Sirio Legramanti

University of Bergamo

Bayesian cumulative shrinkage for infinite factorizations

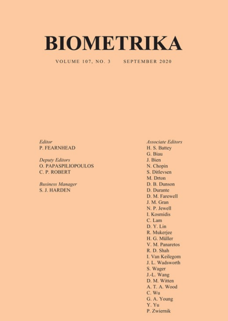
"SELECT" PRIN workshop *Climbing Mortality Models*

Misurina, Italy

August 30 – September 2, 2022

This talk is mainly based on

Legramanti, S., Durante, D., Dunson, D. B.
Bayesian cumulative shrinkage for infinite factorizations
Biometrika, 107(3), 745-752, 2020



and is supported by

- PRIN-MIUR 2017 grant, SELECT project (20177BRJXS)
- U.S. Office of Naval Research and National Institutes of Health

Increasing shrinkage priors

Useful when one expects model dimensions to be **decreasingly important**

Example: Gaussian factor model

$$y_i = \Lambda \eta_i + \epsilon_i$$

$$y_i \in \mathbb{R}^p, \quad \Lambda \in \mathbb{R}^{p \times H}, \quad \eta_i \sim N_H(0, I_H), \quad \epsilon_i \sim N_p(0, \Sigma)$$

$$\lambda_{\cdot h} \sim N_p(0, \theta_h I_p)$$

Parsimony in the number of relevant dimensions can be induced by **a prior that increasingly shrinks $\boldsymbol{\theta_h}$**, e.g.

Multiplicative Gamma Process (Bhattacharya and Dunson, 2011)

$$\theta_h^{-1} = \prod_{\ell=1}^{h} \vartheta_\ell, \qquad \vartheta_1 \sim \mathrm{Ga}(a_1, 1), \quad \vartheta_\ell \sim \mathrm{Ga}(a_2, 1) \;\; (\ell \geq 2),$$

which however has some drawbacks (Durante, 2017)

Cumulative Shrinkage Process (CUSP)

Definition

$\theta = \{\theta_h \in \Theta \subseteq \mathbb{R} : h = 1, 2, \ldots\}$ is distributed according to a CUSP with **parameter** $\boldsymbol{\alpha > 0}$, starting **slab** $\boldsymbol{P_0}$ and target **spike** $\boldsymbol{\delta_{\theta_\infty}}$ if, conditionally on $\pi = \{\pi_h \in [0, 1] : h = 1, 2, \ldots\}$, each θ_h is independent and has the following **spike-and-slab** distribution:

$$(\theta_h \mid \pi_h) \sim P_h = (1 - \pi_h) P_0 + \pi_h \delta_{\theta_\infty}, \tag{1}$$

where

$$\pi_h = \sum_{\ell=1}^{h} \omega_\ell, \qquad \omega_\ell = v_\ell \prod_{m=1}^{\ell-1} (1 - v_m), \tag{2}$$

$$v_1, v_2, \ldots \sim \text{Beta}(1, \alpha), \qquad \text{independent.}$$

- P_0 is tipically a **diffuse** continuous distribution
- δ_{θ_∞} can be replaced with a **concentrated** continuous distribution

Properties of $\pi = \{\pi_1, \pi_2, \ldots\}$

- **Increasing shrinkage**

$$\pi_h \uparrow 1 \quad \text{a.s.}$$

since (2) uses the stick-breaking construction of the Dirichlet process;

- **Expectation**

$$E(\pi_h) = 1 - \left(\frac{\alpha}{1+\alpha}\right)^h;$$

- **Large support**

on non-decreasing sequences taking values in $[0, 1]$;

- **Interpretation**

$$\pi_h = \frac{d_{\mathrm{TV}}(P_0, P_h)}{d_{\mathrm{TV}}(P_0, \delta_{\theta_\infty})}.$$

Increasing shrinkage on $\theta = \{\theta_1, \theta_2, \ldots\}$

Both in **expectation**: $$E(\theta_h) = \theta_\infty + \left(\frac{\alpha}{1+\alpha}\right)^h (\theta_0 - \theta_\infty)$$

and in **probability**: $$\text{pr}(|\theta_h - \theta_\infty| > \varepsilon) = P_0\{\bar{\mathbb{B}}_\varepsilon(\theta_\infty)\} \left(\frac{\alpha}{1+\alpha}\right)^h$$

Larger $\boldsymbol{\alpha}$ = slower shrinkage

In fact, augmenting (1) as

$$(\theta_h \mid c_h) \sim c_h P_0 + (1 - c_h)\delta_{\theta_\infty}, \qquad c_h \sim \text{Bern}(1 - \pi_h),$$

we have

$$\boldsymbol{H^*} = \sum_{h=1}^{\infty} c_h = \textbf{ number of active elements} \text{ in } \theta$$

$$\boldsymbol{E(H^*)} = \sum_{h=1}^{\infty} E(1 - \pi_h) = \sum_{h=1}^{\infty} \left(\frac{\alpha}{1+\alpha}\right)^h \boldsymbol{= \alpha}.$$

Application to Gaussian factor models

$$y_i = \Lambda \eta_i + \epsilon_i,$$
$$\lambda_{\cdot h} \sim N_p(0, \theta_h I_p),$$
$$\eta_i \sim N_H(0, I_H),$$
$$\epsilon_i \sim N_p(0, \Sigma), \quad \Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_p^2), \quad \sigma_j^2 \sim \text{InvGa}(a_\sigma, b_\sigma).$$

We place a **CUSP** on the variances of the loadings columns:

$$(\theta_h \mid \pi_h) \sim (1 - \pi_h)\text{InvGa}(a_\theta, b_\theta) + \pi_h \delta_{\theta_\infty}, \tag{3}$$
$$\pi_h = \sum\nolimits_{\ell=1}^{h} \omega_\ell, \qquad \omega_\ell = v_\ell \prod\nolimits_{m=1}^{\ell-1} (1 - v_m),$$
$$v_1, \dots, v_{H-1} \sim \text{Beta}(1, \alpha), \qquad v_H = 1.$$

This implies that $\quad \pi_H = 1 \quad$ i.e. $\quad \theta_H = \theta_\infty \quad$ a.s.

Properties of CUSP on factor models

The induced probability measure Π on $\Omega = \Lambda\Lambda^{\mathrm{T}} + \Sigma$

- is **well defined**: if $E(\theta_h) < \infty$ for $h = 1, \ldots, H$,

$$\Pi\{\Omega \text{ has } \textbf{finite entries} \text{ and is } \textbf{positive semi-definite}\} = 1$$

- has **large support**: if there exists a decomposition $\Omega_0 = \Lambda_* \Lambda_*^{\mathrm{T}} + \Sigma_0$ with $\Lambda_* \in \mathbb{R}^{p \times H_0}$ and $H_0 \leq H$, then

$$\Pi\{B_\epsilon^\infty(\Omega_0)\} > 0$$

  $\forall \epsilon > 0$ and any covariance matrix $\Omega_0 \in \mathbb{R}^{p \times p}$,
  with $B_\epsilon^\infty(\Omega_0)$ being an ϵ-neighborhood of Ω_0 under the sup-norm.

$\Longrightarrow$ **posterior weak consistency** (Schwartz, 1965).

Moreover, since marginally $\lambda_{jh} \sim (1 - \pi_h) t_{2a_\theta}(0, b_\theta / a_\theta) + \pi_h N(0, \theta_\infty)$,

if $(b_\theta / a_\theta) > \sqrt{\theta_\infty}$, then $\quad \mathrm{pr}(|\lambda_{j,h+1}| \leq \varepsilon) > \mathrm{pr}(|\lambda_{jh}| \leq \varepsilon) \qquad \forall \varepsilon > 0.$

1 **for** *j from 1 to p* **do**
  sample the j-th row of Λ from $N_H(V_j \eta^{\mathrm{T}} \sigma_j^{-2} y_j, V_j)$, with $V_j = (D^{-1} + \sigma_j^{-2} \eta^{\mathrm{T}} \eta)^{-1}$, $D = \mathrm{diag}(\theta_1, \ldots, \theta_H)$, $\eta = (\eta_1, \ldots, \eta_n)^{\mathrm{T}}$ and $y_j = (y_{1j}, \ldots, y_{nj})^{\mathrm{T}}$;

2 **for** *j from 1 to p* **do**
  sample σ_j^2 from $\mathrm{InvGa}\{a_\sigma + 0{\cdot}5n, b_\sigma + 0{\cdot}5 \sum_{i=1}^{n} (y_{ij} - \sum_{h=1}^{H} \lambda_{jh} \eta_{ih})^2\}$;

3 **for** *i from 1 to n* **do**
  sample η_i from $N_H\{(I_H + \Lambda^{\mathrm{T}} \Sigma^{-1} \Lambda)^{-1} \Lambda^{\mathrm{T}} \Sigma^{-1} y_i, (I_H + \Lambda^{\mathrm{T}} \Sigma^{-1} \Lambda)^{-1}\}$;

4 **for** *h from 1 to H* **do**
  sample z_h from the categorical variable with probabilities

$$\mathrm{pr}(z_h = l \mid -) \propto \begin{cases} \omega_l N_p(\lambda_h; 0, \theta_\infty I_p), & \text{for } l = 1, \ldots, h, \\ \omega_l t_{2a_\theta}\{\lambda_h; 0, (b_\theta / a_\theta) I_p\}, & \text{for } l = h+1, \ldots, H, \end{cases}$$

5 **for** *l from 1 to* $(H-1)$ **do**
  update v_l from $\mathrm{Beta}\{1 + \sum_{h=1}^{H} \mathbb{1}(z_h = l), \alpha + \sum_{h=1}^{H} \mathbb{1}(z_h > l)\}$;

set $v_H = 1$ and update $\omega_1, \ldots, \omega_H$ from $v_1, \ldots, v_H$ through (2);

6 **for** *h from 1 to H* **do**
  if $z_h \leq h$ **then** $\theta_h = \theta_\infty$ **else** sample θ_h from $\mathrm{InvGa}(a_\theta + 0{\cdot}5p, b_\theta + 0{\cdot}5 \sum_{j=1}^{p} \lambda_{jh}^2)$;

Output at the end of one cycle: a unit sample from the posterior of $\Omega = \Lambda\Lambda^{\mathrm{T}} + \Sigma$.

Obtained augmenting (3) with z_h s.t. $\mathrm{pr}(z_h = \ell \mid \omega_\ell) = \omega_\ell$, which implies

$$(\theta_h \mid z_h) \sim \{1 - \mathbb{1}(z_h \leq h)\}\mathrm{InvGa}(a_\theta, b_\theta) + \mathbb{1}(z_h \leq h)\delta_{\theta_\infty}.$$

Adaptive Gibbs sampler

Initialize $H = p + 1$
For each iteration $t = \bar{t} + 1, \ldots, T$
with probability $\boldsymbol{p(t) = \exp(-\alpha_0 - \alpha_1 t)}$
if inactive columns of Λ are more than one
replace them with a column sampled from the spike
else, if $H < p + 1$, add a column sampled from the spike

This satisfies the **diminishing adaptation condition** (Roberts and Rosenthal, 2007)

With CUSP, the **inactive columns** of Λ are **naturally identified** as those modeled by the spike, i.e. those such that $\boldsymbol{z_h < h}$.

CUSP vs Multiplicative Gamma Process

On data simulated from a Gaussian factor model ($\lambda_{jh} \overset{iid}{\sim} N(0,1)$, $\Sigma_0 = I_p$)

- CUSP and MGP yield **comparable MSE** on $\Omega = \Lambda\Lambda^{\mathrm{T}} + \Sigma$
- **CUSP exactly recovers H_0** (the true number of factors)
- **CUSP is faster**

(p, H_0)	method	MSE		$E(H^* \mid y)$		averaged ESS	runtime (s)
		median	IQR	median	IQR	median	median
(20,5)	CUSP	0.75	0.29	5.00	0.00	655.04	310.76
	MGP	0.75	0.32	19.69	0.21	547.23	616.61
(50,10)	CUSP	2.25	0.33	10.00	0.00	273.55	716.23
	MGP	2.26	0.28	28.64	1.94	251.35	1845.88
(100,15)	CUSP	3.76	0.40	15.00	0.00	175.26	2284.87
	MGP	3.97	0.45	34.38	2.92	116.10	5002.33

This is **confirmed on real data** (dataset `bfi` from the `R` package `psych`).

Available extensions

Beyond Gaussian factor models, CUSP has been employed in

- singular value decomposition (Tanaka, 2020)
- low-rank **matrix decomposition** (Tanaka, 2021)
- tensor **vector autoregressive models** (Zhang et al., 2021)
- multivariate **categorical data** (Gu and Dunson, 2021)
- infinite basis expansion for **functional data** (Kowal and Canale, 2021)

Computational advances

- mean-field **variational algorithm** (Legramanti, 2020)
- parameter expansion to **improve mixing over H*** (Kowal and Canale, 2021)

Summary

- the cumulative shrinkage process (CUSP) yields **increasing shrinkage** both in expectation and in probability, for any choice of α
- the **shrinkage rate** (regulated by α) can be tuned separately from the distribution for active terms (the slab P_0)
- α is the **expected number of active terms** under the CUSP
- an **adaptive Gibbs sampler**, speeding up computations by discarding inactive dimensions, is provided
- **alternative computational strategies** (parameter expansions, variational algorithms) have been proposed
- the CUSP has been employed in **Gaussian factor models and beyond** (SVD, VAR, categorical and functional data)

Thanks for your attention

For any question, feel free to contact me at

sirio.legramanti@unibg.it

References

A. Bhattacharya and D. B. Dunson. Sparse Bayesian infinite factor models. *Biometrika*, 98: 291–306, 2011.

D. Durante. A note on the multiplicative gamma process. *Stat. Prob. Lett.*, 122:198–204, 2017.

Y. Gu and D. B. Dunson. Identifying interpretable discrete latent structures from discrete data. *arXiv preprint arXiv:2101.10373*, 2021.

D. R. Kowal and A. Canale. Semiparametric functional factor models with Bayesian rank selection. *arXiv preprint arXiv:2108.02151*, 2021.

S. Legramanti. Variational Bayes for gaussian factor models under the cumulative shrinkage process. *Book of short papers SIS 2020*, pages 416–420, 2020.

G. O. Roberts and J. S. Rosenthal. Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms. *J. Appl. Probab.*, 44(2):458–475, 2007.

L. Schwartz. On Bayes procedures. *Probab. Theory Relat. Fields*, 4(1):10–26, 1965.

M. Tanaka. Bayesian singular value regularization via a cumulative shrinkage process. *Communications in Statistics-Theory and Methods*, pages 1–24, 2020.

M. Tanaka. Bayesian matrix completion approach to causal inference with panel data. *Journal of Statistical Theory and Practice*, 15(2):1–22, 2021.

W. Zhang, I. Cribben, M. Guindani, et al. Bayesian time-varying tensor vector autoregressive models for dynamic effective connectivity. *arXiv preprint arXiv:2106.14083*, 2021.