Modern Bayesian Statistics
Part III: high-dimensional modeling
Example 3: Sparse and time-varying covariance modeling

Hedibert Freitas Lopes[1]
hedibert.org

13ª aMostra de Estatística
IME-USP, October 2018

[1]Professor of Statistics and Econometrics at Insper, São Paulo.

Example 3: Sparse and time-varying covariance modeling

Consider the Gaussian linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}),$$

where $\boldsymbol{y}$ is n-length vector and $\boldsymbol{X}$ is a $n \times q$ design matrix.

Ridge Regression (ℓ_2 penalty on $\boldsymbol{\beta}$)

$$\hat{\boldsymbol{\beta}}_{ridge} = \arg\min_{\beta} \left\{ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|_2^2 \right\}, \qquad \lambda \geq 0,$$

leading to $\hat{\boldsymbol{\beta}}_{ridge} = (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y}$.

LASSO (ℓ_1 penalty on $\boldsymbol{\beta}$)

$$\hat{\beta}_{lasso} = \arg\min_{\beta} \{ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|_1 \}, \qquad \lambda \geq 0,$$

which can be solved by using quadratic programming techniques such as *coordinate gradient descent*.

Bayesian regularization in linear regression problems

Hierarchical scale mixture of normals:

$$\beta|\psi \sim \mathcal{N}(0,\psi) \qquad \text{and} \qquad \psi|\theta \sim p(\psi),$$

Maximum a posteriori (MAP): $\arg\max_{\beta}\{p(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2)p(\boldsymbol{\beta}|\boldsymbol{\psi})\}$

A few cases:

Prior	$p(\psi)$	$p(\beta)$
Bayesian Lasso	$\psi \sim \mathcal{E}(\lambda^2/2)$	Laplace
Ridge	$\psi \sim \mathcal{IG}(a,b)$	Scaled-t
Normal-Gamma	$\psi \sim \mathcal{G}(\lambda, 1/(2\gamma^2))$	below

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

where $\text{Var}(\beta|\lambda,\gamma^2) = 2\lambda\gamma^2$ and excess kurtosis $3/\lambda$.

The Normal-Gamma prior

High mass close to zero and heavy tails

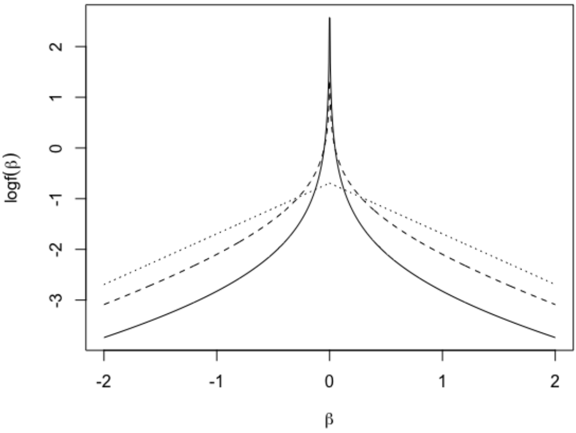

Figure: $\lambda = 0.1$ (dot), $\lambda = 0.33$ (dot-dashed), $\lambda = 1$ (solid).

Spike-and-slab priors

Stochastic search variable selection (SSVS)
SSVS places independent mixture priors directly on the coefficients

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

with $c > 1$ large, $\tau > 0$ small and $J \sim Ber(\omega)$.

SMN representation

$$\beta|\psi \sim \mathcal{N}(0,\psi)$$

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

Spike-and-slab priors on the scale parameter ψ

Normal mixture of Inverse-Gammas (NMIG)

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

SMN representation

$$\beta|\psi \sim \mathcal{N}(0, \psi)$$

and

$$\psi \sim (1-\omega)\mathcal{IG}(a_\tau, \upsilon_0 b_\tau) + \omega\mathcal{IG}(a_\tau, \upsilon_1 b_\tau)$$

The resulting marginal distribution of β is a two component mixture of scaled Student's t distributions.

Spike-and-slab priors on the scale parameter ψ

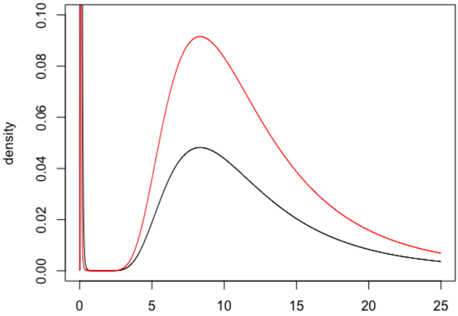

Figure: Conditional density for hypervariance ψ for NMIG mixture prior where $\upsilon_0 = 0.005$, $\upsilon_1 = 1$, $a_\tau = 5$, $b_\tau = 50$ and (a) $\omega = 0.5$ (black line), (b) $\omega = 0.95$ (red line). Note that as ω has a Uniform prior, (a) also corresponds to the marginal density of ψ. Observe that only the height of the density changes as ω varied.

Spike-and-slab priors: summary

Prior	Spike $\psi \vert J=0$	Slab $\psi \vert J=1$	Marginal $\beta \vert \omega$	Constant c
SSVS	$\psi \vert J=0=\delta_{rQ}(.)$	$\psi \vert J=1=\delta_{Q}(.)$	$\omega\mathcal{N}(0,Q)+(1-\omega)\mathcal{N}(0,rQ)$	1
NMIG	$\mathcal{IG}(\nu, rQ)$	$\mathcal{IG}(\nu, Q)$	$\omega t_{2\nu}(0,Q/\nu)+(1-\omega)t_{2\nu}(0,rQ/\nu)$	$1/(\nu-1)$
Mixture of Laplaces	$\mathcal{E}(1/2rQ)$	$\mathcal{E}(1/2Q)$	$\omega Lap(\sqrt{Q})+(1-\omega)Lap(\sqrt{rQ})$	2
Mixture of Normal-Gammas	$\mathcal{G}(a,1/2rQ)$	$\mathcal{G}(a,1/2Q)$	$\omega\mathcal{NG}(\beta_j \vert a,Q)+(1-\omega)\mathcal{NG}(\beta_j \vert a,r,Q)$	$2a$
Laplace-t	$\mathcal{E}(1/2rQ)$	$\mathcal{IG}(\nu, Q)$	$\omega t_{2\nu}(0,Q/\nu)+(1-\omega)Lap(\sqrt{rQ})$	$c_1=2, \quad c_2=1/(\nu-1)$

Gaussian dynamic regression problems

- Consider the univariate Gaussian *dynamic linear model* (DLM) expressed by

$$y_t = \boldsymbol{F}_t'\boldsymbol{\beta}_t + \nu_t, \qquad \nu_t \sim \mathcal{N}(0, V_t) \tag{2}$$

$$\boldsymbol{\beta}_t = \boldsymbol{G}_t\boldsymbol{\beta}_{t-1} + \boldsymbol{\omega}_t, \qquad \boldsymbol{\omega}_t \sim \mathcal{N}(0, \boldsymbol{W}_t), \tag{3}$$

  where $\boldsymbol{\beta}$ is of length q and $\boldsymbol{\beta}_0 \sim \mathcal{N}(\boldsymbol{m}_0, \boldsymbol{C}_0)$.

- Dynamic regression model: $\boldsymbol{F}_t = \boldsymbol{X}_t$ and $\boldsymbol{G}_t = \boldsymbol{I}_q$.

- Static regression model: $\boldsymbol{W}_t = 0$ for all t.

Shrinkage in dynamic regression problems

- Two main obstacles:
  1. Time-varying parameters (states), and
  2. A large number of predictors q.

- Two sources of sparsity:
  1. horizontal sparsity: $\beta_{j,t} = 0, \forall t$ for some coefficients j.
  2. vertical sparsity: $\beta_{j,t} = 0$ for several js at time t.

Shrinkage in dynamic regression problems

- Two main obstacles:
  1. Time-varying parameters (states), and
  2. A large number of predictors q.

- Two sources of sparsity:
  1. horizontal sparsity: $\beta_{j,t} = 0, \forall t$ for some coefficients j.
  2. vertical sparsity: $\beta_{j,t} = 0$ for several js at time t.

Illustration: $q = 5$ and $T = 12$

	jan	*feb*	*mar*	*apr*	*may*	*jun*	*jul*	*aug*	*sep*	*oct*	*nov*
x_0	$\beta_{1,1}$	$\beta_{1,2}$	$\beta_{1,3}$	$\beta_{1,4}$	$\beta_{1,5}$	$\beta_{1,6}$	$\beta_{1,7}$	$\beta_{1,8}$	$\beta_{1,9}$	$\beta_{1,10}$	$\beta_{1,11}$
x_1	0	0	0	0	0	0	0	0	0	0	0
x_2	$\beta_{3,1}$	$\beta_{3,2}$	$\beta_{3,3}$	$\beta_{3,4}$	$\beta_{3,5}$	0	0	0	$\beta_{3,9}$	$\beta_{3,10}$	$\beta_{3,11}$
x_3	0	0	$\beta_{4,3}$	$\beta_{4,4}$	$\beta_{4,5}$	$\beta_{4,6}$	$\beta_{4,7}$	$\beta_{4,8}$	$\beta_{4,9}$	$\beta_{4,10}$	$\beta_{4,11}$
x_4	$\beta_{5,1}$	$\beta_{5,2}$	$\beta_{5,3}$	$\beta_{5,4}$	$\beta_{5,5}$	0	0	0	$\beta_{5,9}$	$\beta_{5,10}$	$\beta_{5,11}$

Our contribution: a dynamic spike-and-slab model

Our contribution is defining a a spike-and-slab prior that not only shrinks time-varying coefficients in dynamic regression problems but allows for dynamic variable selection.

We use a non-centered parametrization:

$$y_t = \boldsymbol{F}_t' \boldsymbol{\beta}_t + \nu_t, \qquad \nu_t \sim \mathcal{N}(0, \sigma_t^2)$$
$$\tilde{\boldsymbol{\beta}}_t = \boldsymbol{G}_t \tilde{\boldsymbol{\beta}}_{t-1} + \boldsymbol{\omega}_t, \qquad \boldsymbol{\omega}_t \sim \mathcal{N}(0, \boldsymbol{W}_t),$$

where

$$\begin{aligned} \tilde{\boldsymbol{\beta}}_t &= \left(\frac{\beta_{1,t}}{\sqrt{\psi_{1,t}}}, \ldots, \frac{\beta_{q,t}}{\sqrt{\psi_{q,t}}} \right)' \\ \boldsymbol{G}_t &= \text{diag}(\varphi_1, \ldots, \varphi_q) \\ \boldsymbol{W}_t &= \text{diag}((1-\varphi_1^2), \ldots, (1-\varphi_q^2)), \\ \boldsymbol{F}_t' &= (X_{1,t}\sqrt{\psi_{1,t}}, \ldots, X_{q,t}\sqrt{\psi_{q,t}}). \end{aligned}$$

Our contribution: a dynamic spike-and-slab model

For shrinking the states $\beta_1, \ldots, \beta_T$, for any j coefficient, we place independent priors for each $\psi_t = \tau^2 K_t$ as

$$\tau^2 \overset{\text{iid}}{\sim} p(\tau^2|\boldsymbol{\theta}),$$

$$(K_t|K_{t-1} = \upsilon_i) \overset{\text{ind}}{\sim} \omega_{1,i}\delta_1(.) + (1-\omega_{1,i})\delta_r(.),$$

$$\omega_{1,i} = p\left(K_t = 1|K_{t-1} = \upsilon_i\right),$$

where $\upsilon_i \in \{r, 1\}$, $p(K_1 = r) = p(K_1 = 1) = 1/2$,
$r = \text{Var}_{spike}(\beta|\boldsymbol{\theta}) / \text{Var}_{slab}(\beta|\boldsymbol{\theta}) \ll 1$ and $p(\tau^2|\boldsymbol{\theta})$ is one of priors from the previous Table.

Markov switching process for K_t

That is, K_t is a binary random latent variable that can assume binary values (regimes) $\upsilon_0 = r$ or $\upsilon_1 = 1$, depending only on the previous value of K_{t-1} and the prior transition probabilities $\{\omega_{0,0}; \omega_{0,1}; \omega_{1,1}; \omega_{1,0}\}$.

Our contribution: direct acyclic graph