Multi-Way, Multi-View Learning

Ilkka Huopaniemi and Tommi Suvitaival
Department of Information and Computer Science
Helsinki University of Technology
Finland
ilkka.huopaniemi@tkk.fi
tommi.suvitaival@tkk.fi

Janne Nikkilä
Department of Basic Veterinary Sciences
Faculty of Veterinary Medicine
University of Helsinki
Finland
janne.nikkila@helsinki.fi

Matej Orešič
VTT Technical Research Centre of Finland
Espoo, Finland
matej.oresic@vtt.fi

Samuel Kaski
Dept of Information and Computer Science
Helsinki University of Technology
Finland
samuel.kaski@tkk.fi

Abstract

We extend multi-way, multivariate ANOVA-type analysis to cases where one covariate is the view, with features of each view coming from different, high-dimensional domains. The different views are assumed to be connected by having paired samples; this is common in our main application, biological experiments integrating data from different sources. Such experiments typically also include a controlled multi-way experimental setup where disease status, medical treatment groups, gender and time of the measurement are usual covariates. We introduce a multi-way latent variable model for this new task, by extending the generative model of Bayesian canonical correlation analysis (CCA) both to take multi-way covariate information into account as population priors, and by reducing the dimensionality by an integrated factor analysis that assumes the features to come in correlated groups.

1 Introduction

Finding disease and treatment effects from populations of biological samples is a prototypical multi-way modeling task, traditionally solved with multivariate ANOVA. The research question is, are there differences in the population that can be explained by either covariate or, more interestingly, their interaction, which would hint at the treatment being effective. It is naturally additionally interesting what the differences are.

A recurring problem in multi-way analyses, especially with modern high-throughput measurements in molecular biology, is the "small n, large p"-problem. The dimensionality p of the measurements is high while the number of samples n is low, and additionally the data may be collinear making estimation of the effects impossible with classical methods, univariate or multivariate linear models solved with multi-way ANOVA techniques. The most promising modern method, Bayesian sparse factor regression model [1], is useful in finding the variables most strongly related to the external covariate and to infer relationships between those variables via common latent factors. Instead of a regression model we will build on a generative multi-way latent factor model [2] which incorporates an assumption of clusteredness of the variables to regularize the model, and makes it possible to extend the model to multi-view factor analysis. Such clusteredness is well justified in biological applications.

Experiments integrating biological data from different sources make measurements of the same sample with different measurement techniques or from different tissues, usually resulting in paired samples from different, unmatched domains. We now assume the different views form one covariate in the multi-way analysis, with the additional problem that the samples come from different domains and cannot be directly compared. We introduce a new hierarchy level of latent variables intended to decompose the views into view-specific and shared components, which is needed for the multi-way analysis. Such a decomposition is possible given that the samples in the different views come in pairs, which we need to assume.

The resulting decomposition between the views turns out to be implementable with Bayesian canonical correlation analysis [3, 4], interpretable as unsupervised multi-view modeling. Hence, in this work we re-interpret unsupervised multi-view modeling as one-way modeling of samples from different domains, and combine it with multi-way modeling. Given that we additionally can work under the *large p, small n* conditions, the model is expected to have widespread applicability in current molecular biological measurements.

2 Model

2.1 Multi-way, multi-view

We will generalize ANOVA to multi-view (multi-domain) analysis, restricting to two covariates and two views for simplicity, although generalization is straightforward. Using ANOVA-style notation and assuming the views to be in the same domain, the multivariate linear model for samples is

$$\mathbf{v}_d = \boldsymbol{\mu}^d + \boldsymbol{\alpha}_a + \boldsymbol{\beta}_b + (\boldsymbol{\alpha}\boldsymbol{\beta})_{ab} + \boldsymbol{\gamma}_d + (\boldsymbol{\alpha}\boldsymbol{\gamma})_{ad} + (\boldsymbol{\beta}\boldsymbol{\gamma})_{bd} + (\boldsymbol{\alpha}\boldsymbol{\beta}\boldsymbol{\gamma})_{abd} + \text{noise}, \tag{1}$$

where $\boldsymbol{\mu}^d$ is the grand mean, a and b ($a = 0, \dots A$ and $b = 0, \dots B$), are the two traditional independent covariates such as disease and treatment, and d denotes the view. The $\boldsymbol{\alpha}_a$, $\boldsymbol{\beta}_b$ and $(\boldsymbol{\alpha}\boldsymbol{\beta})_{ab}$ are the shared main and interaction effects, $\boldsymbol{\gamma}_d$ would be the view-effect, $(\boldsymbol{\alpha}\boldsymbol{\gamma})_{ad}$, $(\boldsymbol{\beta}\boldsymbol{\gamma})_{bd}$ and $(\boldsymbol{\alpha}\boldsymbol{\beta}\boldsymbol{\gamma})_{abd}$ are the view-specific main and interaction effects.

For different values of d the domain of $\mathbf{v}^d$ may vary, meaning different feature spaces with different dimensionalities. We assume the samples of the different views to come in pairs, $\mathbf{v} = [\mathbf{x}, \mathbf{y}]$. For the rest of the paper we will change the notation for clarity to $\mathbf{v}^1 = \mathbf{x}$, $\mathbf{v}^2 = \mathbf{y}$, and assume a mapping f^x from the effects to the domain of $\mathbf{x}$ which is linear for now. Then,

$$\mathbf{x} = \boldsymbol{\mu}^x + f^x(\boldsymbol{\alpha}_a + \boldsymbol{\beta}_b + (\boldsymbol{\alpha}\boldsymbol{\beta})_{ab}) + f^x((\boldsymbol{\alpha})^x_a + (\boldsymbol{\beta})^x_b + (\boldsymbol{\alpha}\boldsymbol{\beta})^x_{ab}) + \text{noise}, \tag{2}$$

assuming $\boldsymbol{\gamma}_d = 0$, because it is impossible to compare means of different domains, and that the view-specific effects are in the same domain as the view-independent effects and hence need to be transformed with the same function. The equation for $\mathbf{y}$ is analogous. To our knowledge, there exists no method capable of studying the shared and view-specific multi-way effects.

2.2 Hierarchical model

We next formulate a hierarchical latent-variable model for the task of multi-way, multi-view learning under "large p, small n" conditions. For this we need three components: (i) regularized dimension reduction, (ii) combination of different data domains, and (iii) multi-way analysis. We formulate each of these as part of an overall generative model, which is solved by Gibbs sampling. In effect the model, shown in Figure 1, consists of two factor analyzers, where the loadings assume cluster memberships (multiplied with scales), a generative model of CCA, and population-specific priors on $\mathbf{z}$ that assume ANOVA-type multi-way structure. We will now introduce the details of each of these parts in turn.

2.2.1 Factor analysis model

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 views into their respective latent variables $\mathbf{x}^{lat}$ and $\mathbf{y}^{lat}$. This can be done by a factor analysis (FA) model

$$\begin{aligned} \mathbf{x}^{lat}_j &\sim \mathcal{N}(0, \boldsymbol{\Psi}^x) \\ \mathbf{x}_j &\sim \mathcal{N}(\boldsymbol{\mu}^x + \mathbf{V}^x \mathbf{x}^{lat}_j, \boldsymbol{\Lambda}^x)\ . \end{aligned} \tag{3}$$

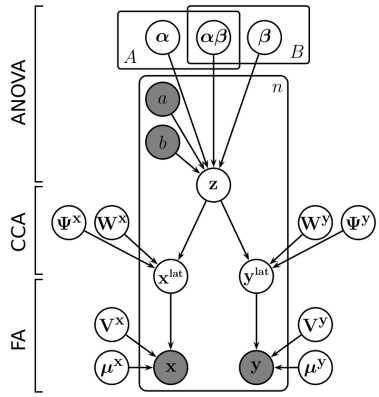

Figure 1: The hierarchical latent-variable model for multi-way, multi-view learning under "large p, small n" conditions.

Here $\boldsymbol{V}^x$ is the projection matrix that is assumed to generate the data vector $\mathbf{x}_j$ from the latent variable $\mathbf{x}_j^{lat}$, $\boldsymbol{\Lambda}^x$ is diagonal noise with elements σ_i. The $\mathbf{x}_j^{lat}$ is a latent variable vector, 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. The covariance matrix of $\mathbf{x}^{lat}$, $\boldsymbol{\Psi}^x$, comes from the CCA. At this point, when $n < p$, $\mathbf{V}^x$ cannot be estimated due to the singularity of the sample covariance matrix.

2.2.2 Regularized projection matrix that assumes grouped variables

We make the structured assumption that there are strongly correlated groups of variables in the data [2]. We regularize the $\mathbf{V}^x$ projection matrix to a clustering matrix such that each variable comes from exactly one factor. The projection matrix is positive-valued, each row having one non-zero element corresponding to the cluster assignment of the variable. We therefore assume that the main correlations are positive correlations between variables belonging to the same cluster.

2.2.3 Generative model of CCA

We now need to search a view shared by the two different domains $\mathbf{x}^{lat}$ and $\mathbf{y}^{lat}$, needed for finding shared multi-way effects. Given paired data, this is a task for Bayesian CCA (BCCA) [3, 4] which 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 as

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

where $\mathbf{z}_j$ is the shared latent variable, $\mathbf{W}^x$ is the projection matrix and $\boldsymbol{\Psi}^x$ is the marginal covariance matrix, and likewise for $\mathbf{y}$.

2.2.4 ANOVA-type model for latent variables.

We assume that the ANOVA-type effects act on the latent variables $\mathbf{z}$, which allows access to shared effects found in both the spaces $\mathbf{x}^{lat}$ and $\mathbf{y}^{lat}$. They are modeled as population priors to the latent variables, which in turn are given Gaussian priors $\boldsymbol{\alpha}_a$, $\boldsymbol{\beta}_b$, $(\boldsymbol{\alpha\beta})_{ab} \sim \mathcal{N}(0, \mathbf{I})$. In normal BCCA the prior is zero-mean. In the K_z-dimensional shared latent variable space we then have

$$\mathbf{z}_j = \boldsymbol{\alpha}_a + \boldsymbol{\beta}_b + (\boldsymbol{\alpha\beta})_{ab} + \text{noise}. \tag{5}$$

In addition, Bayesian CCA assumes that the data is generated by a sum of shared latent variables $\mathbf{z}$ and view-specific latent variables $\mathbf{z}^x$ and $\mathbf{z}^y$. The model then decomposes the ANOVA-type effects to the shared effects and to view-specific effects $\boldsymbol{\alpha}_a^x$, $\boldsymbol{\beta}_b^x$, and $(\boldsymbol{\alpha\beta})_{ab}^x$, and likewise for y.

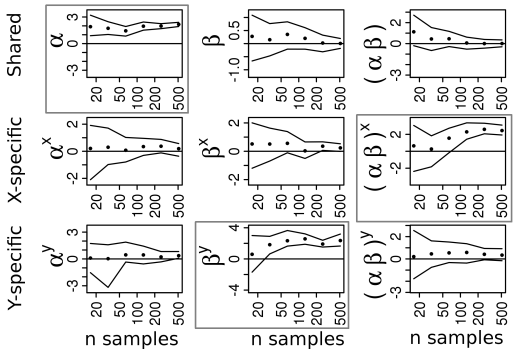

Figure 2: The method finds the generated effects $\alpha = +2$,$\beta^y = +2$, and $(\alpha\beta)^x = +2$ (shown by the boxes). The dots show posterior mean and the thin lines include 95% of posterior mass, as a function of number of observations. A consistently non-zero posterior distribution implies an effect found.

3 Results

We demonstrate the functionality of the method on generated data with a two-way, two-view experimental setup. The generated data has known effects α, β^y, and $(\alpha\beta)^x$, each with strength $+2$. We then study how well the model finds the effects as a function of the number of measurements. Both $\mathbf{x}$ and $\mathbf{y}$ are 200-dimensional and the noise $\sigma_i = 1$. The method finds the three generated effects, shown in Figure 2. 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 views. With small numbers of samples, there is considerable uncertainty in the effects for view-specific components. In typical bioinformatics applications there may be 20-50 samples.

4 Discussion

We have generalized ANOVA-type multi-way analysis to cases where multiple views of samples having a multi-way experimental setup are available. 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 having population-specific priors on the shared latent variable of CCA. Furthermore, the method is able to decompose the covariate effects to shared and view-specific effects, treating the multiple views as one covariate. Finally, 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 correlated groups, which is reasonable for bioinformatics applications.

The modelling 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 with small sample-sizes in the generated multi-view, multi-way dataset. We have also applied the method to metabolomics data, which demonstrated its successful applicability to a real data, and we are currently writing a manuscript about that.

References

[1] Mike West. Bayesian factor regression models in the large p, small n paradigm. *Bayesian Statistics*, 7:723–732, 2003.

[2] Ilkka Huopaniemi, Tommi Suvitaival, Janne Nikkilä, and and Samuel Kaski Matej Orešič. Two-way analysis of high-dimensional collinear data. *Data Mining and Knowledge Discovery*, 19(2):261–276, 2009.

[3] Arto Klami and Samuel Kaski. Local dependent components. In Zoubin Ghahramani, editor, *Proceedings of ICML 2007, the 24th International Conference on Machine Learning*, pages 425–432. Omnipress, 2007.

[4] Cédric Archambeau and Francis Bach. Sparse probabilistic projections. In Daphne Koller, Dale Schuurmans, Yoshua Bengio, and Lon Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 73–80. MIT Press, 2009.