Graphical Multi-way Models

Ilkka Huopaniemi[1,⋆], Tommi Suvitaival[1], Matej Orešič[2], and Samuel Kaski[1]

[1] Aalto University School of Science and Technology,
Department of Information and Computer Science, Helsinki Institute for
Information Technology HIIT, P.O. Box 15400, FI-00076 Aalto, Finland
[2] VTT Technical Research Centre of Finland, P.O. Box 1000, FIN-02044 VTT,
Espoo, Finland
{ilkka.huopaniemi,tommi.suvitaival,samuel.kaski}@tkk.fi,
matej.oresic@vtt.fi
http://www.cis.hut.fi/projects/mi

Abstract. Multivariate multi-way ANOVA-type models are the default tools for analyzing experimental data with multiple independent covariates. However, formulating standard multi-way models is not possible when the data comes from different sources or in cases where some covariates have (partly) unknown structure, such as time with unknown alignment. The "small n, large p", large dimensionality p with small number of samples n, settings bring further problems to the standard multivariate methods. We extend our recent graphical multi-way model to three general setups, with timely applications in biomedicine: (i) multi-view learning with paired samples, (ii) one covariate is time with unknown alignment, and (iii) multi-view learning without paired samples.

Keywords: ANOVA, Bayesian latent variable modeling, data integration, multi-view learning, multi-way learning.

1 Introduction

Multivariate multi-way ANOVA-type methods are the default tool for analyzing data with multiple covariates. A prototypical example in biomedical data analysis is studying the effects of disease and treatment in populations of biological measurements. Formulating the data analysis as a linear model makes it possible to ask if the covariates ("ways", disease and treatment), or more interestingly, their interactions have an effect on the data.

In the two-way case, to explain the covariate-related variation in one data source, say $\mathbf{x}$, the following linear model is usually assumed:

$$\mathbf{x}_j|_{(a,b)} = \boldsymbol{\mu}^x + \boldsymbol{\alpha}_a^x + \boldsymbol{\beta}_b^x + (\boldsymbol{\alpha\beta})_{ab}^x + \epsilon_j. \quad (1)$$

[⋆] I.H., T.S and S.K belong to the Adaptive Informatics Research Centre and Helsinki Institute for Information Technology HIIT. The work was funded by Tekes MASI program and by Tekes Multibio project. 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.

J.L. Balcázar et al. (Eds.): ECML PKDD 2010, Part I, LNAI 6321, pp. 538–553, 2010.

Here $\mathbf{x}_j$ is a continuous-valued data vector, observation number j, and the a and b ($a = 0, \ldots A$ and $b = 0, \ldots B$) are the two independent covariates, such as disease and treatment. The $\boldsymbol{\alpha}_a^x$ and $\boldsymbol{\beta}_b^x$ are parameter vectors describing the covariate-specific effects, called main effects. The $(\boldsymbol{\alpha\beta})_{ab}^x$ denotes the interaction effect; the apparently complicated notation is standard, it simply means a parameter vector. In the biomedical example this interaction is the most interesting parameter, describing if the treatment has disease-specific effects (cures the disease). These effects model the variation from the baseline level (called grand mean) $\boldsymbol{\mu}^x$. The ϵ_j is a noise term. The traditional methods for finding and testing the statistical significance of the effects of the covariates on the data are Analysis of Variance (ANOVA) [4] and its multivariate generalization (MANOVA).

A recurring problem in modern data analyses, especially in biomedical experiments, is that the number of samples n is small and dimensionality p is large. The "small n, large p" has recently gained increasing attention in the machine learning community, whereas only a few methods for multi-way modelling have been reported. The currently popular approaches, multi-task learning and multi-label prediction that attempt to share statistical strength between related tasks help if tasks are assumed related, but are not targeted for studying the effects of multiple independent covariates in the data.

It is evident that with small sample-sizes, harsh dimension reduction is needed and the modelling should be done in a low-dimensional latent factor space, say $\mathbf{x}^{lat}$. In addition to trivial approaches such as a prior PCA dimension reduction, two approaches exist for multivariate multi-way analysis in the case of "small n, large p"-conditions. The first, intended for modelling the effects of multiple covariates, is Sparse factor regression [10,14].

The second, hierarchical generative modelling approach [6] forms factors by assuming the variables are grouped, and the variation of the latent variables is generated by the external covariates $p(\mathbf{x}^{lat}|a, b)$, in the spirit of linear models.

We will now extend multi-way modelling to three novel tasks which cannot be solved by standard ANOVA-models or our earlier [6], and not easily by supervised regression/classification either. The associated machine learning problems are illustrated in Figure 1.

We first consider multi-way analysis when the data comes from different sources ("views") with different domains (has unmatched data spaces). A typical biological example is using two or more measurement techniques or having measurements from several tissues of each individual, the underlying experiment having a multi-way experimental setup. We consider the "view" as an additional "way" (or covariate) in the multi-way analysis. However, since different views have different domains, a standard multi-way model is not applicable. The model we present extends multi-view learning with paired samples into multi-way cases, which has plenty of applications in modern molecular biological experiments in terms of integration of multiple data sources. This first extension has already been described in [5];. we include it for completeness.

We then extend multi-way learning into cases where, for one of the "ways", the covariates have partly unknown structure. We concentrate on time, having an unknown alignment; an intuitive application is learning of unknown alignments with Hidden Markov Models (HMM) having linear chains. We consider "time with unknown alignment" as one covariate in a multi-way model. An example considered in this paper is having time-series measurements with unknown alignment from both healthy and diseased populations. The modelling task is to find, based on data, the effect of time, the effect of the other covariate(s) (disease) and, most interestingly, their interaction. The time alignment is learned at the same time.

As a third extension we consider integrating multiple views in different domains when even the samples are not paired. This almost impossible task becomes weakly possible if the experiments are similar in the sense of having a similar covariate design. An example of "multi-view learning without paired samples" is having a similar healthy-diseased time-series dataset with unknown alignments from two species, man and mouse, with different variable-spaces. We assume and search for some shared covariate-related behavior in the datasets. We propose "view without paired samples" as a covariate for an extended multi-way analysis. This makes it possible to evaluate statistical significance of shared covariate-related behavior, in contrast to having only view-specific effects (interaction effect of "view" and other covariates). We choose the generative approach [6] to extend multi-way modelling to the novel cases, because its hierarchical structuring of the effects acting on latent variables makes the extensions reasonable. The new modelling elements, the generative model of Canonical Correlation Analysis (CCA) [1,7], a standard method for multi-view learning with paired samples and unmatched dimensions, and the HMM-model, for time-dependent covariates turn out to be fully compatible with the generative multi-way modelling approach.

We will call the different types of covariates as follows: Covariates which can in principle be studied with existing ANOVA-type methods are **standard covariates**; examples include disease, treatment, gender. "Time with unknown alignment" is a special case of a **covariate with unknown structure**. "View" is a **view-covariate** in the case of paired (co-occurring) samples, and "view without paired samples" is a **view-covariate** in the case of no pairing between the samples.

The key point why we need to distinguish **view-covariates** from the standard covariates is that it actually does not make sense to define main effects for the view-covariates at all, since the domains of the views are different. However, it is sensible to define interaction effects *b*etween a view-covariate and a standard covariate (including "time with unknown alignment"). This allows us to rigorously decompose standard covariate effects into shared and view-specific effects. Furthermore, this decomposition actually forms the connection between the different views, allowing the multi-way problem to be formulated in the first place.

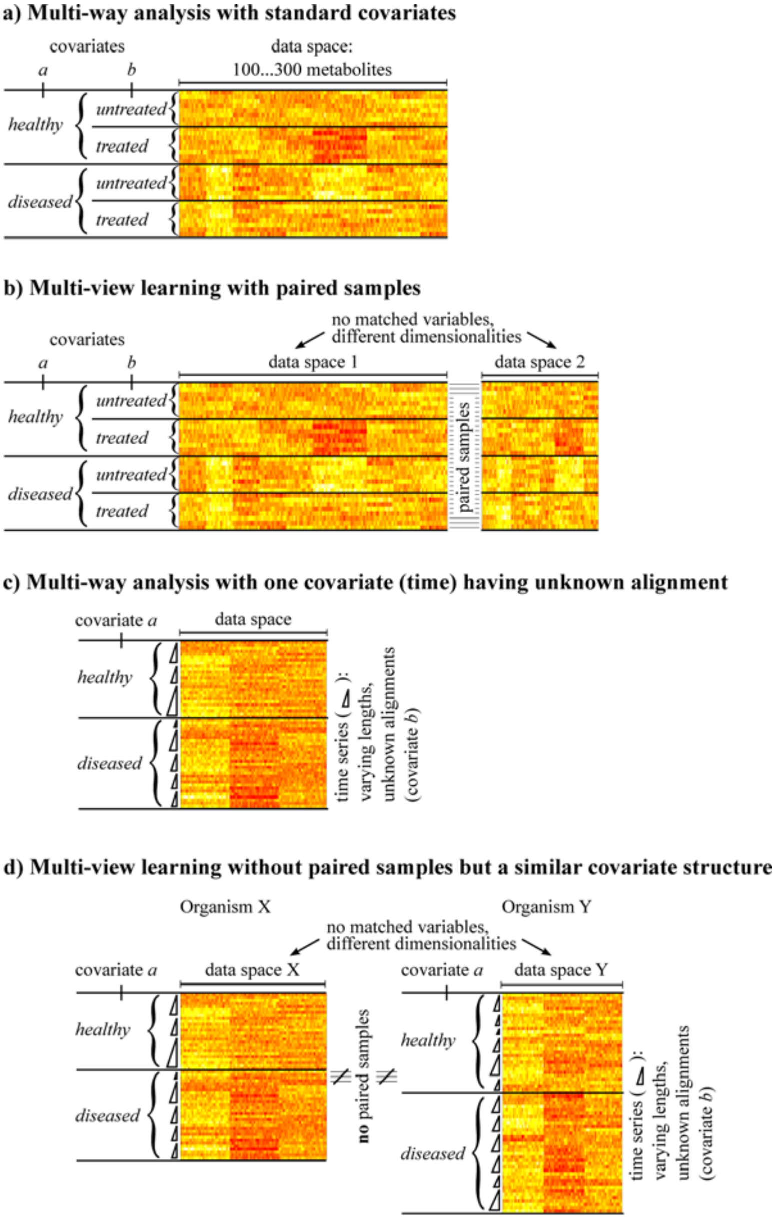

Fig. 1. Illustration of the four data analysis tasks in this paper. (a) Standard ANOVA setup, but with large dimensionality (metabolites) compared to number of samples (rows). (b) Extension to multi-view learning with paired samples. (c) Extension to time with unknown alignment. (d) Extension to multi-view learning without paired samples. The images represent data matrices, where rows are samples and columns are variables. The illustration represents the experimental design of each task, composed of a combination of standard covariates (disease, treatment), time-series information, and integration of multiple views.

The main message and contribution of this paper is that each of the three introduced new machine learning problems is too complicated to be analyzed with any existing method. We will show how to conceptualize each of these problems as an extended multi-way modelling task involving novel covariates. We then introduce a hierarchical generative model for each problem.

2 Model

We now present a unified framework to each of the novel tasks as an extended multi-way model. In each case, it turns out that the model can be formulated as a single hierarchical generative model, which guarantees that uncertainties are propagated properly between the model parts. We use Gibbs sampling for the computations.

The models need three components: (1) a regularized dimension reduction to transform the modelling into low-dimensional latent factor spaces, (2) ANOVA-type modelling of population priors acting on the low-dimensional latent factors, (3) a proper structuring of the analysis setup according to the task. The structure of the tasks and the methodological contributions of this paper are as follows: (i) in the multi-view learning with paired samples case the co-occurring sources are integrated with a generative model of CCA, (ii) in the time-covariate case the means of the emission distributions of an HMM act as one of the latent effects while HMM-alignment is done simultaneously, and (iii) in the case of multi-view learning in different domains without paired samples, the views only share common latent effects.

2.1 Multi-way Learning with Standard Covariates

Multi-way modelling [6] in a low-dimensional factor space requires two parts: regularized dimension reduction and an ANOVA-model formulated as population priors on the latent variables. In our model these parts are integrated into a single generative model, shown in Figure 2 (a). The dimension reduction is done by a factor analyzer that is regularized to find similarly behaving, correlated groups of variables and the ANOVA-effects act on the factors, each representing a cluster of variables.

Regularized Factor analyzer. The basis of the model is a Factor Analyzer (FA). The hierarchical model implementing the factor analyzer is [6]

$$\begin{aligned} \mathbf{x}_j^{lat} &\sim \mathcal{N}(0, \mathbf{I}) \\ \mathbf{x}_j &\sim \mathcal{N}(\boldsymbol{\mu} + \mathbf{V}\mathbf{x}_j^{lat}, \boldsymbol{\Lambda}) \ . \end{aligned} \qquad (2)$$

Here $\mathbf{x}_j$ is a p-dimensional data vector, $\mathbf{V}$ is the projection matrix, and $\mathbf{x}_j^{lat}$ is the latent variable, $\boldsymbol{\Lambda}$ is a diagonal residual variance matrix with diagonal elements σ_i^2, $\boldsymbol{\mu}$ is the mean vector (parameters), $\mathbf{I}$ is the identity matrix and $\mathcal{N}$ denotes the normal distribution with mean being the first argument and covariance matrix being the second.

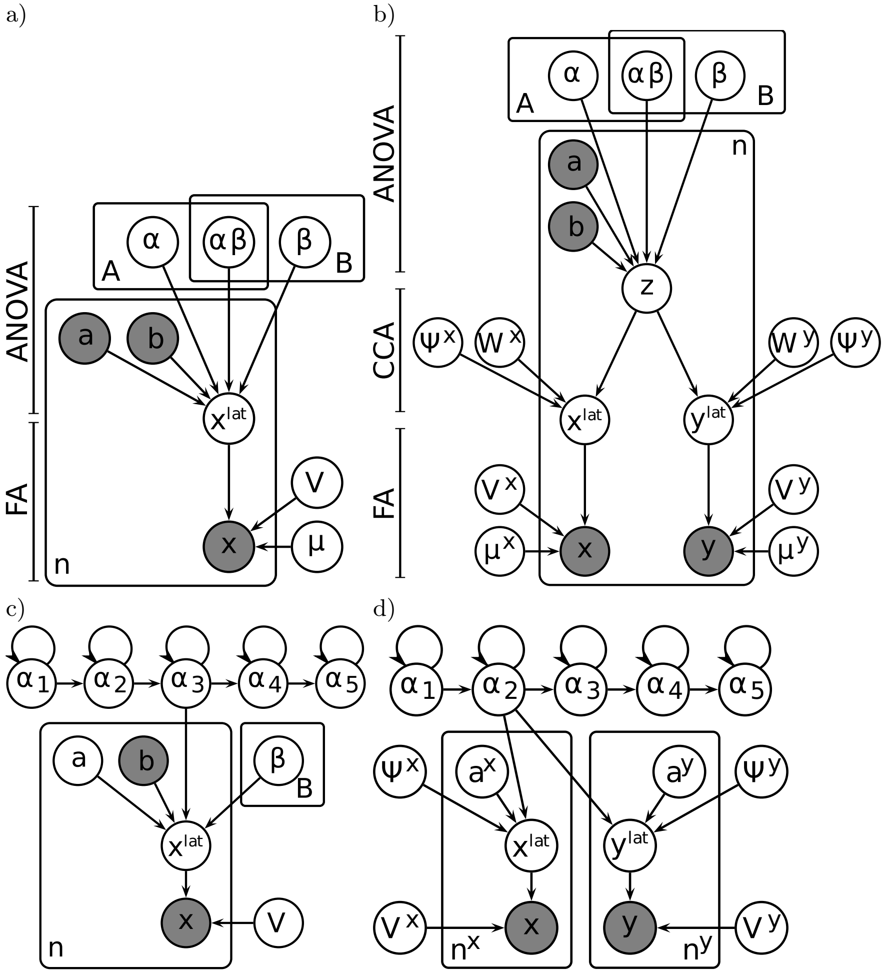

Fig. 2. The introduced model variants. (a) The hierarchical latent-variable model for standard multi-way learning with standard covariates, under "large p, small n" conditions, (b) model for multi-view learning with paired samples, (c) time with unknown alignment, (d) multi-view learning without paired samples, coupled only by shared time-course (with unknown alignment) and shared multi-way experimental design.

Since a standard factor analyzer cannot be used when $n \ll p$, we regularize projection matrix such that each variable comes from one factor only, implying a clustering assumption. The cluster indices are drawn from a multinomial distribution. The ANOVA-effects are then modelled for each cluster of correlated variables. Assuming the scales of the variables can be different, they need to be learned from data as well. For simplicity, we use a point-estimate in this paper for the scales, by scaling the variables to unit variance prior to the analysis. The number of clusters is selected by predictive likelihood [6]. The computational complexity of the models is $O(nKp + pK^2 + K^3 + nZK^2 + KZ^2 + nZ^2 + Z^3)$, where K is the number of clusters and Z is the number of CCA components. The most complex part is the clustering step, being $O(nKp)$. In small n large p conditions, dimensionality p is the main bottleneck.

ANOVA-model on latent factors. In the two-way case, the samples have two observed class covariates, $a = 0, \ldots, A$ and $b = 0, \ldots, B$. The ANOVA-modelling can now be done in the low-dimensional latent factor space,

$$\mathbf{x}_j^{lat}|_{(a,b)} = \boldsymbol{\alpha}_a + \boldsymbol{\beta}_b + (\boldsymbol{\alpha\beta})_{ab} + \text{noise}. \tag{3}$$

The ANOVA effects are set as population priors to the latent variables, which in turn are given Gaussian priors $\boldsymbol{\alpha}_a$, $\boldsymbol{\beta}_b$, $(\boldsymbol{\alpha\beta})_{ab} \sim \mathcal{N}(0, \mathbf{I})$. Note that the mean $\boldsymbol{\mu}$ is modelled in the actual data space (Equation (2)) and does not appear here.

We are now at the point where ANOVA-modelling is done in the latent factor space where the linear ANOVA-model acts as population priors. We now move into the advanced cases where "view", "time with unknown alignment" and "view without paired samples" are covariates. Gibbs-formulas have been derived analogously to [5,6,7] and the standard HMM-formalism, and are omitted due to space constraints.

2.2 Multi-view Learning with Paired Samples

We consider an ANOVA-type analysis when data comes from different views. If the data domains of the two views were the same, one might want to write a linear model

$$\mathbf{x}_d = \boldsymbol{\mu}_d + \boldsymbol{\alpha}_a + \boldsymbol{\beta}_b + (\boldsymbol{\alpha\beta})_{ab} + \boldsymbol{\gamma}_d + (\boldsymbol{\alpha\gamma})_{ad} + (\boldsymbol{\beta\gamma})_{bd} + (\boldsymbol{\alpha\beta\gamma})_{abd} + \text{noise},$$

where a and b are the two standard independent covariates, and d denotes the view. However, since the different views have different domains in general, a model cannot be written as such. It turns out that if the samples are paired (co-occur), it is possible to map the effects from latent effects to the actual data spaces $\mathbf{x}$ and $\mathbf{y}$ with unknown (estimated from the data) projections f^x and f^y as

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

Here the f^x and f^y represent a chain of projections from latent variables into the actual data spaces, shown in Figure 2 (b), for which the projection matrices are estimated from the data; we will define them implicitly in Equation (5) below.

This model now presents a desired decomposition into shared main and interaction effects $\boldsymbol{\alpha}_a$, $\boldsymbol{\beta}_b$, $(\boldsymbol{\alpha\beta})_{ab}$, and to view-specific main and interaction effects $\boldsymbol{\alpha}_a^x$, $\boldsymbol{\beta}_b^x$, $(\boldsymbol{\alpha\beta})_{ab}^x$. Equations are similar for $\mathbf{y}$. Note in particular that it is not meaningful to define a main effect for d since it is a view-effect (as discussed in the introduction), but the possibility to define interaction effects of a standard covariate and a view-covariate, such as $\boldsymbol{\alpha}_a^x$, allows ultimately the decomposition of effects into shared and view-specific ones. To our knowledge, there exist no methods capable of decomposing the covariate effects into shared and view-specific effects in a multi-way scenario.

We now fit the model into the extended multi-way modelling framework, depicted in Figure 2 (b). The integration of different domains takes place in the low-dimensional latent factor spaces $\mathbf{x}^{lat}$ and $\mathbf{y}^{lat}$. These factor spaces can be integrated by combining the factor analyzers into a generative model of Bayesian CCA [1,7]. This introduces a new hierarchy level where a latent variable $\mathbf{z}$ captures the shared variation between the views.

The generative model of BCCA has been formulated [1,7] for sample j as

$$\begin{aligned} \mathbf{z}_j &\sim \mathcal{N}(0, \mathbf{I}), \\ \mathbf{x}_j^{lat} &\sim \mathcal{N}(\mathbf{W}^x \mathbf{z}_j, \boldsymbol{\Psi}^x) \;, \end{aligned} \tag{4}$$

and likewise for $\mathbf{y}$. Note that 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 $\boldsymbol{\Psi}^x$ is a matrix of marginal variances modelling the source-specific effects not responding to external covariates. The prior distributions were chosen as in [7]; $\mathbf{W}^x$ has an Automatic Relevance Determination (ARD) prior [2]; $\boldsymbol{\Psi}^x$ has an inverse Wishart prior.

Decomposition into shared and view-specific effects. The decomposition into shared and view-specific effects is done by adding view-specific latent variables in addition to the shared ones, and the latent effects acting as population-specific priors on shared and specific latent variables identify the effects. The Bayesian CCA assumes that the data is generated by a sum of view-specific $\mathbf{z}^x$ and $\mathbf{z}^y$, and shared latent variables $\mathbf{z}$, as shown in Figure 3. In practice, the decomposition in Figure 3 can be implemented easily by restricting a column of $\mathbf{W}^x$ to be zero for the y-specific components and vice versa for x. As a summary the complete generative model is

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

$$\mathbf{x}_j \sim \mathcal{N}(\boldsymbol{\mu}^x + \mathbf{V}^x \mathbf{x}_j^{lat}, \boldsymbol{\Lambda}^x). \tag{5}$$

2.3 Time with Unknown Alignment

We concentrate here on the case of a small number ($\sim$ 10) of replicate time-series from multiple populations, in the unfavorable conditions of short ($\sim$ 10-20) time-series and high dimensionality.

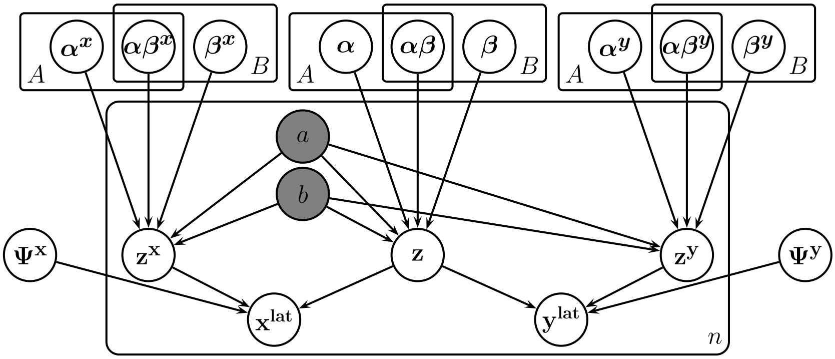

Fig. 3. The graphical model describing the decomposition of covariate effects into shared and view-specific ones. The figure expands the top part of Figure 2 (b).

We consider "time with unknown alignment" as one covariate in an extended multi-way model; a particular case is HMM-alignment. The extended multi-way model with HMM-time as a covariate, is shown in Figure 2 (c). We assume that the time operates on the latent variables as the other covariates, with the unknown alignment modelled by HMM. This can be accomplished by having the HMM emit values for the latent factors. In addition, there is another covariate effect $\boldsymbol{\beta}_b$. The model becomes

$$\mathbf{x}_j^{lat}|_{state(j,t)=s,b} \sim \mathcal{N}(\boldsymbol{\alpha}_s + \boldsymbol{\beta}_b + (\boldsymbol{\alpha\beta})_{sb}, \mathbf{I}). \tag{6}$$

Here $\boldsymbol{\alpha}_s$ is the effect of HMM-time in the multi-way model, that is the mean in the Gaussian emission distribution of HMM-state s. The $\boldsymbol{\beta}_b$ is the effect of the other, observed, covariate b, and $(\boldsymbol{\alpha\beta})_{sb}$ is the interaction effect. Here $state(j,t) = s$ means that time-point t of sample j belongs to state s.

Assignments to HMM states are sampled according to a standard Bayesian HMM formalism, the prior for the transition matrix of the linear HMM allowing only self-transitions and transitions to the next state.

In biological case studies where time-series measurements are taken from multiple populations, e.g. healthy and diseased, there is a need for HMM alignment when intervals between measurements are long and irregular within- and between patients. In addition, patients are assumed to develop to different biological states at individual times/ages. Previous works [3,11] have resorted to training a separate HMM for each population and comparing them afterwards, additionally restricting to strong feature selection, only allowing favorable $n > p$-conditions.

In our experiments, we will have 5 states, $b = 0, ..., 4$ of β_b-effects for the diseased, corresponding to the observed disease-development states in the time-series. For simplicity, we do not consider the interaction effect, restricting to $\mathbf{x}_j^{lat}|_{state(j,t)=s,b} \sim \mathcal{N}(\boldsymbol{\alpha}_s + \boldsymbol{\beta}_b, \mathrm{I})$.As a summary, "time with unknown alignment", such as "HMM-time", can be seen as one covariate in an extended multi-way model, where the covariate assignments (alignment to HMM-states), are inferred from the observed data. A main benefit of building a unified model is that after explaining away the effect of "aligned time", $\boldsymbol{\alpha}_s$, one can answer the following statistical question: is there a difference in the populations, that is; is $\boldsymbol{\beta}_b$ statistically significant for some b? Earlier HMM approaches training a separate model for each population cannot fully rigorously answer this question.

2.4 Multi-view Learning without Paired Samples

Finally, we consider integrating data sources in different domains, without paired samples, which is a much more difficult problem. In a similar case in [13], the underlying assumption was an unobserved pairing between the samples, and the pairing was found by iteratively alternating between searching for pairing and maximizing dependencies between the sources by CCA. However, the assumption of latent unknown pairing might be too restrictive in many cases, and the non-generative solution in [13] cannot easily be extended to the present tasks.

We propose an alternative assumption, allowing to integrate multiple unmatched data sources under the assumption of shared, underlying multi-way covariate-related behavior. For brevity, we concentrate in this article on one standard covariate, say, time with unknown alignment $\boldsymbol{\alpha}$, and the "view without paired samples" is the other covariate. Again, since "view without paired samples" is a **view-covariate**, it cannot be defined at all as a main effect due to data domains being different. However, we can define an interaction effect of time and "view without paired samples". The model becomes

$$\begin{aligned} \mathbf{x} &= \boldsymbol{\mu}^x + f^x(\boldsymbol{\alpha}_a) + f^x(\boldsymbol{\alpha}_a^x) + \epsilon, \\ \mathbf{y} &= \boldsymbol{\mu}^y + f^y(\boldsymbol{\alpha}_a) + f^y(\boldsymbol{\alpha}_a^y) + \epsilon \ , \end{aligned} \tag{7}$$

where $\boldsymbol{\alpha}_a$ is the shared effect of time, and $\boldsymbol{\alpha}_a^x$ and $\boldsymbol{\alpha}_a^y$ are the view-specific time-effects, and f^x and f^y are again functional mappings from latent states to actual data spaces. The possibility to decompose the time-related behavior in the two datasets into shared and view-specific covariate-related behaviors connects the views.

To make the translational problem (presented below) more realistic, we consider the time-covariate to be "HMM-time". This allows us to study more flexible translational cases with time-measurement having irregular intervals, and time-spans being different [9], important in cross-species biological applications.

The model can be formulated as a graphical multi-way model with the techniques presented in the previous sections. For simplicity, the view-specific time-behavior is integrated out with $\boldsymbol{\Psi}_x$ and $\boldsymbol{\Psi}_y$ following [7], and we only search for the shared effects in the simulations. The graphical model of the problem is shown in Figure 2 (d). The key difference to the multi-view learning with paired samples case in Figure 2 (b) is that since there is no known pairing of samples, there is no latent variable $\mathbf{z}_j$ shared by the samples from different views.

The learning algorithm faces a matching problem since a shared time behavior might be identified in, e.g., cluster 1 of $\mathbf{x}$ and cluster 3 of $\mathbf{y}$. For the model to identify the effect as a shared effect, it should be found for the same cluster identity. We include a Metropolis-Hastings step in our Gibbs sampler that proposes to switch identities of two clusters, attempting to maximize similar time-related behavior.

Related work in translational studies. A main application is translating biological findings between experiments on model organisms and actual human experiments [9,10]. The common setup is doing a similar experiment (time-series, same disease) to the two different organisms and comparing the results. High-dimensional biological measurements usually have different unmatched domains in different species. Most multi-species approaches [9] are restricted to the subset of variables that are a priori matched between the species. Since this assumption is restrictive, we have wanted to consider the more general case where the domains are different, making it possible to use all the data, and while doing so to actually search for the matching of variables.

3 Results

3.1 Multi-view Learning with Paired Samples

Generated data. We integrate three data-sources, $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{u}$, with pairing between the samples, which have a two-way experimental setup, generated from the model of Figure 2 (b). The datasets are 200-dimensional, there are three clusters of variables in each dataset. The $\sigma_i = 1$ for each variable. The model is learned by Gibbs sampling, with 2000 samples and 2000 burn-in samples. The optimal number of clusters is found for each data source separately as explained in [6], and always correctly recovered. Unless otherwise stated, these parameters are the same throughout the results section. Effects $\alpha = +2$, $\beta^y = +2$ and $(\alpha\beta)^x = +2$ have been generated. We learned 4 components: one shared and three source-specific. Shared and source-specific α, β and $(\alpha\beta)$ are therefore to be estimated. The model always finds the correct clusterings (data not shown). The results in Figure 4 show how the model finds the generated effects as a function of number of samples. According to the results, the model finds the generated effects with relatively small sample-sizes, and the uncertainty decreases with increasing sample-size. The shared effect is found with considerably less uncertainty since there is evidence from both sources. In a typical biological dataset there may be 20-60 samples.

Lipidomic multi-tissue data. We now apply the method on an unpublished lipidomic lung cancer study, where lipidomic measurements have been taken from several tissues of mice. There are cancerous and healthy mice, and additionally half of both populations have been given a test anti-cancer drug. This is a typical two-way setup with healthy untreated (10 mice), diseased untreated (10), healthy treated (9), diseased treated (10) mice. The tissues have different lipids. We first integrated the lung tissue (68 lipids) with spleen tissue (44 lipids). We learned 3 components, one shared and one for each view. According to the results in Figure 5 (left), the model finds a shared disease effect α and a shared treatment effect β. The result shows that the treatment enhances, not

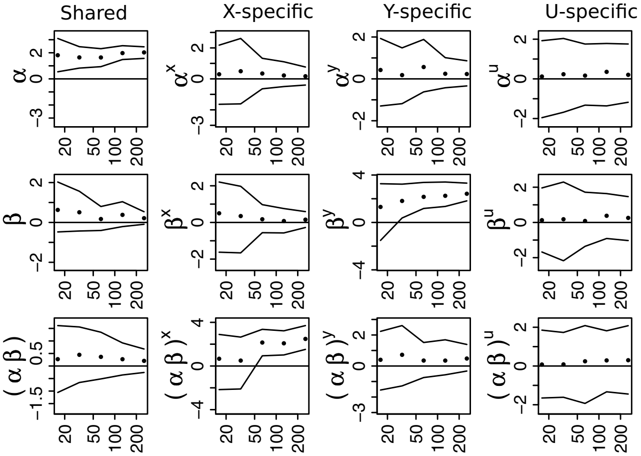

Fig. 4. The method finds the generated effects $\alpha = +2$, $\beta^y = +2$ and $(\alpha\beta)^x = +2$ in a three-view, two-way study. The points show posterior means, and lines the 95% posterior mass of the effects. The posterior distributions have been mirrored to have a positive mean. A consistently non-zero posterior of the effects indicates a statistically significant effect. This corresponds to a classical p-value being $p < 0.05$.

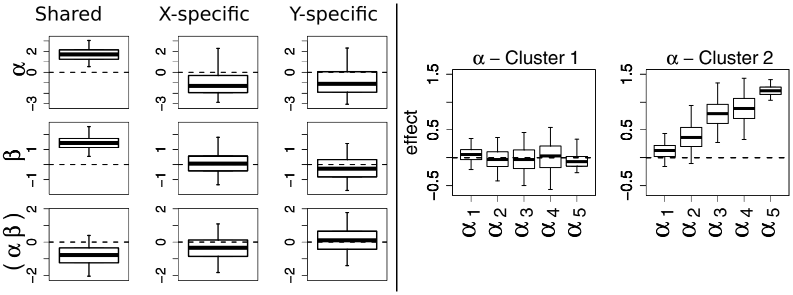

Fig. 5. In the experiment on multi-tissue data (left), the method finds a disease effect α and a treatment effect β shared between the two views, spleen ($\mathbf{x}$) and lung ($\mathbf{y}$) tissues. (right) The shared HMM finds shared effects in two generated datasets (Section 3.3) in different domain with no paired samples. A growing HMM-time effect was generated in cluster 2. A consistently non-zero posterior implies an effect found.

diminishes the effect of the disease, therefore not being effective. In lung, for instance, a cluster of 12 lipids containing ether lipids known to be co-regulated, was coherently up-regulated due to disease, and additionally up-regulated by the treatment. Another cluster of 13 lipids in lung was found down-regulated due to the disease and additionally down-regulated due to treatment. The lipids of the down-regulated cluster are thus negatively correlated with the up-regulated clusters. The effect can be traced back to the clusters of lipids by identifying the responsible elements in $\mathbf{W}^x$, and to the actual lipids from $\mathbf{V}^x$.

No existing ANOVA-type methods are capable of decomposing covariate effects into shared and source-specific effects, when sources have different domains. The possible comparison methods are 1) separate MANOVA-analysis for each source including a dimension reduction, 2) concatenation of the sources and MANOVA-analysis. These methods give only an overall p-value for the statistical significance of the effects. We compare the biological result to concatenating the sources and using 50-50 MANOVA [8], which includes a prior PCA-dimension reduction. The method gives p-values 0.01, 0.71 and 0.071, for α, β and $(\alpha\beta)$, respectively. The method only finds a statistically significant disease effect, not finding the effect of treatment, showing the superior behavior of an integrated dimension reduction in our model. The main difference is, however, that the method cannot distinguish whether the effect is shared or source-specific.

3.2 Time with Unknown Alignment

Generated data. We show results on data generated from the model in Figure 2 (c). There are 5 HMM-states α, 23 replicate time-series from healthy and 21 from diseased population for which there are 3 disease states β. Each time-series has a length of 5-15 time-points at random times (no matching of time-points), dimensionality is $p = 400$. Disease state-type covariates $b_{jt} = \{1, 2, 3\}$ are observed for the diseased patients, healthy patients only have HMM-states. Effects $\boldsymbol{\alpha} = 0, +0.5, +1, +1.5, +2$ have been generated in the consecutive HMM-states in the first cluster of α. In the disease states of β, effects $\boldsymbol{\beta} = -0.5, -1, -2$ have been generated in the consecutive disease states, equally in the first cluster. In the other clusters, there are no covariate-related effects, only structured noise from the model. The model is able to identify the clusters correctly. The results in Figure 6 (left) show that a proper HMM-alignment is achieved, and the model found the generated, growing time-behavior $\boldsymbol{\alpha}$ in cluster 1 and especially is able to separate the correct descending disease-state behavior $\boldsymbol{\beta}$ related to the known covariates.

Lipidomic time-series data. We then applied the model to a recently published lipidomic dataset [12]. There are 71 healthy patients and 53 patients that later developed into type 1 diabetes, there are 3-29 time-points in each time-series, measured at irregular intervals. In addition, for the patients that later developed into type 1 diabetes, the progress of the disease (disease state) is observed at each time point and used here as a covariate b, with 5 disease states. There are 53 lipids. We show results on 2 clusters in Figure 6 (right). The model was capable of identifying the normal aging effects for clusters of similarly behaving lipids (HMM-time effects), and it was able to separate disease state-related effects for each cluster. The model found consistent clusters of lipids known to be co-regulated. In Figure 3 of [12], the data analysis was done by univariate t-tests for each lipid, and time was separated to bins of length 1 year. Our multivariate modelling was able to take into account that different individuals enter age-related metabolic states at largely varying, individual times. In addition, the model could separate the disease-state related behavior from the normal aging effects, all done for similarly behaving groups of lipids. Our results were consistent

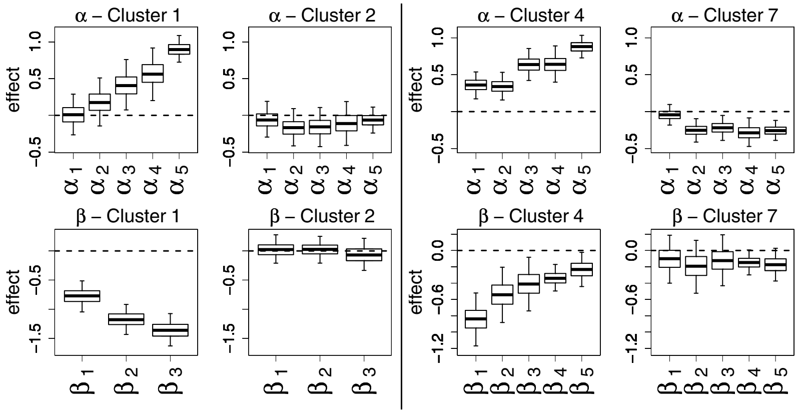

Fig. 6. (left) The HMM-model finds the effect of time with unknown alignment α_s in cluster 1 from two populations of generated data, and is able to separate a disease-progression type effets β generated in the other population. (right) The HMM-model separates normal aging α_s for clusters of similarly behaving lipids, from effects related to known disease progression-states in a real lipidomics type 1 diabetes study. A consistently non-zero posterior shown by box-plots implies an effect found.

with those in the paper. In addition, our model suggested that PC(14:0/18:2) and PC(18:2/16:1) in cluster 4 have a strong down-regulation in the early disease development states, and might act as early biomarkers of a developing disease. This was not revealed by the prior analysis. Existing methods are limited to training a separate HMM for each population, which is an appropriate approach for classification, but cannot be used to rigorously compare the effects of disease states in the data under the assumption that normal aging effects have to be modelled (away) by a HMM-alignment.

3.3 Multi-view Learning without Paired Samples

We now show results of multi-way learning when the domains of multiple data sources are different and samples have no pairing. We consider a non-trivial case where we have two generated time-series datasets, with irregular lengths and measurement times where we assume, however, that behavior of HMM-states is similar. We can now search for similar HMM-behavior in two unpaired datasets in different domains. We can make the assumption that in addition to view-specific effects, there is a shared HMM-chain that emits latent variables $\mathbf{x}^{lat}$ and $\mathbf{y}^{lat}$, which in turn generate the actual data to different domains.

We have generated two data sets from the model with 10 and 11 replicate time-series of irregular lengths between 8-12, and datasets have 100 and 110 variables, respectively. There are 3 clusters in each, and the corresponding $\mathbf{x}^{lat}$ and $\mathbf{y}^{lat}$ have been generated according to the shared HMM-chain where the effects 0,+0.5,+1,+1.5,+2 have been generated to the second factor (cluster) of $\boldsymbol{\alpha}$ in the five HMM-states of the shared HMM-chain, an x-specific time-effect in cluster 1, the third cluster (not shown) does not have effects. In this case study, the specific time effect is integrated out with covariance matrices $\mathbf{\Psi}_x$ and $\mathbf{\Psi}_y$. According to the results in Figure 5 (right), the model was able to find the shared HMM-time-related behavior from different domains without paired samples, while x-specific effects were integrated out successfully.

The results of the simple case study show that in the case of underlying shared HMM-states, connection between two views, even without paired samples, can be formed by formulating the analysis as an ANOVA-type model over views. This makes it possible to rigorously evaluate statistical significance of a similar covariate-related behavior. To our knowledge, such a possibility has not been proposed in any previous studies. To our knowledge, there exists no comparable method to this modelling task, except training a separate HMM for each view, which allows only qualitative comparison of the results.

4 Conclusions

We have extended multi-way learning to three novel cases: (i) multi-view learning with paired samples, where data comes from different domains, (ii) one of the covariates has an unknown structure (iii) data comes from different domains with no pairing of samples, but covariates are shared. In (i) we have shown how covariate-related behavior can be decomposed into shared and view-specific effects, when integrating data sources with paired samples. In (ii) we have presented a multi-way model where one of the covariates has an unknown structure which can be learned jointly. In (iii) we have shown that it is possible to integrate multiple data sources without paired samples, if the datasets have a similar covariate-structure. We have shown that unified hierarchical graphical models can be used to structure each case as a graphical multi-way model.

Each of the presented multi-way models has direct applications for biological experiments, but they also offer novel possibilities for other application domains such as brain signal analysis (multi-way time-series in fMRI), detection problems in sensor fusion and cold-start problem in content-based retrieval given second content descriptors etc. We showed that the models are capable of finding ANOVA-type effects from real and simulated high-dimensional data, even with small sample-sizes. The biological results were plausible, comparing to previous studies.

References

1. Bach, F.R., Jordan, M.I.: A probabilistic interpretation of canonical correlation analysis. Tech. Rep. 688, Department of Statistics, University of California, Berkeley (2005)
2. Bishop, C.M.: Bayesian PCA. In: Kearns, M.S., Solla, S., Cohn, D. (eds.) Advances in Neural Information Processing Systems, vol. 11, pp. 382–388. MIT Press, Cambridge (1999)

3. Costa, I.G., Schonhuth, A., Hafemeister, C., Schliep, A.: Constrained mixture estimation for analysis and robust classification of clinical time series. Bioinformatics 25(12), i6–i14 (2009)
4. Fisher, R.: The correlation between relatives on the supposition of mendelian inheritance. Royal Society of Edinburgh from Transactions of the Society 52, 399–433 (1918)
5. Huopaniemi, I., Suvitaival, T., Nikkilä, J., Orešič, M., Kaski, S.: Multivariate multiway analysis of multi-source data. Bioinformatics 26, i391–i398 (2010)
6. Huopaniemi, I., Suvitaival, T., Nikkilä, J., Orešič, M., Kaski, S.: Two-way analysis of high-dimensional collinear data. Data Mining and Knowledge Discovery 19(2), 261–276 (2009)
7. Klami, A., Kaski, S.: Local dependent components. In: Ghahramani, Z. (ed.) Proceedings of ICML 2007, the 24th International Conference on Machine Learning, pp. 425–432. Omni Press (2007)
8. Langsrud, O.: 50-50 multivariate analysis of variance for collinear responses. Journal of the Royal Statistical Society Series D-the Statistician 51, 305–317 (2002)
9. Lu, Y., Huggins, P., Bar-Joseph, Z.: Cross species analysis of microarray expression data. Bioinformatics 25(12), 1476–1483 (2009)
10. Lucas, J., Carvalho, C., West, M.: A bayesian analysis strategy for cross-study translation of gene expression biomarkers. Statistical Applications in Genetics and Molecular Biology 8(1), 11 (2009)
11. Nikkilä, J., Sysi-Aho, M., Ermolov, A., Seppänen-Laakso, T., Simell, O., Kaski, S., Orešič, M.: Gender dependent progression of systemic metabolic states in early childhood. Molecular Systems Biology 4, 197 (2008)
12. Orešič, M., et al.: Dysregulation of lipid and amino acid metabolism precedes islet autoimmunity in children who later progress to type 1 diabetes. Journal of Experimental Medicine 205(13), 2975–2984 (2008)
13. Tripathi, A., Klami, A., Kaski, S.: Using dependencies to pair samples for multiview learning. In: Proceedings of ICASSP 2009, the International Conference on Acoustics, Speech, and Signal Processing, pp. 1561–1564 (2009)
14. West, M.: Bayesian factor regression models in the large p, small n paradigm. Bayesian Statistics 7, 723–732 (2003)