BIOINFORMATICS

Vol. 26 ISMB 2010, pages i391–i398
doi:10.1093/bioinformatics/btq174

Multivariate multi-way analysis of multi-source data

Ilkka Huopaniemi[1,*], Tommi Suvitaival[1], Janne Nikkilä[1,2], Matej Orešič[3] and Samuel Kaski[1,*]

[1]Aalto University School of Science and Technology, Department of Information and Computer Science, Helsinki Institute for Information Technology HIIT, PO Box 15400, FI-00076 Aalto, Espoo, [2]Department of Veterinary Biosciences, Faculty of Veterinary Medicine, University of Helsinki, PO Box 66, FIN-00014, Helsinki and [3]VTT Technical Research Centre of Finland, PO Box 1000, FIN-02044 VTT, Espoo, Finland

ABSTRACT

Motivation: Analysis of variance (ANOVA)-type methods are the default tool for the analysis of data with multiple covariates. These tools have been generalized to the multivariate analysis of high-throughput biological datasets, where the main challenge is the problem of small sample size and high dimensionality. However, the existing multi-way analysis methods are not designed for the currently increasingly important experiments where data is obtained from multiple sources. Common examples of such settings include integrated analysis of metabolic and gene expression profiles, or metabolic profiles from several tissues in our case, in a controlled multi-way experimental setup where disease status, medical treatment, gender and time-series are usual covariates.

Results: We extend the applicability area of multivariate, multi-way ANOVA-type methods to multi-source cases by introducing a novel Bayesian model. The method is capable of finding covariate-related dependencies between the sources. It assumes the measurements consist of groups of similarly behaving variables, and estimates the multivariate covariate effects and their interaction effects for the discovered groups of variables. In particular, the method partitions the effects to those shared between the sources and to source-specific ones. The method is specifically designed for datasets with small sample sizes and high dimensionality.

We apply the method to a lipidomics dataset from a lung cancer study with two-way experimental setup, where measurements from several tissues with mostly distinct lipids have been taken. The method is also directly applicable to gene expression and proteomics.

Availability: An R-implementation is available at http://www.cis.hut.fi/projects/mi/software/multiWayCCA/

Contact: ilkka.huopaniemi@tkk.fi; samuel.kaski@tkk.fi

1 INTRODUCTION

Data from experiments with multiple covariates are usually analyzed with multi-way analysis of variance (ANOVA)-type methods. A typical one-way analysis setup in experiments looking for potential biomarkers for disease is the diseased–healthy differential analysis. Biological experiments often contain additional covariates, such as drug treatment groups, gender or time-series, resulting in a multi-way experimental setup.

The traditional methods for multi-way analysis are univariate multi-way ANOVA, and its multivariate generalization MANOVA. In the two-way case, to explain the covariate-related variation of the data in one data source, say $\mathbf{x}$, the following linear model is usually assumed:

$$\mathbf{x} = \boldsymbol{\mu}^x + \boldsymbol{\alpha}_a^x + \boldsymbol{\beta}_b^x + (\boldsymbol{\alpha}\boldsymbol{\beta})_{ab}^x + \epsilon. \quad (1)$$

Here the a and b ($a=0,\ldots,A$ and $b=0,\ldots,B$) are the two independent covariates, and the main effects $\boldsymbol{\alpha}_a^x$ and $\boldsymbol{\beta}_b^x$ and the interaction effect $(\boldsymbol{\alpha}\boldsymbol{\beta})_{ab}^x$ model the variation from the baseline level (grand mean) $\boldsymbol{\mu}^x$. The ϵ is a noise term.

A recurring problem in 'omics' measurements, in particular in gene expression and metabolomics, is the small number of samples versus high dimensionality; the traditional multivariate methods break down due to the singularity of the covariance matrix. A further disadvantage of MANOVA is that it gives only a P-value describing the statistical significance of the effects, and the location (which variables) and direction (up/down) of the effects have to be deduced afterwards separately by other methods, such as t-tests. The latter holds for ANOVA as well. Dealing with the small sample size problem is currently an active research topic, and has already led to working solutions.

Multi-way, multivariate ANOVA-type analysis can be done in the small sample size cases with simple two-step approaches relying on a prior principal component analysis (PCA) dimension reduction (Langsrud, 2002; Smilde *et al.*, 2005). Another approach is forming sparse latent factors (West, 2003), where some of the variables but not all are associated with each factor. In this approach, covariate information is used for factor regression and the model has been extended with univariate ANOVA models with a joint sparsity prior (Carvalho *et al.*, 2008; Lucas *et al.*, 2009; Seo *et al.*, 2007).

We have recently introduced a unified Bayesian machine learning model (Huopaniemi *et al.*, 2009) especially designed for metabolomic experiments with a two-way setup and small sample size. This approach assumes similarly behaving correlated groups of variables, a valid assumption for metabolites, and models the multi-way covariate-related variation for the groups. The model is an extension of a factor analyzer and models the statistical significance of multivariate covariate effects on the low-dimensional factor space, representing the discovered clusters of variables.

However, the multi-way data analysis problem becomes complicated when heterogeneous data with multiple covariates are integrated from multiple sources. Different data sources usually have distinct, unmatched variable spaces with different dimensionalities; this becomes evident when considering integration of transcriptomics and metabolomics data, or metabolic profiles from different tissues which usually have different metabolites.

If the variable spaces were the same, one might want to consider the 'source' as an additional covariate in a standard ANOVA-type analysis. However, this is usually impossible since variable spaces

*To whom correspondence should be addressed.

are in general distinct and unmatched, and covariate effects shared between the sources cannot be defined a priori.

The desired goal would be to write a linear model for the sources **x** and **y** in the two-source, two-way case as

$$\mathbf{x} = \boldsymbol{\mu}^x + \boldsymbol{\alpha}_a + \boldsymbol{\beta}_b + (\boldsymbol{\alpha\beta})_{ab} + \boldsymbol{\alpha}^x_a + \boldsymbol{\beta}^x_b + (\boldsymbol{\alpha\beta})^x_{ab} + \epsilon,$$
$$\mathbf{y} = \boldsymbol{\mu}^y + \boldsymbol{\alpha}_a + \boldsymbol{\beta}_b + (\boldsymbol{\alpha\beta})_{ab} + \boldsymbol{\alpha}^y_a + \boldsymbol{\beta}^y_b + (\boldsymbol{\alpha\beta})^y_{ab} + \epsilon, \quad (2)$$

where, in addition to the source-specific effects denoted by superscripts ^{x} and ^{y}, there are effects shared between the sources: $\boldsymbol{\alpha}_a$, $\boldsymbol{\beta}_b$ and $(\boldsymbol{\alpha\beta})_{ab}$. Unfortunately, since **x** and **y** have distinct variable spaces, a model including shared effects cannot be defined as simply as in Equation (2).

It turns out that the problem is accessible under the additional assumption that the observations **x** and **y** from the two different sources come in pairs (co-occur). Then, we can include into the model an unknown functional mapping f^x and f^y from the shared effects to each source, and estimate it from data. The complete model reads

$$\mathbf{x} = \boldsymbol{\mu}^x + f^x(\boldsymbol{\alpha}_a + \boldsymbol{\beta}_b + (\boldsymbol{\alpha\beta})_{ab}) + f^x(\boldsymbol{\alpha}^x_a + \boldsymbol{\beta}^x_b + (\boldsymbol{\alpha\beta})^x_{ab}) + \epsilon,$$
$$\mathbf{y} = \boldsymbol{\mu}^y + f^y(\boldsymbol{\alpha}_a + \boldsymbol{\beta}_b + (\boldsymbol{\alpha\beta})_{ab}) + f^y(\boldsymbol{\alpha}^y_a + \boldsymbol{\beta}^y_b + (\boldsymbol{\alpha\beta})^y_{ab}) + \epsilon\,. \quad (3)$$

The model allows decomposing the effects to shared and source-specific ones, although they are in different variable spaces, which has not been possible with existing ANOVA-type methods.

This article concentrates on the multivariate analysis of multi-way, multi-source datasets. The focus is on cases where two or more data sources have been measured from each biological sample (paired samples), but no a priori known matching of the variables in different sources needs to be assumed and sources can have different numbers of variables.

No methods currently exist for the analysis of data from multi-way, multi-source experimental setups. Since the focus of biological research has moved toward multi-source experiments, there is a need for multi-way, multi-source analysis methods.

In order to connect different kinds of sources we need paired samples. Setups with paired samples are becoming increasingly relevant as experiments that use two or more measurement techniques for each patient, or study a set of tissues of each patient, are becoming more and more common. When for example a metabolomic and a gene expression (or proteomic) profile have been measured from each patient, the samples (profiles) of the two sources are paired, allowing for modeling of dependencies between them.

A widely known classical statistical method for finding dependencies between datasets is canonical correlation analysis [CCA; Hotelling, 1936], for which there exist recent sparse variants (Archambeau and Bach, 2009; Parkhomenko *et al.*, 2007; Waaijenborg and Zwinderman, 2007; Witten and Tibshirani, 2009). The CCA-type two-source analysis does not take covariate information into account and has therefore evolved as a separate field from ANOVA-type methods, although both are equally based on traditional multivariate statistics. In the CCA, there are at least two data-sources, **x** and **y**, with paired samples and distinct variable spaces, and the task is to find canonical latent variables in each source such that the dependency is maximized. The relationship of these two analysis setups and our new task, multi-way, multi-source analysis, is illustrated in Figure 1.

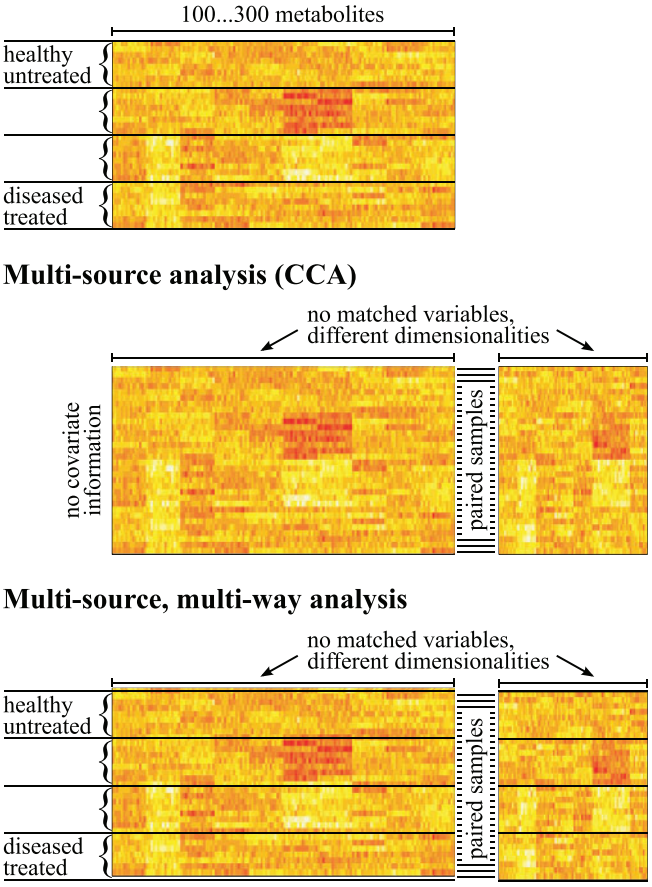

Fig. 1. (top) Multi-way analysis studies datasets with two or more covariates for each sample. The task is to find the effects of the covariates in the data. (middle) Multi-source analysis studies dependencies between two or more datasets with paired samples without covariate information. (bottom) Multi-way, multi-source analysis combines both tasks. The task is to find shared and source-specific covariate effects. In the data matrices, rows represent samples, and columns represent variables.

An initial step for utilizing covariate information in multi-source analysis was taken in a sparse supervised CCA method (Witten and Tibshirani, 2009). The method can handle multi-source setups with one-way covariate information approximately by a simple two-step approach where a coarse dimension reduction for each data source is first performed by choosing variables with a high enough univariate correlation with the outcome variable, and then performing sparse CCA.

Common data integration methods also include partial least squares-based classification (Webb-Robertson *et al.*, 2009) and regression (Le Cao *et al.*, 2009) studies. It is very promising that integrating multiple sources has been shown (Girolami and Zhong, 2007) to increase classification accuracy compared to using a single source. Here, we extend from classification studies to actual latent effect models.

In this article, we present a unified Bayesian machine learning method that can solve the multi-way, multi-source analysis task in a single model. The multiple sources with different variable spaces are integrated by an extended Bayesian CCA that solves the multi-way analysis in a space shared by the sources. The method solves the problem of high dimensionality and small sample size by finding groups of similarly behaving variables and estimating the main covariate effects and the interaction effects in a low-dimensional latent-variable space representing the clusters found. Furthermore, the method partitions these effects to shared effects between the sources and to source-specific effects. Correlations between the groups of variables between and within sources are detected as well. The method also finds groups of variables behaving similarly in each source, but not responding to the external covariates and determines whether they have a dependency shared between the sources.

The method is specifically designed for small sample size, high-dimensional datasets. Given the Bayesian treatment, a rigorous estimate for the uncertainty of the effects is inherently obtained. In addition to estimating the significance of the shared and source-specific multi-way effects, the method directly pinpoints (in contrast to MANOVA) which group(s) of features are up/downregulated.

We demonstrate the method on simulated data, and show how it finds shared and source-specific two-way effects even with small sample sizes. We then apply the method on a lipidomics lung cancer dataset where lipidomic profiles have been measured from several tissues with mostly distinct lipids. In addition to the diseased–healthy division, half of both populations have been given a test anticancer drug.

The method is not restricted to metabolomics; it is generally applicable when the variables can be assumed to form mutually correlated groups, for instance in gene expression and proteomics.

2 METHODS

We formulate the new model for the multivariate analysis of multi-way, multi-source datasets under the so-called 'large p, small n' conditions (high dimensionality p, small number of observations n) as a hierarchical Bayesian machine learning model.

To solve the modeling task, we need three components: (i) regularized dimension reduction, (ii) combination of different data sources, and (iii) multi-way analysis. Following the basic idea of hierarchical Bayesian modeling, these components are formulated as parts of an overall generative model that is assumed to have generated the observed data $\mathbf{x}$ and $\mathbf{y}$. The model parameters are learned jointly with Gibbs sampling; the Gibbs-formulas are presented in Section 2.6. We will first summarize the main components of the model shown in Figure 2, and then describe each part in detail.

To deal with the small sample size ($n \ll p$) problem, we reduce the dimensionality of the data $\mathbf{x}$ and $\mathbf{y}$ from the two sources into their respective latent variables $\mathbf{x}^{lat}$ and $\mathbf{y}^{lat}$. This is done with factor analyzers (FA), which are additionally regularized by assuming that the variables come in groups and each group comes from one factor only. This is a strongly regularizing assumption effective for solving the 'large p, small n' problem. The clustering assumption is particularly sensible under the assumption that metabolomics data, our main application, contains strongly correlated groups of variables (Steuer, 2006) associated to the same biochemical networks.

The second necessary element is search for components shared by the two different sources $\mathbf{x}^{lat}$ and $\mathbf{y}^{lat}$, needed for finding shared multi-way effects. Given paired data, this is a task for Bayesian CCA (BCCA; Archambeau and Bach, 2009; Klami and Kaski, 2007; Wang, 2007), which introduces a new hierarchy level where a latent variable $\mathbf{z}$ captures the shared variation between the sources.

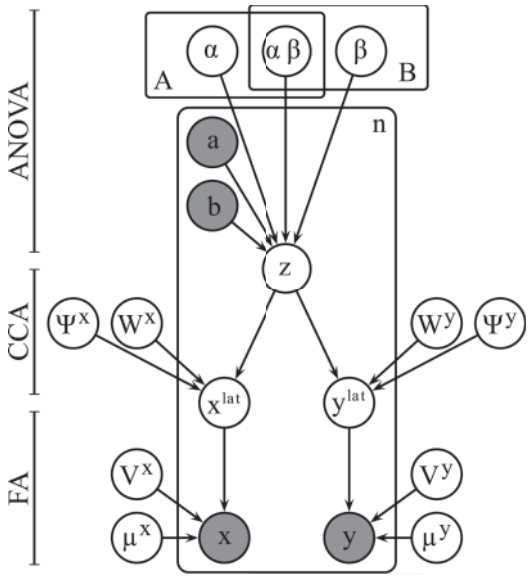

Fig. 2. Plate diagram of the graphical model.

The third necessary element, the ANOVA-type multi-way analysis, is supplemented by assigning the effect terms describing the linear ANOVA model as priors on the latent variables $\mathbf{z}$. The observed covariates a and b choose the correct effects for each sample.

In effect, the model consists of two FA, where the loadings assume cluster memberships (multiplied with scales), a CCA-type generative model for combining the sources, and population-specific priors on $\mathbf{z}$ that assume ANOVA-type multi-way structure. The details of the model are now presented for the two-way, two-source case for simplicity of presentation; generalization to multiple 'ways' and sources is straightforward.

2.1 Dimension reduction by FA

Dimension reduction is done by a FA (analogous to PCA), one for each source, $\mathbf{x}$ and $\mathbf{y}$. Factor analysis assumes that the high-dimensional data spaces $\mathbf{x}$ and $\mathbf{y}$ have been generated by low-dimensional latent variable spaces $\mathbf{x}^{lat}$ and $\mathbf{x}^{lat}$, respectively.

To overcome the $n \ll p$ problem, the FA are regularized by clustering. We assume similarly behaving groups of variables and search for them in the Gibbs sampling. Each factor now represents one cluster of variables. All this is done in a single unified model.

2.1.1 Factor analysis model The factor analysis model (Roweis and Ghahramani, 1999) for n exchangeable replicates is

$$\mathbf{x}_j^{lat} \sim \mathcal{N}(0, \boldsymbol{\Psi}^x),$$
$$\mathbf{x}_j \sim \mathcal{N}(\boldsymbol{\mu}^x + \mathbf{V}^x \mathbf{x}_j^{lat}, \boldsymbol{\Lambda}^x). \quad (4)$$

Here $\boldsymbol{V}^x$ is the projection matrix that is assumed to generate the data vector $\mathbf{x}_j$ from the latent variable $\mathbf{x}_j^{lat}$, whose elements are known as factor scores. The $\mathbf{V}^x \mathbf{x}_j^{lat}$ models such common variance of the data around the variable means $\boldsymbol{\mu}^x$ that can be explained by factors common to all or many variables, effectively estimated based on the sample covariance matrix of the dataset. The sample covariance becomes decomposed into $\hat{\boldsymbol{\Sigma}} = \mathbf{V}^x \mathbf{V}^{xT} + \boldsymbol{\Lambda}^x$, where $\boldsymbol{\Lambda}^x$ is a diagonal residual-variance matrix with diagonal elements σ_i^2 modeling the variable-specific noise not explained by the latent factors. The covariance matrix of $\mathbf{x}^{lat}$, $\boldsymbol{\Psi}^x$, comes here from the Bayesian CCA. Note that the baseline means of variables $\boldsymbol{\mu}^x$ are estimated directly in the $\mathbf{x}$-space, whereas the ANOVA effects will be estimated in the shared latent variable $\mathbf{z}$-space, higher in the hierarchy.

At this point, when $n<p$, $\mathbf{V}^x$ cannot be estimated due to the singularity of the sample covariance matrix. To overcome the $n \ll p$ problem, we now restrict $\mathbf{V}^x$ to a non-singular clustering matrix, suitable for data containing groups of similarly behaving variables.

2.1.2 Projection matrix that assumes grouped variables We make the assumption that there are correlated groups of variables in the data, and restrict the projection matrix $\mathbf{V}^x$ to a clustering-type of matrix, where each variable comes from exactly one factor. This means that in the model the generated values within the whole cluster are being governed by one latent variable. The projection matrix $\mathbf{V}^x$ is positive-valued, each row having one non-zero element corresponding to the cluster assignment of the variable, multiplied by the variable-specific scale λ_i:

$$\mathbf{V}^x = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & 0 & \lambda_2 \\ \vdots & \vdots & \vdots \\ 0 & \lambda_i & 0 \\ 0 & \lambda_{i+1} & 0 \\ \vdots & \vdots & \vdots \end{bmatrix}. \quad (5)$$

The location of the only non-zero value, λ_i on row i, is denoted by v_i. It follows a multinomial distribution with one observation, with an uninformative prior distribution $\boldsymbol{\pi}_i$. The $\boldsymbol{\pi}_i$ could also 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 λ_i may differ. Estimating the λ_i is explained in Section 2.5. The variable-specific residual variances σ_i^2, that are the diagonal elements of $\boldsymbol{\Lambda}^x$, follow a scaled Inv-χ^2 with an uninformative prior.

The information in each high-dimensional sample is now represented by a low-dimensional latent variable corresponding to a vector of factor scores, one score for each cluster. The integration of the data sources including the decomposition into shared and source-specific effects is now done in the low-dimensional latent variable space.

2.2 Integrating the sources by Bayesian CCA

When modeling dependencies between datasets, CCA searches the variation shared between the datasets and separates it from the dataset-specific variation. In the probabilistic Bayesian formulation of CCA (BCCA; Bach and Jordan, 2005; Klami and Kaski, 2007), common variation is modelled by a latent variable $\mathbf{z}_j$ common to both sources. Note that modelling dependencies by a shared low-dimensional latent variable $\mathbf{z}_j$ for each sample j allows to correctly utilize the pairing information and model dependencies between the sources, crucial in integrating the data-sources.

The $\mathbf{z}$ now model dependencies between the low-dimensional latent variable spaces $\mathbf{x}^{lat}$ and $\mathbf{y}^{lat}$.

The generative model of CCA has been formulated (Bach and Jordan, 2005; Klami and Kaski, 2007) for sample j as

$$\begin{aligned} \mathbf{z}_j &\sim N(0, \mathbf{I}), \\ \mathbf{x}_j^{lat} &\sim N(\mathbf{W}^x \mathbf{z}_j, \Psi^x), \end{aligned} \quad (6)$$

and likewise for y. Here, we have assumed no mean parameter since the mean of the data is estimated in the factor analysis part. The $\mathbf{W}^x$ is a projection matrix from the latent variables $\mathbf{z}_j$, and Ψ^x is a matrix of marginal variances. The crucial thing is that the latent variables $\mathbf{z}$ are shared between the two datasets, while everything else is independent. The prior distributions for the Bayesian CCA were chosen (Klami and Kaski, 2007) as

$$\begin{aligned} \mathbf{w}_l &\sim N(0, \beta_l \mathbf{I}), \\ \beta_l &\sim IG(\alpha_0, \beta_0), \\ \Psi^x, \Psi^y &\sim IW(\mathbf{S}_0, \nu_0). \end{aligned} \quad (7)$$

Here, $\mathbf{w}_l$ denotes the lth column of $\mathbf{W}$, and IG and IW are shorthand notations for the inverse Gamma and inverse Wishart distributions. The priors for the covariance matrices Ψ^x and Ψ^y are conventional conjugate priors, and the prior for the projection matrices is the so-called automatic relevance determination (ARD) prior used for example in Bayesian principal component analysis (Bishop, 1999).

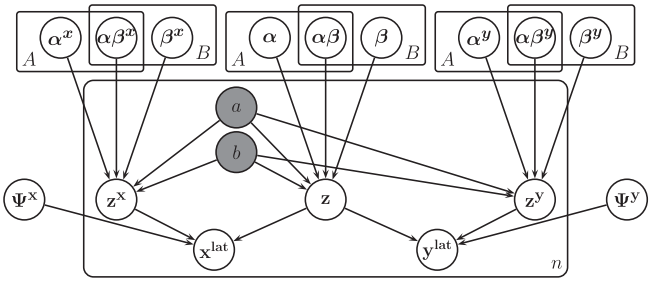

Fig. 3. The plate diagram for the decomposition of the covariate effects into shared and source-specific components.

2.3 Estimating the multivariate, two-way ANOVA-type effects

The linear model for the two-way covariate effects is set on the shared latent variables $\mathbf{z}_j$, in order to have access to effects shared between both spaces $\mathbf{x}^{lat}$ and $\mathbf{y}^{lat}$. In the K_z-dimensional latent variable space we then have

$$\mathbf{z}_j = \boldsymbol{\alpha}_a + \boldsymbol{\beta}_b + (\boldsymbol{\alpha\beta})_{ab} + \boldsymbol{\epsilon}_j, \quad (8)$$

where $\boldsymbol{\alpha}_a$ and $\boldsymbol{\beta}_b$ are the shared main effects, $(\boldsymbol{\alpha\beta})_{ab}$ is the shared interaction effect and $\boldsymbol{\epsilon}_j$ is a noise term. The effects are modeled as population priors on the latent variables, which in turn are given Gaussian priors $\boldsymbol{\alpha}_a$, $\boldsymbol{\beta}_b$, $(\boldsymbol{\alpha\beta})_{ab} \sim \mathcal{N}(0, \mathbf{I})$.

Note that the baselines (grand means) $\boldsymbol{\mu}^x$ and $\boldsymbol{\mu}^y$ are estimated in the lower level of hierarchy, that is, directly in the $\mathbf{x}$ and $\mathbf{y}$ spaces, and do not appear here.

To simplify the interpretation of the effects we center the grand means to the mean of one control population. A similar choice of baseline has been done successfully in other ANOVA studies (Lucas *et al.*, 2009; Seo *et al.*, 2007), and it does not significantly sacrifice generality. We set the parameter vector $\boldsymbol{\mu}^x$, describing variable-specific means, to the mean of the control group. One group, denoted by $(a,b)=(0,0)$, now becomes the baseline to which other classes are compared by adding main and interaction effects.

Due to the new centering, the terms $\boldsymbol{\alpha}_0$, $\boldsymbol{\beta}_0$, $(\boldsymbol{\alpha\beta})_{00}$, $(\boldsymbol{\alpha\beta})_{0b}$ and $(\boldsymbol{\alpha\beta})_{a0}$ become zero and are not estimated. The differences between the populations are now modeled directly with $\mathbf{x}^{lat}$ and $\mathbf{y}^{lat}$, and hierarchically by the main effects $\boldsymbol{\alpha}_a$, $\boldsymbol{\beta}_b$, $(\boldsymbol{\alpha\beta})_{ab}$, $a,b>0$.

In our case study, the factors a (healthy diseased) and b (untreated treated) have only two levels and we have populations $(a,b)=(0,0),(1,0),(0,1),(1,1)$, and there are hence three terms $\boldsymbol{\alpha}_1$ (disease), $\boldsymbol{\beta}_1$ (treatment) and $(\boldsymbol{\alpha\beta})_{11}$ (interaction), that model the difference to the control population $(a,b)=(0,0)$. The subscripts $_1$ and $_{11}$ will be dropped in the results section. In this case study, the healthy untreated is the natural choice for an intuitive control population.

2.4 Decomposition into shared and source-specific effects

The decomposition of the effects into shared and source-specific effects is illustrated in Figure 3. The variation of the data in the latent variable spaces $\mathbf{x}^{lat}$ and $\mathbf{y}^{lat}$ is modeled by a common term $\mathbf{z}_j$ and, in addition, the source-specific terms $\mathbf{z}_j^x$ and $\mathbf{z}_j^y$ model the variation that cannot be explained by the

shared term. The model now decomposes to

$$\mathbf{z}_j = \boldsymbol{\alpha}_a + \boldsymbol{\beta}_b + (\boldsymbol{\alpha\beta})_{ab} + \boldsymbol{\epsilon}_j, \quad (9)$$

$$\mathbf{z}_j^x = \boldsymbol{\alpha}_a^x + \boldsymbol{\beta}_b^x + (\boldsymbol{\alpha\beta})_{ab}^x + \boldsymbol{\epsilon}_j^x, \quad (10)$$

$$\mathbf{x}_j^{lat} \sim \mathcal{N}(\mathbf{W}_{\text{shared}}^x \mathbf{z}_j + \mathbf{W}_{\text{specific}}^x \mathbf{z}_j^x, \Psi^x). \quad (11)$$

The $\boldsymbol{\alpha}_a^x$ and $\boldsymbol{\beta}_b^x$ are the source-specific main effects, $(\boldsymbol{\alpha\beta})_{ab}^x$ is the source-specific interaction effect, $\boldsymbol{\epsilon}_j$ and $\boldsymbol{\epsilon}_j^x$ are noise terms and $\mathbf{W}_{\text{shared}}^x$ and $\mathbf{W}_{\text{specific}}^x$ are the shared and specific components of the projection matrix $\mathbf{W}^x$, respectively. The equations are analogous for y. In practice, the decomposition is implemented by restricting a column of $\mathbf{W}^x$ in Equation. (6) be zero for the y-specific components and vice versa for x.

In summary, the complete hierarchical model of Figure 2 is

$$\begin{aligned}
\boldsymbol{\alpha}_0 = 0, \boldsymbol{\beta}_0 = 0, (\boldsymbol{\alpha\beta})_{a0} = 0, (\boldsymbol{\alpha\beta})_{0b} = 0 \\
\boldsymbol{\alpha}_a, \boldsymbol{\beta}_b, (\boldsymbol{\alpha\beta})_{ab}, \boldsymbol{\alpha}_a^x, \boldsymbol{\beta}_b^x, (\boldsymbol{\alpha\beta})_{ab}^x \sim \mathcal{N}(0, \mathbf{I}) \\
\mathbf{z}_j|_{j\in a,b} \sim N(\boldsymbol{\alpha}_a + \boldsymbol{\beta}_b + (\boldsymbol{\alpha\beta})_{ab}, \mathbf{I}) \\
\mathbf{z}_j^x|_{j\in a,b} \sim N(\boldsymbol{\alpha}_a^x + \boldsymbol{\beta}_b^x + (\boldsymbol{\alpha\beta})_{ab}^x, \mathbf{I}) \\
\mathbf{x}_j^{lat} \sim N(\mathbf{W}_{\text{shared}}^x \mathbf{z}_j + \mathbf{W}_{\text{specific}}^x \mathbf{z}_j^x, \boldsymbol{\Psi}^x) \\
\mathbf{x}_j \sim \mathcal{N}(\boldsymbol{\mu}^x + \mathbf{V}^x \mathbf{x}_j^{lat}, \boldsymbol{\Lambda}^x). \quad (12)
\end{aligned}$$

2.5 Data preprocessing and model complexity selection

For simplicity and to reduce the number of parameters of the model, the data are preprocessed such that for each variable the mean of the control population ($a=0$, $b=0$) is subtracted and the variable is scaled by the SD of the control population. This fixes the scales λ_i to one and the $\boldsymbol{\mu}^x$ and $\boldsymbol{\mu}^y$ to zero. The factor analysis part now models correlations of the variables. The possible covariate effects are now comparable with the control population as discussed in Chapter 2.3.

Model complexity, that is, the number of clusters and latent variables, is chosen separately for both $\mathbf{x}^{lat}$ and $\mathbf{y}^{lat}$ by predictive likelihood in 10-fold cross-validation as in Huopaniemi *et al.* (2009).

2.6 Gibbs-formulas

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:

$$\mathbf{z}_j \sim \mathcal{N}(\hat{\boldsymbol{\mu}}_j^z, \hat{\boldsymbol{\Sigma}}^z), \quad (13)$$

where

$$\hat{\boldsymbol{\mu}}_j^z = \hat{\boldsymbol{\Sigma}}^z (\mathbf{W}^T \boldsymbol{\Psi}^{-1} \mathbf{y}_j^{lat} + \boldsymbol{\alpha}_a + \boldsymbol{\beta}_b + (\boldsymbol{\alpha\beta})_{ab}), \quad (14)$$

$$\hat{\boldsymbol{\Sigma}}^z = (\mathbf{W}^T \boldsymbol{\Psi}^{-1} \mathbf{W} + \mathbf{I})^{-1}. \quad (15)$$

Here $\mathbf{v}_j^{lat} = [\mathbf{x}_j^{lat}; \mathbf{y}_j^{lat}]$, $\mathbf{W} = [\mathbf{W}^x; \mathbf{W}^y]$ and $\boldsymbol{\Psi} = [\boldsymbol{\Psi}^x \ \mathbf{0}; \ \mathbf{0} \ \boldsymbol{\Psi}^y]$. In this Subsection, we denote $\boldsymbol{\alpha}_a = [\boldsymbol{\alpha}_a; \boldsymbol{\alpha}_a^x; \boldsymbol{\alpha}_a^y]$. For $\mathbf{x}$,

$$\mathbf{x}_j^{lat} \sim \mathcal{N}(\hat{\boldsymbol{\mu}}_j^{lat}, \hat{\boldsymbol{\Sigma}}^{lat}), \quad (16)$$

where

$$\hat{\boldsymbol{\mu}}_j^{lat} = \hat{\boldsymbol{\Sigma}}^{lat} ((\mathbf{V}^x)^T (\boldsymbol{\Lambda}^x)^{-1} \mathbf{x}_j + \boldsymbol{\Psi}^x \mathbf{W}^x \mathbf{z}_j), \quad (17)$$

$$\hat{\boldsymbol{\Sigma}}^{lat} = ((\mathbf{V}^x)^T (\boldsymbol{\Lambda}^x)^{-1} \mathbf{V}^x + (\boldsymbol{\Psi}^x)^{-1})^{-1}. \quad (18)$$

$$\boldsymbol{\alpha}_a \sim \mathcal{N}(\frac{1}{n_a+1} \sum_{j\in a} (\mathbf{z}_j - \boldsymbol{\beta}_{b_j} - (\boldsymbol{\alpha\beta})_{ab_j}), \frac{1}{n_a+1} \boldsymbol{I}), \quad (19)$$

$$\boldsymbol{\beta}_b \sim \mathcal{N}(\frac{1}{n_b+1} \sum_{j\in b} (\mathbf{z}_j - \boldsymbol{\alpha}_{a_j} - (\boldsymbol{\alpha\beta})_{a_j b}), \frac{1}{n_b+1} \boldsymbol{I}), \quad (20)$$

$$(\boldsymbol{\alpha\beta})_{ab} \sim \mathcal{N}(\frac{1}{n_{ab}+1} \sum_{j\in ab} (\mathbf{z}_j - \boldsymbol{\alpha}_{a_j} - \boldsymbol{\beta}_{b_j}), \frac{1}{n_{ab}+1} \boldsymbol{I}), \quad (21)$$

and similarly for $\mathbf{y}$. Here 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 regularized projection matrix, or clustering matrix $\mathbf{V}^x$ is sampled one row at time. The cluster assignment of variable i, v_i, is the location of the non-zero scale parameter λ_i. The formula is

$$p(v_i = k) = \frac{\pi_k \Pi_j p(x_{ji} | \lambda_i x_{jk}^{lat}, \sigma_i)}{\Sigma_k \pi_k \Pi_j p(x_{ji} | \lambda_i x_{jk}^{lat}, \sigma_i)}, \quad (22)$$

$$\sigma_i^2 \sim \text{Inv-}\chi^2(n, \Sigma_j (x_{ji} - \lambda_i x_{jk}^{lat})^2). \quad (23)$$

Sampling of the CCA-projection matrices is explained in Klami and Kaski (2007).

2.7 Study design

The sample series corresponds to 28 mice divided into four groups. For the two tumor-bearing groups, Lewis lung cancer cell suspension containing 8106 cells in 0.2 ml saline was injected subcutaneously. The other two groups were non-tumor bearing normal mice. Twenty-four hours after tumor inoculation, 100 mg/kg/d Rh2 in 0.5% sodium carboxymethyl cellulose (CMC-Na) was given by gavage to one tumor-bearing group (TR, diseased treated) and one non-tumor bearing group (NR, healthy treated). The other two groups were given blank CMC-Na as controls, they are named group NS for healthy-untreated and TS diseased-untreated for non-tumor bearing and tumor-bearing groups, respectively. Both Rh2 and CMC-Na groups were treated daily for 21 days. By the end of experiment, blood was collected from the orbital sinus and animals were killed by cervical dislocation. The brain, heart, lung, liver, spleen, thymus and tumor were collected and frozen in liquid nitrogen immediately.

Samples were analyzed by ultra performance liquid chromatography (UPLC)/Mass Spectrometry (MS) as described previously (Kotronen *et al.*, 2010), and the raw data was preprocessed by MZmine (Katajamaa *et al.*, 2006). Lipids were then identified according to RT and m/z values, resulting in 168 lipids in blood plasma, 68 in lung and 58 in heart.

3 RESULTS

We first demonstrate the functioning of the method on simulated data, focusing on how the generated effects are found when the sample size is small. We then apply the method to a lung cancer study where lipidomic profiles have been measured from several tissues of model mouse samples, under a two-way experimental setup (disease and treatment). Different tissues have distinct lipids. Finally, the method is compared with a standard (one-source) MANOVA-approach including dimension reduction with PCA.

3.1 Simulated data

We generate data having known effects, and then study how well the model finds the effects as a function of the number of measurements. There are three generated effects, in α, β^y and $(\alpha\beta)^x$.

Each of the three effects has strength +2, the $\mathbf{x}^{lat}$ and $\mathbf{y}^{lat}$ are both 3D, and the $\mathbf{x}$ and $\mathbf{y}$ are 200D. The marginal covariance matrices $\boldsymbol{\Psi}^x$ and $\boldsymbol{\Psi}^y$ have diagonal variance of magnitude 5 and off-diagonal covariance of magnitude 4. The $\sigma_i = 1$ for each variable i in $\mathbf{x}$ and $\mathbf{y}$, and $p(v_i = k)$ is uniform for each i and k. The $\mathbf{W}_{\text{shared}}^x = [1, 0, -1]$, $\mathbf{W}_{\text{specific}}^x = [0, 1, 0]$, $\mathbf{W}_{\text{shared}}^y = [1, 0, -1]$ and $\mathbf{W}_{\text{specific}}^y = [0, 1, 0]$. Three components are estimated, one shared and two source-specific components. The model is computed by Gibbs sampling, discarding 1000 burn-in samples, and collecting 1000 samples for inference. To fix the sign of the effects without affecting the results, each posterior distribution is mirrored, if necessary, to have a positive mean, i.e. multiplied by the sign of the posterior mean.

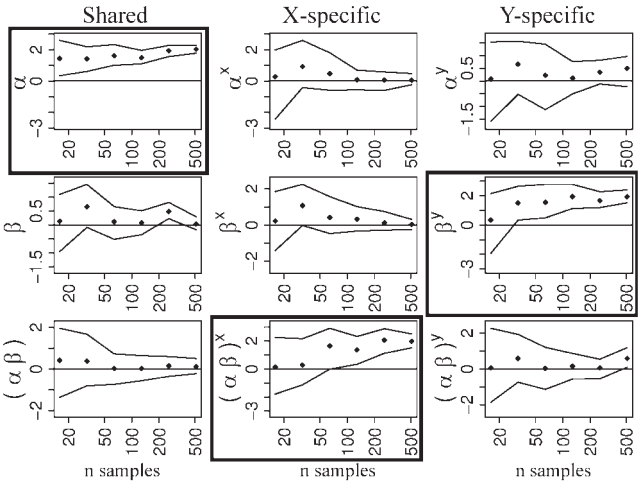

Fig. 4. The method finds the generated effects $\alpha = +2$, $\beta^y = +2$ and $(\alpha\beta)^x = +2$, encircled by the thick black boxes. The dots show posterior mean and the thin lines include 95% of posterior mass, at each number of observations. A consistently non-zero posterior distribution implies a found effect.

The method finds the three generated effects, shown with black boxes in Figure 4. The uncertainty decreases with increasing number of observations. Note that the shared effect is found with much less uncertainty since there is evidence from both sources. With low numbers of samples, there is considerable uncertainty in the effects for source-specific components, which can be interpreted as follows: with such a low number of samples, there is not enough evidence for the effect. In typical bioinformatics applications there may be 20–50 samples. The model also found the correct clusters of variables (data not shown).

3.2 Lung cancer study

We then study data from a two-way, two-source, $n \ll p$, so far unpublished lung cancer mouse model experiment. The diseased mice are compared with healthy control samples and, in addition, some mice from both groups have been given a test anticancer drug treatment. There are thus healthy untreated ($n = 9$ mice), diseased untreated ($n = 7$), healthy treated ($n = 6$) and diseased treated ($n = 6$) samples. Lipidomic profiles have been measured by UPLC/MS. The study has a two-way experimental setup, such that the disease effect α, treatment effect β and an interaction effect ($\alpha\beta$) on lipid groups are to be estimated. The high-dimensional lipidomic profiles have been measured from several tissues of each mouse; the tissues have partly different lipids that have not been matched, and even the roles of the matched lipids may be different in different tissues. Hence, the tissues have different feature spaces with paired samples, implying a two-source study. We specifically focused on the relationship between blood and lung tissue profiles. This is particularly relevant for drug efficacy and safety studies in non-clinical and particularly in clinical studies, given that plasma samples can be easily collected in such investigations.

3.2.1 Experiment 1: effects shared by blood and disease tissue Blood plasma (168 lipids) and lung tissue (68 lipids) were integrated with the method. The optimal number of clusters for plasma was six and for lung five, found based on predictive likelihood.

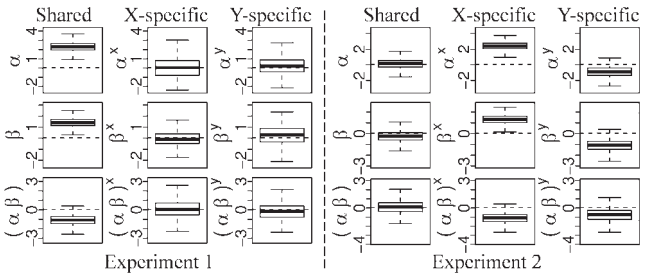

Fig. 5. In Experiment 1 (left), the method finds a disease effect α and a treatment effect β shared between the two sources, plasma (**x**) and lung (**y**) tissues. In Experiment 2 (right), the method finds only source-specific effects in plasma (**x**) when integrating with the heart tissue (**y**). No effects are found in heart. The boxplots show quartiles and 95% intervals of posterior mass of the effects; a consistently non-zero posterior distribution implies an effect is found.

Three components are learned, one shared and two source-specific. The method finds (Fig. 5, left) a disease effect α and treatment effect β shared by both sources; the shared interaction effect is close but not significant (95% confidence).

The results imply that a cluster of 12 lipids in lung and a cluster of 20 lipids in blood are mutually coherently upregulated due to disease, and additionally upregulated by the treatment. Another cluster of 13 lipids in lung was found to be downregulated due to the disease and additionally downregulated due to treatment. The lipids of the downregulated cluster are thus negatively correlated with the upregulated clusters. The results indicate that there might be a shared interaction effect ($\alpha\beta$) with opposite direction, that might indicate a cure for the disease. However, the effect is not significant using the common 95% rule. The 28 samples used in the analysis thus do not give enough evidence for this effect being consistent, and there is no confirmation that the treatment would cure the cancer effects. This confirms our prior concern that the specific treatment might not be efficient. The treatment does, however, affect the same groups of lipids as the disease, so investigating it as a potential cure was not a far-fetched hypothesis.

The effects are traced back to the metabolite groups, by first identifying the responsible row of $\mathbf{W}^x$ and hence component of $\mathbf{x}^{lat}$, and then the lipid cluster from the $\mathbf{V}^x$ corresponding to the $\mathbf{x}^{lat}$ component.

The upregulated cluster of plasma contains abundant triacylglycerols known to be coregulated (Kotronen *et al.*, 2009), the upregulated cluster of lung contains lipotoxic ceramides (Summers, 2006) and proinflammatory lysophosphatidylcholines (Mehta, 2005), while the downregulated cluster of lung contains ether lipids, known as endogenous antioxidants (Brites *et al.*, 2004). Our analysis reveals that the drug treatment enhances, not diminishes, the proinflammatory lipid profile found in the disease. Our findings thus indicate that the tissue-specific changes due to cancer can be monitored by specific plasma lipids. Clearly more studies are needed to establish a mechanistic link between the plasma triacylglycerols and the observed inflammatory changes in the lung.

3.2.2 Experiment 2: when integrated with a non-diseased tissue, only source-specific effects are found We then integrate plasma **x** with another tissue, heart (58 lipids) **y**. The results in Figure 5 (right), show that the disease effect and the treatment effect are found only

Table 1. The comparison method finds only one effect from the real dataset

Experiment 1 Plasma and lung			Experiment 2 Plasma and heart		
α	β	$\alpha\beta$	α	β	$\alpha\beta$
0.02	0.82	0.67	**0.01**	0.78	0.28

P-value from two-way 50–50 MANOVA where lipidomics profile from the tissues have been integrated by naive concatenation. The bold values indicate a statistically significant effect ($P < 0.05$).

in the source-specific component of plasma. This implies that there is no evidence of shared effects between plasma and heart, and in fact no consistent effects are found for the heart tissue. The method finds, however, the same effects, disease α and treatment β in the plasma as in Experiment 1 and for the same cluster of lipids, which signifies that the method works well.

In summary, Experiments 1 and 2 combined provide strong evidence towards the validity of the model: shared effects were found between tissues having a functional link (plasma and tissue containing tumor), whereas no link was found when there was a more remote functional link (between plasma and some other tissue). Then disease effects were only found in plasma. The specific found effects need to be studied further.

3.3 Comparison to existing ANOVA-type methods

No earlier multi-way ANOVA-type method can be properly used to analyze data from multi-way, multi-source experiments; they do not allow access to shared multi-way effects and therefore cannot be used for decomposing covariate effects into shared and source-specific ones. Three imperfect workarounds are possible, presented below in the two-way, two-source case:

(1) Separate single-source two-way analysis for each source.

(2) Concatenation of the datavectors **x** and **y** from the different sources to a longer vector, followed by a single-source two-way MANOVA (with PCA dimension reduction). Such standard MANOVA naturally cannot separate the source-specific effects from the shared ones at all.

(3) In the rare special case, where the variables between the sources **x** and **y** have the same dimensionality and the variables of the two sources have a full, a priori known matching, a naive alternative is to ignore the pairing and consider the source as a covariate in a three-way MANOVA, as in Equation 2. This is not possible in our case study since variable spaces are different.

We compare our method to concatenating the sources and performing a two-way 50-50 MANOVA (Langsrud, 2002), MANOVA based on a PCA-dimension reduction, with the real lipidomics data and with simulated data.

3.3.1 Comparison with real lipidomics data The lipidomic dataset, with results presented in Figure 5, was analyzed by concatenating the two data sources and performing 50–50 MANOVA.

The results in Table 1 point out several interesting differences to the results obtained with our proposed method (Fig. 5). First, the

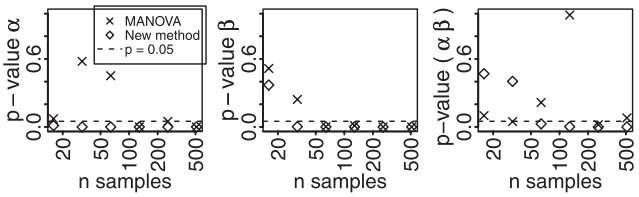

Fig. 6. The proposed new method finds the main effects already with 32 samples and the interaction effect from 64 samples. The comparison method, 50–50 MANOVA for concatenated data sources, does not find the effects reliably for less than 200 samples. The comparison method naturally cannot distinguish in which data sources the effects are. Results are from the simulated data of Figure 4 as a function of number of samples. The lines mark the $P = 0.05$ threshold; the effects below the threshold are considered statistically significant.

P-values of the covariate effects are given to the overall contatenated sources, and the method has no means of differentiating shared and source-specific components. Single-source or univariate analyses would need to be done for the specific components, but they would lose the access to dependencies between the sources and therefore to the shared effects. Second, as a result of using the unoptimal PCA-dimension reduction, only the disease effect α is found in each experiment, and the treatment effect β is not found. This demonstrates the superior behavior of the multi-way, multi-source analysis method that includes a dimension reduction as an integrated clustering of similarly behaving variables. Third, MANOVA gives only a *P*-value, and the locations (which variables) and direction (up/down) of the effects have to be deduced by other methods, such as univariate ANOVA and *t*-tests.

3.3.2 Comparison with simulated data We compare the simulated data presented in Figure 4 with 50–50 MANOVA analysis with concatenated data sources as a function of sample size. The confidences for the effects, shown as distributions in Figure 4, are converted to a numerical confidence value comparable with the *P*-values given by MANOVA; a distribution above or below zero with 95% confidence is considered significant. In addition, the decomposition to shared and source-specific effects is ignored since MANOVA is unable to perform it. The comparison results are shown in Figure 6. The proposed method outperforms MANOVA, especially with small sample sizes, which is the case of interest. The proposed method finds the main effects with 32 samples and the interaction effect with 64 samples, whereas the comparison method needs 256 samples to reliably find all the effects from the data.

4 DISCUSSION

We have generalized ANOVA-type multi-way analysis to cases where measurements from multiple sources are available for samples having a multi-way experimental setup. Furthermore, the method is able to decompose the covariate effects to shared and source-specific effects, unlike any existing methods. The problem is solved by a hierarchical latent variable model that extends the generative model of Bayesian CCA to model multi-way covariate information of samples, by assigning population-specific priors on the shared latent variable of CCA.

The method is designed for cases with high dimensionality and small sample size, common in bioinformatics applications.

The small sample size problem was solved by assuming that the variables come in similarly behaving groups, which is reasonable for the 'omics' applications such as metabolomics in this article. An alternative approach could be using sparse approaches, for instance L_1-regularization or point-mass mixture priors (West, 2003), in applications where the clustering assumption is unrealistic.

The modeling task is extremely difficult due to the complexity of the task and small sample size. Hence, it was striking that the method was capable of finding covariate effects in a real-world lipidomic multi-way, multi-source dataset.

In practice, the model is applicable for finding disease- or other covariate-related effects across multiple paired measurements, from different sources having different variables, which is an increasingly common and important data-analysis task. The key assumption needed for connecting different types of measurements is sample pairing, which allows formulating shared latent variables between the sources.

The study on lung cancer showed that the model is capable of separating effects shared by the sources from effects in one source only. When integrating metabolic profiles from two tissues (blood and lung) which both had disease and treatment effects, the model was able to model these effects as shared disease and treatment effects, establishing a covariate-related dependency between the tissues.

When the same blood tissue was integrated with heart tissue not expected to show disease or treatment effects, the method found no significant shared effects but modeled the effects in blood as source-specific disease and treatment effects. This result provides evidence in support of the method and its modeling assumptions.

In the case of simulated data, the generated shared and source-specific covariate effects were found with relatively small sample sizes, clearly outperforming the alternative MANOVA approach. The shared effect was found already with considerably smaller number of samples, since there is evidence from both sources.

For simplicity, the method was presented in the two-way, two-source case. Generalization to multiple 'ways' is straightforward by adding more main and interaction terms to Equation 8, and to multiple sources by adding sources $(x, y, \ldots)$ to Equation 6. Both generalizations have been tested with simulated data (details not reported here) where effects were found equally well.

Finally, the method is not restricted to metabolomics; it is generally applicable to other data types and especially well suited for combinations of heterogeneous data sources, whenever the variables can be assumed to form similarly behaving functional groups, such as in gene expression and proteomics.

As more and more biological multi-source datasets with paired samples integrating data from different measurement techniques and/or tissues are becoming available, we expect the method to find wide applicability.

ACKNOWLEDGEMENTS

I.H., T.S. and S.K. belong to the Adaptive Informatics Research Centre. Special thanks to Arto Klami for helping with the model development.

Funding: This work was supported by Tekes [MASI program, Multibio project, grant number 40274/06 (to J.N.)]; Graduate School of Computer Science and Engineering [(to I.H.)]; and European Union [FP7 NoE PASCAL2, ICT 216886 (to S.K.)].

Conflict of Interest: none declared.

REFERENCES

Archambeau,C. and Bach,F. (2009) Sparse probabilistic projections. In Koller,D. *et al*. (eds) *Advances in Neural Information Processing Systems 21*, MIT Press, Cambridge, MA, pp. 73–80.

Bach,F.R. and Jordan,M.I. (2005) A probabilistic interpretation of canonical correlation analysis. *Technical Report 688*, Department of Statistics, University of California, Berkeley.

Bishop,C.M. (1999) Bayesian PCA. In Kearns,M.S. *et al*. (eds) *Advances in Neural Information Processing Systems*. Vol. 11, MIT Press, Cambridge, MA, pp. 382–388.

Brites,P. *et al.* (2004) Functions and biosynthesis of plasmalogens in health and disease. *Biochim. Biophys. Acta*, **1636**, 219–231.

Carvalho,C. *et al.* (2008) High-dimensional sparse factor modeling: applications in gene expression genomics. *J. Am. Stat. Assoc.*, **103**, 1438–1456.

Girolami,M. and Zhong,M. (2007) Data integration for classification problems employing gaussian process priors. In Schölkopf,B. *et al*. (eds) *Advances in Neural Information Processing Systems.*, vol. 19, MIT Press, Cambridge, MA, pp. 465–472.

Hotelling,H. (1936) Relations between two sets of variates. *Biometrika*, **28**, 321–377.

Huopaniemi,I. *et al.* (2009) Two-way analysis of high-dimensional collinear data. *Data Min. Knowl. Discov.*, **19**, 261–276.

Katajamaa,M. *et al.* (2006) Mzmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics*, **22**, 634–636.

Klami,A. and Kaski,S. (2007) Local dependent components. In Ghahramani,Z. (ed.), *Proceedings of ICML 2007, the 24th International Conference on Machine Learning*, Omnipress, Madison, Wisconsin, pp. 425–433.

Kotronen,A. *et al.* (2009) Serum saturated fatty acids containing triacylglycerols are better markers of insulin resistance than total serum triacylglycerol concentrations. *Diabetologia*, **52**, 684–690.

Kotronen,A. *et al.* (2010) Comparison of lipid and fatty acid composition of the liver, subcutaneous and intra-abdominal adipose tissue, and serum. *Obesity*. Available at: http://www.nature.com/oby/journal/vaop/ncurrent/pdf/oby2009326a.pdf.

Langsrud,O. (2002) 50-50 multivariate analysis of variance for collinear responses. *J. Roy. Stat. Soc. Series D-the Statistician*, **51**, 305–317.

Le Cao,K.-A. *et al.* (2009) Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC Bioinformatics*, **10**, 34.

Lucas,J.E. *et al.* (2009) Cross-study projections of genomic biomarkers: an evaluation in cancer genomics. *PLoS ONE*, **4**, e4523.

Mehta,D. (2005) Lysophosphatidylcholine: an enigmatic lysolipid. *Am. J. Physiol. Lung Cell. Mol. Physiol.*, **289**, 174–175.

Parkhomenko,E. *et al.* (2007) Genome-wide sparse canonical correlation of gene expression with genotypes. *BMC Proc.*, **1**(Suppl. 1), S119.

Roweis,S. and Ghahramani,Z. (1999) A unifying review of linear gaussian models. *Neural Comput.*, **11**, 305–345.

Seo,D.M. *et al.* (2007) Of mice and men: Sparse statistical modelling in cardiovascular genomics. *Ann. Appl. Stat.*, **1**, 152–178.

Smilde,A.K., *et al.* (2005) ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data. *Bioinformatics*, **21**, 3043–3048.

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

Summers,S.A. (2006) Ceramides in insulin resistance and lipotoxicity. *Prog. Lipid Res.*, **45**, 42–72.

Waaijenborg,S. and Zwinderman,A. (2007) Penalized canonical correlation analysis to quantify the association between gene expression and DNA markers. *BMC Proc.*, **1**(Suppl. 1), S122.

Wang,C. (2007) Variational Bayesian approach to canonical correlation analysis. *IEEE Trans. Neural Net.*, **18**, 905–910.

Webb-Robertson,B.-J.M. *et al.* (2009) A Bayesian integration model of high-throughput proteomics and metabolomics data for improved early detection of microbial infections. In *Pacific Symposium on Biocomputing*, pp. 451–463.

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

Witten,D. and Tibshirani,R. (2009) Extensions of sparse canonical correlation analysis, with applications to genomic data. *Stat. Appl. Genet. Mol. Biol.*, **8**, Article 28.