Data Min Knowl Disc (2009) 19:261–276
DOI 10.1007/s10618-009-0142-5

Two-way analysis of high-dimensional collinear data

Ilkka Huopaniemi · Tommi Suvitaival · Janne Nikkilä · Matej Orešič · Samuel Kaski

Received: 12 June 2009 / Accepted: 24 June 2009 / Published online: 21 July 2009

Abstract We present a Bayesian model for two-way ANOVA-type analysis of high-dimensional, small sample-size datasets with highly correlated groups of variables. Modern cellular measurement methods are a main application area; typically the task is differential analysis between diseased and healthy samples, complicated by additional covariates requiring a multi-way analysis. The main complication is the combination of high dimensionality and low sample size, which renders classical multivariate techniques useless. We introduce a hierarchical model which does dimensionality reduction by assuming that the input variables come in similarly-behaving groups, and performs an ANOVA-type decomposition for the set of reduced-dimensional latent variables.

Responsible editors: Aleksander Kołcz, Wray Buntine, Marko Grobelnik, Dunja Mladenic, and John Shawe-Taylor.

I. Huopaniemi (✉) · T. Suvitaival · J. Nikkilä · S. Kaski
Department of Information and Computer Science, Helsinki University of Technology (TKK), P.O. Box 5400, 02015 Espoo, Finland
e-mail: ilkka.huopaniemi@tkk.fi
URL: http://www.cis.hut.fi/projects/mi/

T. Suvitaival
e-mail: tommi.suvitaival@tkk.fi

S. Kaski
e-mail: samuel.kaski@tkk.fi

J. Nikkilä
Department of Basic Veterinary Sciences (Division of Microbiology and Epidemiology), Faculty of Veterinary Medicine, University of Helsinki, P.O. Box 66, 00014 Helsinki, Finland
e-mail: janne.nikkila@tkk.fi

M. Orešič
VTT Technical Research Centre of Finland (VTT), P.O. Box 1000, 02044 Espoo, Finland
e-mail: matej.oresic@vtt.com

We apply the methods to study lipidomic profiles of a recent large-cohort human diabetes study.

Keywords ANOVA · Factor analysis · Hierarchical model · Metabolomics · Multi-way analysis · Small sample-size

1 Introduction

Although two-way linear models have been thoroughly studied in classical statistics and modern data analysis tasks often involve two-way covariate information, the two-way modelling task has gained little attention in the machine learning literature. A particular, currently active application area where two-way, or in general multi-way experimental setups are ubiquitous, is modern high-throughput bioinformatics. Motivated by research problems in metabolomics, which is an emerging field requiring bioinformatics methods, we introduce a Bayesian hierarchical model capable of two-way analysis of high-dimensional datasets with small sample-size. We present the method as an extension of Bayesian Factor Analysis to maintain a connection to the probabilistic multivariate linear models currently under active research.

1.1 Metabolomics

Metabolomics is rapidly gaining popularity as an application field of bioinformatics. Typically mass spectrometry combined with a chromatography method, such as Liquid Chromatography is used to measure concentrations of metabolic products from tissue samples. Datavectors are then typically 20–200-dimensional metabolic profiles over metabolites, some identified and many not.

In typical experiments, the main interest is in comparing metabolite concentrations between diseased case samples and healthy control samples. When case samples have a consistently higher(lower) concentration, the effect is called up(down)-regulation and considered as a potential biomarker for disease. In addition to the main effect of disease, there are usually additional covariates, such as treatment groups or measurement times that need to be taken into account, resulting in a need of a two-way or in general a multi-way analysis. Additional independent variables and cross-effects between them are introduced and the problem becomes considerably more complicated than a simple differential analysis. There are trivial solutions to convert a two-way problem to one-way problem(s), such as a series of independent one-way analyses or pooling, but they would naturally lose information.

A main complication in metabolomics, and practically in all branches of high-throughput bioinformatics, is that the number of samples n is often much lower than the number of variables p, known as the $n \ll p$ problem. The classical multivariate methods break down due to the singularity of the sample covariance matrix and therefore methods using the full covariance matrix cannot be used directly. Dealing with this problem is currently an active field of research.

1.2 ANOVA for $n \ll p$

ANOVA, analysis of variance, is a well-established univariate method for multi-way analysis in classical statistics. In multivariate cases, especially in bioinformatics, a usual solution is to fit an independent ANOVA-model for each variable. The problem with such multiple testing when $n \ll p$ is a greatly increasing risk of false positives with increasing dimensionality, commonly dealt with more or less heuristic means such as the false discovery rate (FDR) (Benjamini and Hochberg 1995). Typically the data is highly collinear, and information about correlations between metabolites is biologically very relevant; it makes sense to take the collinearity explicitly into account.

The straightforward multivariate generalization of ANOVA, MANOVA, is unfortunately useless when $n \ll p$, since the sample covariance matrix becomes singular. A further technical complication is that (M)ANOVA does not directly reveal the location or direction of the effect (up or down), and these have to be deduced by other methods. There are three common ways for tackling the small sample size problem: dimensionality reduction, regularization of the covariance matrix, and clustering of similarly behaving variables.

Dimensionality reduction is most often done with principal component analysis (PCA). A simple solution is to fit independent ANOVA-models on the principal component scores of each component. Another approach (Langsrud 2002) is to carry out MANOVA on principal component scores. The well-known problem of PCA here is that because of the arbitrariness of the rotation of the components, there is no reason why it should find biologically relevant components.

Partial least-squares (PLS) is a commonly used method for regression and classification and can deal with collinear $n \ll p$ datasets. However, PLS can overfit badly, and component scores are not necessarily reliable for interpretation (Westerhuis et al. 2008).

For studying multivariate $n \ll p$ metabolomic datasets with 2-way experimental setup, a method called Anova-Simultaneous Component analysis or ASCA (Smilde et al. 2005) has been proposed. In the model, both of the one-way effects and the interaction effect were solved independently, assuming a separate basis estimated with principal component analysis. While this is a working solution, it involves major simplifications.

A sparse Bayesian ANOVA model has been proposed for the $n \ll p$ case (Seo et al. 2007). A linear four-way ANOVA model was applied to each gene, using a shared point-mass mixture prior to allow only a small fraction of effects to be non-zero. The sparsity helps in controlling against false discoveries in multiple testing, and also in interpreting the results.

1.3 Covariance regularization

Regularization of the covariance matrix is another way to deal with $n \ll p$. The covariance matrix has to be made non-singular to use traditional statistical multivariate methods, such as Factor Analysis, MANOVA, Linear Discriminant Analysis, or Canonical

Correlation Analysis (CCA). The simplest approach is to use a diagonal correlation matrix, which can be interpreted as assuming the variables to be (conditionally) independent. Lots of less drastic regularization methods have been proposed for shrinking the singular sample covariance matrix towards a positive definite matrix, usually a diagonal matrix; for instance (Cao and Bouman 2009; Tai and Pan 2007). A usual procedure for restricting the projection matrix for Bayesian PCA (Bishop 1999) and FA (Ghahramani and Beal 2000) is by using an Automatic Relevance Determination prior (ARD). Recently sparsity has been imposed in Bayesian PCA and CCA (Archambeau and Bach 2009), resulting in additional advantages in interpretability.

Bayesian sparse factor regression models (West 2003), developed for gene expression data, are suitable for $n \ll p$ regression tasks. Sparsity is enforced by a heavy point-mass mixture prior allowing only a small fraction of regression coefficients to be non-zero. The method is useful in finding only the variables (genes) most strongly related to the external covariate and to infer relationships between the variables via common latent factors. The sparsity also helps in interpreting the components. The model was used for a binary regression, corresponding to a one-way experimental setup.

1.4 Linear mixed models and clustering

It is common to assume that metabolites (as well as mRNAs) form strongly correlated groups, and then to study group-wise differential expression. Studying genes or metabolites one at a time results in a high risk of false positives when $n \ll p$, and the risk can be reduced by studying groups. This has been done on known groups of genes (Wang et al. 2008); other usual approaches include clustering variables according to p-values or choosing only variables with a small enough p-value prior to doing multivariate analyses.

Several methods have been proposed for clustering gene-expression profiles with Linear Mixed Models, usually with a time-dependent experimental design (Ng et al. 2006; Celeux et al. 2005). In a particularly interesting study (Ng et al. 2006), a model-based clustering algorithm was set up by assigning each cluster a subject and cluster-specific random effect common to genes in the cluster. The effect allows modeling correlations and clustering correlated genes, and the clustering solution was computed as the maximum likelihood estimate of the linear model additionally utilizing one-way covariate information as fixed effects. The primary interest of this method was clustering rather than the interpretation of the fixed effects, but it gives inspiration for us to progress to analyzing 2-way effects in a model regularized by assuming a cluster structure for the metabolites.

1.5 Modelling metabolomic datasets

Metabolomic data has certain properties that we want the model to take into account. Fortunately it turns out that the resulting model will still be a reasonably general multi-way factor analysis model.

Due to the existing biochemical pathways where metabolites are converted to one another by chemical reactions, metabolomics data contains correlations caused by tiny fluctuations in metabolic concentrations being transmitted through the pathway. Groups of metabolites are strongly correlated even over biological replicates having the same experimental treatment, a feature not apparent in for instance gene expression data where the correlations mainly result from responses of the genes to the external perturbations (Steuer 2006). Another peculiar feature of metabolomics data is that mean concentrations and scales of different metabolites vary by orders of magnitude; they can be modelled by metabolite-specific means and scales of, say, a healthy control group.

Factor analysis models where latent factor(s) fluctuating around zero are assumed to generate correlated fluctuations around the variable-specific means, fit well the above assumptions. To solve the $n \ll p$ limitation of factor analysis and to simplify the interpretation of the results, we make the simplifying assumption that each variable is generated by exactly one factor. The factor analysis task can now be interpreted to include model-based clustering of variables as a subtask. Biologically the task is related to finding sub-parts of linear pathways which is a current research trend in bioinformatics (Sanguinetti et al. 2008).

We now assume that effects of covariates, such as disease, are visible in the same factors describing the activity of parts of the biochemical network, as up- or down-regulations of the factors. The healthy control biological replicates are assumed to fix the "coordinate basis" of the problem, from which the up- and down-regulations deviate the means of factor values.

As far as we know, the multi-way modeling in high-dimensional metabolomics data, with grouping assumptions made to regularize the problem, is a new approach for generative modeling of the measurement data. Sparse latent factor models (West 2003), being regression-type approaches, can only be used to discover variation of the data that is explained by external covariates. This is reasonable for gene expression data also considering that it has been claimed (Steuer 2006) that for gene expression data, correlations between variables arise mainly due to responses to external variation. However, in metabolomics, fluctuations due to biochemical pathways themselves are another important source of variation that can be useful for instance in finding biological pathways not responding to external covariates. This motivates us in modelling the whole dataset with hierarchical generative modelling.

The clustering methods based on Linear mixed models (Ng et al. 2006; Celeux et al. 2005) have so far not been used for multi-way experimental setups, and because their main goal is clustering they have not considered estimation of statistical significance of the effects.

As PCA in general, ASCA, the only currently existing method addressing the multi-way generative $n \ll p$ metabolomics data, can only be considered an exploratory visualization of PCA scores of one effect at a time. It does not estimate the statistical significance of the effects, although an approach based on permutation tests was later proposed (Vis et al. 2007).

In summary, we introduce a method that combines central aspects needed to model metabolomic datasets in a single, hierarchical generative model. The two-way experimental setup of the research problem is included as population-specific priors on the

latent variables. As a projection matrix we use a clustering matrix enabling $n \ll p$ cases, which also allows an easy interpretation of the clusters related to the different latent factors. Inference on the statistical significance of the effects of external covariates is done by studying the confidence intervals of the posterior distribution. The method is additionally capable of finding clusters of correlated metabolites that are not related to external covariates, but can be interpreted as sub-parts of biochemical pathways. The method generalizes directly to a general multi-way analysis, but for simplicity of presentation we introduce it in the two-way case.

2 Model

We next formulate the two-way analysis model as a factor analysis model where the ANOVA-type two-way effect terms are assigned as the priors of the latent factors. To deal with the small sample size, the projection matrix is formed as a sparse clustering matrix containing only one non-zero element for each variable; this is particularly sensible under the assumption that metabolomics data contains strongly correlated groups of variables. The projection matrix is now non-singular even in the $n \ll p$ cases. The posterior is computed with Gibbs sampling.

In effect the model, shown in Fig. 1 consists of a factor analyzer, where the loadings assume cluster memberships (multiplied with scales) and population-specific priors assume ANOVA-type multi-way structure. We will now introduce each of these parts in turn.

2.1 Factor analysis model

Factor analysis (FA) model (Roweis and Ghahramani 1999) for n exchangeable replicates of the control group is $p(\mathbf{x}_j|\mathbf{V}, \mathbf{x}_j^{\text{lat}}, \boldsymbol{\mu}, \boldsymbol{\Psi}) = \mathcal{N}(\mathbf{x}_j|\boldsymbol{\mu} + \mathbf{V}\mathbf{x}_j^{\text{lat}}, \boldsymbol{\Psi})$, where

$$\begin{aligned} \mathbf{x}_j^{\text{lat}} &\sim \mathcal{N}(0, \mathbf{I}) \\ \mathbf{x}_j &\sim \mathcal{N}(\boldsymbol{\mu} + \mathbf{V}\mathbf{x}_j^{\text{lat}}, \boldsymbol{\Psi}). \end{aligned} \tag{1}$$

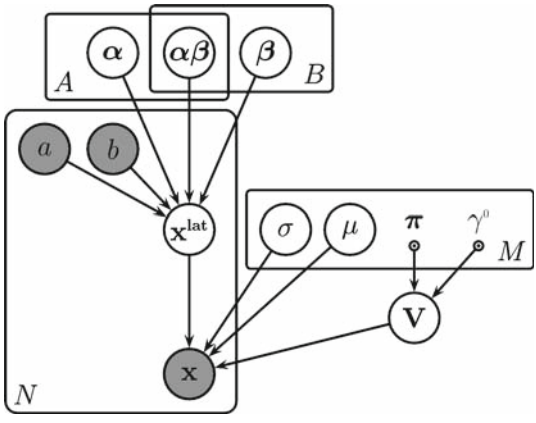

Fig. 1 Plate diagram of the two-way clustering factor analysis model

Here $\mathbf{x}_j$ is a p-dimensional observation vector, $j = 1, \ldots, n$, and $\mathbf{V}$ is the projection matrix that is assumed to generate the datavector $\mathbf{x}_j$ from the latent variable $\mathbf{x}_j^{\text{lat}}$. The $\mathbf{V}\mathbf{x}_j^{\text{lat}}$ models such common variance of the data around the variable-means $\boldsymbol{\mu}$ that can be explained by factors common to all or many variables, effectively estimated from the sample covariance matrix of the dataset. The sample covariance becomes decomposed into $\hat{\boldsymbol{\Sigma}} = \mathbf{V}\mathbf{V}^T + \boldsymbol{\Psi}$, where $\boldsymbol{\Psi}$ is a diagonal residual variance matrix with diagonal elements σ_i^2, modelling the variable-specific noise not explained by the latent factors. The $\mathbf{x}_j^{\text{lat}}$ is a latent variable vector, whose elements are known as factor scores. Following the discussion on unidentifiability problems in (Roweis and Ghahramani 1999), the covariance matrix of $\mathbf{x}^{\text{lat}}$ is set to be the identity matrix.

At this point, the covariates are not yet assumed to induce any special effects, and when $n < p$, $\mathbf{V}$ cannot be estimated due to the singularity of the sample covariance matrix.

2.2 Extending the factor analysis model to low sample-size cases and to two-way analysis

We now extend the model to two complementary directions. We restrict first $\mathbf{V}$ to a non-singular sparse clustering matrix, suitable for data containing highly correlated groups of variables. We then extend the model to include a two-way experimental setup in the latent variable space.

2.2.1 Forming the sparse clustering matrix

We make the structured assumption that there are strongly correlated groups of metabolites in the data, the actual values of group being governed by one latent variable. The projection matrix $\mathbf{V}$ is a positive-valued clustering matrix where each row has one non-zero element corresponding to the cluster assignment of the variable:

$$\mathbf{V} = \begin{bmatrix} \gamma_1 & 0 & 0 \\ 0 & 0 & \gamma_2 \\ \vdots & \vdots & \vdots \\ 0 & \gamma_j & 0 \\ 0 & \gamma_{j+1} & 0 \\ \vdots & \vdots & \vdots \end{bmatrix}. \tag{2}$$

The location of the non-zero value on row i, v_i follows a categorical distribution (multinomial with a single observation), with an uninformative prior distribution $\boldsymbol{\pi}_i$ that does not depend on the size of the cluster. However, the $\boldsymbol{\pi}_i$ could be used to encode prior information on the known grouping of variables.

The variation of each variable within a cluster is assumed to be modeled by the same latent variable, but the scales may differ. The scales γ_i are assigned heavy empirical priors γ_i^0 that keep them close to the values of the control group, to make the γ_i and the population prior-effects identifiable. We follow (Gelman et al. 2003) in parametrizing

the distribution as a scaled Inv-χ^2 distribution with a degrees-of-freedom weighted sum of empirical prior and data scale.

The variable-specific residual variances σ_i^2, that are the diagonal elements of $\boldsymbol{\Psi}$, follow a scaled Inv-χ^2 with an uninformative prior.

In summary, we regularize the covariance matrix by assuming that the main correlations are positive correlations between variables belonging to the same cluster. This correlation is mediated through a common latent variable; this is a reasonable assumption for metabolomics data.

2.2.2 ANOVA-type model for latent variables

For two-way analysis we assume that the samples have been classified into two sets of classes, $a = 0, \ldots, A$ and $b = 0, \ldots, B$. A traditional two-way (M)ANOVA model would be

$$\mathbf{x}_j|_{\text{class}(j)=(a,b)} = \boldsymbol{\mu} + \boldsymbol{\alpha}_a + \boldsymbol{\beta}_b + (\boldsymbol{\alpha\beta})_{ab} + \boldsymbol{\epsilon}_j, \tag{3}$$

where $\boldsymbol{\mu}$ is the grand mean over all samples, $\boldsymbol{\alpha}_a$ and $\boldsymbol{\beta}_b$ are the main effects of the two directions and $(\boldsymbol{\alpha\beta})_{ab}$ are the interaction effects for $a = 0, \ldots, A$ and $b = 0, \ldots, B$.

We assume that the ANOVA-type effects act on the latent variable space, which makes sense both in terms of the interpretation of the latent variables as activities of metabolic pathway parts, and in making it possible to estimate the model for small sample sizes. In the K-dimensional latent variable space we have

$$\mathbf{x}_j^{\text{lat}}|_{\text{class}(j)=(a,b)} = \boldsymbol{\mu}^K + \boldsymbol{\alpha}_a + \boldsymbol{\beta}_b + (\boldsymbol{\alpha\beta})_{ab} + \boldsymbol{\epsilon}_j^K, \tag{4}$$

where class(j) denotes the class labels of sample j, and K denotes lower dimensionality.

The ANOVA effects are set as population priors to the latent variables, which in turn are given Gaussian priors $\boldsymbol{\alpha}_a$, $\boldsymbol{\beta}_b$, $(\boldsymbol{\alpha\beta})_{ab} \sim \mathcal{N}(0, \mathbf{I})$.

To simplify the interpretation of the effects we now deviate from the standard ANOVA convention. Similar choice has been done successfully in other ANOVA studies (Seo et al. 2007), and it does not significantly sacrifice generality. We set the parameter vector $\boldsymbol{\mu}$ describing variable-specific means to the mean of one class, the control group, instead of the grand mean. One group now becomes the baseline to which other classes are compared by adding main and interaction effects. The terms $\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0, (\boldsymbol{\alpha\beta})_{00}, (\boldsymbol{\alpha\beta})_{a0}, (\boldsymbol{\alpha\beta})_{0b}$ become therefore zero. The difference between the classes is now modelled directly with $\mathbf{x}^{\text{lat}}$ and hierarchically by the main effects $\boldsymbol{\alpha_a}$, $\boldsymbol{\beta_b}$ and $(\boldsymbol{\alpha\beta})_{ab}$.

As a simple example consider 2×2 ANOVA analysis. The classes are now $(a, b) = (0, 0), (1, 0), (0, 1), (1, 1)$. The ANOVA terms for samples belonging to different classes are

$$\mathbf{x}_j^{\text{lat}}|_{(a,b)=(0,0)} \sim \mathcal{N}(0, \mathbf{I}) \qquad \mathbf{x}_j^{\text{lat}}|_{(a,b)=(1,0)} \sim \mathcal{N}(\boldsymbol{\alpha}_1, \mathbf{I}) \tag{5}$$
$$\mathbf{x}_j^{\text{lat}}|_{(a,b)=(0,1)} \sim \mathcal{N}(\boldsymbol{\beta}_1, \mathbf{I}) \qquad \mathbf{x}_j^{\text{lat}}|_{(a,b)=(1,1)} \sim \mathcal{N}(\boldsymbol{\alpha}_1 + \boldsymbol{\beta}_1 + (\boldsymbol{\alpha\beta})_{11}, \mathbf{I})$$

There is no effect estimated for the control class $(a, b) = (0, 0)$. The terms $\boldsymbol{\alpha}_1$ and $\boldsymbol{\beta}_1$ now directly model the difference of the samples in the two directions to the control group, and the interaction term $(\boldsymbol{\alpha\beta})_{11}$ models the interactions of the two directions. In standard ANOVA four main effects and four interaction effects would have to be estimated and compared. The inference on the statistical significance of the ANOVA effects now reduces to inferring whether the posterior distribution of these effects is above(up-regulation) or below(down-regulation) zero with, say 95% probability. Each component of the terms, representing different clusters is estimated individually. Note that having only one class would reduce the problem to factor analysis.

The hierarchical model is summarized as

$$\begin{aligned}
&\boldsymbol{\alpha}_0 = 0, \boldsymbol{\beta}_0 = 0, (\boldsymbol{\alpha\beta})_{a0} = 0, (\boldsymbol{\alpha\beta})_{0b} = 0, (\boldsymbol{\alpha\beta})_{00} = 0 \\
&\boldsymbol{\alpha}_a, \boldsymbol{\beta}_b, (\boldsymbol{\alpha\beta})_{ab} \sim \mathcal{N}(0, \mathbf{I}) \\
&\mathbf{x}_j^{\text{lat}}|_{j\in a,b} \sim N(\boldsymbol{\alpha}_a + \boldsymbol{\beta}_b + (\boldsymbol{\alpha\beta})_{ab}, \mathbf{I}) \\
&\mathbf{x}_j \sim \mathcal{N}(\boldsymbol{\mu} + \mathbf{V}\mathbf{x}_j^{\text{lat}}, \boldsymbol{\Psi}).
\end{aligned} \tag{6}$$

2.3 Gibbs-equations

Let us index samples by $j = 1, \ldots, n$, variables by $i = 1, \ldots, p$, and clusters by $k = 1, \ldots, K$. The Gibbs sampling formulas for the model are as follows:

$$\mathbf{x}_j^{\text{lat}} \sim \mathcal{N}(\hat{\boldsymbol{\mu}}_j^{\text{lat}}, \hat{\boldsymbol{\Sigma}}^{\text{lat}}), \tag{7}$$

where

$$\hat{\boldsymbol{\mu}}_j^{\text{lat}} = \hat{\boldsymbol{\Sigma}}^{\text{lat}}(\mathbf{V}^T\boldsymbol{\Psi}^{-1}\mathbf{x}_j + \boldsymbol{\alpha}_a + \boldsymbol{\beta}_b + (\boldsymbol{\alpha\beta})_{ab}), \tag{8}$$
$$\hat{\boldsymbol{\Sigma}}^{\text{lat}} = (\mathbf{V}^T\boldsymbol{\Psi}^{-1}\mathbf{V} + \mathbf{I})^{-1}. \tag{9}$$

The effects are sampled as

$$\boldsymbol{\alpha}_a \sim \mathcal{N}\left(\frac{1}{n_a + 1}\sum_{j\in a}(\mathbf{x}_j^{\text{lat}} - \boldsymbol{\beta}_{b_j} - (\boldsymbol{\alpha\beta})_{ab_j}), \frac{1}{n_a + 1}\mathbf{I}\right), \tag{10}$$

$$\boldsymbol{\beta}_b \sim \mathcal{N}\left(\frac{1}{n_b + 1}\sum_{j\in b}(\mathbf{x}_j^{\text{lat}} - \boldsymbol{\alpha}_{a_j} - (\boldsymbol{\alpha\beta})_{a_j b}), \frac{1}{n_b + 1}\mathbf{I}\right), \tag{11}$$

$$(\boldsymbol{\alpha\beta})_{ab} \sim \mathcal{N}\left(\frac{1}{n_{ab} + 1}\sum_{j\in ab}(\mathbf{x}_j^{\text{lat}} - \boldsymbol{\alpha}_{a_j} - \boldsymbol{\beta}_{b_j}), \frac{1}{n_{ab} + 1}\mathbf{I}\right), \tag{12}$$

where n_a, n_b and n_{ab} denote the number of samples belonging to group a, b, and both a and b, respectively. Finally, the equation for clustering is

$$p(v_i = k) = \frac{\pi_k \prod_j p(x_{ji}|\mu_i + \gamma_i x_{jk}^{\text{lat}}, \sigma_i)}{\Sigma_k \pi_k \prod_j p(x_{ji}|\mu_i + \gamma_i x_{jk}^{\text{lat}}, \sigma_i)}, \tag{13}$$

and for the residual variance and scale parameter

$$\sigma_i^2 \sim \text{Inv-}\chi^2(n, \Sigma_j (x_{ij} - \mu_i - \gamma_i z_{jk})^2), \tag{14}$$

$$\gamma_i^2 \sim \text{Inv-}\chi^2\left(n + n_0, \frac{n\hat{\gamma}_i^2 + n_0 \gamma_i^{0\,2}}{n + n_0}\right), \tag{15}$$

where

$$\hat{\gamma}_i^2 = \frac{\Sigma_j (x_{ji} x_{jk}^{\text{lat}})}{\Sigma_j (x_{jk}^{\text{lat}})^2}. \tag{16}$$

2.4 Empirical prior

To fix the factor analysis to model the correlations of the control group, strong empirical priors are used for $\boldsymbol{\mu}$ and γ_i. The γ_i^0 is the standard deviation of the control group, and n_0 controls the strength of the prior. We use $n_0 = n$. The $\boldsymbol{\mu}$ is the mean vector calculated over the control group. For simplicity and following the results of (Rowe 2006), $\boldsymbol{\mu}$ is subtracted from the whole data and is not sampled, corresponding to the centering discussed in Chap. 2.2.2.

2.5 Model complexity selection

Model complexity, that is, the number of clusters and latent variables is chosen by predictive likelihood with 5-fold cross-validation.

3 Results

We study the performance of the method on simulated data and on data from a recent large-scale empirical study. The simulated data is first used to study how the method copes with a decreasing number of samples in the task of finding ANOVA-type effects. The use of the method is then illustrated on a Lipidomics dataset from a recent Type 1 diabetes study (Oresic et al. 2008). In this study, lipidomic profiles of healthy human patients and patients developing into type 1 diabetes were measured at variable intervals. We first carry out a 2×2 cross-sectional analysis in one time point, on the treatment variables healthy-diseased vs female-male. Finally, we consider the time index as one of the experimental variables, the other being healthy vs diseased samples.

3.1 Simulated data

3.1.1 Effect of sample size

We demonstrate how well the method finds up-regulation effects as a function of the number of samples. The data is generated with the following parameters: There are four classes within a 2-way experimental setup as in Eq. 5. There are $K = 4$ clusters in which the following effects are generated: $\boldsymbol{\alpha}_1 = (+2, 0, 0, 0)$, $\boldsymbol{\beta}_1 = (0, +2, 0, 0)$ $(\boldsymbol{\alpha\beta})_{11} = (0, 0, +2, 0)$. Dimensionality of the dataset is $p = 200$. The optimal number of clusters is chosen by predictive likelihood, recovering the correct number of clusters $K = 4$ (Fig. 2).

The sample size now varies from $n = 20$ to $n = 1{,}000$, such that the four classes have an equal number of samples (e.g., $n = 20$ means 5 samples in each class). The noise parameters are set to $\sigma_i = 1$, scale parameters to $\gamma_i = 1$, and mean parameters to $\mu_i = 0$ for $i = 1, \ldots, p$. The prior n_0 is fixed to $n_0 = 20$. In each run, 1,000 Gibbs samples are collected after 1,000 burn-in iterations. For each sample size, 10 independent datasets with the same parameters are generated and Gibbs sampling repeated for each. The posterior intervals and means of the pooled posterior distributions of the effects are plotted for each found cluster in Fig. 3. We intentionally computed the model with a slightly misplaced number of clusters to demonstrate effects of minor misspecification, having $K = 5$ clusters instead of the optimal $K = 4$.

The results show that the model finds the generated effect in each cluster and does not find false-positive effects in clusters where none were generated (although there is a fair measure of uncertainty in the estimates for small sample sizes). Uncertainty of the effects, that is, the width of the posterior interval diminishes as the number of samples grows, as expected. Correct clustering is found from the posterior of $\mathbf{V}$ each time. In metabolomics experiments, usually 20–100 samples are available. These

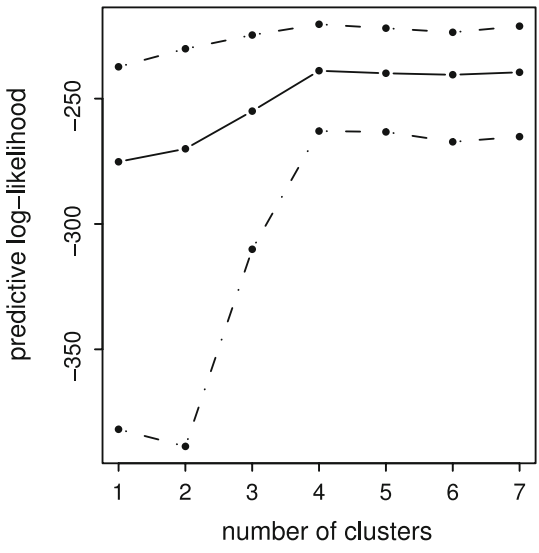

Fig. 2 The correct number of clusters $K = 4$ is found for generated data in model complexity selection. Average predictive likelihood of left-out data is shown as a function of number of clusters. Increasing the number of clusters after $K = 4$ does not increase the likelihood

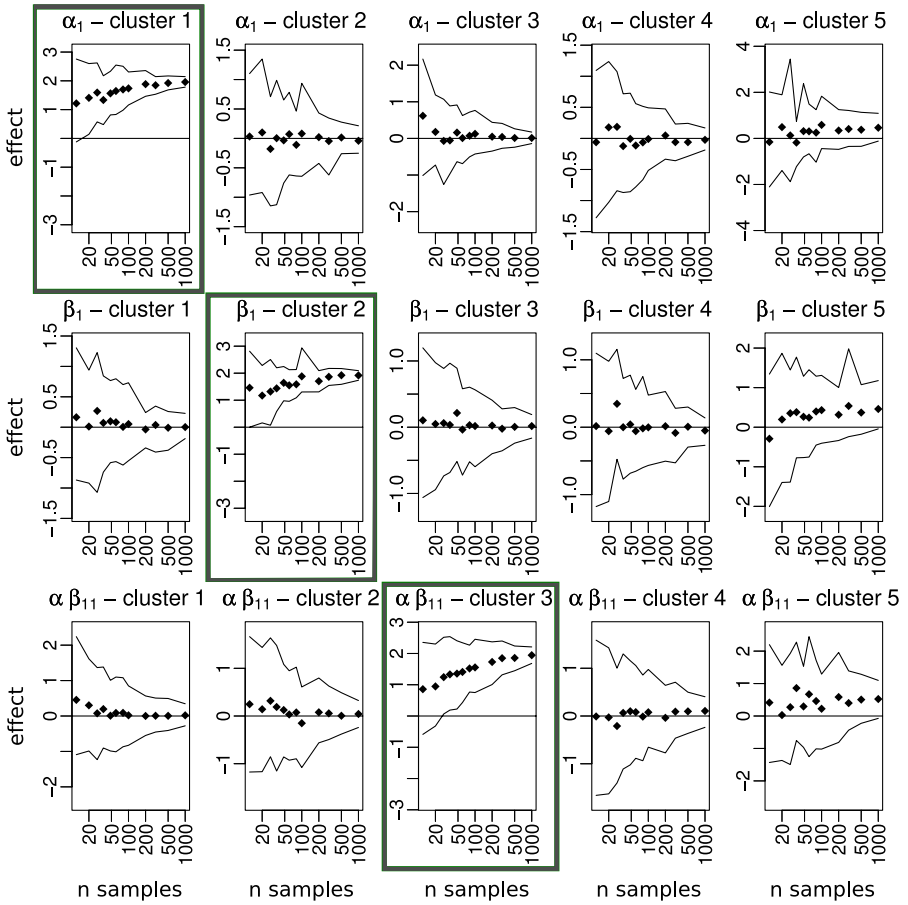

Fig. 3 The method finds the generated effects $\alpha_{\text{cluster1}} = +2$, $\beta_{\text{cluster2}} = +2$, $(\alpha\beta)_{\text{cluster3}} = +2$. In the other clusters, no effects are found. The 95% posterior intervals of the main and interaction effects are plotted for each cluster

amounts are on the borderline of the posterior interval of the effects differing from above zero.

3.2 Lipidomic diabetes data set

3.2.1 Cross-sectional study of healthy-diseased, male-females

We study the two-way experimental setup of a single time point (avg. time 750) in subjects who later progressed to type 1 diabetes (Oresic et al. 2008). The classes are healthy female (18 samples, subjects who have not progressed to diabetes, chosen as the control group), healthy male (17), diabetic female (11 who have later progressed to diabetes), diabetic male (8), and there are 53 lipids. Following the notation of the

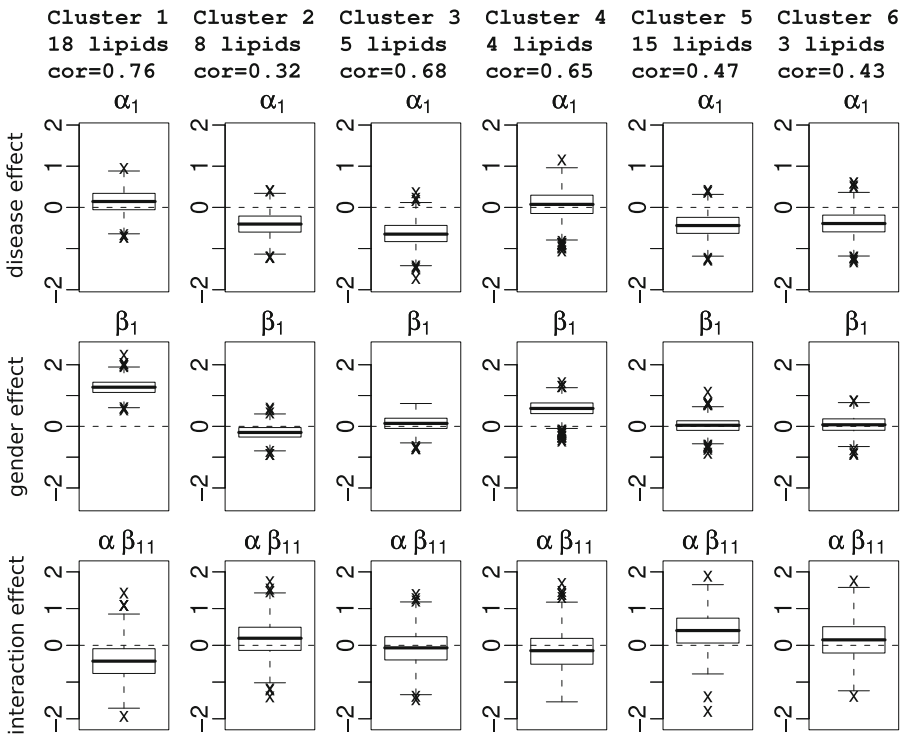

Fig. 4 The method finds statistically significant effects for the human diabetes cross-sectional healthy-diseased, male-female comparison. Effects are found for α_{cluster3}, β_{cluster1} and β_{cluster4}. The figure shows posterior intervals of the main and interaction effects for each cluster. In addition, average correlation coefficients between lipids within each cluster are given

example of Eq. 5, the disease effect is estimated with the $\boldsymbol{\alpha}_1$-parameter for each cluster, $\boldsymbol{\beta}_1$ models the gender-effect and $(\boldsymbol{\alpha\beta})_{11}$ models the interaction of these two effects. The optimal number of clusters is found to be $K = 6$. According to the results shown in Fig. 4, there is a positive, statistically significant gender effect found for clusters 1 and 4 signifying that males have a higher concentration for 18 and 4 lipids, respectively. A negative disease effect is found for cluster 3, signifying that diabetic patients have a lower concentration for 5 lipids. Note that the other effects are not statistically significant, but we still find clusters of strongly correlated lipids.

3.2.2 Time development of healthy and diabetic patients

Finally, we demonstrate the performance of the model for a simple time-series analysis of the human diabetes. The time indices are treated as independent values of the covariate; later the model will be extended by taking the time order into account, for instance by assuming a Hidden Markov Model structure (Beal and Krishnamurthy 2006; Nikkila et al. 2008).

In the diabetes data, lipidomic profiles of healthy human patients and patients developing into type 1 diabetes had been measured at variable intervals. The measurements were aligned to six time-points. The two-way setup now contains time effects and a healthy-diseased categorization. We now assume that there is no static disease effect, but instead disease effects change in time. Therefore, only time effects $\boldsymbol{\alpha}_{1,\ldots,5}$ and time-disease interaction effects $(\boldsymbol{\alpha\beta})_{(0,\ldots,5)1}$ are estimated. The latter now indicate, for each time point, the deviation caused by the disease from the normal time-development. The optimal number of clusters was found to be $K = 5$. The results shown in Fig. 5 reveal clear time-dependent behavior, estimated by the $\boldsymbol{\alpha}$, that is distinct for all clusters. Statistically significant interactions of time and disease ($\boldsymbol{\alpha\beta}$) are found at timepoint 0 for clusters 4 and 5 (disease up-regulation), at timepoint 4 for cluster 3 (disease down-regulation) and at timepoint 5 for clusters 2 and 3 (disease down-regulation). In this machine learning paper we do not analyze the biological implications

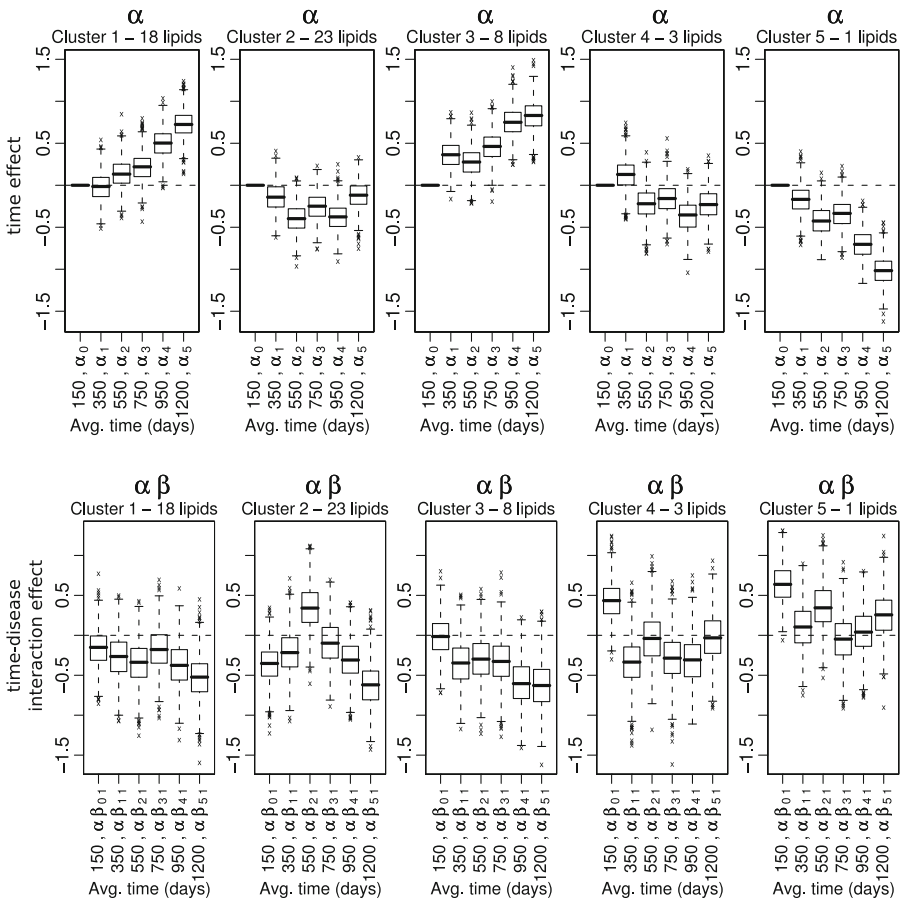

Fig. 5 Statistically significant time-varying behavior is found for each cluster in the human diabetes data (*upper figures*). Time-disease interaction effects are found as well for clusters 3, 4 and 5 (*lower figures*). Posterior intervals of the main effect (time) and interaction effects (time, disease state) are plotted

further; some interesting findings were made and they are being worked on for a biological paper.

4 Conclusion

We introduced a Bayesian hierarchical model that can be used to model two-way experimental data even when $n \ll p$. The model was formulated as a factor analyzer where population-specific priors were set on the latent variables. This can be interpreted as an ANOVA-type model.

The model finds a clustering of correlated variables, and the clustering assumption helps solve the $n \ll p$ problem. There are strong justifications why clusteredness would be a good assumption particularly for modeling metabolomics data. Clustering factor analysis can be easily replaced by a simpler component model such as PCA, sparse PCA or exponential PCA if they are considered more appropriate for other applications.

Prior knowledge on clustering the variables, often available in bioinformatics applications, could directly be taken into account as prior probabilities in the clustering matrix. The model can also be extended to take the time-series nature of the data into account.

Acknowledgments The project was funded by Tekes MASI program. I. H., T. S and S. K belong to the Adaptive Informatics Research Centre and Helsinki Institute for Information Technology. I. H. is funded by the Graduate School of Computer Science and Engineering. S. K is partially supported by EU FP7 NoE PASCAL2, ICT 216886.

References

Archambeau C, Bach F (2009) Sparse probabilistic projections. In: Koller D, Schuurmans D, Bengio Y, Bottou L (eds) Advances in neural information processing systems, vol 21. MIT Press, Cambridge, pp 73–80

Beal M, Krishnamurthy P (2006) Gene expression time course clustering with countably infinite hidden markov models. In: Proceedings of the 22nd annual conference on uncertainty in artificial intelligence (UAI-06), Arlington, Virginia. AUAI Press

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B (Methodological) 57(1):289–300

Bishop CM (1999) Bayesian PCA. In: Proceedings of the 1998 conference on advances in neural information processing systems II. MIT Press, Cambridge, pp 382–388

Cao G, Bouman CA (2009) Covariance estimation for high dimensional data vectors using the sparse matrix transform. In: Koller D, Schuurmans D, Bengio Y, Bottou L (eds) Advances in neural information processing systems, vol 21. MIT Press, Cambridge, pp 225–232

Celeux G, Martin O, Lavergne C (2005) Mixture of linear mixed models for clustering gene expression profiles from repeated microarray experiments. Stat Model 5(3):243–267

Gelman A, Carlin JB, Stern HS, Rubin DB (2003) Bayesian data analysis, 2nd edn. Chapman & Hall/CRC, London

Ghahramani Z, Beal MJ (2000) Variational inference for Bayesian mixtures of factor analysers. In: Advances in neural information processing systems, vol 12. MIT Press, Cambridge, pp 449–455

Langsrud O (2002) 50–50 multivariate analysis of variance for collinear responses. J R Stat Soc Ser D-the Statistician 51:305–317

Ng SK, McLachlan GJ, Wang K, Ben-Tovim Jones L, Ng SW (2006) A mixture model with random-effects components for clustering correlated gene-expression profiles. Bioinformatics 22(14):1745–1752

Nikkila J, Sysi-Aho M, Ermolov A, Seppnen-Laakso T, Simell O, Kaski S, Oresic M (2008) Gender-dependent progression of systemic metabolic states in early childhood. Mol Syst Biol 4(197). doi:10.1038/msb.2008.34

Oresic M, Simell S, Sysi-Aho M, Nanto-Salonen K, Seppanen-Laakso T, Parikka V, Katajamaa M, Hekkala A, Mattila I, Keskinen P, Yetukuri L, Reinikainen A, Lahde J, Suortti T, Hakalax J, Simell T, Hyoty H, Veijola R, Ilonen J, Lahesmaa R, Knip M, Simell O (2008) Dysregulation of lipid and amino acid metabolism precedes islet autoimmunity in children who later progress to type 1 diabetes. J Exp Med 205(13):2975–2984

Rowe DB (2000) On estimating the mean in Bayesian factor analysis. In: Social science working paper 1096, division of humanities and social sciences, Caltech, Pasadena, CA 91125

Roweis S, Ghahramani Z (1999) A unifying review of linear Gaussian models. Neural Comput 11(2):305–345

Sanguinetti G, Noirel J, Wright PC (2008) MMG: a probabilistic tool to identify submodules of metabolic pathways. Bioinformatics 24(8):1078–1084

Seo DM, Goldschmidt-Clermont PJ, West M (2007) Of mice and men: sparse statistical modelling in cardiovascular genomics. Ann Appl Stat 1(1):152–178

Smilde AK, Jansen JJ, Hoefsloot HCJ, Lamers RJAN, van der Greef J, Timmerman ME (2005) ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data. Bioinformatics 21(13):3043–3048

Steuer R (2006) Review: On the analysis and interpretation of correlations in metabolomic data. Brief Bioinform 7(2):151–158

Tai F, Pan W (2007) Incorporating prior knowledge of gene functional groups into regularized discriminant analysis of microarray data. Bioinformatics 23(23):3170–3177

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

Wang L, Zhang B, Wolfinger RD, Chen X (2008) An integrated approach for the analysis of biological pathways using mixed models. PLoS Genet 4(7):e1000115

West M (2003) Bayesian factor regression models in the large p, small n paradigm. Bayesian Stat 7:723–732

Westerhuis J, Hoefsloot H, Smit S, Vis D, Smilde A, van Velzen E, van Duijnhoven J, van Dorsten F (2008) Assessment of plsda cross validation. Metabolomics 4(1):81–89