BIOINFORMATICS ORIGINAL PAPER
Vol. 21 no. 13 2005, pages 3043–3048
doi:10.1093/bioinformatics/bti476

Systems biology

ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data

Age K. Smilde[1,2,*], Jeroen J. Jansen[1], Huub C. J. Hoefsloot[1], Robert-Jan A. N. Lamers[2], Jan van der Greef[2,3] and Marieke E. Timmerman[4]

[1]Biosystems Data Analysis, Faculty of Sciences, University of Amsterdam, Nieuwe Achtergracht 166, 1018 WV Amsterdam, The Netherlands, [2]TNO Quality of life, PO Box 360, 3700 AJ Zeist, The Netherlands, [3]Center for Medical Systems Biology, LACDR, Leiden University, Gorleaus Laboratories, 2300 RA Leiden, The Netherlands and [4]Heymans Institute of Psychology, DPMG, University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands

Received on December 23, 2004; revised on April 21, 2005; accepted on April 27, 2005
Advance Access publication May 12, 2005

ABSTRACT

Motivation: Datasets resulting from metabolomics or metabolic profiling experiments are becoming increasingly complex. Such datasets may contain underlying factors, such as time (time-resolved or longitudinal measurements), doses or combinations thereof. Currently used biostatistics methods do not take the structure of such complex datasets into account. However, incorporating this structure into the data analysis is important for understanding the biological information in these datasets.

Results: We describe ASCA, a new method that can deal with complex multivariate datasets containing an underlying experimental design, such as metabolomics datasets. It is a direct generalization of analysis of variance (ANOVA) for univariate data to the multivariate case. The method allows for easy interpretation of the variation induced by the different factors of the design. The method is illustrated with a dataset from a metabolomics experiment with time and dose factors.

Availability: M-files for MATLAB for the algorithm used in this research are available at: http://www-its.chem.uva.nl/research/pac/Software/ or at http://www.bdagroup.nl

Contact: asmilde@science.uva.nl

INTRODUCTION

Recent developments in genomics and human systems biology have shown the importance of metabolomics (Clish *et al*., 2004; Lindon *et al*., 2004; van der Greef *et al*., 2003). This is understandable, since metabolomics is a crucial element in bridging the difference between the genotype and phenotype of an organism (Fiehn, 2002). Considerable effort has gone into the development of instrumental methods for metabolite profiling, since it emerged in the 1960s and 1970s in clinical chemistry (Gates and Sweeley, 1978; Jellum, 2001), especially focusing on inborn errors of metabolism. The combination with mass spectrometry (GC–MS) and in particular chemometrics created the basis of the technology of today in metabolomics for mammalian systems (Gaspari *et al*., 2001; van der Greef *et al*., 1983). Nuclear Magnetic Resonance (NMR) spectroscopy has become an important component in the field for screening of biofluids (Bales *et al*., 1984) and an important step was the combination with chemometrics (Gartland *et al*., 1990). Moreover, methods for handling metabolomics data are receiving increased attention, with respect to both the preprocessing of metabolomics data (Keun *et al*., 2003; Vogels *et al*., 1996) and the analysis of data itself (Antti *et al*., 2002; Jansen *et al*., 2004; Keun *et al*., 2004).

Metabolomics datasets are becoming more and more complex. It is not uncommon to measure a multiple of metabolites in body fluids of several animals at different points in time with an underlying experimental design, e.g. different dose groups (Antti *et al*., 2004; Keun *et al*., 2004; Lamers *et al*., 2003). This calls for data analysis methods specifically suited for time-resolved (or 'longitudinal'), multigroup, multisubject (containing data of multiple animals) and multivariate data.

When a single variable (e.g. a metabolite) is measured as a function of design factors, analysis of variance (ANOVA) is a well established technique to analyze the data (Searle, 1971). When measuring many metabolites simultaneously, generalizations of ANOVA are necessary. In the statistics literature, the classical generalization of ANOVA to multiple variables is multivariate-ANOVA (MANOVA) (Mardia *et al*., 1979). For the large number of measured variables in a metabolomics experiment, however, MANOVA breaks down owing to problems of singularity of covariance matrices and assumptions that are not fulfilled (Ståhle and Wold, 1990).

In the data analysis literature, mixtures of multivariate analysis and ANOVA have also been reported. One approach performs a principal component analysis of the entire dataset first, and then uses ANOVA on the component score values to test the effects (Bratchell, 1989). This approach has been criticized, since the separate ANOVAs on the score values are not independent (Jackson, 1991). Moreover, the initial principal component analysis does not necessarily distinguish between the groups in the data. Another approach is suggested by using PLS (partial least squares; a popular regression technique for collinear data) to solve the problem (Ståhle and Wold, 1990). However, different suggestions have been made to implement this method depending on whether the coded design variables are regressors (Martens and Martens, 2001) or regressands (Ståhle and Wold, 1990). Moreover, the exact properties of these methods are unknown, since

*To whom correspondence should be addressed.

the criterion that is maximized or minimized is not clear. Redundancy analysis using a coded design matrix seems to be a better alternative than the PLS based approach (Van den Brink and Ter Braak, 1999).

In this paper, a new method is presented that can deal with a temporal and/or design structure of complex multivariate datasets, such as those emerging from metabolomics experiments. However, the problems to which ANOVA-simultaneous component analysis (ASCA) can be applied are certainly not limited to metabolomics. Complex multivariate datasets are abundant in the other post-genomic technologies (e.g. transcriptomics, proteomics) and also in many more fields of biological and non-biological research.

ASCA builds upon and generalizes some earlier proposed methods. Two early papers in pomology and botanics realized the importance of distinguishing 'between' and 'within' factor treatments (Jeffers, 1962; Pearce and Holland, 1960). In the metabolomics literature, the SMART method (Keun *et al*., 2004) also makes this distinction, but is less general than the method proposed in the current paper. The method proposed in this paper builds on the multilevel component analysis method that was developed in psychometrics (M.E. Timmerman, submitted for publication) and metabolomics (Jansen *et al*., 2004) and generalizes it for a situation with any design structure underlying the metabolomics data.

The ASCA method will be explained and illustrated with an example from a metabolomics intervention study, where guinea pigs from a strain developing osteoarthritis (OA) were treated with several dosage levels of vitamin C and their urine was analyzed using NMR spectroscopy at several points in time. This study is therefore a typical example of a designed metabolomics experiment. OA is a multifactorial chronic joint disease that is characterized by the progressive destruction of articular cartilage, resulting in impaired movement, pain and ultimately disability (Creamer and Hochberg, 1997). The Hartley outbred strain guinea pig develops spontaneous progressive knee OA starting when they are ∼10 months of age, with features similar to the human disease (Bendele, 2001; Huebner *et al*., 2001). Ascorbic acid has been associated with the slowing of OA progression in guinea pig and human (McAlindon *et al*., 1996). However, recent studies indicate that vitamin C increases the severity of development of OA in the guinea pig (Kraus *et al*., 2004). The details of the biological questions regarding this study are published elsewhere (Lamers *et al*., 2003).

SYSTEMS AND METHODS

Urine samples and data acquisition

This dataset contains information about male Hartley guinea pigs, which develop OA during aging. Beginning at 4 months of age, the guinea pigs are divided randomly into three dose groups to which varying doses of vitamin C are provided: low dose (2.5–3 mg/day), medium dose (30 mg/day) and high dose (150 mg/day). The doses were chosen such that the low dose exceeds the minimum amount to prevent scurvy and the medium dose corresponds to the normal intake of vitamin C. The high dose of vitamin C corresponds to the amount that was shown in previous studies to slow the development of surgically induced OA.

Each dose group consists of six animals. Urine samples are collected at 4, 7, 10 and 12 months, where the samples collected after 4 months are predose samples. Each urine collection was performed for 24 h, to eliminate the influence of diurnal variation of the metabolite composition of the urine. The total dataset consists of 72 samples. These samples were analyzed with NMR spectroscopy and the dataset was prepared as peak listings (NMR spectra) using the standard Varian software (Varian Inc., Palo Alto, CA). The dataset contains spectra with peaks listed at 253 chemical shifts, expressed in parts per million (p.p.m.), that are equal for all spectra. A typical spectrum in this dataset is given in Figure 1.

The dataset is a subset of a larger dataset. The acquisition of this larger dataset has been described elsewhere (Lamers *et al*., 2003).

Data analysis

Structure of the dataset The structure of the dataset is shown in Figure 1. The following indices will be used: $j = 1, \ldots, J$ for the chemical shifts, $k = 1, \ldots, K$ for the time-points at which measurements are taken, $h = 1, \ldots, H$ for the dosage groups ($h = 1$, low; $h = 2$, medium and $h = 3$, high dosage) and $i_h = 1, \ldots, I_h$ for the guinea pigs nested within the dosage groups. The guinea pigs the 'low' in dosage group are different from those in the other dosage groups (the subindex h on i is used to emphasize this fact). This will be important to realize for the remainder of the paper.

Analysis of variance An NMR signal at one particular chemical shift j, for one-time point k and for one guinea pig i_h (in dosage group h) will be denoted by the scalar x_{hki_hj}. Collecting such values x_{hki_hj} in matrices $\mathbf{X}_{hi_h}$ of size $(K \times J)$ will be convenient. The construction of these matrices is shown in Figure 1. Considering only an NMR signal at one chemical shift (and therefore dropping the index j for convenience) a reasonable ANOVA model would be

$$x_{hki_h} = \mu + \alpha_k + (\alpha\beta)_{hk} + (\alpha\beta\gamma)_{hki_h} \quad (1)$$

where μ represents an overall offset; α_k represents the effect of the factor 'time' that is common for all guinea pigs; $(\alpha\beta)_{hk}$ represents the interaction of 'time' and 'dose'; $(\alpha\beta\gamma)_{hki_h}$ represents the guinea pig specific contribution. Of these effects, $(\alpha\beta)_{hk}$ is most important for the biological interpretation. It represents the effect of the dosage measured as deviations from the common time effect α_k. The contributions $(\alpha\beta\gamma)_{hki_h}$ represent the variations on the lowest (individual animal-specific) level, and can be used for significance testing. Classical ANOVA techniques can now be used to estimate the factor effects and test significance.

Equation (1) shows a division of variation on several factors. This is the basic idea of ANOVA; variation is separated and assigned to factors. The factor effects can be estimated and tested. ANOVA is capable of doing this by splitting the variations in orthogonal and independent parts (Searle, 1971). This division of the variation into orthogonal contributions is the goal of ASCA also (see below).

Simultaneous component analysis When analyzing the simultaneous underlying variation in several related datasets, simultaneous component analysis is a useful tool. This method was developed in psychometrics (Ten Berge *et al*., 1992), but extensions also find their use now in metabolomics (Jansen *et al*., 2004; M. E. Timmerman, submitted for publication).

Let us suppose that data matrices $\mathbf{X}_i$ $(K_i \times J)$ are available where measurements on J identical chemical shifts are available at K_i time-points on I animals (the subdivision of the individuals i into different dose groups h is omitted from the explanation of SCA for simplicity; therefore, the used indices are simpler in this section compared with the other sections of this paper). Note that the number of measurement time points for individual i, denoted by K_i, can differ between animals in SCA. Then a reasonable model for simultaneously analyzing these data matrices is

$$\mathbf{X}_i = \mathbf{T}_i\mathbf{P}^{\mathrm{T}} + \mathbf{E}_i \quad (2)$$

where $\mathbf{P}$ $(J \times R)$ represents the common basis with R directions (components) and $\mathbf{T}_i$ $(K_i \times R)$ contains the scores of the measurement time-points of the i-th animal. Since the variation across the animals at the various time-points is expressed on the common basis $\mathbf{P}$, the scores contained in $\mathbf{T}_i$ can be compared between animals to explore the data. There exist different versions of simultaneous component analysis depending on the type of constraint put on the covariance of $\mathbf{T}_i$, but such constraints are not discussed in this paper.

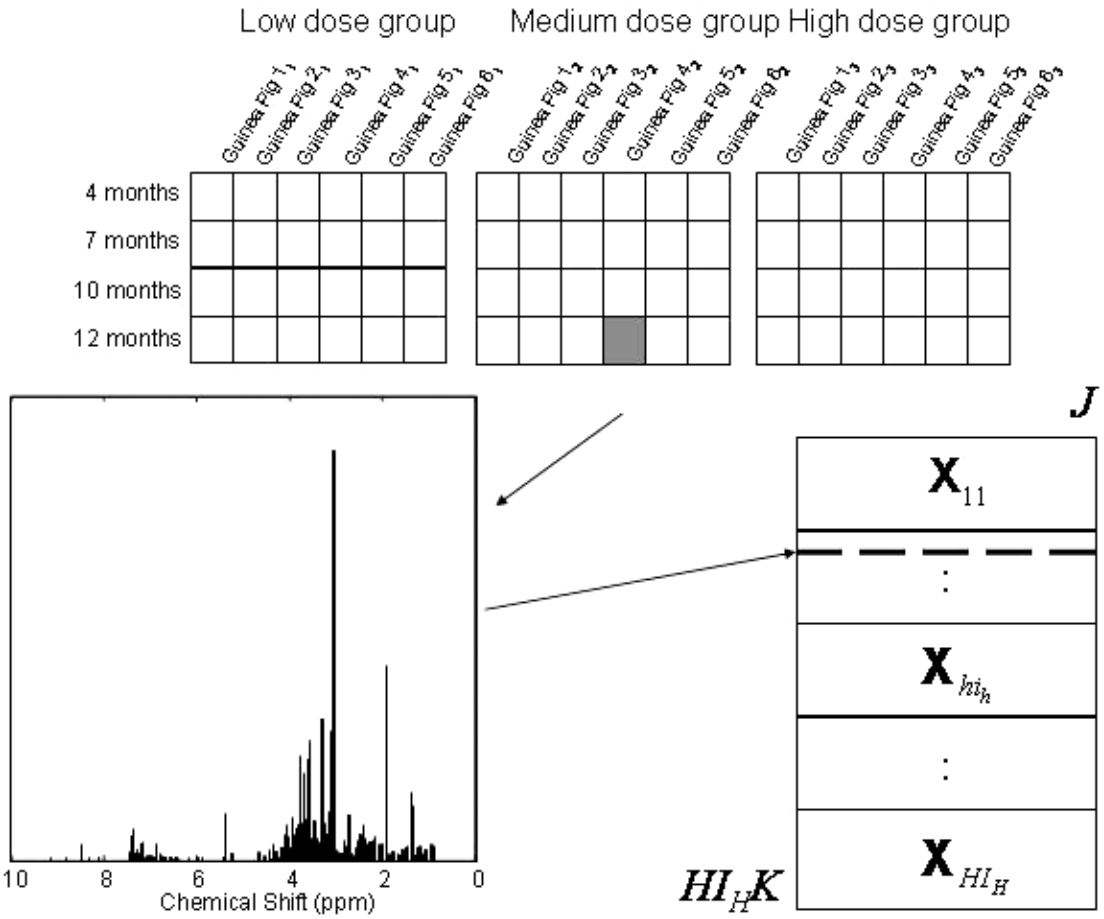

Fig. 1. Structure of the dataset. The relationship between the measurements is given in the top of the figure. The guinea pigs are nested within the dose groups, all other relationships between the factors in the experiment are crossed. Each square in the top panel of the figure represents an NMR spectrum like the one given in the bottom left-hand side. These obtained spectra are arranged into a matrix containing HI_H submatrices, as indicated in the bottom right-hand side.

The model parameters in Equation (2) can be found by solving

$$\min_{\mathbf{T}_i,\mathbf{P}} \sum_{i=1}^{I} \left\| \mathbf{X}_i - \mathbf{T}_i \mathbf{P}^{\mathrm{T}} \right\|^2 \tag{3}$$

which is a standard least squares problem that can be solved by performing a principal component analysis (PCA) on the matrix in which all matrices $\mathbf{X}_i$ are concatenated as $\begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_I \end{bmatrix}$.

ANOVA-simultaneous component analysis (ASCA) The ASCA analog for model (1) is

$$\mathbf{X}_{hi_h} = \mathbf{1}\mathbf{m}^{\mathrm{T}} + \mathbf{T}_{\mathbf{K}}\mathbf{P}_1^{\mathrm{T}} + \mathbf{T}_{\mathbf{K}h}\mathbf{P}_2^{\mathrm{T}} + \mathbf{T}_{\mathbf{K}hi_h}\mathbf{P}_3^{\mathrm{T}} + \mathbf{E}_{hi_h} \tag{4}$$

where $\mathbf{1}$ is a $(K \times 1)$ vector of ones; $\mathbf{m}$ is a $(J \times 1)$ vector of the overall means of the NMR responses (where the mean is taken over all factors and guinea pigs per chemical shift); $\mathbf{T}_{\mathbf{K}}$ $(K \times R_1)$ is the matrix containing the contributions of the factor 'time' expressed on the basis $\mathbf{P}_1$ $(J \times R_1)$; R_1 is the number of components chosen for the basis $\mathbf{P}_1$; $\mathbf{T}_{\mathbf{K}h}$ $(K \times R_2)$ is the matrix containing the dose-specific 'time' contributions $(h = 1, \ldots, H)$ expressed on the basis $\mathbf{P}_2$ $(J \times R_2)$; R_2 is the number of components chosen for the basis $\mathbf{P}_2$; $\mathbf{T}_{\mathbf{K}hi_h}$ $(K \times R_3)$ is the matrix of guinea pig-specific dose–time contributions $(i_h = 1, \ldots, I_h$; all $h)$ expressed on the basis $\mathbf{P_3}$ $(J \times R_3)$; R_3 is the number of components chosen for the basis $\mathbf{P_3}$ and $\mathbf{E}_{hi_h}$ is the matrix of residuals. Note that in Equation (4), $\mathbf{T}_{\mathbf{K}}$ is equal for all animals, $\mathbf{T}_{\mathbf{K}h}$ is equal for all animals belonging to the same dose group and $\mathbf{T}_{\mathbf{K}hi_h}$ is different for all animals. Since each time-point K_{hi_h}, in $\mathbf{T}_{\mathbf{K}}$ and $\mathbf{T}_{\mathbf{K}h}$, is compared between different individuals, the measurement time-points are required to be equal for all animals, such that $K_{hi_h} = K$.

In other words, Equation (4) means that the matrix $\mathbf{X}_{hi_h}$ is separated into contributions from the overall mean ($\mathbf{1}\mathbf{m}^{\mathrm{T}}$), one SCA model ($\mathbf{T}_{\mathbf{K}}\mathbf{P}_{\mathbf{1}}^{\mathrm{T}}$) describing the overall effect of the factor time, an SCA model ($\mathbf{T}_{\mathbf{K}h}\mathbf{P}_{\mathbf{2}}^{\mathrm{T}}$) describing the interaction of dose with time and another SCA model ($\mathbf{T}_{\mathbf{K}hi_h}\mathbf{P}_{\mathbf{3}}^{\mathrm{T}}$) describing the interaction of dose, time and guinea pig, which is the contribution to the variation of each individual guinea pig. This is a direct multivariate generalization of Equation (1). Note that since the number of components in each part of this model is low (which is the basic idea: dimension reduction) there is a residual matrix $\mathbf{E}_{hi_h}$ in Equation (4) that contains the information that is not described by any of the ASCA submodels, whereas such a residual term is not present in Equation (1). To illustrate the basic idea behind ASCA, a toy example is given in Appendix 1 of the supplementary data.

Properties of ASCA The different parts of the model in Equation (4) are orthogonal to each other when proper constraints are imposed on them. These

constraints are

$$\begin{aligned} &(a) \quad \mathbf{1}^{\mathrm{T}}\mathbf{T}_{\mathbf{K}} = \mathbf{0}^{\mathrm{T}} \\ &(b) \quad \sum_{h=1}^{H} \mathbf{T}_{\mathbf{K}h} = \mathbf{0} \\ &(c) \quad \sum_{i_h=1}^{I_h} \mathbf{T}_{\mathbf{K}hi_h} = \mathbf{0} \forall h = 1, \ldots, H \end{aligned} \tag{5}$$

where **1** is a vector of ones of the proper order and **0** is a vector or a matrix of zeros of the proper order. In words, (a) ensures that $\mathbf{T}_{\mathbf{K}}$ is orthogonal to $\mathbf{1}\mathbf{m}^{\mathrm{T}}$, and similarly (b) and (c) ensure orthogonality of the other parts (for a detailed description of the mathematics behind ASCA see Appendix 2 in the supplementary data). This also means that the total variation of the dataset can be separated in parts corresponding to the different factors.

ASCA-algorithms The parameters of the ASCA model can be calculated by solving the following least squares problem

$$\min_{\mathbf{m},\mathbf{T},\mathbf{P}} \sum_{h=1}^{H} \sum_{i_h=1}^{I_h} \left\| \mathbf{X}_{i_h} - \mathbf{1}\mathbf{m}^{\mathrm{T}} - \mathbf{T}_{\mathbf{K}}\mathbf{P}_{\mathbf{1}}^{\mathrm{T}} - \mathbf{T}_{\mathbf{K}h}\mathbf{P}_{\mathbf{2}}^{\mathrm{T}} - \mathbf{T}_{\mathbf{K}hi_h}\mathbf{P}_{\mathbf{3}}^{\mathrm{T}} \right\|^2 \tag{6}$$

under the constraints given in Equation (6). Although this looks like a complicated problem, these constraints actually make the problem relatively simple. Owing to these contraints, the parameter sets corresponding to the different factor combinations ($\mathbf{m}$; $\mathbf{T}_{\mathbf{K}}$, $\mathbf{P}_{\mathbf{1}}$; $\mathbf{T}_{\mathbf{K}h}$, $\mathbf{P}_{\mathbf{2}}$; $\mathbf{T}_{\mathbf{K}hi_h}$, $\mathbf{P}_{\mathbf{3}}$) can be estimated independently. Note that there is rotational freedom in the model parts containing $\mathbf{P}_{\mathbf{1}}$, $\mathbf{P}_{\mathbf{2}}$ and $\mathbf{P}_{\mathbf{3}}$. Hence, the matrices $\mathbf{P}_{\mathbf{1}}$, $\mathbf{P}_{\mathbf{2}}$ and $\mathbf{P}_{\mathbf{3}}$ can be chosen to be orthogonal. For the case of a balanced design (equal number of guinea pigs for each factor combination) it comes down to proper centering and performing PCAs on rearranged data. An algorithm is provided, but standard algorithms for PCA can also be used after the proper centering and rearrangement of the data. For the unbalanced case, a slightly more elaborate algorithm should be used, which is a straightforward generalization of the balanced one. For an explanation of the algorithm, see Appendix 3 in the supplementary data.

RESULTS

Split-up of variation

An impression of the amount of variation related to the design factors can be obtained by separating this variation into contributions from the different factors. Table 1 shows this separation and it is clear that the dominant part of the variation is at the lowest level (guinea pig-specific contributions). This shows the biological variation between the animals used in the study. Note that Table 1 reports sums of squared deviations from the overall mean and not variances.

Factor 'time'

The scores of the first component of the factor 'time' are given in Figure 2 [the $\mathbf{T}_{\mathbf{K}}$ values of Equation (4)]. This first component explains 72% of the variation on the factor 'time'. The maximum number of components that can be fitted for the factor 'time' submodel (and therefore the rank of the factor time variation) is three, since the dataset contains four measurement time-points. Hence, only one component is used to illustrate the variation on this level.

The scores in Figure 2 indicate that all guinea pigs in the data show an initial increasing and a subsequent decreasing behavior. This trajectory is consistent with the biology of growth for the Hartley guinea pig strain (Huebner *et al*., 2001). From 4 to 7 months the metabolism of the guinea pigs changes, because during this time they are in the growing phase. Between 7 and 10 months, the guinea pigs are full-grown, which is shown by the leveling off in the time profile from 7 and 10 months. From 10 months onward, the guinea pigs develop OA. The decrease in score of the 12 months samples is supposed to reflect this effect (Lamers *et al*., 2003).

Table 1. Contributions to the total variation

Level	Percentage of variation in the data
K	24
Kh	10
Khi_h	66

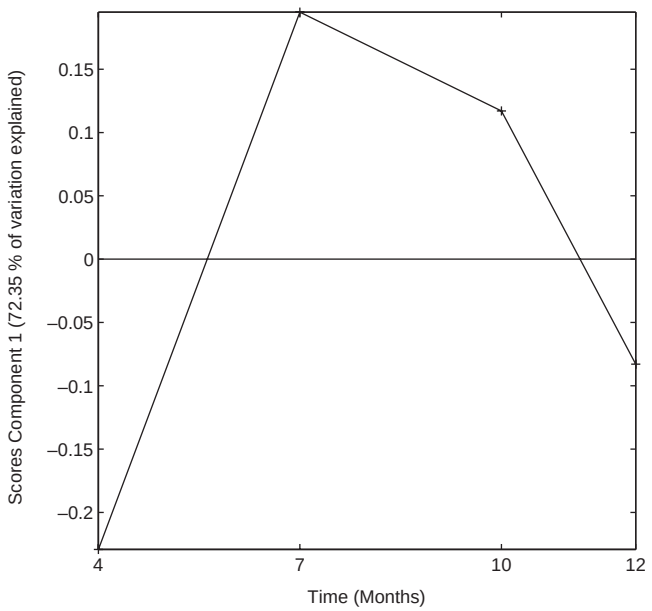

Fig. 2. Factor 'time' scores on the first component. An initially increasing and subsequently decreasing time profile clearly is visible.

The loadings $\mathbf{P}_{\mathbf{1}}$ belonging to the first component are given in Figure 3. These loadings show the chemical shifts and thus the compounds that are corresponding to the behavior observed in the scores, and can be used for biological interpretation. However, urine is a biological fluid that is used by the body for the excretion of waste products and therefore its contents are difficult to trace back to biology. Nevertheless, metabolites like creatinine (δ 3.04 and 4.05 p.p.m.), creatine (δ 3.04 and 3.95 p.p.m.), glucose (δ 3.27, 3.53, 3.60, 3.78, 3.82 and 3.94 p.p.m.), alpha-hydroxybutyrate (δ 1.36 p.p.m.), lactate (δ 1.32 and 1.34 p.p.m.), glycine (δ 3.56 p.p.m.) and acetate (δ 1.92 p.p.m.) that change in time may point at altered energy metabolism owing to growth and disease development. These observations are consistent with results that were described previously (Lamers *et al*., 2003).

Interaction time × dose

The rank required for this submodel is determined using a scree-test (Cattell, 1966). From this test, the rank of this submodel is determined to be two. A model containing two components explains 65% of the variation corresponding to the factor 'interaction time × dose'. The first component explains 50% of this variation and the second component explains 15%.

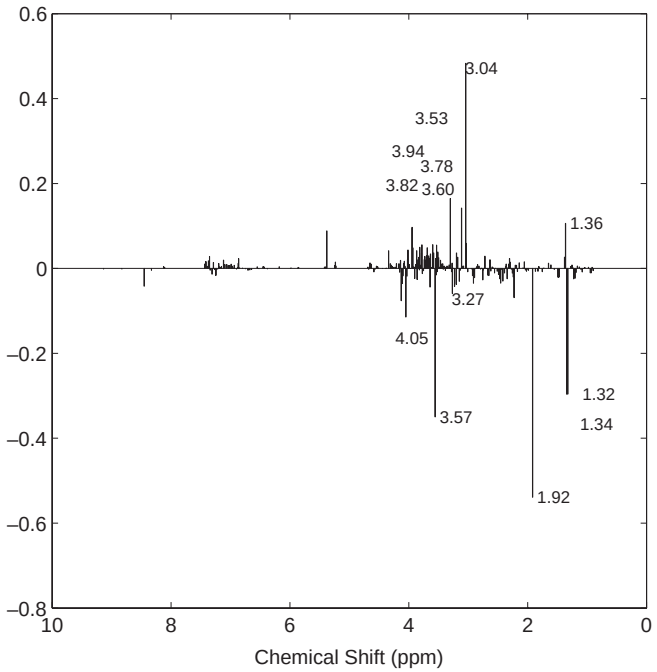

Fig. 3. Factor 'time' loadings on the first component. The chemical shifts that are mentioned in the text are indicated in the figure.

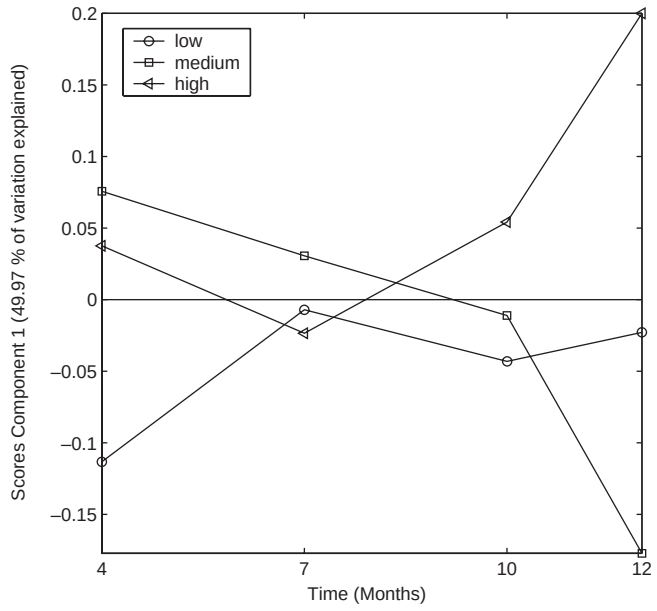

Fig. 4. Interaction 'dose × time' scores on the first component. No quantitative effect is visible in the scores and therefore this model shows that vitamin C has no effect on the development of OA.

The scores of this submodel should be interpreted as the deviation of each dose group from the 'time' factor ($\mathbf{T_K P_1^T}$ or α_k in the ANOVA model). The 'interaction time × dose' scores for the first component are given in Figure 4. On this component no trend related to vitamin C dose is visible: none of the four measurement time-points show an increasing or decreasing score value for differing vitamin C doses. Also the scores of the second component do not show such a quantitative trend.

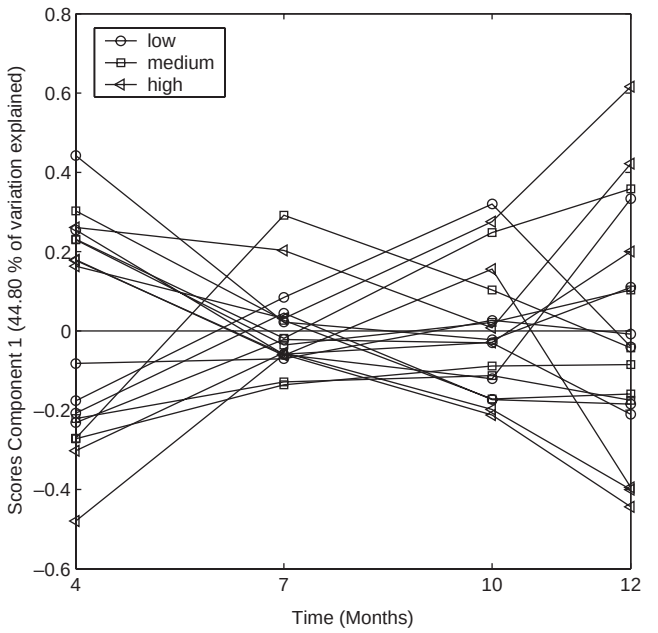

Fig. 5. 'Individual guinea pig contribution' scores on the first component. These scores indicate the deviation of each individual from the dose group-specific variation of the metabolism.

The model results show that the potential of vitamin C in affecting the development of OA is questionable. According to our results, vitamin C has no effect on disease development. Neither the association of vitamin C with the slowing of OA progression in guinea pig and humans (McAlindon *et al*., 1996) nor the observation that vitamin C could increase the severity of development of OA in the guinea pig (Kraus *et al*., 2004) can be corroborated with the results of ASCA on the described dataset. However, the results of the ASCA model of this dataset were in agreement with additional clinical measurements that were performed after the experiment on the guinea pigs used in the study. The severity of OA was determined using 'histology scores' on the Mankin grading system (Mankin *et al*., 1971). These scores did not differ between dose groups, which indicates an equal development of OA.

Individual guinea pig contributions

The rank required for the 'Individual guinea pig contributions' submodel can also be determined by a scree test. From this test, two components are defined for this submodel. This submodel describes 57% of the variation corresponding to the 'Individual guinea pig contributions'. The first component explains 45% and the second component explains 12%. The scores of this submodel must be seen as a deviation of each individual from the dose–time interaction.

The scores for the first component of the 'Individual guinea pig contributions' submodel are given in Figure 5 for the low, medium and high dose groups. From this figure it is clear that the deviation of the individual profile from the group average profile is largest at 4 and 12 months, i.e. at the start and the end of the experiment. The NMR signals that correspond to this behavior are, amongst others,

lactate (δ 1.32 and 1.34 p.p.m.), acetate (δ 1.92 p.p.m.) and glycine (δ 3.56 p.p.m.) that increase and creatinine (δ 3.04 and 4.05 p.p.m.) that decrease. The larger interindividual differences at 4 and 12 months may be explained by the fact that at these time-points growth and disease development occur.

CONCLUSIONS

In metabolomics, an increasing amount of datasets becomes available with an underlying design in factors. Currently, no methods are available to analyze such data. The method proposed in this paper called ANOVA-SCA or ASCA for short, fills this gap. It works by separating the variation in the total dataset by parts that can be assigned to contributions of the different factors and interactions thereof.

The ASCA method is illustrated by a real example of an intervention study examining the effect of vitamin C on the development of OA in guinea pigs. This shows how the method works and that interpretation of the resulting components works in the same way as in an ordinary PCA. In the case of a balanced design, the algorithm is simple and comes down to performing PCAs on properly centered and rearranged data. In the case of unbalanced data, a more elaborate algorithm is necessary, but available.

In a follow-up research, questions regarding validation using resampling methods and significance testing of effects will be treated. This will allow not only for the estimation of factor effects but also for judging their reliability and testing their significance.

ACKNOWLEDGEMENTS

The authors would like to acknowledge Jeroen de Groot of TNO Prevention and Health in Leiden, the Netherlands and Virginia B. Kraus of Duke University Medical Center in Durham, NC for providing the guinea pig dataset.

SUPPLEMENTARY DATA

Supplementary data for this paper is available on Bioinformatics online.

REFERENCES

Antti,H. *et al*. (2002) Batch statistical processing of H-1 NMR-derived urinary spectral data. *J. Chemometr.*, **16**, 461–468.

Antti,H. *et al*. (2004) Statistical experimental design and partial least squares regression analysis of biofluid metabonomic NMR and clinical chemistry data for screening of adverse drug effects. *Chemom. Intell. Lab. Syst.*, **73**, 139–149.

Bales,J.R. *et al*. (1984) Use of high-resolution proton nuclear magnetic resonance spectroscopy for rapid multi-component analysis of urine. *Clin. Chem.*, **30**, 426–432.

Bendele,A.M. (2001) Animal models of osteoarthritis. *J. Musculoskel. Neuron. Interact.*, **1**, 363–376.

Bratchell,N. (1989) Multivaraite response surface modeling by principal component analysis. *J.Chemometr.*, **3**, 579–588.

Cattell,R.B. (1966) The scree test for the number of factors. *Multivariate Behav. Res.*, **1**, 245–276.

Clish,C.B. *et al*. (2004) Integrative biological analysis of the APOE*3-leiden transgenic mouse *OMICS* 2004. *OMICS*, **1**, 3–13.

Creamer,P. and Hochberg,M.C. (1997) Osteoarthritis. *Lancet*, **350**, 503–508.

Fiehn,O. (2002) Metabolomics-the link between genotypes and phenotypes. *Plant Mol. Biol.*, **48**, 155–171.

Gartland,K.P. *et al*. (1990) Pattern recognition analysis of high resolution 1H NMR spectra of urine. A nonlinear mapping approach to the classification of toxicological data. *NMR Biomed.*, **3**, 166–172.

Gaspari,M. *et al*. (2001) Novel strategies in mass spectrometric data handling. *Adv. Mass. Spectrom.*, **15**, 283–296.

Gates,S.C. and Sweeley,Ch.C. (1978) Quantitative metabolic profiling based on gas chromatography. *Clin.Chem.*, **24**, 1663–1673.

Huebner,J.L. *et al*. (2001) Collagenase 1 and collagenase 3 expression in a guinea pig model of osteoarthritis. *Arthritis Rheum.*, **41**, 877–890.

Jackson,J. (1991) *A User's Guide to Principal Components*. Wiley & Sons, New York.

Jansen,J.J. *et al*. (2004) Multilevel component analysis of time-resolved metabolomics data. *Anal. Chim. Acta*, **530**, 173–183.

Jeffers,J.N.R. (1962) Principal component analysis of designed experiments. *Statistician*, **12**, 230–242.

Jellum,E. (2001) Chromatography, mass spectrometry and electrophoresis for diagnosis of human disease, particulary metabolic disorders. In Gehrke,Ch., Wixom,R. and Bayer,E. (eds), *Chromatography—A Century of Discovery 1900–2000*, Elsevier, Amsterdam, pp. 270–277.

Keun,H.C. *et al*. (2004) Geometric trajectory analysis of metabolic responses to toxcicity can define treatment specific profiles. *Chem. Res. Toxicol.*, **17**, 579–587.

Keun,H.C. *et al*. (2003) Improved analysis of multivariate data by variable stability scaling: application to NMR-based metabolic profiling. *Anal. Chim. Acta*, **490**, 265–276.

Kraus,V.B. *et al*. (2004) Ascorbic acid increases the severity of spontaneous knee osteoarthritis in a guinea pig model. *Arthritis Rheum.*, **50**, 1822–1831.

Lamers,R.A.N. *et al*. (2003) Identification of disease and nutrient-related metabolic fingerprints in osteoarthritic guinea pigs, *J. Nutr.*, **133**, 1776–1780.

Lindon,J.C. *et al*. (2004) Metabonomics technologies and their applications in physiological monitoring, drug safety assessment and disease diagnosis. *Biomarkers*, **9**, 1–31.

Mankin,H.J. *et al*. (1971) Biochemical and metabolic abnormalities in articular cartilage from osteo-arthritic human hips. II. Correlation of morphology with biochemical and metabolic data. *J. Bone Joint Surg. Am.*, **53**, 523–537.

Mardia,K.V., Kent,J.T. and Bibby,J.M. (1979) *Multivariate Analysis*. Academic Press, London.

Martens,H. and Martens,M. (2001) *Multivariate Analysis of Quality. An Introduction*. John Wiley & Sons, Chichester.

McAlindon,T.E. *et al*. (1996) Do antioxidant micronutrients protect against the development and progression of knee osteoarthritis? *Arthritis Rheum.*, **39**, 648–656.

Pearce,S.C. and Holland,D.A. (1960) Some applications of multivariate methods in botany. *Applied Stat.*, **9**, 1–7.

Searle,S.R. (1971) *Linear Models*. John Wiley & Sons, New York.

Ståhle,L. and Wold,S. (1990) Multivariate analysis of variance (MANOVA). *Chemom. Intell. Lab. Syst.*, **9**, 127–141.

Ten Berge,J.M.F. *et al*. (1992) Simultaneous component analysis. *Statistica Applicata*, **4**, 277–392.

Van den Brink,P.J. and Ter Braak,C.J.F. (1999) Principal response curves: analysis of time-dependent multivariate responses of biological community to stress. *Environ. Toxicol. Chem.*, **18**, 138–148.

van der Greef,J., Davidov,E., Verheij,E.R., van der Heijden,R., Adourian,A.S., Oresic,M., Marple,E.W. and Naylor,S. (2003) The role of metabolomics in Systems Biology. In Harrigan,G.G. and Goodacre,R. (eds), *Metabolic Profiling: Its Role in Biomarker Discovery and Gene Function Analysis*, Kluwer Academic Publishers, Boston/Dordrecht/London, pp. 170–198.

van der Greef,J. *et al*. (1983) Evaluation of field desorption and fast atom bombardment mass spectrometric profiles by pattern recognition techniques. *Anal. Chim. Acta*, **150**, 45–52.

Vogels,J.T.W.E. *et al*. (1996) Partial linear fit: a new NMR spectroscopy preprocessing tool for pattern recognition applications. *J. Chemometr.*, **10**, 425–438.