Multivariate paired data analysis: multilevel PLSDA versus OPLSDA

Westerhuis, J.A.; van Velzen, E.J.J.; Hoefsloot, H.C.J.; Smilde, A.K.

Published in:
Metabolomics

DOI:
10.1007/s11306-009-0185-z

Metabolomics (2010) 6:119–128
DOI 10.1007/s11306-009-0185-z

ORIGINAL ARTICLE

Multivariate paired data analysis: multilevel PLSDA versus OPLSDA

Johan A. Westerhuis · Ewoud J. J. van Velzen · Huub C. J. Hoefsloot · Age K. Smilde

Received: 27 May 2009 / Accepted: 13 October 2009 / Published online: 28 October 2009

Abstract Metabolomics data obtained from (human) nutritional intervention studies can have a rather complex structure that depends on the underlying experimental design. In this paper we discuss the complex structure in data caused by a cross-over designed experiment. In such a design, each subject in the study population acts as his or her own control and makes the data paired. For a single univariate response a paired *t*-test or repeated measures ANOVA can be used to test the differences between the paired observations. The same principle holds for multivariate data. In the current paper we compare a method that exploits the paired data structure in cross-over multivariate data (multilevel PLSDA) with a method that is often used by default but that ignores the paired structure (OPLSDA). The results from both methods have been evaluated in a small simulated example as well as in a genuine data set from a cross-over designed nutritional metabolomics study. It is shown that exploiting the paired data structure underlying the cross-over design considerably improves the power and the interpretability of the multivariate solution. Furthermore, the multilevel approach provides complementary information about (I) the diversity and abundance of the treatment effects within the different (subsets of) subjects across the study population, and (II) the intrinsic differences between these study subjects.

Keywords Paired data · Multilevel analysis · PLSDA · OPLSDA · Metabolomics

J. A. Westerhuis and E. J. J. van Velzen equally contributed to this work.

J. A. Westerhuis (✉) · E. J. J. van Velzen · H. C. J. Hoefsloot · A. K. Smilde
Biosystems Data Analysis, Swammerdam Institute for Life Sciences, Universiteit van Amsterdam, Amsterdam, The Netherlands
e-mail: j.a.westerhuis@uva.nl

E. J. J. van Velzen
Unilever Research and Development, Vlaardingen, The Netherlands

J. A. Westerhuis
Netherlands Metabolomics Centre, Leiden, The Netherlands

1 Introduction

Metabolomics data from human studies are often characterised by large variations between the subjects. This is different from most animal studies where metabolic variation between the test animals is usually less abundant. A global overview with respect to nutritional metabolomics is provided by Rezzi et al. (2007).

The large variation between human subjects can give rise to two problems in the analysis. The first is that small and subtle treatment effects (e.g. dietary responses) can easily be overlooked, especially when the effect is smaller than the intrinsic variation between the subjects. The second problem is that the response and the impact of the treatment effect may differ between the subjects. This implies that an average treatment effect may not be the most relevant in studies where subsets of subjects respond differently upon a dietary intervention. An often used solution in clinical or nutritional studies is the use of a cross-over design. In a cross-over study all subjects acts as their own control. As a result, multivariate data obtained from a cross-over designed experiment has a paired data structure.

When a cross-over design is used in the study, the treatment effect for each subject can be separated from the

between subject variation. After separating these confounded sources of variation, both can be analyzed separately. The analysis of paired data is usually performed with a paired *t*-test and repeated measures ANOVA in case of univariate responses. Depending on the ratio between the effect size and the variation between the subjects, a paired *t*-test is advantageous over a normal *t*-test due to its increased statistical power. Figure 1 illustrates the principle of both *t*-tests by means of 5 subjects that have been measured in the control period (A) and in the treatment period (B). From these subjects a univariate response was acquired. The columns A and B in Table 1 show the measurement responses collected in the control period and the treatment period respectively. The values in column D represent the differences between A and B, whereas M represents the mean of A and B.

In Fig. 1 the difference between the unpaired analysis and the paired analysis is demonstrated. In Fig. 1b, the paired data structure is accentuated by the connection lines between the two measurements. Without considering the paired structure, a normal unpaired *t*-test does not show a statistically significant difference between A and B. The average of B minus the average of A equals 3 ± 4.61 ($P = 0.172$).

When the paired data structure is used in a paired *t*-test, then the difference D between control group A and treatment group B becomes statistically significant. In that case the difference is 3 ± 1.57 ($P = 0.009$).

Table 1 Univariate example with paired data

Subject	A	B	D	M
1	2	5	3	3.5
2	4	9	5	6.5
3	6	7	1	6.5
4	8	11	3	9.5
5	10	13	3	11.5

The measurement responses of 5 subjects are shown that were collected in the control period (A) and treatment period (B). Columns D and M represent the difference and the mean of A and B for each subject

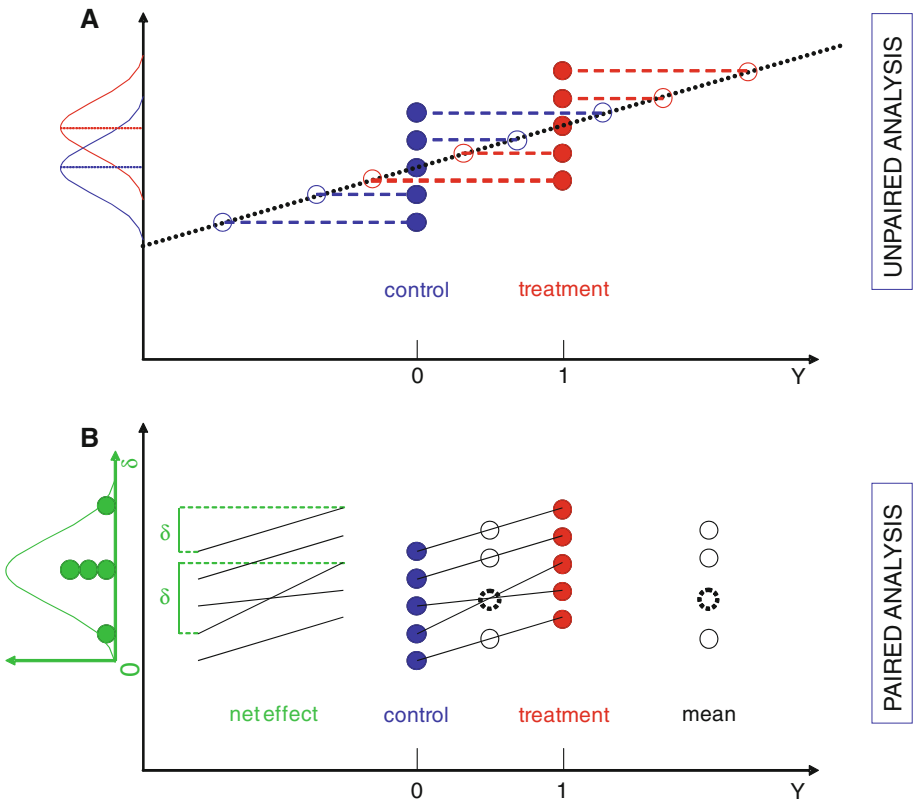

Fig. 1 Basic principles of **a** unpaired and **b** paired data analysis of measurement responses from 5 subjects in the control period (class 0) and in the treatment period (class 1) respectively. In (**a**) no consideration is given to the paired data structure. The effect of the treatment is represented by the dotted line. As shown by the projections of the observations on the dotted line, the discriminant model is not able to separate the intervention classes well. On the Y-axis the overlapping distributions of the two intervention classes are projected. If (**b**) a paired analysis is used instead (illustrated by the lines connecting the 5 data pairs), the data is separated into a mean (black circles) and a difference (δ) per subject. The differences (net treatment effect) are projected on the Y-axis per subject, and are all different from 0. The dotted circle represents two similar mean values. (Color figure online)

By using the paired structure of the data from cross-over studies, statistically significance can be obtained for much smaller treatment effects. However, besides the advantage in power, it is also possible to examine the difference in treatment effect *within* the subjects in a much easier way than without the separation of the different sources of variation. At the same time the differences *between* the subjects can be studied without being confounded with the treatment effect. In the univariate example the treatment effect is not similar for all subjects. This variation simulates the intrinsic differences between subjects.

Nowadays, cross-over designs are also used in combination with 'omics' techniques, resulting in paired multivariate data sets (Bertram et al. 2006; Pohjanen et al. 2007; van Velzen et al. 2008). Time series experiments with different subjects have the same paired data structure (Jansen et al. 2005; Rantalainen et al. 2008), and the analysis of such data sets can also be improved when exploiting the design underlying the study. However, in the analysis of these multivariate paired data, the study design is not always considered. Instead of using a multivariate extension of the paired *t*-test, in general other methods are being applied that particularly focus upon the mean effects over all subjects.

In this paper we will discuss the multivariate extension of the paired *t*-test, which is recently introduced as multilevel data analysis (van Velzen et al. 2008). We will demonstrate the additional benefit of multilevel data analysis in the analysis of multivariate paired (cross-over) data in comparison with a method that does not explicitly consider the cross-over design (OPLSDA). We will examine the differences in a tutorial style by using a small simulation study as well as a genuine cross-over designed (nutritional) study. Both these studies are analysed with OPLSDA and multilevel PLSDA. The results obtained from both analyses are evaluated, compared and discussed. To introduce this methodological evaluation, we will first provide a brief description of OPLSDA and multilevel PLSDA, and in which way these methods deal with paired data.

2 Theory

2.1 OPLSDA

OPLSDA was introduced as an improvement of the PLSDA method to discriminate two or more groups (classes) using multivariate data (Bylesjo et al. 2006; Trygg and Wold 2002). In OPLSDA a regression model is calculated between the multivariate data and a response variable that only contains class information. The advantage of OPLSDA compared to PLSDA is that a single component is used as a predictor for the class, while the other components describe the variation orthogonal to the first predictive component. Wiklund et al. (2008) used the terms *between treatment variation* to describe the average effect of treatment and *within treatment variation* to describe the systematic remainder variation which is not related to the treatment. The treatment effect is considered equal for all subjects although the magnitude is allowed to be different for each subject. Treatment effects that differ from the average treatment effect are referred to as within treatment variation.

The predictive OPLSDA component actually describes the direction of the difference (the treatment effect) between the average of class A and the average of class B according the representation given in Fig. 1a (dotted line). Then all samples are projected on this component to estimate the predictive scores. Although a group-average effect is observed in this example, the projections on the line clearly shows that the classes are not well separated. Furthermore, in OPLSDA only a single predictive component is calculated (in case of a two-class problem). When the treatment effect manifest differently among the subjects in the test population, this will not be observed by the OPLSDA method.

2.2 Multilevel PLSDA

Multilevel PLSDA is another discrimination method that was recently introduced to develop classifications models of multivariate data from cross-over designed studies, i.e. an experimental setup in which each subject underwent a control measurement and a treatment (in a random order) (van Velzen et al. 2008). Multilevel PLSDA can be considered as a multivariate extension of a paired *t*-test. Multilevel data analysis can only be used when the data has a multilevel structure, whereas OPLSDA can be used for any discrimination problem.

In a multilevel PLSDA model, the variation *between* subjects (within treatment variation in OPLSDA) and the variation *within* subjects (total variation due to the treatment) are separated. The within subject variation in multilevel PLSDA is not considered the same for each subject as compared to the between treatment variation in OPLSDA. The between subject variation in multilevel data analysis is performed on the average of the two observations (balck circles in Fig. 1b), whereas the within subject variation is performed on the net differences between the paired observations (δ in Fig. 1b).

The initial step in multilevel PLSDA is to separate the between subject variation from the within subject variation. First, the observations in the control ($\mathbf{A}$) and the treatment ($\mathbf{B}$) periods are concatenated: $\begin{bmatrix}\mathbf{A}\\ \mathbf{B}\end{bmatrix}$. The between subject variation has structure $\begin{bmatrix}\mathbf{M}\\ \mathbf{M}\end{bmatrix}$, where $\mathbf{M} = \frac{1}{2}[\mathbf{A} + \mathbf{B}]$. The

within subject variation is calculated according: $2\left(\begin{bmatrix}\mathbf{A}\\ \mathbf{B}\end{bmatrix} - \begin{bmatrix}\mathbf{M}\\ \mathbf{M}\end{bmatrix}\right)$. Its structure is $\begin{bmatrix}-\mathbf{D}\\ \mathbf{D}\end{bmatrix}$, where $\mathbf{D} = [\mathbf{B} - \mathbf{A}]$. In this two-class problem this just comes down to an analysis on the differences between the data obtained in the two classes. The rank of the within subject variation matrix is usually larger than one, because the effect of the treatment is generally different between the subjects. In that case, more than a single component is needed to provide a good description of the within subject variation in the study population.

Because of its structure, analysis of the within subject variation can be done with several multivariate methods e.g. PCA, PLSDA or even OPLSDA. The within subject variation contains both variation that is equal for all subjects as well as variation that is different between subjects. When MLPLSDA is used to describe the within subject variation, the focus is on the similarity in the treatment effect between the subjects. Therefore the first MLPLSDA component primarily describes the main, corresponding effect, whereas the latter components particularly reflect the differences in treatment effect among the subjects. When MLPCA is used, the focus is simply on the major variation in the within treatment variation.

2.2.1 *Why does this work*

Consider the data measured in a study where I ($i = 1 \ldots I$) individuals are measured at D ($d = 1 \ldots D$) occasions. Then each measurement $\mathbf{x}_{di}$ can be explained partly by a grand mean μ, the group effect $\boldsymbol{\alpha}_d$, while the remainder is an unexplained residual.

$$\mathbf{x}_{di} = \mu + \boldsymbol{\alpha}_d + \mathbf{e}_{di} \quad (1)$$

Here the group effect estimate $\boldsymbol{\alpha}_d$ equals the mean of all $\mathbf{x}_{di}$ averaged over all I individuals. In a one-way ANOVA approach the Mean Square of $\mathbf{x}_{di}$ would be related to the Mean Square of $\mathbf{e}_{di}$ leading to an F-value with $D - 1$ and $D(I - 1)$ degrees of freedom. However in this approach we ignore that besides the fixed effect due to the classes there is also a random effect due to the individual. For a new individual we cannot predict the effect in advance, but we can model it when the data for the new individual is obtained (Sokal and Rohlf 1998). Thus in the case of a cross-over design where the same individuals are measured at D occasions Eq. 1 can be extended with the random individual effect

$$\mathbf{x}_{di} = \mu + \boldsymbol{\alpha}_d + \boldsymbol{\beta}_i + \mathbf{f}_{di} \quad (2)$$

$\boldsymbol{\beta}_i$ is estimated as the mean of all D values for individual i. Note that $\boldsymbol{\alpha}_d$ does not change when the individual effect is included in the model since $\mathbf{e}_{di} = \boldsymbol{\beta}_i + \mathbf{f}_{di}$, thus the random effect is a part of the variation that was first collected in the residual $\mathbf{e}_{di}$. This means that the new residual $\mathbf{f}_{di}$ is smaller, and thus the estimated effect of MS $\boldsymbol{\alpha}_d$ over MS $\mathbf{f}_{di}$ will be larger than the previous ANOVA estimate. Thus the paired data analysis will have a higher power.

Note that $\mathbf{x}_{di} - \boldsymbol{\beta}_i = \mu + \boldsymbol{\alpha}_d + \mathbf{f}_{di}$, i.e. the original data minus the mean over all treatments equals exactly $\left(\begin{bmatrix}\mathbf{A}\\ \mathbf{B}\end{bmatrix} - \begin{bmatrix}\mathbf{M}\\ \mathbf{M}\end{bmatrix}\right)$ for the case when $D = 2$, where $\mathbf{M}$ contains the means $\boldsymbol{\beta}_i$ for all individuals. For a three class problem this would lead to $\left(\begin{bmatrix}\mathbf{A}\\ \mathbf{B}\\ \mathbf{C}\end{bmatrix} - \begin{bmatrix}\mathbf{M}\\ \mathbf{M}\\ \mathbf{M}\end{bmatrix}\right)$.

In the multilevel PLSDA context $\left(\begin{bmatrix}\mathbf{A}\\ \mathbf{B}\\ \mathbf{C}\end{bmatrix} - \begin{bmatrix}\mathbf{M}\\ \mathbf{M}\\ \mathbf{M}\end{bmatrix}\right)$ as $\mathbf{X}$ data is then related to a dummy $\mathbf{Y}$-matrix $\left(\begin{bmatrix}\mathbf{1} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{1} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{1}\end{bmatrix}\right)$, where $\mathbf{1}$ and $\mathbf{0}$ represent vectors of I ones and zeros respectively. Note however that a PLSDA model with more than 2 classes is not straightforward (Indahl et al. 2007; Barker and Rayens 2003; Nocairi et al. 2005).

The ANOVA model described here suits the study design described in the simulation study as well as in the real data example. For different type of studies other ANOVA models apply, but the extension to multivariate multilevel classification models is similar to the situation discussed above.

2.3 Data pretreatment

In multilevel analysis all sources of variation (in this case the between individual variation as well as the between individual variation) are of interest. However, the information that can be obtained from these data may be different. Therefore the type of scaling used can be adjusted for each subset of variation. This approach of scaling after variation splitting is considered as an important benefit of the multilevel approach, as the preferred scaling technique can explicitly be adapted to the part of the data that is examined and the data analysis technique used.

2.4 Score plots

In our previous work on the assessment of PLSDA validation (Westerhuis et al. 2008a) we concluded that score plots should not be used for assessing and interpreting the class separation since the PLSDA model may highly overfit the data. This problem is especially related to multivariate or high dimensional data where the number of variables is much higher than the number of samples. A possible solution to this problem is the use of cross-validated scores in a score plot (Wiklund et al. 2008). However, a problem

with this approach is that the cross-validated scores are all based on different loadings. For that reason the scores cannot be drawn in a similar figure. Only if the differences between the various loadings obtained from the different models in the cross-validation are small, then a composite figure with cross-validated scores may be useful. In the current work we will use double cross-validated scores (Smit et al. 2007; Westerhuis et al. 2008a) to evaluate the difference between the OPLSDA model and the multilevel PLSDA model. We will address the aspects of class separation and the score distribution in relation to the different sources of variation.

3 Analysis of simulated data

3.1 A small simulated example

The properties of paired data and the consecutive data analysis will be explained in a brief example using a small, simulated data. Let's consider the measurement responses of three variables A, B and C (e.g. metabolite concentrations) in 10 subjects, which were collected in the control period as well as in the treatment period (Table 2).

In this example variable B increases +2 for all subjects after the treatment. Variable A increased +1 for the odd subjects (males) and +3 for the even subjects (females). Variable C did not change. The effect of the treatment is clearly visible for variables A and B (see columns D_A and D_B) and not for C (see column D_C). A small fraction of random normally distributed noise was added to the data before analysis.

3.2 OPLSDA analysis of simulated example

To perform OPLSDA analysis a **y** vector was constituted which include the class information for each subject. This column vector contains the class label value −1 for the control group, and the class label value +1 for the treatment group. An OPLSDA model (with 3 components) between the original data [**ABC**] and the **y** vector was calculated. It should be noted that the estimated scores changed upon the number of OPLSDA components calculated. These changes, however, were small and did not influence the conclusions derived from this simulated experiment.

In Fig. 2 the predictive scores $\mathbf{t_P}$ versus the orthogonal scores $\mathbf{t_O}$ are illustrated. The score plot shows that the control group (red squares) is not well separated from the treatment group (blue circles). Although a clear and systematic difference was simulated across the intervention periods, the OPLSDA model was not able to detect this treatment effect. The problem is that methods that do not

Table 2 Simulated measurement responses of three variables A, B and C in 10 subjects collected in the control period and the treatment period

Occasion	Subject	A	B	C	D_A	D_B	D_C	M_A	M_B	M_C
Control period	1	20	10	20	−1	−2	0	20.5	11	20
	2	18	12	17	−3	−2	0	19.5	13	17
	3	16	15	14	−1	−2	0	16.5	16	14
	4	14	16	11	−3	−2	0	15.5	17	11
	5	10	2	8	−1	−2	0	10.5	3	8
	6	9	3	5	−3	−2	0	10.5	4	5
	7	7	7	2	−1	−2	0	7.5	8	2
	8	7	7	8	−3	−2	0	8.5	8	8
	9	3	9	14	−1	−2	0	3.5	10	14
	10	2	9	17	−3	−2	0	3.5	10	17
Treatment period	1	21	12	20	+1	+2	0	20.5	11	20
	2	21	14	17	+3	+2	0	19.5	13	17
	3	17	17	14	+1	+2	0	16.5	16	14
	4	17	18	11	+3	+2	0	15.5	17	11
	5	11	4	8	+1	+2	0	10.5	3	8
	6	12	5	5	+3	+2	0	10.5	4	5
	7	8	9	2	+1	+2	0	7.5	8	2
	8	10	9	8	+3	+2	0	8.5	8	8
	9	4	11	14	+1	+2	0	3.5	10	14
	10	5	11	17	+3	+2	0	3.5	10	17

The difference (D) and mean (M) for each subject and for each variable are given

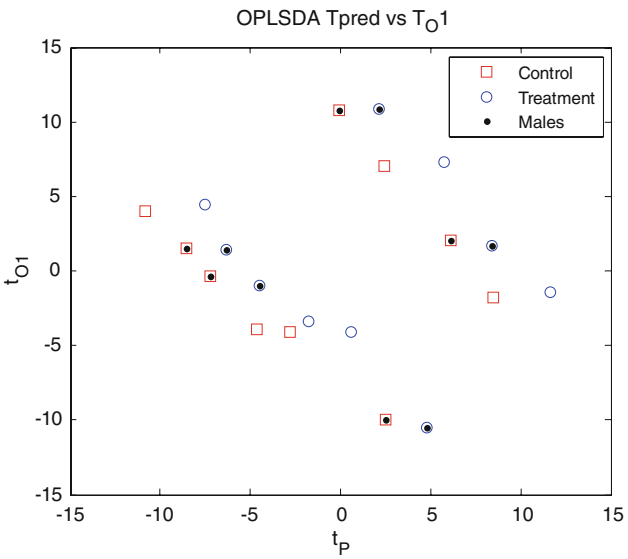

Fig. 2 Double cross validated OPLSDA score plot of simulated data. The predicted scores $\mathbf{t_P}$ versus the orthogonal scores $\mathbf{t_O}$**1** of the control group (*red squares*) and the treatment group (*blue circles*) are shown. The black dots highlights the male subjects in the study population. No class separation between the control and treatment groups as well as between the males and the females could be observed. (Color figure online)

use the paired data structure, only focus on the difference between the class means and the ranges of the classes. When the between class difference is small compared to the range of the observed responses, this difference drowns in the total variation and will not be detected. The loadings of the OPLSDA model $[\mathbf{0.73} \quad \mathbf{0.72} \quad \mathbf{0.01}]$ indicate that particularly the first two variables are important for the predictive component. Nevertheless, the obtained model does not allow a good discrimination between the two classes.

Another limitation that appears when ignoring the paired data structure is that no emphasis can be given to the variation in effect between the subjects. This is illustrated in the simulated example (Table 2) where the male subjects (odd sample numbers) have a relative small increase in variable A as compared to the females (even sample numbers). Methods that do not use the paired data structure only focuses on the difference between the average value of the controls and the treated samples and do not consider a variation in the treatment effect. The score distribution of the males and the females in Fig. 2 clearly demonstrates that the OPLSDA model is not able to discriminate between these systematic, gender-related, response differences.

Note that in Fig. 2, the effect of the treatment is clearly visible on the predictive score (t_P) however it drowns in the large variation caused by the different individuals. Thus OPLSDA estimates the loadings well, but the effect is not statistically significant

3.3 Multilevel PLSDA analysis of the within subject variation of the simulated data

The multilevel approach takes the paired data structure into account. The total variation in the data is divided into between subject variation and within subject variation. The within subject variation exclusively describes the net differences in each of the measured variables for each subject, i.e. $[-\mathbf{1}, -\mathbf{3}, -\mathbf{1}, -\mathbf{3}, \ldots]$ for variable A, $[-\mathbf{2}, -\mathbf{2}, -\mathbf{2}, -\mathbf{2}, \ldots]$ for variable B and $[\mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \ldots]$ for variable C. The large variation between the subjects is completely ignored when the within subject variation is analyzed. Furthermore, the gender-related difference in effect that manifest in variable A remain clearly present in the within variation.

In the estimation of the between subject variation, the mean observations are used, i.e. $[\mathbf{20.5}, \mathbf{19.5}, \mathbf{16.5}, \mathbf{15.5}, \ldots]$ for variable A, $[\mathbf{11}, \mathbf{13}, \mathbf{16}, \mathbf{17}, \ldots]$ for variable B and $[\mathbf{20}, \mathbf{17}, \mathbf{14}, \mathbf{11}, \ldots]$ for variable C. After variation splitting, this source of variation can be analyzed without being confounded with the treatment-related variation.

In Fig. 3 the multilevel PLSDA scores $\mathbf{t_W1}$ and $\mathbf{t_W2}$ of the within subject variation are shown. A clear separation is observed between the control group and the treatment group. Since multilevel PLSDA particularly focuses on the within subject variation that is similar among the subjects,

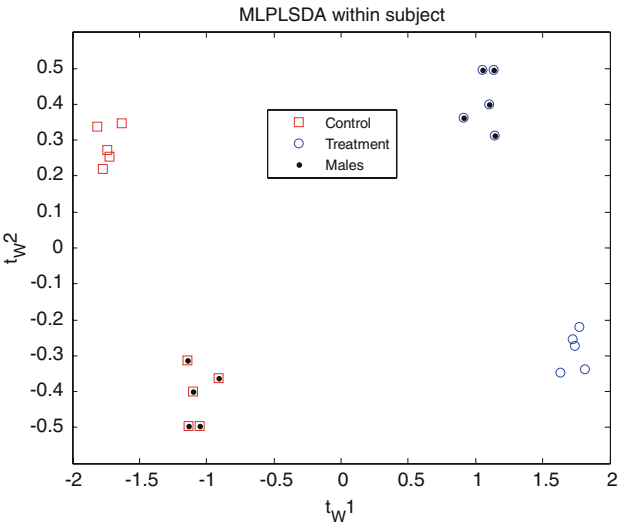

Fig. 3 The multilevel PLSDA scores ($\mathbf{t_W1}$, $\mathbf{t_W2}$) of the within subject variation in simulated data on the first two components. The red squares represents the subjects in the control group. The blue circles represents the same subjects after the treatment. The males (*black dots*) experienced a smaller increase in variable A relative to the females. (Color figure online)

the first component mainly describes the difference between the classes. The second component on the other hand describes the within subject variation that is different between the subjects. We therefore observe a notable separation between the males and females in the test population.

3.4 Multilevel PCA analysis of the between subject variation of the simulated data

Besides the within subject variation, also the between subject variation was examined. Note that the between subject data only consists of 10 subjects, and that variation due to the treatment has been removed. To investigate the main variation between the subjects, a multilevel PCA analysis was performed. Figure 4 shows the score plot of the between subject variation for the first two principal components. Three groups of subjects appear, i.e., (I) subjects 1–4, (II) subjects 5–8 and (III) subjects 9 and 10. These groups are primarily associated with the absolute abundances of the variables.

4 Analysis of experimental data

4.1 Study design

The intervention study has a double-blind, placebo-controlled cross-over design with a single, oral intervention of black tea solids. The black tea solids contained 800 mg

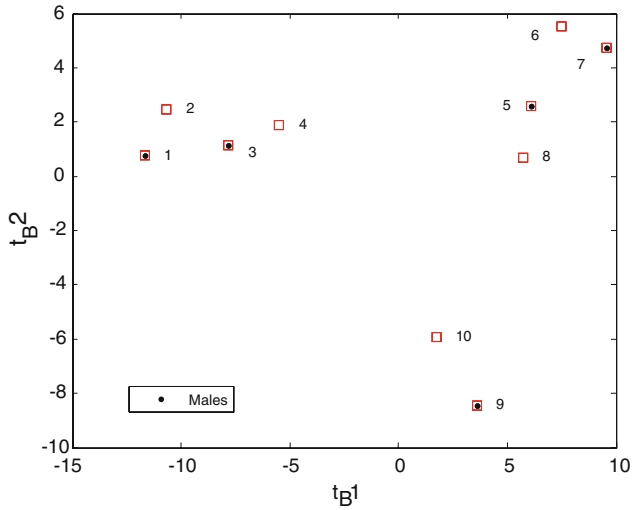

Fig. 4 Score plot of multilevel PCA model representing the between subject variation on the first two principal components. Three main clusters could be identified. The separation of the score clusters is related to the absolute abundances of the variables

polyphenols based on gallic acid equivalents and were administered as non-transparent capsules. In total, 20 male subjects participated in the study (18–40 years of age and Body Mass Index between 19 and 29 kg/m^2). Urine samples were collected (and weighted) from all subjects over a time span of 48 h after the intervention.

4.2 Sample pre-preparation

To investigate the main effects in the data, pooled 48 h urines were prepared. This was done by adding aliquots of the urines together in exactly the same mass-ratio as the collected fractions. Then, to 450 µl of each pooled urine sample 200 µl phosphate buffer solution (0.6 M Na_2HPO_4/NaH_2PO_4, pH 6.5) and 50 µl deuterium oxide (D_2O) was added. The phosphate buffer solution contained 0.05 mg ml^{-1} 3-(trimethylsilyl)propionic acid-d_4 sodium salt (TSP) as an internal standard. After homogenization and centrifugation 650 µl of the clear supernatant was transferred into a 5-mm NMR tube.

4.3 Data acquisition and data pre-processing

600 MHz ^{1}H NMR spectra were acquired at 300 K on a Bruker Avance 600 MHz NMR spectrometer. The data were collected into 64 K points (128 scans) using a spectral width of 9000 Hz, an acquisition time of 3 s and a relaxation delay of 3 s (with suppression of the water signal). The Fourier transformed NMR spectra were manually phase- and baseline corrected, calibrated and normalized against the methyl resonance of TSP at δ 0.0 ppm. Finally, the intensities of the NMR signals were expressed in molar equivalents by multiplying the TSP normalized spectra with the total volumes of the 48 h urines. The resulting NMR spectra were then subdivided in discrete regions ('buckets') of equal width (δ 0.00225 ppm). As bucketing could not completely compensate for line broadening effects and positional shifts (due to differences in pH, ion strength, etc.), also Correlation Optimized Warping (COW) (Skov and Bro 2008; Wu et al. 2006) was applied on the bucketed data. A detailed description of the study and the analytical procedure was recently reported by the authors (van Velzen et al. 2009).

4.4 Data analysis

Based on our previous findings with the same dataset, we only consider the aromatic region of the NMR spectrum (δ 6–9 ppm) in the data analysis. Calculations involving data pre-treatment (bucketing, normalization, volume correction), multilevel PLSDA, OPLSDA, double cross validation (2CV), permutation testing, and Discriminant Q^2 (DQ^2) (Westerhuis et al. 2008b) estimations were performed using Matlab (version 2009a, The MathWorks, USA) with in-house written Matlab routines. These routines (together with a tutorial) are available at http://www.bdagroup.nl/.

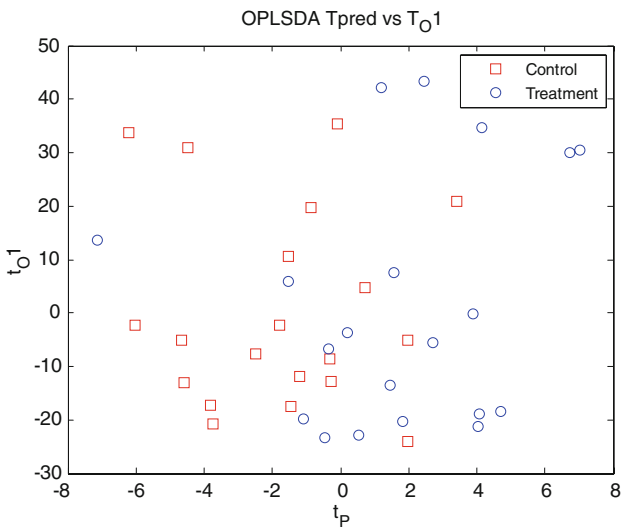

Fig. 5 Double cross-validated OPLSDA scores representing the autoscaled urinary NMR spectra of 20 male subjects. The observations obtained in the control period (*red squares*) and the treatment period (*blue circles*) have the tendency to form two separate classes. Note that there is only one predictive score and one orthogonal score. The predictive score indicates the separation between the two intervention groups. Different from Wiklund et al. (2008) the cross validated scores give different values for t_P as well as for t_O1 compared to the non-crossvalidated scores. (Color figure online)

4.5 OPLSDA analysis of the experimental data

Before analysis, autoscaling of the data was used to improve the weight of the smaller intensities. The (double) cross-validated scores in Fig. 5 show that the OPLSDA model was able to provide a reasonable class estimation of the autoscaled data. The DQ^2 of the classification model was 0.25 (based on the average result of 20 2CV runs), which was higher than all 2500 DQ^2 values obtained from models from permuted data (Lindgren et al. 1996; Westerhuis et al. 2008a). Even though the distribution of cross-validated scores of the subjects in the control group is different from the treatment group, still we can observe a large dispersion across the intervention groups in the direction of predictive score component (t_P).

4.6 Multilevel analysis of the experimental data

In the first step of the multilevel analysis the variation is separated into the between individuals contribution and the within individuals contribution (the treatment effect). The latter is autoscaled before analysis give extra weight to the smaller intensities. The between individual variation is not scaled because we were interested in the larger effects in this data.

When the paired data structure is used in the multilevel PLSDA analysis of the experimental data, a systematic difference revealed between the control group and the treatment group. This is demonstrated in Fig. 6a, d where the double cross-validated scores reflects the within subject variation in the 48 h urine samples of all subjects between the intervention periods. The DQ^2 of the multilevel model was 0.54 (the average result of 20 2CV runs), which was statistically significant in a permutation test. Again 2500 permutations were performed. The DQ^2 was also higher than the DQ^2 obtained in the previously described OPLSDA analysis.

In Fig. 6b the main treatment effect was observed along the first component. Some individuals with high scores on the first component are indicated with colored markers and

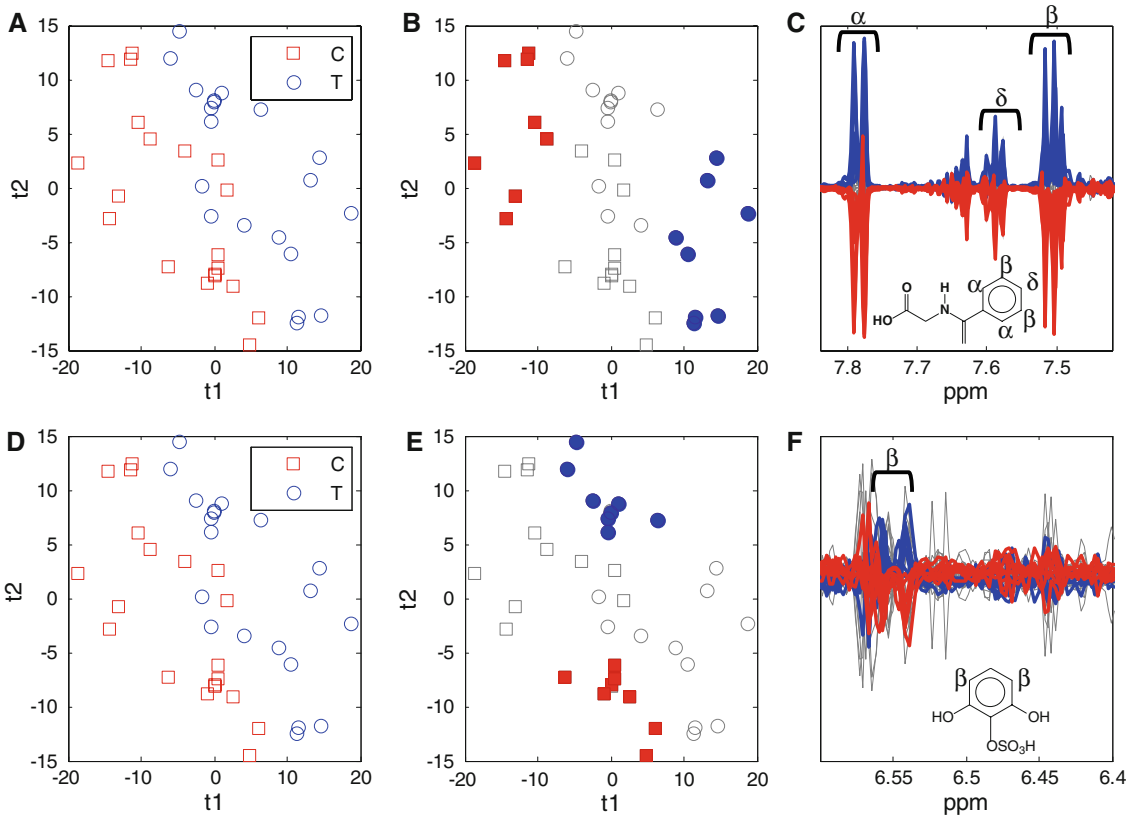

Fig. 6 Multilevel PLSDA double cross-validated scores which represent the urinary NMR spectra of 20 subjects after black tea intake. The (**a** and **d**) scores on the first two components (t_1, t_2) reflects the within subject variation in the control period (*red squares*) and the treatment period (*blue circles*). Two different treatment effects could be identified. The (**b**) first effect along the first component point towards (**c**) increasing hippuric acids levels and increasing 1,3-dihydroxyphenyl-2-*O*-sulphate levels. The (**e**) second effect along the second component is basically described by (**f**) 1,3-dihydroxyphenyl-2-*O*-sulphate, whereas the increase of hippuric acid is less pronounced. (Color figure online)

their corresponding spectra are shown in Fig. 6c. The first component is largely influenced by elevated levels of hippuric acid (Fig. 6c), 1,3-dihydroxyphenyl-2-*O*-sulphate and some other aromatic signals. These observations match our previous findings on the same data. However, this main effect was not equal for all subjects. As shown in Fig. 6e, another subset of individuals responds differently upon the black tea intervention. In this subpopulation are individuals that score low on the first component as their increase of the urinary hippuric acid levels was less pronounced. However this subpopulation score much higher on the second multilevel PLS component which is dominated by the 1,3-dihydroxyphenyl-2-*O*-sulphate levels (Fig. 6f). The current observations lead to the conclusion that not all subjects respond equally to the treatment, and demonstrate that a single component for such a classification model is not sufficient to assess the different treatment effects within a study population.

Besides the within variation, also the between subject variation was explored. The combination of both (multilevel) analyses will then allow a comprehensive interpretation of all major variation sources in the data. Similarly to the simulated example, a multilevel PCA analysis was appropriate to assess the main intrinsic variations between the subjects (on mean-centred data). As shown in Fig. 7a, the scores of 4 subjects on the second principal component t_B2 appear to be different from the other subjects. Whereas the first principal component (Fig. 7b, black profile) is a generic representation of all NMR signal intensities, different variations among the NMR resonances were observed on the second principal component (Fig. 7b, red profile). The loadings show that the variation between the subjects particularly depends on the ratio between the NMR signals of hippuric acid (δ 7.78 ppm, d; δ 7.59 ppm, t and δ 7.50 ppm, t) and the NMR signals of an unknown aromatic compound, U (δ 7.17 ppm, s; δ 7.24 ppm, s and δ 7.31 ppm, s). This unknown compound was observed in a spectral region were several other resonance patterns of aromatic amino acids, (conjugated) polyphenolic acids, (indole) alkaloids etc. come together. For now this complicates a straightforward identification of component U. Four subjects appear to have a higher signal ratio between U and hippuric acid then the other subjects in the study population.

5 Conclusion

In this paper we have shown that when the paired data structure of metabolomics data obtained from a cross-over designed experiment, is taken into account during the multivariate data analysis, the power and the interpretability of the results greatly improves. Furthermore the multilevel approach provides information about the

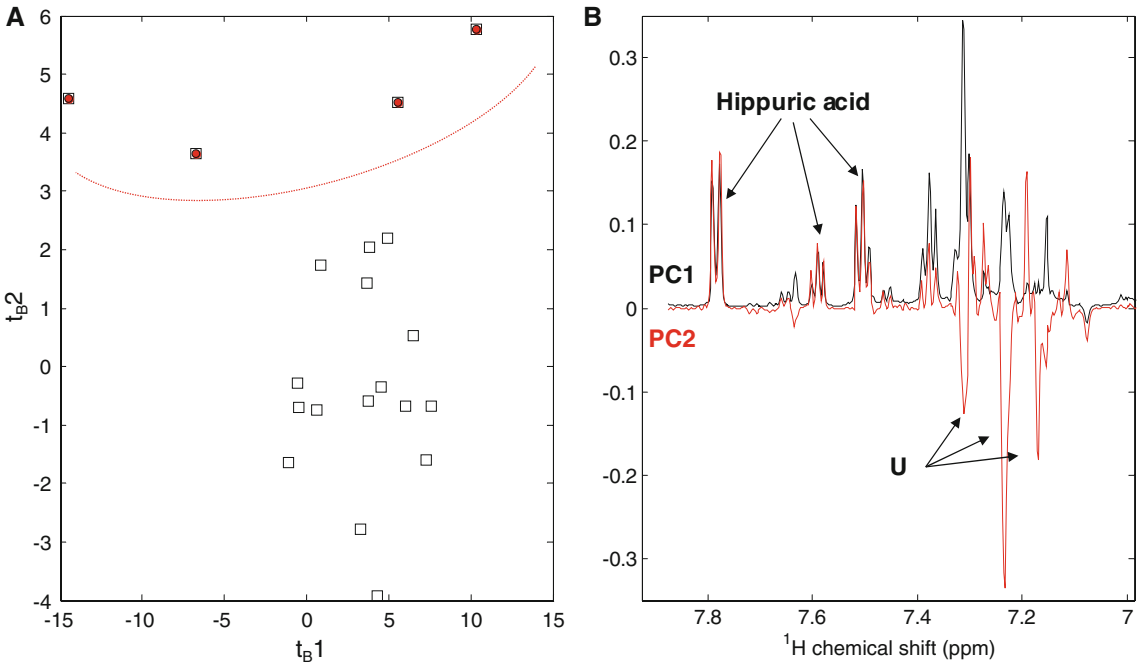

Fig. 7 Variation between the mean-centered 24 h urinary NMR profiles of 20 subjects as represented by (**a**) the t_B1 and t_B2 scores in the multilevel PCA score plot. The (**b**) associated loadings reflect the intensity depended variation along the spectral axis (PC1) and variations between hippuric acid and component U (PC2). The ratio between hippuric acid and U is different in four subjects (*red dots*). (Color figure online)

diversity and abundance of the treatment effects across the study population. Finally, the multilevel analysis allows investigation of the between subject variation which is completely separated from the within subject variation. However, often this paired data structure is ignored during the analysis and default methods that do not consider the paired data structure are used to analyze the data, leading to suboptimal results. In this paper we have discussed the difference between the paired analysis and the non-paired analysis approaches and used a simulated example as well as a real experiment in which a human test panel was given black tea solids. In the latter study we observed two subsets in the human test population that responded differently upon the intake of black tea solids. These subsets show differences in urinary excretion of hippuric acid and 1,3-dihydroxyphenyl-2-*O*-sulphate. We also observed intrinsic differences between the subjects. These variations are mainly described by the relative levels (or molar ratio) of hippuric acid and an unknown aromatic component (U).

Acknowledgements We are grateful to the European Commission for their financial support of the GutSystem project (MTKI-CT-2006-042786) under the framework 6 Marie-Curie Transfer of Knowledge Industry-Academia Strategic Partnership scheme. This project was carried out within the research programme of the Netherlands Metabolomics Centre (NMC) which is part of the Netherlands Genomics Initiative/Netherlands Organization for Scientific Research.

References

Barker, M., & Rayens, W. (2003). Partial least squares for discrimination. *Journal of Chemometrics, 17*(3), 166–173.

Bertram, H. C., Knudsen, K. E. B., Serena, A., et al. (2006). NMR-based metabonomic studies reveal changes in the biochemical profile of plasma and urine from pigs fed high-fibre rye bread. *British Journal of Nutrition, 95*(5), 955–962.

Bylesjo, M., Rantalainen, M., Cloarec, O., et al. (2006). OPLS discriminant analysis: Combining the strengths of PLS-DA and SIMCA classification. *Journal of Chemometrics, 20*(8–10), 341–351.

Indahl, U. G., Martens, H., & Naes, T. (2007). From dummy regression to prior probabilities in PLS-DA. *Journal of Chemometrics, 21*(12), 529–536.

Jansen, J. J., Hoefsloot, H. C. J., van der Greef, J., Timmerman, M. E., & Smilde, A. K. (2005). Multilevel component analysis of time-resolved metabolic fingerprinting data. *Analytica Chimica Acta, 530*(2), 173–183.

Lindgren, F., Hansen, B., Karcher, W., Sjostrom, M., & Eriksson, L. (1996). Model validation by permutation tests: Applications to variable selection. *Journal of Chemometrics, 10*(5–6), 521–532.

Nocairi, H., Qannari, E. M., Vigneau, E., & Bertrand, D. (2005). Discrimination on latent components with respect to patterns. Application to multicollinear data. *Computational Statistics and Data Analysis, 48*(1), 139–147.

Pohjanen, E., Thysell, E., Jonsson, P., et al. (2007). A multivariate screening strategy for investigating metabolic effects of strenuous physical exercise in human serum. *Journal of Proteome Research, 6*(6), 2113–2120.

Rantalainen, M., Cloarec, O., Ebbels, T. M. D., et al. (2008). Piecewise multivariate modelling of sequential metabolic profiling data. *BMC Bioinformatics, 9*, article no. 105.

Rezzi, S., Ramadan, Z., Fay, L. B., & Kochhar, S. (2007). Nutritional metabonomics: Applications and perspectives. *Journal of Proteome Research, 6*(2), 513–525.

Skov, T., & Bro, R. (2008). Solving fundamental problems in chromatographic analysis. *Analytical and Bioanalytical Chemistry, 390*(1), 281–285.

Smit, S., van Breemen, M. J., Hoefsloot, H. C. J., et al. (2007). Assessing the statistical validity of proteomics based biomarkers. *Analytica Chimica Acta, 592*(2), 210–217.

Sokal, R. R., & Rohlf, F. J. (1998). *Biometry*. New York: W.H. Freeman and Company.

Trygg, J., & Wold, S. (2002). Orthogonal projections to latent structures (O-PLS). *Journal of Chemometrics, 16*(3), 119–128.

van Velzen, E. J. J., Westerhuis, J. A., van Duynhoven, J. P. M., et al. (2008). Multilevel data analysis of a crossover designed human nutritional intervention study. *Journal of Proteome Research, 7*(10), 4483–4491.

van Velzen, E. J. J., Westerhuis, J. A., van Duynhoven, J. P. M., et al. (2009). Phenotyping tea consumers by nutrikinetic analysis of polyphenolic end-metabolites. *Journal of Proteome Research, 8*(7), 3317–3330.

Westerhuis, J. A., Hoefsloot, H. C. J., Smit, S., et al. (2008a). Assessment of PLSDA cross validation. *Metabolomics, 4*(1), 81–89.

Westerhuis, J. A., van Velzen, E. J. J., Hoefsloot, H. C. J., & Smilde, A. K. (2008b). Discriminant Q(2) (DQ(2)) for improved discrimination in PLSDA models. *Metabolomics, 4*(4), 293–296.

Wiklund, S., Johansson, E., Sjostrom, L., et al. (2008). Visualization of GC/TOF-MS-based metabolomics data for identification of biochemically interesting compounds using OPLS class models. *Analytical Chemistry, 80*(1), 115–122.

Wu, W., Daszykowski, M., Walczak, B., et al. (2006). Peak alignment of urine NMR spectra using fuzzy warping. *Journal of Chemical Information and Modeling, 46*(2), 863–875.