Generic framework for high-dimensional fixed-effects ANOVA

Age K. Smilde, Marieke E. Timmerman, Margriet M.W.B. Hendriks, Jeroen J. Jansen and Huub C.J. Hoefsloot

Abstract

In functional genomics it is more rule than exception that experimental designs are used to generate the data. The samples of the resulting data sets are thus organized according to this design and for each sample many biochemical compounds are measured, e.g. typically thousands of gene-expressions or hundreds of metabolites. This results in high-dimensional data sets with an underlying experimental design. Several methods have recently become available for analyzing such data while utilizing the underlying design. We review these methods by putting them in a unifying and general framework to facilitate understanding the (dis-)similarities between the methods. The biological question dictates which method to use and the framework allows for building new methods to accommodate a range of such biological questions. The framework is built on well known fixed-effect ANOVA models and subsequent dimension reduction. We present the framework both in matrix algebra as well as in more insightful geometrical terms. We show the workings of the different special cases of our framework with a real-life metabolomics example from nutritional research and a gene-expression example from the field of virology.

Keywords: high-dimensional data; designed experiments; ASCA; PRC; SMART

INTRODUCTION

In functional genomics — transcriptomics, metabolomics, proteomics — very often measurements are performed according to a predefined experimental design [1–3]. Experimental factors involved are typically (combinations of) dose, time and treatment but in principle not restricted to those. Staying in the univariate domain, ANOVA-based methods are typically used [4, 5]. Moving to the multivariate domain, tools like principal component analysis (PCA) or hierarchical cluster analysis are used but these do not reveal the underlying biology sufficiently clear in such designed data and more advanced methods have to be used to highlight all aspects of the experimental design [6]. A set of such advanced methods is based on the idea of separating sources of variation based on the underlying experimental design, extending the ideas used in univariate analysis. Currently, there are four such methods, namely ANOVA-Simultaneous Component Analysis (ASCA [7]); ANOVA-PCA (APCA [3]; Scaled-to-Maximum, Aligned, and Reduced Trajectories (SMART, also known as Geometric Trajectory Analysis [1]) and Principal Response Curves (PRC, [8]). All these methods are very related, as will be shown here. Applications of these methods are starting to appear showing their usefulness [9–17]. Moreover, extensions based on these methods have also appeared [18–21]. APCA equals ASCA, apart from a special way for testing
factor effects. Therefore, we will not discuss APCA separately. The original implementation of SMART uses a special scaling step which is of minor importance in our context. For reasons of simplicity we do not include this scaling.

All methods separate the sources of variation in the data by univariate linear models, albeit that these linear models differ. Subsequently, a dimension reduction is used to summarize the variation. This is then also the general framework we propose: (i) select a linear model suitable for the data and biological question, (ii) estimate the factor effects and collect these in matrices, (iii) select a dimension reduction method to summarize these factor effect matrices. This framework has a nice geometrical interpretation and is flexible enough to select alternatives depending on the biological question. We restrict ourselves to fixed-effects linear models as there are no extensions yet of mixed-effects models to the high-dimensional response case using dimension reducing methods. We also restrict ourselves to balanced data but this is only to simplify our presentation.

We start by describing the framework both in equations and geometry. Along the way, we will put the different methods in this framework as special cases. We finalize by showing the results of the different methods on a metabolomics data set from virology research and a gene-expression data set from nutritional research and a gene-expression data set from virology. In the Supplementary Data, we give the proofs.

FRAMEWORK
Step 1: The model
For didactical purposes the framework will be explained for a simple case: balanced data and the two factors treatment \((j=1,\ldots,J)\) and time \((k=1,\ldots,K)\). This is not a limitation because the framework can easily be generalized to more factors. The treatment of unbalanced data is not considered here explicitly. However, a shortcut which does fit in our framework can be taken, namely by considering the cell-means approach to unbalanced data. In the remaining text, we assume the balanced case and will remark upon unbalancedness where appropriate.

The model which serves as the basis for ASCA for this simple case reads

\[
x_{ijkl} = \hat{\mu}_j + \hat{\delta}_j + \hat{\beta}_{kl} + (\vec{\alpha}\vec{\beta})_{jkl} + e_{ijkl}, \tag{1}
\]

in the notation of [22] (see Supplementary Data), and is identical to a two-way ANOVA model. Per group, \(N\) individuals (rats, mice, humans, etc.) are present. In total, \(I\) individuals are present where \(I=NJK\). In the sequel, we will use the index \(i\) to indicate an individual and for notational simplicity not distinguish explicitly to which group the individual belongs. The index \(l=1,\ldots,L\) refers to the metabolites or genes.

The basic idea behind PRC is to express all variation relative to the control group, which we denote by \(j=1\). This results in the model

\[
x_{ijkl} = \hat{\delta}_{ijkl} + (\vec{\alpha}\vec{\beta})_{jkl} + e_{ijkl}, \tag{2}
\]

There is no direct relationship between this model and the one of ASCA, Equation (1).

For SMART, the idea is to write the effects relative to the first time point, which we denote by \(k=1\) (e.g. a pre-dose reference point). This results in the model

\[
x_{ijkl} = \hat{\beta}_{j1l} + (\vec{\alpha}\vec{\beta})_{jkl} + e_{ijkl}, \tag{3}
\]

Note that for convenience the same symbols are used in Equations (1), (2) and (3) but their estimates are different. Again, there is no direct relationship between this model and the one of ASCA, Equation (1) or PRC, Equation (2).

It is possible to combine PRC and SMART: first express everything relative to the control group \((j=1)\) and, subsequently, relative to the first time point \((k=1)\), or the other way around which is equivalent for the two linear operations on the data. This results in the model

\[
x_{ijkl} = \hat{\mu}_{11l} + \hat{\delta}_{1kl} + \hat{\beta}_{j1l} + (\vec{\alpha}\vec{\beta})_{jkl} + e_{ijkl}. \tag{4}
\]

Step 2: Estimated factor effects
The model as stated in Equation (1) cannot be estimated uniquely without extra constraints; these constraints serve as identifiability restrictions [22]. The usual constraints for an ANOVA model in terms of estimated parameters are

\[
\begin{align*}
\sum_j \hat{\alpha}_j = 0 (\forall \ell); & \quad \sum_k^K \hat{\beta}_{k\ell} = 0 (\forall \ell); \\
\sum_j (\vec{\alpha}\vec{\beta})_{j\ell} = 0 (\forall \ell, \ell); & \quad \sum_k^K (\vec{\alpha}\vec{\beta})_{\ell k} = 0 (\forall \ell, \ell).
\end{align*}
\tag{5}
\]

which are the constraints used in ASCA. Under these constraints, the factor effects can be estimated as.
\[ x_{ijkl} = \bar{x}_{...} + (\bar{x}_{jkl} - \bar{x}_{...}) + (\bar{x}_{ikl} - \bar{x}_{...}) \]
\[ + (\bar{x}_{jkl} - \bar{x}_{jkl} - \bar{x}_{ikl} + \bar{x}_{...}) + (x_{ijkl} - \bar{x}_{jkl}), \] (6)

where the dot-notation in the subscripts indicates over which index the mean is taken. On the right-hand side of Equation (6), the mean over all \( I \), the main treatment effect, the main time effect, the interaction treatment \( \times \) time, and the individual contributions, respectively, can be seen. In short, Equation (6) shows the decomposition of a measured contribution which is the same for all methods [Equations (1–4)].

The systematic variation \( x_{sys,l} \) can be decomposed in different ways leading to the different models. PRC focuses on differences relative to the control group, SMART focuses on differences relative to the first time point and ASCA takes the more traditional approach of relating everything to mean values.

**Step 3: Dimension reduction**

**General**

The starting point of any data analysis is the data itself. For the simple case used above, this will be explained in terms of metabolites which refers to the first example. It holds equally well for gene-expressions or proteomics data as will be shown in the second example. For two metabolites the raw data may look like the presentation in Figure 1. This figure displays three treatment groups (indicated by color and superscript) and three time points (indicated by subscript). For simplicity, biological repeats are not shown, but these are centered around each individual point in the graph. Each dot in the figure is a data point (i.e. the average of the rows of \( \mathbf{X} \) pertaining to the same time/treatment combination).

All methods as explained below can be understood in simple geometrical terms as translations, means...
and linear subspaces. Figure 1 will serve as the geometrical example throughout.

After having selected an appropriate linear model, the next step is to collect estimates for all \( l = 1, \ldots, L \) metabolite responses in matrices where each column represents the contribution of one metabolite. For ASCA, this means that all contributions in Equation (6) are written as

\[
X = X_M + X_K + X_J + X_{JK} + X_{JK},
\]

(14)

where the naming of the matrices of contributions is according to the terms in Equation (6) with obvious subscripts for the matrices and all of them are of size \( I \times L \). Note that some of the rows of, e.g., \( X_J \) are repeated, because their values are the same for different \( k \)'s at the same \( j \). The geometry is shown in Figure 2A. The rows of \( X_M + X_K \) are indicated with \( t_1, t_2, t_3 \) which represent the mean overall time trajectories and the vectors \( a_k^l \) are the deviations of the original data points \( x_{jk}^l \) from these \( t_1, t_2, t_3 \). After this translation operation the rows of \( X_J + X_{JK} \) are shown in Figure 2B. Note that the matrix \( X_{JK} \) is not visualized in Figure 2 but is indeed present in the decomposition.

For the individual matrices on the right-hand side of Equation (14) it can be shown that their column
Spaces are orthogonal, e.g. \( X_j X_K = 0 \) [2]. This is a desirable property for several reasons, one of them being the resulting partitioning of sums-of-squares:

\[
\|X\|^2 = \|X_M\|^2 + \|X_j\|^2 + \|X_K\|^2 + \|X_{jk}\|^2 + \|X_{ijk}\|^2.
\]

(15)

where the symbol \( |X|^2 \) refers to the squared Frobenius norm of \( X \) and equals the sum of squared entries of \( X \). ASCA proceeds now by performing PCA’s on the separate matrices or on sums of matrices; e.g., a PCA on \( X_j + X_{jk} \) (the data shown in Figure 2B) gives a view on the dose effect on the metabolites [2] or, alternatively, a PARAFAC model can be used on these matrices as they can also be considered matricized three-way arrays [6]. For unbalanced data, this orthogonality does not hold automatically anymore and special provisions have to be taken [21].

For a PRC analysis, the starting point is Equation (8). The following decomposition is made:

\[
X = X_{1K} + X_{PRC} + X_{ijk},
\]

(16)

where \( X_{1K} \) collects the terms \( \tilde{x}_{1kl} \) and \( X_{PRC} \) collects the terms \( (\tilde{x}_{jkl} - \tilde{x}_{1kl}) \). The matrix \( X_{ijk} \) containing the residuals is the same in Equations (14) and (16). All of the matrices are of size \((I \times L)\). PRC now proceeds by performing a PCA on \( X_{PRC} \). The geometry of PRC is illustrated in Figure 2C and 2D. Treatment group 1 is defined as the control group and all data points are expressed relative to this control group. The vectors \( p_j^k \) (Figure 2C) illustrate this and the rows of \( X_{PRC} \) are shown in Figure 2D. These rows are translations relative to the vectors \( p_j^k \), the rows of \( X_{1K} \).

The column spaces of \( X_{1K} \) and \( X_{PRC} \) are not orthogonal. This holds already for the univariate linear models [i.e. in Equation (8)] and it carries over to the multivariate case. Hence, a partitioning of variation as in Equation (15) is not possible anymore. If such a partitioning is needed, then this can be ensured by enforcing the loadings of the subsequent PCA of \( X_{PRC} \) to be orthogonal to the row space of \( X_{1K} \). The sum-of-squares due to \( X_{ijk} \) is already separated because this matrix is column-orthogonal to the others (see Supplementary Data for proof). This does not hold for unbalanced data.

For SMART, the starting point is Equation (10). The following decomposition is made:

\[
X = X_{J1} + X_{SMART} + X_{ijk},
\]

(17)

where \( X_{J1} \) collects the terms \( \tilde{x}_{j1l} \) and \( X_{SMART} \) collects the terms \( (\tilde{x}_{jkl} - \tilde{x}_{j1l}) \). The matrix \( X_{ijk} \) containing the residuals is the same in Equations (14), (16) and (17). All of the matrices are of size \((I \times L)\). The geometry of SMART can be visualized in a similar way as for ASCA and PRC and will not be done here.

The column spaces of \( X_{J1} \) and \( X_{SMART} \) are not orthogonal, likewise as for PRC. Hence, a partitioning of variation as in Equation (15) is not possible anymore. SMART now proceeds by performing a PCA on \( X_{SMART} \) possibly after a special scaling step within a treatment group to correct for large differences between treatment effect sizes.

Figure 3: The raw data of 5-(3'-methoxy-4'-hydroxyphenyl)-g-valerolactone and 5-(3',4'-dihydroxyphenyl)-g-valerolactone; gut-mediated metabolites. Legend: blue (dark grey) is placebo extract; red (light grey) is tea extract and black is wine extract. Vertical bars indicate standard deviations of individual differences.
Also the combined SMART/PRC model can be subjected to a dimension reduction starting from the decomposition

\[ X = X_{11} + X_{SP} + X_{IJK}, \]  

where \( X_{11} \) collects the terms \( \bar{x}_{1} + (\bar{x}_{j} - \bar{x}_{1}) \) + (\( \bar{x}_{kl} - \bar{x}_{1kl} \) + \( \bar{x}_{11} \)) of Equation (12) and \( X_{IJK} \) is the same as before. Analysis can now proceed by performing a PCA on \( X_{SP} \).

**Relationships between ASCA, PRC and SMART**

There are clear relationships between the methods which will be exemplified geometrically for ASCA and PRC (Figure 2E). PRC relates everything to the control (treatment 1). This means, for example, that the row of \( X_{PRC} \) pertaining to treatment 2 at time point 1 equals \( p_{1} = a_{2} - a_{1} \) where the data for treatment 2 and time point 1 \( (a_{2}) \) are expressed as deviations from treatment 1 and time point 1 \( (a_{1}) \), see Figure 4E. Also in terms of subspaces there are relationships between the methods. These are:

1. \( R(X'_{PRC}) = R((X_{j} + X_{JK})') \)
2. \( R(X'_{SMART}) = R((X_{K} + X_{JK})') \)
3. \( R(X'_{SP}) = R(X'_{JK}) \)
4. \( R(X'_{PRC}) \neq R(X'_{SMART}) \)
5. \( R(X'_{PRC}) \cap R(X'_{SMART}) = \emptyset \) if \( L < 2JK - K - J \)

where, e.g. \( R(X'_{PRC}) \) indicates the row-space of \( X_{PRC} \) (for proofs: see Supplementary Data). These relationships have repercussions for analyzing the same data set with ASCA, PRC or SMART and they also hold for unbalanced data (see Supplementary Data).

From item 1, it can be concluded that \( X_{PRC} \) and \( (X_{j} + X_{JK}) \) can be expressed on the same orthogonal basis \( V \). Hence, it is possible to express \( X_{PRC} \) as \( T_{PRC}V \) and simultaneously \( (X_{j} + X_{JK}) \) as \( T_{ASCA}V \). Then \( T_{ASCA} = QT_{PRC} \) where \( Q \) is a non-singular matrix.
transformation matrix. A similar line of reasoning can be made for SMART (item 2) in terms of $X_{\text{SMART}}$ and $(X_K + X_{JK})$. This means that if no dimension reduction is used on neither $X_{\text{PRC}}$ nor $(X_{J} + X_{JK})$, then the results of the methods are essentially the same.

When only a reduced set of components is taken in ASCA, PRC or SMART then there is no guarantee that the corresponding loading matrices span the same space and such transformations as above will generally not exist. The ASCA, PRC and SMART scores are then also essentially different and focus on different aspects of the data: deviations relative to a control group (PRC), deviations relative to the first time point (SMART) or differences between groups across time (ASCA). These are simply different views on the same data set and the choice to be made depends on the biological question at hand. For the case that the reduced set of components describe a large amount of the variation in the associated matrices, it is expected that the methods give similar results because then the relationships of items 1 and 2 approximately hold.

Obviously, many generalizations are possible (e.g. the proposed SP generalization), depending on which linear model step is chosen. This will lead to new methods, tailor made for specific biological questions. For practical purposes, it is informative to investigate the ranks of the different matrices involved because this has repercussions for the maximum number of expected principal components in the different contribution matrices associated with PRC, ASCA and SMART (see Supplementary Data).

**EXAMPLES**

The examples are chosen to reflect different types of applications. The first example is from metabolomics with a limited set of metabolites. This limited set allows for a detailed explanation of the results of

---

**Figure 5:** Scores of the three methods on the third component. Legend: (a) ASCA, (b) post-processed ASCA, (c) PRC and (d) SMART. Colors: blue (dark grey) is placebo extract; red (light grey) is tea extract and black is wine extract.
the different methods. The second example is typical for the field of gene-expression analysis where very many genes are measured simultaneously. This example shows the working of the methods for very high-dimensional data.

**A nutritional metabolomics example**

The example is from a nutrikinetics study where 20 male subjects underwent a treatment with a tea extract, a wine extract or a placebo extract in a cross-over design. Metabolites were measured with GC-MS in plasma and of those, 11 were selected for further study. Blood samples were collected just before (0 h) and at 1, 2, 6, 6, 9, 12, 24 and 36h after the intake (for experimental details, see van Velzen et al., manuscript submitted to publication).

A typical example of the raw data is shown in Figure 3 where gut-mediated metabolites (valero-lactones) are presented entering the plasma with a time lag. Clearly, there is large inter-individual variation which is often the case in nutritional studies. These data are now subjected to the different analysis methods. No scaling was performed.

All three methods were applied to the data set according to the theory described above. In all cases three components were extracted from the relevant matrices, explaining 95–96% of the variation in the respective matrices (see “Theory Section”). The results are shown in terms of the scores and loadings of those matrices. The first score vectors are shown in Figure 4A, C and D. The percentages explained variation of these first components for ASCA, PRC and SMART range between 93.6 and 94.1 but note that these percentages cannot be compared directly to each other since they pertain to different matrices ($X_J + X_{JK}$, $X_{PRC}$ and $X_{SMART}$, respectively). To ease the comparison with PRC and SMART, we performed a postprocessing step on the ASCA solutions. That is, we expressed the ASCA scores as deviations from the estimated ASCA scores of the
control group, to be denoted as ASCA postprocessed (Figure 4B). This approach is similar to PRC, but performed on the estimated scores, rather than on the observed values themselves. The first component (Figure 4) shows the overall trend of all the metabolites (see Supplementary Data). The trends of PRC, ASCA postprocessed and SMART are similar and this is revealed in this component. As can be seen in Figure 5, representing the third components, the results start to deviate in the third component. Although the amount of explained variation in this third component is low, the patterns are not random and can be interpreted. There is an extra peak in the scores of SMART and PRC, especially for the wine extract results. This peak describes the behavior of the gut-mediated metabolites 5-(3′,4′-dihydroxyphenyl)-g-valerolactone and 5-(3′,4′-dihydroxyphenyl)-g-valerolactone (see Figure 3 and Supplementary Data). These metabolites behave differently than the other ones and require an extra component. The loadings on the third component (see Figure 6) show this also: the most remarkable differences in these three plots are the loadings on metabolite 11 (5-(3′,4′-dihydroxyphenyl)-g-valerolactone). This loading is high for SMART and PRC whereas this is not the case for ASCA. This can be explained by the least squares properties of principal component analysis. SMART and PRC need to describe the variation of this metabolite measured from baseline, but ASCA first centers this metabolite around its mean value in the time direction. Hence, the amount of variation to be explained is for ASCA lower than for PRC and SMART. Metabolite 5-(3′,4′-dihydroxyphenyl)-g-valerolactone has much higher values than metabolite 5-(3′-methoxy-4′hydroxyphenyl)-g-valerolactone, explaining why the latter metabolite does not load that high on the third component.

**Figure 7**: Gene-expression example. Scores of the three methods on the first component. Legend: (a) ASCA, (b) post-processed ASCA, (c) PRC and (d) SMART. Colors: blue (dark grey) is wild-type, red (light grey) is R80A strain and black is F72A/R73A strain.
The amount of explained variation by the first component is high in all cases (ASCA, PRC and SMART). Hence, the similarity between the methods was already expected in this case as explained in the Theory. All three methods cover much of the same variation. Nevertheless, for illustrative purposes the third component was also shown. With other data sets, the amount of explained variation can be lower and differences are expected to appear earlier.

A gene expression example from virology

This example is from the field of virology using publicly available gene-expression data [23]. In cell lines the role of the Vpr protein in advancing HIV pathogenesis is studied by means of comparing the wild type with two mutated lines derived from the wild type. After induction with doxycyclin, RNA was collected at 0, 1, 2, 4, 6, 8, 12, 16 and 24 h. Microarrays containing 22,434 oligonucleotide spots were subsequently used to measure gene-expression. For each time point, there were at least three biological replicates. This is an example of a high-dimensional data set with a control, two treatments and nine time points and thus fits in our framework. The three methods were applied to the preprocessed gene-expression data. Details regarding the data are given elsewhere [23].

Figure 7 shows the scores on the first component of the different models. ASCA and PRC yield similar results because there is no pronounced time profile and both methods focus on the differences between strains. The first component in SMART is completely different from the other two methods. Since in SMART the strains are forced to be similar on \( t = 0 \), the between strain differences are smaller and the first component in SMART focuses on the treatment induced difference for the wild type. This can be...
seen from the large difference on the blue line for \( t = 0 \) compared to \( t = 1 \). In the second component (Figure 8), ASCA and PRC start to differ. A more striking result is that the second component of SMART picks up the dynamic difference after treatment of strain R80A, that can be seen as the large jump in the red line from \( t = 0 \) to \( t = 1 \).

CONCLUSIONS

In this article, a framework has been given for fixed-effects high-dimensional ANOVA. Theoretical relationships have been derived for methods falling under this framework and these relationships aided to interpret the results. Simple geometric arguments can be used to support the framework and understand the working of the methods. There is no preferred method: all methods simply give another view on the complex data and the choice also depends on the biological question. Mathematical differences between the methods cause differences between their properties, e.g. whether or not sums-of-squares of variation can be split in separate parts. This might influence the choice of the particular method. In a practical example, the differences between the methods are explained and illustrated. The framework allows for generating other methods depending on the biological question to be solved such as the SP model that combines SMART and PRC.

SUPPLEMENTARY DATA

Supplementary data are available online at http://bib.oxfordjournals.org/.

**Key Points**

- Common bioinformatic methods to analyze functional genomics data with an underlying design are special cases of a general framework.
- Functional genomics data with an underlying design can be analyzed in a systematic way following the steps in the presented framework.
- The presented framework allows for interaction between bioinformaticians and biologists.

Acknowledgements

Ewoud van Velzen and John van Duynhoven (Unilever Research Vlaardingen) are gratefully acknowledged for making available the metabolomics data set. This project was carried out within the research programme of the Netherlands Metabolomics Centre (NMC) which is part of the Netherlands Genomics Initiative / Netherlands Organization for Scientific Research.

References


