Horseshoe Regularization for Machine Learning in Complex and Deep Models

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

Purdue University

Overview

- Global-local regularization is now well established for high-dimensional Bayesian statistics.
- Most existing results are in linear, Gaussian models.
- This talk: Advances in complex and deep models.
  - Nonlinear models
  - Non-Gaussian models
  - Deep models
- *Joint work with Jyotishka Datta (Arkansas); Yunfan Li (Purdue); Nick Polson and Brandon Willard (Chicago Booth). Supported by NSF Grant DMS-1613063.*

Global-local (g-l) priors

- Consider the sparse "normal means" model $(y_i \mid \theta_i) \sim \mathcal{N}(\theta_i, 1)$ for $i = 1, \ldots, n$; such that $\#(\theta_i \neq 0) \leq p_n$ with $p_n = o(n)$.

- Carvalho, Polson and Scott (2010) introduced "global-local" normal scale mixture priors for sparsity

$$(\theta_i \mid \lambda_i, \tau) \sim \mathcal{N}(0, \lambda_i^2 \tau^2); \quad \lambda_i \sim p(\lambda_i); \quad \tau \sim p(\tau).$$

  - The "global" term τ should provide substantial shrinkage towards zero.

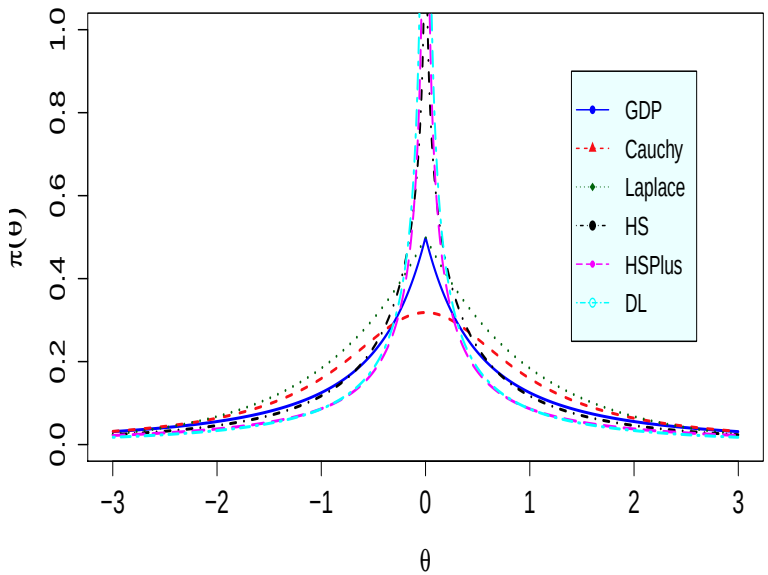
  - The "local" λ_i terms should have heavy tails so that "signals" are not shrunk too much. One option is $p(\lambda_i) \propto (1 + \lambda_i^2)^{-1}$, which induces the "horseshoe prior" on θ.

Some examples of global-local priors

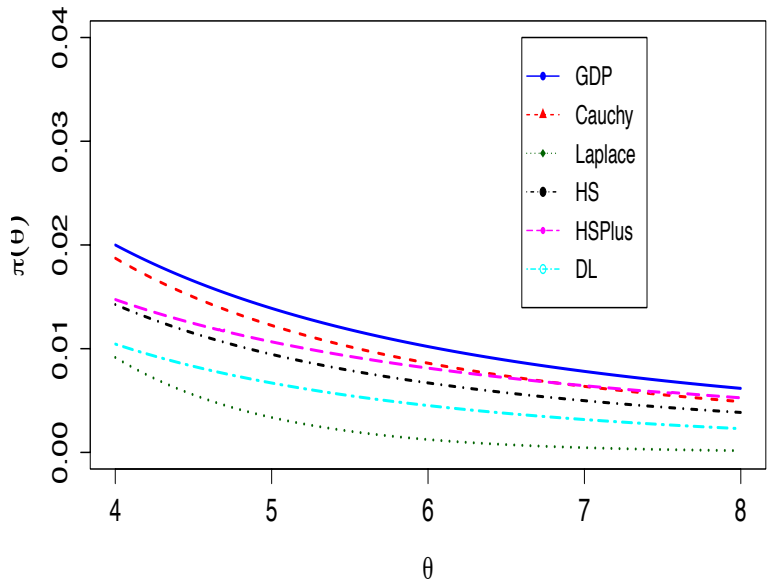
- The horseshoe prior (Carvalho, Polson and Scott, 2010, Biometrika).
- The horseshoe+ prior (Bhadra et al., 2017, Bayesian Anal.).
- The generalized double Pareto prior (Armagan, Dunson and Lee, 2013, Stat. Sinica).
- The Dirichlet-Laplace prior (Bhattacharya et al., 2015, JASA).
- The Inverse Gamma–Gamma Prior (Bai and Ghosh, 2017, arXiv:1710.04369).

Some examples of global-local priors

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



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



Some properties of g-l priors in linear models

- Datta and Ghosh (2013, Bayesian Anal.): proved that the decision rule induced by the horseshoe estimator is asymptotically Bayes optimal for multiple testing under 0-1 loss.

- van der Pas et al. (2014, 2017, EJS): showed the horseshoe, horsehsoe+ and several other g-l estimators are minimax in ℓ_2 up to a constant.

- Bhadra et al. (2019, JMLR): Improved finite sample predictive risk results in linear regression.

- Summary of available results for linear models in Bhadra et al. (2019, Stats. Sci.):

Default Bayes analysis for nonlinear estimation problems using g-l priors

- Bhadra et al. (2016, Biometrika):
  - $\psi_1 = \sum \theta_i^2$ (sum of squares)
  - $\psi_2 = \max \theta_i$ (max)
  - $\psi_3 = \theta_1 \theta_2$ (product)
  - $\psi_4 = \theta_1 / \theta_2$ (ratio or Fieller-Creasy).
- The local heavy-tailed λ_i terms leave large signals un-shrunk, even for nonlinear functions!
- The global term helps shrink the noise components, even for nonlinear functions!

A key property: regular variation

Key property: half-Cauchy (Gelman) has regularly-varying tails.

- Regular variation is closed under many nonlinear transformations (including four on the previous slide).
- The regularly varying tails of θ_is translate to regularly varying tails for the prior of ψ.
- Since the likelihood is light-tailed (normal), the heavy tailed priors on ψ help in non-informative analysis (Dawid, 1973).

Key findings

- Resolves a long-standing difficulty noted by Efron (1973) on designing a non-informative prior that works for all four problems.
- Performance is competitive with the reference priors, when they exist.
- Global-local priors are proper, no difficulties with model selection using usual techniques.

Nonparametric function estimation using g-l priors

- Shin et al. (2019, JASA) consider the standard nonparametric model $Y_i = F(x_i) + \epsilon_i$, for $i = 1, \ldots, n$, where $F(x) = E(Y \mid x)$.
- A natural representation is to use a basis function representation of the form $f(x_i) = \sum_{k=1}^{K} \phi_k(x_i)\beta_k$.
- Possible to shrink the basis coefficients, but not clear what that means.
- If the a-priori belief is $F(\cdot)$ is close to certain parametric families (e.g., linear or quadratic), more reasonable to shrink toward that shape.

The functional horseshoe prior Shin et al. (2019, JASA)

- Define ϕ_0 to be the column space of parametric function one desires to shrink to.

- For example, if the parametric form is assumed to be close to linear then $\phi_0 = \{1, x\} \in \mathbb{R}^{n \times 2}$.

- The prior on β is defined as

$$p(\beta \mid \tau) \propto (\tau^2)^{-(k-d_0)/2} \exp\left\{ -\frac{1}{2\sigma^2\tau^2} \beta^T \phi^T (I - Q_0)\phi\beta \right\},$$

$$p(\tau) \propto \frac{(\tau^2)^{b-1/2}}{(1+\tau^2)^{(a+b)}};\ \tau, a, b > 0,$$

where $d_0 = \mathrm{rank}(\phi_0)$ and $Q_0 = \phi_0(\phi_0^T\phi_0)^{-1}\phi_0^T$, is the projection matrix of ϕ_0.

The functional horseshoe prior Shin et al. (2019, JASA)

- The marginal prior on τ is half-cauchy if $a = b = 1/2$.
- The term $(I - Q_0)$ in the prior inverse covariance enables shrinkage of ϕ towards ϕ_0.
- Model selection consistency is demonstrated, along with good empirical results.

Global-local priors for dependent data

- Define $h_i = \log(\tau^2 \lambda_i^2)$.

- Kowal et al. (2019, JRSSB) show the horseshoe hierarchy up to $(\theta_i \mid \tau)$ is achieved by the model

$$h_i = \mu + \eta_i,\ \eta_i \sim Z(1/2, 1/2, 0, 1),$$

  where $\mu = \log(\tau^2), \eta_i = \log(\lambda_i^2)$ and Z denotes the Fisher-Z distribution.

- Dependence in h_i is introduced using an autoregressive structure as

$$h_i = \mu + \phi(h_{i-1} - \mu) + \eta_i,\ \eta_i \sim Z(1/2, 1/2, 0, 1),$$

Global-local priors for dependent data

- For sampling Kowal et al. (2019) use the Pólya-gamma augmentation scheme.
- Application in Bayesian trend filtering.
- Very similar formulation by Bitto and Frühwirth-Schnatter (2019, J of Econometrics), who use a double gamma prior.

Global-local priors for multivariate models

- Li, Craig and Bhadra (2019, JCGS) consider the multivariate Gaussian model $y_i \mid \Omega \sim \text{Normal}(0, \Omega^{-1})$ for $i = 1, \ldots, n$.

- Prior on Ω is termed the graphical horseshoe (GHS), defined as

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

  with prior mass truncated to the space of positive definite matrices.

- Non-informative prior on the diagonal terms, independent horseshoe priors on the off-diagonal terms.

Posterior sampling under the graphical horseshoe (Li, Craig and Bhadra, 2019, JCGS)

- Additional challenge: need to maintain symmetry and positive definiteness.

- Partition the matrix Ω as:

$$\Omega = \begin{pmatrix} \Omega_{(-p)(-p)} & \boldsymbol{\omega}_{(-p)p} \\ \boldsymbol{\omega}'_{(-p)p} & \omega_{pp} \end{pmatrix},$$

  where $(-p)$ denotes the set of all indices except for p.

- Following Wang (2012, BA) reparameterize:

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

Posterior sampling under the graphical horseshoe (Li, Craig and Bhadra, 2019, JCGS)

- Wang (2012, BA) showed for Bayesian graphical lasso (BGL):

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

  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 gamma.

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

Posterior sampling under the graphical horseshoe (Li, Craig and Bhadra, 2019, JCGS)

- 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 sampling under the graphical horseshoe (Li, Craig and Bhadra, 2019, JCGS)

- 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 $O(p^3)$.
- MATLAB code on github at `http://github.com/liyf1988/GHS`.
- Li, Datta, Craig and Bhadra (2019, arXiv:1903.06768) proposed a similar algorithm for seemingly unrelated regression models.

Global-local priors for non-Gaussian models

- So far: the likelihood is Gaussian (either univariate, or multivariate).
- How do these priors do when the likelihood is changed?
- Datta and Dunson (2016, Biometrika) use global-local priors in a quasi-sparse gamma-Poisson glm. The model is:

$$y_i \mid \theta_i \sim \text{Poisson}(\theta_i), \quad \theta_i \mid \lambda_i, \tau \sim \text{Gamma}(\alpha, \lambda_i^2 \tau^2),$$

  with heavy-tailed prior densities on λ_i and τ.

- Useful alternative to zero-inflated models for count data.

Global-local priors for non-Gaussian models

- Related work in classification problems by Magnusson et al. (2016, arXiv:1602.00260) and Piironen and Vehtari (2017, AISTATS), who use g-l priors in probit and logistic regression model.
- Terenin et al. (2019, Statistics and Computing) classify a million data points in several thousand dimensions in several minutes of running time.
- Their work uses GPU to parallelize sampling the local scale parameters in horseshoe probit regression.

Global-local priors for deep models

- Far fewer works compared to shallow models (nonlinear, non-Gaussian).
- Surprising since deep neural networks are heavily overparameterized with respect to the observed data.
- Full Bayesian inference (Radford Neal style) is non-existent. Computation is too costly. Usual workaround: variational Bayes.

Global-local priors for deep neural networks

- Ghosh and Doshi-Velez (2017, NeurIPS Workshop on DL) and Ghosh et al. (2018, ICML) use horseshoe priors for model selection in deep neural networks.
- Inference is using a variational approximation, first identified by Neville et al. (2014, EJS).
- The trick of writing a half-Cauchy random variable as mixture of two inverse gammas plays an important role.
- Louizos et al. (2017, NeurIPS) establish a connection between horseshoe shrinkage and dropout.

Data augmentation strategies for deep neural networks

- Gan et al. (2015, AISTATS) use Polya-gamma augmentation strategy for deep sigmoidal network.

- Wang, Polson and Sokolov (2019, arXiv:1903.09668) expand the data augmentation strategy for several types of nonlinearities using Gaussian scale mixture techniques.

Nonlinearity	Latent variable
ReLU	Generalized inverse Gaussian
Logistic	Polya-gamma
Check loss	Generalized inverse Gaussian

Horseshoe shrinkage for deep glms

- A conventional glm has $g^{-1}\{E(y \mid X)\} = X\beta$, where $g(\cdot)$ is the link function.
- Thus, the conditional mean of the responses is still a *linear* function of X.
- Tran et al. (2018, arXiv:1805.10157) introduce nonlinearity by replacing X with the output of a deep neural network.
- The model is called DeepGLM and horseshoe priors are used on β to achieve sparsity.

Computational Strategies

- Full Bayes horseshoe shrinkage in linear, Gaussian models is now feasible due to recent breakthrough by Bhattacharya et al. (2016, Biometrika). Complexity reduced from $O(p^3)$ to $O(n^2 p)$.
- Full Bayes is still very hard in deep models, and it is likely to remain that way into the foreseeable future.
- Almost all papers attempting global-local shrinkage in deep models that are known to us use variational approximations or point estimation strategies.

Available software implementations: shallow models

Software with hyperlinked github URL	Relevant Papers	Brief Description of Functionality
MATLAB code: GHS	Li et al. (2017)	Precision matrix estimation in GGMs
MATLAB code: HS-GHS	Li et al. (2019)	Joint mean-covariance estimation in SUR models
Scala code using CUDA: GPUHorseshoe	Terenin et al. (2019)	GPU accelerated Gibbs sampling in probit models
R package: GGMprojpred	Williams et al. (2018)	Projection predictive estimation of GGMs
R package: dsp	Kowal et al. (2019)	Dynamic shrinkage processes

Available software implementations: deep models

Software with hyperlinked github URL	Relevant Papers	Brief Description of Functionality
MATLAB & R code: DeepGLM	Tran et al. (2018)	Fitting DeepGLMs with horseshoe regularization
Python code: HS-BNN	Ghosh and Doshi-Velez (2017)	Horseshoe regularization for Bayesian neural nets
MATLAB code: dsbn	Gan et al. (2015)	Global-local shrinkage in deep sigmoid belief nets
Python code: Bayesian Compression	Louizos et al. (2017)	Bayesian compression for deep learning

References

- **Bhadra, A.**, Datta, J., Li, Y. and Polson, N. G. (2019+). Horseshoe regularization for machine learning in complex and deep models. *(under revision, International Statistical Review)*. [arXiv:1904.10939]
- **Bhadra, A.**, Datta, J., Li, Y., Polson, N. G., and Willard, B. (2019). Prediction risk for the horseshoe regression. *Journal of Machine Learning Research* **20** (78), 1–39.
- **Bhadra, A.**, Datta, J., Polson, N. G., and Willard, B. (2016). Default Bayesian analysis with global-local shrinkage priors. *Biometrika* **103**, 955–969.
- **Bhadra, A.**, Datta, J., Polson, N. G., and Willard, B. (2017). The Horseshoe+ Estimator of Ultra-Sparse Signals. *Bayesian Analysis* **12**, 1105–1131.
- **Bhadra, A.**, Datta, J., Polson, N. G., and Willard, B. (2019). Lasso meets horseshoe: a survey. *Statistical Science* to appear,.
- Li, Y., Craig, B. A., and **Bhadra, A.** (2019). The graphical horseshoe estimator for inverse covariance matrices. *Journal of Computational and Graphical Statistics* to appear,.
- Li, Y., Datta, J., Craig, B. A., and **Bhadra, A.** (2019). Joint mean-covariance estimation via the horseshoe with an application in genomic data analysis. *arXiv preprint* arXiv:1903.06768.