Confidence ellipsoids for ASCA models based on multivariate regression theory

Liland, Kristian Hovde[1,2]; Smilde, Age[3]; Marini, Federico[4] and Næs, Tormod[1]

[1]Nofima, Ås, Norway

[2]Norwegian University of Life Sciences, Ås, Norway

[3]Univ. Amsterdam, The Netherlands

[4]Univ. Roma, La Sapienza, Italy

Correspondence: kristian.liland@nofima.no

Abstract

In Analysis of variance Simultaneous Component Analysis (ASCA), permutation testing is the standard way of assessing uncertainty of effect level estimates. This article introduces an analytical solution to the assessment of uncertainty through classical multivariate regression theory. We visualize the uncertainty as ellipsoids, contrasting these to data ellipsoids. This is further extended to multiple testing of effect level differences. Confirmatory and intuitive results are observed when applying the theory to previously published data and simulations.

Keywords: ASCA; confidence; ellipsoids; CLD

1. Introduction

In modern science, analysing designed experiments with a multivariate output has become a major issue[1–3]. The reason for this is the easy availability of instruments and measurement techniques that provide large amounts of data. Standard and well-established methods from classical statistics (multivariate analysis of variance, MANOVA, see e.g. [4]) can be useful in this context for assessing significance, but they provide little additional information for interpretation and they break down with high-dimensional data. Since modern instruments in most cases provide highly collinear data and one of the main interests lies in understanding how the different design factors influence the multivariate response, a number of methods have been developed for gaining improved insight.

The simplest of these methods is PC-ANOVA[5,6] based on just using principal components analysis (PCA) of the output vectors and then relating the first few principal components to the design variables. This is a simple approach and has the advantage that it can use all the tools available for analysis of variance (ANOVA), like multiple comparisons, random effects, split plot structures etc. A possible drawback of the method is that in cases where the different design factors generate very different multivariate structures, the joint PCA solution will contain very many principal components. Also, PCA is blind to the design and so may find principal components that are of less relevance to the experiment.

A refinement of the PC-ANOVA is the 50/50 ANOVA put forward by [7]. In addition to formal tests of significance, the method also provides an explicit way of selecting the

number of principal component and which design factors to finally incorporate in the model. However, the method suffers from the same drawbacks as the PC-ANOVA.

A third possibility is the ASCA method[8] which reverses the use of PCA and ANOVA as compared to the PC-ANOVA above. First, a regular ANOVA is used for each of the output variables and the coefficients for each variable and design factor are calculated using the standard restriction of setting the sum of the parameters for each design factors equal to 0. Then the coefficients for each factor are submitted to a separate PCA for each factor. The method has been extended for more detailed analysis of the interactions in PARAFASCA[9] and for unbalanced data[10]. Comparisons of ASCA and PCA-ANOVA can be found in [11,12,1]. ASCA fits into a general framework for high-dimensional fixed effect ANOVA which also contains other methods [13].

Whereas the ASCA method has been developed as a high-dimensional fixed effect ANOVA visualization tool, it would be useful to also include statistical inference. A first attempt was made by including permutation testing[14] and by back-projecting the residuals on the loadings of the ASCA model[12]. None of these approaches, however, give a full and rigorous inference of the estimated multivariate effects.

The focus of the present paper is to develop a more general statistical inference procedure in the form of confidence ellipsoids with corresponding visualization that can be used for balanced designs. The method is based on standard distribution theory from multivariate regression. Focus here will be on establishing a closer link

between standard linear model theory[4] and ASCA[8], with the hope that it is easier to generalise than the methods proposed earlier[12,14]. The method will be explained and illustrated with three examples: one from sensory science, one from spectroscopy and one simulation. This new approach gives a visual presentation of the differences between the levels of each factor in the design.

Notation

a – a scalar (lower-case, italic letter)

$\mathbf{X}$ – a matrix (capital, bold letter)

$\mathbf{x}_{i,*}$, $\mathbf{x}_{*,j}$ – a single row or column in $\mathbf{X}$, respectively

$\mathbf{X}_g$ – a subset of columns/rows in $\mathbf{X}$ corresponding to a design factor

$\mathbf{X}^t$ – the transpose of $\mathbf{X}$

2. Standard ANOVA and the need for multivariate extensions

Analysis of variance (ANOVA) is one of the most used and well established methods in statistics. It was developed for determining the effect of various factors, typically varied according to an experimental design, on a response variable. The model used is

$$\mathbf{y} = \mathbf{1}b_0 + \mathbf{X}_1\mathbf{b}_1 + \ldots + \mathbf{X}_g\mathbf{b}_g + \ldots + \mathbf{X}_G\mathbf{b}_G + \mathbf{e} \qquad (1)$$

where $\mathbf{y}$ (nxp) is the vector of responses, the $\mathbf{X}_g$ (nxq_g) represents the design matrix for design factor g, $\mathbf{b}_g$ (q_gxp) is the vector of model coefficients for design

factor g, and **e** is the random error, i.e. the response variation not explained by the linear model. Typically, one is interested in determining which of the factors have a significant effect on the response and also to determine how large the effect is. Of special importance is to estimate the differences of effect sizes between the different levels/values for the significant factors. Is for instance a detected significant effect due to the differences between level 1 and level 2 only or between all levels?

For determining significance of a factor, one will typically use F-tests, with corresponding p-values, that compare a measure (a sum of squares) for the differences between the levels with a measure of the random error. Confidence intervals for important aspects like effect levels also come out naturally from the methodology. A number of extensions of the basic methodology that incorporate random effects and different error structures have been developed.

In modern science, it has become more and more common to measure many Y-variables instead of only one, ending up with a model of the form

$$\mathbf{Y} = \mathbf{1b}_0 + \mathbf{X}_1\mathbf{B}_1 + \ldots + \mathbf{X}_g\mathbf{B}_g + \ldots + \mathbf{X}_G\mathbf{B}_G + \mathbf{E} \qquad (2)$$

where **Y** (nxp) is the matrix of all responses, the $\mathbf{X}_m$ (nxq_m) is the same as above, $\mathbf{B}_g$ (q_gxp) is the matrix of model coefficients for design factor g, and **E** is the error. Here, q_g is the number of effect levels minus one, as in the univariate case. The

rows of **E** have mean equal to 0 and covariance matrix equal to **Σ**. Both main effect factors and interactions are allowed.

One can of course analyse each **Y**-column separately, but that is cumbersome and time-consuming and also in many cases rather meaningless. The estimates of the factor effects are the same with this strategy, but the inference will not take into account the remaining Y-columns. In for instance spectroscopy, each single wavelength is of minor interest. In addition, the correlation structure between all variable is lost if analysed individually. The standard multivariate ANOVA (MANOVA) is an early attempt for solving this, but due to high collinearity among responses and often few samples, this method can often not be used. ASCA is a method which attempts to handle these problems.

3. The standard ASCA method

Conceptually, ASCA is an exploratory subspace analysis of multivariate least squares means (LS-means). The basis is a fixed, balanced multivariate ANOVA (model (2) above) with no complicating additions like repeated measures, covariates or random effects.

The first step of ASCA is to calculate the LS-means in the traditional way for each of the design factors (here represented by $\mathbf{X}_g$) following Equation 2. The sum-to-zero coding (column centred dummy coding of design with last column of each level removed) of the design matrices is in practice used in order to obtain unique

estimates of the coefficients, but this is not needed for the uniqueness of the LS-means. The LS-means are computed for one factor at the time using estimated coefficients $\mathbf{1}\mathbf{b}_0 + \mathbf{X}_g\widehat{\mathbf{B}}_g$.

The next step of ASCA is to centre the LS means for each factor and perform a PCA on the results. Note that the intercept part of the LS means is then essentially eliminated. Since the PCA is performed on LS-means without individual variation, the rank is limited to the number of levels minus one for the main effects and corresponding products of factor ranks for interactions. For instance, a $\mathbf{X}_g\widehat{\mathbf{B}}_g$ based on a three level factor will be completely exhausted after two principal components have been extracted. For factors with two, three, and possibly four (3D plot) levels, this means that score plots will only show the LS-means with a change of basis, not a truncated space. Note that the explained variances reported in the PCA plots below are related only to the actual factor effect matrix and not to the explained variance of the response.

Data ellipses are commonly used in ASCA to emphasize variation patterns in the factor levels in score plots. However, these ellipses are so called data ellipses[15], which only capture the observed variation in the displayed score dimensions. A $1-\alpha$ data ellipsoid is created using a unit circle, the mean $\bar{\mathbf{M}}_{g,r}$ and sample covariance matrices $\mathbf{S}_{g,r}$ of the r-th level of the g-th design factor, and a scaling constant $c = p(n-1)/(n-p)F_{1-\alpha,p,n-p}$. The unit circle is scaled and rotated by c and ${\mathbf{S}_{g,r}}^{1/2}$ (the Cholesky decomposition of $\mathbf{S}_{g,r}$) and translated to be centred in $\bar{\mathbf{M}}_{g,r}$.

Another attempt to assessing uncertainty that is put forward is based on permutation testing[14]. One of the drawbacks with this method is that, while permutation test provide "exact" estimates for main effects, they are only approximate for interactions, and in the latter case, their implementation is not straightforward.

The suggested ellipses in this manuscript will be based on the underlying multivariate ANOVA model, thus capturing the uncertainty in the modelling. In contrast to the data ellipsoids, the model ellipsoids will grow and shrink with the number of samples and the covariance of the design factors.

4. New method based on confidence ellipsoids

In this paper, a new approach is proposed which is based on using standard results from multivariate regression combined with T^2 confidence ellipsoids to assess uncertainty of the points in the ASCA plots. In more detail, the confidence ellipses are first developed for the whole vector of predicted responses before we present corresponding results when the ellipses are projected down onto the most dominating directions in the space.

As with all ANOVA based methods, assuming balanced data makes the theory simpler. This is also the case for ASCA, and as such this assumption underlies the basic ASCA and its implementation in this work. One of the reasons why imbalance

makes modelling more complex is that the dummy coded factor level matrix tends to give covariance between levels and factors that are otherwise orthogonal. We will therefore here concentrate on balanced data, but the aim is to develop a structure that can be extended to more complex cases.

The standard dummy design matrices for the different factors in ANOVA are singular, and the easiest way of establishing the methodology presented is to work with a reparametrization of full rank. In this paper we choose the "sum-to-zero" parametrization (also known as sum coding or deviation coding[16]) as this ensures orthogonal factors in the design matrix in the balanced case. This implies that one column from each of the design matrices is eliminated and the corresponding rows are set to -1. Note that the choice of parametrisation has no effect on the predicted results. However, a non-centred design would lead to more complex LS-means calculations and different regression coefficients. We refer to the appendix for further detail on the coding problem and how it is solved in this paper.

4.1. Confidence ellipses for LS-means in original space.

We will here first take a general approach where we show how to create confidence ellipses for LS means for a full model. Then we discuss how to use these results for each of the design factors and their interactions separately and for points in PCA reduced space.

4.1.1. Confidence ellipsoids for LS means

The model considered first is the full multivariate regression model:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}\ (n \text{ x } p), \qquad (3)$$

The matrix **E** is assumed normally distributed with mean **0** and covariance matrix equal to **Σ**, **B** is the matrix of regression coefficients and **X** full rank. The first column of **X** is a vector of 1's in order to accommodate the intercept. The covariance matrix $Cov(\mathbf{e}_{i,*}) = \mathbf{\Sigma}$, where $\mathbf{e}_{i,*}$ corresponds to a row in **E**, is here assumed to have full rank for **Σ** and to be independent of the remaining rows in **E**. The regression coefficient matrix can be written as:

$$\mathbf{B} = \begin{pmatrix} b_{11} & b_{12} & & \\ & . & & \\ & & . & \\ b_{q1} & & & b_{qp} \end{pmatrix} = (\mathbf{b}_{*,1}, \ldots, \mathbf{b}_{*,p}),$$ i.e. p responses and q columns in **X**.

Note that the q here is equal to $1 + q_1 + \ldots + q_g + \ldots + q_G$ as defined for equation (1). The covariance matrix between $\hat{\mathbf{b}}_{*,k}$ and $\hat{\mathbf{b}}_{*,l}$ (two columns/responses, in $\widehat{\mathbf{B}}$) is equal to

$$cov(\hat{\mathbf{b}}_{*,k}, \hat{\mathbf{b}}_{*,l}) = E(\hat{\mathbf{b}}_{*,k} - \mathbf{b}_{*,k})(\hat{\mathbf{b}}_{*,l} - \mathbf{b}_{*,l})^t = \sigma_{kl}(\mathbf{X^t X})^{-1} \qquad (4)$$

(see Theorem 6.2.3. in Mardia et al.[4]), where σ_{kl} is an element of the covariance matrix **Σ**. Then, the covariance between the two linear functions $\mathbf{x}^t\hat{\mathbf{b}}_{*,k}$ and $\mathbf{x}^t\hat{\mathbf{b}}_{*,l}$ is equal to

$$E\left(\mathbf{x}^t(\hat{\mathbf{b}}_{*,k}-\mathbf{b}_{*,k})(\hat{\mathbf{b}}_{*,l}-\mathbf{b}_{*,l})^t\mathbf{x}\right)=\mathbf{x}^t\sigma_{kl}(\mathbf{X^tX})^{-1}\mathbf{x}. \qquad (5)$$

If $\mathbf{x}^t$ is the column vector corresponding to row *i* in **X**, i.e. equal to $\mathbf{x}_{i,*}$, the two linear functions $\mathbf{x}^t\hat{\mathbf{b}}_{*,k}$ and $\mathbf{x}^t\hat{\mathbf{b}}_{*,l}$ represent estimates of for row *i* in **XB** for the columns (responses) *k* and *l*.

Equation (4) shows the covariance for two linear functions representing two different responses. Putting this together for all responses, the covariance matrix of the whole vector $\mathbf{x}_{i,*}^t\widehat{\mathbf{B}}$, i.e. all the elements in the row, is then equal to $a_i\mathbf{\Sigma}$ where $a_i=\mathbf{x}'_{i,*}(\mathbf{X'X})^{-1}\mathbf{x}_{i,*}$. Normality and unbiasedness follow as usual in linear functions. The distribution of $\mathbf{x}_{i,*}^t\widehat{\mathbf{B}}/\sqrt{a_i}$ is then multivariate normal with a covariance matrix equal to **Σ**.

According to Theorem 6.2.3 in Mardia et al.[4] $n\widehat{\mathbf{\Sigma}}$ is Wishart distributed, $\boldsymbol{W}_p(\Sigma,n-q)$ where $\widehat{\mathbf{\Sigma}}=\frac{1}{n}\hat{\mathbf{E}}^t\hat{\mathbf{E}}$**.** Theorem 3.5.1 in Mardia et al.[4] then immediately gives us that

$$\tfrac{(n-q)}{n}(\mathbf{x}_{i,*}^t\widehat{\mathbf{B}}/\sqrt{a_i}-\mathbf{x}_{i,*}^t\mathbf{B}/\sqrt{a_i})\hat{\Sigma}^{-1}(\mathbf{x}_{i,*}^t\widehat{\mathbf{B}}/\sqrt{a_i}-\mathbf{x}_{i,*}^t\mathbf{B}/\sqrt{a_i})^t \quad (6)$$

is Hotelling's T^2 distributed with parameters *p* and (*n*-*q*). Alternatively, we can write the ellipsoid function with coefficient $1-\alpha$ for $\mathbf{x}_{i,*}^t\mathbf{B}$ as:

$$\tfrac{(n-q)}{na_i}\left(\mathbf{x}_{i,*}^t\widehat{\mathbf{B}}-\mathbf{x}_{i,*}^t\mathbf{B}\right)\hat{\Sigma}^{-1}\left(\mathbf{x}_{i,*}^t\widehat{\mathbf{B}}-\mathbf{x}_{i,*}^t\mathbf{B}\right)^t=T^2_{1-\alpha,p,n-q} \qquad (7).$$

4.2. Confidence ellipsoids for LS means in projected space

The next step is to extend this to linear functions of rows in **XB**, i.e. $\mathbf{x}_{i,*}^t\mathbf{B}\mathbf{L}^t$ where **L** has dimension (*dxp*). Below we will only consider **L** as linear functions defined by the principal component projects, but the following results are general. In this paper d (the number of principal components) is typically set to either 2 or 3. The covariance of $\mathbf{x}_{i,*}^t\mathbf{B}\mathbf{L}^t/\sqrt{a_i}$ is, by the results above, equal to $\mathbf{L}\boldsymbol{\Sigma}\mathbf{L}^t$. Since $n\widehat{\boldsymbol{\Sigma}}$ is Wishart distributed, the distribution of the transformed covariance $n\mathbf{L}\widehat{\boldsymbol{\Sigma}}\mathbf{L}^t$ matrix is still Wishart distributed (Theorem 3.4.1 in Mardia et al.[4]), i.e. $n\mathbf{L}\widehat{\boldsymbol{\Sigma}}\mathbf{L}^t \sim W_d(\mathbf{L}\boldsymbol{\Sigma}\mathbf{L}^t, n-q)$. The same arguments as above lead to the following confidence ellipsoids for the linear functions $\mathbf{x}_{i,*}^t\mathbf{B}\mathbf{L}^t$:

$$\frac{(n-q)}{na_i}\left(\mathbf{x}_{i,*}^t\widehat{\mathbf{B}}\mathbf{L}^t - \mathbf{x}_{i,*}^t\mathbf{B}\mathbf{L}^t\right)\left(\mathbf{L}\widehat{\boldsymbol{\Sigma}}\mathbf{L}^t\right)^{-1}\left(\mathbf{x}_{i,*}^t\widehat{\mathbf{B}}\mathbf{L}^t - \mathbf{x}_{i,*}^t\mathbf{B}\mathbf{L}^t\right)^t = T_{1-\alpha,d,n-q}^2. \quad (8)$$

In this paper, The **L** will be the loadings matrix corresponding to principal components of the rows in **XB**. We will consider the PCA projection as a fixed linear transform meaning that our results are conditioned on the actual PCA projection.

4.3. Confidence ellipsoids for each experimental factor

If we go back to model (1), the different **X**-blocks represent either main effects matrices or interaction matrices obtained by multiplication as described in the Appendix. The number of columns in $\mathbf{X}_g$ is equal to q_g. For the balanced case each of the blocks are estimated and treated independently. This means that the same theory as described in Section 3.1 and 3.2 above holds for each of the blocks with

the corresponding adjustment of the factor a_i (all values are set to zero in $\mathbf{x}_i$, except those that represent the block) calculated for the block of interest only. In other words, one can calculate confidence ellipsoids for rows in $\mathbf{X}_g\mathbf{B}_g$ using the same equation as presented in equation (8) setting q equal to q_g. The principal components used for projection are calculated for each block separately, which means that the **L** is different for each block.

4.4. Implementation

Drawing of the model ellipsoids follows the same pattern as drawing data ellipsoids. The pooled covariance matrix is exchanged with the scaled, projected model covariance matrix $\mathbf{L}\widehat{\mathbf{\Sigma}}\mathbf{L}^t n a_i/(n-q)$, where **L** is a matrix of loadings for the current factor/interaction with dimension (dxp). d is 2 or 3 in the case of ellipsoids. In the balanced case with the chosen design coding $a_i = q_g/n$. This is also the leverage of the i-th design point (diagonal element of the hat matrix[17], $\mathbf{X}(\mathbf{X^tX})^{-1}\mathbf{X^t}$), where q_g is the number of columns in the selected factor/interaction design block. The scaling constant c ensures $1-\alpha$ coverage.

We translate from Hotelling's T^2 distribution to the F-distribution for simple table lookups $T^2(p,m) = \{mp/(m-p+1)\}F_{p,m-p+1}$. In our case $m = n - q_g$ and $p = d$, which means we must scale our ellipsoids with

$$c = \sqrt{\frac{(n-q_g)d}{n-q_g-d+1}\mathcal{F}(1-\alpha, d, n-q_g-d+1)}, \qquad (9)$$

to obtain the desired level of confidence.

4.5. Pairwise comparisons of effect levels means

A direct consequence of the model ellipsoids is the possibility of performing pairwise comparisons of effect levels. As with pairwise testing in ANOVA, all pairwise, standardized distances between effect levels are computed using a Mahalanobis distance with the appropriately scaled projected covariance matrix. The only modification of the method in equation (7) is that now the $\mathbf{x}_{i,*}^t\mathbf{B}$ is replaced by the linear combination $(\mathbf{x}_{r,*}^t - \mathbf{x}_{s,*}^t)\mathbf{B}$. Significance is then judged according to whether the ellipse, with the appropriate significance level, covers 0 or not. When this is done for all combinations, a compact letter display of effect level differences can be created as shown in the example below. Compensation for multiple comparisons can for instance be performed by Bonferroni adjustment of the significance level.

4.6. Relations to other methods – extensions

In multivariate statistical process control (MSPC) similar ideas are used. In MSPC multiple process measurements collected under Normal Operating Conditions (NOC) are subjected to a PCA. The scores of that model are subsequently used to build a T^2 statistic in this reduced space, called a D-statistics, and control limits are derived from that[18,19]. Often in the visualization of MSPC models, ellipsoids are shown reflecting these control limits.

Several aspects of multivariate regression have not been discussed in this work. For instance, the effect of imbalancedness is an issue that needs to be tackled to ensure meaningful ellipsoids. Here the problem is that using an unbalanced design will lead

to non-orthogonal effects which will again affect the ellipsoids. If the ellipsoids are to be interpreted as exact, this may be problematic, while as an explorative tool, slight imbalancedness may be disregarded. The type of sums-of-squares is important in imbalanced analyses, e.g. using Type I (sequential) or Type III (disregarding marginality). This topic is beyond the scope of this article and is therefore left for a later study.

5. Examples.

5.1. Example 1. Candies

The sensory data[20] contains assessments of 11 different candies by 5 assessors that have judged 9 sensory attributes: transparency, acidity, sweet taste, raspberry flavour, sugar coated texture (tested with a spoon), biting strength in the mouth, hardness, elasticity in the mouth, stick to teeth in the mouth. All assessments have been performed three times. The model has two factors: the assessors and the candies, while the sensory attributes are responses. We also include the interaction between assessors and candies, as in the original publication.

Looking at Figure 1, we observe that the assessor ellipsoids mostly overlap to a large extent. This is a good sign, showing that the assessor panel is quite well calibrated. An exception is the lower right assessor having almost no overlap with any of the other assessors. If we take into account the loadings of Figure 2, we see that the attributes acid and sweet are furthest away from the origin and thus hardest to give

exactly the same assessment across assessors. The assessors contribute with 4.4% of the total variation of the ASCA model.

Moving on to the candies, there are two candies that are very well separated from the remaining three, which overlap both in the first and second principal component directions. The first principal component is dominated by physical characteristics, while the second (smaller) principal component is dominated by taste characteristics. In total, the candy components contribute 74.5% of the model variation. The interaction seems chaotic, revealing no interesting patterns of variation and only contributes with 7.7% of the model variation (not shown).

If we perform permutation testing on this model using 10000 permutations we get $P < 1/10000$ for both main effects and $P = 1$ for the interaction. This is in line with our observations as there is at least one level of each main effect that significantly differs from the others, while there is large overlap between all levels of the interaction.

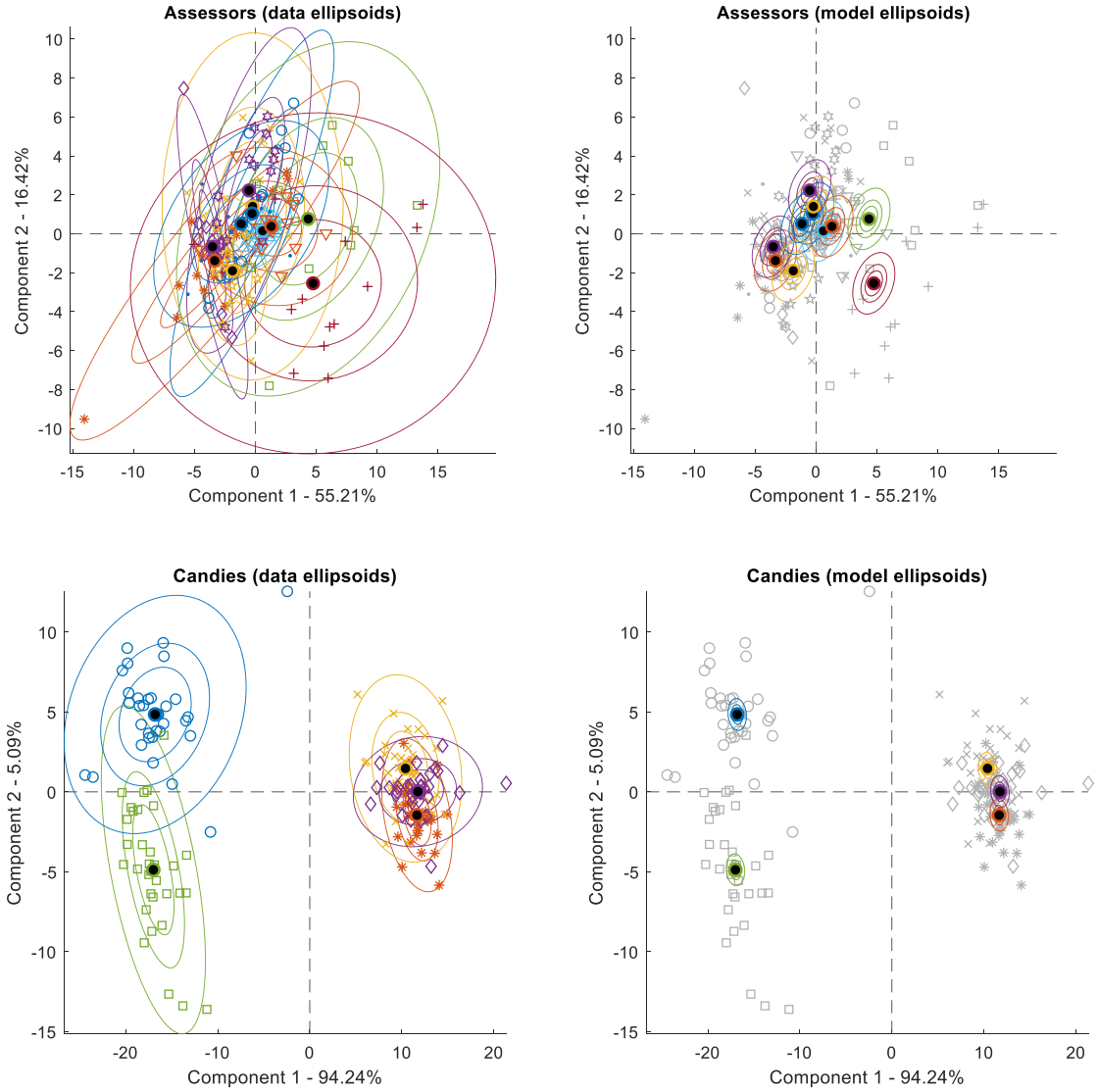

Figure 1 Sensory data score plots for the main effects with confidence ellipsoids. From inner to outer ellipsoid, these represent 40%, 68%, and 95% of the variation of the data or model factor levels, respectively.

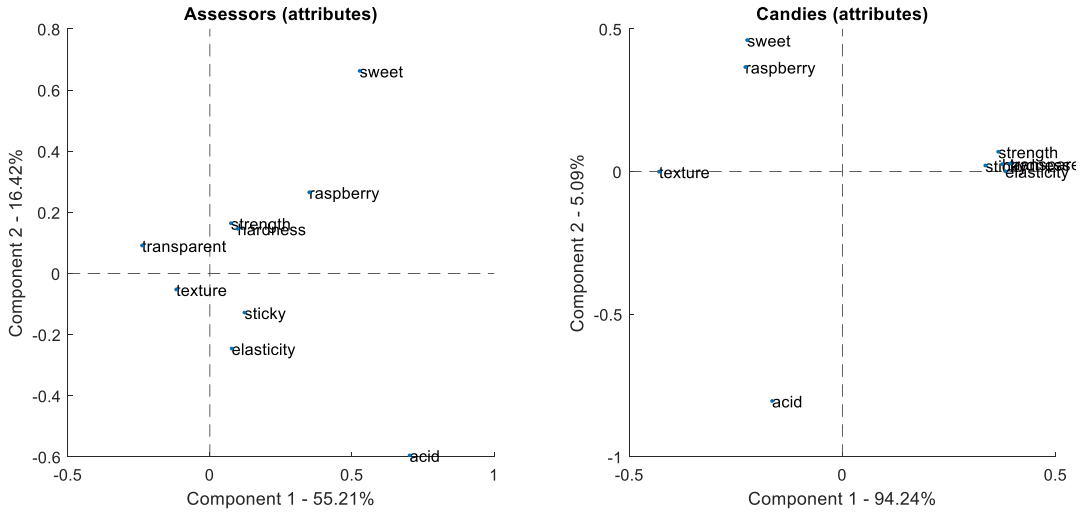

Figure 2 Sensory data loading plots displaying sensory attributes for the candy assessments. Some of the major contributors show similar patterns in the loadings, except the 180 degrees switch along the first axis.

Applying pairwise comparisons of means to the assessor effect in two dimensions, we obtain the compact letter displays (CLD) shown in Figure 3 together with a map of the assessors. This gives us a condensed view of which assessors can be seen as different or not as the latter share a letter in the CLD. For instance, when using Bonferroni correction assessors G is different from all the rest, while assessor A cannot be distinguished statistically from assessors I, F, C, H, or K. The former is true since assessor G has a unique grouping letter. The latter is true since assessor A shares group 'b' with assessors I, C, F and H, and group 'c' with assessors K, C, F and H. Here the insert-absorb algorithm [21] for CLDs is used.

Assessor	Uncorrected	Bonferroni
E	a	a
I	b	ab
K	c	c
A	cde	bc
C	cd	bc
F	b e	bc e
H	d	bcd
J	f	de
B	g	d
D	g	d
G	h	f

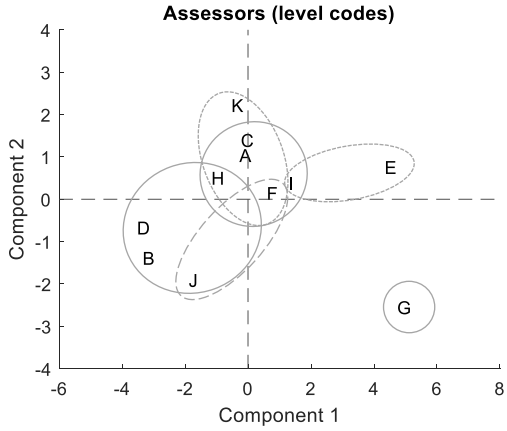

Figure 3 Sensory data compact letter display for mean assessor effects and map of the assessor level means corresponding to Figure 1 with Bonferroni corrected grouping indicated.

5.2. Example 2, Egg-pasta

The spectral data [22] contains NIR measurements (3112 wavelengths) of 540 egg-pasta samples. Three factors have been varied in the recipes: temperature (3), time (3) and concentration of egg (6). The model contains all main effects and second order interactions.

As can be seen in Figure 4, most of the levels in the main effects have non-overlapping model ellipsoids. The model ellipsoids are significantly smaller than the data ellipsoids because of the large number of replicates.

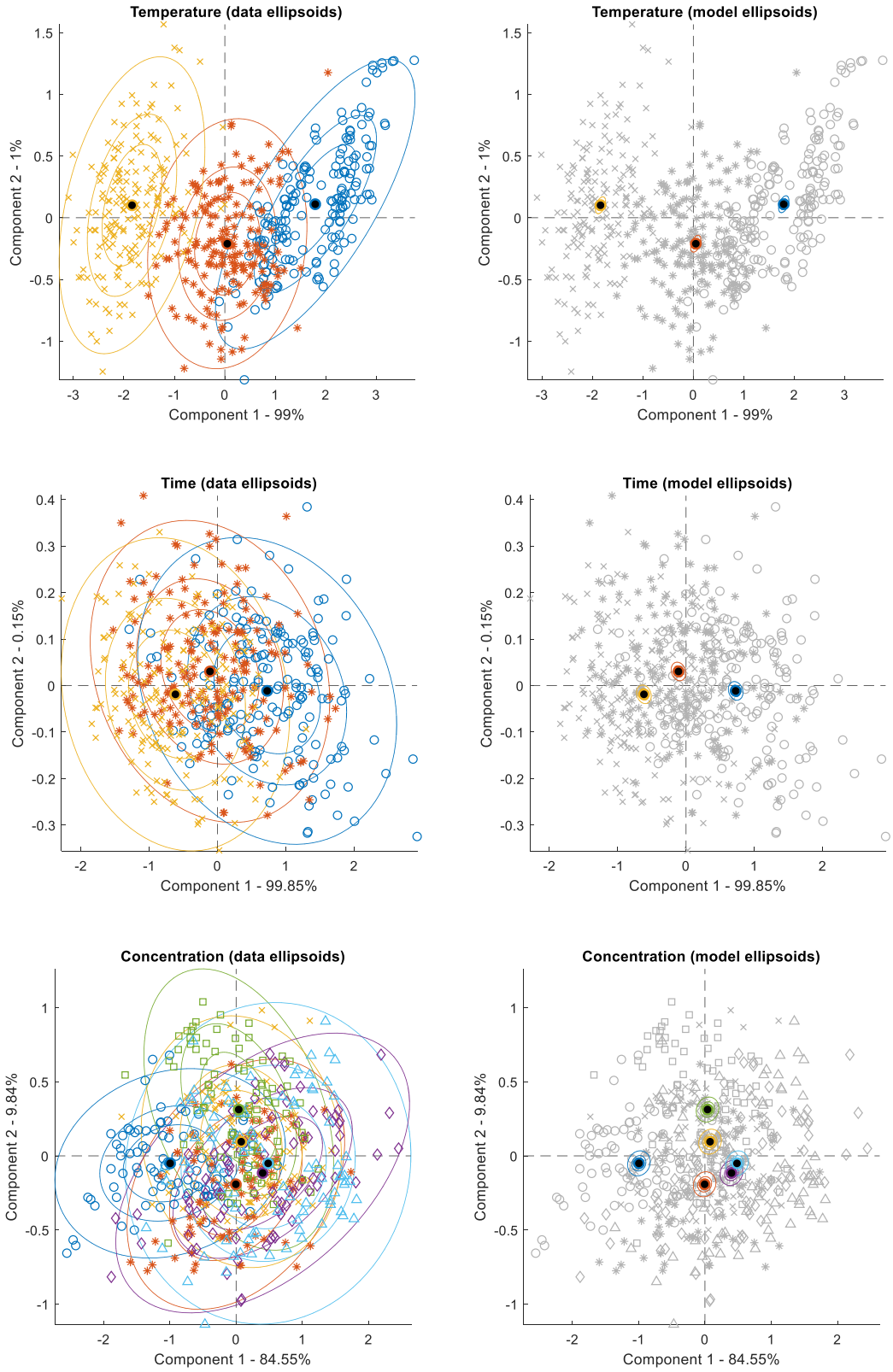

Figure 4 Spectral data score plots for the main effects with confidence ellipsoids. From inner to outer ellipsoid, these represent 40%, 68%, and 95% of the variation of the data or model factor levels, respectively.

In the original article all main effects were found to be significant using permutation testing, which confirms our findings. If we plot the ellipsoids in three dimensions, also the concentration effects can be separated in Figure 5. Also the results for the interactions are in concordance with some overlap between the levels of the temperature:concentration interaction, full overlap for time:concentration and least overlap for temperature:time. We refer to the original article for plots of interactions and loadings.

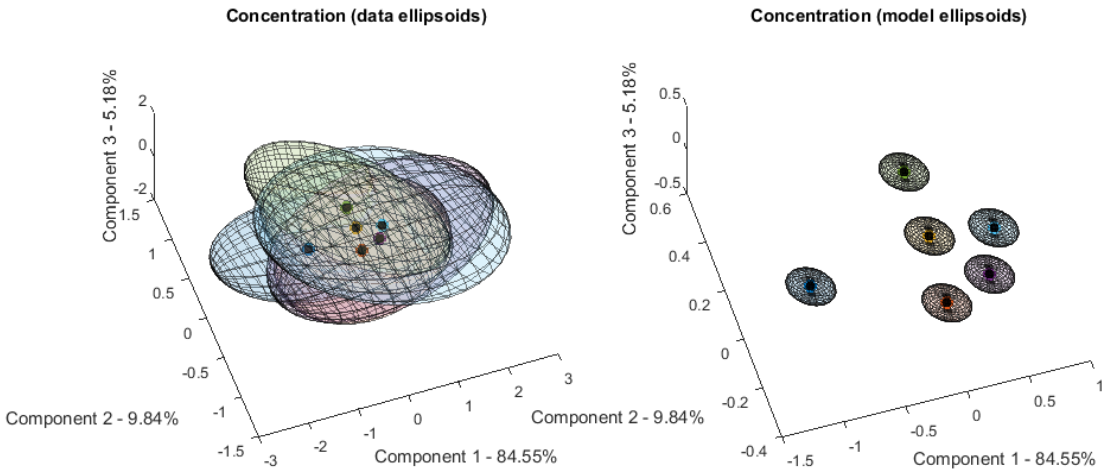

Figure 5 Spectral data score plots for the main effects with three dimensional confidence ellipsoids. Ellipsoids represent 95% of the variation of the data or model factor levels, respectively.

5.3. Simulated data

We have conducted several simulations with various numbers of effects and effect levels, with few responses and spectral like responses. The simulations allow us to explore the effects of changing the numbers of replicates and the amount of noise in each group in a controlled manner. A typical example is to simulate multinormally distributed variation around fixed factor levels. Repeated simulations are

performed with many repetitions while systematically varying the scaling of the multinormal distributions and the number of replicates.

An example of such simulations is shown in Figure 6 where standard binormal noise is added to observations around fixed factor level means in two dimensions. Two factors of four and five levels are simulated and analysed by ASCA, but only the first factor is plotted to show the effect of varying the number of replicates, N=2/N=8, and the variance scaling, $\sigma^2=1/\sigma^2=4$. The observations in the left-hand subfigures are copied four times in the right-hand subfigures to produce exactly the same covariance structures with increased number of samples.

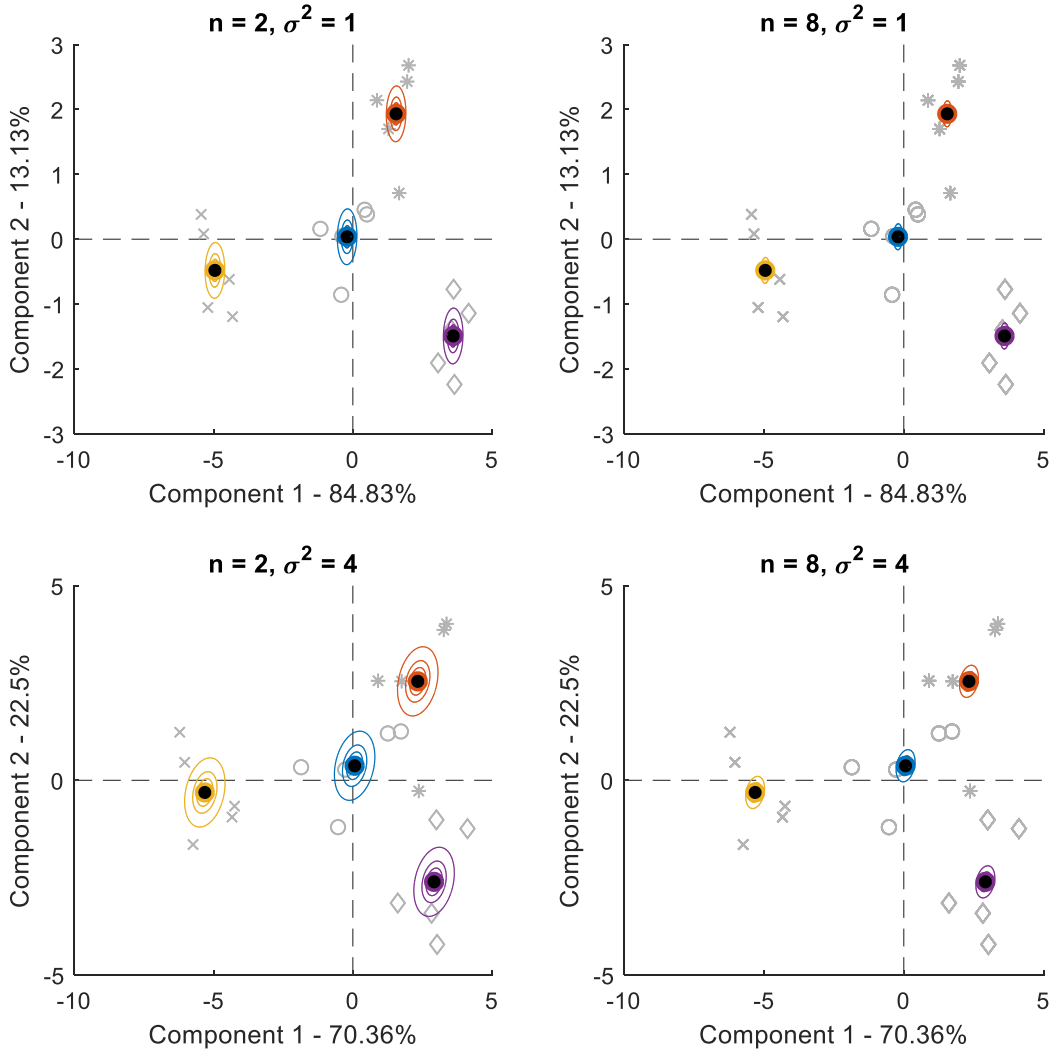

Figure 6 Simulated data score plots for the main effect with model based confidence ellipsoids. From inner to outer ellipsoid, these represent 40%, 68%, and 95% of the variation of the model factor levels, respectively. 'N' is the number of replicates per factor level combination, and 'σ²' is the variance scaling of the standard binormal distributions around the level means.

From these sets of simulations, we always observe that increasing the number of replicates with a factor of v causes the area of the ellipsoids to shrink with a factor of approximately $\sqrt{v^d}$, where d is the number of dimensions. This means that when replicates increase from 2 to 8 per factor combination, v=4 and the areas of the two-

dimensional ellipsoids shrink with a factor of $\sqrt{4^2} = 4$. The shrinkage factor is slightly inflated when only a few replicates is the basis, while quickly converging to $\sqrt{v^d}$ when the number of replicates increases. The effect of increasing the variance of the noise around of the effect levels scales exactly inversely to the number of replicates.

6. Discussion

In the work presented in this article, we have introduced an analytical solution to the estimation of uncertainty of design factor levels in ASCA. It is based on classical, linear multivariate regression theory and is therefore generalizable, e.g. to unbalanced data and more complex models. In contrast to resampling based uncertainty estimations, our approach handles interactions of any order correctly. We consider the PCA of the LS-means fixed, and thus the ellipsoids are conditional on the basis spanned by the PCA. This is similar to the way effective degrees of freedom in ridge regression are conditional on the ridge parameter[23].

In the examples we have chosen to apply the model ellipsoids to previously published data for relevant comparison and to simulated data to illustrate the properties of the theory when the number of samples and variance is changed. Our results concur with the published results, but also add more interpretation possibilities because of the focus on effect levels rather than only on whole effects. The simulations show that the ellipsoids scale in an intuitive way when the number of samples and the noise levels are changed.

With an analytical framework, pairwise comparisons of means can easily be performed. As shown in the sensory example, this can help in getting an overview of the effect levels, i.e. the assessors. Groups of similarly performing assessors can easily be pinpointed and outliers be detected.

Acknowledgements

We would like to thank the Norwegian FFL 'Research Levy on Agricultural Products' for financial support.

References

1. Næs T, Tomic O, Greiff K, Thyholt K. A comparison of methods for analyzing multivariate sensory data in designed experiments - A case study of salt reduction in liver paste. *Food Qual Prefer*. 2014;**33**:64-73.
2. Coulier L, Wopereis S, Rubingh C, Hendriks H, Radonjić M, Jellema RH. *4.09 - Systems Biology*. (Steven D. Brown RT and BW, ed.). Oxford: Elsevier; 2009.
3. Wang X, Baumgartner C, Shields DC, Deng H-W, Beckmann JS, eds. *Application of Clinical Bioinformatics*. Springer Netherlands; 2016.
4. Mardia K, Kent J, Bibby J. Multivariate Analysis. *London Acad Press*. 1979.
5. Bratchell N. Multivariate response surface modelling by principal components analysis. *J Chemom*. 1989;**3**(August 1988):579-588.
6. Ellekjær MR, Ilseng MA, Næs T. A case study of the use of experimental design and multivariate analysis in product improvement. *Food Qual Prefer*. 1996;**7**(1):29-36.
7. Langsrud Ø. 50-50 multivariate analysis of variance for collinear responses. *J R Stat Soc Ser D Stat*. 2002;**51**(3):305-317.
8. Jansen JJ, Hoefsloot HCJ, Van Der Greef J, Timmerman ME, Westerhuis JA, Smilde AK. ASCA: Analysis of multivariate data obtained from an experimental design. *J Chemom*. 2005;**19**(9):469-481.

9. Jansen JJ, Bro R, Hoefsloot HCJ, Van Den Berg FWJ, Westerhuis JA, Smilde AK. PARAFASCA: ASCA combined with PARAFAC for the analysis of metabolic fingerprinting data. *J Chemom*. 2008;**22**(2):114-121.
10. Stanimirova I, Michalik K, Drzazga Z, Trzeciak H, Wentzell PD, Walczak B. Interpretation of analysis of variance models using principal component analysis to assess the effect of a maternal anticancer treatment on the mineralization of rat bones. *Anal Chim Acta*. 2011;**689**(1):1-7.
11. Luciano G, Næs T. Interpreting sensory data by combining principal component analysis and analysis of variance. *Food Qual Prefer*. 2009;**20**(3):167-175.
12. Zwanenburg G, Hoefsloot HCJ, Westerhuis JA, Jansen JJ, Smilde AK. ANOVA – principal component analysis and ANOVA – simultaneous component analysis : a comparison. *J Chemom*. 2011;**25**:561-567.
13. Smilde AK, Hoefsloot HCJ, Westerhuis JA. The geometry of ASCA. *J Chemom*. 2008;**22**(8):464-471.
14. Vis DJ, Westerhuis J a, Smilde AK, van der Greef J. Statistical validation of megavariate effects in ASCA. *BMC Bioinformatics*. 2007;**8**:322.
15. Friendly M, Monette G, Fox J. Elliptical Insights: Understanding Statistical Methods through Elliptical Geometry. *Stat Sci*. 2013;**28**(1):1-39.
16. Thiel M, Féraud B, Govaerts B. ASCA+ and APCA+: Extensions of ASCA and APCA in the analysis of unbalanced multifactorial designs. *J Chemom*. 2017;**31**(6).
17. Hoaglin DC, Welsch RE. The Hat Matrix in Regression and ANOVA. *Am Stat*. 1978;**32**(1):17-22.
18. Nomikos P, MacGregor JF. Monitoring batch processes using multiway principal component analysis. *AIChE J*. 1994;**40**(8):1361-1375.
19. Nomikos P, MacGregor JF. Multivariate SPC Charts for Monitoring Batch Processes. *Technometrics*. 1995;**37**(1):41-59.
20. Luciano G, Næs T. Interpreting sensory data by combining principal component analysis and analysis of variance. *Food Qual Prefer*. 2009;**20**(3):167-175.
21. Piepho H. An Algorithm for a Letter-Based Representation of All-Pairwise

Comparisons. *J Comput Graph Stat*. 2004;**13**(2):456-466.

22. Bevilacqua M, Bucci R, Materazzi S, Marini F. Application of near infrared (NIR) spectroscopy coupled to chemometrics for dried egg-pasta characterization and egg content quantification. *Food Chem*. 2013;**140**(4):726-734.
23. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning - Data Mining, Inference and Prediction*. Second edi. Stanford, CA: Springer; 2008.

Appendix.

We will us a simple example based on two factors with 3 and 2 levels respectively (balanced), to illustrate the coding problem and how it is approached in this paper.

$$\begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 2 & 1 \\ 2 & 2 \\ 3 & 1 \\ 3 & 2 \end{bmatrix} \xrightarrow[code]{Dummy} \left[\begin{array}{ccc|cc} 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \end{array}\right] \xrightarrow[\substack{last\ level \\ and\ subtract \\ corresponding \\ rows}]{Drop} \left[\begin{array}{cc|c} 1 & 0 & 1 \\ 1 & 0 & -1 \\ 0 & 1 & 1 \\ 0 & 1 & -1 \\ -1 & -1 & 1 \\ -1 & -1 & -1 \end{array}\right].$$

To the left, actual levels, then dummy, before centred dummy with last columns eliminated (sum-to-zero coding).

Interactions are obtained by column-wise multiplications which gives us (the two last are interactions):

$$\left[\begin{array}{cc|c|cc} 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & -1 & -1 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & -1 & 0 & -1 \\ -1 & -1 & 1 & -1 & -1 \\ -1 & -1 & -1 & 1 & 1 \end{array}\right].$$

The columns are orthogonal between the three blocks, main effects, and interactions, but not within each effect. This means that the effects can be handled independently. Calling the above mentioned matrix **X**, we can compute **X'X** to be:

$$\begin{bmatrix} 4 & 2 & 0 & 0 & 0 \\ 2 & 4 & 0 & 0 & 0 \\ 0 & 0 & 6 & 0 & 0 \\ 0 & 0 & 0 & 4 & 2 \\ 0 & 0 & 0 & 2 & 4 \end{bmatrix}.$$

This is also illustrated in Figure 7. The block diagonal form also ensures a block diagonal inverse, $(\mathbf{X'X})^{-1}$ for any number of levels or orders of interactions:

$$\begin{bmatrix} 1/3 & -1/6 & 0 & 0 & 0 \\ -1/6 & 1/3 & 0 & 0 & 0 \\ 0 & 0 & 1/6 & 0 & 0 \\ 0 & 0 & 0 & 1/3 & -1/6 \\ 0 & 0 & 0 & -1/6 & 1/3 \end{bmatrix}.$$

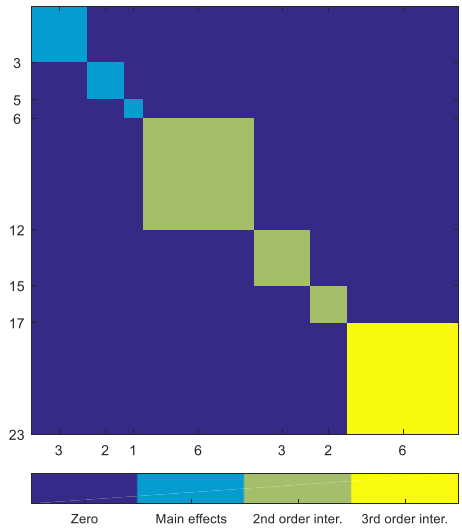

Figure 7 Block structure of X'X when using the desired dummy coding for a design having main effects of 4, 3, and 2 levels, respectively, plus 2nd and 3rd order interactions.