Sparse statistical health monitoring: A novel variable selection approach to diagnosis and follow-up of individual patients


ARTICLE INFO

Keywords:
Sparse Mahalanobis distance
Multivariate statistical process monitoring
Variable smearing
Metabolomics
Precision medicine
Human disease diagnosis

ABSTRACT

The –omics technologies are becoming increasingly important in health care and are expected to contribute to personalized health care. In a typical experiment, cases and controls are compared as a two-class classification problem. This approach is often unsuitable, for example, because the classes are not well defined due to associated populations being biologically too heterogeneous. Recently, statistical health monitoring (SHM) was introduced as a complementary approach to allow for predictions at the individual level. This approach could be of use in all sorts of applications such as diagnosis of rare diseases, analysis of individual patterns in disease manifestation, disease monitoring, or personalized therapy.

SHM uses the framework of statistical process monitoring (SPM) in a clinical setting. The method essentially combines estimation of Mahalanobis distances (MD) with principal component analysis (PCA) to evaluate the difference in the –omics data of an individual subject to a normal reference range (normal operating conditions). It is well known from SPM, however, that reliable identification of the variables primarily responsible for this difference is hampered by the smearing effect, which is a result of the PCA step. To avoid this problem, we propose to combine estimation of the MD with variable selection via an l1-norm penalty instead of using dimension reduction. This way a sparse MD metric is obtained.

The effectiveness of this method is illustrated by several simulation studies and its application to urine 1H-NMR metabolomics data for diagnosis of multiple inborn errors of metabolism.

1. Introduction

Metabolomics refers to the use of high-dimensional analytical technologies for the global and unbiased measurement of metabolites in a single biological sample such as urine or blood [1–3]. Through measuring these small molecules, the metabolic phenotype (metabotype) of individuals is studied to inform on their health status. The metabotype is dependent on the complex interplay between the genetic profile and environmental factors such as the gut microbial composition, lifestyle and diet, and varies greatly between individuals and populations. Because of this metabolomics has (in concert with other ‒omics techniques) an application in population-based health care as well as precision medicine [2,3]. For example, metabolomics has been used to diagnose cancer states, diabetes, cardiovascular diseases, neurological diseases and inborn errors of metabolism (IEM) amongst others [2,4–7].

Commonly, a metabolomics experiment involves hundreds to thousands of measured variables per patient. Therefore, analysis of the acquired data with (multivariate) statistical approaches is a crucial step e.g. to diagnose a disease or monitor the healthy state [7,8]. Typically, a case-control approach is used where two-class classification techniques such as LDA, (O)-PLS-DA, and SVM are used to model the metabolic differences between the groups of samples [7,8]. Despite many successes, this approach might be impractical in a clinical setting. It assumes that both groups are well defined biologically and that the metabolic differences observed at the group (population) level carry over to the level of the individual patient [9]. Often, however, this is not the case due to individual exposures and response characteristics to e.g. disease or treatment [10]. In other words, the case-control comparison may not be a two-class problem at all. The individual responses are hypothesized to overlap in two-class classification models and may therefore be very challenging to detect this way [11]. Moreover, in the
coined the approach statistical health monitoring (SHM). It industrial processes to the clinical approaches beyond population-based case-control models are required. not be detected or falsely interpreted. Clearly, statistical modeling case of diagnosis of rare diseases typically too few samples are available for a useful two-class comparison between patients and controls; a separate model is required for each disease; and unknown metabolic perturbations, i.e. caused by an unexpected response to treatment may not be detected or falsely interpreted. Clearly, statistical modeling approaches beyond population-based case-control models are required.

Recently, Engel et al. transferred the class-modeling framework used in e.g. statistical process monitoring (SPM) for monitoring of industrial processes to the clinical field as an alternative or complementary tool for comparing cases to controls in metabolomics [4]. They coined the approach statistical health monitoring (SHM). It offers an intuitive way to assess the disease response in individual patients, where disease is seen as an extreme of the normal range (normal operating conditions; NOC) or as an abnormal deviation from normal functioning within this range. Note that SHM is not limited to this application and can be used for identification of any abnormal metabotype compared to the control samples (e.g. due to treatment). We use disease diagnosis as a running example throughout this text. An example is shown in Fig. 1. A set of control samples is used to multivariately define the NOC represented by the black ellipse in the figure. The NOC can be defined for the population, or for the individual.

In this example, the patient sample falls outside the black ellipse and is marked as significantly different. A second step is used to determine which pattern of variables is primarily responsible for this difference (the abnormal variables). This way, disease responses are assessed in an individual manner. Engel et al. successfully used SHM for diagnosis of orphan diseases [4]. The method was used by Del Carratore et al. for biomarker discovery for hepatocellular carcinoma [9]. They showed that by means of the individual predictions of SHM several metabolites that were identified as a possible biomarker by a two-class classification approach were false positive identifications. Recently, Marquand et al. proposed a similar approach (normative modeling) to study brain functioning on the basis of functional magnetic resonance imaging (fMRI) data [12].

The two-dimensional example above uses the Mahalanobis distance (MD) to compare the patient to the NOC [4]. However, it is well known that the MD cannot be estimated reliably for high-dimensional metabolomics data where the number of control samples is (much) smaller than the number of variables [13]. SHM circumvents these issues by projecting the data to a lower dimensional space using principal component analysis (PCA), similar to soft independent modeling of class analogy (SIMCA) and several SPM approaches [4,14–18]. Hotelling’s $T^2$ statistic and the Q-statistic are used to determine whether a given patient deviates from the NOC in the space spanned by the selected principal components and its orthogonal complement, respectively. Contribution plots based on the relative partial decomposition criterion are used to identify the variables (metabolites) that mainly contributed to this deviation. A major drawback of the PCA step is, however, that it allows for interaction between the abnormal variables (i.e. variables on which the patient differs from NOC) and the normal variables [19,20]. Because of this, reliable identification of the abnormal variables is challenging (normal variables appear to be abnormal). This is sometimes referred to as the smearing effect and hampers disease diagnosis. Additionally, selection of the correct number of principal components is a crucial step [21]. In our experience, standard selection approaches such as screeplots do not provide a clear solution for metabolomics data.

In this paper we propose an alternative SHM method, that is, sparse SHM (sSHM). The key idea consists of combining the MD with variable selection through an $l_1$-norm constraint. This way a sparse MD metric is obtained. Variable selection offers multiple advantages. Firstly, subtle differences between the patient and NOC can be detected this way while they could be masked by the accumulative noise effect of irrelevant variables when no variable selection is applied (i.e. when the MD is used). Secondly, dimension reduction by PCA is not required anymore for analysis of high-dimensional data. This way the smearing effect is avoided. Because of this, sSHM identifies abnormal variables more reliably compared to SHM (which uses PCA) while having similar power to detect patients that deviate from NOC as will be shown by simulation. A fast algorithm is employed to compute the entire solution path of the model, i.e. it finds all solutions with $1, 2, \ldots, p$ selected variables, where $p$ corresponds to the total number of variables in the data. This path can be graphically depicted and provides an intuitive way to identify those variables on which the patient differs most from the NOC, i.e. the disease response of the patient.

The next section will outline the concepts of sSHM and its mathematical background. We expect the method to also be applicable to problems encountered in other fields such as industrial process monitoring and food authentication. Therefore, this section also highlights the connections sSHM to some of the class-modeling approaches that are used in these fields [14–18]. The subsequent sections detail the simulation study that was carried out to compare the properties of SHM to those of sSHM, and the application of sSHM to a case study involving the diagnosis of orphan diseases based on the metabolic profile of individual patients.

### 2. Theory

In sSHM an individual patient sample is compared to the NOC in two steps. In the first step the patient data is matched against the NOC and marked as normal (healthy) or possibly abnormal (possibly ill). Note that a similar strategy is used in industrial process monitoring (’healthy’ data is obtained when the process is in-control) and food authentication (’healthy’ samples correspond to a certain product or food type) [14,17]. When a sample is marked as abnormal a second step is used to identify the abnormal variables, i.e. the disease response in this individual.

Below, it is first described how the Mahalanobis distance can be used for analysis of low-dimensional data. Next, the sSHM model, which combines the MD-statistic with variable selection, is introduced for analysis of high-dimensional problems. Due to variable selection more subtle differences between the patient and NOC can be detected in the first step. Additionally, the disease response can be better identified in the second step. Finally, the SHM model and its associated issue of variable smearing are briefly reviewed.

#### 2.1. The Mahalanobis distance

To be able to compare a patient to NOC, a set of healthy control

![Fig. 1. Comparison of a patient (marked red) to a set of controls (marked blue) by the Mahalanobis distance. Note that the solid black line corresponds to the 95% confidence sphere, i.e. the normal range. The dotted arrow indicates the direction of canonical variate a, which is used to determine on which variables the patient differed most from the normal range. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).](image-url)
samples \( (X_n) \) that represent the NOC well has to be selected. Note that the choice of control samples defines what kind of patterns in the patient sample will be marked as abnormal. For example, if the patient data contains signal due to paracetamol intake, but the NOC data does not, this signal will be marked as abnormal. More details regarding the choice of NOC samples in the context of SHM are provided in [4]. We assume that the distribution of the NOC samples is multivariate normal (possibly after a suitable transformation of the data).

First, the patient is compared to the NOC by the squared Mahalanobis distance \( (MD^2) \) [22]:

\[
MD^2 = (x_p - \mu_a)\Sigma^{-1}(x_p - \mu_a)^T > c
\]

where \( x_p, \mu_a, \) and \( \Sigma \) indicate the row vector of patient data, the mean vector of the control samples, and covariance matrix of the control samples, respectively. Geometrically, expression 1 tests whether the patient sample falls inside the confidence sphere of the control samples, where the upper limit \( c \) is traditionally derived from a scaled \( F \)-distribution [22]. Note that the sphere defines the NOC: any sample that falls inside the sphere is marked as normal. The location and shape of the NOC sphere are defined by \( \mu_a \) and \( \Sigma \). An example is shown in Fig. 1. Note that Eq. (1) defines a (one-class) class model, which is also known as unequal dispersed classes (UNEQ) [15,17].

If a sample falls outside the NOC it is abnormal and must be further inspected. The MD can also be used for this step. For this purpose, another interpretation of the MD is used. More specifically, it has been shown that the squared Mahalanobis distance is the weighted distance between the scalar projections of \( x_p \) and \( \mu_a \) onto the vector \( a = [x_p - \mu_a]\Sigma^{-1} \), i.e. \( MD^2 = \{x_p - \mu_a\}^T \text{var}(X_a) \Sigma^{-1} \text{var}(X_a)^T \} \) [23]. Vector \( a \) is a canonical variate (CV) known from techniques such as Fisher LDA and MANOVA. The CV is indicated by the dotted line in Fig. 1. It is a linear combination of the variables that maximizes the difference between \( x_p \) and \( \mu_a \) relative to the variance of the controls [23]. Therefore, the coefficients of \( a \) can be studied to determine on which variables the patient sample differed most from NOC, where large absolute coefficients indicate the abnormal variables. The abnormal variables are interpreted as the individual disease response of the patient. A clinical practitioner can use this information for example for disease diagnosis.

Typically, only a few of the measured variables are related to a disease. However, all measured variables are taken into account by \( MD^2 \). As shown by Zimek et al., and also in Appendix A, the effect caused by the disease in a few variables may be masked by the normal variation in the other variables [24]. This can result in a considerable loss in power and hamper identification of the abnormal variables when inspecting direction \( a \), especially when the effect is small. Additionally, Eq. (1) cannot be applied when the number of control samples is smaller than the number of variables because the inverse of \( \Sigma \) cannot be computed.

In this work a data-driven approach is introduced that automatically takes only the most abnormal variables into account when comparing a patient to the NOC. This increases the power of the \( MD^2 \)-test and improves identification of the abnormal variables. Additionally, the method is applicable to high-dimensional data.

2.2. Sparse statistical health monitoring

Sparse SHM combines estimation of the MD with variable selection. A flowchart highlighting the main steps of the approach is shown in Fig. 2.

The defining feature of our approach is that we first estimate the canonical variate \( a \) (step 1a) before the MD is computed (step 1b). Additionally, we regularize estimation of direction \( a \) with an \( \ell_1 \)-norm constraint. This constraint is well known from techniques such as the LASSO and the elastic net [25]. It has the sparsity property in the sense that it will force some coefficients in \( a \) to be exactly zero indicating variables on which the patient was similar to the NOC. A (sparse) MD is calculated using sparse estimate(s) of \( a \) to determine if the patient differs significantly from the NOC (step 2). Note that the \( \ell_1 \)-norm constraint effectively introduces a variable selection step since only the variables with a nonzero coefficient in \( a \) contribute to the estimate of the MD. Because of this sSHM can be directly applied to high-dimensional data, and no dimension reduction using e.g. PCA is required. Sparse estimation of \( a \) is also very useful to identify the abnormal variables (step 3) since they are usually among the first variables that are selected.

2.2.1. Step 1a: sparse estimation of canonical vector \( a \)

To be able to combine estimation of canonical variate \( a \) with a constraint, we first note that \( a \) can be estimated by maximizing the Rayleigh quotient \( \{a^T(d',d)/a^T\Sigma a\} \), where \( d = x_p - \mu_a \). Wu et al. showed that maximizing the Rayleigh coefficient is equal to the following expression [26]:

\[
\hat{a} = \arg\min_a s \text{tr}(a^T\Sigma a) \quad \text{s. t.} \quad (x_p - \mu_a)^T a = 1
\]

Inspired by sparse linear discriminant analysis (LDA) [26], we propose to include an \( \ell_1 \)-norm constraint in expression 2 to obtain a sparse solution of the direction \( a \), i.e. incorporate variable selection (step 1a):

\[
\hat{a} = \arg\min_a (\Sigma a)^T a \quad \text{s. t.} \quad (x_p - \mu_a)^T a = 1, \sum_{i=1}^p |a_i| \leq \lambda
\]

where \( a_i \) indicates the \( i \)-th coefficient of \( a \), and \( \lambda \) is a fixed constant. The value of \( \lambda \) controls the amount of variables that are included in the model; when \( \lambda \) is small, most of the \( a_i \) will be exactly zero. Expression 3 can be solved in different ways. Here a fast algorithm was used as developed by Wu et al. to solve a sparse LDA problem [26,27]. To stabilize the algorithm and allow for selection of more than \( n \) variables when \( n < p \), \( 2\log(p)/n \) was added to the diagonal of \( \Sigma \) [26]. Here, \( p \) and \( n \) are the number of variables and training samples in matrix \( X_p \) respectively. This step will be referred to as regularization of \( \Sigma \).

2.2.2. Steps 1b and 2: comparison of a patient to the NOC

In sSHM, the dissimilarity between the patient and the NOC is evaluated by the squared MD using a sparse estimate of \( a \) (step 1b):

\[
MD^2 = \{(x_p - \mu_a)^T\Sigma^{-1}(x_p - \mu_a)^T\}
\]

where \( MD^2 \) indicates a sparse squared Mahalanobis distance based on \( i \) selected variables. As mentioned above, the power of the method is increased by variable selection. The variables are selected in a data-driven fashion by Eq. (3), where \( a_i \) is its solution with \( \lambda = \lambda_h \) chosen such that \( i \) variables are selected. Typically, a range of constants \( \lambda_1 < \lambda_2 < \cdots < \lambda_h \) exists such that Eq. (3) gives a solution with \( i \) nonzero coefficients (\( i \) selected variables). The value for \( \lambda_i \) in (4) is always set to the highest constant in this sequence, i.e. \( \lambda_i = \lambda_h \).

For each patient there is an optimal number of selected variables such that a possible difference with the NOC can be best observed (step 2). This number depends on how much the patient differs from the NOC in the abnormal variables with respect to the accumulating noise due to normal variation in the other variables (see Appendix A). In practice, however, this number is unknown. In this work we use two approaches to resolve this issue (see box parameter optimization in Fig. 2). The first option is a practical approach that was suggested by Wang et al. in the context of industrial process monitoring where they used domain knowledge to restrict the solution of their sparse model to \( i \) variables [28]. In other words, Eq. (4) is used for \( i \) selected variables. In the context of sSHM this means that the clinical practitioner must have a rough idea regarding how many variables are affected by e.g. a disease in the patient data. Our simulation studies show that often a significant difference from NOC can be detected by selecting roughly this expected number of variables. The second option is to determine \( i \) from the data by selecting that number of variables for which the largest distance
$M_D^2$ is observed. However, this is not a straightforward task since distances based on different amounts of selected variables are not directly comparable [24]. Therefore, a normalization step is used to fairly compare the $M_D^2$-values:

$$MD_{\text{max}}^2 = \max_{i=1-p} \frac{MD_i^2 - E(MD_i^2)}{\text{Var}(MD_i^2)}$$

(5)

where $E(MD_i^2)$ and $\text{Var}(MD_i^2)$ are the mean and variance of $MD_i^2$ for control samples, respectively. The values for $E(MD_i^2)$ and $\text{Var}(MD_i^2)$ are estimated from the control data by means of leave-one-out cross-validation (LOO-CV). Zou et al. and Capizzi et al. also used this approach to select the optimal number of variables in their sparse industrial process control methods [29,30].

After selecting a specific number of variables a significance test is required to determine if the patient indeed differs from NOC. For this purpose, the upper limit of $MD^2$ and $MD_{\text{max}}^2$ is estimated from a generalized extreme value distribution. Note that the significance test does not test whether the sample is significantly different from the NOC on the specific variables selected by Eq. (3) for that sample. The fact that each sample (control and patient) may differ most from NOC on completely different variables is taken into account. More details are provided in Appendix B.

2.2.3. Step 3: identification of abnormal variables

As described above, the first steps of sSHM are to use Eqs. (3)–(5) to detect a significant difference between the patient and the controls. The third step involves interpretation: the goal is to identify most (all) abnormal variables, e.g. the individual disease response. Again, variable selection is useful in this respect. In principle, the abnormal variables should be selected first by the model and can thereby be identified this way. However, it is unclear how many abnormal variables are present in a sample and should therefore be selected. Expression 5 seems to be useful to automatically select all relevant variables. However, our simulations show that this test (that is designed to best detect a difference between a patient and NOC) does not always reliably identify all of the abnormal variables (see e.g. Fig. 7). Similarly, when the coefficients of $a_\lambda$ for a specific number of selected variables are inspected it can be that too many noise variables are included masking the relevant variables (see e.g. Fig. 4a), or that not all relevant variables are selected. Therefore, we propose to inspect multiple solutions with different numbers of selected variables together in the so-called solution path figure to identify most abnormal variables. Examples will be provided in the results section in Figs. 5 and 7 and in Appendix F.

The solution path offers an intuitive way to see in which order the variables were selected by sSHM. Based on this information the variables can be ranked, where it is assumed that the most abnormal variables are selected first (step 3 in Fig. 2). A clinical practitioner can inspect the top ranked variables to determine the patient’s response to e.g. a disease, where the number of variables to inspect can be chosen based on practical considerations (i.e. it is not feasible to inspect more than 50 variables). Additionally, as soon as it is clear that a number of variables corresponding to a specific diagnostic marker have been...
selected, the solution path offers an intuitive way to quickly assess the selection rank of other variables that are biologically related to this marker.

2.3. Statistical health monitoring

As mentioned above, SHM can also be used to compare a patient sample to the NOC. Below, we briefly review the main steps in SHM. For more details we refer the reader to [4].

SHM combines the MD (Eq. (1)) with dimension reduction instead of variable selection (as in sSHM, see Section 2.2) [4]. More specifically, PCA is used to reduce the dimension of the data and compare a patient sample to the NOC in two different subspaces:

$$MD^2 = (x_p - \mu_p)\Sigma^{-1}(x_p - \mu_p)^T = (x_p - \mu_p)\Sigma^{-1}P^T(x_p - \mu_p)^T + (x_p - \mu_p)\Sigma^{-1}P^TP^T(x_p - \mu_p)^T = T^2 + T^2_{\text{res}}$$

(6)

where the columns in $P$ and the diagonal elements in matrix $\Lambda$ indicate the $k$ eigenvectors (PC) and eigenvalues of $\Sigma$ that are retained in the model, respectively. The matrices $\Sigma$ and $\Lambda$ indicate the $p-k$ residual eigenvectors and eigenvalues. The statistic $T^2$ is used to monitor the principal component subspace spanned by the $k$ PCs. The residual space is typically monitored by the $Q$-statistic since $T_{\text{res}}^2$ cannot be applied when the control data has more variables than sample $[4]$:

$$Q = |\|p|p|^2$$

(7)

where $e_p = (x_p - \mu_p) - (x_p - \mu_p)PP^T = (x_p - \mu_p)(I - PP^T)$ is the residual information in the patient data that is not captured by the first $k$ principal components, and $I$ indicates the identity matrix.

Identification of abnormal variables in SHM (for outlying samples as judged by the $T^2$ and/or $Q$-statistic) is based on inspection of contribution plots. These highlight the contribution of each variable to the $T^2$ or $Q$-statistic $[4,14,19]$. For example, contributions to the $Q$-statistic can be estimated using the complete decomposition criterion:

$$q_i = \left| (x_p - \mu_p) - (x_p - \mu_p)PP^T \right|^T (x_p - \mu_p)\Sigma^{-1}P^T (x_p - \mu_p)\Sigma^{-1}P^T (x_p - \mu_p)\Sigma^{-1}P^T$$

(8)

where $\Sigma^{-1}P^T$ is the $i$th column of matrix $\Sigma$ and $q_i$ indicates the contribution of variable $i$ to the $Q$-statistic $[4,19]$. Variables with the highest contribution values are flagged as most abnormal. From Eq. (8) it can be seen that contributions to the $Q$-statistic are computed by projection of the data onto a lower dimensional space spanned by the selected principal components $(x_p - \mu_p)P$ (data compression) and subsequent expansion to the original measurement space $(x_p - \mu_p)PP^T$ $[20]$. This allows for the ‘interaction’ between the abnormal variables and the normal variables $[19,20,31]$. Because of this, reliable identification of the abnormal variables is challenging (e.g. normal variables can have high contributions and appear to be abnormal while abnormal variables can be obscured). This is sometimes referred to as the variable smearing effect $[19,20,31]$. Note that variable smearing also occurs for the $T^2$-statistic and other contribution criteria (e.g. partial decomposition), and is also expected to occur when dimension reduction is carried out using other PCA-based approaches (e.g. sparse PCA) $[19,20]$. Variable smearing greatly hampers reliable identification of abnormal variables, as will be shown in Section 2.4 and was also shown numerous times in the industrial process monitoring literature $[19,20,31]$. This is a major drawback of the SHM approach. The variable smearing effect was the main motivation of the work developed in Section 2.2. It is avoided in sSHM since no dimension reduction by PCA is used.

3. Method

We use simulated and real data to investigate the properties of the sSHM model under different structures of the NOC. The simulation study was also used to compare sSHM to competitive methods such as SHM. All methods were applied according to the protocols described in the original papers $[4,14,30]$. In the main text of this paper, we focus on the comparison between sSHM and SHM with respect to identification of the abnormal variables in a patient (step 3). We refer the reader to appendices D and E for a comparison of the classification accuracy (as NOC or non-NOC), sensitivity and specificity of the approaches (step 2).

3.1. Simulation design

Throughout the simulation it was assumed that the distribution of healthy controls was multivariate normal $N(\theta, \Sigma)$. The covariance matrix $\Sigma$ was constructed by multiplication of a predefined correlation structure $R$ with variance values drawn from the uniform distribution $U(0,1,16)$.

It is difficult to simulate correlation structures $R$ that closely resemble the complex structures of metabolomics data. Therefore, we studied three “simple” structures to be able to systematically explore the properties of the sSHM method. Note that the simulated data contained some aspects of real metabolomics data such as correlations between variables, grouping of variables (e.g. metabolic pathways), and irrelevant noise variables. Additionally, we applied sSHM to real NMR metabolomics data to study its performance for more complicated correlations (see Section 3.2 below). The following correlation structures were considered in the simulation study:

1. R1 Common correlation: all variables were correlated to each other with value $\rho$
2. R2 Block-diagonal: blocks of variables were correlated to each other with value $\rho$. The different blocks were uncorrelated. Each block contained 10 variables. The first 10 variables corresponded to the first block, variables 11 to 20 to the second block, etc.
3. R3 Toeplitz structure: variables close to each other were more highly correlated compared to variables that were far apart. The distance between variables with indices $p1$ and $p2$ was defined as $|p1-p2|$. For example, the distance between the first and fourth variable in the data is equal to three. The correlation between variables was given by $\rho^{2(1-p2)}$

The correlation $\rho$ was set to 0, 0.4, 0.6, and 0.8. The correlation structures are visualized in Fig. 3.

In the simulation the number of variables in the data set was varied between 10 and 100. Additionally, the number of control samples that was used to train the sSHM model (i.e. used to estimate $\mu_i$ and $\Sigma$ in Eq. (3)) was 50 or 1000. These samples were drawn from $N(\theta, \Sigma)$ with a specific covariance matrix as defined above. Next, test samples were simulated. These samples were different from NOC on two variables. Correct identification by SHM is not guaranteed in this case due to variable smearing $[14]$. Therefore, this simulation design was enough to highlight the main difference between SHM and sSHM for identification of the disease response in individual patients. The amount of difference was varied by constant $f$, which ranged from 0 to 3 in steps of 0.25. We will refer to this constant as the fault magnitude. Note that the test samples were similar to the controls when the fault magnitude was 0. In this case sSHM should not mark them as significantly different from NOC. A thousand samples were simulated for each value of $f$ as follows to ensure that the canonical vector $a$ was (approximately) sparse (had two non-zero coefficients):

$$x_p = x_0 + f(\xi + \zeta)\Sigma$$

(9)

where $x_0$ indicates a row vector that was drawn from $N(\theta, \Sigma)$, i.e. a control sample. Note that half of the samples had a negative sign in $f$. The two abnormal variables were defined by unit vectors $\xi$ and $\zeta$, where $i$ and $j$ indicated their indices. For each simulated sample, the
indices \( i \) and \( j \) were drawn from a uniform distribution. Prior to construction of the models, all simulated samples were autoscaled to the mean and standard deviation of their corresponding control samples.

3.2. Diagnosis of inborn errors of metabolism

To assess the value of sSHM for disease diagnosis, a set of urine samples of 193 healthy children and a set of 24 patients with an IEM was measured using proton NMR spectroscopy [4]. Eighteen patients were known to suffer from one of seven different IEM. For the other six patients, no IEM was diagnosed, but signals related to commonly prescribed drugs such as depakine and paracetamol were found in the NMR spectra by visual inspection of the data by a clinical expert. A subject had to be between 4 and 12 years old to participate in the study and be of Dutch ancestry. No other selection criteria such as lifestyle or diet were imposed. The study was approved by the medical ethical committee of the Radboud University Medical Centre in Nijmegen, The Netherlands. More details regarding the data, including information regarding the measurement and subsequent processing of the NMR spectra, can be found in [4]. Briefly, the regions 0.2–4.7 ppm and 5.0–10.0 ppm were selected for further analysis. Next, the urine NMR spectra were normalized to the creatinine concentration to correct for dilution effects. Equidistant binning with a bin size of 0.04 ppm was used to reduce the normalized data from 30,888 measurements to 246 bins. Finally, Pareto scaling data was applied to the data for reasons justified in Section 4.3.

Since outliers in the set of controls can heavily influence an sSHM model, the spectra of the 193 healthy children were inspected using robust PCA [32]. Seventeen samples with abnormal patterns related to dietary influences and drug intake were identified. These samples were marked as abnormal and used to validate the sSHM model since detection of abnormal patterns due to diet and drugs is in principle no different from the detection of abnormalities related to disease. The set of 24 patients was also used for this purpose. Additionally, another set of 56 samples from the remaining 176 healthy controls was used to validate the model. The remaining healthy control samples were used to define the NOC and train the sSHM model. Based on their value in Eq. (5), it was observed during cross-validation of the 120 training samples that 2 of these samples greatly differed from the others. Visual inspection of the data showed that this was due to bad baseline correction and water suppression. Therefore, these samples were excluded from further analysis.

To clearly demonstrate the advantage of variable selection by sSHM the analysis was repeated 15 times where each time the data was concatenated with a block of 100 additional random variables (noise bins). The bins were normally distributed with zero mean and a standard deviation chosen such that the intensity of the noise bins was roughly equal to the median peak intensity of the control samples.

4. Results

4.1. Individual analysis of the simulated samples

As mentioned in Section 3.1, the main properties of the sSHM model were studied by simulation. A large number of control and patient samples were simulated for different structures of the NOC. Below, the analysis of a single simulated sample is described to demonstrate how the method can be used on a case-by-case basis. Next, sSHM is compared to SHM where we focus on identification of abnormal variables (step 3) since the main differences between the methods become visible in this step. Additional details of the simulation study

Fig. 3. A graphical representation of (a) the common variance, (b) the block diagonal, and (c) the Toeplitz, correlation structures with \( \rho = 0.6 \) that were used in the simulation study. Structures (a–c) are equal to an independent structure when the correlation is zero \( \rho = 0 \). This is shown in panel d.
Table 1: Accuracy of different sSHM tests for health monitoring of a simulated patient sample.

<table>
<thead>
<tr>
<th>Number of selected variables</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manual</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.00</td>
</tr>
<tr>
<td>10</td>
<td>0.01</td>
</tr>
<tr>
<td>25</td>
<td>0.03</td>
</tr>
<tr>
<td>75</td>
<td>0.26</td>
</tr>
<tr>
<td>100</td>
<td>0.55</td>
</tr>
<tr>
<td>Automatic</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.00</td>
</tr>
</tbody>
</table>

such as the empirical type I (1-sensitivity) and type II (1-specificity) error rates of the sSHM model (for the detection of normal and abnormal samples, respectively) are presented in Appendix C. We refer the reader to Appendix C for a comparison of the sensitivity and specificity of SHM and sSHM.

4.1.1. Steps 1 and 2: detection of abnormal samples

Here, we consider a single simulated sample \( (f=2) \) for the case where the correlation matrix of the control samples had a block diagonal structure with \( \rho = 0.4 \) and 50 training samples were available. The simulated sample was automatically compared to the NOC by sSHM (step 2). As shown in Table 1, a significant difference could be observed due to variable selection. When using a significance level of 5%, for example, no significant difference was observed when all variables were taken into account (i.e. when the normal MD was used): the effect of the abnormal variables was masked by the “noise” of the other variables. In contrast a clear significant difference was observed when only 2 variables were selected. The difference became less clearly visible with additional selected variables. This was expected since the simulated sample contained only two abnormal variables. In real applications the number of variables that should be selected is unknown. As shown in the table, the simple standardization suggested in Eq. (5) (marked as automatic in the table) ignored enough irrelevant variables such that a difference between the patient and the NOC was observed: the observed p-value was similar as the p-value found for the solution with 2 selected variables. Similar results were obtained for the other simulations (see appendix C).

4.1.2. Step 3: identification of abnormal variables

The observed significant difference between the simulated sample and the NOC suggested that the sample should be further investigated to identify which variables were abnormal. Fig. 4a shows the estimates of the coefficients of the canonical vector \( a \) that were obtained when all variables were selected. The abnormal variables (solid red circle, solid red line) could barely be distinguished from the irrelevant variables (blue empty circle, blue dotted line). Identification was greatly improved by the variable selection step in sSHM as shown in Fig. 4b. In this case the abnormal variables were easily identified since they were the only two variables that were selected.

In this simulation it was known that only 2 variables were abnormal. In real applications, however, this number is not exactly known and it can be difficult to select the right solution to inspect. Therefore, it is preferred to inspect multiple solutions together in the so-called solution path figure (step 3). The path found for this simulation is shown in Fig. 5. The solution path displays the estimated coefficients of \( a \) for all possible values of the \( \ell_1 \)-norm constraint (see Eq. (3)). The lines in the figure are the paths of the coefficients shrinking towards zero as a function of the penalty. This allows a clinical practitioner to see how specific variables behaved as a function of the constraint. Note that the x-axis has been scaled to have a maximum value of one.\(^1\) A value close to 0 means that the \( \ell_1 \)-norm constraint was large and only a few variables were selected (had nonzero coefficient), while a value of 1 indicates that all variables were selected. The solutions corresponding to Fig. 4a and 4b can be found at 0.22 and 1, respectively. Moving from left to right in the solution path figure provides an intuitive way to see at which points a variable had a nonzero coefficient, i.e. was selected. In this case, the solution path clearly shows that the abnormal variables (red solid lines) were the first two variables that were selected. This demonstrates that the information in the solution path can be used to rank the variables, where it is assumed that the abnormal variables are among the first selected ones.

Note that the solution found by automatic variable selection with expression 5 selected only one of the two relevant variables. Therefore, inspection of the solution path is preferred for identification of abnormal variables.

4.2. Comparison between SHM and sSHM

The main motivation for the development of sSHM was the unreliable identification of abnormal variables by SHM. Below, the SHM and sSHM model are compared with respect to analysis of the simulated data. The analysis of the data by SHM was carried out as described in Section 2.3 and in [4]. It was observed that SHM and sSHM had similar type I (1 – sensitivity) and type II (1 – specificity) error rates with respect to identification of significant differences between the simulated samples and the NOC (step 2). More details are provided in Appendix E.

As expected, sSHM greatly outperformed SHM with respect to identification of the abnormal variables in the samples (step 3). This is shown in Fig. 6 for the case that 50 training samples were available to define the NOC and 100 variables. To allow for a fair comparison between the different approaches it was assumed that all samples were correctly marked as significantly different from NOC. The x-axis in the figure expresses the difference between the NOC and the simulated patient as defined by the fault magnitude \( (f) \) in Eq. (9). The y-axis shows the percentage of simulated samples in which the abnormal variables were correctly identified. Correct identification was achieved by sSHM when the first two selected variables were the abnormal ones. Identification of abnormal variables in SHM was achieved by studying the relative contribution values (of the Hotelling \( T^2 \) and Q-statistics) [4]. Correct identification was achieved when the two abnormal variables had the highest contribution values.

From Fig. 6 it is clear that for sSHM and SHM the identification rate improved when the difference between the patient and the NOC was larger. However, as shown in Fig. 6 the identification rate of SHM remained below 40% for the values of \( f \) considered. This poor identification rate is attributed to the smearing effect (see Section 2.3). As mentioned earlier, sSHM does not use dimension reduction with PCA and does therefore not suffer from the smearing effect (see Section 2.3). Therefore, considerable higher identification rates were achieved with sSHM. This was also observed for other variable and sample sizes (see Appendix E).

4.3. Diagnosis of inborn errors of metabolism

Inborn errors of metabolism are a group of rare genetic defects that collectively occur in roughly 1 out of every 2500 individuals [33]. Therefore, they are an important group of diseases to consider. Unfortunately, standard classification models such as Partial Least Squares – Discriminant Analysis cannot be used to diagnose these diseases since the number of training samples is extremely limited. Recently, SHM was successfully used to analyze NMR data and diagnose several IEM [4]. Below, we will use the same data to show the value of sSHM in a practical example. As shown in Table 2, the data did not only contain abnormal patient samples related to IEM, but also abnormalities related to diet and medication. These abnormalities had already

\(^1\) The x-axis corresponds to the proportion of shrinkage \( (|A|_1/|A|) \) instead of the size of the \( \ell_1 \)-norm constraint.
been observed by visual inspection of the data by a clinical expert and were also used to validate the sSHM approach since detection of abnormal patterns due to diet and medication is in principle no different from detection of metabolites related to a disease.

First, the data were analyzed after autoscaling. An improvement of roughly 10% in terms of the percentage of correctly identified patient samples (specificity) was observed due to variable selection by sSHM. However, subsequent correct identification of the abnormal metabolites was hampered by the large influence of baseline signal on the model due to the autoscaling. Therefore, it was decided to apply Pareto scaling to the data instead. Additionally, the regularization constant (see Section 2.2.1) was increased to \( \frac{p_n \log(n)}{n} \). Sensitivity and specificity of the model were not affected by this stronger regularization. However, based on inspection of the solution path of a single sample, fewer baseline bins were amongst the first variables selected (see Section 5 for more details).

As shown in Table 3, the effect of variable selection on the number of patient samples that were correctly flagged as such (specificity) was improved by using sSHM. However, the sensitivity of the model decreased in some cases due to the removal of baseline bins from the model.

### Table 2

<table>
<thead>
<tr>
<th>Disease (IEM)</th>
<th>n</th>
<th>Dietary</th>
<th>n</th>
<th>Medication</th>
<th>n</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>3ß-Hydroxy-D-(\Delta_2)-steroid dehydrogenase deficiency</td>
<td>1</td>
<td>Cycamate (artificial sweetener)</td>
<td>3</td>
<td>Depakine(^a)</td>
<td>2</td>
<td>Bacterial contamination</td>
</tr>
<tr>
<td>3-Methylcrotonyl CoA carboxylase deficiency</td>
<td>1</td>
<td>Fish</td>
<td>5</td>
<td>Paracetamol</td>
<td>5</td>
<td>High Taurine signal (cause unknown)</td>
</tr>
<tr>
<td>5-Oxoprolinuria</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>Unknown (possibly due to nutrition)</td>
</tr>
<tr>
<td>Alkaptonuria</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Cystinuria</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Formiminotransferase deficiency</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Isovaleric aciduria</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(^a\) Intake of piracetam or sabril medication was also detected in these samples.
less clear for the Pareto-scaled data. sSHM in combination with a large number of selected variables could also be used to identify the abnormal samples. This is attributed to the reduced influence on the model of noise signal close to the baseline of the NMR spectrum. To clearly show the advantage of the variable selection step for the Pareto scaled data, the analysis was repeated 15 times. In each repetition the data was concatenated with a block of 100 irrelevant variables. As shown in Table 3, the percentage of correctly classified samples was clearly improved by variable selection in this case. Note that the highest accuracy and specificity were obtained when 2 variables were selected, although the automatic selection by Eq. (5) was very competitive. All abnormal samples due to an IEM were correctly identified by sSHM. Four other samples were not correctly identified. The abnormality in these samples was related to cyclamate (2x), paracetamol, and bacterial contamination.

After a sample had been marked as abnormal, the next step (step 3) was to identify the abnormal metabolites. Again, it was observed that expression 5 could not be used in this step as indicated in e.g. Fig. 7a. Via inspection of the solution path figures, however, all IEM were successfully diagnosed and all dietary and medication abnormalities were successfully identified. Two examples are presented in Fig. 7. The solution paths of the other samples are shown in Appendix F. In these figures the solid red lines indicate the variables (resonances in the NMR spectrum) that were known to be related to the IEM and other abnormalities in the data. For clarity the relevant features are indicated in the NMR spectrum in panels b and d. In each case the variable selection rank (VS) indicates when the variable was selected by the sSHM model, where “1” corresponds to the first variable that was selected (the first nonzero coefficient in the left of panel a), etc. The IEM and other abnormalities were correctly diagnosed when enough of the relevant (red) variables had low VS values.

In the solution path of the first example in Fig. 7a the first four bins that were selected (VS 1 – 4) corresponded to resonances around 6.76 ppm and 3.64 ppm. This indicated that the metabolic profile of this patient contained abnormal amounts of homogentisic acid. Thanks to this, the patient was diagnosed with the IEM alkaptonuria, an inborn error of metabolism affecting tyrosine catabolism. In this example all known resonances (red lines) were selected first by the model and the diagnosis was clearly made. The second example is shown in Fig. 7c. This example is considered more difficult since it involved many

<table>
<thead>
<tr>
<th>Number of irrelevant bins</th>
<th>Number of selected bins</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Manual</td>
<td>2</td>
<td>95.1</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>95.1</td>
<td>100</td>
<td>90.2</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>95.1</td>
<td>100</td>
<td>90.2</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>93.9</td>
<td>100</td>
<td>87.8</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>93.9</td>
<td>100</td>
<td>87.8</td>
</tr>
<tr>
<td></td>
<td>Automatic</td>
<td>1–246</td>
<td>95.1</td>
<td>100</td>
</tr>
<tr>
<td>100</td>
<td>Manual</td>
<td>2</td>
<td>95.1</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>93.7</td>
<td>99.6</td>
<td>87.8</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>92.8</td>
<td>98.8</td>
<td>86.8</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>92.7</td>
<td>99.6</td>
<td>85.7</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>92.3</td>
<td>99.9</td>
<td>84.7</td>
</tr>
<tr>
<td></td>
<td>Automatic</td>
<td>1–346</td>
<td>94.3</td>
<td>99.8</td>
</tr>
</tbody>
</table>

Fig. 7. (a, c) Variable selection paths of sSHM models applied to the NMR data of a urine sample of two patients. The relevant diagnostic resonances are indicated by the solid red lines. The vertical dotted line indicates the solution of the maximum test (Eq. (5)). (b, d) Plot of the NMR spectra highlighting the relevant diagnostic resonances and their variable selection rank (VS). The metabolites are markers for (a, b) Alkaptonuria disease, and (c, d) intake of paracetamol by the patient.
resonances in many different parts of the NMR spectrum. The bins centred around 2.16, 5.14, 7.13, 7.31, 7.45, 9.67 and 9.78 ppm were clearly marked by the model as abnormal. All these features had VS values lower than 20. This suggested that the metabolites acetaminophen, acetylsalicylic acid-glucuronide, and acetylsalicylic acid-sulphate were present in high concentrations. This was caused by intake of paracetamol by this individual.

5. Discussion

In this work, sSHM was introduced for (1) assessment of the deviation from the normal range and (2) identification of the variables related to this difference in individual patients. The properties of the method were investigated by simulation and compared to those of SHM. Additionally, sSHM was successfully applied to real NMR metabolomics for diagnosis of several IEM. The same data was previously investigated by the SHM model in chapter 5 [4]. Therefore, sSHM and SHM can be compared with respect to this application.

The methods had similar power to identify normal and abnormal samples. Compared to SHM, the specificity (correct identification of patient samples) of SHM was slightly higher for this dataset. The specificity of SHM, however, dropped considerably after the addition of additional noise variables while that of sSHM remained very robust as shown in Table 3. Additionally, interpretation of the SHM contribution plots for identification of abnormal variables was not straightforward because of two reasons: (1) the contribution values couldn’t be reliably interpreted due to the so-called smearing effect (see Section 2.3 and Appendix E), and (2) the contributions of the relevant variables differed by orders of magnitude making it difficult to visually inspect the results. It must be remarked, though, that for this data set inspection of the highest contributions identified enough key variables as abnormal to make a correct diagnosis. In contrast, the sSHM solution path offered an easy tool to identify the abnormal variables as shown in Fig. 7. The method does not suffer from the smearing effect allowing for much more reliable identification of these variables. This was confirmed by repeating the simulation study described in Section 4.3 with the correlation structure of the $^1$H-NMR data. This clearly shows that sSHM is a useful alternative to SHM in a clinical setting.

The idea of combining the Mahalanobis distance with variable selection is not new. Several other methods have been proposed in the context of industrial process monitoring [28–30]. These methods apply variable selection directly to the difference between the patient and the centre of the controls, i.e. $x_p - \mu$. However, this does not guarantee that the canonical variate $a = (x_p - \mu)^\top \Sigma^{-1}$ is sparse. The CV is directly penalized by sSHM. Preliminary simulation studies shown in Appendix E suggest that sSHM is therefore better able to identify abnormal variables in many situations. Similar results have been found for the closely related problem of sparse LDA [34].

It has been shown for sparse linear regression that addition of an additional $l_1$-norm (ridge-type) penalty can improve prediction and identification of important variables in cases of highly collinear and/or ultra-high dimensional data [25]. Such a penalty results in SHM in $\Sigma$ in expression 3 being replaced by $\Sigma_{\delta} = \Sigma + \delta I$, with tuning parameter $\delta$ [26]. This form of regularization decorrelates the data; matrix $\Sigma_{\delta}$ approaches a diagonal matrix for large $\delta$ and the sSHM procedure will essentially select the largest absolute values of $(x_p - \mu)$. The algorithm used to solve expression 3 uses $\delta = 2 \log (p)/n$ as suggested by Wu et al. in the context of sparse LDA [26]. This was justified by the empirical observation that $l_1$-norm regularization seemed to stabilize the algorithm, while the power was fairly robust to the choice of $\delta$ [26]. This was also observed for SHM in the simulation study in Section 4.2. For analysis of real NMR metabolomics data, however, variable selection was improved by increasing $\delta$ to $4 \log (p)/n$ (see Section 4.3). Additionally, we note that $\delta = 2 \log (p)/n$ is suggested for autoscaled data (where the diagonal elements of $\Sigma$ all equal 1), but might be suboptimal for differently processed data (with much larger diagonal elements of $\Sigma$, for example). This suggests that the choice of tuning parameter $\delta$ (next to the number of variables to select) offers an interesting direction for further improvement of sSHM. However, it is not immediately clear how such tuning should be achieved, since the optimal values of $\delta$ for detection of patients that are outside of NOC and selection of a limited set of important abnormal variables in these patients do not need to be the same. As mentioned in Section 4.3, the solution path might offer some insight regarding sensible values for $\delta$ since variables that clearly correspond to baseline signal in e.g. an NMR spectrum should not be among the first selected variables. An interesting direction for future research is offered by stability selection [35]. Essentially, this method aggregates the results of multiple variable selections applied to subsamples of the data. For sparse linear regression (the LASSO) it has been shown that variable selection may be markedly improved in this way [35]. In a preliminary investigation we randomly selected values for $\delta$ from a range of pre-defined values, each time fitting an sSHM model to a subsample of the data presented in Section 4.3. Most of the variables known to be abnormal were consistently among the first selected variables, while the baseline signal had a lower rank.

In this study it was shown that sSHM ranks the variables from abnormal to normal by variable selection. A clinical practitioner can use this information, which is visualized in the solution path figure, to diagnose a disease. Currently, however, no clear guidelines exist on how many of the top-ranked variables should be inspected. Eq. (5) could clearly not be used for this purpose, nor is this expression based on firm statistical theory. It would be interesting to develop another strategy to obtain an upper bound. Stability selection also offers an interesting direction for future research in this respect since it allows for error control on the expected number of falsely selected variables [35]. Another interesting direction for future research is to employ an initial variable screening method to narrow the number of variables before application of sSHM to a sample [36]. This could extend the applicability of the method to larger data sets such as LC-MS data.

6. Conclusion

In this work an $l_1$-norm penalized Mahalanobis distance was proposed for identification of disease biomarkers in individual patients. Here, a disease is seen as an extreme from the normal range. We demonstrated improved identification of abnormal variables compared to the use of PCA-based contribution plots (well known from statistical process monitoring), which suffer from the smearing effect. It was shown how the method can be used for analysis of metabolomics data for rapid screening of individual patients for a multitude of (rare) diseases.

The method also offers perspectives in the framework of precision medicine and in non-life science applications such as industrial process monitoring and food authentication.

The source code of the method will be made available at the following github repository: https://github.com/JasperE/sSHM.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.chemolab.2017.03.003.

References


