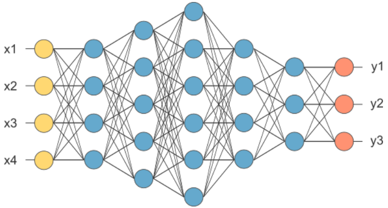
Bayesian Regularization for High Dimensional Models

Lingrui Gan, Naveen N. Narisetty, and Feng Liang

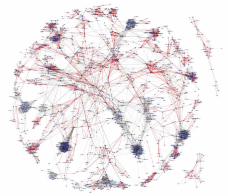
Department of Statistics
University of Illinois at Urbana-Champaign

April 9, 2019
Banff International Research Station

In modern applications in business, science and engineering, statistical models usually have a large number of parameters (high-dimensional models).

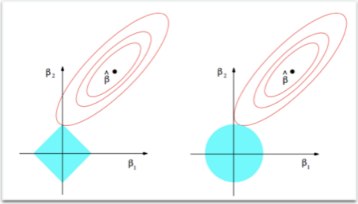


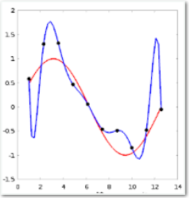
(a) Image source: Quora

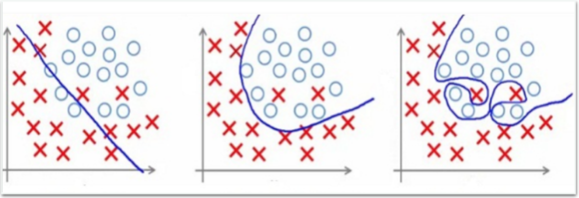


(b) Image source: www.john.ranola.org

Regularization







Penalized Likelihood

Penalized Likelihood Framework

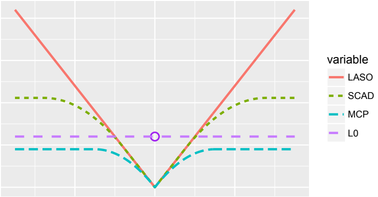
The penalized likelihood framework has the following form:

$$\underbrace{\hat{\Theta}}_{\text{Estimate}} \in \arg\min_{\beta \in \Omega} \Big\{ \underbrace{-\log p(\text{Data} \mid \Theta)}_{\text{Loss function}} + \underbrace{\Omega_\lambda(\Theta)}_{\text{Penalty function}} \Big\}$$

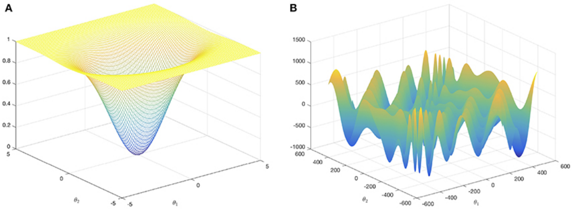
Penalty Functions

- L_0 penalty (aka subset selection) : ideal choice but hard to compute.
- L_1 penalty (aka Lasso)[Tibshirani, 1996]: easy to compute, but biased.
- SCAD [Fan and Li, 2001], MCP [Zhang, 2010]: unbiased, but non-convex.

Popular forms of penalty functions on θ



- Multiple local solutions $\Longrightarrow$ computational and theoretical challenges.



A convex objective function

A non-convex objective function

Image Source: www.frontiersin.org

- [Fan et al., 2014, Wang et al., 2014] studied estimation accuracy of solutions returned by specific algorithms, such as local linear approximation (LLA) algorithm [Zou and Li, 2008].
- [Loh and Wainwright, 2015, Loh and Wainwright, 2017] studied statistical properties of all local solutions satisfying $\|\Theta\|_1 \leq R$.

The Bayesian Approach

In the Bayesian framework, we have a generative model for both data and parameter:

$$\begin{aligned} \text{Prior} \quad &: \quad \pi(\Theta) \\ \text{Likelihood} \quad &: \quad P(\text{Data} \mid \Theta) \end{aligned}$$

where the prior $\pi(\Theta)$ plays the role of a penalty function. In fact,

$$\text{Penalty} = -\log \text{Prior}$$

- The MAP estimate of Θ is the value that maximizes $\pi(\Theta \mid \text{Data})$. Recall

$$\begin{aligned}\pi(\Theta \mid \text{Data}) &= \frac{P(\text{Data} \mid \Theta) \times \pi(\Theta)}{\int P(\text{Data} \mid \Theta) \times \pi(\Theta) d\Theta} \\ &\propto P(\text{Data} \mid \Theta) \times \pi(\Theta)\end{aligned}$$

- So finding MAP is equivalent to minimizing

$$-\log P(\text{Data} \mid \Theta) + \underbrace{\left[-\log \pi(\Theta)\right]}_{\text{Bayesian Penalty}},$$

that is, Prior $= \exp(-\Omega_\lambda(\boldsymbol{\Theta}))$.

- Lasso $\rightarrow \exp(-\lambda|\theta|) \rightarrow$ MAP of Double Exponential Prior.

The priors used in the Bayesian approach can broadly be classified as[1]:

- A single continuous shrinkage prior, such as the Double Exponential prior [Park and Casella, 2008] and the Horseshoe prior [Carvalho et al., 2009];
- Two-group spike-and-slab prior, such as the spike-and-slab Normal prior [George and McCulloch, 1993, Rocková and George, 2014] and spike-and-slab Lasso prior [Rocková and George, 2016b].

There is a lack of unified framework studying the theoretical properties of the aforementioned Bayesian regularization in a general setting.

[1] Here we focus on continuous priors so priors involving point masses are not not discussed.

Outline

- We consider a general class of prior distributions that are scale mixtures of Laplace distributions which includes specific cases of both continuous shrinkage priors and spike-and-slab priors.

- We study the maximum a posteriori (MAP) estimator to obtain insights about the shrinkage corresponding to these priors.

- We show that the regularization induced by these priors is concave (and non-convex) and yet under certain conditions, the MAP estimator is unique and has an optimal rate of convergence in ℓ_∞ norm.

- Although the proposed Bayesian regularization induces a family of non-convex penalty functions, the theoretical results from [Loh and Wainwright, 2017] are not applicable to our study.

  In addition, we do not require the beta-min condition which is required for the estimation accuracy result in [Loh and Wainwright, 2017].

Scale Mixture of Laplace Distributions

$$\pi(\theta) = \int_0^\infty \frac{1}{2v} \exp\Big\{ -|\theta|/v \Big\} dF(v)$$

$$\Longleftrightarrow \begin{cases} \theta \mid v \sim \mathsf{LP}(\cdot \mid v) \\ v \sim F \end{cases}$$

where F is a general (discrete or continuous) distribution function on the positive line.

- **Spike-and-slab Lasso** [Rocková and George, 2016b, Rocková and George, 2016a, Deshpande et al., 2017, Gan et al., 2018]

$$-\log\left(\frac{\eta}{2v_1}\exp\left\{-\frac{|\theta|}{v_1}\right\}+\frac{1-\eta}{2v_0}\exp\left\{-\frac{|\theta|}{v_0}\right\}\right),$$

when $F(v)$ is a discrete distribution with probability mass η on v_1 and $(1-\eta)$ on v_0.

- **Double Pareto** [Armagan et al., 2013]

$$\log\left(1+\frac{|\theta|}{\sigma}\right)^a = a\log\left(1+\frac{|\theta|}{\sigma}\right),$$

when F is an inverse Gamma distribution.

Examples (cont.)

- **Log-shift penalty (LSP)** [Candes et al., 2008]

$$a \log \left(1 + \frac{|\theta|}{\sigma}\right)$$

The marginal prior distribution $\pi(\theta)$ is a **double Pareto** distribution used by [Armagan et al., 2013].

- **Smooth integration of counting and absolute deviation (SICA)** [Lv and Fan, 2009]

$$b\frac{(a+1)|\theta|}{a+|\theta|} = b\frac{|\theta|}{a+|\theta|}I(\theta \neq 0) + b\frac{a}{a+|\theta|}|\theta|$$

Bayesian Regularization Function

The corresponding Bayesian regularization function is given by

$$\rho(\theta) = -\log \pi(\theta) = -\log \left(\int \mathsf{LP}(\theta \mid v) dF(v) \right).$$

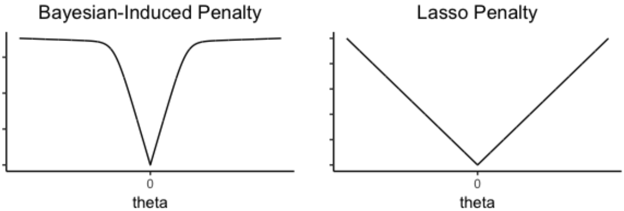


Figure: Figure on the left is from the spike-and-slab Lasso prior.

The Bayesian regularization function is folded concave

Proposition

Let $\eta = 1/v$. When $\theta > 0$, the derivatives of the Bayesian regularization function $\rho(\theta)$ satisfy

$$\begin{cases} \rho'(\theta) = \mathbb{E}(\eta \mid \theta) \\ \rho''(\theta) = -\mathsf{Var}(\eta \mid \theta) \end{cases}$$

provided that the mean and variance exist.

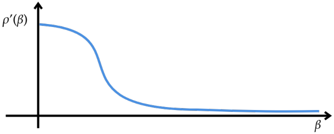


Figure: Gradient of the Bayesian regularization function on the positive real line.

Proof of the Proposition

Throughout assume $\theta \geq 0$ and write $\eta = 1/v$.

$$\begin{aligned}
\pi(\theta) &= \int \frac{\eta}{2} e^{-\eta|\theta|} dF(\frac{1}{\eta}) \\
\pi'(\theta) &= \int (-\eta) \frac{\eta}{2} e^{-\eta|\theta|} dF(\frac{1}{\eta}) \\
\pi''(\theta) &= \int \eta^2 \frac{\eta}{2} e^{-\eta|\theta|} dF(\frac{1}{\eta})
\end{aligned}$$

Then

$$\rho'(\theta) = (-\log \pi(\theta))' = -\frac{\pi'(\theta)}{\pi(\theta)} = \mathbb{E}(\eta|\theta).$$

Similarly

$$\rho''(\theta) = \left[\frac{\pi'(\theta)}{\pi(\theta)}\right]^2 - \frac{\pi''(\theta)}{\pi(\theta)} = -\mathbb{E}(\eta^2|\theta) + \mathbb{E}(\eta|\theta)^2 = -\mathsf{Var}(\eta|\theta).$$

Consider the classical one-dimensional normal mean problem:

$$Z_1, \ldots, Z_n \overset{iid}{\sim} N(\beta, 1) \text{ with prior } \pi(\beta) = \exp\{-\rho(\beta)\}.$$

To find the MAP estimator of the mean parameter β, we minimize

$$\frac{n}{2}(\bar{z} - \beta)^2 + \rho(\beta),$$

Uniqueness

If Var$(\eta \mid \beta) < n$, the objective function is strictly convex:

$$\frac{d^2}{d\beta^2}\left[\frac{n}{2}(\bar{z} - \beta)^2 + \rho(\beta)\right] = n + \rho''(\beta) \geq 0,$$

Sparsity & Adaptive Shrinkage

If $\mathrm{Var}(\eta \mid \beta) < n$, the unique MAP estimator is given by

$$\hat{\beta} = \begin{cases} 0, & \text{when } |\bar{z}| \leq \lambda/n, \\ \left[|\bar{z}| - \frac{\rho'(\hat{\beta})}{n}\right]\mathrm{sign}(\bar{z}), & \text{when } |\bar{z}| > \lambda/n, \end{cases}$$

where $\lambda = \lim_{\beta \to 0+} \rho'(\beta) = \mathbb{E}(1/v|\beta = 0)$.

It leads to desirable shrinkage and selection behavior.

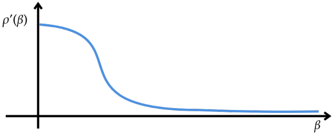


Figure: Gradient of the Bayesian regularization function on the positive real.

A caveat in high dimensions

- One dimensional normal model: with some conditions on the penalty function $\rho(\beta)$, the objective $L_n(\beta) + \rho(\beta)$ becomes convex.
- However, in high-dimensions, conditions on $\rho(\beta)$ alone do not lead to convexity of the objective function.
- For example, for linear regression

$$\hat{\beta} = \arg\min \frac{1}{2}\|Y - X\beta\|^2 + \rho(\beta),$$

the Hessian of the loss function $L_n(\beta)$ is $X^t X$. When $p > n$, the matrix $X^t X$ is at most rank n, i.e., the Hessian matrix has a null space of dimension $p - n$.

Nonconvexity and Uniqueness

In order to study the theoretical properties of our MAP estimator, we adopt the side constraint from [Loh and Wainwright, 2017]:

$$\arg \min_{\|\beta\|_1 \leq R} L_n(\beta) + \sum_{i=1}^{p} \rho(\beta_i). \tag{1}$$

Note: the upper bound R is allowed to increase with n, and the L_1 norm can be replaced by other norms.

Findings

In this constrained space, for a large class of statistical models, the MAP estimator $\hat{\beta}$ is well-behaved.

Theoretical Results

With the following assumptions:

- Assumptions on the likelihood function[a]:
  - Restricted strong convexity
  - Locally Bounded Gradient
  - Locally Bounded Second-order Gradient
  - Conditions on the sampling error $\nabla L_n(\beta^0)$
- Assumptions on the Bayesian regularization function $\rho(\cdot)$[b]

[a] satisfied by linear regression, generalized linear regression, and graphical models

[b] satisfied by the aforementioned priors.

we can show that the MAP estimator $\hat{\beta}$ is unique and

$$\|\hat{\beta} - \beta^0\|_\infty \sim \sqrt{\frac{\log p}{n}},$$

and $\text{supp}(\hat{\beta}) \subseteq S$.

Computation

- (Variational) EM algorithm treating the scale parameters v_j's as latent. [Rocková and George, 2014, Rocková and George, 2016b, Gan et al., 2018]

- Composite gradient descent algorithm [Nesterov, 2013, Loh and Wainwright, 2017].

Conclusion

- We propose a novel class of Bayesian regularization induced from scale mixtures of Laplace priors that include spike-and-slab Lasso priors and the double Pareto priors considered in the Bayesian literature, as well as the LSP and SICA regularization considered in the penalization literature as special cases.
- Our theoretical results proved that the proposed Bayesian regularization enjoys optimal theoretical properties in terms of ℓ_∞-estimation accuracy for a large class of statistical models.

Conclusion

- We propose a novel class of Bayesian regularization induced from scale mixtures of Laplace priors that include spike-and-slab Lasso priors and the double Pareto priors considered in the Bayesian literature, as well as the LSP and SICA regularization considered in the penalization literature as special cases.
- Our theoretical results proved that the proposed Bayesian regularization enjoys optimal theoretical properties in terms of ℓ_∞-estimation accuracy for a large class of statistical models.
- Personal recommendation for Bayesian regularization: spike-and-slab Lasso.

References I

Armagan, A., Dunson, D. B., and Lee, J. (2013).
Generalized double pareto shrinkage.
Statistica Sinica, 23(1):119.

Candes, E. J., Wakin, M. B., and Boyd, S. P. (2008).
Enhancing sparsity by reweighted ℓ_1 minimization.
Journal of Fourier analysis and applications, 14(5-6):877–905.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009).
Handling sparsity via the horseshoe.
In *Artificial Intelligence and Statistics*, pages 73–80.

Deshpande, S. K., Rockova, V., and George, E. I. (2017).
Simultaneous variable and covariance selection with the multivariate spike-and-slab lasso.
arXiv preprint arXiv:1708.08911.

Fan, J. and Li, R. (2001).
Variable selection via nonconcave penalized likelihood and its oracle properties.
Journal of the American statistical Association, 96(456):1348–1360.

Fan, J., Xue, L., and Zou, H. (2014).
Strong oracle optimality of folded concave penalized estimation.
Annals of statistics, 42(3):819.

Gan, L., Narisetty, N. N., and Liang, F. (2018).
Bayesian regularization for graphical models with unequal shrinkage.
Journal of the American Statistical Association, (just-accepted).

George, E. I. and McCulloch, R. E. (1993).
Variable selection via Gibbs sampling.
Journal of the American Statistical Association, 88:881–889.

References II

Loh, P.-L. and Wainwright, M. J. (2015).
Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima.
Journal of Machine Learning Research, 16:559–616.

Loh, P.-L. and Wainwright, M. J. (2017).
Support recovery without incoherence: A case for nonconvex regularization.
The Annals of Statistics, 45(6):2455–2482.

Lv, J. and Fan, Y. (2009).
A unified approach to model selection and sparse recovery using regularized least squares.
The Annals of Statistics, pages 3498–3528.

Nesterov, Y. (2013).
Gradient methods for minimizing composite functions.
Mathematical Programming, 140(1):125–161.

Park, T. and Casella, G. (2008).
The bayesian lasso.
Journal of the American Statistical Association, 103(482):681–686.

Rocková, V. and George, E. I. (2014).
EMVS: The EM approach to Bayesian variable selection.
Journal of the American Statistical Association, 109(506):828–846.

Rocková, V. and George, E. I. (2016a).
Fast Bayesian factor analysis via automatic rotations to sparsity.
Journal of the American Statistical Association, 111(516):1608–1622.

Rocková, V. and George, E. I. (2016b).
The spike-and-slab lasso.
Journal of the American Statistical Association, (just-accepted).

References III

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

Wang, Z., Liu, H., and Zhang, T. (2014).
Optimal computational and statistical rates of convergence for sparse nonconvex learning problems.
Annals of statistics, 42(6):2164.

Zhang, C.-H. (2010).
Nearly unbiased variable selection under minimax concave penalty.
The Annals of statistics, 38(2):894–942.

Zou, H. and Li, R. (2008).
One-step sparse estimates in nonconcave penalized likelihood models.
Annals of statistics, 36(4):1509.