Anne-Laure Boulesteix · Athanassios Kondylis · Nicole Krämer

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

Received: date / Accepted: date

Svitlana Tyekucheva and Francesca Chiaromonte provide an attractive solution to the problem of the estimation of the inverse covariance matrix with high-dimensional data and small samples, which is an important challenge in modern bioinformatics.

1 Optimizing the noise parameter

Our first comment is on the optimization of the model parameter τ controlling the amount of noise in the augmented bootstrap method. In a supervised prediction problem, τ can and should be optimized using, e.g., a cross-validation (CV) procedure, as suggested by the authors. If the prediction accuracy is itself evaluated by cross-validation or a related approach, this yields a *nested* cross-validation procedure involving an inner-loop (in which the parameter is tuned) and an outer-loop (in which the prediction rule with tuned parameter is evaluated), see Statnikov et al (2005) and Boulesteix (2007). Note that different cross-validation schemes can yield different results due to, e.g., the difference in the size of the considered training subsets. In leave-one-out cross-validation, the training subsets have size $n-1$, whereas they have size $n/2$ in 2-fold cross-validation. In the case considered here, it is conceivable that cross-validation schemes with many folds (i.e. with large

A.-L. Boulesteix
Sylvia Lawry Centre for Multiple Sclerosis Research, Hohenlindenerstr. 1, D-81677 Munich, Germany, E-mail: boulesteix@slcmsr.org
A. Kondylis
Institute of Statistics, University of Neuchâtel, Pierre-à-Mazel 7, CH-2000 Neuchâtel, Switzerland
N. Krämer
Machine Learning/Intelligent Data Analysis Group, Technical University of Berlin FR 6-9, Franklinstr. 28-29, D-10587 Berlin, Germany

training subsets) yield smaller optimal values of τ (i.e. less noise) than cross-validation schemes with less folds, as indirectly suggested by Tyekucheva and Chiaromonte in the right panel of Figure 1.

Cross-validation is a standard procedure for tuning parameters in supervised problems. For example, it may be used to optimize the model parameter u of the method by Cook et al (2007). In general, an optimized parameter is expected to yield better performance than a parameter value fixed without consideration of the data, but the obtained accuracy may be worse than if the parameter value is chosen optimally a posteriori, i.e. without inner cross-validation. In the case of an unsupervised problem such as the estimation of the partial correlation matrix (Schäfer and Strimmer, 2005), the choice of the "best" τ is a more difficult and hazardous task. As suggested by Figure 1, the estimation accuracy may depend heavily on τ. The classical cross-validation procedure can not be applied here. Since the value of the optimal τ is expected to depend, among others, on the covariances to be estimated, we have in a way to do with a chicken and egg paradox.

One option is to turn this unsupervised problem into p supervised problems by successively considering each of the p variables as a response to be predicted by regression based on the $p-1$ remaining variables. This approach is related to the regression-based estimation of partial correlation coefficients. For the i-th variable, the ordinary least squares estimator of the regression coefficients is given as $(X_{-i}^T X_{-i})^+ X_{-i}^T X_i$. For each of these linear regressions, the term $(X_{-i}^T X_{-i})^+$ can be replaced by the result of the augmented bootstrap procedure, using different values of τ successively. One can then select the value of τ minimizing the mean squared prediction error over the p regressions in cross-validation settings. This approach, though probably suboptimal and computationally intensive, could at least give a valuable approximation of the optimal τ value to be used in the unsupervised problem.

2 Bootstrap

The second point that we would like to discuss is the comparison between bootstrap samples of size n drawn with replacement and subsamples of size $< n$ drawn without replacement. Bickel and Ren (2001) point out that bootstrap hypothesis testing fails when performed based on bootstrap samples. As outlined by Strobl et al (2007), the use of bootstrap samples also potentially leads to substantial biases when used for variable selection or for the calculation of variable importance measures in random forests. Hence, Strobl et al (2007) recommend drawing subsamples rather than bootstrap samples when building a random forest. In the same vein, Binder and Schumacher (2007) show that complexity selection in bootstrap samples drawn with replacement is biased towards more complex models in many settings. It would be interesing to check whether similar biases occur in the situation considered by Tyekucheva and Chiaromonte, where bootstrap samples are used for estimation. In case they occur, one could consider data sets composed of m noised versions of the original sample instead of noised bootstrap samples.

3 The CLC approach for supervised problems

The following comments refer to the approximation of the solution of

$$\Sigma x = \nu \tag{1}$$

in terms of matrix-vector multiplications of the form $\Sigma^i \nu$ (Cook et al, 2007). As the authors point out, this method is similar to Partial Least Squares (PLS) (Wold, 1975). We would like to briefly clarify this connection, which is established via the conjugate gradient (cg) algorithm (Hestenes and Stiefel, 1952). It follows directly from Eq. (9) in the paper by Tyekucheva and Chiaromonte that the method proposed by Cook et al (2007) is equivalent to the cg method, a classical approach in numerical linear algebra. It is an iterative procedure to compute approximate solutions of linear equations by minimizing the loss function $x^T \Sigma x - 2\nu^T x$ on the Krylov subspace of dimension m spanned by the vectors $\Sigma^i \nu$ for $i = 0, \ldots, m-1$. The relationship between Krylov spaces and PLS is already outlined in Helland (1988) and Helland (1990), and it can be shown that the solution of PLS regression (with m components) equals the approximate solution of Eq. (1) found via cg after m iterations, with Σ denoting the sample covariance matrix and ν denoting the vector of sample cross-covariance between the predictor variables and the response variable.

When the response is binary, SIR sets the term ν to the difference of the two sample class means modulo normalization, which equals the cross-covariance between the predictor variables and the response variable up to the weighting of the class means. Hence, PLS is almost equivalent to the CLC approach by Cook et al (2007) outlined in Eq. (9) in the case of a binary response, and even perfectly equivalent if the two classes are equally sized. Note that the optimization criterion for PLS is usually modified in the case of multiple class prediction (Barker and Rayens, 2003; Rosipal and Krämer, 2006), which leads to orthogonal PLS.

Although commonly used in numerical linear algebra, the cg approach is not well-known as a statistical tool. However, we strongly believe that transferring these methods into a statistical framework will enrich the field of data analysis for very high dimensional data, and we truly appreciate the contribution of the authors to this field. The merits of cg as a *regularized* estimation technique have been exploited for statistical analysis only recently. For instance, Ide and Tsuda (2007) use Krylov subspace learning for change point detection. Krämer et al (2007) introduce penalized PLS and relate it to a preconditioned conjugate gradient method. Kondylis and Whittaker (2007) set up spectral preconditioners in order to combine Principal Component Analysis and PLS in a unified statistical framework, and use preconditioned conjugate gradients to improve prediction and dimensionality reduction. The penalization/preconditioning approach might also be applied to the analysis of high-dimensional data with a functional structure. As pointed out in Krämer et al (2007), PLS with a penalization term is equivalent to solving the normal equation with a preconditioning matrix that is defined as $M = (I + P)^{-1}$, where P denotes a penalty matrix. This approach might also be beneficial for the classification task discussed in the

paper by Tyekucheva and Chiaromonte if the data have a longitudinal or a time structure.

We conclude with a comment regarding the implementation of the proposed method. For high-dimensional data, the size of the linear equations (1) is huge, since the dimension equals the number of variables. The matrix-vector multiplications $\Sigma^i \nu$ scale quadratically with the number of variables, leading to heavy computations, as noted by the authors. Note however that we can apply the kernel trick (Schölkopf et al, 1998) to solve this problem. In a nutshell, we use the fact that the (approximate) solution of Eq. (1) is a linear combination of the n observations. We can then derive the dual representation of Eq. (1), which is a linear equation involving the $n \times n$ kernel matrix of pairwise inner products between observations. This leads to a so-called kernel representation of Eq. (1) and a kernel representation of the algorithm proposed by Cook et al (2007). Since it scales with the number of observations, which is typically much smaller than the number of variables, the kernel trick can reduce the computation time dramatically. In the context of the proposed cg method (Cook et al, 2007), this leads to two alternatives. We can either use the conjugate gradient algorithm for the dual representation of Eq. (1) or the kernel representation of the initial algorithm, which both scale quadratically with respect to n. Note that these two approaches are different (Krämer and Braun, 2007).

References

Barker M, Rayens WS (2003) Partial Least Squares for discrimination, Journal of Chemometrics, 17:166-173

Bickel PJ, Ren JJ (2001) The bootstrap in hypothesis testing. In *State of the Art in Probability and Statistics, Festschrift for Willem R. van Zwet, IMS Lecture Notes Monograph Series, Beachwood, OH, USA*. Edited by: de Gunst M, Klaassen C, van der Vaart A, 36:91-112

Binder H, Schumacher M (2007) Adapting prediction error estimates for biased complexity selection in high-dimensional bootstrap samples, FDM-Preprint 100, University of Freiburg

Boulesteix AL (2006) Reader's reaction to "Dimension reduction for classification with microarray gene expression data". Statistical Applications in Genetics and Molecular Biology 5:16

Boulesteix AL (2007) WilcoxCV: an R package for fast variable selection in cross-validation. Bioinformatics 23:1702-1704

Boulesteix AL, Strimmer K (2007) Partial Least Squares: A versatile tool for the analysis of high dimensional genomic data. Brief. Bioinf. 8:24-32

Cook RD, Li B, Chiaromonte F (2007) Dimension reduction in regression without matrix inversion. Biometrika 94(3) 569 – 584

Helland IS (1988) On the structure of partial least squares regression, Communications in Statistics - Simulation and Computation, 17:581-607

Helland IS, (1990) Partial Least Squares Regression and statistical models, Scandinavian Journal of Statistics, 17:97-114

Hestenes M, Stiefel E (1952) Methods for conjugate gradients for solving linear systems, Journal of Research of the National Bureau of Standards 49, 409 - 436

Ide T, Tsuda K (2007) Change-point detection using Krylov subspace learning, Proceedings of the SIAM Internations Conference on data mining, 515-520

Kondylis A, Whittaker J (2007) Spectral preconditioning of Krylov spaces: combining PLS and PC regression, Computational Statistics and Data Analysis, to appear

Krämer N, Boulesteix AL, Tutz G (2007) Penalized partial least squares with applications to B-splines and functional data, submitted

Krämer N, Braun ML (2007) Kernelizing PLS, degrees of freedom, and efficient model selection, Proceedings of the 24th International Conference on Machine Learning, 441 - 448

Rosipal R, Krämer N (2006) Overview and recent advances in PLS.In: Subspace, Latent Structure and Feature Selection Techniques, Springer, 34–51

Schäfer J, Strimmer K (2005) A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statistical Applications in Genetics and Molecular Biology 4:32

Schölkopf B, Smola A, Müller KR (1998). Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation 10 (5), 1299-1319

Statnikov A, Aliferis CF, Tsamardinos I, Hardin D, Levy S (2005) A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. Bioinformatics 21:631-643

Strobl C, Boulesteix AL, Zeileis A, Hothorn T (2007) Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinformatics 8:25

Wold H (1975) Path models with latent variables: the NIPALS approach. In: Quantitative Sociology: International Perspectives on Mathematical and Statistical Model Building, Academic Press, 307–357