Available online at www.sciencedirect.com

ScienceDirect

Computational Statistics & Data Analysis 51 (2006) 1232–1242

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Bayesian sigmoid shrinkage with improper variance priors and an application to wavelet denoising

Cajo J.F. ter Braak*

Biometris, Wageningen University and Research Centre, Wageningen, 6700 AC, The Netherlands

Received 21 October 2005; received in revised form 18 April 2006; accepted 12 June 2006
Available online 7 July 2006

Abstract

The normal Bayesian linear model is extended by assigning a flat prior to the δth power of the variance components of the regression coefficients $\left(0<\delta\leqslant\frac{1}{2}\right)$ in order to improve prediction accuracy. In the case of orthonormal regressors, easy-to-compute analytic expressions are derived for the posterior distribution of the shrinkage and regression coefficients. The expected shrinkage is a sigmoid function of the squared value of the least-squares estimate divided by its standard error. This gives a small amount of shrinkage for large values and, provided δ is small, heavy shrinkage for small values. The limit behavior for both small and large values approaches that of the ideal coordinatewise shrinker in terms of the expected squared error of prediction, when δ is close to 0. In a simulation study of wavelet denoising, the proposed Bayesian shrinkage model yielded a lower mean squared error than soft thresholding (lasso), and was competitive with two recent wavelet shrinkage methods based on mixture prior distributions.

Keywords: Bayesian model; Improper prior; Shrinkage rule; Wavelet denoising

1. Introduction

Shrinkage of the coefficients in regression models aims at improving the predictive performance of such models. The best known shrinkage methods (Brown et al., 2002; Hastie et al., 2001) are proportional shrinkage (Breiman and Friedman, 1997; Copas, 1983), and methods based on ridge regression (Hoerl and Kennard, 1970) and the lasso (Tibshirani, 1996). In the orthonormal case, the lasso gives the soft thresholding rule and ridge leads to proportional shrinkage (Hastie et al., 2001). These methods share the feature that all coefficients are shrunken regardless of whether the coefficients are small or large in a statistical sense. This is counterintuitive. It appears that this form of shrinkage is not a prerequisite for good predictive power. Bayesian methods that shrink the small coefficients, but shrink the large coefficients only slightly have been devised for wavelet smoothing and have excellent predictive power (Abramovich et al., 2004, 1998; Clyde et al., 1998; De Canditiis and Vidakovic, 2004; Johnstone and Silverman, 2004). These methods are based on the prior belief that only a few coefficients contain the main part of the signal. The prior of the coefficients is specified as a mixture of two distributions. These Bayesian models require at least two hyperparameters (one for the mixing weight and one or more for the component distributions), whereas ridge regression and the lasso have only a single tuning parameter. Extending a proposal by Xu (2003), ter Braak et al. (2005) proposed a Bayesian

* Tel.: +31 317476929; fax: +31 317 483554.
E-mail address: cajo.terbraak@wur.nl.

0167-9473/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2006.06.011

linear model with a single hyperparameter to tune the shrinkage. In this model, each coefficient has a normal prior with its own specific variance. Each variance is given an improper prior that depends on a single hyperparameter δ. This prior is not a limiting case of the common inverse Gamma prior. In an application in genetics the model successfully detected the influential regressors from a large pool (ter Braak et al., 2005).

In this paper I study the shrinkage properties of the model in the case of orthogonal regressors and known error variance. In Section 2, I derive analytic expressions for the posterior distribution of the shrinkage and regression coefficients and show that the expected shrinkage in this model is a sigmoid function of z-ratio (estimate/standard error), with effectively no shrinkage for large z-ratios. As an illustration of the utility of Bayesian sigmoid shrinkage, I apply it in Section 3 to the problem of wavelet shrinkage. I show by simulation that the method yields lower mean squared prediction error than the well-known Sure–Hybrid shrinkage (Donoho and Johnstone, 1995), which uses the lasso. The method is competitive with two recent methods, Bayesian block shrinkage (De Canditiis and Vidakovic, 2004), a convex combination of two ridge shrinkage estimators, and empirical Bayes thresholding (Johnstone and Silverman, 2005). In the discussion shrinkage methods are compared with the ideal shrinker. Bayesian sigmoid shrinkage is the only parameterized method in the comparison that approaches the ideal shrinkage for both very small and very large z-ratios.

2. Bayesian sigmoid shrinkage

Consider the linear regression model

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e} \tag{1}$$

with $\mathbf{y}$ an n-vector of responses of n individuals and $\mathbf{X}$ an $n \times p$ matrix with fixed known values of p predictors for these individuals, $\mathbf{b}$ a p-vector of regression coefficients and $\mathbf{e}$ an n-vector of independent normal noise with zero mean and variance σ^2, which we assume known for now. The jth column of $\mathbf{X}$ is denoted by $\mathbf{x}_j$. The model is made Bayesian by adding a prior distribution for $\mathbf{b}$. We assume that $\mathbf{b}|\mathbf{G} \sim N\left(\mathbf{0}_p, \mathbf{G}\right)$ with $\mathbf{0}_p$ a zero vector and $\mathbf{G} = \operatorname{diag}\left(\sigma_1^2, \ldots, \sigma_p^2\right)$. The new feature of our model, which makes it different from Bayesian ridge regression, is to set independent improper priors for the variance components σ_j^2 $(j = 1, \ldots, p)$ with, for the jth component, $p\left(\sigma_j^2\right) \propto \left(\sigma_j^2\right)^{\delta_j - 1}$ or, equivalently, $p\left(\sigma_j^{2\delta_j}\right) \propto 1$ with $0 < \delta_j \leqslant \frac{1}{2}$. The range restriction to the hyperparameters δ_j guarantees that the posterior distribution for $\mathbf{b}$ is proper, as we will see shortly. Indeed, $\delta_j = 0$ gives $p\left(\log\left(\sigma_j^2\right)\right) \propto 1$, which is known to yield an improper posterior distribution(O'Hagan, 1994) and $\delta_j = \frac{1}{2}$ gives $p\left(\sigma_j\right) \propto 1$, which yields an improper posterior distribution for σ_j^2, but a proper posterior for b_j (see below). The hyperparameters can be taken equal $\left(\delta = \delta_1 = \cdots = \delta_p\right)$ or divided into groups of equal values, as in the wavelet application in Section 3. Because $p(\mathbf{b}|\mathbf{y}, \mathbf{G})$ is multivariate normal, the posterior distribution of the variance components, can be shown (O'Hagan, 1994) to have density

$$p\left(\sigma_1^2, \ldots, \sigma_p^2 \middle| \mathbf{y}\right) \propto \sigma_1^{2(\delta_1 - 1)} \times \cdots \times \sigma_p^{2(\delta_p - 1)} \left|\mathbf{I} + \sigma^{-2}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{G}\right|^{-1/2} \exp\left(\frac{\mathbf{y}^{\mathrm{T}}\mathbf{X}\mathbf{V}^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}}{2\sigma^2}\right) \tag{2}$$

with $\mathbf{V} = \mathbf{X}^{\mathrm{T}}\mathbf{X} + \sigma^2\mathbf{G}^{-1}$. Standard Markov chain Monte Carlo techniques can be used to obtain a sample of this posterior, either by a Metropolis Hasting algorithms or by Gibbs sampling (ter Braak et al., 2005; Xu, 2003).

The posterior mean of $\mathbf{b}$, given $\mathbf{G}$, is $\tilde{\mathbf{b}} = \mathbf{S}\hat{\mathbf{b}}$ with $\hat{\mathbf{b}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$, the least-squares estimator of **b,** and $\mathbf{S}$ the shrinkage matrix (Brown et al., 2002)

$$\mathbf{S} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \sigma^2\mathbf{G}^{-1}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}.$$

To gain insight in the shrinkage properties of the model we consider the special case of orthogonal predictors where $\mathbf{S}$ is a diagonal matrix with shrinkage coefficients

$$s_j = \left(\mathbf{x}_j^{\mathrm{T}}\mathbf{x}_j + \sigma^2 \Big/ \sigma_j^2\right)^{-1} \mathbf{x}_j^{\mathrm{T}}\mathbf{x}_j = f_j\sigma_j^2 \Big/ \left(1 + f_j\sigma_j^2\right), \tag{3}$$