Cross-Species Translation of Multi-Way Biomarkers

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

[1] Aalto University School of Science
Department of Information and Computer Science,
Helsinki Institute for Information Technology HIIT
firstname.lastname@tkk.fi
http://research.ics.tkk.fi/mi/

[2] VTT Technical Research Centre of Finland
firstname.lastname@vtt.fi
http://sysbio.vtt.fi/

[3] University of Helsinki
Department of Computer Science,
Helsinki Institute for Information Technology HIIT

Abstract. We present a Bayesian translational model for matching patterns in data sets which have neither co-occurring samples nor variables, but only a similar experiment design dividing the samples into two or more categories. The model estimates covariate effects related to this design and separates the factors that are shared across the data sets from those specific to one data set. The model is designed to find similarities in medical studies, where there is great need for methods for linking laboratory experiments with model organisms to studies of human diseases and new treatments.

Keywords: Bayesian inference, cross-species modeling, multi-way modeling, translational modeling

1 Introduction

We study the translational modeling problem, where the aim is to integrate data sets which have neither co-occurring samples nor variables. The only known commonality between the sets is that they have been collected from experiments with a similar design.

Translational modeling has an increasingly important application in cross-species analysis of biological experiments, where treatments to human diseases are studied using model organisms. In cross-species analysis, the question is how to integrate data sets with high dimensionality, small sample-size, and potentially structured covariates, as illustrated in Figure 1a.

The basic experimental design in the search for disease biomarkers is one-way comparison of healthy and diseased patient groups. At the simplest, biomarkers can be translated across species by comparing lists of p-values of differential

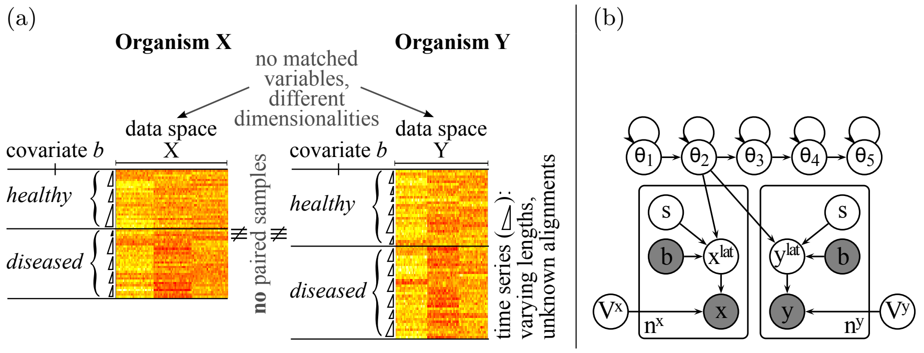

Fig. 1. (a) Data matrix representation of the translational problem. (b) Plate diagram of the proposed Bayesian graphical model. The sets $\boldsymbol{\theta}_s = \{\boldsymbol{\alpha}_s^{\mathrm{sh}}, \boldsymbol{\alpha}_s^{\mathrm{x}}, \boldsymbol{\alpha}_s^{\mathrm{y}}, (\boldsymbol{\alpha\beta})_{s,b}^{\mathrm{sh}}, (\boldsymbol{\alpha\beta})_{s,b}^{\mathrm{x}}, (\boldsymbol{\alpha\beta})_{s,b}^{\mathrm{y}}\}$ contain all latent variables describing the corresponding HMM states. The state (category) of each sample j is determined by an observed covariate b_j and an unobserved covariate s_j.

expression from a t-test. Most existing cross-species analysis tools are limited to these simple designs [6].

Most biological experiments have, however, a multi-way experiment design, where healthy and diseased groups are further divided into subgroups according to additional covariates, such as treatment, gender, age, measurement time, etc. The basic standard statistical methods capable of properly dealing with the multi-way design are analysis of variance (ANOVA) and its multivariate generalization (MANOVA) [8].

Taking all the covariates into account complicates the analysis only slightly, but also allows us to extract considerably more information from the data. There are no earlier tools for utilizing multiple covariates and estimating their effect across data sets with neither co-occurring samples nor variables.

Time series experiments are becoming more and more common in clinical studies searching for disease biomarkers. In our multi-way design, time is one of the covariates, having a special structure. In a clinical follow-up study, such as the Type 1 Diabetes prediction and prevention study [9], measurement times are irregular due to practical reasons of data collection, and there are missing time points. In addition, life spans of organisms, such as human and mouse, are very different, resulting in very different measurement intervals. These complications cause challenges for cross-species data analysis, and call for a possibility to align the time series using machine learning techniques.

In this paper, we show how it is possible to integrate data sets with neither co-occurring samples nor variables, only based on a similar experiment design. We separate and identify shared covariate effects from data set-specific effects. We do this by building on our recent work on high-dimensional multi-way modeling and time series alignment [4]. We test the method on simulated data, and on lipidomic and metabolomic data sets.

2 Previous Work

A few iterative approaches have matched samples without taking covariate information into account. One of the methods matched only samples [10], and another matched both samples and variables [1].

For cross-species analysis, there are methods that use and require side information about the possible matchings of the variables between the data sets. Le & Bar-Joseph [5] utilized sequence similarities as a prior for clustering and matching genes across data sets of two species. Lucas *et al.* [7] inferred a set of factors that are active in one data set and used that as a starting point for the inference in the other data set, requiring at least a subset of variables to be the same across data sets. The model that we present next, does not require any prior match across neither samples nor variables.

3 Model

We address the problem of translating covariate effects across two data sets which have neither co-occurring samples nor variables. We develop a method that handles traditional multi-way experimental designs, where samples have been divided, for instance, into healthy-diseased and treated-untreated categories, or more categories with possibly more levels. In addition, the model extends to time series designs, where one covariate, the time point, is not necessarily matched across the two data sets. Irregular time points are handled by aligning the time series into latent states, which are then matchable across the data sets.

In our previous work [4], we were only able to estimate the covariate effects shared by the data sets. In this paper we present a novel matching algorithm for separating shared covariate effects from effects specific to one data set.

3.1 Dimensionality Reduction and Covariate Effects

We construct a unified multivariate model, where the inference is carried out with Gibbs sampling. It is a single hierarchical Bayesian model capable of handling uncertainty across the levels, in contrast to a straightforward successive dimensionality reduction and MANOVA. In terms of estimation of multi-way covariate effects and dimensionality reduction, the new approach builds on our earlier work on high-dimensional multi-way modeling [3]: we assume that a single latent factor vector $\mathbf{x}^{\mathrm{lat}}$ generates a group of correlated variables in the observed data $\mathbf{x}$, and the latent factors have a covariate-dependent prior structure for each sample. These factors can thus be called clusters (of variables).

The model for sample j explained by K latent factors is

$$\begin{aligned} \mathbf{x}_j &\sim \mathcal{N}\left(\boldsymbol{\mu} + \mathbf{V}\mathbf{x}_j^{\mathrm{lat}}, \boldsymbol{\Lambda}\right) \\ \mathbf{x}_j^{\mathrm{lat}} \,|\, (a_j, b_j) &\sim \mathcal{N}\left(\boldsymbol{\alpha}_{a_j} + \boldsymbol{\beta}_{b_j} + (\boldsymbol{\alpha\beta})_{a_j,b_j}, \mathbf{I}\right) \ , \end{aligned} \tag{1}$$

where $\mathbf{x}_j$ is a p-dimensional data sample from the $n \times p$ data matrix, $\boldsymbol{\mu}$ is a p-vector of variable means, $\mathbf{V}$ is a $p \times K$ projection matrix, $\mathbf{x}_j^{\mathrm{lat}}$ the K-vector of latent factors from the $K \times n$ latent space matrix, and $\boldsymbol{\Lambda}$ is a diagonal residual variance matrix with diagonal elements σ_i^2. Covariate effects are estimated in the K-dimensional latent space, and in Equation 1 the prior is presented for the two-way case with particular covariate values a_j and b_j selecting the main effects $\boldsymbol{\alpha}_{a_j}$ and $\boldsymbol{\beta}_{b_j}$, and an interaction effect $(\boldsymbol{\alpha\beta})_{a_j,b_j}$. In the notation, covariates a_j and b_j independently select a corresponding row from the main effect matrices $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, respectively, and $(\boldsymbol{\alpha\beta})_{a_j,b_j}$ is an interaction effect vector of the combination a_j, b_j.

3.2 Alignment of Irregular Time Series

When one of the "ways" is irregularly sampled time, underlying states in the time series are inferred in the model by a hidden Markov model (HMM)-type state projection. The learned state allocations $\mathbf{s}$ are used as a covariate and the corresponding HMM latent variable is interpreted as the covariate effect for the sample group [4].

Now, $\mathbf{x}_j^{\mathrm{lat}}$ is assumed to be generated by using the learned covariate s_j instead of a fixed covariate a_j:

$$\mathbf{x}_j^{\mathrm{lat}} | (s_j, b_j) \sim \mathcal{N}\left(\boldsymbol{\alpha}_{s_j} + \boldsymbol{\beta}_{b_j} + (\boldsymbol{\alpha\beta})_{s_j,b_j}, \mathbf{I}\right) , \tag{2}$$

where $\boldsymbol{\alpha}_{s_j}$ is the HMM-aligned time effect. We restrict the HMM to a linear chain structure, which is reasonable for the biological patient progression data of our experiment.

3.3 Estimation of Shared and Specific Covariate Effects

Now we have presented the model for dimensionality reduction and estimation of covariate effects in the case of a single data set. Next, we will show how this framework can be extended to the analysis of multiple data sets, and how to identify latent factors that have a match across the data sets. We not only estimate covariate effects of a single data set, but also probabilities of each latent factor being generated either by data set-specific covariate effects or by effects shared with a factor from the other data set. A plate diagram of the model is shown in Figure 1b.

The model makes a flexible assumption [4] that the observed data vectors in the two data sets X and Y are generated by the covariate effects through a transformation f^{x} and f^{y}, respectively:

$$\begin{aligned}\mathbf{x}_j | (s_j, b_j) &= \boldsymbol{\mu}^{\mathrm{x}} + f^{\mathrm{x}}\left(\boldsymbol{\alpha}_{s_j}^{\mathrm{sh}} + \boldsymbol{\beta}_{b_j}^{\mathrm{sh}} + (\boldsymbol{\alpha\beta})_{s_j,b_j}^{\mathrm{sh}}\right) + f^{\mathrm{x}}\left(\boldsymbol{\alpha}_{s_j}^{\mathrm{x}} + \boldsymbol{\beta}_{b_j}^{\mathrm{x}} + (\boldsymbol{\alpha\beta})_{s_j,b_j}^{\mathrm{x}}\right) + \boldsymbol{\varepsilon}^{\mathrm{x}} \\ \mathbf{y}_i | (s_i, b_i) &= \boldsymbol{\mu}^{\mathrm{y}} + f^{\mathrm{y}}\left(\boldsymbol{\alpha}_{s_i}^{\mathrm{sh}} + \boldsymbol{\beta}_{b_i}^{\mathrm{sh}} + (\boldsymbol{\alpha\beta})_{s_i,b_i}^{\mathrm{sh}}\right) + f^{\mathrm{y}}\left(\boldsymbol{\alpha}_{s_i}^{\mathrm{y}} + \boldsymbol{\beta}_{b_i}^{\mathrm{y}} + (\boldsymbol{\alpha\beta})_{s_i,b_i}^{\mathrm{y}}\right) + \boldsymbol{\varepsilon}^{\mathrm{y}},\end{aligned} \tag{3}$$

where symbols with superscript $^{\mathrm{sh}}$ represent covariate effects shared by the two data sets, and symbols with superscripts $^{\mathrm{x}}$ and $^{\mathrm{y}}$ represent data set X and Y-specific covariate effects, respectively. The variable spaces of data sets X and Y are different, and therefore also the latent factor spaces $\mathbf{x}^{\mathrm{lat}}$ and $\mathbf{y}^{\mathrm{lat}}$ representing groups of correlated variables need not match. For this reason, the covariate effects have to be projected into the actual observed data spaces $\mathbf{x}$ and $\mathbf{y}$ through the previously unknown projections f^{x} and f^{y}, which will be learned jointly.

Earlier, we have learned covariate effects from multiple data sets, where samples co-occur across the sets (views) [2]. The translational problem is now more complicated, and we have to solve it in a different way.

The modeling question for two non-co-occurring data sets with a multi-way experiment design becomes the following: Does some dimension of $\mathbf{x}^{\mathrm{lat}}$ respond to the covariates $\mathbf{s}$ and $\mathbf{b}$ similarly as one of $\mathbf{y}^{\mathrm{lat}}$? If it does, one can represent this pattern with *shared* covariate effects $\boldsymbol{\theta}^{\mathrm{sh}} = \{\boldsymbol{\alpha}^{\mathrm{sh}}, \boldsymbol{\beta}^{\mathrm{sh}}, (\boldsymbol{\alpha\beta})^{\mathrm{sh}}\}$. The interpretation is that a group of correlated variables in data set X matches with a group in data set Y, represented by a dimension of $\mathbf{x}^{\mathrm{lat}}$ and $\mathbf{y}^{\mathrm{lat}}$, respectively. In biology, such factors can be considered as multi-species biomarkers. If there is no match, the response to the covariates is modeled by species-specific covariate effects $\boldsymbol{\theta}^{\mathrm{x}} = \{\boldsymbol{\alpha}^{\mathrm{x}}, \boldsymbol{\beta}^{\mathrm{x}}, (\boldsymbol{\alpha\beta})^{\mathrm{x}}\}$, and similarly for Y. Our modeling framework estimates the confidence of the shared effects.

3.4 Matching

We propose the following measure for quantifying the quality of the match between two factors from different data sets: whether the matching is better than an average matching (over other pairs). On a meta-level the measure is intuitively appealing in the spirit of permutation tests, and it can be formulated more exactly by specifying what we mean by "better." We will use probabilistic modeling to measure the relative goodness below.

The matching problem of the clusters is a combinatorial problem, where possible configurations of pairs need to be evaluated, judging for each pair how similarly they respond to multi-way covariates. We resort to an iterative algorithm that attempts to change the matching of one cluster at a time.

After selecting a candidate pair, we compare its goodness to an average pair (uniformly selected having one same endpoint), and accept forming a link between them by a Metropolis criterion that compares the likelihoods of the two pairings. A reverse operation is to attempt to break a link by comparing an existing link between two clusters to an average (random) pair. The goodness (likelihood) of the linked pair is evaluated by comparing likelihoods of the two shared covariate effect structures. Factors with no pairs are modeled by data set-specific covariate effects. Averaging over sampling iterations, we can estimate the probability for matchings and the patterns of the covariate effects. High probability of a particular pair indicates a found matching. Low probability of any pair indicates that there might not be suitable match for the factor in the other data set.

4 Experiments

In this section, we demonstrate how the model works on high-dimensional toy data, and on biological data from human blood samples.

4.1 Generated Data

We generated from the model two data sets X and Y with no pairing of samples but only a shared two-way covariate structure. There are 11 separate time series ("patients") in both of the two data matrices, each series consisting of 5 to 15 time points. This results in 100 and 112 samples in total, and data matrices are 200- and 210-dimensional. The latent factors $\mathbf{x}_j^{\text{lat}}$ and $\mathbf{y}_j^{\text{lat}}$ are 3- and 4-dimensional, respectively. Two latent factors in each data set were generated from a shared HMM chain with five states.

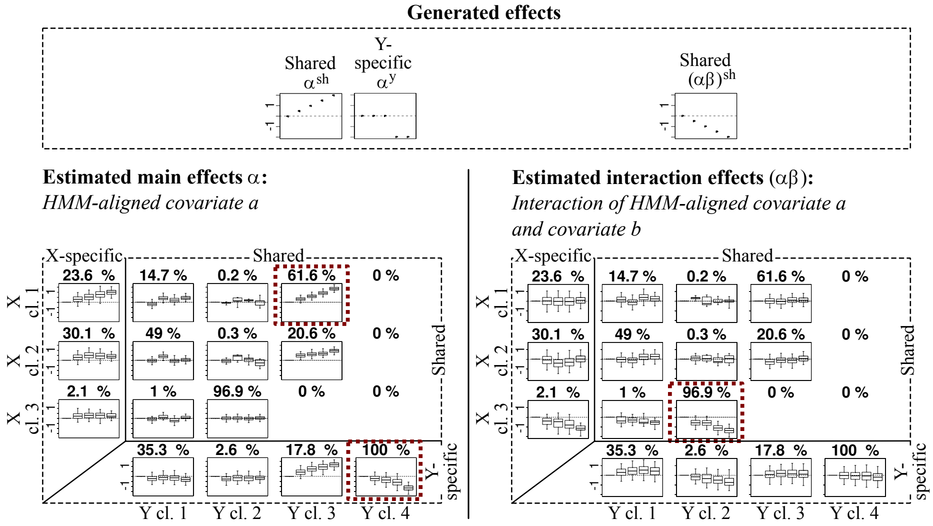

Fig. 2. Matching results from generated time-series data. Shown are the main effects of the HMM-aligned covariate a ($\boldsymbol{\alpha}$; left), and interaction effects of covariates a and b (($\boldsymbol{\alpha\beta}$); right). Topmost, the generated effects are illustrated. In both the lower parts, the table of estimated covariate effects shows shared (top-right area) and data set-specific (left column and bottom row) effects for both $\boldsymbol{\alpha}$ and ($\boldsymbol{\alpha\beta}$). Rows and columns in the area of shared effects correspond to clusters in data sets X and Y, respectively. The found true pairing is highlighted by a red box. The value on top of each plot shows the percentage of posterior samples, where the matching was found. The boxplots within each subplot represent posterior distributions of effects at different levels of the covariate. A distribution above or below zero with 95 % confidence is considered significant.

We used the proposed model to simultaneously align the samples into matchable HMM states, learn the clusters of variables, search for the possible matches of the clusters between the two data sets, and model the ANOVA-type covariate effects acting on the found clusters. We *a priori* chose a model with five HMM states. During sampling, 150,000 burn-in samples and 150,000 posterior Gibbs samples were collected, and every 50th sample was collected. The generated effects and the results are shown in Figure 2. Our model found the previously generated clusters without mistakes and matched clusters across the datasets correctly.

4.2 Biological Data

We analysed biological data from a follow-up study of type 1 diabetes, where 53 lipid and 74 metabolite concentrations from blood samples were measured from two sets of human patients, respectively [9]. In total, we had 1153 and 417 samples from 124 and 37 patients, respectively.

We separated the normal development of young individuals from progression of the disease by labeling samples of patients, who acquired the disease, into four stages of progression of the disease using additional information of the antibody levels in blood. These stages were fixed as the levels of covariate b_j, while the temporal alignment a_j of all patients was learned within the model by the HMM. We used a five-state HMM, and 6- and 15-dimensional latent variables to explain the correlated groups of lipids and metabolites in the data, respectively.

Comparison of Matchings of Lipids to the Ground Truth. First, we tested how the model finds matching, when the variables are actually co-occurring across the data sets. We split the lipidomic data set into two groups of patients and used the groups as data sets X and Y.

As a result, we found out that the three strongest matches out of the six were correct.

Integration of Lipidomic and Metabolomic Data Sets. Next, we searched for matching groups between the lipidomic and metabolomic data sets. Some of the patients were the same in the two data sets, but we did not utilize this information to help the model.

The main result was that the best match was a group of three glycerophosphocholine (GPCho) lipids to a group of four metabolites with probability of 19.7 % (see Table 1). Three first of the metabolites in the list are fatty acids, which are building blocks for GPCho lipids. The found lipid and metabolite groups had a similar covariate effect pattern in time (up-regulation) and in the stages of the disease (down-regulation).

5 Conclusion

We presented a novel method for translating biomarkers between multiple species from multi-way, time series experiments, which is applicable even in the extremely hard case of no *a priori* known matching between neither variables nor

Table 1. The best-matched pair of a lipid and a metabolite cluster.

Lipids	**Metabolites**
GPCho(14:0/18:2)	X4.7.10.13.16.19.Docosahexaenoic.acid
GPCho(18:2/16:1)	X9.Octadecenoic.acid..Z.
GPCho(16:0/20:5)	Hexadecanoic.acid
	Phosphoric.acid

samples across the two data sets, but only a similar experiment design. The method estimates ANOVA-type multi-way covariate effects for clusters of variables, and identifies and separates covariate effects that are shared between the data sets and effects that are specific to one data set.

Acknowledgements. T.S., I.H. and S.K belong to the Finnish Centre of Excellence in Adaptive Informatics Research. 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.

References

1. Gholami, A.M., Fellenberg, K.: Cross-species common regulatory network inference without requirement for prior gene affiliation. Bioinformatics 26(8), 1082–1090 (2010)
2. Huopaniemi, I., Suvitaival, T., Nikkilä, J., Orešič, M., Kaski, S.: Multivariate multi-way analysis of multi-source data. Bioinformatics 26, i391–i398 (2010)
3. 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)
4. Huopaniemi, I., Suvitaival, T., Orešič, M., Kaski, S.: Graphical multi-way models. In: Balcazar, J., *et al.* (eds.) ECML PKDD 2010, Part I, LNAI 6321, pp. 538–553. Springer-Verlag, Berlin Heidelberg (2010)
5. Le, H.S., Bar-Joseph, Z.: Cross species expression analysis using a Dirichlet process mixture model with latent matchings. In: Lafferty, J., *et al.* (eds.) Advances in Neural Information Processing Systems 23, pp. 1270–1278 (2010)
6. Lu, Y., Huggins, P., Bar-Joseph, Z.: Cross species analysis of microarray expression data. Bioinformatics 25(12), 1476–1483 (2009)
7. 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)
8. Mardia, K.V., Bibby, J.M., Kent, J.T.: Multivariate analysis. Academic Press, London; New York (1979)
9. 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)
10. Tripathi, A., Klami, A., Orešič, M., Kaski, S.: Matching samples of multiple views. Data Mining and Knowledge Discovery (2011)