Metabolomics (2018) 14:73
https://doi.org/10.1007/s11306-018-1369-1

ORIGINAL ARTICLE

# Group-wise ANOVA simultaneous component analysis for designed *omics* experiments

**Edoardo Saccenti**[1] · **Age K. Smilde**[2] · **José Camacho**[3]

Received: 2 March 2018 / Accepted: 5 May 2018

**Abstract**

**Introduction** Modern *omics* experiments pertain not only to the measurement of many variables but also follow complex experimental designs where many factors are manipulated at the same time. This data can be conveniently analyzed using multivariate tools like ANOVA-simultaneous component analysis (ASCA) which allows interpretation of the variation induced by the different factors in a principal component analysis fashion. However, while in general only a subset of the measured variables may be related to the problem studied, all variables contribute to the final model and this may hamper interpretation.
**Objectives** We introduce here a sparse implementation of ASCA termed group-wise ANOVA-simultaneous component analysis (GASCA) with the aim of obtaining models that are easier to interpret.
**Methods** GASCA is based on the concept of group-wise sparsity introduced in group-wise principal components analysis where structure to impose sparsity is defined in terms of groups of correlated variables found in the correlation matrices calculated from the effect matrices.
**Results** The GASCA model, containing only selected subsets of the original variables, is easier to interpret and describes relevant biological processes.
**Conclusions** GASCA is applicable to any kind of *omics* data obtained through designed experiments such as, but not limited to, metabolomic, proteomic and gene expression data.

**Keywords** Analysis of variance · Designed experiments · Principal component analysis · Sparsity

## 1 Introduction

In systems biology and functional genomics designed experiments are nowadays very common: this refers to research situations in which a dependent variable $x$, measured on a (biological) system, constitutes the response to $I$ independent variables, called factors (or treatments), whose levels are controlled by the experimenter. The variety of designed experiments ranges from simple case–control settings to complex scenarios where many factors are manipulated simultaneously.

When a measured variable is a function of (several) factors, analysis of variance (ANOVA) is a well establish technique to analyze the data (Searle and Gruber 2016). However, in functional genomics and systems biology many variables $(x_1, x_2, \dots, x_J)$ are usually measured on a system, like in metabolomics, proteomics and transcriptomics experiments where hundreds to thousands of variables are acquired. In these cases a single ANOVA model can be fitted separately on each variable. Although effective (this is what is usually done in the case of gene expression data), this approach does not take into account the relationships existing among the variables, i.e. the multivariate nature of the problem is discarded (Saccenti et al. 2014). Because biological variables, such as metabolites or genes, are often interrelated, it is desirable, in many occasions, to analyze all the $J$

✉ Edoardo Saccenti
esaccenti@gmail.com

Age K. Smilde
a.k.smilde@uva.nl

José Camacho
josecamacho@ugr.es

[1] Wageningen University & Research, Wageningen, the Netherlands

[2] University of Amsterdam, Amsterdam, the Netherlands

[3] University of Granada, Granada, Spain

Published online: 21 May 2018

variables simultaneously (Saccenti et al. 2014). ANOVA can be generalized to the multivariate case through multivariate-ANOVA (MANOVA, see Eq. 1) (Bibby et al. 1979). In contrast with running several separated ANOVAs, MANOVA takes into account the correlation among the $x_1, x_2, \ldots, x_J$ dependent variables when testing for the significance of the factor effects; moreover, since all variables are used simultaneously it also reduces the risk of Type I error (O'Brien and Kaiser 1985).

Unfortunately, in the case of high dimensional *omics* data the MANOVA model often breaks down because the number of variables $J$ is larger than the number of observations $N$, and this results in the covariance matrices involved in the calculations for significance testing and data visualization to be singular, which makes the MANOVA solution non achievable. A classical remedy to this problem is regularization which involves numerical manipulation of the covariances matrices to remove singularity: this is the so-called regularized-MANOVA for which several solutions have been proposed (Engel et al. 2015; Ullah and Jones 2015). Other solutions regard non-parametric reformulations of the problem (Legendre and Anderson 1999; Anderson 2001).

A different approach is to combine ANOVA with principal component analysis (PCA) to reduce the dimensionality of the data to be analyzed. Earlier applications involved performing MANOVA on a reduced data set consisting of the low dimensional scores of a PCA model fitted to the data (Bratchell 1989). However this approach suffers from the limitation that PCA is not able to resolve the different type of variation induced by the different factors which may be confounded by the initial PCA model. These limitations can be overcome by using ANOVA-simultaneous component analysis (ASCA) (Smilde et al. 2005; Jansen et al. 2005).

ASCA uses an ANOVA model to first decompose the data matrix into factor effect (and interaction) matrices containing the average values for each factor level (and interaction thereof); then a PCA model is fitted separately on each effect matrix to extract and assess the contribution of each variable to the systematic variability induced by each experimental factor. Hence, an ASCA model can be explored and interpreted like a standard PCA model. A similar but less powerful approach is ANOVA-PCA (Harrington et al. 2005; Zwanenburg et al. 2011).

While ASCA retains both the flexibility of the ANOVA framework to account for (possibly) very complicated experimental designs and the versatility of PCA as a data dimensionality reduction method, it also inherits the limitations of the classical principal component analysis.

PCA is a valuable tool for data reduction and exploration: however since it is essentially a data factorization based on variance maximization, it presents two main shortcomings when data understanding and interpretation are the goal of the analysis. First, it cannot distinguish between variance which is unique for a single variable and variance which is shared among several variables and this can seriously hamper the unveiling of (possibly) hidden relationships existing among variables (Jolliffe 2002). Second, the principal components are linear combinations of all the variables simultaneously and this greatly complicates data interpretation since all variables contribute to the PCA model (Jolliffe et al. 2003).

While the first limitation can be addressed by using methods that focus on shared variance, like Factor analysis (Fabrigar et al. 1999), better interpretability of the PCA solution can be obtained by imposing a simple structure on the components, in such a way that the components are combinations of a smaller number of original variables. This is the realm of sparse methods and many formulations have been proposed (see for instance sparse implementations using LASSO (Jolliffe et al. 2003; Zou et al. 2006), group LASSO (Jacob et al. 2009) or structure-based regularization criteria (Jenatton et al. 2009).)

We recently proposed a new sparse implementation of PCA where sparsity is defined in terms of groups of (correlated) variables identified from the data to be analyzed, called group-wise PCA (GPCA) (Camacho et al. 2017). The GPCA solution is such that every principal component contains loadings different from zero only for a group of variables. This grounds on the framework of simplivariate models (Hageman et al. 2008; Saccenti et al. 2011) which aim to retain both the comprehensiveness of a multivariate model and the simplicity of interpretation of a univariate one, under the assumption that a given (biological) phenomenon may not be accounted by all measured variables but only by one, or more, subsets of variables. This kind of sparsity is natural in biological problems: examples are sets of metabolites participating in the same metabolic network or co-expressed and co-regulated genes which are expected to exhibit a correlative behavior. Thus, the sparsity exploited in GPCA is different from the one used in sparse PCA implementations based on regularization: in the latter sparsity is obtained by forcing to zero the loadings corresponding to some variables by controlling one or more regularization parameters which must be algorithmically optimized. In GPCA the parameter controlling sparsity is immediately related to the strength of association among (groups of) variables as expressed, for instance, by their correlation. In addition the relationship between the threshold on the correlation value and the level of sparsity, i.e. the size and the number of groups selected, can be graphically visualized and explored, consistently with a data exploratory philosophy.

In this paper we propose to replace the PCA step in ASCA with GPCA, arriving to a group-wise sparse version of ASCA termed Group-wise ANOVA-simultaneous component analysis (GASCA). The aim is to improve the interpretability of the ASCA solution when analyzing complex data sets. The characteristics of this approach are illustrated with simulations and by comparing the GASCA model with both PCA and the original ASCA and by the analysis of a designed plant and human metabolomics experiments. The paper is organized as follows: The Sect. 2 introduces the ANOVA–MANOVA framework and details the mathematics and the properties of the PCA, GPCA, ASCA and GASCA models. The Sect. 3 presents the data description and details on software used. Finally, Results and discussion of PCA, ASCA and GASCA modeling of simulated and experimental data are given: in particular, the fitting of the GASCA model is illustrated step by step using real experimental data. Some final considerations are offered in the Sect. 5.

## 2 Theory

### 2.1 (M)ANOVA model

We consider here a study design involving two factors $\alpha$ and $\beta$ with $A$ and $B$ levels, respectively, where $J$ variables are measured. For a balanced design, in which every measurement is replicated $R$ times for each combination of factor levels, there are in total $N = ABR$ observations. The multivariate ANOVA model (MANOVA) is given by

$$\mathbf{X} = \mathbf{1}\mathbf{m}^{\mathrm{T}} + \mathbf{X}_{\alpha} + \mathbf{X}_{\beta} + \mathbf{X}_{(\alpha\beta)} + \mathbf{E} \tag{1}$$

where the first term $\mathbf{1}\mathbf{m}^{\mathrm{T}}$ represents the overall mean for the data, $\mathbf{1}$ is a column vector of ones of length $N$ and $\mathbf{m}^{\mathrm{T}}$ is a row vector of size $J$ with the averages over the data for each variable. The effect matrices $\mathbf{X}_{\alpha}$ and $\mathbf{X}_{\beta}$ contain the level averages for each factor and the $\mathbf{X}_{(\alpha\beta)}$ describes the interaction between the two factors. The variation that cannot be represented by the model is collected in the residual matrix $\mathbf{E}$. Equation (1) is the starting point of both ASCA and the newly proposed GASCA.

### 2.2 The PCA model

Given a data matrix $\mathbf{X}$ of size $N \times J$ (observations × variables), the standard PCA model follows the expression:

$$\mathbf{X} = \mathbf{T}_{H}\mathbf{P}_{H}^{\mathrm{T}} + \mathbf{E}_{H}, \tag{2}$$

where $\mathbf{T}_{H}$ is the $N \times H$ score matrix containing the projection of the objects onto the $H$ principal components subspace, $\mathbf{P}_{H}$ is the $J \times H$ loading matrix containing the linear combination of the variables represented in each principal component, and $\mathbf{E}_{H}$ is the $N \times J$ matrix of the residuals. Usually $H$ is chosen to be much smaller than $J$.

### 2.3 The group-wise PCA model

The Group-wise Principal component analysis (GPCA) (Camacho et al. 2017) is a sparse formulation of the PCA algorithm where sparsity is defined in terms of groups of correlated variables: every component contains non-zero loadings for a single group of correlated variables which simplifies the interpretation of the model. The GPCA approach consists of three steps:

1. Computation the association map $\mathbf{M}$ form the data
2. Identification of the groups of associated variables
3. Calibration and fitting of the GPCA model

The GPCA modeling starts with the definition of a $J \times J$ association map $\mathbf{M}$ computed from the data and describing the relationship among the variables. In the original formulation of GPCA (Camacho et al. 2017) the MEDA approach (Missing-data for Exploratory Data analysis) (Camacho 2011) was used to define $\mathbf{M}$. Briefly, MEDA consists of a post-processing step after the PCA factorization to infer the relationships among variables using missing data imputation (Arteaga and Ferrer 2002; Arteaga and Ferrer 2005). The $l, j$-th element $m_{lj}$ (for variables $l = 1, \ldots, J$ and $j = 1, \ldots, J$) of the MEDA map $\mathbf{M}$ can be expressed as (Arteaga, 2011):

$$m_{lj} = \frac{\left\{ \mathbf{x}_l^{\mathrm{T}}\mathbf{x}_j + (\mathbf{e}_l^{Q})^{\mathrm{T}}\mathbf{e}_j^{Q} \right\} \cdot \mathrm{abs}\left\{ \mathbf{x}_l^{\mathrm{T}}\mathbf{x}_j - (\mathbf{e}_l^{Q})^{\mathrm{T}}\mathbf{e}_j^{Q} \right\}}{\sigma^2_{\mathbf{x}_l}\sigma^2_{\mathbf{x}_j}} \tag{3}$$

where $\mathbf{e}_l^{Q}$ is the vector of residuals for the $l$-th variable in the PCA model with $Q$ latent variables; data is assumed to be centered. Practically, this approach uses a missing data strategy to estimate the correlation between any two variables: this approach has been found to be effective in filtering out noise when estimating correlations (Camacho 2010). Here we set $Q$ using the *ckf* cross-validation algorithm (Saccenti and Camacho 2015b) as in the original GPCA formulation (Camacho et al. 2017) but other approaches are possible (Saccenti and Camacho 2015a). Note that if we set $Q = \mathrm{rank}(X)$ the MEDA map from Eq. (3) reduces to a standard Pearson correlation map where the original magnitudes are replaced by their squared values while the sign is retained. In place of the MEDA map any square symmetric matrix describing mutual relationship among the variables can be used as an input for GPCA (like mutual information, as often done in the case of gene expression data): since metabolomic data is usually analyzed in term of correlations (Saccenti et al. 2014; Saccenti 2016) we will present also a GASCA implementation based on correlations. Because relationships among metabolites cannot be assumed to be linear, we use here Spearman's rank correlation, which equal to the Pearson's correlation of the ranks of the variables. In addition, to reduce the risk of including chance associations we force to 0 correlations for which the associated $P$ value is larger than 0.01. Summarizing the elements $m_{lj}$ of the association map $\mathbf{M}$ based on correlations are:

$$m_{lj} = \begin{cases} r_{lj} & \text{if } P\text{-val} \leq 0.01 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Once the association map $\mathbf{M}$ is defined, the $K$ groups of correlated/associated variables are identified using the so called group identification algorithm (GIA). Briefly, the GIA works as follows: given $m_{lj} \in [-1, 1]$ the $l, j$-th element of $\mathbf{M}$, and $|\gamma| < 1$, the group $S_k$ is built in such a way that $|m_{lj}| > \gamma$ for all $l, j$ in $S_k$ with maximum cardinality. The $j$-th variable is not in group $S_k$ when $|m_{lj}| \leq \gamma$ with at least one variable $l$-th in $S_k$. The GIA algorithm is fully detailed in the Appendix of reference (Camacho et al. 2017). As in the case of the definition of the association map $\mathbf{M}$, other strategies can be implemented to define the $S_k$ groups, such as hierarchical clustering (Langfelder et al. 2007), which can be a convenient approach when dealing with high-dimensional genomic data. The smallest possible size for $S_k$ is 1: however since the goal is to identify groups of correlated variables and not singletons or pairs, the minimal groups size can be user-defined. Here we set this to be equal to $\sqrt{J}$, where $J$ is the total number of variables; this is a longstanding common choice in statistics and machine learning practice to select the minimum size of subset of variables (Summerfield and Lubin 1951; Guyon and Elisseeff 2003).

The GIA is an ordination algorithm and takes as input precomputed correlations, hence its performance is not affected by the noise in the data (only the estimation of correlation is) nor by the number of variables considered, although the computational cost increases with the number of variables. We remark that (group-wise) sparsity is a property of the data and not of the method used for the analysis of data. The GPCA approach has the advantage that situations where sparse modeling is not suitable can be easily detected through the correlation (association) map $\mathbf{M}$: in absence of correlation structure or when the $S$ groups contain mostly singletons or very few variables GPCA (and GASCA) are not recommended to analyze the data.

Once the $S_k$ have been defined, the GPCA algorithm first computes $K$ candidate loading vectors, where the $k$-th loading vector has non zero elements associated to the variables in the $k$-th group $S_k$ (the loadings for variables which are not in $S_k$ are then set to zero.) From these, only the loading with the largest explained variance is retained in the model and it is used to deflate data matrix $\mathbf{X}$. The complete GPCA algorithm is outlined in the Appendix.

The parameter $\gamma$ is user defined and can be determined by visually inspecting the correlation or the MEDA map $\mathbf{M}$ (or any other association map) and the output of the GIA, since $\gamma$ simultaneously controls both the size and the number of groups of correlated variables. This approach is consistent with the exploratory data analysis philosophy under with the group-wise PCA was developed. Moreover, $\gamma$ has a direct interpretation as a threshold on the strength of the correlation, and it may be easier to tune than the regularization parameters that characterize other sparse implementations of PCA. In the Sect. 4 we guide the reader through the selection process of the $\gamma$ parameter during the analysis of simulated and experimental data.

## 2.4 The ANOVA simultaneous components model

ANOVA-simultaneous component analysis (ASCA) (Smilde et al. 2005) aims to overcome the limitations of MANOVA reducing the original number $J$ of variables in the effect (interaction) matrices $\mathbf{X}_i$ by replacing them with a lower number ($H << J$) of principal components. In this way it is possible to explore the relationship among variables and their contribution to variability observed in the data even in the case of singular data covariance matrices. This is accomplished by fitting a PCA model (see Eq. 2) to each of the effect (interaction) matrices $\mathbf{X}_i$ in the model given by Eq. (1). The ASCA decomposition in principal components is given by

$$\mathbf{X}_i = \mathbf{T}_i^{\mathrm{PCA}} \left(\mathbf{P}_i^{\mathrm{PCA}}\right)^{\mathrm{T}} \qquad i \in \{\alpha, \beta, \alpha\beta\} \tag{5}$$

where $\mathbf{T}_i^{\mathrm{PCA}}$ and $\mathbf{P}_i^{\mathrm{PCA}}$ are the scores and loadings matrices of a PCA model fitted on the effect (interaction) matrix $i$.

Note that here we have dropped for convenience the subscript $H$ referring to the dimensionality of the PCA model as given in Eq. (2). It is intended that $H$ components are retained to fit the ASCA model and that the number of components used can be different for different effect matrices (thus $H = H_i$).

Since the PCA models are fitted to the effect matrices which contain the averages of variables within the same factor levels, the variation between replicates in each level is lost. However this information can be retrieved by projecting the effect matrix ($\mathbf{X}_i$) plus the residual matrix ($\mathbf{E}$) onto the space defined by the loading matrix $\mathbf{P}_i$ for the PCA model for $\mathbf{X}_i$ as first proposed by Zwanenburg *at al.*(Zwanenburg et al. 2011):

$$\mathbf{Y}_i = (\mathbf{X}_i + \mathbf{E})\mathbf{P}_i^{\mathrm{PCA}} = \mathbf{T}_i^{\mathrm{PCA}} + \mathbf{E}\mathbf{P}_i^{\mathrm{PCA}} \qquad i \in \{\alpha, \beta, \alpha\beta\}. \tag{6}$$

The projection $\mathbf{Y}_i$ represents the variability of the replicates in terms of the loadings $\mathbf{P}_i$ of the PCA model for $\mathbf{X}_i$.

*Important note*: for simplicity of illustration we show here a 2-way ASCA model but ASCA can be applied to designed experiments with an arbitrary number of factors and levels.

## 2.5 The group-wise ANOVA simultaneous component analysis

While ASCA is well suited for analyzing designed experiments, the ASCA model may be complicated to interpret since the principal components are linear combinations of all the variables: to overcome this limitation we propose here to replace the PCA step in ASCA with GPCA to arrive at an ASCA solution which is sparse in a group-wise sense.

With respect to the 2-way model considered in Eq. (1) the GASCA model assumes the form

$$\mathbf{X}_i = \mathbf{T}_i^{\mathrm{GPCA}}\left(\mathbf{P}_i^{\mathrm{GPCA}}\right)^{\mathrm{T}} \qquad i \in \{\alpha, \beta, \alpha\beta\} \tag{7}$$

where $\mathbf{T}_i^{\mathrm{GPCA}}$ and $\mathbf{P}_i^{\mathrm{GPCA}}$ are the scores and loadings matrices for the *i*-th factor effect (or factor interaction matrix) obtained using GPCA (see Eqs. (12) and (11) in the Appendix). Operatively, a group-wise ASCA consists of three steps:

1. Definition of the correlation/association maps (matrices) $\mathbf{M}_i$, one for each effect and interaction matrices
2. Setting the convenient $\gamma$ parameter to define the number and the size of the $S_1, S_2, \dots, S_k$ groups of correlated variable for each maps
3. Fitting of the GASCA model to obtain one set of loadings for each effect and interaction matrices.

We consider here two approaches to define the association maps $\mathbf{M}_i$. The first is to derive $\mathbf{M}_i$ starting from correlation (or any other association measure) calculated from the effect matrices $\mathbf{X}_i$ which are intrinsically low noise: this is fully consistent with the ASCA framework. An alternative approach is to obtain $\mathbf{M}_i$ from the sum of the factor (and interaction) matrices and residual matrix. As an example, for factor $\alpha$ in the two-way MANOVA design from Eq. (1), $\mathbf{M}_\alpha$ can be obtained from

$$\mathbf{E}_\alpha = \mathbf{X} - \mathbf{1}\mathbf{m}^{\mathrm{T}} - \mathbf{X}_\beta - \mathbf{X}_{(\alpha\beta)} \tag{8}$$

Note that this is the same matrix used to calculate the scores. When this method is used, the MEDA approach is better suited to define association among variables than standard correlations. This approach should be used in the case of a design with factors with two levels. This is because with two levels, the correlation cannot be computed from the effect matrix $\mathbf{X}_i$, since this will always result in a matrix containing only − 1 and 1 values arising from the design and not from the biology of the data. In the Sect. 4 simulated and real data are analyzed using both approaches and which method to use depends on the nature of the data.

Finally, these three steps should be preceded by a statistical validation of the multivariate effect: this is illustrated in the next section. It should be noted that when the design is not balanced (i.e. when there is not the same number of observations for each factor level) the effects estimates are not orthogonal and fitting the model becomes cumbersome and requires *ad-hoc* approaches (Rawlings et al. 2001). The ASCA framework has been extended to work with unbalanced design (Thiel et al. 2017); an equivalent approach can be used for GASCA. For the sake of simplicity GASCA has been illustrated with a 2-way ANOVA design, but it is generalizable to any number of factors and levels and the software code provided (see Sect. 3.3) will work with a general *N*-way design.

## 2.6 Validation of multivariate effects

Since GASCA is designed to obtain sparse models of the effect matrices obtained from designed *omics* experiments, it is necessary, before fitting a GASCA (or an ASCA) model, to validate whether the levels observed in the sample reflect effects specific in the population or originate by sampling fluctuations. This problem has been addressed in the ASCA context (Vis et al. 2007) and the solution proposed transfers directly to the GASCA case. Following (Vis et al. 2007) we employ a permutation approach to assess the statistical significance of the high-dimensional effects observed in GASCA since the standard MANOVA approach based on the multivariate extension *F*-test can not be applied in this framework because the number of variables is larger than the number of samples. The permutation approach has several advantages: it is optimal for small data sets, is free of distributional assumptions, and gives exact probability values (Berry et al. 2016).

The procedure validates the ANOVA partitioning of the data and should be performed before fitting the GASCA model to the data since it makes no sense to fit a model to effect (interaction) matrices which do not contain significant factor effects but are likely to contain sampling and/or measurement noise.

# 3 Material and methods

## 3.1 Experimental plant data

### 3.1.1 Experimental design

This data set contains the time-resolved metabolomic response of *Arabidopsis thaliana* towards changing light and/or temperature (Caldana et al. 2011). The original data comprises both metabolomic and transcriptomic data measured under four different light conditions (D: dark, LL: low light, L: light, and HL: high light) at three different temperatures (4°C, 21°C, 32°C) with different growth time from 0 to 360 min for a total of 19 time points. We consider here only data acquired at 21° and at time points (t = 0, 5, 10, 20, 40, 80, 160 min) under the four light conditions. The data here analyzed has a design with two factors (light condition and time) with 7 and 4 levels, respectively. The complete data is available through the original publication (Caldana et al. 2011).

### 3.1.2 Missing data imputation

There were 147 missing values in the data set: since removing observations with a missing value would drastically reduce the number of observations, we imputed the missing values replacing them with the average value of the cell and adding a random number drawn from a normal distribution with 0 mean and variance equal to the data cell variance.

### 3.1.3 Data cleaning

There are six biological replicates for each factor level, except for the LL level which has only 5. Since a balanced design is need for both ASCA and GASCA, we randomly removed 1 observation from the factors with 6 replicates. Data for starting condition ($t = 0$ min) was given once only for level $L$ of the light condition factor and was replicated for all the remaining level (hence, the data for first time point is identical for all light conditions) to remove imbalance. Two variables (raffinose and glycine) were removed from the data set because we suspected that something went wrong with the measurement and/or the quantification (data analyzed using PCA): several observations where characterized by disproportionately high values (up to 2 order of magnitudes) for these metabolites which were discarded. A problem with the measurement of glycerol for light condition L and $t = 10$ min was also detected. The value was replaced with the cell average. After this correction no more outliers were evident. The final data matrix has dimensions $140(= 4 \times 7 \times 5) \times 67$.

### 3.1.4 Data pre-processing

Metabolite abundances were normalized by dividing each raw value by the median of all measurements of the experiment for one metabolite.

### 3.1.5 Experimental details

For convenience of the reader we give a short summary of the experimental setup. We refer to the original publication (Caldana et al. 2011) for more details. Plants grown at 21 °C with a light intensity of 150 $\mu\text{E} \times \text{m}^{-2} \times \text{s}^{-1}$ were either kept at this condition or transferred into seven different environments (4 °C, darkness; 21°, darkness; 32 °C, darkness; 4 °C, 85 $\mu\text{E} \times \text{m}^{-2} \times \text{s}^{-1}$; 21 °C, 75 $\mu\text{E} \times \text{m}^{-2} \times \text{s}^{-1}$; 21 °C, 300 $\mu\text{E} \times \text{m}^{-2} \times \text{s}^{-1}$; 32 °C, 150 $\mu\text{E} \times \text{m}^{-2} \times \text{s}^{-1}$.

Metabolites were extracted from single rosettes in a total of six replicates. Extraction and derivatization of metabolites from leaves using GC–MS were performed as previously reported (Lisec et al. 2006). GC–MS data were acquired on a Agilent 7683 series autosampler coupled to an Agilent 6890 gas chromatograph Leco Pegasus two time-of-flight mass spectrometer; acquisition parameters were as reported in (Weckwerth et al. 2004). Peak detection, retention time alignment and library matching were obtained using the TargetSearch package (Cuadros-Inostroza et al. 2009).

## 3.2 Experimental human data

### 3.2.1 Experimental design

We randomly selected two subjects from the METREF study (Assfalg et al. 2008) where 22 healthy subjects were sampled for their urinary profile on 40 consecutive days. The data has a 1-way ANOVA design with two level (Subject 1 and Subject 2). The data is available through the KODAMA R package (Cacciatore et al. 2017).

### 3.2.2 Data pre-processing

Bucketing was applied to the NMR spectra after the removal of region with $\delta > 9.5$ ppm, $4.5 < \delta < 6.0$ppm, and $\delta < 0.5$ ppm, containing water and urea signals. each spectrum was divided into sequential bins of 0.02 ppm width, which were integrated using AMIX software (Bruker BioSpin). Finally, total area normalization was carried out on all the spectra. Further, bins corresponding to noise and empty spectral areas were removed to reduce dimensionality. To make the design balanced, 37 spectra for each subject were considered: the final dataset has size $74 \times 206$.

### 3.2.3 Experimental details

$^1$H NMR spectra were acquired using a Bruker 600 MHz metabolic profiler (Bruker BioSpin) operating at 600.13 MHz proton Larmor frequency and equipped with a 5 mm CPTCI cryoprobe. Each urine sample was acquired with a NOESY-presaturation pulse sequence. Details on sample preparation and further information on the NMR experimental setup can be found in the original publication (Assfalg et al. 2008) and in other publications where the data has been analyzed (Bernini et al. 2009; Ghini et al. 2015).

## 3.3 Software

The GPCA, ASCA and GASCA and GIA algorithms are freely available in the Matlab MEDA toolbox (Camacho et al. 2015) at the address: github.com/josecamachop/MEDA-Toolbox. The GASCA code is based on the original Matlab code for ASCA by G. Zwanenbourg (Zwanenburg et al. 2011). The function to call is `gasca`: typing `help gasca` in the Matlab command windows will prompt instructions and a worked out example to perform GASCA.

# 4 Results and discussion

## 4.1 Simulations

We begin presenting the GASCA of a simple simulated data set to show how GASCA models data which is sparse in a group-wise fashion: the data follows a two factor design ($\alpha$ and $\beta$), with four and three levels, respectively, and no interaction, with one group of correlated variables contributing only to the first factor, and another group of variables contributing only to the second factor; both groups consists of five variables. There are 100 observations and 50 variables; the two factors are both significant at the 0.01 level.

The analysis starts with the construction of the association matrices $\mathbf{M}_\alpha$ and $\mathbf{M}_\beta$ for the effect matrices $\mathbf{X}_\alpha$ and $\mathbf{X}_\beta$. As discussed in the Sect. 2.3, there are several strategies to construct such matrices: for this example we build MEDA maps (see Eq. 3) starting from the matrices $\mathbf{E}_\alpha$ and $\mathbf{E}_\beta$ defined in Eq. (8). Another approach, based on correlations, will be shown in the analysis of the plant metabolomics data (see Sect. 4.4). The MEDA maps for the two factor are shown in Fig. 1 panels a and d, respectively. The two groups of associated variables are evident: the threshold $\gamma$ controlling the sparsity of the solution can be chosen by inspecting the MEDA maps and we set $\gamma = 0.8$ for both factors. This is a rather straightforward situation; association maps for real data, especially for metabolomics data, are usually more complicate: a guided procedure to select $\gamma$ will be shown in Sect. 4.4. The score plots resulting GASCA models for the two factors are shown in Fig. 1 in panel b and e, while the loadings are given in panels c and f. The GASCA solution is sparse in a group-wise fashion, with just one group of variables contributing to each factor, which greatly facilitate interpretation, correctly retrieved by the model.

## 4.2 PCA modeling of the plant data

In the following we present a comparison of PCA, ASCA and GASCA models obtained on a designed plant metabolomic experiment. We begin by fitting a standard PCA model (see Eq. 2) to the data. The scatter plot of the first two principal components and the corresponding loading vectors are given in Fig. 2. It appears that a simple PCA is not well suited for analyzing this data since it does not distinguish between the groups in the data: factors and levels are mixed up in the score plot. Moreover, loadings are complicate to interpret since all variables contribute to the final model.

## 4.3 ASCA modeling of the plant data

Before applying ASCA (and, of course, GASCA) we test the significance of effects for the two factors of the experimental design (light conditions and time) and their interaction. Applying a permutation test with $n_{perm} = 10^4$ to test the significance of the factors the calculated $P$ values are 0.0001, 0.0001 and 0.0278 for light condition, Time and their interaction, respectively. Since all factors and interactions are significant we will fit the ASCA (and later GASCA) model on all effect and interaction matrices, *i.e* $\mathbf{X}_\alpha$, $\mathbf{X}_\beta$ and $\mathbf{X}_{\alpha\beta}$.

The *Arabidopsis* data follows a two factors design (time and light condition) with 4 and 7 levels respectively, thus there are two matrices for the effects ($\mathbf{X}_\alpha$ and $\mathbf{X}_\beta$ and one matrix for the interaction $\mathbf{X}_{\alpha\beta}$.) The overall mean explains the 86.7% of the total sum of squares, the two factors 0.86% and 1.3%, the interaction 2.1% and the residuals 9.1%.

A scatter plot of the first two ASCA components and the corresponding loading vectors are given in Fig. 3 for the factor 1 (light conditions); ASCA is able to resolve the different levels of the treatment. but the interpretation is not straightforward. As almost all metabolites contribute to the model (i.e. have non zero loadings), this makes hard to identify which metabolites are important to explain the systematic variation induced on the system by manipulating the light condition. In the next section we show how GASCA can simplify data understanding and interpretation.

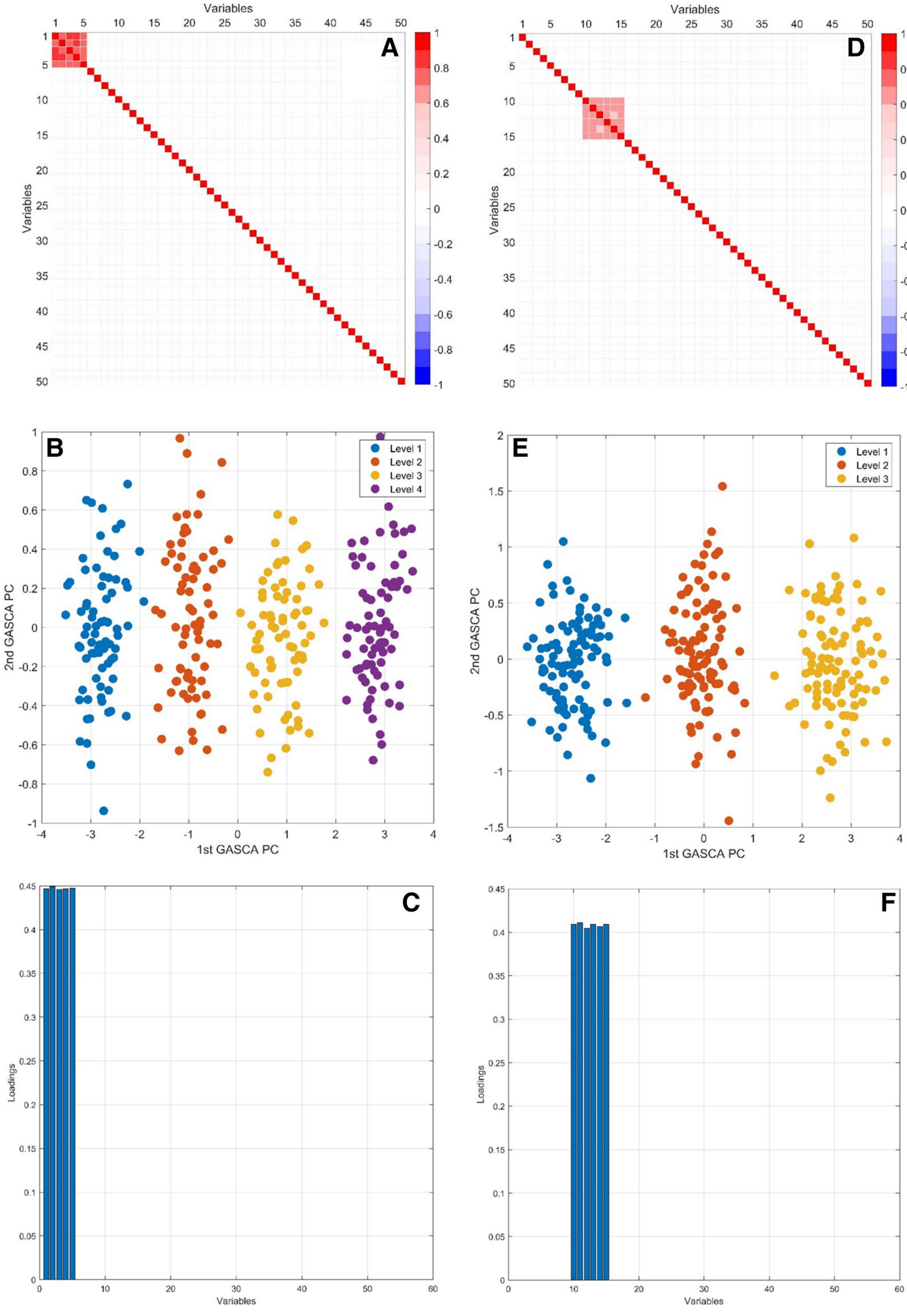


**Fig. 1** GASCA modeling for the first factor of the simulated examples: **a** MEDA map built from the residuals; **b** score plot **c** loadings. GASCA modeling for the second factor of the simulated examples: **d** MEDA map built from the residuals; **e** score plot and **f** loadings

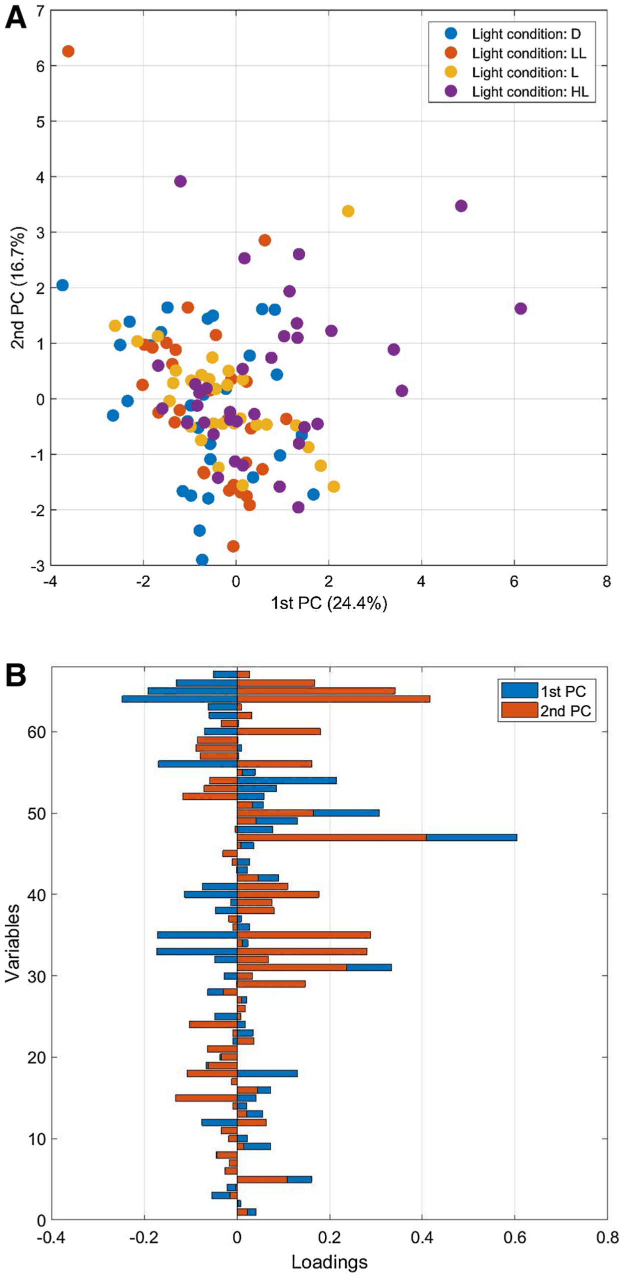


**Fig. 2** PCA model for the *Arabidopsis* data: **a** scores and **b** loadings. Only factor 1 (light condition) is color coded. The levels for factor 1 are: dark (D), light (L), low light (LL) and high light (HL)

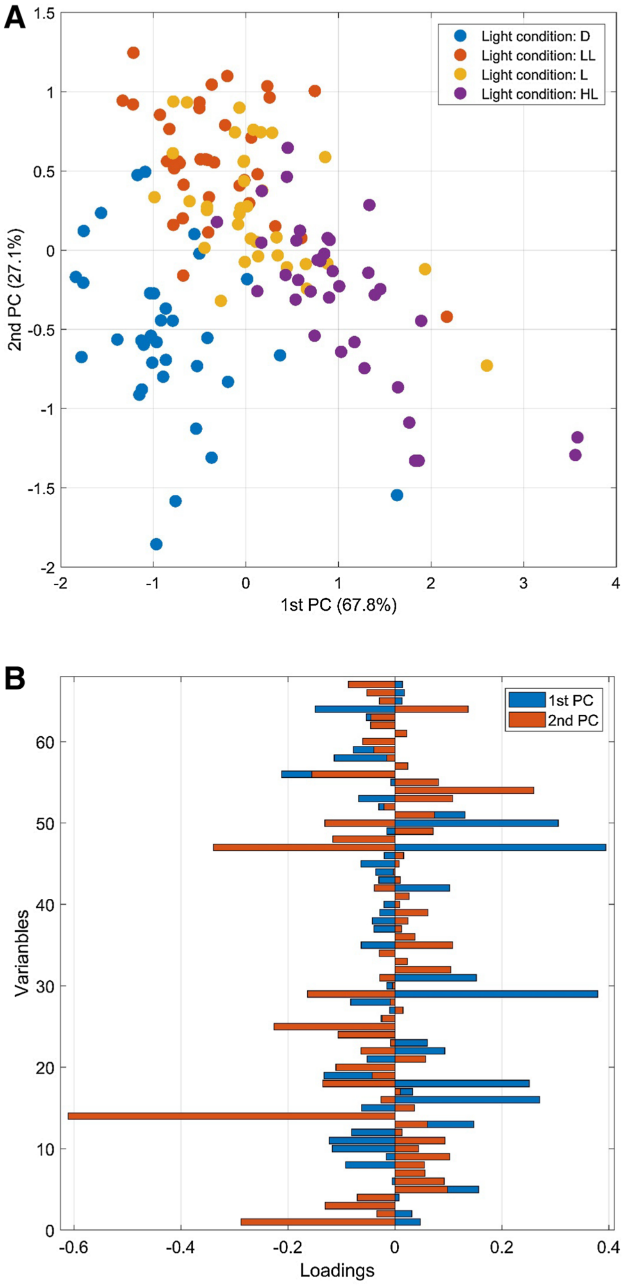


**Fig. 3** ASCA model for the *Arabidopsis* data: **a** scores and **b** loadings. Only factor 1 (light condition) is color coded. The levels for factor 1 are: dark (D), light (L), low light (LL) and high light (HL)

## 4.4 Group-wise ASCA modeling of the plant data set

Once the significance of the effects is assessed (this has been already done in the ASCA modeling showed in the previous section), the group-wise ASCA modeling starts from the construction of the variable association map **M** from the original data. Since there is one GASCA model for each effect matrix and for each interaction (see model in Eq. 1) there are three maps ($\mathbf{M}_i$ with $i \in \{\alpha, \beta, \alpha\beta\}$) to be built. As detailed in the Sect. 2 we use Spearman correlation to quantify the relationships among the metabolites, retaining only those which are statistically significant (see Eq. 4)

The Spearman correlation maps $\mathbf{M}_i$ are shown in Fig. 4. Several groups of (highly) correlated variables are evident

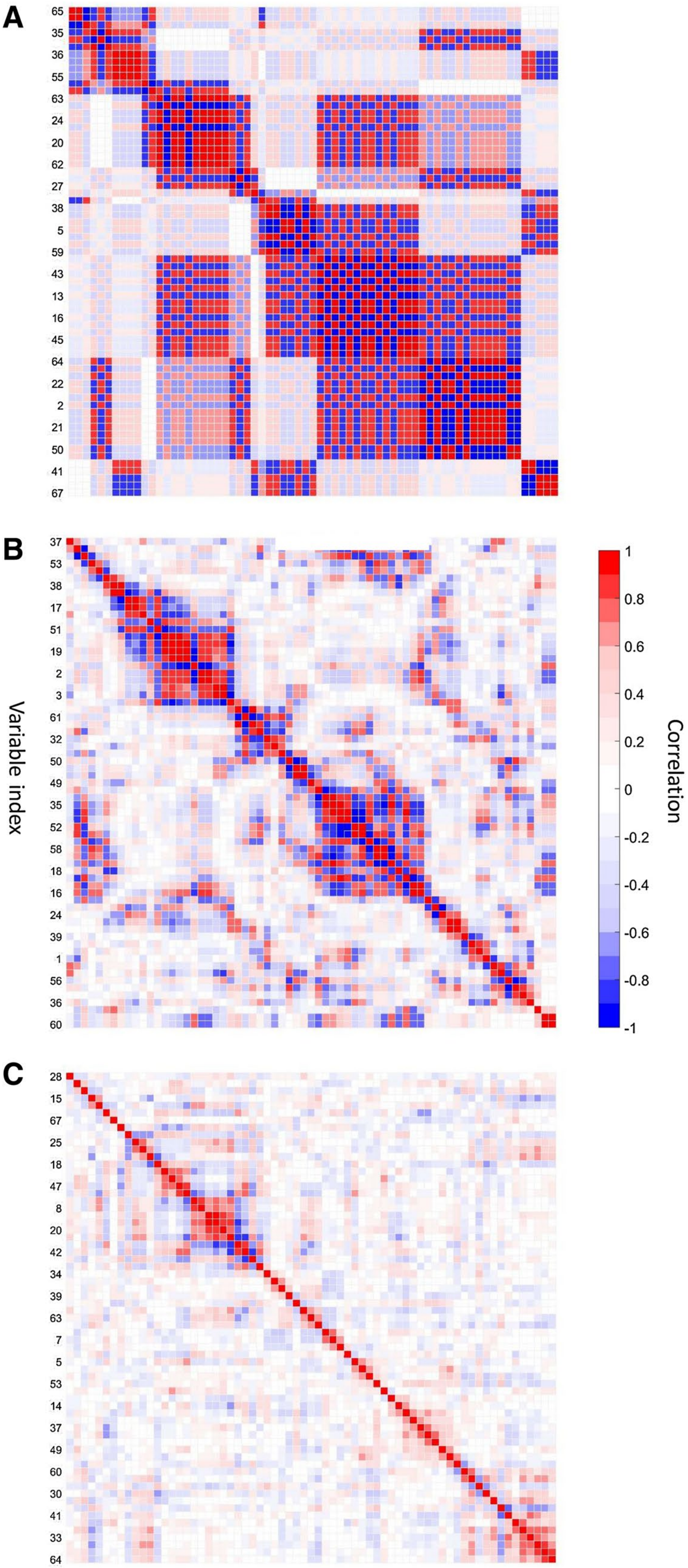
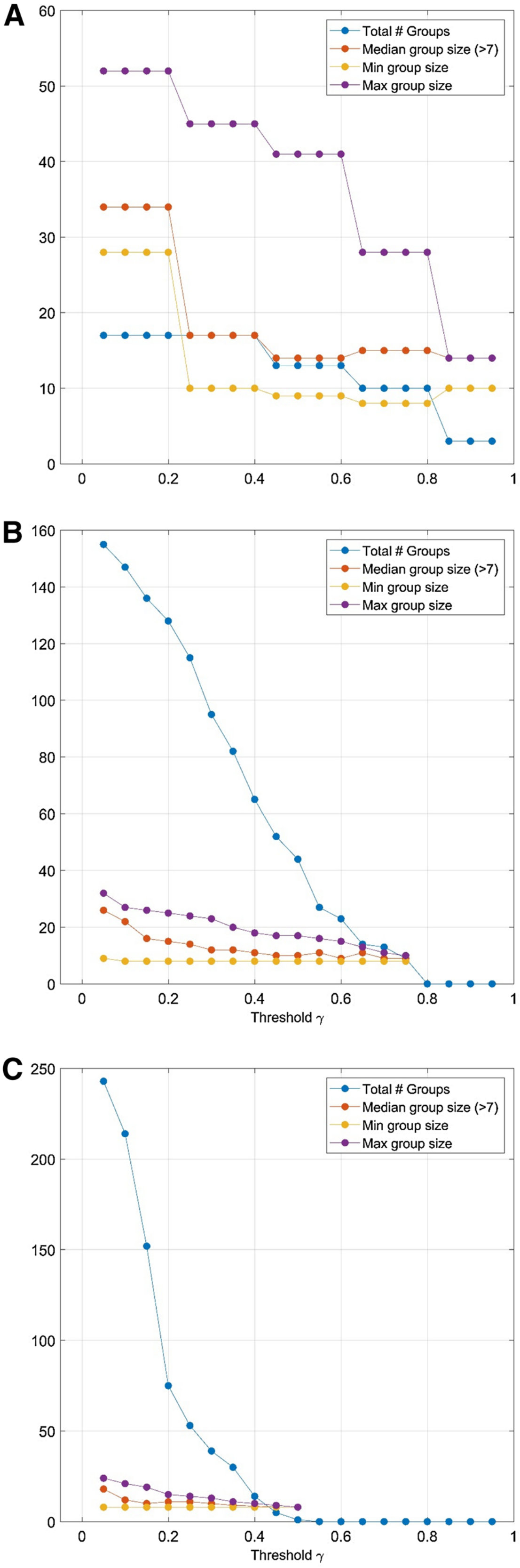


**Fig. 4** MEDA maps for **a** factor 1, light condition; **b** factor 2, iime; **c** interaction light×Time

**Fig. 5** Number of groups $S_k$ and median number of variables per ▶ group as a function of $\gamma$ for the correlation maps (see Eq. 4 and accompanying text for more details) of: **a** factor 1, light condition; **b** factor 2, iime; **c** factor interaction (light condition × time)



in all three cases; however there are many variables not contributing to the correlation structure of the data. This is a situation in which a GPCA implementation in ASCA is adequate because data is sparse in a group-wise fashion

The second step in GASCA is the selection of appropriate $\gamma$ values to control the sparsity of the solution. Since there are three maps, three values need to be chosen. Values can be selected by visually inspecting the map or by exploring the number of groups and their size as a function of $\gamma$: this is shown in Fig. 5. For the correlation map for factor 1 (light condition, panel a in Fig. 5) the median size of the $S_k$ groups is roughly constant for $\gamma > 0.4$; however the number of variables in the groups decreases sharply when increasing $\gamma$, as expected. We choose $\gamma_\alpha = 0.85$ which gives a good compromise between the number of groups and their size, with not too many groups of moderate size. We use the subscript $\alpha$ to emphasize that this value of $\gamma$ is specific for the first factor, indicated with $\alpha$ in the models given by Eq. (7)

For factor 2 (Time, panel b in Fig. 5) the number of groups sharply decreases with $\gamma$ ( which indicates lower correlation among the variables) while the median, maximum and minim size remains approximately constant. We set $\gamma_\beta = 0.7$ not to have too many groups. Both $\gamma_\alpha$ and $\gamma_\beta$ values are also in line with what can be inferred by visually inspecting the correlation plots from Fig. 4, like first suggested in the original publication of GPCA (Camacho et al. 2017)

For the interaction (light condition × Time, panel c in Fig. 5) the total number of groups decrease with $\gamma$ while the median, maximum and minimum size remains approximately constant. From the visual inspection of the correlation map in Fig. 4 panel c, it can be seen that there are very few groups of correlated variables, so we set $\gamma_{\alpha\beta} = 0.45$.

In general $\gamma$ should be set in such a way not to have too many groups containing only one or two variables. Because sparsity is an inherent property of the data also $\gamma$ is data specific and there is not a general rule to define the appropriate values which need to be specified with respect with the data at hand. However, since $\gamma$ is a threshold on the correlation magnitude, its value can be seen in context with what observed in metabolomics studies: Camacho (Camacho et al. 2005) suggested to divide correlations values into three levels: low ($|\rho| \leq 0.6$), medium ($0.6 < |\rho| < 0.8$), and high ($|| \geq 0.8$) based on metabolic modeling considerations. In general, metabolomics data are abundant in low

**Table 1** Loadings for the two first components of GASCA model for the light condition and time effect matrices and their interaction

| # | Metabolites | Factors and interactions | | | | | |
|---|---|---|---|---|---|---|---|
| | | Light | | Time | | Light × Time | |
| | | PC1 | PC2 | PC1 | PC2 | PC1 | PC2 |
| 1 | 4-Hydroxy-benzoic acid | | | | 0.360 | | |
| 2 | 4-Hydroxycinnamic acid | | | | 0.365 | | |
| 3 | Alanine | | | | | | 0.648 |
| 4 | Arabinose | | | | | | |
| 5 | Arabitol | 0.186 | | | | | |
| 6 | Ascorbic acid | | − 0.061 | | | | |
| 7 | Asparagine | 0.128 | | − 0.156 | | | |
| 8 | Aspartate | | − 0.549 | | − 0.195 | | |
| 9 | Benzoic acid | | − 0.155 | 0.117 | | | |
| 10 | Beta-alanine | | | − 0.491 | | − 0.268 | − 0.071 |
| 11 | Citramalate | | − 0.256 | | | | |
| 12 | Citric acid | | | | | | |
| 13 | Citrulline/arginine | | | − 0.084 | | | |
| 14 | Dehydroascorbic acid | | 0.293 | | | | |
| 15 | Dehydroascorbic acid dimer | | | | | | |
| 16 | Docosanoic acid | | | | 0.258 | | |
| 17 | Erythritol | 0.114 | | | | | |
| 18 | Ethanolamine | 0.184 | | | − 0.188 | | |
| 19 | Fructose | | 0.262 | | | | 0.445 |
| 20 | Fucose | | − 0.124 | | | | |
| 21 | Fumaric acid | 0.084 | | | | | |
| 22 | Gaba | 0.162 | | | | − 0.320 | − 0.074 |
| 23 | Galactinol | | | | − 0.109 | | |
| 24 | Galactose | | | | | | |
| 25 | Gluconic acid | | − 0.228 | | | | |
| 26 | Glucose | | 0.315 | | | | 0.580 |
| 27 | Glutamate | | − 0.164 | | | | |
| 28 | Glutamine | | 0.445 | | | | |
| 29 | Glycerol | | | 0.202 | | | |
| 30 | Glycolic acid | − 0.475 | | | | | |
| 31 | Hexacosanoic acid | | 0.051 | | | | |
| 32 | Hydroxyproline | | | | | | |
| 33 | Indole-3-acetonitrile | | | | | | |
| 34 | Isoleucine | | | − 0.344 | | − 0.334 | |
| 35 | Itaconic acid | | | | 0.336 | | |
| 36 | Lactic acid | | | | | | |
| 37 | Leucine | | | − 0.371 | | − 0.369 | |
| 38 | Lysine | 0.214 | | − 0.352 | | − 0.449 | − 0.005 |
| 39 | Maleic acid | | | | | | |
| 40 | Malic acid | | | | | | |
| 41 | Maltose | 0.061 | | | − 0.355 | | |
| 42 | Mannitol | | | | | | |
| 43 | Methionine | − 0.340 | | | | 0.308 | 0.092 |
| 44 | Myo-inositol | | | | | | |
| 45 | Nicotinic acid | | | | | | |
| 46 | O-acetyl-serine | | | | 0.254 | | |
| 47 | Octacosanoic acid | | | | | | |
| 48 | Octadecanoic acid | | | | | | |

**Table 1** (continued)

| # | Metabolites | Factors and interactions | | | | | |
|---|---|---|---|---|---|---|---|
| | | Light | | Time | | Light × Time | |
| | | PC1 | PC2 | PC1 | PC2 | PC1 | PC2 |
| 49 | Ornithine | | | | | | |
| 50 | Palmitic acid | | | | | | |
| 51 | Phenylalanine | – 0.505 | | | – 0.178 | | |
| 52 | Proline | | | | | | |
| 53 | Putrescine | | | | – 0.284 | | |
| 54 | Pyruvic acid | | | | | | |
| 55 | Serine | | | | | | |
| 56 | Shikimate | – 0.236 | | | | 0.335 | 0.163 |
| 57 | Similar to adenine | | | | | | |
| 58 | Sinapic acid | | | | | | |
| 59 | Succinic acid | | | | | – 0.251 | |
| 60 | Sucrose | – 0.328 | | 0.167 | | | |
| 61 | Tetracosanoic acid | | | | 0.320 | | |
| 62 | Threonic acid | | – 0.149 | | | | |
| 63 | Threonine | | | | | | |
| 64 | Trehalose | | | | | | |
| 65 | Tyrosine | | | – 0.405 | | – 0.325 | |
| 66 | Uracil | 0.244 | | | | | |
| 67 | Valine | | 0.185 | – 0.318 | | | |

Null loadings are omitted

correlations as a result of the systemic nature of metabolic control (Camacho et al. 2005) and thus the low values observed for correlation map of the interaction terms are not unexpected.

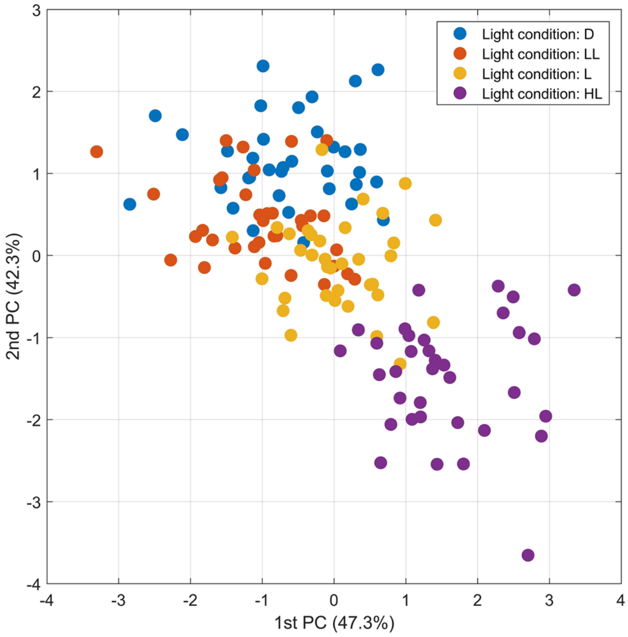


**Fig. 6** Score plot for the GASCA model for the first factor (light condition) of the plant data

Once the $\gamma$ values have been specified the GASCA model can be finally obtained: the loadings of the model (see Eqs. 7 and 11) are given in Table 1. Comparing the loadings of ASCA (Fig. 3) and GASCA models it is evident the gain in simplicity and interpretability of the solutions.

The panel of measured metabolites analyzed here covers only a tiny fraction of the thousands of low molecular weight compounds produced by plants. However, several interesting observations can be made by analyzing the loadings (see Table 1), which describe the relative contribution of each variable to explain the variation observed in the data, associated to the different metabolites in the GASCA model. The score plot for the light condition factor is given in Fig. 6. The plot is slightly dissimilar from the one obtained for the standard ASCA model (see Fig. 3): however, separation among the different factor levels is evident, it should be remember that only subsets of variables are used in GASCA, hence differences will be observed among ASCA and GASCA score plots while interpretability is increased.

The loadings of the first component for the GASCA model for the light condition factor indicate substantial contribution of phenylalanine and shikimate whose pathways are indeed strongly interlinked (Tohge et al. 2013) and found to be affected by light (Caldana et al. 2011). Glycolic acid, a product of photosynthesis (Jensen and Bassham 1966), has also a high loading and can be an indicator of

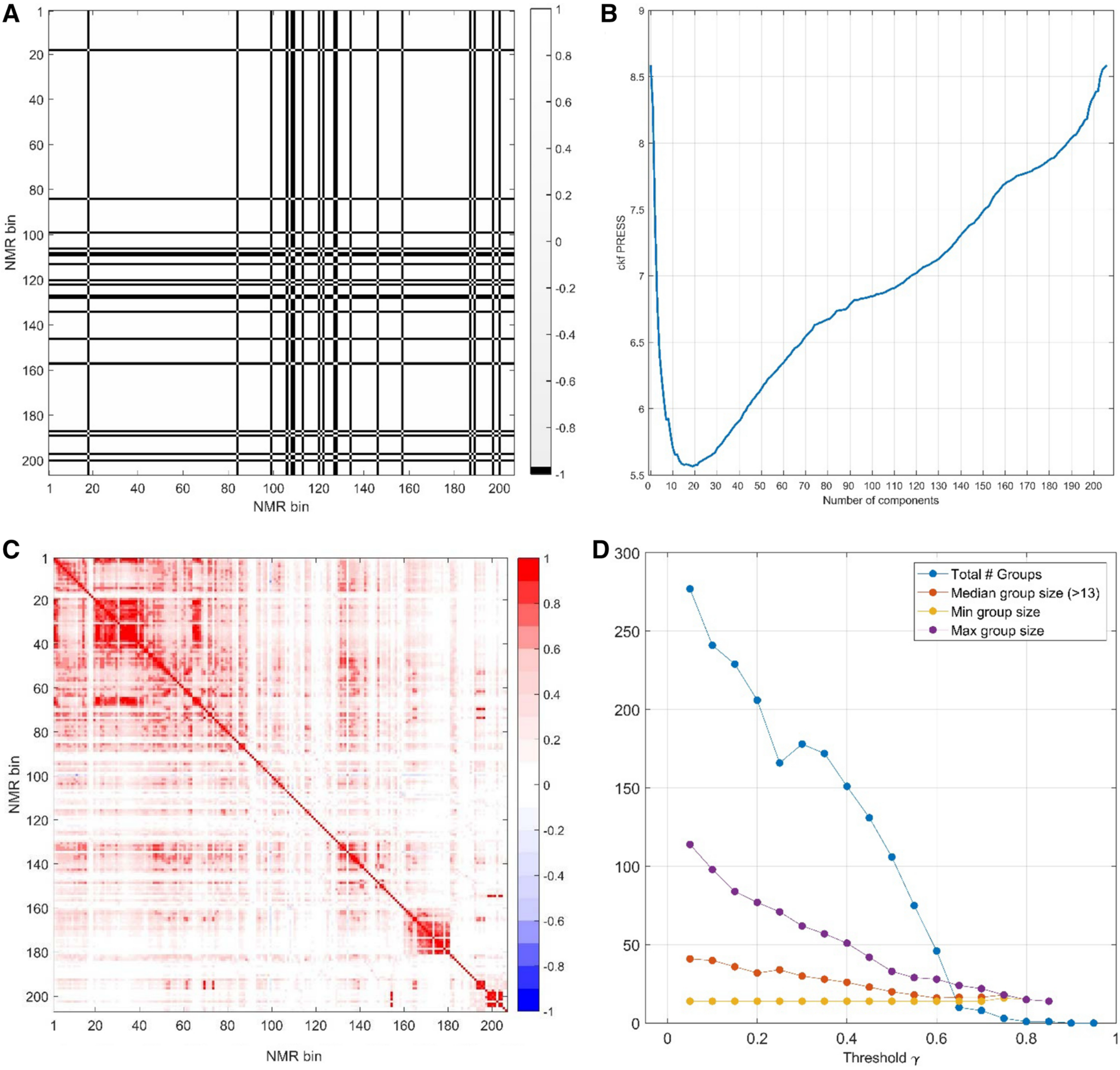


**Fig. 7** **a** Correlation matrix obtained from the effect matrix for the human data. **b** Cross-validation plot for the residual matrix (see Eq. in the main text.) **c** MEDA map fitted with 20 components. **d** Number of groups $S_k$ and median number of variables per group as a function of $\gamma$

varying photosynthetic activity depending on the light levels.

The second components contains contributions from sugars (glucose, fructose, fucose), which is not surprising since plants transform carbon dioxide into sugars which are then used as energy source. Notably glutamine has here the larger loading, and this may indicate reduced glutamine synthetase activity, which plays a role in the control of photosynthetic responses to high light (Brestic et al. 2014).

Arginine is metabolically connected to glutamate: glutamate is used to synthesize ornithine from which arginine synthesis follows with citrulline as intermediate (Winter et al. 2015). Interestingly, ornithine does not contribute to the model indicating that likely the variance in the enzymes that control the reaction affects both metabolites in equal amounts and different directions resulting in very low correlation among these metabolites. Overall, the arginine biosynthesis in plant is poorly understood (Winter et al. 2015) and light modulating effects have been suggested (Frémont et al. 2013).

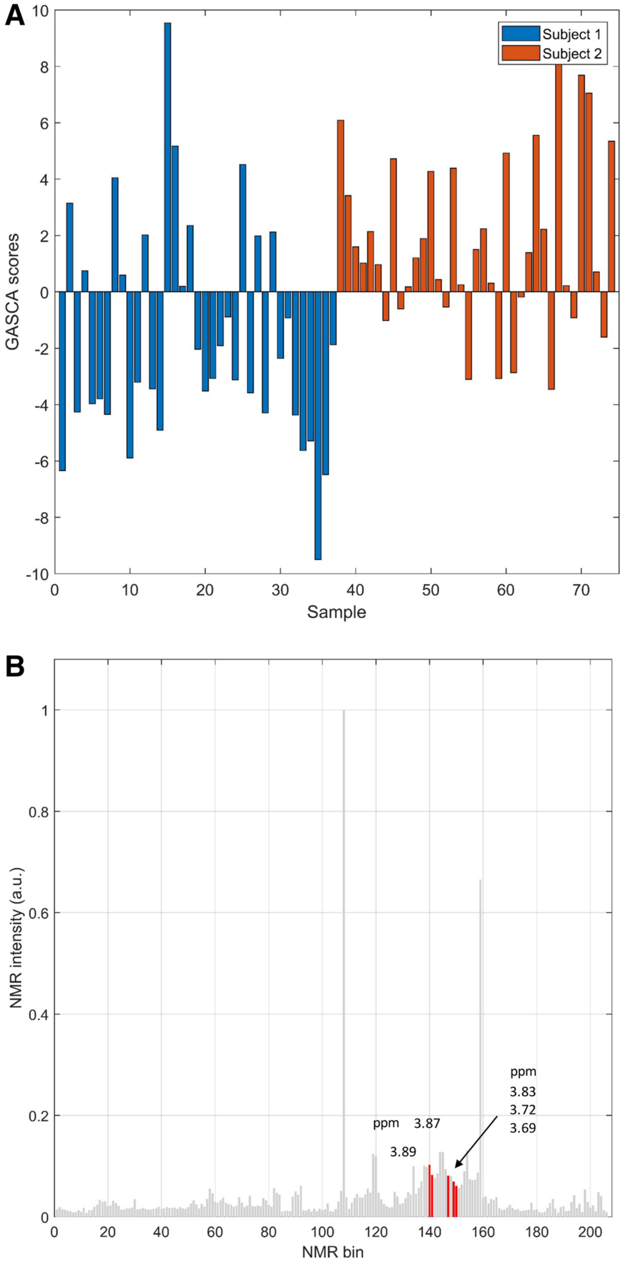


**Fig. 8** **a** Scores for the GASCA model for the human data. **b** Average NMR spectrum: the bins/ppm corresponding to the non-zero loadings for the GASCA component are highlighted in red

Concerning the factor Time (minutes of growth under a given light condition), the first component is dominated by benzoic acid whose synthesis pathway in plants has yet to be fully characterized (Moerkercke et al. 2009). This component contains also the branched amino acid leucine, isoleucine and valine whose biosynthesis in plants follows the reaction pathways established in microorganisms (Binder 2010) and play a pivotal role in plant development (Singh 1998)

Analysis of the loadings for the first two components for the interaction effect light condition × Time shows that succinate and leucine, isoleucine and methionine have high loading contribution and this indicates a link between amino acids and the tricarboxylic acid (TCA) cycle via succinate which is both light and time dependent. This suggests that amino acids produced by an increased protein biosynthesis may be used to fuel central metabolism (Caldana et al. 2011). Interestingly, also the 4-aminobutyric acid (GABA) has also high loading which suggests a possible role of this compound in fueling the TCA cycle. Using a differential network approach, Caldana and coworkers also highlighted GABA and suggested, building on a previous study (Taylor et al. 2004), that branched amino acids may promote their own degradation to acetyl-CoA, propionyl-CoA and acetoacetate that can subsequently enter the TCA cycle.

It is interesting to note that lysine is the only metabolite that appears in the models for all three effect matrices, thus providing a link between the plant response to light condition, plant development and their interaction. Indeed, it has been shown that lysine metabolism is strongly associated with the functioning of the tricarboxylic acid cycle while being largely disconnected from other metabolic networks (Angelovici et al. 2009): lysine catabolism into the TCA cycle seems to be fundamental for seed and plant development (Galili et al. 2014).

## 4.5 GASCA analysis of the human data set

We present here the analysis of the second experimental data set using GASCA. The experimental design follow a 1-factor ANOVA model with just two levels (subject 1 and 2). In this case is not possible to define a meaningful correlation matrix starting from the effect matrix as noted in the Sect. 2, which is shown in Fig. 7 panel a. For this reason we built the variable association map **M** starting from the matrix

$$\mathbf{E}_{\alpha} = \mathbf{X} - \mathbf{1}\mathbf{m}^{\mathrm{T}}, \tag{9}$$

which the analogue for one-way design of Eq. (8), using the MEDA approach. To determine the optimal number of components to fit the MEDA map we use cross-validation, but other approaches are possible. Figure 7 panel b shows the cross-validation plot, from which we infer 20 to be the optimal dimensionality. This is used to obtain the MEDA map (see Eq. 3) shown in Fig. 7 panel c. As typical for NMR data sets, there is a high degree of correlations: the number of groups sharply decreases with the threshold $\gamma$ and we opt here for a rather sparse model by selecting $\gamma = 0.6$, as shown in Fig. 7 panel d.

Since the design has only two levels, there is only one component in the GASCA model for this data. The monodimensional scores are shown in Fig. 8 panel a where it is evident the separation among the scores corresponding to the NMR spectra belonging to the Subject 1 and 2. The loadings for this component are shown in Fig. 8 panel b: a few ppm are selected (3.69, 3.71, 3.83, 3.87, and 3.89) which correspond to signal from dimethylglycine, citrate, trimethylamine and $\alpha$-ketoglutarate. Citrate and $\alpha$-ketoglutarate are intermediate of the TCA cycle. Dimethylglycine and trimethylamine are two metabolites associated, among others, with the activity of gut microflora, confirming the role of gut microflora activity to the shaping of the individual urinary metabolic phenotype (Bernini et al. 2009; Saccenti et al. 2016).

## 5 Conclusions

Designed *omics* experiments are becoming increasingly complex with many factors considered simultaneously and having high dimensional multivariate responses. We have proposed here Group-wise ANOVA simultaneous component analysis (GASCA), an extension of the well established ANOVA-simultaneous component analysis (ASCA), which implements the idea of group-wise sparsity to arrive to solutions which are easier to interpret. The use of GASCA is advisable when data is sparse in a group-wise fashion, that is when there are groups of correlated/associated variables: this can be easily checked by visually inspecting the association maps built from the data. In this case GASCA models are easier to interpret than the ASCA counterpart.

The characteristics of the method are shown through the analysis of a real-life metabolomics experiments concerning the growth of *Arabidopsis thaliana* under different light conditions and phenotyping of healthy subjects. Results are compared with those of classical PCA and ASCA. It is shown that the GASCA models, containing only selected subsets of the original variables, are easier to interpret and describes relevant biological processes. We showed how the selection of closely related variable points to biologically relevant effects that are otherwise lost when all variables are considered. Finally, GASCA is applicable to any kind of *omics* data obtained through designed experiments such as, (but not limited to) gene expression and proteomic data.

**Acknowledgements** The authors wish to thank Gooitzen Zwanenburg for making the data ready for analysis. This work was partially funded by the EU Commission through the FP7 project INFECT (Contract No. 305340) and by the Spanish Ministry of Economy and Competitiveness and FEDER funds through the TIN2014-60346-R project.

## Compliance with ethical standards

**Conflict of interest** The authors declare that they have no conflict of interest.

**Ethical approval** We declare that all authors comply with Springer's ethical policies.

**Research involving human and animal rights** No human participants or animals were involved in the study.

## Appendix

Once the set of groups of variables $S_1, S_2, \dots S_k, \dots S_K$ have been identified, for instance with MEDA and GIA, the algorithmic procedure to obtain a GPCA model from the data matrix $\mathbf{X}$ is as follows:

Step 1 Initialize

$$\mathbf{C} = \mathbf{X}^{\mathrm{T}}\mathbf{X}$$

$$\mathbf{B} = \mathbf{I}$$

where $\mathbf{C}$ has dimension $J \times J$ and $\mathbf{I}$ is a $J \times J$ identity matrix.

Step 2 For the $a$-th component with $a = 1, 2, \dots, A$

Step 2.1 For the $k$-th group $S_k$ of correlated variables $S_k$, with $k = 1, 2, \dots, K$

Step 2.1.1 Build the matrix $\mathbf{C}^k$ from $\mathbf{C}$ by setting to zero the variables not belonging to $S_k$ group.

$$c^k_{lm} = 0,\ \forall l \notin S_k \ or\ \forall m \notin S_k \tag{10}$$

where $c^k_{lm}$ is the $l$-th, $m$-th element of $\mathbf{C}^k$

Step 2.1.2 Perform eigendecomposition of $\mathbf{C}^k$ and select the first eigenvector.

$$\mathbf{C}^k = \mathbf{p}^k(\sigma^k)^2(\mathbf{p}^k)^{\mathrm{T}} + \mathbf{E}^k$$

Step 2.2: Choose the loadings and scores of component $a$ from the group capturing the most variance.

$$\mathbf{p}_a = \mathrm{argmin}_{\mathbf{p}^k} \|\mathbf{E}^k\|_F \tag{11}$$

$$\mathbf{t}_a = \mathbf{X}\mathbf{p}_a \tag{12}$$

Step 2.3: Perform the deflation according to (Mackey, 2008).

$$
\begin{aligned}
\mathbf{q} &= \mathbf{B}\mathbf{p}_a \\
\mathbf{C} &= (\mathbf{I} - \mathbf{q}\mathbf{q}^{\mathrm{T}})\mathbf{C}(\mathbf{I} - \mathbf{q}\mathbf{q}^{\mathrm{T}}) \\
\mathbf{X} &= \mathbf{X}(\mathbf{I} - \mathbf{q}\mathbf{q}^{\mathrm{T}}) \\
\mathbf{B} &= \mathbf{B}(\mathbf{I} - \mathbf{q}\mathbf{q}^{\mathrm{T}})
\end{aligned}
$$

Per each component, the GPCA algorithm computes $K$ potential loading vectors, each of them with non-zero elements only for the set of variables corresponding to group $S_k$. To do that, it discards all the elements of the covariance matrix that do not correspond to variables in $S_k$ and performs a rank-1 eigendecomposition on the resulting covariance. Comparing the resulting $K$ eigenvectors, it selects the one with the highest variance, discarding the rest. Using this loading vector, the complete matrix **C** and the data matrix is deflated following (Mackey, 2008) to continue with successive components.

## References

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, *26*(1), 32–46.

Angelovici, R., Fait, A., Zhu, X., Szymanski, J., Feldmesser, E., Fernie, A. R., et al. (2009). Deciphering transcriptional and metabolic networks associated with lysine metabolism during arabidopsis seed development. *Plant Physiology*, *151*(4), 2058–2072.

Arteaga F (2011) A note on “missing-data theory in the context of exploratory data analysis”. Technical Report MEDA Toolbox

Arteaga, F., & Ferrer, A. (2002). Dealing with missing data in MSPC: Several methods, different interpretations, some examples. *Journal of Chemometrics*, *16*, 408–418.

Arteaga, F., & Ferrer, A. (2005). Framework for regression-based missing data imputation methods in on-line mspc. *Journal of Chemometrics*, *19*, 439–447.

Assfalg, M., Bertini, I., Colangiuli, D., Luchinat, C., Schäfer, H., Schütz, B., et al. (2008). Evidence of different metabolic phenotypes in humans. *Proceedings of the National Academy of Sciences United States of America*, *105*(5), 1420–1424.

Bernini, P., Bertini, I., Luchinat, C., Nepi, S., Saccenti, E., Schafer, H., et al. (2009). Individual human phenotypes in metabolic space and time. *Journal of Proteome Research*, *8*(9), 4264–4271.

Berry, K. J., Mielke, P. W, Jr., & Johnston, J. E. (2016). *Permutation statistical methods: An integrated approach*. Cham: Springer.

Bibby, J., Kent, J., & Mardia, K. (1979). *Multivariate analysis*. London: Academic Press.

Binder S (2010) Branched-chain amino acid metabolism in *Arabidopsis thaliana*. *The Arabidopsis Book*, *8*, e0137

Bratchell, N. (1989). Multivariate response surface modelling by principal components analysis. *Journal of Chemometrics*, *3*(4), 579–588.

Brestic, M., Zivcak, M., Olsovska, K., Shao, H. B., Kalaji, H. M., & Allakhverdiev, S. I. (2014). Reduced glutamine synthetase activity plays a role in control of photosynthetic responses to high light in barley leaves. *Plant Physiology and Biochemistry*, *81*, 74–83.

Cacciatore, S., Tenori, L., Luchinat, C., Bennett, P. R., & MacIntyre, D. A. (2017). KODAMA: An R package for knowledge discovery and data mining. *Bioinformatics*, *33*(4), 621–623.

Caldana, C., Degenkolbe, T., Cuadros-Inostroza, A., Klie, S., Sulpice, R., Leisse, A., et al. (2011). High-density kinetic analysis of the metabolomic and transcriptomic response of arabidopsis to eight environmental conditions. *The Plant Journal*, *67*(5), 869–884.

Camacho, D., De La Fuente, A., & Mendes, P. (2005). The origin of correlations in metabolomics data. *Metabolomics*, *1*(1), 53–63.

Camacho, J. (2010). Missing-data theory in the context of exploratory data analysis. *Chemometrics and Intelligent Laboratory Systems*, *103*, 8–18.

Camacho, J. (2011). Observation-based missing data methods for exploratory data analysis to unveil the connection between observations and variables in latent subspace models. *Journal of Chemometrics*, *25*(11), 592–600. https://doi.org/10.1002/cem.1405.

Camacho, J., Pérez-Villegas, A., Rodríguez-Gómez, R. A., & Jiménez-Manas, E. (2015). Multivariate exploratory data analysis (meda) toolbox for matlab. *Chemometrics and Intelligent Laboratory Systems*, *143*, 49–57.

Camacho, J., Rodríguez-Gómez, R. A., & Saccenti, E. (2017). Group-wise principal component analysis for exploratory data analysis. *Journal of Computational and Graphical Statistics*, *26*, 501–512.

Cuadros-Inostroza, Á., Caldana, C., Redestig, H., Kusano, M., Lisec, J., Peña-Cortés, H., et al. (2009). Targetsearch—A bioconductor package for the efficient preprocessing of GC-MS metabolite profiling data. *BMC Bioinformatics*, *10*(1), 428.

Engel, J., Blanchet, L., Bloemen, B., Van den Heuvel, L., Engelke, U., Wevers, R., et al. (2015). Regularized MANOVA (rMANOVA) in untargeted metabolomics. *Analytica Chimica Acta*, *899*, 1–12.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*(3), 272.

Frémont, N., Riefler, M., Stolz, A., & Schmülling, T. (2013). The Arabidopsis TUMOR PRONE5 gene encodes an acetylornithine aminotransferase required for arginine biosynthesis and root meristem maintenance in blue light. *Plant Physiology*, *161*(3), 1127–1140.

Galili, G., Avin-Wittenberg, T., Angelovici, R., & Fernie, A. R. (2014). The role of photosynthesis and amino acid metabolism in the energy status during seed development. *Frontiers in Plant Science*, *5*, 447.

Ghini, V., Saccenti, E., Tenori, L., Assfalg, M., & Luchinat, C. (2015). Allostasis and resilience of the human individual metabolic phenotype. *Journal of Proteome Research*, *14*(7), 2951–2962.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*, 1157–1182.

Hageman, J. A., Hendriks, M. M., Westerhuis, J. A., Van Der Werf, M. J., Berger, R., & Smilde, A. K. (2008). Simplivariate models: Ideas and first examples. *PLoS ONE*, *3*(9), e3259.

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

Jacob L, Obozinski G, Vert JP (2009) Group Lasso with Overlaps and Graph Lasso. Proceedings of the 26 th International Conference on Machine Learning, Montreal, Canada 10.1145/1553374.1553431, http://eprints.pascal-network.org/archive/00006439/, arXiv:1110.0413v1

Jansen, J. J., Hoefsloot, H. C., van der Greef, J., Timmerman, M. E., Westerhuis, J. A., & Smilde, A. K. (2005). ASCA: Analysis of multivariate data obtained from an experimental design. *Journal of Chemometrics*, *19*(9), 469–481.

Jenatton R, Obozinski G, Bach F (2009) Structured Sparse Principal Component Analysis. Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS) 9:366–373, 1553374, http://arxiv.org/abs/0909.1440, 0909.1440

Jensen, R., & Bassham, J. (1966). Photosynthesis by isolated chloroplasts. *Proceedings of the National Academy of Sciences United States of America*, *56*(4), 1095–1101.

Jolliffe, I. (2002). *Principal component analysis*. New York: Springer.

Jolliffe, I. T., Trendafilov, N. T., & Uddin, M. (2003). A modified principal component technique based on the LASSO. *Journal of Computational and Graphical Statistics*, *12*(3), 531–547.

Langfelder, P., Zhang, B., & Horvath, S. (2007). Defining clusters from a hierarchical cluster tree: The dynamic tree cut package for R. *Bioinformatics*, *24*(5), 719–720.

Legendre, P., & Anderson, M. J. (1999). Distance-based redundancy analysis: Testing multispecies responses in multifactorial ecological experiments. *Ecological Monographs*, *69*(1), 1–24.

Lisec, J., Schauer, N., Kopka, J., Willmitzer, L., & Fernie, A. R. (2006). Gas chromatography mass spectrometry-based metabolite profiling in plants. *Nature Protocols*, *1*(1), 387–396.

Mackey L (2008) Deflation methods for sparse PCA. Nips (pp. 1–8)

Moerkercke, A. V., Schauvinhold, I., Pichersky, E., Haring, M. A., & Schuurink, R. C. (2009). A plant thiolase involved in benzoic acid biosynthesis and volatile benzenoid production. *The Plant Journal*, *60*(2), 292–302.

O'Brien, R. G., & Kaiser, M. K. (1985). MANOVA method for analyzing repeated measures designs: An extensive primer. *Psychological Bulletin*, *97*(2), 316.

Rawlings, J. O., Pantula, S. G., & Dickey, D. A. (2001). *Applied regression analysis: A research tool*. New York: Springer.

Saccenti, E. (2016). Correlation patterns in experimental data are affected by normalization procedures: Consequences for data analysis and network inference. *Journal of Proteome Research*, *16*(2), 619–634.

Saccenti, E., & Camacho, J. (2015a). Determining the number of components in principal components analysis: A comparison of statistical, cross validation and approximated methods. *Chemometrics and Intelligent Laboratory Systems*, *149*, 99–116.

Saccenti, E., & Camacho, J. (2015b). On the use of the observation-wise k-fold operation in PCA cross-validation. *Journal of Chemometrics*, *29*(8), 467–478.

Saccenti, E., Westerhuis, J. A., Smilde, A. K., van der Werf, M. J., & Hageman, J. A. (2011). Simplivariate models: Uncovering the underlying biology in functional genomics data. *PLoS ONE*, *6*(6), e20747.

Saccenti, E., Hoefsloot, H. C., Smilde, A. K., Westerhuis, J. A., & Hendriks, M. M. (2014). Reflections on univariate and multivariate analysis of metabolomics data. *Metabolomics*, *10*(3), 361–374.

Saccenti, E., Menichetti, G., Ghini, V., Remondini, D., Tenori, L., & Luchinat, C. (2016). Entropy-based network representation of the individual metabolic phenotype. *Journal of Proteome Research*, *15*(9), 3298–3307.

Searle, S. R., & Gruber, M. H. (2016). *Linear models*. New York: Wiley.

Singh, B. K. (1998). *Plant amino acids: Biochemistry and biotechnology*. Boca Raton: CRC Press.

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

Summerfield, A., & Lubin, A. (1951). A square root method of selecting a minimum set of variables in multiple regression: I. The method. *Psychometrika*, *16*(3), 271–284.

Taylor, N. L., Heazlewood, J. L., Day, D. A., & Millar, A. H. (2004). Lipoic acid-dependent oxidative catabolism of $\alpha$-keto acids in mitochondria provides evidence for branched-chain amino acid catabolism in Arabidopsis. *Plant Physiology*, *134*(2), 838–848.

Thiel, M., Féraud, B., & Govaerts, B. (2017). ASCA+ and APCA+: Extensions of ASCA and APCA in the analysis of unbalanced multifactorial designs. *Journal of Chemometrics*, *31*(6), e2895.

Tohge, T., Watanabe, M., Hoefgen, R., & Fernie, A. R. (2013). Shikimate and phenylalanine biosynthesis in the green lineage. *Frontiers in Plant Science*, *4*, 62.

Ullah, I., & Jones, B. (2015). Regularised MANOVA for high-dimensional data. *Australian and New Zealand Journal of Statistics*, *57*(3), 377–389. https://doi.org/10.1111/anzs.12126.

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

Weckwerth, W., Wenzel, K., & Fiehn, O. (2004). Process for the integrated extraction, identification and quantification of metabolites, proteins and RNA to reveal their co-regulation in biochemical networks. *Proteomics*, *4*(1), 78–83.

Winter, G., Todd, C. D., Trovato, M., Forlani, G., & Funck, D. (2015). Physiological implications of arginine metabolism in plants. *Frontiers in Plant Science*, *6*, 534.

Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, *15*(2), 265–286.

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