     3609.79    -4131.25      3269.36
x1        10.00         5.00        5.42        0.37        5.45         0.37
x2      5000.00      2500.00     2824.61     1217.06     3224.68      1068.05
x3     10000.00      5000.00    17105.17     1737.65    16123.06      1624.84
x4     10000.00      5000.00     7634.90     1009.84     7691.85       974.03

Sig2      4e+07        6e+07       3e+08       2e+07       3e+08        2e+07
```


Posterior density estimate, Gaussian kernel, scale = 1
Posterior_density
lot_size
Posterior density estimate, Gaussian kernel, scale = 1
Posterior_density
n_bedroom
Posterior density estimate, Gaussian kernel, scale = 1
Posterior_density
n_bathroom
Posterior density estimate, Gaussian kernel, scale = 1
Posterior_density
n_storeys

# Shrinkage regression

Suppose to have a sample of $n = 300$ observations ($n_0 = 200$ train and $n_1 = 100$ test) of a given $\underset{n\times 1}{y}$ and $\underset{n\times k}{X}$, with $k = 300$.

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

**Under these circumstances common OLS fails when using $X$**

- $\ell 1$ norm - LASSO (Tibshirani, 1996)
- $\ell 2$ norm - ridge

# Ridge regression

**Frequentist framework**:

$$\min_{\beta}\left(\frac{1}{n_0}||y - X\beta||_2^2 + \lambda\sum_{j=1}^{k}\beta_j^2\right) \to \hat{\beta}_r = (X^\top X + \lambda I)^{-1}X^\top y$$

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

$$\beta|\sigma^2 \sim \mathrm{MVN}(0, \sigma^2\gamma I) \to \mathrm{E}(\beta|y) = \left(X^\top X + \frac{1}{\gamma}I\right)^{-1} X^\top y$$

# 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):

$$\beta|\sigma^2, \tau_1^2, \dots, \tau_k^2 \sim \mathrm{MVN}(0, \sigma^2 D), \quad D = diag\{\tau_1^2, \dots, \tau_k^2\}$$

$$P(\tau_1^2, \dots, \tau_k^2) = \prod_{j=1}^{k} \frac{\lambda^2}{2} \exp\left(-\lambda^2 \tau_j^2 / 2\right)$$

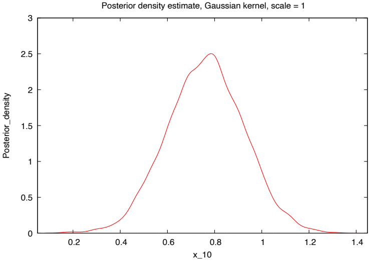
# Comparing predictions on test set

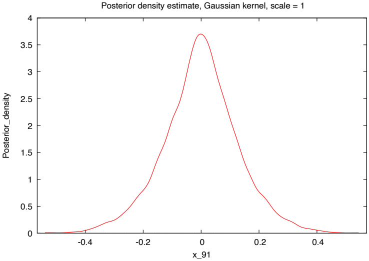
| | MSE($\hat{y}$) | CPU (sec) |
|---|---|---|
| ridge (cv) | 10.45 | 0.42 |
| Bay. ridge ($\gamma = 0.01$) | 20.55 | 81.81 |
| Bay. ridge ($\gamma = 0.05$) | 11.17 | 79.19 |
| Bay. ridge ($\gamma = 0.10$) | 9.49 | 85.76 |
| Bay. ridge ($\gamma = 1$) | 9.50 | 78.86 |
| Bay. ridge ($\gamma = 10$) | 9.45 | 80.18 |
| LASSO (cv) | 3.37 | 5.56 |
| Bay. LASSO (hyper) | 4.80 | 77.74 |

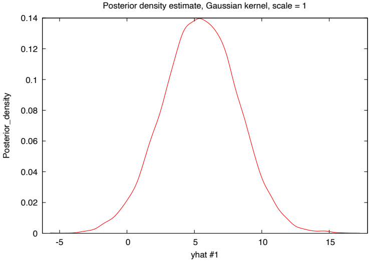
Notes : cv = cross-validation with 10 folds, 50 $\lambda$ sequences; hyper = hyperprior set-up. Posterior mean of sampled predictions used in Bayesian experiments with 10000 Gibbs iterations with 1000 burn-in (MPI with 4 threads).

# Inference

In the Bayesian context deriving inference tools is straightforward.



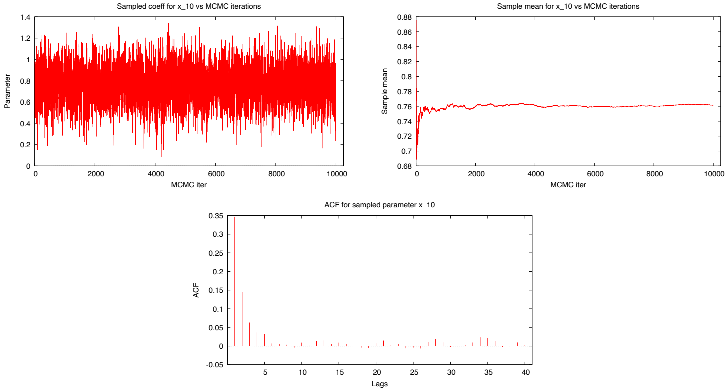



# MCMC diagnostics

- 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);
  - Gelman and Rubin (1992); Brooks and Gelman (1998).

# 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

Anglin, P. M. and Gencay, R. (1996). Semiparametric estimation of a hedonic price function. *Journal of applied econometrics*, 11(6):633–648.

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, pages 457–472.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian Statistics (J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith eds.)*, volume 4, pages 169–193. Clarendon Press, Oxford, UK.

Heidelberger, P. and Welch, P. D. (1983). Simulation run length control in the presence of an initial transient. *Operations Research*, 31(6):1109–1144.

Koop, G. (2003). *Bayesian Econometrics*. John Wiley & Sons Ltd.

Park, T. and Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Vats, D., Flegal, J. M., and Jones, G. L. (2019). Multivariate output analysis for markov chain monte carlo. *Biometrika*, 106(2):321–337.