Journal of the Korean Statistical Society 45 (2016) 307–313

Contents lists available at ScienceDirect

Journal of the Korean Statistical Society

journal homepage: www.elsevier.com/locate/jkss

Scaled ridge estimator and its application to multimodel ensemble approaches for climate prediction

Yongdai Kim *, Young-Oh Kim, Jaeseok Kim, Woosung Kim, Ilsang Ohn

Seoul National University, Republic of Korea

ARTICLE INFO

Article history:
Received 10 September 2015
Accepted 26 November 2015
Available online 6 January 2016

AMS 2000 subject classifications:
primary 62J07
secondary 62P12

Keywords:
Ensemble climate prediction
Global circulation model
Highly correlated covariates
Regularization
Ridge estimator

ABSTRACT

We propose a new regularized estimator called the *scaled ridge* estimator. The scaled ridge estimator is a modified version of the ridge estimator devised to reduce the bias of the ridge estimator by multiplying a positive constant to the ridge estimator. We show theoretically as well as numerically that the scaled ridge estimator performs better than the ridge estimator when the covariates are highly correlated and the true regression coefficients are similar. A motivational example is an ensemble approach for climate prediction based on the global circulation model. By analyzing data sets of monthly precipitation of 10 cites in South Korea, we illustrate that the scaled ridge estimator is a useful and efficient alternative to other competitors for ensemble climate prediction.

© 2015 The Korean Statistical Society. Published by Elsevier B.V. All rights reserved.

1. Introduction

The ridge estimator of Hoerl and Kennard (1970) removes instability of the ordinary least square (OLS) estimator successively, in particular when covariates are highly correlated. Even though it sacrifices the unbiasedness, the ridge estimator reduces the variance much to improve the overall predictability. It is well known that the ridge estimator can be obtained by the penalized least square estimator with the ridge penalty that is the square of the l_2 norm of the regression coefficients.

Even though it is simple, however, the ridge estimator might not be optimal in prediction and could be improved further. In this paper, we propose a modification of the ridge estimator so called the scaled ridge estimator which is obtained by multiplying a constant larger than 1 to the ridge estimator. Note that the ridge estimator has a bias since it is a shrinkage estimator toward 0. By multiplying a constant larger than 1, the amount of the shrinkage is reduced and hence the scaled ridge estimator may have less bias than the ridge estimator. We prove theoretically as well as numerically that the mean squared error (MSE) of the scaled ridge estimator is less than that of the ridge estimator for certain cases.

A motivational example of the scaled ridge estimator is an ensemble approach for climate prediction based on the global circulation model (GCM). Global circulation models are the models that can generate meteorological variables under various emission scenarios. They have well explained the past variations of climate and are used in predicting future climate. Assessment reports of the Intergovernmental Panel on Climate Change (IPCC, 2007) is a main reference for GCMs.

One of the important problems in using GCMs for predicting future climate is that large uncertainties exist in GCMs. For example, GCMs are sensitive to the change of emission scenarios (Mearns et al., 2001). During the last decades, ensemble

* Corresponding author.
E-mail address: ydkim0903@gmail.com (Y. Kim).

http://dx.doi.org/10.1016/j.jkss.2015.11.006
1226-3192/© 2015 The Korean Statistical Society. Published by Elsevier B.V. All rights reserved.

prediction based on GCMs has became an important part of climate prediction as a tool of reducing the uncertainty. There are various ways to combine different GCMs to predict. The simplest approach is to assign the equal weights to GCMs and take the simple average (Lambert & Boer, 2001; Sperber et al., 2004). A more sophisticated approach is to use a multivariate linear regression model where the values simulated by GCMs are treated as covariates and observed data as responses. The regression coefficients obtained from the regression model are used to assign weights to GCMs. In general, ensemble prediction based on multivariate linear regression approaches tend to outperform other ensemble approaches and has been studied extensively by Gneiting, Raftery, Westveld, and Goldman (2005), Kharin and Zwiers (2002), Krishnamurti et al. (1999), Krishnamurti et al. (2000) and Unger, Van Den Dool, O'Lenic and Collins (2009).

However, the standard least square estimator would not be optimal for ensemble prediction since values generated from GCMs tend to be highly correlated. High correlations are expected since various GCMs simulate the common meteorological variable such as temperature or precipitation. It is well known that the variance of the least square estimator is large when the covariates are highly correlated. To reduce the variance of the least square estimator, the ridge estimator is a useful alternative, which is known to outperform the least square estimator in particular when the covariates are highly correlated. However, the ridge regression itself is not fully satisfactory because it produces seemingly unnecessary bias in prediction, and this is the main motivation for developing the scaled ridge estimator. By analyzing climate data, we empirically show that the scale ridge estimator is superior to the OLS estimator as well as the ridge estimator for ensemble prediction.

The paper is organized as follows. We propose the scaled ridge estimator and study various properties in Section 2, and simulation results are given in Section 3. Analysis of real climate data is presented in Section 4 and concluding remarks follow in Section 5.

2. Scaled ridge estimator

Let $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ be n covariate-response variable pairs where $\mathbf{x}_i \in \mathcal{X}$ and $y_i \in \mathbb{R}$, where $\mathcal{X}$ is a subset of $\mathbb{R}^p$. We assume that the response is centered and the predictors are standardized: $\sum_{i=1}^n y_i = 0$, $\sum_{i=1}^n x_{ij} = 0$, $\sum_{i=1}^n x_{ij}^2 = n$ for $j = 1, \ldots, p$.

2.1. Definition

Recall that the ridge estimator is defined as

$$\hat{\boldsymbol{\beta}}^r(\lambda) = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^n (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 + \lambda \sum_{j=1}^p \beta_j^2.$$

The main idea of the ridge estimator is to reduce the variance while sacrificing the bias. Hence, the gain of the ridge estimator compared to the OLS estimator is large when the variance of the OLS estimator is large. However, since there is only one regularization parameter, the ridge estimator may not optimally control the variance and bias simultaneously. Thus, it would be useful to have an estimator in between the ridge and OLS estimator. For this purpose, we propose a new regularized estimator so called the *scaled ridge* estimator, which is defined as

$$\hat{\boldsymbol{\beta}}^s(\lambda, a) = (1 + a)\hat{\boldsymbol{\beta}}^r(\lambda)$$

for $a > 0$. The scaled ridge estimator reduces the variance compared to the OLS estimator by using the ridge estimator, but it reduces the bias compared to the ridge estimator by rescaling the ridge estimator by multiplying $(1 + a)$.

The scaled ridge estimator is related to the elastic net (Enet) estimator of Zou and Hastie (2005). Without the l_1 penalty, the Enet estimator is equal to the scaled ridge estimator with $a = \lambda$. Zou and Hastie (2005) showed that the Enet estimator converges to the univariate estimator as $\lambda \to \infty$, where the univariate regression estimator defined as

$$\begin{aligned}\hat{\beta}_j^{ur} &= \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^n (y_i - \gamma x_{ij})^2 \\ &= \sum_{i=1}^n y_i x_{ij}.\end{aligned}$$

The univariate regression estimator, however, may not be optimal in cases where covariates are highly correlated. Consider an extreme situation where $p = 2$ and $x_{i1} = x_{i2}$ for all i. Suppose that the true model is $y_i = x_{i1} + x_{i2} + \epsilon_i$. That is, the true regression coefficient vector $\boldsymbol{\beta}^*$ is $\boldsymbol{\beta}^* = (1, 1)'$. Since $x_{i1} = x_{i2}$, it is easy to see that $\mathrm{E}(\hat{\beta}_j^{ur}) = 2$ for $j = 1, 2$. Hence, the predictor $\hat{y}_i = \hat{\beta}_1^{ur} x_{i1} + \hat{\beta}_2^{ur} x_{i2}$ based on the univariate regression estimator has $\mathrm{E}(\hat{y}_i^{ur}) = 2(x_{i1} + x_{i2})$, which is twice larger than the expectation of the optimal prediction that is $x_{i1} + x_{i2}$. In fact, the optimal estimator of $\boldsymbol{\beta}$ is given as $\hat{\beta}_j^{ur}/2$ for $j = 1, 2$, which is the OLS estimator assuming equal regression coefficients. The scaled ridge estimator resolves this problem by setting a at other than λ.

Download English Version:

https://daneshyari.com/en/article/1144568

Download Persian Version:

https://daneshyari.com/article/1144568

Daneshyari.com