Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v8n2p214

AoV-PLS: a new method for the analysis of multivariate data depending on several factors

By El Ghaziri *et al.*

Published: 14 October 2015

Electronic Journal of Applied Statistical Analysis
Vol. 08, Issue 02, October 2015, 214-235
DOI: 10.1285/i20705948v8n2p214

AoV-PLS: a new method for the analysis of multivariate data depending on several factors

El Ghaziri Angélina*[a], El Mostafa Qannari[a], Thomas Moyon[b], and Marie-Cécile Alexandre-Gouabau[b]

[a]*LUNAM Université, ONIRIS, Unité de sensométrie et Chimiométrie, Nantes, F-44322, France; INRA, 44307, Nantes, France,*
[b]*INRA, UMR1280, Physiologie des Adaptations Nutritionnelles, Nantes, France,*

Published: 14 October 2015

A new method for the analysis of a multivariate dataset depending on several factors is proposed. It is called AoV-PLS (Analysis of Variance-PLS). It is based on the decomposition of the dataset into the main effects, the interactions effects and possibly the residual matrix using a model akin to analysis of variance (ANOVA). Each effect is considered in turn and assessed through the use of a Partial Least Square regression (PLS-regression). The connection of AoV-PLS to competing methods such as ANOVA-PCA and ANOVA-Simultaneous Component Analysis (ASCA) is emphasized and these methods are compared on the basis of a dataset pertaining to metabolomics field.

keywords: ANOVA, ANOVA-Simultaneous Component Analysis, ANOVA-PCA, PLS regression, PLS-DA, Metabolomics data.

1 Introduction

It often occurs that we are faced with a problem that involves the collection of multivariate data that depend on several factors of variation. These are categorical variables measured on the individuals. For instance, in sensory analysis, the so-called conventional sensory profiling consists in assessing the sensory characteristics of a set of products by

*Corresponding author: angelina.el-ghaziri@oniris-nantes.fr

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

means of trained assessors. Therefore, these multivariate data depend on several factors: products, assessors, sessions of evaluation and the interaction between these factors. In metabolomics, Smilde et al. (2005) expand on this kind of data and discuss the example of several metabolites (variables) measured on animals at different points of time according to an experimental design. One of the aims in these studies is to assess the effect of the various factors on the variability of the data. When we dispose of only one variable whose variations depend on several factors, we usually have recourse to ANOVA. Likewise, when we dispose of multivariate data and one factor of variation, one of the appropriate methods to use is discriminant analysis or one of its variants such as PLS-Discriminant analysis (PLS-DA; Barker and Rayens, 2003; Nocairi et al., 2005). We aim at extending PLS-DA to the case of several factors.

Since the situation considered herein (*i.e.* multivariate data depending on several factors) is relatively common, several statistical methods to deal with this kind of data have been proposed. Among these methods of analysis, we single out the following:

– Multivariate analysis of variance (MANOVA): typically, this method of analysis is concerned with the assessment of the effect of several factors on a multivariate dataset. However, it is focused on the assessment of the significance of the effects of the various factors and does not yield an appropriate framework for an exploratory data analysis. Moreover, it tends to be unstable in presence of quasi-colinearity among the multivariate data which is a common occurrence in chemometrics (Smilde et al., 2005).

– ANOVA-Simultaneous Component Analysis or ASCA was introduced by Smilde et al. (2005) in the context of metabolomics data. As discussed below, this method is based on a generalization of the ANOVA decomposition of a single variable into its main and interaction effects to the case of a multivariate dataset depending on several factors.

– ANOVA-PCA was first introduced in a proteomic data analysis framework (Harrington et al., 2005, 2006), and later used in studies on reference materials (Sarembaud et al., 2007). Similarly to ASCA, it is based on the same decomposition of the multivariate dataset into its main and interaction effects.

We propose a method of analysis which bears similarities to ASCA and ANOVA-PCA. Indeed, it is based on the same decomposition of the data matrix at hand according to the various sources of variations and the matrix of residuals. However, instead of successively performing a principal components analysis (PCA) on the matrices highlighting the effects of the various factors, we advocate using PLS regression. The advantages of this choice are twofold: (i) the method of analysis yields components that are likely to better highlight the effect of the factors under study than methods based on PCA; (ii) It is a straightforward extension of PLS-DA to the case of multiple factors.

The paper is organized as follows. In section 2, we discuss the proposed method of analysis and its connections to ANOVA-PCA and ASCA. We also introduce the case study pertaining to metabolomics field that is used to illustrate the approach of analysis. In section 3, we outline the results concerning the case study and we compare the performance of our approach of analysis to alternative approaches. We end the paper by concluding remarks.

2 Material and methods

2.1 The case of one factor

Suppose that we dispose of a multivariate dataset, X, depending on a factor F with q levels. This case is known as the case of multivariate measurements on a sample of individuals divided into q known groups. It pertains to the general framework of discriminant analysis where methods such as Fisher's linear discriminant analysis (LDA, McLachlan, 2004) or PLS-discriminant analysis (Barker and Rayens, 2003; Kemsley, 1996) are used. We discuss a strategy of analysis that leads to a variant of PLS-DA. The interest of this strategy is to hint to a general approach to analyze the data when more than one factor are available.

Following the classical decomposition used in ANOVA, matrix X can be decomposed as follows (Smilde et al., 2005; Harrington et al., 2005):

$$X = \bar{X} + X_F + E$$

where $\bar{X}$ and X_F have the same dimensions as X (say, n individuals by p variables). The rows of $\bar{X}$ are all similar and each contains the average values of the variables in X. The rows of X_F associated with the same level of factor F are also identical and contain the average values of the variables in $X - \bar{X}$ restricted to the individuals associated with the level of F under consideration. Finally, E is the matrix of residuals: $E = X - \bar{X} - X_F$. In a matrix form, if we denote by the vector of dimension n whose components are equal to 1, we have $\bar{X} = \frac{1}{n}{}^{T}X$ (*i.e.* the projection of X on the space spanned by). Let us denote by $_k$ $(k = 1, ..., q)$ the vector whose component i $(i = 1, ..., n)$ is equal to 1 if individual i has the level k and 0, otherwise. We have $X_F = P_F(X - \bar{X})$, where P_F is the projector upon the space spanned by the vectors $(_k)_{k=1...q}$. From these developments it follows that E and X_F are orthogonal: $X_F^T E = 0$ since $E = (I - P_F)(X - \bar{X})$, I being the identity matrix. As a consequence, the covariance matrix of X can be decomposed as follows

$$V = B + W$$

Where $V = \frac{1}{n}(X - \bar{X})^T(X - \bar{X})$, $B = \frac{1}{n}X_F^T X_F$ and $W = \frac{1}{n}E^T E$. B and W are, respectively, known as the between groups and the within groups covariance matrix.

In the following, we will consider that X is centered. This entails that $\bar{X} = 0_{n\times p}$ (*i.e.* zero matrix).

The aim of the following is to show that it is of high interest to investigate the relationships between X_F on the one hand and $X = X_F + E$ on the other hand. Let us consider a strategy of investigation of the relationships between X and X_F akin to PLS regression analysis. More precisely, we seek a linear combination of X: $t = Xw$ and a linear combination of X_F: $u = X_F\nu$ so as to maximize $cov(Xw, X_F\nu)$ under the constraint $\|w\| = \|\nu\| = 1$, where $cov(.,.)$ stands for the covariance between two variables. We have:

$$cov(Xw, X_F\nu) = \frac{1}{n}w^T X^T X_F\nu = \frac{1}{n}w^T(X_F + E)^T X_F\nu = \frac{1}{n}w^T X_F^T X_F\nu + \frac{1}{n}w^T E^T X_F\nu$$

Since $E^T X_F = 0$, it follows that:

$$cov(Xw, X_F\nu) = \frac{1}{n} w^T X_F^T X_F \nu = w^T B\nu$$

Since B is symmetric semi-definite positive, by applying Cauchy-Schwartz' inequality, we have:

$$w^T B\nu \leq \sqrt{w^T Bw}\sqrt{\nu^T B\nu}$$

Moreover, we know that the equality holds if and only if w and ν are collinear. This means that $w = \nu$ since both these vectors are assumed to be of unit length. The implication of these properties is that the maximum of $cov(Xw, X_F\nu)$ over w and ν (with $||w|| = ||\nu|| = 1$) is equal to the maximum of $w^T Bw$.

It follows that this maximum is achieved for w, eigenvector of B associated with the largest eigenvalue, λ. We also have the following property:

$$\lambda = \frac{1}{n} w^T X_F^T X_F w = \frac{1}{n} w^T (P_F X)^T (P_F X) w = \frac{1}{n} (P_F t)^T (P_F t)$$

Where as stated above $t = Xw$ and P_F is the projector upon the space spanned by the indicator variables associated with the levels of factor F. It follows that λ is equal to the variance of $P_F t$, which is, as a matter of fact, the between-group variance of component t. Thus, the maximization criterion stated above seeks to determine a latent variable, t, that sets the groups centroids associated with this latent variable apart as far as possible. This is precisely the aim of PLS-discriminant analysis (Kemsley, 1996). However, for technical reasons, the usual PLS-DA procedure leads to considering the eigenvectors of a matrix B^* which is, although very close to B, not interpretable in terms of between-groups variation (Barker and Rayens, 2003; Nocairi et al., 2005).

Subsequent components can be determined following the deflation strategy that consists in starting anew the same analysis after deflating the variables in X with respect to the components determined at an earlier stage.

2.2 The case of several factors

The procedure outlined for the case of one factor can be easily extended to the case of several factors. For this purpose, we draw from the approach of analysis followed by ANOVA-PCA (Harrington et al., 2005) and ASCA (Smilde et al., 2005). For the sake of simplicity, we consider the case of two factors G and L. We are interested in investigating the effect of these two factors and their interaction. X which is, as previously, assumed to be centered, can be decomposed as follows:

$$X = X_G + X_L + X_{GL} + E$$

where X_G and X_L are defined in a very similar manner as discussed in the previous section. X_{GL} follows the same pattern since this matrix has the same number of rows and columns as matrix X and the rows corresponding to each combination of the levels of factor G and factor L are replaced by the average values of the variables in $X - X_G - X_L$ restricted to the individuals associated with this combination.

It is known that, in the case of balanced experimental designs, we have the following decomposition of the variance covariance matrix:

$$V = V_G + V_L + V_{GL} + V_E$$

Where V is the total variance covariance matrix, V_G, V_L, V_{GL} and V_E are respectively the variance covariance matrices associated with factors G, L, the interaction and the error.

In order to investigate the effect of factor G, we propose to undergo a PLS regression of X_G (matrix to be predicted) on $X_G + E$ (predictor matrix). The effect of factor L can be investigated by performing a PLS-regression of X_L on $X_L + E$. Likewise, in order to investigate the effect of the interaction, we propose to perform a PLS regression of X_{GL} on $X_{GL} + E$.

The rationale behind this strategy of analysis is the following. When confronting X_G (for instance) to $X_G + E$ there are two configurations: (i) factor G is not significant and, in this case, the variations among the levels of this factor (*i.e.* dataset X_G) will be completely diluted in the noise (*i.e.* dataset E). Therefore, the PLS components will not succeed in distinguishing the signal X_G from the noise E; (ii) factor G is significant and, in this case, the PLS components will successfully set apart the levels of this factor.

An additional advantage of using the strategy of analysis proposed herein is that, by performing PLS regression, we are led to a wide range of tools to assess the significance of the factors and depict the outcomes graphically. Among these tools, we single out the PLS latent components (scores), the Root Mean Square Error of Prediction (RMSEP), the criterion Q^2 and the Variables Importance in the Projection (VIP). We refer to the book by Tenenhaus (1998) for a comprehensive exposition of these tools. Furthermore, one can perform an ANOVA on each PLS component in order to assess its ability to significantly discriminate the levels of the factor under study.

2.3 Comparison of methods

By way of comparing methods, we recall that both ASCA and ANOVA-PCA are based on the same decomposition of X as in AoV-PLS, namely:

$$X = X_G + X_L + X_{GL} + E$$

In ASCA it is advocated performing a PCA on X_G in order to assess the significance of the effect of factor G. The rationale behind this strategy of analysis is to seek components that recover the variations among the levels of factor G. Formally, we seek, in a first step, a component $t = X_G w$ such that $var(t)$ is maximized, where $var(.)$ stands for the variance. However, as pointed out in the literature (Vis et al., 2007; Zwanenburg et al., 2011), with this approach we are unable to assess the significance of a factor since the variation between the factor levels is maximized but the within group variations are missing. Permutation tests were proposed to overcome this problem (Vis et al., 2007; Zwanenburg et al., 2011).

For the assessment of the effect of factor G, ANOVA-PCA amounts to performing PCA of $X_G + E$. The rationale behind this strategy of analysis is that if factor G is significant then it is likely to overcome the noise and emerge on the first principal components of PCA performed on $X_G + E$. However, in practice it may occur that, although the factor G is significant, the dataset X_G may be so overwhelmed by the noise that it does not show up on the first components (Climaco-Pinto et al., 2009). To counteract this problem, Climaco-Pinto et al. (2009) proposed a strategy of analysis which consists in progressively reducing the impact of the noise. These authors also proposed a permutation test to assess the significance of the effect of the factor under consideration.

Formally, in ANOVA-PCA, we seek, in a first step a component $u = (X_G + E)\nu$ with $||u|| = 1$, such that $var(u)$ is maximized.

In AoV-PLS, we stated above that we seek to maximize $cov(X_G w, (X_G + E)\nu)$. We have:

$$cov^2(X_G w, (X_G + E)\nu) = var(X_G w) \times var((X_G + E)\nu) \times cor^2(X_G w, (X_G + E)\nu)$$

Clearly, the first term, $var(X_G w)$, is the same term as in ASCA. The second term, $var((X_G+E)\nu)$ is the same term that is maximized in ANOVA-PCA. The term $cor^2(X_G w, (X_G+E)\nu)$ ensures that the directions outlined by $X_G w$ and $(X_G + E)\nu$ agree with each other as much as possible. Therefore, AoV-PLS realizes a compromise between these two methods of analysis.

2.4 Metabolomics dataset

We illustrate AoV-PLS and compare its outcomes to those of ASCA and ANOVA-PCA on the basis of metabolomics data pertaining to a nutritional experiment in a rodent programming model (Agnoux et al., 2014). The global aim of the experiment is to assess the impact of maternal nutrition during both gestation and lactation periods (foetal and post-natal nutrition) on the metabolic status of the suckled offsprings and later at adulthood. The experimental design involved two factors related to maternal feeding patterns:
– Factor G ("gestation"): rat pups were born from dams submitted to a protein-restricted diet (level: R) or from control dams (level: C).
– Factor L ("lactation"): rat pups were weaned by dams submitted to protein-restricted diet (R) or by control dams (C).
This results in four groups of rat pups (labeled with a first letter related to gestation and a second latter to lactation):

-RC: rat pups born from protein-restricted dams and weaned by control dams (51 rat pups).

-RR: rat pups born and weaned by protein-restricted dams (55 rat pups).

-CR: rat pups born from control dams and weaned by protein-restricted dams (26 rat pups).

-CC: rat pups born and weaned by control dams (56 rat pups).

All in all, we dispose of 188 experimental unit (rat pups) on which 1031 metabolomics variables were measured.

AoV-PLS was run on these data by considering successively the factor gestation denoted by G, the factor lactation denoted by L and the interaction denoted by GL.

For each case, we determine in a first step, the AoV-PLS components that turn out to be significant. The relevance of theses components was assessed by means of the mean squared errors of prediction (MSEP) based on a leave-one-out cross-validation procedure. We also performed a one-way ANOVA on each latent component considering the factor under study as a source of variation. Once the appropriate number of AoV-PLS components was selected, these components were subjected to a Fisher's linear discriminant analysis (LDA).

3 Results

3.1 Factor "gestation"

In order to assess the impact of factor G, we performed a PLS regression of X_G on $X_G + E$. Table 1 shows the percentage of total variance in X_G and $X_G + E$ explained by the first ten PLS components. We also show in this table the p-values associated with a one way ANOVA performed on each PLS component.

Table 1: The percentages of total variance in X_G and $X_G + E$ explained by the ten first AoV-PLS components. P-values associated with one way-ANOVA performed on each component.

	percentage of explained variation		p-values
	X_G	$X_G + E$	Factor G
comp 1	17.83	10.70	0.00
comp 2	18.61	9.67	0.00
comp 3	19.01	4.66	0.00
comp 4	15.00	2.77	0.00
comp 5	10.54	2.57	0.00
comp 6	6.79	2.31	0.00
comp 7	3.57	2.96	0.01
comp 8	1.62	5.08	0.08
comp 9	1.84	2.57	0.06
comp 10	1.37	2.22	0.11

Not surprisingly considering the large number of variables involved in this case study, the total variance recovered by the successive PLS components is not very large. Yet, the first seven PLS components seem to significantly discriminate the two levels of factor G (significance level $\alpha = 5\%$).

These findings are corroborated by the evolution of the mean squared error of prediction (MSEP) as a function of the number of PLS components introduced in the model

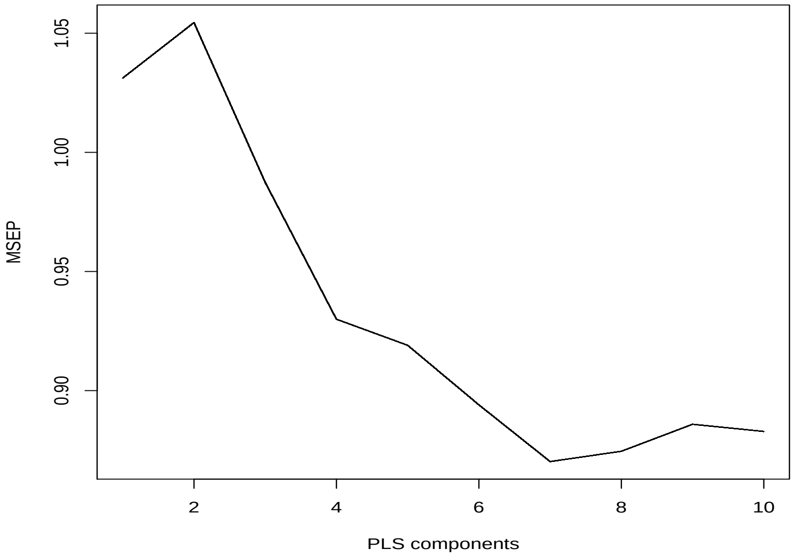

Figure 1: MSEP associated with factor “gestation”

(figure 1). As stated above, the MSEP values were obtained from a leave-one-out cross-validation procedure. It can be seen in figure 1 that the MSEP curve steadily decreases from component 2 down to component 7 and starts slightly to increase. The scores associated with the first two PLS components are depicted in figure 2 where each individual in $X_G + E$ is labeled by the group to which it belongs (R for born from “restricted” dams and C for born from “control” dams). This figure shows a fair separation of the two groups along the first main diagonal.

From the MSEP curve and the p-values associated with the one-way ANOVA performed on the PLS components, we decided to retain seven components. These components were subjected to Fisher’s LDA considering factor G as the group variable. This resulted in a single canonical variate whose values are depicted as box plots for both the groups of rat pups (*i.e.* pups born from restricted (R) dams and from control (C) dams) (figure 3). The two groups are clearly discriminated. Moreover, the p-value of the one-way ANOVA performed on this canonical variate led to a significant discrimination (p-value < 0.001).

By way of comparing methods, we performed ANOVA-PCA and ASCA on the same data. Table 2 shows the percentage of total variance in X_G and $Z_G = X_G + E$ recovered by the components derived from these two methods. It also shows the p-values associated with a one-way ANOVA performed on each component, considering G as a factor.

For ANOVA-PCA, it turns out that none of the first ten components is related to X_G since, on the one hand, the total variances in X_G recovered by these components are very small and, on the other hand, the p-values do not indicate that the two groups are

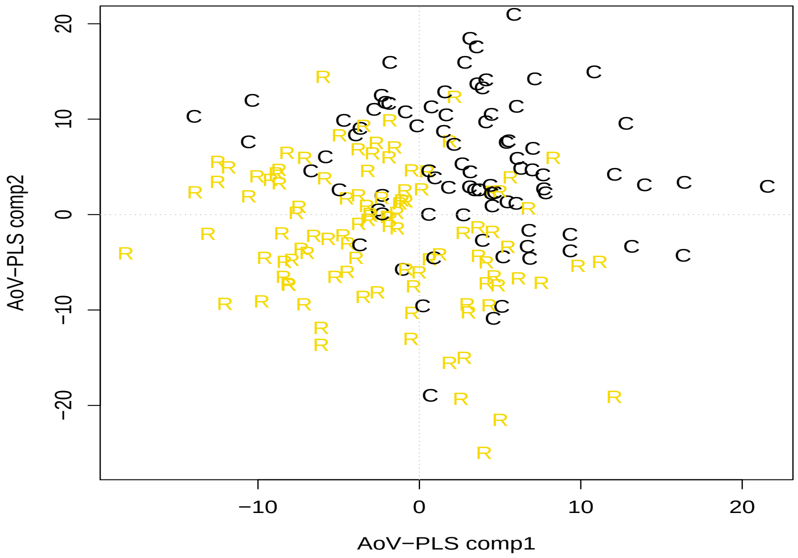

Figure 2: AoV-PLS on "gestation" factor. Representation of the individuals on the first two PLS components. Each individual is labeled by the group it belongs to (R for 'protein-restricted' and C for 'control'.)

significantly discriminated.

Regarding ASCA, not surprisingly the PCA performed on X_G led to a single component that explained 100% of the variation in X_G. This is because X_G contains only two different rows, each repeated as many times as there are individuals in the associated group (R or C). The rows of $X_G + E$ were superimposed by projection on this component (Zwanenburg et al., 2011) and the scores thus obtained were depicted in figure 4 as box plots associated to the two groups (R and C). It can be seen that the two groups of individuals are not clearly discriminated. However, one-way ANOVA performed on the first ASCA component considering G as a factor led to a significant discrimination (p-value < 0.001).

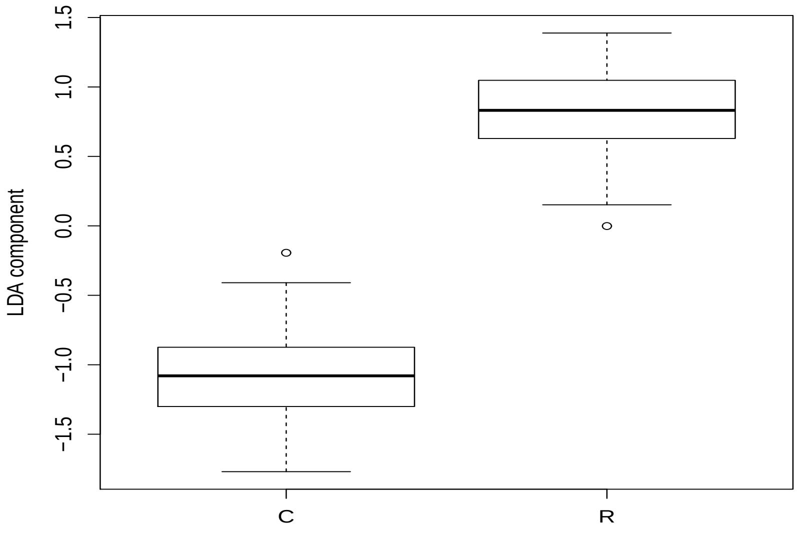

Figure 3: Factor "gestation". Box plot of the LDA canonical variate obtained by performing LDA on the first seven AoV-PLS components. Group C contains individuals born from control dams (CC or CR) and group R contains individuals born from protein-restricted dams (RC or RR).

3.2 Factor "lactation"

We proceeded in a very similar way for factor L as for factor G. Table 3 shows the percentages of total variances in X_L and $X_L + E$ explained by the first ten components. It also shows the p-values associated with a one-way ANOVA performed on each component. It turns out that the first six components seem to be relevant for discriminating the two levels of factor L.

The scores associated with the first two AoV-PLS components are depicted in figure 5 where each individual in $X_L + E$ is labeled by the group to which it belongs (R for suckled by protein-restricted dams and C for nursed by control dams). This figure shows a good separation of the two groups along the first main diagonal.

The one way ANOVAs performed on each component (table 3) together with the evolution of the MSEP as a function of the number of components introduced in the model indicate to retain 5 components (figure 6). These five components where subjected to Fisher's LDA and resulted in a single canonical variate. It can be seen in figure 7 that this canonical variate perfectly separates the two groups of factor L.

ANOVA-PCA performed on these data showed a much better performance as when it was applied to the data concerning factor G. From table 4, we can see that among the first ten PCA components, components 2, 3, 4, 5, 6 and 10 seem to significantly discriminate the two groups of factor L. The fourth component is the component that

Table 2: The percentages of total variance in X_G and $X_G + E$ explained by the ten first ANOVA-PCA and ASCA components. P-values associated with one way-ANOVA performed on each component.

	ANOVA-PCA			ASCA		
	X_G	$X_G + E$	p-values	X_G	$X_G + E$	p-values
comp 1	0.52	17.09	0.32	100	0.73	0
comp 2	0.00	7.08	0.98	0	0.10	1
comp 3	1.82	6.77	0.07	0	1.25	1
comp 4	0.60	4.39	0.29	0	0.17	1
comp 5	0.21	3.86	0.53	0	0.54	1
comp 6	1.07	3.09	0.16	0	0.35	1
comp 7	0.36	2.71	0.42	0	0.74	1
comp 8	1.07	2.44	0.16	0	0.84	1
comp 9	0.82	2.32	0.22	0	0.59	1
comp 10	0.07	1.92	0.72	0	0.50	1

explains the most variation in X_L (about 15%). This should be contrasted with the outcomes from AoV-PLS where the first component alone explains up to 47.69 % of the variation in X_L. Up to 20 ANOVA-PCA components were retained and submitted to LDA. Figure 8 shows the box plot associated with the canonical variate derived from this analysis. We can see that this canonical variate operates a fair discrimination of the two groups but this discrimination is less marked than that obtained by means of AoV-PLS (figure 7).

Regarding the outcomes of ASCA, PCA performed on X_L naturally led to a single component. The data from $Z_L = X_L + E$ were superimposed on this component and the scores thus obtained were depicted as box plots associated with the two groups (R and C). As shown in figure 9, the two groups are not at all discriminated. However, one-way ANOVA performed on the first ASCA component considering L as a factor led to a significant discrimination (p-value < 0.001).

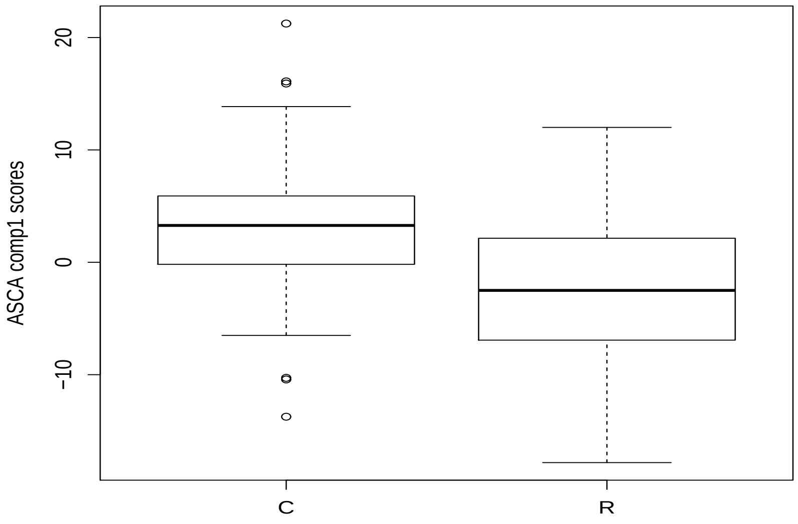

Figure 4: Factor "gestation". Box plot of the first ASCA component for groups C (Control) and R (Restricted). Group C contains individuals born from control dams (CC or CR) and group R contains individuals born from protein-restricted dams (RC or RR).

3.3 Interaction

The strategy of analysis advocated by AoV-PLS was applied to investigate the interactions of factors G and L. We recall that, in this case, this amounts to performing a PLS regression of dataset X_{GL} on $Z_{GL} = X_{GL} + E$. Table 5 shows the percentages of total variances in X_{GL} and Z_{GL} recovered by the first ten components. It also shows the p-values associated with a one-way ANOVA performed on each component considering the interaction GL as a factor. It turns out that the first nine components seem to be relevant for discriminating the four groups RC, RR, CR and CC.

In a subsequent stage, the first nine PLS components were submitted to a Fisher's discriminant analysis which resulted in three canonical variates. In figure 10, we depict the individuals in Z_{GL} on the basis of the first two canonical variates. It appears that the two groups that were not submitted to a nutritional switch, namely CC and RR are fairly separated from the two groups that were submitted to a nutritional transition, namely RC and CR.

By way of comparing methods, we performed ANOVA-PCA and ASCA on the same data. Table 6 shows the percentages of total variances in X_{GL} and Z_{GL} recovered by the first ten components associated to each method together with the p-values associated with a one-way ANOVA performed on each component. It is clear that the outcomes from ANOVA-PCA are not relevant or would necessitate to investigate more than ten

Table 3: The percentages of total variance in X_L and $X_L + E$ explained by the first ten AoV-PLS components. P-values associated with the one way ANOVA performed on each component.

	percentage of explained variation		p-values
	X_L	$X_L + E$	Factor L
comp 1	47.69	8.05	0.00
comp 2	13.13	13.01	0.00
comp 3	14.25	5.01	0.00
comp 4	8.91	3.57	0.00
comp 5	5.91	2.42	0.00
comp 6	2.96	2.65	0.02
comp 7	1.37	4.47	0.11
comp 8	1.38	3.49	0.11
comp 9	0.88	2.71	0.20
comp 10	0.77	1.80	0.23

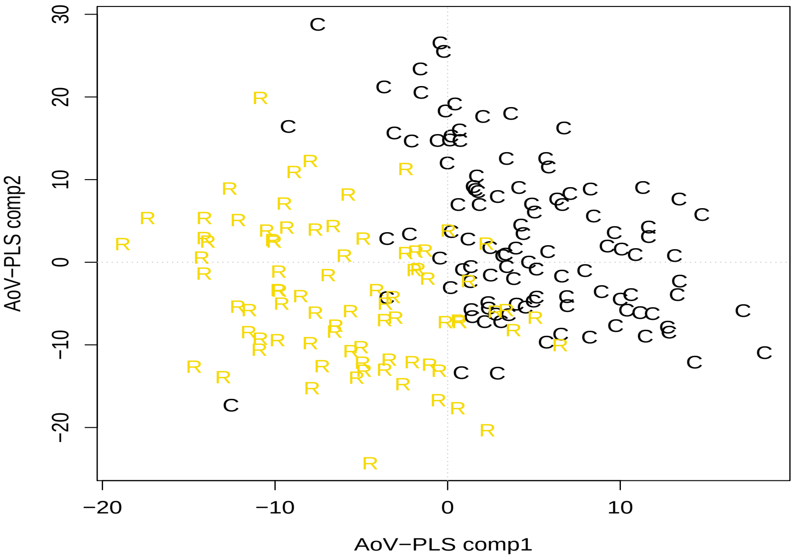

Figure 5: AoV-PLS on "lactation" factor. Representation of the individuals on the first two PLS components. Each individual is labeled by the group it belongs to (R for suckled by protein-restricted dams and C for nursed by control dams)

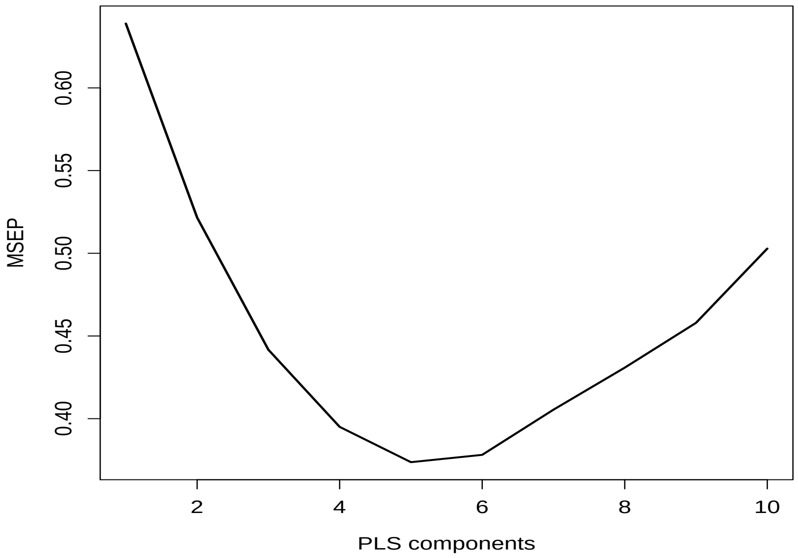

Figure 6: MSEP associated with AoV-PLS (factor “lactation”)

components. ASCA naturally led to three significant components. The data in Z_{GL} were superimposed on these components. The scores thus obtained were subjected to Fisher's LDA. We depict in figure 11 the values of the three canonical variates obtained by Fisher's LDA as box plots for each group of rat pups. We can see that no separation of the groups is achieved by any of the Fisher's LDA canonical variates. The p-values associated with a one-way ANOVA performed on each of the three canonical variates considering the interaction GL as a factor, indicated that only the first canonical variate was significant.

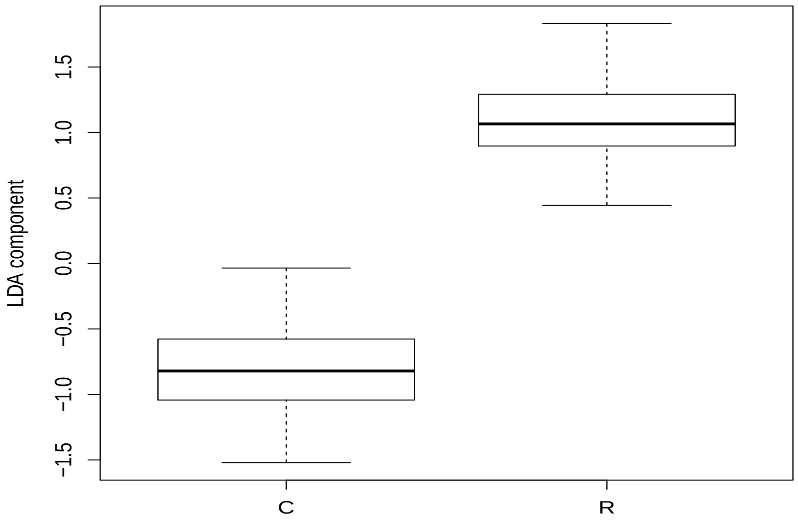

Figure 7: Factor “lactation”. Box plot of the LDA canonical variate associated with the first five AoV-PLS components. Group C contains individual nursed by control dams and group R contains individuals suckled by protein-restricted dams.

4 Conclusion

For the analysis of multivariate data depending on several factors, we proposed a method of analysis to investigate the effect of the various factors and their interactions. Similarly to ANOVA-PCA and ASCA, AoV-PLS starts from the well known ANOVA model which decomposes a data matrix into a sum of matrices that reflect the effects of the various factors and their interactions. However, whereas ASCA and ANOVA-PCA are based on a PCA of each factor effect or the factor effect augmented by the matrix of residuals, we advocate performing a PLS regression on each factor effect upon this factor effect augmented with the residuals. The rationale behind this strategy is very clear since it consists in assessing whether the factor effect overpowers the noise (significant factor) or whether it is diffused in the noise (non significant factor). By using PLS regression, we access to a wide range of tools that make it possible to undertake an exploratory study (graphical displays, indicators...) and a confirmatory study (regression/discrimination, cross-validation...). We have emphasized the link between AoV-PLS and ANOVA-PCA, on the one hand, and ASCA, on the other hand. In particular, we showed that AoV-PLS appears as a compromise between these two methods since it aims at recovering the variation in each effect matrix (similarly to ASCA) and in the effect matrix augmented with the noise matrix (similarly to ANOVA-PCA). The comparison of these three methods on the basis of the metabolomics data showed a much better performance of AoV-PLS

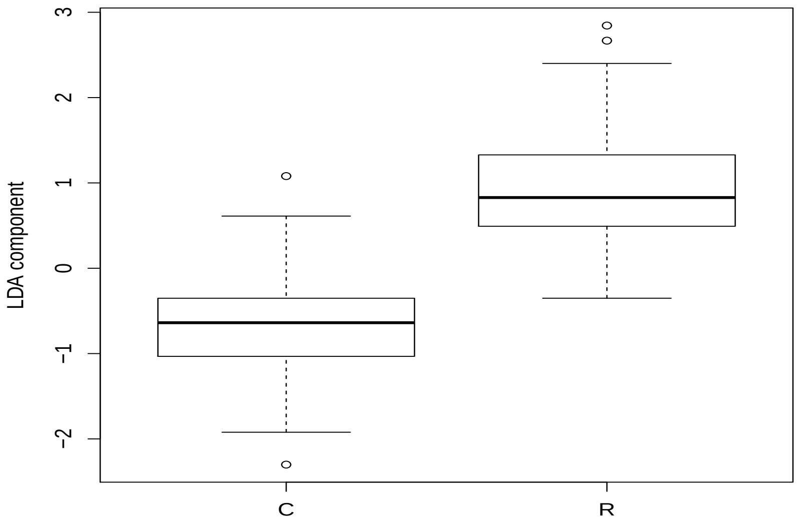

Figure 8: Factor “lactation”. Box plot of the LDA canonical variate associated with the first 20 ANOVA-PCA components. Group C contains individual nursed by control dams and group R contains individuals suckled by protein-restricted dams.

than the other two alternative methods.

As stated above, AoV-PLS was applied to a metabolomics dataset which involved a relatively large number of variables. In the course of this case study, we showed how AoV-PLS could be used in conjunction with Fisher’s LDA. The former method makes it possible to exhibit directions of interest for a discrimination purpose and the latter method focuses on discriminating the groups on the basis of a small number of canonical variates.

Within the context of discriminant analysis using a partial least squares strategy, several variants have been proposed (Sabatier et al., 2003; Lombardo et al., 2012). It is clear that these variants could be applied instead of PLS-DA as presented herein. The interest of using these alternative strategies will be investigated in a future work.

An important aspect that we have overlooked concerns the impact of the experimental design on the outcomes of the strategy of analysis proposed herein. This aspect was also overlooked when introducing ASCA and ANOVA-PCA. We believe that this aspect deserves more attention.

As a final remark, we point out that there exists a method of analysis called ANOVA-PLS which was proposed by Thissen et al. (2009). This method is different from the approach proposed herein since the ANOVA decomposition of matrix X is used to predict other response variables by means of PLS-regression whereas in our approach PLS-regression is used to assess the significance of the various factors at hand and no addi-

Table 4: The percentages of total variance in X_L and $X_L + E$ explained by the ten first ANOVA-PCA and ASCA components. P-values associated with one way-ANOVA performed on each component.

	ANOVA-PCA			ASCA		
	X_L	$X_L + E$	p-values	X_L	$X_L + E$	p-values
comp 1	1.13	16.84	0.15	100	2.61	0
comp 2	9.15	7.06	0.00	0	0.13	1
comp 3	2.25	6.91	0.04	0	0.5	1
comp 4	15.75	4.78	0.00	0	0.2	1
comp 5	3.28	4.00	0.01	0	0.31	1
comp 6	5.57	3.27	0.00	0	0.43	1
comp 7	0.75	2.68	0.24	0	0.48	1
comp 8	1.08	2.40	0.16	0	0.68	1
comp 9	1.38	2.30	0.11	0	0.31	1
comp 10	6.27	1.98	0.00	0	0.55	1

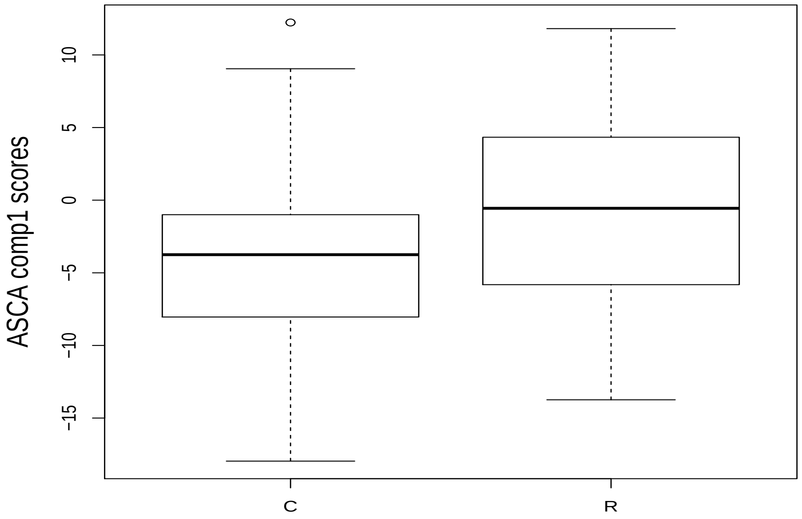

Figure 9: Factor "Lactation". Box plot of ASCA component for groups C (Control) and R (Restricted). Group C contains individual nursed by control dams and group R contains individuals suckled by protein-restricted dams.

Table 5: The percentages of total variance in X_{GL} and $X_{GL} + E$ explained by the first ten AoV-PLS components. P-values associated with the one way ANOVA performed on each component.

	percentage of explained variation		p-values
	X_{GL}	$X_{GL} + E$	interaction GL
comp 1	16.06	11.47	0.00
comp 2	16.32	9.11	0.00
comp 3	15.34	4.42	0.00
comp 4	9.22	3.57	0.00
comp 5	7.33	2.53	0.00
comp 6	4.12	3.64	0.00
comp 7	2.87	4.29	0.01
comp 8	2.25	2.77	0.03
comp 9	2.20	2.05	0.02
comp 10	1.24	2.10	0.10

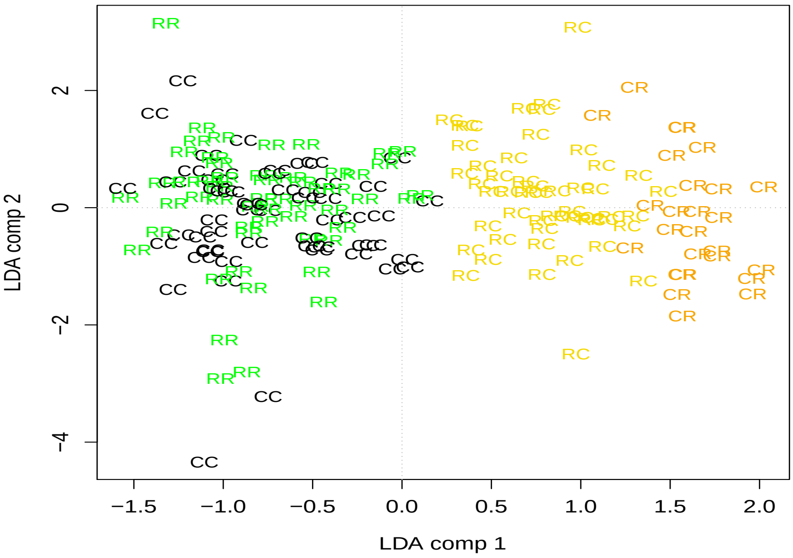

Figure 10: Interaction, representation of the individuals on the first two canonical variates of LDA performed on the first nine AoV-PLS components.

Table 6: The percentages of total variance in X_{GL} and $X_{GL}+E$ explained by the ten first ANOVA-PCA and ASCA components. P-values associated with one way-ANOVA performed on each component.

	ANOVA-PCA			ASCA		
	X_{GL}	$X_{GL}+E$	p-values	X_{GL}	$X_{GL}+E$	p-values
comp 1	0.79	17.10	0.21	87.46	0.92	0
comp 2	0.61	7.09	0.26	10.00	0.11	0
comp 3	1.09	6.73	0.12	2.54	0.03	0
comp 4	1.60	4.43	0.08	0.00	0.05	1
comp 5	0.83	3.88	0.21	0.00	1.13	1
comp 6	0.93	3.09	0.20	0.00	0.84	1
comp 7	0.65	2.72	0.23	0.00	0.25	1
comp 8	0.80	2.43	0.20	0.00	0.42	1
comp 9	0.89	2.32	0.19	0.00	0.48	1
comp 10	2.81	1.97	0.02	0.00	0.72	1

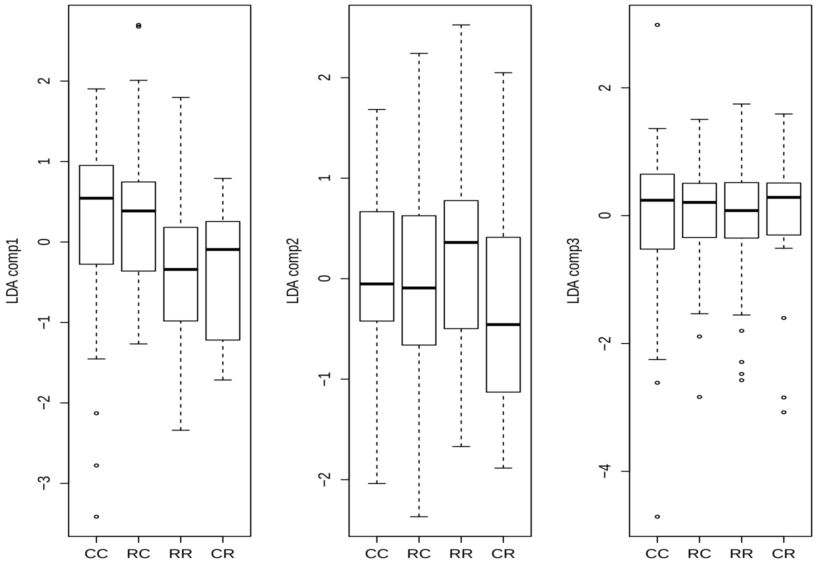

Figure 11: Interaction. Box plots of the three LDA canonical variates associated with the three ASCA components.

tional (response) variables are involved.

References

Agnoux, A. M., Antignac, J.-P., Simard, G., Poupeau, G., Darmaun, D., Parnet, P., and Alexandre-Gouabau, M.-C. (2014). Time window-dependent effect of perinatal maternal protein restriction on insulin sensitivity and energy substrate oxidation in adult male offspring. *American Journal of Physiology - Regulatory, Integrative and Comparative Physiology*, 307(2):R184–R197.

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

Climaco-Pinto, R., Barros, A. S., Locquet, N., Schmidtke, L., and Rutledge, D. N. (2009). Improving the detection of significant factors using ANOVA-PCA by selective reduction of residual variability. *Analytica Chimica Acta*, 653(2):131–142.

Harrington, P. d. B., Vieira, N. E., Chen, P., Espinoza, J., Nien, J. K., Romero, R., and Yergey, A. L. (2006). Proteomic analysis of amniotic fluids using analysis of variance-principal component analysis and fuzzy rule-building expert systems applied to matrix-assisted laser desorption/ionization mass spectrometry. *Chemometrics and Intelligent Laboratory Systems*, 82(1-2):283–293.

Harrington, P. d. B., Vieira, N. E., Espinoza, J., Nien, J. K., Romero, R., and Yergey, A. L. (2005). Analysis of variance-principal component analysis: A soft tool for proteomic discovery. *Analytica Chimica Acta*, 544(1-2):118–127.

Kemsley, E. K. (1996). Discriminant analysis of high-dimensional data: a comparison of principal components analysis and partial least squares data reduction methods. *Chemometrics and Intelligent Laboratory Systems*, 33(1):47–61.

Lombardo, R., Durand, J.-F., and P. Leone, A. (2012). Multivariate additive PLS Spline boosting in agro-chemistry studies. *Current Analytical Chemistry*, 8(2):236–253.

McLachlan, G. J. (2004). *Discriminant Analysis and Statistical Pattern Recognition*. Wiley Series in Probability and Statistics.

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

Sabatier, R., Vivien, M., and Amenta, P. (2003). Two approaches for discriminant partial least squares. In Schader, M., Gaul, W., and Vichi, M., editors, *Between Data Science and Applied Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 100–108. Springer Berlin Heidelberg.

Sarembaud, J., Pinto, R., Rutledge, D., and Feinberg, M. (2007). Application of the ANOVA-PCA method to stability studies of reference materials. *Analytica Chimica Acta*, 603(2):147–154.

Smilde, A. K., Jansen, J. J., Hoefsloot, H., Lamers, R.-J., Greef, J., and Timmerman, M. (2005). ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data. *Bioinformatics*, 21(13):3043–3048.

Tenenhaus, M. (1998). *La régression PLS, théorie et pratique*. Technip. Paris.

Thissen, U., Wopereis, S., van den Berg, S., Bobeldijk, I., Kleemann, R., Kooistra, T., Willems van Dijk, K., van Ommen, B., and Smilde, A. (2009). Improving the analysis of designed studies by combining statistical modelling with study design information. *BMC Bioinformatics*, 10(1):52.

Vis, D., Westerhuis, J., Smilde, A. K., and van der Greef, J. (2007). Statistical validation of megavariate effects in ASCA. *BMC Bioinformatics*, 8(1):322.

Zwanenburg, G., Hoefsloot, H. C. J., Westerhuis, J. A., Jansen, J. J., and Smilde, A. K. (2011). ANOVA-principal component analysis and ANOVA-simultaneous component analysis: a comparison. *Journal of Chemometrics*, 25(10):561–567.