Comments on: Augmenting the bootstrap to analyze high dimensional genomic data

Connections to the ridge regularized covariance estimator with bagging

Sündüz Keleş · Hyonho Chun

Published online: 21 March 2008
© Sociedad de Estadística e Investigación Operativa 2008

1 Introduction

In a timely and stimulating paper, Tyekucheva and Chiaromonte propose an augmented bootstrap (AB) procedure for estimating inverse covariance matrices within the “large p, small n” paradigm. As authors elaborate in detail, such an estimation problem frequently arises in various areas of genomic research. It is our pleasure to discuss this paper and connect it to a related idea of regularized estimation of covariance matrices motivated by ridge regression (Hoerl and Kennard 1970).

Let $X_1, \ldots, X_n$ denote a random sample of $X \in \mathbb{R}^p$ with probability distribution function $F$. Let $F_n^{(j)}(h, \tau^2)$ denote the empirical distribution of $j$th nonparametric bootstrap sample of size $h$ with a noising parameter of $\tau^2 \geq 0$. Tyekucheva and Chiaromonte’ noising procedure is motivated by the Smoothed Bootstrap procedure (Efron 1979) and adds to each resampled data point an independent draw from $N(0, \tau^2 I_p)$. Consistent with this notation, we will denote the empirical distribution of the original dataset by $F_n(n, 0)$, the sample covariance operator by $\hat{\Sigma}(\cdot)$ and its inverse by $\hat{\Sigma}^{-1}(\cdot)$. Tyekucheva and Chiaromonte’ paper focuses on four different estimators of the inverse of the covariance matrix:

This comment refers to the invited paper available at: http://dx.doi.org/10.1007/s11749-008-0098-6.

S. Keleş (✉)
Departments of Statistics and of Biostatistics & Medical Informatics, University of Wisconsin, Madison, WI 53706, USA
e-mail: keles@stat.wisc.edu

H. Chun
Department of Statistics, University of Wisconsin, Madison, WI 53706, USA
Augmenting the bootstrap to analyze high Moore–Penrose inverse of the sample covariance matrix: 
\[ \hat{\Sigma}^{-1} = \hat{\Sigma}^{-1}(F_n(n, 0)) \]

Bagged Moore–Penrose inverse of the sample covariance matrix by Schäfer and Strimmer (2005): 
\[ \hat{\Sigma}_{BB}^{-1} = \frac{1}{m} \sum_{j=1}^{m} \hat{\Sigma}^{-1}(F_n^{(j)}(n, 0)) \]

Augmented bootstrap estimator: 
\[ \hat{\Sigma}_{AB}^{-1} = \hat{\Sigma}^{-1}(F_n^{(1)}(nm, \tau^2)) \]

Shrinkage estimator of Schäfer and Strimmer (2005): 
\[ \hat{\Sigma}_{SH}^{-1} = [\psi \Psi + (1 - \psi) \hat{\Sigma}(F_n(n, 0))]^{-1} \]

Here, \( m \) denotes the total number of bootstrap samples, \( \psi \in [0, 1] \) is the shrinkage parameter, and \( \Psi \) is a \( p \times p \) positive definite matrix. In what follows, we connect the augmented bootstrap estimator to a regularized estimator of the covariance matrix utilized in ridge regression and augment the simulation results provided by the authors by considering two additional alternatives.

2 Bagged regularized inverse covariance estimator

A well known technique for making a singular matrix non-singular is the addition of a large enough positive number to its diagonals. This forms the basis of the Ridge Regression method (Hoerl and Kennard 1970). Formally, the regularized estimate of the sample covariance matrix is 
\[ \hat{\Sigma}(F_n(n, 0)) + \tau^2 I_p, \]
where \( \tau^2 \) is a tuning parameter. Therefore, in a way, the augmented bootstrap procedure corresponds to a resampling version of this regularized estimator. In addition, the shrinkage estimator \( \hat{\Sigma}_{SH} \) of Schäfer and Strimmer (2005) can be reorganized as:
\[ \hat{\Sigma}_{SH} \propto \hat{\Sigma}(F_n(n, 0)) + \frac{\psi}{1 - \psi} \Psi. \]

When \( \Psi = I_p \) and \( \tau^2 = \psi/(1 - \psi) \in [0, 1] \), this coincides with the ridge regularization of the sample covariance matrix. This indicates that there is a close connection between this estimator and the proposed augmented bootstrap estimator. As both Schäfer and Strimmer (2005) and Tyekucheva and Chiaromonte provide clear empirical evidence for improvement achieved by bagging, we consider a bagged version of the ridge regularized estimator. In addition, the shrinkage estimator \( \hat{\Sigma}_{SH} \) of Schäfer and Strimmer (2005) can be reorganized as:
\[ \hat{\Sigma}_{SH} \propto \hat{\Sigma}(F_n(n, 0)) + \frac{\psi}{1 - \psi} \Psi. \]

Ridge regularized estimator: 
\[ \hat{\Sigma}_{RR}^{-1} = (\hat{\Sigma}(F_n(n, 0)) + \tau^2 I_p)^{-1}. \]

Bagged ridge regularized estimator: 
\[ \hat{\Sigma}_{BRR}^{-1} = \frac{1}{m} \sum_{j=1}^{m} (\hat{\Sigma}(F_n^{(j)}(n, 0)) + \tau^2 I_p)^{-1}. \]

In Table 1, relative squared errors (RSEs) are computed over 500 simulated datasets. The \( \tau^2 \) parameter in this setting is set according to Tyekucheva and Chiaromonte’ Table 3 reflecting the optimal choices for the AB method. Since tuning the smoothing parameter without a specific problem at hand is very difficult, we compare the AB and BRR estimators for various values of \( \tau^2 \in [0.01, 1.2] \). The re-