Chemometrics and Intelligent Laboratory Systems 174 (2018) 142–148

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Better interpretable models after correcting for natural variation: Residual approaches examined

Mike Koeman [a], Jasper Engel [a,b], Jeroen Jansen [a,*], Lutgarde Buydens [a]

[a] Department of Analytical Chemistry, Institute for Molecules and Materials (IMM), Radboud University, Heyendaalseweg 135, 6525 AJ Nijmegen, The Netherlands
[b] Biometris, Wageningen UR, Droevendaalsesteeg 1, 6708 PB Wageningen, The Netherlands

ARTICLE INFO

Keywords:
PCA
Residuals
Smearing
Interpretation
Metabolomics
Disease diagnosis

ABSTRACT

The interpretation of estimates of model parameters in terms of biological information is often just as important as the predictions of the model itself. In this study we consider the identification of metabolites in a possibly biologically heterogeneous case group that show abnormal patterns with respect to a set of (healthy) control observations. For this purpose, we filter normal (baseline) natural variation from the data by projection of the data on a control sample model: the residual approach. This step should more easily highlight the abnormal metabolites. Interpretation is, however, hindered by a problem we named the 'residual bias' effect, which may lead to the identification of the wrong metabolites as 'abnormal'. This effect is related to the smearing effect.

We propose to alleviate residual bias by considering a weighted average of the filtered and raw data. This way, a compromise is found between excluding irrelevant natural variation from the data and the amount of residual bias that occurs. We show for simulated and real-world examples that this compromise may outperform inspection of the raw or filtered data. The method holds promise in numerous applications such as disease diagnoses, personalized healthcare, and industrial process control.

1. Introduction

Untargeted metabolomics is becoming increasingly important in an extensive range of applications such as food science [1,2], environmental science [3], forensics [4], and healthcare [5,6]. Comprehensive profiling with metabolomics therefore has become a household approach in many branches of quantitative research and many societally relevant topics. Oftentimes, a set of control observations and a set of cases are measured by high-throughput techniques (*e.g.* ^{1}H NMR or LC-MS) in such studies. This leads to the case-control studies that we focus on in this work. Next, based on (multivariate) statistical analysis of the acquired data, hypotheses on the mechanism that may be responsible for biological phenomena are generated. Such a mechanism generally influences multiple metabolites at the same time, with the desired result being a series of biomarker metabolites that together may be specific for that process and may possibly be used for prediction. Multivariate chemometric approaches are widely used for this, as these may extract relevant information using all variables at once, as opposed to one feature at a time. One challenge in analysing experiments like these is the large amount of (possibly confounding) natural variation such as a subjects diet, genotype or gut microbiome. These variations cannot be completely known and are beyond control of the experimental researcher. It hinders the analysis as this variation is inherently non-informative.

Our goal is to separate this irrelevant natural variation from the biologically interesting information (related to the phenomenon of interest) in a case-control experiment. Our focus here is on interpretation rather than prediction: we want to find the systematic metabolic differences between the two groups so that we can interpret them to learn more about the biological phenomenon investigated in the experiment. The most common way to tackle this in case-control studies, is to pose it as a two-class classification problem and analyse it with a method such as PLS-DA [7,8]. A shortcoming of this approach, is that this assumes a homogenous response to a disease. This assumption is often not met in practice. Using multiple classes to model the heterogeneity of the disease would be possible, but requires both sufficient data and the class labels.

A method without this shortcoming is Statistical Health Monitoring (SHM) [9] that builds on principles from analysis of industrial process monitoring. SHM is based on describing, using principal component analysis (PCA), the variation common to most of the samples in the control group, the natural variation. This is referred to as modelling the Normal Operating Conditions (NOC) of metabolic variability. Subsequently, patient data can be matched to the NOC. Individuals that do not

* Corresponding author.
E-mail address: chemometrics@science.ru.nl (J. Jansen).

https://doi.org/10.1016/j.chemolab.2018.01.007
Received 3 January 2017; Received in revised form 17 January 2018; Accepted 18 January 2018
Available online 1 February 2018

match this NOC are abnormal and should be further inspected with the use of 'contribution plots' that provide the measured metabolites that are most 'abnormal' and can be used for root cause analysis. A notable characteristic is that SHM regards each sample separately, as opposed to the classification approach. Recently, SHM been successfully applied in a liver study in Ref. [10], where it is shown that SHM found metabolites that have been confirmed to play a role in relevant pathways, as opposed to the classification approach which found other metabolites.

SHM suffers from two limitations however. One is that is only able to process individual samples, each sample is analysed individually to see if and where it deviates from the NOC. This is of course not a concern if there is only one sample available, but when multiple samples are available, methods that use these at the same time could be inherently more powerful than the methods that do not. The second limitation is the smearing effect [11,12] that contribution plots are known to suffer from. This can cause the identification of the wrong metabolites.

The 'Residual Approach' we investigate here takes the ideas of Statistical Health Monitoring further, and extends them to a multi-sample situation. This makes the Residual Approach applicable for both single sample and multi sample situations. It calculates residuals by removing the natural variation from the data. These information-rich residuals can then be further investigated to identify the metabolites (variables) affected by the experiment. Analysis of these residuals, by for example PCA, makes it possible to reveal those variables. Calculating these residuals can thus be seen as a form of pre-processing to rid the data of natural variability unrelated to the case-related metabolism. Other residual-based approaches with different goals can be found in Refs. [13–15]. The approach we present here can also be used for these goals.

The square of these residuals are equivalent to a specific type of contribution plot: the complete decomposition of the Q-statistic [12]. As contribution plots are known to suffer from the 'smearing effect', it can be expected that this affects the residuals in much the same way. A model trained on these residuals cannot be completely trusted as a result, as the deviations compared to the NOC model may express themselves on features unrelated to the response. Furthermore, this residual approach limits itself to the residual space and disregards the NOC space as we explicitly remove the entire NOC space from the data. This again limits its interpretability. We name these effects together the 'residual bias effect'. Other methods, such as ICA [16] or MCR-ALS [17], could in theory also be used to analyse this type of data but have their own associated challenges. ICA, for example, assumes that components are statistically independent. This assumption may not be valid, as disease may not necessarily manifest themselves as statistically independent components. MCR-ALS requires sufficient constraints on the determined components to come to a meaningful solution, these constraints follow from prior information that may not necessarily be available for many diseases.

In this work we investigate the Residual Approach and the associated 'residual bias effect' and propose a new method to alleviate this effect. We show that this new method combined with PCA analysis can be more reliable in terms of interpretability than PLS-DA.

2. Theory

2.1. Normal Operating Conditions

The Normal Operating Conditions (NOC) describes a group of healthy individuals, for example ^{1}H NMR spectra of their urine. Typically, this NOC is represented by a dataset where each of the samples is a measurement from the situation that is 'under control', *i.e.* healthy or at least non-diseased. This is analogous to process control, where the NOC is a situation where the industrial plant is under control and generates products within specifications. We denote this dataset with $\mathbf{X}_{\mathrm{NOC}}$. This $\mathbf{X}_{\mathrm{NOC}}$ is often modeled by latent variable models like PCA or ICA [16]. Such component-based models may be generically represented by eq. (1)

$$\mathbf{X}_{\mathrm{NOC}} = \mathbf{T}_{\mathrm{NOC}}\mathbf{P}_{\mathrm{NOC}}^{\mathrm{T}} + \mathbf{E}, \tag{1}$$

where $\mathbf{P}_{\mathrm{NOC}}$ are the NOC loadings, $\mathbf{T}_{\mathrm{NOC}}$ the scores and $\mathbf{E}$ the residuals, note that $\mathbf{T}_{\mathrm{NOC}}\mathbf{P}_{\mathrm{NOC}}^{\mathrm{T}}$ is the reconstruction of our data matrix $\mathbf{X}_{\mathrm{NOC}}$. While models like these typically describe the data well, the number of components needs to be estimated. Choosing the appropriate number of components is critical to the model as the incorrect number may cause the model to over- or underfit. Selecting an appropriate number of components is as challenging as in most chemometric methods based on dimensionality reduction, especially without a good objective criterion to optimize. Here we have opted to use the NUMFACT approach [18]. In the case where the NOC is a group of healthy individuals, subgroups can be present, for example males and females. If the data within these subgroups is very distinct, multiple NOCs could be used, leading to a SIMCA [19]-like approach. Here we only consider the situation where no distinct subgroups are present.

There is of course also the group under investigation: the case group. This group may be no longer in control and is in a state that may not be *completely* described by the model created on the NOC, due to an effect of a disease or experiment has on their metabolic profile, analogous to products from an industrial plant that has a fault of some sort. One key difference between industrial process control and the Residual Approach we discuss here, is that we regard these samples as a group as we expect there to be similarities between them which we would like to exploit.

The case group we will be using here is a patient group with a specific disease. This group is described by a series of measurements which we shall denote by $\mathbf{X}_{\mathrm{case}}$. The group can still be described *partially* by the NOC model, since a disease might manifest itself in the urine as an *additional* contribution to the 'healthy' metabolism they share with the control group. Another part of the urine composition (either over- or underrepresented metabolites) can however not be described by the NOC: this is the contribution to the urine composition of most interest, as this contains the biomarkers of disease. This contribution might be similar for each individual, leading to a two-class problem. In practice however, some people react more strongly than others to an experiment; people might even react by changes in different combinations between metabolites. If we describe this group with a latent variable model the combined model would look like

$$\mathbf{X}_{\mathrm{case}} = \mathbf{T}_{\mathrm{NOC}}^{*}\mathbf{P}_{\mathrm{NOC}}^{\mathrm{T}} + \mathbf{T}_{\mathbf{case}}\mathbf{P}_{\mathrm{case}}^{\mathrm{T}} + \mathbf{E}, \tag{2}$$

where $\mathbf{T}_{\mathrm{NOC}}^{*}$ is the score of the NOC component, $\mathbf{T}_{\mathrm{case}}$ the score of the case component, $\mathbf{P}_{\mathrm{case}}$ is the loading of the case component(s) we are looking for and $\mathbf{E}$ are residuals. Eq. (2) can describe both a homogeneous or heterogeneous group. $\mathbf{P}_{\mathrm{NOC}}$ and $\mathbf{P}_{\mathrm{case}}$ are often orthogonal matrices, $\mathbf{P}^{\mathrm{T}}\mathbf{P} = \mathbf{I}$. The spaces spanned by $\mathbf{P}_{\mathrm{NOC}}$ and $\mathbf{P}_{\mathrm{case}}$ are however typically mutually non-orthogonal $\mathbf{P}_{\mathrm{NOC}}^{\mathrm{T}}\mathbf{P}_{\mathrm{case}} \neq \mathbf{I}$, as disease or experimental manipulations may affect several endogenous metabolites that are already present in $\mathbf{P}_{\mathrm{NOC}}$—hence there is no biological foundation for both spaces to be orthogonal.

Our goal is to find $\mathbf{P}_{\mathrm{case}}$ as accurately as possible, to find the most information-rich metabolites as clues for the mechanism responsible for the disease under investigation.

It should be noted that not all deviations from the NOC will necessarily manifest as an additional effect as in eq. (2). It could also be possible that $\mathbf{X}_{\mathrm{case}}$ will have higher values for $\mathbf{T}_{\mathrm{NOC}}^{*}$ compared with the NOC. This should be evident from these values.

2.2. Residual-based approach

If the data indeed follows the model in eq. (1) and $\mathbf{X}_{\mathrm{case}}$ can be described partly by the NOC and partly by a latent variable corresponding to the disease, we should be able to remove the variation that can be explained by the NOC. After the NOC variation has been removed, the residuals should contain only information relevant to the disease. Mathematically this corresponds to: