The graphical horseshoe estimator for inverse covariance matrices

Anindya Bhadra
www.stat.purdue.edu/∼bhadra

Purdue University

Overview

- Goal: To develop a new Bayes estimator for sparse precision matrices for multivariate Gaussian data.
- A full Gibbs sampler for efficient sampling.
- Theoretical contrast with popular existing methods such as the graphical lasso and graphical SCAD.
- Numerical examples.
- *Joint work with Yunfan Li and Bruce Craig at Purdue. Supported by NSF Grant DMS-1613063.*

Some existing estimators

- Estimators that assume an unstructured precision matrix usually make an assumption of sparsity.
  - Frequentist: graphical lasso (Friedman et al. 2008, Biostatistics); graphical SCAD (Lam and Fan, 2009, AoS).
  - Bayesian: the Bayesian graphical lasso (Wang, 2012, BA); the proposed graphical horseshoe estimator.

- Estimators that assume an underlying structure (banding, latent factors, low rank)
  - Frequentist: Bickel and Levina (2008, AoS) and many others
  - Bayesian: Banerjee and Ghosal (2014, EJS (banded)); Pati et al. (2014, AoS (sparse latent factors))

Some existing estimators

- Let Ω be $p \times p$ positive definite. Observe i.i.d.

$$\mathbf{y}_k \sim \text{Normal}(\mathbf{0}, \Omega^{-1}),$$

  for $k = 1, \ldots, n$ and let $p > n$.
- Let $S = \sum_{i=1}^{n} \mathbf{y}_i \mathbf{y}_i'$. Penalized likelihood approaches maximize

$$\log(\det \Omega) - \text{tr}(S\Omega/n) - \sum_{i,j} \phi_\lambda(|\omega_{ij}|),$$

  - Graphical lasso: $\phi_\lambda(|x|) = \lambda|x|;\ \lambda > 0$.
  - Graphical SCAD:

$$\phi'_\lambda(|x|) = \lambda \left\{ 1_{\{|x| \le \lambda\}} + \frac{(a\lambda - |x|)_+}{(a-1)\lambda} 1_{\{|x| > \lambda\}} \right\};\ \lambda > 0, a > 2.$$

Some existing estimators

- Wang (2012, BA) showed the frequentist glasso estimate is posterior mode under the prior

$$p(\Omega \mid \lambda) \propto \prod_{i<j} \{\mathrm{DE}(\omega_{ij} \mid \lambda)\} \prod_{i=1}^{p} \{\mathrm{EXP}(\omega_{ii} \mid \lambda/2)\} 1_{\Omega \in \mathcal{S}_p} \ ,$$

  where S_p is the set of positive definite matrices.

- He also developed a block Gibbs sampler for a full Bayes solution - a strategy we will closely follow.

- The posterior mean estimate under this prior is known as the Bayesian graphical lasso (BGL).

Issues with the existing estimators

- Graphical lasso introduces bias in estimating large signals due to soft thresholding.
- Graphical SCAD penalty is non-convex and unbiased for large signals, but the estimate is not guaranteed to be positive definite in finite samples.
- We will also argue neither graphical lasso nor graphical SCAD provides strong enough shrinkage towards zero, resulting in poorer information-theoretic properties.

Our proposal

- With the constraint $\Omega \in S_p$, define element-wise $(i, j = 1, \ldots, p)$

$$\begin{aligned} \omega_{ii} &\propto 1, \\ \omega_{ij:i<j} &\sim \text{Normal}(0, \lambda_{ij}^2 \tau^2), \\ \lambda_{ij:i<j} &\sim C^+(0, 1), \end{aligned}$$

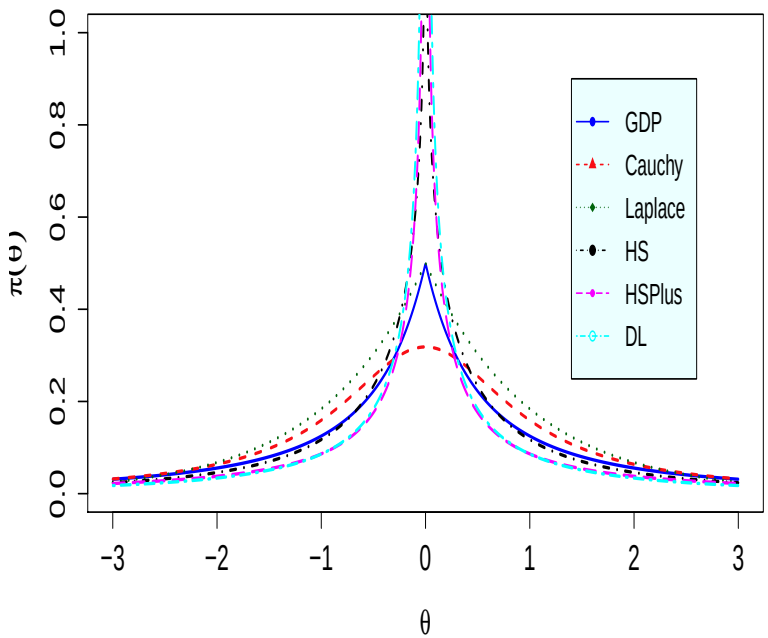
- That is, the prior on Ω is

$$p(\Omega \,|\, \tau) \propto \prod_{i<j} \text{Normal}(\omega_{ij} \,|0, \lambda_{ij}^2 \tau^2) \prod_{i<j} \text{C}^+(\lambda_{ij} \,|\, 0, 1) 1_{\Omega \in \mathcal{S}_p},$$

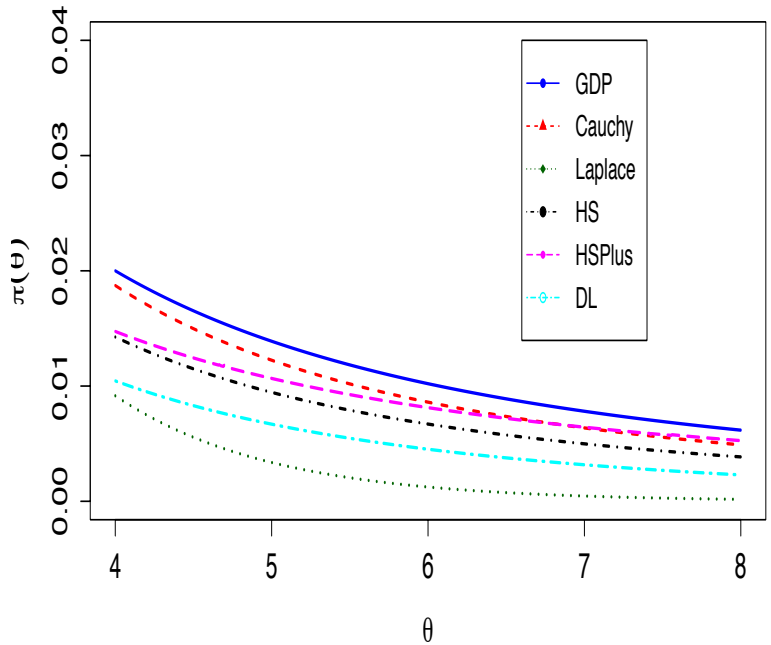
- A non-informative prior on the diagonal terms and independent global-local horseshoe priors on off-diagonal terms to ensure (i) efficient shrinkage of noise terms and (ii) not shrinking the signals.

Some examples of global-local priors

- The order of peakedness near zero: $HS+ \approx DL > HS > GDP = Laplace > Cauchy$



- The order of tail heaviness: $GDP > Cauchy > HS+ > HS > DL > Laplace$



Posterior simulation

- The posterior is

$$p(\Omega \mid Y, \Lambda, \tau) \propto |\Omega|^{\frac{n}{2}} \exp\Big\{ -\operatorname{tr}\Big(\frac{1}{2} S\Omega\Big)\Big\} \prod_{i<j} \exp\Big(-\frac{\omega_{ij}^2}{2\lambda_{ij}^2 \tau^2}\Big) 1_{\Omega \in \mathcal{S}_p}.$$

- Write

$$\Omega = \begin{pmatrix} \Omega_{(-p)(-p)} & \boldsymbol{\omega}_{(-p)p} \\ \boldsymbol{\omega}'_{(-p)p} & \omega_{pp} \end{pmatrix}, \; S = \begin{pmatrix} S_{(-p)(-p)} & \mathbf{s}_{(-p)p} \\ \mathbf{s}'_{(-p)p} & s_{pp} \end{pmatrix}.$$

- Define $\boldsymbol{\beta} = \boldsymbol{\omega}_{(-p)p}$ and $\gamma = \omega_{pp} - \boldsymbol{\omega}'_{(-p)p} \Omega^{-1}_{(-p)(-p)} \boldsymbol{\omega}_{(-p)p}$.

Posterior simulation

- Wang (2012, BA) showed

  $$p(\gamma, \beta \,|\, \Omega_{(-p)(-p)}, Y, \Lambda, \tau) \sim \text{Gamma}(n/2+1, s_{pp}/2) \times \text{Normal}(-C\mathbf{s}_{(-p)p}, C),$$

  where $C = \{s_{pp}\Omega^{-1}_{(-p)(-p)} + (\Lambda^* \tau^2)^{-1}\}^{-1}$.

- Conditional posteriors of off-diagonal terms are normal, those of diagonal terms are inverse gamma.

- If the initial Ω is positive definite, ensures all subsequent iterations are also positive definite.

Posterior simulation

- The difference is, for BGL, $\lambda_{ij} \sim \text{Exp}(1)$ but for GHS $\lambda_{ij} \sim C^+(0,1)$, a priori.

- Here, we follow the key data augmentation technique proposed by Makalic and Schmidt (2016, IEEE Sig. Proc. Letters):

$$\text{if } x^2 \mid a \sim \text{InvGamma}(1/2, 1/a) \quad \text{and} \quad a \sim \text{InvGamma}(1/2, 1),$$
$$\text{then marginally,} \quad x \sim C^+(0,1)$$

- That is, a half-Cauchy is a mixture of two inverse gammas.

- BUT! Inverse gamma is conjugate to itself and to the variance parameter in a normal linear regression model.

Posterior simulation

- All conditional posteriors are available in closed form, leading to a full Gibbs sampler.
- They are either multivariate normal, gamma or inverse gamma.
- Computational complexity is $O(p^3)$.
- For full details, see Algorithm 1 of Li, Craig and Bhadra (2017, arXiv: 1707.06661).
- MATLAB code on github at `http://github.com/liyf1988/GHS`.

A useful lemma (Li, Craig and Bhadra, 2017)

- Modification of Barron and Clarke (1990, IEEE Trans. Inf. Theory).
- Let the true parameter be Ω_0 and $A_\epsilon = \{\Omega : D(p_{\Omega_0} || p_\Omega) \leq \epsilon\}$.
- Let $\nu(d\Omega)$ be the prior measure of Ω and $\nu_n(d\Omega) \propto \prod_{i=1}^n p_\Omega(y_i)\nu(\mathrm{d}\Omega)$ be the posterior measure.
- $\hat{p}_n = \int p_\Omega \nu_n(d\Omega)$ be the posterior mean estimate of the density.
- The Cesàro-average risk R_n of the estimator $\hat{p}_n$ admits the following lower and upper bounds for all $\epsilon > 0$:

$$-\epsilon - \frac{1}{n}\log \nu(A_\epsilon) \leq R_n = \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}D(p_{\Omega_0} || \hat{p}_j) \leq \epsilon - \frac{1}{n}\log \nu(A_\epsilon),$$

where $\mathbb{E}$ denotes an expectation with respect to the data distribution.

K–L risk bounds

- If we take $\epsilon = 1/n$, then R_n is a function of two things: (i) the sample size n and (ii) $\nu(A_{1/n})$, the prior measure of K–L information neighborhood of true Ω_0.

- BUT! Recall the horseshoe density is unbounded at zero. When the true parameter is zero, it places more mass in the K–L neighborhood than any prior that is continuous and bounded above.

- This immediately includes the double exponential prior in BGL.

- The graphical SCAD estimate does not admit an interpretation as a posterior mode similar to BGL, but if we view the corresponding penalty as negative of the log of prior, it's easy to see that prior will be bounded above.

K–L risk bounds (Thm. 3.2; Li, Craig and Bhadra, 2017)

- For $\hat{p}_n$ under the graphical horseshoe prior, $p_0 \log\{\frac{C_1}{Mn^{1/2}p}\log(2Mn^{1/2}p)\} + p_1 \log\frac{C_2}{n^{1/2}p} < \log\nu(A_{1/n}) < p_0 \log\{\frac{C_3}{Mn^{1/2}p}\log(2^{1/2}Mn^{1/2}p)\} + p_1 \log\frac{C_4}{n^{1/2}p}$, where p_0 is the number of zero elements in Ω_0, p_1 is the number of nonzero elements in Ω_0, and C_1, C_2, C_3, C_4 are constants.

- Suppose $p(\omega_{ij})$ is any other prior density that is continuous, bounded above, and strictly positive on a neighborhood of the true value ω_{ij0}. Then $p^2 \log\frac{K_1}{n^{1/2}p} < \log\nu(A_{1/n}) < p^2 \log\frac{K_2}{n^{1/2}p}$, where K_1 and K_2 are constants.

K–L risk bounds: summary

- The bounds for all methods diverge when $p^2/n \to \infty$.
- But the upper and lower bounds for GHS diverge slower when true Ω is sparse.
- For finite n and finite $p > n$, we see a nontrivial improvement over BGL and graphical SCAD.

Bias

- We are also able to show the graphical horseshoe estimate is nearly unbiased in estimating large signals.
- This is at a contrast with the BGL, which will leave a constant bias regardless of the signal strength due to soft thresholding.
- These results mirror the findings of lasso vs. horseshoe in the normal means model.

Numerical examples

We consider the following structures for true Ω:

- *Random.* A sparse matrix is generated. Then each off-diagonal element is randomly assigned a sign.

- *Hubs.* The rows/columns are partitioned into disjoint groups $\{G_k\}_1^K$. Within each group a sparse structure is generated, but no connections between groups.

- *Cliques.* A sparse decomposable graph that admits clique-separator decomposition.

Numerical examples ($p = 100, n = 50$)

nonzero pairs nonzero elem.	Random 35/4950 $\sim -\mathrm{Unif}(0.2, 1)$					Hubs 90/4950 0.25				
	GL1	GL2	GSCAD	BGL	GHS	GL1	GL2	GSCAD	BGL	GHS
Stein's loss	10.20	13.42	10.05	80.92	**6.44**	10.12	12.78	**10.01**	77.85	12.56
	(0.53)	(1.06)	(0.55)	(1.63)	(0.85)	(0.53)	(0.96)	(0.50)	(1.66)	(1.04)
F norm	4.33	5.30	4.31	5.58	**3.31**	3.95	4.63	**3.94**	5.97	3.96
	(0.18)	(0.20)	(0.16)	(0.26)	(0.29)	(0.13)	(0.18)	(0.14)	(0.30)	(0.27)
TPR	.8246	.7097	**.9977**	.8709	.5903	.8649	.7333	**.9987**	.8513	.2687
	(.0520)	(.0620)	(.0078)	(.0470)	(.0537)	(.0443)	(.0751)	(.0053)	(.0378)	(.0764)
FPR	.0947	.0374	.9955	.1055	**.0004**	.0919	.0281	.9976	.1189	**.0013**
	(.0141)	(.0070)	(.0102)	(.0059)	(.0003)	(.0130)	(.0086)	(.0069)	(.0058)	(.0005)
Avg CPU time	0.30	0.35	6.24	40.94	38.32	0.14	0.16	4.01	35.44	41.58

nonzero pairs nonzero elem.	Cliques positive 30/4950 -0.45					Cliques negative 30/4950 0.75				
	GL1	GL2	GSCAD	BGL	GHS	GL1	GL2	GSCAD	BGL	GHS
Stein's loss	9.16	14.16	8.99	81.58	**5.87**	11.00	14.37	10.90	81.27	**6.28**
	(0.55)	(1.06)	(0.52)	(2.51)	(0.93)	(0.43)	(1.02)	(0.43)	(1.98)	(1.09)
F norm	3.75	5.01	**3.71**	5.44	3.81	6.00	6.86	5.99	6.51	**3.64**
	(0.16)	(0.16)	(0.17)	(0.33)	(0.41)	(0.14)	(0.16)	(0.14)	(0.20)	(0.36)
TPR	**1**	**1**	**1**	**1**	.7487	.9993	.9880	**1**	.9993	.9733
	(0)	(0)	(0)	(0)	(.0427)	(.0047)	(.0221)	(0)	(.0047)	(.0421)
FPR	.0900	.0255	.9901	.1014	**.0003**	.0922	.0279	.9752	.1161	**.0010**
	(.0098)	(.0056)	(.0177)	(.0052)	(.0003)	(.0135)	(.0084)	(.0219)	(.0051)	(.0005)
Avg CPU time	0.24	0.28	4.52	34.45	41.65	0.18	0.20	6.91	33.88	41.05

Summary and conclusions

- The proposed estimator performs better than competing methods in terms of Stein's loss (a finite sample estimate of the K–L risk). Also performs well in terms of estimation and variable selection.
- The full Gibbs sampler helps avoiding issues such as tuning an M–H sampler, notoriously difficult in high dimensions.
- We have provided some preliminary theory, but the rich theory developed for horseshoe priors in the normal means model is waiting to be applied here!
- Preprint: arXiv: 1707.06661; Code: http://github.com/liyf1988/GHS