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 Biometrics, 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. 1H 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 subject’s 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
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 has 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 it 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 $X_{NOC}$. This $X_{NOC}$ is often modeled by latent variable models like PCA or ICA [16]. Such component-based models may be generically represented by eq. (1)

$$X_{NOC} = T_{NOC}P_{NOC}^T + E, \tag{1}$$

where $P_{NOC}$ are the NOC loadings, $T_{NOC}$ the scores and $E$ the residuals, note that $T_{NOC}^TP_{NOC}$ is the reconstruction of our data matrix $X_{NOC}$. 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 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 $X_{case}$. 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

$$X_{case} = T_{case}^TP_{case}^T + T_{case}^P_{case} + E, \tag{2}$$

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

Our goal is to find $P_{case}$ 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 $X_{case}$ will have higher values for $T_{NOC}$ 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 $X_{case}$ 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:
\[ \tilde{E}_{\text{case}} = X_{\text{case}} - X_{\text{case}} \tilde{P}_{\text{NOC}} \tilde{P}_{\text{NOC}}^T, \]  
(3)

where the hat denotes that we estimated the matrix from the data. This intends to remove all variation that is due to \( P_{\text{NOC}} \) from the data, leaving us with only the effect we are looking for in the residuals. Technically, the residuals can then be treated as any other dataset and can be analysed by suitable methods like for example PCA or ICA, in the case of PCA we can for example use biplots to see all relevant information in one picture. In this work we will only consider application of PCA to \( E_{\text{case}} \) for the sake of simplicity. This way of modelling allows a \( \tilde{P}_{\text{case}} \) of more than one component, which allows for heterogeneity in the case group. This is in contrast with classification methods such as PLS-DA, which assume a homogeneous response in only one component.

### 2.3. Residual bias effect

The columns of the residuals, \( \tilde{E}_{\text{case}} \), are orthogonal to \( \tilde{P}_{\text{NOC}} \) so we are not hindered by that healthy variation anymore. However, if the effect we are looking for, \( P_{\text{case}} \), is not actually orthogonal to \( \tilde{P}_{\text{NOC}} \) we will still find a \( \tilde{P}_{\text{case}} \) orthogonal to the space spanned by \( \tilde{P}_{\text{NOC}} \) and the wrong direction in the data will be found as a result. This forced orthogonality occurs because of the projection on \( \tilde{P}_{\text{NOC}} \). All information in \( \tilde{P}_{\text{case}} \) that is not orthogonal to \( \tilde{P}_{\text{NOC}} \) can be explained by \( P_{\text{NOC}} \). An example of an effect like this is shown in Fig. 1. This figure shows a NOC set in green that mainly varies along the horizontal axis, variable 1. The case group in red has variation in both variables and its mean has moved to the top right compared to the NOC group. If we model the NOC group with PCA we will find that the first loading will almost exclusively contain variable 1, which is to be expected. If we then subtract the NOC components from \( X_{\text{case}} \) according to eq. (3), we remove all variation that can be explained by \( P_{\text{NOC}} \) and thus end up with residuals consisting only of variable 2. This phenomena happens because all variation in variable one can be explained by \( P_{\text{NOC}} \) and will be subtracted as a result. \( P_{\text{case}} \), the true value, cannot be found in this manner.

### 2.4. Compromise approach

Analysing \( X_{\text{case}} \) directly using PCA is hindered by the NOC variation, which leads to loadings that are a mix of the NOC and the case-related effect. Analysing the residuals of the projection of the case group on the PCA space of the NOC, \( \tilde{E}_{\text{case}} \), leads to the residual bias effect. Both are demonstrated for simulated data below (see Fig. 3 and the appendix). So on one hand we are hindered by the NOC and on the other we introduce a bias in the loadings. The most accurate approach is somewhere in the middle. This middle ground allows balance between the two, which in turn leads to a more reliable estimate of the effect we are attempting to find. We therefore propose a compromise between \( X_{\text{case}} \) and \( \tilde{E}_{\text{case}} \) which is calculated as follows:

\[ \tilde{E}_{\text{comp}} = X_{\text{case}} - \lambda X_{\text{case}} \tilde{P}_{\text{NOC}} \tilde{P}_{\text{NOC}}^T, \]  
(4)

where \( \lambda \) is the weight that determines the compromise. Note that this is essentially eq. (3) with an additional factor \( \lambda \). In the case that \( \lambda \) equals 1, \( \tilde{E}_{\text{comp}} = \tilde{E}_{\text{case}} \), with \( \lambda \) equals 0, nothing is done and \( \tilde{E}_{\text{comp}} = X_{\text{case}} \). This allows us to balance out the residual bias effect with the NOC components which leads to a better analysis. In fact, by not completely removing the NOC variation, we can obtain a \( \tilde{P}_{\text{case}} \) that is not orthogonal to \( \tilde{P}_{\text{NOC}} \). This leads to a more reliable estimation of \( P_{\text{case}} \). Note that since eq. (4), depends on the principal components estimated by eq. (1), the data should be processed in the same manner as was done for this PCA model. As a consequence, \( \tilde{E}_{\text{comp}} \) is also mean-centered and possibly scaled using the values from \( X_{\text{noc}} \).
2.5. Optimal lambda for Compromise Approach

The performance of the Compromise Approach is greatly dependent on the choice for lambda. We have found that for less challenging problems a low value for lambda should be selected, while for challenging problems a high value is required (see results section). Therefore, we propose the following approach for automatic selection of lambda:

\[
\lambda_{opt} = \min \left\{ \frac{\sum \text{SS}_\text{NOC}}{\sum \text{SS}_\text{case}} \right\}
\]

(5)

where \(\lambda_{opt}\) is the determined value, \(\sum \text{SS}_\text{NOC}\) and \(\sum \text{SS}_\text{case}\) are the mean sum of squares of the residuals of the NOC data and case data respectively. In other words, the value for lambda is chosen as the ratio of the amount of variance that is not explained by the model (eq.(1)) of the NOC data to the case data. If the model does not describe the variance in the case data well, the variance in the residuals of the cases will be much larger compared to that of the NOC. In this case a high value for lambda will be selected. In the opposite case the value for lambda will be small.

3. Methods

3.1. Residual bias effect simulation

To demonstrate the residual bias effect and the problems that it can cause a data set was simulated that can visualize it. The steps of this simulation have been outlined below.

1. Create \(P_{\text{NOC}}\) and \(P_{\text{case}}\) with the same 8 variables. \(P_{\text{NOC}}\) has a non-zero value in the first 4 variables while \(P_{\text{case}}\) has a non-zero value in variables 3–8. The loadings overlap in variable 3 and 4 which causes \(P_{\text{NOC}}\) and \(P_{\text{case}}\) to correlate. Detailed plots of these loadings are given in Fig. 2 in the results section.

2. Create \(T_{\text{NOC}}\) for the NOC set as well as for the case group and \(T_{\text{case}}\) for the case group. These values are drawn from a normal distribution with mean 0 and standard deviation 1. These values determine the contributions of each of the NOC loadings in \(P_{\text{NOC}}\).

3. Simulate \(X_{\text{NOC}}\) and \(X_{\text{case}}\) using eqs. (6) and (7) respectively, \(\epsilon\) is normally distributed noise with \(\mu = 0\) and \(\sigma = 0.4\).

\[
X_{\text{NOC}} = T_{\text{NOC}}P_{\text{NOC}}^T + \epsilon,
\]

(6)

\[
X_{\text{case}} = T_{\text{case}}P_{\text{case}}^T + \epsilon.
\]

(7)

We have now simulated a NOC set and a case set, so that we may calculate the residuals based on the procedure described on the previous page. We have used a single principal component for the NOC here, although the approach is not limited to this. There are a number of alternative ways to analyse this data. One approach is to simply take \(X_{\text{case}}\) and analyse it directly with the use of PCA. One (or more) component(s) could then correspond to the NOC and the other to the case-related effect we are attempting to find. Another might be to take \(X_{\text{case}}\) and \(X_{\text{NOC}}\) together and apply PCA on the combination with the same intent. We used these approaches as well as the residual approach on this example to characterize the residual bias effect and show it in an intuitive figure.

3.2. Applying Compromise Approach

The following scheme has been used to apply the method to the datasets.

1. Center both \(X_{\text{NOC}}\) and \(X_{\text{case}}\) using the means of \(X_{\text{NOC}}\).
2. PCA on \(X_{\text{NOC}}\) obtaining \(P_{\text{NOC}}\), the estimate of \(P_{\text{NOC}}\).
3. Project \(X_{\text{case}}\) on \(P_{\text{NOC}}\) obtaining \(T_{\text{NOC}}\) and calculate \(\hat{E}_{\text{comp}}\).

\[
\hat{E}_{\text{comp}} = X_{\text{case}} - \lambda X_{\text{case}}P_{\text{NOC}}P_{\text{NOC}}^T.
\]

(8)

With \(0 \leq \lambda \leq 1\) where \(\lambda = 0\) corresponds to pure \(X_{\text{case}}\) and \(\lambda = 1\) to \(E_{\text{case}}\) where all NOC variation has been removed.

4. PCA analysis on \(E_{\text{comp}}\) obtaining \(\hat{P}_{\text{case}}\)

\[
\hat{E}_{\text{comp}} = T_{\text{case}}\hat{P}_{\text{case}} + E.
\]

(9)

To compare this Compromise Approach with other approaches such as the pure residuals or PLS-DA we make use of simulated data. The estimated loadings can be compared to the true loadings by calculating the Pearson correlation coefficient between the two. The higher this correlation, the more similar the estimated loading is to the true loading and the better that method performs.

In the case of PLS-DA, the case group is taken as one class and the control group as another class. Centering was performed differently as the data was mean centered as a whole. The PLS weights are then our estimate for \(P_{\text{case}}\). Note that in this simulation we have constrained all case samples to behave in the same way, by imposing identical values for \(P_{\text{case}}\) for each sample. In real data this is not necessarily the case as the same disease may manifest differently on each individual.

4.1. Data simulation

To investigate how the Compromise Approach works for varying values of \(\lambda\), two datasets have been considered. One is an extended version of the simulation described in the previous section and the other is an experimental \(^1\)H NMR dataset on healthy individuals that has been split into a NOC set and a case set in which the case component was artificially introduced.

The construction of the simulated data has been outlined below:

1. Create \(Q_{\text{NOC}}\) and \(Q_{\text{case}}\), each consisting of 50 variables. \(Q_{\text{NOC}}\) has 3 components and comes from a uniform distribution with values between –0.5 and 1.5 a. \(Q_{\text{case}}\) has 1 component and also comes from a uniform distribution but now with values between 0 and 1. As only the direction of \(Q_{\text{case}}\) is relevant, will refer to it as \(P_{\text{case}}\) in the result section to make the notation more consistent.

2. Create \(T_{\text{NOC}}\) and \(T_{\text{case}}\) for the case set as well as \(T_{\text{case}}\) for the NOC set. These values are drawn from a normal distribution with \(\mu = 1\) and \(\sigma = 1\).

3. Simulate \(X_{\text{NOC}}\) and \(X_{\text{case}}\) using eqs. (10) and (11) respectively, \(\epsilon\) is normally distributed noise with \(\mu = 0\) and \(\sigma = 1\) and \(\alpha\) is a parameter that controls how large the contribution of the case-related effect is, this controls the difficulty of the problem. A low value of \(\alpha\) corresponds to a challenging problem.

\[
X_{\text{NOC}} = T_{\text{NOC}}Q_{\text{NOC}}^T + \epsilon.
\]

(10)

\[
X_{\text{case}} = T_{\text{case}}Q_{\text{case}}^T + \alpha T_{\text{case}}Q_{\text{case}}^T + \epsilon.
\]

(11)

This procedure was repeated 100 times and the results reported are averages. The number of principal components has been fixed on 3, the simulated number of NOC components.

4.2. Digitally spiked real data

To test the methodology on a more realistic dataset while still being able compare the findings to a known true value \(^1\)H NMR data on healthy individuals was used. This data [20] consists of 22 healthy individuals (9 males, 11 females in the age range 25–55) whose urine has been measured during multiple occasions in 2005 during a period of about 3 months. We believe this to be a good representation of the healthy phenotype. For more information about the data and the collection
This data was split into two parts, where 8 samples per person have been used as a NOC set and the remaining samples have been spiked with a known (digital) multivariate pattern to function as an case set. This component, essentially $Q_{case}$, has been simulated analogous to the simulated data described before. The number of components used for the NOC model was determined as 24 using NUMFACT.

5. Results and discussion

5.1. Demonstration residual bias effect

Figs. 2 and 3 show the results on the simulated data (see section 3.1) that we used to visualize the residual bias effect. The loadings that are used for simulation are given in Fig. 2 and the results of the analysis in Fig. 3. The results of PCA on this simulated data are given in the supplements. We can see from these figures that the NOC components indeed hinder the interpretation of the biomarkers, as the resulting loadings for $P_{case}$ are essentially a linear combination of $P_{noc}$ and $P_{case}$, which results in the wrong direction in the data being found. This is analogous with Fig. 1. The PCA analysis on $E_{case}$ shown in Fig. 3 gives essentially the right shape but underestimates the effect for variable 1 to 4. What actually happens, is that the NOC component explains part of the variance of the case component, because they are not orthogonal to each other. In practice, this means that the amount of natural variation is overestimated, therefore too much is subtracted from the data by eq. (8). We can see this effect here, because the two components have overlap in metabolites 3 and 4 which causes them to be non-orthogonal. This causes first 4 variables in $E_{case}$ to be too small in Fig. 3 and is what causes the residual bias effect. Another insight from this figure is that choosing the correct amount of components is critical: with too many components the residual bias effect would be even more pronounced while with too few components the NOC variation would be too dominant. Fig. 4 also shows loadings estimated with the Compromise Approach. This figure shows that the compromise does not underestimate variable 1 to 4 and gives a more reliable estimate of $P_{case}$.

5.2. Simulated data

The results for the simulated data with different values of $\lambda$ are shown in Fig. 5. To compare the calculated value $P_{case}$ with the true value $P_{case}$, we calculate their Pearson correlation. The higher this correlation is, the more accurate $P_{case}$ has been found. We first examine the situation for $\lambda = 1$, the pure residuals. On the left side of the plot, where the problem of finding the correct loading is relatively challenging because of the low effect size, $\lambda = 1$ seems to perform relatively well. As the problem gets easier (corresponding to a high value of $\alpha$ in eq. (11)), however, the correlation coefficient does not exceed 0.75. This is caused by the residual bias effect and is further proof that the residuals of the NOC model alone may indeed have seriously limited interpretability. The analysis does show however, that mathematically removing natural variation may indeed aid the analysis where the natural variation is truly dominant, i.e. the left part of the plot. The situation for $\lambda = 0$, pure $X_{case}$, performs poorly for the highly challenging situations on the left while it performs progressively better as the problem difficulty is decreasing, which is in contrast with high levels of lambda. This behaviour can be explained by the presence of the natural variation: in the more challenging situations the variance of the natural variation is dominant which causes PCA to inaccurately estimate the sought component, resulting in a very low correlation. With a less challenging problem, the variance of the natural variation is less dominant. The sought loading is found relatively reliably, especially since the residual bias by definition cannot occur since no actual residuals are calculated.

The differences between analyses with $\lambda = 1$ and $\lambda = 0$ emerge as a continuous pattern for intermediate values of $\lambda$: higher values lead to a better performance on challenging problems. The $\lambda$ therefore is a tunable parameter to compromise between the residual bias effect on one hand and the hindering natural variation on the other.

In this particular simulation the situation in the middle, between an effect size of 3 and 4, is perhaps the most interesting: the loadings can be estimated accurately enough that they can be useful but not accurately enough that they are always useful. In other words, improving an already poor result or slightly improving a good result is less important than improving an average result. There is a balance between the residual bias effect and the natural variation and consequently, the results can vary greatly depending of the value of $\lambda$. Especially in this middle area of the figure, the added value of this the Compromise Approach is shown as these are the type of situations that one encounters the most in practice.

Another striking observation from this figure is that the results for $\lambda = 0$ or $\lambda = 1$, so $X_{case}$ and $E_{case}$ respectively, always perform relatively poorly in this simulation, except for the most challenging problems. Subtracting the NOC only partly leads to higher performances than subtracting everything or nothing, this shows that the Compromise Approach has
added value. The optimal value for $\lambda$, as selected using eq. (5), estimates the relative amount of variance in the case set. It thereby sets the parameter according to the problem difficulty. The effectiveness of this approach is shown by the superiority of the optimal lambda for almost all difficulty settings.

Fig. 5 also shows the results for PLS-DA, where the weights have been taken as an estimate of $P_{\text{case}}$. Only 1 latent variable has been found to be optimal so this latent variable is equal to the PLS weights. We can see that, while the PLS-DA weights seems to estimate $P_{\text{case}}$ reasonably well, the Compromise Approach performs better. This further shows the added value of our approach, especially since PLS-DA is the most widely used method for multivariate analysis of the data from such experiments. PLS-DA assumes that the two classes are separated by a difference in their means. Here, this is not the case here as the case group is somewhat heterogeneous. Therefore PLS-DA is not the optimal approach. We attempted the Variable Importance in the Projection (VIP) [21] methodology and this did not improve the results in a meaningful fashion (results not shown here).

Fig. 5 also shows the results for PLS-DA, where the weights have been taken as an estimate of $P_{\text{case}}$. Only 1 latent variable has been found to be optimal so this latent variable is equal to the PLS weights. We can see that, while the PLS-DA weights seems to estimate $P_{\text{case}}$ reasonably well, the Compromise Approach performs better. This further shows the added value of our approach, especially since PLS-DA is the most widely used method for multivariate analysis of the data from such experiments. PLS-DA assumes that the two classes are separated by a difference in their means. Here, this is not the case here as the case group is somewhat heterogeneous. Therefore PLS-DA is not the optimal approach. We attempted the Variable Importance in the Projection (VIP) [21] methodology and this did not improve the results in a meaningful fashion (results not shown here).

Fig. 6 shows this same simulation but for a single representative run. The ten highest loading weights are shown for different values of $\lambda$ as well as the simulated values and the PLS weights. The figure shows the same pattern as Fig. 5 as the PLS weights for $\lambda = 0.5$ are the most similar to the simulated values.

### 5.3. Digitally spiked data

Fig. 7 shows the results on this data where the performance again has been expressed as the Pearson correlation coefficient between the estimated and true (simulated) loading. This shows a slightly different

![Fig. 6. Ten highest loading weights for a typical simulation.](image)

In this figure the ten highest loading weights for the true simulated values as well as for three different values for $\lambda$ are shown for an effect size of 6. This corresponds to a less challenging problem where the main objective is to find $P_{\text{case}}$ as accurately as possible. It is evident from the plot that for $\lambda = 1$, the pure residuals, the weights have a rather big difference with the true values. This is due to the residuals bias effect. At $\lambda = 0$ the similarity with the true loadings is much higher but still misses a number of the peaks. $\lambda = 0.5$, the Compromise Approach, gives the best performance. Finally, the PLS weights in the bottom are a good approximation of the true loadings but not as accurately as those of the Compromise Approach. These results are in line with the results shown in Fig. 5.

![Fig. 7. Correlation with true effect against effect size for spiked real data.](image)

This figure shows the same pattern as Fig. 5 where the value of lambda seems independent of the problem difficulty and the Compromise Approach performs best except for effect sizes $< 2$. Much lower values of $\lambda$ seem optimal here however. The lowest effect size corresponds to about ten times as much NOC variance than the effect variance and two times as much for the high effect size.
picture than Fig. 5. The Compromise Approach still gives the most accurate results for most effect sizes, but the PLS weights are more accurate on the very challenging side. Also, lower values of \( \lambda \) seem optimal. Finally, the automatic procedure for selection lambda is again performing well. This plot shows that the added value of the Compromise Approach is not limited to purely simulated data but can also occur in the case of a real-world NOC dataset. When the simulation parameters are changed to make the case group more heterogeneous the PLS weights become less and less accurate, which further shows that PLS-DA is not optimal for heterogeneous diseases (results not shown here).

Example applications of our methodology are cohort studies with a group of healthy controls and a group of individuals in various stages of a disease. Asthma [22] might be a good example of this, as it involves a complex and heterogeneous response where methods that rely on response homogeneity such as PLS-DA may fall short. Removing the natural variation present in individuals will allow increased focus on the more relevant disease-related metabolome.

6. Conclusion

Natural variation may be a hindrance when analysing an experiment consisting of a control group and a case group. We examine the residual approach where we project the case group on a model for the control group and analyse the residuals, in essence subtracting the control variation from that data. We show that an effect we named the residual bias effect occurs and that it may be a serious problem. This effect causes misinterpretations, as the residuals are orthogonal to this model, especially if the problem at hand is otherwise easy. If an underlying effect that is being sought is partly being explained by the model, that part will be lost when analysing the residuals.

We propose a method to alleviate this effect in which we remove only a part of the model. This creates a compromise between the residuals and the raw data, which leads to a more reliable analysis in the sense that it gives better interpretable models. We have shown the effectiveness of this approach on both simulated data as well as well as digitally spiked real-world data. We show that in the specific situation where we subtract all of the variation, corresponding to a pure residual approach, the residual bias effect may cause grave misinterpretations of the biomarker metabolites relevant for a given effect.

The Compromise Approach could be used on its own, to find an effect as accurately as possible. Another option would be to incorporate it in other, already existing residual-based approaches. This way, it could improve the interpretation in Statistical Health Monitoring, process control and other fields of quantitative analysis.

Funding

This work was supported by an NWO/ECHO grant [grantnr: 711.014.007].

Acknowledgements

We would like to thank Claudio Luchinat and Leonardo Tenori of the CERM Department of the University of Florence for kindly allowing us to use the urine spectra data.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.chemolab.2018.01.007.

References