Chemometrics and Intelligent Laboratory Systems 98 (2009) 88–96

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Exploring the analysis of structured metabolomics data

Maikel P.H. Verouden [a], Johan A. Westerhuis [a,*], Mariët J. van der Werf [b], Age K. Smilde [a]

[a] Biosystems Data Analysis, Swammerdam Institute for Life Sciences, University of Amsterdam, Nieuwe Achtergracht 166, 1018 WV Amsterdam, The Netherlands
[b] TNO Quality of Life, P.O.Box 360, 3700 AJ Zeist, The Netherlands

ARTICLE INFO

Article history:
Received 9 December 2008
Received in revised form 24 April 2009
Accepted 13 May 2009
Available online 27 May 2009

Keywords:
PCA
ASCA
Experimental design
Overfit
Microbial metabolomics

ABSTRACT

In metabolomics research a large number of metabolites are measured that reflect the cellular state under the experimental conditions studied. In many occasions the experiments are performed according to an experimental design to make sure that sufficient variation is induced in the metabolite concentrations. However, as metabolomics is a holistic approach, also a large number of metabolites are measured in which no variation is induced by the experimental design. The presence of such non-induced metabolites hampers traditional data analysis methods as PCA to estimate the true model of the induced variation. The greediness of PCA leads to a clear overfit of the metabolomics data and can lead to a bad selection of important metabolites. In this paper we explore how, why and how severe PCA overfits data with an underlying experimental design. Recently new data analysis methods have been introduced that can use prior information of the system to reduce the overfit. We show that incorporation of prior knowledge of the system under investigation leads to a better estimation of the true underlying structure and to less overfit. The experimental design information together with ASCA is used to improve the analysis of metabolomics data. To show the improved model estimation property of ASCA a thorough simulation study is used and the results are extended to a microbial metabolomics batch fermentation study. The ASCA model is much less affected by the non-induced variation and measurement error than PCA, leading to a much better model of the induced variation.

1. Introduction

Advances in (bio-)analytical techniques enable scientists to use more and more variables to characterize their samples. The fact that the number of experiments is often low leads to high dimensional data with the number of variables greatly exceeding the number of experiments. This type of data is often referred to as high dimensional or megavariate. Microbial metabolomics data, as an example of megavariate data, emerges from the 'omics' field that focuses on low molecular weight compounds, so-called metabolites, present in and around microbial cells at a given time during their growth or production cycle [1]. The metabolome, i.e. the concentration of all metabolites, is a reflection of the phenotype of the sample under the studied experimental conditions [2]. The experimental conditions are changed or perturbed such that sufficient variation is induced in the metabolome, that responds to these changes or perturbations in the experimental conditions. Since metabolomics is a holistic approach, covering as many metabolites as possible, there are also always many metabolites in the data set in which no variation is induced by the change or pertubation of experimental conditions. The reason for this is that a change or perturbation in an experimental condition 'hits' or excites only part(s) of the biochemical network, while the rest of the network operates as if under normal operating conditions. These, so-called non-induced, metabolites can still have a large variation in their concentration, i.e. the metabolites are not tightly regulated, but this variation is not caused by a change or a perturbation of the experimental condition. Furthermore, there is also always some random variation in the data set due to measurement error [3].

Ideally a data analysis method used for analyzing this type of data should only model the induced variation leaving all other variation for the residuals. Here we define incorporation of all variation other than the induced variation as overfit. Principal Component Analysis (PCA) is often used for explorative data analysis and focuses on describing the maximum variation in the data by modeling it into scores, that provide information on the samples, and loadings, that provide information on the metabolites. By focusing on explaining as much of the variation in the data as possible PCA tends, especially with only few experiments in the data set, to be greedy and, therefore, to overfit the data by incorporating random sampling variation and the variation of the non-induced metabolites into the model.

The use of prior information can help to focus the explorative data analysis. In curve resolution, nonnegativity, unimodality and smoothness constraints help to identify chemical compounds in complex mixtures [4,5]. In biology knowledge of transcription factors can be used to unravel complex gene expression data [6,7]. Recently new methods were introduced that are able to incorporate various types of prior information to focus the data analysis [8,9]. By using these more

* Corresponding author. Tel.: +31 20 525 6546; fax: +31 20 525 6971.
E-mail address: j.a.westerhuis@uva.nl (J.A. Westerhuis).

0169-7439/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2009.05.004

advanced explorative analysis methods, focus is on the relevant part of the variation and thus overfit can be reduced and sometimes even additional information can be obtained [10].

An underlying structure, that is nowadays often present in many of the collected megavariate 'omics' data sets, is an experimental design in which experimental factors are varied to study their effect [11–13]. Several methods for analyzing metabolomics data with an underlying experimental design exist, that focus the analysis onto the induced variation by the design by taking it into account [11,14–18]. As a method, that uses the underlying design in the microbial metabolomics data to focus the analysis, we have chosen to use ANOVA-simultaneous component analysis (ASCA) [19]. In ASCA the induced variation can be separated from the non-induced variation and measurement error by creating orthogonal partitions. Subsequent simultaneous component analysis of the individual partitions may elucidate the relation between the samples and the metabolic profile for the effects. The orthogonality between the data partitions allows for individual estimation of effects without mixing of effects. Recently a method was developed that allows statistical validation of megavariate effects in ASCA [20]. It also creates the possibility of testing whether the experimental design induces any sources of variation in the data. Although many metabolomics data sets have an underlying experimental design, in the analysis this is still often neglected [21].

The major goal of this paper is to show by comparison of PCA and ASCA how, why and how severe PCA overfits data with an underlying experimental design. By comparison to ASCA the effect of incorporating prior knowledge with respect to experimental design into the explorative analysis can be shown. In a thorough simulation study we will show how PCA and ASCA behave in terms of fit (how well are the induced metabolites that vary according to the underlying design modeled?) and overfit (how much is modeled of the non-induced metabolites that do not vary according to the design?) when modeling metabolomics data in which induced and non-induced metabolites are present. The results of the simulation study will be discussed in terms of the row and column space of the data [22] in order to show why and how incorporation of design information helps to focus the explorative analysis. Furthermore both methods will be used to model microbial metabolomics data obtained from *Escherichia coli* batch fermentations with an underlying design.

Section 2 of this paper describes PCA, ASCA and their differences. It also describes how the simulated data and the *E. coli* batch fermentation metabolomics data have been created and which measures were used to assess the ability of modeling the induced and non-induced variation. Section 3 describes the results and at the end some important findings are concluded.

2. Materials and methods

In the following text bold uppercase characters (e.g. $\mathbf{X}$) represent matrices, bold lowercase characters (e.g. $\mathbf{x}$) represent vectors and scalars are displayed as italic characters (e.g. I).

2.1. Principal Component Analysis (PCA)

PCA [23] decomposes the data $\mathbf{X}$ $[I\times J]$, consisting of I samples with J measured variables, into a bilinear model of scores $\mathbf{T}$ $[I\times R]$ and loadings $\mathbf{P}$ $[J\times R]$ according to

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \tag{1}$$

Here $\mathbf{T}\mathbf{P}^T$, the PCA model ($\hat{\mathbf{X}}_{\mathrm{PCA}}$), represents a lower dimensional approximation of $\mathbf{X}$ and $\mathbf{E}$ contains the residuals. The number of principal components R, with $R \ll \min(I, J)$, can be chosen by means of cross-validation or by using a scree graph [24]. The calculation of the scores $\mathbf{T}$ and loadings $\mathbf{P}$ by PCA is performed in such a manner, that the sum of squares of the residuals Q, as shown in Eq. (2), is minimized.

$$Q(\mathbf{T}, \mathbf{P}|\mathbf{X}) = \|\mathbf{X} - \mathbf{T}\mathbf{P}^T\|^2 \tag{2}$$

PCA restricts the scores and loadings with the requirements of $\mathbf{T}^T\mathbf{T}$ being a diagonal matrix and $\mathbf{P}^T\mathbf{P}$ being an identity matrix. This restriction does not decrease the explained variation but serves to arrive at easy interpretable graphs.

2.2. ANOVA-Simultaneous Component Analysis (ASCA)

A recently developed data analytical method for analyzing megavariate metabolomics data with an underlying experimental design is ASCA [19]. This method starts with the ANOVA decomposition of the data $\mathbf{X}$ $[I\times J]$ and partitions the variation in the data into orthogonal parts per effect according to the experimental design. The variation partitioning for the effects with ANOVA is achieved by averaging the experiments, that have been performed with the same level-setting of the corresponding factors [8]. If for instance the underlying experimental design is a full factorial design with three factors, the partitioning for the main effects can be represented by Eq. (3).

$$\mathbf{X} - \mathbf{X}_{\mathrm{M}} = \mathbf{X}_1 + \mathbf{X}_2 + \mathbf{X}_3 + \mathbf{X}_{\mathrm{res}} \tag{3}$$

In Eq. (3), $\mathbf{X}_{\mathrm{M}}$ represents the matrix of means, which is calculated as

$$\mathbf{X}_{\mathrm{M}} = \frac{1}{I} 1_I 1_I^T \mathbf{X},$$

with $\mathbf{1}_I$ $[I\times 1]$ denoting a vector of ones. The matrices $\mathbf{X}_1$, $\mathbf{X}_2$ and $\mathbf{X}_3$ represent the variation partitions of the three main effects (the three design factors) and $\mathbf{X}_{\mathrm{res}}$ contains the remainder variation consisting of all interaction effects and all other sources of non-induced variation and measurement error. Of course one can choose to also decompose the interaction partitions or to combine different partitions into a single matrix [8]. In Eq. (4) the orthogonality between the variation partitions is shown, which means that the column spaces of the individual matrices on the right side of the equality sign in Eq. (3) are orthogonal. Proof of Eq. (4) is supplied elsewhere [8].

$$\mathbf{X}_\alpha^T \mathbf{X}_\beta = 0 \forall \{\alpha, \beta\} \subset \{1, 2, 3, \mathrm{res}\} : \alpha \neq \beta \tag{4}$$

The orthogonality between the variation partitions is a desirable property, because it shows that each partition per effect is calculated independently without mixing of effects. Because all variation partitions are orthogonal to each other the following statement is also true,

$$[\mathbf{X}_1 + \mathbf{X}_2 + \mathbf{X}_3]^T \mathbf{X}_{\mathrm{res}} = 0,$$

and shows that the combined variation partitions for the main effects are orthogonal to and, thus, independent of the remainder variation.

To describe ASCA as a multivariate regression model here we focus on the variation partitioning for the effects within mean centered data ($\mathbf{X}_{\mathrm{mc}} = \mathbf{X} - \mathbf{X}_{\mathrm{M}}$), having an underlying two level full factorial design. The various multivariate effects can also be achieved by least squares fitting of a linear model [25,26]. The linear model, for an experiment