BIOCYBERNETICS AND BIOMEDICAL ENGINEERING 35 (2015) 64–67

Available online at www.sciencedirect.com

ScienceDirect

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

Original Research Article

Exploiting Stein's paradox in analysing sparse data from genome-wide association studies

Zdeněk Valenta *, *Jan Kalina*

Institute of Computer Science, Academy of Sciences of the Czech Republic, Prague, Czech Republic[1]

ARTICLE INFO

Article history:
Received 31 July 2014
Accepted 19 October 2014
Available online 4 November 2014

Keywords:
Multivariate analysis
Shrinkage
Biased estimation
Risk
Squared-error loss
Bias-variance trade-off

MSC:
00-01
99-00

ABSTRACT

Unbiased estimation appeared to be an accepted golden standard of statistical analysis ever until the Stein's discovery of a surprising phenomenon attributable to multivariate spaces. So called Stein's paradox arises in estimating the mean of a multivariate standard normal random variable. Stein showed that both natural and intuitive estimate of a multivariate mean given by the observed vector itself is not even admissible and may be improved upon under the squared-error loss when the dimension is greater or equal to three. Later Stein and his student James developed so called 'James–Stein estimator', a shrunken estimate of the mean, which had uniformly smaller risk for all values in the parameter space. The paradox first appeared both unintuitive and even unacceptable, but later it was recognised as one of the most influential discoveries of all times in statistical science. Today the 'shrinkage principle' literally permeates the statistical technology for analysing multivariate data, and in its application is not exclusively confined to estimating the mean, but also the covariance structure of multivariate data. We develop shrinkage versions of both the linear and quadratic discriminant analysis and apply them to sparse multivariate gene expression data obtained at the Centre for Biomedical Informatics (CBI) in Prague.

1. Introduction

Gene expression data obtained from Genome-Wide Association Studies (GWAS) are characterised by both its high dimensionality and sparseness. That means that the number of dependent variables involved in statistical modelling (p) is very large, typically tens of thousands of variables corresponding to genes and their transcripts, and is greatly exceeding the number of observations (n), typically only a few hundreds of records ($p \gg n$). In analysing such data we exploit the consequences of Stein's paradox in estimating the covariance structure and overcome singularity of the empirical covariance matrix in defining regularised (shrinkage) versions of both the Linear (LDA*) and Quadratic Discriminant Analysis (QDA*).

In 1956 Stein [1] presented his discovery proving inadmissibility of the usual estimator for the mean of multivariate normal distribution when the dimension of multivariate

[1] http://www.ustavinformatiky.cz

* *Corresponding author at:* Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 2, 182 07 Prague, Czech Republic

http://dx.doi.org/10.1016/j.bbe.2014.10.004
0208-5216/

normal space is $p > 2$. Five years later, James and Stein [2] published their 'shrinkage estimator' of the mean which shrinks estimated mean of the multivariate normal distribution X towards 0. Assuming $X \sim N(\theta, I_p)$, they obtained:

$$\delta_{JS}(X) = \left(1 - \frac{p-2}{||X||^2}\right)X, \tag{1}$$

Results of Baranchik [3] and Bock [4] rendered the shrinkage estimator of the mean of the p-variate normal distribution $X \sim N(\theta, \Sigma)$, where Σ is a positive definite covariance matrix:

$$\delta_{pJS}(X) = \left(1 - \frac{\hat{p}-2}{X^T S^{-1} X}\right)^+ X, \tag{2}$$

where $S = \hat{\Sigma}$ and $\hat{p}$ is the effective dimension of Σ which equals the trace of S divided by its maximum eigenvalue.

Particularly with the development of statistical methodology for analysing multivariate genomic data there appeared a growing interest in applying Stein's principles at the level of the covariance structure of sparse multivariate data. In most situations, empirical covariance and correlation estimators are ill-suited for this purpose. When estimating the inverse of the covariance matrix in the context of sparse multivariate data one needs to overcome natural singularity of the empirical covariance matrix. In the 'small n, large p' data setting growing number of eigenvalues become zero and estimated covariance matrix cannot be inverted. As noted by Schäfer and Strimmer [5], inferring large-scale covariance matrices from sparse genomic data is a difficult and ever-present problem in bioinformatics. They introduced their improved estimator S^* of the $p \times p$ covariance matrix Σ by adopting the linear shrinkage approach combined with analytic determination of the shrinkage intensity according to the Ledoit-Wolf theorem [6].

The linear shrinkage estimator S^* of Σ and the corresponding risk function $R(.)$ are shown in Eqs. (3) and (4), respectively. The S^* represents a weighted average of a target matrix T representing a biased constrained estimator of Σ with simplified covariance structure, and an unbiased unconstrained estimate of Σ, the conventional empirical covariance estimator S.

$$S^* = \lambda T + (1-\lambda)S \tag{3}$$

$$\begin{aligned} R(\lambda) &= E(L(\lambda)) = E||S* - \Sigma||_F^2 = E||\lambda T + (1-\lambda)S - \Sigma||_F^2 \\ &= E\sum_{i=1}^{p}\sum_{j=1}^{p}[\lambda t_{ij} + (1-\lambda)s_{ij} - \sigma_{ij}]^2, \end{aligned} \tag{4}$$

where $\lambda \in [0, 1]$ and t_{ij}, s_{ij} and σ_{ij} are the corresponding elements of matrices T, S and Σ. An optimal linear combination of a constrained (biased) estimator T and an unconstrained (unbiased) estimator S of Σ minimises the risk using the squared error loss function $L(.)$ under Frobenius norm $||A||_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{i,j}|^2}$. Thus the unbiasedness principle is here traded-off with that of the smallest mean square error.

In line with Stein's original finding we wanted to show that exploiting shrinkage estimators of the covariance matrix is not only meaningful when the empirical covariance matrix becomes singular, so there is no other remedy in obtaining an invertible estimator, but also when it is non-singular while its dimension is much higher than two, say 10 or more.

For the class prediction purposes we use the gene expression microarray data obtained from the CBI's GWAS study of stroke. The data involves gene expression profiles of patients with the stroke history and their paired controls matched on major clinical risk factors. We compare the performance of two regularised class prediction methods (LDA* and QDA*) with their ordinary counterparts (LDA and QDA, respectively) and several competing alternatives.

We use Schäfer's and Strimmer's shrinkage estimator [5] of the covariance matrix in deriving the shrinkage version of the Mahalanobis distance [7] which is inherent in the discriminant scores of LDA and QDA. Linear discriminant scores of the shrunken LDA* are shown in Eq. (5), where π_k stands for a prior probability of x belonging to the k-th population and $\bar{x}_k$ represents the mean within that population:

$$\delta_k^*(x) = (x - \bar{x}_k)\prime S*^{-1}(x - \bar{x}_k) - 2\log\pi_k \tag{5}$$

We focus on the dimensionality reduction and class prediction and demonstrate improved classification properties of the shrinkage versions LDA* and QDA* with respect to their ordinary counterparts. We also compare the class prediction performance of these methods with several established alternatives.

2. Methodology

Study samples constituted of cerebrovascular stroke (CVS) cases admitted to the ICU of the Municipal hospital in Èáslav, Czech Republic, between September 2006 and January 2011, and their paired controls matched on age, gender and hypertension status. In analysing our data we first reduced their dimensionality, i.e. 'the genes dimension'. Samples from a genome-wide study included almost fifty thousands gene transcripts available for use in the class prediction. Randomly selected two thirds of available samples were used as the training dataset for the classification algorithms under consideration. The remaining third of the data served as a validation dataset. These two steps proceeded over 100 simulations.

We analysed changes in predictive performance of the linear and quadratic discriminant analysis stemming from incorporating the shrinkage estimates of the covariance.

In our applications the empirical covariance matrices were shrunken towards a diagonal matrix with generally unequal variances. We also compared the performance with several alternative class prediction algorithms used in this context.

Dimensionality reduction was achieved using the 'Prediction Analysis for Microarrays' (PAM [8]) and 'Linear Models for Microarray Analysis' (*limma* [9]). For the purpose of class prediction we used the LDA*, QDA* and their ordinary counterparts, logistic regression (LR) and PAM. In order to achieve the study goals we performed a simulation study which involved evaluating the sensitivity, specificity and Youden's index of the class prediction techniques over 100