Test (2008) 17: 25–27
DOI 10.1007/s11749-008-0101-2

DISCUSSION

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

Connections between the augmented bootstrap and the shrinkage covariance estimator

Korbinian Strimmer

Published online: 21 March 2008

1 Introduction

In their enlightening and stimulating paper Svitlana Tyekucheva and Francesca Chiaromonte propose an "augmented bootstrap" (AB) approach to estimate covariance structure in high-dimensional data. They show that the AB estimator performs well in a catalog of examples. Moreover, according to the authors no assumption of a sparsity rationale is made. This is in contrast to a competing and computationally less expensive Stein-type "shrinkage" (SH) approach.

In my comments I address the relationship between the AB and the SH estimators. Perhaps surprisingly, it turns out that there is a very close connection between the two approaches. This leaves questions concerning their relative performance in the examples presented in the paper, an issue which I also discuss below.

2 Relationship between the augmented bootstrap and the shrinkage covariance estimator

The augmented bootstrap (AB) covariance estimator is an extension of the usual empirical estimate and works as follows. Computing the covariance from the empirical distribution F_n leads to the empirical covariance $\hat{\Sigma} = \mathrm{Cov}(F_n)$. However, in a setting with sample size n smaller than the dimension p, this estimator is singular. The problem is circumvented by first generating an augmented bootstrap sample of size

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

K. Strimmer (✉)
Institute for Medical Informatics, Statistics, and Epidemiology, University of Leipzig, Leipzig, Germany
e-mail: strimmer@uni-leipzig.de

$h = mn > p$ from F_n and subsequently adding spherical noise from $N_p(0, \tau^2 I)$ to the sample. The AB covariance estimate is the covariance obtained from the convoluted distribution, i.e.

$$\hat{\Sigma}_{\text{AB}} = \text{Cov}\big(F_{n(mn)} \circ N_p\big(0, \tau^2 I\big)\big).$$

An interesting point of the paper of Tyekucheva and Chiaromonte is that the AB estimator is (nearly) equivalent to bagging using the smoothed bootstrap.

To elucidate another connection, consider the two independent p-dimensional random variates $X \sim F_n$ and $Y \sim N_p(0, I)$, and their sum $Z = X + \tau Y$. We are interested in $\text{Cov}(Z)$. The AB estimator infers $\text{Cov}(Z)$ from nm realizations of Z, i.e. $\hat{\Sigma}_{\text{AB}} = \widehat{\text{Cov}}(z_{(mn)})$. However, there is no need for resampling here, since $\text{Cov}(Z)$ can be computed directly. Recalling that X and Y are independent and that $\text{Cov}(Y)$ is known, we get

$$\widehat{\text{Cov}}(Z) = \text{Cov}(F_n) + \text{Cov}(\tau\, Y) = \hat{\Sigma} + \tau^2 I.$$

Intriguingly, this is a linear shrinkage or ridge estimator with a diagonal "sparse" target and the empirical covariance as unconstrained estimator. It is convenient to substitute $\tau^2 = \frac{\lambda}{1-\lambda}$, so that the intensity $\lambda \in [0, 1]$ controls the amount of shrinkage.

Hence, the AB approach is essentially a Monte Carlo approximation to a linear shrinkage estimator of the covariance. The benefits of resampling over analytic calculations, especially if the latter are straightforward, are not obvious.

3 Importance of the choice of shrinkage intensity and smoothing variance

Having understood that both the AB and the SH procedures essentially rely on the very same underlying shrinkage principle, one needs to ask why the two estimators perform so notably different in the applications of the paper. In my view, the reason for this discrepancy is related to the choice of the parameters τ and λ.

This is particularly evident in the microarray data analysis. For the AB estimator the smoothing parameter is *fixed* at $\tau = 0.5$, whereas in the case of the SH estimator the shrinkage intensity λ is *estimated* from the data. This is a key methodological difference, as the first procedure essentially amounts to employing a Bayesian approach with fixed prior chosen in advance, whereas the latter corresponds to an *empirical* Bayes method. It is well know that estimation of smoothing parameters or shrinkage intensities is a hard problem.

When computing partial correlations by inversion of the shrinkage covariance matrix, different choices of smoothing parameter do lead to changes in the ordering of the partial correlation coefficients. Hence, it is not surprising that the choice of $\tau = 0.5$ allows a set of genes to be identified that are top ranking for the AB but not for the SH estimator. Specifically, for the analyzed data it appears that the SH estimator has applied a higher shrinkage intensity than $\tau = 0.5$ or, equivalently, $\lambda = 0.2$. Therefore, I suggest that for a more even comparison of the SH and AB estimators both need be employed with equivalent settings for λ and τ (or at least the shrinkage intensity for the SH estimator should be reported).

The picture is similar in the simulation study. The AB estimator is employed using a smoothing variance τ^2 optimized versus the truth on a grid of possible values, whereas the SH estimator estimates λ from the data—without reference to the true values. In any case it is clear why the SH estimator performs best when there is little misspecification between diagonal target and the data—this reflects exactly how the shrinkage intensities are being estimated, see the corresponding discussion in Schäfer and Strimmer (2005).

4 The shrinkage covariance estimator implemented in "corpcor"

Finally, I'd like to note that the estimator implemented in the R package "corpcor" does *not* shrink the covariance matrix towards the unit diagonal target, as implied by Eq. (3) in the paper of Tyekucheva and Chiaromonte, or as in the above formula for the shrinkage pendant of the AB estimator.

Rather, our procedure follows Barnard et al. (2000) in that it *separately* shrinks *correlations* (to zero, Schäfer and Strimmer 2005) and *variances* (towards their median, Opgen-Rhein and Strimmer 2007), which gives an estimator of the form $s_{kl}^\star = r_{kl}^\star \sqrt{v_k^\star v_l^\star}$ with $k, l = 1, \dots, p$ and components

$$r_{kl}^\star = (1 - \hat{\lambda}_1^\star) r_{kl},$$
$$v_k^\star = \hat{\lambda}_2^\star v_{\text{median}} + (1 - \hat{\lambda}_2^\star) v_k.$$

The shrinkage intensities are estimated via

$$\hat{\lambda}_1^\star = \min\left(1, \frac{\sum_{k \neq l} \widehat{\text{Var}}(r_{kl})}{\sum_{k \neq l} r_{kl}^2}\right),$$
$$\hat{\lambda}_2^\star = \min\left(1, \frac{\sum_{k=1}^p \widehat{\text{Var}}(v_k)}{\sum_{k=1}^p (v_k - v_{\text{median}})^2}\right).$$

Note that all functions for shrinkage estimation available in "corpcor" report the intensities $\hat{\lambda}_1^\star$ and $\hat{\lambda}_2^\star$.

References

Barnard J, McCulloch R, Meng XL (2000) Modeling covariance matrices in terms of standard deviations and correlations, with applications to shrinkage. Stat Sin 10:1281–1311

Opgen-Rhein R, Strimmer K (2007) Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach. Stat Appl Genet Mol Biol 6:9

Schäfer J, Strimmer K (2005) A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Stat Appl Genet Mol Biol 4(1):32