# Probabilistic Factor Analysis - Assignment 1B (Dependency modeling)

Suleiman Ali Khan

May 16, 2017

## Assignment 1B

Read the text below carefully and answer the following questions:

1. A large online retail company is recording the search items a user clicks and the items he purchases. The company wants to know for which items, the clicks are *not* associated with the purchase. Which type of model will you use, how and why?

2. What is the relation between Factor Analysis and Group Factor Analysis?

Submission:

- Submit your answers on paper at 9:15 am in the class on 18th May 2017. Write your student number on top.
- No late submissions.
- If you are not coming to the lecture send it via email to suleiman.khan@helsinki.fi

# 1 Canonical Correlation Analysis (CCA)

**Classical and regularised CCA** [1] is an unsupervised method that decomposes two paired matrices into a shared low-dimensional representation. The paired matrices are characterized by having a common identity of the samples. Therefore, unlike FA that finds dependencies between two or more variables, CCA aims to capture the correlated patterns between two matrices. CCA linearly transforms the matrices into a maximally correlated subspace of components, such that any two components are uncorrelated with each other. This way, it can find distinct components that are common to both of the matrices.

Methodologically, CCA is an established data integration approach that maximally explains the dependency between two data sets [1]. The method linearly projects the data sets to obtain a maximally correlated low-dimensional representation. This low-dimensional representation aka the *shared space* or the *components* capture the statistically shared patterns between the two data sets, whereas patterns specific to any one of the data set are considered noise and ignored.

In data sets that have more dimensions than the number of samples with several highly correlated variables, there is a potential for the CCA covariance matrices to become ill-conditioned, which could result in numerical inaccuracies while computing the inverse. This is a classical problem when $D > N$, and regularized solutions are available [2–4].

Given two data sets $\mathbf{X} \in \mathcal{R}^{N \times D_1}$ and $\mathbf{Y} \in \mathcal{R}^{N \times D_2}$, with $N$ paired occurrences of samples in the two views, regularized CCA finds $K$ linear projections of the data sets, $\mathbf{X}\mathbf{w}_k$ and $\mathbf{Y}\mathbf{v}_k$, such that their correlation $P_k$ is maximized as,

$$\begin{aligned} P_k &= \underset{\mathbf{w}_k, \mathbf{v}_k}{\arg\max} \operatorname{cor}(\mathbf{X}\mathbf{w}_k, \mathbf{Y}\mathbf{v}_k) \\ &= \underset{\mathbf{w}_k, \mathbf{v}_k}{\arg\max} \frac{\mathbf{w}_k^T \mathbf{C}_{\mathbf{xy}} \mathbf{v}_k}{\sqrt{\mathbf{w}_k^T \mathbf{C}_{\mathbf{xx}} \mathbf{w}_k + L_1 \|w_k\|^2} . \sqrt{\mathbf{v}_k^T \mathbf{C}_{\mathbf{yy}} \mathbf{v}_k + L_2 \|v_k\|^2}} . \end{aligned} \tag{1}$$

The vectors $\mathbf{w}_k$ and $\mathbf{v}_k$ are the *projection weights*, or *loadings* when normalized, while the projected space of data sets $\mathbf{X}\mathbf{w}_k$ and $\mathbf{Y}\mathbf{v}_k$ constitutes the *CCA components* or *canonical covariates*, capturing the shared patterns between the two data sets. The first component $k = 1$ is found such that the correlation $P_k$ is the largest possible. All other components $k = 2, 3...$ are computed analogously, but with the additional constraint that they are uncorrelated with the previously obtained components. These constraints ensure that the first $K$ components capture the strongest shared effects, which are also distinct from each other. The regularization replaces the empirical covariances $\mathbf{C}_{\mathbf{xx}}$ and $\mathbf{C}_{\mathbf{yy}}$ by their regularized estimates $\mathbf{C}_{\mathbf{xx}} + L_1\mathbf{I}$ and $\mathbf{C}_{\mathbf{yy}} + L_2\mathbf{I}$, and also acts as a penalizer on the projection weights $L_1\|w_k\|^2$, $L_2\|v_k\|^2$, preferring a simpler solution.

**Probabilistic Canonical Correlation Analysis** For two data vectors $\mathbf{x}_n^{(1)}$ and $\mathbf{x}_n^{(2)}$, CCA can be represented as a generative process [5, 6]:

$$\begin{aligned} \mathbf{x}_n^{(m)} &\sim N(\mathbf{W}^{(m)}\mathbf{z}_n, \Psi^{(m)}) \qquad m = 1, 2 \\ \mathbf{z}_n &\sim N(0, \mathbf{I}) \ , \end{aligned} \tag{2}$$

where $\mathbf{z}_n$ is the latent vector common to both matrices, $\mathbf{W}^{(m)}$ are loadings for each matrix, and $\Psi^{(m)}$ the corresponding noise covariance matrix. The shared latent representation $\mathbf{z}_n$ of CCA models the covariation patterns between the two matrices, while $\Psi^{(m)}$ models the variation specific to each matrix. This division implies that $\mathbf{z}_n$ can capture the dependencies between the two matrices. An efficient CCA solution using group-sparse priors was recently presented by [6]. Their formulation uses the latent variables to represent both the correlated patterns between the matrices as well as the matrix specific variation, while the noise covariance is assumed isotropic for each of the data sets. Their formulation efficiently scales to large data sets.

CCA (eqn 2) can also be seen as two parallel FA on two paired and whitened matrices, such that $\mathbf{z}_n$ is made common between the two. For a comprehensive review of Bayesian canonical correlation analysis see [6], while [7] for classical CCA. CCA has been successfully used for modeling dependencies between data sets. For example, in genomics it has been used to identify chromosomal regions showing dependencies in copy number and gene expression of a set of samples [8, 9].

# 2 Group Factor Analysis (GFA)

The dependency modeling task now becomes the identification of common patterns from multiple paired data sets. Existing multi-set extensions of CCA formulate the task for more than two data sets [10, 11]; however, they only model components common to all views, and hence are unable to identify patterns shared by only a subset of the views.

[12, 13] presents a novel latent component model *Group Factor Analysis* (GFA) that not only extracts the statistical dependencies between all the data sets but also identifies dependencies between any subset of them.

Group Factor Analysis (GFA) is a model designed to capture relationships (statistical dependencies) by reducing the collection of data sets (views) into a combined set of low-dimensional factors (components). A component can be active in one or more of the data views, representing that it captures the underlying relationships between the corresponding views only. For example, a component active in all the views captures the shared dependency structure between all the views while one active in a single view identifies variation and features specific to that particular view only. GFA learns the activity of the components in a data-driven manner making it possible to identify dependencies that exist between a subset of the views.

Formally, given a collection of M data sets $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, ...\mathbf{X}^{(M)} \in \mathcal{R}^{N \times D_m}$, having $N$ paired samples, and a separate set of dimensions $D_m$ in each view, GFA searches for a $K$-dimensional matrix factorization for the entire collection. The model is formulated in a Bayesian setting as a product of a Gaussian latent component matrix $\mathbf{Z} \in \mathcal{R}^{N \times K}$ containing the $K$ components, and a projection weight matrix $\mathbf{W}^{(m)} \in \mathcal{R}^{D_m \times K}$ for each data set $m$. The model is represented as

$$\begin{aligned}
\mathbf{x}_n^{(m)} &\sim \mathcal{N}\left(\mathbf{W}^{(m)}\mathbf{z}_n, \mathbf{I}(\tau^{(m)})^{-1}\right) \\
\mathbf{z}_n &\sim \mathcal{N}(0, I) \\
\mathbf{w}_{:,k}^{(m)} &\sim \mathcal{N}\left(0, (\alpha_k^{(m)})^{-1}\right) \\
\alpha_k^{(m)} &\sim Gamma(a^\alpha, b^\alpha) \\
\tau^{(m)} &\sim Gamma(a^\tau, b^\tau)\ ,
\end{aligned} \tag{3}$$

where $\tau^{(m)}$ denotes the view-specific noise precision. The latent variables $\mathbf{z}_n$ are common between all the views, representing the statistical patterns. GFA solves the joint decomposition problem using the *group-wise sparse* matrix factorization of all data sets, where each data set is considered a group. The group-wise projections $\mathbf{W}^{(m)}$ capture both group-specific variation (activity seen exclusively in one view), and dependencies between the groups (activity in more than one views). This is achieved by modeling the total variation in the data while constrained by the group-sparse prior. The group-sparse prior (via $\alpha_k^{(m)}$) controls the scale of the projection weights $\mathbf{w}_{:,k}^{(m)}$ for each of the component-view pairs. Higher values of $\alpha_k^{(m)}$ shrink the corresponding $\mathbf{w}_{:,k}^{(m)}$ towards zero switching the component off, while smaller values of $\alpha_k^{(m)}$ increase the scale of $\mathbf{w}_{:,k}^{(m)}$, making the component active in the view. GFA learns the $\alpha_k^{(m)}$ in a data-driven way, yielding dependency patterns between the views. As a practical step, $\alpha_k^{(m)}$ is thresholded with respect to the captured variance to obtain active and non-active status of the components.

# References

[1] H. Hotelling, “Relations between two sets of variates,” *Biometrika*, vol. 28, pp. 321–377, 1936.

[2] H. Vinod, “Canonical ridge and econometrics of joint production,” *Journal of Econometrics*, vol. 4, no. 2, pp. 147 – 166, 1976.

[3] S. E. Leurgans, R. A. Moyeed, and B. W. Silverman, “Canonical correlation analysis when the data are curves,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 55, pp. 725–740, 1993.

[4] I. González, S. Déjean, P. G. Martin, A. Baccini, *et al.*, “Cca: An R package to extend Canonical correlation analysis,” *Journal of Statistical Software*, vol. 23, pp. 1–14, 2008.

[5] F. R. Bach and M. I. Jordan, “A probabilistic interpretation of canonical correlation analysis,” Tech. Rep. 688, Department of Statistics, University of California, Berkeley, 2005.

[6] A. Klami, S. Virtanen, and S. Kaski, “Bayesian canonical correlation analysis,” *Journal of Machine Learning Research*, vol. 14, pp. 965–1003, 2013.

[7] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, “Canonical correlation analysis: An overview with application to learning methods,” *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[8] D. M. Witten and R. J. Tibshirani, “Extensions of sparse canonical correlation analysis with applications to genomic data,” *Statistical applications in genetics and molecular biology*, vol. 8, no. 1, pp. 1–27.

[9] L. Lahti, S. Myllykangas, S. Knuutila, and S. Kaski, "Dependency detection with similarity constraints," in *Machine Learning for Signal Processing, 2009. MLSP 2009. IEEE International Workshop on*, pp. 1–6, IEEE, 2009.

[10] C. Archambeau and F. R. Bach, "Sparse probabilistic projections," in *Advances in neural information processing systems*, pp. 73–80, 2009.

[11] F. Deleus and M. M. Van Hulle, "Functional connectivity analysis of fmri data based on regularized multiset canonical correlation analysis," *Journal of Neuroscience methods*, vol. 197, no. 1, pp. 143–157, 2011.

[12] S. Virtanen, A. Klami, S. A. Khan, and S. Kaski, "Bayesian group factor analysis," in *Proceedings of AISTATS, JMLR W&CP 22*, pp. 1269–1277, 2012.

[13] S. A. Khan, S. Virtanen, O. P. Kallioniemi, K. Wennerberg, A. Poso, and S. Kaski, "Identification of structural features in chemicals associated with cancer drug response: a systematic data-driven analysis," *Bioinformatics*, vol. 30, no. 17, pp. i497–i504, 2014.