Available online at www.sciencedirect.com

ScienceDirect

Stochastic Processes and their Applications 126 (2016) 3828–3842

stochastic processes and their applications

www.elsevier.com/locate/spa

Sub-optimality of some continuous shrinkage priors

Anirban Bhattacharya[a], David B. Dunson[b], Debdeep Pati[c], Natesh S. Pillai[d,*]

[a] *Texas A& M University, United States*
[b] *Duke University, United States*
[c] *Florida State University, United States*
[d] *Harvard University, United States*

Received 1 August 2015; accepted 18 May 2016
Available online 26 August 2016

Abstract

Two-component mixture priors provide a traditional way to induce sparsity in high-dimensional Bayes models. However, several aspects of such a prior, including computational complexities in high-dimensions, interpretation of exact zeros and non-sparse posterior summaries under standard loss functions, have motivated an amazing variety of continuous shrinkage priors, which can be expressed as global–local scale mixtures of Gaussians. Interestingly, we demonstrate that many commonly used shrinkage priors, including the Bayesian Lasso, do not have adequate posterior concentration in high-dimensional settings.

Keywords: Bayesian; Convergence rate; High dimensional; Lasso; ℓ_1; Lower bound; Penalized regression; Regularization; Shrinkage prior; Sub-optimal

1. Introduction

With the recent flurry of activities in high-throughput data, taking advantage of sparsity to perform statistical inference is a common theme in situations where the number of model parameters (p) increases with the sample-size (n). In such scenarios, penalization methods [3] can yield a point estimate very quickly. There is a rich theoretical literature justifying the optimality

* Corresponding author.
E-mail address: pillai@fas.harvard.edu (N.S. Pillai).

http://dx.doi.org/10.1016/j.spa.2016.08.007
0304-4149/

properties of such penalization approaches [31,26,30,16,23,17], with fast algorithms [10] and compelling applied results leading to their routine use.

On the other hand, statistical theory for characterizing the uncertainty of model parameters using penalization methods in high dimensions has received comparatively less attention [18]. Bayesian approaches provide a natural measure of uncertainty through the induced posterior distribution. Most penalization methods have a Bayesian counterpart. For instance, ℓ_1 and ℓ_2 regularization methods are equivalent to placing zero-mean double-exponential and Gaussian priors respectively on the parameter vector and the solutions of the corresponding optimization problems are precisely the mode of the Bayesian posterior distribution. Moreover, a Bayesian approach has distinct advantages in terms of choice of tuning parameters, allowing key penalty parameters to be marginalized over the posterior distribution instead of relying on cross-validation. Thus a fruitful line of research is to investigate the behavior of the entire posterior distribution of the Bayesian models corresponding to penalization methods.

The process of eliciting prior distributions can be very tricky in high-dimensions. Two-component mixture priors with a point mass at zero are traditionally used in high-dimensional settings because of their ability to produce exact zeros and ease of eliciting hyperparameters based on the prior knowledge about the level of sparsity and the size of the signal coefficients. In [7,6], the authors showed optimality properties for carefully chosen point mass mixture priors in high-dimensional settings. Recently, in an insightful article [22], several arguments were raised against the point mass priors concerning interpretation of exact zeros and computational issues arising from exploring a very high-dimensional model space. This prompted the authors of [22] to seek for continuous analogues of point mass priors based on Gaussian scale mixtures which obviates the need to search over the huge model space. These scale mixtures of Gaussian priors are designed to have a sharp peak near zero with heavy tails so as to emulate the point mass mixture priors. In the last few years, a huge variety of shrinkage priors have been proposed in the Bayesian literature [19,25,13,4,1]. In [1] the authors studied shrinkage priors and provided simple sufficient conditions for posterior consistency in $p \leq n$ settings. However, results on quantifying posterior concentration using continuous shrinkage priors are scarce.

Even from a purely practical point of view, considerable difficulties have arisen when attempts have been made to reflect prior beliefs on sparsity through the associated hyperparameters of these distributions. For example, suppose we wish to estimate $\theta_0 \in \mathbb{R}^n$ from $y \sim \mathrm{N}_n(\theta_0, I_n)$ under the prior knowledge that only a fraction of the coordinates of θ_0 are non-zero. What are the appropriate parameters one should choose in the Bayes Lasso formulation [19] to ensure efficient estimation of θ_0? A first step towards answering such questions is to understand the concentration of shrinkage priors around sparse vectors. This is critically important in two aspects. First, optimal prior concentration is almost necessary for optimal posterior contraction rates under a variety of loss functions. Second, studying the concentration of shrinkage priors around sparse vectors will yield insights into the geometry of shrinkage priors which can then be harnessed for prior elicitation for a broad class of models.

Our contribution in this paper is two fold. First, we obtain sharp bounds for the concentration of continuous shrinkage priors around sparse high dimensional vectors. This is quite challenging because the joint distributions of such priors obtained through integrating several latent hyperparameters are often unwieldy to work with. One of the reasons why the point mass priors enjoy theoretical optimality properties is because they have optimal concentration around sparse vectors [7,6]. We show that the concentration of some of the commonly used continuous shrinkage priors can sometimes be smaller than that of the point mass priors by several orders of magnitude. Second, using these results, we show that for the normal means problem, the Bayesian Lasso [19]