Big Data, Biostatistics and Complexity Reduction

Jan Kalina*

Institute of Computer Science of the Czech Academy of Sciences, Prague, Czech Republic

Abstract

The aim of this paper is to overview challenges and principles of Big Data analysis in biomedicine. Recent multivariate statistical approaches to complexity reduction represent a useful (and often irreplaceable) methodology allowing performing a reliable Big Data analysis. Attention is paid to principal component analysis, partial least squares, and variable selection based on maximizing conditional entropy. Some important problems as well as ideas of complexity reduction are illustrated on examples from biomedical research tasks. These include high-dimensional data in the form of facial images or gene expression measurements from a cardiovascular genetic study.

Keywords

Biostatistics; Big data; Multivariate statistics; Dimensionality; Variable selection

Correspondence to:

Jan Kalina

Department of Machine Learning, Institute of Computer Science of the Czech Academy of Sciences, Praha 8, Czech Republic.

E-mail: kalina@cs.cas.cz

EJBI 2018; 14(2):24-32

Received: February 19, 2018

Accepted: March 16, 2018

Published: April 30, 2018

1 Introduction

Big Data in biomedicine represent an important and perspective but still not sufficiently utilized capital with a potential to improve the diagnosis, prognosis and therapy for individual patients. A proper biostatistical analysis of Big Data is one of key components (or even accelerators) of the development of reliable clinical decision support tools and has to face difficult challenges [1, 2].

Exploiting Big Data in biomedical research and practice also allows to contribute to improving the decision making process of clinical decision support, which requires to solve classification tasks. The aim is to learn a classification rule over a training dataset allowing to assign a new sample (individual) to one of the groups, e.g. according to the diagnosis and thus to decide for a particular diagnosis. It is however not so common that biomedical data have a very larger number of observations n (i.e, the number of samples or patients). More commonly, Big Data in biomedicine have the form of high-dimensional data with a small or moderate n, but a large number of variables p (symptoms and signs, results of biochemical or laboratory measurements etc.).

If Big Data are also contaminated by noise, which is a typical situation, then a pre-processing and cleaning the data together with a consequent complexity reduction represent crucial preliminary steps of each analysis [3, 4]. Typical examples of such applications, which cannot be appropriately analyzed by standard methods, include molecular genetic studie, functional magnetic resonance imaging (fMRI) of brains [5], or longitudinal data.

This paper can be understood as an attempt to formulate our experience with analyzing biomedical Big Data, especially from the point of view of reducing their complexity, which makes the process of data analysis more accessible by means of multivariate statistical tools. We recall basic challenges of Big Data analysis, overview major approaches to their complexity reduction and illustrate their principles on some recent biomedical research tasks. Section 2 introduces the problem of Big Data and its analysis in biomedicine. Section 3 presents basic principles of complexity reduction. Particular methods are described in the consequent sections accompanied by examples, namely the principal component analysis in Section 4 for variable selection by maximal conditional entropy in Section 5. Finally, partial least squares for regression are described in Section 6, which are often transferred also to classification problems. Hypothesis tests are discussed in Section 7.

2 Big Data in Biomedicine

Origins of multivariate statistics trace back to Karl Pearson (1857-1936), who developed the first multivariate statistical methods for the needs of anthropology and forensic science. The first applied statisticians of the beginning of the 20th century are at the same time well known by biologists. In 1911, Pearson founded the first statistical department at the world in London called Department of Applied Statistics, where

he was appointed as professor. The interdisciplinary department included also biometric and eugenic laboratories and the boundary between statistics and biological sciences was not clear at that time. Pearson introduced also the concept of contingency tables, which is until now a general statistical concept for tables of counts (i.e, of a discrete variable). Pearson analyzed them when trying to prove the evolution theory by means of statistical methods. Pearson defined contingency as a broadly discussed phenomenon describing randomness or unpredictability within evolution with influence on the whole species of organisms. The concept of contingency tables remains in statistics as one statistical concepts with a biologically motivated name.

The amount of data with a potential to improve healthcare for an individual patient rises enormously. Such big data represent a valuable capital with an opportunity for a dramatic change of current practices of health care, accelerating the development of information-based medicine [3, 4]. So far, big data in the psychiatric context are measured primarily for the research purposes, while they have the potential to contribute to improving the efficiency of clinical decision making and patient safety.

Low-level computer tasks applied on the new types of data (including Big Data) have been described in recent monographs on health informatics [6]. So far, intensive attention has paid to technological aspects concerning the storage of big medical data in large databases and their transfer, protection (data security issues), sharing, lossless compression, information retrieval, and appropriate visualization. An important issue is also integration of various e-health systems allowing integrating individual data with data about the current care, brain imaging results, presence of risk gene variants etc.

The analysis of the clinical data by means of methods of multivariate statistics and data mining becomes a necessity as the volume of data namely grows not only rapidly but also much faster than the ability to analyze and interpret them. Unfortunately, the crucial important question how to acquire new medical knowledge by a proper analysis of big medical data reliably has obtained less attention. An example of an improper interpretation of statistical result is the paper by Nordahl H et al. [7], where low education is denoted as a risk factor of cerebrovascular stroke, while it is only an instrument associated with true risk factors (e.g. lifestyle or stress).

Traditional statistical methods are unsuitable for any form of Big Data. Therefore, dimensionality reduction (complexity reduction, variable selection) is generally recommended as the initial step of the analysis of data with a large number p of variables observed over n samples. It can actually improve the result of a subsequent analysis in spite of losing some relevant information.

On the other hand, the idea of parsimony (i.e, reducing the set of variables to a too small number of relevant ones) has been also criticized [8] and variable selection may be optimized for a classification or clustering context [9]. Experience of applied researchers is critical for not being as good as presented in theoretical papers on simulated data [10]. We also should not leave out that in some clinical fields, data analysis has to face their own challenges and fulfill field-specific requirements.

3 Overview of Complexity Reduction

Complexity reduction is a general concept including any approach to simplifying the analysis of data of various forms, e.g. finding suitable relevant features from medical images of the brain, voice records, narrative text of health reports etc. Various types and formats of biomedical data require a broad spectrum of sophisticated methods for their analysis.

In recent years, new specific complexity reduction approaches have proposed within the fields of multivariate statistics, computer science (machine learning) or information theory [11, 12, 13, 14]. This section recalls the most common methods used in biomedical applications, however only for the context of numerical data. Then, the concept of complexity reduction is usually replaced by dimensionality reduction, which can be understood as a more specific version of a general complexity reduction.

In the whole paper, we consider numerical (discrete or continuous) data with the dimensionality denoted as p, i.e, with p variables corresponding to e.g. measure symptoms or laboratory measurements measured over n samples (individuals). In general, dimensionality reduction may bring several important benefits:

- Simplification of subsequent computations
- Comprehensibility (e.g. allowing to divide variables to clusters)
- Reduction or removing correlation among variables
- A possible improvement of the classification performance (which happens however only occasionally).

If p is large, and especially if p largely exceeds the number of observations n, numerous standard classification methods suffer from the so-called curse of dimensionality. They are either computationally infeasible or at least numerically unstable for such high-dimensional data [15, 16]. In such a case, dimensionality reduction is a necessity. We distinguish between supervised and unsupervised complexity reduction methods, where supervised ones are tailor-made for data coming from two or more groups, while the information about the group belonging is taken into account. None of the approach is uniformly the best across all datasets.

Variable selection methods extract a relevant subset of the set of the original variables. Their important examples include:

- Statistical hypothesis tests (which are however used only to rank variables in order of evidence against the null hypothesis, instead of computing a p-value).

- Variable selection based on maximal conditional entropy (Section 5).
- MRMR variable selection (Maximum Relevance Minimum Redundancy) [17].
- Bayesian variable selection methods.
- Wrappers or filters (or embedded methods).
- *t*-scores corrected for marginal correlations.

Tailor-made variable selection approaches for regression models include:

- Lasso estimation.
- Partial least squares (Section 6).
- Linear Models for Microarrays (limma).
- Sliced inverse regression.
- Elastic net.
- Regularized discriminant analysis (RDA).
- Shrunken centroid regularized discriminant analysis (SCRDA).
- Smoothly clipped absolute deviation (SCAD).

Feature extraction methods search for linear (or nonlinear) combinations of variables, while retaining all variables in the model.

Prominent examples of linear methods include:

- Principal component analysis (Section 4).
- Robust versions of principal component analysis [18].
- Factor analysis (FA).
- Linear discriminant analysis (LDA, which is however aimed primarily at classification).

While nonlinear include:

- Independent component analysis (ICA).
- Correspondence analysis.
- Methods of information theory.

4 Principal Component Analysis

Principal component analysis (PCA) represents the most commonly used complexity reduction method in biomedical applications. The examples show that PCA is very often used for data observed in groups, although this is not suitable due to its unsupervised nature as investigated already by Mertens BJA [19].

4.1 Method

The aim of PCA is to replace the total number of n observations, which are p-variate, by a set of transformed n observations with a smaller number of variables (dimensions). Thus, the original variables are replaced by a small number of (say s) principal components, where the user may choose a suitable s fulfilling $s<min(n,p)$. New s-dimensional observations represent mutually uncorrelated (orthogonal) linear combinations of the original variables with the ability to explain a large (more precisely the largest possible) portion of variability of the data [11].

The empirical covariance matrix S is ensured to be symmetric and positive semi definite with non-negative eigenvalues and its rank does not exceed $mim(n,p)$. Because the sum of eigenvalues of a general square matrix is equal to the sum of its diagonal elements (i.e, its trace), this is for the case of a covariance matrix equal to the sum of variances of individual variables.

PCA may bring a remarkable reduction of computational costs, especially for small values of the constant s. The contribution of the i-th principal component ($i=1,\ldots,p$), i.e, the component corresponding the i-th largest eigenvalue to the explanation of the total variability in the data can be expressed as the relative contribution of the corresponding eigenvalue. A different (not equivalent) approach may be based on computing principal components from the empirical correlation matrix, which is recommended in case of big differences in the variability of individual variables.

Formally, PCA projects individual observations to the subspace generated by s eigenvectors of the matrix S, which belong to the largest eigenvalues. Then, consequent computations are performed in a space generated by these eigenvectors and the computations replace each observation by the resulting linear combinations. A popular tool for selecting a proper value of s is the scree plot, which is shown in Figure 1 for a dataset described later in Section 4.2. It exploits the fact that the total variability in the data is equal to the trace of D and thus also to the sum of the eigenvalues of S.

Commonly, the user demands the selected principal components to explain at least a given percentage of the total variability, which formulates a requirement on the eigenvalues. Particularly, if the selected principal components should explain e.g. 80% of the total data variability, this means to select such number s of principal components so that the sum of s largest eigenvalues exceeds 80% of the sum of all eigenvalues.

Standard dimensionality reduction methods suffer from the presence of measurement errors or outlying measurements (outliers) in the data [20, 21]. We may recommend performing multivariate methods including PCA by robust alternative of standard approaches. Robust versions of PCA, which are resistant (insensitive) to outliers, have been developed [18]. If robust PCA is based on eigen-decomposition of a robust covariance matrix estimator, the resulting robust principal components are uncorrelated.

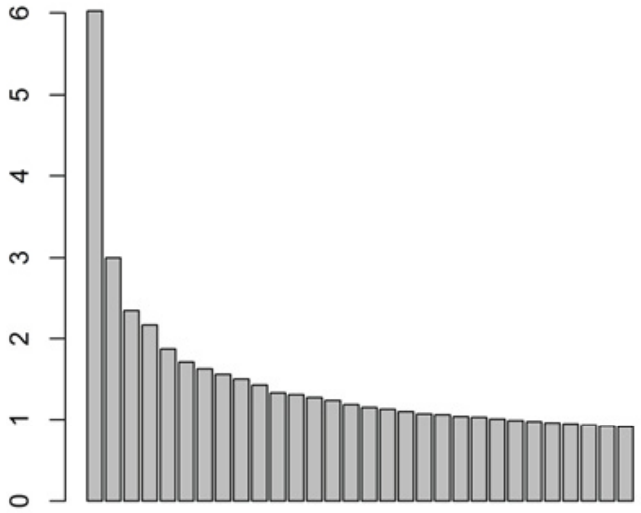

Figure 1. 30 largest eigenvalues of the matrix S in the example of Section 4.2.

4.2 Example: Diagnostics of Cardiovascular Diseases

In a cardiovascular genetic study performed at the Center of Biomedical Informatics in Prague, headed by Prof. Jana Zvárová, a research of gene expressions was performed in 2006-2011 to construct a decision support system based on clinical and gene expressions data [22]. The microarray technology was used to measure average gene expressions of more than 39 thousands gene transcripts across the whole genome. The aim was finding sets of genes, which are useful in the process of diagnostics of (new) individuals.

As it is a typical situation in molecular genetics that there are thousands or tens of thousands of variables (gene expressions) measured on a sample of tens or hundreds (at maximum) of patients, we perform a dimensionality reduction at first and proceed to constructing a classification rule only afterwards. PCA was performed for various values of s and the results reveal that there is no remarkable small group of variables responsible for a large portion of variability of the data and the first few principal components seem rather arbitrary. We used the simplest regularized version of linear discriminant analysis (LDA) [23] to learn a classification rule allowing assigning a new individual to one of the given categories according to the diagnosis.

If s=10, the constructed classification rule was not able to overcome a classification accuracy of 75%. Only if the number of selected principal components was raised to the maximal possible value, which is equal to the number of observations in the data set, the classification accuracy in a leave-one-out cross validation study was able to further increase above 90%. This is the situation with infeasible standard LDA, but the regularized version does not suffer from curse of dimensionality and represents a reliable tool with no tendencies to overfit [23]. Results with a large s allowed obtaining results with a clear interpretation, because there seems no small set of very dominant genes, which would be sufficient for the subsequent classification task. On the other hand, there is a large number of genes with only a small influence on the classification task, which cannot be however neglected.

The results without reducing the dimensionality to a small number of principal components allowed to predict the risk of a manifestation of acute myocardial infarction or cerebrovascular stroke in the next 5 years for a particular patient. If he/she has already undergone an acute myocardial infarction, then the resulting principal components of gene expressions are able to predict the risk of a more severe prognosis or a relapse [22]. Patients with a high risk of a future manifestation of a cardiovascular disease can be consequently monitored, which can increase the patient safety and lead to a more effective and safer care for patients with a life-threatening risk.

4.3 Example: Face Detection

Another example is devoted to the face detection task in a database of images coming from the Institute of Human Genetics, University of Duisburg-Essen, Germany (projects BO 1955/2-1 and WU 314/2-1 of the German Research Council) [24]. This database contains 212 grey-scale images of the size 192 times 256 pixels, each image corresponding to a different person. The persons are volunteers in the age between 18 and 35 years of German origin without a manifested genetic disease. The images were photographed under standardized conditions; the faces do not differ much in size and are also rotated in the plane by small angles. Therefore, eyes are not in a perfectly horizontal position in such images.

The work aimed at constructing a decision support system [25]. The aim was to propose a mouth detection method for the sake of a decision support system for the diagnostics of genetic diseases in children with dysmorphic faces. Thus, it was required to have a method which is comprehensible and useful also for genetic patients with a facial dysmorphia. From the given database of images of the whole faces, we manually localized and selected a database of 212 mouths and 212 non-mouths of size 26x56 pixels. Particularly, a non-mouth was selected within each image, which has the largest similarity with the mouth in the same image by means of the correlation coefficient with a bearded template [24]. These images are however transformed to vectors, i.e, with length p=1456.

We use the projection pursuit (PP) algorithm for the robust PCA of [18] implemented in library pcaPP of the R software. The PP is a general approach for finding the most informative directions or components for multivariate (high-dimensional) data. Such dimensionality reduction is based

on a robust measure of spread of the data, taking into account the outlyingness of each data point. Candidate directions for the principal components are selected by a grid algorithm optimizing such objective function only in a plane, while the subsequent components are added in the later steps.

We computed 5 main principal components from the mouths and non-mouths by the PP algorithm. As the robust method allows to identify the outliers, we have revealed more reliable data point in the top part of the images, corresponding to the face parts above and aside from the lips. On the other hand the (potential) outliers are located on the boundary of the mouth or in the bottom part of the images in the area between the mouth and the chin.

Further, the classification task itself is solved by the standard quadratic discriminant analysis (QDA), which would not be otherwise feasible for high-dimensional data with a number of variables exceeding the number of observations. The classification with QDA yields a correct performance of 100% in a leave-one-out cross validation study, which represents a standard attempt for an independent validation [26].

5 Variable Selection based on Maximal Conditional Entropy

An important class of supervised variable selection procedures is based on principles of the information theory. This section recalls a stepwise variable selection approach based on maximizing conditional entropy. Such approach was applied within a prototype of a system for clinical decision support of [22]. We investigated the performance of the system again on the molecular genetic data from the Center of Biomedical Informatics (2006-2011).

5.1 Method

Data observed in two groups are considered. The method is able to reduce the set of all variables by a forward procedure optimizing a decision-making criterion. We consider a set of variables and the classification rule should be based on them over a training dataset. We define Y as an indicator variable, assuming that it equals 1 if and only if a given sample belongs to the first group. We understand Y as a binary response of the observed variables, which play the role of regressors; these must be however categorized, i.e, replaced by categorical variables with at most 4 categories.

It will be necessary to measure the contribution of a given variable (say X) to explaining the uncertainty in the response. This will be quantified by means of the conditional Shannon information. The first selected variable maximixes the conditional Shannon information with the response among all variables i.e, is the most relevant variable for the classification task. Further on, selecting the variables may be described in the following way. If variables $X_1,..,X_s$ have been already selected as the most relevant s variables, the next variable (say X_{s+1}) is selected as that variable fulfilling the requirement.

$$d(Y \mid X_1,\ldots,X_s,X_{s+1}) = \max d(Y \mid X_1,\ldots,X_s,X), \quad (1)$$

where all variables X not present in the set $\{X_1,\ldots,X_s\}$ are considered. The expression d in (1) is the conditional Shannon information. Thus, a variable very relevant for the classification task is chosen taking into account the dependence of the selected variables. Finally, only such variables for the consequent classification analysis are considered, which contribute to explaining more than a given percentage of the inter-class variability of the data; the choice for this percentage will be discussed in the example of Section 5.2.

5.2 Example: System SIR

A prototype of a clinical decision support system called SIR (System for selecting relevant Information for decision suppoRt) was proposed and implemented in [22], exploiting a sophisticated variable selection component. It contains various tools of supervised learning methods to learn the sophisticated classification rule in order to support a diagnostic decision making The main advantage of the system is suitability also for high-dimensional data obtained e.g. in molecular genetic studies.

The system SIR can be described as an easy-to-use web-based generic service devoted to data collection and decision support with a sophisticated information extraction component. It is proposed for being used mainly for general practitioners in the primary care, but it is able to handle data from any area of medicine. The decision making of the SIR requires data from a (sufficiently large) clinical study in order to construct the optimal classification rule for the decision making problem.

Data collected within a clinical study represent the training database of the SIR, which can import the whole data set from a clinical study automatically together with a data model. The maximum entropy variable selection of Section 5.1 is used. All variables selected by the variable selection procedure are required to enter the decision support system, which can be performed through the automatically generated interface from an electronic health record (EHR) or health information system (HIS), although a manual input of data is also possible.

The clinician must specify the prior diagnosis before entering the data to the SIR, because he/she is the only one to carry the legal responsibility for the clinical decision. Now the SIR can be used through the web service to obtain a diagnosis support. Then, the clinician is asked to manually select his/her final decision and only if it is not in accordance with the SIR, the clinician writes a short text justifying the decision. The system allows quantifying the influence of an additional examination (variable) on the diagnostic decision. Additionally, the dimension reduction procedure

may be extended to consider also costs of obtaining each clinical or laboratory measurement.

The prototype of the system SIR was verified on a different data set from set from the previously described cardiovascular genetic study [22]. Clinical and gene expression measurements were measured on 59 patients with infarction, 45 patients having a cerebrovascular stroke, and 77 control persons without a manifested cardiovascular disease.

If no dimensionality reduction is performed, a regularized LDA yields the classification accuracy 0.85, which is defined as the percentage of correctly classified samples.

We applied the variable selection described in Section 5.1 to the set of 8 personal and clinical variables. Requiring that the selected variables contribute to at least 90% of the intra-class variability of the whole set, we selected 5 variables. At the same time, the variable selection was applied to the set of more than 39 000 gene transcripts, where 245 of them were selected again based on the requirement to contribute to more than 90% of the variability. The classification accuracy in a leave-one-out cross validation study with the 5 variables and 245 genes is equal to 0.85, while it drops to 0.65 if 5 variables are used with 10 genes selected by a MRMR [17] variable selection. These results were obtained with a support vector machine classifier with a Gaussian kernel, which outperforms a number of other standard classifiers.

6 Partial Least Squares

Partial least squares (PLS, also projection to latent structures) can be presented as a supervised dimensionality reduction connected with a regression or classification method [12]. While the PLS is a common method in biomedical or chemometric applications, the method for parameter estimation is more complicated compared to other standard methods of multivariate statistics. The method replaces original variables by new ones, which will be denoted as latent variables, although they are commonly denoted also as principal components or predictive components. A real data set will be described first, which was selected as an example of a study, for which the PLS represents a suitable tool, while general principles of the method will be overviewed afterwards.

6.1 Example: Toxicity of Rat Liver

Let us consider gene expression data from the liver toxicity experiment of [27] in which the total number of n=64 rats was exposed to acetaminophen. The structure of the data is shown in Table 1. The rats were divided to 4 groups. A necropsy was performed on the liver of each rat, while it was performed 6 hours after exposure in the first group, 18 hours in the second group, 24 in the third and finally 48 hours in the fourth group. The data set contains gene expressions measured for p=3116 selected genes on each of the rats. These data were already pre-processed and normalized in a standard way and the remains to learn a classification rule based on the data.

The analysis of the data observed in this experiment requires learning a classification rule allowing to assign a new observation to one of the four given groups according the time interval between the exposition and necropsy. Its intrinsic dimensionality reduction remarkably simplifies the classification, which is illustrated in Figure 2 depicting two major latent variables computed by the PLS-DA. The PLS method allows to visualize the results. In regression tasks, the contribution of individual latent variables to the variability of the response may be evaluated. In classification, the contribution of latent variables to the separation among the groups may be evaluated. This is a similar property with the PCA, where the contribution of individual principal components to the variability of the original data may be evaluated, as explained in Section 4.1.

In the example only two of these latent variables are sufficient to construct a reliable classification rule practically with any classification method. The classification to four groups with QDA attains 100% in a leave-one-out cross-validation study.

Table 1. Data on the toxicity of rat liver in the example of section 6.1.

Rat	Time to necropsy	Gene 1	…	Gene 3116
1	6	0.051	…	-0.034
2	6	0.015	…	-0.079
…	…	…	…	…
64	48	-0.014	…	-0.017

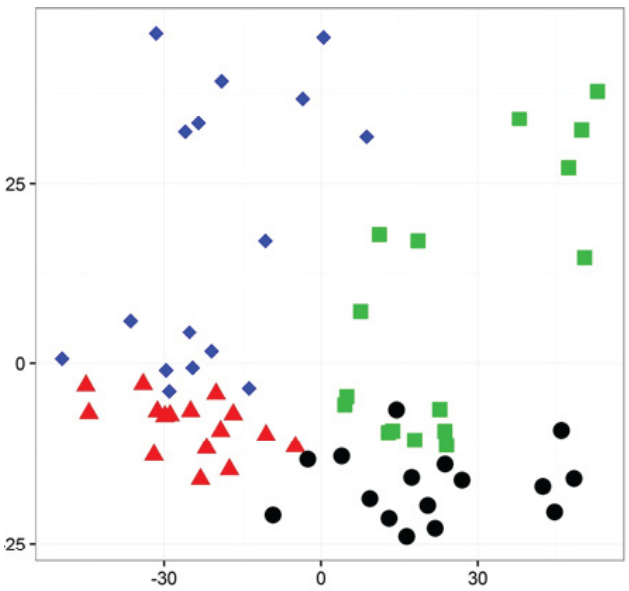

Figure 2. Graph of the association between two latent (predictive) components computed by PLS-DA in the example of Section 6.1, where the contribution of individual observations to one of the groups is given by the shape (circle, square, triangle, and rhombus).

6.2 Method

The PLS was originally proposed for the linear regression model. Thus, the regression version of the PLS remains to be commonly denoted as PLS-R. Formally, PLS-R exploits the standard linear regression model, where Y can be a matrix corresponding to a multivariate response explained by regressors X. The PLS-R method combines the regression task (parameter estimation) with dimensionality reduction in the following way. It searches for the optimal set of latent variables for regressors and also an analogous set for the multivariate response. Instead of X and Y, their linear combinations are considered, which have a smaller dimensionality but the maximal mutual covariance.

Various numerical studies indicated the suitability of the PLS-R method in some applications, especially if one or more of the following situations are true:

- The model suffers from multicollinearity.
- p is large.
- The errors e has a large variability.

Commonly, the PLS method in the regression version is used also for solving classification tasks with the aim to construct a classification rule allowing to assign a new observation to one of K ($K>2$) groups. In the situation $K>2$, however, it is unsuitable to consider the response in the form of a single variable with values in the set $\{1,2,\ldots,K\}$.

Therefore, a special PLS version combining dimensionality reduction with classification into $K>2$ groups. It has been denoted as PLS-DA or D-PLS to stress the discrimination (i.e, classification) context [16, 28]. The training p-dimensional observations are considered. The (multivariate) response is considered as a block of indicators, while the are only $K-1$ of them for K groups. A given observations, if belonging to the k-th group with $k \leq K-1$, has only the k-th coordinate of the response to be equal to 1 and the remaining coordinates are zero. If $k=K$, then all its $K-1$ coordinates are equal to 0. The resulting model remains to be again the standard linear regression model, where X plays the role of the matrix of regressors and Y represents the multivariate response with the total number of $K-1$ indicators.

The PLS-DA method searches for the optimal transform (linear combination) of regressors as well as responses so that the resulting latent variables allow the maximal possible separation among the K groups. To estimate the parameters of the model requires solving an optimization problem, which maximizes the covariance between the set of regressors and the set of responses. The resulting latent variables are predictive, i.e, able to discriminate among the groups in the optimal way. At the same time, the contribution of individual original variables to the construction of the classification rule has a clear interpretation.

Important properties of the PLS are common for the regression and classification version:

- The result of the computation depends to some extent to the choice of the algorithm
- A suitable number of latent variables are commonly found by a cross-validation, although it may have a tendency to overfitting.

Some special versions of the PLS have been proposed more recently, including PLS-EDA (PLS-enhanced discriminant analysis) or OPLS (orthogonal PLS), where the latter offers the same prediction ability as the standard PLS but improves the interpretation. Intensive attention has been paid to the study of assumptions under which the PLS yields better results compared to those obtained with a combination of the PCA with one of standard classifiers.

The PLS methodology resembles that of canonical correlation analysis, while the first maximizes the covariance but the latter is focuses on correlations. Their relationship was investigated by Sun L et al. [29], who showed their equivalence in a special case with orthonormalized PLS.

7 Hypothesis Testing

Hypothesis testing (e.g. a two-sample test) is often desirable for molecular genetic data to find a set of differentially expressed genes. Some recent tests for high-dimensional were overviewed in [16]. Here, we discuss briefly some important approaches to testing high-dimensional data. It is nevertheless useful to point out that testing may not be always the aim of the analysis (if the user prefers a classification rule from a test). Another drawback of simple testing by a repeated using of standard tests is their increase in the probability of type I error due to repeating testing. If there is a large number of samples n in the data, there is also a clear tendency for the power of the tests to increase and thus nearly every hypothesis test yields a significant result. Let us now review three important classes of tests for high-dimensional data.

One class includes tests based on regularization (shrinkage estimation), including the approach of [30]. It replaces all high-dimensional matrices (mainly the covariance matrices) by regularized counterparts and thus a shrinkage Hotelling test is based on a regularized version of the Mahalanobis distance.

Another class of tests of [31] represents a combination of testing with a linear dimensionality reduction. Tests based on linear scores or principal components (i.e, performed on results of LDA or PCA, respectively) are exact tests for normally distributed data. Using the theory of spherical distributions, the tests keep the significance level on the selected 5 % and follow exactly the t- or F-distribution if the variable selection based appropriately performed on the unsupervised data.

The most recent class of tests is based on interpoint distances. Tests based on a nonparametric combination of dependent interpoint distances are consistent and unbiased for high-dimensional data even without the assumption of normally distributed data [32, 33].

8 Acknowledgement

The work has been supported by the project NV15-29835A of the Czech Health Research Council. Preliminary results were presented as an invited lecture at the 17th Summer school of mass spectrometry in Luhačovice in 2016. The author is thankful to an anonymous referee for constructive advice.

References

[1] Zvárová J, Veselý A, Vajda I. Data, information and knowledge. In: Berka P, Rauch J, Zighed D. (eds.) Data mining and medical knowledge management: Cases and applications standards, Hershey: IGI Global; 2009. p. 1-36.

[2] He M, Petoukhov S. Mathematics of bioinformatics: Theory, methods and applications. Pan Y, Zomaya AY. (eds.) Hoboken: Wiley; 2011.

[3] Kalina J, Zvárová J. Decision support for mental health: Towards the information-based psychiatry. In: Clarke S, Jennex ME, Becker A, Anttiroiko AV. (eds.) Psychology and Mental Health: Concepts, Methodologies, Tools, and Applications. Hershey: IGI Global; 2016. p. 1-14.

[4] Kalina J, Zvárová J. Decision support systems in the process of improving patient safety. In: Moumtzoglou A, Kastania A.(eds.) E-health technologies and improving patient safety: Exploring organizational factors. Hershey: IGI Global; 2013. p. 71-83.

[5] Kalina J, Hlinka J. Implicitly weighted robust classification applied to brain activity research. Communications in Computer and Information Sciences. 2017; 690: 87-107.

[6] Hanson A, Levin BL. Mental health informatics. Oxford: Oxford University Press; 2013.

[7] Nordahl H, Osler M, Frederiksen BL, Andersen I, Prescott E, Overvad K, et al. Combined effects of socioeconomic position, smoking, and hypertension on risk of ischemic and hemorrhagic stroke. Stroke. 2014; 45: 2582-2587.

[8] Harrell F. Regression modeling strategies with applications to linear models, logistic and ordinal regression, and survival analysis. 2nd ed. Cham: Springer; 2015.

[9] Kalina J, Vlčková K. Robust regularized cluster analysis for high-dimensional data. Proceedings of 32nd International Conference Mathematical Methods in Economics MME 2014; 2014 Sep 10-12; Olomouc: Palacký University; 2014.

[10] Maroco J, Silva D, Rodrigues A, Guerreiro M, Santana I, de Mendonca A. Data mining methods in the prediction of dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. BMC Res Notes. 2011; 4: 299.

[11] Rencher AC. Methods of multivariate analysis. 2nd ed. New York: Wiley; 2002.

[12] Dziuda DM. Data mining for genomics and proteomics: Analysis of gene and protein expression data. New York: Wiley; 2010.

[13] Fan J, Lv J. A selective overview of variable selection in high dimensional feature space. Statistica Sinica. 2010; 20: 101-148.

[14] Kalina J, Rensová D. How to reduce dimensionality of data: Robustness point of view. Serbian Journal of Management. 2015; 10: 131-140.

[15] Martinez WL, Martinez AR, Solka JL. Exploratory data analysis with MATLAB. 2nd ed. Boca Raton: Chapman & Hall/CRC; 2011.

[16] Kalina J. Classification methods for high-dimensional genetic data. Biocybern Biomed Eng. 2014; 34: 10-18.

[17] Kalina J, Schlenker A. A robust supervised variable selection for noisy high-dimensional data. BioMed Research International. 2015; 320385.

[18] Croux C, Filzmoser P, Oliveira MR. Algorithms for projection-pursuit robust principal component analysis. Chemom Intell Lab Syst. 2007; 87: 218-225.

[19] Mertens BJA. Microarrays, pattern recognition and exploratory data analysis. Stat Med. 2003; 22: 1879-1899.

[20] Heritier S, Cantoni E, Copt S, Victoria-Feser MP. Robust methods in biostatistics. New York: Wiley; 2009.

[21] Saleh AKME, Picek J, Kalina J. R-estimation of the parameters of a multiple regression model with measurement errors. Metrika. 2012; 75: 311-328.

[22] Kalina J, Seidl L, Zvára K, Grünfeldová H, Slovák D, Zvárová J. System for selecting relevant information for decision support. Stud Health Tech Inform. 2013; 186: 83-87.

[23] Hastie T, Tibshirani R, Wainwright M. Statistical learning with sparsity: The lasso and generalizations. Boca Raton: CRC Press; 2015.

[24] Kalina J. Highly robust statistical methods in medical image analysis. Biocybern Biomed Eng. 2012; 32: 3-16.

[25] Böhringer S, Vollmar T, Tasse C, Würtz RP, Gillessen-Kaesbach G, Horsthemke B,et al. Syndrome identification based on 2D analysis software. Eur J Hum Genet. 2006; 14: 1082-1089.

[26] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. Data mining, inference, and prediction. 2nd ed. New York: Springer; 2009.

[27] Bushel P, Wolfinger RD, Gibson G. Simultaneous clustering of gene expression data with clinical chemistry and pathological evaluations reveals phenotypic prototypes. BMC Syst Biol. 2007; 1: 15.

[28] Tan Y, Shi L, Tong W, Hwang GTG, Wang C. Multi-class tumor classification by discriminant partial least squares using microarray gene expression data and assessment of classification models. Comput Biol Chem. 2004; 28: 235-244.

[29] Sun L, Ji S, Yu S, Ye J. On the equivalence between canonical correlation analysis and orthonormalized partial least squares. Proceedings of the 21st international joint conference on artificial Intelligence IJCAI'09; 2009 11-17; Pasadena: Morgan Kaufmann Publishers Inc; 2009.

[30] Dong K, Pang H, Tong T, Genton MG. Shrinkage-based diagonal Hotelling's tests for high-dimensional small sample size data. J Multivar Anal. 2016; 143: 127-142.

[31] Läuter J, Glimm E, Kropf S. Multivariate tests based on left-spherically distributed linear scores. Ann Stat. 1998; 26: 1972-1988.

[32] Marozzi M. Tests for comparison of multiple endpoints with application to omics data. Stat Appl Genet Mol Biol. 2018; 17: 13.

[33] Murakami H. Power comparison of multivariate Wilcoxon-type tests based on the Jurečková-Kalina's ranks of distances. Commun Stat Simul Comput. 2015; 44: 2176-2194.