Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics and Data Analysis 52 (2008) 3658–3669
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Predictive performance of Dirichlet process shrinkage methods in linear regression

David J. Nott*

Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546, Singapore

Received 12 June 2007; received in revised form 28 November 2007; accepted 9 December 2007
Available online 9 January 2008

Abstract

An obvious Bayesian nonparametric generalization of ridge regression assumes that coefficients are exchangeable, from a prior distribution of unknown form, which is given a Dirichlet process prior with a normal base measure. The purpose of this paper is to explore predictive performance of this generalization, which does not seem to have received any detailed attention, despite related applications of the Dirichlet process for shrinkage estimation in multivariate normal means, analysis of randomized block experiments and nonparametric extensions of random effects models in longitudinal data analysis. We consider issues of prior specification and computation, as well as applications in penalized spline smoothing. With a normal base measure in the Dirichlet process and letting the precision parameter approach infinity the procedure is equivalent to ridge regression, whereas for finite values of the precision parameter the discreteness of the Dirichlet process means that some predictors can be estimated as having the same coefficient. Estimating the precision parameter from the data gives a flexible method for shrinkage estimation of mean parameters which can work well when ridge regression does, but also adapts well to sparse situations. We compare our approach with ridge regression, the lasso and the recently proposed elastic net in simulation studies and also consider applications to penalized spline smoothing.

1. Introduction

Two commonly used approaches to estimation of mean parameters in linear regression are ridge regression and subset selection. These approaches can achieve reduced mean squared error in estimation and prediction compared to least squares estimation. Ridge regression (Hoerl and Kennard, 1970) works by assuming that the coefficients are exchangeable with a normal prior, where the prior variance controls the degree of shrinkage of parameter estimates towards zero. Shrinkage achieves variance reduction, which may also help to reduce mean squared error of estimation and prediction. On the other hand, subset selection and related approaches (Miller, 1990; Breiman, 1996) achieve variance reduction through dimension reduction by estimating parameters in such a way that some coefficients are exactly zero, eliminating some predictors from the model. In Bayesian variants we may average over different models in prediction (Raftery et al., 1997; Smith and Kohn, 1996) and a prior distribution is used which assigns positive probability to a coefficient being exactly zero.

* Tel.: +65 6516 2744; fax: +65 6872 3919.
E-mail address: standj@nus.edu.sg.

0167-9473/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2007.12.005

In this paper we consider a rather obvious Bayes nonparametric generalization of ridge regression where we model the coefficients in their prior distribution as coming independently from some unknown distribution P which is given a Dirichlet process prior with zero-mean normal base measure. As the precision parameter in the Dirichlet process approaches infinity the method reduces to ridge regression. On the other hand, for finite values of the precision parameter the discreteness of the Dirichlet process means that there is positive probability that coefficients of distinct predictors will be estimated as being equal. If a group of predictors are assigned the same coefficient then effectively we are replacing a group of predictors by their sum, achieving a dimension reduction not unlike variable selection. The precision parameter can be estimated from the data, offering a flexible prior for the regression coefficients which can behave like a ridge-type normal prior or adapt to coefficient sparsity as required. The generalization of ridge regression that we consider here does not seem to have been studied in detail before, despite related applications of the Dirichlet process to shrinkage in estimation of multivariate normal means (Escobar, 1994; MacEachern, 1994), the analysis of randomized block designs (Bush and MacEachern, 1996), and extensions of random effects models in the analysis of longitudinal data (Kleinman and Ibrahim, 1998; Müller and Rosner, 1997). Leslie et al. (2007) consider a scale family constructed from a distribution modelled through a Dirichlet process mixture as an error distribution in applications to heteroscedastic regression.

In recent years there has been renewed interest in methods for shrinkage and variable selection in linear regression due to important applications such as in analysis of microarray gene expression data where the number of predictors is large compared to the number of observations. A recent innovation is the elastic net of Zou and Hastie (2005) which generalizes both ridge regression and the lasso of Tibshirani (1996). The lasso is a shrinkage method which does automatic variable selection but can never select more variables than the number of observations, which may be a disadvantage in applications where the number of predictors exceeds the number of observations. The elastic net overcomes this deficiency by combining ridge-type and lasso-type penalties on coefficients in estimation. Later we show that our Dirichlet process approach retains the good performance of ridge regression for prediction when ridge regression works well, while improving on ridge regression when ridge regression performs relatively poorly. A general approach to shrinkage estimation in linear regression which clusters coefficients may be of considerable interest in gene expression analysis — methods of estimation which cluster coefficients have recently been considered in this context (Tibshirani et al., 2005; Park et al., 2007).

This paper makes two main contributions. First, we investigate a generalization of ridge regression in the linear model using the Dirichlet process prior, and suggest a suitable prior specification on the crucial precision parameter of the Dirichlet process for this application. We also compare our implementation with other approaches to shrinkage and variable selection such as the elastic net and the lasso in simulation studies. Second, we consider flexible function estimation using penalized splines where we replace the usual normal prior on a set of basis function coefficients with an unknown prior which is then given a Dirichlet process prior with normal base measure. We investigate the extent to which the more flexible prior is helpful in function estimation and find that in applications where the smoothness of the function to be estimated is very different in different parts of the predictor space the nonparametric prior can be helpful.

In the next section we describe our Dirichlet process shrinkage regression model and also provide a brief introduction to the Dirichlet process. Section 3 deals with computation, Section 4 discusses some simulation studies comparing our approach to the elastic net and Section 5 discusses application of our approach to flexible regression with penalized splines. Section 6 discusses our conclusions and future work.

2. Dirichlet process shrinkage regression

Consider a linear regression model

$$y = X\beta + \epsilon$$

where y is an n-vector of responses, X is an $n \times p$ design matrix, β is a p-vector of unknown mean parameters and $\epsilon \sim N(0, \sigma^2 I)$ where I denotes the identity matrix and σ^2 is the error variance. We assume columns of X are centred and have length one and that y is centred. Furthermore, X does not contain a column for an intercept (which is sensible if columns of X and y are centred). In a Bayesian analysis of this model we require a prior distribution for (β, σ^2). Here we assume that β and σ^2 are independent in the prior. In the examples later we take $\sigma^2 \sim IG(0.001, 0.001)$ where $IG(a, b)$ denotes an inverse gamma distribution with parameters a and b. This is a noninformative prior given