The Illusion of the *Illusion of Sparsity*[2]

Bruno Fava[1] Hedibert F. Lopes[2]

[1]Northwestern University, Illinois, USA

[2]Professor of Statistics and Econometrics
Head of the Center of Statistics, Data Science and Decision
INSPER, São Paulo, Brazil

August/September 2020

[2]Giannone, Lenza and Primiceri (2020) *Economic predictions with big data: the illusion of sparsity*. Our manuscript and these slides can be found in my page at `hedibert.org`

Outline

Motivation

Sparsity in static regressions

- Ridge and lasso regressions
- Spike and slab model (or SMN model)
- SSVS and scaled SSVS priors
- Other mixture priors
- Toy example: `R` package `Bayeslm`

Revisiting GLP

- The sparse-inducing linear model
- Their findings
- An important drawback

Experiments

- I. Adding meaningless variables
- II. Fatter tails via Student's t
- III. A simulation exercise

Outline

Motivation

Sparsity in static regressions

- Ridge and lasso regressions
- Spike and slab model (or SMN model)
- SSVS and scaled SSVS priors
- Other mixture priors
- Toy example: `R` package `Bayeslm`

Revisiting GLP

- The sparse-inducing linear model
- Their findings
- An important drawback

Experiments

- I. Adding meaningless variables
- II. Fatter tails via Student's t
- III. A simulation exercise

Sparsity in Economics

We revisit the paper *Economic predictions with big data: the illusion of sparsity* by Giannone, Lenza and Primiceri, whose July 2020 abstract says:

> *We compare sparse and dense representations of predictive models in macroeconomics, microeconomics and finance. To deal with a large number of possible predictors, we specify a prior that allows for both variable selection and shrinkage. The posterior distribution does not typically concentrate on a single sparse model, but on a wide set of models that often include many predictors.*

They conclude the paper saying:

> *In economics, there is no theoretical argument suggesting that predictive models should in general include only a handful of predictors. As a consequence, the use of low-dimensional model representations can be justified only when supported by strong statistical evidence.*

They add that:

> *Empirical support for low-dimensional models is generally weak. Predictive model uncertainty seems too pervasive to be treated as statistically negligible. The right approach to scientific reporting is thus to assess and fully convey this uncertainty, rather than understating it through the use of dogmatic (prior) assumptions favoring low dimensional models.*

Our contribution

We proposes a revision of the methods adopted by Giannone, Lenza and Primiceri.

- We analyze the posterior distribution of the included coefficients of the linear model. This was not explored by Giannone, Lenza and Primiceri.
- We add bogus predictors and observe correct exclusion only in a subset of the data sets.
- We extend their analysis with Student's t prior for the regression coefficients. The heavier-tailed distribution was more restrictive in selecting possible predictors, and results once again corroborate with the thesis that the original Spike-and-Slab prior is unable to correctly allow and distinguish between shrinkage or sparsity.
- We developed a simulation exercise to check the performance of the original model and with the t-student modification in a totally controlled environment. Posterior inference reinforces the belief that their prior incorrectly induces shrinkage.

Overall conclusion: Their Spike-and-Slab approach does not seem to be robust, leading to the illusion that sparsity is nonexistent, when it might in fact exist.

Outline

Ridge and lasso regressions

Throughout, we consider the standard Gaussian linear model,

$$y_t = \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_q x_{qt} + \nu_t,$$

where RSS$= (y - X\beta)'(y - X\beta)$ is the residual sum of squares.

- *Ridge regression* Hoerl and Kennard [1970] - ℓ_2 penalty on β:

  $$\hat{\beta}_{ridge} = \arg\min_{\beta} \left\{ RSS + \lambda_r^2 \sum_{j=1}^{q} \beta_j^2 \right\}, \qquad \lambda_r^2 \geq 0,$$

  leading to $\hat{\beta}_{ridge} = (X'X + \lambda_r^2 I_q)^{-1} X'y$.

- *Lasso regression* Tibshirani [1996] - ℓ_1 penalty on β:

  $$\hat{\beta}_{lasso} = \arg\min_{\beta} \left\{ RSS + \lambda_l \sum_{j=1}^{q} |\beta_j| \right\}, \qquad \lambda_l \geq 0,$$

  which can be solved by a *coordinate gradient descent* algorithm.

Ridge and lasso estimates are posterior modes!

The posterior mode or the maximum a posteriori (MAP) is given by

$$\tilde{\beta}_{\text{mode}} = \arg\min_{\beta}\{-2\log p(y|\beta) - 2\log p(\beta)\}$$

The $\hat{\beta}_{ridge}$ estimate equals the posterior mode of the normal linear model with

$$p(\beta_j) \propto \exp\{-0.5\lambda_r^2\beta_j^2\},$$

which is a Gaussian distribution with location 0 and scale $1/\lambda_r^2$, $N(0, 1/\lambda_r^2)$.
The mean is 0, the variance is $1/\lambda_r^2$ and the excess kurtosis is 0.

The $\hat{\beta}_{lasso}$ estimate equals the posterior mode of the normal linear model with

$$p(\beta_j) \propto \exp\{-0.5\lambda_l|\beta_j|\},$$

which is a Laplace distribution with location 0 and scale $2/\lambda_l$, Laplace$(0, 2/\lambda_l)$.
The mean is 0, the variance is $8/\lambda_l^2$ and excess kurtosis is 3.

Spike and slab model (or scale mixture of normals)

Ishwaran and Rao [2005] define a spike and slab model as a Bayesian model specified by the following prior hierarchy:

$$\begin{array}{rcll}
(y_t|x_t,\beta,\sigma^2) & \sim & N(x_t'\beta,\sigma^2), & t=1,\ldots,n \\
(\beta|\psi) & \sim & N(0,\mathrm{diag}(\psi)) & \\
\psi & \sim & \pi(d\psi) & \\
\sigma^2 & \sim & \mu(d\sigma^2) &
\end{array}$$

They go to say that

> *"Lempers [1988] and Mitchell and Beauchamp [1988] were among the earliest to pioneer the spike and slab method. The expression 'spike and slab' referred to the prior for β used in their hierarchical formulation."*

Spike and slab model (or scale mixture of normals model)

Regularization and variable selection are done by assuming independent prior distributions from the SMN class to each coefficient β_j:

$$\beta_j|\psi_j \sim N(0,\psi_j) \qquad \text{and} \qquad \psi_j \sim p(\psi_j)$$

so

$$p(\beta_j) = \int p(\beta_j|\psi_j)p(\psi_j)d\psi_j.$$

Mixing density $p(\psi_j)$	Marginal density $p(\beta_j)$	$V(\beta_j)$	Ex.kurtosis(β_j)
$\psi_j = 1/\lambda_r^2$	$N(0,1/\lambda_r^2)$ - (ridge)	$1/\lambda_r^2$	0
$IG(\eta/2,\eta\tau^2/2)$	$t_\eta(0,\tau^2)$	$\eta/(\eta-2)\tau^2$	$6/(\eta-4)$
$G(1,\lambda_l^2/8)$	Laplace$(0,2/\lambda_l)$ - (blasso)	$8/\lambda_l^2$	3
$G(\zeta,1/(2\gamma^2))$	$NG(\zeta,\gamma^2)$	$2\zeta\gamma^2$	$3/\zeta$

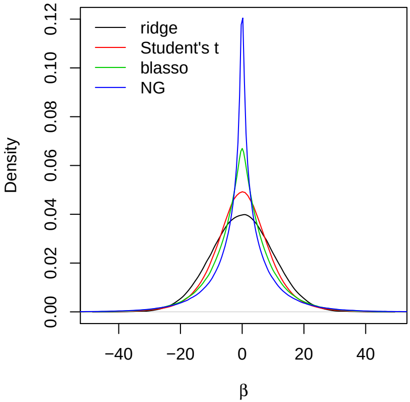
Griffin and Brown [2010] Normal-Gamma prior:

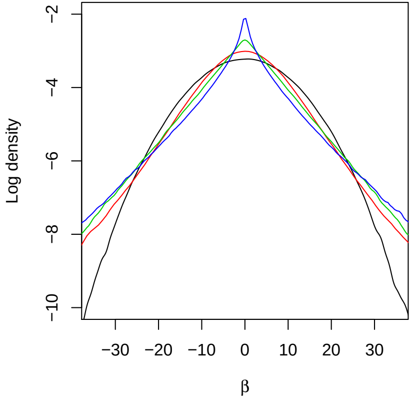
$$p(\beta|\zeta,\gamma^2) = \frac{1}{\sqrt{\pi}2^{\zeta-1/2}\gamma^{\zeta+1/2}\Gamma(\zeta)}|\beta|^{\zeta-1/2}K_{\zeta-1/2}(|\beta|/\gamma),$$

where K is the modified Bessel function of the 3rd kind.

Illustration

Ridge: $\lambda_r^2 = 0.01$ $\Rightarrow$ Excess kurtosis=0
Student's t: $\eta = 5$, $\tau^2 = 60$ $\Rightarrow$ Excess kurtosis=6
Blasso: $\lambda_l^2 = 0.08$ $\Rightarrow$ Excess kurtosis=3
NG: $\xi = 0.5$, $\gamma^2 = 100$ $\Rightarrow$ Excess kurtosis=6
All variances are equal to 100.





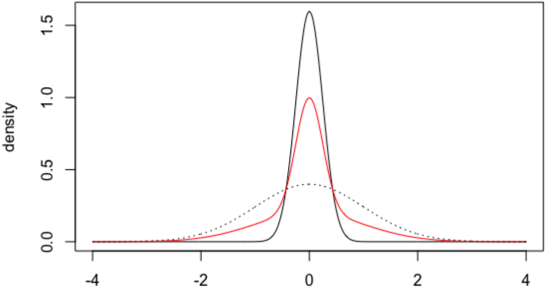
Stochastic search variable selection (SSVS) prior

SSVS George and McCulloch [1993]: For small $\tau > 0$ and $c >> 1$,

$$\beta|\omega, \tau^2, c^2 \sim (1-\omega) \underbrace{N(0, \tau^2)}_{spike} + \omega \underbrace{N(0, c^2\tau^2)}_{slab}.$$

SMN representation: $\beta|\psi \sim N(0, \psi)$ and

$$\psi|\omega, \tau^2, c^2 \sim (1-\omega)\delta_{\tau^2}(\psi) + \omega\delta_{c^2\tau^2}(\psi)$$

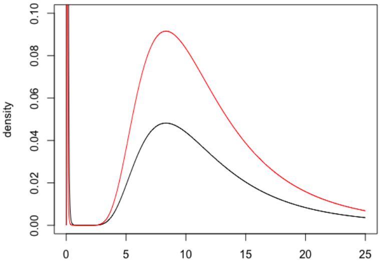


Scaled SSVS prior = normal mixture of IG prior

NMIG prior of Ishwaran and Rao [2005]: For $\upsilon_0 \ll \upsilon_1$,

$$\begin{aligned} \beta|K,\tau^2 &\sim N(0, K\tau^2), \\ K|\omega,\upsilon_0,\upsilon_1 &\sim (1-\omega)\delta_{\upsilon_0}(K) + \omega\delta_{\upsilon_1}(K), \\ \tau^2 &\sim IG(a_\tau, b_\tau). \end{aligned} \quad (1)$$

- Large ω implies non-negligible effects.
- The scale $\psi = K\tau^2 \sim (1-\omega)IG(a_\tau, \upsilon_0 b_\tau) + \omega IG(a_\tau, \upsilon_1 b_\tau)$.
- $p(\beta)$ is a two component mixture of scaled Student's t distributions.



Other mixture priors

Frühwirth-Schnatter and Wagner [2011]: absolutely continuous priors

$$\beta \sim (1-\omega)p_{spike}(\beta) + \omega p_{slab}(\beta), \tag{2}$$

Let $Q > 0$ a scale parameter and

$$r = \frac{\mathrm{Var}_{spike}(\beta)}{\mathrm{Var}_{slab}(\beta)} \ll 1,$$

then the mixing densities for ψ,

1. IG: $\psi \sim (1-\omega)IG(\nu, rQ) + \omega IG(\nu, Q)$,
2. Exp: $\psi \sim (1-\omega)Exp(1/2rQ) + \omega Exp(1/2Q)$,
3. Gamma: $\psi \sim (1-\omega)G(a, 1/2rQ) + \omega G(a, 1/2Q)$,

leads to the marginal densities for β,

1. Scaled-t: $\beta \sim (1-\omega)t_{2\nu}(0, rQ/\nu) + \omega t_{2\nu}(0, Q/\nu)$,
2. Laplace: $\beta \sim (1-\omega)Lap(\sqrt{rQ}) + \omega Lap(\sqrt{Q})$,
3. NG: $\beta \sim (1-\omega)NG(a, r, Q) + \omega NG(a, Q)$.

Inverted-Gamma prior for the variance of β

It is easy to see that, for a constant c,

$$\mathrm{Var}_{spike}(\beta) = cQr \quad \text{and} \quad \mathrm{Var}_{slab}(\beta) = cQ.$$

Therefore, when

$$v_\beta = \mathrm{Var}(\beta) = (1-\omega)\,\mathrm{Var}_{spike}(\beta) + \omega\,\mathrm{Var}_{slab}(\beta) \sim IG(c_0, C_0),$$

the implied distribution of Q is

$$Q \sim IG\left(c_0, \frac{C_0}{c((1-\omega)r+\omega)}\right).$$

Spike-and-slab priors:

Prior	Spike	Slab	$p(\beta)$	Constant c
SSVS	$\psi = rQ$	$\psi = Q$	$(1-\omega)N(0,rQ) + \omega N(0,Q)$	1
NMIG	$IG(\nu, rQ)$	$IG(\nu, Q)$	$(1-\omega)t_{2\nu}(0, rQ/\nu) + \omega t_{2\nu}(0, Q/\nu)$	$1/(\nu-1)$
Laplaces	$Exp(1/2rQ)$	$Exp(1/2Q)$	$(1-\omega)Lap(\sqrt{rQ}) + \omega Lap(\sqrt{Q})$	2
Normal-Gammas	$G(a, 1/2rQ)$	$G(a, 1/2Q)$	$(1-\omega)NG(\beta_j\|a, r, Q) + \omega NG(\beta_j\|a, Q)$	$2a$
Laplace-t	$Exp(1/2rQ)$	$IG(\nu, Q)$	$(1-\omega)Lap(\sqrt{rQ}) + \omega t_{2\nu}(0, Q/\nu)$	$c_1 = 2, c_2 = 1/(\nu-1)$

Toy example: `R` package `Bayeslm`

For observation $i = 1, \dots, n = 68$ and predictor $j = 1, \dots, k = 16$, we simulate

$$x_{ij} \sim N(0,1) \quad \text{and} \quad \varepsilon_i^* \sim N(0,1)$$

We also fix $\beta_1 = -0.86$, $\beta_2 = 0.64$ and $\beta_3 = 0.89$, while the response variable is:

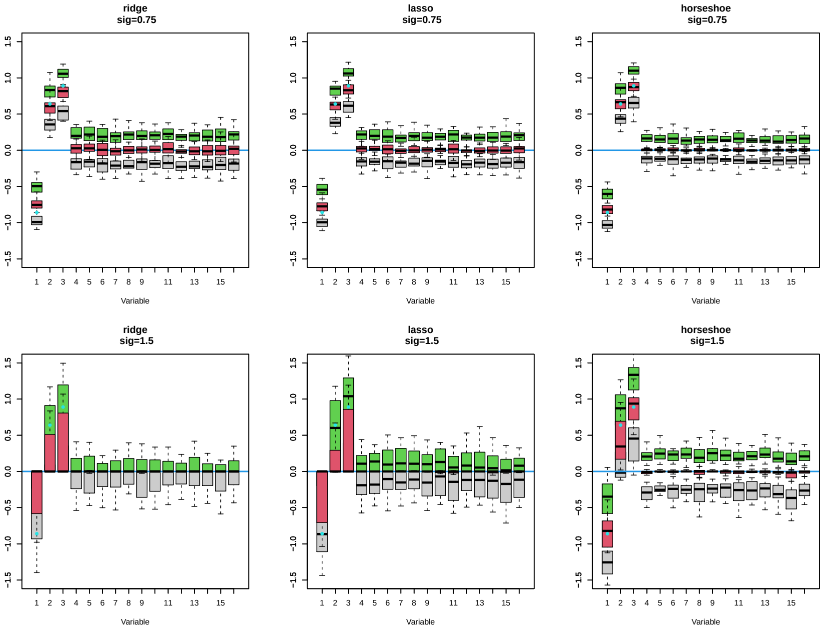
$$y_i^{(s)} = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \sigma_\varepsilon^{(s)} \varepsilon_i^*,$$

and $\sigma_\varepsilon^{(s)} = 0.75s$, for $s = 1, 2$.

MCMC set-up: $N = 2000$ draws, `burnin`$= 10000$ burn-in

Monte Carlo error: $R = 20$ replicates

Ridge, Laplace and horseshoe priors



Toy example: R script

```
install.packages("bayeslm");library("bayeslm")
n=68;k=16;betas=c(-0.86,0.64,0.89,rep(0,k-3));sigs=c(0.75,1.5)
N=2000;burnin=10000;R=20
qs=c(0.025,0.5,0.975)
J=length(sigs);quants=array(0,c(R,J,3,k,3))
set.seed(54321)
for (r in 1:R){
  for (j in 1:J){
    X = matrix(rnorm(n*k),n,k)
    y = rnorm(n,X%*%betas,sigs[j])
    fit.hs = bayeslm(y,x,prior='horseshoe',N=N,burnin=burnin,icept=FALSE)
    fit.ridge = bayeslm(y,x,prior='ridge',N=N,burnin=burnin,icept=FALSE)
    fit.lasso = bayeslm(y,x,prior='laplace',N=N,burnin=burnin,icept=FALSE)
    quants[r,j,1,,] = t(apply(fit.hs$beta,2,quantile,qs))
    quants[r,j,2,,] = t(apply(fit.ridge$beta,2,quantile,qs))
    quants[r,j,3,,] = t(apply(fit.lasso$beta,2,quantile,qs))
  }
}
method = c("horseshoe","ridge","lasso")
par(mfrow=c(2,3))
for (i in 1:2)
  for (j in c(2,3,1)){
    boxplot(quants[,i,j,,1],names=1:k,ylim=c(-1.5,1.5),outline=FALSE,col=gray(0.8),
            xlab="Variable",main=paste(method[j],"\n sig=",sigs[i],sep=""))
    abline(h=0,col=4,lwd=2)
    for (l in 3:2)
      boxplot(quants[,i,j,,l],names=rep("",k),outline=FALSE,col=l,add=TRUE)
    points(1:3,betas[1:3],col=5,pch=16)
  }
```

A few additional references

Park and Casella (2008) The Bayesian lasso. *JASA*, 103(482), 681-686.

Carvalho, Polson and Scott (2010) The horseshoe estimator for sparse signals. *Biometrika*, 97(2)465-480.

Polson and Scott (2010) Shrink globally, act locally: Sparse Bayesian regularization and prediction, *Bayesian Statistics*, Volume 9, 501–538.
Polson and Scott (2012) Local shrinkage rules, Lévy processes and regularized regression, *JRSS-B*, 74(2), 287-311.

van der Pas, Kleijn and van der Vaart (2014) The horseshoe estimator: Posterior concentration around nearly black vectors. *Electronic Journal of Statistics*, 8, 2585-2618.

Bhattacharya, Pati, Pillai and Dunson (2015) Dirichlet–Laplace priors for optimal shrinkage, *JASA*, 110, 1479–1490.

Makalic and Schmidt (2016) A Simple Sampler for the Horseshoe Estimator. IEEE Signal Processing Letters, 23(1), 179-182.

Bhadra, Datta, Polson and Willard (2017) The Horseshoe+ Estimator of Ultra-Sparse Signals, *Bayesian Analysis*, 12(4), 1105–1131.

Rocková and George (2018) The Spike-and-Slab LASSO, *JASA*, 113(521), 431-444.

Hahn, He and Lopes (2019) Efficient sampling for Gaussian linear regression with arbitrary priors, *JCGS*, 28, 142-154.

Outline

GLP spike-and-slab prior

Let y_t be the response variable and x_t the k-dimensional vector of potential explanatory variables. The Gaussian linear model is

$$y_t = x_t'\beta + \epsilon_t, \qquad \epsilon_t \sim N(0, \sigma^2)$$

The prior specification for σ^2 is

$$p(\sigma^2) \propto \frac{1}{\sigma^2}$$

and the prior for β_i is

$$\beta_i | \sigma^2, \gamma^2, q \sim \begin{cases} N(0, \sigma^2\gamma^2) & \text{with prob. } q \\ 0 & \text{with prob. } 1 - q \end{cases} \qquad i = 1, \ldots, k.$$

q governs the degree of sparsity.

γ governs the degree of shrinkage.

Hyperprior of (q, γ^2)

Instead of setting a hyperprior for (q, γ^2), GLP defined a prior for the pair (q, R^2), where

$$R^2(\gamma^2, q) \equiv \frac{qk\gamma^2}{qk\gamma^2 + 1},$$

is the coefficient of determination.

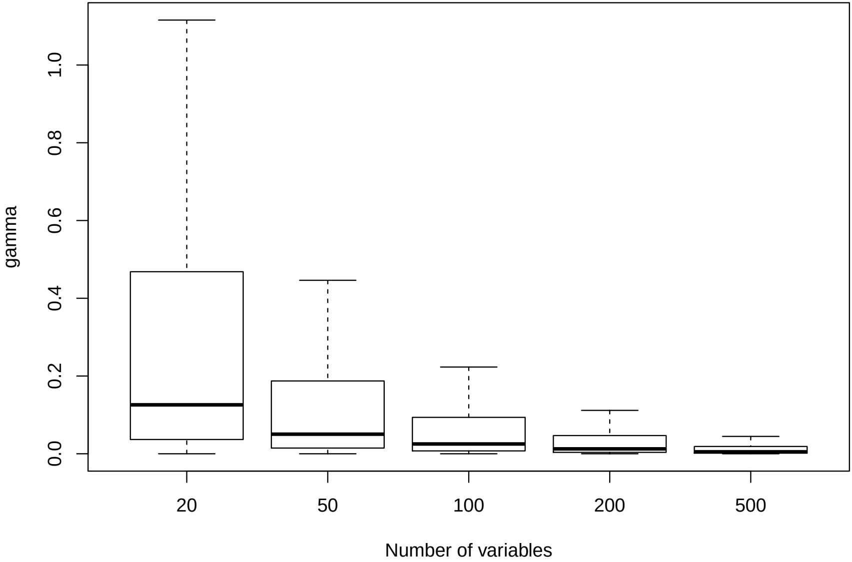
The hyperprior distributions are:

$$q \sim Beta(1, 1)$$

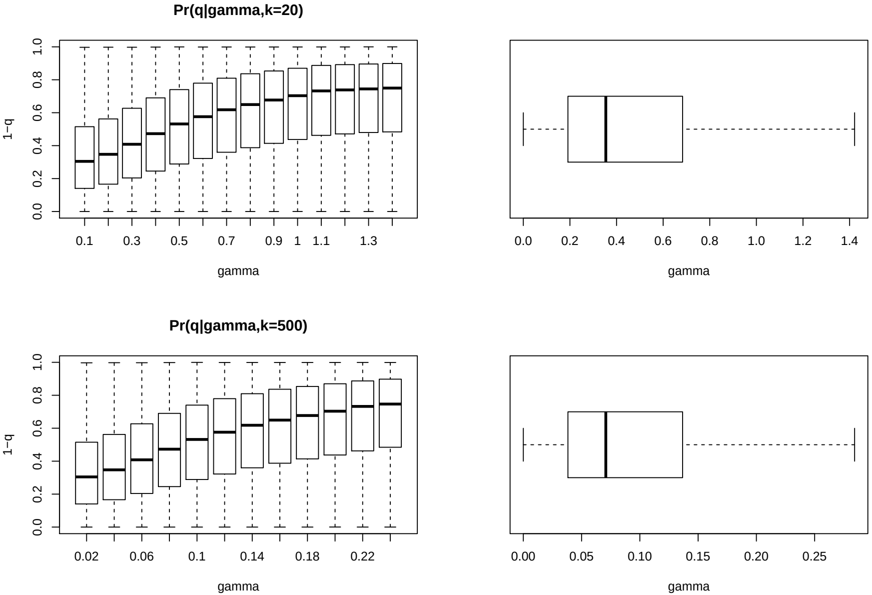
and

$$R^2 \sim Beta(1, 1)$$

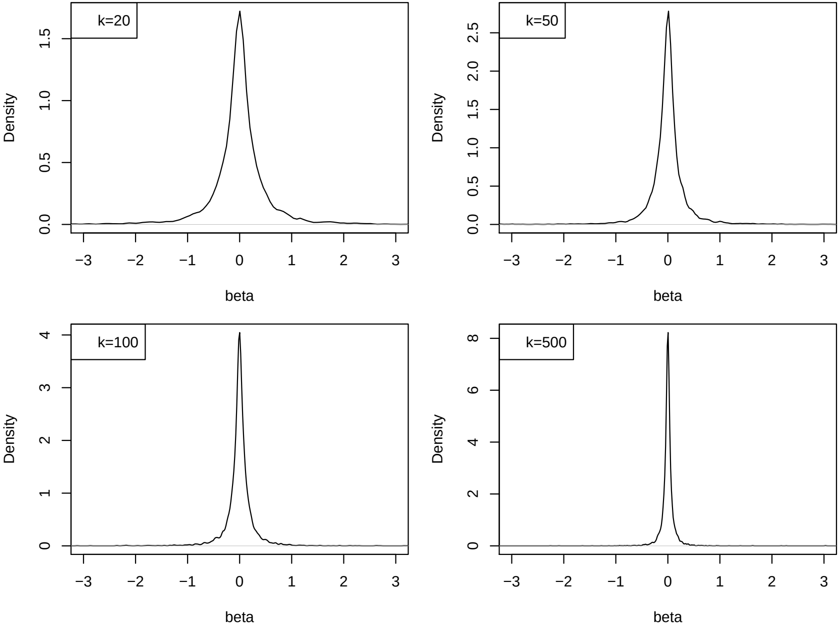
Marginal prior of γ: $p(\gamma|k)$



$p(1 - q|\gamma)$ and $p(\gamma)$



$p(\beta|k, \sigma = 1)$



GLP Macro and finance data sets

Table 1: Description of the datasets.

	Dependent variable	Possible predictors	Sample
Macro 1	Monthly growth rate of US industrial production	130 lagged macroeconomic indicators	659 monthly time-series observations, from February 1960 to December 2014
Macro 2	Average growth rate of GDP over the sample 1960-1985	60 socio-economic, institutional and geographical characteristics, measured at pre-60s value	90 cross-sectional country observations
Finance 1	US equity premium (S&P 500)	16 lagged financial and macroeconomic indicators	68 annual time-series observations, from 1948 to 2015
Finance 2	Stock returns of US firms	144 dummies classifying stock as very low, low, high or very high in terms of 36 lagged characteristics	1400k panel observations for an average of 2250 stocks over a span of 624 months, from July 1963 to June 2015

Source: [Giannone et al., 2020, p. 15]

GLP Micro data sets

Table 1: Description of the datasets.

	Dependent variable	Possible predictors	Sample
Micro 1	Per-capita crime (murder) rates	Effective abortion rate and 284 controls including possible covariate of crime and their transformations	576 panel observations for 48 US states over a span of 144 months, from January 1986 to December 1997
Micro 2	Number of pro-plaintiff eminent domain decisions in a specific circuit and in a specific year	Characteristics of judicial panels capturing aspects related to gender, race, religion, political affiliation, education and professional history of the judges, together with some interactions among the latter, for a total of 138 regressors	312 panel circuit/year observations, from 1975 to 2008

Source: [Giannone et al., 2020, p. 15]

Their main remarks

The conclusion is that a clear pattern of sparsity is found only on the Micro 1 data set, in which only one variable is included most of the times.

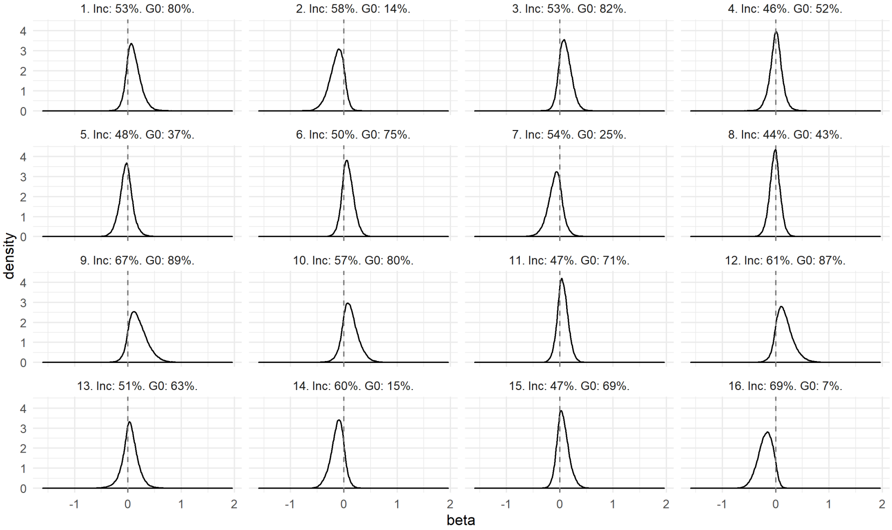
For all other data sets, one is incapable of determining which variables should be included, as many have a high estimated probability of inclusion $\Rightarrow$ dense models.

Their conclusion: Ssparsity cannot be assumed for any economic data set, unless in the presence of strong statistical evidence, and suggest an "illusion of sparsity" when using statistical models that assume (and force) sparsity.

An important drawback

Finance 1 data set -

Inc: Probability of inclusion. G0: Probability above zero.



A drawback of their approach

The spike-and-slab prior, as defined, seems to be inducing shrinkage by including predictors with a near-zero coefficient.

Example: β_5 and $(\beta_9, \beta_{12}, \beta_{16})$

- Probability of inclusion near 0.5, but also about 0.4/0.6 probability above/below zero.
- It could be, for example, that an economist trying to make inference on the regression would very easily exclude variable 5, but keep, for example, variables 9, 12 and 16.

Outline

I. Adding meaningless variables

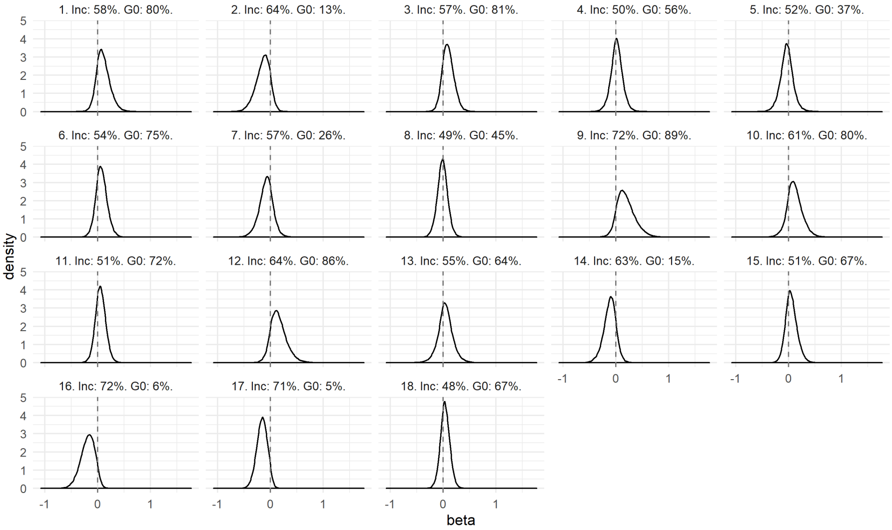
We re-run the estimation algorithm for all the five datasets but now include two additional regressors that were completely randomly generated.

Micro 1: 1.6% and 3.9%
Macro 1: 12.2% and 21.1%

Micro 2: 20.0% and 18.7%
Macro 2 56.1% and 55.2% (57th and 58th most included out of 62)
Finance 1: 71.0% and 48.4% (3rd and 18th most included out of 18)

I. Adding meaningless variables

Finance 1 data set ($n = 68$): Here x_{17} and x_{18} are meaningless.



Similar shapes: β_{18} and $(\beta_4, \beta_5, \beta_{15})$.
High inclusion: x_{17} included 71% of times.

II. Fatter tails via Student's t

New prior:

$$\beta_i|\sigma^2, \gamma^2, \lambda_i^2, q \sim \begin{cases} N(0, \sigma^2\gamma^2\lambda_i^2) & \text{with prob. } q \\ 0 & \text{with prob. } 1-q \end{cases} \qquad i = 1, \ldots, k \ ,$$

with an Inverse-Gamma prior for λ_i^2:

$$\lambda_i^2 \sim IG\left(\frac{\nu}{2}, \frac{\nu}{2}\right)$$

Therefore, β_i follows a Student's t distribution:

$$\beta_i|\sigma^2, \gamma^2, q \sim \begin{cases} t_\nu(0, \sigma^2\gamma^2) & \text{with prob. } q \\ 0 & \text{with prob. } 1-q \end{cases} \qquad i = 1, \ldots, k \ ,$$

where

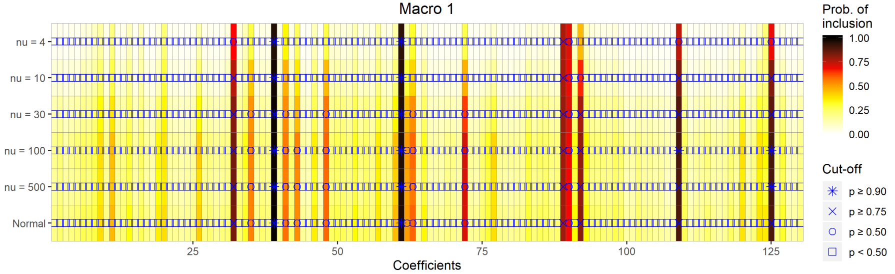
$$V(\beta_i|\sigma^2, \gamma^2, q) = \frac{\nu}{\nu-2}\sigma^2\gamma^2$$

II. Fatter tails via Student's t - Macro 1

x_{72} and x_{90} are both relevant for $\nu > 10$.

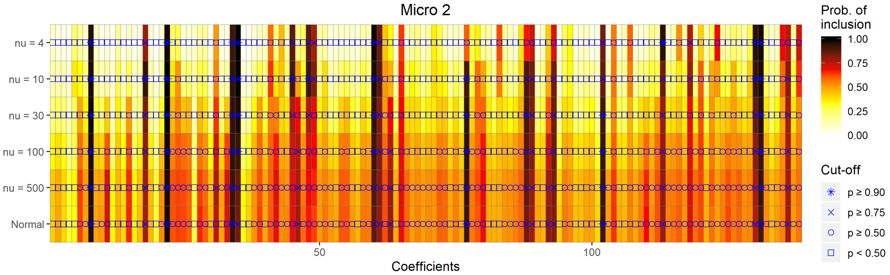
Only x_{90} for $\nu <= 10$ (sparsity reemerges).

Prob. inclusion $\downarrow$ as $\nu \uparrow$.



Argument: Spike-and-Slab, as originally defined, induces selection and shrinkage, since for $\nu = 4$ only 7 of 130 available predictors are relevant - that is, included more than 50% of the times.

II. Fatter tails via Student's t - Micro 2

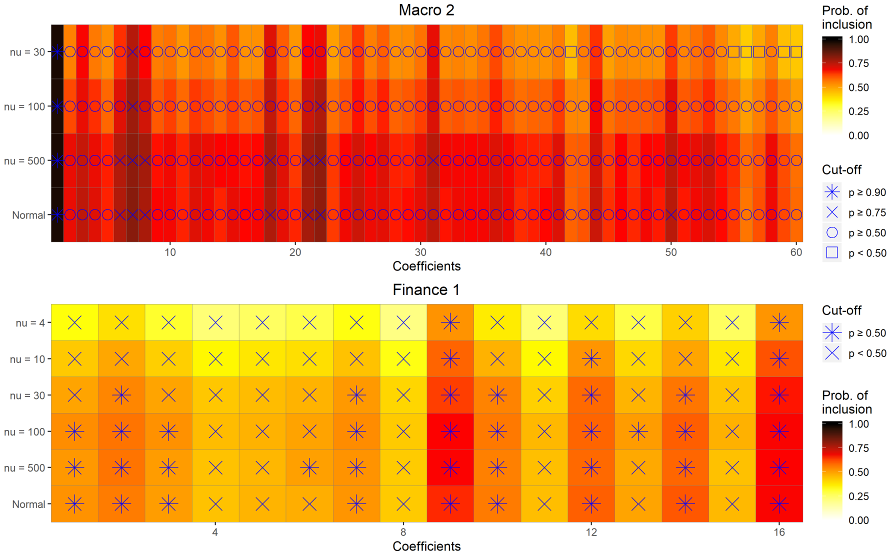


Gaussian: no pattern of variable selection.
106 of 138 predictors are selected more than 50% of the times.

Student's t: Sparsity in action.
For $\nu = 4$, only 30 predictors are selected.
For $\nu = 10$, only 34 predictors are selected.

II. Fatter tails via Student's t - Macro 2 & Finance 1

Similarity across ν



III. A simulation exercise

For observation $i = 1, \ldots, n = 68$ and predictor $j = 1, \ldots, k = 16$, we simulate

$$x_{ij} \sim N(0,1) \quad \text{and} \quad \varepsilon_i^* \sim N(0,1)$$

We also fix $\beta_1 = -0.86$, $\beta_2 = 0.64$ and $\beta_3 = 0.89$, while the response variable is:

$$y_i^{(s)} = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \sigma_\varepsilon^{(s)} \varepsilon_i^*,$$

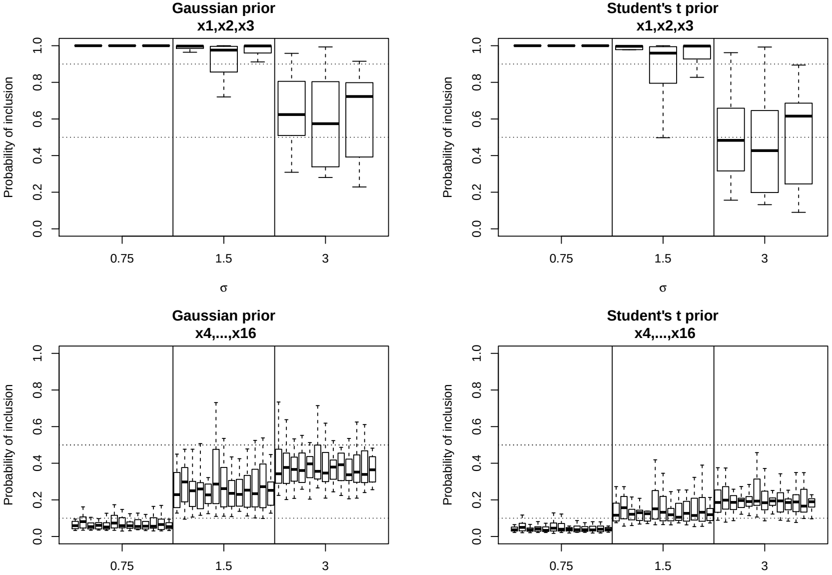
and $\sigma_\varepsilon^{(s)} = 0.75s$, for $s = 1, 2, 3$.

The prior for β are Gaussian or Student's t with $\nu = 4$ degrees of freedom.

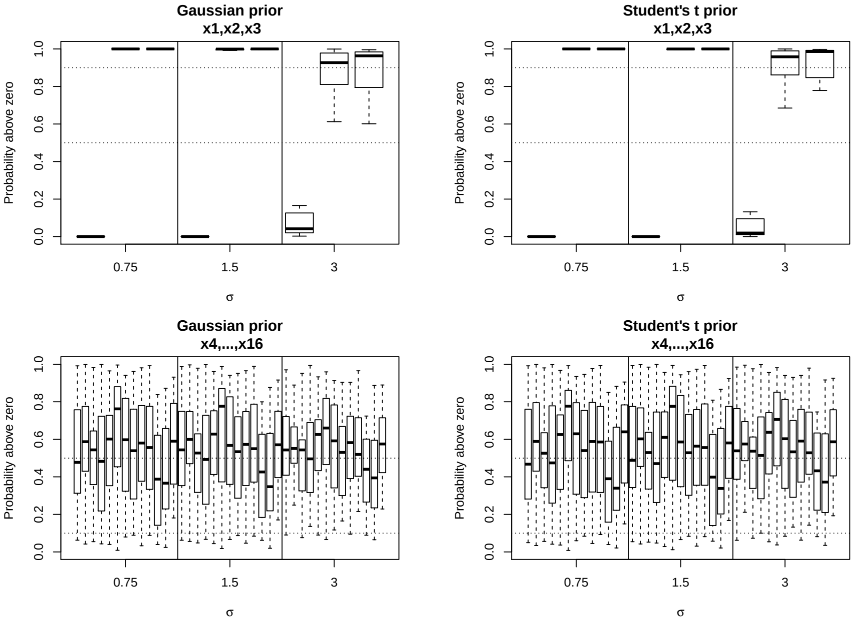
We replicate the above simulation $R = 20$ times.

III. Probability of inclusion

- $\sigma \uparrow$: inclusion of x_1, x_2, x_3 decreases. More so for the Student's t case.
- $\sigma \uparrow$: inclusion of $x_4, \ldots, x_{16}$ increases. More so for the Gaussian case.

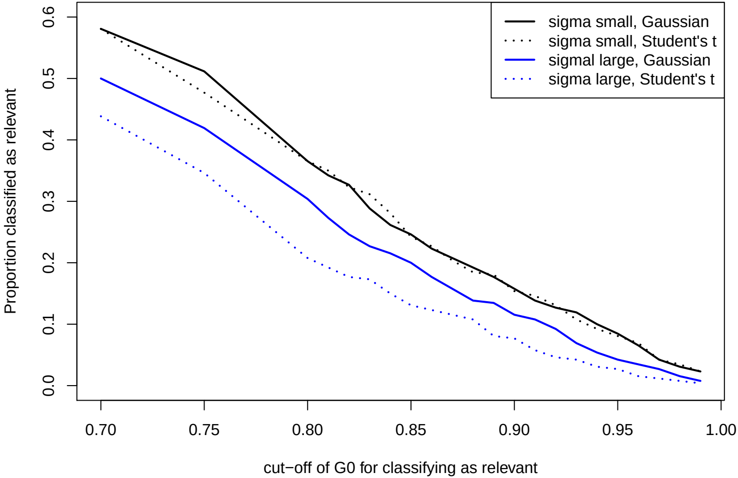


III. Probability above zero



III. Proportion of $\beta_4, \ldots, \beta_{16}$ classified as relevant

For σ large, Student's t prior performs better at shrinking towards zero.



References

Sylvia Frühwirth-Schnatter and Hedibert F. Lopes. Sparse Bayesian factor analysis when the number of factors is unknown. Technical report, 2018.

Sylvia Frühwirth-Schnatter and Helga Wagner. Bayesian variable selection for random intercept modeling of gaussian and non-gaussian data. *Bayesian Statistics 9*, 9:165, 2011.

Edward I George and Robert E McCulloch. Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

Domenico Giannone, Michele Lenza, and Giorgio Primiceri. Economic predictions with big data: The illusion of sparsity. *SSRN Electronic Journal*, 07 2020. doi: 10.2139/ssrn.3166281.

Jim Griffin and Philip Brown. Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, 5(1):171–188, 2010.

Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Hemant Ishwaran and J Sunil Rao. Spike and slab variable selection: frequentist and bayesian strategies. *Annals of Statistics*, pages 730–773, 2005.

Gregor Kastner, Sylvia Frühwirth-Schnatter, and Hedibert F. Lopes. Efficient Bayesian inference for multivariate factor stochastic volatility models. *Journal of Computational and Graphical Statistics*, 26: 905–917, 2017.

F. B. Lempers. *Posterior Probabilities of Alternative Linear Models*. Rotterdam University Press, 1988.

T. J. Mitchell and J. J. Beauchamp. Bayesian variable selection in linear regression (with discussion). *Journal of the American Statistical Association*, 83:1023–1036, 1988.

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