Metabolomics (2012) 8:S17–S28
DOI 10.1007/s11306-011-0327-y

ORIGINAL ARTICLE

Concurrent class analysis identifies discriminatory variables from metabolomics data on isovaleric acidemia

Gerhard Koekemoer · Marli Dercksen · James Allison · Leonard Santana · Carolus J. Reinecke

Received: 1 April 2011 / Accepted: 15 June 2011 / Published online: 7 July 2011

Abstract Metabolomics data are typically complex and high dimensional. Multivariate dimension-reducing techniques have thus been developed for analysing metabolomics data to disclose underlying relationships, with principal component analysis (PCA) as the technique mostly applied. Despite its widespread use in metabolomics, PCA has shortcomings that limit its applicability. Several approaches have been made to overcome these limitations and we describe an advanced disjoint PCA (DPCA) model, termed concurrent class analysis and abbreviated as CONCA. CONCA is a new model, and is unique in linking DPCA models to a traditional PCA model. This is accomplished by restructuring the input data matrix, applying DPCA group models to the restructured data, and combining the DPCA models in order to replicate a traditional PCA. We applied the CONCA model to a metabolomics data set on isovaleric acidaemia (IVA), a rare inherited metabolic disorder. The outcome showed that three of the variables with high discrimination value identified through the CONCA analysis are prominent organic acid biomarkers for IVA. Moreover, three further minor metabolites associated with the disease, and two as a consequence of treatment, were likewise identified as important discriminatory variables. The benefit of the CONCA model thus is its ability to disclose information concerning each individual group and to identify the variables important in discrimination (VIDs) which are also responsible for group separation.

Keywords Metabolomics · Biomarker identification · Variables important in discrimination · Concurrent class analysis · Isovaleric acidemia

1 Introduction

The general aim of metabolomics is to identify, measure and interpret the complex time-related or steady-state concentration of low molecular weight substances, called metabolites (designated as variables in bioinformatics applications), in cells, tissues, biofluids, such as blood, serum, saliva or cerebrospinal fluid, and in air, like breath (reviewed in Harrigan and Goodacre 2003). These substances are predominantly of endogenous origin, such as products of intermediate or secondary metabolism, collectively known as the metabolome, defining the between-subject variations and thus the normal ranges of excretion of metabolites in man (Chalmers et al. 1976). Variables may, however, also include exogenous substances derived from diet, gut flora, medication or from contaminated air, water or any other environmental source, reflecting possible within-subject variations in excretion profiles. Variables that may be measured objectively and are indicators of normal biological processes, pathological conditions or therapeutic interventions, are designated as biomarkers

G. Koekemoer
Statistical Consultation Services, North-West University, Potchefstroom Campus, Potchefstroom, South Africa

M. Dercksen · C. J. Reinecke
Centre for Human Metabolomics, North-West University, Potchefstroom Campus, Potchefstroom, South Africa

J. Allison · L. Santana (✉)
Division for Statistics, School for Mathematics, Statistics and Computer Science, North-West University, Potchefstroom Campus, Potchefstroom, South Africa
e-mail: leonard.santana@nwu.ac.za

(Mamas et al. 2011). Metabolomics studies in humans seek to define biomarkers related to the diagnosis and the efficacy of therapeutic interventions or prognosis of disease.

Untargeted metabolomics analyses, or metabolic/metabolite profiling, attempt to identify as many variables as possible, following a well-defined procedure for data generation. Various statistical and bioinformatic methods are concomitantly applied in the analysis of the complex data sets generated in untargeted metabolomics and metabolite profiling (Van den Berg et al. 2006). Gas chromatography-mass spectrometry (GC-MS) is one method frequently used in untargeted metabolomics investigations and produced the illustrative data that are used here.

Generation of GC-MS data involves complex upstream experimental protocols, including the choice of analytical procedure to isolate the section of the metabolome to be investigated, derivatization of the isolated substances to increase their volatility and temperature resistance, and the separation and identification procedures to generate the raw data (Styczynski et al. 2007). Subsequently, alternative ways of data preprocessing are used to produce a clean data set of normalized peak areas that reflect the concentrations of variables of the sets of experimental groups under investigation. That is followed by data pretreatment to produce a data matrix of observations (n) and variables (p) suitable for data analysis (Van den Berg et al. 2006). The purpose of the data analysis is to disclose relevant biological or other information related to the object under investigation and to reduce the influence of unrelated interfering factors such as measurement noise. The aim of this paper is to describe an approach for this final stage of downstream metabolomics analysis, directed to the identification of biomarkers that discriminate between normal and perturbed metabolic conditions and on therapeutic interventions aimed at normalization. For this purpose a GC-MS-derived data matrix was chosen which reflected the perturbations due to isovaleric acidemia (IVA), a well-defined inherited metabolic disease (reviewed by Sweetman and Williams 2001). The observations were of untreated and treated IVA patients and matched controls; the upstream experimental protocol that was followed to generate the data was not taken into consideration here.

From the information contained in a data matrix, it is possible to establish a relationship between metabolite levels and physiological responses, which provides a powerful means of exploring the biochemical consequences of disease and treatment (Weckwerth and Morgenthal 2005). Given the structure of a metabolomics matrix ($n << p$), standard parametric statistical methods such as regression are not easily applicable as there are insufficient data for parameter estimation. As implied, uninduced biological variation may produce a metabolite profile unrelated to the perturbation under investigation and statistical data analysis methods are typically not able to make this distinction (Van den Berg et al. 2006). Multivariate dimension-reducing techniques have thus been developed as standard practice for analysing metabolomics data to disclose underlying relationships. Principal component analysis (PCA) (Jolliffe 2002) evolved into the technique most used.

Despite its widespread use in metabolomics, PCA has shortcomings that limit its applicability. The results of PCA (the loadings) are often not intuitive, and the optimized function (which captures variance) is not necessarily ideal for biomarker discovery (Styczynski et al. 2007).

Several approaches have been made to overcome these limitations (Wold et al. 1996; Van den Berg et al. 2009; Nyamundanda et al. 2010). In this paper we briefly discuss one of these approaches, the disjoint PCA (DPCA) described by Wold (1976). We then propose a related, but more advanced model, termed concurrent class analysis (CONCA). The traditional PCA provides one model for all groups under consideration. The DPCA expands the traditional PCA by using one model for each group under consideration. CONCA is a new model, as described in Sect. 3 of this paper. It is unique in linking DPCA models to a traditional PCA model. This is accomplished by restructuring the input data matrix, applying DPCA group models to the restructured data, and combining the DPCA models in order to replicate a traditional PCA. It is a novel approach that is similar to a traditional PCA model with the added benefit of providing further information concerning each individual group and an ability to identify which variables have discriminatory power and are responsible for group separation. The application of the CONCA model is evaluated by using a well-defined metabolomics data set, generated from controls as well from untreated and treated IVA cases, and illustrated by a comparison of the DPCA and CONCA models by using the same data set. An advantage of the IVA metabolomics data is that they are derived from three classes of experimental subjects with distinctly different physiological and pathological characteristics. This thus provides an appropriate and interesting data set for assessment of the applicability of the CONCA model.

The purpose of our paper is not to present CONCA as a classification model, but rather as a PCA model. The best performance (in terms of separation between groups and variance extracted) of our model can never exceed the performance one can expect from a traditional unsupervised PCA. Hence our method attempts to explore the natural separation of groups given the metabolic profile. The added benefit of using the CONCA model (when compared to traditional PCA) is that, when natural separation occurs, we are able to

- obtain additional information concerning the discriminatory role of each variable in the separation by interpreting the "inner" and "outer" loadings, and
- rank the variables according to their discrimination power a measure that addresses the variables important in the separation more directly.

2 Materials and method

2.1 Experimental subjects and samples used

Urine samples from 10 homozygous untreated and treated IVA patients (all South African Caucasian cases) and 22 control cases (matched according to age, gender and ethnicity) were used to generate the metabolomics matrix. The patient samples were those referred to the Laboratory for Inherited Metabolic Disorders at North-West University (Potchefstroom Campus) from paediatricians requesting analyses to detect a possible metabolic disorder in their patients. A clear clinical picture, including information on clinical biochemical analyses and treatment with medication, were obtained for each patient from their clinician. The patients were diagnosed with IVA from interpretation of the standard organic acid and L-carnitine assays. A high-carbohydrate diet treatment supplemented with L-carnitine and glycine was applied as a treatment regime to stimulate detoxification of toxic secondary metabolites. Repeat samples were collected after initiation of this treatment to access the detoxification process of each patient. A complete organic acid analysis for the metabolomics study was conducted on the original urine sample from each patient (designated the untreated cases) and one sample from each of the same patients after reaching a successful level of clinical and metabolic improvement following the treatment (designated the treated cases). Informed consent was given by all parents of patients in accordance with the ethical requirements of North-West University.

2.2 Generation of the metabolomics data matrix

The organic substances were isolated from the urine, and consisted mostly of organic acids, derivatized and separated by GC according to a procedure that we have refined to a standardized GC-MS protocol (Reinecke et al. 2011), since our first description of a South African patient with organic aciduria (Erasmus et al. 1985). Automated Mass Spectral Deconvolution and Identification System software (AMDIS, version 2.66 from the National Institute for Standards and Technology) was used to perform component peak identification and spectral deconvolution. The semi-quantitative identification of the organic acids was conducted according to Chen et al. (2009). All organic acids identified above the detection limit of the equipment used were expressed as a relative concentration of mmol organic acid per mol creatinine.

The original data consisted of 212 metabolites that occurred in at least one of the control or patient (untreated or treated) samples. Variable reduction was performed using a "60%-rule". Any variable which occurred at least 60% of the time in any one of the three experimental groups (Group 1: Controls (C); Group 2: Original samples used for the first diagnosis of untreated IVA patients (P); Group 3: Samples collected after treatment of an IVA patient, taken at a metabolic stable condition (T)) were retained in the data matrix (compare also the approach of Bijlsma et al. 2006). This yielded a total of 113 variables, of which 86 were derived from human metabolism and are designated as endogenous substances. The remaining 27 were not directly related to human metabolism as they originated from the diet, medication or other external source, and are designated as exogenous substances. The endogenous variables (essential for comparing the consequences of the inherited genotype aberration of IVA for the phenotype, thereby separating groups 1 and 2) as well as the exogenous variables (essential for determining the effect of an external dietary treatment intervention, so separating groups 1 and 3 as well as groups 2 and 3), was retained in the subsequent analysis. The final data matrix, which we could use for traditional PCA, a DPCA and the CONCA analyses, thus consisted of observations from the experimental subjects ($n = 21 + 10 + 10 = 41$) and the reduced number of variables derived from the untargeted urinary organic analysis ($p = 113$).

2.3 Development of the CONCA model

PCA is often used as a dimension reduction technique. In metabolomic experiments the researcher is often interested in discrimination between various defined groups. As a first step a PCA is often employed as an unsupervised pattern recognition procedure, and to identify initial important variables.

Let

$$\underset{(n\times p)}{\mathbf{X}} = \begin{bmatrix} X_{1,1} & \cdots & X_{1,p} \\ \vdots & \ddots & \vdots \\ X_{n,1} & \cdots & X_{n,p} \end{bmatrix}$$

be the observed (scaled, but not centered) data with p variables and n cases. Furthermore, let the mean matrix be given by

$$\underset{(n\times p)}{\bar{\mathbf{X}}} = \begin{bmatrix} \bar{X}_1 & \cdots & \bar{X}_p \\ \vdots & \ddots & \vdots \\ \bar{X}_1 & \cdots & \bar{X}_p \end{bmatrix},$$

where $\bar{X}_i = \frac{1}{n}\sum_{j=1}^{n} X_{j,i}$. A principal component (PC) model for the centered data is given by:

$$\underset{(n\times p)}{\mathbf{X} - \bar{\mathbf{X}}} = \underset{(n\times p)}{(\mathbf{X} - \bar{\mathbf{X}})}\ \underset{(p\times k)}{\mathbf{P}}\ \underset{(k\times p)}{\mathbf{P}'} + \underset{(n\times p)}{\mathbf{E}}$$

where $\mathbf{P}$ is the loadings matrix, k is the number of components extracted, and $\mathbf{E}$ is a matrix of error terms.

DPCA is a variation of the simple PCA and involves the construction of separate PC models within each class under consideration. This partitioning of the PCA allows for an enhanced ability to recognize patterns in a data set. Wold (1976) shows that this partitioning also improves one's ability to interpret the relevance of variables, identify outliers among objects and distances between different classes.

Lamboy (1990) discusses the use of DPCA in the identification of unknown plant specimens and also provides a step-by-step procedure for applying the technique. Five desirable attributes associated with the technique can be found in this article.

For other applications of DPCA see, e.g., Bicciato et al. (2003) and Su et al. (2007). Su et al. (2007) also developed an improvement on the DPCA by incorporating a genetic algorithm. They applied the technique to identify differentially expressed genes based on microarray gene expression profiles by assessing the ability of combinations of the genes to distinguish groups. A genetic algorithm was used to solve the combinatorial optimization problem associated with this method.

Consider the following notation related to the DPCA models:

Let G denote the number of groups (classes) and $n_1, n_2, \ldots, n_G$ be the size of each class, so that $n = n_1 + n_2 + \cdots + n_G$. Also denote the data matrix of the ith class, $i = 1, 2, \ldots, G$, by:

$$\underset{(n_i\times p)}{\mathbf{X}^{(i)}} = \begin{bmatrix} X^{(i)}_{1,1} & \cdots & X^{(i)}_{1,p} \\ \vdots & \ddots & \vdots \\ X^{(i)}_{n_i,1} & \cdots & X^{(i)}_{n_i,p} \end{bmatrix},$$

and the mean matrix of the ith class by:

$$\underset{(n_i\times p)}{\bar{\mathbf{X}}^{(i)}} = \begin{bmatrix} \bar{X}^{(i)}_1 & \cdots & \bar{X}^{(i)}_p \\ \vdots & \ddots & \vdots \\ \bar{X}^{(i)}_1 & \cdots & \bar{X}^{(i)}_p \end{bmatrix},$$

where $\bar{X}^{(i)}_j = \frac{1}{n_i}\sum_{l=1}^{n_i} X^{(i)}_{l,j}$.

The resulting disjoint PC model for class i is then given by:

$$\underset{(n_i\times p)}{\mathbf{X}^{(i)} - \bar{\mathbf{X}}^{(i)}} = \underset{(n_i\times p)}{(\mathbf{X}^{(i)} - \bar{\mathbf{X}}^{(i)})}\ \underset{(p\times k_i)}{\mathbf{P}_i}\ \underset{(k_i\times p)}{\mathbf{P}_i'} + \underset{(n_i\times p)}{\mathbf{E}_i} \quad (1)$$

for $i = 1, 2, \ldots, G$, where k_i is the number of PCs extracted for class model i.

Wold uses the following measure of discriminating power for variable r ($r = 1, 2, \ldots, p$):

$$D_r = \left\{ \sum_{i=1}^{G} \sum_{\substack{j=1 \\ j\neq i}}^{G} \sum_{c=1}^{n_j} \left(\varepsilon^{(i,r)}_{c,j}\right)^2 \right\} \Bigg/ \left\{ (G-1) \sum_{i=1}^{G} \sum_{c=1}^{n_i} \left(\varepsilon^{(i,r)}_{c,i}\right)^2 \right\}, \quad (2)$$

where $\varepsilon^{(i,r)}_{c,j}$ denotes the error of the cth case in class j when fitting data from class j to the model of class i. In other words, this measure involves the calculation of the ratio of the sum of the squared errors where $i \neq j$ (i.e., the sum of all the squared errors when fitting model i on class j) to the sum of squared errors when fitting data from class i to the model of class i (i.e., the sum of all the squared errors when fitting model i on class i).

Having a model for individual classes is advantageous, because the coefficients (loadings) of the models can be compared (since the input data were scaled). However, the model proposed by Wold consists of separate PCA class models which do not link the class models together. The CONCA model, however, is able to concurrently accommodate all classes and their inter-relationships and is also able to provide information concerning each separate class. The CONCA model, where a PC model for each class is developed and then combined to model the total data structure, will now be described.

Consider the following partitioning of the data matrix $\mathbf{X}$ into $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_G$:

$$\underset{(n\times p)}{\mathbf{X}} = \underset{(n\times p)}{\mathbf{Y}_1} + \cdots + \underset{(n\times p)}{\mathbf{Y}_i} + \cdots + \underset{(n\times p)}{\mathbf{Y}_G},$$

where

$$\underset{(n\times p)}{\mathbf{Y}_i} = \begin{bmatrix} \mathbf{0}_1 \\ \vdots \\ \mathbf{0}_{i-1} \\ \mathbf{X}^{(i)} \\ \mathbf{0}_{i+1} \\ \vdots \\ \mathbf{0}_G \end{bmatrix}$$

and $\underset{(n_i\times p)}{\mathbf{0}_i} = \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix}$, for $i = 1, 2, \ldots, G$.

In addition, the mean matrix of $\mathbf{X}$ can be written as:

$$
\begin{aligned}
\underset{(n\times p)}{\bar{\mathbf{X}}} &= \begin{bmatrix} \bar{X}_1 & \cdots & \bar{X}_p \\ \vdots & \ddots & \vdots \\ \bar{X}_1 & \cdots & \bar{X}_p \end{bmatrix} \\
&= \frac{n_1}{n}\begin{bmatrix} \bar{\mathbf{X}}^{(1,1)} \\ \bar{\mathbf{X}}^{(1,2)} \\ \bar{\mathbf{X}}^{(1,3)} \\ \vdots \\ \vdots \\ \vdots \\ \bar{\mathbf{X}}^{(1,G)} \end{bmatrix} + \cdots + \frac{n_i}{n}\begin{bmatrix} \bar{\mathbf{X}}^{(i,1)} \\ \bar{\mathbf{X}}^{(i,2)} \\ \bar{\mathbf{X}}^{(i,3)} \\ \vdots \\ \vdots \\ \vdots \\ \bar{\mathbf{X}}^{(i,G)} \end{bmatrix} + \cdots \\
&\quad + \frac{n_G}{n}\begin{bmatrix} \bar{\mathbf{X}}^{(G,1)} \\ \bar{\mathbf{X}}^{(G,2)} \\ \bar{\mathbf{X}}^{(G,3)} \\ \vdots \\ \vdots \\ \vdots \\ \bar{\mathbf{X}}^{(G,G)} \end{bmatrix} \\
&= \underset{(n\times p)}{\bar{\mathbf{Y}}_1} + \cdots + \underset{(n\times p)}{\bar{\mathbf{Y}}_i} + \cdots + \underset{(n\times p)}{\bar{\mathbf{Y}}_G},
\end{aligned}
$$

where $\bar{\mathbf{X}}^{(i,j)} = \underset{(n_j\times p)}{\bar{\mathbf{X}}^{(i)}}$. Our proposed CONCA model is then derived as follows:

$$
\begin{aligned}
\underset{(n\times p)}{\mathbf{X}} - \underset{(n\times p)}{\bar{\mathbf{X}}} &= \underset{(n\times p)}{(\mathbf{X}-\bar{\mathbf{X}})}\,\underset{(p\times k)}{\mathbf{P}}\,\underset{(k\times p)}{\mathbf{P}'} + \underset{(n\times p)}{\mathbf{E}} \\
&= [(\mathbf{Y}_1-\bar{\mathbf{Y}}_1)+\cdots+(\mathbf{Y}_i-\bar{\mathbf{Y}}_i)+\cdots \\
&\quad +(\mathbf{Y}_G-\bar{\mathbf{Y}}_G)]\mathbf{P}\mathbf{P}'+\mathbf{E} \\
&= \big[\big\{(\mathbf{Y}_1-\bar{\mathbf{Y}}_1)\mathbf{P}_1\mathbf{P}_1'+\mathbf{E}_1\big\}+\cdots \\
&\quad +\big\{(\mathbf{Y}_i-\bar{\mathbf{Y}}_i)\mathbf{P}_i\mathbf{P}_i'+\mathbf{E}_i\big\}+\cdots \\
&\quad +\big\{(\mathbf{Y}_G-\bar{\mathbf{Y}}_G)\mathbf{P}_G\mathbf{P}_G'+\mathbf{E}_G\big\}\big]\mathbf{P}\mathbf{P}'+\mathbf{E} \\
&= \big[(\mathbf{Y}_1-\bar{\mathbf{Y}}_1)\mathbf{P}_1\mathbf{P}_1'+\cdots+(\mathbf{Y}_i-\bar{\mathbf{Y}}_i)\mathbf{P}_i\mathbf{P}_i'+\cdots \\
&\quad +(\mathbf{Y}_G-\bar{\mathbf{Y}}_G)\mathbf{P}_G\mathbf{P}_G']\mathbf{P}\mathbf{P}'+(\mathbf{E}_1+\cdots+\mathbf{E}_i+\cdots \\
&\quad +\mathbf{E}_G)\mathbf{P}\mathbf{P}'+\mathbf{E}. \\
&= \mathbf{T}\mathbf{P}'+(\mathbf{E}_1+\cdots+\mathbf{E}_i+\cdots+\mathbf{E}_G)\mathbf{P}\mathbf{P}'+\mathbf{E}.
\end{aligned} \tag{3}
$$

In the model above the terms $\mathbf{P}_i$ and $\mathbf{E}_i$ refer to the loadings matrix and error matrix for the "inner" models, i.e., the models for each class, whereas $\mathbf{P}$ and $\mathbf{E}$ are the loadings matrix and error matrix for the "outer" model. $\mathbf{T}$ denotes the scores matrix of the CONCA model; we will denote the ith column of this matrix by $\mathbf{t}_i$. Each class can have a different number of components extracted, i.e., $k_1, k_2, \ldots, k_G$. An advantage of the CONCA model is that when the number of components extracted in each class is equal to the number of variables, p, then $\mathbf{E}_1 = \cdots = \mathbf{E}_i = \cdots = \mathbf{E}_G = \mathbf{0}$ and the model reduces to an ordinary PC model with k components extracted.

To evaluate the performance of the model we will calculate the percentage of variance extracted. Let $\mathbf{\Sigma}_I$ denote the variance/covariance matrix of the total of the "inner" models, i.e., $(\mathbf{Y}_1-\bar{\mathbf{Y}}_1)\mathbf{P}_1\mathbf{P}_1{}' + \cdots + (\mathbf{Y}_G-\bar{\mathbf{Y}}_G)\mathbf{P}_G\mathbf{P}_G{}'$, then the variance extracted by the ith component in the CONCA model is given by V_{ii}, the ith diagonal entry of $\mathbf{V} = \mathbf{P}'\mathbf{\Sigma}_I\mathbf{P}$. Let $\mathbf{\Sigma}_X$ denote the variance/covariance matrix of $\mathbf{X}$, then the percentage variance extracted of the total variation by the ith CONCA component is given by: $V_{ii}/tr(\mathbf{\Sigma}_X) \times 100$ and the cumulative variance extracted by the CONCA model is given by $tr(\mathbf{V})/tr(\mathbf{\Sigma}_X) \times 100$. The operator $tr(\cdot)$ refers to the *trace* of a matrix and is defined as the sum of the diagonal elements of the matrix. In the case where $\mathbf{E}_1 = \mathbf{E}_2 = \cdots = \mathbf{E}_G = \mathbf{0}$, the variances extracted by the CONCA model and the traditional PCA model are equal. Hence, a criterion that can be used to select the optimal number of "inner" and "outer" components in a CONCA model is to select the minimum number of components (possibly different for each group) that best approximates a traditional PCA model, where the components in the latter case can be selected via cross validation. Alternatively, the practitioner can make use of scree-plots to determine the optimal number of components.

For the purpose of discriminating between two classes l and m, we define the following two error matrices:

$$
\begin{aligned}
\underset{(n\times p)}{\mathbf{E}^{TOT}} &= (\mathbf{E}_1+\cdots+\mathbf{E}_G)\mathbf{P}\mathbf{P}'+\mathbf{E} \\
&= (\mathbf{X}-\bar{\mathbf{X}}) - ((\mathbf{Y}_1-\bar{\mathbf{Y}}_1)\mathbf{P}_1\mathbf{P}_1{}'+\cdots \\
&\quad +(\mathbf{Y}_G-\bar{\mathbf{Y}}_G)\mathbf{P}_G\mathbf{P}_G{}')\mathbf{P}\mathbf{P}'
\end{aligned}
$$

and

$$
\begin{aligned}
\underset{(n\times p)}{\mathbf{E}^{(l,m)}} &= (\mathbf{X}-\bar{\mathbf{X}}) - ((\mathbf{Y}_1-\bar{\mathbf{Y}}_1)\mathbf{P}_1\mathbf{P}_1{}'+\cdots \\
&\quad +(\mathbf{Y}_l-\bar{\mathbf{Y}}_m)\mathbf{P}_m\mathbf{P}_m{}'+\cdots+(\mathbf{Y}_m-\bar{\mathbf{Y}}_l)\mathbf{P}_l\mathbf{P}_l{}'+\cdots \\
&\quad +(\mathbf{Y}_G-\bar{\mathbf{Y}}_G)\mathbf{P}_G\mathbf{P}_G{}')\mathbf{P}\mathbf{P}'
\end{aligned}
$$

where $m < l$. Note that when the model is evaluated using data from class l, the mean of the m^{th} class is used because it is the intercept term of that class's model. We thus interpret $\mathbf{E}^{(l,m)}$ as the error matrix that results from evaluating the model (in Eq. 3) by switching the data of classes l and m. From these two error matrices we now propose using the following measure of discriminating power between two classes l and m for variable r:

$$
D_r^{(l,m)} = \frac{\sum_{i=1}^{n}\left(\varepsilon_{i,r}^{(l,m)}\right)^2}{\sum_{i=1}^{n}\left(\varepsilon_{i,r}^{TOT}\right)^2}, \tag{4}
$$

where $\varepsilon_{i,j}^{(l,m)}$ is the element found at the ith row and jth column of the matrix $\mathbf{E}^{(l,m)}$ and $\varepsilon_{i,j}^{TOT}$ is the element found at the ith row and jth column of the matrix $\mathbf{E}^{TOT}$. An overall measure of discriminating power for all the classes is then given by:

$$\sum_{l=1}^{G-1} \sum_{m=l+1}^{G} D_r^{(l,m)}. \quad (5)$$

Henceforth, the term "CONCA All" will refer to the list of variables obtained using Eq. 5 and "CONCA l, m" will refer to the list of variables obtained using Eq. 4.

A starting point for the practitioner is to use Eq. 5 to identify those variables that are globally important for discrimination across all the classes. The measure in Eq. 4 can then be used to determine the discriminatory role of the variables identified, i.e., between which classes the specific variables play a leading role in discriminating between these classes. This will become clear in the following section.

3 Results and discussion

To explore the advantages of the proposed CONCA model, the model was first fitted to the complete set of data generated from controls, untreated and treated IVA patients, and compared with a traditional PCA. These results are shown in Fig. 1. Figure 1a–c represent the scores plots of the proposed CONCA model with 3 "outer" components and the number of components for each "inner" model selected as 1, 2 and 3, respectively (see Eq. 3). Figure 1d represents the scores plot of the traditional PCA for PCs one and two (PC1, PC2). When the number of components for the "inner" model is small (Fig. 1a), the model shows a strong linear trend for each group. However, as the number of "inner" components increases (Fig. 1b, c respectively), the CONCA model expands to a traditional PCA model (Fig. 1c in comparison with Fig. 1d).

The variances extracted are 40.70, 44.99 and 46.59% for the CONCA models with 1, 2, and 3 "inner" components extracted respectively. The variance extracted by the traditional PCA with 3 components is 48.30%. Hence, in terms of variation extracted the CONCA model with 3 "inner" components for all classes compares favourably to a traditional PCA.

Thus we select 3 components based on the above considerations as well as on the biological interpretation value. Note, however, one experiences a loss of variance extracted of 1.71% compared to a usual PCA, but CONCA has the added benefit of containing information concerning the groups (classes). All subsequent results in this paper are

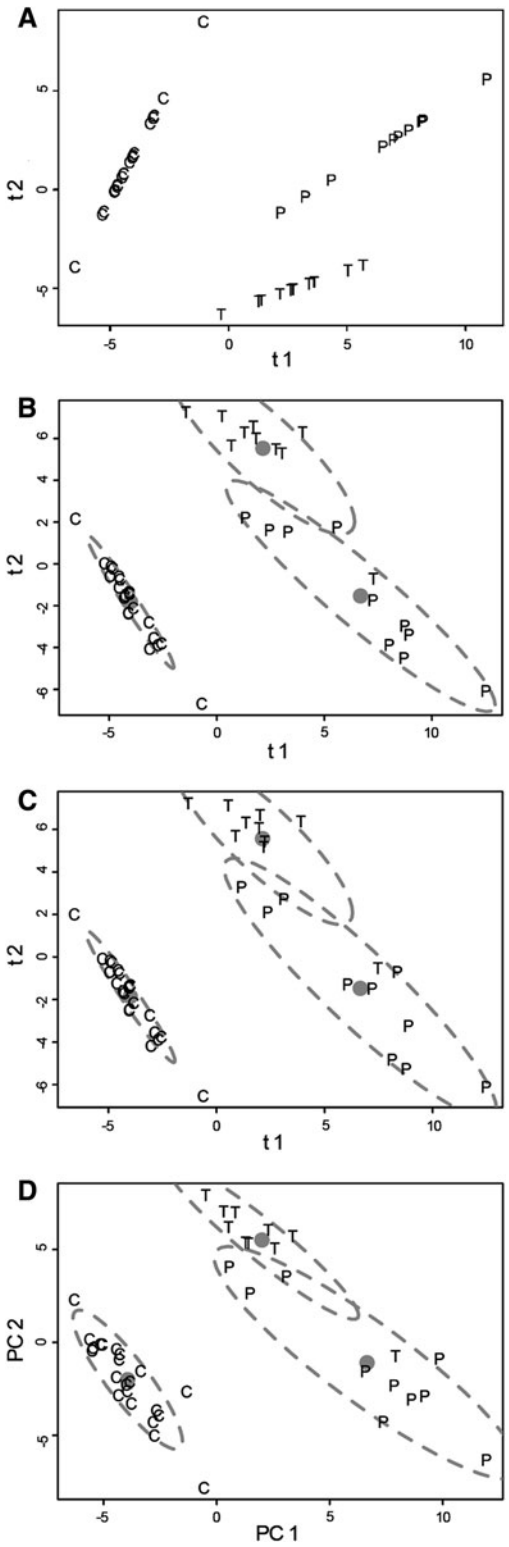

Fig. 1 Score plots for the CONCA model (**a**–**c**) and the traditional PC model (**d**). Figure 1a–c represent the CONCA score plots for the choice of 3 "outer" components and 1, 2, and 3 "inner" components for each class, respectively. Figure 1d displays the score plot for the traditional PC model with 3 components extracted. The locations of the control (C), patient (P), and treated (T) scores are represented by *elliptical contours* centered at the *dots*

based on the model with 3 "inner" components for each group and 3 "outer" components.

Second, an attempt was made to assess the applicability of the CONCA model to identify those variables important in discrimination (abbreviated as VIDs) between the various groups included in the data matrix. This should pave the way for the biological interpretation of the information on IVA patients encapsulated in the metabolomics data. The potential global VIDs across all classes were identified by the application of Eq. 5. Subsequently the discriminatory power of each variable for the various combinations of groups (1 & 2, 1 & 3, and 2 & 3) were calculated using Eq. 4. These results were finally compared with the global discriminatory power of the DPCA model (Eq. 2); shown in Table 1.

It appeared that the numerical values of the loadings of the "CONCA All" data decreased exponentially, reaching a lower plateau level after approximately 8 variables ($\sim$7% of the total number of variables in this data set). Of these 8 variables, 5 are well-known metabolites that are observed as perturbations due to IVA (*N*-isovalerylglycine, *N*-isovalerylglutamic acid, methylfumaric acid, *N*-isovalerylserine and 3-hydroxyisovaleric acid; see Sweetman and Williams 2001). The remaining three variables are also known human metabolites derived from an endogenous (3-hydroxycaproic acid) or exogenous (2,3,4-trihydroxybutyryllactone and gluco-6,1-lactone) origin, as will be described below. These three metabolites have, however, not been observed previously in IVA investigations. The relatively small list of eight variables may be regarded as potential VIDs and provides an interesting combination of metabolites for biological interpretation and for a first assessment of the CONCA model. We therefore used the first eight variables as the cut-off point for all group analyses for comparative purposes.

Application of the DPCA method to the same data set revealed four variables from the first eight variables that coincide with those from the CONCA analysis [three IVA markers (*N*-isovalerylglycine, *N*-isovalerylglutamic acid and *N*-isovalerylserine) and the exogenous 2,3,4-trihydroxy-butyryllactone]. The remaining four variables are known endogenous or exogenous metabolites, but have not been described previously for, or associated with IVA (these are 2-ethyl-3-hydroxyisocaproic acid, 4,7-decadienedioic acid, 2,3,5-trihydroxyvaleryllactone, methylmalonic acid and 2,3,4-trihydroxybutyric acid). The biological significance of this observation will not be discussed further here, as the primary focus of this paper is on the CONCA method. The variables observed for the CONCA 1,2, CONCA 1,3 and CONCA 2,3 groups, which overlap with the CONCA All list, are clearly suggestive of a discriminatory value directly associated with the phenotypic condition of the cases included in these three groups.

Furthermore, Table 2 provides information concerning the link between the variables identified in Table 1 and the "outer" component loadings, since the "outer" components are responsible for linking the "inner" PCA models together. Table 2 indicates that variables with larger "outer" loadings correspond with most of the variables identified in Table 1, but this tendency decreases as more components are added. From this we conclude that the outer loadings also contain information regarding the discriminatory role of individual variables.

The third step in the assessment of the CONCA method and the potential significance of the VIDs was to investigate the biological significance of the metabolites

Table 1 A comparison of VIDs generated using the CONCA model

Metabolite (VIDs)	Numerical value found for each VID				
	CONCA All	CONCA 1,2	CONCA 1,3	CONCA 2,3	DPCA
N-isovalerylglycine (*N*-ivgly)	100.78	47.39 (1)	51.52 (1)	NI (15)	233.24 (1)
N-isovalerylglutamic acid (*N*-ivglu)	85.45	46.08 (2)	38.29 (2)	NI (60)	172.95 (3)
Methylfumaric acid (mfa)	17.80	11.20 (3)	NI (13)	3.94 (1)	NI (11)
N-isovalerylserine (ivser)	14.43	7.94 (4)	5.12 (5)	NI (28)	37.25 (5)
2,3,4-Trihydroxy-butyryllactone	10.57	NI (28)	6.21 (3)	2.75 (4)	212.58 (2)
Gluco-6,1-lactone	10.46	NI (27)	6.05 (4)	2.74 (5)	NI (12)
3-Hydroxycaproic acid	9.90	5.50 (5)	NI (18)	NI (10)	NI (10)
3-Hydroxyisovaleric acid (3-hiva)	9.20	3.89 (7)	NI (42)	3.84 (2)	NI (16)

NI A variable not included in the first eight or the respective list of variables for that group. The values in the CONCA All list are the numerical values found for the respective variables. The values in the DPCA, CONCA 1,2, CONCA 1,3 and CONCA 2,3 lists are the numerical values found for the variables included in the CONCA All list, as well as the position of each of those variables in their respective lists shown in *brackets*. Variables observed among the first eight of the DPCA list, followed by their numerical value and position in the list in *brackets* are: 2-ethyl-3-hydroxyisocaproic acid—60.89 (4); 4,7-decadienedioic acid—58.49 (5); 2,3,5-trihydroxyvaleryllactone—55.78 (6); methylmalonic acid—51.88 (7); 2,3,4-trihydroxybutyric acid—45.14 (8). The first four of these metabolites in the DPCA list are not related to IVA. Metabolite 2,3,4-trihydroxybutyric acid derives from a high synthetic carbohydrate diet (Chalmers et al. 1976)

Table 2 Variables associated with the outer component CONCA loadings

Outer component 1		Outer component 2		Outer component 3	
Ranked variables	Loadings	Ranked variables	Loadings	Ranked variables	Loadings
N-isovalerylglycine	0.587	*N*-isovalerylglycine	0.405	3-Hydroxyphenylhydracrylic acid	0.320
N-isovalerylglutamic acid	0.392	Isocitric acid	−0.309	Citric acid	0.284
3-Hydroxybutyric acid	0.236	3-Hydroxyisovaleric acid	−0.238	Aconitic acid	0.279
3-Hydroxyisovaleric acid	0.215	3-Hydroxybutyric acid	−0.209	Oxalic acid	0.259
Methylfumaric acid	0.201	3-Hydroxyphenylhydracrylic acid	0.190	Vanillic acid	0.229
Lactic acid	0.179	Lactic acid	−0.189	Hippuric acid	0.226
Methylsuccinic acid	0.174	*N*-isovalerylglutamic acid	0.170	Isocitric acid	0.206
N-isovalerylserine	0.169	2,3,4-Trihydroxybuteryllactone	0.170	Phenol	0.171

Variables from each of the outer components were identified by the use of the loadings for each variable, ranked according to their absolute values, and thus providing the three lists of ranked variables as shown in the table. The loadings included in the table are the actual values, as determined for each variable in the respective outer components

summarized in Table 3. For this we compared the respective loadings of these metabolites. The loadings for the selected metabolites (Table 1) derived from the CONCA model were accordingly compared with those for the DPCA model, and for groups 1 (controls), 2 (untreated IVA patients) and 3 (treated IVA patients). We also include some relevant descriptive statistics in Table 3 (means and standard deviations for each variable) as well as the normal values of the variable where available. Table 3 includes aconitic acid, an intermediate of the Krebs cycle which is not regarded as a biomarker of IVA or being significantly affected by this disorder. The other variables in Table 3 are those eight found in the CONCA All variable list. A result of this comparison is that important variables tend to have higher values within a group for the respective loadings derived from the CONCA analysis, but not for the DPCA analysis. These changes in the loadings from the CONCA analysis coincide with the changes in the mean values of the respective metabolites listed in Table 3. The beneficial effects of treatment are expected to result in the opposite tendency.

Table 3 A comparison of the loadings from groups 1, 2 and 3 from the DPCA and CONCA analyses, descriptive statistics and normal values for aconitic acid and the eight variables identified by the CONCA analysis

Metabolite	Component 1 loadings of the metabolites						Descriptives						Normal values
	CONCA			DPCA			Mean (SD)						
	C	P	T	C	P	T	C		P		T		
Aconitic acid	0.214	0.155	0.208	0.205	0.087	0.051	45.08	(39.03)	46.74	(41.46)	41.43	(17.30)	26.8–189
N-isovalerylglycine	0.009	0.313	0.374	0.013	−0.002	0.028	0.19	(0.42)	1042.33	(482.67)	895.38	(400.67)	nd
N-isovalerylglutamic acid	0.000	0.211	0.199	0.000	−0.055	0.119	0	(0)	112.05	(87.07)	45.01	(49.75)	nr
Methylfumaric acid	0.002	0.112	0.034	0.003	−0.224	0.096	0.30	(0.09)	17.99	(19.78)	1.48	(2.82)	nr
N-isovalerylserine	0.000	0.088	0.066	0.000	−0.104	0.122	0	(0)	8.09	(7.71)	3.46	(6.09)	nr
2,3,4-Tri-OH-but-lac	0.000	0.000	0.109	0.000	0.000	0.040	0	(0)	0	(0)	7.92	(9.88)	Exogenous
Gluco-1,6-lactone	0.000	0.000	0.069	0.000	0.000	−0.082	0	(0)	0	(0)	3.40	(3.21)	Exogenous
3-OH-caproic acid	0.005	0.089	0.047	−0.023	−0.245	0.149	0.17	(0.35)	12.71	(16.45)	2.42	(5.12)	nr
3-OH-isovaleric acid	0.131	0.209	0.094	0.098	−0.339	0.155	8.40	(5.33)	273.45	(307.26)	6.93	(12.28)	10–67

The information is organized in 14 columns and 9 rows. The first column indicates the metabolite under consideration. Aconitic acid was added to the tables as an example of a metabolite not affected by IVA. The other eight metabolites were identified as important discriminatory variables by the CONCA-Global analysis. The numerical values of the first component for each of the three groups as identified by the CONCA analyses are listed in columns 2–4. The corresponding DPCA first component loadings are listed in columns 5–7. Columns 8–13 contain the following descriptive statistics: (1) The mean value of each metabolite as found in this investigation, expressed as mmol per mol creatinine. (2) The standard deviation for these data is displayed in *brackets* for each metabolite and The normal urinary values for each metabolite for the age group of children below the age of five, as reported by Hoffman & Feyh (2005), are given in column 14

nd Not detected—below the detection limit

nr Not reported by Hoffman and Feyh (2005), nor in the Human Metabolome Database (2010)

We subsequently describe the biological role of the 9 variables in IVA listed in Table 3. The aim of this discussion is to demonstrate the discriminatory power of the 8 CONCA All variables.

IVA is a catabolic deficiency in the degradation of the amino acid leucine, culminating in the accumulation of isovaleryl-CoA (isovaleric acid) (Budd et al. 1967). The excess isovaleric acid results in the activation of secondary pathways, e.g. glycine conjugation, L-carnitine detoxification, ω-oxidation and inhibition of the urea cycle. Secondary metabolites observed from these perturbations of the normal leucine metabolism serve as diagnostic markers for the disease. Many of these abnormal metabolites are also responsible for some of the clinical manifestations in these patients. IVA patients are consequently treated with a low-protein, high-carbohydrate diet and the additional intake of L-carnitine and glycine to facilitate detoxification of isovaleric acid and products of ω-oxidation (Sweetman and Williams 2001). The metabolic fingerprints of the variables identified by the CONCA analysis (Table 1) for the controls, untreated IVA and treated IVA groups give a clear indication of the consequences of IVA, as well as of its treatment (Table 3). For a more functional approach, these metabolic fingerprints are presented in three categories:

(1) Metabolites not significantly affected by IVA
Aconitic acid is included in Table 3 as an example to indicate the numerical values and possible changes in the loadings of the three groups in the case of a metabolite that is not significantly affected by IVA and by its treatment. This is also observed for the individual cases shown in Fig. 2a. Aconitic acid is an intermediate in the Krebs cycle, and a normal constituent of human urine. Its levels in urine vary with age, but are not specifically indicated for the disease studied in this case (Gates et al. 1988; Boulat et al. 2003). This is reflected by the numerical values of the loadings which are $\sim$ 0.2 in all three cases.

(2) Markers of IVA and its treatment

(a) Isovalerylglycine
The most important detoxification response to the accumulated isovaleric acid in IVA is through glycine conjugation, catalysed by glycine-*N*-acylase. Isovaleryl-CoA is a preferred substrate for glycine-N-acylase, resulting in highly elevated levels of isovalerylglycine in untreated patients (Fig. 2b). The additional glycine is used in treatment of the disease, and maintains the elevated levels of isovalerylglycine in the urine of treated IVA patients. Moreover, this is used to indicate which patients react to glycine supplement in the diet used for

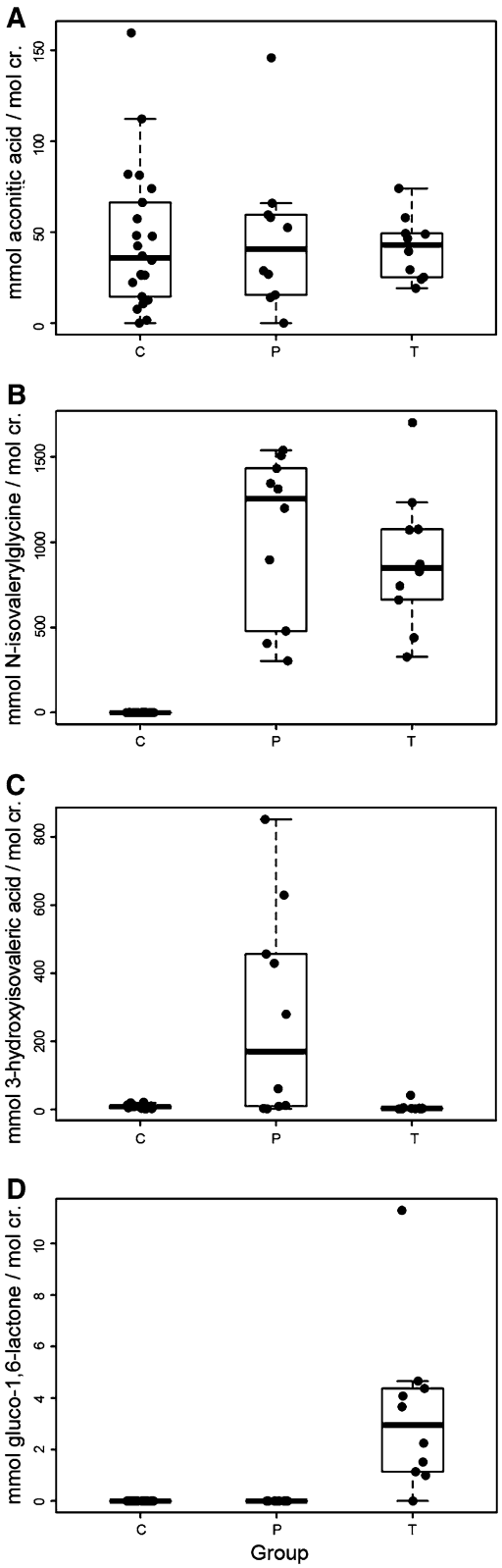

Fig. 2 A graphical representation of some descriptive statistics with superimposed data points. Box plots constructed from a five point summary of the relative concentrations (minimum, quartiles 1 to 3, and maximum) for 4 metabolites (Fig. 2a–d) for each control (C), patient (P), and treated (T) case in the three groups

treatment (Naglak et al. 1988). Higher numerical values of the loadings for isovalerylglycine are thus expected to be observed relative to the controls, as was observed for the CONCA model and shown in Table 3.

(b) Isovalerylglutamic acid
Isovalerylglutamic acid is also a detoxification conjugate and is prominent in IVA patients, linked to the inhibition of the urea cycle by cumulative isovaleryl-CoA (details of this mechanism are not presented here). The presence of this metabolite often does not differ between treated and untreated patients. This emphasizes the fact that current treatments do not necessarily improve the efficiency of the urea cycle. Ammonia levels of untreated and treated IVA patients should be monitored and dietary changes needs to be made in some patients (Coude et al. 1979; Lehnert 1981). The numerical values for the respective loadings (Table 3) obtained from the CONCA model again reflect these biological phenomena.

(c) Methylfumaric acid
Truscott et al. (1981) reported the formation of 4-hydroxyisovaleric acid due to the ω-oxidation of isovaleric acid. A cascade of secondary reactions results in the formation of methylsuccinic acid and subsequently methylfumaric acid through oxidation and dehydrogenase mechanisms (not discussed in this paper). Methylfumaric acid is evident on the profiles of untreated IVA patients. As isovaleryl-CoA is expected to be highly reduced due to its improved detoxification, it is likely that the urinary concentration of methylfumaric acid will be high in the untreated patients (with acute metabolic acidosis) and low in the treated IVA patients (Sweetman and Williams 2001).

(d) *N*-isovalerylserine
N-Isovalerylserine is only one of the isovaleric amino-acid conjugates described by Loots et al. (2005) as seen in urinary profiles of patients with IVA. Isovaleric acid is prone to taking part in acetylation reactions with various amino acids due to the cell's natural metabolic shift to detoxification. The alternative amino acid conjugation is lower in treated IVA patients due to primary detoxification through glycine and L-carnitine conjugation

(e) 3-Hydroxyisovaleric acid
3-Hydroxyisovaleric acid is a product of ω-1-oxidation of isovaleric acid. This metabolite is usually present in first-time diagnostic IVA samples as well as in those patients who are in a state of metabolic crisis (treatment is insufficient at this stage). The 3-hydroxy-isovaleric acid levels normalize if treatment is efficient with L-carnitine and glycine as seen in Fig. 2c, shifting the secondary metabolism to primary detoxification (Tanaka et al. 1968; Sweetman and Williams 2001).

(f) 3-Hydroxycaproic acid
The significance of 3-hydroxycaproic acid (also called 3-hydroxy-hexanoic acid) in IVA patients still needs to be investigated. Our hypothesis is that ketosis in IVA patients may be responsible for the formation of this metabolite during an acute metabolic crisis. Niwa and Yamada (1985) identified this compound in patients with ketoacidosis and explained its presence as a result of an impaired β-oxidation pathway, which causes a deficiency of coenzyme A, free L-carnitine and NAD^+ in the mitochondria, which is also observed in IVA patients. The control group and treated IVA patients do not indicate clear evidence for this metabolite, which corresponds with the current hypothesis.

The VIDs identified through the CONCA model thus clearly disclose the relevant metabolic events in the untreated and treated IVA patients versus the controls. Acute metabolic decomposition in IVA patients results in the significant additional formation of secondary metabolites, e.g. 3-hydroxyisovaleric acid and methylfumaric acid. With treatment, the metabolic status of the patient shifts to detoxification mode via glycine and L-carnitine conjugation, with a less serious impact on the patient's health. A secondary effect of the high-carbohydrate, low-protein diet is also observed in this model, as discussed below. This was not as evident when other differential PCA models were used.

(3) Variables reflecting treatment of IVA
At least two metabolites were included among the variables identified as being strongly discriminatory for the groups by the loadings obtained from the CONCA model (Table 1). These metabolites were not previously described for IVA, and both seem to be a consequence of the diet prescribed for the IVA patients.

(a) 2,3,4-Trihydroxybutryllactone
2,3,4-Trihydroxybutryllactone is a component of substances collectively known as the tetronics group. It is a metabolite that is not present in controls and untreated IVA patients, but the treated IVA group indicates an elevation of this

compound (Fig. 2d). This may be explained as being due to the treatment regime of IVA patients who were given a synthetic high-carbohydrate and low-protein diet, similar to observations reported some time ago on high-carbohydrate diets by Chalmers et al. (1976).

(b) Glucopyrano-1,6-lactone
Glucopyrano-1,6-lactone is part of the collective gluconolactone group. It is also a metabolite not present in controls and untreated IVA patients. The treated IVA group indicates an elevation of this metabolite. This may likewise be explained as being due to the treatment regime of IVA patients with a synthetic high-carbohydrate and low-protein diet by Chalmers et al. (1976).

These final biological interpretations of the variables identified by CONCA All indicate that the two VIDs served as indicators of the therapeutic intervention on IVA. These results might pave the way for further developments in assessing the efficacy of treatment in this disease.

4 Concluding remarks

We developed an advanced DPCA model, termed concurrent class analysis and abbreviated as CONCA. We applied this model to a data set on IVA, a rare inherited metabolic disorder. Five of the variables identified by the CONCA analysis are regarded as prominent organic acid biomarkers for IVA (Sweetman and Williams 2001). Their ranking as well as the numerical values of the PC1 loadings for group 2 (the IVA patients), are:

	N-ivgly	>*N*-ivglu	>3hiva	>>mfa	>>ivser
CONCA:	0.313	0.211	0.209	0.112	0.088
DPCA:	−0.002	−0.055	−0.339	−0.224	−0.104

It is of interest to note that for this application of the CONCA model, the order of the loadings from the CONCA analysis decreased in the same order as the decrease in the diagnostic value of the biomarkers as found in clinical studies. This observation does not imply, however, that this related description of metabolites and their loadings is an intrinsic characteristic or the model. More applications should be done to gain further insight into this phenomenon. The benefit of the CONCA model to disclose information concerning each individual group and to identify which variables are responsible for group separation and important in discriminatory power is, however, clear from this application.

Acknowledgements This study formed part of BioPAD Project BPP007, funded by the South African Department of Science and Technology. Additional financial support from North-West University and the Royal Netherlands Academy of Arts and Sciences for a Carolina MacGillavry PhD Fellowship to M. Dercksen are likewise acknowledged.

References

Bicciato, S., Luchini, A., & Di Bello, C. (2003). PCA disjoint models for multiclass cancer analysis using gene expression data. *Bioinformatics 19*, 571–578.

Bijlsma, S., Bobeldijk, I., Verheij, E., Ramaker, R., Kochhar, S., Macdonald, I., van Ommen, B., & Smilde, A. (2006). Large-scale human metabolomics studies: A strategy for data (pre-) processing and validation. *Analytical Chemistry 78*, 567–574.

Boulat, O., Gradwohl, M., Matos, V., Guignard, J., & Bachmann, C. (2003). Organic acids in the second morning urine in healthy Swiss paediatric population. *Clinical Chemistry and Laboratory Medicine 41*, 1642–1658.

Budd, M., Tanaka, K., Holmes, L., Efron, M., Crawford, J., & Isselbacher, K. (1967). Isovaleric acidemia. Clinical features of a new genetic defect of leucine metabolism.*The New England Journal of Medicine 277*, 321–327.

Chalmers, R., Healy, M., Lawson, A., & Watts, R. (1976). Urinary organic acid in man. II. Effects of individual variation on diet on the urinary excretion of acidic metabolites. *Clinical Chemistry 22*, 1288–1291.

Chen, J., Meng, C., Narayan, S., Luan, W., & Bennett, M. (2009). The use of Deconvolution Reporting Software© and backflush improves the speed and accuracy of data processing for urinary organic acid analysis.*Clinica Chimica Acta 405*, 53–59.

Coude, F., Sweetman, L., & Nyhan, W. (1979). Inhibition by propionyl-coenzyme A of *N*-acetylglutamate synthase in rat liver mitochondria: A possible explanation for hyperammonemia in propionic and methylmalonic academia. *The Journal of Clinical Investigation 64*, 1544–1551.

Erasmus, C., Mienie, L., Reinecke, C., & Wadman, S. (1985). Organic aciduria in late-onset biotin-responsive multiple carboxylase deficiency. *Journal of Inherited Metabolic Disease 8*, 105–106.

Gates, S., Sweeley, C., Krivit, W., DeWitt, D., & Blaisdell, B. (1988). On-line classification and updating of disjoint principal component models.*Chemometrics and Intelligent Laboratory Systems 3*(3), 243–247.

Harrigan, G., & Goodacre, R. (2003). *Metabolic profiling: iIs role in biomarker discovery and gene function analysis*. Boston, MA: Kluwer Academic Publishers.

Hoffman, G., & Feyh, P. (2005). Organic acid analysis. In N. Blau, M. Duran & M. Blaskovics (Eds.), *Physician 's guide to the laboratory diagnosis of metabolic diseases*, revised 2nd edition (pp. 27–44). Heidelberg: Springer-Verlag.

Human Metabolome Database. (2010). Human Metabolome Database version 2.5. Retrieved October 23, 2010 from http://www.hmdb.ca..

Jolliffe, I. (2002). *Principal component analysis* (2nd ed.). Springer, New York.

Lamboy, W. (1990). Disjoint principal component analysis: A statistical method of botanical identification. *Systematic Botany 15*(1), 3–12.

Lehnert, W. (1981). Excretion of N-isovalerylglutamic acid in isovaleric acidemia. *Clinica Chimica Acta 116*, 249–253.

Loots, D., Erasmus, E., & Mienie, L. (2005). Identification of 19 new metabolites induces by abnormal amino acid conjugation in isovaleric acidemia.*Clinical Chemistry 51*, 1510–1511.

Mamas, M., Dunn, W.B., Neyses, L., & Goodacre, R. (2011). The role of metabolites and metabolomics in clinically applicable biomarkers of disease. *Archives of Toxicology 85*, 5–17.

Naglak, M., Salvo, R., Madsen, K., Dembure, P., & Elsas, L. (1988). The treatment of isovaleric acidemia with glycine supplement. *Pediatric Research 24*, 9–13.

Niwa, T., & Yamada, K. (1985). 3-Hydroxyhexanoic acid: An abnormal metabolite in urine and serum of diabetic ketoacidotic patients. *Journal of Chromatography B: Biomedical Sciences and Applications 337*, 1–7.

Nyamundanda, G., Brennan, L., & Gormley, I. (2010). Probabilistic principal component analysis for metabolomics data. *BMC Bioinformatics 11*, 571–582.

Reinecke, C., Koekemoer, G., Van der Westhuizen, F., Louw, R., Lindeque, J., Mienie, L., & Smuts, I. (2011). Metabolomics of urinary organic acids in respiratory chain deficiencies in children. *Metabolomics*. doi:10.1007/s11306-011-0309-0.

Styczynski, M., Moxley, J., Tong, L., Walther, J., Jensen, K., & Stephanopoulos, G. (2007). Systematic identification of conserved metabolites in GC/MS data for metabolomics and biomarker discovery. *Analytical Chemistry 79*, 966–973.

Su, Z., Hong, H., Tong, W., Perkins, R., Shao, X., & Cai, W. (2007). Identification of differently expressed genes using disjoint principal component analysis coupled with genetic algorithm. *Chemical Journal of Chinese Universities 28*(9), 1640–1644.

Sweetman, L., & Williams, J. (2001). Branched chain organic acidurias. In S. Scriver, A. Beaudet, W. Sly, & D. Valle (Eds.), *The metabolic and molecular basis of inherited disease* (8th ed., pp. 2125–2164). New York, NY: McGraw-Hill.

Tanaka, K., Orr, J., & Isselbacher, K. (1968). The identification of 3-hydroxyisovaleric acid in the urine of a patient with isovaleric acidemia. *Biochimica et Biophysica Acta 152*(152), 638–641.

Truscott, R., Malegan, D., McCairns, E., Burke, D., Hick, L., Sims, P., Halpern, B., Tanaka, K., Sweetman, L., Nyhan, W., Hammond, J., Bumack, C., Haan, E., & Danks, D. (1981). New metabolites in isovaleric acidemia.*Clinica Chimica Acta 110*, 187–203.

Van den Berg, R., Hoefsloot, H., Westerhuis, J., Smilde, A., & Van der Werf, M. (2006). Centering, scaling, and transformations: Improving the biological information content of metabolomics data. *BMC Genomics 7*, 141–157.

Van den Berg, R., Rubingh, C., Westerhuis, J., Van der Werf, M., & Smilde, A. (2009). Metabolomics data exploration guided by prior knowledge. *Analytica Chimica Acta 651*, 173–181.

Weckwerth, W., & Morgenthal, K. (2005). Metabolomics: From pattern recognition to biological interpretation.*Drug Discovery Today 10*, 1551–1558.

Wold, S. (1976). Pattern recognition by means of disjoint principal component models. *Pattern Recognition 8*, 127–139.

Wold, S., Kettaneh, N., & Tjessem, K. (1996). Hierarchical multiblock PLS and PC models for easier model interpretation and as an alternative to variable selection. *Journal of Chemometrics 10*, 463–482.